

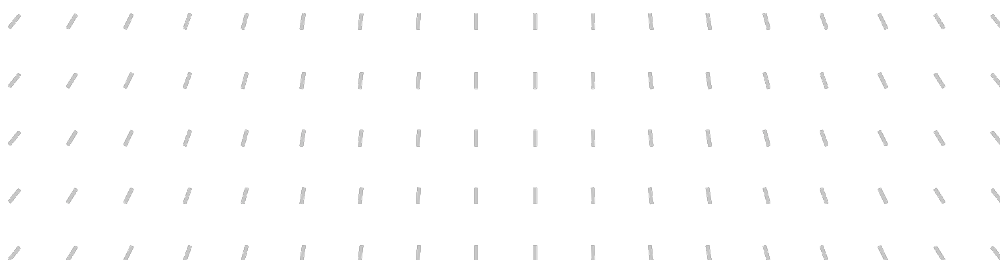
2025 Activity Report

RESEARCH CENTRE: Inria Centre at Université Côte d'Azur


Project-Team

ABS

Algorithms - Biology - Structure

Project-Team ABS

Creation of the Project-Team: 2021 August 01

Each year, Inria research teams publish an Activity Report presenting their work and results over the reporting period. These reports follow a common structure, with some optional sections depending on the specific team. They typically begin by outlining the overall objectives and research programme, including the main research themes, goals, and methodological approaches. They also describe the application domains targeted by the team, highlighting the scientific or societal contexts in which their work is situated. The reports then present the highlights of the year, covering major scientific achievements, software developments, or teaching contributions. When relevant, they include sections on software, platforms, and open data, detailing the tools developed and how they are shared. A substantial part is dedicated to new results, where scientific contributions are described in detail, often with subsections specifying participants and associated keywords. Finally, the Activity Report addresses funding, contracts, partnerships, and collaborations at various levels, from industrial agreements to international cooperations. It also covers dissemination and teaching activities, such as participation in scientific events, outreach, and supervision. The document concludes with a presentation of scientific production, including major publications and those produced during the year.

Keywords

Computer sciences and digital sciences

- A2.5. – Software engineering
- A3.3.2. – Data mining
- A6.1.4. – Multiscale modeling
- A6.2.4. – Statistical methods
- A6.2.8. – Computational geometry and meshes
- A8.1. – Discrete mathematics, combinatorics
- A8.3. – Geometry, Topology
- A8.7. – Graph theory
- A9.2. – Machine learning
- A9.2.1. – Supervised learning
- A9.2.2. – Unsupervised learning

Other research topics and application domains

- B1.1.1. – Structural biology
- B1.1.5. – Immunology
- B1.1.7. – Bioinformatics

Contents

| | |
|--|-----------|
| Project-Team ABS | 1 |
| 1 Team members, visitors, external collaborators | 5 |
| 2 Overall objectives | 6 |
| 3 Research program | 9 |
| 3.1 Modeling the dynamics of proteins | 9 |
| 3.2 Algorithmic foundations: geometry, optimization, machine learning | 9 |
| 3.3 Software: the Structural Bioinformatics Library | 10 |
| 3.4 Applications: modeling interfaces, contacts, and interactions | 10 |
| 4 Application domains | 11 |
| 5 Social and environmental responsibility | 11 |
| 5.1 Footprint of research activities | 11 |
| 5.2 Impact of research results | 11 |
| 6 Highlights of the year | 11 |
| 7 Latest software developments, platforms, open data | 12 |
| 7.1 Latest software developments | 12 |
| 7.1.1 SBL | 12 |
| 8 New results | 12 |
| 8.1 Modeling the dynamics of proteins | 13 |
| 8.1.1 Simpler protein domain identification using spectral clustering | 13 |
| 8.2 Algorithmic foundations | 13 |
| 8.2.1 Improved seeding strategies for k-means and k-GMM | 13 |
| 8.2.2 Modeling high dimensional point clouds with the spherical cluster model | 14 |
| 8.3 Applications in structural bioinformatics and beyond | 14 |
| 8.3.1 Fold or flop: quality assessment of AlphaFold predictions on whole proteomes | 14 |
| 8.3.2 Characterizing the fragmentation of AlphaFold predictions | 15 |
| 8.3.3 Orphan genes survey | 15 |
| 8.3.4 Orphan genes detection and classification | 15 |
| 9 Bilateral contracts and grants with industry | 15 |
| 9.1 Bilateral contracts with industry | 16 |
| 10 Partnerships and cooperations | 16 |
| 10.1 National initiatives | 16 |
| 10.2 Regional initiatives | 17 |
| 11 Dissemination | 17 |
| 11.1 Promoting scientific activities | 17 |
| 11.1.1 Scientific events: organization | 17 |
| 11.1.2 Invited talks | 17 |
| 11.1.3 Leadership within the scientific community | 18 |
| 11.2 Teaching - Supervision - Juries - Educational and pedagogical outreach | 18 |
| 11.2.1 Supervision | 18 |
| 11.2.2 Juries | 18 |

| | |
|---|-----------|
| 12 Scientific production | 19 |
| 12.1 Major publications | 19 |
| 12.2 Publications of the year | 20 |
| 12.3 Cited publications | 20 |

1 Team members, visitors, external collaborators

Research Scientists

- Frédéric Cazals [Team leader, INRIA, HDR]
- Dorian Mazaauric [INRIA, Researcher, 20%, HDR]
- Edoardo Sarti [INRIA, Researcher]

Post-Doctoral Fellow

- Antoine Commaret [INRIA, Post-Doctoral Fellow]

PhD Students

- Guillaume Carriere [INRIA]
- Antoine Hauser [VECT-HORUS, CIFRE, from Sep 2025]
- Simon Queric [INRIA, until Oct 2025]
- Ercan Seckin [INRAE]

Technical Staff

- Michal Boniecki [INRIA, Engineer, from Jul 2025 until Aug 2025]
- Nelson Feyeux [INRIA, Engineer, from Dec 2025]
- Sikao Guo [INRIA, Engineer, from Jul 2025]
- Come Le Breton [INRIA, Engineer]

Interns and Apprentices

- Abhaas Aggarwal [INRIA, Intern, from May 2025 until Jul 2025]
- Destiny Hanna [UNIV COTE D'AZUR, Apprentice, until Sep 2025]
- Anoop Singh [INRIA, Intern, from May 2025 until Jul 2025]
- Arinjay Singhal [INRIA, Intern, from May 2025 until Jul 2025]

Administrative Assistant

- Vanessa Wallet [INRIA]

External Collaborator

- Alix Lhéritier [AMADEUS]

2 Overall objectives

Biomolecules and their function(s). Computational Structural Biology (CSB) is the scientific domain concerned with the development of algorithms and software to understand and predict the structure and function of biological macromolecules. This research field is inherently multi-disciplinary. On the experimental side, biology and medicine provide the objects studied, while biophysics and bioinformatics supply experimental data, which are of two main kinds. On the one hand, genome sequencing projects give supply protein sequences, and ~200 millions of sequences have been archived in UniProtKB/TrEMBL – which collects the protein sequences yielded by genome sequencing projects. On the other hand, structure determination experiments (notably X-ray crystallography, nuclear magnetic resonance, and cryo-electron microscopy) give access to geometric models of molecules – atomic coordinates. Alas, only ~150,000 structures have been solved and deposited in the Protein Data Bank (PDB), a number to be compared against the $\sim 10^8$ sequences found in UniProtKB/TrEMBL. With one structure for ~1000 sequences, we hardly know anything about biological functions at the atomic/structural level. Complementing experiments, physical chemistry/chemical physics supply the required models (energies, thermodynamics, etc). More specifically, let us recall that proteins with n atoms has $d = 3n$ Cartesian coordinates, and fixing these (up to rigid motions) defines a conformation. As conveyed by the iconic *lock-and-key* metaphor for interacting molecules, Biology is based on the interactions stable conformations make with each other. Turning these intuitive notions into quantitative ones requires delving into statistical physics, as macroscopic properties are average properties computed over ensembles of conformations. Developing effective algorithms to perform accurate simulations is especially challenging for two main reasons. The first one is the high dimension of conformational spaces – see $d = 3n$ above, typically several tens of thousands, and the non linearity of the energy functionals used. The second one is the multiscale nature of the phenomena studied: with biologically relevant time scales beyond the millisecond, and atomic vibrations periods of the order of femto-seconds, simulating such phenomena typically requires $\gg 10^{12}$ conformations/frames, a (brute) *tour de force* rarely achieved [38].

Computational Structural Biology: three main challenges. The first challenge, *sequence-to-structure prediction*, aims to infer the possible structure(s) of a protein from its amino acid sequence. While recent progress has been made recently using in particular deep learning techniques [37], the models obtained so far are static and coarse-grained.

The second one is *protein function prediction*. Given a protein with known structure, *i.e.*, 3D coordinates, the goal is to predict the partners of this protein, in terms of stability and specificity. This understanding is fundamental to biology and medicine, as illustrated by the example of the SARS-CoV-2 virus responsible of the Covid19 pandemic. To infect a host, the virus first fuses its envelope with the membrane of a target cell, and then injects its genetic material into that cell. Fusion is achieved by a so-called class I fusion protein, also found in other viruses (influenza, SARS-CoV-1, HIV, etc). The fusion process is a highly dynamic process involving large amplitude conformational changes of the molecules. It is poorly understood, which hinders our ability to design therapeutics to block it.

Finally, the third one, *large assembly reconstruction*, aims at solving (coarse-grain) structures of molecular machines involving tens or even hundreds of subunits. This research vein was promoted about 15 years back by the work on the nuclear pore complex [26]. It is often referred to as *reconstruction by data integration*, as it necessitates to combine coarse-grain models (notably from cryo-electron microscopy (cryo-EM) and native mass spectrometry) with atomic models of subunits obtained from X ray crystallography. Fitting the latter into the former requires exploring the conformation space of subunits, whence the importance of protein dynamics.

As an illustration of these three challenges, consider the problem of designing proteins blocking the entry of SARS-CoV-2 into our cells (Fig. 1). The first challenge is illustrated by the problem of predicting the structure of a blocker protein from its sequence of amino-acids – a tractable problem here since the mini proteins used only comprise of the order of 50 amino-acids (Fig. 1(A), [29]). The second challenge is illustrated by the calculation of the binding modes and the binding affinity of the designed proteins for the RBD of SARS-CoV-2 (Fig. 1(B)). Finally, the last challenge is illustrated by the problem of solving structures of the virus with a cell, to understand how many spikes are involved in the fusion mechanism leading to infection. In [29], the promising designs suggested by modeling have been assessed by an array of

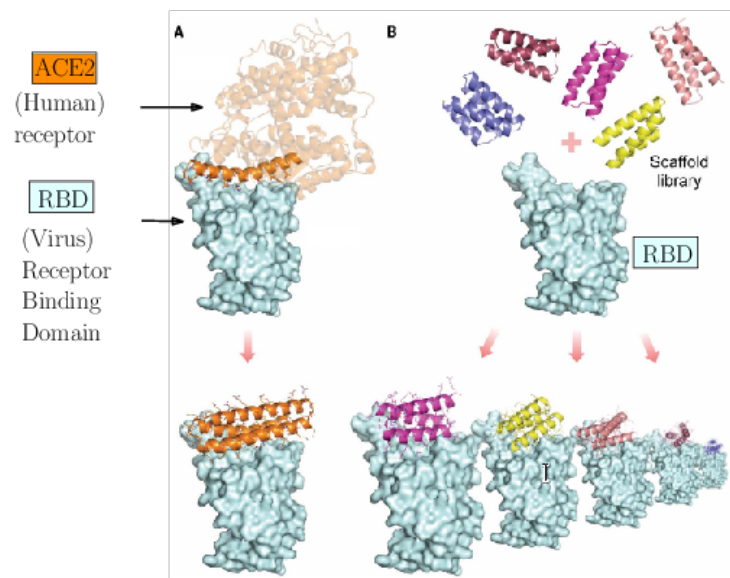


Figure 1: **The synergy modeling - experiments, and challenges faced in CSB: illustration on the problem of designing miniproteins blocking the entry of SARS-CoV-2 into cells. From [29].** Of note: the first step of the infection by SARS-CoV-2 is the attachment of its receptor binding domain of its spike (RBD, blue molecule), to a target protein found on the membrane of our cells, ACE2 (orange molecule). A strategy to block infection is therefore to engineer a molecule binding the RBD, preventing its attachment to ACE2. **(A)** Design of a helical protein (orange) mimicking a region of the ACE2 protein. **(B)** Assessment of binding modes (conformation, binding energies) of candidate miniproteins neutralizing the RBD.

wet lab experiments (affinity measurements, circular dichroism for thermal stability assessment, structure resolution by cryo-EM). The *hyperstable* minibinders identified provide starting points for SARS-CoV-2 therapeutics [29]. We note in passing that this is truly remarkable work, yet, the designed proteins stem from a template (the *bottom* helix from ACE2), and are rather small.

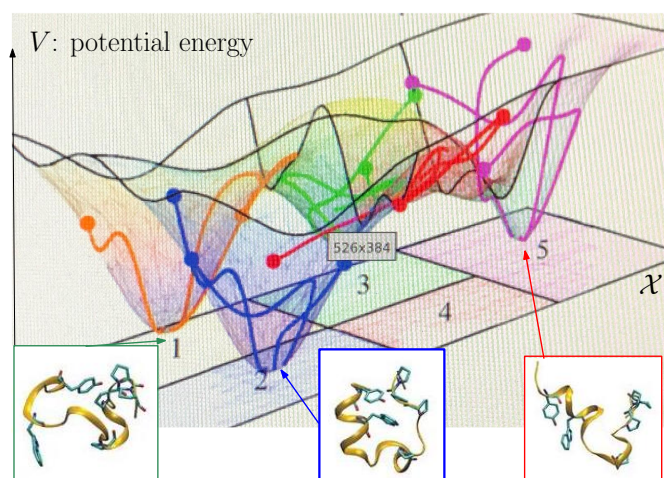


Figure 2: The main challenges of molecular simulation: Finding significant local minima of the energy landscape, computing statistical weights of catchment basins by integrating Boltzmann’s factor, and identifying transitions. Practically, $d > 100$.

Protein dynamics: core CS - maths challenges. To present challenges in structural modeling, let us recall the following ingredients (Fig. 2). First, a molecular model with n atoms is parameterized over a conformational space \mathcal{X} of dimension $d = 3n$ in Cartesian coordinates, or $d = 3n - 6$ in internal coordinate—upon removing rigid motions, also called degree of freedom (*d.o.f.*). Second, recall that the *potential energy landscape* (PEL) is the mapping $V(\cdot)$ from \mathbb{R}^d to \mathbb{R} providing a potential energy for each conformation [39, 36]. Example potential energies (PE) are CHARMM, AMBER, MARTINI, etc. Such PE belong to the realm of molecular mechanics, and implement atomic or coarse-grain models. They may embark a solvent model, either explicit or implicit. Their definition requires a significant number of parameters (up to $\sim 1,000$), fitted to reproduce physico-chemical properties of (bio-)molecules [40].

These PE are usually considered good enough to study non covalent interactions – our focus, even though they do not cover the modification of chemical bonds. In any case, we take such a function for granted ¹.

The PEL codes all **structural**, **thermodynamic**, and **kinetic** properties, which can be obtained by averaging properties of conformations over so-called *thermodynamic ensembles*. The **structure** of a macromolecular system requires the characterization of active conformations and important intermediates in functional pathways involving significant basins. In assigning occupation probabilities to these conformations by integrating Boltzmann’s distribution, one treats **thermodynamics**. Finally, transitions between the states, modeled, say, by a master equation (a continuous-time Markov process), correspond to **kinetics**. Classical simulation methods based on molecular dynamics (MD) and Monte Carlo sampling (MC) are developed in the lineage of the seminal work by the 2013 recipients of the Nobel prize in chemistry (Karplus, Levitt, Warshel), which was awarded “*for the development of multiscale models for complex chemical systems*”. However, except for highly specialized cases where massive calculations have been used [38], neither MD nor MC give access to the aforementioned time scales. In fact, the main limitation of such methods is that they treat structural, thermodynamic and kinetic aspects at once [32]. The absence of specific insights on these three complementary pieces of the puzzle makes it impossible to optimize simulation methods, and results in general in the inability to obtain converged simulations on biologically relevant time-scales.

¹We note passing that the PE model currently implemented in the SBL is a classical one with particle-particle interactions, see [Potential Energy](#). But it could be easily extended to accommodate dipole - charge interactions for polarizable force fields (amoeba).

The hardness of structural modeling owes to three intertwined reasons.

First, PELs of biomolecules usually exhibit a number of critical points exponential in the dimension [27]; fortunately, they enjoy a multi-scale structure [30]. Intuitively, the significant local minima/basins are those which are *deep* or *isolated/wide*, two notions which are mathematically qualified by the concepts of persistence and prominence. Mathematically, problems are plagued with the curse of dimensionality and measure concentration phenomena. Second, biomolecular processes are inherently multi-scale, with motions spanning ~ 15 and ~ 4 orders of magnitude in time and amplitude respectively [25]. Developing methods able to exploit this multi-scale structure has remained elusive. Third, macroscopic properties of biomolecules, *i.e.*, observables, are average properties computed over ensembles of conformations, which calls for a multi-scale statistical treatment both of thermodynamics and kinetics.

Validating models. A natural and critical question naturally concerns the validation of models proposed in structural bioinformatics. For all three types of questions of interest (structures, thermodynamics, kinetics), there exist experiments to which the models must be confronted – when the experiments can be conducted.

For structures, the models proposed can readily be compared against experimental results stemming from X ray crystallography, NMR, or cryo electron microscopy. For thermodynamics, which we illustrate here with binding affinities, predictions can be compared against measurements provided by calorimetry or surface plasmon resonance. Lastly, kinetic predictions can also be assessed by various experiments such as binding affinity measurements (for the prediction of K_{on} and K_{off}), or fluorescence based methods (for kinetics of folding).

3 Research program

Our research program ambitions to develop a comprehensive set of novel concepts and algorithms to study protein dynamics, based on the modular framework of PEL.

3.1 Modeling the dynamics of proteins

Keywords: Molecular conformations, conformational exploration, energy landscapes, thermodynamics, kinetics.

As noticed while discussing *Protein dynamics: core CS - maths challenges*, the integrated nature of simulation methods such as MD or MC is such that these methods do not in general give access to biologically relevant time scales. The framework of energy landscapes [39, 36] (Fig. 2) is much more modular, yet, large biomolecular systems remain out of reach.

To make a definitive step towards solving the prediction of protein dynamics, we will serialize the discovery and the exploitation of a PEL [4, 16, 3]. Ideas and concepts from computational geometry/geometric motion planning, machine learning, probabilistic algorithms, and numerical probability will be used to develop two classes of probabilistic algorithms. The first deals with algorithms to discover/sketch PELs, *i.e.*, enumerate all significant (persistent or prominent) local minima and their connections across saddles, a difficult task since the number of all local minima/critical points is generally exponential in the dimension. To this end, we will develop a hierarchical data structure coding PELs as well as multi-scale proposals to explore molecular conformations. (NB: in Monte Carlo methods, a proposal generates a new conformation from an existing one.) The second focuses on methods to exploit/sample PELs, *i.e.*, compute so-called densities of states, from which all thermodynamic quantities are given by standard relations [28][35]. This is a hard problem akin to high-dimensional numerical integration. To solve this problem, we will develop a learning based strategy for the Wang-Landau algorithm [34]—an adaptive Monte Carlo Markov Chain (MCMC) algorithm, as well as a generalization of multi-phase Monte Carlo methods for convex/polytope volume calculations [33, 31], for non convex strata of PELs.

3.2 Algorithmic foundations: geometry, optimization, machine learning

Keywords: Geometry, optimization, machine learning, randomized algorithms, sampling, optimization.

As discussed in the previous Section, the study of PEL and protein dynamics raises difficult algorithmic /

mathematical questions. As an illustration, one may consider our recent work on the comparison of high dimensional distribution [7], statistical tests / two-sample tests [8, 13], the comparison of clustering [9], the complexity study of graph inference problems for low-resolution reconstruction of assemblies [12], the analysis of partition (or clustering) stability in large networks, the complexity of the representation of simplicial complexes [2]. Making progress on such questions is fundamental to advance the state-of-the-art on protein dynamics.

We will continue to work on such questions, motivated by CSB / theoretical biophysics, both in the continuous (geometric) and discrete settings. The developments will be based on a combination of ideas and concepts from computational geometry, machine learning (notably on non linear dimensionality reduction, the reconstruction of cell complexes, and sampling methods), graph algorithms, probabilistic algorithms, optimization, numerical probability, and also biophysics.

3.3 Software: the Structural Bioinformatics Library

Keywords: Scientific software, generic programming, molecular modeling.

While our main ambition is to advance the algorithmic foundations of molecular simulation, a major challenge will be to ensure that the theoretical and algorithmic developments will change the fate of applications, as illustrated by our case studies. To foster such a symbiotic relationship between theory, algorithms and simulation, we will pursue high quality software development and integration within the SBL, and will also take the appropriate measures for the software to be widely adopted.

Software in structural bioinformatics. Software development for structural bioinformatics is especially challenging, combining advanced geometric, numerical and combinatorial algorithms, with complex biophysical models for PEL and related thermodynamic/kinetic properties. Specific features of the proteins studied must also be accommodated. About 50 years after the development of force fields and simulation methods (see the 2013 Nobel prize in chemistry), the software implementing such methods has a profound impact on molecular science at large. One can indeed cite packages such as CHARMM, AMBER, gromacs, gmin, MODELLER, Rosetta, VMD, PyMol, On the other hand, these packages are goal oriented, each tackling a (small set of) specific goal(s). In fact, no real modular software design and integration has taken place. As a result, despite the high quality software packages available, inter-operability between algorithmic building blocks has remained very limited.

The SBL. Predicting the dynamics of large molecular systems requires the integration of advanced algorithmic building blocks / complex software components. To achieve a sufficient level of integration, we undertook the development of the Structural Bioinformatics Library (SBL, SB) [6], a generic C++/python cross-platform library providing software to solve complex problems in structural bioinformatics. For end-users, the SBL provides ready to use, state-of-the-art applications to model macro-molecules and their complexes at various resolutions, and also to store results in perennial and easy to use data formats (SBL Applications). For developers, the SBL provides a broad C++/python toolbox with modular design (SBL Doc). This hybrid status targeting both end-users and developers stems from an advanced software design involving four software components, namely applications, core algorithms, biophysical models, and modules (SBL Modules). This modular design makes it possible to optimize robustness and the performance of individual components, which can then be assembled within a goal oriented application.

3.4 Applications: modeling interfaces, contacts, and interactions

Keywords: Protein interactions, protein complexes, structure/thermodynamics/kinetics prediction.

Our methods will be validated on various systems for which flexibility operates at various scales. Examples of such systems are antibody-antigen complexes, (viral) polymerases, (membrane) transporters.

Even very complex biomolecular systems are deterministic in prescribed conditions (temperature, pH, etc), demonstrating that despite their high dimensionality, all *d.o.f.* are not at play at the same time. This insight suggests three classes of systems of particular interest. The first class consists of systems defined from (essentially) rigid blocks whose relative positions change thanks to conformational changes of linkers; a

Newton cradle provides an interesting way to envision such as system. We have recently worked on one such system, a membrane proteins involve in antibiotic resistance (AcrB, see [17]). The second class consists of cases where relative positions of subdomains do not significantly change, yet, their intrinsic dynamics are significantly altered. A classical illustration is provided by antibodies, whose binding affinity owes to dynamics localized in six specific loops [14, 15]. The third class, consisting of composite cases, will greatly benefit from insights on the first two classes. As an example, we may consider the spikes of the SARS-CoV-2 virus, whose function (performing infection) involves both large amplitude conformational changes and subtle dynamics of the so-called receptor binding domain. We have started to investigate this system, in collaboration with B. Delmas (INRAE).

In ABS, we will investigate systems in these three tiers, in collaboration with expert collaborators, to hopefully open new perspectives in biology and medicine. Along the way, we will also collaborate on selected questions at the interface between CSB and systems biology, as it is now clear that the structural level and the systems level (pathways of interacting molecules) can benefit from one another.

4 Application domains

The main application domain is Computational Structural Biology, as underlined in the *Research Program*.

5 Social and environmental responsibility

5.1 Footprint of research activities

A tenet of ABS is to carefully analyze the performances of the algorithms designed—either formally or experimentally, so as to avoid massive calculations. Therefore, the footprint of our research activities has remained limited.

5.2 Impact of research results

The scientific agenda of ABS is geared towards a fine understanding of complex phenomena at the atomic/molecular level. While the current focus is rather fundamental, as explained in *Research program*, an overarching goal for the current period (i.e. 12 years) is to make significant contributions to important problems in biology and medicine.

6 Highlights of the year

We would like to comment in three directions.

Science. Regarding the algorithmic foundations of our activities, we have made significant progress on various clustering procedures which are needed to model molecular conformations in high dimensional spaces. In particular, our improvement of the k-means seeding procedure [18] is currently being used to define complex mixtures of periodic functions to model joint probabilities of torsion angles in proteins.

On the applied side, we developed more thorough and rigorous analysis of structure predictions made by the *Nobel prize winning* program AlphaFold, see [23] and [22]. We hope some of these analysis will become standard and made available on portal such as the [AlphaFold Protein Structure Database](#).

Software. We have continued our efforts on the development of the Structural Bioinformatics Library, adding no less than eight new packages – see updates about the Bioinformatics Library.

Most importantly, the `Plugin_manager` package was developed in the context of the PIQ project (Programme InriaQuadrant) Eminence headed by F. Cazals, which started in June 2025. For each application package, this specific package eases the development of plugins for various target platforms, including VMD, PyMol, and web servers. A vast campaign advertising plugins for six applications will be orchestrated early 2026, both at the national and international levels.

Teaching. Attracting bright students is a notoriously difficult challenge. In 2025, F. Cazals started the class [Algorithms and learning for protein science](#), within the Master Mathématiques, Vision, Apprentissage, ENS Paris-Saclay. This first season involved 25 students.

7 Latest software developments, platforms, open data

7.1 Latest software developments

7.1.1 SBL

Name: Structural Bioinformatics Library

Keywords: Structural Biology, Biophysics, Software architecture

Functional Description: The SBL is a generic C++/python cross-platform software library targeting complex problems in structural bioinformatics. Its tenet is based on a modular design offering a rich and versatile framework allowing the development of novel applications requiring well specified complex operations, without compromising robustness and performances.

More specifically, the SBL involves four software components (1-4 thereafter). For end-users, the SBL provides ready to use, state-of-the-art (1) applications to handle molecular models defined by unions of balls, to deal with molecular flexibility, to model macro-molecular assemblies. These applications can also be combined to tackle integrated analysis problems. For developers, the SBL provides a broad C++ toolbox with modular design, involving core (2) algorithms, (3) biophysical models, and (4) modules, the latter being especially suited to develop novel applications. The SBL comes with a thorough documentation consisting of user and reference manuals, and a bugzilla platform to handle community feedback.

Release Contributions: In 2025, several new packages were added to the distribution.

On the computer science side, the package Seeding (<https://sbl.inria.fr/doc/Seeding-user-manual.html>) provides new seeding methods for Kmeans++ like algorithms, while the package Cluster spherical introduces the first algorithm to fit so-called spherical clusters in high dimensional spaces (https://sbl.inria.fr/doc/Cluster_spherical-user-manual.html)

On the computational structural biology side, the package AlphaFold analysis provides novel quality analysis for reconstructions delivered by the AlphaFold program (https://sbl.inria.fr/doc/Alphafold_analysis-user-manual.html). Also, three packages were added to represent nucleic acids in addition to protein structures: the package Linear_polymer_representation abstract commonalities (https://sbl.inria.fr/doc/Linear_polymer_representation-user-manual.html), the package Nucleic_acid_representation specifically encodes nucleic acids (https://sbl.inria.fr/doc/Nucleic_acid_representation-user-manual.html), the package Biomolecule_representation offers a common interface to represent proteins and nucleic acids (https://sbl.inria.fr/doc/Biomolecule_representation-user-manual.html).

Finally, on the software engineering side, the package Script design provides guidelines to automate the development of efficient python scripts (https://sbl.inria.fr/doc/Script_design-user-manual.html), while the package Plugin manager introduces a novel framework to automate the design of scripts targeting several platforms (https://sbl.inria.fr/doc/Plugin_manager-user-manual.html).

URL: <https://sbl.inria.fr/>

Publication: hal-01570848

Contact: Frédéric Cazals

8 New results

Participants: Frédéric Cazals, Dorian Mazaauric, Edoardo Sarti.

8.1 Modeling the dynamics of proteins

Keywords: Protein flexibility, protein conformations, collective coordinates, conformational sampling, loop closure, kinematics, dimensionality reduction.

8.1.1 Simpler protein domain identification using spectral clustering

Participant: Frédéric Cazals, Edoardo Sarti.

The decomposition of a biomolecular complex into domains is an important step to investigate biological functions and ease structure determination. A successful approach to do so is the SPECTRUS algorithm, which provides a segmentation based on spectral clustering applied to a graph coding inter-atomic fluctuations derived from an elastic network model.

We present SPECTRALDOM [19], which makes three straightforward and useful additions to SPECTRUS. For single structures, we show that high quality partitionings can be obtained from a graph Laplacian derived from pairwise interactions—without normal modes. For sets of homologous structures, we introduce a Multiple Sequence Alignment mode, exploiting both the sequence based information (MSA) and the geometric information embodied in experimental structures. Finally, we propose to analyze the clusters/domains delivered using the so-called D-Family matching algorithm, which establishes a correspondence between domains yielded by two decompositions, and can be used to handle fragmentation issues.

Our domains compare favorably to those of the original SPECTRUS, and those of the deep learning based method Chainsaw. Using two complex cases, we show in particular that SPECTRALDOM is the only method handling complex conformational changes involving several sub-domains. Finally, a comparison of SPECTRALDOM and Chainsaw on the manually curated domain classification ECOD as a reference shows that high quality domains are obtained without using any evolutionary related piece of information.

SPECTRALDOM is provided in the Structural Bioinformatics Library, see [SBL](#) and [Spectral domain explorer](#).

8.2 Algorithmic foundations

Keywords: Computational geometry, computational topology, optimization, graph theory, data analysis, statistical physics.

8.2.1 Improved seeding strategies for k-means and k-GMM

Participant: Guillaume Carrière, Frédéric Cazals.

In [18], we revisit the randomized seeding techniques for k-means clustering and k-GMM (Gaussian Mixture model fitting with Expectation-Maximization), formalizing their three key ingredients: the metric used for seed sampling, the number of candidate seeds, and the metric used for seed selection. This analysis yields novel families of initialization methods exploiting a *lookahead* principle—conditioning the seed selection to an enhanced coherence with the final metric used to assess the algorithm, and a *multipass strategy* to tame down the effect of randomization.

Experiments show a significant improvement over classical contenders. In particular, for k-means, our methods improve on the recently designed multi-swap strategy (similar results in terms of sum of square errors (SSE), seeding $\sim \times 6$ faster), which was the first one to outperform the greedy k-means++ seeding.

Our experimental analysis also shed light on subtle properties of k-means often overlooked, including the (lack of) correlations between the SSE upon seeding and the final SSE, the variance reduction phenomena observed in iterative seeding methods, and the sensitivity of the final SSE to the pool size for greedy methods.

Practically, our most effective seeding methods are strong candidates to become one of the—if not the—standard technique(s). From a theoretical perspective, our formalization of seeding opens the door to a new line of analytical approaches.

8.2.2 Modeling high dimensional point clouds with the spherical cluster model

Participant: Frédéric Cazals, Antoine Commaret.

In collaboration with L. Goldenberg (former Inria intern).

A parametric cluster model is a statistical model providing geometric insights onto the points defining a cluster. The *spherical cluster model* (SC) approximates a finite point set $P \subset \mathbb{R}^d$ by a sphere $S(c, r)$ as follows. Taking r as a fraction $\eta \in (0, 1)$ (hyper-parameter) of the standard deviation of distances between the center c and the data points, the cost of the SC model is the sum over all data points lying outside the sphere S of their power distance with respect to S . The center c of the SC model is the point minimizing this cost. Note that $\eta = 0$ yields the celebrated center of mass used in KMeans clustering. We make three contributions [21].

First, we show that fitting a spherical cluster yields a strictly convex but not smooth combinatorial optimization problem. Second, we present an exact solver using the Clarke gradient on a suitable stratified cell complex defined from an arrangement of hyper-spheres. Finally, we present experiments on a variety of datasets ranging in dimension from $d = 9$ to $d = 10,000$, with two main observations. First, the exact algorithm is orders of magnitude faster than Broyden-Fletcher-Goldfarb-Shanno (BFGS) based heuristics for datasets of small/intermediate dimension and small values of η , and for high dimensional datasets (say $d > 100$) whatever the value of η . Second, the center of the SC model behaves as a parameterized high-dimensional median.

The SC model is of direct interest for high dimensional multivariate data analysis, and the application to the design of mixtures of SC will be reported in a companion paper.

8.3 Applications in structural bioinformatics and beyond

Keywords: Docking, scoring, interfaces, protein complexes, phylogeny, evolution.

8.3.1 Fold or flop: quality assessment of AlphaFold predictions on whole proteomes

Participant: Frédéric Cazals, Edoardo Sarti.

Reliability of AlphaFold predictions is primarily assessed by the method's self-reported score predicted Local Distance Difference Test (pLDDT). For model organisms, AlphaFold predictions show that 30% to 40% of all amino acids fall into the low-confidence range of pLDDT. Moreover, pLDDT has occasionally failed to flag predictions that are physically implausible. This raises two fundamental questions: can we identify more robust indicators of reliability? And do unreliable predictions exhibit shared structural or biophysical traits?

To address these questions, we introduce semi-global statistics characterizing packing properties at multiple scales, and performing dimensionality reduction and clustering at once [23]. We use these to perform a systematic whole-proteome structural quality assessment of prediction contained in the AlphaFold Database (AFDB), investigating connections between unreliable predictions, fold classification, and intrinsic disorder propensity.

Our results reveal consistent relationships between low-confidence predictions, clustering of intrinsically disordered regions (IDRs), and distinctive packing properties, thereby highlighting both strengths and limitations of current self-assessment metrics. This work provides a framework for deeper confidence assessment of AlphaFold predictions and offers generalizable strategies for distinguishing reliable from unreliable structural models.

8.3.2 Characterizing the fragmentation of AlphaFold predictions

Participant: Frédéric Cazals, Edoardo Sarti.

The Nobel prize winning program AlphaFold computes plausible structures of (well) folded proteins. The main quality assessment is based on the *predicted Local Distance Difference Test* (pLDDT), a per amino acid confidence score. To enhance quality assessment, we provide novel quantitative measures to identify *coherent* amino acid (a.a.) stretches along the sequence in terms of pLDDT values [22]. These measures, which rely on standard tools from topological data analysis and combinatorics, qualify the coherence / fragmentation of AlphaFold predictions. The outcome of our analysis can readily be used to select reliable regions/domains within proteins whose pLDDT values span the entire pLDDT range.

8.3.3 Orphan genes survey

Participant: Edoardo Sarti, Ercan Seçkin.

Orphan genes are protein-coding genes that lack detectable homologs in other species, making them lineage-specific and evolutionarily enigmatic. This review [20] synthesizes research on orphan genes in animals and fungi, summarizing their prevalence, proposed origins (including divergence and de novo emergence), and biological roles. Orphan genes are implicated in diverse processes such as reproduction, development, adaptation, and disease, highlighting their functional importance. They are especially interesting for computational biology because identifying them challenges homology-based annotation methods and requires novel comparative and statistical approaches. By consolidating scattered knowledge, this work provides a foundation for developing better computational tools to detect, classify, and model the evolution and function of orphan genes.

8.3.4 Orphan genes detection and classification

Participant: Edoardo Sarti, Ercan Seçkin.

Building on the broader synthesis of orphan gene prevalence and function, we provide a focused, data-driven case in plant-parasitic nematodes of the genus *Meloidogyne*. Using comparative genomics across 85 nematode species, we show that orphan genes are not rare anomalies but constitute 18% of the genome, with strong transcriptional support [24]. By integrating synteny and ancestral sequence reconstruction, the work quantifies the relative contributions of divergence and de novo gene birth, directly addressing questions raised in the earlier review. Proteomic and translomic evidence further validates these genes as bona fide coding sequences with distinctive molecular features. Together, this study builds a new and effective pipeline for detecting and classifying orphan genes, and exemplifies how computational approaches can move from cataloging orphan genes to dissecting their origins and linking them to lineage-specific adaptations such as parasitism.

9 Bilateral contracts and grants with industry

Participants: Frédéric Cazals.

9.1 Bilateral contracts with industry

Since Septembre 2025, Frédéric Cazals co-supervises the CIFRE PhD project of Antoine Hauser, in collaboration with the **Vect-Horus** company. The goal of this project is to design new molecules crossing the blood-brain barrier, a notoriously difficult challenge to address the brain—i.e. to send medicines into the brain or molecules used for imaging purposes.

10 Partnerships and cooperations

Participants: Frédéric Cazals, Edoardo Sarti.

10.1 National initiatives

ANR Innuendo. This ANR project (running from 01-2024 to 12-2027) is a joint project with INRAE Jouy-en-Josas (B. Delmas) and IBS Grenoble (W. Weissenhorn), and combines two goals: the first is methodological, and the second is applied.

Methods-wise, our project ambitions to advance the state-of-the-art of flexible computational protein design and binding affinity estimations, which raise difficult high dimensional geometric problems. The novel algorithms will make it possible to explore a larger design space, while at the same time reducing the experimental burden, via superior binding affinity estimates. All methods are made available to the community in the Structural Bioinformatics Library (SBL), a unique software environment providing both low level algorithms and applications for end-users.

Application-wise, we will develop high affinity neutralizing biosynthetic proteins, called α repeat proteins (α Reps), with broad spectrum of recognition for circulating sarbecoviruses and limited sensitivity to immune escape mutations. This will be achieved by a virtuous cycle combining our novel computational protein design methods, as well as experiments for structure (cryoEM, X-ray crystallography) and thermodynamics (binding affinity measurements.)

Action Exploratoire Inria. The AEx DEFINE, involving Inria **ABS** and **Laboratory of Computational and Quantitative Biology** (LCQB) from Sorbonne University started in September 2023, for a period of four years.

ABS develops novel methods to study protein structure and dynamics, using computational geometry/topology and machine learning. LCQB is a leading lab addressing core questions at the heart of modern biology, with a unique synergy between quantitative models and experiments. The goal of DEFINE is to provide a synergy between ABS and LCQB, with a focus on the prediction of protein functions, at the genome scale and for two specific applications (photosynthesis, DNA repair).

Co-supervised PhD thesis Inria-INRAE. The PhD thesis of Ercan Seckin started in October 2023 is co-supervised by Etienne Danchin (supervisor) and Dominique Colinet at the INRAE **GAME** team and Edoardo Sarti at **ABS**.

The thesis title is: "Détection, histoire évolutive et relations structure - fonction des gènes orphelins chez les bioagresseurs des plantes". The two teams are closely collaborating for advancing current knowledge on the emergence of orphan genes/proteins in the *Meloidogyne* genus as well as their structural and functional characterization. Notably, the ABS team will focus on the structural and functional inference, and the interplay between structure and function in the process of gene formation.

PIQ project Eminence. Frédéric Cazals was laureate of the Programme Inria Quadrant project Eminence. Its goal is to revisit the foundations of molecular dynamics, ultimately leading to the precise prediction of protein interactions. This objective relies on novel algorithms exploiting inverse problems, new sampling

techniques, and their integration within the Structural Bioinformatics Library. In doing so, this project will contribute to ushering in a new era for the design of active biomolecules. This project also aims at reaching new heights for the diffusion of our library, the Structural Bioinformatics Library.

10.2 Regional initiatives

Co-supervised master internship with CHU Nice. In the framework of a collaboration with the Inria EPIONE team, Edoardo Sarti and Cécile Rouzier have received funding from Alliance Maladies Rares (AMR) for a 6-month Master 2 internship, with the aim of studying the structure and interaction network of the human Wolfram protein WFS1 in presence of its pathogenic mutations. The expected outcomes are an improved set of criteria to determine the pathogenicities of never-observed mutations, and a clarification on the functional mechanism of the protein.

11 Dissemination

Participants: Frédéric Cazals, Edoardo Sarti, Guillaume Carrière.

11.1 Promoting scientific activities

11.1.1 Scientific events: organization

- Frédéric Cazals was involved in the organization of:
 - Winter School Algorithms in Structural Bioinformatics: *Structure modeling and design towards RNA-based therapeutics*. CIRM, Marseille, 8-12th December, 2024. Web: [AlgoSB](#).
- Edoardo Sarti was involved in the organization of:
 - Multi-Omics and Data Integration conference. CHU L'Archet 2, Nice, 16-17th October, 2025. Web: [MODI](#).

Member of the conference program committees Frédéric Cazals participated to the following program committees:

- Symposium on Solid and Physical Modeling
- Intelligent Systems for Molecular Biology (ISMB)

11.1.2 Invited talks

- Frédéric Cazals:
 - *AlphaFold predictions on whole genomes at a glance: towards a finer characterization of the models' reliability*, Univ. of Athens, April 2025.
 - *Clustering and mixture models: two geometric insights*, Ecole des Mines de Paris, March 2025.
- Edoardo Sarti:
 - *AlphaFold predictions at a glance*, Sorbonne Université, campus Jussieu, Paris, April 2025.

11.1.3 Leadership within the scientific community

- Frédéric Cazals:
 - 2010-. . . : Member of the steering committee of the GDR Bioinformatique Moléculaire, for the Structure and macro-molecular interactions theme.
 - 2017-. . . : Co-chair, with Yann Ponty, of the working group / groupe de travail (GT MASIM - Méthodes Algorithmiques pour les Structures et Interactions Macromoléculaires), within the GDR de Bioinformatique Moléculaire (GDR BIM, [GDR BIM](#)).

11.2 Teaching - Supervision - Juries - Educational and pedagogical outreach

- 2024-. . . : Master Mathématiques, Vision, Apprentissage, ENS Paris-Saclay, *Algorithms and learning for protein science*; F. Cazals (24h). See [Official web site](#) and [Lecture notes](#).
- 2021-. . . : Master Data Sciences & Artificial Intelligence (M2), Université Côte d'Azur; *Geometric and topological methods in machine learning*; F. Cazals, J-D. Boissonnat and M. Carrière, Inria Sophia / (ABS, DataShape, DataShape); Web: [GTML](#).

11.2.1 Supervision

- Frédéric Cazals
 - **PhD thesis, ongoing, October 2023-. . .** : Guillaume Carrière. *Attention mechanisms for graphical models, with applications to protein structure analysis*. Advisor: F. Cazals.
 - **PhD thesis, ongoing, October 2025-. . .** : Antoine Hauser. *Optimization of ligand design with application to therapeutic delivery across the blood-brain barrier*. Co-advised with Gauthier Dangla-Pellisier from [Vect-Horus](#) .
 - **PhD thesis, ongoing, October 2025-. . .** : Victor Gertner. *Multimodal encoding of proteins geared towards the prediction of interactions*. Co-advised with Vincent Mallet, Mines ParisTech.
 - **PDoc, January 2025 – PhD in Mathematics**: Antoine Commaret. Topic: *Geometric analysis for data science*.
 - **Engineer, July 2025-. . . . PhD in physics**: Sikao Guo. *Software engineering for the Structural Bioinformatics Library*.
 - **Engineer, July 2025-. . . . PhD in computer science**: Nelson Feyeux. *Software engineering for the Structural Bioinformatics Library*.
- Edoardo Sarti
 - **PhD thesis, ongoing, October 2023-. . .** : Ercan Seçkin. *Detection, evolutionary history and structure-function relationships of orphan genes in plant parasitic nematodes*. Advisor: E. Danchin (Inrae), D. Colinet, E. Sarti (co-supervision)

11.2.2 Juries

F. Cazals was involved in the following juries:

1. Riccardo Pellarin, Université of Lyon, November 2025. Rapporteur on the Habilitation thesis *Components of the cell: simulation, modeling and design*.
2. Clement Bernard, Université Paris-Saclay, October 2025. Rapporteur of the PhD thesis *Computational methods based on deep learning for the prediction of RNA 3D structure*. Advisor: Fariza Tahy.
3. Hamed Khakzad, Sorbonne University, Mars 2025. Rapporteur on the Habilitation thesis *Machine learning driven integrative structural biology and protein design*.

12 Scientific production

12.1 Major publications

- [1] J.-C. Bermond, D. Mazauric, V. Misra and P. Nain. ‘Distributed Link Scheduling in Wireless Networks’. In: *Discrete Mathematics, Algorithms and Applications* 12.5 (2020), pp. 1–38. DOI: [10.1142/S1793830920500585](https://doi.org/10.1142/S1793830920500585). URL: <https://hal.inria.fr/hal-01977266>.
- [2] J.-D. Boissonnat and D. Mazauric. ‘On the complexity of the representation of simplicial complexes by trees’. In: *Theoretical Computer Science* 617 (29th Feb. 2016), p. 17. DOI: [10.1016/j.tcs.2015.12.034](https://doi.org/10.1016/j.tcs.2015.12.034). URL: <https://hal.inria.fr/hal-01259806> (cit. on p. 10).
- [3] J. Carr, D. Mazauric, F. Cazals and D. J. Wales. ‘Energy landscapes and persistent minima’. In: *The Journal of Chemical Physics* 144.5 (2016), p. 4. DOI: [10.1063/1.4941052](https://doi.org/10.1063/1.4941052). URL: <https://www.repository.cam.ac.uk/handle/1810/253412> (cit. on p. 9).
- [4] F. Cazals, T. Dreyfus, D. Mazauric, A. Roth and C. Robert. ‘Conformational Ensembles and Sampled Energy Landscapes: Analysis and Comparison’. In: *J. of Computational Chemistry* 36.16 (2015), pp. 1213–1231. DOI: [10.1002/jcc.23913](https://doi.org/10.1002/jcc.23913). URL: <https://hal.archives-ouvertes.fr/hal-01076317> (cit. on p. 9).
- [5] F. Cazals and T. Dreyfus. *The Structural Bioinformatics Library: modeling in biomolecular science and beyond*. RR-8957. Inria, 11th Oct. 2016. URL: <https://hal.inria.fr/hal-01379635>.
- [6] F. Cazals and T. Dreyfus. ‘The Structural Bioinformatics Library: modeling in biomolecular science and beyond’. In: *Bioinformatics* 33.8 (1st Apr. 2017). DOI: [10.1093/bioinformatics/btw752](https://doi.org/10.1093/bioinformatics/btw752). URL: <https://hal.inria.fr/hal-01570848> (cit. on p. 10).
- [7] F. Cazals and A. Lhéritier. ‘Beyond Two-sample-tests: Localizing Data Discrepancies in High-dimensional Spaces’. In: *IEEE/ACM International Conference on Data Science and Advanced Analytics*. IEEE/ACM International Conference on Data Science and Advanced Analytics. IEEE/ACM International Conference on Data Science and Advanced Analytics. Paris, France, Mar. 2015, p. 29. URL: <https://hal.inria.fr/hal-01245408> (cit. on p. 10).
- [8] F. Cazals and A. Lhéritier. ‘Low-Complexity Nonparametric Bayesian Online Prediction with Universal Guarantees’. In: *NeurIPS 2019 - Thirty-third Conference on Neural Information Processing Systems*. Vancouver, Canada, 8th Dec. 2019. URL: <https://hal.inria.fr/hal-02425602> (cit. on p. 10).
- [9] F. Cazals, D. Mazauric, R. Tetley and R. Watrigant. ‘Comparing Two Clusterings Using Matchings between Clusters of Clusters’. In: *ACM Journal of Experimental Algorithmics* 24.1 (17th Dec. 2019), pp. 1–41. DOI: [10.1145/3345951](https://doi.org/10.1145/3345951). URL: <https://hal.inria.fr/hal-02425599> (cit. on p. 10).
- [10] A. Chevallier and F. Cazals. ‘Wang-Landau Algorithm: an adapted random walk to boost convergence’. In: *Journal of Computational Physics* 410 (2020), p. 109366. DOI: [10.1016/j.jcp.2020.109366](https://doi.org/10.1016/j.jcp.2020.109366). URL: <https://hal.science/hal-01919860>.
- [11] A. Chevallier, F. Cazals and P. Fearnhead. ‘Efficient computation of the volume of a polytope in high-dimensions using Piecewise Deterministic Markov Processes’. In: *AISTATS 2022 - 25th International Conference on Artificial Intelligence and Statistics*. Virtual, France, 28th Mar. 2022. URL: <https://inria.hal.science/hal-03918039>.
- [12] N. Cohen, F. Havet, D. Mazauric, I. Sau Valls and R. Watrigant. ‘Complexity dichotomies for the Minimum F-Overlay problem’. In: *Journal of Discrete Algorithms* 52-53 (Sept. 2018), pp. 133–142. DOI: [10.1016/j.jda.2018.11.010](https://doi.org/10.1016/j.jda.2018.11.010). URL: <https://hal.inria.fr/hal-01947563> (cit. on p. 10).
- [13] A. Lhéritier and F. Cazals. ‘A Sequential Non-Parametric Multivariate Two-Sample Test’. In: *IEEE Transactions on Information Theory* 64.5 (May 2018), pp. 3361–3370. URL: <https://hal.inria.fr/hal-01968190> (cit. on p. 10).
- [14] S. Marillet, P. Boudinot and F. Cazals. *High Resolution Crystal Structures Leverage Protein Binding Affinity Predictions*. RR-8733. Inria, Mar. 2015. URL: <https://hal.inria.fr/hal-01159641> (cit. on p. 11).

- [15] S. Marillet, M.-P. Lefranc, P. Boudinot and F. Cazals. ‘Novel Structural Parameters of Ig–Ag Complexes Yield a Quantitative Description of Interaction Specificity and Binding Affinity’. In: *Frontiers in Immunology* 8 (9th Feb. 2017), p. 34. DOI: [10.3389/fimmu.2017.00034](https://doi.org/10.3389/fimmu.2017.00034). URL: <https://hal.archives-ouvertes.fr/hal-01675467> (cit. on p. 11).
- [16] A. Roth, T. Dreyfus, C. Robert and F. Cazals. ‘Hybridizing rapidly growing random trees and basin hopping yields an improved exploration of energy landscapes’. In: *J. Comp. Chem.* 37.8 (2016), pp. 739–752. DOI: [10.1002/jcc.24256](https://doi.org/10.1002/jcc.24256). URL: <https://hal.inria.fr/hal-01191028> (cit. on p. 9).
- [17] M. Simsir, I. Broutin, I. Mus-Veteau and F. Cazals. ‘Studying dynamics without explicit dynamics: A structure-based study of the export mechanism by AcrB’. In: *Proteins - Structure, Function and Bioinformatics* (22nd Sept. 2020). DOI: [10.1002/prot.26012](https://doi.org/10.1002/prot.26012). URL: <https://hal.archives-ouvertes.fr/hal-03006981> (cit. on p. 11).

12.2 Publications of the year

International journals

- [18] G. Carrière and F. Cazals. ‘Improved seeding strategies for k-means and k-GMM’. In: *Transactions on Machine Learning Research Journal* (1st Oct. 2025). URL: <https://hal.science/hal-05441325> (cit. on pp. 11, 13).
- [19] F. Cazals, J. Herrmann and E. Sarti. ‘Simpler protein domain identification using spectral clustering’. In: *Proteins - Structure, Function and Bioinformatics* (13th Feb. 2025), pp. 1212–1225. DOI: [10.1002/prot.26808](https://doi.org/10.1002/prot.26808). URL: <https://inria.hal.science/hal-04504447> (cit. on p. 13).
- [20] E. Seçkin, D. Colinet, E. Sarti and E. Danchin. ‘Orphan and de novo Genes in Fungi and Animals: Identification, Origins and Functions’. In: *Genome Biology and Evolution* 17.12 (25th Nov. 2025). DOI: [10.1093/gbe/evaf220](https://doi.org/10.1093/gbe/evaf220). URL: <https://inria.hal.science/hal-05455139> (cit. on p. 15).

Reports & preprints

- [21] F. Cazals, A. Commaret and L. Goldenberg. *Modeling high dimensional point clouds with the spherical cluster model*. 29th Dec. 2025. URL: <https://hal.science/hal-05442010> (cit. on p. 14).
- [22] F. Cazals and E. Sarti. *Characterizing the fragmentation of AlphaFold predictions*. 21st Dec. 2025. URL: <https://hal.science/hal-05438856> (cit. on pp. 11, 15).
- [23] E. Sarti and F. Cazals. *Fold or flop: quality assessment of AlphaFold predictions on whole proteomes*. 19th Dec. 2025. DOI: [10.64898/2025.12.19.695427](https://doi.org/10.64898/2025.12.19.695427). URL: <https://hal.science/hal-05438855> (cit. on pp. 11, 14).
- [24] E. Seçkin, D. Colinet, M. Bailly-Bechet, A. Seassau, S. Bottini, E. Sarti and E. G. Danchin. *Identification, evolutionary history and characteristics of orphan genes in root-knot nematodes*. 21st Dec. 2025. URL: <https://hal.science/hal-05438858> (cit. on p. 15).

12.3 Cited publications

- [25] S. Adcock and A. McCammon. ‘Molecular dynamics: survey of methods for simulating the activity of proteins’. In: *Chemical reviews* 106.5 (2006), pp. 1589–1615 (cit. on p. 9).
- [26] F. Alber, S. Dokudovskaya, L. Veenhoff, W. Zhang, J. Kipper, D. Devos, A. Suprpto, O. Karni-Schmidt, R. Williams, B. Chait, A. Sali and M. Rout. ‘The molecular architecture of the nuclear pore complex’. In: *Nature* 450.7170 (2007), pp. 695–701 (cit. on p. 6).
- [27] K. Ball and R. Berry. ‘Dynamics on statistical samples of potential energy surfaces’. In: *The Journal of chemical physics* 111.5 (1999), pp. 2060–2070 (cit. on p. 9).
- [28] H. Callen. *Thermodynamics and an Introduction to Thermostatistics*. Wiley, 1985 (cit. on p. 9).

- [29] L. Cao, I. Goreshnik, B. Coventry, J. Case, L. Miller, L. Kozodoy, R. Chen, L. Carter, A. Walls, Y.-J. Park, E.-M. Strauch, L. Stewart, M. Diamond, D. Veessler and D. Baker. ‘De novo design of picomolar SARS-CoV-2 miniprotein inhibitors’. In: *Science* 370.6515 (2020), pp. 426–431 (cit. on pp. 6–8).
- [30] J. Carr, D. Mazauric, F. Cazals and D. J. Wales. ‘Energy landscapes and persistent minima’. In: *The Journal of Chemical Physics* 144.5 (2016), p. 4. DOI: [10.1063/1.4941052](https://doi.org/10.1063/1.4941052). URL: <https://www.repository.cam.ac.uk/handle/1810/253412> (cit. on p. 9).
- [31] B. Cousins and S. Vempala. ‘A practical volume algorithm’. In: *Mathematical Programming Computation* 8.2 (2016), pp. 133–160 (cit. on p. 9).
- [32] D. Frenkel and B. Smit. *Understanding molecular simulation*. Academic Press, 2002 (cit. on p. 8).
- [33] R. Kannan, L. Lovász and M. Simonovits. ‘Random walks and an $O^*(n^5)$ volume algorithm for convex bodies’. In: *Random Structures & Algorithms* 11.1 (1997), pp. 1–50 (cit. on p. 9).
- [34] D. Landau and K. Binder. *A guide to Monte Carlo simulations in statistical physics*. Cambridge university press, 2014 (cit. on p. 9).
- [35] T. Lelièvre, G. Stoltz and M. Rousset. *Free energy computations: A mathematical perspective*. World Scientific, 2010 (cit. on p. 9).
- [36] C. Schön and M. Jansen. ‘Prediction, determination and validation of phase diagrams via the global study of energy landscapes’. In: *Int. J. of Materials Research* 100.2 (2009), p. 135 (cit. on pp. 8, 9).
- [37] A. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Židek, A. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, K. Pushmeet, D. Jones, D. Silver, K. Kavukcuoglu and D. Hassabis. ‘Improved protein structure prediction using potentials from deep learning’. In: *Nature* (2020), pp. 1–5 (cit. on p. 6).
- [38] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan and W. Wriggers. ‘Atomic-level characterization of the structural dynamics of proteins.’ In: *Science* 330.6002 (2010), pp. 341–346. URL: <http://dx.doi.org/10.1126/science.1187409> (cit. on pp. 6, 8).
- [39] D. J. Wales. *Energy Landscapes*. Cambridge University Press, 2003 (cit. on pp. 8, 9).
- [40] L.-P. Wang, T. J. Martinez and V. S. Pande. ‘Building force fields: an automatic, systematic, and reproducible approach’. In: *The journal of physical chemistry letters* 5.11 (2014), pp. 1885–1891 (cit. on p. 8).