ERC - StG

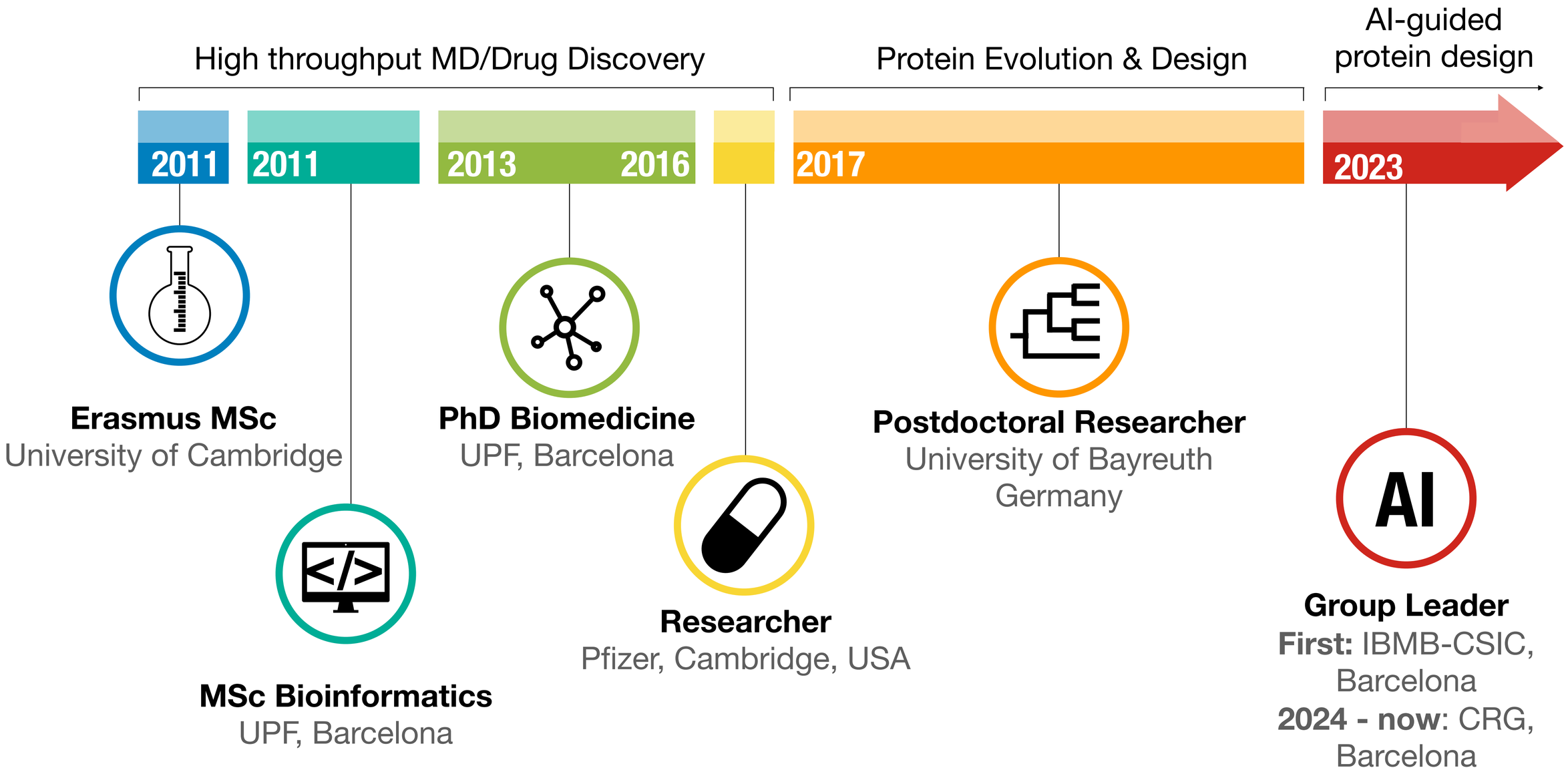
My own experience (and some tips)

Noelia Ferruz
Group Leader
Centre for Genomic Regulation

CRG
Centre
for Genomic
Regulation

My Career

Timeline



The project

Timeline

Postdoc not going well

Maternity leave

Group Leader

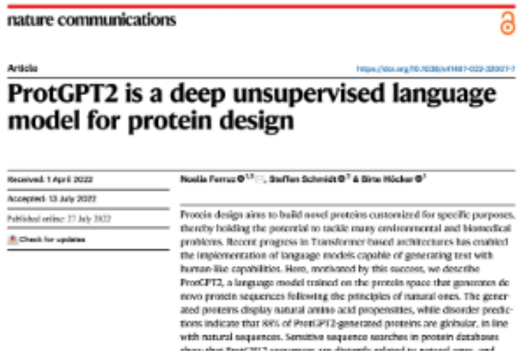
Ramon y Cajal

2017 2018 2019 2020 2021 2022 2023

Started having a very rough idea of things that got me excited
(AI, proteins)

Emmy Noether
Volkswagen-Stiftung
Juan de la Cierva
Max Planck Institut
W2 Regensburg
La Caixa Junior Leader

Proof of concept


nature communications

Article

ProtGPT2 is a deep unsupervised language model for protein design

Received: 1 April 2022

Accepted: 13 July 2022

Published online: 27 July 2022

Check for updates

Noelia Ferruz, Steffen Schmidt & Birte Höcker

Protein design aims to build novel proteins customized for specific purposes, thereby holding the potential to tackle many environmental and biomedical problems. Recent progress in Transformer based architectures has enabled the implementation of language models capable of generating text with human-like capabilities. Here, motivated by this success, we describe ProtGPT2, a language model trained on the protein space that generates de novo protein sequences following the principles of natural ones. The generated proteins display natural amino acid propensities, while disorder predictions indicate that 88% of ProtGPT2-generated proteins are globular, in line with natural sequences. Sensitive sequence searches in protein databases show that ProtGPT2 sequences are dissimilar to natural ones, and

At the time of writing:
20 papers
A group of 1 person (XD)

At the time of interview:
24 papers
A group of 5-6 people
Funding >1.2M

Postdoc not going well

Maternity leave

Group Leader

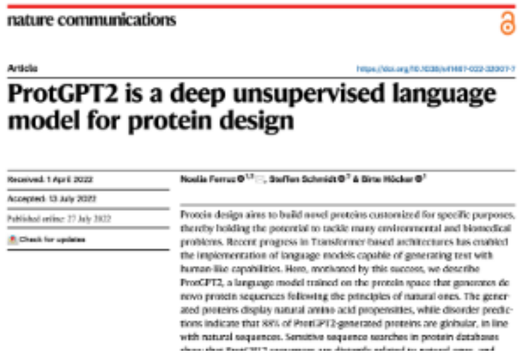
Ramon y Cajal

2017 2018 2019 2020 2021 2022 2023

Started having a very rough idea of things that got me excited
(AI, proteins)

Emmy Noether
Volkswagen-Stiftung
Juan de la Cierva
Max Planck Institut
W2 Regensburg
La Caixa Junior Leader

Proof of concept


nature communications

Article

ProtGPT2 is a deep unsupervised language model for protein design

The Writing

May: I decided on one idea. Dropped it. Decided on a new idea in June.

Wrote it for computer science (PE6). Didn’t feel right. Rewrote it for LS9

Finished writing part B1 around late July (everyone had advised writing B2 first)

I asked around 10-12 people to read my part B1.

Did start writing B2 until around late August/early September, but I had read the methods and everything was VERY clear in my mind.

Nobody read my part B2. (Felt bad asking to read something so long). Don’t be me!

Prepared submission for late September (due to baby coming). In the end I spent 4-5 months full time on writing. I ended up having more time in October/November (and I could polish a lot)

Tip 1: Include a graphic abstract. Why not?

ERC Starting Grant 2024
Research proposal [Part B1]

An intelligent agent for general-purpose protein engineering

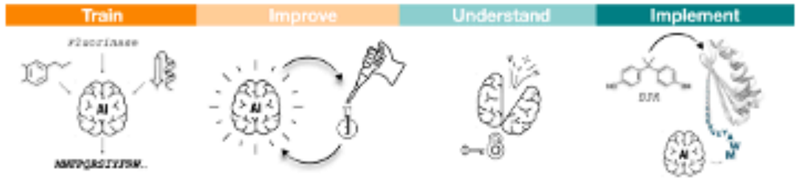
ATHENA

- Name of the Principal Investigator (PI): Noelia Ferruz
- Host Institution for the project: Molecular Biology Institute of Barcelona (IBMB-CSIC)
- Proposal duration in months: 60

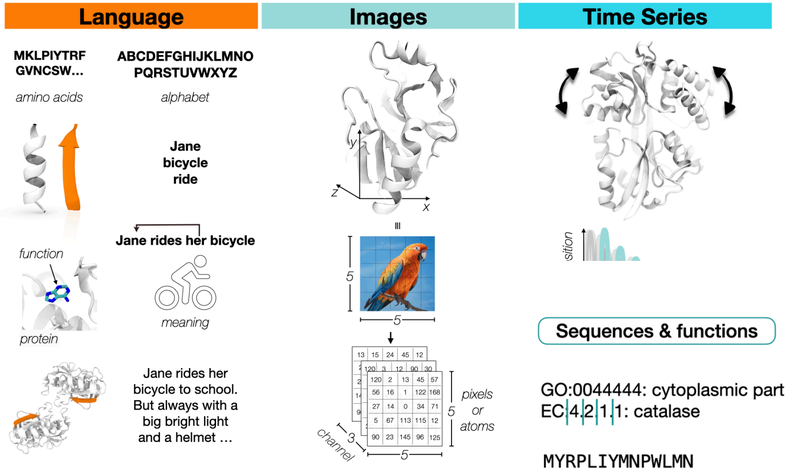
Proteins offer an exciting path to address a multitude of biotechnological challenges. Functionally encoded by amino acid sequences and capable of working under non-toxic, mild conditions, their controllable design has been pursued for decades. Even though Artificial Intelligence (AI)-based models have recently made giant strides in computational protein design, it is evident that we cannot yet design proteins as efficient as those present in nature. To gain a technological advantage in a world with pressing demands in sustainability and healthcare, we must **improve** and **accelerate** the development of **tailored**, **proficient** proteins.

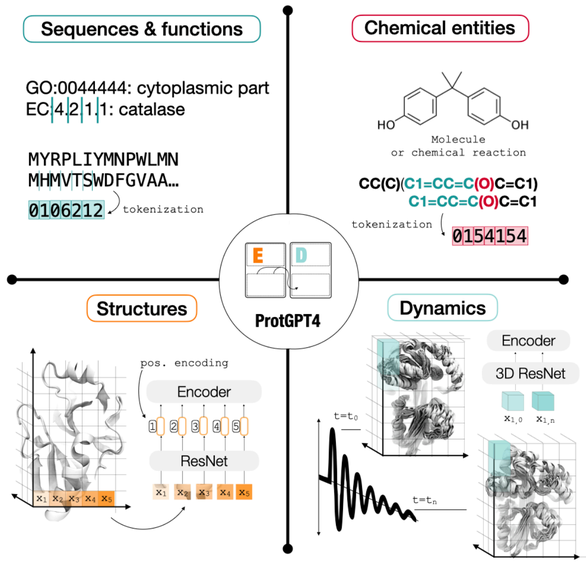
In ATHENA, I aim to develop a **multi-task intelligent agent** capable of efficiently engineering functional proteins tailored to user-defined specifications. To achieve this goal, the agent will be trained using available sequence, structural, functional, and dynamic data to perform multiple protein engineering tasks. The agent will iteratively improve from experimental feedback using **Reinforcement Learning**, and **explainable AI** will allow us to '*open the black box*' and understand its decision-making process. A vital component of this work will be its rigorous **experimental validation**, tackling traditionally challenging tasks with biotechnological applications.

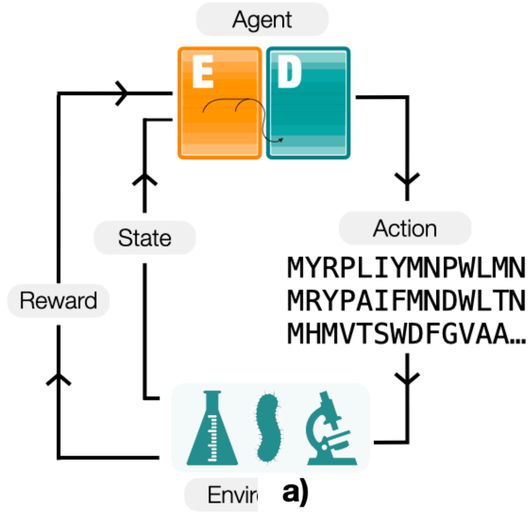
ATHENA will deliver an intelligent agent with continuous learning capabilities, freely accessible through user-friendly interfaces, empowering researchers worldwide with an easy-to-use tool for tailored protein design. Through incorporating XAI, this project will provide a new angle to understanding fundamental **sequence-to-function relationships**. Moreover, a comprehensive experimental validation process will offer innovative solutions to biotechnological and environmental challenges complex to address with currently available approaches.

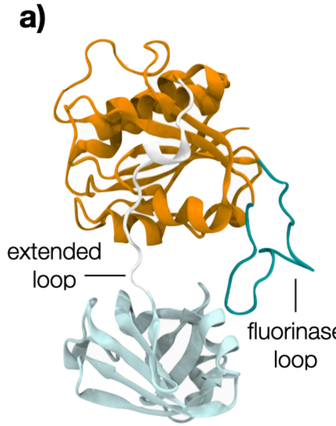


Tip 2: Make figures consistent among themselves (palette)

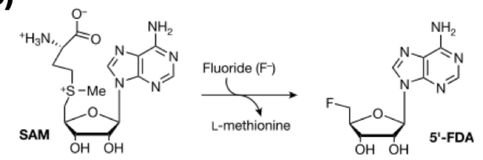








b)



c)

Organism	kM (μM)	Kcat (min⁻¹)
S. Cattleya	29.2	0.083
S. Sp MA37	82.4	0.262
N. Brasiliensis	27.8	0.122
Actinoplanes	45.8	0.204
Methanosaeta sp.	6.99	0.41
S. Xinghaeinsis	3.76	0.22
S sp. SAJ15	9.62	0.34

Tip 3: Block time for writing and put yourself an earlier deadline

Good luck!