UNIVERSITÀ DEGLI STUDI DI MILANO

Scuola di Dottorato in Fisica, Astrofisica e Fisica Applicata

Dipartimento di Fisica

Corso di Dottorato in Fisica, Astrofisica e Fisica Applicata

Ciclo XXX

Computational modeling of proteins: from statistical mechanics to immunology

Settore Scientifico Disciplinare FIS/03

Supervisori:

Prof. Guido TIANA

Dott. Giorgio COLOMBO

Coordinatore:

Prof. Francesco RAGUSA

Tesi di Dottorato di:

Riccardo Capelli

Anno Accademico 2016-2017

Commission of the final examination:

External Referee:
Prof. Martin Zacharias

External Member:
Prof. Giovanni Bussi

External Member:
Prof. Laura Bonati

Final examination:

Date: *November 24th, 2017*

Università degli Studi di Milano, Dipartimento di Fisica, Milano, Italy

Cover illustration:

"Controluce" – A. Capelli

MIUR subjects:

FIS/03 - Fisica della Materia

PACS:

87.10.Tf Molecular dynamics simulation
87.15.km Protein-protein interactions

"And the feeling is that there's something wrong,
because I can't find the words and I can't find the songs."

RADIOHEAD, Stop Whispering

Contents

List of Figures

List of Tables

Motivation

"Muovo le molecole immobili."

AFTERHOURS

One of the biggest revolutions occurred during the second half of the 20th century in physics was the introduction of computers in research. In particular, the use of fast computing machines opened the possibility to study complex systems by simulating their dynamics, without the need to pursue analytical solutions [1], otherwise impossible to tackle. The consequences of this breakthrough were huge both in the study of equilibrium and non-equilibrium many-body problems, with the strong limitation given by the number of atoms involved in the calculation.
During the same period, molecular biology moved its first steps. During the 1950s the α-helix [2] and the β-hairpin [3] were identified as the principal motifs in protein structures. The discovery of these small (in the order of the hundreds of atoms) structures opened the possibility to apply this computational many-body approach to molecular biology.
The first technique used in biology-related problems was the Monte Carlo Method [4, 5], and some years later Molecular Dynamics (MD) [6] was formalized. In MD, for each atom of the system one can solve its Newton equations of motion, obtaining a trajectory in the phase space for the entire system, and study its behavior in equilibrium and non-equilibrium conditions. The constant rise in computational power gave the possibility to scientists to study larger and larger systems, while the advances in experimental techniques enhanced the possibility for direct comparisons between wet and *in silico* data at similar levels of resolution. Despite the validity of Moore's Law (*i.e.*, the exponential growth of the computing power due to transistors miniaturization) until now [7], the timescale of the events that can be simulated has an upper limit of the millisecond with tailor-made computers [8], which is not enough to study all the biologically-relevant phenomena. Since the birth of computational chemistry, a huge number of different statistical mechanics-based methods has been implemented to permit, given the computing power limit, an effective reliable use of MD simulations in biochemistry.

One of the most relevant problems tackled by MD is the calculation of free energy differences, both in conformational changes and in sequence mutations of a protein. The main reason of this difficulty is represented by the frustrated nature of interactions in proteins and the size of these systems: this leads to a complex energy landscape which in principle needs very long sampling times to overcome all possible energy barriers. A lot of different methods have been developed to overcome this problem, but most of them need the *a priori* knowledge of a meaningful collective variable for the system (like Umbrella Sampling [9] or Metadynamics[10]), or, like in the case of free energy calculation techiques (Free Energy Perturbation [11]) the calculation strategy and its convergence strongly depend on the choice of pathway in the potential space, which is also in this case a system-dependent problem.
Moved by this reason, we studied and improved a path-independent and system-independent free energy calculation techique, called Simplified Confinement Method [12, 13]. We describe this work in Chapter 1.

Although MD has been successful in most of its applications, there are still many open problems: as mentioned before, the available parametrizations of interaction potentials (called force fields) are not completely reliable [14]. In particular, the choice of force field parameters is performed comparing experimental data on a fixed set of (usually small) molecules with computed data on the same molecules. This raises a significant problem: large molecules can have a more complex behavior, and using these potentials can lead to a systematic error; furthermore, the timescale in which the force field is tested needs to be limited. Another strong limitation of MD depends on the equilibrium experiments used for parametrization: the kinetic properties of a system are not considered [15].
Given the impossibility to reparametrize a general force field with non-equilibrium experimental data, we implemented a technique that uses equilibrium-based force fields, adding a potential term based on time series resulting from kinetic experiments. This approach, based on the principle of Maximum Caliber, restrains the system with an experimental-based bias, returning a more realistic behavior of the simulation in condition where the usual force fields show their limitations. We describe this work in Chapter 2.

The application of computational methods in the study of proteins confirms its efficacy in other fields of life sciences: an actual and emerging topic is represented by vaccinology. With techniques developed by Louis Pasteur at the end of the 19^{th} century (isolation of the pathogen, its inactivation and subsequent inoculation in the host), various scientists developed vaccines for deadly diseases like poliomyelitis, diphterite and measles. None of the mentioned was developed with molecular biology-based approaches.
Almost 50 years after the birth of molecular biology, the Human Genome Project decoded human DNA [16] and, at the same time, the genome of the most dangerous pathogen was screened. This has laid the foundation of Reverse Vaccinology (RV)[17], where the proteins responsible for immune reaction can be identified from the pathogen DNA and tested directly on animal models, obtaining a new vaccine candidate with little or no risk for the host, having removed the pathogen itself. At the beginning of the 21^{st}

century the first vaccine against *Meningococcus B*, responsible for the 50% of the meningococcal meningitis, was developed using this protocol [18]. Since then, crystallographic data was inserted in RV workflow to exploit conformational data, creating the so-called Structural Vaccinology (SV) [19]. To enhance its efficacy, SV exploits all the aspects of molecular modeling like computer-aided drug/protein design and MD to integrate information that come from experimental sources. One of the most promising technique in this field is the grafting of an immunogenic sequence (*i.e.*, a portion of a protein recognized by the immune system) on a foreign protein; this approach could lead to a new vaccine component which have no risk for the patient. To date, the grafting technique has been carried out by human-driven workflows.
Motivated by this reason, we studied immunogenic peptides from a family of pathogens involved in respiratory diseases, exploiting Structural Vaccinology principles with both computational and experimental approach. Furthermore, we developed and implemented an unsupervisionated automated tool to design grafted protein sequences. We describe this work in Chapter 3.

Organizational note

The present Thesis consists of 3 Chapters.

Chapter 1. Calculation of free energy differences in biomolecules: we describe our original work based on the Confined Method (CM) proposed by Tyka *et al.* and refined as Simplified Confinement Method (SCM) by Ovchinnikov *et al.*. The work on single-point mutants has been completed in collaboration with F. Villemot, A. van der Vaart, G. Tiana, E. Moroni, and G. Colombo and published in *The Journal of Physical Chemistry Letters* (ref. 1 in Refereed Publications); the work on conformational free energy differences using MBAR interpolation has been completed in collaboration with F. Villemot, A. van der Vaart and G. Colombo and published in the *Journal of Chemical Theory and Computation* (ref. 2 in Refereed Publications).

Chapter 2. Non-equilibrium sampling: we describe the application of principle of Maximum Caliber to non-equilibrium sampling of biomolecules in MD. This work has been completed in collaboration with G. Tiana and C. Camilloni and a paper is currently in preparation (ref 1 of Publications in preparation).

Chapter 3. Peptide and protein design for immunology: we describe our original study on peptides used as a probe for immunodiagnostic purposes. This work has been completed in collaboration with E. Matterazzo, M. Amabili, C. Peri, A. Gori, P. Gagni, M. Chiari, G. Lertmemongkolchai, M. Cretich, M. Bolognesi, G. Colombo and L. J. Gourlay and published on *ACS Infectious Diseases* (ref. 4 in Refereed Publications). The work on automated grafting of epitopes on foreign scaffolds has been completed in collaboration with F. Marchetti, G. Tiana and G. Colombo and published in the *Journal of Chemical Information and Modeling* (ref. 3 in Refereed Publications). The first ongoing application of the SAGE workflow is also discussed; this part involved M. Amabili, L. J. Gourlay, M. Bolognesi and G. Colombo and a paper is currently in preparation (ref 2 of Publications in preparation).

CHAPTER 1

Calculation of free energy differences in biomolecules

"L'improbabilità di un'ipotesi è esponenzialmente proporzionale alle menzogne che invento per farla verificare."

UOCHI TOKI, il Non-illuminato

Proteins are linear polymers with an extremely small number of equilibrium conformations at biological conditions, which depends on their sequence [20]. To estimate the stability of a protein the Gibbs free energy difference between two conformations (ΔG) is used, which is related with the probability to find the system in a given configuration. In the case of sequence mutation, the variation of the Gibbs free energy difference ($\Delta\Delta G$) tells us if the change in protein sequence is stabilizing or destabilizing the structure the system. In biochemistry, the computation of conformational ΔG or mutational $\Delta\Delta G$ of a protein is then crucial to understand the stability of a biomolecule, and, from an applicative point of view, it is fundamental in protein engineering and drug design.
Experimentally, it is possible to obtain the variation of Gibbs free energy difference using fluorescence [21], calorimetry [22] or kinetics experiments [23].
From the computational point of view, the calculation of free energy difference is far from trivial. During my PhD, I have studied and improved a novel technique, called Simplified Confinement Method, and I have applied it to the computation of free energy difference in peptides and small proteins.

1.1 Introduction

Molecular Dynamics (MD) is the simulation technique devoted to solve numerically the Newton equation of motion for a system of N particles, where their potential is described by an approximated (derived from experimental data or with *ab-initio* calculations) force field. In most of the applications of computational biophysics, one is interested in studying by means of MD the equilibrium properties of a system. This leads to a clear problem: MD is not the most efficient way to target the equilibrium properties of a biomolecule in its solvent, because it returns a trajectory which is a single realization of all the possible paths in the phase space of that system.

Following the Ergodic Theorem, if one can perform a *very long* MD simulation, it is possible to achieve equilibrium and compute directly all the thermodynamic properties for such a system. This could lead to a comparison with experimental data which are, by construction, an average over $\sim 10^{23}$ molecules.
One of the most fundamental (and thus more interesting) system observables is the free energy difference between two states of the system or between two independent systems. This quantity cannot be obtained directly from averaging, but can be obtain inverting the equilibrium distribution of the system. As in the case of equilibrium observables, the accuracy of free energy computation strongly depends on the number of conformation visited, making its (brute-force) computation a burdensome problem. Gibbs free energy difference ΔG can be related to fundamental properties of biochemical interest like solubility [24, 25], binding affinity [26] and biological activity [27].
For this reason, free energy calculation has been a central topic for a wide range of different sciences with different approaches, from biology to statistical physics [28].

In the canonical ensemble, the Helmholtz free energy is defined as

$$\begin{aligned} F(N,V,T) &\equiv -k_B T \log Z(N,V,T) \\ &= -k_B T \log \iint \exp\left(-\frac{H(p,q)}{k_B T}\right) d^N p \, d^N q \end{aligned}$$

where $Z(N,V,T)$ is the partition function of the system in the canonical ensemble, and $H(p,q)$ is the hamiltonian that describes the dynamics of our system.
In molecular dynamics classical potentials, The hamiltonian is defined as $H(p,q) = U(q) + K(p)$ and thus the potential is completely independent from the momenta, and the integral over p is then a constant (at fixed T).
Neglecting the "non-informative" kinetic part, the Helmholtz free energy results

$$F(N,V,T) = -k_B T \log \int \exp\left(-\frac{U(q)}{k_B T}\right) d^N q. \tag{1.1}$$

All the derivations in this chapter for Helmholtz free energy which is based on canonical partition function (*i.e.*, number of particle N, volume V and temperature T are constant) are equivalent in Gibbs free energy $G(N,P,T)$ apart for the change of ensemble, from canonical to NPT (*i.e.*, number of particle N, pressure P and temperature T are constant). Furthermore, the Gibbs free energy is the experimental measurable state function.

Given the definition of free energy and considering an incredibly long sampling of the phase space for a system, the most trivial technique to obtain free energy differences between two states A and B, defined setting a threshold on a meaningful order parameter (*i.e.* a reaction coordinate) of the system, is its naïve application: we can count the number N_{A} of configurations in a state A and the number of configurations N_{B} in a state B during the sampling

$$\Delta F_{A\to B} = -k_B T \ln\left(\frac{N_{\mathrm{B}}}{N_{\mathrm{A}}}\right). \tag{1.2}$$

As previously said, a sampling which gives a reasonable statistics to use the equation 1.2 is not reachable in a human being lifetime. In particular, one has to consider the free energy barriers that have to be crossed by the system and the enormous dimensionality of phase space, making overlap between the two states A and B completely unrealistic in real-world applications.

During the last 50 years a plethora of different techniques based on statistical mechanics assumptions has been developed to compute free energies given a limited sampling. There are two major schemes to tackle this problem:

Enhanced Sampling: The first approach consists in studying the free energy variation along a reaction coordinate $\xi(q)$ of interest; in this way, the dimensionality of the problem is lowered to a small number of collective variables, instead of $3N$ independent atomic coordinates. It is possible to add a bias term to $U(q)$ which depends on $\xi(q)$ and then subdivide the phase space in different parts which can reach equilibrium in a reasonable time (like in Umbrella Sampling technique [9]). Another way is to introduce in the potential a memory term for the reaction coordinate $\xi(q,t)$ (like in Local Elevation [29], Metadynamics [10] and in its well-tempered variant [30]) in order to "lower" the free energy barriers and perform a more efficient sampling.

Free Energy Calculation: The second approach consists in computing directly the difference between a state A and a state B starting from a definition of a derivative of free energy with respect to some parametrization. Having two different states of interest A and B, we switch from a potential U_{A} that describes the state A as a global minimum to a potential U_{B} that describes the state B following the variation of a parameter λ. Given the freedom in choosing the two potentials, this permits also the so-called computational alchemy [31]. To this class of techniques belong Thermodynamic Intergration [32, 33] and Free Energy Perturbation [11].

During my PhD, I worked towards enhancing the performance of a Free Energy Calculation technique based on Thermodynamic Integration.

1.2 Thermodynamic Integration

Given a classical system formed by N particles subjected to a potential $U(q)$, the goal of Thermodynamic Integration (TI) is the calculation of free energy difference between two different states (or even different systems) A and B. We introduce a new $U(q)$

$$U(q,\lambda) = \lambda U_{\mathrm{A}}(q) + (1-\lambda)U_{\mathrm{B}}(q) \tag{1.3}$$

where λ is a decoupling parameter in the potential such that, for $\lambda = 0$, $U(q) = U_{\mathrm{A}}(q)$, which is the potential function of the state A, while for $\lambda = 1$, $U(q) = U_{\mathrm{B}}(q)$, which is the potential function of the state B.

Inserting this potential in the expression of Helmholtz free energy (1.1) and differentiat-

ing with respect to λ

$$\begin{aligned}\frac{\partial F(N,V,T,\lambda)}{\partial \lambda} &= -k_B T \frac{\partial}{\partial \lambda} \log Z(N,V,T,\lambda) \\ &= -k_B T \frac{\partial Z(N,V,T,\lambda)}{\partial \lambda} \cdot \frac{1}{Z(N,V,T,\lambda)} \\ &= \frac{\int \frac{\partial U(q,\lambda)}{\partial \lambda} \exp\left(-k_B T\, U(q,\lambda)\right)\, d^N q}{\int \exp\left(-k_B T\, U(q,\lambda)\right)\, d^N q} \\ &= \left\langle \frac{\partial U(q,\lambda)}{\partial \lambda} \right\rangle_\lambda \end{aligned} \tag{1.4}$$

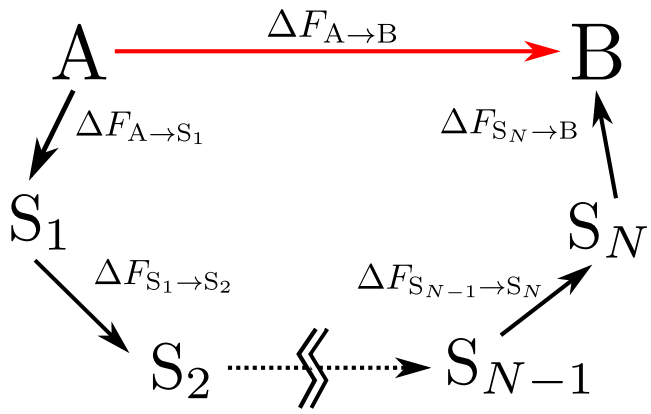
Integrating equation (1.4), the free energy difference between the state A and the state B results

$$\Delta F_{\mathrm{A}\to\mathrm{B}} = \int_0^1 \left\langle \frac{\partial U(q,\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda \tag{1.5}$$

This integral can be carried out numerically, simulating the system of interest with different values of λ (called "windows"), computing directly the value of the $\left\langle \frac{\partial U}{\partial \lambda} \right\rangle_\lambda$ after the equilibration of the system with the new potential.

The major issue of the Thermodynamic Integration technique (apart from statistical error in the evaluation of the average, which can be resolved with longer window simulations) is the convergence of the integral in equation (1.5) [34]. In particular, the lack of phase-space overlap for subsequent values of λ (approximately, we can set a maximum ΔF between two consequent windows of about 2 kcal/mol) can results in large errors, and thus in an unphysical value for the free energy difference between the two states. This problem can be resolved in two ways: the naïve one is the simulation of more λ windows; the wiser way is to choose a different path in the phase space which makes TI avoid the crossing of high free energy barriers, to limit the presence of huge free energy difference between windows. In particular, free energy at equilibrium is a state function, thus $\Delta F_{\mathrm{A}\to\mathrm{B}}$ is independent from the followed path.
Operatively, it is possible to break the transition between the two states inserting some intermediate (even unphysical) states, that make the calculation convergence easier and building a thermodynamic cycle (see Figure 1.1).

The use of a thermodynamic cycle is computationally necessary in some fundamental applications of Thermodynamic Integration, such as the computation of solvation [35] or binding free energies [36].
Given the possibility to split the pathway in different TI steps, it is however far from trivial to build a path that converges in a reasonable time to the correct value of free energy difference, and the wise choice of the intermediate states is the key point to get a convergent simulation. During the years some solutions, like soft-core potentials [37, 38, 39], have been developed to overcome these problems. Another approach is the creation of intermediate states which are independent from the force field free energy landscape, and thus not affected by the overlap problem, like in the Confinement Techniques.



$$\Delta F_{\mathrm{A}\to\mathrm{B}} = \Delta F_{\mathrm{A}\to\mathrm{S}_1} + \Delta F_{\mathrm{S}_1\to\mathrm{S}_2} + \cdots + \Delta F_{\mathrm{S}_{N-1}\to\mathrm{S}_N} + \Delta F_{\mathrm{S}_N\to\mathrm{B}}$$

Figure 1.1: Example of a thermodynamic cycle. The free energy difference between states A and B can be computed using an alternative pathway via N intermediate states $\mathrm{S}_1 \ldots \mathrm{S}_N$.

1.3 Simplified Confinement Method

One of the techniques that uses thermodynamical cycles with TI is the Confinement Method (CM), proposed by Tyka *et al.* [12] and subsequently improved by Cecchini *et al.* [40] and Ovchinnikov *et al.* [13] into the Simplified Confinement Method (SCM).
To avoid the overlap problem explained in section 1.2, each state of interest has to be "confined" (harmonically restrained) to a reference state which is equivalent to a system of $3N$ non-interacting harmonic oscillators. Using TI, it is possible to compute ΔF from the "real" state and the confined one, while the free energy difference between the two reference states can be obtained analytically. The free energy between state A and state B results then

$$\Delta F_{\mathrm{A}\to\mathrm{B}} = \Delta F_{\mathrm{A}\to\mathrm{A}^{\mathrm{HO}}} + \Delta F_{\mathrm{A}^{\mathrm{HO}}\to\mathrm{B}^{\mathrm{HO}}} + \Delta F_{\mathrm{B}^{\mathrm{HO}}\to\mathrm{B}}$$

as shown in the thermodynamic cycle in Figure 1.2.
For the TI transition in the Confinement Method, we have a slightly different expression of the potential $U(q,\lambda)$ with respect to the one in (1.3)

$$\begin{aligned} U(q,\lambda) &= U_{\mathrm{ff}}(q) + \lambda U_{\mathrm{HO}}(q) \\ &= U_{\mathrm{ff}}(q) + \frac{1}{2}\lambda k_f \left|q - q_0\right|^2, \end{aligned} \tag{1.6}$$

where $U_{\mathrm{ff}}(q)$ is the original force field, $U_{\mathrm{HO}}(q)$ is the confinement potential, k_f is the maximum harmonic constant used to constrained the system, and it should be big enough so that $U(q,\lambda) \simeq \frac{1}{2}\lambda k_f \left|q - q_0\right|^2$ for $\lambda = 1$ (*i.e.*, the force field has to be only a perturbation of the harmonic confinement). In this way, it is possible to avoid the switching-off of the force field for different values of λ.
For practical reason explained below, it is possible to switch to a rappresentation based

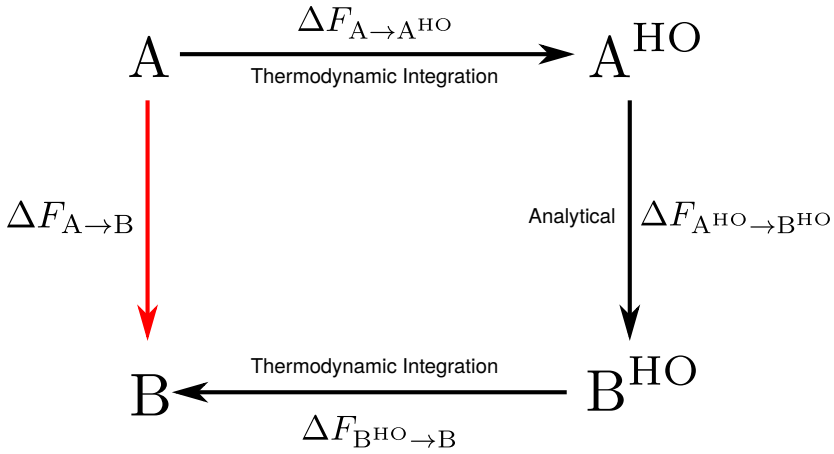


Figure 1.2: Simplified Confinement Method thermodynamic cycle. The ΔF between the force field states and the harmonic ones are obtained via TI, while the ΔF between the two confined states is obtained analytically. The target free energy difference is represented by a red line.

on typical oscillation frequency

$$4\pi^2\nu^2 = \frac{k_f}{m}.$$

And the potential (1.6) becomes

$$U(q,\lambda) = U_{\mathrm{ff}}(q) + 2\pi^2 m\lambda\nu^2 \left|q - q_0\right|^2. \tag{1.7}$$

Inserting the new potential (1.7) in the expression of ΔF for the TI (1.5) and changing the integration variable to $\zeta = \lambda\nu^2$ we have

$$\Delta F_{\mathrm{A}\to\mathrm{A}^{\mathrm{HO}}} = 2\pi^2 M \int_0^{\nu^2} \langle \left|q - q_0^{\mathrm{A}}\right|^2 \rangle_\zeta \, d\zeta,$$

where M is the total mass of the system. The squared average displacement can be expressed in terms of average of the potential energy of the confinement, obtaining

$$\Delta F_{\mathrm{A}\to\mathrm{A}^{\mathrm{HO}}} = M \int_0^{\nu^2} \frac{\langle U_{\mathrm{HO}} \rangle_\zeta}{\zeta} \, d\zeta. \tag{1.8}$$

And the integral is carried out numerically using a logarithmic fit $\langle \left|q - q_0\right|^2 \rangle_k \simeq ak^b$. Details on numerical integral calculation are in appendix A.
We choose ν^2 as integration endpoint because we can define a physical reasonable boundary, which is in the order of the typical vibrational frequency of hydrogen atoms (~ 90 ps^{-1}).

The analytical part of the thermodynamic cycle is based on the idea that the harmonically confined state subject to the potential defined in (1.6) is a system of $3N$ non-interacting harmonic oscillators. Introducing a best-fit on the configuration based on the reference

initial state in the MD, we remove all the roto-translational entropy contribution in free energy. In this way, the free energy of a confined system (which is represented only by vibrational and roto-translational entropy) results

$$F_{\text{HO}} = E_0 + 3Nk_BT\log\left(\frac{h\nu}{k_BT}\right) + F_{\text{rot+tr}},$$

where $E_0 \equiv U_{\text{ff}}(q_0)$, h is the Planck constant, ν is the typical frequency of the oscillators, and $F_{\text{rot+tr}}$ is the roto-translational contribution. In the case of the same number of atoms N, the translational contribution is equal to 0 (see appendix A).
The free energy difference between two different states of the same system results

$$\begin{aligned}\Delta F_{\text{A}^{HO}\to\text{B}^{HO}} &= F_{\text{B}^{\text{HO}}} - F_{\text{A}^{\text{HO}}} \\ &= E_0^{\text{B}} + 3Nk_BT\log\left(\frac{h\nu}{k_BT}\right) + F_{\text{rot+tr}}^{\text{A}} - E_0^{\text{A}} - 3Nk_BT\log\left(\frac{h\nu}{k_BT}\right) - F_{\text{rot+tr}}^{\text{B}} \\ &= E_0^{\text{B}} + F_{\text{rot}}^{\text{B}} - E_0^{\text{A}} - F_{\text{rot}}^{\text{A}}\end{aligned}$$

Considering all the thermodynamical cycle, the free energy difference results

$$\begin{aligned}\Delta F_{\text{A}\to\text{B}} &= \Delta F_{\text{A}\to\text{A}^{\text{OH}}} + \Delta F_{\text{A}^{\text{OH}}\to\text{B}^{\text{OH}}} + \Delta F_{\text{B}^{\text{OH}}\to\text{B}} \\ &= M\int_0^{\nu^2}\frac{\langle U_{\text{HO}}^{\text{A}}\rangle_\zeta}{\zeta}\,d\zeta + E_0^{\text{B}} - E_0^{\text{A}} - M\int_0^{\nu^2}\frac{\langle U_{\text{HO}}^{\text{B}}\rangle_\zeta}{\zeta}\,d\zeta + \Delta F_{\text{A}\to\text{B}}^{\text{rot}}\end{aligned} \tag{1.9}$$

We applied this technique to compute $\Delta\Delta G$ relative to single-point mutations of a hairpin (section 1.4) and we improved SCM performance in ΔG calculation using a interpolation/extrapolation scheme for confinement energies (section 1.5).

1.4 Computation of mutational $\Delta\Delta G$

The $\Delta\Delta G$ of mutation for a protein (or a peptide) is defined as

$$\begin{aligned}\Delta\Delta G_{\text{WT}\to\text{mut}} &= \Delta G_{\text{F}_{\text{mut}}\to\text{U}_{\text{mut}}} - \Delta G_{\text{F}_{\text{WT}}\to\text{U}_{\text{WT}}} \\ &= \Delta G_{\text{U}_{\text{WT}}\to\text{U}_{\text{mut}}} - \Delta G_{\text{F}_{\text{WT}}\to\text{F}_{\text{mut}}}\end{aligned}$$

where WT refers to the wild type protein, mut to the mutant of interest, U for the unfolded state and F for the folded state of the protein.

In the context of SCM, the calculation of mutational $\Delta\Delta G$ can be expressed in the double thermodynamic cycle in Figure 1.3.
The right cycle ($\Delta G_{\text{F}_{\text{WT}}\to\text{F}_{\text{mut}}}$) can be computed using as reference state the global minimum of the force field in biological conditions, while for the left cycle we have a problem in the reference state definition. The unfolded state is a huge collection of possible conformational states dominated by entropy [21]; for this reason it is impossible to cover all the phase space belonging to the unfolded state considering only a single conformation.

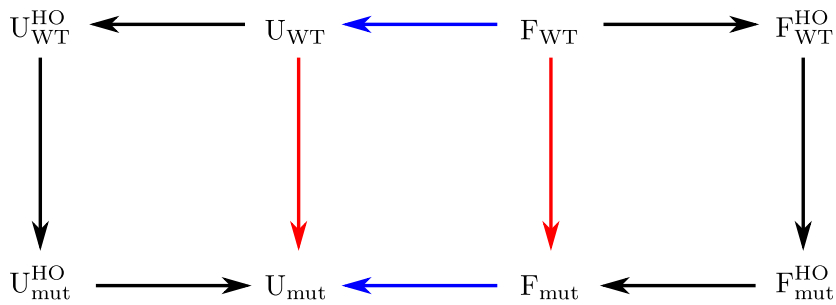


Figure 1.3: $\Delta\Delta G$ SCM thermodynamic cycle. The difference in ΔG which defines mutational $\Delta\Delta G$ regards blue arrows, while the equivalent difference is carried out on the red transitions, that can be computed using two SCM cycles black arrows).

Following the work of Seeliger and De Groot [41], we will approximate the unfolded state as a random coil where every residue is in contact with the solvent and with negligible long-range interaction between distant amino acids. By means of this approximation we can consider only local changes in environment. For this reason, we compute $\Delta G_{U_{WT} \to U_{mut}}$ in GXG tripeptides, where X is the mutated amino acid (G is the glycine). This permits us to consider all the chemical bonds (the covalent bonds at backbone terminals of the amino acid of interest), neglecting all the interaction with other residues (glycine have the simplest possible sidechain, an hydrogen atom).

In both the thermodynamic cycles we have a fundamental difference with respect to the conformational ΔG calculation: the starting state and the final state have a different number of atoms. In the analytical part of the cycle we have to add the vibrational entropy due to the change in N and, furthermore, add a new term for roto–translational free energy term (extensive derivation for this last term is in appendix A):

$$\Delta G_{WT^{HO} \to mut^{HO}} = E_0^{mut} - E_0^{WT} + 3(N^{mut} + N^{WT}) k_B T \log\left(\frac{h\nu}{k_B T}\right) + \Delta G^{rot+tr}_{WT^{HO} \to mut^{HO}} \quad (1.10)$$

The application of the GXG tripeptides approximation gives us an advantage: we have to compute only a single thermodynamic integration for every amino acid kind; in principle, it is possible to compute all the possible left cycles after 20 different thermodynamic integrations: one for every natural residue. This leads us also to a possible systematic error due to the neglect of long-range interactions; for instance, the presence of a transient secondary structure in the denatured state or the existence of a metastable intermediate state [42] can not be taken into account.

We applied this technique to study the thermostability and compute the $\Delta\Delta G$ of 8 different mutants to alanine (W43A, Y45A, D46A, D47A, T49A, K50A, F52A, and V54A, see Figure 1.4) of streptococcus protein G immunoglobulin binding domain hairpin (PDB code: 1GB1 [43] residues 41-56), comparing them with experimental data [44]. Only 5 mutants (D46A, D47A, T49A, K50A, and V54A) have an experimental $\Delta\Delta G$ value, while the remaining 3 (W43A, Y45A, and F52A) were shown to be too destabilizing to obtain

accurate numerical results.

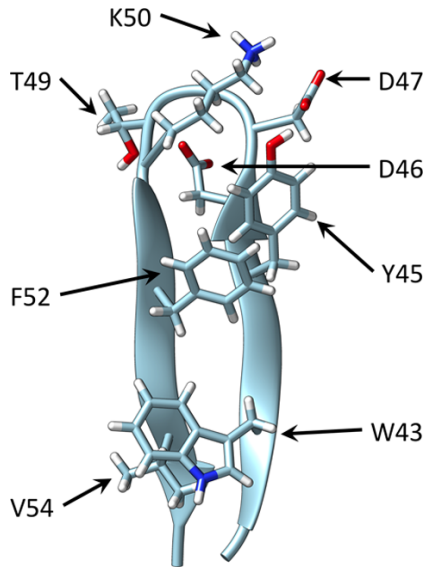


Figure 1.4: Wild type streptococcus hairpin of protein G immunoglobulin binding domain. Mutated residues are highlighted.

1.4.1 Computational Implementation

All the minimizations and MD simulations were carried out with GROMACS 4.5.7 [45] using the AMBER99SB force field [46]. The MD engine was patched with PLUMED 2.1 [47] for the biasing potential and for the roto-traslational best fit.
Starting from the crystallographic structure from Protein Data Bank, we designed the tripeptides and all the single-point mutant for the hairpin with the PyMOL framework [48]. Every structure (original crystallographic structure included) was then minimized with a three-step protocol:

1. $2 \cdot 10^4$ steps with steepest-descent algorithm, with a convergence tolerance of 1 kJ/(mol nm);

2. $2 \cdot 10^4$ steps with conjugate gradient algorithm, with a convergence tolerance of 1 kJ/(mol nm);

3. $2 \cdot 10^4$ steps with L-BFGS algorithm [49, 50], with a convergence tolerance of 10^{-3} kJ/(mol nm).

obtaining in this way the reference structures for the SCM protocol.

All the MD simulations were performed at $T = 298$ K in GBSA implicit solvent [51] with a leapfrog stochastic integrator [52] with a friction coefficient of 1 ps^{-1}, using two

Hamiltonian Replica Exchange simulations [53] for the lowest restraining, each one involving 4 consecutive frequencies. Our choice on the use of implicit instead of explicit solvent, while not necessary for the SCM approach (see [54], where the authors show a SCM technique with a further explicit solvation calculation), permits us to speed up simulations and sample better the conformational space of the system. Furthermore, we found out that the main source of systematic error is related to the tripeptide approximation of the denatured state (see below), making this choice not crucial. To avoid noise in harmonic oscillators we did not use LINCS [55] bond constraining in our simulations. We simulated a total of 16 different frequencies exponentially spaced for the thermodynamic integration, where the total simulation time was 10 ns for every replica while the timestep and the number of steps were adapted to the restraint frequency to avoid integrator errors. The parameters for the restraining is in table 1.1.

Table 1.1: Simulation parameters for the SCM windows

Simulation	ν [ps^{-1}]	Δt [fs]	n_{step}	Simulation	ν [ps^{-1}]	Δt [fs]	n_{step}
1	0.020	1	10^7	9	3.474	1	10^7
2	0.039	1	10^7	10	6.601	1	10^7
3	0.074	1	10^7	11	12.541	1	10^7
4	0.140	1	10^7	12	23.828	1	10^7
5	0.267	1	10^7	13	45.273	0.25	$4{\cdot}10^7$
6	0.506	1	10^7	14	86.019	0.25	$4{\cdot}10^7$
7	0.962	1	10^7	15	163.435	0.25	$4{\cdot}10^7$
8	1.828	1	10^7	16	310.527	0.25	$4{\cdot}10^7$

To decrease the statistical error, the multistate Bennett acceptance ratio (MBAR) estimator [56] was used to calculate $\langle U_{\mathrm{HO}} \rangle_\zeta$ in all the windows. The last four simulated windows have shorter time steps to avoid integrator problems in the proximity of the resonance frequency of hydrogen atoms. A longer discussion on energy reweighting using multiple simulation is in section 1.5.

1.4.2 Results

Results in Figure 1.5 show that $\Delta\Delta G$ values for D46A, T49A, V54A, and D47A strongly correlate with the experimentally measured values when using the unfolded tripeptide reference structure of Figure 1.6. The calculated $\Delta\Delta G$ values of these mutants fit the experimental values according to

$$\Delta\Delta G_{\mathrm{calc}} = 1.08 \Delta\Delta G_{\mathrm{exp}} - 4.79 \text{ kcal/mol}$$

shown by the dashed black line labeled "A" in Figure 1.5, with a correlation coefficient of 0.987. The K50A mutant strongly deviated from the trend, however. A possible explanation for this observation is the strong mobility of the charged side chain, which could be stabilized in the denatured state through nonlocal interactions with the rest of the chain. If that is the case, the unfolded state configuration of the wild type differs from that of

the mutant, violating our initial assumption.

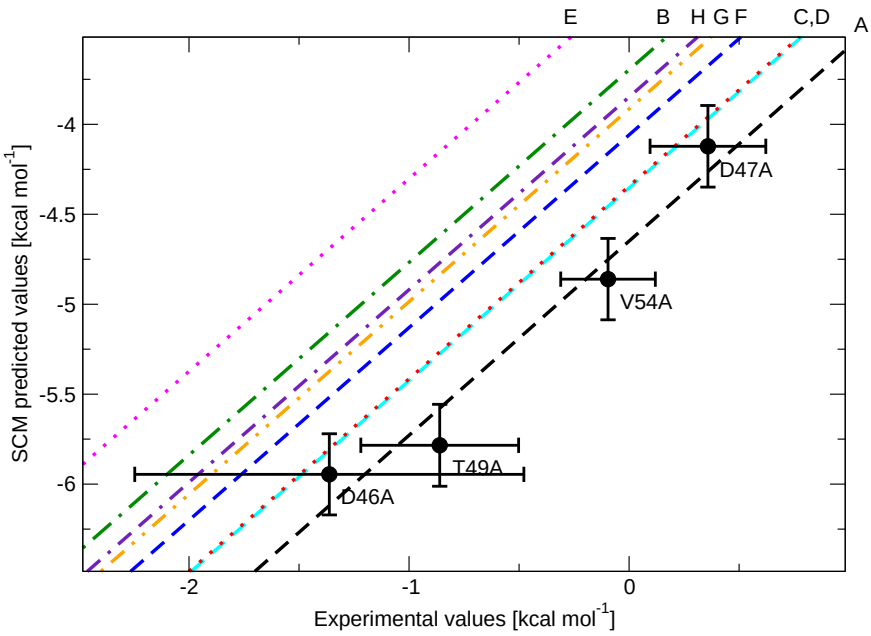


Figure 1.5: Comparison between experimental and calculated $\Delta\Delta G$ values for GB1 domain hairpin mutants. Trend line labels refer to different GAG tripeptide conformations (Figure 4). Not shown is K50A, which had a calculated $\Delta\Delta G$ of -13.87 $\pm$ 0.24 kcal/mol and an experimental $\Delta\Delta G$ of -0.45 $\pm$ 0.26 kcal/mol. Also not shown are W43A (calculated: -3.72 $\pm$ 0.24 kcal/mol), Y45A (calculated: -7.66 $\pm$ 0.24 kcal/mol) and F52A (calculated: -8.82 $\pm$ 0.24 kcal/mol), which were also unstable in experiments but lack experimental $\Delta\Delta G$ values.

In this framework, because of the use of an implicit solvent model and the extended tripeptide-model for the unfolded state, systematic errors in the calculation of $\Delta\Delta G$ are to be expected. However, if the mutations do not affect the conformations of the unfolded and folded states, calculated $\Delta\Delta G$ values should strongly correlate with experimental values.

Considering that all the left SCM cycle used the same $\Delta G_{U_{mut}^{HO} \rightarrow U_{mut}}$ computed on GAG tripeptide (all the considered amino acids mutates to alanine), we used 8 different reference conformations (Figure 1.6, where conformation A is the one that followed the original minimization protocol explained in Methods section, and for the other an explanation is given below) to evaluate the possibility of a systematic error due to the initial conformation.

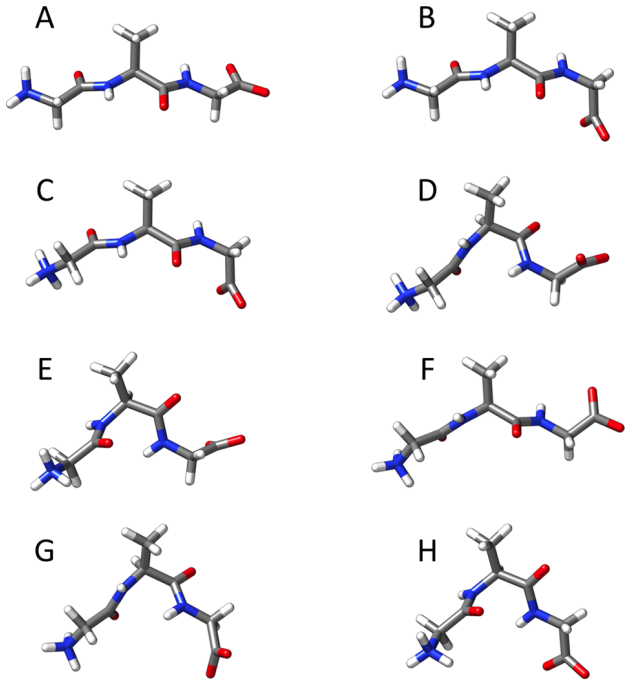


Figure 1.6: GAG tripeptide different reference conformations.

Although no experimental $\Delta\Delta G$ values were available for the W43A, Y45A, and F52A mutants, experiments showed that these mutations were medium (for W43A) to strongly destabilizing [44]. In accord with the experiments, Y45A and F52A showed the most negative $\Delta\Delta G$ values, but W43A showed the least negative value of all mutants. In fact, for W43A $\Delta\Delta G$ was higher than for the stabilizing D47A mutation. This deviation from experimental results could be ascribed to the presence of the large hydrophobic side chain of tryptophan which can form nonlocal interactions in the denatured state to avoid solvent exposure: these factors might be poorly represented in our minimalistic model of the unfolded state.

Despite a good correlation between the predicted and experimental data for most mutants, we observed a shift in $\Delta\Delta G$ values (-4.79 kcal/mol) compared to experimental values. To verify that this was indeed a systematic error, we changed the reference structures. We first changed the reference structures for the WT. Two new WT references for the folded state were used: the first taken after a 5 ns MD simulation of the WT at 298 K and a second by subjecting this MD structure to the three-step minimization protocol. After repeating the confinement simulations, we observed that the shift did not change substantially when using either native wild type reference states, (the first WT had a $\Delta\Delta G$ drift of 4.98 kcal/mol; the second 4.58 kcal/mol).
As mentioned above, we also reran the calculations using different GAG tripeptide reference structures for the confinement simulations of the unfolded mutant. For each local

minimum in conformational space (obtained from metadynamics [10] simulations), two reference structures were generated: the first by MD and the second by minimizing the MD structure using the three-step protocol. All GAG tripeptide reference structures are shown in Figure 1.6, and the effect on the calculated $\Delta\Delta G$ values is shown by the various dashed lines in Figure 1.5, labeled according to the structures of Figure 1.6. Also in these cases, the shift in $\Delta\Delta G$ did not change in a significant way.

Overall, the approach highly correlates with experimental $\Delta\Delta G$ values. Except for two outliers, the approach correctly predicted the order of thermodynamic stability of the single-point mutants: $\Delta\Delta G_{\mathrm{D47A}} > \Delta\Delta G_{\mathrm{V54A}} > \Delta\Delta G_{\mathrm{T49A}} > \Delta\Delta G_{\mathrm{D46A}} > \Delta\Delta G_{\mathrm{Y45A}} > \Delta\Delta G_{\mathrm{F52A}}$.

Summarizing, we have shown that confinement simulations can be used as an alternative to alchemical free energy simulations when calculating mutational free energies and that this method can be used to provide a practical scanning tool to evaluate the direction of free energy changes upon mutations. We have illustrated the method by evaluating the differences in unfolding free energies for mutants of a small β-hairpin. Despite the use of additional approximations with respect to the unfolded state representation, the approach largely predicted the correct order of thermodynamic stabilities for the mutants.

1.5 SCM efficiency enhancement

As we have shown in the previous 2 sections, the confinement method provides an efficient and robust way to calculate conformational free energy differences, even for states that are highly dissimilar in structure, but the method has other computational benefits that can be exploited. The free energy of transforming the system to a set of independent harmonic oscillators is obtained through a series of restrained simulations, each with a different strength of the harmonic restraint. While the strengths differ, the center of the restraint is the same in all of these simulations and corresponds to the equilibrium structure of interest. This means that all configurational space accessible to the high restraint simulation is contained in the accessible configurational space of the low constraint simulation. Naturally, the numerical value of this overlap will be very small when there is a large difference in restraints (because the amount of accessible configurational space shrinks as the restraint increases in strength); however, the overlap can be particularly large for "neighboring" simulations for which the restraint force constants are closest in value. The overlap in configurational space is closely associated with the overlap of energy distributions, which is crucial for the accurate estimate of thermodynamic properties [57].

Here we exploit this overlap in order to maximize the efficiency and minimize the statistical error of confinement simulations. We will show that one can use the overlap in configurational space to accurately predict confinement energies at unsampled strengths and that this interpolation significantly decreases the error in calculated free energies.

We will also show that instead of sampling at given intervals, one can use the overlap to predict at which restraining strength to sample next for simultaneously optimal errors and costs. Finally, by coupling these interpolations and extrapolations to relaxation time analyses, we will introduce a robust protocol for optimal confinement simulations, which is illustrated by applications to the alanine n-peptide and lactoferricin.

Because the system is confined to the vicinity of the same reference structure $\{q_0\}$ in each simulation, there is large spatial overlap between these sets of configurations. This means that the configurations obtained at a given frequency can be used to estimate ensemble averages at a different frequency. Consider N_i configurations obtained from a simulation with restraint frequency ν_i.
The ensemble average of an observable A at frequency ν_j is given by

$$\langle A \rangle_{\nu_j} = \langle A\, e^{(-\beta(U_j - U_i))} \rangle_{\nu_i}\, e^{-\beta \Delta G^{\mathrm{HO}}_{i \to j}} \tag{1.11}$$

where $\beta = (k_B T)^{-1}$, $\Delta G^{\mathrm{HO}}_{i \to j} = G^{\mathrm{HO}}_j - G^{\mathrm{HO}}_i$ is the free energy difference between two states restrained with typical frequencies ν_i and ν_j, and U_i and U_j are the biased potential for such frequencies.
Remembering the potential expression (1.7), the difference between two biased states with different frequencies ν_i and ν_j results

$$U_j(q) - U_i(q) = 2\pi^2 m \left| q - q_0 \right|^2 \left(\frac{\nu_j^2}{\nu_i^2} - 1 \right),$$

where the accuracy in observable average estimation increases for smaller $|\nu_i - \nu_j|$. Another way to improve accuracy in average estimation is to perform a longer sampling; this can be done combining the conformations obtained from all the simulations at all frequencies. As we did in the previous work (section 1.4), we applied Multistate Bennett Acceptance Ratio (MBAR) [56] as estimator for confinement energy. MBAR is a technique that, given an ensemble of simulations performed in different conditions (in our case, different confinement potentials), returns thermodynamic averages for each (even not simulated) condition, reweighting all the available data. Its main advantage, compared with other state-of-the-art methods (*i.e.* Weighted Histogram Analysis Method), consists in associating an error to each computed average.

The application of the MBAR estimation has been performed also in frequencies higher than the highest simulated during the TI calculation, giving an extrapolation that can be used to assess the most convenient frequency for a new TI window. Operatively, we start from a set of simulations performed with different restraints, up to an arbitrary frequency $\nu_{\max}$. We subsequently employ the extrapolation with MBAR to estimate the confinement energy for a set of unsampled frequencies $\nu_i > \nu_{\max}$. The computational cost of this extrapolation is low (much lower than the actual sampling) and nearly independent of the number of unsampled frequencies.
We chose this method because, in extrapolating toward higher frequencies, phase space is compressed. This means that all relevant areas of space for the higher frequency restraint were sampled in the lower frequency simulation (but insufficiently). If we were to

choose the other direction, that is, sampling at the higher frequency followed by extrapolation to the lower frequency, certain regions of space important for the low frequency restraint would be left unsampled. After the extrapolation, a confinement simulation is performed for each of the new frequencies in order to obtain the actual value of the confinement energy, and these calculated values are compared with those obtained from the extrapolation. The ratio between the extrapolated and actual confinement energy is a measure of the error of the extrapolation. We express this ratio as a function of the free energy difference ΔG^{HO} between the simulations at ν_i and $\nu_{\max}$. ΔG^{HO} can be obtained from equation (1.11) or, if simulations at multiple frequencies are used, from MBAR. The latter approach would yield somewhat more accurate extrapolations because more data is used. However, here we used data from only one simulation and the former approach in order to base all comparisons on the same amount of data. For a given $\nu_{\max}$, ΔG^{HO} increases with $|\nu_{\max} - \nu_i|$ and represents a meaningful quantity that can be compared across systems of different sizes.

The efficiency and accuracy can be improved further by considering the correlation time of the system. This correlation time is affected by the addition of harmonic restraints, especially at high frequencies, when the confinement energy accounts for a large portion of the total potential energy. In addition, these restraints limit the configurations accessible to the system to the ones close to the reference structure q_0, and the phase space to sample gets smaller as the frequency gets higher. To attain comparable sampling for each frequency, different sampling times are therefore needed, which can be estimated from the correlation time. These were estimated by block-averaging the confinement energies [58] and also by calculating the autocorrelation function of the confinement energy.

1.5.1 Computational Implementation

Alanine *n*-peptide

We performed confinement simulations of capped alanine n-peptides (n = 2, 4, 6, 8, and 10, see Figure), with the general formula $CH_3CO\text{-}Ala_{n-1}\text{-}NHCH_3$. These simulations were carried out with the CHARMM program [59], using the CHARMM polar hydrogen parameter set param19 [60] and the ACE implicit solvent model [61]. The two lowest-energy conformations of the alanine dipeptide are $c7_{ax}$ and $c7_{eq}$, which for the force field and implicit solvent method used, correspond to backbone dihedral angles of $(\varphi, \psi) = (61.4, -71.4)$ and $(78.0, 138.7)$ degrees, respectively. The $c7_{ax}$ and $c7_{eq}$ conformations were used as the reference structures for the alanine dipeptide. Larger alanine n-peptide systems behave nearly like independently linked alanine dipeptides when the peptides are in the $c7_{ax}$ and $c7_{eq}$ states [62, 63]. For instance, the energy minima $(c7_{ax}, c7_{ax})$ and $(c7_{eq}, c7_{eq})$ of the alanine 3-peptide correspond to $(\varphi_1, \psi_1, \varphi_2, \psi_2) = (61.1, -72.1, 59.6, -71.6)$ and $(-77.4, 137.7, -76.7, 137.8)$ degrees, respectively.
For all $n \geq 3$ alanine systems, the confinement reference structures were obtained by setting all the (φ, ψ) dihedral angles to the values of $c7_{ax}$ and $c7_{eq}$ of the alanine dipeptide and performing an energy minimization. In the following, these configurations are simply named $c7_{ax}$ and $c7_{eq}$, independently of the number of dihedral angles (see Figure

1.7).

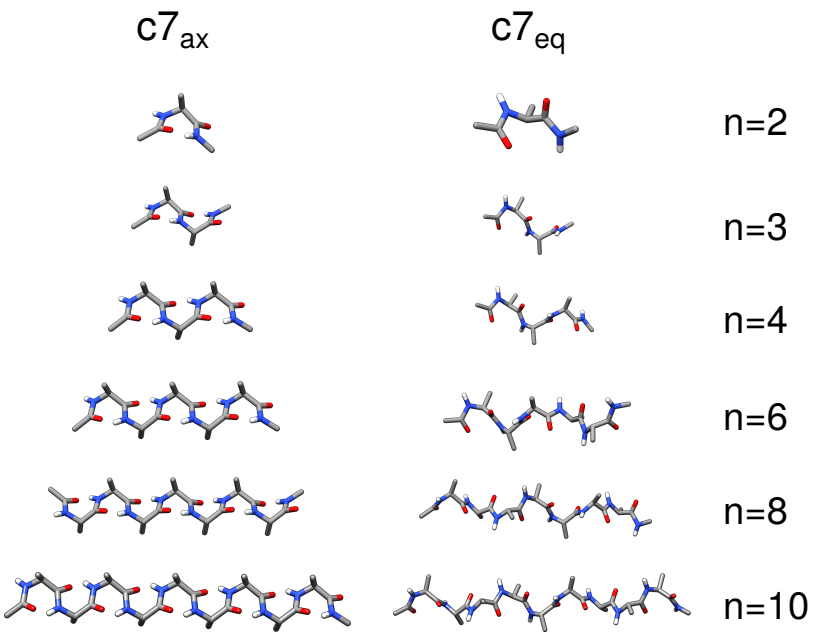


Figure 1.7: The different analyzed alanine n-peptides in $c7_{ax}$ and $c7_{eq}$ conformations.

Also in this case, the analytical transition of the free energy difference contains a non-vibrational term, which is the rotational entropy contribution. The moments of inertia were obtained for the reference structures. The corresponding contribution to the free energy difference between $c7_{ax}$ and $c7_{eq}$ equals

$$\Delta G^{\text{rot}}_{\text{A}^{\text{HO}}\rightarrow\text{B}^{\text{HO}}} = -\frac{1}{2}Nk_BT\log\left(\frac{I_1^{\text{B}}I_2^{\text{B}}I_3^{\text{B}}}{I_1^{\text{A}}I_2^{\text{A}}I_3^{\text{A}}}\right)$$

where I_i are the momenta of inertia and N is the number of molecules (1 in our case). All confinement simulations were performed using Langevin dynamics at 300 K, with friction coefficients of 1, 5, 10, or 20 ps^{-1}. The time step had a maximum value of 1 fs and was adjusted depending upon the restraint frequency. It was chosen so there are at least 80 time steps per harmonic oscillator period, resulting in smaller time steps for higher frequencies. Different time steps were tested, which showed that at least 40 steps per period are required to obtain accurate estimation of the confinement energy. A conservative value of 80 steps/period was then chosen. SHAKE [64] was not used in the simulations. To further restrict sampling to the state of interest, especially at the lowest frequencies, we also added flat-bottom dihedral restraints. These were centered on the energy-minimized values, with a force constant of 10 kcal/mol/rad^2 and a width of 2.5°. This value was chosen so that the states of interest are the same as the ones defined in the umbrella sampling. Interpolation of the confinement energy was done for 10 frequencies, equally spaced in log-space, between consecutive simulations. Adding more points did not change the final free energy difference or the error bars.

For comparison, free energy differences between the $c7_{ax}$ and $c7_{eq}$, conformations were also obtained from one-dimensional umbrella sampling simulations [9]. For the alanine n-peptide, the transformation from $c7_{ax}$ and $c7_{eq}$ involves $(n-1)$ φ and $(n-1)$ ψ dihedral transformations. These angles were treated as reaction coordinates and changed one at a time (in the order $(\varphi_1, \psi_1) \rightarrow (\varphi_{n-1}, \psi_{n-1})$) while keeping the others constant. For each dihedral angle, 50 equally spaced umbrella windows were used, with a force constant of 150 kcal/mol/rad^2 and a simulation time of 100 ns per window. To maintain the trans peptide configuration, flat-bottom dihedral restraining potentials were used for the ω backbone dihedral angles with a force constant of 10 kcal/mol/rad^2 and a width of 90°. Simulations were performed with Langevin dynamics at 300 K, using a 1 fs time step, no SHAKE [64], param19 [60], and ACE [61]. Potentials of mean force (PMF) were obtained from MBAR [56].

Lactoferricin

Bovine lactoferricin is 25-residue peptide cleaved from lactoferrin with antimicrobial properties [65]. In lactoferrin, the sequence is folded into an α-helix followed by a β-strand, while the cleaved peptide adopts a β-hairpin fold; the peptide contains one disulfide bond (Figure 1.8) [66, 67].

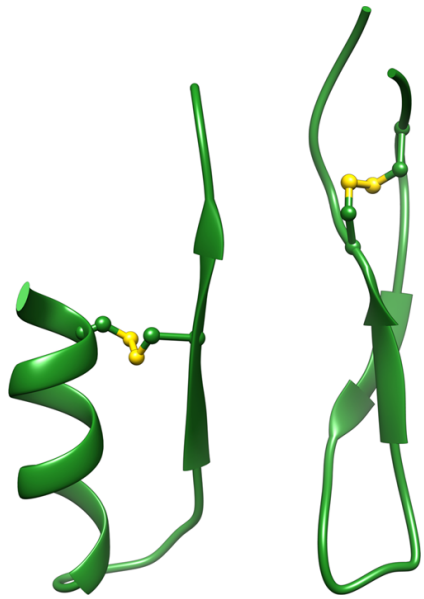


Figure 1.8: Structure of lactoferricin. Solution structure on right, structure of the lactoferricin sequence within the lactoferrin protein on left, with disulfide bond highlighted.

No spontaneous conformational transitions were observed in long unbiased MD simulations [68]. Because of its size and the complexity of the transition, lactoferricin is a good test system for the confinement method and representative of the more challenging biological systems that are the ultimate target for the method.
The $\alpha + \beta$ conformation was obtained from residues 17-41 of lactoferrin (PDB: 1BLF

[67]), while the β-hairpin conformation was taken from lactoferricin in solution (PDB: 1LFC [66]). We used the CHARMM36 force field [69] with the GBMV implicit solvent model [70], Langevin dynamics, and no SHAKE [64]. A friction coefficient of 1 ps^{-1} was used for simulations with a confinement frequency lower than 2 ps^{-1} and a friction coefficient of 20 ps^{-1} for frequencies above. Interpolation of the confinement energy was done for 10 equally log-spaced frequencies between consecutive simulations. After an energy minimization, each conformation of the peptide was heated and equilibrated at 300 K. The reference structures used in the confinement simulations were obtained from rmsd-based clustering with a cut off of 3.5 Å of a 25 ns unrestrained trajectory. All simulations were conducted with the CHARMM program [59].

1.5.2 Results

Alanine n-Peptide

Figure 1.9 shows the ratio between the extrapolated and actual confinement energies as a function of ΔG^{HO} for multiple values of $\nu_{\max}$. Curves for all alanine systems are also provided; a value of one indicates that the extrapolation perfectly predicted the confinement energy. As expected, deviations from one strongly increased with the free energy difference and the ratio was very close to one for small ΔG^{HO}. A notable feature is that the extrapolation stayed accurate for larger free energy differences as $\nu_{\max}$ increases. In other words, as the system becomes more harmonic, it becomes easier to predict the result of a new simulation. This is due to the fact that at larger frequencies, the harmonic restraints represent a larger portion of the total energy and the configurational space is compactly distributed around q_0 in a predictable manner. In addition, we observed that the extrapolation is more accurate as the size of the system increases. Larger systems have narrower energy distributions, so that the weights in equation (1.11) are closer to one another. More configurations will therefore contribute significantly to the ensemble average at another frequency, thus lowering the error. This is a particularly encouraging feature which will facilitate the application of the Simplified Confinement Method to larger and more complex systems.

The information on Figure 1.9 can be used to extract the maximum value of ΔG^{HO} for which the extrapolation error is below a desired threshold. We chose 5%, a fairly conservative value for this error, which corresponds to an extrapolated/actual confinement energy ratio of 0.95 or 1.05. In the following, we will refer to this spacing as $\Delta G^{\mathrm{HO}}_{\mathrm{ext}}$.
Figure 1.10 shows $\Delta G^{\mathrm{HO}}_{\mathrm{ext}}$ as a function of frequency for the alanine n-peptides. While the curves are bumpy (due to the fact that the ratios switched between 0.95 and 1.05, discretization of ν, and finite sampling), the graph shows several clear trends: consistent with the results of Figure 1.9, $\Delta G^{\mathrm{HO}}_{\mathrm{ext}}$ increased with both frequency and system size.

The physical relevance of $\Delta G^{\mathrm{HO}}_{\mathrm{ext}}$ is the following. When the free energy difference between the sampled system at $\nu_{\max}$ and the unsampled system at higher frequency is $\Delta G^{\mathrm{HO}}_{\mathrm{ext}}$, there is sufficient overlap in distribution functions to estimate, within some preselected error bound (here 5%), the confinement energy at the unsampled frequency from

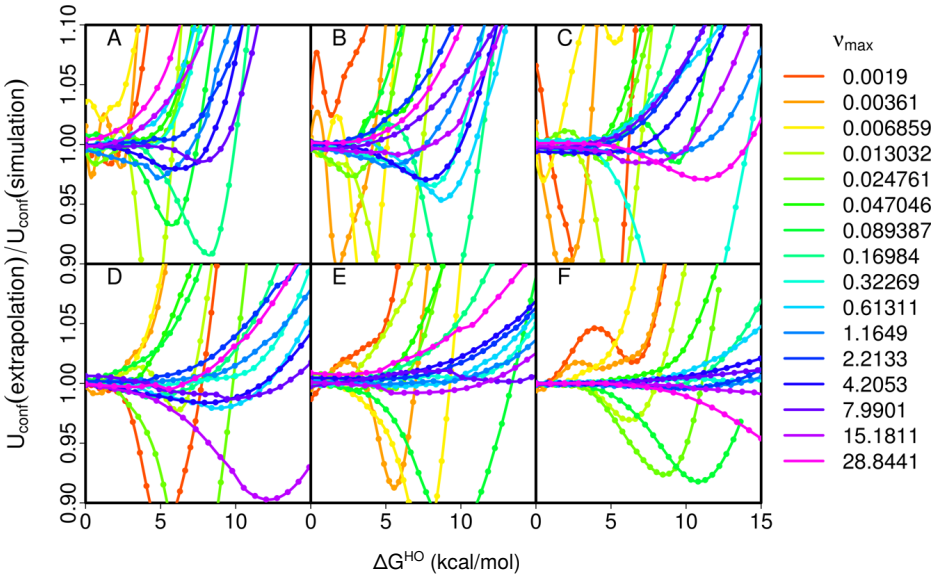


Figure 1.9: Ratio between the confinement energy computed from extrapolation and from an actual simulation for alanine 2-peptide (A), 3-peptide (B), 4-peptide (C), 6-peptide (D), 8-peptide (E), and 10-peptide (F). Simulations with different restraints up to a frequency $\nu_{\max}$ were used to extrapolate the confinement energy at higher frequencies. These extrapolated states have a higher free energy than the state corresponding to $\nu_{\max}$, with a difference of ΔG^{HO}.

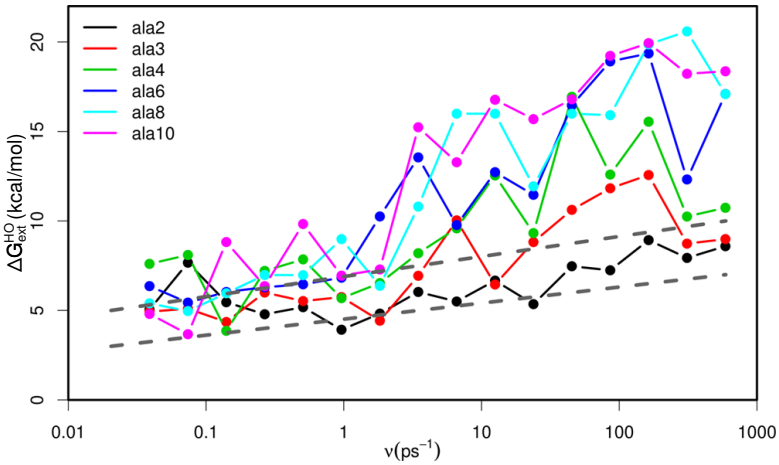


Figure 1.10: Free energy difference corresponding to an error of 5% on the confinement potential obtained by extrapolation. This free energy difference is between the simulated system having the highest harmonic restraints and the one corresponding to the extrapolated frequency.

simulated data at ν_{max}. This means that after sampling at both frequencies, there will be sufficient overlap in distribution functions to accurately calculate confinement energies at frequencies in between. The accuracy of this interpolation will be higher than the accuracy of the extrapolation because more data is available for the interpolation (one extra set of simulations). Furthermore, the error of interpolation can be reduced further by taking into account all simulated data at all simulated frequencies. As shown below, this interpolation significantly reduced the overall error in calculating the configurational free energies. Thus, we propose to exploit ΔG_{ext}^{HO} as guideline for selecting the frequency spacing of the simulations. The goal of this procedure is to pick the maximum spacing at which high quality interpolations remain feasible, thereby obtaining high accuracy at minimal computational costs. If we have a set of simulations up to frequency ν_{max}, the next frequency of simulation will be picked such that its free energy difference with the ν_{max} simulation is ΔG_{ext}^{HO}.

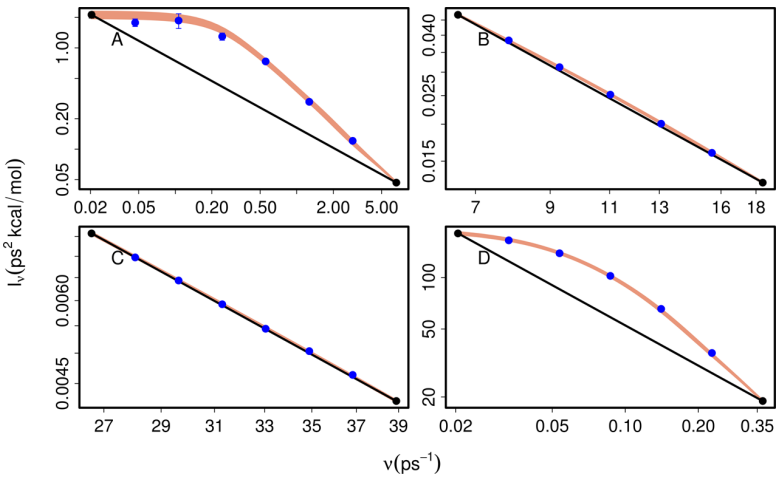


Figure 1.11: Interpolation of the confinement energy for alanine 2-peptide for different restraint frequency ranges: (A) low frequencies (0.02-6.6 ps^{-1}), (B) intermediate frequencies (6.6-18.4 ps^{-1}), and (C) high frequencies (26.6-38.8 ps^{-1}), as well as (D) for alanine 10-peptide at low frequencies (0.02-0.37 ps^{-1}). The value of the confinement energy is interpolated (red) for frequencies between two initial simulations (black). The thickness of the red line represents the error bar. Additional simulations (blue) are added thereafter to estimate the accuracy of the interpolation.

Figure 1.11 shows that interpolation can be performed to obtain confinement energies at non-simulated frequencies. In Figure 1.11A, the value of the integrand function U/ν^2 is shown in black for the alanine dipeptide at simulated frequencies of 0.021 and 6.6 ps^{-1}. The free energy difference between these simulations was 4.2 kcal/mol. The black line represents the value of the integrand that would vary linearly with the logarithm of the frequency, which is the assumption made when performing the integration of equation (1.9) in logarithmic space (see appendix A) and also the analytical solution for harmonic oscillators. The red curve corresponds to interpolated values using MBAR. Additional simulations at intermediate frequencies confirm the accuracy of the interpolation.

The simulated values (blue dots) show that the integrand does not follow a straight line in logarithmic space but falls on the interpolated curve instead. The observed non-logarithmic/linear behavior is expected for this frequency range because the system is far from being purely harmonic. In fact, the harmonic terms contribute only 24% of the total energy at a frequency of 6.6 ps^{-1}. The interpolation accurately reproduced the observed behavior, which demonstrates that meaningful information about the system can be obtained through interpolation. It also shows how the interpolation can greatly increase the efficiency of the method: while all simulations (represented by black and blue symbols) would be needed to accurately compute the free energy difference over that frequency range, just two initial simulations (black point) are sufficient if the interpolation is used. Greater accuracy can also be achieved by reducing the discretization error arising from the frequency spacing, but this comes at additional computational costs. Figure 1.11B,C illustrate how the interpolation for the alanine dipeptide performs at higher frequencies. The free energy differences between these two frequencies are comparable to the one corresponding to Figure 1.11A (5.9 and 4.2 kcal/mol versus 4.2 kcal/mol). Again, the interpolation correctly estimated the integrand for frequencies that were not simulated. At higher frequencies, the integrand varied more linearly with the log of the frequency, as expected for more harmonic system. The same behavior was observed for the other alanine systems, such as alanine 10-peptide (Figure 1.11D).

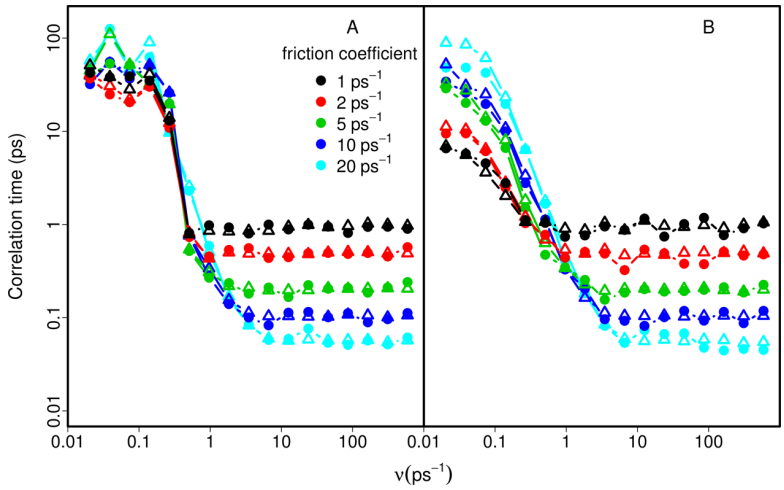


Figure 1.12: Correlation times for (A) alanine 2-peptide and (B) alanine 10-peptide for different restraint frequencies, calculated from the autocorrelation function of the confinement energy (triangle symbols) and from block analysis (circle symbols). Multiple friction coefficients (from 1 to 20 ps^{-1}) were used as parameters for the Langevin thermostat.

Figure 1.12 shows the correlation time of the confinement energy for the alanine dipeptide and decapeptide as a function of the restraint frequency. The correlation times were calculated by block-averaging [58] (indicated by circles) and from the autocorrelation function of the confinement energy (triangles). The two methods gave similar results, which indicates that the correlation time could be properly estimated. For the dipeptide,

the correlation time was similar for all frequencies $<$0.2 ps^{-1} and irrespective of the friction coefficient, while for the decapaptide, higher friction coefficients led to higher correlation times in this frequency range. This is likely due to the more complex landscape of the decapeptide, which has subbasins; visiting the various subbasins is hindered by large friction terms. Near a frequency of 0.2 ps^{-1}, the correlation times dropped significantly for all systems. At this frequency, U_{conf} represents between 2 and 4% of the kinetic energy. Apparently, this energy is sufficient to limit the system to one subbasin, which explains the precipitous decline in correlation time. At high frequencies, low correlation times were observed, inversely proportional to the friction coefficient. We checked that the average values of the confinement energies were not affected by the friction coefficient so that higher friction coefficients indeed led to faster sampling. The simulation time needed to obtain a given number of independent measurements is proportional to the correlation time. Figure 1.12 shows that in order to obtain uniform sampling across the frequencies, much smaller simulation times are needed at higher frequencies. In addition, by increasing the friction coefficient at high frequency, the simulation time can be reduced, thereby further increasing the efficiency of the calculation.

To demonstrate the increased performance of the confinement method through extrapolation, interpolation, and assessment of correlation times, we computed the free energy difference between the c7$_{\text{ax}}$ and c7$_{\text{eq}}$ conformations of the alanine n-peptides. For each frequency, the correlation time of the confinement energy was estimated at regular time intervals and the simulation was stopped when the number of independent measurements (which is the simulation time divided by the correlation time) was at least 1000. The confinement simulations were run in an iterative manner. The first simulation was performed at a frequency $\nu_1 = 0.02$ ps^{-1}.
The frequency of the next simulation was calculated by extrapolation. 4 different strategies were employed:

1. The first used a constant free energy spacing of 5 kcal/mol. The other three strategies used the information on Figure 1.10 to vary this spacing as a function of ν.
2. In the second strategy, the spacing was system-dependent and obtained from a log-linear best fit of $\Delta G^{\text{HO}}_{\text{ext}}$ to ν.
3. The third strategy was system independent, and given by $\Delta G^{\text{HO}}_{\text{ext}} = 3{+}0.388\log(\nu/\nu_1)$ (indicated by the lower dashed line in Figure 1.10), a conservative estimate of $\Delta G^{\text{HO}}_{\text{ext}}(\nu)$.
4. In the last strategy, a more aggressive estimate was chosen (indicated by the upper dashed line in Figure 1.10): $\Delta G^{\text{HO}}_{\text{ext}}(\nu) = 5 + 0.485\log(\nu/\nu_1)$. This iterative process of extrapolation and simulation was repeated until the value U_{conf} is equal to the purely harmonic value of $3Nk_BT/2$ was met.

Table 1.2 summarizes the free energy differences obtained with these strategies. For comparison, the table also shows the free energies obtained from one-dimensional umbrella sampling (ΔG_{US}), and from confinement simulations according to the setup of Ovchinnikov *et al.* [13] (ΔG_{Hom}), which involved 17 simulations at frequencies equally spaced

in log space, with a simulation time of 20 ns per simulation. Finally, the total cost of the simulations are shown relative to the total cost of the simulations using homogeneous spacing in frequency space.

Figure 1.13 shows the frequency spacing and number of steps for the alanine 10-peptide for each of the simulation setups; the number of steps is indicated by the length of the bars (but the unit length represents 10^8 steps for the homogeneous and 10^6 steps for the other setups). Because of the small time step, the high frequency simulations are particularly costly in the homogeneous frequency setup. For this reason, Ovchinnikov *et al.* recommended simulating up to a frequency of 86 ps^{-1} because the free energy difference for the alanine dipeptide is already converged at that frequency (even though the absolute free energies of the $c7_{ax}$ and $c7_{eq}$ configurations are not). A converged free energy difference at a lower frequency implies that the anharmonicity of the system at higher frequencies is the same for both configurations. However, this is not necessarily the case for large conformational changes, especially if new interactions were formed. While we also observed a convergence of the free energy difference for the alanine dipeptide at 86 ps^{-1}, omitting the high frequency portion led to an error in ΔG of between 0.10 and 0.31 kcal/mol for the alanine decapeptide and 0.98 kcal/mol for lactoferricin. Because of these errors, we included all frequencies until the absolute free energies of the $c7_{ax}$ and $c7_{eq}$ states were converged.

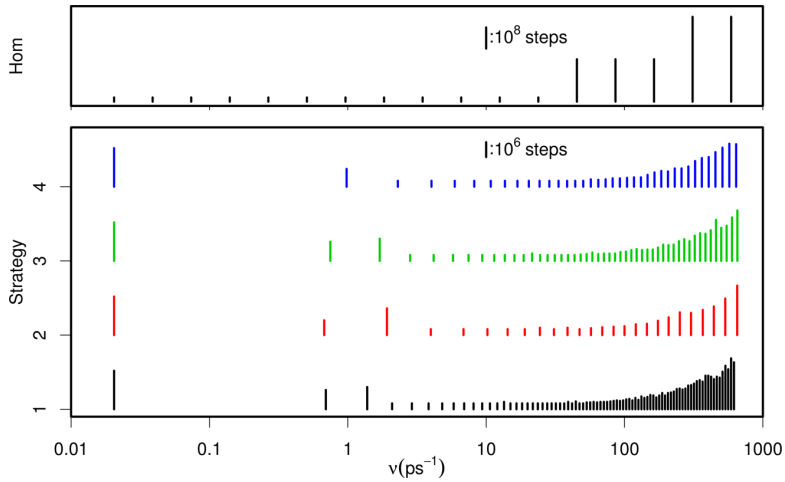


Figure 1.13: Number of MD steps used in the confinement simulations of the alanine 10-peptide in the $c7_{eq}$ conformation. Each vertical bar represents a simulation, and its length is linearly proportional to the number of steps. Different strategies were employed (see text), and led to different number of simulations.

Table 1.2: Free energy differences between the $c7_{ax}$ and $c7_{eq}$ conformations of the alanine n-peptides. All the errors are computed via error propagation (details are in Appendix A).
[a]) ΔG^{HO} = 5 kcal/mol.
[b]) $\Delta G^{\mathrm{HO}}_{\mathrm{ext}}$ from system-dependent best fits of to $\Delta G^{\mathrm{HO}}_{\mathrm{ext}}$ (Figure 1.9). These fits are: $\Delta G^{\mathrm{HO}}_{\mathrm{Ala_2}} = 3.19 + 0.54\log(\nu/\nu_1)$, $\Delta G^{\mathrm{HO}}_{\mathrm{Ala_3}} = 1.44 + 1.02\log(\nu/\nu_1)$, $\Delta G^{\mathrm{HO}}_{\mathrm{Ala_4}} = 0.95 + 1.27\log(\nu/\nu_1)$, $\Delta G^{\mathrm{HO}}_{\mathrm{Ala_6}} = 3.36 + 1.12\log(\nu/\nu_1)$, $\Delta G^{\mathrm{HO}}_{\mathrm{Ala_8}} = 3.27 + 1.38\log(\nu/\nu_1)$, $\Delta G^{\mathrm{HO}}_{\mathrm{Ala_{10}}} 4.87 + 1.55\log(\nu/\nu_1)$.
[c]) $\Delta G^{\mathrm{HO}}_{\mathrm{ext}} = 3 + 0.388\log(\nu/\nu_1)$, shown by the lower dotted line in Figure 1.9.
[d]) $\Delta G^{\mathrm{HO}}_{\mathrm{ext}} = 5 + 0.485\log(\nu/\nu_1)$, shown by the upper dotted line in Figure 1.9.
[e]) ΔG_{US} corresponds to the free energy difference calculated from one-dimensional umbrella sampling. ΔG_{Hom} corresponds the free energy obtained from confinement at frequencies that are equally spaced in log space. A total of 17 of these were performed per conformation, with a constant simulation time of 20 ns per simulation, according to the setup of the work of Ovchinnikov *et al.* [13]. In the strategies 1-4, the simulations are spaced in free energy space according to a given relation (see text). All free energies are in kcal/mol.
[f]) Total computational cost as a percentage of the total computational cost when using homogeneous spacing in log frequency.

			Strategy							
			1^a		2^b		3^c		4^d	
System	ΔG_{US} [e] (kcal/mol)	ΔG_{Hom} (kcal/mol)	ΔG (kcal/mol)	cost[f] (%)	ΔG (kcal/mol)	cost[f] (%)	ΔG (kcal/mol)	cost[f] (%)	ΔG (kcal/mol)	cost[f] (%)
Ala_2	-3.42 $\pm$ 0.06	-3.6 $\pm$ 0.3	-3.25 $\pm$ 0.06	2.57	-3.15 $\pm$ 0.06	2.41	-3.77 $\pm$ 0.07	2.19	-3.6 $\pm$ 0.1	1.50
Ala_3	-7.01 $\pm$ 0.09	-6.8 $\pm$ 0.4	-6.91 $\pm$ 0.05	3.54	-7.03 $\pm$ 0.06	2.20	-6.99 $\pm$ 0.07	2.29	-6.67 $\pm$ 0.07	1.98
Ala_4	-9.4 $\pm$ 0.2	-10.0 $\pm$ 0.4	-9.72 $\pm$ 0.05	4.94	-10.15 $\pm$ 0.08	2.33	-10.38 $\pm$ 0.07	3.12	-10.10 $\pm$ 0.08	2.86
Ala_6	-16.3 $\pm$ 0.3	-17.2 $\pm$ 0.5	-16.96 $\pm$ 0.05	6.72	-16.75 $\pm$ 0.09	2.88	-17.20 $\pm$ 0.06	4.46	-17.5 $\pm$ 0.1	3.67
Ala_8	-22.6 $\pm$ 0.3	-23.4 $\pm$ 0.5	-23.47 $\pm$ 0.05	9.24	-23.68 $\pm$ 0.09	3.27	-23.85 $\pm$ 0.06	5.50	-23.74 $\pm$ 0.09	4.47
Ala_{10}	-29.0 $\pm$ 0.4	-29.9 $\pm$ 0.3	-30.94 $\pm$ 0.05	12.28	-30.69 $\pm$ 0.09	3.67	-29.6 $\pm$ 0.2	6.87	-30.30 $\pm$ 0.09	5.39

The free energies obtained by umbrella sampling showed good agreement with the confinement free energies for all the alanine systems. The confinement simulations gave free energies of $\sim 3.3 - 3.5$ kcal/mol per (φ, ψ) dihedral angles, which shows the lack of correlations between (φ, ψ) backbone angles. Backbone rotation in the larger systems act as backbone rotation in independent alanine dipeptide systems, as observed before for the alanine tripeptide [62, 63]. The four extrapolation strategies gave similar ΔG values, with error bars 2-10 times smaller than with the homogeneous setup, which shows that all strategies could be used with good accuracy. The low error bars came from interpolation, which reduced the discretization error, the use of correlation times which ensured sufficient sampling, and the use of MBAR. While the use of interpolation does not significantly affect the free energy differences, it significantly contributes to the low error bars. With these strategies, the free energy spacing between consecutive simulations was small enough so that configurations from multiple simulations could be used to increase the statistics at a given frequency. If the free energy spacing would be too high, the weights in eq 5 would be very small, which would then effectively prevent the mixing of configurations and increase the error on the interpolated values. Use of MBAR and interpolation is therefore only useful when the frequencies are chosen judiciously. While all strategies gave values that were relatively close with low error bars, not all free energies of the various strategies overlap within their error bars, which indicates that the error bars are underestimated. This is likely due to insufficient sampling, which is not taken into account by the error bars. The problem of insufficient sampling cannot be easily solved, as one cannot quantify missing information.
Because each simulation was run until a fixed number of independent frames was obtained, the simulation time was different for each frequency. The low frequency simulations required the highest number of simulation steps because of large correlation times (Figure 1.12, 1.13). This correlation time was system-dependent because at low frequencies the harmonic restraints were fairly weak and the system dynamics were only slightly affected by the restraints. Upon increasing the frequency, the simulation time dropped significantly because of a drop in correlation times. At frequencies above $\sim$12.5 ps^{-1}, smaller time steps were required so that even though correlation times were roughly constant at high frequencies, the required number of steps increased (Figure 1.13). The various extrapolation strategies had non-constant frequency spacings that were larger than the homogeneous setup at low frequencies but smaller at high frequencies. The difference in spacing is due to the free energy difference between neighboring simulations (ΔG^{HO} of equation 1.11) which increases with frequency for a given frequency spacing. In addition, the total number of simulations increased with the size of the system. This makes sense because for a purely harmonic system, the free energy difference between two frequencies is proportional to the number of degrees of freedom. The number of simulations at high frequencies, where the system is largely harmonic, will therefore scale approximately linearly with the number of atoms. The cost of the first strategy, which is based on a constant free energy spacing of 5 kcal/mol between consecutive simulations, indeed increased with system size (Table 1.2).

The four extrapolation strategies led to much lower computational costs than the setup

with homogeneous sampling in log-frequency space (between 1.5 and 12.3% of the cost). This was mostly due to much shorter simulation lengths at high frequencies. The computational cost of the three system-independent setups (strategies 1, 3, and 4 in Table 1.2) increased with the size of the system (as discussed above). The setup with a constant free energy spacing of 5 kcal/mol was the most expensive of the four strategies, as it required the most simulations at high frequency. The setup based on best fits of $\Delta G_{\text{ext}}^{\text{HO}}$ was the cheapest overall as it used the largest free energy spacing. This advantage was especially pronounced for the alanine decapeptide, for which extrapolations to high free energy differences were possible (Figure 1.10). For future applications, obtaining system-dependent expressions for ΔG^{HO} is not practical due to the simulation costs associated with estimating this expression. System-independent strategies are much more practical, and even the most aggressive strategy (strategy 4) presented here was accurate as well as cost efficient.

While the optimized protocol consists of a combination of interpolation, extrapolation, optimized friction coefficients, and correlation analysis to determine simulation lengths, the contribution of the interpolation and extrapolation to the decrease in error was estimated for the alanine decapeptide by calculating ΔG using the 17 windows of the homogeneous setup and optimizing the friction coefficients, simulation length, and time steps only. This resulted in a free energy difference of -31.55 $\pm$ 0.44 kcal/mol, at 1.0% of the cost of the homogeneous setup. Relative to umbrella sampling results, the partially optimized 17-window strategy led to a larger shift in the free energy than the fully optimized strategies while the statistical error was also significantly larger (8.8, 4.9, 2.8, and 4.9 times larger than the fully optimized schemes, respectively).
When taking the difference in simulation lengths into account, these statistics suggest that the interpolation/extrapolation strategy reduces the error 2-fold for the alanine decapeptide. In fact, when calculating ΔG using the fully optimized schemes but at exactly the same cost of the partially optimized 17-window strategy (by using less frames), free energies of -31.18 $\pm$ 0.17, -31.61 $\pm$ 0.18, -29.58 $\pm$ 0.28, and -30.24 $\pm$ 0.22 kcal/mol were obtained for the four schemes, respectively.
Thus, for the alanine decapeptide, errors were about factor of 2 (2.6, 2.4, 1.6, and 2.0, respectively) lower when interpolation/extrapolation was used. Lactoferricin results suggest that interpolation/extrapolation might become more important for larger systems though.

Lactoferricin

As shown by the alanine systems, larger molecules require simulations at more frequencies to maintain high accuracy. Performing these simulations sequentially, as was done for the alanine n-peptides, can be impractical if a high number of CPUs is not available: running multiple simulations at the same time is then usually more wall-clock time efficient. In this case, two strategies can be employed. One can either use the extrapolation protocol to estimate the optimal frequencies from short test simulations and then extend these simulations in a parallel fashion, or one can start multiple simulations at prede-

termined frequencies (estimated from experience), calculate the free energy differences between these simulations using MBAR, and insert extra simulations to obtain the desired free energy spacing between consecutive simulations. This second approach was used here in order to compute the free energy difference between the two lactoferricin conformations. The free energy spacing used corresponds to strategy 4 of Table 1.2, which is the most aggressive.

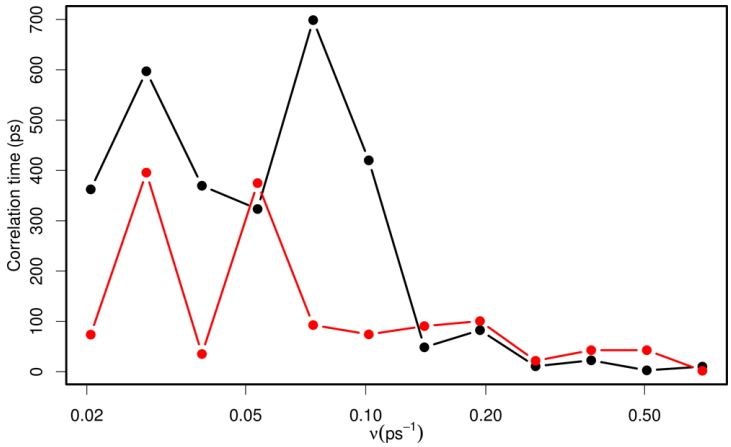


Figure 1.14: Correlation times for lactoferricin in the $\alpha+\beta$ conformation, using temperature replica exchange and a reference structure obtained from energy minimization (black) and clustering (red). Clustering had a modest effect on low frequency correlation times.

Because we observed that extrapolation and interpolation are more accurate for larger system, we expect this strategy to be very accurate for lactoferricin. Simulations were run until 1000 independent frames were obtained, except when using replica exchange, for which we used 100 frames per replica (500 frames total). Similar to the alanine systems, the correlation time of the confinement energy for lactoferricin was much higher at low frequencies than at high frequencies (Figure 1.14) because at low frequencies the harmonic restraints had a small impact on the overall dynamics of the molecule. For lactoferricin, this correlation time was as high as 10 ns for some frequencies, which would require extremely long simulations to obtain sufficient uncorrelated data. The long correlation times are likely a general feature for more complex biomolecules. To gain efficiency, it is therefore highly desirable to lower the correlation times at low frequencies. Multiple strategies can be employed: a careful choice of the reference structure, the use of additional restraints to eliminate subbasin hopping, or the use of replica exchange.

While any configuration that belongs to the basin of interest can be used as a reference structure, depending on the free energy landscape, different configurations may result in different correlation times. An energy-minimized configuration is a straightforward choice, but there is no guarantee that this structure is most representative of the free energy basin because the energy-minimized configuration corresponds to zero temperature and excludes entropic effects. A more representative configuration can easily be

obtained by performing rmsd-based clustering of an unrestrained MD trajectory, which may then yield lower correlation times (Figure 1.14). This procedure was used here to obtain the reference structures for both lactoferricin conformations. At low frequencies, the dynamics of the system are only slightly affected by the harmonic restraints. The correlation time of the confinement energy is therefore very close to the correlation time of the rmsdfor an unrestrained simulation of the same molecule. Restricting the motion of the molecule, by using additional restraints, will therefore lower the correlation time. However, the free energy might also be affected, depending on whether the definition of each basin is modified.

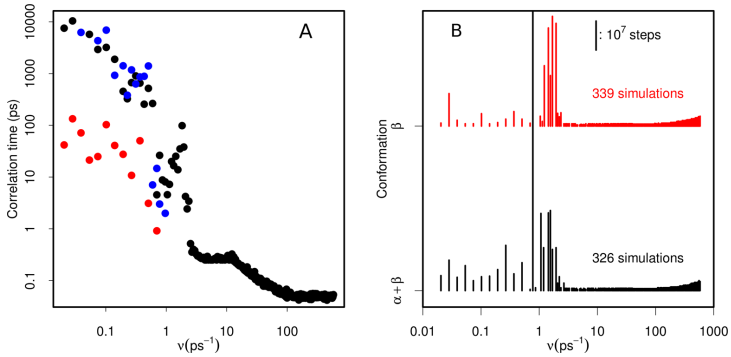


Figure 1.15: (A) Correlation times for lactoferricin in the β conformation, using a reference structure obtained from clustering (blue), the same reference with the addition of restraints (black), and temperature replica exchange in addition to the restraints (red).
(B) Number of MD steps used in the confinement simulations. Each vertical bar represents a simulation, and its length is linearly proportional to the number of steps.

Analysis of the $\alpha + \beta$ conformation trajectories showed that some backbone dihedral angles switched between two different values with a long correlation time. We therefore added flat-bottom dihedral angle restraints to confine these angles near the values of the reference state. Fraying was observed in simulations of the β-hairpin conformation, which was prevented by the addition of NOE restraints to maintain hydrogen bonding between residues 2 and 24. While these restraints indeed restricted the peptide to a single basin, the effect on the correlation times was marginal (Figure 1.15A).

Finally, correlation times can be broken and sampling can be enhanced by using temperature replica exchange. In replica exchange, multiple independent simulations are run at different temperatures, and at given time intervals, coordinates between the different simulations are swapped based on a criterion that preserves detailed balance [53]. Sampling is enhanced by the use of elevated temperatures, while correlations times are broken after swapping. In order not to unfold the peptide, we only used a small temperature range. For each simulation with a harmonic oscillator frequency lower than 0.72 ps^{-1}, replica exchange with five replicas at temperatures of 300, 312, 324, 337, and 350 K was used. This setup reduced the correlation times by a factor ∼10-100 (Figure 1.15A). In addition, extra efficiency was gained because data from all the temperatures could be

combined using MBAR.

Using these approaches, we obtained a free energy difference of 2.13 $\pm$ 0.05 kcal/mol in favor of the β-hairpin, which is indeed the stable form in solution. While many simulations were needed for each conformation (Figure 1.15B), most of these were at high frequencies and only required short simulation times (typically 10^5-10^7 steps per simulation). Most simulation steps were needed at frequencies between 0.7 and 2 ps^{-1}, but simulation times could likely have been reduced by also using replica exchange for these frequencies. When excluding interpolation/extrapolation from the optimization protocol (*i.e.*, using the 17 frequencies of the homogeneous setup, but optimizing friction coefficients, lengths, and time steps), a free energy difference of of 4.67 $\pm$ 4.19 kcal/mol was obtained, indicating the importance of the interpolation and extrapolation strategy for larger systems.

1.5.3 Discussion

We showed that the accuracy and efficiency of the confinement method can be greatly increased by the use of interpolation and extrapolation of the confinement energies and the careful consideration of correlation times. Interpolation can be used to obtain confinement energies at unsampled frequencies, which significantly reduces the discretization error. The free energy difference between two consecutive simulations must stay below a certain value for accurate interpolations; however, this difference can be increased for larger systems and at higher frequencies. Extrapolated free energy differences between simulated and unsimulated frequencies can also be used as a guide to select the optimal frequencies of the simulations. Cost and accuracy can be further optimized by basing the duration of each simulation on correlation times, costs can be decreased by increasing the friction coefficient at high frequencies, and accuracy can be increased by combining all data from multiple simulations. This setup proved to be efficient, as it led to proper estimations of conformation free energy differences for alanine n-peptides, with significantly increased accuracy (factor of 2-10) and greatly decreased computational costs (factor of 8-67) compared to homogeneous sampling. Additional techniques were used to speed up sampling for lactoferricin, a much more complex system with very long correlation times at low frequencies. Correlation times were significantly reduced by the use of temperature replica exchange (factor $\sim$10-100). They were also slightly reduced by using a reference structure obtained from rmsd-based clustering of unrestrained simulations and by the application of additional restraints to restrict the configurational space.

Our analysis revealed promising features for application of the confinement method to large systems. As illustrated by our alanine n-peptide and lactoferricin simulations, large systems will clearly take longer sampling times because their configurational space is larger. To maintain accuracy, the total number of simulations grows with system size, but most of these simulations are at high frequencies where sampling is relatively short (Figure 1.15B). Moreover, the growth in the number of simulations is partly counteracted by the fact that at a given accuracy, the spacing in free energy can be larger for larger sys-

tems. It is likely that large and complex systems will suffer from long correlation times at low frequencies, as also observed for lactoferricin. We showed, however, that sampling at low frequencies can be significantly reduced by temperature replica exchange and other strategies to reduce the correlation times. While treatment of large systems will be computationally expensive, our study provides effective ways by which costs and accuracy can be managed and controlled.

CHAPTER 2

Biasing non-equilibrium simulations

"Yes, sir! Prepare ship for light speed!"
"No-no-no, light speed is too slow!"
"Light speed too slow?"
"Yes, we're gonna have to go right to... Ludicrous speed!"

SPACEBALLS

In life sciences, most of the interesting phenomena is represented by non-equilibrium events. Biomolecules, in particular, perform most of their biological function in a non-equilibrium context; for example folding, the most fundamental transition in a protein, is a non-equilibrium process.
Both experimentally and computationally, the study of non-equilibrium is a challenging problem: in both cases we have to withstand the restrictions given by our instruments, whether they are due to physical limits in the experiment or in the lack of computational power.
During my PhD, I have implemented an approach based on the principle of Maximum Caliber to simulate the kinetic behavior of a biomolecule using MD, starting from experimental data. I applied this technique to the correction of approximated force fields used in MD, and to accelerate molecular dynamics simulations.

2.1 Introduction

The experimental investigation of the dynamic, time–dependent properties of biomolecules is usually much more difficult than the study of their equilibrium, time–independent features. The measurement of dynamic properties is affected by instrumental dead times, necessity of concentrated samples to compensate the lack of signal–accumulation time, experimental noise. Except for few types of single–molecule techniques, like those performed with optical tweezers or based on Förster resonance energy transfer, common biochemical experiments report the time course of properties averaged over very many molecules. Moreover, techniques suitable to detect kinetic properties, like fluorescence spectroscopy, circular dichroism and small–angle X-ray scattering (SAXS), usually depend on overall molecular features (like size, shape, internal symmetries, etc.), and are

difficult to be mapped to an atomic level.

Molecular dynamics (MD) simulations are a popular way to complement the information provided by kinetic experiments in solution. With the available computational power it is relatively easy to simulate multiple trajectories followed by small biomolecules on the time scale of hundreds of nanoseconds, and using tailor–made computers, to milliseconds[8]; in this way one can investigate some, but not all, the time scales associated with conformational changes in proteins and nucleic acids.

However, there are two major problems in using MD simulations to investigate biomolecular kinetics. First of all, portable force fields have been greatly refined over the last years, but show serious limitations when one attempts to reproduce quantitatively specific properties of well–defined molecules. In fact, while it was shown that standard force fields are all reasonably good at reproducing equilibrium properties of small proteins, they fail much more often with kinetic properties [15].

Furthermore, MD simulation are still limited to a range of time scales that does not cover all phenomena taking place in biomolecules. While several algorithms have been developed to enhance the equilibrium sampling of complex systems, little was done to foster the simulation of their dynamic trajectories, often restricting the attempts to the brute–force solution of the equations of motion, and thus relying only on the increasing power of nowadays processors. Among the algorithms developed to enhance the study of kinetic molecular properties are Markov State Models [71, 72], methods based on the identification of the most likely trajectories [73, 74, 75, 76] and milestoning [77].

Herein, we present a computational scheme to generate MD trajectories guided by time series resulting from kinetic experiments. The algorithm is based on the Principle of Maximum Caliber (pMaxCal), that is the dynamic version of the principle of maximum entropy [78] and that was so far used to study basic aspects of non-equilibrium systems [79, 80] and to model chemical reactions [81].

2.2 Principle of Maximum Entropy

A large part of the problems in statistical mechanics are represented by *inverse problems*, where one wants to obtain the microscopic parameters of the system (*e.g.* the interaction parameters in a Ising model) starting from the measurement of some meaningful observables. In this kind of problems, typically the number of known variables is lower with respect to the number of variables we want to find. One of the main issues in tackling this kind of problems is the possibility to add information that does not come from experimental data, introducing a *bias*. The principle of maximum entropy [82] is the most honest way to make a guess in our solution.
Let be ξ a random variable, which can assume the discrete value $\xi_i = 1, 2, ..., n$ with some probability distribution $p_i = \{p_1, p_2, ..., p_n\}$ with its natural normalization $\sum_i p_i = 1$, and we want to infere that probability distribution p_i, knowing only the expectation value of

a known function $f(\xi)$

$$\langle f(\xi)\rangle = \sum_{i=1}^{n} p_i f(\xi_i).$$

we need $n-2$ constraints more to obtain the exact evaluation of the probability distribution.
Given this problem, the principle of maximum entropy (pMaxEnt) [82] says that

> *From among all the probability distributions compatible with empirical data, the only unbiased distribution is the one with the highest information entropy.*

We then have to maximize the Shannon entropy, defined as

$$S(p) = -\sum_{i=1}^{n} p_i \log p_i. \tag{2.1}$$

Introducing the Lagrange multipliers λ and μ, the probability distribution results

$$p_i = \exp\left(-\lambda - \mu f(\xi_i)\right), \tag{2.2}$$

f and inserting this expression in the normalization and in the definition of $\langle f(\xi)\rangle$ we can obtain the values of the Lagrange multipliers from the constraint equations

$$\langle f(\xi)\rangle = -\frac{\partial}{\partial \mu} \log Z(\mu) \tag{2.3}$$

$$\lambda = \log Z(\mu) \tag{2.4}$$

where $Z(\mu)$ is the partition function defined as

$$Z(\mu) = \sum_{i=1}^{n} \exp\left(-\mu f(\xi_i)\right) \tag{2.5}$$

This approach can be generalized to any number m of functions $f(\xi)$ that constraints the probability distribution, redefining it as

$$p_i = \exp\left(-\lambda_0 - \lambda_1 f_1(\xi_i) - ... - \lambda_m f_m(\xi_i)\right), \tag{2.6}$$

and we obtain a system of $m+1$ equation which defines the Lagrange multipliers

$$\langle f_k(\xi)\rangle = -\frac{\partial}{\partial \lambda_k} \log Z \tag{2.7}$$

$$\lambda_0 = \log Z \tag{2.8}$$

for any $k \in [1, m]$, being the partition function

$$Z(\lambda_1, ..., \lambda_m) = \sum_{i=1}^{n} \exp\left(-\sum_{k=1}^{m} \lambda_k f_k(\xi_i)\right), \tag{2.9}$$

Choosing the distribution obtained with pMaxEnt, we obtain the most uninformative answer to our initial question. Any other distribution containing more information (thus less entropy), will introduce a bias in our guess, making the distribution obtained by pMaxEnt the only consistent choice.

2.3 Principle of Maximum Caliber

In the previous section we showed a technique to obtain unbiased probability distributions for equilibrium states maximizing their information entropy. It is possible to extend this approach, looking for probability distribution over paths in non-equilibrium conditions. This generalization, called principle of Maximum Caliber (pMaxCal), was introduced by Jaynes in 1980 [78].
We can define a path entropy for a non-equilibrium transition

$$S[p(\gamma)] = -\sum_{\gamma} p(\gamma) \log p(\gamma), \tag{2.10}$$

where the sum is over all the possible γ paths of the system, and $p(\gamma)$ is the probability to follow a particular path. If we have m different functions $f_{(k)}$, with $k = 1, ..., m$, and we know their ensemble averages along the path $f_{(k)}^{\text{exp}}$, it is possible to apply m linear constraints [83] in the form

$$\sum_{\gamma} p(\gamma) f_{(k)}(\gamma) - f_{(k)}^{\text{exp}} = 0. \tag{2.11}$$

Inserting the Lagrange multipliers λ_k to maximize our path entropy, we obtain a definition for the probability distributions

$$p(\gamma) = Z^{-1} \exp\left(\sum_{k=1}^{m} \lambda_k f_{(k)}(\gamma)\right), \tag{2.12}$$

where the dynamical partition function Z is

$$Z(\lambda_1, ..., \lambda_m) = \sum_{\gamma} \exp\left(\sum_{k=1}^{m} \lambda_k f_{(k)}(\gamma)\right). \tag{2.13}$$

To date, the maximum caliber approach was exploited in systems with discrete dynamical states on stochastic dynamics. We decided to apply the pMaxCal on non-equilibrium trajectories that came from equations of motion.

2.4 Force field correction and accelerating of non-equilibrium sampling

Similarly to what done in equilibrium statistical mechanics [82], it is possible to use Lagrange multipliers to constrain the optimization of $S[p]$ in such a way that the average $\sum_{\gamma} p(\gamma) f(\gamma)$ of some conformational property f of the system matches at each time any function of time, and in particular that which report the time course of some experimental data. The resulting distribution $p(\gamma)$, besides being in agreement with the experimental data, guarantees to minimize the amount of further, arbitrary information we provide to the model.
To implement the pMaxCal, we started from an approach which proved successful to correct force fields to reproduce known experimental data under equilibrium conditions

[84, 85], extending it to kinetic simulations. In brief, we performed replica MD simulations controlled by a suboptimal force field and a potential which drives the *average* of the conformational properties over all replicas to match the experimental data. We showed in some test cases that not only the very quantity which is biased in the simulation follows the biasing curve, but also quantities which are weakly correlated with it and their fluctuations follow the correct dynamics.
Implementing the pMaxCal can also be useful to accelerate MD simulations. The time scale associated with the biasing data can be different from that of the underlying unbiased simulation. Consequently, rescaling the time scale of the biasing data to lower nominal values can force the dynamics to take place on a faster nominal rate, basically modifying the time units of the biasing data. We showed that in this way we can accelerate the simulation of a factor between 10 and 100, maintaining unchanged (in the new time units) the dynamics of the biasing quantity and of the quantities perpendicular to it.

2.4.1 Theoretical framework

The goal of the algorithm we developed is to simulate the ensemble of kinetic trajectories that initiate from a given conformation (or ensemble of conformations), match as ensemble average the time course of a set of time–dependent experimental data and minimize the subjective bias we introduce into the system, maximizing the associated caliber. We define $\{\gamma\}$ as the set of trajectories of the system, where the trajectories are regarded as discrete set of conformations $\gamma \equiv \{r_0, r_1, ..., r_T\}$, as those usually generated in MD simulations. Kinetic experiments usually return time–dependent quantities that depend on the conformations visited along the trajectory. We define f_t^{exp} the time-course of the quantity monitored in the available experiment, indexed by the discrete time t; this can be one– or higher–dimensional. We assume to know the forward model associated with the experiment, that is the function $f(r_t)$ that maps a conformation r_t visited along a trajectory into the ideal result that the experiment would give if applied to an ensemble of identical conformations r_t.
The probability $p(\gamma)$ of a given trajectory that maximizes the caliber under the constrain

$$\sum_{\gamma} p(\gamma) f(r_t) = f_t^{\text{exp}} \tag{2.14}$$

that the average of the forward model over all possible trajectories is equal to the experimental value at each time t, and that the microscopic diffusion coefficient is D is

$$p(\gamma) = \frac{1}{Z_d} \exp\left[-\sum_t \left(\nu_t [r_{t+1} - r_t]^2 + \lambda_t f(r_t) \right) \right], \tag{2.15}$$

where Z_d is a dynamic partition function, λ_t and ν_t are the Lagrange multipliers that implement the experimental data and the diffusion coefficient, respectively. In principle, the numerical values of the Lagrange multipliers can be found from $\partial \log Z_d / \partial \lambda_t = f_t^{\text{exp}}$ and $\partial \log Z_d / \partial \nu_t = D$. However, these are implicit equations involving the sum Z_d over all trajectories, and are thus computationally useless.

Similarly to the case of equilibrium simulations [86], it can be shown (see the Appendix B for a detailed derivation) that the maximum–caliber distribution of trajectories of Eq. (2.15) is automatically sampled by coupled replica MD simulations, each replica (identified by greek letters) biased by a time–dependent potential

$$U(r^{\alpha},t) = \frac{k}{2}\left(\frac{1}{n}\sum_{\beta=1}^{n} f(r^{\beta}) - f_t^{\text{exp}}\right)^2 \tag{2.16}$$

that forces the *average* conformation, averaged over the n replicas, to match the experimental data in the limit of large k. The main drawback of this method is that the effective diffusion coefficient that the molecules experience is not the same as the nominal one D, this effect being more marked the more different is the generated trajectory from that one would generate in absence of an experimental bias.

2.4.2 Validation strategy

To test the validity of the replica–averaging scheme on molecular models, we performed some sand–box studies selecting some protein systems and defining for each of them two different interaction potentials. One of the two (U_{true}) is regarded as the "true" potential that controls the dynamics of the system in the experiment (which is unknown in real life) while the other (U_{approx}) is regarded as the approximated potential we known and we can use in real–life MD simulations. The two potentials are chosen in such a way that the system displays markedly different kinetic properties when interacting with each of them, but similar equilibrium properties.
We performed multiple simulations with U_{true} that serve as reference for the tests. We also defined some low–dimensional functions $f(r)$ of the conformation r of the system (the forward model) to mimic the experimental observables. Some of them (like the RMSD or the fraction of native contacts) are good approximations of the reaction coordinates of the system, while others (like the SAXS intensities) are closer to what one can obtain in real experiments. The time courses $f(r_t)$ of the forward model applied to the different trajectories are averaged together at each time to obtain the putative experimental data f_t^{exp}.
We then applied the pMaxCal to the system interacting with the potential U_{approx}, performing MD simulations of n replicas of the system biased by f_t^{exp} through the potential described in Eq. (2.16) (fig. 2.1). The dynamics of the biasing variable averaged over the replicas, of its fluctuations over the replicas and of other variables weakly coupled to it are then compared with the reference dynamics.

2.4.3 Computational Implementation

MD simulations are performed with Gromacs 4.5.7 [45, 87] coupled to Plumed 2 [47]. We implemented a `CALIBER` module into Plumed to apply the potential describe in Eq. (2.16). The simulations biased by SAXS data were carried out with the Plumed-ISDB module[88]. Simulations were performed with a Langevin integrator with $\gamma = 1$ ps^{-1} and a time-step of 0.1 fs.

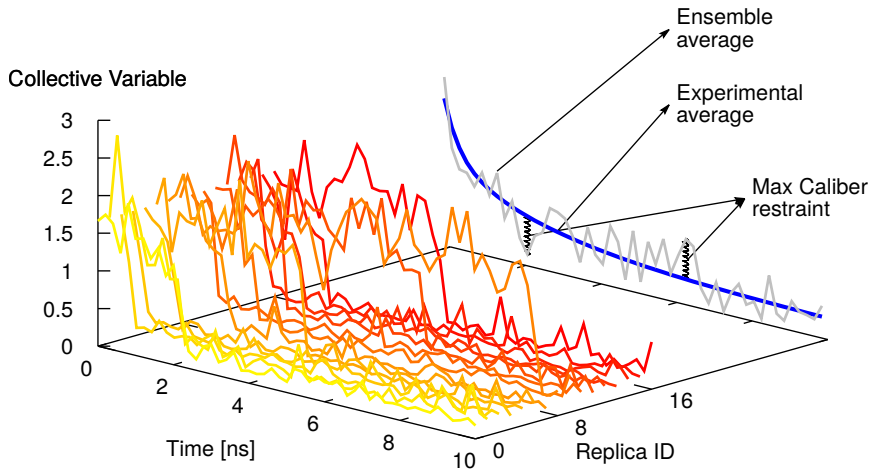


Figure 2.1: A sketch of the MD simulations, where n replicas of the system evolve in time coupled by Eq. (2.16). Lines colored in different hues of red and yellow represent the time evolution of the biasing variable in the various replicas. The grey line is the average of the biasing variable over the replicas. The biasing potential is an harmonic spring acting on this average, centred at the value of the experimental value (blue line) at the corresponding time.

We tested different quantities to bias the simulations, such as the root mean square deviation (RMSD) of the position of the C_α from those of the crystallographic conformation, and the fraction Q of native contacts, defined as [89]

$$Q(r) = \frac{1}{N} \sum_{i \neq j} \frac{1}{1 + \exp(\beta(r_{ij} - \lambda r_{ij}^0))}, \tag{2.17}$$

where N is the total number of pairs in the potential, $r_{ij} \equiv |r_i - r_j|$ is the distance between the i-th and j-th atom, r_{ij}^0 is the distance between the two atoms in the crystallographic structure, $\beta = 50$ nm^{-1} and $\lambda = 1.8$ are two switching parameters.
More realistic variables are the SAXS intensities, whose forward model is [90]

$$I(q) = \sum_i \sum_{j \neq i} f_i(q) f_j(q) \frac{\sin(q r_{ij})}{q r_{ij}}, \tag{2.18}$$

where q is the scattering vector, $f_k(q)$ is the atomic form factor of the k-th atom, and r_{ij} is the distance between the i-th and the j-th atom. Operatively, we selected 15 equispaced values of q from 0.02 Å to 0.4 Å and added 15 corresponding biasing terms to the interaction potentials in the form of Eq. (2.16).

The values of the harmonic constant k are chosen to be as large as possible, compatibly with the time step of the simulation.

2.4.4 Results

GB1 hairpin

The first test to verify the ability of replica–averaged simulations to correct the dynamics of a molecular system were carried out on an all–atom model of the second hairpin of protein G B1 domain (residues 41–65, pdb code 1PGB [91]) *in vacuo*. We built two different structure-based Gō potentials [92], these potentials stabilize by definition a reference conformation. The choice of *in vacuo* condition is a huge simplification of the system kinetics. In fact, we start assuming any potential "wrong", because the main point of our technique is the correction of a force field which is, by definition, approximated, using real data. The potential U_{head} is obtained rescaling the interactions between the pairs of atoms of a factor which is proportional to the distance from the turn of the hairpin, from 1.5 for pairs close to the turn, to 0.5 for pairs close to the termini (Fig. 2.2). In this way, we expect to induce a folding dynamics that nucleates from the turn. The other potential U_{tail} is obtained inverting the scaling factors to weaken by a factor 0.5 the interactions close to the turn and strengthen by 1.5 those close to the termini, in order to induce a different folding dynamics while keeping comparable stability between the folded and the unfolded state. In Fig. 2.3 is shown the heat capacity for both potentials showing a comparable melting temperature. The dynamics of the hairpin interacting with both potentials was simulated starting from an unfolded conformation at $T = 50$K (note that in a Gō model energy units, and consequently temperature units, are arbitrary), generating 500 folding trajectories for each of them. In Fig. 2.4 is displayed the average value $\overline{Q}(t)$ of the fractio of native contacts as a function of time, which result qualitatively different for the two systems.

The test consisted in biasing the system interacting with U_{head} (regarded as U_{approx}) to display the dynamics of the system interacting with U_{tail} (regarded as U_{true}). For this purpose, we used the function $\overline{Q}(t)$ of the latter as putative experimental data $f^{\text{exp}}(t)$, and simulated the dynamics of the hairpin with the potential $U_{\text{tail}} + U_{bias}$, varying the number of replicas from $n = 4$ to $n = 128$ and using a harmonic constant for U_{bias} equal to $k = 2.5 \cdot 10^4 \cdot n$. The behavior of $\overline{Q}(t)$ for the resulting simulations is essentially indistinguishable from that of the simulations we wanted to target for any n, indicating that the two dynamics are identical at least when projected over the space defined by the biasing variable.

To compare in more detail the biased to the target trajectories, we plotted in the left panel of Fig. 2.5 the dynamics of other unbiased conformational variables of the system which, although not orthogonal to Q, report different features of the system. Also in this case, the biased curves match reasonably well the target dynamics, quite independently on the number of replicas (cf. also the χ^2 displayed in Figs. 2.6 and 2.7).

In the right panel of Fig. 2.5 we plotted the fluctuations, defined as the standard deviation of these quantities over the replicas as a function of time. In spite of their noisy behavior, the bias is able to push the system interacting with U_{head} to display fluctuations

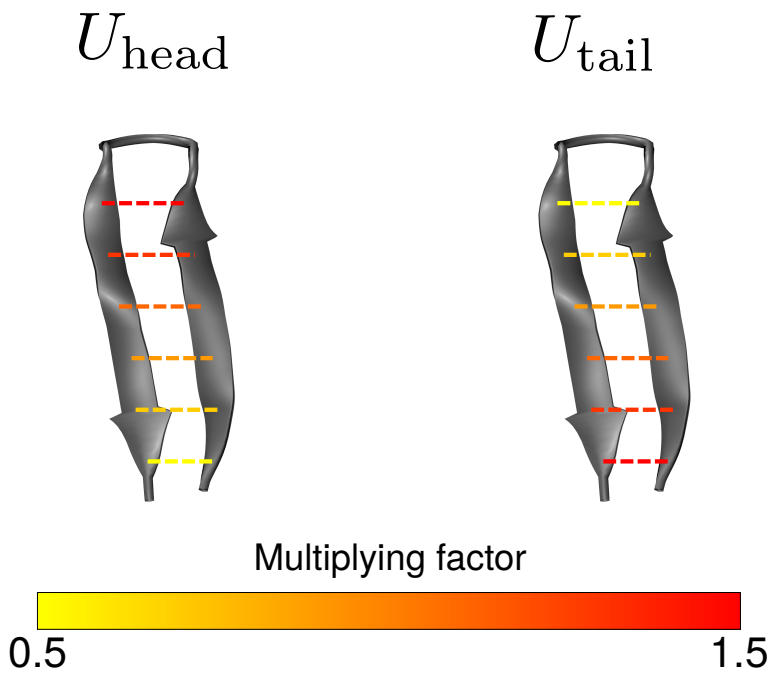


Figure 2.2: Cartoon representation of protein G GB1 hairpin (PDB: 1PGB, rr. 41-56). The colored lines show how the potential was modified.

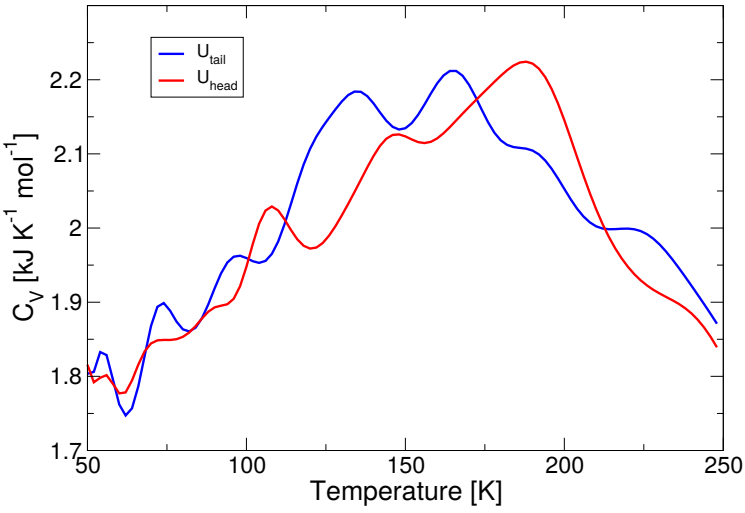


Figure 2.3: Heat capacity vs. temperature for GB1 hairpin under U_{tail} (blue) and U_{head} (red). There is a single broad transition between 120 and 200 K in both the potentials.

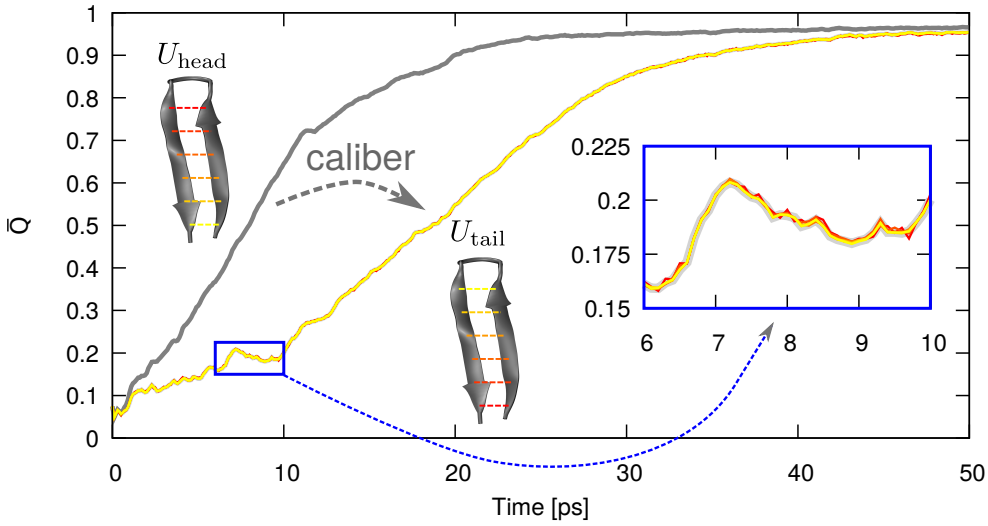


Figure 2.4: Average fraction of native contacts ($\overline{Q}$) in function of time for unbiased U_{head} (dark grey), unbiased U_{tail} (light grey, covered by caliber-restrained simulations), and caliber-restrained simulations from U_{head} to U_{tail}, from 4 (red) to 128 replicas (yellow) in color scale.

similar to those of the system interacting with U_{tail}. Also for them there is not a clear behavior as a function of the number n of replicas, except for the fact that $n = 4$ gives an agreement much worse than the case for larger n (see also Figs.). Finally, as a control, similar results are obtained by using U_{head} as putative "true" potential and biasing the system interacting with U_{tail} to follow its dynamics (see Appendix B).

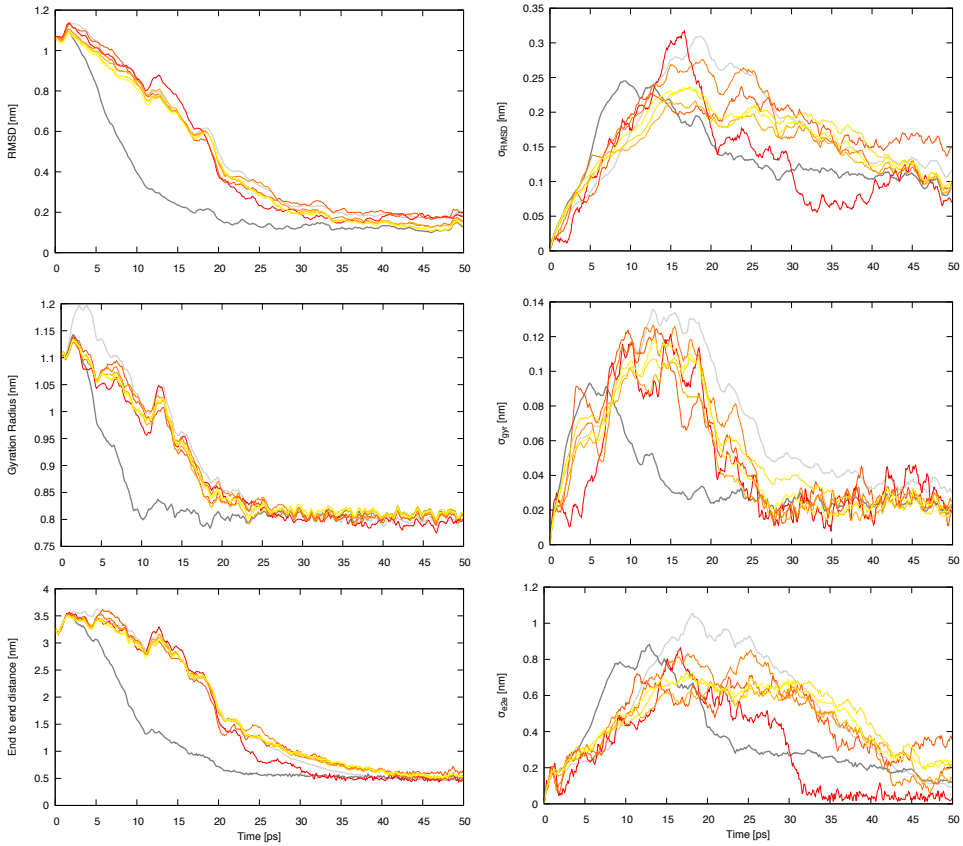


Figure 2.5: To the left, the dynamics of the RMSD (top), of the gyration radius (middle) and of the end–to–end distance (bottom) of the hairpin. The dark–grey line indicate the dynamics generated with U_{head}, the light–grey line is the target dynamics generated with U_{tail} and the colored lines are the simulations performed with U_{head} and biased using from 4 (red) to 128 replicas (yellow). To the right, the standard deviations over the replicas of the same quantities.

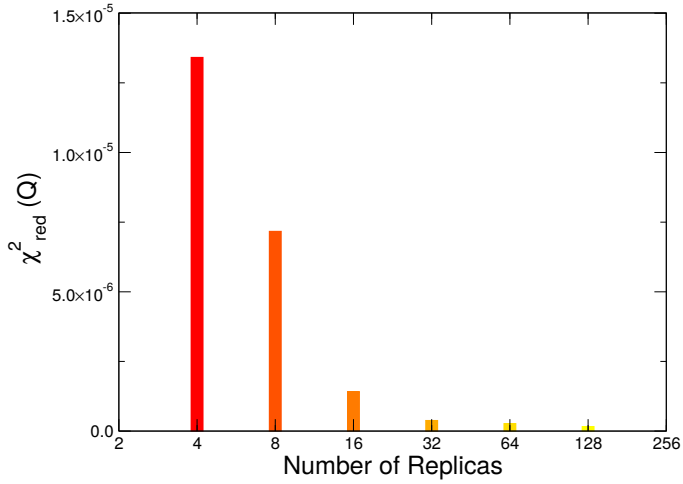


Figure 2.6: The χ^2_{red}, defined as $\chi^2_{\text{red}} = \frac{1}{N}\sum_t \frac{(\overline{Q}^{\text{bias}(t)} - \overline{Q}^{\text{exp}}(t))^2}{\overline{Q}^{\text{exp}(t)}}$, between the points of the function $\overline{Q}$ of the system interacting with U_{head} and biased in simulations with a variable number of replicas and that of the system interacting with U_{tail}, regarded as the "true" system.

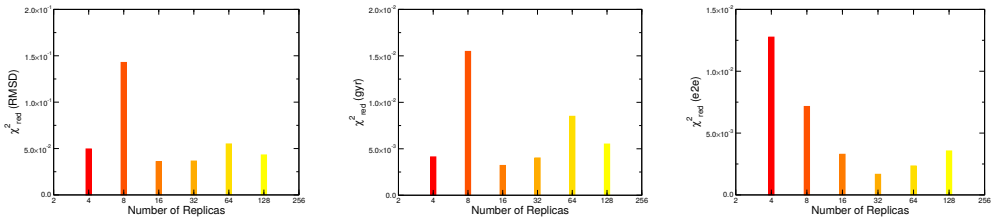


Figure 2.7: The χ^2_{red} (defined as in Fig. 2.6) for the averages displayed in Fig. 2.5.

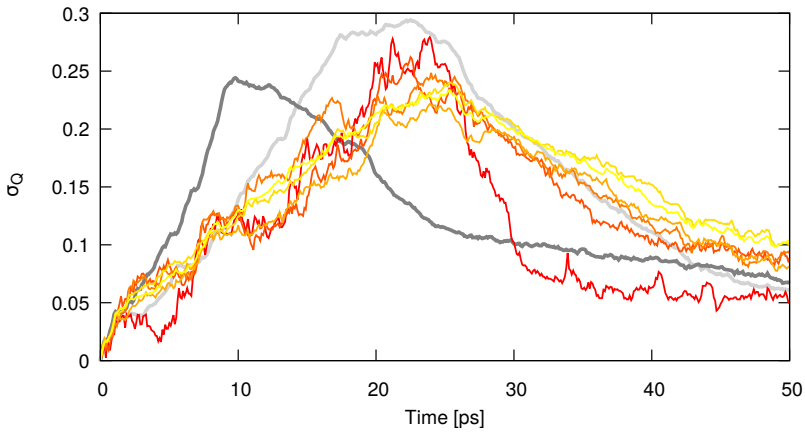


Figure 2.8: Same as Fig. 2.4, but displaying the standard deviation of Q over the replicas.

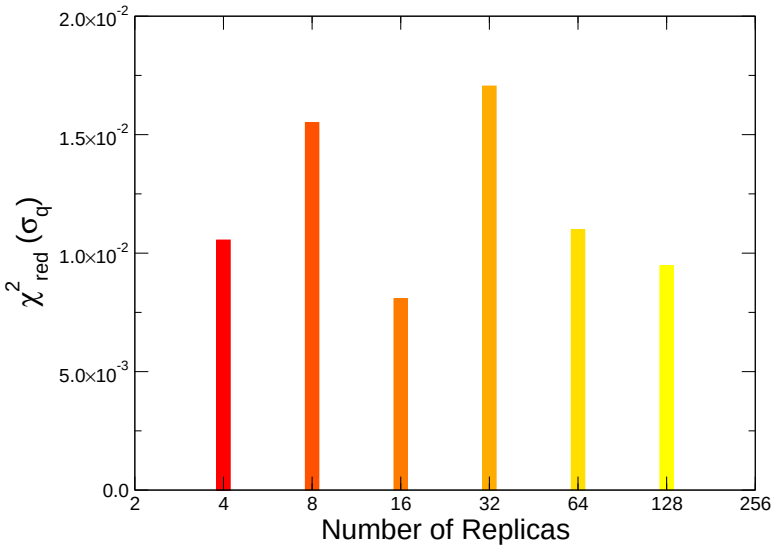


Figure 2.9: The $\chi^2_{\rm red}$ (defined as in Fig. 2.6) for the curves displayed in Fig. 2.8.

Protein G - Q-biased

Given the ability of pMaxCal replica-simulation to correct the dynamics of a simple system, we challenged the algorithm with a larger system. We defined two models for the full protein G B1 domain. The first described by the standard Gō potential $U_{\text{Gō}}$ and the second in which the Gō potential is modified strengthening the intra-helix interactions by a factor of 2 (we shall label the latter as U_α). The equilibrium properties of the two models are similar (Fig. 2.10), but their folding dynamics, starting from a disordered conformation, is different (cf. the shapes of $\overline{Q}$ displayed as dark–grey and light–grey curves in Fig. 2.11). A simulation, carried out over 32 replicas, biasing the molecule interacting with the potential U_α to follow the dynamics of the mean fraction of native contacts $\overline{Q}$ of the molecule interacting with $U_{\text{Gō}}$ is almost indistinguishable from the dynamics of its target simulation when comparing the biasing variable (cf. the red curve in Fig. 2.11). Importantly, the dynamics of other conformational variables, like the total RMSD, the gyration radius, the RMSD restricted to the two β–hairpins and to the whole β–sheet are very similar to those of the target system (see Fig. 2.12).

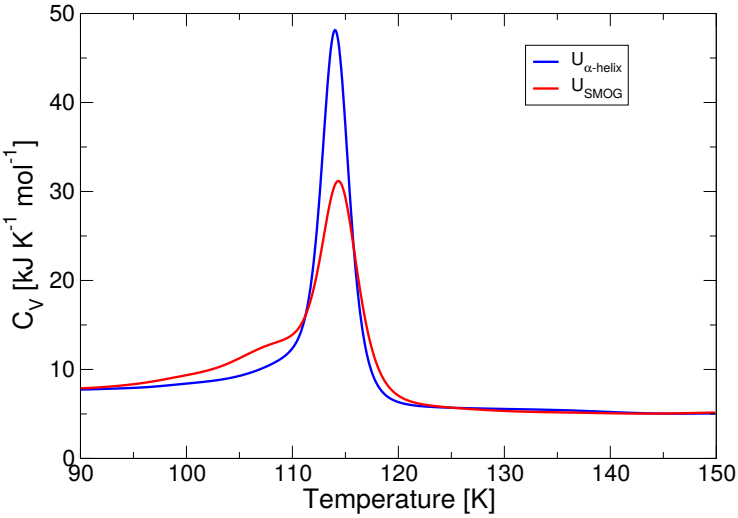


Figure 2.10: Heat capacity vs. temperature for protein G under U_α (blue) and U_{SMOG} (red). There is a single folding transition between 110 and 120 K in both the potentials.

In principle, the biasing procedure based on the pMaxCal can be used, besides correcting a potential, to speed up simultaneously the simulation. To reach this goal it is enough to modify the time scale at which the experimental data change as a function of time, rescaling its time units to a smaller value by a factor λ_s. In other words, one "pretends" that the data evolve following a dynamics that is faster than in reality, restoring *a pos-*

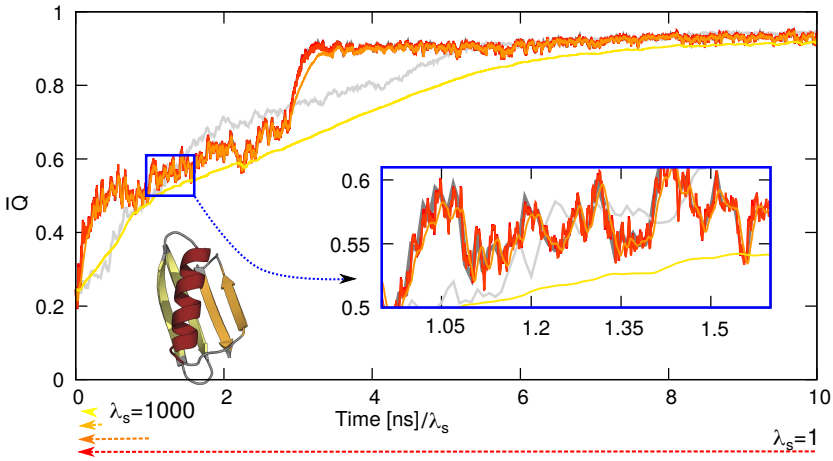


Figure 2.11: Average fraction of native contacts $\overline{Q}$ as function of time calculated from the unbiased simulations of protein G interacting with $U_{\text{G}\tilde{\text{o}}}$ (dark grey), interacting with U_α (light grey), and calculated from biased simulations of the molecule interacting with U_α on 32 replicas, with a time compression of $\lambda_s = 1$ (red), $\lambda_s = 10$ (dark orange), $\lambda_s = 100$ (light orange), and $\lambda_s = 1000$ (yellow). Simulations are performed at $T = 106$K starting from a conformation denatured at 400K.

teriori the correct time units. This feature can be critical for future applications of the algorithm because a number of non-equilibrium experiments report on time-scales that are often longer than those that is possible to reach with MD.

To test the correctness and the efficiency of this scheme, we repeated the above simulations compressing time of the target "experimental"-data by factors $\lambda_s = 10$, $\lambda_s = 100$ and $\lambda_s = 1000$. In Figs. 2.11 and 2.12 we compare the dynamics of the biasing coordinate and of other coordinates, respectively, with that of the target system interacting with $U_{\text{G}\tilde{\text{o}}}$, rescaling back the time axis to the original time scale to allow a clear comparison. A time compression of a factor $\lambda_s = 10$ gives results which are essentially identical to the case without time compression. With a time compression of a factor $\lambda_s = 100$ the qualitative agreement is still good, but the two curves are no longer perfectly overlapping, while a factor $\lambda_s = 1000$ gives a dynamics which is completely different from both the unbiased and the target–molecule ones (see Fig 2.13).

To compare the behavior of the system kinetics under the different potentials, we performed a tICA analysis [93, 94] on unbiased and biased simulations, obtaining a qualitative estimate of the relaxation times of the tICA-derived slow collective variables (cf. Fig S18 in the Supplementary Materials). The two original potentials $U_{\text{G}\tilde{\text{o}}}$ and U_α show significantly different relaxation times, and the caliber-biased simulation with $\lambda_s = 1$ dis-

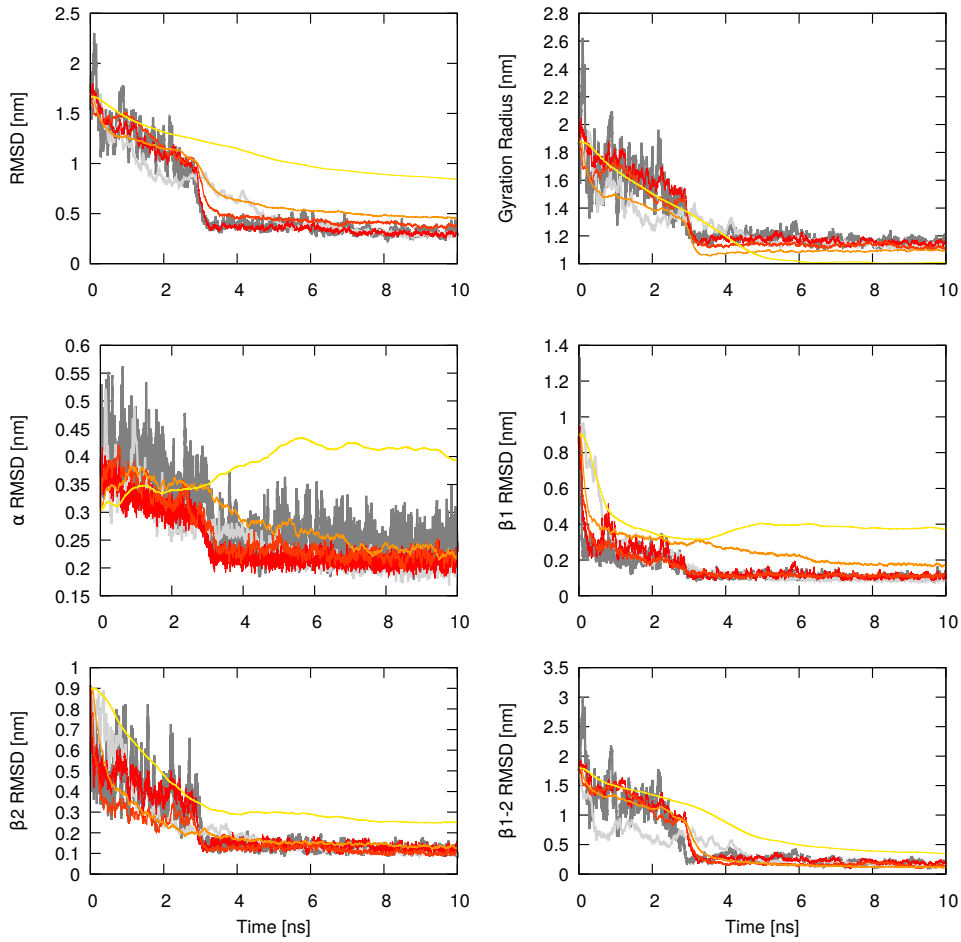


Figure 2.12: The dynamics of the average C_α-RMSD (top left), gyration radius (top right), α RMSD (center left), β-hairpin–1 RMSD (center right), β-hairpin–2 RMSD (bottom left), and the RMSD of the interface between β-hairpins 1-2 (bottom right) for the same simulations (and with the same color code) as those displayed in Fig. 2.11.

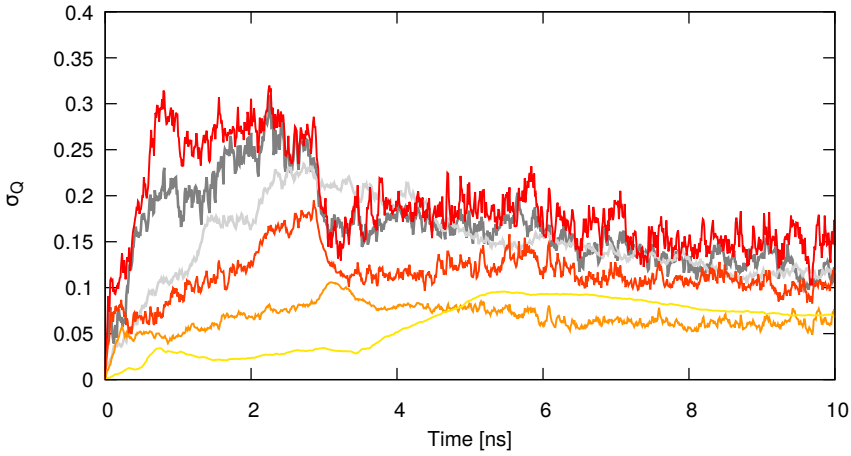


Figure 2.13: The fluctuations over replicas of Q in the simulation described in Fig. 2.11.

plays a good agreement with the target potential relaxation times, demonstrating once again that replica-averaged simulations can be used to include time dependent data in MD. As expected, with the increase of λ_s the system shows a speed up in all the slow variables. The worse behavior of the simulations with $\lambda_s = 100$ and 1000 can be explained considering the system diffusion time, which is in the order of 1 ps: With a too strong time acceleration, the resulting relaxation time is in the order of the ps, and thus the system cannot follow the caliber bias.

Protein G - SAXS-biased

All the former simulations have been biased to follow a quantity that is not experimentally accessible, this to test if at least in principle pMaxCal replica-simulations could work. To test our approach in the case of more realistic biasing quantities, we used the same two models described in the previous section and used SAXS intensities as the source of information. Indeed, SAXS is routinely use to follow conformational changes of biomolecules[95]. We calculated the SAXS intensities from the target system interacting with $U_{\mathrm{G\tilde{o}}}$ and used the dynamics of the SAXS intensities at 15 equispaced values of the scattering vector as putative experimental data (see Sect. 2.4.3 for details) to bias the model interacting with U_α.

The dynamics of the SAXS obtained from the target simulations applying Eq. (2.18) is displayed in the upper panel of Fig. 2.14, while in the lower panel it is shown the dynamics of the SAXS intensities at the values of Q (0.08Å^{-1}, 0.25Å^{-1} and 0.35Å^{-1}), chosen as an example. For these q and for all the others which are not shown here, the biased dynamics can follow perfectly well the dynamics of the target system. In Fig. 2.15 it is shown the dynamics of other conformational variables not used for biasing the

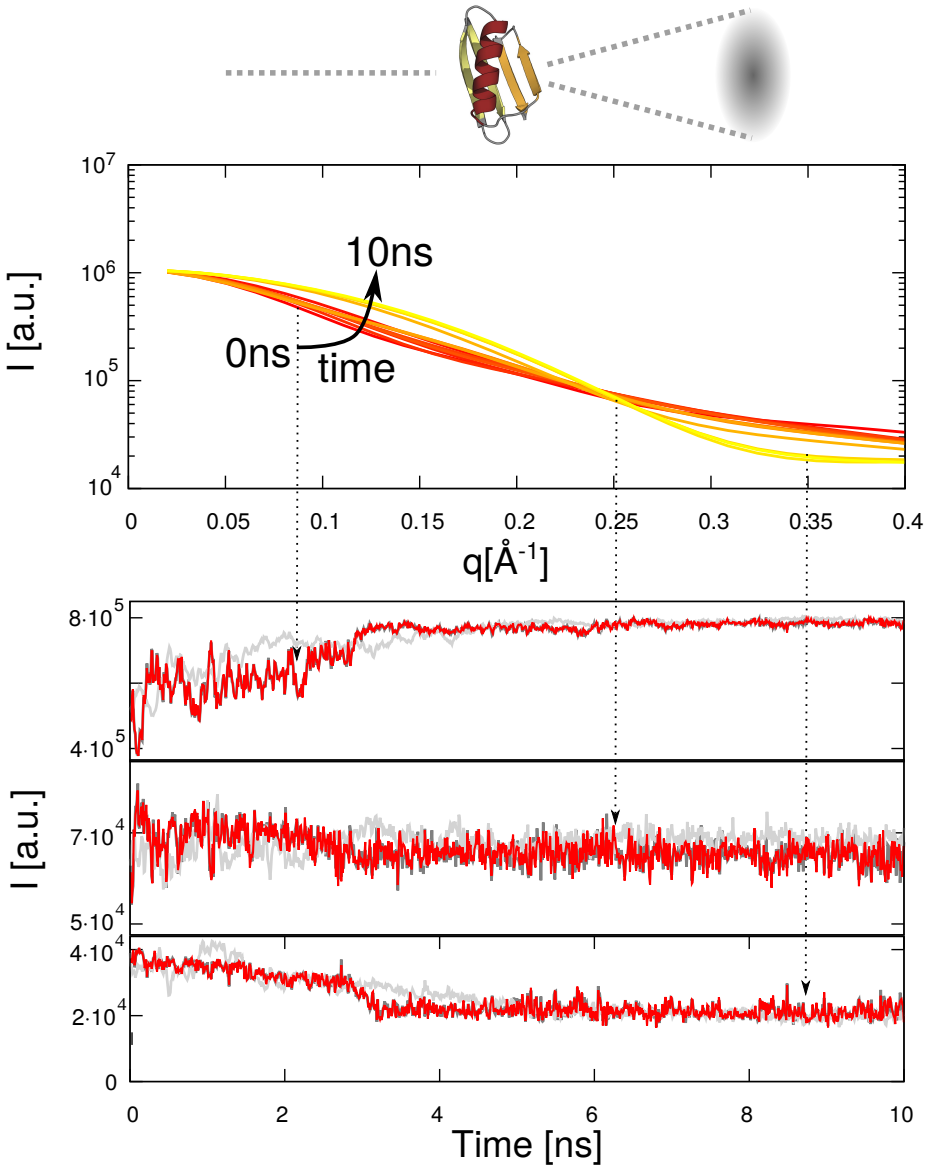


Figure 2.14: In the upper panel, the dynamics of the SAXS spectrum simulated for the model of protein G interacting with $U_{G\delta}$. In the lower panel, the dynamics of the SAXS intensities at $q = 0.08\text{Å}^{-1}$, at $q = 0.25\text{Å}^{-1}$ and $q = 0.35\text{Å}^{-1}$. The light grey curve is the unbiased dynamics, the dark-grey curve is the target dynamics and the red curve is the biased dynamics.

simulation that are also in good agreement with the target dynamics. Finally, also the tICA-derived slow variables relaxation times are in good agreement with the ones of the unbiased target potential (Fig. 2.16).

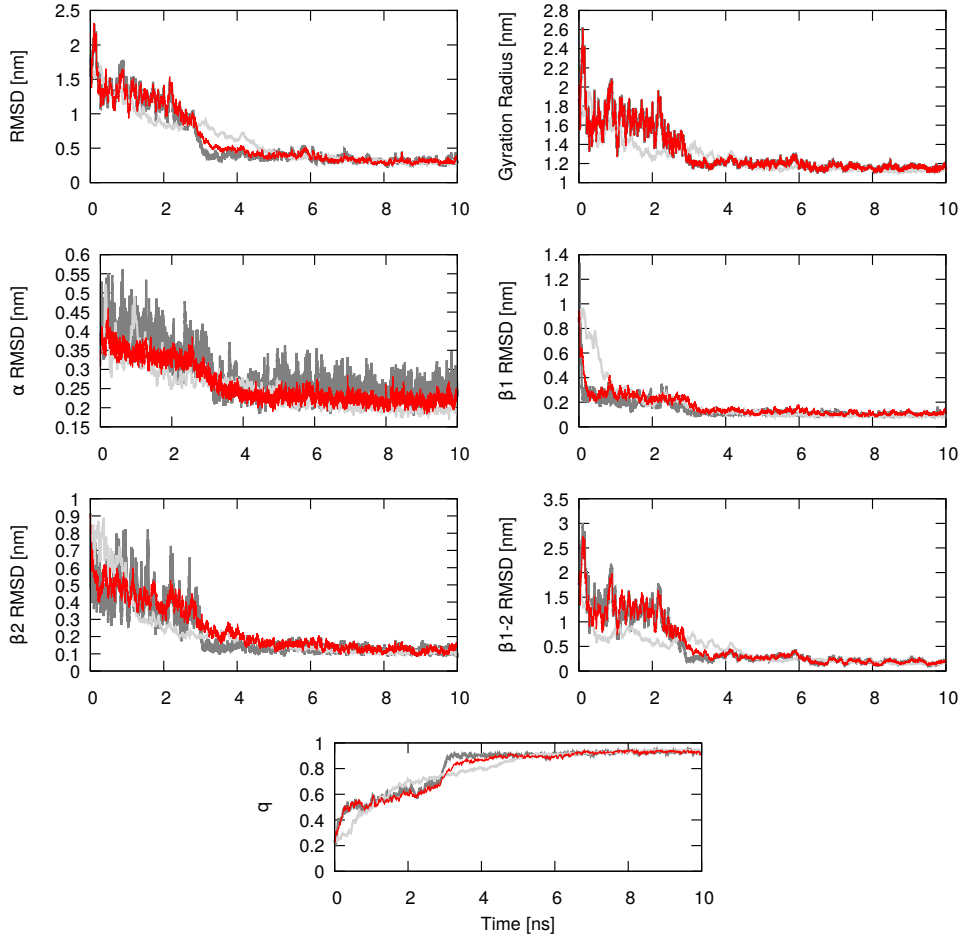


Figure 2.15: The dynamics of some conformational coordinates of protein G obtained biasing by means of the SAXS intensities. The light grey curves are obtained from the unbiased simulations of the model interacting with U_{α}, the dark grey come from the target model interacting with $U_{\text{G\=o}}$ and the red lines from the biased simulations.

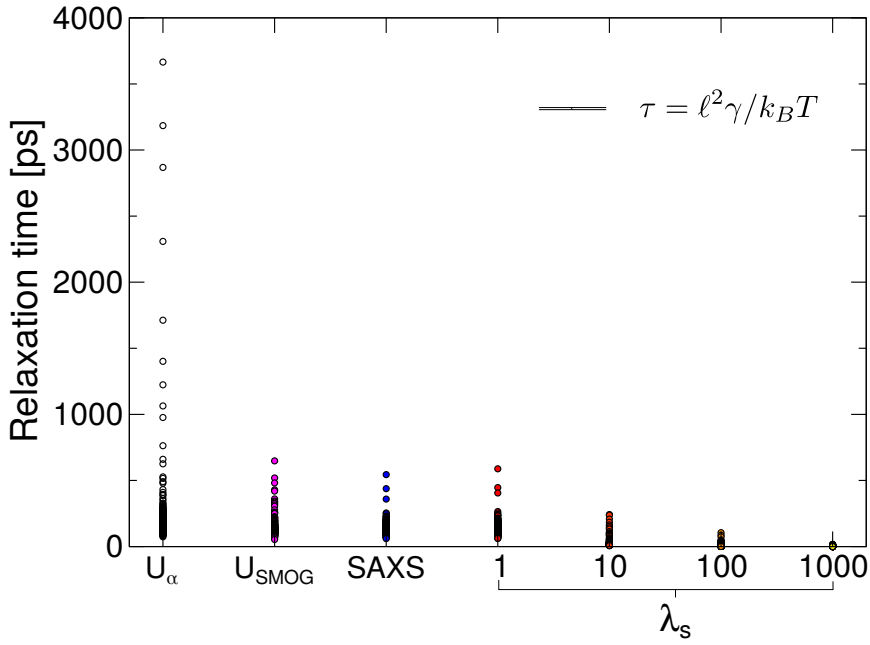


Figure 2.16: Relaxation times for all the variables obtained by tICA analysis on the C_α positions. The original unbiased potential U_α (white dots) shows a longer relaxation time with respect to the unbiased U_{SMOG} potential (magenta dots). Without time acceleration, both the caliber-biased simulations shows a good agreement in relaxation time (blue dots for the SAXS-biased one and red dot for the Q-biased one) with the target potential. Varying the acceleration parameter λ_s, we obtain a decrease in relaxation times, although it is not sufficient to reconstruct the correct kinetics for $\lambda_s = 100, 1000$. To corroborate our hypotesis, we show the typical diffusion times of the system on the plot.

2.4.5 Discussion

The quality of molecular mechanics force–fields is generally improving[96, 97], but these improvements, even if significant, are limited by the difficult of training them on systems with size comparable to the one of interest (e.g. small to medium sized proteins) and by the approximations that are intrinsic in the functional form. To overcome these limitations, system dependent solutions have been developed to model equilibrium ensembles of structures based on experimental data either by reweighing a posteriori an MD simulation or by adding a bias to the force field[88]. Among these, replica-averaged simulations[98], based on the maximum entropy principle[86] and recently extended to include a Bayesian treatment of the errors[99], have been particularly successful[100, 88].

In the present work, we used replica-averaged MD to perform out–of–equilibrium simulations, by introducing a time-dependent bias. We showed that simulating multiple replicas of a system and coupling them with an harmonic potential acting on the average over the replicas of some conformational variable and centred around the time course of the corresponding experimental observable, is a way to implement the principle of maximum caliber. The equivalence is formally showed if the simulation is driven only by the biasing potential, while it remains to be studied what is the relation with the pMaxEnt when the simulation is driven both by a force field and by the biasing potential, furthermore a consideration of the errors is also currently missing. Nonetheless, using some test–box cases, we could show that biasing the folding of model proteins, we can recover a target dynamics not only for the biasing variablfe, but also for other standard conformational variables of the protein and, importantly, their fluctuations. Importantly, we obtained good results not only biasing an ideal function, but also simulating the outcome of a SAXS experiment. In this case, the lack of the actual reaction coordinate of the system was compensated by the fact that all length scales of the protein were under control at the same time.

A powerful byproduct of the algorithm is that it allows speeding up MD simulations, simply rescaling time in the reference time course used to bias the simulation. In this way one can easily gain a factor of 10 to 100 in computer time, being thus able to carry out simulations of unprecedented duration. This is particularly relevant given the fact that most real–time experiments (H/D exchange[101], real-time NMR[102] as well as time-resolved SAXS/WAXS[103, 95]) are performed on time scales that are longer than those usually accessible by MD (i.e. on the order of milliseconds). In this case the choice of the biasing variable plays an important role to ensure the realism of the resulting trajectories. The biasing scheme (independently on its equivalence with the pMaxCal) affects the time–dependent probability distribution of conformations only along the direction defined by the biasing variable. No direct effect is exerted in the directions perpendicular to it; these are only controlled by the molecular force field. Consequently, if one chooses a biasing variable which is correlated with the slowly–varying reaction coordinate of the system, the macroscopic dynamic will be correct, and this will strongly constrain the faster degrees of freedom perpendicular to it. The dynamics of these fast variables depends on the force field, but they are constrained to the subspace identified by a given value of the reaction coordinate. On the other hand, if one biases the dynamic with a

fast–varying variable, perpendicular to the actual reaction coordinate, the macroscopic dynamics of the system will only rely on the force field. Now the biasing mechanism will force the fast–varying variable to follow an apparently–correct dynamics but on a possibly wrong subspace. This poses a theoretical limit to the time–compression factor λ_s one can use to speed–up MD simulations. In fact, if one biases the reaction coordinate and simultaneously accelerates too much its dynamics, the next–to–slowest degree of freedom perpendicular to the reaction coordinate, whose motional time scale is not affected by λ_s, will be promoted to new reaction coordinate, and its dynamics will only depend on the force field.

In conclusion with the present work we introduced a hybrid scheme for the integration of time–resolved experimental data into molecular dynamics simulation. This with the aim of improving at the same time MD accuracy and efficiency in generating ensembles of trajectories corresponding to experimentally accessible processes.

CHAPTER 3

Peptide and protein design for immunology

"E poi prendo proteine e le tiro, le tiro, le tiro, finché faccio opere che rendano questo mondo migliore, come anziani imbottiti di tritolo."

DAW, Brullonulla

Novel immunological tools for efficient diagnosis and treatment of emerging infections are an urgent necessity. From the point of view of an efficient medical treatment, the rise in the number of drug-resistance pathogens [104] and the lack of effective and/or new antibiotics [105] poses major challenges for effective management of infections. In diagnostics, a large number of pathogens cannot be detected with existing tools, and consequently those disease are underreported and/or misdiagnosed.

A way to tackle this problem is the use of peptide-based diagnostic tools and vaccines that use engineered proteins or peptides[106]. At the molecular level, the recognition and the immune response against pathogens is driven by protein-protein interactions and the design (or partial redesign) of those molecules can provide new solution to detect pathogenic infection and to trigger the correct immune response in an individual.

During my PhD I have worked on immodiagnostic tools design (Section 3.5) and on a new unsupervised algorithm aimed to redesigning part of a protein to enhance its immunoreacting activity with a vaccine-oriented purpose (Sections 3.6 and 3.7).

3.1 Introduction

The immune system is a complex machinery that uses cells and different types of molecules to defend the organism against pathogenic invasions[107]. Inside the immune system, we can find many complex activities, such as recognition tasks, learning, and memory storage of previous infections.

One of the main actors in the immune system are lymphocites, a class of white blood cells. Those cells are created in the bone marrow, are are transported along the body via the blood stream. Furthermore, they can exit the blood stream passing trough the

capillaries and the body, searching for foreign cells (recognized binding some exposed proteins of the pathogen called *antigens*) and then return back via the lymphatic system. Lymphocites are subdivided in two classes:

B cells: this kind of lymphocites produces *antibodies*, which are large proteins that bind to the target pathogen (see below). Antibodies can block the life cycle of the pathogen or enhance the activity of the T cells against them.

T cells: this kind of lymphocytes are subdivided in two classes: T helper cell[1], which marks the foreign cell with a marker called CD4 that promotes the production of antibodies in the B cells, and the cytotoxic T cells, which are responsible to physically kill the pathogen via a marker called CD8 that activates the NK (natural killer) cells, also secreted by cytotoxic T cells.

Both types of lymphocites have a pattern recognition mechanism formed by receptor molecules on their membrane which can identify antigens. In the case of B cells those molecules are immunoglobulins (antibodies), while for T cells are called T Cell Receptors (TCR).

Operatively, the antigen-recognition is performed by the immune system at molecular level: the receptor has a complementary 3D structure with respect to a particular binding site (called *epitope*) of the antigen. The interaction is usually due to van der Waals interaction, electrostatic and hydrogen bonds.

The molecules responsible of the first recognition of foreign objects are the antibodies. At molecular level, an antibody is a big protein tetramer formed by two simmetric dimers. Every dimer is composed by two different chains called heavy chain and light chain ("heavy" and "light" refers to the relative molecular weigth of those proteins). Antibodies have a distinctive "Y" shape (see Figure 3.1)
The binding site of the antibodies is located at the end of the Y arms, in a pocket formed by the two chains. This region is called "hypervariable loop" and corresponds to the quasi-random part in the antibody sequence, which optimally adapts to a specific antigen (see below). The rest of the antibody, that forms the stable structural part, is uniform in structure; this characteristic guarantees the recognition of the antibody by the other cells of the immune system, like NK cells, macrophages and other lymphocites for further processing of the antigen and completion of the immune response.

Each lymphocyte presents 10^4 to 10^5 random receptors on its surface. B cells, when stimulated, produce a soluble version of their surface antibodies: antibodies are extremely specific and adapt to every possible change in the antigen structure and chemical properties. This leads to a clear problem: if the organism is under attack, the lymphocites which can recognize the pathogen antigen will be in a small number. To get around this issue, the immune system applies the so called *clonal selection*: only the lymphocites that

[1]This kind of cells are the main target of HIV, the retrovirus responsible of Acquired Immune Deficiency Syndrome (AIDS).

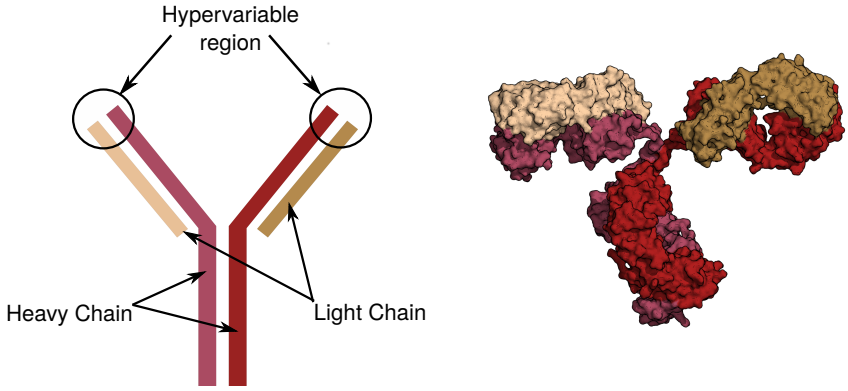


Figure 3.1: Schematic and surface representation of an antibody. On the left we can see the structure of an antibody: two symmetric dimers formed by a light chain (yellow) and a heavy chain (red); the binding site for antigens, called hypervariable region for its capacity to change its in function of the target antigen, is highlighed. On the right we can see the surface representation of an antibody, with the same colors of the left sketch.

are activated by antigen recognition proliferate, creating a huge number of clones of the original "random" lymphocite and trigger an immune response. The possibility to activate a response with clonal selection is however not sufficient. During the life of an individual, the selection and differentiation of the antibodies repertoire reflects the interaction with the environment. Evolution favors the creation of a learning system that biases the creation of random sequences in antibodies, with the aim to be more efficient in immune response. For example, if a particular antigen has been detected in the past, the immune system responds to subsequent encounters with a larger number of lymphocites presenting the corresponding antibody, amplifying the reaction. This behavior is called *secondary immune response*, and is the cellular basis of immune system memory and, of course, the principal motivation for vaccination.

The practice of vaccination begun with a scientific approach in 1796, when Edward Jenner, noticing the absence of deadly smallpox cases in milkmaids, inoculated pus from a cowpox lesion to a 8-year-old boy, having seen the similarity between the two diseases. After 3 weeks, Jenner inoculated material coming from a smallpox lesion in the same patient, which showed no clinical consequences [108]. Jenner, after this first result, performed a clinical trial on 16 additional cases formalizing the variolation process, which consisted in the inoculation of the material coming from a cowpox lesion to the arm of a healthy patient using a lancet.
After the discovery of microbes in late 19^{th} century, Louis Pasteur estabilished the fundamental principles of rational design of vaccines, called the "3I" approach:

Isolate the microorganism responsible of the disease,

Inactivate the microorganism by means of physical of chemical intervention,

Inoculate the inactivated microorganism in the patient.

With this approach, he developed a rabies vaccine in 1885. The same rules were used for more than a century, developing vaccines for a then-deadly diseases like diphtheria and tethanus (Ramon and Descombey, 1920s), poliomyelitis (Salk and subsequently Sabin in 1955), and measles (Enders and Pebbles in 1963). All those vaccines contained attenuated living pathogens, which could lead in a small minority of cases, to the disease.

By the end of the 20th century, the introduction of recombinant DNA technologies in live-attenuated vaccines greatly enhanced the reliability of vaccines and reduced in a significant way the appearance of adverse effects in vaccinated patients [110].
At this point, the original 3I approach was starting to show its limitations. For example, some pathogens can not be grown *in vitro*, some others have an intracellular cycle and thus their infection is blocked by T cells rather than humoral response, and some others present antigenic hypervariability (the most famous –and problematic– case is HIV).

The change of paradigm needed to tackle some of those pathogen came in 1995, when the first entire genome of an organism was published [111]. Knowing all the proteins expressed by an organism, it was possible for the first time to rationally design a vaccine without the need of the entire microorganism, using only the antigens responsible of the immunitary response. This new approach is called Reverse Vaccinology (RV) [112] and the first target was the Meningococcus B, resposible of the 50% of the meningococcal meningitis worldwide [113]. The Pasteur paradigm did not work in this case because its capsular polysaccharide is identical to a human self-antigen, whereas the bacterial surface proteins are extremely variable [17]. From the genome, 600 different possibly antigenic surface proteins were identified, and a small number of candidates with high conservation and sequence similarity in all the strains of the pathogen ($\sim$20) was selected *in silico* and then tested in mice [18, 114]. In 2014, the vaccine (under the commercial name of BEXSERO^{TM}) was approved for human use in the United States, Canada, Australia, and European Union.

3.2 Structural Vaccinology

A further step forward was represented by the inclusion of the structural information on antigens from crystallography experiments in the context of Reverse Vaccinology during the first decade of 21st century, which led to the introduction of Structural Vaccinology (SV) [115]. The enhancement in experimerimental techniques in the field of crystallography made possible a new revolution in vaccinology: the possibility to express and crystallize a protein permits to study, besides sequence information (exploited by means of bioinformatics tecniques), also the physicochemical properties of a target antigen.

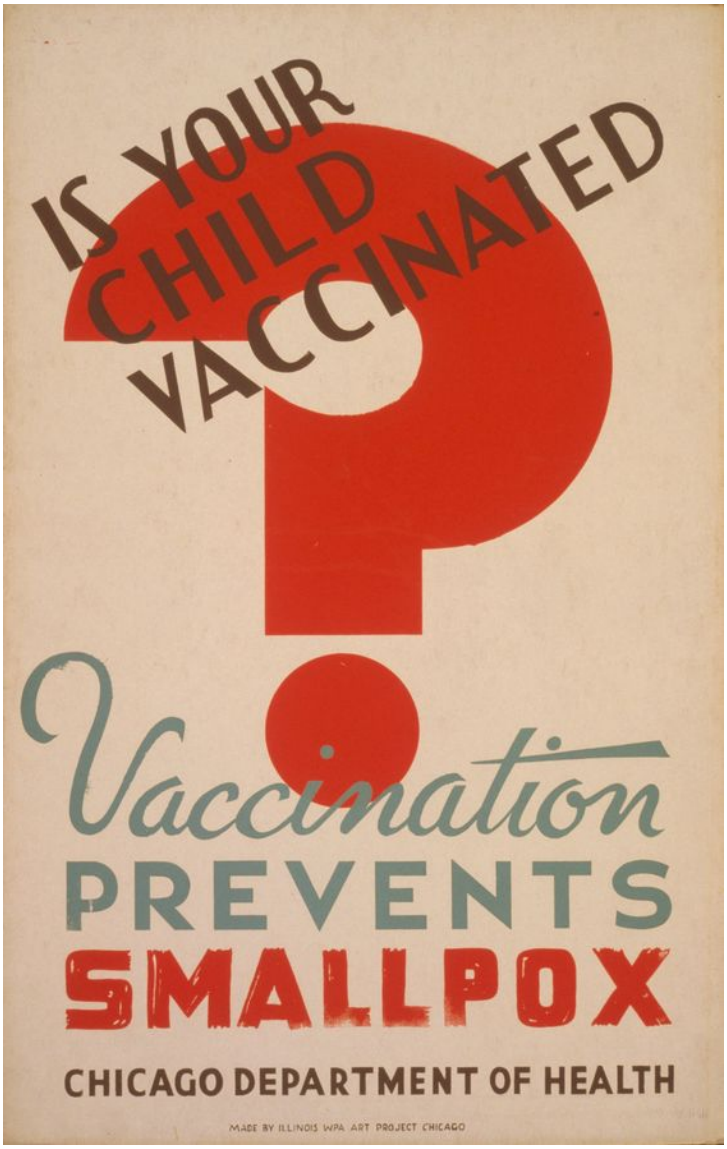


Figure 3.2: 1941 poster of Chicago Department of Health encouraging smallpox vaccination. Smallpox was the first disease successfully eradicated thanks to mass vaccination: the last natural case was reported in 1977. The global eradication was certified by WHO in 1980 [109].

The interacting part of antigens, the epitopes, interact with the antibodies thanks to their particular 3D conformation and physico-chemical properties: in principle, it is not necessary to use the whole immunogenic protein to elicit an immune response, but it is sufficient to present the interacting part alone [116]. To date, the design of a peptide-only vaccine has been not so successful, though. The mechanism of *in vivo* protease-mediated peptide degradation, combined with the difficulties to mantain the peptide in a correct 3D structure in the patient strongly limited the efficacy of this approach.

To overcome the limitation of epitope presentation *in vivo*, a series of different solution were presented in the last 10 years, where the two most promising are nanoparticles carriers and epitope grafting.
A correct epitope presentation can be achieved covering inorganic nanoparticles (gold nanoparticles) with peptides or inserting them on vescicular nanoparticles [117] (like liposomes, already used in commercialy available vaccines for influenza and hepatitis A [118]). Furthermore, it is possible to present multiple antigens on a single carrier, eliciting in this way a broader immune response or, like in case of influenza, obtain a response versus different strain of the same pathogen.
The epitope grafting approach consists in inserting in a non-immunogenic protein, or in a protein already carrying other reactive epitopes, the recognized part of the antigen of a pathogen of interest. The interaction of those proteins with a foreign functional motif (or multiple motifs) can elicit an immune response and can be adapted to address antigenic variability in different pathogenic species. This approach will be discussed with our automated grafted technique in section 3.6.

Another important application of the knowledge of the antibody-antigen interaction at molecular level is related to diagnostics. The presence of antibodies in patient serum is the molecular signature of a pathogen. As we explained earlier, if a host has been attacked by an external organism, a immune reaction starts, and the specific antibodies are spread through the bloodstream. This brings us to the *serodiagnostics*: it is possible to diagnose the presence of a pathogen in an indirect way, *i.e.* detecting the presence of the specific antibodies related to the pathogen. To do so, we need to identify the antibody-binding region of the antigens (the epitopes) and use them as a probe in experiments involving serum of patients (see section 3.3).

3.3 Peptides for immunodiagnostics

In 21st century, despite the tremendous advance in pharmaceutics, pathogen infections remain one of the major cause of death and economic loss (especially –but not only– in developing countries) [119]. In particular, the emergence of drug-resistance [120] and novel pathogens [121] represents a challenge to the modern biological and medical research.

In this context, diagnostics plays a key role to minimize the effect of a novel or drug-resistant disease. The state of the art is represented by enzyme immune assays (EIA),

carried out in the ELISA (enzyme-linked immunosorbent assay) tests, that uses parts of antigens bound to a surface that "capture" antibodies from a patient. Despite its efficiency, ELISA tests show some limitations: is impossible to bind an entire antigen to the surface, and this can rise a problem of epitope presentation. Furthermore, it is possible to use only a limited number of probes, with the result that a "complete" screen on antigens can have a high cost, making it not sustainable, especially in developing countries.

One of the most promising alternatives to ELISA test is the use of microarray technology, which would allow a huge increase in number of probes in a single test and make the diagnostics more statistically significant [122, 123]. A microarray is a plastic planar slide on a solid substrate in glass or silicon. On this surface, a molecule (in our case an entire antibody or an epitope) can be chemically linked and, in particular, tens of different probes can be displayed on the same chip, allowing a parallel, high-throughput screening. When a microarray is covered with probes, it has to be washed with a serum that contains a primary antibody, that recognizes and binds the probes. After this step, a secondary antibody with a fluorescent label which recognizes the first one is applied. Measuring the fluorescence given by the latter, one can have a quantitative evaluation on the binding affinity of the probe with respect to the serum.
The main issue for the widespread use of this technique is the cost due to the expression of the recombinant proteins and their conservation on the microarray.
A possible solution to overcome this limitation is represented by peptide microarrays, where entire libraries of antigenic peptides are bound on the microarray [124]. Peptides synthesis and/or expression is cheaper, and the mantainance of a peptide bound on a microarray is way simpler with respect to a whole antigen.
One of the major advantages of peptides is the possibility to insert and evaluate small chemical modification in the probes to enhance binding zone presentation. Moreover, the use of cross-reactive epitopes provides the possibility to design a single immunological test that can be used in very diverse context.

3.4 Epitope prediction

In both the two delivering techniques discussed above, we need to know the position of the epitope in the immunogenic protein. Experimentally, it is extremely difficult to identify the antibody-binding site of an antigen as a whole, and computational prediction methods play an important role in epitope search.
Given the exponential growth of computational power and immunology databases (like IEDB [125], AntiJen [126] and IMGT [127]), now it is possible to follow a statistical approach, like in epitope predictors based on sequence information (MULTIPRED [128], TEPITOPE [129] and ProPred [130]), or on physicochemical properties of the epitopic region, like hydrophobicity, flexibility and charge (ElliPro [131] and SEPPA [132]).

In our work, we used a structural prediction method called MLCE (Matrix of the Local Coupling Energies) [133]. This technique is based on a simple assumption: all the residues that guarantee structural stability cannot be involved in inter-protein interac-

tions, like antigen-antibody recognition.
The algorithm analyzes the interaction between all the amino acids via a MM-GBSA calculation considering only non-bonded interaction (van der Waals, electrostatic interactions, solvent), obtaining, for a protein composed by N residues, a M_{ij} interaction matrix of dimension $N \times N$. This symmetric matrix can be diagonalized and reconstructed using the resulting eigenvalues and eigenvectors

$$M_{ij} = \sum_{\alpha=1}^{N} \lambda_\alpha v_i^\alpha v_j^\alpha$$

Where λ_α is the α-th eigenvalue and v_k^α are the k-th components of the corresponding eigenvector. Sorting and labeling the eigenvalues from the most negative to the most positive, we can assume [134, 135] that the first eigenvector labeled in this way contains most of the information regarding the stabilizing interactions between amino acids. In this way, we can obtain an approximated interaction matrix $\tilde{M}_{ij}$ that does not contain noise and ignores weaker interactions

$$M_{ij} \simeq \tilde{M}_{ij} = \lambda_1 v_i^1 v_j^1$$

Knowing the structure of the antigen of interest, it is possible to compute a boolean contact matrix C_{ij} defining a distance threshold to consider two amino acids in contact. The Hadamard product between the approximated and the contact matrices returns the Matrix of the Local Coupling Energies L_{ij}

$$L_{ij} = C_{ij} \cdot \tilde{M}_{ij}$$

which contains only the non-bonded interaction between residues which are close to each other.
Remembering our assumption, the residues with a strong interaction will be ones responsible of structure stability, while the residues with weak interaction will be prone to interact with an external protein [133]. The non-stabilized amino acids identified in this way are labeled as part of a candidate epitopic zone. This technique has been recently made available to the public on a webserver [136].

3.5 Design of a probe for *Burkholderia* family diagnostics

In our work, efforts have focused on the analysis of epitope conservation between *B. pseudomallei* (Bp) and *B. cenocepacia* (Bc). The former is the etiologic agent of melioidosis, a severe endemic disease in Southeast Asia and an emerging threat in Australia, on the Indian subcontinent, and in South America. Melioidosis can cause septicemia and organ failure, with a high mortality rate; treatment with antibiotics is largely ineffective because of multidrug resistance [137, 138]. *B. cenocepacia* is an opportunistic pathogen that colonizes the respiratory apparatus of cystic fibrosis (CF) patients, causing lung infections that often have fatal consequences [139, 140]. Genomic similarities between the two bacteria raise the possibility of designing cross-reactive epitopes for the simultaneous diagnosis of *Burkholderia* species. Such designed molecules, once shown

to be immunoreactive in serological tests, may be further developed into components of protective vaccines, thus opening new and long-sought perspectives for the therapeutic treatment of Burkholderia infections.

In this context, detailed structural information on the antigens and epitopes of the pathogen offers prime opportunities to engineer biomolecules with specific immunological and recognition properties. In previous studies, starting from the X-ray structure of the peptidoglycan-associated lipoprotein from B. pseudomallei (Pal_{Bp}), an epitope peptide (BpEp3 comprising Pal_{Bp} residues 72 to 91) was predicted and designed using *in silico* methods [141]. BpEp3 showed improved immunological properties with respect to the initial recombinant antigen as well as cross-reactivity and significant diagnostic performances for *B. cenocepacia* infections in CF patients, demonstrating that rational epitope engineering can be an effective strategy for delivering better immunoreagents [142, 143]. The conformational flexibility of BpEp3 was found to be a key property affecting its diagnostic potential; a cyclic, more rigid form of the peptide performed better [144].

In our work, we targeted the Pal antigen from *B. cenocepacia* (Pal_{Bc}) as the basis for comparative and structure- based epitope discovery, design, and immunodiagnostic studies. Pal_{Bc} and Pal_{Bp} are highly conserved proteins (sequence identity of 84%) with only 27/170 differing amino acids. Four such residue differences fall in the region corresponding to Ep3 in Pal_{Bp}, with the most significant substitution from a structural point of view being an Ala to Pro replacement (residue 81). The presence of Pro81 raises the possibility that the region corresponding to BpEp3 in Bc may influence the dynamic states of the epitope when synthesized and used as an isolated peptide, as previously carried out on BpEp3 [142]. Such residue substitution could ultimately provide different diagnostic properties for BpEp3 and BcEp3, despite two otherwise very similar epitope sequences, thus providing new insights into the design requirements for efficient immunoreactive probes.

To investigate the above issues, we solved the crystal structure of Pal_{Bc} at 1.8 Åresolution for subsequent in silico epitope predictions and epitope design (Figure 3.3). Epitope predictions carried out on the crystal structure identified the region of Pal_{Bc} corresponding to BpEp3 as a potential candidate for epitope design. We synthesized BcEp3 as a free peptide and carried out comparative functional and structural analyses with its Bp counterpart. Specifically, we probed sera from individuals affected by Bp and Bc infections on a microarray platform, comparing the immunodiagnostic performances of the BcEp3 and BpEp3 epitopes both in the context of their corresponding recombinant full antigens (Pal_{Bc} and Pal_{Bp}) and as isolated synthetic peptides.

Our analyses suggest that sequence- and structure-based conservation of the full-length antigen alone may not be enough to correlate with immunodiagnostic properties. Attention must be paid to the conformational dynamics of the epitope sequence and to the main ensembles that this may determine.

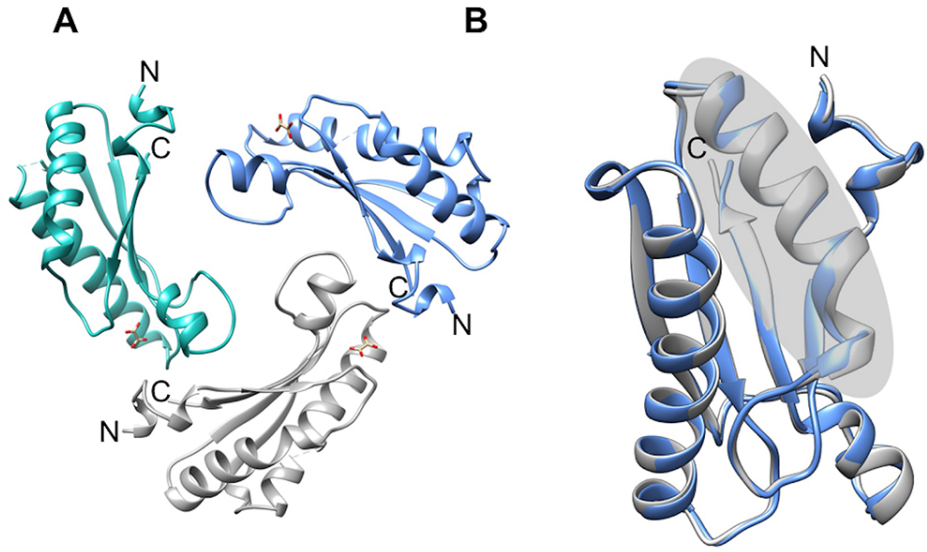


Figure 3.3: 3D structure of Pal_{Bc}. (A) The three Pal_{Bc} chains present in the asymmetric unit are shown by gray (chain A), blue (chain B), and green ribbons (chain C). (B) Secondary structure representation of superposed 3D structures of Pal_{Bc} chain A (blue) and Pal_{Bp} chain A (gray, PDB entry 4B5C [142]). The Ep3 α-helix is indicated by shading.

3.5.1 Results

Pal_{Bc} crystal structure

Pal_{Bc} (residues 19-170) was expressed as an N-terminal His-tagged fusion protein and crystallized at the Structural Biology Lab of University of Milan. A single crystal was used to collect data at a resolution of 1.8 Å, and the structure of Pal_{Bc} was solved by molecular replacement. Three identical Pal_{Bc} chains (A-C; RMSD values of 0.24-0.42 Åover the traced C_α chains) were present in the crystal asymmetric unit; such a trimeric arrangement is not proposed to be biologically relevant because there are no significantly extended interaction surfaces between chains (Figure 3.3A). Electron density was visible for residues 52 to 170 (for chains A and B) and for residues 52 to 169 (chain C) but was absent for the first 62 (29 pertaining to the vector) N-terminal residues, indicating that this region is flexible and lost to the solvent.

The overall 3D structure conforms to the canonical Pal $\alpha - \beta$ sandwich fold, organized in helix-strand-helix topology (Figure 3.3B). Structural comparisons were made between the A chains of Pal_{Bp} and Pal_{Bc} using the Superpose module available under the CCP4 suite [145]. As expected, the two structures are very similar (RMSD 0.43 Å; sequence identity 84%) (Figure 3.3B). Among the 27 residues that differ between the primary structures of Pal_{Bp} and Pal_{Bc}, 4 (positions 76, 78, 81, and 83) are located in the region corresponding to the highly antigenic BpEp3; at residue 81, an Ala to Pro replacement occurs in the region encompassing an α-helix in BpEp3. Despite the conformational restraints

posed by Pro residues that typically cause distortions in polypeptide chains, such as kinks in α-helices, BpEp3 and BcEp3 maintain the same backbone conformation as in the full-length proteins (Figure 3.3B).
There is one oxalate ion bound to each Pal_{Bc} chain. Hydrogen bonds are formed between backbone and sidechain atoms contributed from residues D71, D105, R107, N113, and R120. As reported in the literature, an acetate ion was bound to each Pal Bp chain, interacting with equivalent residues present in Pal_{Bc}. The cavity that houses these anions is the proposed peptidoglycan-binding pocket [142, 146].

Structure-Based Computational Epitope Mapping of Pal_{Bc} Epitopes

We previously identified B-cell epitope BpEp3 using the matrix of local coupling energies (MLCE) computational epitope prediction method applied to the crystal structure of Pal_{Bp} [142, 133]. In peptide form, BpEp3 elicited bactericidal antibodies (Abs), triggered Ab-dependent agglutination, and was preferentially recognized by recovery melioidosis IgGs, in comparison to healthy controls and seropositive individuals, thus implying potential diagnostic and therapeutic applications.

T-cell epitopes conserved between Bc and Bp have been observed for the flagellin antigen [147]. Furthermore, B-cell epitope conservation was also demonstrated in a separate microarray study that showed that a panel of Bp epitope peptides could specifically detect Bc infections in CF patients [143]. These results suggested the possibility of rationally designing cross-reactive probes for the diagnosis of *Burkholderia* species infections in general.

In this context, to investigate the potential effects of sequence and structural properties on immunoreactivity, an analogous strategy using the MLCE epitope prediction approach was applied to the Pal_{Bc} crystal structure [142, 133]. Following in silico analysis, two main epitope regions were identified largely overlapping with the previously determined Pal_{Bp} putative epitopes; one is equivalent to BpEp3, and here labeled BcEp3, and the second region maps a conformational epitope (Figure 3.4), strongly suggesting that Pal_{Bc} and Pal_{Bp} share immunoreactive regions.
The BcEp3 was then synthesized as isolated peptide and tested for immunitary response.

Probing the Human Antibody Response to BpEp3 and BcEp3 in *Burkholderia*-Affected Individuals

As previously mentioned, the BpEp3 isolated peptide was found to be cross-reactive and immunodiagnostic for Bc infections in CF patients [143]. Here we investigate the immunoreactivity of the Bc counterpart against Bc patient sera and assess its cross-reactivity against melioidosis patient sera samples. To this aim, we compared the serodiagnostic capability of the two epitopes to identify individuals affected by Bc and Bp infections, both in the context of Pal full antigens and as-synthesized free peptides. All the experimental work in this section was performed at the AμS lab at ICRM-CNR.

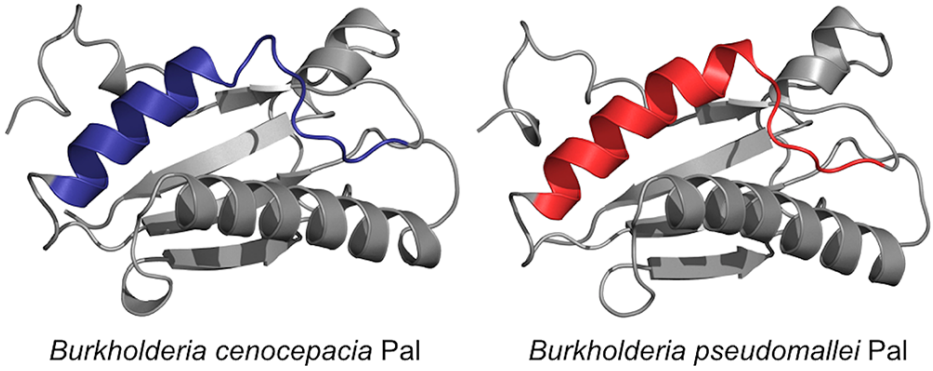


Figure 3.4: Location of predicted Pal_{Bc} epitopes. The location of predicted reactive epitopes on the two Pal structures, from *B. cenocepacia* (left) and *B. pseudomallei* (right), is highlighted in color.

A protein microarray displaying recombinant Pal_{Bc} and Pal_{Bp} antigens was probed with 12 serum samples from Bc-positive CF patients, diagnosed by microbiological culture and MALDI-TOF spectrometry, and with 20 melioidosis patient serum samples (10 healthy seropositive sera and 10 recovered melioidosis cases, as judged by indirect hemagglutinin assay (IHA) Ab titers (Khon Kaen University and Srinakarin Hospital, Thailand) [143]. As a control group, seronegative serum samples from healthy donors were used (12 samples for the Bc patients and 10 sera samples for melioidosis patients). The antigen-specific IgG content in each serum was evaluated by fluorescence detection using an anti-human IgG labeled with the Cy3 dye. The ability of each antigen to distinguish controls and patients was evaluated by performing the unpaired t test (significant if p values are < 0.05) on the protein-specific fluorescence signals detected in the groups (Figure 3.5).
As expected, Pal Bp and Pal Bc effectively captured patient Abs elicited by Bc and Bp infections and, as predicted by sequence and structure conservation between the two protein antigens, showed significant cross-reactivity. Specifically, Bc-affected individuals (upper panel) were correctly diagnosed (p values <0.01) by both antigens, whereas recovered melioidosis patients (lower panel) were successfully distinguished from control individuals. A slightly higher sensitivity of the directly related Pal Bp antigen was observed, thus allowing the detection of low Ab titers in seropositive individuals. The immunodiagnostic capability of corresponding isolated peptides BpEp3 and BcEp3 were assessed in analogous serological tests on the same set of serum samples. The peptides were chemoselectively immobilized by a terminal cysteine residue via specific thiol addition to maleimides on microarray chips coated with a polymer bearing maleimido groups [144]. The ability of each peptide to distinguish controls and patients was evaluated performing the unpaired t-test (significant if p values were <0.05) on the peptide-specific fluorescence signals detected in the groups. A summary of the diagnostic performance for each peptide is detailed by representations of the t-test analysis (Figure 3.6).
The BpEp3 peptide was found to be significantly serodiagnostic and cross-reactive for both Bp and Bc infections, being able to detect Bc-positive individuals versus healthy controls and allowing one to distinguish Bp seropositive and recovery individuals from

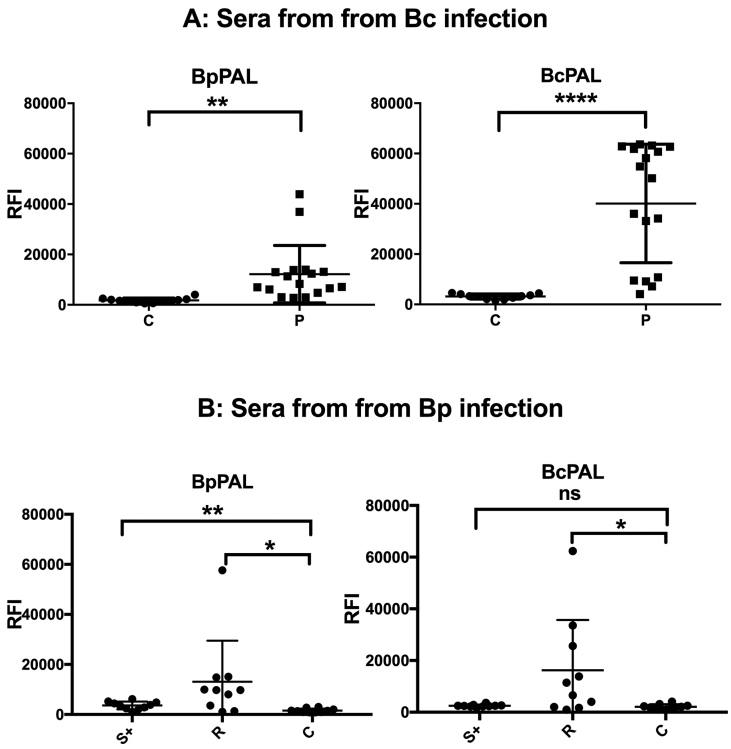


Figure 3.5: (Upper panel) Results of the unpaired t test for Bc infection. The Bc positive patient group is labeled as P. The healthy control group is labeled as C. (Lower panel) Results of the unpaired t test for *B. pseudomallei* infection (Bp). The seropositive patient group is labeled as S+, recovery patients are labeled as R, and the seronegative control group is labeled as C. (ns means not significant. Significant: $p < 0.05$, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, and **** = $p < 0.0001$).

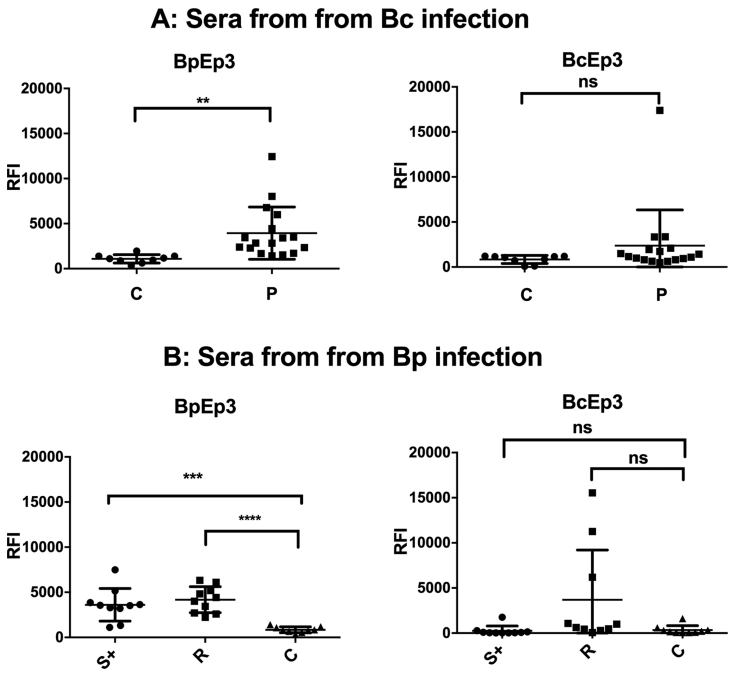


Figure 3.6: (Upper panel) Results of the unpaired t test for Bc infection. The Bc-positive patients group is labeled as P. The healthy control group is labeled as C. (Lower panel) Results of the unpaired t test for B. pseudomallei infection (Bp). The seropositive patients group is labeled S+, recovery patients are labeled R, and the seronegative control group is labeled C. (ns means not significant. Significant: $p < 0.05$, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, and **** = $p < 0.0001$).

seronegative controls (all p values <0.05). On the contrary, BcEp3 failed to discriminate patients from healthy individuals for both types of infections, indicating that mutations in the epitope sequence may be reflected by completely different immunoreactivity of the free peptide epitope, most likely linked to different conformational ensembles resulting from differences in amino acid composition, and in particular from the presence of Pro81. To address these issues, we carried out comparative MD analyses of BpEp3 and BcEp3, in the context of both their cognate proteins and as isolated peptides in solution.

Comparative MD Study of BpEp3 and BcEp3 within $\mathbf{Pal_{Bc}}$ and $\mathbf{Pal_{Bp}}$ and in Isolation

In light of the different immunoreactivities shown by BpEp3 and BcEp3 epitope peptides, extensive MD simulations were carried out to gain insight into the conformational preferences of the two Pal epitope sequences. The most significant sequence difference is the Ala (BpEp3) to Pro (BcEp3) substitution at position 81. Other changes entail BpEp3-Lys to BcEp3-Gln (residue 76), BpEp3-Glu to BcEp3-Gln (residue 78), and BpEp3-Glu to BcEp3-Met (residue 83). Despite such substitutions, the crystal structures show no significant conformational difference for the polypeptide backbones of the two full-length proteins (Figure 3.3B). This observation holds true also when focusing specifically on the epitope regions: in the context of their cognate proteins, both epitope stretches populate analogous helical conformations, which are stabilized by the structural and packing constraints imposed by the rest of the protein in the native 3D fold.
Notable differences, however, emerge when the two epitopes are investigated in isolation. Figure 3.7 shows, as a function of time, the RMSD of the conformations sampled by the free epitope peptides during the MD simulations having the epitope structure adopted in the cognate full-length proteins as the reference structure.
In general, BcEp3 and BpEp3 tend to show different dynamic behavior, with the former recursively populating conformations with low RMSD vs the same polypeptide stretch in the full-length protein context at 300 K. In contrast, time-dependent RMSD plots show that this is not the case for BpEp3: higher RMSD values consistently emerge, and more varied ensembles of BpEp3 conformations are populated. Additional MD simulations were run at 330 K to speed up sampling, confirming such observations. We next calculated Root Mean Square Fluctuations (RMSF), defined as

$$\mathrm{RMSF}(k) = \sqrt{\frac{1}{N}\sum_{i=i}^{N}\left(r_{k,i} - r_{k,\mathrm{ref}}\right)^2},$$

where k is an amino acid, N is the total number of frames, $r_{k,i}$ is the position of the center of mass of the k-th amino acid at the i-th frame, and $r_{k,\mathrm{ref}}$ is the reference position (in our case, the crystallographic structure) the center of mass of the k-th amino acid.
RMSF gives us an insight into the flexibility properties of both epitope peptides in solution. RMSF values are consistently higher for BpEp3, indicating higher flexibility for this peptide, a property that may in turn favor its ability to explore a wider ensemble of conformations. (Figure 3.8).
In terms of secondary structure content, as expected, the presence of Pro81 in free BcEp3 acts as a helix breaker, forcing the peptide toward alternative turn conformations (Figure

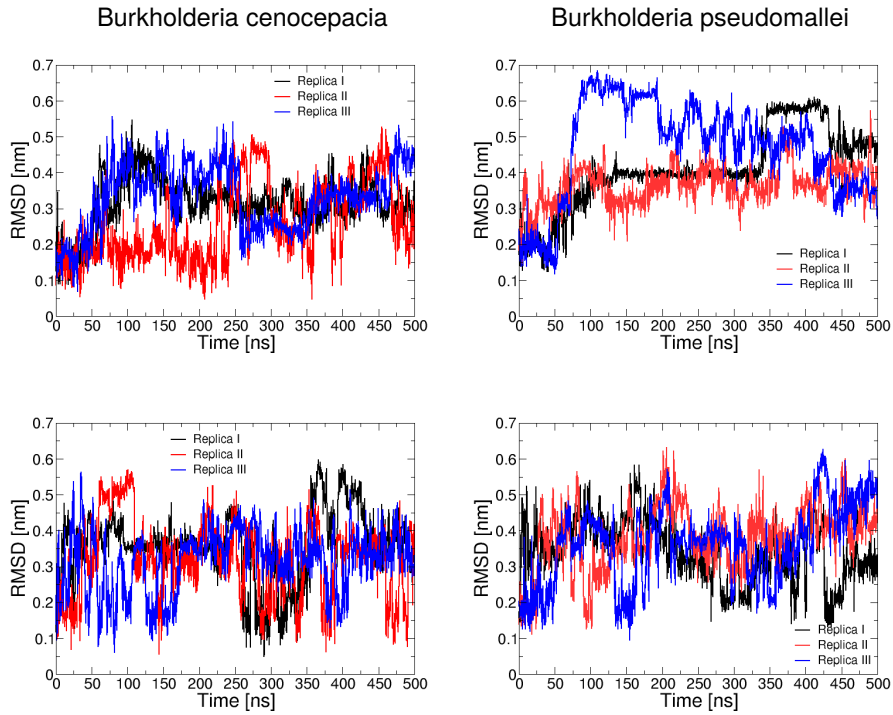


Figure 3.7: Time dependent Root Mean Square Deviation (RMSD) of the backbone atoms for the simulations of BcEp3 (left panel) and BpEp3 (right panel) vs. the corresponding experimental crystallographic structure of the epitope in the whole protein. Upper lane - simulations at 300 K; lower lane - simulations at 330 K. Different colors refer to different replicas.

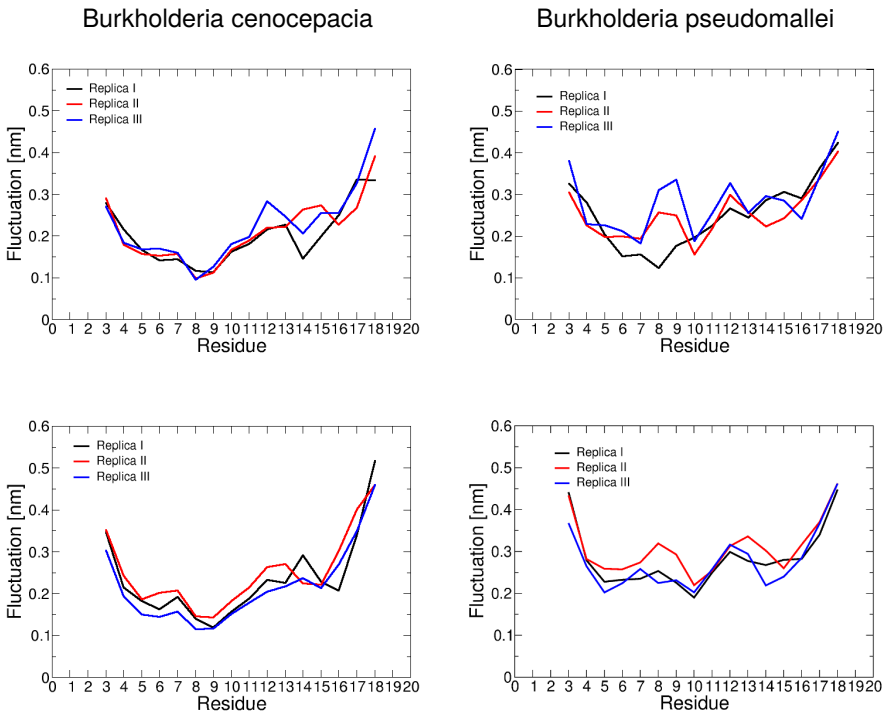


Figure 3.8: Root Mean Square Fluctuations (RMSF) for the simulations of BcEp3 (left panel) and BpEp3 (right panel) having the crystallographic structure of the epitopes in the whole protein. Upper lane - simulations at 300 K; lower lane - simulations at 330 K. Different colors refer to different replicas.

3.9). Interestingly, no significant secondary structural differences were evident in the context of simulations run on the respective full-length proteins.

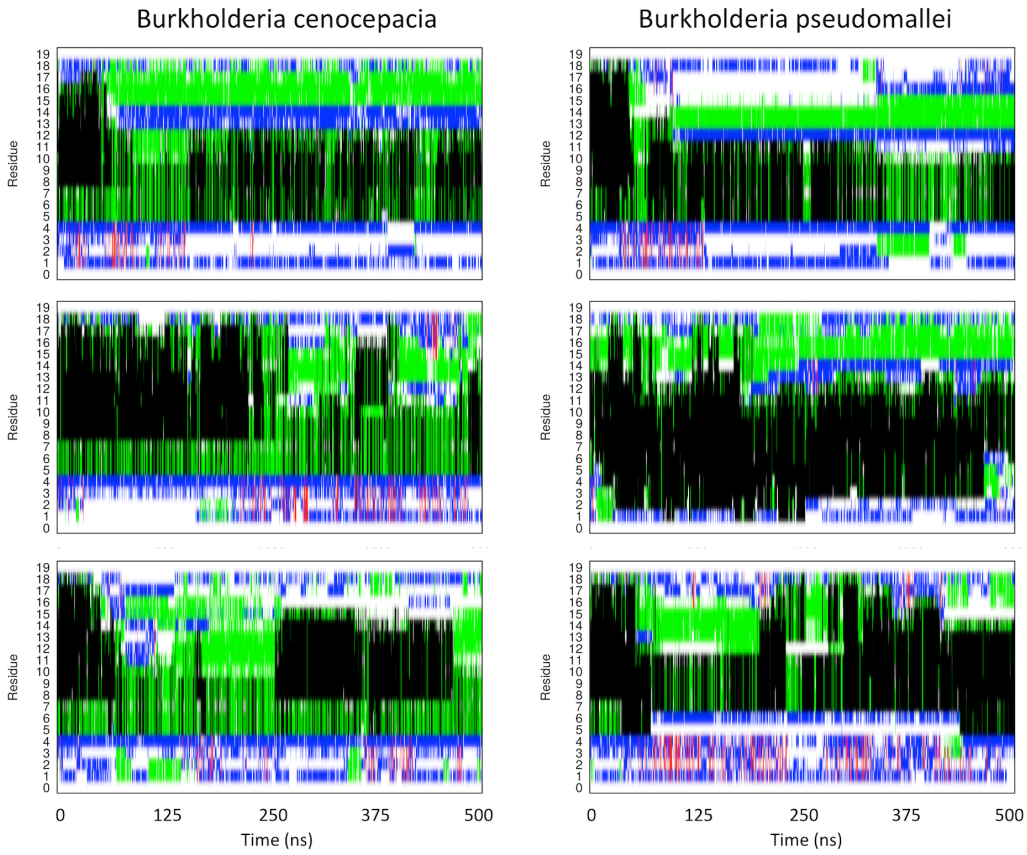


Figure 3.9: Time-dependent evolution of the secondary structure content of BcEp3 (left panel) and BpEp3 (right panel) in isolation, in solution. Results from three independent 500 ns simulations per system are shown. Color code: White, coil; black, α-helix; red,β-sheet; green, β-turn; and blue, polyproline (II).

To corroborate our computational calculations, the tendency of BpEp3 and BcEp3 peptides to populate helical conformations was assessed by circular dichroism spectroscopy (CD). Although both epitopes did not adopt characteristic helical conformations in absence of TFE, BpEp3 displayed a remarkable propensity to fold into an α-helical structure, as revealed by spectral minima in the 222 and 208 nm regions, upon addition of low concentrations of structure-inducing cosolvent trifluoroethanol (TFE) (Figure 3.10). This behavior was not observed for BcEp3 (Figure 3.11).

Overall, the BpEp3 sequence shows a more pronounced tendency to populate helical conformations, which also characterize the sequence in the full-length structure of the cognate protein, while displaying higher conformational flexibility. The presence of Pro diminishes the tendency of the peptide to populate helical conformational ensembles

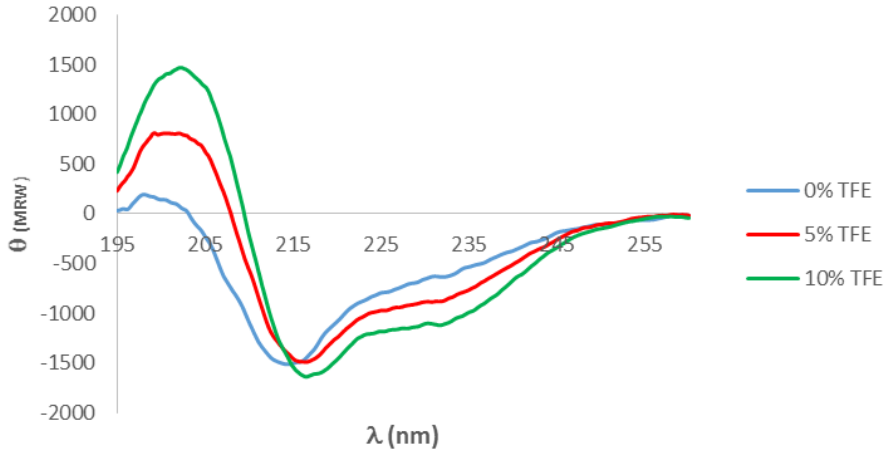


Figure 3.10: BpEp3 circular dichroism spectra comparison at variable trifluoethanol (TFE) concentration. While the BpEp3 peptide alone is not shaped into a canonical α-helix conformation, the typical α-helical CD signal is promptly observed upon increasing concentration of TFE. This analysis was carried out at PPC lab at ICRM-CNR.

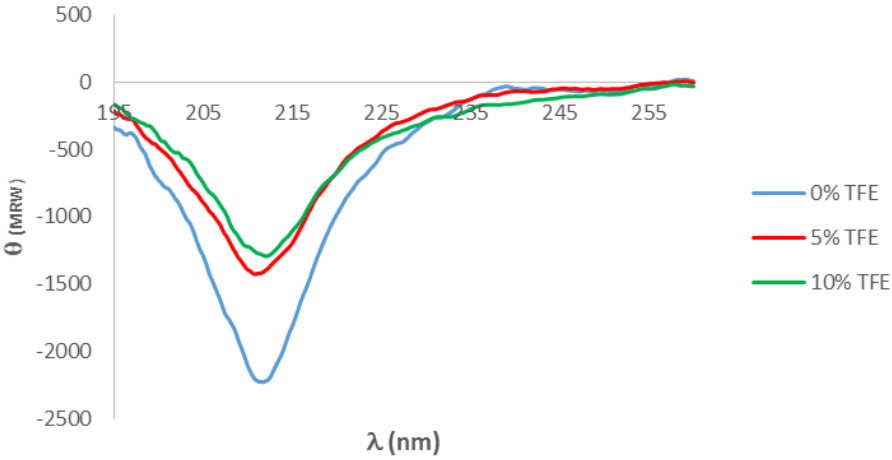


Figure 3.11: BcEp3 circular dichroism spectra comparison at variable trifluoethanol (TFE) concentration. In contrast to BpEp3, the typical α-helical CD signal is not observed upon increasing concentration of TFE, indicating a lower propensity to fold into an α-helical structure. This analysis was carried out at PPC lab at ICRM-CNR.

while decreasing its flexibility, trapping the isolated BcEp3 peptide in conformations that are poorly recognized by patient sera Abs that were ultimately generated against the original protein antigen.

In this context, we examined the free-energy landscape of the two peptides, projecting the trajectories on the collective variable represented by the RMSD from the α-helical structure, using the weighted histogram analysis method (WHAM) [148] approach to evaluate the free energies of the conformational basins visited (Figure 3.12).

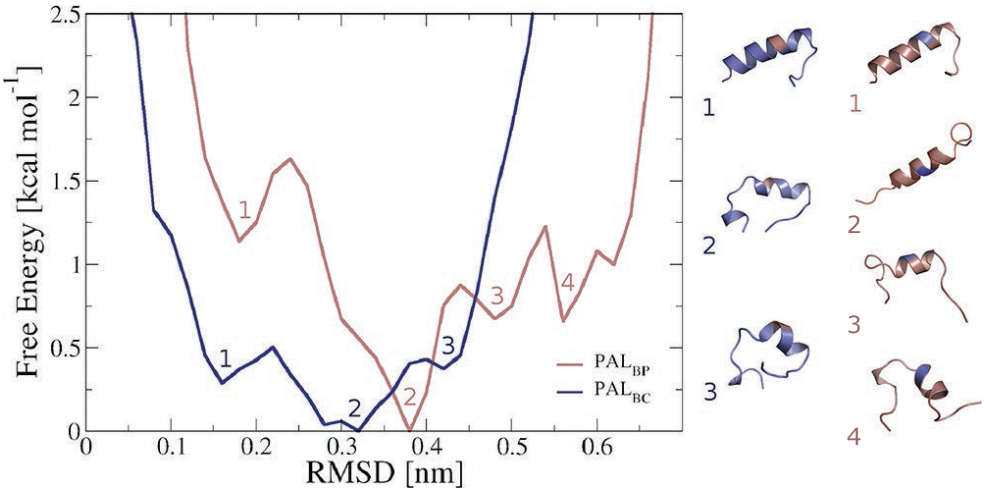


Figure 3.12: The results of the WHAM free-energy analysis from the combined trajectories of BcEp3 and BpEp3, together with representative structures from each basin.

BcEp3 shows a broader free-energy minimum in which the helical conformation is separated by relatively low barriers corresponding to turnlike structures. Qualitatively, we have 4 different conformers for BpEp3 and 3 conformers for BcEp3. The helical basin for BpEp3 is separated from alternative conformations by higher barriers, compared to those in the previous case, confirming the possibility of the peptide to be locally trapped in a helical conformation. Nevertheless, alternative conformations are also accessible, consistent with the observed increased flexibility for BpEp3.

A minimalistic model to rationalize the overall differences characterizing BcEp3 versus BpEp3 thus originates from the consideration of the differences in conformational dynamics observed for the two synthetic epitopes. One can in fact hypothesize that the more structurally flexible BpEp3, while conserving a higher tendency than its Bc counterpart to populate helical conformations, can adapt to a larger and more diverse pool of Abs, providing broader cross-reactivity, capturing Abs elicited against homologous antigens from different species. In this framework, because the aim is to identify pools of polyclonal Abs able to detect general Burkholderia infections and not to discriminate between them, the ability of BpEp3 to explore several sets of conformations may effectively facilitate the binding of several diverse Abs by expanding the dimensions of the

conformational ensemble accessible to the epitope. This property may also be partially reflected in the context of the full-length protein, whereby the more flexible epitope in Pal_{Bp} may allow local conformational changes or unfolding events that prompt adaptation and binding to diverse Ab pools. In this context, the restriction of the conformational freedom by the Ala81 to Pro mutation (in Pal_{Bc}), combined with the lower tendency to populate helical conformations reminiscent of the epitope structure in the full-length parent antigen, translates into a decrease in diagnostic performance in this particular application, where the aim is to detect infections caused by different *Burkholderia* strains.

Along these lines it was previously shown that favoring BpEp3 helical structuring, through conformational restriction by chemical stapling, improved the diagnostic performance when targeting specific Abs within a patients group infected with Bp only, with the capability of distinguishing between seropositive and recovered patients [144]. Preorganization to a α-helix is expected to increase the affinity for specific Abs raised against a given structural subpopulation, which would translate into stronger signals. On the other hand, the stapled peptide was shown to elicit Abs that were significantly less bactericidal than those generated against the unrestricted BpEp3 [144]. We speculate that conformational restriction of the immunogen, combined with the fact that such conformations may be quite different from those in the 3D structure of the full antigen, translates into the generation of Abs capable of recognizing only a limited ensemble of specific epitope conformations at any time. The effect of such a conformational limitation primes the Abs to target (*in vivo*) only epitopes that present structures very similar to that of the stapled epitope, thus decreasing their bactericidal potential compared to Abs raised against the more flexible BpEp3.

3.5.2 Computational Implementation

We performed energy minimization of all of the structures with AMBER16 [149] in explicit solvent (TIP3P water model [150]) using the FF14SB force field [151] with a steepest-descent method ($3 \cdot 10^3$ steps), followed by a run in a conjugate gradient algorithm ($7 \cdot 10^3$ steps). After minimization, we heated the system from 0 to 300 K (or 330 K) in the NVT ensemble over $2.5 \cdot 10^4$ MD step with a time step of 2 fs. After the heating process, the simulation was run for 50 ps (2 fs time step) in the NPT ensemble (Berendsen barostat [152], Langevin temperature control with $\gamma = 2\text{ps}^{-1}$) using the SHAKE algorithm [64] to constrain all of the hydrogen-containing bonds.
All production simulations were performed under the same conditions of the last equilibration part; we carried out MD simulations at 300 and 330 K (three replicas per temperature, 250 ns each) of the crystal structures of Pal_{Bp} and Pal_{Bc}. For both epitopes, we followed the same protocol at 300 and 330 K, performing a longer MD run (500 ns) on three different replicas for both systems.

All trajectory analyses (RMSD, RMSF, radius of gyration, and structural clustering) were performed using the GROMACS suite (version 4.5.5) [45]. To prove, in an unsupervised way, the different behavior of the two epitopes, we performed a cluster analysis on a

trajectory obtained by concatenating all of the C_α positions (to maintain a fixed number of atoms for all the frames) from MD runs carried out on both epitopes. If this analysis separates the two different epitopes in different clusters, then we have statistical proof of the dynamical difference of the two sequences due to the four amino acid mutations. Secondary structure analysis was carried out using the STRIDE algorithm [153]. Finally, we studied the free-energy landscape of all of the systems reweighting data from all of the trajectories using WHAM in the implementation provided in the SMOG suite [154]. Epitope predictions were carried out on representative structures of the most populated structural clusters obtained during MD by applying the matrix of local coupling energies (MLCE) method, as previously described [142, 133].

3.5.3 Discussion

Our work presents a simple viable approach to the characterization of the molecular determinants that guide recognition in designed immunodiagnostic probes. The protocol was applied to study the potential of two highly related epitope sequences for the detection of related B. pseudomallei and B. cenocepacia infections. Present treatment of both Burkholderia species relies on intravenous and/or oral antibiotic administration that is often inefficient because of the resistance of both bacteria to most common antibiotic classes [155]. Immunodiagnostic tests would accelerate the selection of the most appropriate antibiotic to be used, avoiding laborious diagnosis based on bacterial cultures. Our approach allowed us to explore the impact of sequence variation and structural flexibility on Ab binding and to identify BpEp3 from Bp (a more conformationally flexible epitope) as the probe with the higher cross-reactivity and better diagnostic performances to simultaneously reveal infections from both Bp and Bc. These characteristics make BpEp3 a lead candidate for further refinement for use in the diagnosis of *Burkholderia* species infections. In this context, epitopes with cross-reactivity against different species of the same pathogen may represent optimal candidate components for the development of broadly protective vaccines, thus generating interesting therapeutic applications. Under such perspectives, the results presented here highlight the potential offered by current experimental and simulative methods to expand and modulate the molecular and conformational diversity space of reactive *Burkholderia* epitopes using rational, structure-based approaches.

3.6 SAGE: automated epitope grafing

A possible improvement to the direct use of peptidic epitopes consists in implanting them onto a scaffold of interest to improve their stability. Moreover, to maximize the efficacy and the durability of the immune response, multiple epitopes can be inserted into the starting structure. This procedure, termed epitope grafting, has been widely used on viruslike particles (VLPs) since the early 1980s [156, 157] and was pioneered for immunogenic proteins by the Schief group[158, 159, 160, 161].

Epitope grafting involves the transplantation of a structural/functional motif onto a

structurally homologous region of an unrelated protein target, possibly already hosting other reactive sequences. The antigen target protein must display one or more regions that are conformationally compatible with those of the epitope to be grafted. Once the epitope is transplanted, its conformation should be stable in the new context, to support optimal presentation for the binding of the antibody and for processing by the immune system. To date, all the design part is carried out with a strongly limited use of computational tools (with the only exception represented by ROSETTA [162]), and an eventual lenghtly and costly experimental validation, which restricts the possibility to efficiently obtain viable grafted candidates.
For this reason, the availability of an automated tool for the design of multiepitope antigens is expected to be a relevant support for vaccine development and for their testing in different contexts, from structural biology to immunology. In particular, it permits to rapidly select stable protein constructs containing a foreign antigen among an ensemble of alternative solutions, without resorting to complex and lengthy techniques of structural biology and modeling. Our computational pipeline, called SAGE (strategy for alignment and grafting of epitopes), was tested analyzing the results of blind linear epitope grafting predictions and benchmarking against known cases of successful design reported in the literature [158, 160, 161, 163, 164]. It is found that SAGE, despite its simple approach, identifies successfully the best experimentally validated candidates among its top scoring solutions without any preliminary knowledge-based input.

3.6.1 Workflow

The prediction of the grafting position requires the knowledge of the structure of the target, of the structure of the whole immunogenic cognate protein that contains the epitope to be transplanted and of the position of the epitope along the sequence. It consists of four phases:

1. Structural/sequence alignment
2. Secondary structure prediction
3. Structure scoring
4. Exposed surface scoring

SAGE is written as a Python 2.7 package that performs the analysis by accessing external resources such as PSIBLAST and Naccess.

Alignment phase

The program takes the user-defined linear epitope from the original crystallographic structure of the cognate protein and performs three different types of alignment onto the target protein. For this step PyMOL [48] scripts are used, namely a pure sequence alignment (Pymol script `align`), a structure-based alignment (Pymol script `super`) and hybrid alignment (Pymol script `CEAlign` [165]). In every alignment run, the script returns the three best candidates. To obtain a larger set of suboptimal candidates, the alignment

is repeated for several cycles, constraining the search to segments of decreasing length which cover exhaustively the protein, as shown in Figure 3.13.
At the end of this first part, SAGE generates a set of FASTA files containing the sequences of all the grafting candidates.

Secondary Structure Scoring

Since the secondary structure of the cognate fragment and of the scaffold can change upon grafting, we evaluated the secondary-structure propensities of the two based solely on their sequence, independently of the actual secondary structure they display. For this purpose we applied the s2D method [166], which returns a per-residue α/β formation probability based on sequence information. Applying the prediction to a scaffold of N amino acids, the α/β propensities can be seen as two different points $a^{\alpha,\beta}$ in an N-dimensional euclidean space.
After the alignment, we obtained M different sequences of N amino acids each. Thus, we can define a distance

$$d_{\alpha,\beta}(a^{\alpha,\beta}, b^{\alpha,\beta}) = \sqrt{\sum_{i=1}^{N} \left(a_i^{\alpha,\beta} - b_i^{\alpha,\beta}\right)^2}$$

between such points that quantifies the difference in each of the two types of secondary structure between any two sequences. These distances are calculated between the scaffold and the sequences obtained from the alignment to evaluate how similar the grafting candidates are expected to be to the original structure. The scores on the two types of secondary structures are then merged in a score for secondary-structure scaffold compatibility, defined as

$$S_{\text{sc}}(b) = \frac{1}{d_\alpha(a,b)} + \frac{1}{d_\beta(a,b)}.$$

To take into account the similarity between the epitope in its original protein and in the candidates, we also evaluated the similarity in secondary-structure propensity in the epitope-containing segment. If this is composed by L residues, we compare L-dimensional vector $a^{\alpha,\beta}$ for the epitope in its cognate protein and two different L- dimensional vectors $b^{\alpha,\beta}$ for every candidate. Therefore, we define another score which reports the epitope secondary structure compatibility

$$S_{\text{ec}}(b) = \frac{1}{d_\alpha(a,b)} + \frac{1}{d_\beta(a,b)}.$$

Exposed Surface Scoring

The structural information available is used to evaluate how grafting affects the exposed surface of the protein. The exposed surface area is measured using Naccess [167]. For every candidate, SAGE, via the PyMOL mutagenesis tool, creates a new structure mutating the residues of the original scaffold with those of the new epitope using the most probable rotamer. The exposed surface is evaluated for all epitope residues in the cognate protein and for the corresponding grafted residues on all the putative structures.

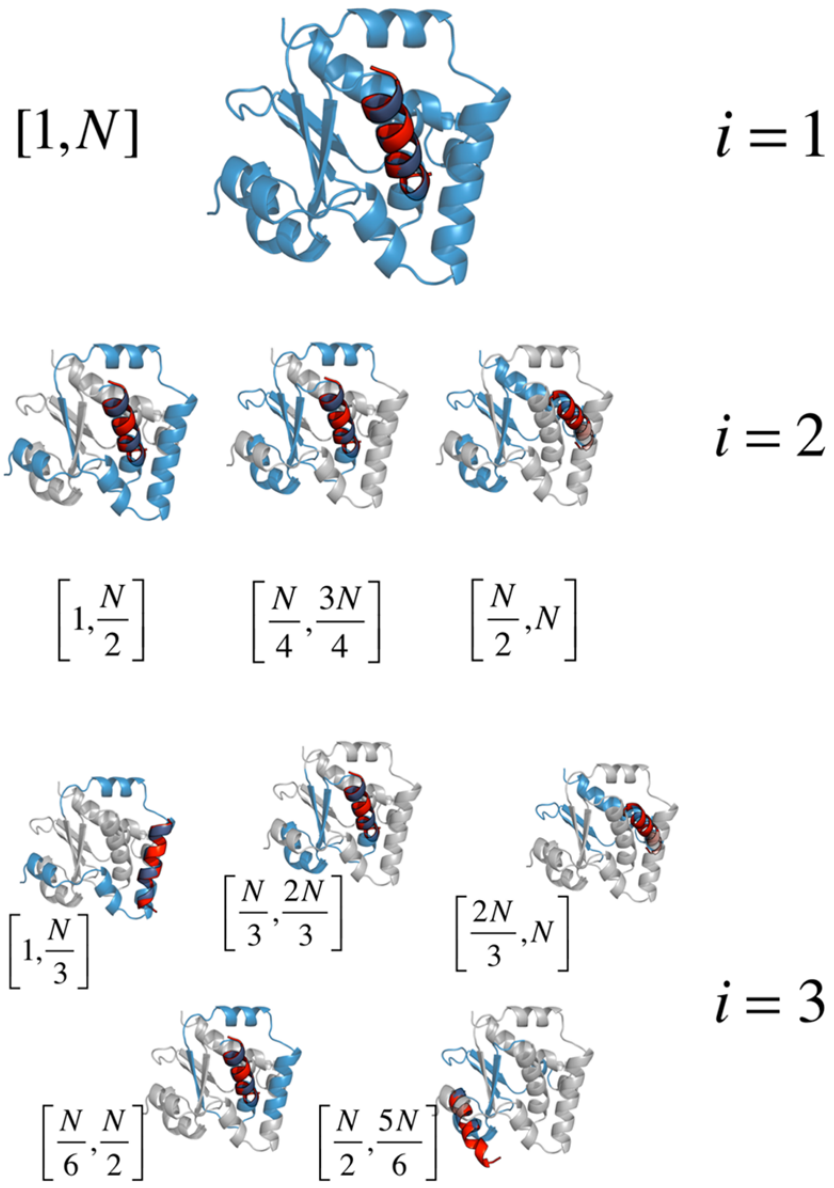


Figure 3.13: Scheme of the structure/sequence alignment. In red, the superposed epitope (4e10) and in blue the part of the scaffold aligned (1Z6N). In the first step ($i = 1$) the epitope is aligned on the full scaffold, from residue 1 to N. In the second step, it is aligned first on the first half (residues [1, $N/2$]), then on the second half [$N/2$, N], in the middle part [$N/4$, $3N/4$], and so on. The alignment process can be performed with a user-defined number of steps.

We obtain a L-dimensional vector a with the reference exposure and, having M candidates, a collection of M L-dimensional vectors for the exposure of the grafted fragment. We can compute again an Euclidean distance

$$d(a,b) = \sqrt{\sum_{i=1}^{L} (a_i - b_i)^2}$$

between the per-residue exposed area a measured on the epitope in the cognate protein and b measured on the grafted epitope, and an exposure score is defined as

$$S_{\text{exp}}(b) = \frac{1}{d(a,b)}.$$

Selection of the candidates

To restrict the number of viable candidates, we need to define a single score. The main problem with the three scores defined above is that they are incommensurable, being defined on different scales. To make them comparable, every score is normalized between 0 and 1 dividing it by the highest score found in the whole candidate pool.
A reactive epitope in an immunogenic context has to be exposed to the solvent to be recognized by the immune system. To exploit the structural information given by the exposed surface analysis and to penalize the candidates with a hidden grafted epitope we use the normalized exposed surface score as a weight for the remaining scores

$$\tilde{S}_{\text{sc,ec}}(b) = \frac{S_{\text{exp}}(b)}{\max_b(S_{\text{exp}}(b))} \cdot \frac{S_{\text{sc,ec}}(b)}{\max_b(S_{\text{sc,ec}}(b))}.$$

Finally, we generate the total score as the average of the scaffold and the epitope similarity reweighted scores defined above

$$S_{\text{tot}}(b) = \frac{1}{2}\left(\tilde{S}_{\text{sc}}(b) + \tilde{S}_{\text{ec}}(b)\right).$$

3.6.2 Validation of the algorithm

To validate the algorithm, we compared the grafting candidates obtained by SAGE with the structure shown in previous structural epitope grafting works. It is important to note here that a limited number of grafting reports have appeared in the literature, in particular with regards to cases where the final structure of the protein could be crystallized/obtained by homology modeling. We therefore retrieved the most possible available cases with structural information, running our analysis for viral sequences taken from HIV-1, RSV, and snake toxin [160, 161, 163, 168], in which the authors grafted a relevant linear epitope onto a set of different scaffolds. We also tested SAGE performance for epitope grafting on viruslike particles [164, 169], although the graft length in one case [164] does not correspond to the length of the scaffold deletion. The authors used fragments of different length, ranging from extremely short [161, 164, 169] (<10 residues), to short [160, 163] (10-12 residues), and to long [168] (24 residues). The number of cycles in

the alignment phase was chosen considering the ratio between the length of the epitope and the length of the smallest aligned scaffold fraction.
Given a N residues long scaffold and a L residues long epitope, the number of cycles k is given by $k = \frac{M}{L}$. For short epitopes, the number of cycles can be decreased to speed up the candidate generation.

Extremely short epitope

In the work of Azoitei *et al.* [161], the authors expressed four sequences grafting the 2F5 binding site from the gp41 HIV-1 glycoprotein (PDB: 1TJI/P) on two different scaffolds (PDB: 1WNU and 2CX5). We performed the grafting candidate search using three different section of the 2F5 binding site, corresponding to residues 661-667, 661-666, and 662-667. In every run the correct original grafting zone was found by the algorithm (see Table 3.1).

Table 3.1: SAGE grafting site prediction results for extremely short epitopes

Scaffold	Epitope	Graft Position	Original Graft	no. of cycles	1st	2nd	3rd	4th	5th	Rank
1WNU	2F5	661-666	48-53	13	56-61	**48-53**	75-80	80-85	61-66	2nd
		661-667	48-54	11	56-62	80-86	135-141	86-92	73-79	6th
		662-667	49-54	13	57-62	81-86	94-99	**49-54**	87-92	4th
2CX5	2F5	661-666	85-90	14	71-76	145-150	25-30	84-89	**85-90**	5th
		661-667	85-91	12	31-37	25-31	71-77	84-90	145-151	7th
		662-667	86-91	14	32-37	85-90	72-77	70-75	26-31	6th
3J2V	MAGE-3	168-176	76-84	9	**76-84**	2-10	12-20	134-142	30-38	1st
hamVP1	CEA	571-579	80-89 288-295	20	**82-90**	241-249	83-91	193-201	338-346	1st 14th

In three cases out of six, the position chosen by the authors (obtained from the PDB structures 3RFN and 3RI0) was in the top five of our selection. In the remaining cases the original graft was predicted as the sixth (two cases) and the seventh candidate (one case).

In the work of Kazaks *et al.* [169], a melanoma-specific antigen (MAGE-3, PDB: 4V0P/A) was grafted onto a Hepatitis B virus core protein (PDB: 3J2V/A). In the original paper, the epitope was grafted in the middle of the scaffold and inserted at the end of the C terminal region of the scaffold. While the insertion was not detectable by SAGE workflow, the grafted part was correctly predicted as the first candidate (see Table 3.1).

In the work of Lawatscheck *et al.* the authors delete residues from 4 sites (residues 80-89, 222-225, 243-247, and 288-295) of the hamster polyomavirus and inserted a carcinoembryonic epitope (CEA) characterized by a different length than the excised sequence, thus varying the overall length of the final protein. Both the structures needed reconstruction: the immunogenic protein described in the original paper was a homology model (PDB: 10E7) that we completed using PULCHRA [170], while the structure for

the scaffold was reconstructed using MODELLER [171] having as template the murine polyomavirus capsid (PDB: 1SID) and the simian virus S40 capsid (PDB: 1SVA)[172]. Since the lengths of insertion and deletion do not match, SAGE could only find the zone of insertion in the sites that differed less than two amino acids in length from the original epitope: the first and the fourth site (residues 80-89 and 288-295). Importantly, SAGE found site 1 as the first candidate (residues 82-90) and the site 4 as the 14th candidate (residues 288-295) (see Table 3.1).

Short epitope

In the work of Correia, Ban, *et al.* [160] the authors expressed 103 different designed protein obtained by a side chain grafting method, using the 4E10 epitope of the gp41 HIV-1 glycoprotein (PDB: 2FX7/P) on six different scaffolds (1EZ3, 1ISE, 1IS1, 1VI7, 1XIZ, 1Z6N). Six of them were deposited in the Protein Data Bank. In contrast to the previous case, the epitope (residues 671-680) is not grafted entirely on the scaffold, but it contains a gap of 2 residues (NWFDIT–LW instead of NWFDITNWLW). Furthermore, the authors changed the sequence both in the scaffold and in the epitope by single-point mutations to favor protein stability. SAGE was applied to reproduce this grafting, with the difference that only the wild-type sequences were used and no mutation was applied. In Table 3.2 we show the redesign results, comparing them with the grafting position in the original paper.

Table 3.2: SAGE grafting site prediction results for short epitopes

Scaffold	Epitope	Graft Position	Original Graft	no. of cycles	1st	2nd	3rd	4th	5th	Rank
1EZ3	4E10	671-680	68-77	7	**68-77**	27-36	110-119	32-41	132-141	1st
1IS1	4E10	671-680	149-158	10	144-153	106-115	**149-158**	33-42	124-133	3rd
1ISE	4E10	671-680	149-158	10	144-153	68-77	**149-158**	33-42	106-115	3rd
1VI7	4E10	671-680	149-158	11	108-117	135-144	100-119	34-43	147-156	14th
1XIZ	4E10	671-680	111-120	8	**111-120**	88-97	139-148	38-47	114-123	1st
1Z6N	4E10	671-680	138-147	9	119-128	**138-147**	22-31	33-42	3-12	2nd
2CRD	1NEA	26-37	25-36	4	3-14	4-15	**25-36**	19-30	12-23	3rd

Importantly, the original grafting zone was found in all the candidates. In five out of six calculations, the original grafting position was in the first three suggested candidates. In the remaining case of the 1VI7 candidate, the original grafting position was in the 14th position. This is probably due to the massive residue deletion in the scaffold (135 out of 206 amino acids), which eliminated possible grafting sites detected by SAGE. In fact, neglecting the predictions obtained on the deleted part, the original position scores third.

In the work of Drakopoulou *et al.* [163], the authors grafted a 12 residue long epitope from a snake toxin (PDB: 1NEA) onto a 37 residue long scaffold from a scorpion toxin (PDB: 2CRD). In this case, the grafted sequence was mutated with respect to the original one by inserting two cysteine residues to stabilize the final construct through disulfide

bridges; furthermore, the first six amino acids in the final chimeric protein were deleted. SAGE predicts the correct position at the third place, but the first two candidates are partially in the deleted portion of the sequence (see Table 3.2).

Long epitope

A comparison is performed with the work of Correia, Bates, *et al.* [168]. The authors, starting from a single structure (3LHP), graft a long epitope from the F1 glycoprotein of RSV (3IXT/P), obtaining four different candidates with the same final position of the epitope sequence. The scaffold in this case is rather short (116 residues), considering the epitope length (24 residues). Results are in Table 3.3.

Table 3.3: SAGE grafting site prediction results for long epitope

Scaffold	Epitope	Graft Position	Original Graft	no. of cycles	1st	2nd	3rd	4th	5th	Rank
3LHP	3IXT	254-277	74-97	5	**74-97**	76-99	85-105	71-94	94-117	1st

The first suggested position obtained by SAGE is the one chosen by the authors for experimental grafting.

3.6.3 Discussion

We presented a new method for redesign of antigens with the goal of creating immunoreactive proteins displaying multiple epitopes. Presenting multiple immune-reactive sequences on a single biomolecule is expected to elicit a more efficient and durable immune response. This concept was first introduced using nanoparticle-based constructs and multivalent synthetic systems [173] but can aptly be applied to entirely biomolecular systems. In this respect, SAGE provides a novel solution for the initial screening of sequences coding for potentially highly reactive protein antigens, alleviating the need of producing and testing high numbers of candidates. A graphical summary of SAGE performance is in Figure 3.14).

As a caveat it must be stated that currently SAGE works only on linear epitopes. However, since many antibody mediated recognition phenomena involve conformational epitopes, we foresee the inclusion of procedures for discontinuous epitope grafting as a natural extension of our approach. Due to several technical hurdles, this implementation is currently under development.

Operatively, we believe SAGE is a useful tool for structural vaccinology studies, at the computational as well as at the experimental level. It provides a platform which is suitable for integration with other approaches of sequence/structure optimization, such as Rosetta-based methods or MD refinement and checks of the stabilities of resulting constructs [174, 168, 162].

Furthermore, SAGE will be implemented in a general-purpose web server to provide access to structural vaccinology to a diverse community of scientists. At this moment, SAGE is available as a Python 2.7 package which performs the explained analysis by

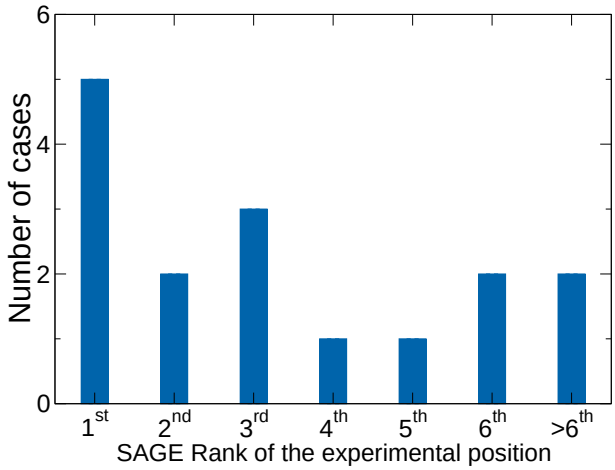


Figure 3.14: Prediction performance in SAGE validation.

accessing external programs that can be downloaded from the Internet: PSIBLAST [175] (via s2D [166]) and Naccess [167].

3.7 Applications of SAGE and preliminary experimental results

After the validation (see Section 3.6), we applied the SAGE algorithm to transplant the known Pal_{Bp} epitope BpEp3 onto two immunogenic proteins from *Burkholderia pseudomallei*, already known to host other immunoreactive sequences.

3.7.1 Grafting of Pal_{Bp} BpEp3 on $FliC_{Bp}$

$FliC_{Bp}$ is the flagellin of *Burkholderia pseudomallei*, which assembles to form the flagellar filament responsible for the bacterium motility (see figure 3.15). FliC has been shown to induce T-cell responses in mice and humans [176]. In mice, FliC inoculation induces partial protection against *Burkholderia* reinfection, thus reducing mortality and morbidity. Sterile protection, however, has not been achieved to date [177].
In a previous study [178], 3 principal epitopes (the green, red and blue regions in Figure 3.15) have proven to be immunogenic in seropositive patients.

We took the original crystallographic structure presented in the work of Nithichanon *et al.* [178] (PDB: 4CFI) and reconstructed the missing parts with an homology model using the homologue flagellin from *Salmonella typhimurium*. Later, we refined the structure carrying out a minimization. After a MLCE analysis of the reconstructed protein, one new epitope was detected in position 51-69 (the yellow region in Figure 3.15); however, this epitope can be considered as an "incidental" epitope, because that part of the protein

in the pathogen is buried into the flagellum structure.

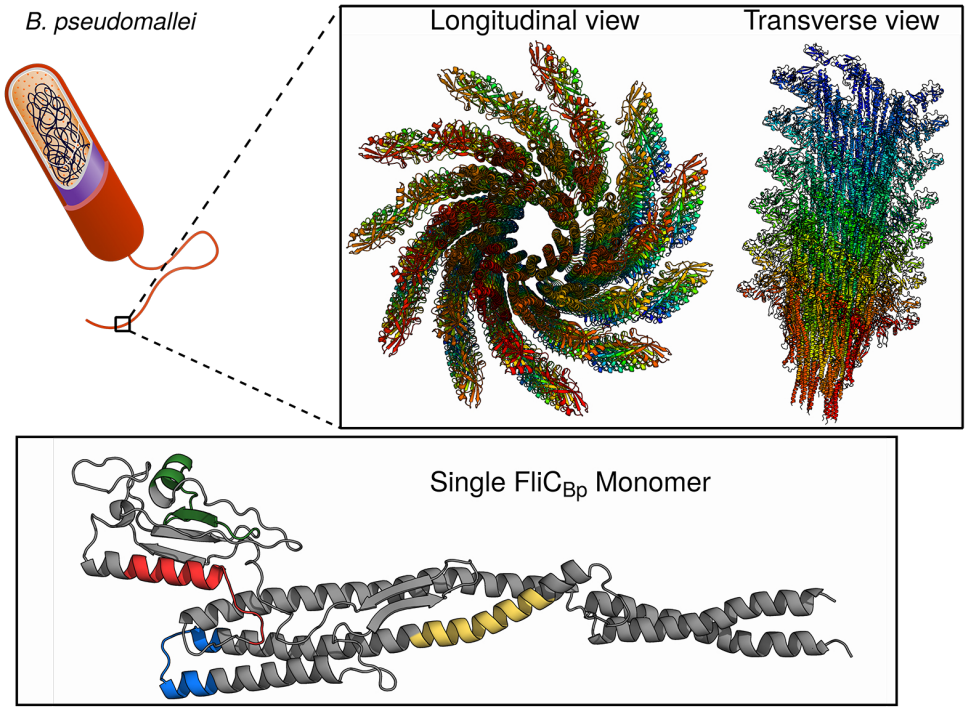


Figure 3.15: Schematic representation of *B. pseudomallei* flagellum at microscopic and molecular level. In the top-left part we have a scketch of a procariotic cell like *B. pseudomallei* (source: Wikipedia). In the top-right panel we have a molecular representation of the motility flagellum of the bacterium. Every color represent a different FliC$_{Bp}$ monomer. In the lower panel we can find the ribbon representation of FliC$_{Bp}$. The 4 colored regions of the structure was identified as natural epitopes of the monomer in solution (taken from [178]).

In our work, we designed using SAGE a list of possible grafts of the Pal$_{Bp}$ Bp3Ep epitope onto the reconstructed and refined FliC$_{Bp}$ scaffold as a proof of concept of the "superantigen" principle. Herein, we consider a superantigen a protein that contains known reactive epitopes, whise reactivity is augmented by the insertion of a foreign epitope from a different pathogen antigen. We run SAGE with a number of cycles $L = 9$, obtaining 46 candidates. The first 10 candidates predicted by SAGE are in Table 3.4 (The complete prediction are in Appendix C, Table C.14).

The first two candidates (101-120 and 100-119) were selected, for their distance from the N- and C-terminal zones of the protein, and for their partial superposition on an existing almost non-immunogenic epitope (the blue one in Figure 3.15). In fact, the presence of a native epitope suggests that this region of the protein does not contain any crucial residue for protein stability, and therefore it is an ideal position where insert a foreign epitope.

Rank	Insertion position	Exposure score	Structure score		Epitope score		Final score
			Raw	Weighted	Raw	Weighted	
1	101-120	0.749	0.835	0.625	1.000	0.749	0.687
2	100-119	0.702	0.983	0.690	0.927	0.651	0.670
3	70-89	0.963	0.790	0.761	0.482	0.464	0.612
4	319-338	0.928	0.725	0.673	0.562	0.522	0.597
5	294-313	0.711	0.784	0.557	0.756	0.538	0.548
6	81-100	0.787	0.915	0.720	0.432	0.340	0.530
7	13-32	0.741	0.956	0.708	0.448	0.332	0.520
8	105-124	0.621	1.000	0.621	0.597	0.371	0.496
9	66-85	0.898	0.676	0.608	0.422	0.380	0.494
10	46-65	0.778	0.873	0.679	0.354	0.275	0.477

Table 3.4: SAGE prediction for Ep3Bp epitope grafting on $FliC_{Bp}$.

The two selected candidates have been expressed at the Structural Biology Lab at University of Milan, and crystallization experiments are currently being performed.

3.7.2 Grafting of Pal_{Bp} BpEp3 on BPSL2520

BPSL2520 is a symmetric homodimer with a horseshoe form (see Figure 3.16), with unknown function. This particular protein was chosen for grafting for its ease in expression, purification and crystallization.

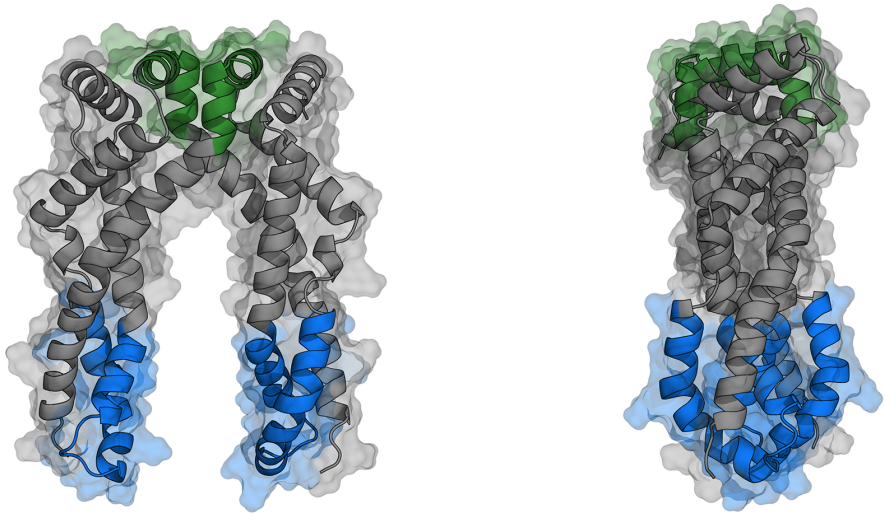


Figure 3.16: Frontal and lateral view of BPSL2520 dimer. The zone highlighted in green and blue are the epitopes detected by MLCE [133].

Preliminary computational analyses on the wildtype protein

We analyzed via MD simulation the behavior of the dimer; in particular, we focused on the stability of the crystallographic open configuration and the possibility for the protein to populate alternative, namely closed, conformation. This expectation was based on the observation that most of the amino acids in the interior of the horsehoe conformation are hydrophobic.

After a two step minimization ($3{\cdot}10^3$ steps with steepest descent algorithm and $7{\cdot}10^3$ steps with conjugate gradient algorithm), we heated the system to 300 K with a 50 ps simulation (TIP3P water model [150], Langevin thermostat with $\gamma = 2$ ps^{-1}, 2 fs timestep, Berendsen barostat [152], SHAKE [64] constraints on hydrogen-containing bonds) and then equilibrated with further 50 ps of MD at same conditions. The production run was performed for 3 replicas, 500 ns per replica at the same conditions of the equilibration. All the simulations were performed with AMBER16 [149] using FF14SB force field [151]. All analyses performed on MD trajectories were carried out using GROMACS 4.5.5 [45].

In Figure 3.17 2 replicas out of 3 reach the closed conformation within 100 ns of simulation, while the last replica shows a slighly different behavior, reaching the closed state at $\sim$ 350 ns of simulation.
We also studied the largest amplitude motions in the essential space of the wildtype protein: operatively, we computed a PCA of the covariance matrix of the position of all the residues during the trajectory, applyng a projection of all the trajectory along the 2 principal eigenvectors (Figure 3.18). The essential space analysis confirms the behavior seen in RMSD analysis: we can find a stable (closed) structure due to the hydrophobic interaction in the middle of the horsehoe (upper left part of Figure 3.18). From the last replica trajectory (blue dots in Figure 3.18), which was open for most of the MD run, we cannot find a localized open conformation in the eigenvector space, suggesting that the open conformation is a metastable state for the dimer.

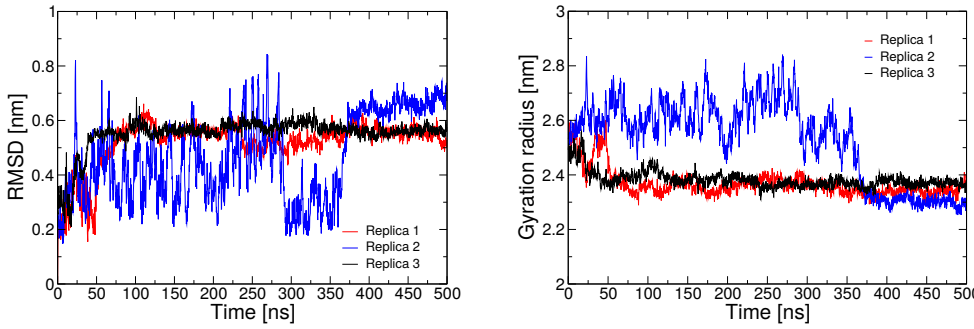


Figure 3.17: RMSD with respect to the crystallographic structure and gyration radius for BPSL2520 MD simulations

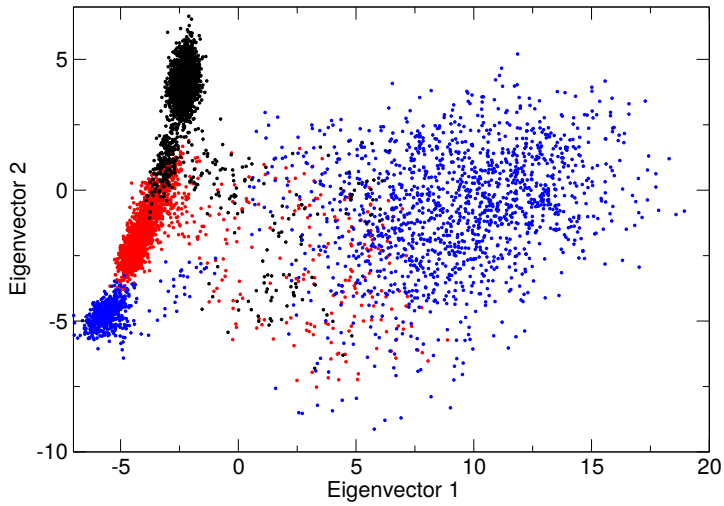


Figure 3.18: Representation of the essential modes from MD of BPSL2520. The replica 2 (blue) shows a different behavior in the essential space with respect to replica 1 (red) and 3 (black).

SAGE design and selection of candidates

SAGE was applied on BPSL2520 and on Ep3Bp epitope from Pal_{Bp} (residue 72-91) with a number of cycles $L = 8$, obtaining 30 candidates. The first 10 grafting candidates selected are in Table 3.5 (complete prediction results are in Appendix C, Table C.15).

Rank	Insertion position	Exposure score	Structure score		Epitope score		Final score
			Raw	Weighted	Raw	Weighted	
1*	66-85	0.989	0.838	0.828	0.790	0.781	0.805
2†	85-104	0.779	0.850	0.662	1.000	0.779	0.721
3*	62-81	1.000	0.827	0.827	0.611	0.611	0.719
4*	156-175	0.768	0.841	0.646	0.936	0.719	0.682
5*	115-134	0.945	0.704	0.665	0.740	0.699	0.682
6†	168-187	0.847	1.000	0.847	0.606	0.513	0.680
7*	118-137	0.901	0.789	0.710	0.720	0.649	0.680
8	89-108	0.855	0.761	0.651	0.753	0.644	0.647
9	169-188	0.829	0.958	0.794	0.590	0.489	0.642
10	88-107	0.803	0.792	0.636	0.797	0.640	0.638

Table 3.5: First 10 candidates from SAGE prediction for Ep3Bp epitope grafting on BPSL2520. Candidates marked with † cover part of the existing epitopes, thus was removed from the ranking. Candidates marked with $*$ were the 5 best candidates.

Some of the candidates cannot be considered: SAGE does not manage multichain proteins, and we performed the grafting protocol on a single monomer. In this way, the exposure score is not completely reliable and we need to check manually the surface of the epitope on the grafted protein. Fortunately, none of the first 10 candidates results buried in the dimer interface.
The 2nd and the 6th candidates share a large part of their position with an existing epitope, and thus were eliminated from the viable candidates.

We selected the best 5 candidates (see Figure 3.19) with smallest possible epitope overlap with the natural ones and carried out further *in silico* analyses to identify the more stable construct with the aim to express and crystallize it.

Computational analyses on candidates

Like in the case of wildype protein, we carried out MD simulations for 3 replicas with shorter production runs (250 ns instead of 500 ns).

All the grafting candidates show a similar behavior with respect to the wildtype scaffold: the majority of the replicas start from an open conformation and reach the closed state. Both the RMSD and the radius of gyration analyses do not give us a clear hint on which are the most promising grafts (see Figures C.1, C.2).

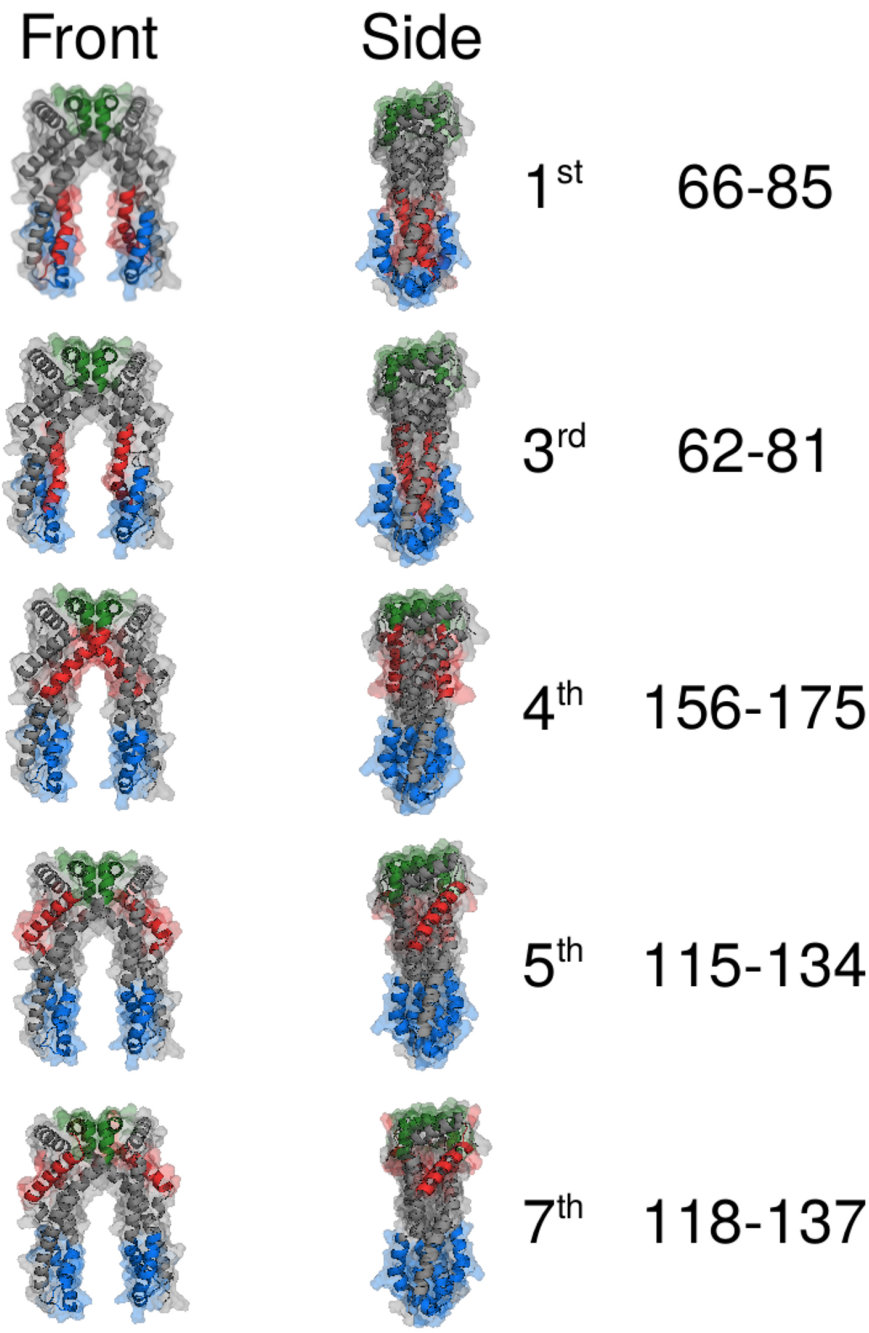


Figure 3.19: 5 selected candidates for Ep3Bp grafting on BPSL2520. In green and blue we higlighted the natural epitopes of BPSL2520, in red the grafted epitope.

On the other hand, analyzing the projection of the graft trajectories on the essential space defined by the first two eigenvector of the wildtype trajectories, we found the candidates most similar to the wildtype scaffold (see Figure 3.20) are the 1st and the 7th.

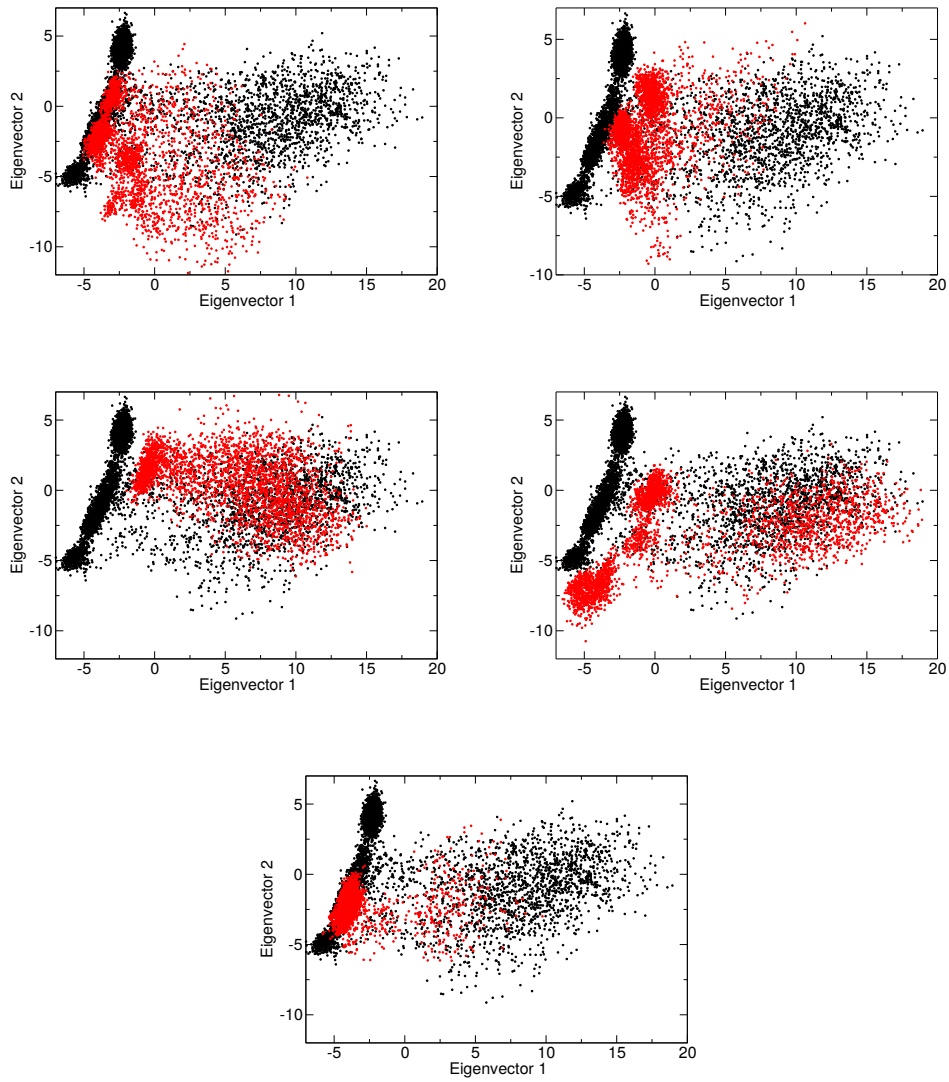


Figure 3.20: Representation of the essential modes of the graft trajectories. Black dots are relative to the wildtype trajectories, red dots are the one from the grafting candidates. Plots order is: first candidate (top left), third candidate (top right), fourth candidate (center left), fifth candidate (center right) and seventh candidate (bottom).

Experimental validation

All the 5 candidates was sent to Structural Biology Lab at University of Milan, to express and crystallize them.

To date, the first SAGE candidate (graft in position 66-85) crystallized (see Figure 3.21) and was sent to ESRF in Grenoble to perform X-ray scattering experiments to resolve its structure. Optimization of the crystal resolution conditions are currently ongoing. If the experimental validation will confirm the *in silico* 3D structure prediction, we plan to perfom immunological tests on patients sera using this superantigen as a probe.

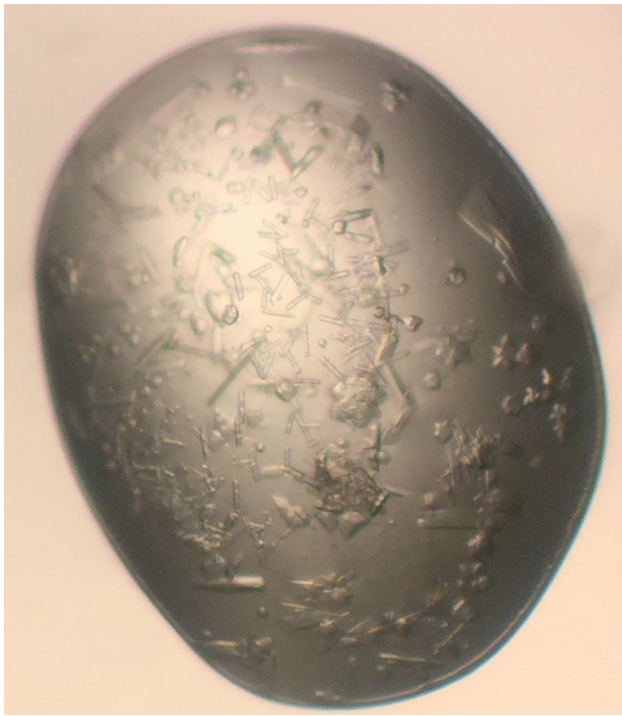


Figure 3.21: Crystal of the 1st SAGE-designed superantigen

Conclusions and future directions

"Ci fu una grande battaglia di idee e alla fine non ci furono né vincitori, né vinti, né idee."

STEFANO BENNI, Elianto

Computational modeling of proteins, despite the enormous progress made in both theoretical and technical approaches, remains an extremely challenging research field. This topic has been addressed in the present thesis from a theoretical point of view, applying statistical mechanics-based principles to molecular mechanics, and from a more applicative point of view studying the conformational and energetic properties of peptides with the aim to use them in a diagnostic tool and/or in a vaccine.

In the first chapter of the present thesis we proposed a path-independent free energy calculation technique, called Simplified Confinement Method, applying it to a new problem (single point mutants $\Delta\Delta G$ calculation) and enhancing its efficiency with respect to its original formulation. For the single-point mutants calculations, despite the very good agreement in thermostability order, we found a systematic drift in our free energy estimate, probably due to the oversimplified approximation applied to model denatured state (GXG tripeptide approximation). In the future, we plan to focus on this problem, trying to remove (or at least limit) the effect of this approximation.
In the SCM efficiency enhancement section, we exploited interpolation and extrapolation of confinement energies to sensibly reduce the computational effort needed to obtain reliable conformational free energy differences in biomolecules. After a proof of principle on alanine-n-peptides, we applied interpolation and extrapolation on a larger system, lactoferricin, showing the capability of the method to tackle also biologically-relevant biomolecules in a reasonable time.

In the second chapter, we studied non-equilibrium transitions in biomolecules using MD, correcting a force field following the principle of Maximum Caliber. In particular, we showed that, adding a bias which follows reference time series to an inaccurate force field, it is possible to obtain a good kinetic behavior also in unbiased collective

variables. Interestingly, in our simulation time becomes a bias parameter, and we also showed the possibility to "accelerate" the non-equilibrium simulation, keeping a good agreement with the experimental behavior of the system. As a last test, we biased our system with SAXS intensity data generated *in silico*, reproducing also in this case the non-equilibrium transition of the system. For the future, we plan to apply this technique using time-resolved data that comes from real experiments (NMR, SAXS, CD), directly integrating experimental information to MD.

In the first part of the third chapter we applied the principles of Structural Vaccinology to study two immunogenic peptides extracted from homologous proteins that comes from a bacterial family of pathogens to study their molecular determinants, with the aim to obtain a good diagnostic candidate which can be used in both the diseases. Our approach enlightened the role of small sequence mutations and the resulting structural flexibility change. The exploitation of this cross-reactivity can enhance the diagnostic capability of a tools as well as represent a good vaccine candidate.
In the second part, we presented an automated tool to graft linear immunogenic sequences onto foreign proteins, called SAGE. We applied this algorithm to graft epitopes on already immunogenic proteins, with the aim to obtain a broader response in the host. Currently we are waiting for experimental structural determination for our designed candidates and the subsequent tests to verify their immunogenic response. In the future, we plan to extend the reliability of SAGE predictions adding some new features, like coevolutionary evaluation of grafted sequences, or their solubility. Furthermore, we want to add the possibility to design grafting candidates starting from conformational epitopes.

Overall, my thesis shows the relevance and reach of atomistic simulations and theoretical methods in different realms of computational biophysics. I believe that the increase of structural and functional data we are currently witnessing will benefit from and spur the optimization and integration of methods such as the ones I developed into actual integrative biology approaches.
Pushing structure-based design and atomic resolution kinetic understanding of experiments into the analysis of biology will likely be the key to capture the physical basis of fundamental phenomena.

Appendices

APPENDIX A

Derivation of Simplified Confinement Method

A.1 Numerical integral calculation for TI

Numerical integration of the windows' average of the restraining energy was performed using the trapezoidal rule in a double log scale, fitting the consecutive points using a power law, as explained by Tyka *et al.* [12] and in Cecchini *et al.* [40].
Defining $y_i \equiv \frac{\langle U_{\text{HO}} \rangle_{\zeta_i}}{\zeta_i}$ and $x_i \equiv \zeta_i$, the power law is

$$y_i = a x_i^b \tag{A.1}$$

Given two consecutive points (x_i, y_i) and (x_j, y_j) with $j = i + 1$, the area A_i under the curve results

$$A_i = \int_{x_i}^{x_j} a x^b dx = \frac{a}{b+1} \left[x^{b+1} \right]_{x_i}^{x_j} = \frac{1}{b+1} (y_j x_j - y_i x_i) \tag{A.2}$$

And, computing the logarithm of the power law (A.1) for both the points, and subtracting one to the other, we can obtain an expression for the parameter b

$$\begin{cases} \log(y_i) & = \log a + b \log(x_i) \\ \log(y_j) & = \log a + b \log(x_j) \end{cases} \qquad \Rightarrow \qquad b = \frac{\log(y_j) - \log(y_i)}{\log(x_j) - \log(x_i)}$$

And the total free energy difference results

$$\Delta G_{\text{A} \to \text{A}^{\text{HO}}} = \sum_{i=1}^{n-1} A_i \tag{A.3}$$

where n is the total number of the simulated windows.

For the integral error calculation, we considered the statistical uncertainities on energy interpolation with MBAR reweighting [56] and applied the error propagation, considering null the error on the frequency. The error on every step of the integral then results

$$\begin{aligned} \sigma_{A_i} &= \sqrt{\left(\frac{\partial A_i}{\partial y_i} \right)^2 \sigma_{y_i}^2 + \left(\frac{\partial A_i}{\partial y_j} \right)^2 \sigma_{y_j}^2} \\ &= \sigma_{y_{i+1}} \left(k \cdot l \cdot \left(x_{i+1}^2 - \frac{l \cdot m}{y_{i+1}^2} \right) \right)^2 + \sigma_{y_i} \left(k \cdot l \cdot \left(x_i^2 - \frac{l \cdot m}{y_i^2} \right) \right)^2 \end{aligned} \tag{A.4}$$

where the parameters k, l, and m are

$$\begin{aligned} k &= \log(x_{i+1}) - \log(x_i), \\ l &= \frac{1}{\log(y_{i+1}) - \log(y_i) - k}, \\ m &= y_{i+1}x_{i+1} - y_i x_i. \end{aligned}$$

A.2 Derivation of roto-translational free energy

In the confined state, as we explained in section 1.3, the free energy of the system is represented only by entropy contributions. The roto-translational free energy for a N molecules in a box is defined as

$$G_{\text{rot+tr}} = -k_B T \log Z = -k_B T \log\left(\frac{q^N}{N!}\right)$$

where $q = q_{\text{rot}} + q_{\text{tr}}$.
Applying the Stirling approximation

$$\begin{aligned} G_{\text{rot+tr}} &= -k_B T \log\left(\frac{e\, q_{\text{rot}} q_{\text{tr}}}{N}\right) \\ &= -k_B T \log\left(\frac{e\, q_{\text{tr}}}{N}\right) - k_B T \log\left(q_{\text{rot}}\right) \\ &= G_{\text{tr}} + G_{\text{rot}} \end{aligned}$$

For the translational part we have [179]

$$\begin{aligned} G_{\text{tr}} &= -N k_B T \left(\log\left(\frac{V}{N}\right) + 1 + \frac{3}{2}\log\left(\frac{2\pi m k_B T}{h^2}\right)\right) \\ &= -N k_B T \left(-\log\left(\frac{N}{V}\right) + 1 + \frac{3}{2}\log\left(\frac{2\pi m k_B T}{h^2}\right)\right) \end{aligned} \tag{A.5}$$

where N/V is the molecule concentration in the box, and we consider a 1 M standard concentration. Note that here we assume that the ideal gas equation holds, and thus there is no interactions between molecules (in our case we simulate only one molecule, without periodic boundary conditions).

For the rotational part we have [179]

$$G_{\text{rot}} = -N k_B T \left(\frac{1}{2}\log\left(\pi I_1 I_2 I_3\right) + \frac{3}{2}\log\left(\frac{8\pi^2 k_B T}{h^2}\right) - \log(\sigma)\right) \tag{A.6}$$

where I_i are the 3 moments of inertia along the principal axes and σ is the symmetry number, which is equal to 1 for asymmetrical molecules.
The moments of inertia are computed on the confined trajectories using the `g_gyrate` routine of the GROMACS suite, or the `coor` module in CHARMM.

Considering two systems A and B in their confined states, the roto-translational free energy difference results

$$\Delta G^{\text{rot+tr}}_{\text{A}^{\text{HO}}\rightarrow\text{B}^{\text{HO}}} = -\frac{1}{2}Nk_BT\left(\log\left(\frac{I_1^{\text{B}}I_2^{\text{B}}I_3^{\text{B}}}{I_1^{\text{A}}I_2^{\text{A}}I_3^{\text{A}}}\right)+3\log\left(\frac{m_{\text{B}}}{m_{\text{A}}}\right)\right) \tag{A.7}$$

APPENDIX B

Derivation of Maximum Caliber

B.1 Proof of the equivalence between pMaxCal and biased replica simulations

If $\{\gamma\}$ is the set of stochastic trajectories of the N–particle system starting at point r_0, we are interested in the probability $p(\gamma)$. Since computational algorithms usually produce trajectories over discrete time steps and kinetic experiments are usually recorded at discrete times, we assume that $p(\gamma) = p(r_0, r_1, ..., r_T)$. Defining f_t^{exp} as the experimental time course of some conformational property of the system, and $f(r)$ the associated forward model, and assuming to know the diffusion coefficient D of the system, the principle of maximum caliber requires that $p(r_0, r_1, ..., r_T)$ maximizes

$$S[p] = -\sum_{r_1 r_2 ... r_T} p(r_0, r_1, ..., r_T) \log p(r_0, r_1, ..., r_T) \tag{B.1}$$

with the constraints

$$\sum_{r_1 r_2 ... r_T} p(r_0, r_1, ..., r_T) f(r_t) = f_t^{exp} \tag{B.2}$$

and

$$\frac{1}{2\Delta t} \sum_{r_1 r_2 ... r_T} p(r_0, r_1, ..., r_T)[r_{t+1} - r_t]^2 = D \tag{B.3}$$

at each discrete time t, and that $\sum p(r_0, r_1, ..., r_T) = 1$. The constrained maximization gives

$$p(r_0, r_1, ..., r_T) = \frac{1}{Z_d} \exp\left[-\sum_t \left(\nu_t [r_{t+1} - r_t]^2 + \lambda_t f(r_t)\right)\right], \tag{B.4}$$

where Z_d is the normalization constant and ν_t is the set of Lagrange multipliers which implement the average of Eq. (B.3) and λ_t that implementing Eq. (B.2). In principle, λ_t can be obtained by $d(\log Z_d)/d\lambda_t = f_t^{exp}$, but in practice this is useless because it is an implicit equation and it involves the sum Z_d over all possible paths.

It is useful to extend the expression found in Eq. (B.4) in two ways, which so far are just formally correct, and whose use will be clear later. First, let us consider n independent,

identical replicas of the system, each defined by coordinates $\{r_t^\alpha\}$ with $\alpha = 1, ..., n$ and $t = 0, ..., T$. The maximum-caliber probability distribution can be easily extendend to

$$p(\{r_t^\alpha\}) = \frac{1}{Z_d} \exp\left[-\sum_{t,\alpha} \left(\nu_t^\alpha [r_{t+1}^\alpha - r_t^\alpha]^2 + \lambda_t^\alpha f(r_t^\alpha) \right) \right]. \tag{B.5}$$

Moreover, one can require that

$$\sum_{\{r_t^\alpha\}} p(\{r_t^\alpha\}) \left[\frac{1}{n} \sum_\alpha f(r_t^\alpha) - f_t^{exp} \right]^2 = \sigma_{nt}^2, \tag{B.6}$$

that is the standard error of the *average* of f over the replicas is some value σ_n. For sake of compactness, let's define

$$\xi_t \equiv \frac{1}{n} \sum_\beta f(r_t^\beta) - f_t^{exp}, \tag{B.7}$$

implying that the experimental data are matched if $\xi_t = 0$ for all t. Applying the Lagrange–multipliers method also to this constrain, the maximum–caliber distribution becomes

$$p(\{r_t^\alpha\}) = \frac{1}{Z_d} \exp\left[-\sum_{t,\alpha} \left(\nu_t^\alpha [r_{t+1}^\alpha - r_t^\alpha]^2 + \lambda_t^\alpha f(r_t^\alpha) + \mu_{nt}^\alpha \xi_t^2 \right) \right]. \tag{B.8}$$

In the limit $n \to \infty$, $\sigma_{nt} \to 0$ for every t because of the law of large numbers, and consequently one can set $\mu_{nt}^\alpha \to \infty$ for each t and α. In particular, $\sigma_n \sim n^{-1/2}$ and consequently $\mu_{nt}^\alpha \sim \log n$.
To generate trajectories distributed according to Eq. (B.4), we follow a computational strategy based on coupled replica MD simulations, similarly to what done to correct force fields to match equilibrium data [86]. The replicas are controlled by the time-dependent potential

$$U(\{r^\alpha\}, t) = \frac{nk}{2} \left(\frac{1}{n} \sum_\alpha f(r^\alpha) - f_t^{exp} \right)^2, \tag{B.9}$$

where r^α is the conformation of the system in the replica α, n is the number of replicas and k is an harmonic constant. The associated stochastic process in the $(3N \times n)$–dimensional replica space can be regarded as a Markov chain

$$p_n(\{r_t^\alpha\}) = p_N(r_0^\alpha) w(r_0^\alpha \to r_1^\alpha) w(r_1^\alpha \to r_2^\alpha) ... w(r_{T-1}^\alpha \to r_T^\alpha) \tag{B.10}$$

which can be written according to the simplest form of the Onsager–Machlup function, corresponding to an over–damped stochastic dynamics discretized according to Ito prescription [180]

$$p_n(\{r_t^\alpha\}) = c \cdot \exp\left[-\sum_{t\alpha} \frac{\left(r_{t+1}^\alpha - r_t^\alpha + k\Delta t \xi_t \right)^2}{2D'\Delta t} \right], \tag{B.11}$$

recalling that by definition the initial point r_0^α is fixed for all replicas. Here the diffusion coefficient is $D' = T/\gamma'$, where γ' is the friction coefficient chosen as an input of the simulation. In the limit of large k this can be approximated as

$$p_n(\{r_t^\alpha\}) = c \cdot \exp\left[-\sum_{t\alpha} \frac{\left(r_{t+1}^\alpha - r_t^\alpha\right)^2}{2D'\Delta t}\right] \cdot \prod_t \delta\left(\xi_t\right) \tag{B.12}$$

because of the definition of Dirac's delta, that is for any distribution $\varphi(\xi)$ and any t

$$c\int d\xi_t \ \exp\left[-\sum_\alpha \frac{\left(r_{t+1}^\alpha - r_t^\alpha + k\Delta t\xi_t\right)^2}{2D'\Delta t}\right]\varphi(\xi_t) = c\cdot\exp\left[-\sum_\alpha \frac{\left(r_{t+1}^\alpha - r_t^\alpha\right)^2}{2D'\Delta t}\right]\cdot\varphi(0) \tag{B.13}$$

in the limit $k \to \infty$.

Equation (B.12) can be rewritten multiplying its r.h.s. by the exponential of a linear function of ξ_t, that is which is equivalent to

$$p_n(\{r_t^\alpha\}) = c \cdot \exp\left[-\sum_{t\alpha} \frac{\left(r_{t+1}^\alpha - r_t^\alpha\right)^2}{2D'\Delta t} - \sum_t \gamma_t\xi_t\right] \cdot \prod_t \delta\left(\xi_t\right) \tag{B.14}$$

for any γ_t. In fact, for any distribution $\varphi(\xi)$ and any t

$$\begin{aligned} &c\int d\xi_t \ \exp\left[-\sum_\alpha \frac{\left(r_{t+1}^\alpha - r_t^\alpha\right)^2}{2D'\Delta t}\right]\delta\left(\xi_t\right)\varphi(\xi_t) = \\ &= c\int d\xi_t \ \exp\left[-\sum_\alpha \frac{\left(r_{t+1}^\alpha - r_t^\alpha\right)^2}{2D'\Delta t} - \gamma_t\xi_t\right]\delta\left(\xi_t\right)\varphi(\xi_t), \end{aligned} \tag{B.15}$$

meaning that Eq. (B.12) is equivalent to Eq. (B.14).

Using the Gaussian representation of Dirac's delta $\delta(\xi_t) = \lim_{\kappa\to\infty}\exp(-\kappa\xi_t^2)$, Eq. (B.14) becomes

$$p_n(\{r_t^\alpha\}) = c \cdot \exp\left[-\sum_{t\alpha} \frac{\left(r_{t+1}^\alpha - r_t^\alpha\right)^2}{2D'\Delta t} - \sum_t \gamma_t\xi_t - \sum_t \kappa_t\left(\xi_t\right)^2\right] \tag{B.16}$$

in the limit $\kappa_t \to \infty$ for any t. Choosing $\gamma_t = \lambda_t$, remembering that both μ_{nt}^α and $\kappa_t \to \infty$ for large k, then Eq. (B.16) is equivalent to the maximum–caliber distribution of Eq. (B.8). However, there is a further difficulty involving the diffusion coefficient. If the experimental data are not taken into account, i.e. $\lambda_t^\alpha = \mu_{nt}^\alpha = 0$, then the partition function in Eq. (B.8) is a Gaussian integral and the condition $\partial \log Z_d/\partial\nu_t^\alpha = D$ defining the Lagrange multipliers gives $\nu_t^\alpha = 1/D$ and thus $D = D'$. In this case, the diffusion coefficient used as an input to the replica simulation is the same required by the maximum–caliber principle.

On the other hand, if one accounts for the experimental data, then $\nu_t^\alpha \neq 1/D$ and the simulated diffusion of the particles becomes different from that required by the principle of maximum caliber. If the constraining effect of the experimental data is mild, one

can expect that λ_t^α are small and the dynamical partition function in Eq. (B.8) can be approximated as

$$Z_d = \sum_{\{r_t^\alpha\}} \exp\left[-\sum_{t,\alpha} \nu_t^\alpha [r_{t+1}^\alpha - r_t^\alpha]^2\right] \left(1 - \sum_{t,\alpha} \lambda_t^\alpha f(r_t^\alpha)\right), \tag{B.17}$$

and consequently to the first order in λ_t^α

$$D = \frac{1}{\nu_t^\alpha} - \lambda_t^\alpha \frac{\partial}{\partial \nu_t^\alpha} \langle f(r_t^\alpha)\rangle_d \,, \tag{B.18}$$

where $\langle\cdot\rangle_d$ is the unperturbed average over paths. Comparing this with Eq. (B.16) gives

$$D = D' + \lambda_t^\alpha (D')^2 \frac{\partial}{\partial D'} \langle f(r_t^\alpha)\rangle_d \,, \tag{B.19}$$

suggesting that the actual diffusion coefficient is modified by the bias with respect to that used as input to the simulation.

B.2 $U_{\text{tail}} \rightarrow U_{\text{head}}$ results

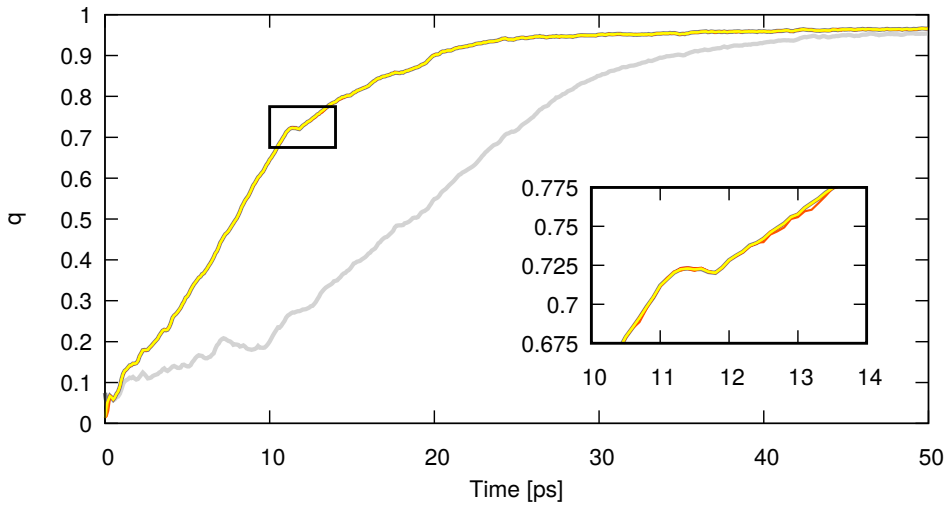


Figure B.1: Average fraction of native contacts (q) in function of time for unbiased U_{head} (dark grey), unbiased U_{tail} (light grey, covered by caliber-restrained simulations), and caliber-restrained simulations from U_{tail} to U_{head}, from 4 (red) to 128 replicas (yellow) in color scale.

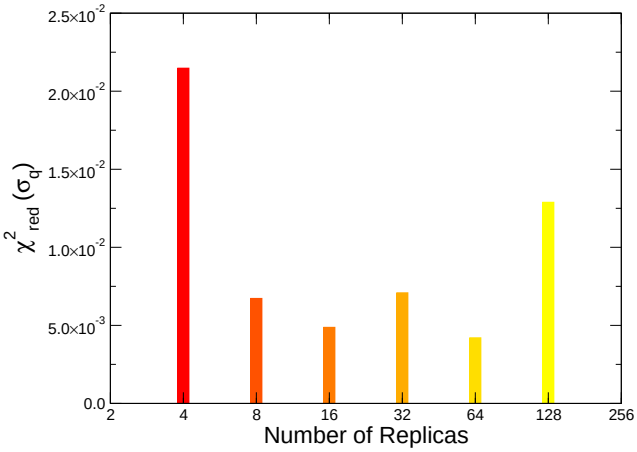


Figure B.2: χ^2_{red} between experimental U_{tail} and biased U_{tail} to U_{head} fluctuations of fraction of native contacts (q) during the first half of the simulation. Despite some small differences in χ^2_{red} value, all the simulations are almost identical to the original one and the number of replicas does not change the result of the bias

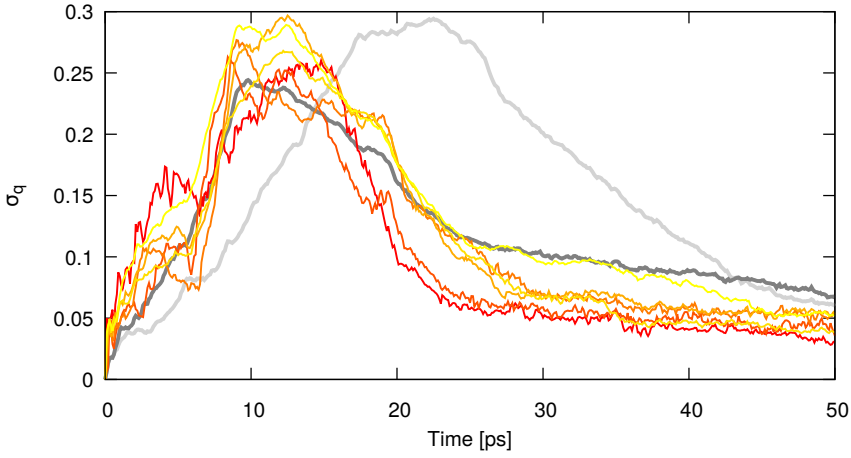


Figure B.3: Fluctuations of average fraction of native contacts (q) in function of time for unbiased U_{head} (dark grey), unbiased U_{tail} (light grey), and caliber restrained simulations from U_{tail} to U_{head}, from 4 (red) to 128 replicas (yellow) in color scale.

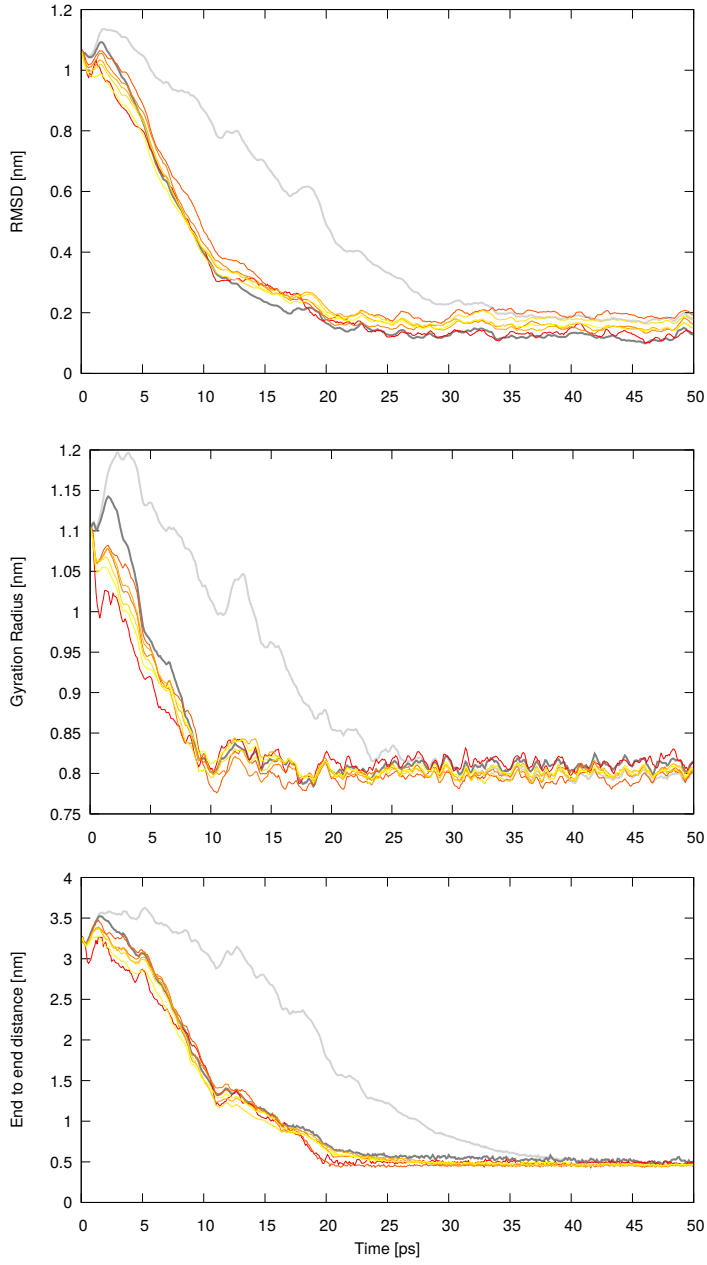


Figure B.4: Average C_α-RMSD (top), gyration radius (center), and end-to-end distance (bottom) in function of time for unbiased U_{head} (dark grey), unbiased U_{tail} (light grey), and caliber restrained simulations from U_{tail} to U_{head}, from 4 (red) to 128 replicas (yellow) in color scale.

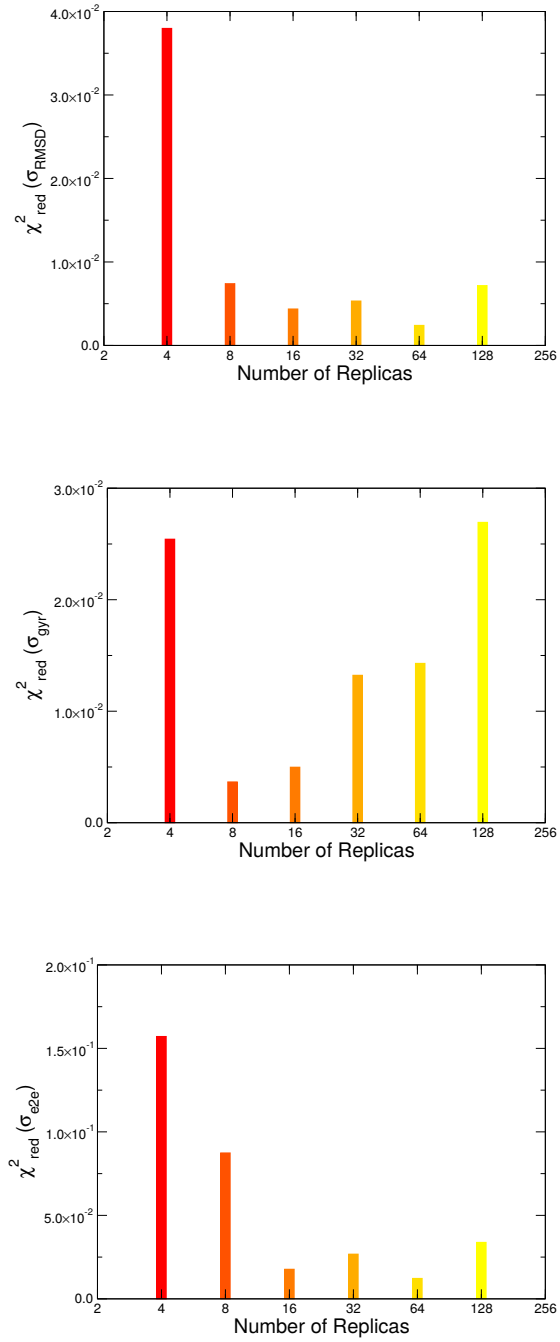


Figure B.5: χ^2_{red} between time series from experimental U_{tail} and biased U_{tail} to U_{head} fluctuations of RMSD (left), gyration radius (center), and end to end distance during the first half of the simulation.

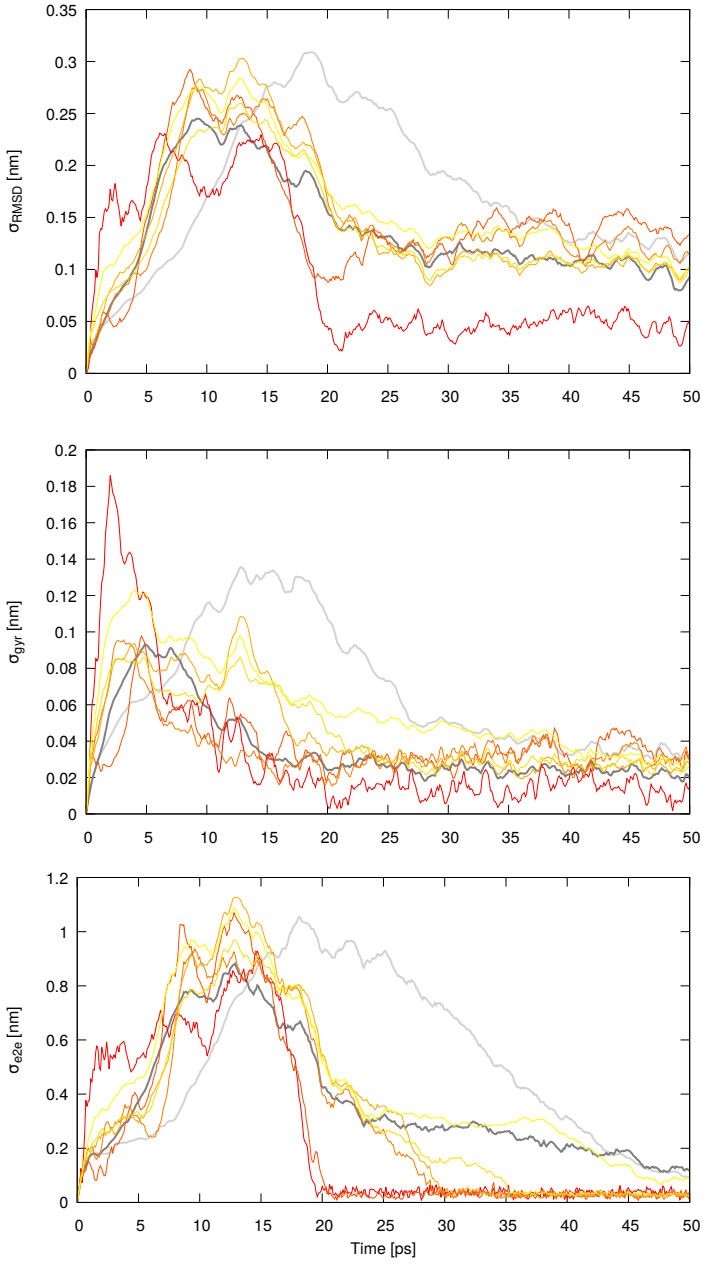


Figure B.6: Fluctuations of average C_α-RMSD (top), gyration radius (center), and end-to-end distance (bottom) in function of time for unbiased U_{head} (dark grey), unbiased U_{tail} (light grey), and caliber restrained simulations from U_{tail} to U_{head}, from 4 (red) to 128 replicas (yellow) in color scale.

APPENDIX C

Complete data from SAGE predictions

Rank	Insertion position	Exposure score	Structure score Raw	Structure score Weighted	Epitope score Raw	Epitope score Weighted	Average score
1	56-61	0.679	0.937	0.636	1.000	0.679	0.657
2*	48-53	0.880	0.627	0.552	0.344	0.303	0.427
3	75-80	0.831	0.826	0.687	0.199	0.165	0.426
4	80-85	0.724	0.562	0.407	0.599	0.434	0.421
5	61-66	0.461	1.000	0.462	0.738	0.341	0.402
6	118-123	0.676	0.801	0.541	0.316	0.213	0.377
7	86-91	0.649	0.638	0.414	0.393	0.255	0.335
8	4-9	0.531	0.890	0.472	0.340	0.180	0.326
9	70-75	0.704	0.708	0.499	0.218	0.153	0.326
10	94-99	1.000	0.367	0.367	0.278	0.278	0.322
11	44-49	0.610	0.828	0.505	0.205	0.125	0.315
12	12-17	0.410	0.792	0.325	0.632	0.259	0.292
13	131-136	0.607	0.715	0.434	0.160	0.097	0.265
14	149-154	0.548	0.642	0.352	0.227	0.125	0.238
15	89-94	0.508	0.455	0.264	0.345	0.200	0.232
16	29-34	0.531	0.545	0.289	0.307	0.163	0.226
17	106-111	0.535	0.655	0.350	0.166	0.089	0.220
18	71-76	0.707	0.295	0.209	0.274	0.194	0.201
19	43-48	0.494	0.363	0.179	0.164	0.081	0.130

Table C.1: SAGE prediction for 2F5 epitope (residues 661-666) on 1WNU scaffold. Candidate marked with $*$ is the one chosen in the original paper.

Rank	Insertion position	Exposure score	Structure score Raw	Structure score Weighted	Epitope score Raw	Epitope score Weighted	Average score
1	56-62	0.631	0.861	0.543	1.000	0.631	0.587
2	80-86	0.737	0.680	0.501	0.565	0.416	0.459
3	135-141	0.695	1.000	0.695	0.232	0.161	0.428
4	86-92	0.651	0.453	0.295	0.551	0.359	0.327
5	73-79	0.874	0.558	0.488	0.180	0.157	0.323
6*	48-54	0.876	0.415	0.364	0.316	0.277	0.320
7	118-124	0.628	0.670	0.420	0.325	0.204	0.312
8	70-76	0.719	0.623	0.448	0.194	0.140	0.294
9	94-100	1.000	0.305	0.305	0.260	0.260	0.283
10	4-10	0.510	0.764	0.390	0.341	0.174	0.282
11	130-136	0.610	0.771	0.471	0.152	0.093	0.282
12	75-81	0.844	0.387	0.326	0.270	0.228	0.277
13	12-18	0.405	0.728	0.295	0.624	0.253	0.274
14	44-50	0.584	0.646	0.377	0.202	0.118	0.248
15	32-38	0.822	0.380	0.312	0.217	0.179	0.245
16	106-112	0.534	0.724	0.387	0.165	0.088	0.238
17	29-35	0.541	0.512	0.277	0.278	0.150	0.214
18	119-125	0.530	0.509	0.270	0.274	0.145	0.208
19	71-77	0.709	0.255	0.181	0.250	0.177	0.179
20	148-154	0.466	0.281	0.131	0.320	0.149	0.140
21	105-111	0.490	0.354	0.173	0.156	0.077	0.125
22	42-48	0.452	0.337	0.152	0.167	0.075	0.114

Table C.2: SAGE prediction for 2F5 epitope (residues 661-667) on 1WNU scaffold. Candidate marked with $*$ is the one chosen in the original paper.

Rank	Insertion position	Exposure score	Structure score Raw	Structure score Weighted	Epitope score Raw	Epitope score Weighted	Average score
1	57-62	0.421	1.000	0.421	1.000	0.421	0.421
2	81-86	0.482	0.654	0.316	0.572	0.276	0.296
3	94-99	1.000	0.355	0.355	0.209	0.209	0.282
4*	49-54	0.578	0.523	0.302	0.362	0.209	0.256
5	87-92	0.481	0.475	0.229	0.375	0.180	0.204
6	13-18	0.287	0.589	0.169	0.735	0.211	0.190
7	48-53	0.607	0.290	0.176	0.300	0.182	0.179
8	45-50	0.475	0.556	0.264	0.181	0.086	0.175
9	119-124	0.416	0.539	0.224	0.262	0.109	0.167
10	76-81	0.597	0.300	0.179	0.242	0.145	0.162
11	107-112	0.411	0.599	0.246	0.145	0.059	0.153
12	95-100	0.654	0.240	0.157	0.215	0.141	0.149
13	90-95	0.401	0.384	0.154	0.234	0.094	0.124
14	28-33	0.398	0.358	0.142	0.209	0.083	0.113

Table C.3: SAGE prediction for 2F5 epitope (residues 662-667) on 1WNU scaffold. Candidate marked with $*$ is the one chosen in the original paper.

Rank	Insertion position	Exposure score	Structure score		Epitope score		Average score
			Raw	Weighted	Raw	Weighted	
1	71-76	0.587	1.000	0.587	0.744	0.437	0.512
2	145-150	0.630	0.573	0.361	1.000	0.630	0.496
3	25-30	0.932	0.749	0.698	0.293	0.274	0.486
4	84-89	0.827	0.516	0.427	0.606	0.501	0.464
5*	85-90	1.000	0.297	0.297	0.578	0.578	0.438
6	41-46	0.739	0.369	0.272	0.789	0.583	0.427
7	69-74	0.727	0.459	0.334	0.670	0.487	0.411
8	122-127	0.580	0.540	0.313	0.713	0.414	0.363
9	110-115	0.717	0.506	0.362	0.342	0.245	0.304
10	46-51	0.513	0.346	0.178	0.837	0.429	0.304
11	83-88	0.682	0.282	0.192	0.599	0.408	0.300
12	64-69	0.491	0.884	0.434	0.286	0.140	0.287
13	129-134	0.512	0.452	0.231	0.548	0.281	0.256
14	27-32	0.760	0.288	0.219	0.366	0.278	0.248
15	137-142	0.556	0.474	0.263	0.386	0.214	0.239
16	126-131	0.435	0.553	0.240	0.512	0.223	0.232
17	114-119	0.521	0.538	0.280	0.342	0.178	0.229
18	23-28	0.653	0.390	0.255	0.278	0.182	0.218
19	101-106	0.466	0.576	0.268	0.358	0.167	0.217
20	125-130	0.445	0.372	0.166	0.558	0.248	0.207
21	24-29	0.734	0.264	0.194	0.282	0.207	0.200
22	63-68	0.461	0.509	0.235	0.258	0.119	0.177

Table C.4: SAGE prediction for 2F5 epitope (residues 661-666) on 2CX5 scaffold. Candidate marked with $*$ is the one chosen in the original paper.

Rank	Insertion position	Exposure score	Structure score		Epitope score		Average score
			Raw	Weighted	Raw	Weighted	
1	31-37	0.854	0.965	0.825	1.000	0.854	0.839
2	25-31	0.894	0.881	0.787	0.176	0.158	0.473
3	71-77	0.627	1.000	0.627	0.406	0.255	0.441
4	84-90	0.908	0.530	0.481	0.365	0.332	0.407
5	145-151	0.683	0.600	0.410	0.472	0.322	0.366
6	69-75	0.760	0.549	0.417	0.387	0.294	0.356
7*	85-91	1.000	0.319	0.319	0.354	0.354	0.336
8	42-48	0.673	0.365	0.246	0.584	0.393	0.319
9	43-49	0.599	0.453	0.271	0.498	0.298	0.285
10	122-128	0.590	0.560	0.330	0.391	0.231	0.281
11	93-99	0.712	0.458	0.326	0.267	0.190	0.258
12	64-70	0.533	0.753	0.401	0.171	0.091	0.246
13	83-89	0.736	0.314	0.231	0.351	0.258	0.244
14	110-116	0.717	0.484	0.347	0.186	0.133	0.240
15	137-143	0.580	0.534	0.309	0.233	0.135	0.222
16	129-135	0.528	0.469	0.247	0.307	0.162	0.205
17	23-29	0.692	0.404	0.280	0.164	0.114	0.197
18	101-107	0.501	0.568	0.285	0.209	0.105	0.195
19	126-132	0.459	0.513	0.235	0.285	0.131	0.183
20	54-60	0.614	0.395	0.243	0.183	0.112	0.178
21	63-69	0.488	0.556	0.271	0.153	0.075	0.173
22	114-120	0.537	0.423	0.227	0.205	0.110	0.169
23	102-108	0.495	0.302	0.150	0.210	0.104	0.127
24	52-58	0.549	0.260	0.142	0.193	0.106	0.124

Table C.5: SAGE prediction for 2F5 epitope (residues 661-667) on 2CX5 scaffold. Candidate marked with $*$ is the one chosen in the original paper.

Rank	Insertion position	Exposure score	Structure score Raw	Structure score Weighted	Epitope score Raw	Epitope score Weighted	Average score
1	32-37	1.000	1.000	1.000	1.000	1.000	1.000
2	85-90	0.969	0.421	0.408	0.344	0.333	0.371
3	72-77	0.588	0.795	0.467	0.459	0.270	0.369
4	70-75	0.750	0.478	0.358	0.387	0.290	0.324
5	26-31	0.746	0.676	0.504	0.171	0.127	0.316
6*	86-91	0.787	0.325	0.256	0.428	0.337	0.296
7	94-99	0.605	0.613	0.371	0.246	0.149	0.260
8	42-47	0.708	0.239	0.169	0.389	0.275	0.222
9	123-128	0.486	0.541	0.263	0.358	0.174	0.219
10	65-70	0.488	0.599	0.292	0.176	0.086	0.189
11	138-143	0.539	0.432	0.233	0.238	0.128	0.180
12	130-135	0.474	0.445	0.211	0.311	0.147	0.179
13	47-52	0.481	0.314	0.169	0.420	0.202	0.176
14	115-120	0.480	0.440	0.211	0.229	0.110	0.161
15	25-30	0.828	0.222	0.184	0.163	0.135	0.159
16	106-111	0.573	0.294	0.168	0.238	0.136	0.152
17	55-60	0.584	0.283	0.165	0.173	0.101	0.133
18	102-107	0.412	0.430	0.177	0.202	0.083	0.130
19	126-131	0.402	0.275	0.110	0.288	0.116	0.113
20	103-108	0.446	0.255	0.114	0.215	0.096	0.105

Table C.6: SAGE prediction for 2F5 epitope (residues 662-667) on 2CX5 scaffold. Candidate marked with $*$ is the one chosen in the original paper.

Rank	Insertion position	Exposure score	Structure score Raw	Structure score Weighted	Epitope score Raw	Epitope score Weighted	Average score
1*	68-77	1.000	0.756	0.756	0.446	0.446	0.601
2	27-36	0.995	0.669	0.665	0.489	0.487	0.576
3	110-119	0.708	0.571	0.404	1.000	0.708	0.556
4	32-41	0.787	0.996	0.784	0.344	0.271	0.527
5	132-141	0.777	0.874	0.679	0.360	0.279	0.479
6	88-97	0.695	0.877	0.609	0.447	0.310	0.460
7	36-45	0.697	0.970	0.676	0.342	0.238	0.457
8	133-142	0.869	0.675	0.586	0.376	0.327	0.457
9	116-125	0.684	0.609	0.417	0.721	0.494	0.455
10	94-103	0.735	0.485	0.357	0.690	0.508	0.432
11	72-81	0.853	0.654	0.557	0.345	0.294	0.426
12	42-51	0.630	1.000	0.630	0.333	0.210	0.420
13	52-61	0.704	0.730	0.514	0.432	0.304	0.409
14	82-91	0.679	0.691	0.470	0.490	0.333	0.401
15	126-135	0.718	0.711	0.510	0.340	0.244	0.377
16	47-56	0.625	0.777	0.485	0.341	0.213	0.349

Table C.7: SAGE prediction for 4E10 epitope (residues 671-680) on 1EZ3 scaffold. Candidate marked with $*$ is the one chosen in the original paper.

Rank	Insertion position	Exposure score	Structure score		Epitope score		Average score
			Raw	Weighted	Raw	Weighted	
1	144-153	1.000	0.510	0.510	0.896	0.896	0.703
2	106-115	0.796	0.498	0.397	0.774	0.616	0.506
3*	149-158	0.669	0.447	0.299	1.000	0.669	0.484
4	33-42	0.513	0.906	0.465	0.813	0.417	0.441
5	124-133	0.482	0.825	0.397	0.987	0.476	0.436
6	70-79	0.645	0.585	0.377	0.693	0.447	0.412
7	50-59	0.482	1.000	0.482	0.707	0.341	0.412
8	71-80	0.574	0.481	0.276	0.748	0.429	0.352
9	16-25	0.504	0.884	0.446	0.470	0.237	0.341
10	131-140	0.503	0.780	0.392	0.541	0.272	0.332
11	15-24	0.557	0.721	0.401	0.468	0.261	0.331
12	109-118	0.548	0.530	0.291	0.663	0.363	0.327
13	76-85	0.489	0.596	0.291	0.594	0.290	0.291
14	10-19	0.541	0.437	0.237	0.597	0.323	0.280
15	161-170	0.522	0.486	0.254	0.573	0.299	0.276
16	18-27	0.516	0.458	0.236	0.557	0.287	0.262
17	132-141	0.485	0.506	0.245	0.543	0.263	0.254
18	75-84	0.442	0.508	0.225	0.618	0.273	0.249
19	171-180	0.499	0.390	0.195	0.541	0.270	0.232

Table C.8: SAGE prediction for 4E10 epitope (residues 671-680) on 1IS1 scaffold. Candidate marked with $*$ is the one chosen in the original paper.

Rank	Insertion position	Exposure score	Structure score		Epitope score		Average score
			Raw	Weighted	Raw	Weighted	
1	144-153	1.000	0.465	0.465	0.487	0.487	0.476
2	68-77	0.485	0.553	0.268	1.000	0.485	0.376
3*	149-158	0.716	0.475	0.340	0.519	0.372	0.356
4	33-42	0.425	0.958	0.407	0.461	0.196	0.301
5	106-115	0.709	0.424	0.301	0.423	0.300	0.300
6	16-25	0.452	1.000	0.452	0.261	0.118	0.285
7	53-62	0.470	0.633	0.297	0.531	0.250	0.273
8	104-113	0.520	0.515	0.268	0.466	0.242	0.255
9	124-133	0.411	0.713	0.293	0.526	0.216	0.254
10	2-11	0.527	0.237	0.125	0.554	0.292	0.208
11	119-128	0.430	0.555	0.239	0.402	0.173	0.206
12	71-80	0.475	0.477	0.227	0.385	0.183	0.205
13	109-118	0.505	0.443	0.224	0.366	0.185	0.204
14	159-168	0.447	0.642	0.287	0.262	0.117	0.202
15	74-83	0.413	0.624	0.258	0.299	0.124	0.191
16	18-27	0.500	0.397	0.199	0.314	0.157	0.178
17	167-176	0.439	0.507	0.223	0.286	0.126	0.174
18	76-85	0.425	0.468	0.199	0.329	0.140	0.169
19	153-162	0.443	0.389	0.172	0.366	0.162	0.167
20	162-171	0.411	0.470	0.193	0.295	0.121	0.157
21	132-141	0.432	0.426	0.184	0.299	0.129	0.157
22	13-22	0.419	0.394	0.165	0.331	0.139	0.152
23	157-166	0.458	0.286	0.131	0.372	0.171	0.151
24	164-173	0.435	0.372	0.162	0.317	0.138	0.150
25	75-84	0.382	0.426	0.163	0.335	0.128	0.146

Table C.9: SAGE prediction for 4E10 epitope (residues 671-680) on 1ISE scaffold. Candidate marked with $*$ is the one chosen in the original paper.

Rank	Insertion position	Exposure score	Structure score		Epitope score		Average score
			Raw	Weighted	Raw	Weighted	
1	108-117	1.000	0.526	0.526	1.000	1.000	0.763
2	135-144	0.810	1.000	0.810	0.304	0.246	0.528
3	110-119	0.764	0.532	0.407	0.753	0.575	0.491
4	34-43	0.879	0.757	0.666	0.318	0.279	0.473
5	147-156	0.824	0.742	0.612	0.380	0.314	0.463
6	184-193	0.725	0.771	0.559	0.308	0.224	0.392
7	111-120	0.573	0.582	0.333	0.620	0.355	0.344
8	39-48	0.836	0.509	0.425	0.308	0.257	0.341
9	31-40	0.652	0.641	0.418	0.393	0.256	0.337
10	40-49	0.675	0.698	0.471	0.283	0.191	0.331
11	187-196	0.626	0.757	0.474	0.296	0.186	0.330
12	181-190	0.736	0.512	0.377	0.375	0.276	0.327
13	161-170	0.610	0.639	0.390	0.399	0.243	0.316
14*	149-158	0.641	0.628	0.403	0.349	0.224	0.313
15	30-39	0.578	0.641	0.370	0.400	0.231	0.301
16	38-47	0.636	0.653	0.415	0.292	0.185	0.300
17	191-200	0.553	0.753	0.417	0.312	0.173	0.295
18	186-195	0.704	0.449	0.316	0.344	0.242	0.279
19	127-136	0.677	0.509	0.345	0.309	0.209	0.277
20	46-55	0.712	0.392	0.279	0.375	0.267	0.273
21	83-92	0.508	0.634	0.323	0.440	0.224	0.273
22	81-90	0.468	0.645	0.302	0.482	0.225	0.264
23	94-103	0.446	0.827	0.369	0.279	0.125	0.247
24	122-131	0.511	0.584	0.298	0.289	0.148	0.223

Table C.10: SAGE prediction for 4E10 epitope (residues 671-680) on 1VI7 scaffold. Candidate marked with $*$ is the one chosen in the original paper.

Rank	Insertion position	Exposure score	Structure score		Epitope score		Average score
			Raw	Weighted	Raw	Weighted	
1*	111-120	1.000	0.901	0.901	0.800	0.800	0.850
2	88-97	0.782	0.927	0.725	1.000	0.782	0.753
3	139-148	0.743	0.992	0.737	0.640	0.475	0.606
4	38-47	0.761	0.829	0.631	0.763	0.581	0.606
5	114-123	0.759	0.952	0.723	0.635	0.482	0.603
6	121-130	0.752	1.000	0.752	0.576	0.433	0.593
7	128-137	0.761	0.855	0.651	0.677	0.515	0.583
8	44-53	0.646	0.970	0.627	0.751	0.486	0.556
9	19-28	0.664	0.975	0.648	0.586	0.389	0.519
10	133-142	0.694	0.797	0.553	0.655	0.454	0.504
11	40-49	0.618	0.981	0.606	0.611	0.377	0.492
12	25-34	0.649	0.821	0.533	0.574	0.372	0.453
13	42-51	0.654	0.681	0.446	0.701	0.459	0.452
14	119-128	0.674	0.737	0.497	0.552	0.372	0.434

Table C.11: SAGE prediction for 4E10 epitope (residues 671-680) on 1XIZ scaffold. Candidate marked with $*$ is the one chosen in the original paper.

Rank	Insertion position	Exposure score	Structure score Raw	Structure score Weighted	Epitope score Raw	Epitope score Weighted	Average score
1	119-128	0.785	0.813	0.638	0.590	0.463	0.550
2*	138-147	1.000	0.691	0.691	0.335	0.335	0.513
3	22-31	0.851	0.686	0.584	0.396	0.337	0.460
4	33-42	0.747	0.503	0.376	0.676	0.505	0.440
5	3-12	0.729	0.827	0.603	0.370	0.270	0.436
6	6-15	0.626	1.000	0.626	0.365	0.229	0.427
7	42-51	0.896	0.592	0.531	0.356	0.319	0.425
8	63-72	0.541	0.438	0.237	1.000	0.541	0.389
9	131-140	0.659	0.615	0.405	0.500	0.329	0.367
10	94-103	0.691	0.583	0.403	0.437	0.302	0.352
11	147-156	0.615	0.718	0.442	0.397	0.244	0.343
12	41-50	0.762	0.547	0.417	0.339	0.258	0.338
13	43-52	0.741	0.536	0.397	0.345	0.255	0.326
14	95-104	0.612	0.672	0.411	0.395	0.242	0.326
15	27-36	0.628	0.617	0.388	0.329	0.207	0.297
16	74-83	0.475	0.758	0.360	0.411	0.195	0.278
17	66-75	0.452	0.447	0.202	0.720	0.326	0.264
18	72-81	0.438	0.724	0.317	0.453	0.198	0.258
19	56-65	0.476	0.644	0.306	0.394	0.187	0.247

Table C.12: SAGE prediction for 4E10 epitope (residues 671-680) on 1Z6N scaffold. Candidate marked with $*$ is the one chosen in the original paper.

Rank	Insertion position	Exposure score	Structure score Raw	Structure score Weighted	Epitope score Raw	Epitope score Weighted	Average score
1*	74-97	1.000	1.000	1.000	0.887	0.887	0.943
2	76-99	0.804	0.713	0.573	0.941	0.756	0.664
3	82-105	0.759	0.836	0.634	0.834	0.633	0.634
4	71-94	0.777	0.445	0.346	1.000	0.777	0.561
5	94-117	0.690	0.754	0.520	0.797	0.550	0.535
6	97-120	0.693	0.535	0.371	0.931	0.645	0.508
7	90-113	0.688	0.538	0.370	0.885	0.608	0.489
8	21-44	0.942	0.201	0.189	0.822	0.774	0.482
9	12-35	0.689	0.381	0.262	0.821	0.565	0.414
10	60-83	0.855	0.536	0.458	0.386	0.330	0.394
11	7-30	0.885	0.395	0.350	0.398	0.352	0.351
12	31-54	0.991	0.319	0.316	0.356	0.353	0.334
13	43-66	0.761	0.170	0.129	0.594	0.453	0.291
14	36-59	0.741	0.340	0.252	0.429	0.318	0.285
15	29-52	0.802	0.179	0.144	0.487	0.390	0.267
16	54-77	0.686	0.345	0.236	0.408	0.280	0.258
17	30-53	0.797	0.265	0.211	0.371	0.296	0.253

Table C.13: SAGE prediction for F1 glycoprotein epitope (residues 254-277) on 3LHP scaffold. Candidate marked with $*$ is the one chosen in the original paper.

Rank	Insertion position	Exposure score	Structure score		Epitope score		Average score
			Raw	Weighted	Raw	Weighted	
1	101-120	0.749	0.835	0.625	1.000	0.749	0.687
2	100-119	0.702	0.983	0.690	0.927	0.651	0.670
3	70-89	0.963	0.790	0.761	0.482	0.464	0.612
4	319-338	0.928	0.725	0.673	0.562	0.522	0.597
5	294-313	0.711	0.784	0.557	0.756	0.538	0.548
6	81-100	0.787	0.915	0.720	0.432	0.340	0.530
7	13-32	0.741	0.956	0.708	0.448	0.332	0.520
8	105-124	0.621	1.000	0.621	0.597	0.371	0.496
9	66-85	0.898	0.676	0.608	0.422	0.380	0.494
10	46-65	0.778	0.873	0.679	0.354	0.275	0.477
11	260-279	1.000	0.586	0.586	0.214	0.214	0.400
12	348-367	0.736	0.613	0.451	0.454	0.334	0.393
13	329-348	0.617	0.870	0.536	0.398	0.246	0.391
14	292-311	0.539	0.859	0.463	0.561	0.302	0.383
15	302-321	0.599	0.745	0.446	0.506	0.303	0.375
16	50-69	0.700	0.655	0.458	0.396	0.277	0.368
17	353-372	0.603	0.723	0.436	0.490	0.295	0.366
18	115-134	0.562	0.868	0.488	0.308	0.173	0.330
19	290-309	0.704	0.436	0.307	0.477	0.336	0.321
20	369-388	0.476	0.724	0.345	0.506	0.241	0.293
21	76-95	0.519	0.546	0.283	0.561	0.291	0.287
22	151-170	0.591	0.692	0.409	0.276	0.163	0.286
23	252-271	0.557	0.825	0.460	0.183	0.102	0.281
24	124-143	0.544	0.796	0.433	0.223	0.121	0.277
25	293-312	0.575	0.515	0.296	0.435	0.250	0.273
26	109-128	0.509	0.641	0.326	0.419	0.213	0.269
27	324-343	0.469	0.762	0.357	0.385	0.181	0.269
28	64-83	0.539	0.611	0.329	0.346	0.187	0.258
29	157-176	0.532	0.733	0.390	0.235	0.125	0.257
30	79-98	0.427	0.631	0.270	0.557	0.238	0.253
31	296-315	0.440	0.550	0.242	0.601	0.265	0.253
32	145-164	0.575	0.593	0.341	0.218	0.125	0.233
33	215-234	0.512	0.619	0.317	0.237	0.121	0.219
34	154-173	0.530	0.480	0.255	0.346	0.183	0.219
35	262-281	0.490	0.663	0.325	0.219	0.107	0.216
36	110-129	0.458	0.449	0.205	0.477	0.219	0.212
37	217-236	0.590	0.472	0.279	0.230	0.136	0.207
38	184-203	0.508	0.610	0.310	0.186	0.094	0.202
39	161-180	0.496	0.568	0.282	0.220	0.109	0.196
40	68-87	0.424	0.545	0.231	0.360	0.153	0.192
41	310-329	0.478	0.375	0.179	0.419	0.200	0.190
42	185-204	0.439	0.663	0.291	0.178	0.078	0.185
43	219-238	0.572	0.345	0.197	0.261	0.149	0.173
44	222-241	0.488	0.482	0.236	0.224	0.110	0.173
45	141-160	0.530	0.417	0.221	0.227	0.120	0.171
46	182-201	0.411	0.523	0.215	0.193	0.079	0.147

Table C.14: SAGE prediction for Ep3Bp epitope grafting on FliC.

Rank	Insertion position	Exposure score	Structure score		Epitope score		Average score
			Raw	Weighted	Raw	Weighted	
1	66-85	0.989	0.838	0.828	0.790	0.781	0.805
2	85-104	0.779	0.850	0.662	1.000	0.779	0.721
3	62-81	1.000	0.827	0.827	0.611	0.611	0.719
4	156-175	0.768	0.841	0.646	0.936	0.719	0.682
5	115-134	0.945	0.704	0.665	0.740	0.699	0.682
6	168-187	0.847	1.000	0.847	0.606	0.513	0.680
7	118-137	0.901	0.789	0.710	0.720	0.649	0.680
8	89-108	0.855	0.761	0.651	0.753	0.644	0.647
9	169-188	0.829	0.958	0.794	0.590	0.489	0.642
10	88-107	0.803	0.792	0.636	0.797	0.640	0.638
11	84-103	0.718	0.792	0.569	0.959	0.689	0.629
12	161-180	0.791	0.884	0.699	0.695	0.549	0.624
13	157-176	0.792	0.778	0.616	0.750	0.594	0.605
14	92-111	0.791	0.692	0.547	0.828	0.655	0.601
15	174-193	0.780	0.789	0.615	0.706	0.550	0.582
16	90-109	0.754	0.790	0.596	0.734	0.554	0.575
17	113-132	0.824	0.729	0.601	0.650	0.535	0.568
18	47-66	0.876	0.660	0.578	0.602	0.528	0.553
19	105-124	0.726	0.709	0.515	0.668	0.485	0.500
20	65-84	0.694	0.716	0.497	0.719	0.499	0.498
21	77-86	0.827	0.640	0.529	0.550	0.454	0.492
22	172-191	0.765	0.708	0.542	0.563	0.430	0.486
23	122-141	0.755	0.594	0.449	0.621	0.469	0.459
24	49-68	0.723	0.592	0.428	0.633	0.458	0.443
25	61-80	0.654	0.694	0.454	0.628	0.411	0.432
26	42-61	0.669	0.617	0.413	0.616	0.412	0.412
27	131-150	0.867	0.337	0.293	0.605	0.524	0.408
28	53-72	0.731	0.506	0.370	0.600	0.439	0.404
29	133-152	0.734	0.308	0.226	0.760	0.558	0.392
30	135-154	0.716	0.376	0.269	0.591	0.423	0.346

Table C.15: SAGE prediction for Ep3Bp epitope grafting on BPSL2520.

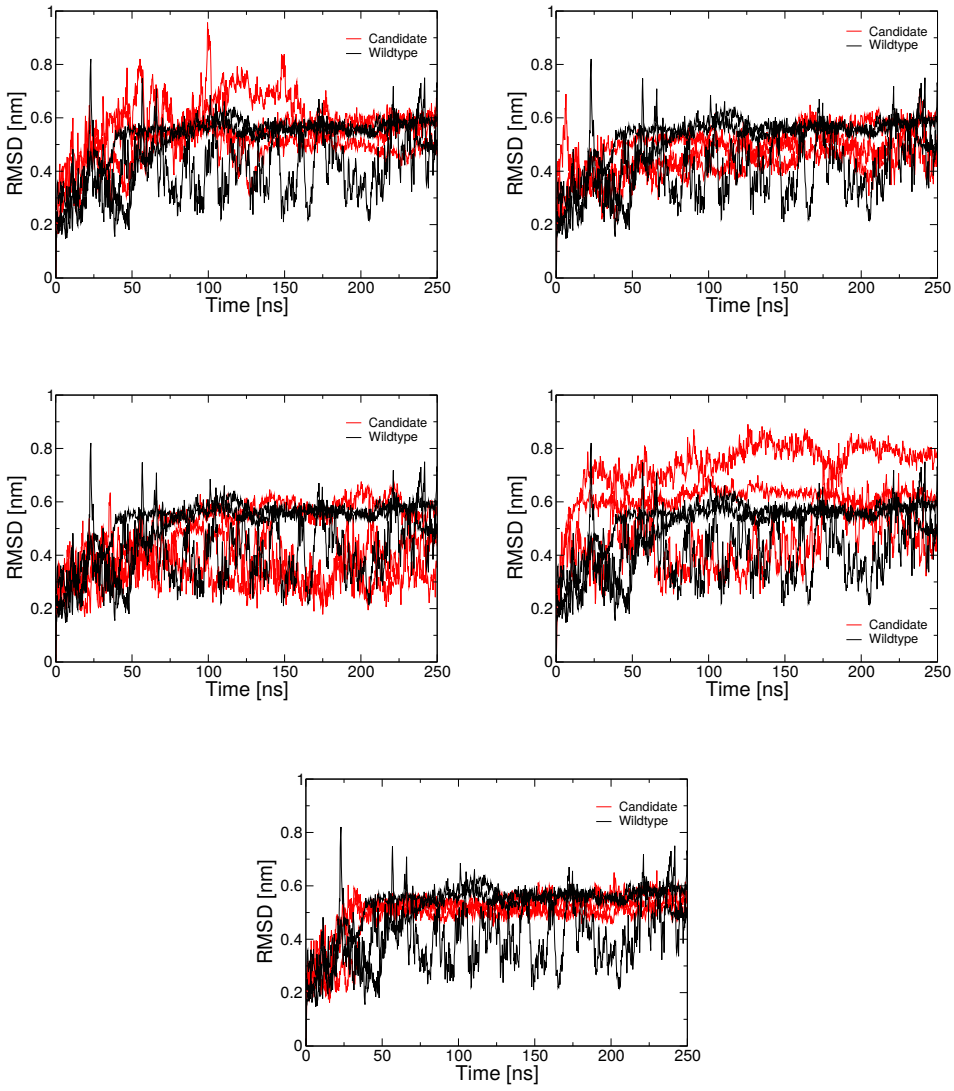


Figure C.1: RMSD for BPSL2520 grafting candidates.

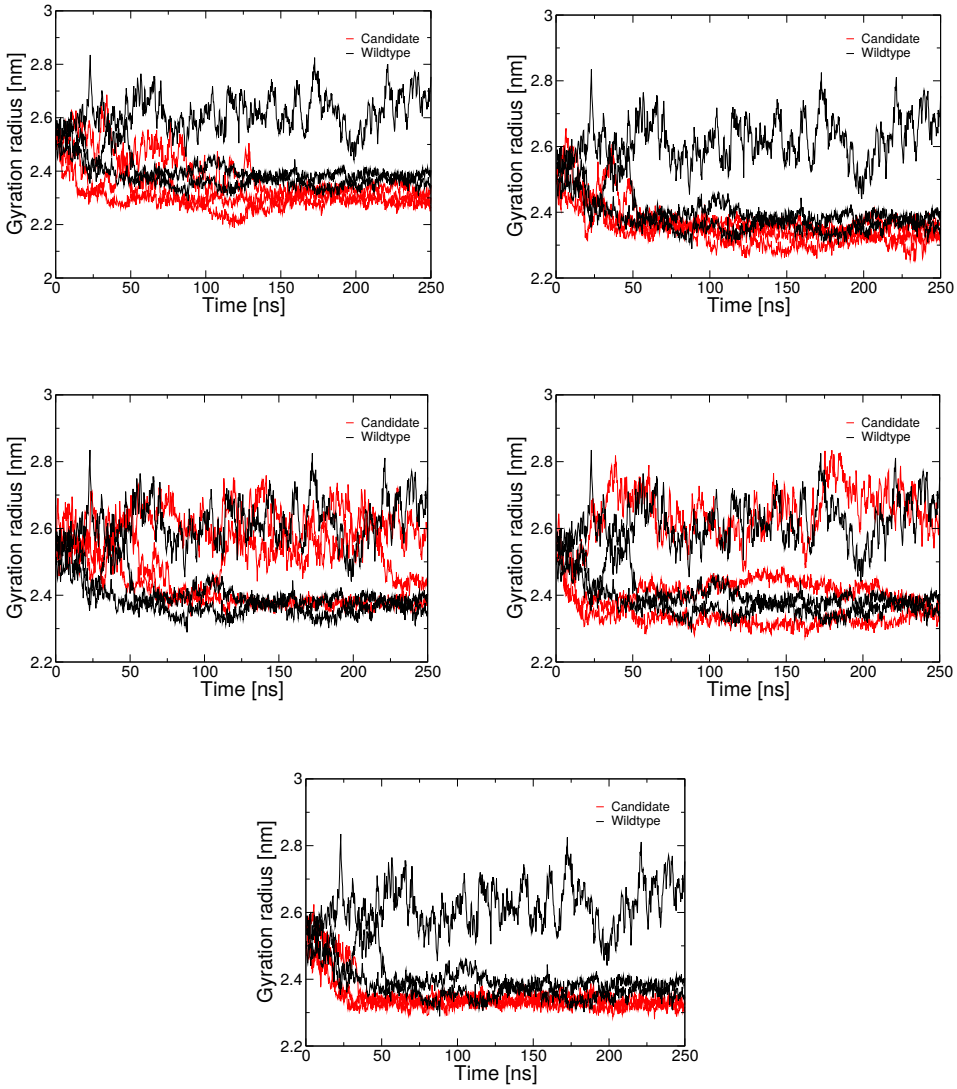


Figure C.2: Gyration radius for BPSL2520 grafting candidates.

Bibliography

[1] Fermi, E.; Pasta, J.; Ulam, S. *Studies of the nonlinear problems*; 1955; pp 1–9.

[2] Pauling, L.; Corey, R. B.; Branson, H. R. *Proceedings of the National Academy of Sciences of the USA* **1951**, *37*, 205–211.

[3] Pauling, L.; Corey, R. B. *Proceedings of the National Academy of Sciences of the USA* **1951**, *37*, 729–40.

[4] Kahn, H.; Harris, T. E. Estimation of particle transmission by random sampling. 1951.

[5] Rosenbluth, M. N.; Rosenbluth, A. W. *The Journal of Chemical Physics* **1955**, *23*, 356–359.

[6] Alder, B. J.; Wainwright, T. E. *Journal of Chemical Physics* **1959**, *31*, 459.

[7] Moore, G. E. *Proceedings of the IEEE* **1998**, *86*, 82–85.

[8] Shaw, D. E. et al. *International Conference for High Performance Computing, Networking, Storage and Analysis, SC* **2014**, *2015-Janua*, 41–53.

[9] Torrie, G. M.; Valleau, J. P. *Journal of Computational Physics* **1977**, *23*, 187–199.

[10] Laio, A.; Parrinello, M. *Proceedings of the National Academy of Sciences of the USA* **2002**, *99*, 12562.

[11] Zwanzig, R. W. *Journal of Chemical Physics* **1954**, *22*, 1420–1426.

[12] Tyka, M. D.; Clarke, A. R.; Sessions, R. B. *Journal of Physical Chemistry B* **2006**, *110*, 17212–17220.

[13] Ovchinnikov, V.; Cecchini, M.; Karplus, M. *Journal of Physical Chemistry B* **2013**, *117*, 750–762.

[14] Allison, J. R.; Bergeler, M.; Hansen, N.; van Gunsteren, W. F. *Biochemistry* **2011**, *50*, 10965–10973.

[15] Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *Biophysical Journal* **2011**, *100*, L47–L49.

[16] Lander, E. S. et al. *Nature* **2001**, *409*, 860–921.

[17] Sette, A.; Rappuoli, R. *Immunity* **2010**, *33*, 530–541.

[18] Pizza, M. et al. *Science* **2000**, *287*, 1816–1820.

[19] Dormitzer, P. R.; Ulmer, J. B.; Rappuoli, R. *Trends in biotechnology* **2008**, *26*, 659–667.

[20] Anfinsen, C. B. *Science* **1973**, *181*, 223–230.

[21] Fersht, A. *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding*; W. H. Freeman: New York, 1999.

[22] Privalov, P. L.; Khechinashvili, N. N. *Journal of Molecular Biology* **1974**, *86*, 665–684.

[23] Kramers, H. A. *Physica* **1940**, *7*, 284–304.

[24] Hansch, C.; Quinlan, J. E.; Lawrence, G. L. *The Journal of Organic Chemistry* **1968**, *33*, 347–350.

[25] Orozco, M.; Luque, F. J. *Chemical Reviews* **2000**, *100*, 4187–4225.

[26] Gohlke, H.; Klebe, G. *Angewandte Chemie - International Edition* **2002**, *41*, 2644–2676.

[27] Kollman, P. A. *Chemical Reviews* **1993**, *93*, 2395–2417.

[28] Van Gunsteren, W. F.; Daura, X.; Mark, A. E. *Helvetica Chimica Acta* **2002**, *85*, 3113–3129.

[29] Huber, T.; Torda, A. E.; van Gunsteren, W. F. *Journal of Computer-Aided Molecular Design* **1994**, *8*, 695–708.

[30] Barducci, A.; Bussi, G.; Parrinello, M. *Physical Review Letters* **2008**, *100*, 1–4.

[31] Straatsma, T. P.; McCammon, J. A. *Annual Review of Physical Chemistry* **1992**, *43*, 407–435.

[32] Kirkwood, J. G. *The Journal of Chemical Physics* **1935**, *3*, 300–313.

[33] Straatsma, T. P.; Berendsen, H. J. C. *The Journal of Chemical Physics* **1988**, *89*, 5876–5886.

[34] Jorge, M.; Garrido, N. M.; Queimada, A. J.; Economou, I. G.; MacEdo, E. A. *Journal of Chemical Theory and Computation* **2010**, *6*, 1018–1027.

[35] Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Dill, K. A. *Journal of Physical Chemistry B* **2009**, *113*, 4533–4537.

[36] Bash, P. A.; Singh, U. C.; Brown, F. K.; Langridge, R.; Kollman, P. A. *Science* **1987**, *235*, 574.

[37] Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. *Chemical Physics Letters* **1994**, *222*, 529–539.

[38] Oostenbrink, C.; Van Gunsteren, W. F. *Proteins: Structure, Function and Genetics* **2004**, *54*, 237–246.

[39] Steinbrecher, T.; Mobley, D. L.; Case, D. A. *Journal of Chemical Physics* **2007**, *127*.

[40] Cecchini, M.; Krivov, S. V.; Spichty, M.; Karplus, M. *Journal of Physical Chemistry B* **2009**, *113*, 9728–9740.

[41] Seeliger, D.; De Groot, B. L. *Biophysical Journal* **2010**, *98*, 2309–2316.

[42] Dill, K. A.; Shortle, D. *Annual review of Biochemistry* **1991**, *60*, 795–825.

[43] Gronenborn, A. M.; Filpula, D. R.; Essig, N. Z.; Achari, A.; Whitlow, M.; Wingfield, P. T.; Clore, G. M. *Science (Washington)* **1991**, *253*, 657–661.

[44] Kobayashi, N.; Honda, S.; Yoshii, H.; Munekata, E. *Biochemistry* **2000**, *39*, 6564–6571.

[45] Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. *Journal of Chemical Theory and Computation* **2008**, *4*, 435–447.

[46] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Structure, Function, and Bioinformatics* **2006**, *65*, 712–725.

[47] Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. *Computer Physics Communications* **2014**, *185*, 604–613.

[48] DeLano, W. L. *Schrödinger, LLC: New York,* **2002**,

[49] Byrd, R. H.; Lu, P.; Nocedal, J.; Zhu, C. *SIAM Journal on Scientific Computing* **1995**, *16*, 1190–1208.

[50] Zhu, C.; Byrd, R. H.; Lu, P.; Nocedal, J. *ACM Transactions on Mathematical Software (TOMS)* **1997**, *23*, 550–560.

[51] Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *Journal of Physical Chemistry A* **1997**, *101*, 3005–3014.

[52] Van Gunsteren, W. F.; Berendsen, H. J. C. *Molecular Simulation* **1988**, *1*, 173–185.

[53] Sugita, Y.; Kitao, A.; Okamoto, Y. *Journal of Chemical Physics* **2000**, *113*, 6042–51.

[54] Esque, J.; Cecchini, M. *Journal of Physical Chemistry B* **2015**, *119*, 5194–5207.

[55] Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *Journal of Computational Chemistry* **1997**, *18*, 1463–1472.

[56] Shirts, M. R.; Chodera, J. D. *Journal of Chemical Physics* **2008**, *129*.

[57] Frenkel, D.; Smit, B. *Understanding Molecular Simulations: from Algorithms to Applications*; San Diego: Academic Press, 2002.

[58] Grossfield, A.; Zuckerman, D. M. *Annual Reports in Computational Chemistry*; Elsevier, 2009; Vol. 5; pp 23–48.

[59] Brooks, B. R. et al. *Journal of Computational Chemistry* **2009**, *30*, 1545–1614.

[60] Neria, E.; Fischer, S.; Karplus, M. *The Journal of Chemical Physics* **1996**, *105*, 1902–1921.

[61] Schaefer, M.; Karplus, M. *Journal of Physical Chemistry* **1996**, *100*, 1578–1599.

[62] Van Der Vaart, A.; Karplus, M. *Journal of Chemical Physics* **2007**, *126*.

[63] Maragakis, P.; van der Vaart, A.; Karplus, M. *The Journal of Physical Chemistry B* **2009**, *113*, 4664–4673.

[64] Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *Journal of Computational Physics* **1977**, *23*, 327–341.

[65] Bellamy, W.; Takase, M.; Yamauchi, K.; Wakabayashi, H.; Kawase, K.; Tomita, M. *Biochimica et Biophysica Acta (BBA)/Protein Structure and Molecular* **1992**, *1121*, 130–136.

[66] Moore, S. a.; Anderson, B. F.; Groom, C. R.; Haridas, M.; Baker, E. N. *Journal of molecular biology* **1997**, *274*, 222–236.

[67] Hwang, P. M.; Zhou, N.; Shan, X.; Arrowsmith, C. H.; Vogel, H. J. *Biochemistry* **1998**, *37*, 4288–4298.

[68] Zhou, N.; Tieleman, D. P.; Vogel, H. J. *BioMetals* **2004**, *17*, 217–223.

[69] Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. *Journal of Chemical Theory and Computation* **2012**, *8*, 3257–3273.

[70] Lee, M. S.; Feig, M.; Salsbury, F. R.; Brooks, C. L. *Journal of Computational Chemistry* **2003**, *24*, 1348–1356.

[71] Noé, F.; Fischer, S. *Current Opinion in Structural Biology* **2008**, *18*, 154–162.

[72] Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. *Journal of Chemical Physics*

2009, *131*.

[73] Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Annual Review of Physical Chemistry* **2002**, *53*, 291–318.

[74] Faccioli, P.; Sega, M.; Pederiva, F.; Orland, H. *Physical Review Letters* **2006**, *97*, 1–4.

[75] Camilloni, C.; Broglia, R. A.; Tiana, G. *Journal of Chemical Physics* **2011**, *134*.

[76] Tiana, G.; Camilloni, C. *Journal of Chemical Physics* **2012**, *137*.

[77] Faradjian, A. K.; Elber, R. *Journal of Chemical Physics* **2004**, *120*, 10880–10889.

[78] Jaynes, E. T. *Annual Review of Physical Chemistry* **1980**, 579–601.

[79] Stock, G.; Ghosh, K.; Dill, K. A. *Journal of Chemical Physics* **2008**, *128*.

[80] Hazoglou, M. J.; Walther, V.; Dixit, P. D.; Dill, K. A. *Journal of Chemical Physics* **2015**, *143*.

[81] Pressé, S.; Ghosh, K.; Dill, K. A. *Journal of Physical Chemistry B* **2011**, *115*, 6202–6212.

[82] Jaynes, E. T. *Physical Review* **1957**, *108*, 171.

[83] Shore, J. E.; Johnson, R. W. *IEEE Transactions on Information Theory* **1980**, *26*, 26–37.

[84] Best, R. B.; Vendruscolo, M. *Journal of the American Chemical Society* **2004**, *126*, 8090–8091.

[85] Camilloni, C.; Robustelli, P.; Simone, A. D.; Cavalli, A.; Vendruscolo, M. *Journal of the American Chemical Society* **2012**, *134*, 3968–3971.

[86] Cavalli, A.; Camilloni, C.; Vendruscolo, M. *Journal of Chemical Physics* **2013**, *138*.

[87] Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; Van Der Spoel, D.; Hess, B.; Lindahl, E. *Bioinformatics* **2013**, *29*, 845–854.

[88] Bonomi, M.; Camilloni, C. *Bioinformatics* **2017**, 1–2.

[89] Best, R. B.; Hummer, G.; Eaton, W. A. *Proceedings of the National Academy of Sciences* **2013**, *110*, 17874–17879.

[90] Debye, P. *Annalen der Physik* **1915**, *351*, 809–823.

[91] Gallagher, T.; Alexander, P.; Bryan, P.; Gilliland, G. L. *Biochemistry* **1994**, *33*, 4721–4729.

[92] Whitford, P. C.; Noel, J. K.; Gosavi, S.; Schug, A.; Sanbonmatsu, K. Y.; Onuchic, J. N. *Proteins: Structure, Function and Bioinformatics* **2009**, *75*, 430–441.

[93] Molgedey, L.; Schuster, H. G. *Physical Review Letters* **1994**, *72*, 3634–3637.

[94] Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. *Journal of Chemical Physics* **2013**, *139*.

[95] Pollack, L. *Biopolymers* **2011**, *95*, 543–549.

[96] Beauchamp, K. A.; Lin, Y.-S.; Das, R.; Pande, V. S. *J. Chem. Theory Comput.* **2012**, *8*, 1409–1414.

[97] Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. *PLoS ONE* **2012**, *7*, e32131.

[98] Lindorff-Larsen, K.; Best, R. B.; Depristo, M. A.; Dobson, C. M.; Vendruscolo, M. *Nature* **2005**, *433*, 128–132.

[99] Bonomi, M.; Camilloni, C.; Cavalli, A.; Vendruscolo, M. *Science Advances* **2016**, *2*, e1501177.

[100] Boomsma, W.; Lindorff-Larsen, K.; Ferkinghoff-Borg, J. *PLoS Comput. Biol.* **2014**, *10*, e1003406.

[101] Rand, K. D.; Zehl, M.; Jørgensen, T. J. D. *Accounts of Chemical Research* **2014**, *47*, 3018–3027.

[102] van Nuland, N. A. J.; Forge, V.; Balback, J.; Dobson, C. M. *Accounts of Chemical Research* **1998**, *31*, 773–780.

[103] Cammarata, M.; Levantino, M.; Schotte, F.; Anfinrud, P. A.; Ewald, F.; Choi, J.; Cupane, A.; Wulff, M.; Ihee, H. *Nature Methods* **2008**, *5*, 881–886.

[104] Goossens, H.; Ferech, M.; Vander Stichele, R.; Elseviers, M. *Lancet* **2005**, *365*, 579–587.

[105] Norrby, S. R.; Nord, C. E.; Finch, R. *The Lancet Infectious diseases* **2005**, *5*, 115–9.

[106] Gourlay, L. J.; Peri, C.; Bolognesi, M.; Colombo, G. *Trends in Biotechnology* **2017**, *xx*, 1–13.

[107] Perelson, A. S.; Weisbuch, G. *Reviews of Modern Physics* **1997**, *69*, 1219–1268.

[108] Jenner, E. *On the origin of the vaccine inoculation*; Shury: London, 1801.

[109] World Health Organization, The Global Eradication of Smallpox. Final Report of the Global Commission for the Certification of Smallpox Eradication. 1980.

[110] Detmer, A.; Glenting, J. *Microbial cell factories* **2006**, *5*, 23.

[111] Fleischmann, R. D. et al. *Science* **1995**, *269*, 496–512.

[112] Rappuoli, R. *Current Opinion in Microbiology* **2000**, *3*, 445–450.

[113] WHO, World Health Organization. Meningococcal meningitis fact sheet. 2015; `http://www.who.int/mediacentre/factsheets/fs141/en/`.

[114] Giuliani, M. M. et al. *Proceedings of the National Academy of Sciences of the USA* **2006**, *103*, 10834–9.

[115] Kulp, D. W.; Schief, W. R. *Current opinion in virology* **2013**, *3*, 322–331.

[116] Skwarczynski, M.; Toth, I. *Chemical Science* **2016**, *7*, 842–854.

[117] Irvine, D. J.; Hanson, M. C.; Rakhra, K.; Tokatlian, T. *Chemical Reviews* **2015**, *115*, 11109–11146.

[118] Bovier, P. A. *Expert review of vaccines* **2008**, *7*, 1141–50.

[119] Philipson, T. *Handbook of Health Economics* **2000**, *1*, 1761–1799.

[120] Spellberg, B.; Guidos, R.; Gilbert, D.; Bradley, J.; Boucher, H. W.; Scheld, W. M.; Bartlett, J. G.; Edwards Jr., J. *Clinical Infectious Diseases* **2008**, *46*, 155–164.

[121] Jones, K. E.; Patel, N. G.; Levy, M. A.; Storeygard, A.; Balk, D.; Gittleman, J. L.; Daszak, P. *Nature* **2008**, *451*, 990–993.

[122] Felgner, P. L. et al. *Proceedings of the National Academy of Sciences of the USA* **2009**, *106*, 13499–504.

[123] Liang, L.; Juarez, S.; Nga, T. V. T.; Dunstan, S.; Nakajima-Sasaki, R.; Davies, D. H.; McSorley, S.; Baker, S.; Felgner, P. L. *Scientific reports* **2013**, *3*, 1043.

[124] Forsstrom, B.; Axnas, B. B.; Stengele, K.-P.; Buhler, J.; Albert, T. J.; Richmond, T. A.; Hu, F. J.; Nilsson, P.; Hudson, E. P.; Rockberg, J.; Uhlen, M. *Molecular & Cellular Proteomics* **2014**, *13*, 1585–1597.

[125] Vita, R.; Overton, J. A.; Greenbaum, J. A.; Ponomarenko, J.; Clark, J. D.;

Cantrell, J. R.; Wheeler, D. K.; Gabbard, J. L.; Hix, D.; Sette, A.; Peters, B. *Nucleic Acids Research* **2015**, *43*, D405–D412.

[126] Toseland, C. P.; Clayton, D. J.; McSparron, H.; Hemsley, S. L.; Blythe, M. J.; Paine, K.; Doytchinova, I. A.; Guan, P.; Hattotuwagama, C. K.; Flower, D. R. *Immunome research* **2005**, *1*, 4.

[127] Lefranc, M. P.; Giudicelli, V.; Ginestoux, C.; Jabado-Michaloud, J.; Folch, G.; Bellahcene, F.; Wu, Y.; Gemrot, E.; Brochet, X.; Lane, J.; Regnier, L.; Ehrenmann, F.; Lefranc, G.; Duroux, P. *Nucleic Acids Research* **2009**, *37*, 1006–1012.

[128] Zhang, G. L.; Khan, A. M.; Srinivasan, K. N.; August, J. T.; Brusic, V. *Nucleic Acids Research* **2005**, *33*, 172–179.

[129] Zhang, L.; Chen, Y.; Wong, H. S.; Zhou, S.; Mamitsuka, H.; Zhu, S. *PLoS ONE* **2012**, *7*.

[130] Singh, H.; Raghava, G. P. S. *Bioinformatics* **2002**, *17*, 1236–1237.

[131] Ponomarenko, J.; Bui, H.-H.; Li, W.; Fusseder, N.; Bourne, P. E.; Sette, A.; Peters, B. *BMC bioinformatics* **2008**, *9*, 514.

[132] Qi, T.; Qiu, T.; Zhang, Q.; Tang, K.; Fan, Y.; Qiu, J.; Wu, D.; Zhang, W.; Chen, Y.; Gao, J.; Zhu, R.; Cao, Z. *Nucleic Acids Research* **2014**, *42*, 59–63.

[133] Scarabelli, G.; Morra, G.; Colombo, G. *Biophysical journal* **2010**, *98*, 1966–1975.

[134] Tiana, G.; Simona, F.; De Mori, G. M. S.; Broglia, R. A.; Colombo, G. *Protein Science* **2004**, *13*, 113–124.

[135] Morra, G.; Colombo, G. *Proteins: Structure, Function and Bioinformatics* **2008**, *72*, 660–672.

[136] Peri, C.; Conchillo-Solé, O.; Corrada, D.; Gori, A.; Daura, X.; Colombo, G. *Methods in Molecular Biology* **2015**, *1348*, 505 pp.

[137] Wiersinga, W. J.; Currie, B. J.; Peacock, S. J. *New England Journal of Medicine* **2012**, *367*, 1035–1044.

[138] Schnetterle, M.; Koch, L.; Gorgé, O.; Valade, E.; Bolla, J.-M.; Biot, F.; Neulat-Ripoll, F. *Current Tropical Medicine Reports* **2017**, *4*, 127–135.

[139] LiPuma, J. J. *Clinical Microbiology Reviews* **2010**, *23*, 299–323.

[140] Bazzini, S.; Udine, C.; Riccardi, G. *Applied Microbiology and Biotechnology* **2011**, *92*, 887–895.

[141] Hara, Y.; Mohamed, R.; Nathan, S. *PLoS ONE* **2009**, *4*.

[142] Gourlay, L. J. et al. *Chemistry and Biology* **2013**, *20*, 1147–1156.

[143] Peri, C.; Gori, A.; Gagni, P.; Sola, L.; Girelli, D.; Sottotetti, S.; Cariani, L.; Chiari, M.; Cretich, M.; Colombo, G. *Scientific reports* **2016**, *6*, 32873.

[144] Gori, A.; Peri, C.; Quilici, G.; Nithichanon, A.; Gaudesi, D.; Longhi, R.; Gourlay, L. J.; Bolognesi, M.; Lertmemongkolchai, G.; Musco, G.; Colombo, G. *ACS Infectious Diseases* **2016**, *2*, 221–230.

[145] Maiti, R.; Van Domselaar, G. H.; Zhang, H.; Wishart, D. S. *Nucleic Acids Research* **2004**, *32*, 590–594.

[146] Godlewska, R.; Wiśniewska, K.; Pietras, Z.; Jagusztyn-Krynicka, E. K. *FEMS Microbiology Letters* **2009**, *298*, 1–11.

[147] Musson, J. A.; Reynolds, C. J.; Rinchai, D.; Nithichanon, A.; Khaenam, P.; Favry, E.; Spink, N.; Chu, K. K. Y.; De Soyza, A.; Bancroft, G. J.; Lertmemongkolchai, G.; Maillere, B.; Boyton, R. J.; Altmann, D. M.; Robinson, J. H. *Journal of immunology (Baltimore, Md. : 1950)* **2014**, *193*, 6041–9.

[148] Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *Journal of Computational Chemistry* **1992**, *13*, 1011–1021.

[149] Case, D. A. et al. *AMBER 2016*; 2016.

[150] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *The Journal of Chemical Physics* **1983**, *79*, 926.

[151] Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. *Journal of Chemical Theory and Computation* **2015**, *11*, 3696–3713.

[152] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, a.; Haak, J. R. *The Journal of Chemical Physics* **1984**, *81*, 3684–3690.

[153] Frishman, D.; Argos, P. *Proteins: Structure, Function, and Bioinformatics* **1995**, *23*, 566–579.

[154] Noel, J. K.; Whitford, P. C.; Sanbonmatsu, K. Y.; Onuchic, J. N. *Nucleic Acids Research* **2010**, *38*, 657–661.

[155] Rhodes, K. A.; Schweizer, H. P. *Drug Resistance Updates* **2016**, *28*, 82–90.

[156] Pumpens, P.; Grens, E. *Artificial DNA*; 2002; pp 249–328.

[157] Pumpens, P.; Grens, E. *Molecular Biology* **2016**, *50*, 26–8933.

[158] Correia, B. E. et al. *Structure* **2010**, *18*, 1116–1126.

[159] Ofek, G.; Guenaga, F. J.; Schief, W. R.; Skinner, J.; Baker, D.; Wyatt, R.; Kwong, P. D. *Proceedings of the National Academy of Sciences of the United States of America* **2010**, *107*, 17880–17887.

[160] Correia, B. E.; Ban, Y. E. A.; Friend, D. J.; Ellingson, K.; Xu, H.; Boni, E.; Bradley-Hewitt, T.; Bruhn-Johannsen, J. F.; Stamatatos, L.; Strong, R. K.; Schief, W. R. *Journal of Molecular Biology* **2011**, *405*, 284–297.

[161] Azoitei, M. L.; Ban, Y. E. A.; Julien, J. P.; Bryson, S.; Schroeter, A.; Kalyuzhniy, O.; Porter, J. R.; Adachi, Y.; Baker, D.; Pai, E. F.; Schief, W. R. *Journal of Molecular Biology* **2012**, *415*, 175–192.

[162] Gront, D.; Kulp, D. W.; Vernon, R. M.; Strauss, C. E. M.; Baker, D. *PLoS ONE* **2011**, *6*.

[163] Drakopoulou, E.; Zinn-Justin, S.; Guenneugues, M.; Gilquin, B.; Ménez, A.; Vita, C. *Journal of Biological Chemistry* **1996**, *271*, 11979–11987.

[164] Lawatscheck, R.; Aleksaite, E.; Schenk, J. a.; Micheel, B.; Jandrig, B.; Holland, G.; Sasnauskas, K.; Gedvilaite, A.; Ulrich, R. G. *Viral immunology* **2007**, *20*, 453–60.

[165] Shindyalov, I. N.; Bourne, P. E. *Protein Engineering Design and Selection* **1998**, *11*, 739–747.

[166] Sormanni, P.; Camilloni, C.; Fariselli, P.; Vendruscolo, M. *Journal of Molecular Biology* **2015**, *427*, 982–996.

[167] Hubbard, S. J.; Thornton, J. M. Naccess. 1993.

[168] Correia, B. E. et al. *Nature* **2014**, *507*, 201–206.

[169] Kazaks, A.; Balmaks, R.; Voronkova, T.; Ose, V.; Pumpens, P. *Biotechnology Journal* **2008**, *3*, 1429–1436.

[170] Rotkiewicz, P.; Skolnick, J. *Journal of computational chemistry* **2008**, *29*, 1460–5.

[171] Webb, B.; Sali, A. *Current Protocols in Bioinformatics*; 2014; Vol. 2014; pp 5.6.1–5.6.32.

[172] Gedvilaite, A.; Frömmel, C.; Sasnauskas, K.; Micheel, B.; Ozel, M.; Behrsing, O.; Staniulis, J.; Jandrig, B.; Scherneck, S.; Ulrich, R. *Virology* **2000**, *273*, 21–35.

[173] Gregory, A. E.; Titball, R.; Williamson, D. *Frontiers in cellular and infection microbiology* **2013**, *3*, 13.

[174] Peri, C.; Gagni, P.; Combi, F.; Gori, A.; Chiari, M.; Longhi, R.; Cretich, M.; Colombo, G. *ACS Chemical Biology* **2013**, *8*, 397–404.

[175] Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L. *BMC Bioinformatics* **2009**, *10*, 421.

[176] Chen, Y. S.; Hsiao, Y. S.; Lin, H. H.; Yen, C. M.; Chen, S. C.; Chen, Y. L. *Vaccine* **2006**, *24*, 750–758.

[177] Suwannasaen, D.; Mahawantung, J.; Chaowagul, W.; Limmathurotsakul, D.; Felgner, P. L.; Davies, H.; Bancroft, G. J.; Titball, R. W.; Lertmemongkolchai, G. *Journal of Infectious Diseases* **2011**, *203*, 1002–1011.

[178] Nithichanon, A.; Rinchai, D.; Gori, A.; Lassaux, P.; Peri, C.; Conchillio-Solé, O.; Ferrer-Navarro, M.; Gourlay, L. J.; Nardini, M.; Vila, J.; Daura, X.; Colombo, G.; Bolognesi, M.; Lertmemonkolchai, G. *PLoS Neglected Tropical Diseases* **2015**, *9*, 1–20.

[179] Finkelstein, A. V.; Janin, J. *Protein Engineering, Design and Selection* **1989**, *3*, 1–3.

[180] Adib, A. B. *Journal of Physical Chemistry B* **2008**, *112*, 5910–5916.

List of Publications

As of September 2017

Refereed publications

1. Assessment of mutational effects on peptide stability through confinement simulations
   R. Capelli, F. Villemot, E. Moroni, G. Tiana, A. van der Vaart and G. Colombo.
   J. Phys. Chem. Lett. 7 (1), 126–130, (2016) DOI:10.1021/acs.jpclett.5b02221, PMID: 26678679

2. Balancing accuracy and cost of confinement simulations by interpolation and extrapolation of confinement energies
   F. Villemot, **R. Capelli**, G. Colombo and A. van der Vaart.
   J. Chem. Theory Comput. 12(6), 2779–2789, (2016) DOI:10.1021/acs.jctc.5b01183, PMID: 27120438

3. SAGE: a fast computational tool for epitope grafting onto a foreign protein scaffold
   R. Capelli, F. Marchetti, G. Tiana and G. Colombo.
   J. Chem. Inf. Model. 57(1), 6–10, (2017) DOI:10.1021/acs.jcim.6b00584, PMID: 27992203

4. Designing probes for immunodiagnostics: structural insights into an epitope targeting *Burkholderia* infections.
   R. Capelli, E. Matterazzo, M. Amabili, C. Peri, A. Gori, P. Gagni, M. Chiari, G. Lertmemongkolchai, M. Cretich, M. Bolognesi, G. Colombo and L. J. Gourlay
   ACS Infect. Dis., Just Accepted DOI: 10.1021/acsinfecdis.7b00080, PMID: 28707874

Publications in preparation

1. Replica implementation of the maximum–caliber principle corrects and accelerates molecular dynamics simulations
   R. Capelli, G. Tiana and C. Camilloni

2. Design and crystallization of a BPSL2520 superantigen
   R. Capelli, M. Amabili, L. J. Gourlay, M. Bolognesi and G.Colombo

Ringraziamenti

Siamo stati cavalieri erranti del nuovo millennio,
antieroi e antipatici un po' persino a noi stessi.
Gridavamo forte contro i grattacieli,
diretti discendenti dei mulini a vento.
CASO, Poco Memorabile

Prima di tutto, devo ringraziare le due persone che, per motivi a me ancora ignoti, hanno deciso di accompagnarmi in questa corsa ad ostacoli che risulta essere il dottorato di ricerca. Scientificamente, ma soprattutto umanamente, sono state due guide fondamentali per questo percorso.
Grazie a Giorgio, che mi ha dato un'immensa fiducia, lasciandomi carta bianca e tonnellate di processori per le mie idee quando io, ad essere sincero, non mi sarei dato in mano le chiavi di un Ciao.
Grazie a Guido e al suo candido pragmatismo[1], che mi ha salvato da dei loop infiniti una quantità non numerabile di volte.
La possibilità di poter lavorare con due approcci diversi – quando non opposti – è stata la più grande risorsa di questi anni di dottorato, rendendomi, spero, un ricercatore migliore.
Un grande grazie va a Carlo per l'estrema pazienza, le discussioni davanti alla macchinetta del caffè e i consigli che mi ha dato in questi ultimi mesi.

Grazie a mio padre, per le poche e preziose parole e per i molti grandi gesti.
Grazie a mia madre, per la pazienza e per l'umiltà che spero di avere imparato.
Grazie alla Maru, per tutte le qualità che ha e che vorrei io.
Grassie. Amò öna ölta. Te te sét ol perchè.

Da qui in poi cercherò di riassumere questi tre anni con una storia.

[1] Vorrei in particolare sottolineare l'uso dell'esoterico simbolo $\mathfrak{Sh}$ nelle bozze della qui presente tesi, e gli scompensi da esso derivati in un povero dottorando.

Il percorso che porta un poco capace studente di dottorato a quell'agognato (e odiato a tratti) titolo di *Philosophiæ Doctor* è da definirsi quantomeno tortuoso. Da lillipuziano, devi salire sulle spalle dei giganti; e devi contrattarci, per avere il permesso di cavalcarli. Concentrandosi sul sottoscritto, ahimè, la risposta standard è stata "Stai violando il secondo principio della termodinamica" o un più serafico `NaN`.
La questione fastidiosa è che non è solo con esseri mitologici che ti tocca avere a che fare. Devi compiere altre imprese epiche: scalare il MonteGrappa, interrogare un vecchio saggio e massimizzare il disordine come se non ci fosse un domani per cercare di capire cosa c'è dentro.

Ad un certo punto, però, scopri che c'è qualcuno che può darti una mano ad arrampicarti sui giganti. L'equipaggiamento fornito per la scalata è rappresentato da una lavagna piena di conti o da un tovagliolo stroppicciato davanti a una birra al beneamato Birrificio di Lambrate. Ça va sans dire, anche il tovagliolo è pieno di conti.
Allo stesso modo, ci si rende conto di avere dei compagni di cordata per il Grappa, oppure qualcuno che può aiutarti a capire cosa chiedere al saggio o, infine, qualcuno che dal caos sa tirare fuori qualcosa di ordinato (ma con espressioni inguardabili). Il mio ringraziamento va a ciascuno di loro[2], senza i quali le pagine che precedono questa non sarebbero mai state scritte.

Un ringraziamento particolare va alle povere anime che mi sono stati affidate (ahiloro) per la loro tesi. Mi avete insegnato – non senza traumi miei e vostri – che non puoi spiegare nulla se non l'hai capito davvero. E ora che tutti abbiamo il pezzo di carta in mano, posso ammettere che in un buon 70% dei casi avevate ragione voi.

Grazie anche agli studenti del corso di Fisica Generale a Scienze Naturali e a Biotecnologie, per avermi fatto capire che stare dall'altra parte era molto più complicato del previsto. In più, grazie a voi ho scoperto che Nettuno ha massa trascurabile e che il principio di Archimede è una trappola mortale per chi vuole fare il bagno.

Ci sono state altre persone presenti davanti a una birra (stavolta senza tovaglioli incriminati) o ad un caffè o, meglio ancora, sopra e sotto ad un palco. Il loro apporto a questa tesi non è assolutamente quantificabile ma, senza ombra di dubbio, è stato fondamentale.

Tornando per un attimo serio, voglio ringraziare l'Università degli Studi di Milano e il Ministero dell'Istruzione, dell'Università e della Ricerca per il sostegno finanziario durante il mio percorso di dottorato.

In ultima analisi, nonostante abbia dimostrato senza ombra di dubbio di essere un *tàmbor* (o più propriamente il suo accrescitivo, *tamburù*), dopo tre anni e qualche mese di fette, posso dire tranquillamente di essere convinto del fatto che fosse polenta[3].

[2]Ma in particolare a Bob, alla cui assenza nell'ufficio in cui lavoro non mi abituerò mai del tutto.
[3]Questo ultimo paragrafo poi lo spiego ai non orobici.