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

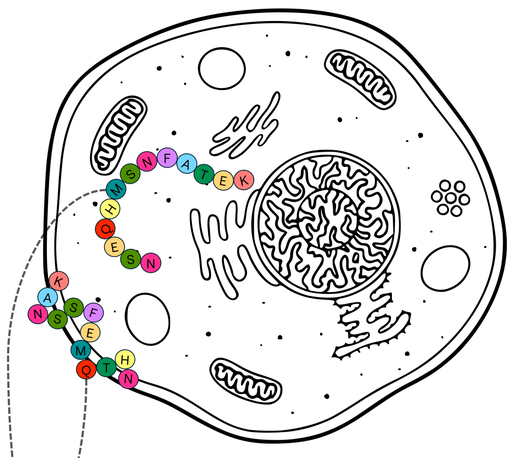
Accelerate
R&D by several years
35-40% faster

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

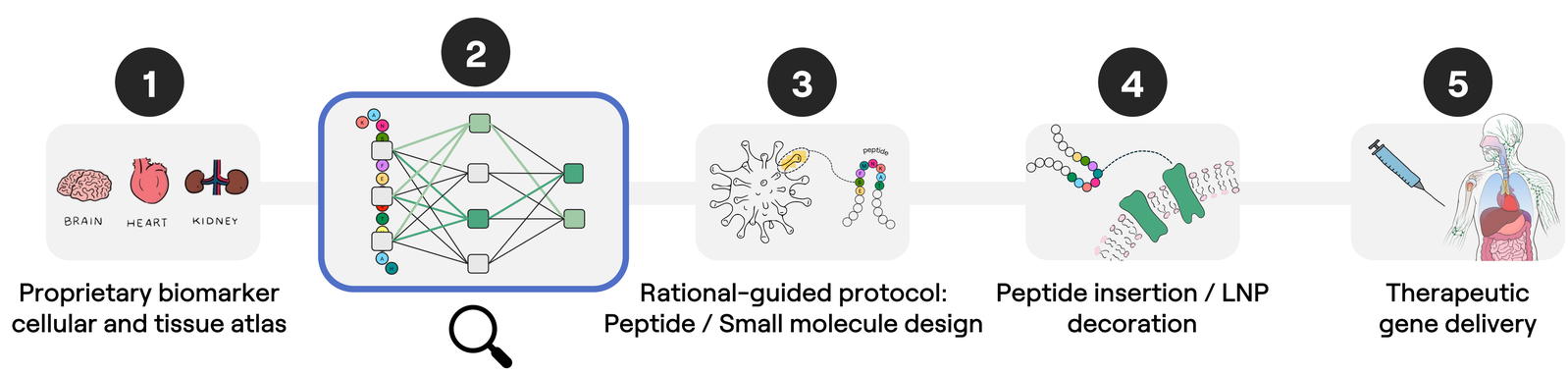
References

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

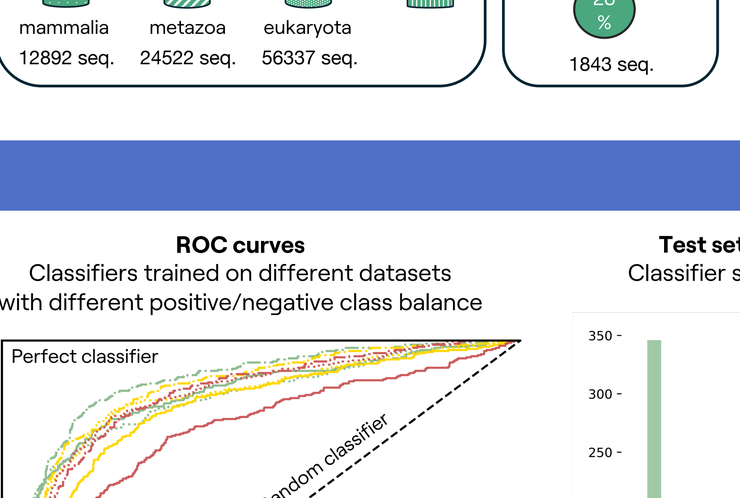
The ML model relies on very specific features computation

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

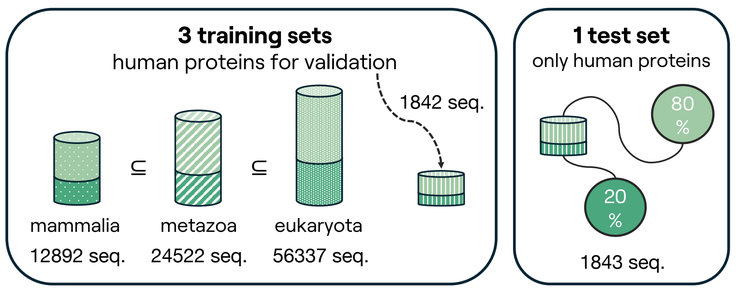
1. Query UniProt database with appropriate keywords and by taxonomy



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

3. Cluster sequences based on their identity — mmseqs2

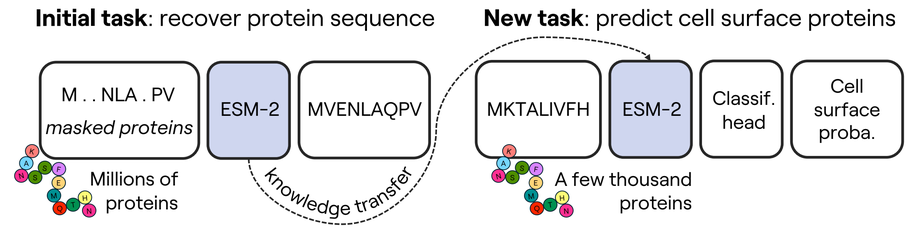
4. Aggregate sequences and custom balance between sets



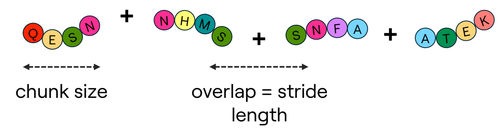
II. Model development

Goal: obtain a feature-free ML model

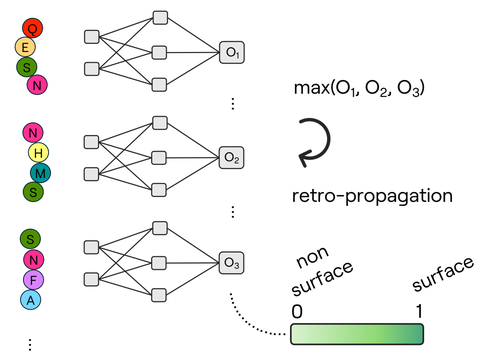
1. Protein language model selection and transfer learning



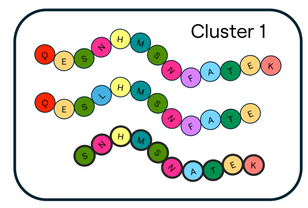
2. Protein split in chunks



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

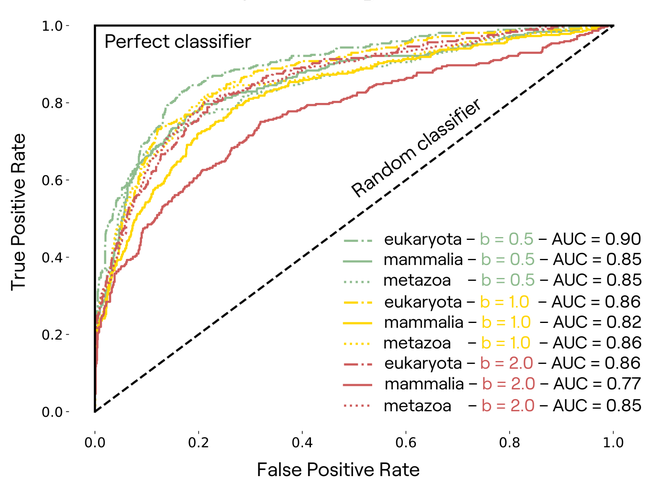
Accuracy
Overall correctness of the model on all proteins

Recall
How well the model finds the surface proteins

Precision
Accuracy of the positive predictions made by the model

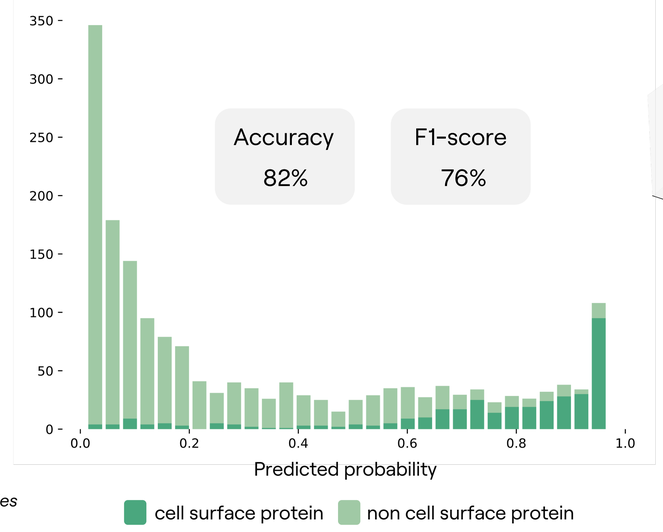
Results

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

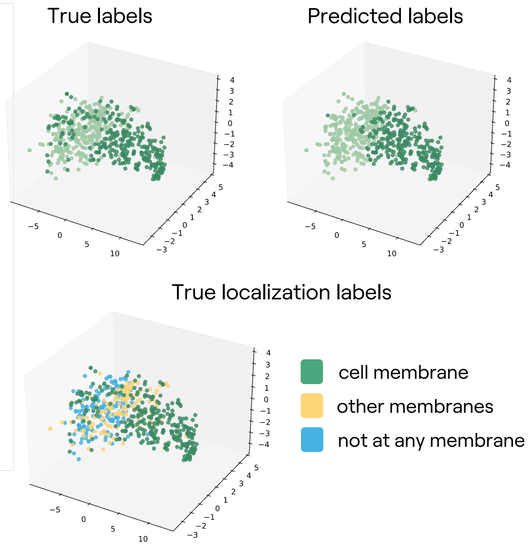


balance (b): 0.5 means there is twice more negative samples than positive samples
1.0 means there is as many negative samples as positive samples
2.0 means there is twice more positive samples than negative samples

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

Novel *in silico* Protocol for Small Molecule Discovery Enhancing Lipid Nanoparticle Targeting Specificity

B. Dafniet[1], C. Alliot[1], J. Maes[1], D. Del Bourgo[1], J. Cottineau[1], A. Attinger[2], S. Vigano[3], L. Marx[3], J. Quartier[3], C. Herkenne[3], P. Garrouste[3], D. Serillon[1]

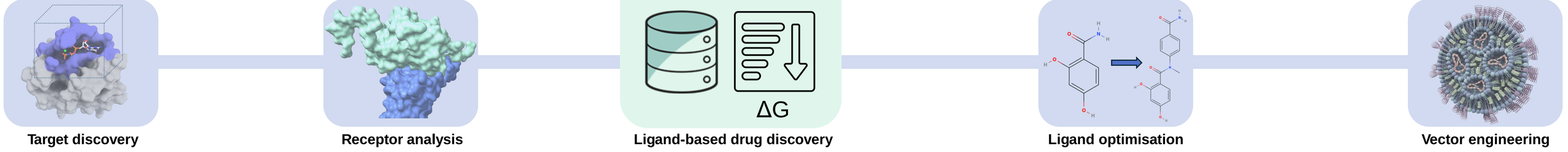
[1]WhiteLab Genomics, FUTURE4CARE, 8 rue Jean Antoine de Baif, 75013 Paris, France
[2]Debiopharm International, 5 Chemin Messidor, 1006 Lausanne, Suisse
[3]Debiopharm Research and Manufacturing, 146 Rue du Levant, 1920 Martigny, Suisse

About us

WhiteLab Genomics is a pioneering in-silico company leveraging Artificial Intelligence to accelerate discovery and mitigate risks in early-stage R&D 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.

Introduction

Lipid nanoparticles (LNPs) are a type of non-viral vectors able to deliver therapeutic payload directly to tissues or organs[1]. In silico approaches can be used to discover small molecules or peptides that will help functionalize the LNP to target receptors of interest, making it more specific and minimizing adverse effects[2-3]. The interest of our approach is twofold: (i) to provide a robust protocol minimizing *in silico* bias in a small molecule discovery context, (ii) to discover promising small molecules binding to our target of interest to enhance LNP specificity. In this study, the protocol was applied to an undisclosed target involved in a specific pathology.



Data Preparation

Enamine REAL database
6.75 billions compounds

- Satisfying biodistribution criteria namely Lipinski and Veber
- Commercially available

Enamine

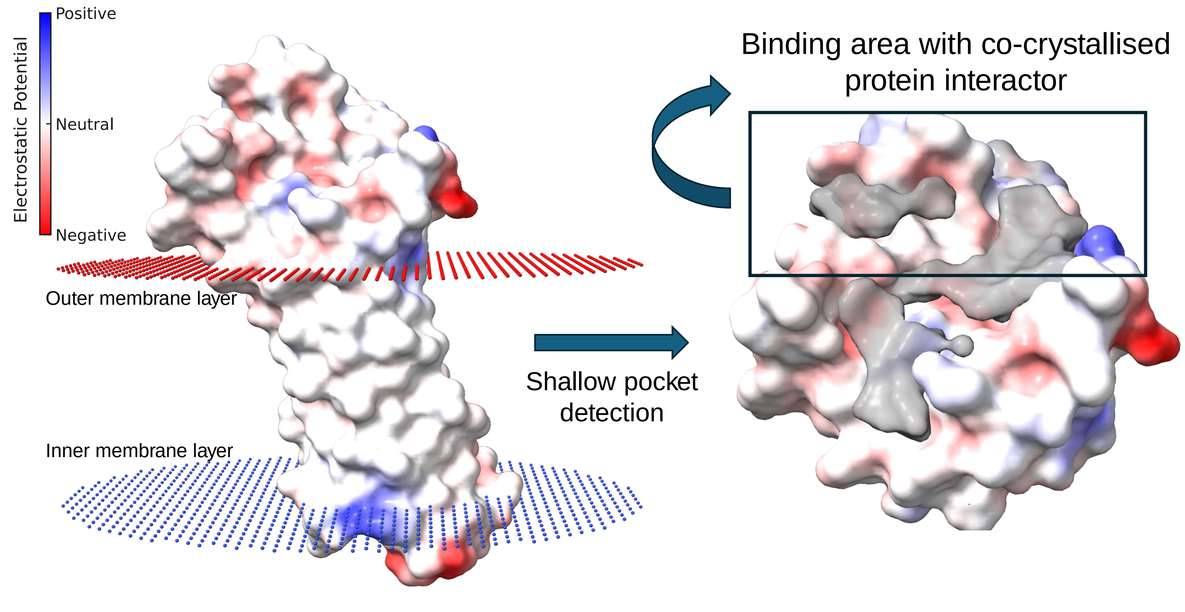
Subset

Diversity set
48.2 millions compounds

- Chemically diverse compounds
- Lack PAINS and toxicophores

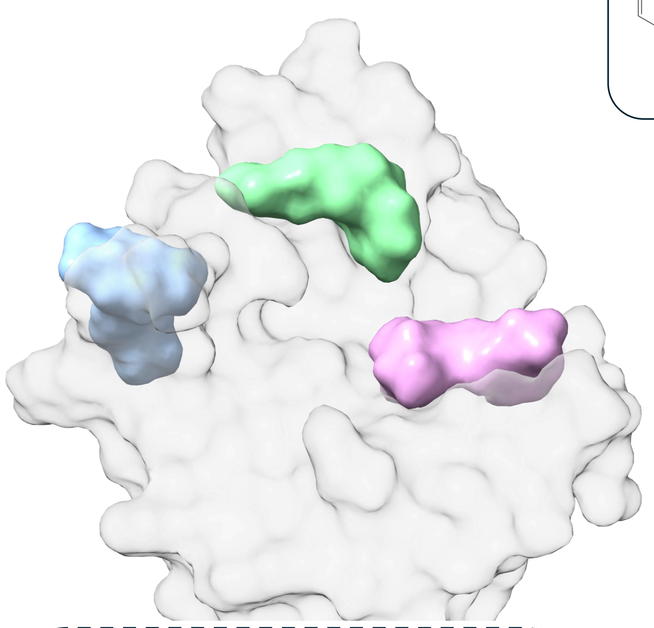
Use of Diversity set to efficiently explore the chemical space of the REAL database

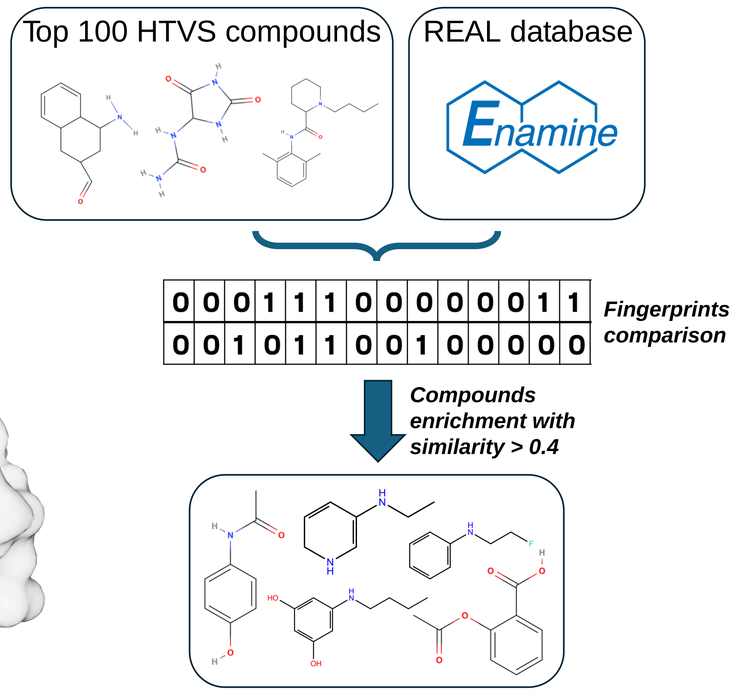
Target Preparation

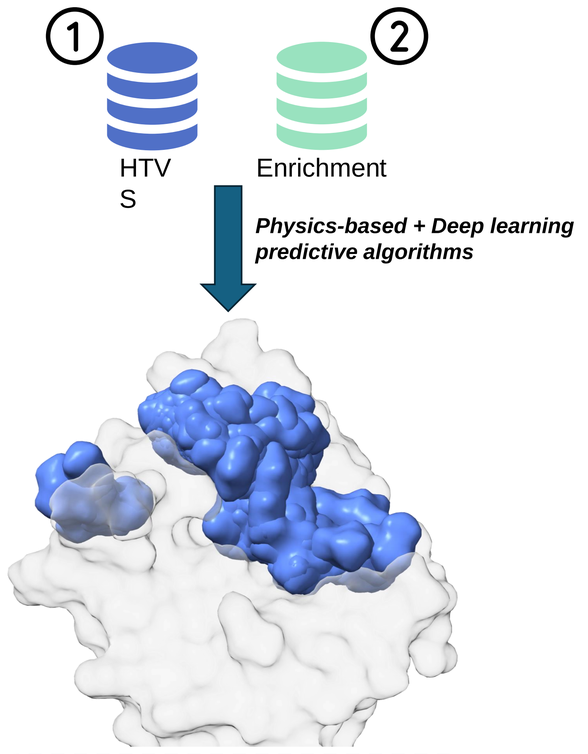


Hit Discovery

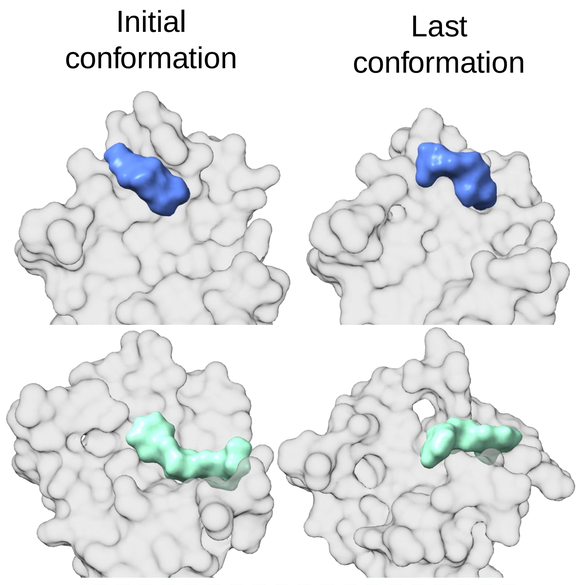
Enamine diversity dataset screened







Top 200 scored compounds for physics-based simulations over time



High-throughput virtual screening (HTVS)

- Three interesting binding sites sampled by the HTVS
- Energy score cut-off to keep predicted good binders

20,000 compounds ①

Tanimoto similarity approach for dataset enrichment

- Morgan fingerprints
- Adds chemical diversity while keeping key structural features

10,000 compounds for enrichment ②

Scoring

- Score to rank them mainly based on a binding mode consensus between algorithms
- 35% of enriched compounds in the top 200 score-wise

Binding stability assessment

- Simulation of binding < 100ns for the top ranking compounds
- Creation of a smaller subset for experimental testing

Conclusions & Perspectives

This study showed a protocol relying on physics-based and deep learning algorithms to create an in-house score maximizing binding consensus in a relevant pocket space. It allows us to generate libraries of predicted good binders for any relevant receptor involved in a pathology of interest. The addition of simulations over time after the consensus algorithms adds a layer of robustness to our protocol, improving the confidence of our predictions. The next steps would be to mimic the hydrophobic behavior of the LNP-small molecules addressing for more in-depth analysis.

Contact Us

whitelabgx.com | bd@whitelabgx.com

Bryan
bdafniet@whitelabgx.com

Camille
calliot@whitelabgx.com

Dylan
dserillon@whitelabgx.com

/bryan-dafniet
/camille-alliot
/dylan-serillon

References

[1] Lipid Nanoparticle Systems for Enabling Gene Therapies Cullis, Pieter R. et al. (2017) *Molecular Therapy*, 25(7), 1467 – 1475
[2] Biotinylated liposomes as potential carriers for the targeted delivery of doxorubicin in cancer therapy. Zhang P. *et al.* (2019). *Applied Biochemistry and Biotechnology*, 189(4), 1179–1194
[3] TAT peptide-conjugated nanoparticles incorporating pemetrexed and cisplatin for the treatment of glioma. Chen Y. *et al.* (2019) *Oncology Reports*, 42(3), 947–956

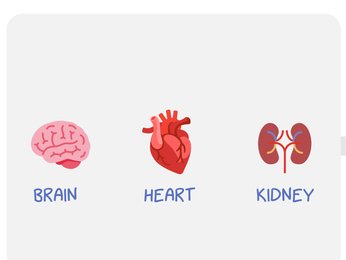
Integration of Protein Language Models and Reinforcement Learning for the Systematic Design of Phenotype-Specific AAV Capsids

O. de Felice[1], D. Serillon[1], D. Del Bourgo[1], J. Cottineau[1]

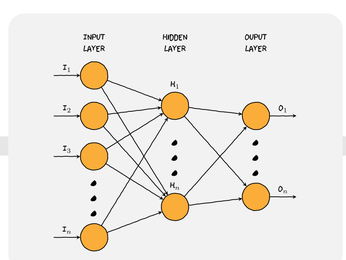
[1]WhiteLab Genomics, FUTURE4CARE, 8 rue Jean Antoine de Baif, 75013 Paris, France

Whitelab Genomics

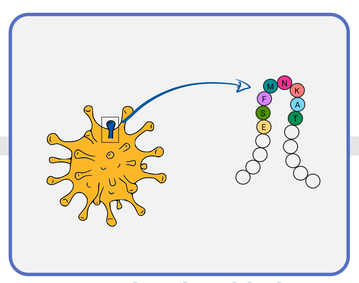
WhiteLab Genomics is a pioneering in-silico company leveraging Artificial Intelligence to accelerate discovery and mitigate risks in early-stage research and development pipelines exclusively within the field of genomic medicine. Founded in 2019, and backed by Y-Combinator, WhiteLab stands at the convergence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



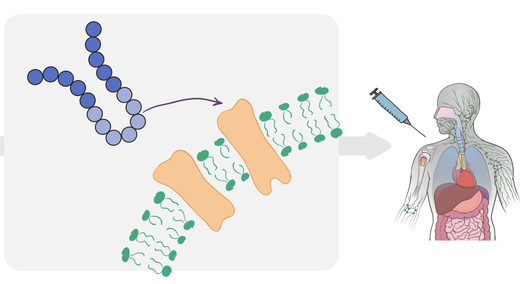
Prioprietary biomarker cellular and tissue atlas



Prioprietary ML models



Rational-guided protocol for peptide/small molecule design



Peptide insertion or LNP decoration

Introduction

Gene therapy effectiveness relies on AAV vectors engineered to deliver genes with precision. Phenotype-guided design of these vectors is crucial for enhancing tissue specificity, minimising immune responses, and optimising gene transduction, which are essential for targeted and safe therapeutic outcomes.

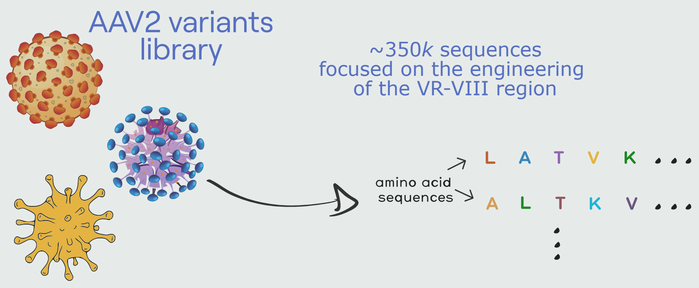
Purpose

The aim of the study was to build a **generative model** to generate capsid mutations with a given phenotype.

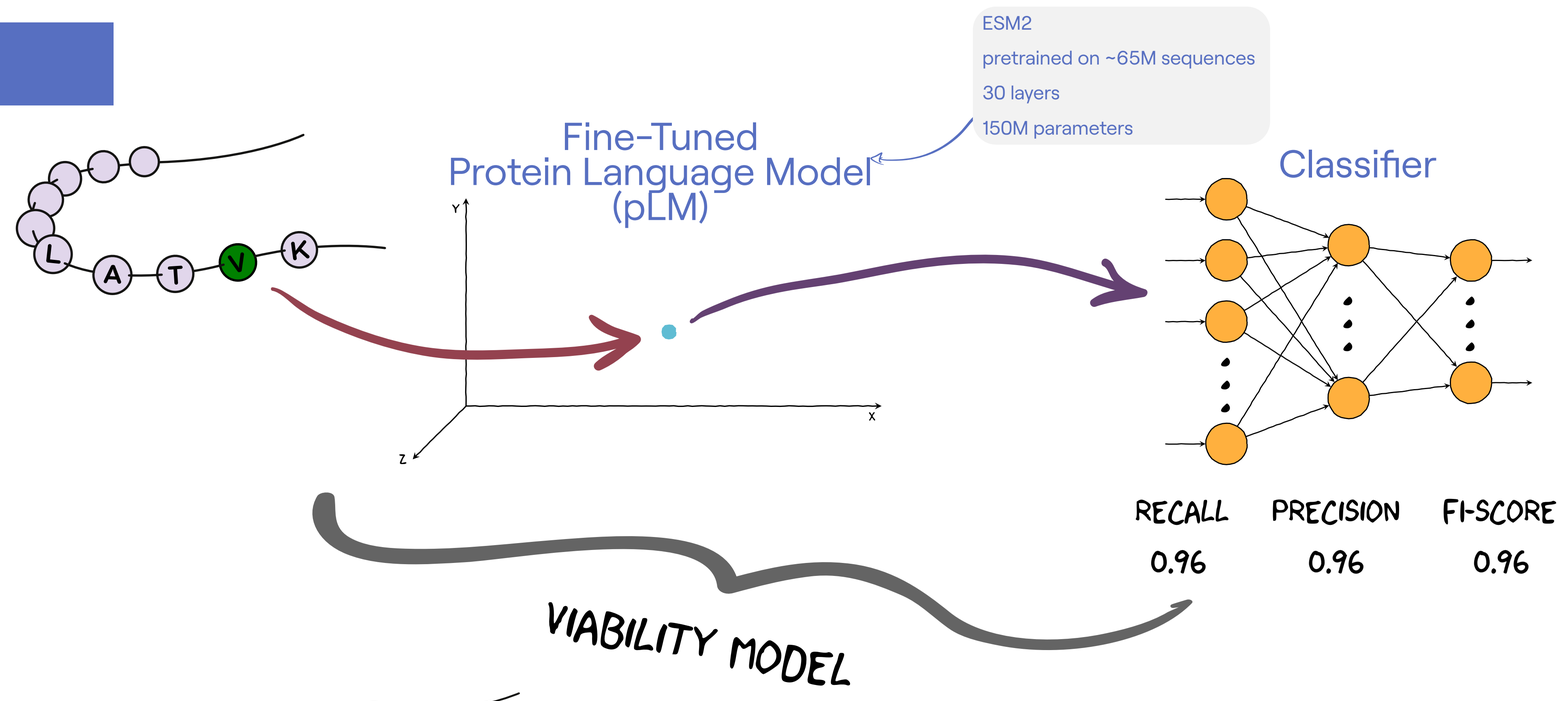
The model was initially trained on a dataset focused on AAV viability phenotypes, leveraging variable regions (VR-VIII) ability to tolerate mutations. The main criterion for success of the model was to predict and generate new mutations that maintain viability.

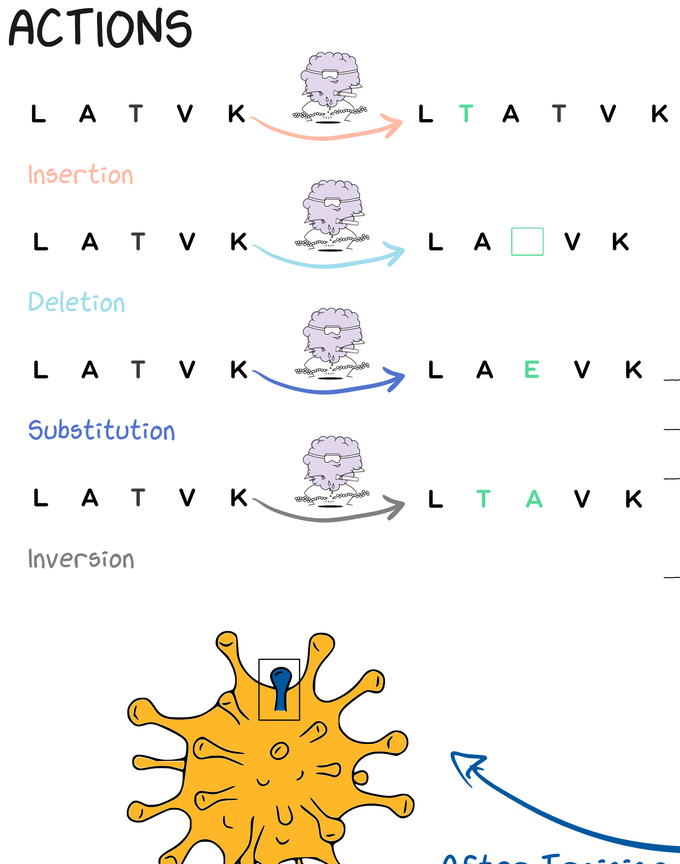
Methods

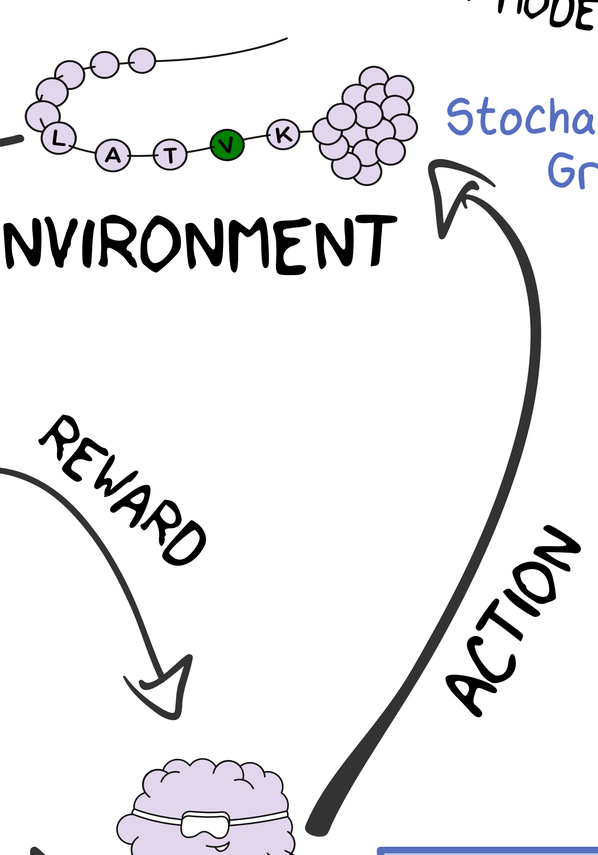
Data Input

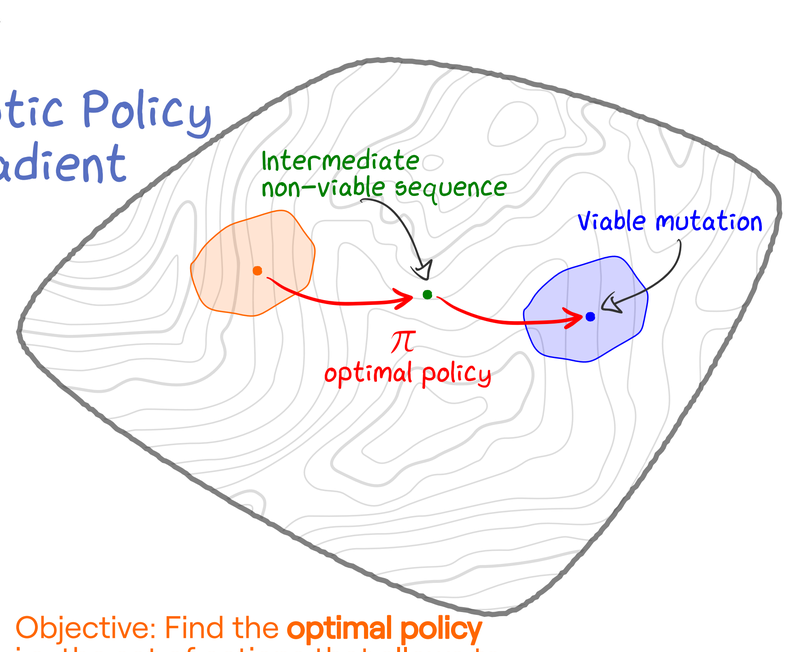


Permutations in the constitutive variable region (VR-VIII)



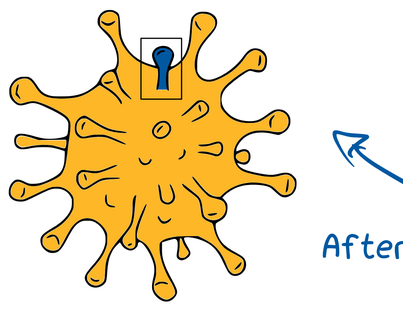




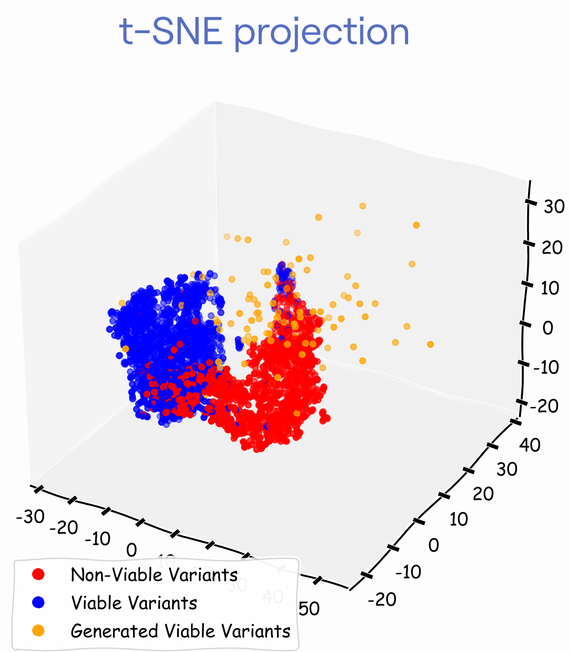


Objective: Find the **optimal policy** i.e. the set of actions that allows to generate viable variants.

Each sequence is evaluated by the model, which rewards or penalises the agent providing information about the current state.



Results



In the results, we observe how the generated sequences, conditioned on specific phenotypic targets, explore previously uncharted regions of the sequence space. This is achieved by integrating a protein Language Model (pLM) with a Reinforcement Learning (RL) framework, where the RL agent optimises sequence generation towards a desired phenotype. The pLM provides a probabilistic model of the sequence structure, capturing both local and global dependencies, while the RL agent iteratively updates the model to maximise the phenotypic reward. The exploration of regions outside the training data indicates that the model is capable of generalising and proposing novel sequence variations that may lead to new phenotypic expressions, thus showcasing its potential for de novo design. This divergence from the training data distribution is a clear indicator of the model's capacity to expand the biological diversity of the sequences, a critical aspect for uncovering new, functionally relevant mutations.

Conclusions

We developed a RL-based generation approach able to generate protein sequence insertions based on a given capsid phenotype. The *exploitation/exploration* dilemma is addressed, at the training level, by setting to 5 the number of allowed actions before evaluating the phenotype of the generated variant.

Further research will explore extending this approach to other phenotypes.
Collaborations with wet lab facilities are essential to validate the designed AAV capsids.

The main advantage of such approach is that it allows the generation of penotype-specific variants, resulting in an AI-driven procedure, instead testing for a huge number of random mutations.

Contacts

Whitelab Genomics

whitelabgx.com

bd@whitelabgx.com

Author

odfelice@whitelabgx.com

/oscar-de-felice

References

1. Some of the images are kindly provided by the SIB, Swiss Institute of Bioinformatics

2. Drew H Bryant et al. Deep diversification of an AAV capsid protein by machine learning, *2021*

3. Fatma-Elzahraa Eid et al. Systematic multi-trait AAV capsid engineering for efficient gene delivery, *2024*

4. Havlik, L.P. et al. Coevolution of Adeno-associated Virus Capsid Antigenicity and Tropism through a Structure-Guided Approach, *2020*

Optimizing AAV Vectors for Precision Gene Therapy: A Rational Design Approach

C Colas, C Alliot, B Dafniet, J Maes, D del Bourgo, J Cottineau, D Serillon

WhiteLab Genomics, 4 rue Pierre Fontaine, 91000 EVRY, / FUTURE4CARE, 8 rue Jean Antoine de Baif, 75013 Paris, France.

About us

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.

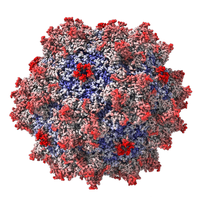
- **Accelerate** R&D by several years 35-40% faster
- **Increase** Productivity of R&D resources Multiplied by 4
- **Reduce** Unnecessary experiment costs 25-30%
- **Improve** Pre-clinical success rate 20-30%

Introduction

Genomic medicines represent groundbreaking therapeutic approaches for treating diseases with limited therapeutic options by delivering genetic material directly to patient cells. Various vectors serve as delivery vehicles, including viral and non-viral vectors such as lipid nanoparticles (LNPs). Adeno-associated viruses (AAV) are particularly favored in genomic medicines due to their extensive tissue biodistribution, which allows them to target a broad range of diseases. However, this broad tropism can result in a lack of tissue specificity. Therefore, optimizing capsid structure is crucial for enhancing AAV specificity, selectivity, manufacturing efficiency, and reducing immunogenicity. Current AAV capsid engineering efforts aim to improve the vector delivery efficiency to target cells.

Objectives

- Enhance AAV capsid specificity
- Improve delivery efficiency to target cells
- Maximize therapeutic potential



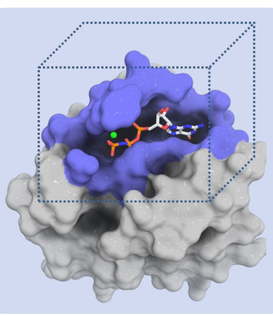
3D representation of an AAV capsid

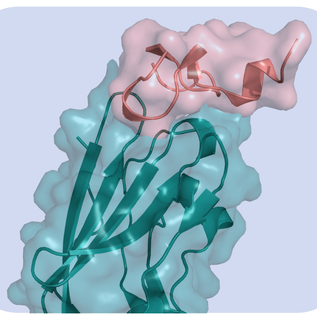
Strategy

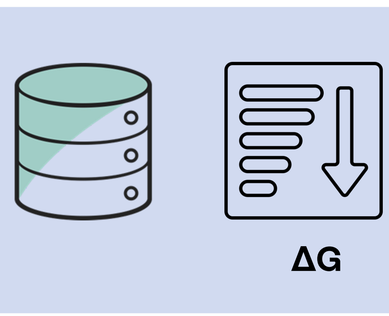
- Design vectors that enhance capsid specificity for target receptors
- Leverage variable regions' (e.g. VR-VIII) inherent ability to tolerate mutations and peptide insertions
- Develop a guided rational approach for peptide design

Protocol

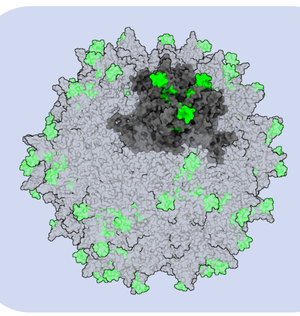
Target discovery → **Receptor analysis** → **Ligand-based drug discovery** → **Ligand optimization** → **Vector engineering**





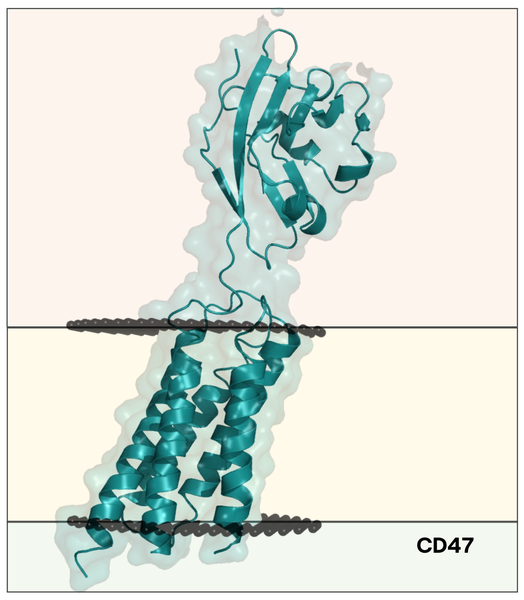


IXSXXXG
XXSRXXG
IXSXXXA
XXSAXXP
RXAXCXP



- Identifying the receptor of interest based on its tissue expression and accessibility at the cell surface
- Assessing the structural feasibility: qualitative experimental structure available, known binders, etc
- Analysis is crucial for understanding the receptor's interactions and functionality under physiological conditions

1. Annotating the target receptor's structure

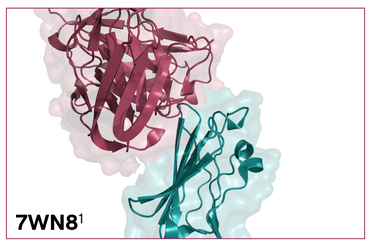


Extracellular domain

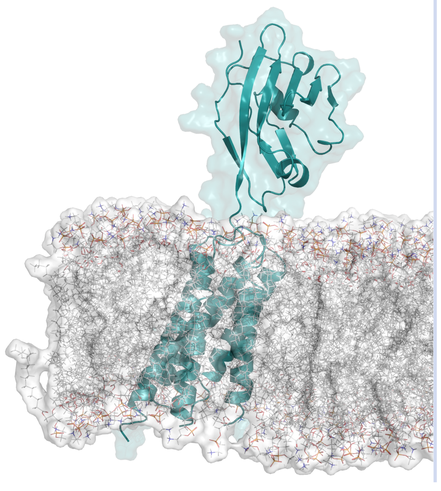
Transmembrane domain

Cytoplasmic domain

2. Analyzing the binding site



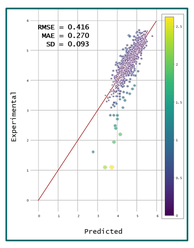
3. Assessing the behavior in a solvated environment

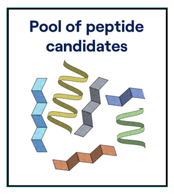


- Predicting the affinity of peptides derived from an enriched large database using in house predictive machine learning models
- Docking and filtering of predicted good binders using various software

Interaction site analysis

AI-based virtual screening





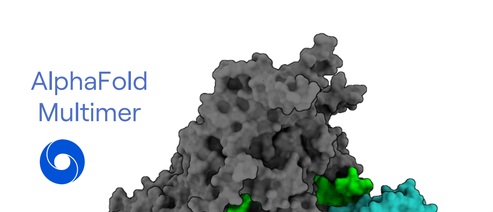
Peptide identification

Large peptide database docking — Schrödinger

First selection of peptides — Ranking and affinity prediction

Interaction-based selection — Smaller optimized peptide database

Final list of peptides

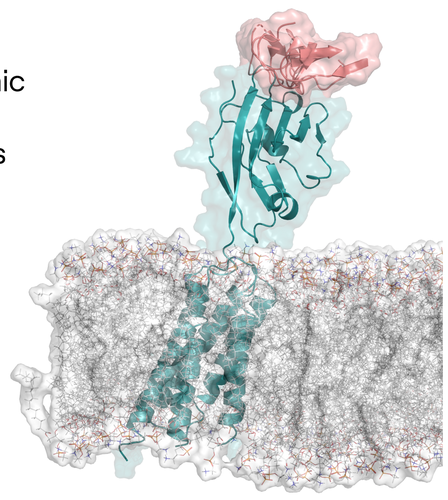


Macrocyclic peptides generation

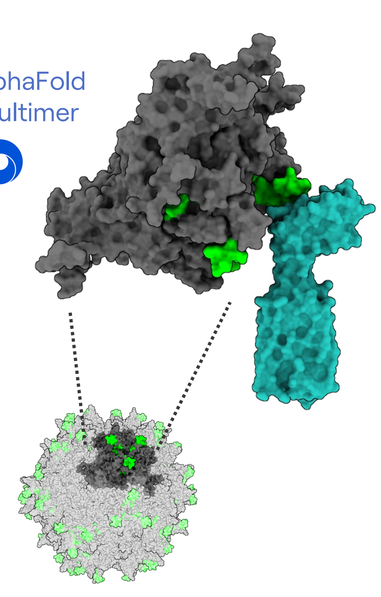
Macrocyclic peptides docking

- Converting the selected peptides into macrocycles
- Sampling multiple conformations
- Restraining peptide conformations to mimic the AAV-inserted peptide behavior

- Performing molecular dynamic simulations of the receptor-selected peptides complexes
- Evaluating the stability of peptides in their binding site
- Assessing the consistency of key interactions
- Detecting potential cryptic pockets, interactions, conformations



AlphaFold Multimer



- Inserting the selected peptides in the AAV trimer
- Evaluating the maintenance of the essential interactions with various methods
- Predicting the viability of the modified capsid with in-house patented models

AAV capsid binding prediction to the receptor of interest

Conclusion

- Impact: The workflow aims at enhancing AAV vector specificity, improving the precision and efficacy of gene therapy.
- Future Work: Further testing and optimization of identified peptide candidates for clinical applications are ongoing.

References

[1] Li, Y. et. al " A pH-dependent anti-CD47 antibody that selectively targets solid tumors and improves therapeutic efficacy and safety. "*J Hematol Oncol* **16:** 2-2 (2023).

[2] Fenalti, G. et al. " Structure of the human marker of self 5-transmembrane receptor CD47 " *Nat Commun* **12**: 5218-5218 (2021)

Contact us

Whitelabgx.com I bd@whitelabgx.com

ccolas@whitelabgx.com
dserillon@whitelabgx.com

/claire-c-colas /camille-alliot /bryan-dafniet /juliette-maes /dylan-serillon