An Enhanced Protein Language Model for Cell Surface Protein Identification in Genomic Therapies

C. Canavate[1], L. Polit[1], O. de Felice[1], D. Zielinski[1], D. Del Bourgo[1], J. Cottineau[1]

[1]WhiteLab Genomics, FUTURE4CARE, 8 rue Jean Antoine de Baif, 75013 Paris, France

AI for Genomic Medicine

major challenges

Specificity — **Immunogenicity** — **Manufacturing complexity**

AI-based solutions by WhiteLab

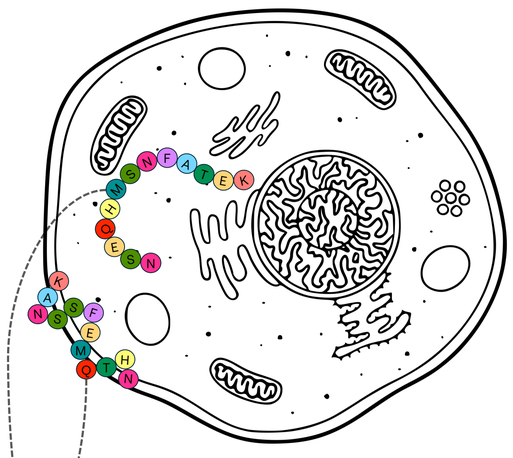
Accelerate
R&D by several years
35-40% faster

Increase
Productivity of R&D resources
Multiplied by 4

Reduce
Unnecessary experiment costs
25-30%

Improve
Pre-clinical success rate
20-30%

In a nutshell



Team & Contact

WhiteLab Genomics is a pioneering in-silico company leveraging Artificial Intelligence to accelerate discovery and mitigate risks in early-stage research and development pipelines exclusively within the field of genomic medicine. Founded in 2019, and backed by Y-Combinator, WhiteLab stands at the convergence of biology and computer science.

Boston — Paris

whitelabgx.com I bd@whitelabgx.com
ccanavate@whitelabgx.com

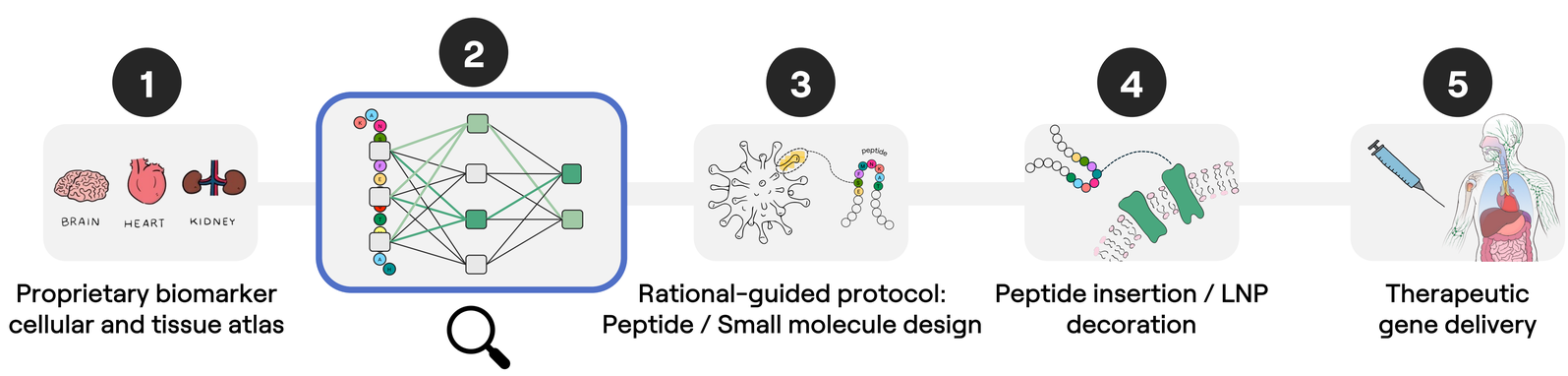
References

This work has led to a patent application.

Bausch-Fluck D. et *al.*, A Mass Spectrometric-Derived Cell Surface Protein Atlas. PLoS One 10: e0121314, 2015.

Bausch-Fluck D. et *al.*, The in silico human surfaceome. PNAS, 13, 115(46), 2018.

Lin Z. *et al.*, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379, 1123-1130, 2023.



A machine learning model for cell surface protein identification

CSPA
Experimental cell surface capture technique
881 proteins found at the cell surface with high confidence

SURFY
In silico ML model trained on pre-computed features
7903 proteins predicted

Low number of proteins characterized or predicted

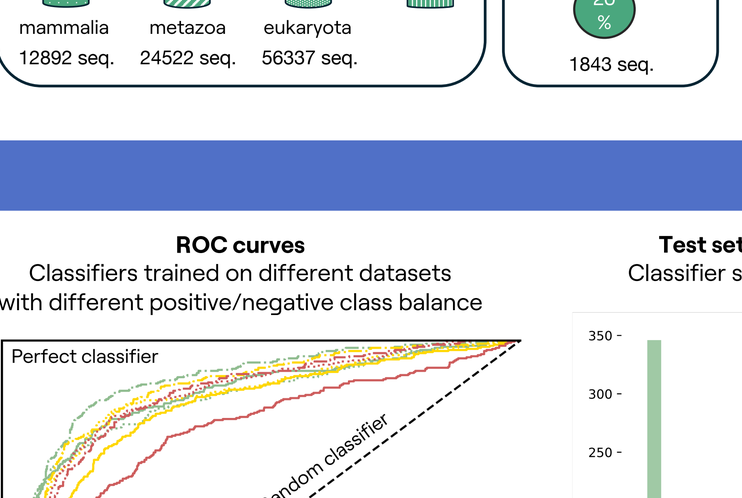
The ML model relies on very specific features computation

There is a need for a **generalizable** model that can **predict all proteins without** preliminary calculation of **specific features**

I. Datasets creation

Goal: generate our own comprehensive training datasets with cell surface and non cell surface proteins

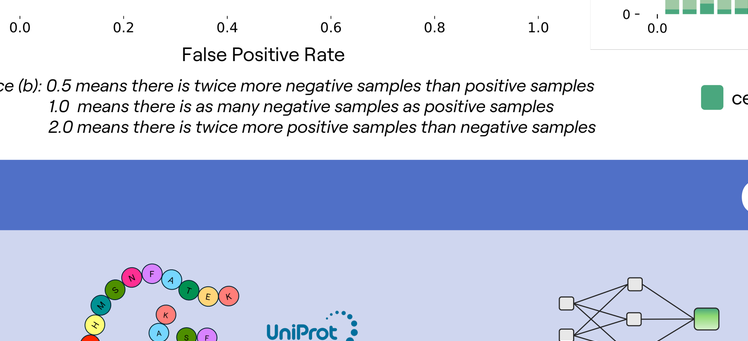
1. Query UniProt database with appropriate keywords and by taxonomy



2. Filter human proteins with experimental sources — COMPARTMENTS subcellular localization db

3. Cluster sequences based on their identity — mmseqs2

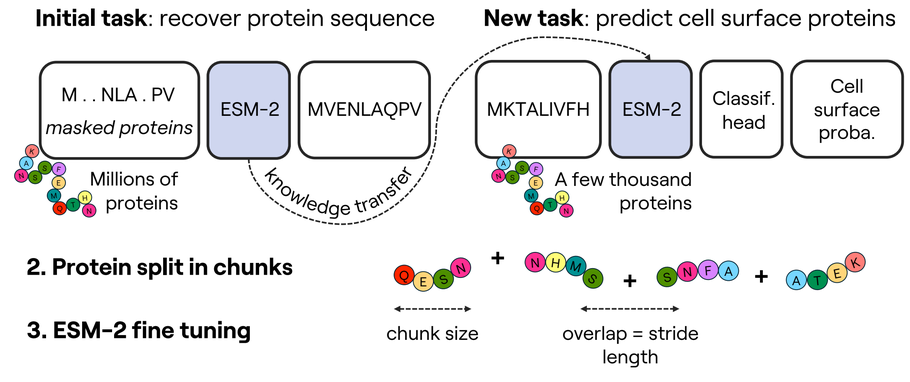
4. Aggregate sequences and custom balance between sets



II. Model development

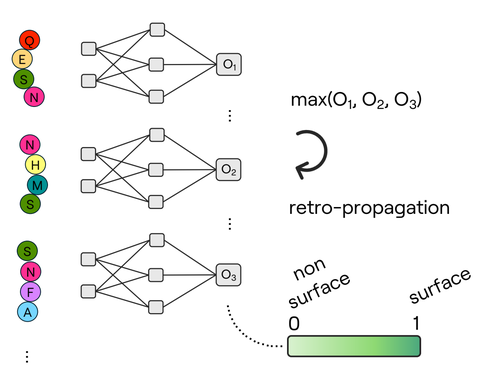
Goal: obtain a feature-free ML model

1. Protein language model selection and transfer learning

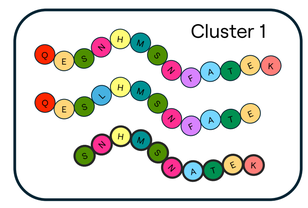


2. Protein split in chunks

3. ESM-2 fine tuning



4. With data augmentation



randomly sample sequence clusters at each epoch

5. And hyperparameters search — 5-fold cross validation

6. Performances assessment

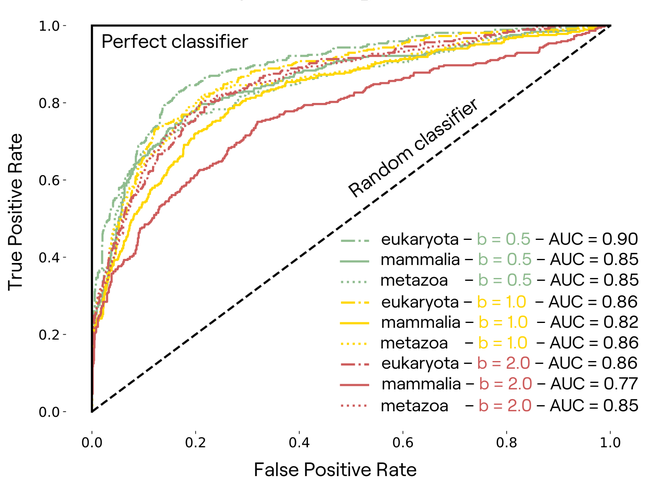
Accuracy
Overall correctness of the model on all proteins

Recall
How well the model finds the surface proteins

Precision
Accuracy of the positive predictions made by the model

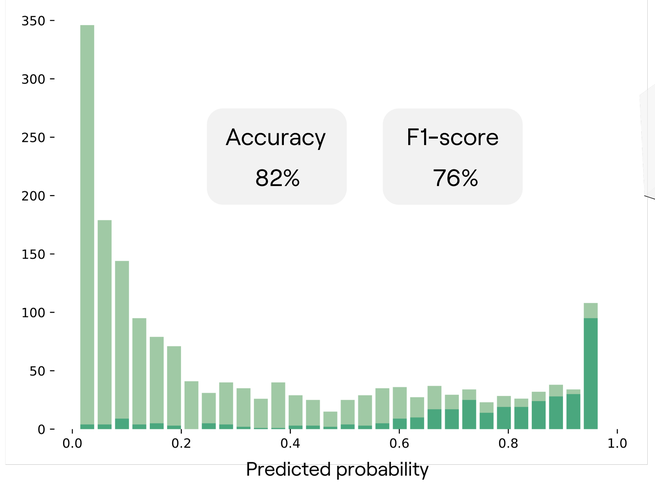
Results

ROC curves
Classifiers trained on different datasets with different positive/negative class balance

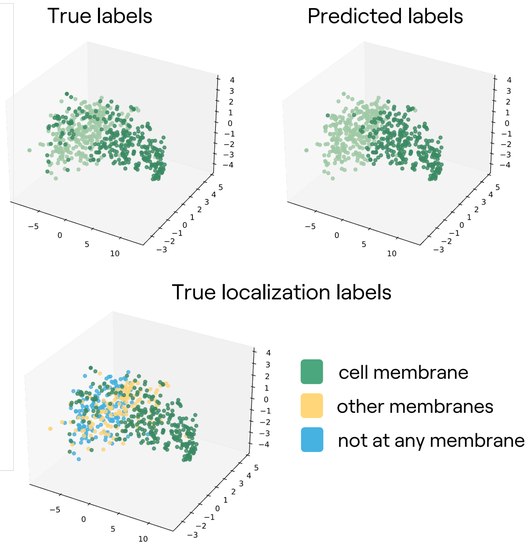


balance (b): 0.5 means there is twice more negative samples than positive samples
1.0 means there is as many negative samples as positive samples
2.0 means there is twice more positive samples than negative samples

Test set predicted probabilities distribution
Classifier selected is the one trained on eukaryota dataset with balance = 0.5



Test set embeddings after training
t-SNE projections



Conclusion

A new methodology to create a protein dataset

A completely feature-free model that can predict from any protein sequence

The classifier trained on the larger dataset (eukaryota) and with the balance closest to the real distribution achieves the best performances

The main challenge is distinguishing cell membrane proteins from those of other membranes inside the cell, with the next focus on refining this