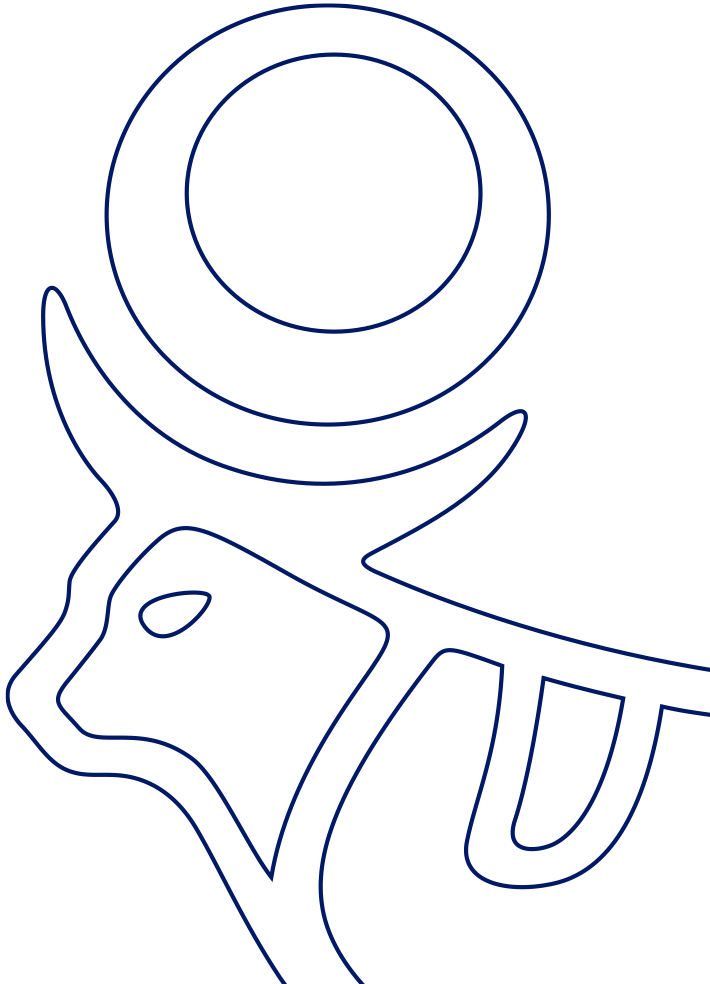
Exploring Gene Embeddings for Biological Analysis

CMP proposal 2026

M. Lisandra Zepeda Mendoza

Feb 2026



Novo Nordisk at a **glance**

Novo Nordisk is a leading global healthcare company, founded in 1923 and headquartered in Denmark.

Our purpose is to drive change to defeat serious chronic diseases, built upon our heritage in diabetes.

We do so by pioneering scientific breakthroughs, expanding access to our medicines, and working to prevent and ultimately cure disease.

Supplier of nearly
50%
of the world's insulin

Net sales
232.3
billion DKK

Affiliates in
80
countries

About
64,319
employees

Total tax contribution
51
billion DKK

R&D centres in China, Denmark, India, UK and US

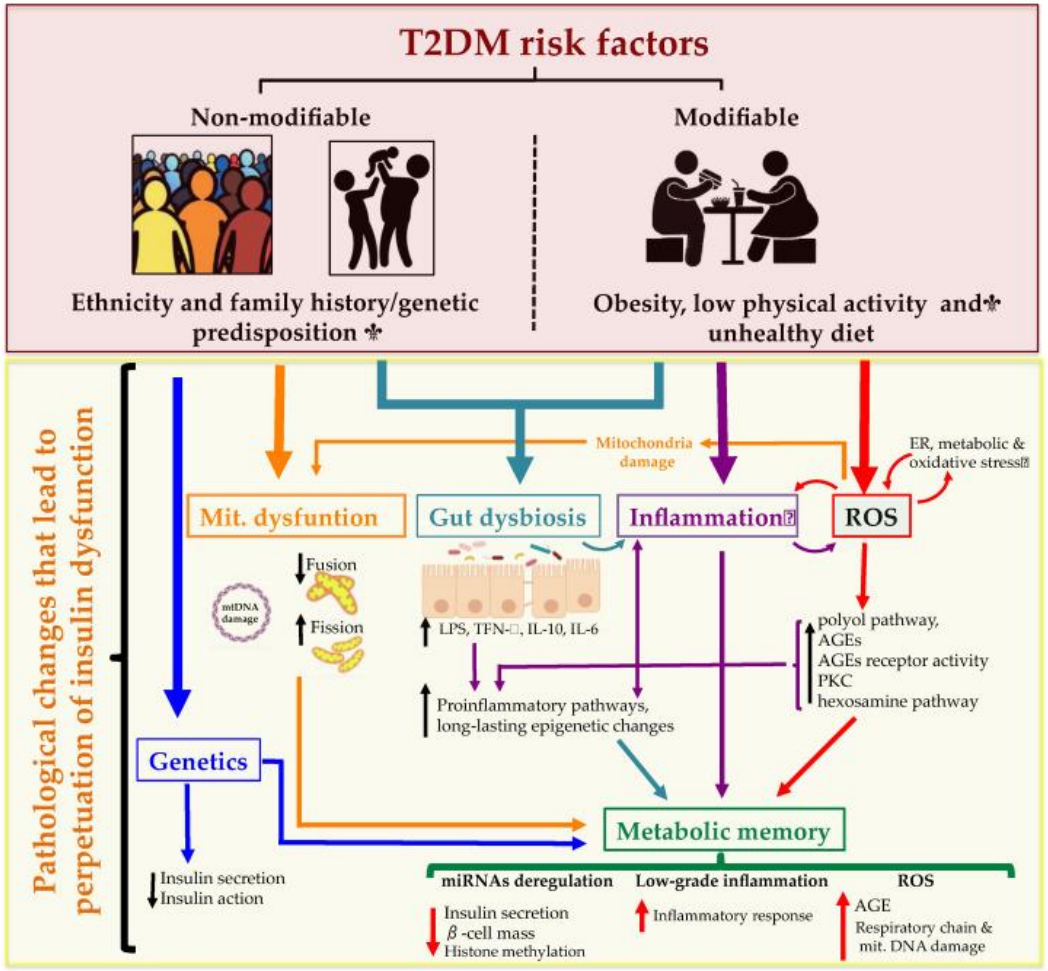
Strategic production sites in Denmark, Brazil, China, France and US

Globally, serving
41.6
million people living with diabetes and obesity

A top ten
pharma company measured by market value in 2025

Novo Nordisk Research Centre Oxford

Why is finding new drugs for CMDs so hard?



Picture taken from DOI:10.3390/ijms21176275

General R&ED steps for drug target discovery

1. Integrate various datasets with various approaches and identify drug target candidates.

2. Further evaluate the target in the lab:
   → Hypothesise target's mechanism of action (MoA)
   → Identify the tissue(s) of action (ToA)

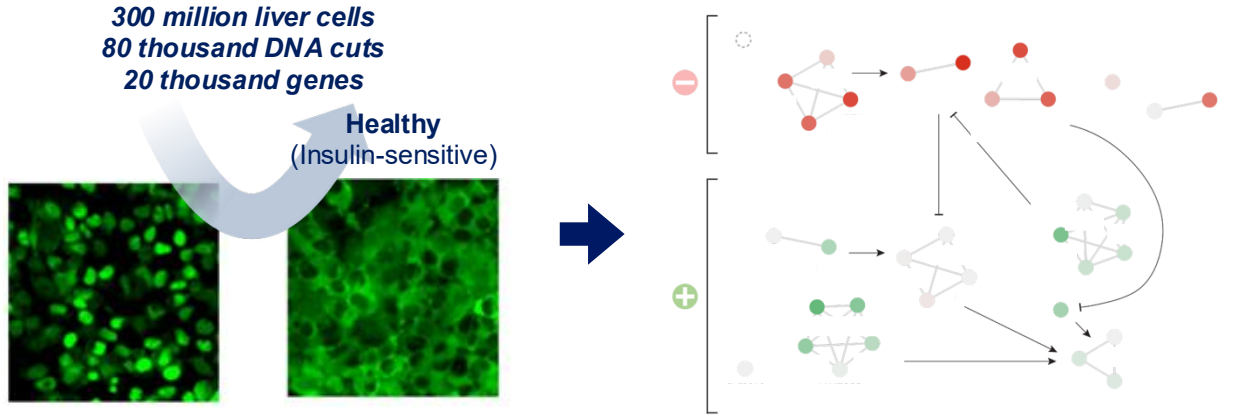
The challenge for this proposal

Gene embeddings from various data sources can harbour all relevant information from a gene's context.

- **How to make use of such gene embeddings?**
- **How to choose the most relevant embeddings?**

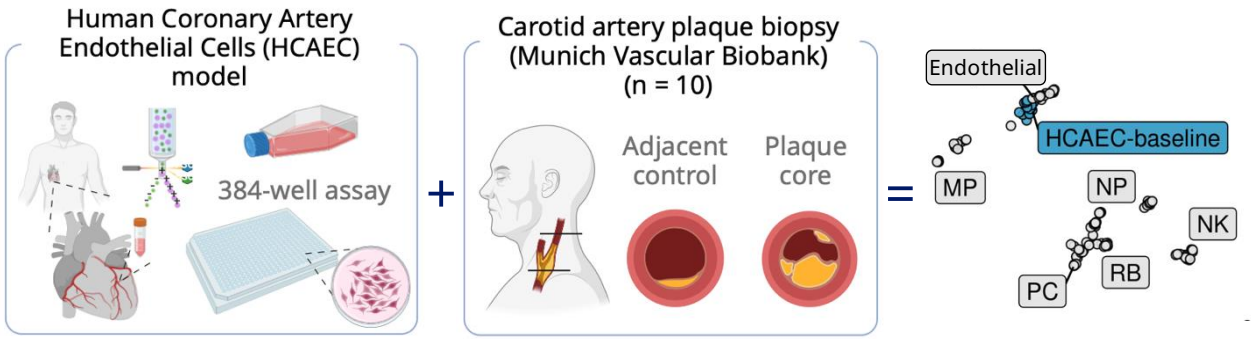
What can you do with Gene embeddings?

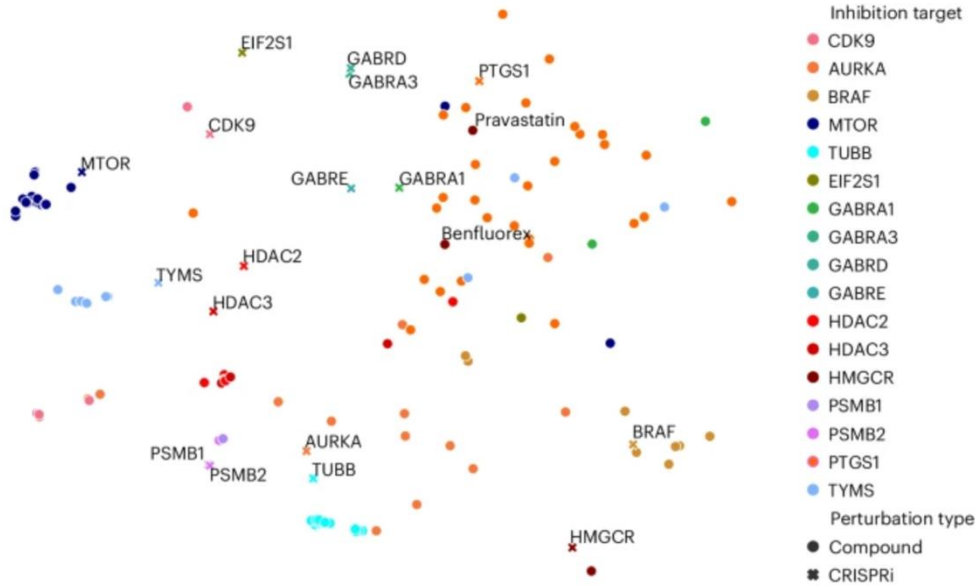
Inhouse applications



In-vitro model/assay evaluation

Cell-type resolved mRNA atlases & DRUGseq bring context to our in-vitro assays





Learning compound-CRISPR perturbation representations

Figure from the large perturbation model (LPM) study*, inhouse we have had similar applications

** https://www.nature.com/articles/s43588-025-00870-1*

What can you do with Gene embeddings?

Public implementations and resources

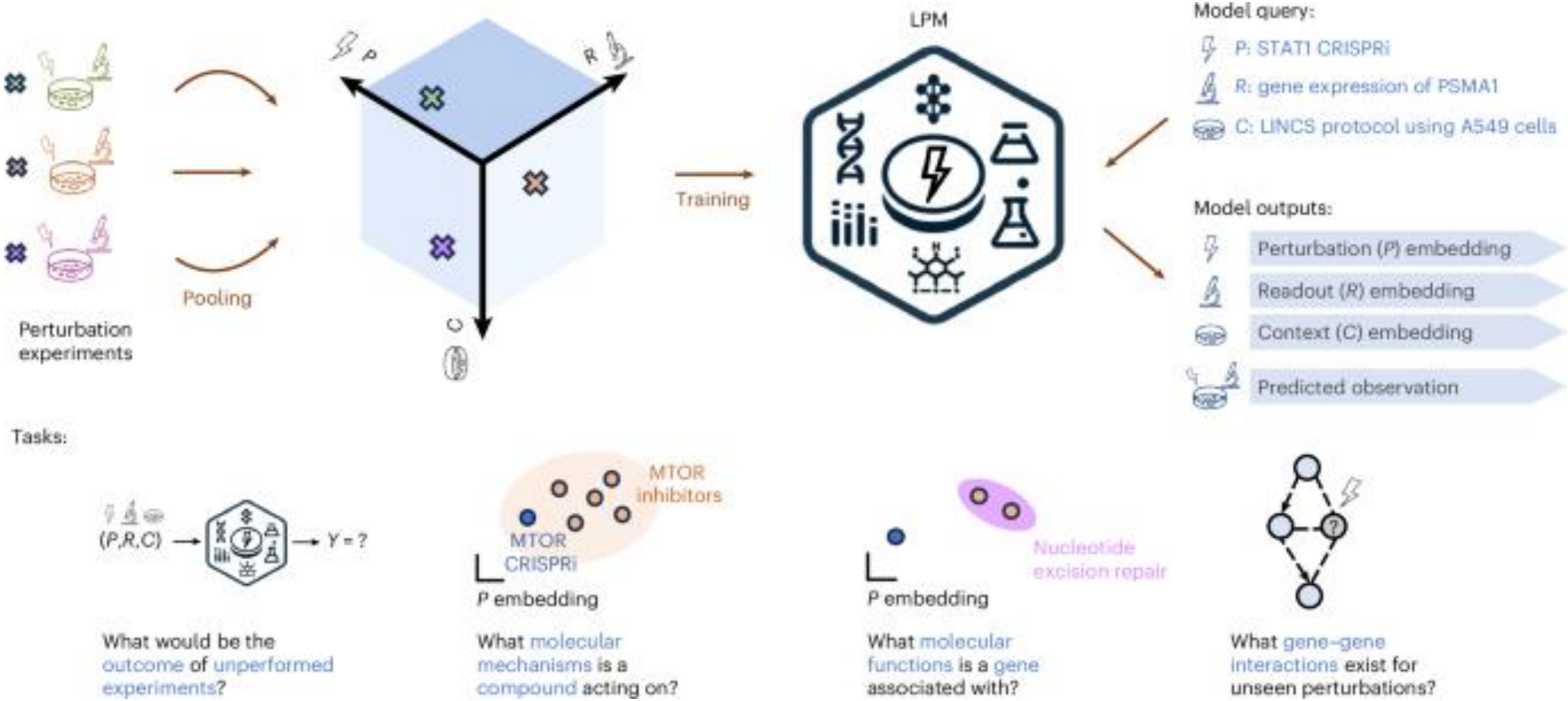
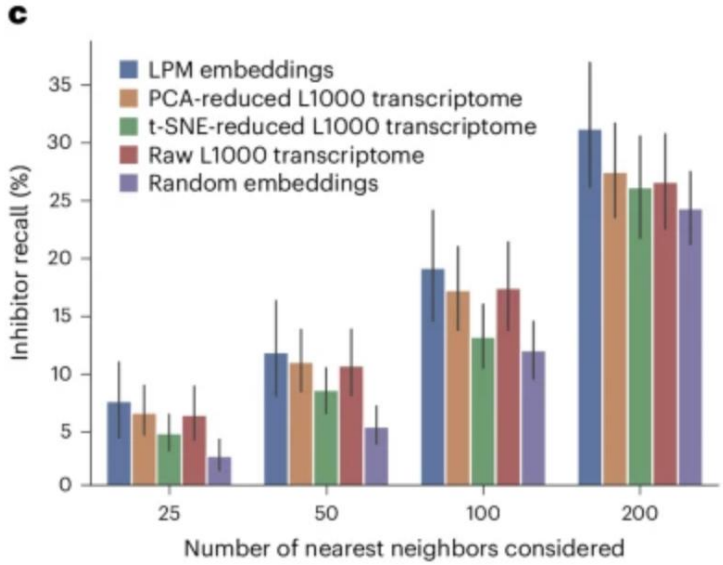


Image from https://www.nature.com/articles/s43588-025-00870-1

Main task of this proposal

Inhouse implementation of the Large Perturbation Model (LPM) to shed light onto our prioritized Modes of Action and benchmarking to other existing approaches



Biological relationships captured by gene embeddings
Each point represents a gene perturbation color-coded by the gene's molecular function.

The project consists of: 1) Implementing LPM inhouse. 2) Comparing its embeddings on the LINCS dataset to those derived from sequence, gene expression, ontology relationships, network interactions, and literature (LLMs).

The reasoning of the goal:
From experience we know that the performance of a FM depends on the dataset. The project is to implement this state-of-the-art approach and compare their biological relevance to other existing approaches, to identify the one that best fits our inhouse data.

https://www.nature.com/articles/s43588-025-00870-1


Qs&Cs feel free to e-mail
vmnz@novonordisk.com


novo nordisk®