# Deep Learning for the prediction of the effects of mutations in proteins and protein-protein interactions

## Context and motivations

Protein-Protein Interactions (PPI) play a key role in biology and medicine in the interpretation of protein functions in cellular processes. The *ab initio* reconstruction of *highly precise* PPI networks of *individual* genomes from human populations (African, Caucasian…) and the *ab initio* reconstruction of the phenotypic mutational landscape of its proteins constitute a fundamental step. *Ab initio* networks are independent from experimental and literature knowledge, allowing to identify proteins whose interactions have not been previously observed. This unbiased approach is fundamental in precision medicine where the reconstruction of PPI network topology changes is of crucial importance. Indeed, each network carries information for an estimated average number of ~3000 nonsynonymous genetic variants per individual genome (1000 Genome Project Consortium, 2012). Only some of these variants disrupt protein functions and are likely disease-causing. Identifying those alterations that change many characteristics of a protein (thermodynamic stability, ligand binding and cellular localization) and, consequently, network topology is crucial to understand diseases. The ensemble of PPI networks for a population will be a mine of new information to enhance medical knowledge and generate new questions.

Recent deep learning approaches based on sequences (mCSM, iSEE, FoldX, MuPIPR) addressed the question of estimating binding affinity changes in PPIs, performed better than more classical approaches, but also showed that they are far from solving the problem, reaching a correlation of only 0.25 with experimental data measuring the changes of binding affinity over reference experimental datasets (SKEMPI v2). A contribution to these fundamental questions will have clearly an impact in Biology and Medicine. The development of original DL approaches will also have an impact in Computer Science by bringing new kinds of data complexity and computational challenges for their treatment to the world of AI.

## Scientific aims

In this thesis we wish to tackle two related problems concerning the effect of protein mutations. They have a different degree of complexity. **First,** we wish to develop an end-to-end deep learning framework to estimate the effects of mutations on PPIs based on sequences only. More precisely, we want to estimate protein binding affinity (BA) change and protein buried surface area (BSA) changes upon mutation for pairs of protein sequences. **Second**, we wish to construct a second end-to-end deep learning framework to estimate the functional and structural effects of mutations in a protein sequence, taken alone, without knowing its partner. This is an independent problem whose solution will likely benefit from what is learned in the first problem, which is explicitly considering physical contacts of protein pairs.

Our two questions are stated for protein sequences. Indeed, primary sequence is the fundamental information to describe a protein. Recently, AlphaFold2 clearly showed that protein sequences contain enough information to successfully reconstruct the 3-dimensional structure of a protein at a few Amstrong resolution and, possibly, of complexes made of several proteins. For more than 20 years now, many studies highlighted the importance of considering evolutionary signals to extract biological information from sequences also concerning the function of a protein, and not only its structure. Here, we go further in the study of proteins by exploring novel patterns in sequences leading to estimate the functional, and not only structural, effects caused by protein mutations.

To better characterize the sequence information and the mutation, **for both problems**, we will adopt several levels of encoding processes respectively on individual amino acids and on the sequence. For the first time, our architecture will incorporate, on the one hand, multiple probabilistic models, called *profile models*, representing a protein sequence, and on the other hand, a contextualized representation mechanism of amino acids in a protein profile to propagate the effects of a point mutation to surrounding amino acid representations, therefore amplifying the subtle change in a long protein sequence.

On top of that, for the first question of the thesis devoted to BA and BSA estimation, the architecture will leverage a Siamese convolutional neural network to encode a wild-type protein pair and its mutation pair.

We expect that multilayer perceptron regressors will be applied to the protein pair representations to successfully predict the quantifiable changes of PPI properties upon mutations.

The second question of the thesis is devoted to the reconstruction of the mutational landscape of a protein, when one or more mutations are applied. A very recent statistical study (Sivley et al 2018), covering the 77% of all human proteins, highlighted that the spatial distribution of genetic variation within protein structures is shaped by evolutionary constraint and provides insight into the functional importance of protein regions and the potential pathogenicity of protein alterations. In parallel, deep mutational scanning (Fowler and Fields 2014) conducted on a few proteins (Hopf et al 2017), opened the way to systematically estimating functional consequences of single-point mutations at every position in a protein and revealed that a relatively small number of positions in a protein are highly deleterious mutations: a substitution of the amino acid at any of these positions by almost any other amino acid produces a deleterious phenotype.

Based on the biological information highlighted in (Sivley et al 2018), on soft disorder which is highly correlated to the interface regions of a protein (see below), and on sequence evolution (see GEMME, IMPRINT and ProfileView below), we shall implement a deep learning CNN architecture predicting phenotypic mutational landscapes with one or more mutations.

## Justifications of the scientific method

Our **first architecture** will pre-train a multi-layer bidirectional long short-term memory (LSTM) language model of a collection of protein profiles, and alter the representation of each amino acid based on the surrounding context captured by the language model. The benefits of this representation learning mechanism are two-fold: (i) it automatically extracts more refined amino-acid-level features that are differentiated between different contexts of the proteins; (ii) it propagates the mutation effects to the representations of surrounding amino acids, therefore amplifying the subtle signal of each mutation in a long protein sequence. On top of the contextualized amino acid representations, a deep neural learning architecture will be designed to subsequently estimate the quantifiable PPI property changes between a wild-type pair and a mutant pair of proteins. The double siamese architecture, one dedicated to the wild-type and the other to the mutant protein pairs, will be inspired by our recent Deep Learning architecture IMPRINT devoted to protein partners identification (L. David thesis, SU; David, Richard, Carbone, 2022, manuscript in preparation) and PPI reconstruction, which proved to outperforms all existing methods, scale to thousands of proteins and allow to extract information on the binding sites at unprecedented fashion.

Given two proteins, IMPRINT decides whether they are partners or not. It is designed as a siamese architecture based on CNN and data augmentation. IMPRINT takes as input multiple profile models for the two sequences. The convolution modules of IMPRINT are designed as a set of 100-200 small filters either randomly defined or encoding small linear motifs (slm) coming from databases of slm in human proteins.

Our **second architecture** will be based on three main results from our group. 1. A current work which is devoted to the identification of soft disordered residues in sequences, that is residues that are either flexible, amorphous or intrinsically disordered, and that have been demonstrated to localize with protein binding regions (Seoane and Carbone 2021). We are using a LSMT architecture for this problem and encouraging results were obtained. These predictions will allow to focus on specific parts of the protein, expected to be especially important for structure and function. 2. We developed “Global Epistatic Model for predicting Mutational Effects” (GEMME) (Laine, Karami, Carbone 2019), an original and fast method that predicts mutational outcomes by explicitly modelling the evolutionary history of natural sequences. GEMME overall performs similarly or better than existing methods and outperforms generative deep learning approaches. 3. We showed that multiple profile models contain information to successfully classify protein families by function, solving a fundamental problem in biology (Vicedomini et al. 2022, in press). The second new deep learning architecture that will be developed in this thesis will be enriched with knowledge on general properties of biological sequences coming from soft disorder (cf.1), evolution (GEMME model, cf.2), multiple probabilistic models associated to protein domains annotating the sequence (ProfileView model, cf.3).

**Thesis advisor:** Alessandra Carbone, Professor SU

**Email**: alessandra.carbone@lip6.fr **Phone** : 01.44.27.73.45

**Research Unit**: Laboratoire de Biologie Computationnelle et Quantitative (LCQB), UMR7238

**Synthetic presentation of the team -** The "Analytical Genomics" team develops mathematical approaches, derived from statistics and combinatorics, and algorithmic methods (Combinatorial optimization, Machine Learning, and more recently Deep Learning) to study the basic principles of cellular functioning from genomic data. Our projects aim at understanding the basic principles of evolution and co-evolution of molecular structures in the cell. They concern: protein sequences and structures, protein phenotypes and genetic mutations, genome organization. The applications are multiple and play a role in directed mutagenesis, synthetic biology, organization of metagenomic data and genome annotation. AC will advise the PhD student on the development of deep learning architectures and the statistical analysis of genomic/clinical data.

**Selected publications linked to the project**

1. R.Vicedomini, J.P. Bouly, E. Laine, A. Falciatore, **A. Carbone**. "Multiple profile models extract features from protein sequence data and resolve functional diversity of very different protein families", *Molecular Biology and Evolution*, 2022. In press.
2. C. Dequeker, Y. Mohseni Behbahani, L. David, E. Laine, **A. Carbone**. "Protein partners discrimination reached with coarse-grain docking and binding sites predictions", *PLoS Computational Biology*, 2022.
3. B. Seoane, **A. Carbone**. "The complexity of protein interactions unravelled from structural disorder", *PLoS Computational Biology*, 2021.
4. E. Laine, Y. Karami, **A. Carbone**. "GEMME: a simple and fast global epistatic model predicting mutational effect", *Molecular Biology and Evolution*, 2019.
5. C. Dequeker, E. Laine, **A. Carbone**. "Multiple binding sites of protein-protein in- teractions predicted by combining sequence analysis and molecular docking", *Proteins: Structure, Function, and Bioinformatics*, 2019.
6. R. Raucci, E. Laine, **A. Carbone**. "Local Interaction Signal Analysis: a new approach for the prediction of protein-protein binding affinity", *Structure*, 2018.
7. F. Douam, F. Fusil, M. Enguehard, L. Dib, F. Nadalin, L. Schwaller, J. Mancip, L. Mailly, T. Baumert, **A. Carbone***, FL. Cosset*, D. Lavillette*. "A protein coevolution method designed for conserved sequences uncovers critical features of the original HCV fusion mechanism and provides molecular basis for the design of effective antiviral strategies", *PLoS Pathogens*, 2018. (* corresponding author)
8. F. Oteri, F. Nadalin, R. Champeimont, **A. Carbone**. "BIS2Analyzer: a server for coevolution analysis of conserved protein families", *Nucleic Acids Research*, 2017.
9. A.Lopes, S.Sacquin-Mora, V.Dimitrova, E.Laine, Y.Ponty, **A.Carbone**. "Protein-protein interactions in a crowded environment: an analysis via cross-docking simulations and evolutionary information", *PLoS Computational Biology*, 2013.

**Place where the thesis will be developed:** Laboratoire de Biologie Computationnelle et Quantitative, Sorbonne Université

**Justification of suitability for *SCAI***:

The project is based on the development of Deep Learning architectures in genomics with strong implications in biology and medicine, of interest for SCAI.

**Profile of the desired student**:

Computer science, mathematics or physics. This profile is required by the processing of large datasets of genomic data, the development of deep learning architectures and, possibly, of efficient methods (algorithms on strings, compression,…) dedicated to GPU architectures for neural networks.