Résumé du projet de thèse *(1 page maximum, en anglais)*
Indiquer la participation de chaque co-directeur et structure dans la gestion du projet. Please indicate explicitely the specific contribution of each supervisor to the PhD project.

Search for protein sequences with adequate functional properties is key to the natural evolution of organisms. A fascinating example of random search in the sequence space is that of **Diversity Generating Retroelements** (DGRs) [1] employed by some bacteriophages to generate variants of their receptor binding protein and thereby change their host range. Through error-prone (at adenine nucleotides) reverse transcription and recombination, highly variable variants of a template region (TR, ~100bp) are produced, resulting into a theoretical 10^{20} unique polypeptide sequences. Such a colossal capacity for exploring the coding sequence space is unmatched by any other biological mechanisms [2], including what is achieved by eukaryotic Immunoglobulin domains.

The purpose of this PhD project is two-fold:
(A) **we will quantitatively characterize the evolutionary potential of DGRs**. It is plausible that DGRs have evolved to optimize the codon usage and adenine positions in the TR to generate diverse, potentially high-quality proteins. It has been for instance shown that the positions of the adenines ensure that stop codons are never introduced. However, to what extent this optimization goes beyond the avoidance of non-sense mutations, and how it shapes the exploration of the sequence space remain largely unknown. To answer this question, we will use machine-learning models [3] trained from homologous protein sequence data developed in the Cocco-Monasson group (LPENS, PSL-SU, hosted on the SU Jussieu campus from May 2024), which were shown to successfully design new functional proteins [4,5]. We will study how the random exploration process defined by DGRs explore the protein sequence space and determined the best putative TR for protein variant generation.
(B) **we will use DGR for machine-guided protein design**. The conception of novel proteins with specified function and biochemical properties is a longstanding goal in bioengineering with applications across medicine and nanotechnology. The Bikard lab at Institut Pasteur has for the first time harnessed DGRs in a compact system that functions in *E. coli*, and allows for massive protein diversification and functional screening in directed evolution setups. As a model system we will apply this method to the generation of Cas9 variants with modified PAM recognition specificities. The computational models developed in (A) will allow us to propose TR, which will be used to produce variants of Cas9. In turn, the results of the functional screening will help us improve the models.

As a conclusion, this project will allow the Cocco-Monasson and Bikard labs to explore the synergy between data-driven modelling and directed evolution strategies by employing DGRs as a novel in vivo targeted mutagenesis tool. It builds on the successful existing collaboration between the two groups, which has resulted in common articles [5,6,7] and the cosupervision of a PhD student (2019-2022).

References related to the project (names of cosupervisors are underlined): [1] Doulatov, S. *et al.* Tropism switching in *Bordetella* bacteriophage defines a family of diversity-generating retroelements. *Nature* **431**, 476–481 (2004); [2] Le Coq, J. & Ghosh, P. Conservation of the C-type lectin fold for massive sequence variation in a Treponema diversity-generating retroelement. *Proc Natl Acad Sci USA* **108**, 14649–14653 (2011); [3] Tubiana, J., Cocco, S. & Monasson, R. Learning protein constitutive motifs from sequence data. *eLife* **8**, e39397 (2019); [4] Russ, W.P., ..., Monasson R., Cocco S., Weigt, M. & Ranganathan R. An evolution-based model for designing chorismate mutase enzymes. *Science* **369**, 440–445 (2020); [5] Malbranke, C., Rostain, W., Depardieu, F., Cocco, S., Monasson, R., & Bikard, D. Computational design of novel Cas9 PAM-interacting domains using evolution-based modelling and structural quality assessment. *PLoS Comp. Biol.* **19**:e1011621 (2023); [6] Malbranke, C., Bikard, D., Cocco, S. & Monasson, R. Improving sequence-based modeling of protein families using secondary structure quality assessment. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab442; [7] Malbranke, C., Bikard, D., Cocco, S., Monasson, R. & Tubiana J. Machine learning for evolutionary-based and physics-inspired protein design: Current and future synergies. *Current Opinion in Structural Biology* **80**:102571 (2023).