Département
D5: Complex Systems, Artificial Intelligence and Robotics

Équipe CAPSID

Computational Algorithms for Protein and RNA Structure and Interaction

Loria

Laboratoire lorrain de recherche en informatique et ses applications

Rapport d'activité 2025

En partenariat avec
cnrs
UNIVERSITÉ DE LORRAINE
Inria
CentraleSupélec

Contents

Team CAPSID

Keywords

Structural bio-informatics, Computational structural biology, molecular modeling.

1 Team members, visitors, external collaborators

Research Scientists

- Isaure Chauvot de Beauchêne [Team leader, CNRS, Researcher]
- Bernard Maigret [CNRS, Emeritus]

Faculty Members

- Malika Smaïl-Tabbone [UL, Associate Professor]

Post-Doctoral Fellows / engineers

- Diego Amaya Ramirez [CNRS, from Oct 2025]

PhD Students

- Tristan Haquard [Cifre CNRS-bYoRNA]
- Victor Pryakhin [UL]

Interns and Apprentices

- Louise Conceiro [CNRS, Intern, Nov 2025 - Apr 2026]
- Paul Malvaux [UL, Intern,]
- Ibrahim Mechaouat [UL, Intern,]

Administrative Assistant

- Antoinette Courrier [CNRS]

Visiting Scientists

- Fabiano Cavalcanti Fernandes [Oct 2024 - Feb 2025]

2 Overall objectives

2.1 Computational Challenges in Structural Biology

Many of the processes within living organisms can be studied and understood in terms of biochemical interactions between biomolecules such as DNA, RNA, proteins, and small ligands. For a first approximation, DNA may be considered to encode the blueprint for life, whereas proteins and RNA make up the three-dimensional (3D) molecular machinery.

Most biological processes are governed by complex systems of proteins and/or RNA that interact with each other, forming molecular complexes, with a range of specificity and binding affinity. It is becoming increasingly feasible to isolate and characterize some of the individual molecular components of such systems, but it remains difficult to achieve detailed models of how these complex systems actually work.

This is especially true for complexes involving highly flexible molecules, such as single-stranded RNA (ssRNA) and intrinsically disordered proteins (IDP). Those can adopt a large variety of 3D conformations, whose propensities (probabilities) depend on biological contexts: cell location, metabolite or drug binding to it, etc.

Consequently, a new multidisciplinary approach has emerged, called integrative structural biology, that aims to bring together experimental data from a wide range of sources and resolution scales to meet this challenge [0, 0].

Understanding how biological systems work dynamically, at the level of 3D molecular structures, presents fascinating challenges for biologists and computer scientists alike. Despite being made from a small set of simple chemical building blocks, protein and nucleic acid (NA) molecules have a remarkable ability to self-assemble into complex molecular machines that carry out very specific biological processes. As such, these molecular machines may be considered complex systems because their properties are much larger than the sum of the properties of their component parts.

2.2 Two Research Axes

Our objective at CAPSID is to develop algorithms and software that advance the study of biological systems and phenomena from a structural point of view. We focus on creating methods to model highly flexible complexes involving proteins, nucleic acids and/or small biomolecules.

Our goals are to:

- Develop computational techniques for representing, analyzing, and comparing biomolecular 3D conformations and interactions.
- Understand structure-dynamic-function relationships of biomolecular complexes.
- Identify or design novel protein binders, such as artificial proteins, RNA, or small-molecule drugs, for therapeutic or biotechnological applications.

The CAPSID team is organised according to two research axes:

- Axis 1: Knowledge discovery and machine learning for understanding bio-molecular interactions
- Axis 2: Modeling of flexible molecular complexes.

In the first axis, our main objective is to design, implement and test new KDD ("Knowledge Discovery in Databases") and ML (Machine Learning) approaches to exploit relevant information contained and sometimes hidden in many biological databases. These approaches will be oriented towards understanding molecular interactions in living organisms under physiological or pathological conditions.

In the second axis, our main objective is to propose new methods to tackle large conformational ensembles of highly flexible molecules, for modeling/designing the 3D structure of their complexes, thanks to molecular dynamics simulation and/or combinatorial approaches.

Finally, the complementarity of the two axes is expressed through the common objective of proposing possible new treatments against diseases, based on the knowledge extracted and on the advances in 3D modeling of flexible molecular interactions. This objective will benefit from our network of biologist and clinician partners.

3 Research program

3.1 Knowledge discovery and machine learning for understanding bio-molecular interactions

3.1.1 Context

In this axis, the CAPSID team develops methods related to knowledge discovery from databases (KDD) and ML. The diversity of biological databases and resources is such today that it is more and more difficult to

consider each database independently from the others [0]. A limited subset of these resources is devoted to the 3D structure of biological objects (proteins, nucleic acids, glycanes...). Structural information is also contained in databases classifying protein domains as building blocks of proteins that can be reused in different proteins sharing the same function (Pfam, CATH and InterPro are well-known examples of such databases) [0, 0, 0]. There are millions of proteins across all living species but only tens of thousands of domains that are combined in proteins. Thus, complex tasks such as predicting protein function or interactions can be simplified when envisaged at the domain level.
Due to the great diversity of databases, Knowledge Graphs (KGs) are more and more used to represent and integrate biological data, information, and knowledge. There is no single definition of KGs as these graphs cover nowadays a large variety of domains and data representation contexts. The main feature that differentiates KGs from classical graphs is the fact that both nodes (or entities) and edges (or relations) in the graph are heterogeneous and belong to various types described in the KG schema (metagraph). The field of biological knowledge discovery and ML in KG is expanding rapidly [0]. Several biological KGs are developed for drug repurposing tasks (e.g. HetioNet [0] or DRKG). Clinicians are also very interested in network science carried out on rich knowledge graphs as a mean to interpret biomarker studies (e.g. CKG [0]). However, there is still a need for building reliable biological KGs and for efficient analysis methods in KGs possibly augmented with supplementary data. Relevant supplementary data in our context are the protein and compound learned representations through the use of pretrained protein/chemical language models [0, 0].

3.1.2 Leveraging on pretrained language models in poor-data settings

Characterising or predicting complex biological phenomena is challenging especially when a small amount of data is available. Indeed, many prediction models when trained on a dataset may show good results in the context of this dataset but show difficulties to generalize to new entities [**ang2020predicting**, 0].

The solution we are investigating is to leverage pretrained models to generate representations of bio-molecules combined with generalizable transfer learning for downstream tasks. Indeed, similarly to large language models, such pretrained models are trained in unsupervised (or self supervised) mode and can exploit huge numbers of proteins or small-molecule sequences. Hence, the learned representations are expected to have captured general biological laws. Nevertheless, training task-specific prediction models using such representations as input needs to be performed on unbiased data to ensure generalization capacity.

3.1.3 Function Annotation in Large Protein Graphs

Knowledge of the functional properties of proteins can shed considerable light on how they might interact. However, many protein sequences in public databases such as UniProt/TrEMBL lack a functional annotation, and retrieving it is a highly challenging problem. We are developing graph-based and machine learning techniques to annotate automatically the available unannotated sequences with functional properties such as Enzyme Commission (EC) numbers and Gene Ontology (GO) terms (note that these terms are organised hierarchically allowing generalization/specialization reasoning). The idea is to transfer annotations from expert-reviewed sequences present in the UniProt/SwissProt database (about 560 thousands entries) to unreviewed sequences present in the UniProt/TrEMBL database (about 80% of 180 millions entries). For this, we have to learn from the UniProt/SwissProt database how to compute the similarity of proteins sharing identical or similar functional annotations. Various similarity measures can be tested using cross-validation approaches in the UniProt/SwissProt database. For instance, we can use primary sequence or domain signature similarities. More complex similarities can be computed with graph-embedding techniques.

3.1.4 Knowledge graph embedding adapted to augmented biomedical Knowledge graphs

KGs are particularly useful and appropriate in biology and medicine, to represent and integrate the heterogeneous contents of biomedical databases [0]. In this context, KGs mostly represent PPIs, integrated with various properties attached to proteins, such as cellular compartment, pathways, binding drugs or relations with diseases. Beside the relations, the KG entities corresponding to bio-molecules (proteins,

RNA, drugs, peptides) can be further described with important intrinsic features such as learned representation (refer to the previous subsection). We aim at assessing and extending GNN embedding methods and alike to such augmented biomedical Knowledge graphs. Our objective is to extract useful knowledge from such graphs in order to better understand and highlight the role of multi-component assemblies in various types of cells or organisms.

3.2 Modelling interactions of flexible molecules

3.2.1 Context

The long-standing challenge of predicting a 3D structure from a protein sequence, known as "the folding problem," was solved in 2021 by methods like AlphaFold2 (DeepMind)[0] and RosettaFold[0]. This breakthrough, demonstrated in the CASP14 (Critical Assessment of Structure Prediction) challenge, was enabled by advances in AI and the growth of experimental 3D structure data in the **Protein Data Bank (PDB)**, the main repository for biomolecular structures.

Similarly, modeling the 3D structure of molecular complexes, referred to as "the docking problem," was addressed in 2022 with AlphaFold-Multimer, at least for well-folded proteins [0, 0]. However, challenges remain in docking disordered proteins or flexible nucleic acids that fold upon binding to proteins.

RNA molecules perform essential biological functions through interactions with proteins. A key challenge in computationally modeling these interactions is the high flexibility of RNA, particularly in single-stranded RNA (ssRNA) regions, that adjust their 3D conformation to bind protein surfaces.

AI-based docking methods still lag behind in modeling RNA-protein interactions involving long ssRNA regions. This is likely due to the scarcity of experimental structural data compare to the multiple conformations a given RNA sequence can adopt. These conformations coexist in equilibrium in the cell, with their prevalence influenced by factors such as chemical and physical conditions (e.g., pH, post-transcriptional modifications, cellular location) and binding partners (e.g., ions, metabolites, proteins).

3.2.2 Discrete combinatorial docking

Conventional docking algorithms typically assume that the 3D structures of proteins or RNA are rigid or only moderately flexible. They often sample a limited number of their conformations before docking, or use normal modes to model flexibility along a few pre-computed harmonic motions during binding. However, these methods are unsuitable for modeling interactions involving disordered molecules that fold upon binding.

The conformational space of a flexible ssRNA is too vast for exhaustive sampling, especially when considering the bound conformation, which may not correspond to a local energy minimum in the free state. Instead, the bound ssRNA conformation must be constructed directly on the protein surface, highlighting the need for specialized folding-docking algorithms.

We use a fragment-based approach that first docks small fragments of ssRNA (typically three nucleotides at a time) onto a protein surface, then combinatorially reassembles those fragments in order to recover a contiguous ssRNA structure on the protein surface [0, 0].

The fragments are docked as rigid, and only the translation and rotation degrees of freedom are explored to produce "poses", i.e. putative positions of a fragment on the protein surface. The diversity of 3D conformations that an RNA fragment can adopt is discretised by a 3D fragment library that must contain all possible internal conformations within a chosen precision. Each docking pose is evaluated by a scoring function, which is an approximation of the interaction energy, statistically derived from all experimental RNA-protein structures. A pool of the best-scored poses is kept for each fragment. Then, chains of geometrically contiguous poses of consecutive fragments are selected as final models.

3.2.3 Integrative (data-driven) modeling

In molecular docking, the search space can be reduced by constraints extracted from additional experimental data beyond the available structures of the ligands. It can restrict the relative position of the molecules, or their conformation preference upon binding. It can be extracted from an in vitro experiment, such as binding assay, or a biophysical experiment, such as NMR. For instance: the impact of a mutation on the binding can indicate that the corresponding residue is at the interface; fluorescence

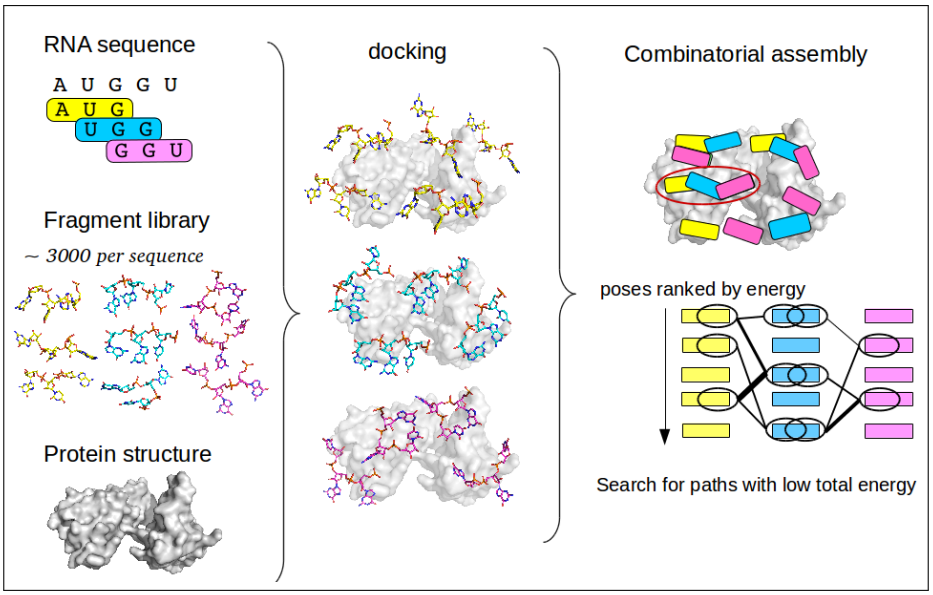

Figure 1: Discrete combinatorial ("fragment-based") docking

transfer experiments can indicate that a pair of residues are located close to each other in the complex; multiple models of a complex can be filtered by fitting to NMR low-resolution data.

Additional constraints can be inferred by learning on all sequences and structures available for a family of protein and/or RNA. For instance, the RNA-binding domains of proteins are well-conserved by evolution, and some of their sequence motifs contain residues that are known to establish specific interactions with nucleic acids, such as "stacking" via aromatic cycles.

The greater the number of such constraints, the more accurate the final models. But at the cost of limiting the modeling to the well-known and well-studied systems, or of costly experiments.

3.3 Mdeling biomolecular interactions for drug discovery

This part of our research focuses on the computational investigation of biomolecular interactions involved in infectious diseases, neurodegenerative disorders, and cancer. The objective is to understand the molecular determinants of protein–protein and protein–ligand interactions and to exploit this knowledge for the discovery of new therapeutic strategies.

The research relies on a combination of computational approaches, including molecular dynamics (MD) simulations, molecular docking, virtual screening, and machine-learning methods. Molecular dynamics simulations are used to characterize the conformational dynamics of biological macromolecules and their complexes, while docking and virtual screening allow the exploration of large chemical spaces to identify potential inhibitors of therapeutic targets.

A particular focus of our work is the integration of large-scale simulations and artificial intelligence approaches in drug discovery pipelines. AI-based prediction methods can accelerate the identification of promising compounds, while physics-based simulations provide a detailed mechanistic understanding of molecular recognition processes.

These approaches are applied to several biological systems of interest, including viral proteins involved in SARS-CoV-2 infection, enzymes essential for bacterial or fungal metabolism, and protein networks implicated in neurodegenerative diseases. The ultimate goal is to identify new molecular targets and design compounds capable of modulating their activity.

4 Application domains

4.1 Biomedical Knowledge Discovery

Participants: Malika Smaïl-Tabbone *(contact person)*, Paul Malvaux, Ibrahim Mechaouat, Fabiano Cavalcanti.

Our main application for Axis 1 : "Knowledge discovery and machine learning for understanding bio-molecular interactions" concerns biomedical knowledge discovery. We intend to develop KDD approaches on preclinical (experimental) or clinical datasets integrated with knowledge graphs with a focus on discovering which PPIs or molecular machines play an essential role in the onset of a disease and/or for personalised medicine.

Using our expertise in receptor-ligand docking, we are investigating possible cross-talks between mineralocorticoid and other nuclear receptors.

A new application has started in 2023 in the context of an interdisciplinary project funded by the CAPES-COFECUB, in collaboration with two Brazilian universities (Catholic University of Brasilia and Universisty of Maringa). It concerns ML approach to identify candidate drugs targeting vital proteins in several pathogenic fungi.

Moreover, simulating biological networks will be important to understand biological systems and test new hypotheses. One major challenge is the identification of perturbations responsible for the transformation of a healthy system to a pathological one and the discovery of therapeutic targets to reverse this transformation. Control theory, which consists in finding interventions on a system in order to prevent it to go in undesirable states or to force it to converge towards a desired state, is of great interest for this challenge. It can be formulated as “How to force a broken system (pathological) to act as it should do (normal state)?”. Many formalisms are used to model biological processes, such as Differential Equations (DE), boolean networks, cellular automata. In her PhD thesis, Athenaïs Vaginay investigated ways to find a boolean network fitting both the knowledge about topology and state transitions “inferred“ from experimental data. This step is known as “boolean function synthesis”. Our aim is to design automated methods for building biological networks and define operators to intervene on them [0]. Our approaches will be driven by knowledge and will keep close connection with experimental data.

4.2 Design of protein-binding oligo-nucleotides

Participants: Isaure Chauvot de Beauchêne *(contact person)*, Tristan Haquard, Louise Monciero, Aurélien Back.

collaborators: Alexandre Bonvin, Anna Kravchenko (Univ Utrecht)

The main current application of Axis 2 is the design of oligonucleotides that can bind to a target protein. This application is of utmost interest to pharmaceutical companies for several usages: creating therapeutical small RNA to compete with a physiological/pathological protein-RNA interaction; facilitating the in vivo production process of vaccine RNA; addressing drug-containing liposomes with membrane-inserted oligoRNA that bind to specific cellular proteins.

In that direction, our collaborators develop chemically modified oligonucleotides to improve their binding specificity, safety as drug, and resistance to enzyme degradation. We published an approach to improve the design of a modified ssRNA sequence targeting a protein of interest using the fragment-based and graph-based approach [**yacoub:hal-03816423**].

This proof-of-principle has interested two companies, Sanofi and the start-up bYoRNA, with whom we are writing two projects for CIFRE theses.

4.3 3D structural differences among HLA antigens

Participants: Malika Smaïl-Tabbone *(contact person)*, Diego Amaya Ramirez, Bernard Maigret.

collaborators: Marie-DominiqueDevignes

This application domain has emerged in Axis 2 through the Inria-Inserm PhD thesis project of Diego Amaya Ramirez, in collaboration with the Immunology and Histocompatibility Laboratory at the APHP Saint-Louis Hospital in Paris. Differences between donor and recipient HLA proteins are one of the major limitations of organ transplant because of HLA ubiquity on cells of tissues and organs [0]. Indeed, in case of incompatibility between the HLA proteins of the donor and those of the patient, an immune response is triggered in the patient that can result in rejection of the transplanted organ. The thesis project aims at deciphering the role played by tiny 3D structure differences between donor and recipient HLA proteins in determining the production of donor-specific antibodies by the recipient. We have been developing methods to compare local structure variations between HLA proteins, taking into account the dynamics of these proteins.

4.4 Allosteric pathways in protein-nucleic acid complexes

Participants: Malika Smaïl-Tabbone *(contact person)*, Viktor Pryakhin.

Novel methodological approaches based on deep learning (AlphaFold2 and RosettaFold) have started to make remarkable advances in protein structure prediction and design. However, our knowledge regarding their dynamical behaviour and function is still highly limited. One important example is the notion of allostery which refers to processes whereby a binding event at one site of a biological macromolecule affects the binding activity at another distinct functional site, enabling the regulation of the corresponding function. The allosteric behavior of a macromolecular system arises from the properties of the native free-energy landscape of the system, and how this landscape is remodeled by various perturbations, such as ligand binding, protonation, mutations, post-translational modifications, or interactions with other molecules. Therefore, understanding allosteric pathways of communication is of high importance and is not well understood yet. All-atom MD simulations could be used to capture subtle dynamical changes that are associated with allosteric signaling. Moreover, graph theory-based methods were developed to investigate the set of trajectories generated by MD simulations and extract allosteric pathways. Y. Karami, Elodie Laine and Alessandra Carbone (LCQB) had previously developed a method called COMMA (COMmunication Mapping) that describes the dynamical architecture of a protein by predicting the network of communications within the system [0].
Yasaman Karami and Malika Smaïl-Tabbone recruited a PhD student, Victor Pryakhin who obtained a doctoral contract from Université de Lorraine and started in October 2023. His PhD subject is to investigate communication networks in protein-RNA complexes using deep learning approaches. They obtained computational resources on Jean Zay super computer to perform MD simulations on a set of protein-RNA complexes. The results of this computations will be used in Viktor Pryakhin's doctoral project.

4.5 Antiviral drug discovery targeting SARS-CoV-2

Participants: Bernard Maigret, Isaure Chauvot de Beauchene.

The COVID-19 pandemic highlighted the importance of understanding the molecular mechanisms governing viral infection and replication. Our work contributes to this effort through computational studies of key viral proteins and the design of molecules capable of inhibiting their activity.

One focus of our research concerns the interaction between the SARS-CoV-2 spike protein and the human ACE2 receptor, which mediates viral entry into host cells. Understanding how mutations affect the stability and dynamics of this complex is crucial for explaining differences in infectivity between viral variants and for guiding the development of therapeutic strategies.

Another important target is the viral main protease (Mpro), an enzyme essential for viral replication. Computational methods are used to screen large chemical libraries and identify potential inhibitors

capable of blocking the protease activity. These approaches contribute to the development of new antiviral strategies against emerging viral variants

4.6 Antimicrobial drug discover

Participants: Bernard Maigret.

The emergence of antimicrobial resistance and the persistence of infectious diseases such as tuberculosis require the continuous development of new therapeutic agents.

Our work contributes to this field through the study of novel synthetic molecules with activity against mycobacteria, as well as the investigation of enzymatic targets involved in microbial metabolism. Computational methods are used to analyze the structural determinants of ligand binding and to guide the design of improved compounds.

The integration of molecular modeling with experimental microbiology allows the identification of promising candidate molecules and provides insights into their mechanisms of action.

4.7 Control of plant pathogens

Participants: Bernard Maigret.

Plant pathogens represent a major challenge for agriculture. Our research addresses this issue by investigating enzymatic targets involved in fungal metabolism.

In particular, we study tryptophan synthase from the fungus Hemileia vastatrix, the causal agent of coffee leaf rust disease. This enzyme is essential for tryptophan biosynthesis and represents a promising target for the development of new fungicides.

Computational approaches are used to characterize the structure and dynamics of the enzyme and to identify natural compounds capable of inhibiting its activity. Such studies contribute to the development of environmentally friendly strategies to control fungal diseases affecting crops.

4.8 Multi-target drug discovery for neurodegenerative diseases

Participants: Bernard Maigret.

Alzheimer's disease is a multifactorial neurodegenerative disorder involving complex molecular pathways. Because therapies targeting a single protein have shown limited success, current research increasingly focuses on multi-target therapeutic strategies.

Our work aims to identify multi-target directed ligands (MTDLs) capable of simultaneously modulating several proteins involved in Alzheimer's disease pathology. In particular, we investigate ligands targeting proteins interacting with Tau, such as GSK3β, FKBP51, and TTBK1.

Computational screening and modeling approaches are used to identify compounds capable of interacting with several targets simultaneously, providing a promising avenue for the development of more effective treatments.

5 Highlights of the year

- Bernard Maigret had his CNRS emeritus prolongated for 5 years.
- Isaure CB accepted the invitation to be co-responsible of the GT ("groupe de travail") *MASIM* ("Méthodes Algorithmiques pour les Structures et Interactions Macromoléculaires") belonging to the GDR BIMMM ("Bioinformatique Moléculaire: Modélisation et Méthodologie")

6 New software, platforms, open data

Section et sous sections optionnelles.

6.1 New software

6.2 New platforms

6.3 Open data

7 New results

7.1 Knowledge graph analysis with embedding-based methods

Participants: Malika Smaïl-Tabbone.

7.2 Protein-ligand interaction prediction in poor-data settings

Participants: Malika Smaïl-Tabbone, Fabiano Cavalcanti Fernandes, Karina Sakita.

7.3 Structural basis of donor-specific antibody response in graft rejection.

Participants: Diego Ramirez, Malika Smail-Tabbone.

7.3.1 Molecular dynamics analysis of SARS-CoV-2 spike–ACE2 interactions

Participants: Bernard Maigret, Isaure Chauvot de Beauchêne.

We performed extensive molecular dynamics simulations of the complexes formed by the receptor-binding domain (RBD) of the SARS-CoV-2 spike protein with the human ACE2 receptor for the wild-type, Beta, Delta, and Omicron variants. Carrying out a comprehensive analysis of residue interactions within and between the two partners allowed us to draw the profile of each variant by using complementary methods. We were therefore able to highlight the residues most involved in electrostatic interactions, which make a strong contribution to the binding with highly stable interactions between spike the protein partners. Moreover, apolar contacts made a substantial and complementary contribution in Omicron with the detection of two hydrophobic patches. Contact networks and cross-correlation matrices were able to detect subtle changes at point mutations as the S375F mutation occurring in all Omicron variants, which is likely to confer an advantage in binding stability. This study brings new highlights on the dynamic binding of spike RBD to hACE2, which may explain the final persistence of Omicron over Delta.

• Akinwumi IA, Bheemireddy S, Chaloin L, Perez S, Khakzad H, Maigret B, Karami Y. Stoichiometric insights into SARS-CoV-2 spike-ACE2 binding across variants. Comput Struct Biotechnol J. 2025 Jul 24;27:3285-3291. • Gheeraert A, Leroux V, Mias-Lucquin D, Karami Y, Vuillon L, Chauvot de Beauchêne I, Devignes MD, Rivalta I, Maigret B, Chaloin L. Subtle Changes at the RBD/hACE2 Interface During SARS-CoV-2 Variant Evolution: A Molecular Dynamics Study. Biomolecules. 2025 Apr 7;15(4):541.

7.3.2 Design of potential inhibitors of the SARS-CoV-2 main protease

Participants: Bernard Maigret.

The global impact of SARS-CoV-2 has highlighted the urgent need for novel antiviral therapies. This study integrates combinatorial chemistry, molecular docking, and deep learning to design, evaluate and synthesize new pyrazole derivatives as potential inhibitors of the SARS-CoV-2 main protease (Mpro). A library of over 60,000 pyrazole-based structures was generated through scaffold decoration to enhance chemical diversity. Virtual screening employed molecular docking (ChemPLP scoring) and deep learning (DeepPurpose), with consensus ranking to identify top candidates. Binding free energy calculations refined the selection, revealing critical structural features such as tryptamine and N-phenyl fragments for Mpro binding. High-temperature solvent-free amidation allowed the synthesis of a selected derivative. Final compounds demonstrated favorable drug-likeness properties based on Lipinski's and Veber's rules. This work highlights the integration of computational and synthetic strategies to accelerate the discovery of Mpro inhibitors and provides a framework for future antiviral development. • Rational design and synthesis of pyrazole derivatives as potential SARS-CoV-2 Mpro inhibitors: An integrated approach merging combinatorial chemistry, molecular docking, and deep learning. Bioorg Med Chem. 2025 Apr 1;120:118095.

7.3.3 Molecular dynamics study of fungal tryptophan synthase

Participants: Bernard Maigret.

This is the first study aimed at determining the activity of two innovative synthetic 1,3,4-oxadiazole molecules, (4-[cyclohexyl(ethyl) sulfamoyl]-N-[5-(furan-2-yl)-1,3,4-oxadiazol-2-yl]benzamide), namely LMM11, and ((N-cyclo-hexyl-N-ethylsulfamoil)-N-(5- (4-fluorophenyl)-1,3,4-oxadiazol-2-il) benzamide), namely LMM6, against Mycobacterium tuberculosis and nontuberculous mycobacteria, and their ability to present synergism in activity against M. tuberculosis when combined with anti-TB drugs. In vitro cytotoxicity studies were conducted showing that the new oxadiazoles showed activity against mycobacteria, in fact, more pronounced against M. tuberculosis, and seem to bring light to the synthesis of new antimicobacterial. • Andriato PM, Baldin VP, de Almeida AL, Sampiron EG, de Vasconcelos SSN, Caleffi-Fercioli KR, Scodro RBL, Meneguello JE, Maigret B, Kioshima ÉS, Cardoso RF. 1,3,4-oxadiazoles with effective anti-mycobacterial activity. Lett Appl Microbiol. 2025 Mar 3;78(3):ovaf029.

7.3.4 Virtual screening of natural compounds targeting fungal TRPS

Participants: Bernard Maigret.

We conducted 1 µs all-atom molecular dynamics simulations on Hemileia vastatrix TRPS to address two questions: (i) the role of the linker segment and (ii) the comparative mode of action. Since there is not an experimental structure, we started our simulations with homology modeling. Based on the results, it seems that TRPS makes use of an already-existing tunnel that can spontaneously move the indole moiety from the α catalytic pocket to the β one. Such behavior was completely disrupted in the simulation without the linker. In light of these results and the $\alpha\beta$ dimer's low stability, the full-working TRPS single genes might be the result of a particular evolution • Martins NF, Viana MJA, Maigret B. Fungi Tryptophan Synthases: What Is the Role of the Linker Connecting the α and β Structural Domains in Hemileia vastatrix TRPS? A Molecular Dynamics Investigation. Molecules. 2024 Feb 6;29(4):756.

Considering the significant losses that Hemileia vastatrix causes to coffee plantations, our next course

of action this 2026 year is to use the TRPS to look for substances that can block tryptophan production and therefore control the disease.

7.4 Modeling of ssRNA-protein interaction

7.4.1 Completeness of RNA conformations in databases

Participants: Isaure Chauvot de Beauchêne, Sjoerd de Vries *(SISR)*.

Our RNA 3D fragment library is constructed by retrieving all experimentally determined RNA 3D structures from the PDB, extracting all fragments of n contiguous nucleotides (so far mainly $n = 3$), and clustering them using approximate geometric similarity criteria. This reduces redundancy, improves computational efficiency, and limits the number of docking models to be evaluated.

The choice of fragment size n must balance two competing requirements:

- **Scoring efficiency**: Larger fragments form more specific interactions with proteins, improving docking discrimination—especially when conserved protein contact points are involved. Previous tests showed that $n \geq 3$ is required for effective scoring.
- **Library completeness**: A near-native docking pose can only be found if a close structural analogue exists in the fragment library. Smaller fragments are easier to cover exhaustively at a given geometric precision. The completeness and precision of the fragment library therefore set an upper limit on the correctness of the final model.

Last year, we quantitatively compared the completeness of dinucleotide (2-nt) and trinucleotide (3-nt) libraries at 1 Å precision, by evaluating which fragments have at least one structural neighbor within 1 Å in the PDB. We observed that 7% of 3-nt lack any neighbor within 1 Å, compared with only 2% of 2-nt. These results motivated a revised strategy using 2-nt fragments.

This year, we extended our completeness analysis across RNA families. We showed that the high completeness level of 2-nt (~ 95%) is maintained even when searching for fragments in RNA families different from the one from which they originate. This indicates that most local 2-nt conformations are transferable across RNA families and suggests that it is possible to model RNAs from newly discovered or experimentally uncharacterized families using already observed 2-nt fragments.

More unexpectedly, we found that more than 99% of the geometrically possible combinations of two 2-nt fragments into a 3-nt fragment (Figure 2) are absent from the PDB, despite being geometrically feasible. Statistically, such depletion is highly improbable. One possible explanation is that these missing conformations correspond to disordered RNA states that are not experimentally resolved. This previously unnoticed contradiction raises fundamental questions about the compatibility criteria between local fragments and the intrinsic stability of RNA local structures.

These results were presented by ICB in an invited lecture at the international congress 3DBioinfo (March 2025, Barcelona).

7.4.2 Statistical analysis of the geometry of protein-RNA stacking interactions

Participants: Isaure Chauvot de Beauchêne, Sjoerd de Vries *(SISR)*, Tristan Hacquard.

During RNA fragment docking, the search space of fragment conformations and positions on the protein can be reduced by introducing constraints derived from prior knowledge about the system under study. This decreases the number of candidate poses and improves the accuracy of the top-ranked solutions. One such source of knowledge is the presence, on the protein surface, of evolutionarily

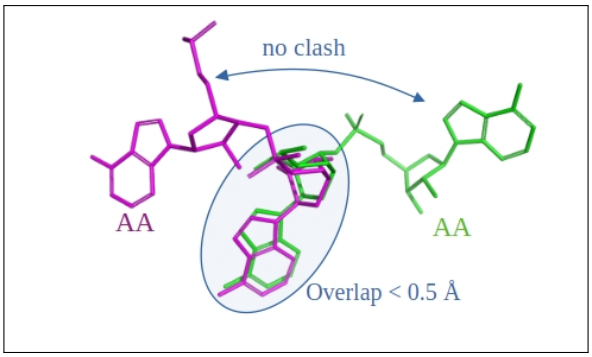

Figure 2: combination of 2-nt into 3-nt fragments

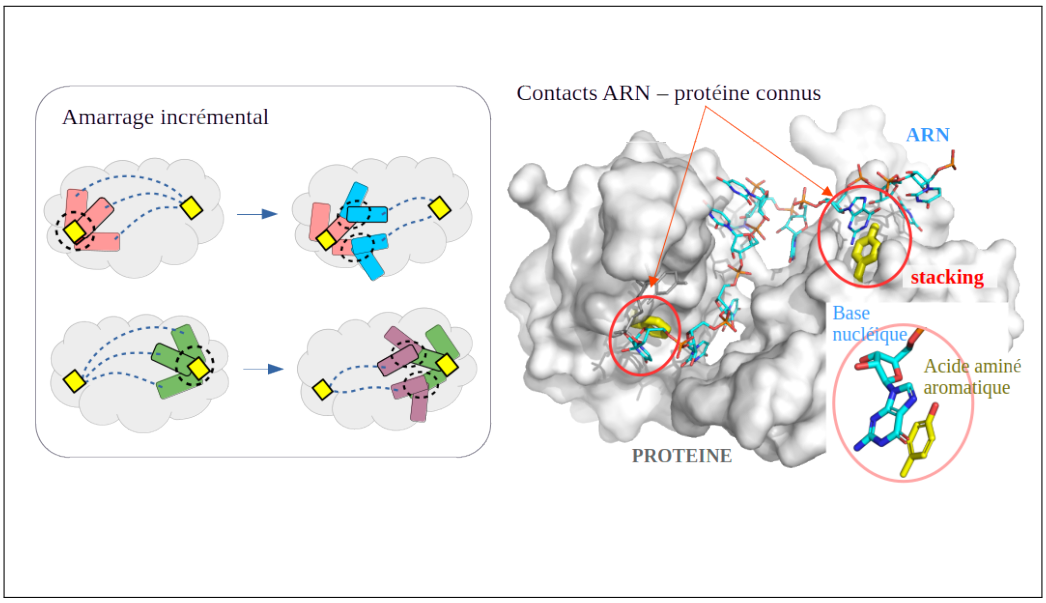

Figure 3: Incremental docking from stacking contacts

conserved aromatic amino acids forming stacking interactions with RNA bases (Figure 3). We had previously provided a proof of principle for exploiting this information in incremental docking, assuming the coordinates of the stacked bases to be known.

For application to real cases, however, these base coordinates must be inferred directly from the protein structure. To this end, we performed a statistical analysis of the relative positioning of aromatic amino acid rings and nucleobases engaged in stacking interactions across all known RNA–protein structures.

First, we observed that when the distance between ring centers exceeds $d > 5.3$ Å, the relative orientation of the rings is indistinguishable from random. In contrast, when $d < 4$ Å, the rings are nearly parallel (angle θ between ring planes smaller than 24°) (Figure 4).

This practical definition of stacking is relatively loose, meaning that many geometries satisfy these criteria but are nevertheless not observed in real stacking interactions. We therefore refined the geometrical definition by performing additional analyses, in particular of the d_{xy} and d_z components of the distance d, where the z axis is defined as normal to the plane of the protein aromatic ring. This led to the following geometric definition of a stacking interaction:

$$(\sin(\theta) < 0.4) \cap \left(d < 5.3\,\text{Å}\right) \cap \left(2.3\,\text{Å} < d_z + 1.78966\,d_{xy} < 3.8\,\text{Å}\right) \cap \left(-45^\circ < \phi_{AB} < +45^\circ\right)$$

This definition can be used to constrain and filter the possible docking poses of nucleotide fragments interacting with a protein through stacking. This work enabled a drastic reduction of the search space

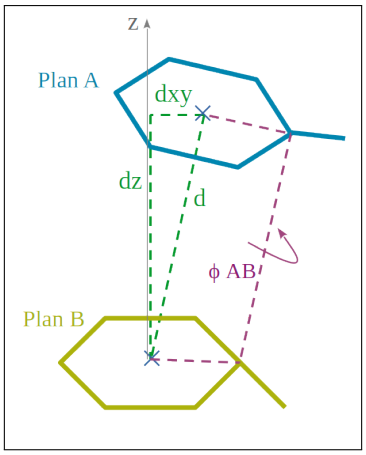

Figure 4: Geometrical features of RNA–protein stacking

during RNA modeling in a specific (but very common) class of RNA-protein interactions.

7.4.3 Modeling of RNA fragments bound through stacking interactions

Participants: Isaure Chauvot de Beauchêne, Sjoerd de Vries *(SISR)*.

This work comprises two main components.

1. High-precision modeling from double stacking constraints. First, we show that if an RNA dinucleotide or trinucleotide is known to form two stacking interactions with a protein, the correct binding mode can be modeled with sub-angstrom accuracy, although some ambiguity may remain.

We introduce a new systematic sampling scheme based on a triple discretization strategy:

- discretization of RNA 3D conformations (using our fragment library),
- rotational discretization (based on systematic enumeration of all fragment rotations),
- translational discretization on a 3D grid.

The sampling space is filtered using the statistical stacking constraints described previously (see Section 3), resulting in approximately 100 million docking poses at 0.5 Å resolution. These poses are then scored using ATTRACT, and the top 1,000 to 10,000 solutions are retained, resulting in an effective sampling precision below 1 Å.

2. Prediction of double stacking interactions from protein structure alone. Second, we demonstrate that, in certain cases, the presence of a double stacking interaction with RNA can be predicted solely from the protein structure.

For the family of proteins containing RNA Recognition Motifs (RRM), an evolutionarily conserved double stacking interaction is observed in 84% of RNA–protein structures available in the PDB. This stacking interaction is strongly correlated with a specific protein structural motif (the 3D conformation of a group of amino acids). In the absence of this motif (defined as a structural deviation > 0.5 Å), only one of the three RRM domains exhibits a double stacking interaction with RNA.

Remarkably, we found that this structural motif can also serve as a predictor in another family of RNA-binding proteins, namely Cold Shock Domain (CSD) proteins. Although the correlation is weaker than for RRMs, it remains statistically significant.

Finally, using 3D models from AlphaFoldDB, we classified all human RRM and CSD domains as competent or incompetent for double-stacking interactions. For all competent domains, and for all possible RNA sequences, we predicted the structure of the bound RNA fragment.

Impact: This work demonstrates that the structure of an RNA fragment bound to a protein can be predicted from two known stacking contacts. This establishes the basis for very high–precision RNA single-strand (ssRNA)–protein modeling.

7.4.4 ssRNA–protein complex models reaching sub-angstrom accuracy

Participants: Isaure Chauvot de Beauchêne, Sjoerd de Vries *(SISR)*.

Impact: In this recent work, we constructed full ssRNA–protein complex models from stacking-bound fragments, achieving an unprecedented level of accuracy. The combined use of our new 2-nt fragment library, of the systematic triple-discretization sampling scheme and of the stacking constraints made it possible to define a sampling space that includes an approximation of the experimental structure at sub-angstrom precision.

Fragments are docked iteratively, starting from anchor points defined by stacked nucleotide bases. The poses of each fragment are filtered according to their geometric compatibility with the previous fragment (or anchor point) and their score (i.e., interaction energy). The entire procedure is exhaustive, meaning that all possible RNA chains within the defined sampling space are considered.

We show that within this sampling space, the model closest to the experimental structure is always a Pareto optimum. This means that there exists no RNA chain that is strictly better in all respects—both in terms of score and geometric overlap of fragments—than this closest model: for any alternative chain, at least one of these criteria is worse for at least one fragment.

The current procedure is not yet fully unbiased. However, as a proof of principle, we demonstrate that manageable numbers of models (suitable for subsequent refinement) are still obtained even when the protocol is made less biased—for example, by using an approximate protein structure, by applying non-Pareto selection thresholds for score and geometric overlap, and/or by using predicted coordinates for the anchor points.

8 Bilateral contracts and grants with industry

8.1 Bilateral contracts with industry

8.1.1 Cifre bYoRNA

Participants: Isaure Chauvot de Beauchene, Tristan Hacquard.

Company: bYoRNA
Duration: 3 years

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 Visits of international scientists

Fabiano Cavalcanti Fernandes

Status: Post-Doc

Institution of origin: Instituto Federal de Brasília

Country: Brazil

Dates: October 6th 2024 - January 31st 2025

Context of the visit: CAPES-COFECUB Project

Mobility program/type of mobility: research stay

9.2 European initiatives

9.2.1 Visits to international teams

INS2I DIALOG programm "RNAdock"

person: Isaure Chauvot de Beauchene

Visited institution: Bonvin's Lab at Utrecht University

Country: Netherland

Dates: July 2025

Context of the visit: Combining ATTRACT's (Loria) sampling and HADDOCK's (Utrecht) scoring for RNA-protein modeling.

type of mobility: research stay

9.3 National initiatives

9.3.1 ANR EPIHLA

Participants: Malika Smaïl-Tabbone, Diego Amaya Ramirez, Bernard Maigret.

Title: HLA compatibility in organ transplantation : from antigens to epitopes (EPIHLA)

Duration:

Coordinator: Pr. Jean-Luc Taupin (Inserm U976, Saint-Louis Hospital, Paris)

LORIA contact: Malika Smaïl-Tabbone.

Partner Institutions:

- Inserm U976 IRSL Saint-Louis Hospital (Paris)
- LORIA CNRS (Nancy)
- INSERM U1016 Cochin Institute (Paris)
- CNRS U144 Institut Curie (Paris)

Summary: The EPIHLA project has two major aims. (1) It aims at correctly representing HLA molecule 3D structure and superimposing predicted conformations in order to identify 3D differences that could constitute epitopes and eplets, targets of donor-specific antibodies. (2) It aims at developing the capacity to isolate and clone anti-HLA antibody genes from patients' B lymphocytes. The results will provide decisive new information on the understanding of humoral alloreactivity and will make it possible to better anticipate transplant rejection. This project was initially based on the Inria-Inserm PhD project of Diego Amaya Ramirez (2019-2022). This thesis ("HLA genetic system and organ transplantation: understanding the basics of immunogenicity to improve donor - receptor compatibility when assigning grafts to recipients") was co-supervised by Marie-Dominique Devignes and Pr. Jean-Luc Taupin. It has been defended on July 11, 2024.

9.4 Regional initiatives

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

Isaure CdB was in the organisation committee of the Winterschool "Algorithms in Structural Bioinformatics (**AlgoSB**) 2025 : Structure modeling and design towards RNA-based therapeutics" (w. Samuela Pasquali (Paris-Cite) and Yann Ponty (LIX)). She obtained for it the "Ecole Thematique CNRS" status and support.

General chair, scientific chair

Member of the organizing committees

10.1.2 Scientific events: selection

Chair of conference program committees

Member of the conference program committees Isaure for ISMB/ECCB 2025 (International Conference) Isaure for Jobim 2025 (National Conference)

Reviewer

10.1.3 Journal

Member of the editorial boards

Reviewer - reviewing activities

10.1.4 Invited talks

Isaure was invited for a talk at the joint "ELIXIR 3DBioinfo Annual General Meeting | ISCB 3DSig 2025", March 2025, Barcelona

10.1.5 Leadership within the scientific community

Isaure is co-responsible of the MASIM GT of the GDR BIMMM

10.1.6 Scientific expertise

10.1.7 Research administration

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

- Malika is an associate professor at the Université de Lorraine with a full service. She is co-responsible with Pascal Moyal of the IMSD track ("Ingénierie Mathématique pour la Science des Données") in the Applied Mathematics Master's degree at the Université de Lorraine. She is

also a member of the pedagogic team of the CMI BSE ("Cursus Master Ingénieur Biologie-Santé-Environnement").

10.2.2 Supervision

10.2.3 Juries

10.3 Popularization

10.3.1 Internal or external responsibilities

10.3.2 Articles and contents

10.3.3 Education

10.3.4 Interventions

11 Scientific production

11.1 Major publications

[0] E. Bresso, J.-P. Ferreira, N. Girerd, M. Kobayashi, G. Preud'homme, P. Rossignol, F. Zannad, M.-D. Devignes and M. Smaïl-Tabbone. 'Inductive database to support iterative data mining: Application to biomarker analysis on patient data in the Fight-HF project'. In: *Journal of Biomedical Informatics* 135 (Nov. 2022), p. 104212. DOI: 10.1016/j.jbi.2022.104212. URL: https://hal.univ-lorraine.fr/hal-03805671.

[0] M. Kobayashi, O. Huttin, M. Magnusson, J. P. Ferreira, E. Bozec, A.-C. Huby, G. Preud'Homme, K. Duarte, Z. Lamiral, K. Dalleau, E. Bresso, M. Smaïl-Tabbone, M.-D. Devignes, P. M. Nilsson, M. Leosdottir, J.-M. Boivin, F. Zannad, P. Rossignol and N. Girerd. 'Machine Learning-Derived Echocardiographic Phenotypes Predict Heart Failure Incidence in Asymptomatic Individuals'. In: *JACC: Cardiovascular Imaging* S1936-878X.21 (Sept. 2021), pp. 00556–8. DOI: 10.1016/j.jcmg.2021.07.004. URL: https://hal.univ-lorraine.fr/hal-03357064.

[0] A. Moniot, Y. Guermeur, S. J. de Vries and I. Chauvot de Beauchêne. 'ProtNAff: Protein-bound Nucleic Acid filters and fragment libraries'. In: *Bioinformatics* 38.162022-07-01 (2022), pp. 3911–3917. DOI: 10.1093/bioinformatics/btac430. URL: https://hal.science/hal-03765772.

[0] M. E. Ruiz Echartea, I. Chauvot de Beauchêne and D. Ritchie. 'EROS-DOCK: Protein-Protein Docking Using Exhaustive Branch-and-Bound Rotational Search'. In: *Bioinformatics* 35.23 (2019), pp. 5003–5010. DOI: 10.1093/bioinformatics/btz434. URL: https://hal.archives-ouvertes.fr/hal-02269812.

11.2 Publications of the year

International journals

[0] A. Gheeraert, V. Leroux, D. Mias-Lucquin, Y. Karami, L. Vuillon, I. Chauvot de BeauchÃªne, M.-D. Devignes, I. Rivalta, B. Maigret and L. Chaloin. 'Subtle Changes at the RBD/hACE2 Interface During SARS-CoV-2 Variant Evolution: A Molecular Dynamics Study'. In: *Biomolecules* 15.4 (7th Apr. 2025), p. 541. DOI: 10.3390/biom15040541. URL: https://hal.science/hal-05041555.

[0] E. Laine, S. Grudinin, R. Klypa and I. C. D. BeauchÃªne. 'Navigating proteinâ€"nucleic acid sequence-structure landscapes with deep learning'. In: *Current Opinion in Structural Biology* 95 (Dec. 2025), p. 103162. DOI: 10.1016/j.sbi.2025.103162. URL: https://hal.sorbonne-universite.fr/hal-05334762.

[0] K. Pats, I. Glukhov, S. Petrosian, M. Mamaeva, A. Sergushichev, M.-D. Devignes and F. MolnÃ¡r. ‘GEODES: Geometric Descriptors for the Assessment of Global and Local Flexibility of Proteins During Molecular Dynamics Simulation’. In: *IEEE Access* 13 (9th Apr. 2025), pp. 64259–64270. DOI: 10.1109/ACCESS.2025.3558781. URL: https://inria.hal.science/hal-05063673.

11.3 Cited publications

[0] A. Santos, A. R. o, A. B. Nielsen, L. Niu, M. Strauss, P. E. Geyer, F. Coscia, N. J. W. Albrechtsen, F. Mundt, L. J. Jensen and M. Mann. ‘A knowledge graph to interpret clinical proteomics data’. In: *Nat Biotechnol* 40.5 (May 2022), pp. 692–702.

[0] G. Uludoğan, E. Ozkirimli, K. O. Ulgen, N. Karalı and A. Özgür. ‘Exploiting pretrained biochemical language models for targeted drug design’. In: *Bioinformatics* 38.Supplement_2 (2022), pp. ii155–ii161.

[0] Z. Zhang, M. Xu, A. Jamasb, V. Chenthamarakshan, A. Lozano, P. Das and J. Tang. ‘Protein representation learning by geometric structure pretraining’. In: *arXiv preprint arXiv:2203.06125* (2022).