SCIENCES & BIOENGINEERING SCIENCES

The Research Group
Artificial Intelligence Lab

has the honor to invite you to the public defence of the PhD thesis of

Konstantina Tzavella

to obtain the degree of Doctor of Bioengineering Sciences

Title of the PhD thesis:

Combining protein Language Models and Evolutionary Information in a constrained Foundation Model to explore the sequence space and predict the effect of mutations on protein behavior

Supervisor:
Prof. dr. Wim Vranken (VUB)
Co-supervisor:
Prof. dr. Catharina Olsen (VUB)

The defence will take place on

Wednesday, November 12, 2025 at 4.30 p.m.

LIC Learning & Innovation Center ULB-VUB, Bd de la Plaine 2, 1050 Bruxelles

The defence can be followed through a live stream on Microsoft Teams.

Members of the jury

Prof. dr. Dominique Maes (VUB, chair)
Prof. dr. Jef Vandemeulebroucke (VUB)
Prof. dr. Eva Hadadi (VUB)
Prof. dr. Yves Moreau (KU Leuven)
Prof. dr. Elodie Laine (Sorbonne Université, FR)

Curriculum vitae

Konstantina Tzavella was born in Akrata, a small coastal village in southwestern Greece. She moved to Athens to pursue her studies in Electrical and Computer Engineering, followed by a master's degree in Biomedical Engineering in Paris in 2015.

After working for several years in the pharmaceutical industry, she returned to academia and began her PhD at the Vrije Universiteit Brussel in 2020, within the interdisciplinary VUB/UZB TumorScope project.

Abstract of the PhD research

Mutations are changes in an organism's genetic material. They are fundamental to evolution but also central to our understanding of disease, as they can alter the function of proteins, disrupt cellular processes and ultimately impair the health of the organism. Despite their importance, the effects of nearly 98% of mutations in humans remain unknown.

Many state-of-the-art (SOTA) predictors have been developed to predict the effect of single mutations, often relying on evolutionary information that captures how the sequence of specific proteins has evolved over time. However, few methods can account for the combined effects of multiple mutations. These interactions - known as *epistasis* - play a crucial role in evolution, genetic stability, and disease, but remain poorly understood.

Language Models (LMs), such as ChatGPT, have in a very short time become ubiquitous in society. LMs provide a powerful framework for modelling statistical dependencies in sequences of words. Just as LMs learn word relationships in natural language, protein LMs (pLMs) infer statistical dependencies between amino acids.

This dissertation addresses four central questions: (1) How do pLMs perform compared to current SOTA in predicting single and multiple mutation effects? (2) Can inclusion of evolutionary information enhance their predictive performance? (3) Do pLMs capture epistasis and reflect underlying biophysics? (4) Can we meaningfully interpret them for biological applications?

To explore these questions, we developed a novel evolutionarily constrained foundation model for predicting mutation effects in biologically and clinically relevant contexts, such as cancer driver mutation classification. The results reveal that, while SOTA predictors have made progress, they often suffer from key limitations, especially overreliance on structural data and biases toward well-studied genes. PLM-based approaches, on the other hand, can mitigate these challenges, even when their encoded knowledge remains only partially understood.

Overall, this work contributes to improving mutation effect prediction as well as providing a deeper biological understanding of pLMs themselves. It also enables further development of diverse clinical and experimental applications.