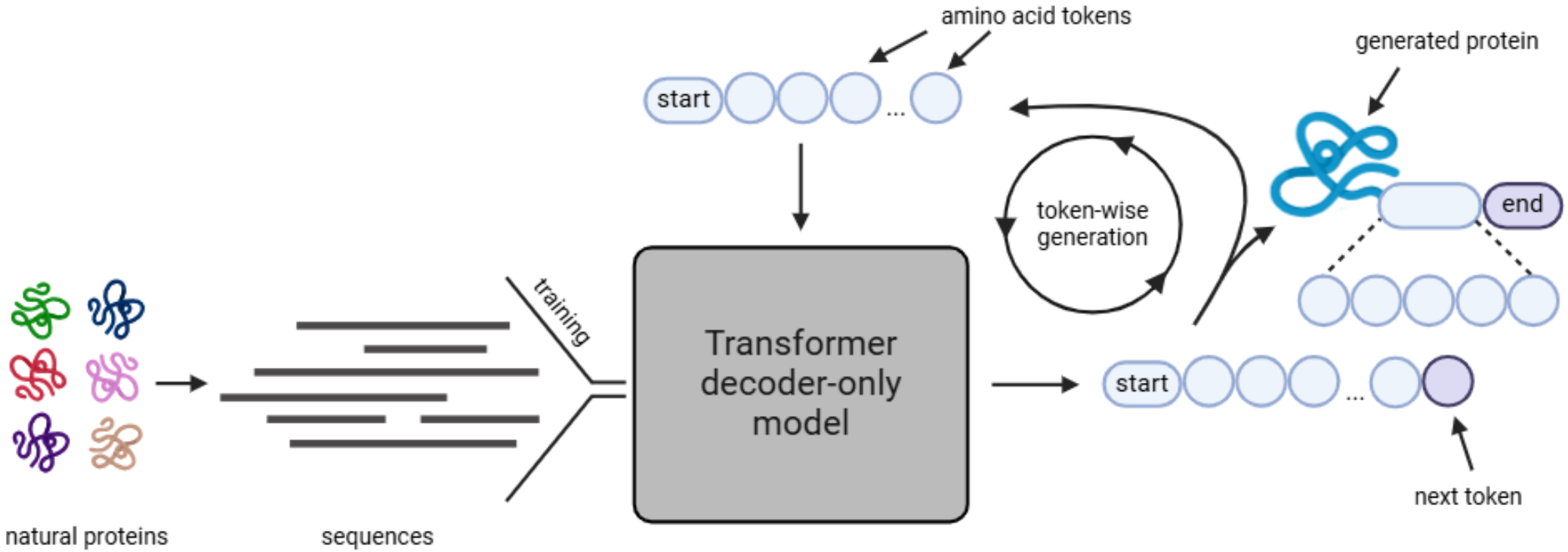
Explainability of Transformer-based models for the design of protein sequences

AiChemist School - Lausanne

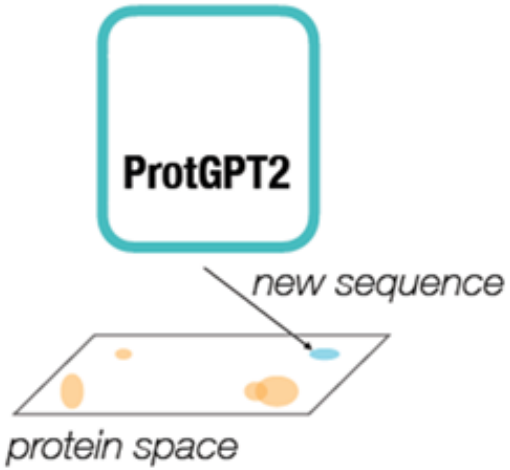
Andrea Hunklinger

25.04.2025

Generative protein language model

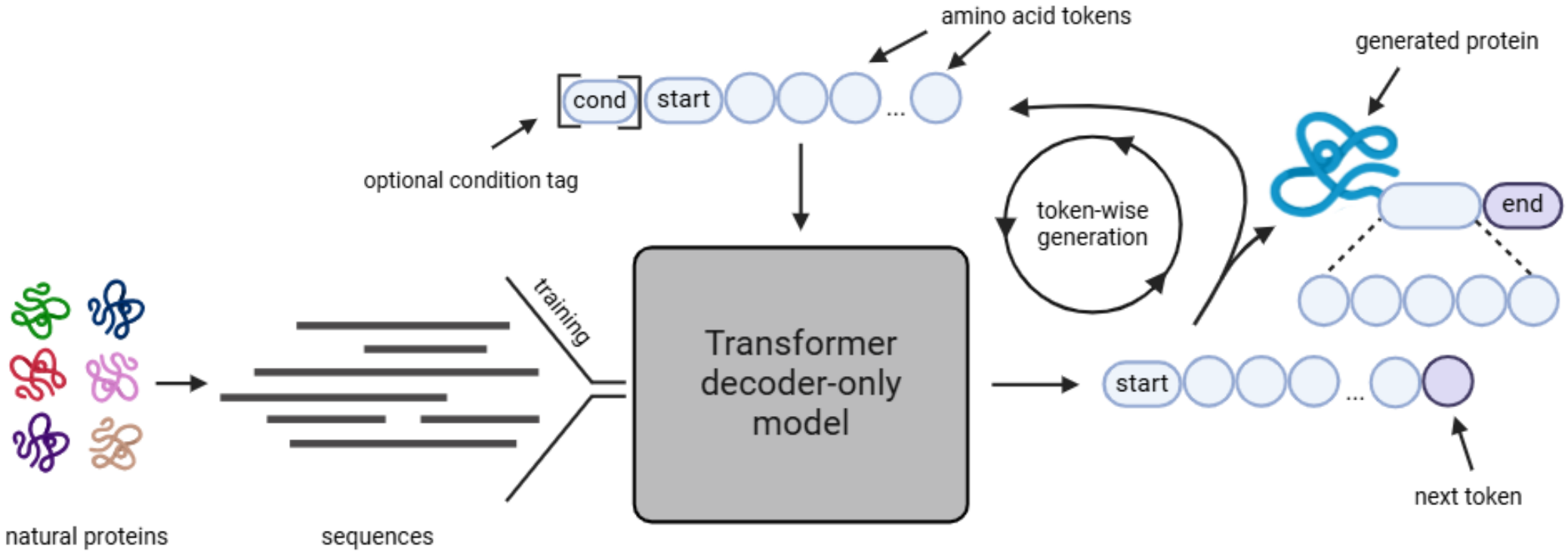


Protein language models

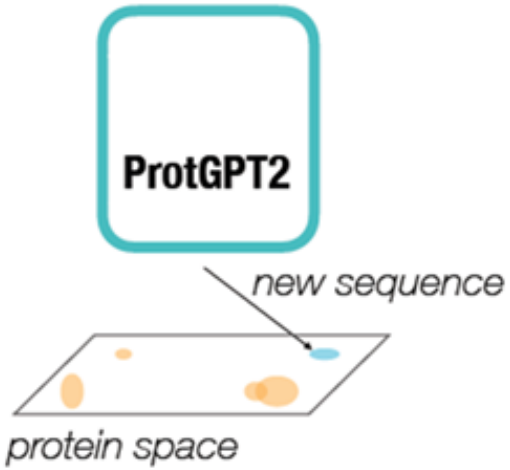


Ferruz N et al. ProtGPT2 Is a Deep Unsupervised Language Model for Protein Design. Nat. Commun., 13, 4348 (2022).

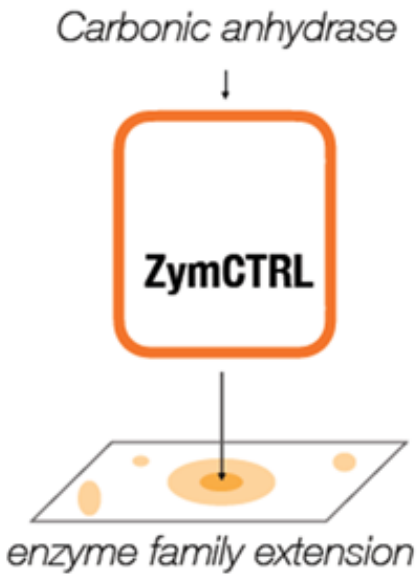
Generative protein language model



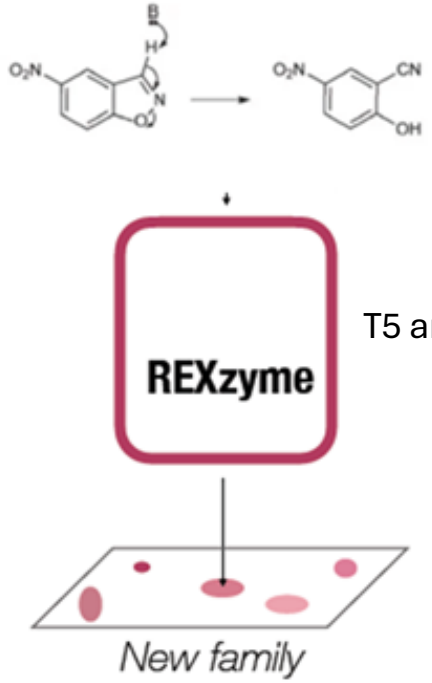
Protein language models



Ferruz N et al. ProtGPT2 Is a Deep Unsupervised Language Model for Protein Design. Nat. Commun., 13, 4348 (2022).



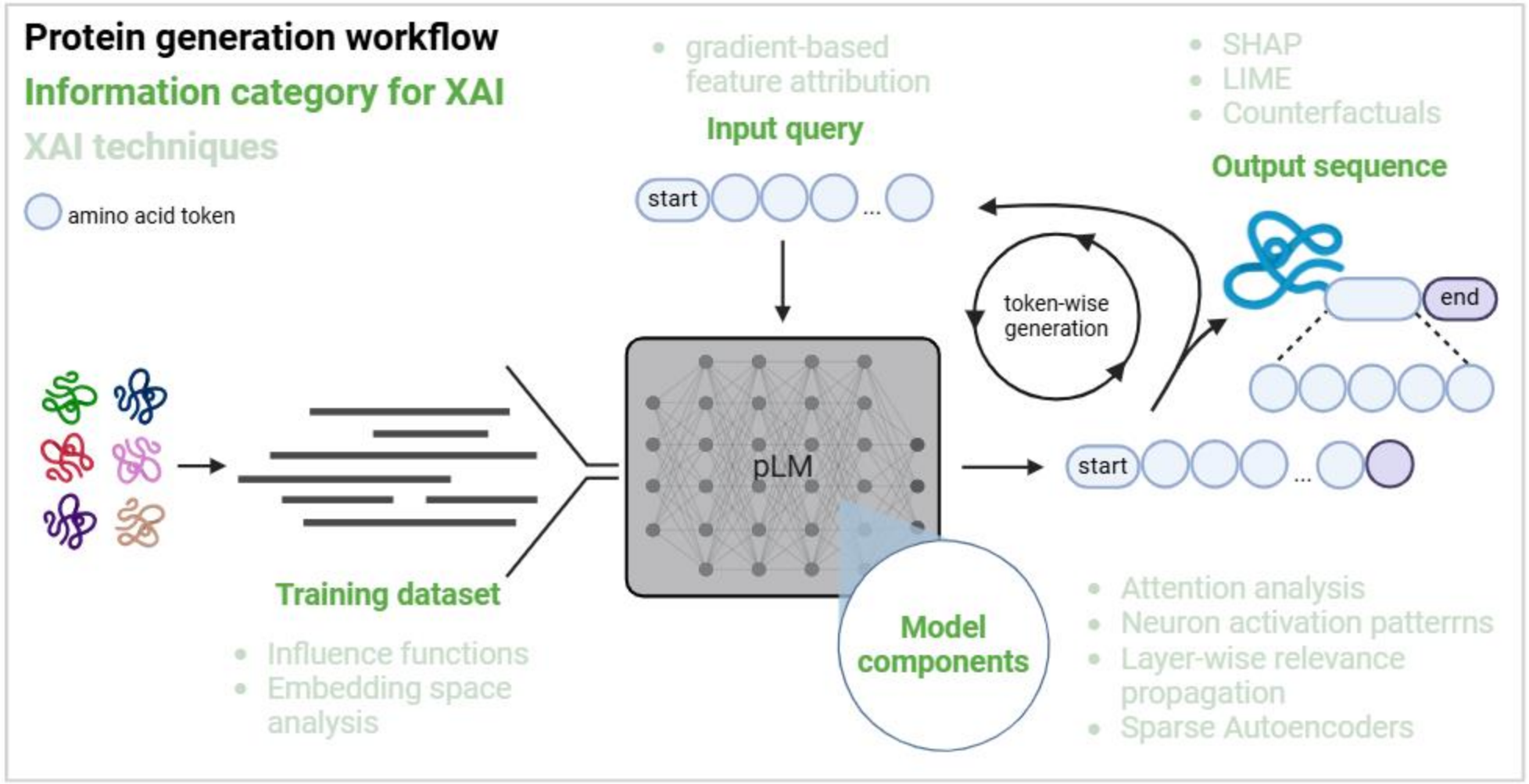
Munsamy G, … Ferruz N. Conditional language models enable the efficient design of proficient enzymes. *BioRxiv preprint:*10.1101/2024.05.03.592223 (2024).



T5 architecture

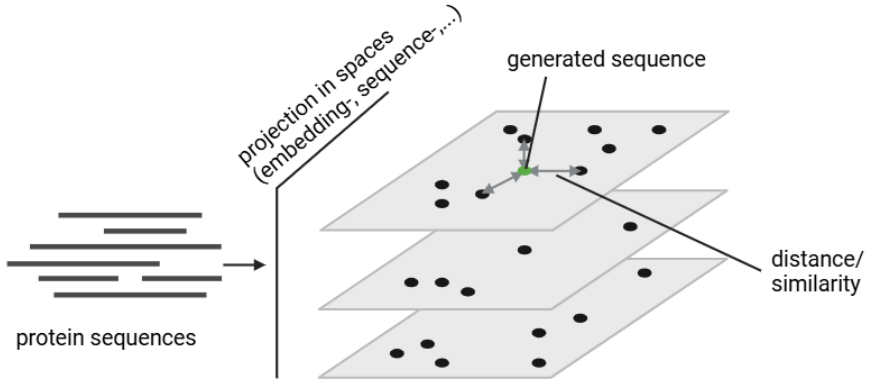
Testing in progess

Categorization of XAI techniques

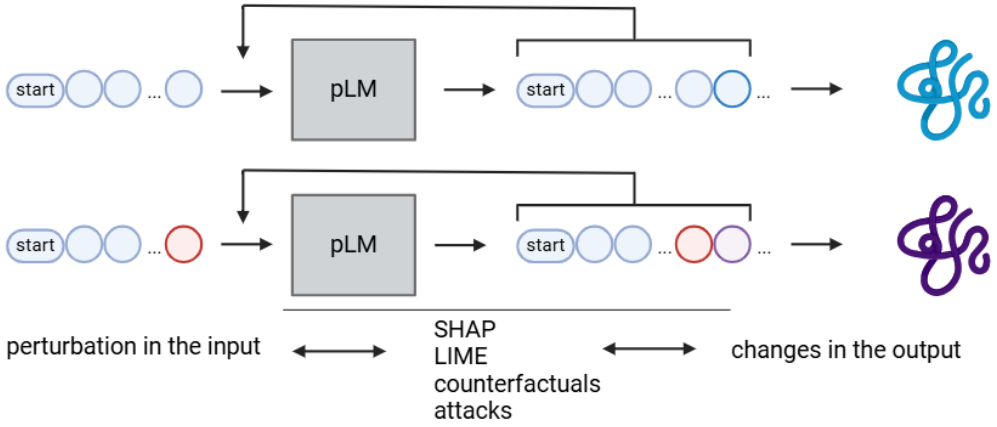


Four information categories

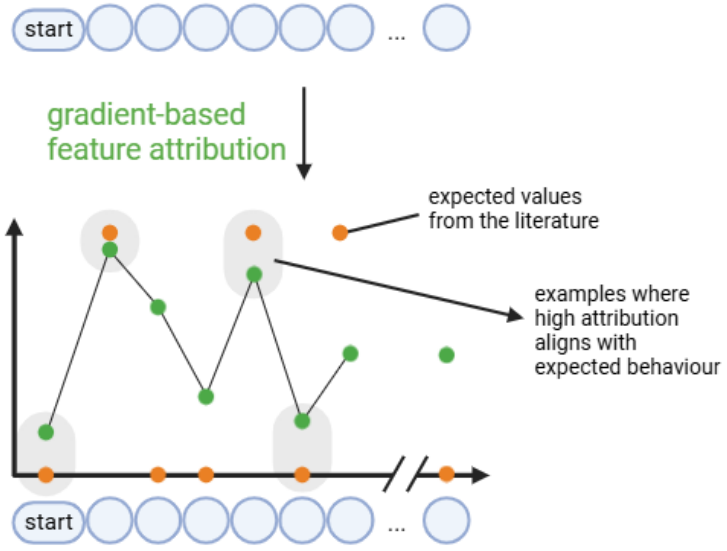
Training dataset



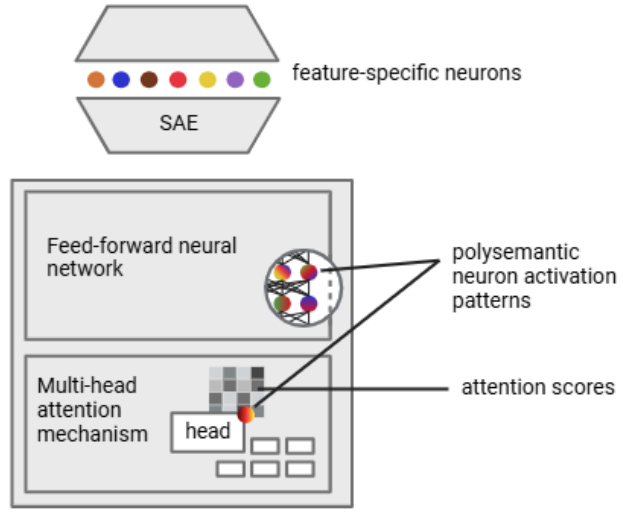
Output sequence



Input query



Model components

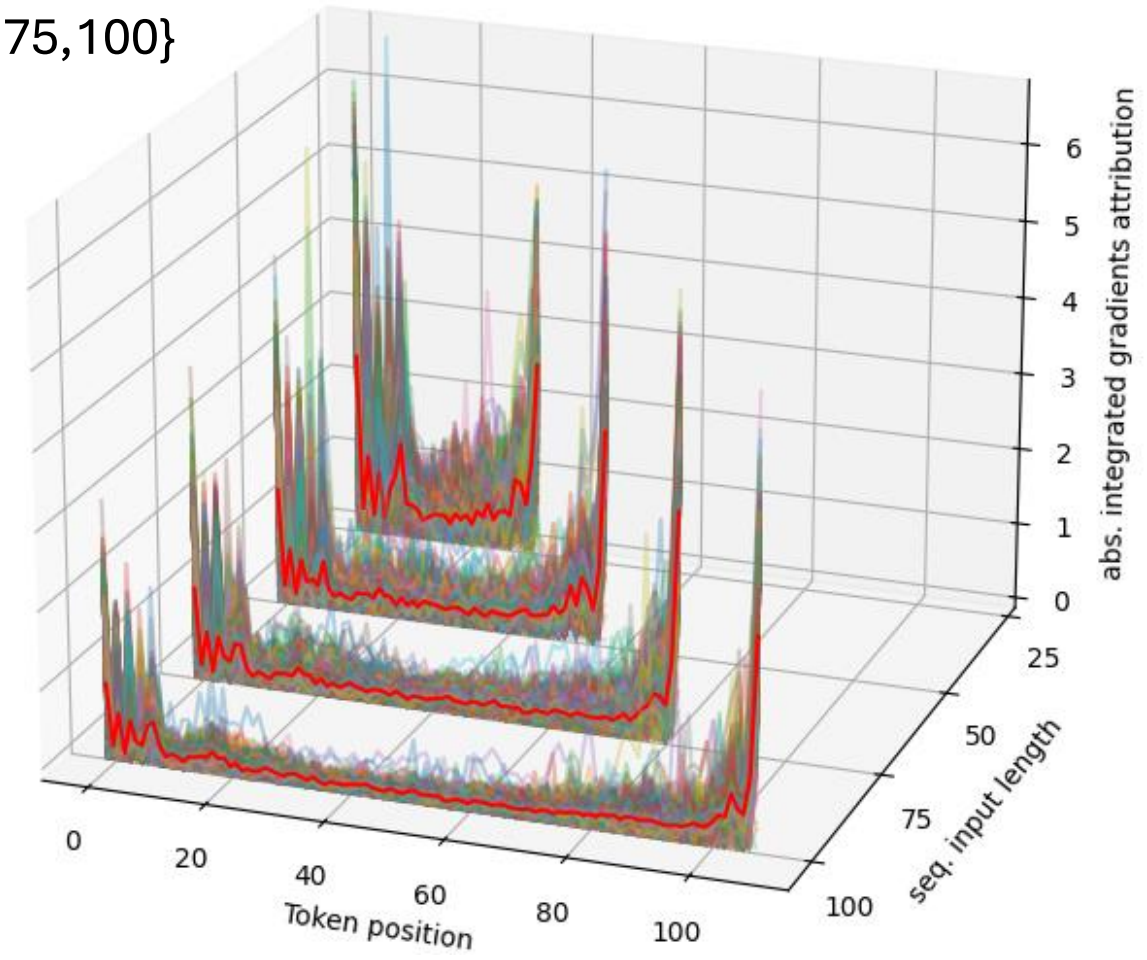


use-cases – Input query

- Input query for ZymCTRL

  x.x.x.x<sep><start>sequence[:n] for n ∈ {25,50,75,100}

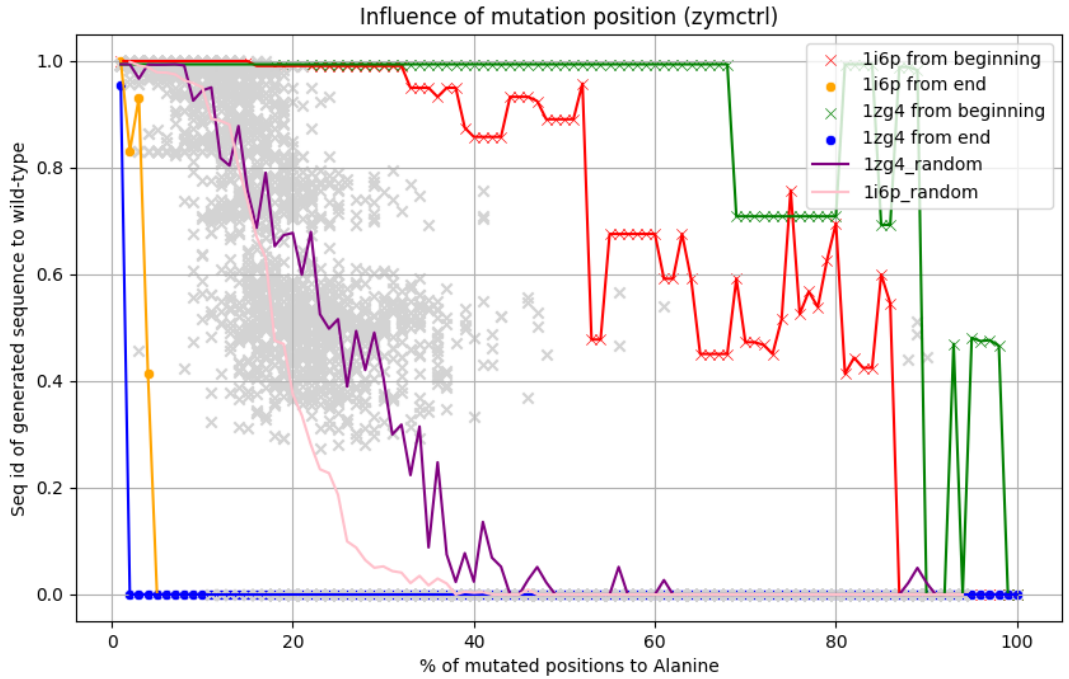
- Calculate attribution (Integrated Gradients) for the generation of residue n+1
- Results showed **enhanced importance** on condition (EC label) and local environment
- Different trend in ProGen[1] (trained on sequences and their reverse)



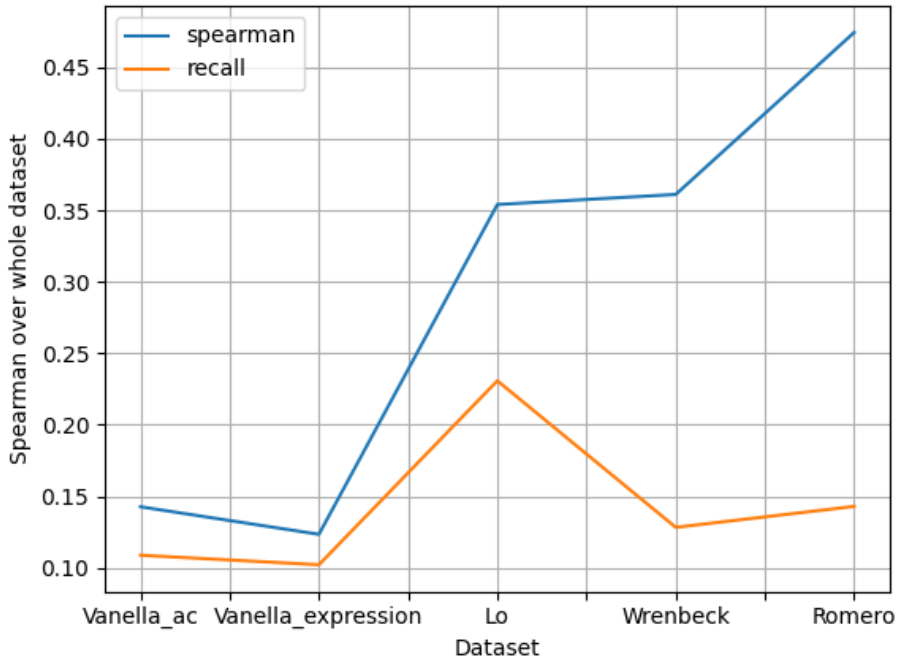
[1]Madani, Ali, et al. Progen: Language modeling for protein generation. arXiv preprint arXiv:2004.03497 (2020).

use-cases – output sequence

- Alanine scanning (generation)
  - Mutation in end of input more dramatic change in output



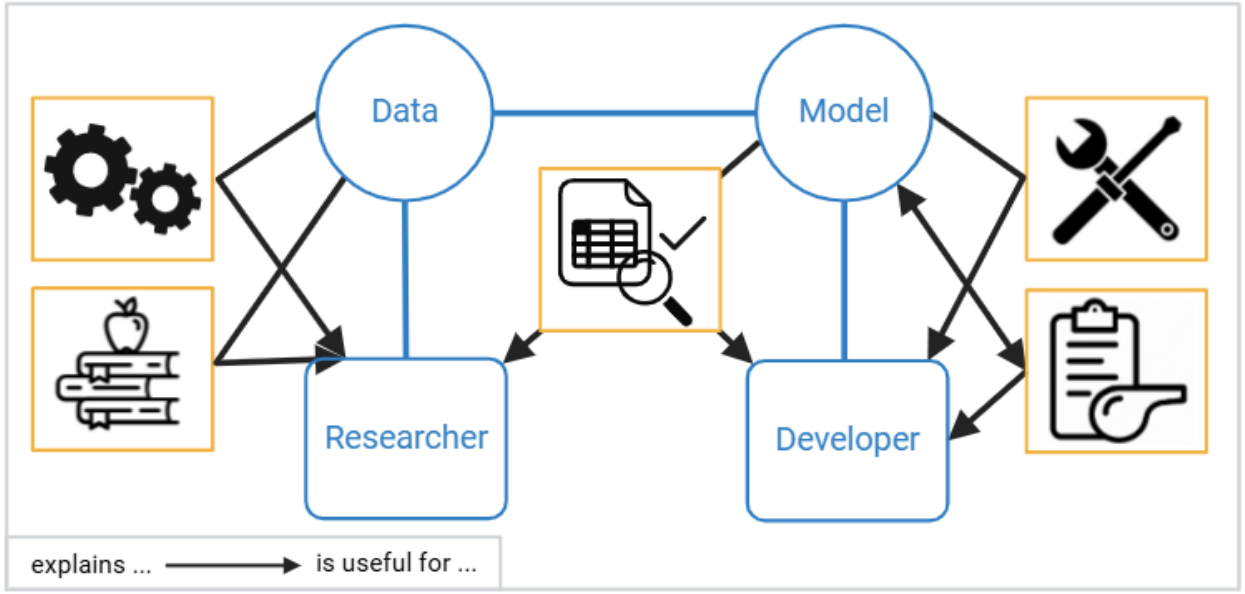
- Mutational effect prediction (token probability)
  - Comparable values with ProteinGym[1] benchmark



[1]Notin, Pascal, et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. Advances in Neural Information Processing Systems, 36, 64331-64379 (2023).

Roles of XAI

- **Evaluator**: compare extracted pattern with expectations
- **Multitasker**: use insights for annotation
- **Teacher**: extract general rules
- **Engineer**: change model efficiency
- **Coach**: improve output



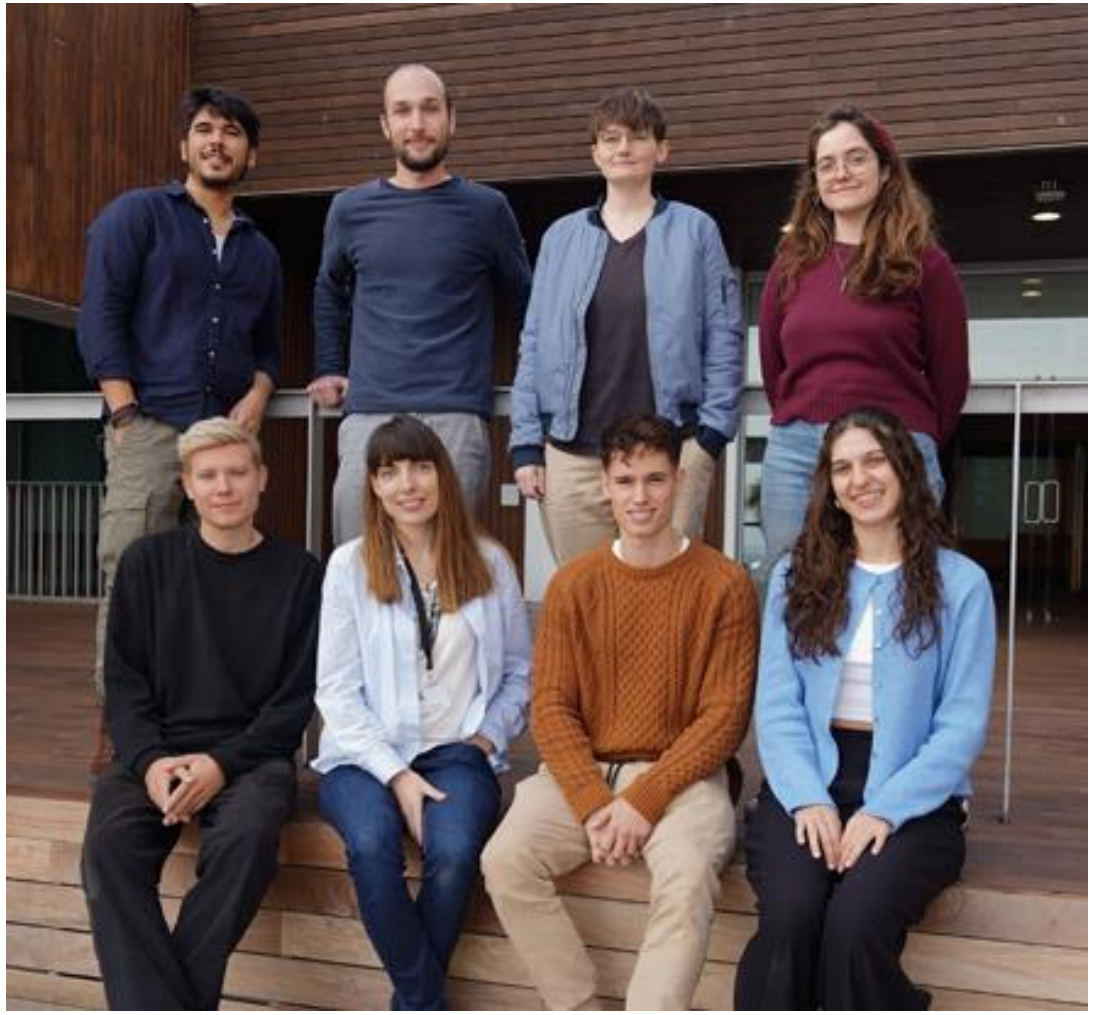
→ Perspective paper

Application of XAI to pLMs

Role	Information category	XAI method	# applications
Evaluator	Training dataset	Influence functions	1
	Input query	Gradient-based feature attribution	2
	Model components	Attention scores	8
		Layer-wise relevance propagation (LRP)	theoretical
		Sparse Autoencoders (SAE)	3
	Output sequence	Shapley Additive Explanations (SHAP)	10
		Local Interpretable Model-Agnostic Explanations (LIME)	2
		Counterfactual-like	5
Multitasker	Training dataset	Embedding space distances	1
	Model components	Attention scores	2
		Sparse Autoencoders (SAE)	3
	Output sequence	Counterfactual-like	1
Coach	Input query	Gradient-based feature attribution	theoretical
	Model components	Sparse Autoencoders (SAE)	1
Engineer	Model components	Attention scores	theoretical
Teacher	unclear	unclear	-

→ Underexplored methods and roles

Thank you! Questions?



Noelia Ferruz
Ramiro Illanes
Filippo Stocco
Maria Artigues
Lasse Middendorf
Alex Vicente
Núria Mimbrero
Michele Garibbo
Lucia Urcelay Ganzabal

CRG Centre for Genomic Regulation

Universitat de Barcelona

AiChemist program

MSCA Doctoral Network Fellowship

Santiago Villalba
Igor Tetko
Katya Ahmad

BAYER