RESEARCH CENTRE

Inria Centre at Université de Lorraine

IN PARTNERSHIP WITH:

Université de Lorraine, CNRS

2023

ACTIVITY REPORT

Project-Team

CAPSID

Computational Algorithms for Protein Structures and Interactions

IN COLLABORATION WITH: Laboratoire lorrain de recherche en informatique et ses applications (LORIA)

DOMAIN

Digital Health, Biology and Earth

THEME

Computational Biology

Inria

Contents

Project-Team CAPSID

Creation of the Project-Team: 2015 July 01

Keywords

Computer sciences and digital sciences

A3.1.1. – Modeling, representation

A3.1.9. – Database

A3.1.10. – Heterogeneous data

A3.1.11. – Structured data

A3.2.1. – Knowledge bases

A3.2.2. – Knowledge extraction, cleaning

A3.2.4. – Semantic Web

A3.2.5. – Ontologies

A3.2.6. – Linked data

A3.3.2. – Data mining

A3.5.1. – Analysis of large graphs

A6.1.4. – Multiscale modeling

A6.2.7. – High performance computing

A6.3.3. – Data processing

A6.5.5. – Chemistry

A8.2. – Optimization

A9.1. – Knowledge

A9.2. – Machine learning

Other research topics and application domains

B1.1.1. – Structural biology

B1.1.2. – Molecular and cellular biology

B1.1.7. – Bioinformatics

B2.2.1. – Cardiovascular and respiratory diseases

B2.2.4. – Infectious diseases, Virology

B2.4.1. – Pharmaco kinetics and dynamics

1 Team members, visitors, external collaborators

Research Scientists

- Marie-Dominique Devignes [Team leader, CNRS, Researcher, HDR]
- Isaure Chauvot de Beauchêne [CNRS, Researcher]
- Yasaman Karami [INRIA, Researcher]
- Hamed Khakzad [INRIA, Advanced Research Position]
- Bernard Maigret [CNRS, Emeritus]

Faculty Members

- Sabeur Aridhi [UL, Associate Professor, from Jul 2023]
- Sabeur Aridhi [UL, Associate Professor Delegation, until Jun 2023]
- Malika Smaïl-Tabbone [UL, Associate Professor, HDR]

Post-Doctoral Fellow

- Dominique Mias-Lucquin [UL, Post-Doctoral Fellow, until Jun 2023]

PhD Students

- Diego Amaya Ramirez [UL, ATER, until Sep 2023]
- Kamrul Islam [UL, until Jan 2023]
- Mohammed Khatbane [UL, from Oct 2023]
- Omid Mokhtari [INRIA, from Oct 2023]
- Victor Pryakhin [UL, from Oct 2023]

Technical Staff

- Hrishikesh Dhondge [CNRS, Engineer, until Aug 2023]
- Anna Kravchenko [CNRS, Engineer]
- Athénaïs Vaginay [CNRS, Engineer, until Mar 2023]
- Taher Yacoub [CNRS, Engineer, from Nov 2023]

Interns and Apprentices

- Benjamin Gottis [CNRS, Intern, from Feb 2023 until Jul 2023]
- Jean-Baptiste Paquin [UL, Intern, from Mar 2023 until Apr 2023]
- Hugo Rimet [UNIV NANTES, Intern, from Mar 2023 until May 2023]

Administrative Assistants

- Antoinette Courrier [CNRS]
- Sophie Drouot [INRIA]

External Collaborators

- Taha Boukhobza [UL, HDR]
- Emmanuel Bresso [CHRU Nancy]
- Pablo Chacon [Université de Madrid]

2 Overall objectives

2.1 Computational Challenges in Structural Biology

NB: This section has been remodeled since the death of Dave Ritchie in 2019.

Many of the processes within living organisms can be studied and understood in terms of biochemical interactions between large macromolecules such as DNA, RNA, and proteins. To 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. Many biological processes are governed by complex systems of proteins and/or RNA which interact cooperatively to regulate the chemical composition within a cell or to carry out a wide range of biochemical processes such as photosynthesis, metabolism, and cell signalling, for example. It is becoming increasingly feasible to isolate and characterise some of the individual molecular components of such systems, but it still remains extremely difficult to achieve detailed models of how these complex systems actually work. Consequently, a new multidisciplinary approach called integrative structural biology has emerged which aims to bring together experimental data from a wide range of sources and resolution scales in order to meet this challenge [58, 72].

Understanding how biological systems work 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 which carry out very specific biological processes. As such, these molecular machines may be considered as complex systems because their properties are much greater than the sum of the properties of their component parts.

2.2 Two Research Axes

The overall objective of the CAPSID team is to develop algorithms and software to help study biological systems and phenomena from a structural point of view. In particular, the team aims to develop algorithms which can help to model the structures of large multi-component biomolecular machines and to develop tools and techniques to represent and mine knowledge of the 3D shapes of proteins, NA and their interactions. Thus, a unifying theme of the team is to tackle the recurring problem of representing and reasoning about large 3D macromolecular shapes. More specifically, our aim is to develop computational techniques to represent, analyse, and compare the shapes and interactions of biomolecules in order to better understand how their 3D structures relate to their biological function. In summary, the CAPSID team is organised according to two research axes whose complementarity constitutes an original contribution to the field of structural bioinformatics:

- Axis 1: Knowledge Discovery in Structural Databases,
- Axis 2: Integrative Multi-Component Assembly and Modeling.

In the first axis, our main objective is to design, implement and test new KDD ("Knowledge Discovery in Databases") approaches to exploit specifically the structural 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 and fast methods to model the 3D structure of multi-component systems and characterize their dynamic behaviour. The challenge here is to integrate molecular flexibility into 3D models, thanks to molecular dynamics simulation and/or combinatorial approaches.

Finally, the complementarity of the two axes will be expressed through a common objective oriented towards the proposal of 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

This section presents the current CAPSID research program. Several subjects initially present at the creation time (2015) or at last evaluation (2017) are no longer presented due to the death of Dave Ritchie.

3.1 Knowledge Discovery in Structural Databases

3.1.1 Context

In this axis, the CAPSID team develops methods related to knowledge discovery from databases (KDD, [36]). 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 [62]. 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) [56, 67, 26]. 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 information. There is no single definition of KGs as these graphs cover a large variety of domains and data representation contexts (for instance the GAFAM companies advertize various KG uses). 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 in KG is expanding rapidly [52]. Most biological KGs today are developed for drug repurposing tasks (e.g. HetioNet [41] or DRKG). Clinicians are also very interested in network science carried out on rich knowledge graphs as a mean to interpret biomarker studies. However, there is still a need for curated, reliable biological KGs and for efficient knowledge discovery methods in KGs.

3.1.2 Knowledge discovery from protein structural databases

Concerning protein structural databases, we aim to explore novel classification paradigms exploiting existing resources about protein folds and domains [21, 22, 56, 67]. In particular it will be interesting to use Kpax, our structural alignment tool [63], to define domain-domain similarity matrices. A non-trivial issue with clustering is how to handle similarity rather than mathematical distance matrices, and how to find the optimal number of clusters, especially when using a non-Euclidean similarity measure. We will adapt the algorithms and the calculation of quality indices to the Kpax similarity measure. More fundamentally, it will be necessary to integrate this classification step in the more general KDD process leading from data to knowledge.

For example, protein domain classification is relevant for studying domain-domain interactions (DDI). Our previous work on Knowledge-Based Docking (KBDOCK, [38, 40]) will be updated and extended using newly published DDIs. Methods for inferring new DDIs from existing protein-protein interactions (PPIs) will be developed. Efforts should be made for validating such inferred DDIs so that they can be used to enrich DDI classification and predict new PPIs.

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, huge numbers of protein sequences in public databases such as UniProt/TrEMBL lack any functional annotation, and the functional annotation of such sequences 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 approches 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 discovery algorithms in large biological knowledge graphs

KGs are particularly useful and appropriate in biology, to represent and integrate the complex contents of biological databases [60]. We intend to design algorithms for leveraging information embedded in biological KGs (also known as complex networks). In biology, KGs mostly represent PPIs, integrated with various properties attached to proteins, such as pathways, drug binding or relation with diseases. Setting up similarity measures for proteins in a knowledge graph is a difficult challenge. 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 cell or organisms. Ultimately, knowledge graphs can be used to model and simulate the functioning of such molecular machinery in the context of the living cell, under physiological or pathological conditions.

3.2 Integrative Multi-Component Assembly and Modelling

3.2.1 Context

This axis deals with 3D protein structure and interactions. In fact, the long-lasting problem of predicting a 3D structure from a protein sequence has been solved in 2021 by the AlphaFold2 (DeepMind) [46] or RosettaFold methods [24]. This success, revealed in the CASP14 (Critical Assessment of Stucture Prediction) challenge, was possible not only thanks to AI methods but also because the amount of experimental 3D structures has reached a sufficient size in the Protein Data Bank (PDB). For the same type of reasons, the rigid docking problem (in which the bodies to dock are rigid) seems to be on the way to being solved as well [28, 35]. However, research is still required to address the problem of docking disordered proteins or flexible nucleic acids that will fold as they bind to proteins. This is the direction taken by the team since the arrival of Isaure Chauvot de Beauchêne, the inventor of a fragment-based approach for RNA docking onto proteins.
Modeling protein - and even more RNA - flexibility accurately during docking is very computationally expensive. This is due to the very large number of internal degrees of freedom in each molecule, associated with twisting motions around covalent bonds. Therefore, it is highly impractical to use detailed force-field or geometric representations in a brute-force docking search. Instead, most docking algorithms use fast heuristic methods to perform an initial rigid-body search in order to locate a relatively small number of candidate binding orientations, and these are then refined using a more expensive interaction potential or force-field model, which might also include flexible refinement using molecular dynamics (MD), for example.

3.2.2 Coarse-Grained Models

Many approaches have been proposed in the literature to take into account protein (and more recently RNA/DNA) flexibility during docking. The most thorough methods rely on expensive atomistic simulations using MD. However, much of a MD trajectory is unlikely to be relevant to a docking encounter unless it is constrained to explore a putative protein-protein/NA interface. Consequently, MD is normally only used to refine a small number of candidate rigid body docking poses. A much faster - but more approximate - method is to use "coarse-grained" (CG) normal mode analysis (NMA) techniques to reduce the number of flexible degrees of freedom to just one or a handful of the most significant vibrational modes [34, 54, 57, 59]. In our experience, docking ensembles of NMA conformations do not give much

improvement over basic FFT-based soft docking [71], and it is very computationally expensive to use side-chain repacking to refine candidate soft docking poses [39].

In the last few years, CG force-field models have become increasingly popular in the MD community because they allow very large biomolecular systems to be simulated using conventional MD programs [23]. Typically, a CG force-field representation replaces the 5-15 atoms in each amino acid with 2-4 "pseudo-atoms" (each pseudo-atom represents few atoms of an amino-acid as a single bead). It then assigns each pseudo-atom a small number of parameters to represent its chemo-physical properties. By directly attacking the quadratic nature of pair-wise energy functions, coarse-graining can speed up MD simulations by up to three orders of magnitude. Nonetheless, such CG models can still produce useful models of very large multi-component assemblies [66]. Furthermore, this type of CG model effectively integrates many internal DOFs to build a smoother but still physically realistic energy surface [42]. We are currently developing a CG scoring function for RNA-protein docking by fragments assembly.

3.2.3 Assembling Multi-Component Complexes and Integrative Structure Modelling

We also want to develop related approaches for integrative structure modeling using cryo-electron microscopy (cryo-EM). Thanks to recent developments in cryo-EM instruments and technologies, it is now feasible to capture low resolution images of very large macromolecular machines. However, while such developments offer the intriguing prospect of being able to trap biological systems in unprecedented levels of detail, there will also come with an increasing need to analyse, annotate, and interpret the enormous volumes of data that will soon flow from the latest instruments. In particular, a new challenge that is emerging is how to fit previously solved high resolution protein structures into low resolution cryo-EM density maps. However, the problem here is that large molecular machines will have multiple sub-components, some of which will be unknown, and many of which will fit each part of the map almost equally well. Thus, the general problem of building high resolution 3D models from cryo-EM data is like building a complex 3D jigsaw puzzle in which several pieces may be unknown or missing, and none of which will fit perfectly. We wish to proceed firstly by putting more emphasis on the single-body terms in the scoring function [33], and secondly by using fast CG representations and knowledge-based distance restraints to prune large regions of the search space, as initiated with the EROS-DOCK software [6, 65].

3.2.4 Protein-Nucleic Acid Interactions

As well as playing an essential role in the translation of DNA into proteins, RNA molecules carry out many other essential biological functions in cells, often through their interactions with proteins. A critical challenge in modeling such interactions computationally is that the RNA is often highly flexible, especially in single-stranded (ssRNA) regions of its structure. These flexible regions are often very important because it is through their flexibility that the RNA can adjust its 3D conformation in order to bind to a protein surface. However, conventional protein-protein docking algorithms generally assume that the 3D structures to be docked are rigid, and so are not suitable for modeling protein-RNA interactions. There is therefore much interest in developing dedicated protein-RNA docking algorithms which can take RNA flexibility into account. This research topic has been initiated with the recruitement of Isaure Chauvot de Beauchêne in 2016 and is becoming a major activity in the team. A novel flexible docking algorithm is currently under development in the team. It first docks small fragments of ssRNA (typically three nucleotides at a time) onto a protein surface, and then combinatorially reassembles those fragments in order to recover a contiguous ssRNA structure on the protein surface [29, 30].

As the correctness of the initial docking of the fragments settles an upper limit to the correctness of the full model, we are now focusing on improving that step. A key component of our docking tool is the energy function of the protein-fragment interactions that is used both to drive the sampling (positioning of the fragments) by minimization, and to discriminate the correct final positions from decoys (i.e., false positives). We are developing a new approach to create knowledge-based parameters for coarse-grain energy functions from public structural data, in collaboration with Sjoerd de Vries (INSERM). Such approach will be applied first to ssRNA-protein complexes, then to other types of complexes such as protein-peptides.

Another key requirement for this approach is an exhaustive but non-redundant library of possible internal conformations of RNA fragments. Our library is built by clustering hundreds of thousands of

experimentally known RNA structures, based on an approximate geometric similarity criteria. We want to develop new algorithms for the clustering of 3D conformations based on internal coordinates and on epsilon-net theory, in order to optimise the representativity and computational cost of the library.

In the future, we will improve the combinatorial algorithm used for reassembling the docked fragments using both experimental constraints and knowledge-based constraints pertaining from the research carried out in Axis 1.

4 Application domains

4.1 Biomedical Knowledge Discovery

Participants: Marie-Dominique Devignes *(contact person)*, Malika Smaïl-Tabbone *(contact person)*, Sabeur Aridhi, Kamrul Islam, Athénaïs Vaginay.

Our main application for Axis 1 : "New Approaches for Knowledge Discovery in Structural Databases", 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 essentiel role in the onset of a disease and/or for personalised medicine.

As a first step we have been involved since 2015 in the ANR RHU “FIGHT-HF” (Fight Heart Failure) project, which is coordinated by the CIC-P (Centre d’Investigation Clinique Plurithématique) at the CHRU Nancy and INSERM U1116. In this project, the molecular mechanisms that underly heart failure (HF) are re-visited at the cellular and tissue levels in order to adapt treatments to patients’ needs in a more personalised way. The CAPSID team is in charge of a workpackage dedicated to network science. A platform has been constructed with the help of a company called Edgeleap (Utrecht, NL) in which biological molecular data and ontologies, available from public sources, are represented in a single integrated complex network also known as knowledge graph. We are developing querying and analysis facilities to help biologists and clinicians interpreting their cohort results in the light of existing interactions and knowledge. We are also currently analysing pre-clinical data produced at the INSERM unit on the comparison of aging process in obese versus lean rats. Using our expertise in receptor-ligand docking, we are investigating possible cross-talks between mineralocorticoid and other nuclear receptors.

Another application is carried out in the context of an interdisciplinary project funded by the Université de Lorraine, in collaboration with the CRAN laboratory. It concerns the study of the role of estrogen receptors in the development of gliobastoma tumors. The available data is high-dimensional but involves rather small numbers of samples. The challenge is to identify relevant sets of genes which are differentially expressed in various phenotyped groups (w.r.t. gender, age, tumor grade). The objectives are to infer pathways involving these genes and to propose candidate models of tumor development which will be experimentally tested thanks to an ex-vivo experimental system available at the CRAN.

Finally, 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 (BN), cellular automata. In her PhD thesis, Athenaïs Vaginay investigates ways to find a BN 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 [70]. Our approaches will be driven by knowledge and will keep close connection with experimental data.

4.2 Prokaryotic Type IV Secretion Systems

Participants: Isaure Chauvot de Beauchêne *(contact person)*, Marie-Dominique Devignes, Bernard Maigret, Dominique Mias-Lucquin.

Concerning Axis 2 : "Integrative Multi-Component Assembly and Modeling", our first application domain is related to prokaryotic Type IV secretion systems.

Prokaryotic type IV secretion systems constitute a fascinating example of a family of nanomachines capable of translocating DNA and protein molecules through the cell membrane from one cell to another [20]. The complete system involves at least 12 proteins. The structure of the core channel involving three of these proteins has recently been determined by cryo-EM experiments for Gram-negative bacteria [37, 64]. However, the detailed nature of the interactions between the other components and the core channel remains to be found. Therefore, these secretion systems represent a family of complex biological systems that call for integrated modeling approaches to fully understand their machinery.

In the framework of the Lorraine Université d'Excellence (LUE-FEDER) "CITRAM" project we are pursuing our collaboration with Nathalie Leblond of the Genome Dynamics and Microbial Adaptation (DynAMic) laboratory (UMR 1128, Université de Lorraine, INRAE) on the mechanism of horizontal transfer by integrative conjugative elements (ICEs) and integrative mobilisable elements (IMEs) in prokaryotic genomes. These elements use type IV secretion systems for transferring DNA horizontally from one cell to another. We have discovered more than 200 new ICEs/IMEs by systematic exploration of 72 Streptococcus genomes and characterised a new class of relaxases [68]. We have modeled the dimer of this relaxase protein by homology with a known structure. For this, we have created a new pipeline to model symmetrical dimers of multi-domains proteins. As one activity of the relaxase is to cut the DNA for its transfer, we are also currently studying the DNA-protein interactions that are involved in this very first step of horizontal transfer (see next section).

4.3 Protein - RNA Interactions

Participants: Isaure Chauvot de Beauchêne *(contact person)*, Antoine Moniot, Anna Kravchenko, Hrishikesh Dhondge, Marie-Dominique Devignes, Malika Smaïl-Tabbone.

The second application domain of Axis 2 concerns protein-nucleic acid interactions. We need to assess and optimise our new algorithms on concrete protein-nucleic acid complexes in close collaboration with external partners coming from the experimental field of structural biology. To facilitate such collaborations, we are creating automated and re-usable protein-nucleic acid docking pipelines.

This is the case for our PEPS collaboration "InterANRIL" with the IMoPA lab (CNRS-Université de Lorraine). We are currently working with biologists to apply our fragment-based docking approach [30] to model complexes of the long non-coding RNA (lncRNA) ANRIL with proteins and DNA [19].

In the framework of our LUE-FEDER CITRAM project (see above), we are adapting this approach and pipeline to single-strand DNA docking, in order to model the complex formed by a bacterial relaxase and its target DNA [68].

In the framework of our H2020 ITN project RNAct, we tackle a defined group of RNA-binding proteins containing RNA-Recognition Motifs (RRM) [53, 31]. We study existing and predicted complexes between various types of RRMs and various RNA sequences in order to infer rules of their sequence-structure-interaction relationship, and to help design new synthetic proteins with targeted RNA specificity. This work is made in tight collaboration with computer scientists and biophysicists of the consortium.

4.4 3D structural differences among HLA antigens

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

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 [27]. 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 are currently developing methods to compare local structure variations between HLA proteins, taking into account the dynamics of these proteins.

4.5 Investigating the dynamical behaviour of protein complexes with MD simulation

Participants: Yasaman Karami *(contact person)*, Malika Smaïl-Tabbone, Isaure Chauvot de Beauchêne, Viktor Pryakhin.

Allosteric pathways in protein-nucleic acid complexes. 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. Yasaman Karami has 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 [48] and she is constantly working at improving it.
In 2023, Benjamin Gottis was hired for an M2 internship (6 months, February-July) to improve COMMA for the analysis of large protein complexes. His internship resulted in the improvement of threshold definitions in COMMA and accelerating the calculations. Moreover, 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. At the same time, Yasaman Karami has 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.

Dynamics of Type IV Pilus. Type IV pili (T4P) are dynamic filaments at the surface of many bacteria that can rapidly extend and retract and withstand strong forces. T4P are responsible for various cellular processes due to their highly dynamic behavior and are crucial for bacterial virulence in many human pathogens, including Enterohemorrhagic *Escherichia coli* (EHEC), *Pseudomonas aeruginosa* (PaK), *Neisseria meningitidis* (Nm), *Neisseria gonorrhoeae* (Ng), and *Myxococcus xanthus* (Mx). They are assembled by complex protein machinery localized in the bacterial envelope, formed by the repeat of major pilin subunits organised in a helical manner. The structure of these T4Ps have been determined by combining NMR with a cryo-EM density map of the pilus, resulting in an atomistic model of the T4Ps filament. Despite the recent progress in cryo-EM and integrative structural biology, the high flexibility of this family of fibers often limits the resolution of the structure. Modeling is therefore a necessary step in structure determination. To unravel the structural basis of different T4Ps dynamics, we performed extensive and

large-scale classical and steered MD simulations of five different T4Ps, EHEC, PaK, Nm, Ng and Mx. The results highlighted key regions for the function of those filaments. Such simulations require an important number of computational resources, which were provided thanks to the Swiss National Supercomputing Centre (CSCS) and GENCI computing center. This project is one of the main activities of Yasaman Karami, in collaboration with Michael Nilges (Institut Pasteur, Paris) and Edward Egelman (University of Virginia, USA).

4.6 Learning from protein surfaces to accurately predict their functional sites.

Participants: Hamed Khakzad *(contact person)*, Marie-Dominique Devignes, Omid Mokhtari.

Today, deep learning from large datasets of sequences and structures (AlphaFold, RosettaFold) enables structure prediction from sequence with remarkable accuracy. However, the performance of these methods on predicting protein-protein interactions (PPIs) is still arguable as in the case of AlphaFold multimer [35, 51]. More importantly, none of these methods can address the interactions between other types of macromolecules such as DNA, RNA, and ligands. Taking a different stance, this project is focused on a different aspect of macromolecular structure; the surface. Extracting and learning surface features, which are critical for function including the interaction with other macromolecules, has been less explored than other aspects of macromolecular structures. The idea here is that the surface geometrical and chemical patterns could be used to understand and predict important aspects of macromolecular functions. Although these features are difficult to determine visually, they can be learned by a deep neural network trained over large-scale datasets. This project will attempt to create a broader methodological framework in comparison to the existing approaches that can be used to capture aspects of surface conformational diversity, and interactions with different types of macromolecules including small molecules (ligands), protein-RNA, protein-DNA, and protein- protein interactions. This project is part of Hamed Khakzad's ANR-CPJ funding, and a PhD student (Omid Mokhtari) was recruited in October 2023 to work on this topic.

5 Social and environmental responsibility

5.1 Environmental Footprint of Research Activities

In structural bioinformatics and deep learning approaches, the computational costs are usually very high. The CAPSID team pays attention to use shared equipment (Platform MBI-DS4H, Grid5K) for running HPC (High Performance Computing) jobs as efficiently as possible. In particular, we use the "best effort" mode for distributed jobs.
When travelling to conferences, members of the CAPSID team prefer the train solution as often as possible.

6 New software, platforms, open data

6.1 New software

6.1.1 CROMAST

Name: Cross-Mapper of domain Structural instances

Keywords: Protein domain, 3D structure, Classification, Databases, Workflow

Scientific Description: see the scientific paper 10.1093/bioadv/vbad081

Functional Description: CroMaSt (Cross Mapper of domain Structural instances) is an automated iterative workflow to clarify the assignment of protein domains to a given domain type of interest, based on their 3D structure and by cross-mapping of domain structural instances between domain

databases. CroMaSt (for Cross-Mapper of domain Structural instances) will classify all structural instances of a given domain type into 4 different categories (Core, True, Domain-like, and Failed)

Release Contributions: Operational version for publication

URL: https://workflowhub.eu/workflows/390

Publication: hal-04210856

Contact: Hrishikesh Dhondge

6.1.2 HIPPO

Name: HIstogram-based Pseudo-POtential

Keywords: Structural Biology, Computational biology

Functional Description: Pipeline to create scoring potentials from docking decoys

Contact: Isaure Chauvot De Beauchêne

6.1.3 RRMpip

Name: RRM_modeling_pipeline

Keywords: Structural Biology, Computational biology

Functional Description: pipeline to create 3D models of RRM protein domains

Contact: Isaure Chauvot De Beauchêne

6.1.4 ssRNATTRACT

Keywords: Computational biology, Structural Biology

Functional Description: 3D modelling of protein-ssRNA complexes by fragment assembly

URL: https://github.com/isaureCdB/ssRNATTRACT

Contact: Isaure Chauvot De Beauchêne

6.2 New platforms

Participants: Marie-Dominique Devignes *(scientific responsible)*, Malika Smaïl-Tabbone *(contact person)*, Sabeur Aridhi, Bernard Maigret, Antoine Moniot, Diego Amaya Ramirez.

The CAPSID team is at the origin of the creation of the LORIA MBI-DS4H research platform that provides a shared environment to the CAPSID and ORPAILLEUR teams for running distributed intensive computation. This platform is also the place for optimizing codes that can be run later on Grid 5K or on the Jean-Zay supercalculator. Moreover, the platform offers opportunities for newcomers in the team to get trained to good practices in development and in sharing code and data.

The technical support of the platform is ensured by the LORIA SISR (Service d'Ingénierie en Soutien de la Recherche) via a private project on gitlab.

6.3 Open data

Benchmark simulated datasets for clustering of mixed data

Description We deposited in the Inria open data repository seven sets of synthetic datasets to be used for benchmarking clustering algorithms for mixed (continuous and categorical) data. The synthetic datasets correspond to 9 simulation designs described in the README file and refer to a 2021 publication in Scientific Reports [61]

Contact Marie-Dominique Devignes

url doi:10.57745/6IFQYQ

Benchmark of proteins bound to ssDNA

Description We explored the Protein Data Bank (PDB) to collect protein-ssDNA structures and create a multi-conformational docking benchmark including both bound and unbound protein structures. Due to ssDNA high flexibility when not bound, no ssDNA unbound structure is included in the benchmark. This benchmark is, to our knowledge, the first one made to peruse available structures of ssDNA-protein interactions to such an extent, aiming to improve computational docking tools dedicated to this kind of molecular interactions. Related publication: [55].

Contact Isaure Chauvot de Beauchêne

url doi:10.57745/3W8CCV

MD simulations and ML dataset of HLA-EpiCheck epitope predictor tool

Description This dataset contains all the data used to implement the B-cell epitope predictor tool called HLA-EpiCheck (preprint available on bioRXiv).

Contact Diego Amaya-Ramirez and Marie-Dominique Devignes

url doi:10.57745/GXZHH8

Experiences with a training DSW knowledge model for early-stage researchers

Description Data management is fast becoming an essential part of scientific practice, driven by open science and FAIR (findable, accessible, interoperable, and reusable) data sharing requirements. Whilst data management plans (DMPs) are clear to data management experts and data stewards, understandings of their purpose and creation are often obscure to the producers of the data, which in academic environments are often PhD students. Within the RNAct EU Horizon 2020 ITN project, we engaged the 10 RNAct early-stage researchers (ESRs) in a training project aimed at formulating a DMP. To do so, we used the Data Stewardship Wizard (DSW) framework and modified the existing Life Sciences Knowledge Model into a simplified version aimed at training young scientists, with computational or experimental backgrounds, in core data management principles. We collected feedback from the ESRs during this exercise. Here, we introduce our new life-sciences training DMP template for young scientists. We report and discuss our experiences as principal investigators (PIs) and ESRs during this project and address the typical difficulties that are encountered in developing and understanding a DMP. We found that the DS-wizard can also be an appropriate tool for DMP training, to get terminology and concepts across to researchers [7].

Contact Malika Smaïl-Tabbone and Marie-Dominique Devignes

url doi:10.12688/openreseurope.15609

7 New results

7.1 Axis 1 : Knowledge Discovery in Structural Databases

Participants: Marie-Dominique Devignes, Malika Smaïl-Tabbone, Sabeur Aridhi, Kamrul Islam, Athénaïs Vaginay.

7.1.1 Knowledge graph mining with embedding-based methods

In the context of Md Kamrul Islam's PhD project, we addressed the problem of link prediction in large knowledge graphs (KGs) using KG embedding methods. These methods aim to learn low dimensional vector representations of entities and relations in a KG. Such representations (in a latent space) facilitate link prediction tasks along with other downstream tasks. In this context, it is important to achieve both an efficient KG embedding and explainable predictions. During learning of efficient embeddings, sampling negative triples is an important step as KGs only come with observed positive triples. We previoulsy proposed an efficient simple negative sampling (SNS) method based on the assumption that the entities which are closer to the corrupted entity in the embedding space are able to provide high-quality negative triples [44]. As for explainability, we also report in the same paper a new rule mining method which exploits the learned embeddings [44].
We then extended this work to propose an integrated drug repurposing, evaluation and explanation pipeline for COVID-19 disease [11]. The workflow starts with collecting and cleaning a COVID-19 centric drug repurposing knowledge graph (DRKG). Then, high-quality and compact ensemble embeddings are learned using three embedding methods. The embeddings are then used to train a deep neural network based model to predict the probability of unobserved triples connecting drugs with 27 COVID-19 proteins (as drug targets). The top-100 predictions are evaluated based on (i) cross-matching with in-trial drugs for COVID-19 and (ii) molecular evaluation based on compound and protein structures. Beside these evaluations, we learn high quality rules from DRKG and provide possible explanations of predictions. This study demonstrates how complementary embedding methods can be used to generate high-quality ensemble embeddings of a KG and how to use embeddings for the drug repurposing task. To the best of our knowledge, it is the first attempt to combine virtual screening methods with KG embedding methods in predicting and evaluating repurposable drugs for COVID-19. Besides the retrieval of many in-trial drugs, both methods show a converging result that the Fosinopril compound could be a new potential nsp13 inhibitor. Experimental validation of Fosinopril compound to treat COVID-19 is a potential perspective of this study. The molecular evaluation results and explanations of the predictions make us confident about the drawn conclusions. Md Kamrul Islam has successfully defended his PhD on December 16, 2022 [43].

7.1.2 Biological network modeling

Boolean Networks (BNs) refer to a simple formalism used to study complex biological systems when the prediction of exact reaction times is not of interest. BNs play a key role in understanding the dynamics of the studied systems and in predicting their disruption in case of complex human diseases. The BioModels database is a well-known repository of peer-reviewed models represented in the Systems Biology Markup Language (SBML). Most of these models are quantitative, but in some use cases, qualitative models—such as BNs—are better suited. In the context of Athénais Vaginay's PhD project, we proposed SBML2BN, a pipeline dedicated to the automatic transformation of quantitative SBML models to Boolean networks [69]. Our approach takes advantage of several SBML elements (reactions, rules, events) as well as a numerical simulation of the concentration of the species over time to constrain both the structure and the dynamics of the Boolean networks to synthesise. Finding all the BNs complying with given structure and dynamics was formalised as an optimisation problem formulated in the answer-set programming framework.
We ran SBML2BN on more than 200 quantitative SBML models, and we could construct Boolean networks which are compatible with the structure and the dynamics of the SBML models [69]. The most recent work relies on abstract simulation of a chemical reaction network (CRN) to avoid the tricky binarization task as we propose to simulate chemical reaction networks with the deterministic semantics abstractly, without any precise knowledge on the initial concentrations. For this, the concentrations of species are abstracted to Booleans stating whether the species is present or absent, and the derivatives of the concentrations are abstracted to signs saying whether the concentration is increasing, decreasing, or unchanged. We use abstract interpretation over the structure of signs for mapping the ODEs of a reaction network to a Boolean network with nondeterministic updates. The abstract state transition graph of such Boolean networks can be computed by finite domain constraint programming over the finite structure of signs. Constraints on the abstraction of the initial concentrations can be added naturally, leading to an abstract simulation algorithm that produces only the part of the abstract state transition graph that is

reachable from the abstraction of the initial state. We proved the soundness of our abstract simulation algorithm, and showed its applicability to reaction networks in the SBML format from the BioModels database [14]. Athénaïs Vaginay has successfully defended her PhD on July, 7 2023 [16].

7.1.3 Machine Learning and Scalable Graph-based Approaches

In the context of a collaboration with researchers from both the University Badji Mokhtar-Annaba (Algeria), the University of Quebec At Montreal (UQAM) and the University of Lille, Sabeur Aridhi proposed a genetic algorithm for random forest [12]. The proposed algorithm has three main objectives: (1) strengthening the classification accuracy of individual decision trees as well as that of the forest, (2) making use of diversity measures among the decision trees to improve the generalization of the constructed model, and (3) minimizing the number of trees in the forest and finding an optimal subset of the random forest.

7.2 Axis 2 : Integrative Multi-Component Assembly and Modeling

Participants: Isaure Chauvot de Beauchêne, Marie-Dominique Devignes, Malika Smaïl-Tabbone, Bernard Maigret, Yasaman Karami, Hamed Khakzad, Dominique Mias-Lucquin, Antoine Moniot, Anna Kravchenko, Hrishikesh Dhondge, Diego Amaya Ramirez.

7.2.1 Modeling and design of RNA-RRM complexes

Our recent H2020 ITN project RNAct (2018-2022) aimed at designing new RNA-binding proteins based on the evolutionary conserved protein domain[1], called RNA Recognition Motif (RRM). In this context and in the continuity of this project, we develop approaches to create 3D models of ssRNA-RRM complexes, building up on our expertise in RNA-protein modeling by fragments assembly [30]. We use the ATTRACT docking software to sample and evaluate possible (low binding energy) positions of RNA fragments on the protein surface, then assemble the geometrically compatible poses with compatible sequences, to build a continuous model for the full RNA sequence.

Empirical inference of an RNA-protein energy function. In the frame of Anna Kravchenko's thesis and in collaboration with Sjoerd de Vries (SISR team, LORIA), we have implemented a new method to create energy parameters for ssRNA-protein interactions in coarse-grained representation. In the ATTRACT procedure, each amino-acid of the protein and each nucleotide of the RNA is represented by 2 to 7 pseudo-atoms ("beads"). For each model of an RNA-protein interaction, the energy is computed as the sum of the bead-bead energies, and the model with the lowest energy is considered as the most probable. Each bead-bead energy depends solely on the 2 bead types (among 17 RNA types and 32 protein types) and the inter-bead distance. The models are then ranked by energy, in order to select the top-ranked poses that are supposed to be enriched in correct positions. We previously used parameters created in 2010 and tailored for double-stranded RNA, but their performance is poor on ssRNA. One main goal of Anna Kravchenko's PhD was to optimize those parameters for a better discrimination of the correct positions for a given ssRNA fragment on an RRM protein domain. In 2022, we have set up a novel "histogram-based" approach, called HIPPO (HIstogram-based Pseudo-POtential). From a training set of solved RRM-ssRNA complexes, we derived a set of non-redundant RRM-fragment complexes (here a fragment is a tri-nucleotide of RNA). From each RRM-fragment complex and for each bead-bead type, we build a log-odds histogram of the occurrences of bead-bead distances (discretized into bins) observed in correct/incorrect poses retrieved from the fragment docking run corresponding to this RRM-fragment complex. The set of histograms obtained for a given RRM-fragment complex is called a "scoring potential" and later used to score the poses retrieved from fragment docking runs performed with RRMs and fragments derived from other solved RRM-ssRNA complexes constituting the test set.
In 2023, we have greatly improved this approach by creating a consensus scoring procedure from 4 sets of

[1] An evolutionary conserved protein domain is a protein domain for which the aligned sequences do not differ much across all possible living species.

histogramms/scoring potentials that were identified to cover at best the diversity of RRM-ssRNA binding modes while avoiding over-fitting. We tested this consensus procedure on a benchmark of RRM-fragment complexes, extracted from 51 experimental structures of RRM-ssRNA complexes. HIPPO achieved a successful enrichment in correct poses (60% of correct poses in the 20% top-ranked poses) for 53% of our RRM-fragment complexes with HIPPO, versus 26% with the old parameters. Most importantly, HIPPO achieved a high enrichment for at least 1 fragment of the full ssRNA for 75% of the RRM-ssRNA complexes, versus 54% with the old parameters. This led us to adapt the assembling of geometrically compatible poses, by retaining less top-ranked poses for one of the fragments (iteratively supposed to be the highly enriched one), thus decreasing the complexity of the whole fragment-based docking procedure.
As a great surprise, we also found out that those parameters trained on ssRNA-RRM structure only, also perform better than the old parameters for ssRNA-protein complexes that do not contain the RRM protein domain. This encourage us to train a general scoring potential for all ssRNA-protein complexes in the near future. We also found that, in mot cases, a large majority of the top-ranked correct poses are selected by only one of the 4 sets of histograms. This work was presented at the ISMB-ECCB 2023 international conference in Lyon in July 2023 [18] and a full paper is actually under revision.
For 2024, a way to improve HIPPO's performance would be to predict which of the 4 sets will perform the best on a given protein-fragment case. This would avoid retaining the false positives returned by the other three. This may be achieved with supervised machine learning techniques based on the sequence of the fragment and the sequence or/and structure of the protein, and/or on the docking poses. Such a pre-trained classifier not only would drastically improve the performance of the scoring but could also give biological insight into the most prevalent protein-ssRNA binding modes.

Search of self-avoiding paths for fragment-based docking. The assembly of compatible poses in our fragment-based modeling of RNA-protein complexes can be formalized as the search of paths of low global energy in a graph representing the pairwise compatibility of poses (fragment positions on the protein), each pose having a binding energy. The algorithm we used so far to sample paths does not take into account possible steric collisions - or clashes - between non-consecutive fragments. Chains containing such collisions have to be filtered a posteriori. In collaboration with Yann Ponty (LIX) and Fabrice Leclerc (IBC), we tried to avoid such collisions more efficiently, by integrating this global constraint during path sampling. We adapted the historical Noga Alon's color-coding algorithm[2] for the search for self-avoiding paths of k vertices in a directed graph, in order to circumvent the NP-completeness of this problem. It consists in randomly assigning one of k colors to each vertex, applying an $O(n2^k)$ algorithm for finding well-colored paths of size k (1 vertex per color) with minimal global energy, and reiterating the method E times, where E is the expectation of the waiting time for a good coloring for a given path ($E = (k^k)/k!$) which can be approximated by e^k. In this way, we obtain the best chain of poses with a high probability in $O(n(2e)^k)$) time, compared to a brute force complexity in n^k.
To further reduce the size of the problem, and improve computation time, we have grouped the poses into cliques of clashing poses, which can therefore be systematically assigned the same color. The goal of cliques is to maximize the number of graph edges in a clique, i.e. edges which will never be chosen during path sampling, thus reducing the total number of possible paths. This work was recently accepted for the international conference RECOMB 2024, to be held April 29 - May2, 2024 in Cambridge MA (USA).

Cross-mapping of protein domain structural instances. To address the issue of inconsistencies of protein domain classification in different domain databases, we developed a tool named Cross-Mapper for domain Structural instances (CroMaSt). This tool is a workflow that uses database querying and structural alignment to assign a confidence level to each 3D structure of a given domain in relation to its possible membership of a given domain type. This workflow was primarily developed for RRMs, but can easily be adapted to any structural domain. The workflow has been formalized using the Common Workflow Language (CWL) [3]. Its rationale is based on the cross-mapping of PDB (Protein Data Bank) entries retrieved as RRM, in the Pfam and CATH domain classifications. Entries that are mapped in both classifications are considered as the "core" RRMs. Those that are specific to only one classification are further analyzed using structural alignment, in order to determine whether they are RRM-like or false

[2] Noga Alon, Raphael Yuster and Uri Zwick. Color-coding. Journal of the ACM, 42:844-856. 1995
[3] https:www.commonwl.org

RRMs [8]. The workflow can be used to create datasets for specific domains with a good confidence level, which are useful for characterizing domain structural diversity or for further analyses such as machine learning, evolutionary studies or synthetic biology. This work was an important contribution of the doctoral thesis of Hrishikesh Dhondge who successfully defended it on 11 July 2023 [15].

7.2.2 3D Modeling of proteins and protein complexes

Modeling protein-DNA complexes to fight antibiotic resistance. In the context of the FEDER-CITRAM project (collaboration with the DynAMic and LPCT labs, Université de Lorraine), we study a new type of relaxase (RelSt3) as a key protein for horizontal DNA transfer, one of the processes responsible for the spread of antibiotic resistance in bacteria. The relaxase cuts parts of the bacterial DNA genome, then brings the cut DNA piece to the "coupling protein" inserted in the bacterial membrane, which transfers this piece to the recipient cell. In 2022, we contributed to the exploration of DNA processing features of RelSt3 [50]. In 2023, we extended this work to the 3D modeling of different types of relaxases and coupling proteins in different bacteria strains, thanks to AlphaFold and including the evaluation of the most reliable parts of each model. This paves the way for usage of those models in studying the relaxase-DNA and relaxase-coupling protein interactions, in order to target them by inhibitors.
In parallel, a new round of virtual screening inhibitors has been carried out targeting the ssDNA binding pocket of RelSt3. The inhibitory activity of the selected comppounds is currently being tested in the DynAMic lab.

Structural basis of donor-specific antibody response in graft rejection. In the context of the Inria-Inserm PhD project of Diego Amaya Ramirez and in collaboration with the Immunology and Histocompatibility Laboratory at the APHP Saint-Louis Hospital in Paris, we study the structural differences between donor and recipient HLA proteins to understand and possibly predict the immune response triggered in the recipient, which can result in rejection of the transplanted organ. A dataset of 207 HLA 3D structures has been created, both from PDB and AlphaFold predictions. Molecular dynamics runs (10 ns) have been analyzed at the level of single residues or surface patches centered on surface residues. Surface patches are described by a set of 18 descriptors, including both static and dynamic properties, such as hydrophobicity, electrostatic charges, relative solvent-accessible surface area and side-chain flexibility. A machine-learning predictor for B-cell epitopes on HLA proteins (HLE-EpiCheck) has been trained using an Extra Trees ensemble learning method to discriminate between positive (confirmed epitopes) and negative (non epitope) surface patches. HLA-EpiCheck prediction performance outperformed the performance achieved by DiscoTope-3.0, a state-of-the-art B-cell epitope predictor in the task of predicting HLA epitopes. HLA-EpiCheck was also used to assess the epitope status of a subset of non-confirmed eplets. The predictions were compared to experimental data and a notable consistency was found. These results suggest that HLA-EpiCheck could be used to better define HLA matching between donor and recipient in order to reduce de novo DSA formation and graft rejection. This work was presented as a poster at the ISMB-ECCB 2023 International conference in July 2023 [17]. A full paper has been submitted for publication in December 2023 (preprint available on bioRXiv).

7.2.3 Investigating the dynamical behaviour of protein complexes with MD simulation

Analysis of COVID-19 spike binding to ACE2. In collaboration with Laurent Chaloin at the IRIM (Institut de Recherche en Infectiologie de Montpellier, CNRS-Université de Montpellier), we investigated the interaction between the RBD domain of SARS-CoV-2 Spike protein and human ACE2 along very long molecular dynamics (MD) simulations (1.5 µs, performed on Jean Zay super computer at GENCI, IDRIS-CNRS in Orsay). A panel of methods was used to analyse the MD trajectories. The goal of this project is to compare the Omicron variants with the Delta variant of SARS-CoV-2. The manuscript is under revision.

Nucleosomes and post-translational modifications. Nucleosomes are made of proteins and DNA and Yasaman Karami has started a new collaboration with Emmanuelle Bignon, a recently recruited CNRS researcher at LPCT (Laboratoire de Physique et Chimie Théorique). Together, they study the effect of

post-translational modifications of proteins on the nucleosome[4], using MD simulations and COMMA analysis [48]. The results of their work are now under consideration at the Computational and Structural Biotechnology journal with a favorable revision (preprint available on bioRXiv).

M1-IgG interaction network and dynamics. In collaboration with Pontus Nordenfelt (Lund University), Yasaman Karami and Hamed Khakzad investigated the role of hinge flexibility for the opsonic function of antibodies[5]. An opsonic IgG1 monoclonal antibody, Ab25 [25], targeting the M protein of the bacteria *Streptococcus pyogenes (S. pyogenes)*, was engineered into the IgG2-4 subclasses. Despite reduced binding, the IgG3 version demonstrated enhanced opsonic function. MD simulations showed that IgG3 Fc region exhibits extensive mobility in 3D space relative to the antigen due to its extended hinge region. The MD simulations also showed altered Fab-antigen interactions, in line with IgG3 diminished affinity. We explored the impact of hinge-engineering by generating a panel of IgG antibodies, IgGh, containing the CH1-3 domains of IgG1 and different segments of IgG3 hinge. Hinge-engineering enhanced opsonic function, with the most potent hinge having 47 amino acids. IgGh47 far exceeded the parent IgG1 and even the IgG3 version. The IgGh47 was protective against *S. pyogenes* in a systemic infection mouse model, contrary to parent IgG3 and IgG1. The *in vitro* phenotype of IgGh47 was generalizable to clinical isolates with different M protein types. Finally, we generated IgGh47 versions of anti-SARS-CoV-2 mAbs, which exhibited strongly enhanced in vitro opsonic function compared to the original IgG1. The improved function of the IgGh47 subclass in two distant biological systems provides new insights into antibody function and how to enhance it for opsonic function. The paper received favorable revision from Nature Communications (preprint available in BioRXiv [45]).

Unraveling the complexity of glycosaminoglycan deficiencies caused by B3GALT6 loss-of-function mutations: a multifaceted approach Deleterious variants in beta-1,3-galactosyltransferase (B3GALT6) compromise the early steps of glycosaminoglycan (GAG) synthesis causing a spondylodysplastic subvariety of Ehlers-Danlos syndromes (spEDS). Unfortunately, the mechanisms by which pathogenic mutations impair B3GALT6 function and translate into a severe, often lethal, connective tissue disorder are poorly defined. In collaboration with Sylvie Fournel-Gigleux at IMoPA, Nancy, Yasaman Karami and Hamed Khakzad dissected the patho-mechanisms of spEDS by a multi-tiered approach from exploring the protein structure alterations by MD simulations to atomic force microscopy-based extracellular matrix (ECM) investigations. We found that pathogenic variants produced a unique structural B3GalT6 alteration resulting in loss of function that was partially rescued by glucuronosyltransferase 1 (GlcAT-1), the next enzyme in the pathway. Transcriptomics revealed that the resulting GAG defects predominantly dysregulated collagen maturation. In line, by leveraging a B3galT6-invalidated ECM model recapitulating the condition, we suggest that the absence of collagen XII glycosylation contributes to altered tissue structure and biomechanics. Our findings uncover a novel link between GAG and collagen defects and shed light on the pathobiology of spEDS. The manuscript has been submitted recently to Journal of Clinical Investigations. In continuity with this work, the team is now involved in the ANR GlycoLink project (2023-2027) coordinated by Sylvie Fournel-Gigleux.

7.2.4 Machine learning methods for proteomics, interactomics and protein design

Deep Learning and *de novo* MS-based peptide sequencing. In the recent years, the application of deep learning represented a breakthrough in the mass spectrometry (MS) field by improving the assignment of the correct sequence of amino acids from observable MS spectra without prior knowledge, also known as *de novo* MS-based peptide sequencing. However, like other modern neural networks, models do not generalize well enough as they perform poorly on highly varying N- and C-termini peptide test sets. To mitigate this generalization problem, Hamed Khakzad and colleagues from Lund University (Sweden) conducted a systematic investigation to unravel the requirements for building generalizable models and boosting the performance on the MS-based *de novo* peptide sequencing task. Several experiments confirmed that the peptide diversity of the training set directly impacts the resulting generalizability of the model. Data showed that the best models were the multienzyme models (MEMs), *i.e.*, models trained

[4] A nucleosome is a region of DNA that is wrapped around a core of proteins, inside the cell nucleus.

[5] Opsonic function: capacity of antibodies to coat pathogens present in the blood and thereby promote their recognition by the phagocytosis system responsible of their elimination.

from a compendium of highly diverse peptides, such as the one generated by digesting samples from a wide variety of species with a group of proteases. The applicability of these MEMs was later established by fully *de novo* sequencing eight of the ten polypeptide chains of five commercial antibodies and extracting over 10,000 proving peptides[10].

Machine learning and protein design. In collaboration with the team of Pierre Tuffery (Université Paris-Cité), Yasaman Karami developed a method to design cyclic peptides and propose candidate linkers. Large-scale data-mining of available protein structures was previously shown as useful for the precise identification of protein loop conformations, even from remote structural classes [47]. This approach was transposed to linkers, allowing head-to-tail peptide cyclization. This project has strong potential for cyclic petpide-based drug design and was published in 2023 in Journal of Chemical Informatics and Modelling [49].
From a more general point of view, Hamed Khakzad has written in collaboration with Bruno Correia's team at Ecole Polytechnique Fédérale de Lausanne and Michael Bronstein from University of Oxford, a review on the recent developments and technologies in deep learning methods with examples of their performance to generate novel functional proteins [13].

7.2.5 Miscellaneous results on structural studies of host-pathogen interactions

Mutational analysis of vinculin and cell adhesion Vinculin is a cytoskeletal linker strengthening cell adhesion. The *Shigella* IpaA invasion effector binds to vinculin to promote vinculin supra-activation associated with head-domain mediated oligomerization. In collaboration with Pr Guy Tran van Nhieu (I2BC, Université Paris-Saclay), Hamed Khakzad has investigated the impact of mutations within the vinculin D1D2 subdomains, which are predicted to interact with IpaA VBS3. These mutations influence the rate of D1D2 trimer formation, with distinct effects on monomer disappearance, consistent with structural modeling of a "closed" and "open" D1D2 conformer induced by IpaA. Notably, mutations targeting the closed D1D2 conformer significantly reduced Shigella invasion of host cells, in contrast to a mutation targeting a putative D2 coiled-coil motif or a cysteine clamp affecting later stages of vinculin head-domain oligomerization. All mutations affected the focal adhesions (FAs) formation. Our findings suggest that IpaA-induced vinculin supra-activation primarily reinforces matrix adhesion in infected cells, rather than promoting bacterial invasion. Consistently, shear stress studies pointed to a key role for IpaA-induced vinculin supra-activation in accelerating and strengthening cell matrix adhesion. Additionally, our results support the involvement of vinculin supra-activation in FAs maturation and cell adhesion. The manuscript received favorable revisions from Life Science Alliance (preprint available in bioRXiv [32]). This work initiated the ANR grant application TRIVIAL under evaluation for the AAPG2024.

Functional proteomics to reveal streptolysin O as a novel plasminogen-binding streptococcal protein The bacteria *S. pyogenes* is a highly adaptive human specific pathogen weaponized with multifunctional bacterial proteins, known to exploit the plasminogen (PLG)-plasmin (PLM) system to promote its dissemination and survival in the human host. In collaboration with Pr Johan Malmström (Lund University, Sweden), Hamed Khakzad applied a series of functional proteomics methods to chart the protein-protein interaction landscape centred on streptolysin O (SLO), revealing that this critical cytolytic toxin binds specifically to PLG. Binding of SLO to PLG potentially alters the shape of PLG from a compact to a partially relaxed conformation, thereby accelerating more PLM production via tissue-type plasminogen activator. Our results reveal a conserved moonlighting pathomechanistic role for SLO carried in all *S. pyogenes* isolates, extending beyond its established cytolytic activity. A deeper investigation is warranted in order to better understand the profound relationships between bacterial adaptation and host haemostasis during different stages of infection. A manuscript is under preparation to describe this work.

8 Partnerships and cooperations

8.1 International initiatives

8.1.1 Participation in other International Programs

NewDAFI

Participants: Malika Smaïl-Tabbone, Sabeur Aridhi, Marie-Dominique Devignes, Bernard Maigret.

Funding: COFECUB - CAPES 2023

Title: New drugs against invasive fungal infections: from hits to optimised leads through machine learning

Partner Institution(s): The Catholic University of Brasília (UCB), Brazil

Date/Duration: 2023-2026

Additionnal info/keywords: The main objective of this project is to convert the previous knowledge acquired, from comparative genomics, selection of new therapeutic targets and identification of a new class of antifungals, into a product that can reach the preclinical phase, and that effectively contributes to the fight against fungal emerging diseases and nosocomial infections. Our ambition is also to perpetuate the international scientific exchange between Brazil and France for training qualified human resources, especially in interdisciplinary areas.

8.2 International research visitors

8.2.1 Visits of international scientists

Other international visits to the team

Pr Edward Egelman

Status Professor

Institution of origin: University of Virginia

Country: USA

Dates: 11-13 July 2023

Context of the visit: Ongoing collaboration with Yasaman Karami and 1st Nancy Computational Structural Biology (NCSB) day

Mobility program/type of mobility: Invitation for a lecture and scientific discussions

Pr Maria Sueli Felipe

Status Professor

Institution of origin: Catholic University of Brasilia

Country: Brazil

Dates: 22-27 October 2023

Context of the visit: Exchange of senior scientists in the frame of a COFECUB-CAPES program (2023-2026) coordinated by Malika Smaïl-Tabbone.

Mobility program/type of mobility: Lecture and scientific discussions

8.2.2 Visits to international teams

Research stays abroad

Marie-Dominique Devignes

Visited institution: Catholic University of Brasilia and State University of Maringa

Country: Brazil

Dates: 20-30 November 2023

Context of the visit: Exchange of senior scientists in the frame of a COFECUB-CAPES program (2023-2026) coordinated by Malika Smaïl-Tabbone.

Mobility program/type of mobility: Lectures and Master Classes

Sabeur Aridhi

Visited institution: University of Trento, Pr Alberto Montresor.

Country: Italy

Dates: 21-26 May 2023

Context of the visit: Inria delegation and short fellowship support (LORIA).

Mobility program/type of mobility: scientific discussions.

8.3 European initiatives

8.3.1 Other european programs/initiatives

Through her implication in the French institute of Bioinformatics (joint coordination of the Open Science and Interoperability taskforce), Marie-Dominique Devignes is a member of the European ELIXIR Interoperability platform.

8.4 National initiatives

ANR EPIHLA

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

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

Duration: October 2022-October 2025

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

Inria contact: Marie-Dominique Devignes

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") is not finished yet and still co-supervised by Marie-Dominique Devignes and Pr. Jean-Luc Taupin.

ANR GlycoLink

Participants: Hamed Khakzad *(contact person)*, Yasaman Karami, Isaure Chauvot de Beauchêne.

Title: Exploring Glycosyltransferases complexes involved in glycosaminoglycan-Linker assembly

Duration: October 2023-October 2026

Coordinator: Pr. Sylvie Fournel-Gigleux (UMR 7365 CNRS-Université de Lorraine)

Inria contact: Hamed Khakzad

Partner Institutions:
- UMR 7365 CNRS-Université de Lorraine (Nancy)
- IBS - UMR 5075 (Grenoble)
- INRIA Center of Université de Lorraine (Nancy)
- ICOA, UMR 7311 CNRS U-Orléans (Orléans)

Summary: The supramolecular arrangement of glycosyltransferases (GT) and accessory enzymes (kinase, phosphatase, sulfotransferases) governs the repertoire of carbohydrates and their multiple biological functions. The assembly of glycosaminoglycans (GAG), a major class of linear glycopolymers, is initiated by a unique tetrasaccharide linker (GlcAb1-3Galb1-3Galb1-4Xyl) synthesized by the coordinated action of five GT. It serves as a unique primer for the elongation of GAG chains. A recent paradigm puts forward a multimolecular complex called “GAGosome" formed of GT and auxiliary enzymes as a major mechanism to guarantee fidelity and efficiency of GAG synthesis. Indeed, association of the heparan-sulfate (HS) polymerases EXT1/EXT2 and of chondroitin-sulfate (CS) synthases (CSS) sustain GAG elongation. Recent evidence also supports complex formation between GAG-linker enzymes. GlycoLink will explore the formation and organization of multimolecular complexes between GT and partners involved in the assembly of the GAG-linker region, and their functional impact in rare genetic diseases.

8.5 Regional initiatives

LUE-FEDER CITRAM (2017-2023) The CITRAM project (Conception d’Inhibiteurs de la Transmission de Résistances AntI-Microbiennes), co-funded by Lorraine Université d’Excellence (LUE) and FEDER was extended until June 2023.
Partners other than CAPSID are:

- DynAMic lab (Genome dynamics and microbial adaptation, INRAE-Université de Lorraine UMR 1128), Team of Nathalie Leblond, coordinator ;
- LPCT (Laboratoire de Physique et Chimie Théoriques, CNRS-Université de Lorraine UMR 7019), Team of Chris Chipot.

MolAI4Cryo (2023-2025) The molAI4Cryo project (Modeling and Artificial Intelligence applied to Cryo-EM 3D structures to fight COVID) has obtained a grant from the Région Grand-Est to equip the IMoPA laboratory with a cutting-edge equipement in cryo-electromicroscopy. The CAPSID team will be involved in proposing algorithms to use cryo-EM results as constraints for modeling protein-RNA complexes.
Partners other than CAPSID are:

- IMoPA (Ingénierie Moléculaire et Physiopathologie Articulaire, CNRS-Université de Lorraine UMR 7365), Team of Xavier Manival, coordinator ;
- DynAMic lab (Genome dynamics and microbial adaptation, INRAE-Université de Lorraine UMR 1128), Team of Nathalie Leblond and Nicolas Soler ;
- LPCT (Laboratoire de Physique et Chimie Théoriques, CNRS-Université de Lorraine UMR 7019), Team of Chris Chipot and François Dehez.

9 Dissemination

Participants: All Team Members.

9.1 Promoting scientific activities

9.1.1 Scientific events: organisation

Yasaman Karami organized in Nancy (12 July 2023) a seminar called NCSB (Nancy Computational Structural Biology) with the help from both Inria and Loria. The event was held successfully with about 50 participants at the Université de Lorraine. We had the pleasure of having Professor Edward Egelman as the keynote speaker. In total we had 7 oral communications by both regional (IMoPA, DynaAMic, LPCT, Inria-Loria) and international (Germany, USA) scientists.

Member of the organizing committees Isaure Chauvot de Beauchêne was member of the organizing committee of AlgoSB Winter school 2023: "From structure resolution to dynamical modeling in cryo-electron microscopy".

Member of the conference program committees Malika Smaïl-Tabbone was member of the program committee of DATA ANALYTICS 2023 international conference and EGC 2023 francophone conference. Sabeur Aridhi was member of the program committee of IEEE Big Data 2023 international conference and of the ICML (International Conference on Machine Learning) Workshop on Computational Biology (WCB) 2023. Isaure Chauvot de Beauchêne was member of the program committee of ISMB-ECCB 2023 international conference (Lyon, 23-27 July 2023).

Member of the editorial boards Malika Smaïl-Tabbone is member of the Scientific Reports Editorial Board (from December 2023).
Marie-Dominique Devignes is guest editor of Bioinformatic Advances.

Reviewer - reviewing activities Marie-Dominique Devignes reviewed articles for Computational Structural Biotechnology Journal.
Yasaman Karami reviewed articles for Nucleic Acid Research (NAR) and Computational Structural Biotechnology journals.

9.1.2 Invited talks

Marie-Dominique Devignes gave an invited talk on "Open Science at the French Institute of Bioinformatics" during the Love Data Week organized at the Université de Lorraine, 13-16 March 2023.
Yasaman Karami gave an invited talk on “Conformational dynamics of Type-IV Pilus” during the Statistical Physics and Low Dimensional systems (SPLDS) Conference organized by the LPCT lab at Pont-a-Mousson, 24-26 May, 2023.

9.1.3 Leadership within the scientific community

Marie-Dominique Devignes is co-leader of the Open Science and Interoperability Task at the French Institute of Bioinformatics (with Alban Gaignard - Institut du Thorax, Nantes, Platform BiRD, and Frederic de Lamotte - INRAE, Department of Biology et Plant Improvement, Montpellier). In this frame, she is member of the ELIXIR European Interoperability platform and she co-authored a paper on automatic FAIR assessment of web resource data [9].
Hamed Khakzad is an active member of the Rosetta Commons scientific community.

9.1.4 Scientific expertise

Sabeur Aridhi reviewed an ANR project in 2023.

9.1.5 Research administration

Malika Smaïl-Tabbone is member of the Scientific Council of the Université de Lorraine. As such, she carries out various scientific expertises within the broad framework of the University of Lorraine.

9.2 Teaching - Supervision - Juries

9.2.1 Teaching

- Malika Smaïl-Tabbone 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").
- Sabeur Aridhi is an assistant professor at the Université de Lorraine with a full service. He is responsible for the major in IAMD ("Ingénierie et Applications des Masses de Données") at TELECOM Nancy.
- Marie-Dominique Devignes teaches every year 10 to 16h in the CMI BSE.
- Diego Amaya Ramirez held an ATER position from January to August 2023.
- Yasaman Karami gave 27 hours on Machine learning with Python as part of the Data Scientist continuing education course at the Institute for Digital Management and Cognition (IDMC), Nancy.
- Hamed Khakzad gave 16 hours on deep learning methods in the frame of the Advanced AI course (3A) at TELECOM Nancy.

9.2.2 Supervision

In 2023, there have been in total 7 PhD students supervised by CAPSID members. Three of them successfully defended their thesis. Three of them have been recruited and started their PhD in October 2023. Hamed Khakzad is also co-supervising with Rebekka Wild a PhD thesis that started in October 2023 at the University of Grenoble.

9.2.3 Juries

Malika Smaïl-Tabbone was reviewer of two PhD theses.
Marie-Dominique Devignes was reviewer of one PhD thesis and examiner of two PhD theses.
(Not counting the participation of team members in the juries of CAPSID PhD students)

9.2.4 Internal or external Inria responsibilities

Yasaman Karami is a member of the Inria Comité de Développement Technologique (CDT). The main task of the committee is to evaluate the application and recruitement of research engineers.

10 Scientific production

10.1 Major publications

[1] S. Z. Alborzi, A. Ahmed Nacer, H. Najjar, D. W. Ritchie and M. D. Devignes. 'PPIDomainMiner: Inferring domain-domain interactions from multiple sources of proteinprotein interactions'. In: *PLoS Computational Biology* 17.8 (2021), e1008844. DOI: 10.1371/journal.pcbi.1008844. URL: https://hal.archives-ouvertes.fr/hal-03435140.

[2] 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.

[3] M. K. Islam, D. Amaya-Ramirez, B. Maigret, M.-D. Devignes, S. Aridhi and M. Smaïl-Tabbone. 'Molecular-evaluated and explainable drug repurposing for COVID-19 using ensemble knowledge graph embedding'. In: *Scientific Reports* 13.1 (Mar. 2023), p. 3643. DOI: 10.1038/s41598-023-30095-z. URL: https://inria.hal.science/hal-04017432.

[4] 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.

[5] D. W. Ritchie and S. Grudinin. 'Spherical polar Fourier assembly of protein complexes with arbitrary point group symmetry'. In: *Journal of Applied Crystallography* 49.1 (Feb. 2016), pp. 158–167. DOI: 10.1107/S1600576715022931. URL: https://hal.inria.fr/hal-01261402.

[6] 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.

10.2 Publications of the year

International journals

[7] M.-D. Devignes, M. Smaïl-Tabbone, H. Dhondge, R. Dolcemascolo, J. Gavaldá-García, R. A. Higuera-Rodriguez, A. Kravchenko, J. Roca Martínez, N. Messini, A. Pérez-Ràfols, G. Pérez Ropero, L. Sperotto, I. Chauvot de Beauchêne and W. Vranken. 'Experiences with a training DSW knowledge model for early-stage researchers'. In: *Open Research Europe* 3 (2023), p. 97. DOI: 10.12688/openreseurope.15609.1. URL: https://hal.science/hal-04234402.

[8] H. Dhondge, I. Chauvot de Beauchêne and M.-D. Devignes. 'CroMaSt: a workflow for assessing protein domain classification by cross-mapping of structural instances between domain databases and structural alignment'. In: *Bioinformatics Advances* 3.1 (1st Jan. 2023). DOI: 10.1093/bioadv/vbad081. URL: https://inria.hal.science/hal-04210856.

[9] A. Gaignard, T. Rosnet, F. de Lamotte, V. Lefort and M.-D. Devignes. 'FAIR-Checker: supporting digital resource findability and reuse with Knowledge Graphs and Semantic Web standards'. In: *Journal of Biomedical Semantics* 14.1 (July 2023), p. 7. DOI: 10.1186/s13326-023-00289-5. URL: https://hal.science/hal-04148181.

[10] C. Gueto-Tettay, D. Tang, L. Happonen, M. Heusel, H. Khakzad, J. Malmström and L. Malmström. 'Multienzyme deep learning models improve peptide de novo sequencing by mass spectrometry proteomics'. In: *PLoS Computational Biology* 19.1 (20th Jan. 2023), e1010457. DOI: 10.1371/journal.pcbi.1010457. URL: https://hal.science/hal-04403345.

[11] M. K. Islam, D. Amaya-Ramirez, B. Maigret, M.-D. Devignes, S. Aridhi and M. Smaïl-Tabbone. 'Molecular-evaluated and explainable drug repurposing for COVID-19 using ensemble knowledge graph embedding'. In: *Scientific Reports* 13.1 (Mar. 2023), p. 3643. DOI: 10.1038/s41598-023-30095-z. URL: https://inria.hal.science/hal-04017432.

[12] N. E. I. Karabadji, A. Amara Korba, A. Assi, H. Seridi, S. Aridhi and W. Dhifli. 'Accuracy and diversity-aware multi-objective approach for random forest construction'. In: *Expert Systems with Applications* 225.1 (Apr. 2023), p. 120138. DOI: 10.1016/j.eswa.2023.120138. URL: https://inria.hal.science/hal-04079595.

[13] H. Khakzad, I. Igashov, A. Schneuing, C. Goverde, M. Bronstein and B. Correia. 'A new age in protein design empowered by deep learning'. In: *Cell Systems* 14.11 (Nov. 2023), pp. 925–939. DOI: 10.1016/j.cels.2023.10.006. URL: https://hal.science/hal-04288627.

International peer-reviewed conferences

[14] J. Niehren, C. Lhoussaine and A. Vaginay. 'Core SBML and its Formal Semantics'. In: CMSB 2023 - 21th International Conference on Formal Methods in Systems Biology. Luxembourg, Luxembourg, 13th Sept. 2023. URL: https://inria.hal.science/hal-04125922.

Doctoral dissertations and habilitation theses

[15] H. Dhondge. 'Structural characterization of RNA binding to RNA recognition motif (RRM) domains using data integration, 3D modeling and molecular dynamic simulation'. Université de Lorraine, 11th July 2023. URL: https://theses.hal.science/tel-04219324.

[16] A. Vaginay. 'Synthesis of Boolean networks from the structure and dynamics of reaction networks'. Université de Lorraine, 7th July 2023. URL: https://hal.univ-lorraine.fr/tel-04257373.

Other scientific publications

[17] D. Amaya-Ramirez, R. Lhotte, C. Usureau, M. Devriese, M. Smaïl-Tabbone, J.-L. Taupin and D. Marie-Dominique. 'HLA-EpiCheck : A B-cell epitope prediction tool on HLA antigens using molecular dynamics simulation data'. In: ISMB-ECCB 2023 - Intelligent System in Molecular Biology and European Conference on Computational Biology merged event. Lyon, France, 23rd July 2023. URL: https://inria.hal.science/hal-04405086.

[18] A. Kravchenko, S. Jacob de Vries, M. Smaïl-Tabbone and I. Chauvot de Beauchene. 'HIPPO: HIstogram-based Pseudo-POtential for scoring ssRNA-protein fragment-based docking poses'. In: The 31st Annual Intelligent Systems For Molecular Biology (ISMB) and the 22nd Annual European Conference on Computational Biology (ECCB). Lyon, France, 23rd July 2023. URL: https://hal.science/hal-04168414.

10.3 Cited publications

[19] C. Alfeghaly, I. Behm-Ansmant and S. Maenner. 'Study of Genome-Wide Occupancy of Long Non-Coding RNAs Using Chromatin Isolation by RNA Purification (ChIRP)'. In: *Methods Mol Biol* 2300 (2021), pp. 107–117.

[20] C. E. Alvarez-Martinez and P. J. Christie. 'Biological diversity of prokaryotic type IV secretion systems'. In: *Microbiology and Molecular Biology Reviews* 73 (2011), pp. 775–808.

[21] A. Andreeva, D. Howorth, C. Chothia, E. Kulesha and A. G. Murzin. 'SCOP2 prototype: a new approach to protein structure mining'. In: *Nucleic Acids Res* 42.Database issue (Jan. 2014), pp. D310–314. DOI: 10.1093/nar/gkt1242.

[22] A. Andreeva, E. Kulesha, J. Gough and A. G. Murzin. 'The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures'. In: *Nucleic Acids Res* 48.D1 (Jan. 2020), pp. D376–D382. DOI: 10.1093/nar/gkz1064.

[23] M. Baaden and S. R. Marrink. 'Coarse-grained modelling of protein-protein interactions'. In: *Current Opinion in Structural Biology* 23 (2013), pp. 878–886.

[24] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read and D. Baker. 'Accurate prediction of protein structures and interactions using a three-track neural network'. In: *Science* 373.6557 (Aug. 2021), pp. 871–876. DOI: 10.1126/science.abj8754.

[25] W. Bahnan, L. Happonen, H. Khakzad, V. Kumra Ahnlide, T. de Neergaard, S. Wrighton, O. André, E. Bratanis, D. Tang, T. Hellmark, L. Björck, O. Shannon, L. Malmström, J. Malmström and P. Nordenfelt. 'A human monoclonal antibody bivalently binding two different epitopes in streptococcal M protein mediates immune function'. In: *EMBO Molecular Medicine* 15.2 (2022), e16208. DOI: https://doi.org/10.15252/emmm.202216208. eprint: https://www.embopress.org/doi/pdf/10.15252/emmm.202216208. URL: https://www.embopress.org/doi/abs/10.15252/emmm.202216208.

[26] M. Blum, H. Y. Chang, S. Chuguransky, T. Grego, S. Kandasaamy, A. Mitchell, G. Nuka, T. Paysan-Lafosse, M. Qureshi, S. Raj, L. Richardson, G. A. Salazar, L. Williams, P. Bork, A. Bridge, J. Gough, D. H. Haft, I. Letunic, A. Marchler-Bauer, H. Mi, D. A. Natale, M. Necci, C. A. Orengo, A. P. Pandurangan, C. Rivoire, C. J. A. Sigrist, I. Sillitoe, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, A. Bateman and R. D. Finn. 'The InterPro protein families and domains database: 20 years on'. In: *Nucleic Acids Res* 49.D1 (Jan. 2021), pp. D344–D354. DOI: 10.1093/nar/gkaa977.

[27] A. Bonneau and C. Monchaud. 'La transplantation d'organes en France'. In: *Actualités Pharmaceutiques* 60.605 (2021), pp. 18–20. DOI: https://doi.org/10.1016/j.actpha.2021.02.005. URL: https://www.sciencedirect.com/science/article/pii/S0515370021000562.

[28] D. F. Burke, P. Bryant, I. Barrio-Hernandez, D. Memon, G. Pozzati, A. Shenoy, W. Zhu, A. S. Dunham, P. Albanese, A. Keller, R. A. Scheltema, J. E. Bruce, A. Leitner, P. Kundrotas, P. Beltrao and A. Elofsson. 'Towards a structurally resolved human protein interaction network'. In: *Nat Struct Mol Biol* 30.2 (Feb. 2023), pp. 216–225.

[29] I. J. Chauvot De Beauchene, S. J. De Vries and M. J. Zacharias. *Fragment-based modeling of protein-bound ssRNA*. ECCB 2016: The 15th European Conference on Computational Biology. Poster. Sept. 2016. URL: https://hal.archives-ouvertes.fr/hal-01573352.

[30] I. Chauvot de Beauchêne, S. J. De Vries and M. Zacharias. 'Fragment-based modelling of single stranded RNA bound to RNA recognition motif containing proteins'. In: *Nucleic Acids Research* (June 2016). DOI: 10.1093/nar/gkw328. URL: https://hal.archives-ouvertes.fr/hal-01505862.

[31] A. Clery and F. Allain. 'RNA binding proteins'. In: ed. by L. Zdravko. Landes Bioscience and Springer Science+Business Media, 2011. Chap. From Structure to Function of RNA Binding domains.

[32] B. Cocom-Chan, H. Khakzad, M. Konate, D. I. Aguilar, C. Bello, C. Valencia-Gallardo, Y. Zarrouk, J. Fattaccioli, A. Mauviel, D. Javelaud and G. T. V. Nhieu. 'IpaA reveals distinct modes of vinculin activation during Shigella invasion and cell-matrix adhesion'. In: *bioRxiv* (2023). DOI: 10.1101/2023.03.23.533139. eprint: https://www.biorxiv.org/content/early/2023/10/09/2023.03.23.533139.full.pdf. URL: https://www.biorxiv.org/content/early/2023/10/09/2023.03.23.533139.

[33] S. J. De Vries, I. Chauvot de Beauchêne, C. E. M. Schindler and M. Zacharias. 'Cryo-EM Data Are Superior to Contact and Interface Information in Integrative Modeling'. In: *Biophysical Journal* (Feb. 2016). DOI: 10.1016/j.bpj.2015.12.038. URL: https://hal.archives-ouvertes.fr/hal-01505863.

[34] S. E. Dobbins, V. I. Lesk and M. J. E. Sternberg. 'Insights into protein flexibility: The relationship between normal modes and conformational change upon protein–protein docking'. In: *Proceedings of National Academiy of Sciences* 105.30 (2008), pp. 10390–10395.

[35] R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper and D. Hassabis. 'Protein complex prediction with AlphaFold-Multimer'. In: *bioRxiv* (2021). DOI: 10.1101/2021.10.04.463034. eprint: https://www.biorxiv.org/content/early/2021/10/04/2021.10.04.463034.full.pdf. URL: https://www.biorxiv.org/content/early/2021/10/04/2021.10.04.463034.

[36] W. J. Frawley, G. Piatetsky-Shapiro and C. J. Matheus. 'Knowledge Discovery in Databases: An Overview'. In: *AI Magazine* 13 (1992), pp. 57–70.

[37] R. Fronzes, E. Schäfer, L. Wang, H. R. Saibil, E. V. Orlova and G. Waksman. 'Structure of a type IV secretion system core complex'. In: *Science* 323 (2011), pp. 266–268.

[38] A. Ghoorah, M.-D. Devignes, M. Smaïl-Tabbone and D. Ritchie. 'KBDOCK 2013: A spatial classification of 3D protein domain family interactions'. In: *Nucleic Acids Research* 42.D1 (Jan. 2014), pp. 389–395. DOI: 10.1093/nar/gkt1199. URL: https://hal.inria.fr/hal-00920612.

[39] A. Ghoorah, M.-D. Devignes, M. Smaïl-Tabbone and D. Ritchie. 'Protein Docking Using Case-Based Reasoning'. In: *Proteins - Structure, Function and Bioinformatics* 81.12 (Oct. 2013), pp. 2150–2158. DOI: 10.1002/prot.24433. URL: https://hal.inria.fr/hal-00880341.

[40] A. Ghoorah, M.-D. Devignes, M. Smaïl-Tabbone and D. Ritchie. 'Spatial clustering of protein binding sites for template based protein docking'. In: *Bioinformatics* 27.20 (Aug. 2011), pp. 2820–2827. DOI: 10.1093/bioinformatics/btr493. URL: https://hal.inria.fr/inria-00617921.

[41] D. S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S. L. Chen, D. Hadley, A. Green, P. Khankhanian and S. E. Baranzini. 'Systematic integration of biomedical knowledge prioritizes drugs for repurposing'. In: *Elife* 6 (Sept. 2017).

[42] H. I. Ingólfsson, C. A. Lopez, J. J. Uusitalo, D. H. de Jong, S. M. Gopal, X. Periole and S. R. Marrink. 'The power of coarse graining in biomolecular simulations'. In: *WIRES Comput. Mol. Sci.* 4 (2013), pp. 225–248. URL: http://dx.doi.org/10.1002/wcms.1169.

[43] K. Islam. 'Explainable link prediction in large complex graphs - application to drug repurposing'. Theses. Université de Lorraine, Dec. 2022. URL: https://hal.univ-lorraine.fr/tel-04027361.

[44] K. Islam, S. Aridhi and M. Smaïl-Tabbone. 'Negative Sampling and Rule Mining for Explainable Link Prediction in Knowledge Graphs'. In: *Knowledge-Based Systems* 250 (Aug. 2022), p. 109083. DOI: 10.1016/j.knosys.2022.109083. URL: https://hal.science/hal-03684205.

[45] A. Izadi, Y. Karami, E. Bratanis, S. Wrighton, H. Khakzad, M. Nyblom, B. Olofsson, L. Happonen, D. Tang, M. Nilges, J. Malmström, W. Bahnan, O. Shannon, L. Malmström and P. Nordenfelt. 'The increased hinge flexibility of an IgG1-IgG3 hybrid monoclonal enhances Fc-mediated protection against group A streptococci'. In: *bioRxiv* (2023). DOI: 10.1101/2023.10.14.562368. eprint: https://www.biorxiv.org/content/early/2023/10/18/2023.10.14.562368.full.pdf. URL: https://www.biorxiv.org/content/early/2023/10/18/2023.10.14.562368.

[46] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis. 'Highly accurate protein structure prediction with AlphaFold'. In: *Nature* 596.7873 (Aug. 2021), pp. 583–589. DOI: 10.1038/s41586-021-03819-2.

[47] Y. Karami, F. Guyon, S. De Vries and P. ry. 'DaReUS-Loop: accurate loop modeling using fragments from remote or unrelated proteins'. In: *Sci Rep* 8.1 (Sept. 2018), p. 13673.

[48] Y. Karami, E. Laine and A. Carbone. 'Dissecting protein architecture with communication blocks and communicating segment pairs'. In: *BMC Bioinformatics* 17 Suppl 2.Suppl 2 (Jan. 2016), p. 13.

[49] Y. Karami, S. Murail, J. Giribaldi, B. Lefranc, F. Defontaine, O. Lesouhaitier, J. Leprince, S. de Vries and P. Tuffery. 'Exploring a Structural Data Mining Approach to Design Linkers for Head-to-Tail Peptide Cyclization'. In: *J Chem Inf Model* 63.20 (Oct. 2023), pp. 6436–6450.

[50] H. Laroussi, Y. Aoudache, E. Robert, V. Libante, L. Thiriet, D. Mias-Lucquin, B. Douzi, Y. Roussel, I. ne, N. Soler and N. Leblond-Bourget. 'Exploration of DNA processing features unravels novel properties of ICE conjugation in Gram-positive bacteria'. In: *Nucleic Acids Res* 50.14 (Aug. 2022), pp. 8127–8142.

[51] J. Liu, Z. Guo, T. Wu, R. S. Roy, F. Quadir, C. Chen and J. Cheng. 'Enhancing alphafold-multimer-based protein complex structure prediction with MULTICOM in CASP15'. In: *Commun Biol* 6.1 (Nov. 2023), p. 1140.

[52] F. MacLean. 'Knowledge graphs and their applications in drug discovery'. In: *Expert Opin Drug Discov* 16.9 (Sept. 2021), pp. 1057–1069.

[53] C. Maris, C. Dominguez and F. H. Allain. 'The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression'. In: *FEBS J* 272.9 (May 2005), pp. 2118–2131.

[54] A. May and M. Zacharias. 'Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein-protein docking'. In: *Proteins* 70 (2008), pp. 794–809.

[55] D. Mias-Lucquin and I. Chauvot de Beauchene. 'Conformational variability in proteins bound to single-stranded DNA: A new benchmark for new docking perspectives'. In: *Proteins: Structure, Function, and Bioinformatics* 90.3 (2022), pp. 625–631. DOI: https://doi.org/10.1002/prot.26258. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.26258. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.26258.

[56] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn and A. Bateman. 'Pfam: The protein families database in 2021'. In: *Nucleic Acids Res* 49.D1 (Jan. 2021), pp. D412–D419. DOI: 10.1093/nar/gkaa913.

[57] I. H. Moal and P. A. Bates. 'SwarmDock and the Use of Normal Modes in Protein-Protein Docking'. In: *International Journal of Molecular Sciences* 11.10 (2010), pp. 3623–3648.

[58] C. Morris. 'Towards a structural biology work bench'. In: *Acta Crystallographica* PD69 (2013), pp. 681–682.

[59] D. Mustard and D. Ritchie. 'Docking essential dynamics eigenstructures'. In: *Proteins: Structure, Function, and Genetics* 60 (2005), pp. 269–274. DOI: 10.1002/prot.20569. URL: https://hal.inria.fr/inria-00434271.

[60] D. N. Nicholson and C. S. Greene. 'Constructing knowledge graphs and their biomedical applications'. In: *Comput Struct Biotechnol J* 18 (2020), pp. 1414–1428. DOI: 10.1016/j.csbj.2020.05.017.

[61] G. Preud'homme, K. Duarte, K. Dalleau, C. Lacomblez, E. Bresso, M. l-Tabbone, M. Couceiro, M. D. Devignes, M. Kobayashi, O. Huttin, J. P. Ferreira, F. Zannad, P. Rossignol and N. Girerd. 'Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark'. In: *Sci Rep* 11.1 (Feb. 2021), p. 4202.

[62] D. J. Rigden and X. M. Fernández. 'The 2021 Nucleic Acids Research database issue and the online molecular biology database collection'. In: *Nucleic Acids Research* 49.D1 (Dec. 2020), pp. D1–D9. DOI: 10.1093/nar/gkaa1216. eprint: https://academic.oup.com/nar/article-pdf/49/D1/D1/35364664/gkaa1216.pdf. URL: https://doi.org/10.1093/nar/gkaa1216.

[63] D. W. Ritchie. 'Calculating and scoring high quality multiple flexible protein structure alignments'. In: *Bioinformatics* 32.17 (May 2016), pp. 2650–2658. DOI: 10.1093/bioinformatics/btw300. eprint: https://academic.oup.com/bioinformatics/article-pdf/32/17/2650/49021295/bioinformatics_32_17_2650.pdf. URL: https://doi.org/10.1093/bioinformatics/btw300.

[64] A. Rivera-Calzada, R. Fronzes, C. G. Savva, V. Chandran, P. W. Lian, T. Laeremans, E. Pardon, J. Steyaert, H. Remaut, G. Waksman and E. V. Orlova. 'Structure of a bacterial type IV secretion core complex at subnanometre resolution'. In: *EMBO Journal* 32 (2013), pp. 1195–1204.

[65] M. E. Ruiz Echartea, D. Ritchie and I. Chauvot de Beauchêne. 'Using Restraints in EROS-Dock Improves Model Quality in Pairwise and Multicomponent Protein Docking'. In: *Proteins - Structure, Function and Bioinformatics* 88.8 (Aug. 2020), pp. 1121–1128. DOI: 10.1002/prot.25959. URL: https://hal.science/hal-02930827.

[66] M. G. Saunders and G. A. Voth. 'Coarse-graining of multiprotein assemblies'. In: *Current Opinion in Structural Biology* 22 (2012), pp. 144–150.

[67] I. Sillitoe, N. Bordin, N. Dawson, V. P. Waman, P. Ashford, H. M. Scholes, C. S. M. Pang, L. Woodridge, C. Rauer, N. Sen, M. Abbasian, S. Le Cornu, S. D. Lam, K. Berka, I. H. Varekova, R. Svobodova, J. Lees and C. A. Orengo. 'CATH: increased structural coverage of functional space'. In: *Nucleic Acids Res* 49.D1 (Jan. 2021), pp. D266–D273. DOI: 10.1093/nar/gkaa1079.

[68] N. Soler, E. Robert, I. Chauvot de Beauchêne, P. Monteiro, V. Libante, B. Maigret, J. Staub, D. W. Ritchie, G. Guédon, S. Payot, M.-D. Devignes and N. N. Leblond-Bourget. 'Characterization of a relaxase belonging to the MOBT family, a widespread family in Firmicutes mediating the transfer of ICEs'. In: *Mobile DNA* 10.1 (Dec. 2019), pp. 1–16. DOI: 10.1186/s13100-019-0160-9. URL: https://hal.inria.fr/hal-02138843.

[69] A. Vaginay, T. Boukhobza and M. Smaïl-Tabbone. 'From quantitative SBML models to Boolean networks'. In: *Applied Network Science* 7.1 (Dec. 2022), p. 73. DOI: 10.1007/s41109-022-00505-8. URL: https://hal.science/hal-03902922.

[70] A. Vaginay, M. Smail-Tabbone and T. Boukhobza. 'Towards an automatic conversion from SBML core to SBML qual'. In: *JOBIM 2019 - Journées Ouvertes Biologie, Informatique et Mathématiques*. Présentation Poster. Nantes, France, July 2019. URL: https://hal.archives-ouvertes.fr/hal-02407443.

[71] V. Venkatraman and D. Ritchie. 'Flexible protein docking refinement using pose-dependent normal mode analysis'. In: *Proteins* 80.9 (June 2012), pp. 2262–2274. DOI: 10.1002/prot.24115. URL: https://hal.inria.fr/hal-00756809.

[72] A. B. Ward, A. Sali and I. A. Wilson. 'Integrative Structural Biology'. In: *Biochemistry* 6122 (2013), pp. 913–915.