**Controllable protein design with unsupervised language models**

Noelia Ferruz

Department of Structural and Molecular Biology, Molecular Biology Institute of Barcelona (IBMB-CSIC), Barcelona Science Park, 08028 Barcelona, Spain.

*Corresponding author:* noelia.ferruz@ibmb.csic.es

Artificial Intelligence (AI) methods are emerging as powerful tools in fields such as Natural Language Processing (NLP) and Computer Vision (CV), impacting the tools and applications we use in our daily lives. Language models have recently shown incredible performance at understanding and generating human text, producing text often indistinguishable from that written by humans. Inspired by these recent advances, we trained a language model, ProtGTP2, which effectively learned the protein language and generated sequences in unexplored regions of the protein space. A desirable critical feature in protein design is having control over the design process, i.e., designing proteins with specific properties. For this reason, we trained ZymCTRL, a model trained on enzyme sequences and their associated Enzymatic Commission (EC) numbers. ZymCTRL generates enzymes upon user-defined specific catalytic reactions, which show natural-like catalytic activities in wet lab experiments. Lastly, we have trained REXzyme, a translation machine capable of designing enzyme sequences for user-defined chemical reactions.