Computational Molecular Biology and Bioinformatics

PDGrapher

Malay Bhattacharyya

Associate Professor

Machine Intelligence Unit
Indian Statistical Institute, Kolkata

September, 2025

1 Introduction

2 The PDGrapher Method

3 References

What is PDGrapher?

`Perturbagens` are chemical, biological, or environmental substances (e.g., drugs) which may disrupt the normal state of a biological system. They provide important information about the operation of pathways and networks within a cell. The perturbagens are used as therapeutic targets.

PDGrapher is a causally inspired graph neural network (GNN) model that predicts combinatorial perturbagens capable of reversing disease phenotypes [1]. By embedding disease cell states into networks, learning a latent representation of these states, and identifying optimal combinatorial perturbations, PDGrapher solves the inverse problem of directly predicting the perturbagens required to achieve a desired response, in contrast to methods that learn how perturbations alter phenotypes.

Problem formulation

Let $\mathtt{M} = < \mathbf{E}, \mathbf{V}, \mathcal{F}, P(\mathbf{E}) >$ be a structural causal model (SCM) associated with causal graph G, where $\mathbf{E}$ is a set of exogenous variables affecting the system, $\mathbf{V}$ is a set of system variables, $\mathcal{F}$ denote structural equations encoding causal relations between variables, and $P(\mathbf{E})$ is a probability distribution over exogenous variables.

Let $\mathcal{T} = \{\mathtt{T}_1, \ldots, \mathtt{T}_m\}$ be a dataset of paired healthy and diseased samples (namely disease intervention data), where each element $\mathtt{T}_i$ is a triplet $< \mathbf{v}^h, \mathbf{U}, \mathbf{v}^d >$ with $\mathbf{v}^h \in [0,1]^N$ being normalized gene expression values of healthy cell line (variable states before perturbation), $\mathbf{U}$ being the disease-causing perturbed variable (gene) set in $\mathbf{V}$, and $\mathbf{v}^d \in [0,1]^N$ being gene expression values of diseased cell line (variable states after perturbation).

Problem formulation

The goal is to find, for each sample $\mathtt{T}_i = < \mathbf{v}^h, \mathbf{U}, \mathbf{v}^d >$, the variable set **U**' with the highest likelihood of shifting variable states from diseased $\mathbf{v}^d$ to healthy $\mathbf{v}^h$ state. To increase generality, we refer to the desired variable states as *treated* ($\mathbf{v}^t$). The goal can then be expressed, under the assumption of no unobserved confounders, as follows:

$$\arg\max_{\mathbf{U}'} P^{G^U}(\mathbf{V} = \mathbf{v}^t | \mathbf{U}),$$

where P^{G^U} denotes the probability on the graph G mutilated by perturbations in variables within **U**.

Note: In causal graph theory, a `mutilated` graph is typically a Directed Acyclic Graph (DAG) from which edges or nodes are removed to represent an intervention or the removal of information in a causal system.

Problem reformulation with representation learning

To address the limitations of real-world setting of noisy and incomplete causal graphs, a reformulation of this problem is done with representation learning to approximate the queries of interest.

Suppose $G = (\mathcal{V}, \mathcal{E})$ denotes a proxy causal graph with $|\mathcal{V}| = n$ nodes and $|\mathcal{E}|$ edges, which contains partial information on causal relationships between nodes in $\mathcal{V}$ and some noisy relationships. Let $\mathcal{T} = \{\mathrm{T}_1, \ldots, \mathrm{T}_m\}$ be a dataset with each element T_i represented as a triplet $< \mathbf{x}^h, \mathrm{U}, \mathbf{x}^d >$ with $\mathbf{x}^h \in [0,1]^N$ being the node states (attributes) of a healthy cell sample (before perturbation), $\mathcal{U}$ being the disease-causing perturbed nodes in $\mathcal{V}$, and $\mathbf{x}^d \in [0,1]^N$ being the node states (attributes) of a diseased cell sample (after perturbation).

Problem reformulation with representation learning

By referring to the desired variable states as treated ($\mathbf{x}^t$), the goal becomes learning the following function:

$$f : G^{\mathcal{U}'}, \mathbf{x}^d, \mathbf{x}^t \rightarrow \arg \max_{\mathcal{U}, P^{G^{\mathcal{U}'}}} (\mathbf{x} = \mathbf{x}^t | \mathbf{x}^d, \mathcal{U}).$$

Thus, given the graph $G^{\mathcal{U}}$, the diseased node states $\mathbf{x}^d$ and treated node states $\mathbf{x}^t$, we plan to predict the combinatorial set of nodes $\mathcal{U}$ that if perturbed have the highest chance of shifting the node states to the treated state $\mathbf{x}^t$. Note that $P^{G^{\mathcal{U}'}}$ represents probabilities over graph $G^{\mathcal{U}}$ mutilated upon perturbations in nodes in $\mathcal{U}'$.

Problem reformulation as a graph prediction task

Given a graph $G = (\mathcal{V}, \mathcal{E})$, with paired sets of node attributes $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m\}$ and node labels $\mathcal{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m\}$, where each $\mathbf{Y}_i = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, with $\mathbf{y}_i \in [0, 1]$, we aim at training a neural message-passing architecture that given node attributes $\mathbf{X}_i$ predicts the corresponding node labels $\mathbf{Y}_i$. Note that this reformulated problem is not principally the same as that of conventional graph prediction tasks (e.g., node classification).

Given that the prediction for each variable is dependent only on its parents in a graph, we can address the problem on Graph Neural Networks (GNNs).

Overview of PDGrapher

PDGrapher is an approach for combinatorial prediction of therapeutic targets composed of two modules.

1. **Perturbagen discovery:** The function f_p searches the space of potential gene sets to predict a suitable candidate $\mathcal{U}'$.

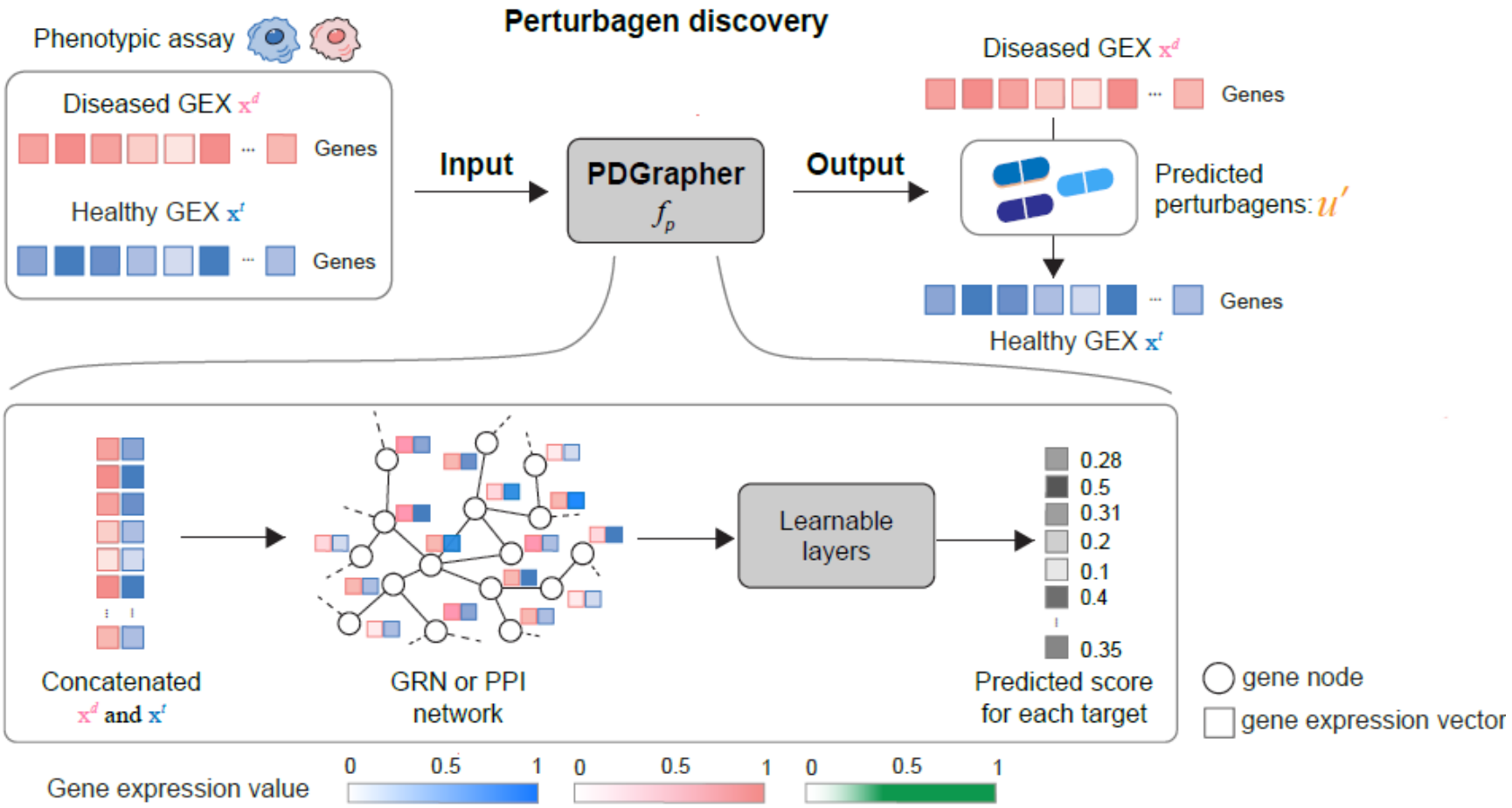
$$\mathbf{x}^d, \mathbf{x}^t \rightarrow \hat{\mathcal{U}}'$$

2. **Response prediction:** The function f_r checks the effectiveness of the predicted set $\mathcal{U}'$, i.e., how intervening on variables in $\mathcal{U}'$ is to shift node states to the desired treated state $\mathbf{x}^t$.

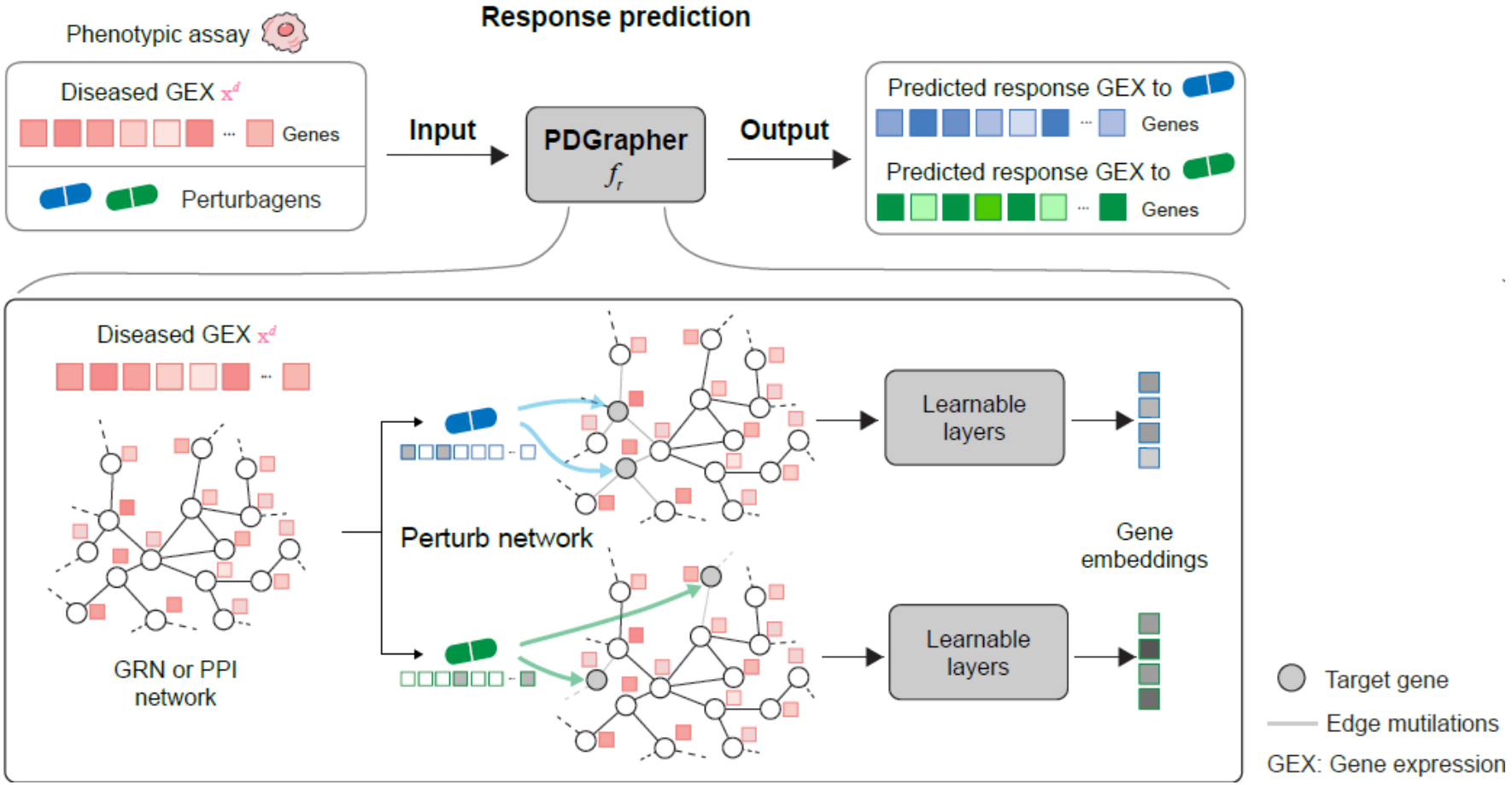
$$\mathbf{x}^d, \hat{\mathcal{U}}' \rightarrow \hat{\mathbf{x}}^t$$

We optimize the response prediction module using cross-entropy loss.

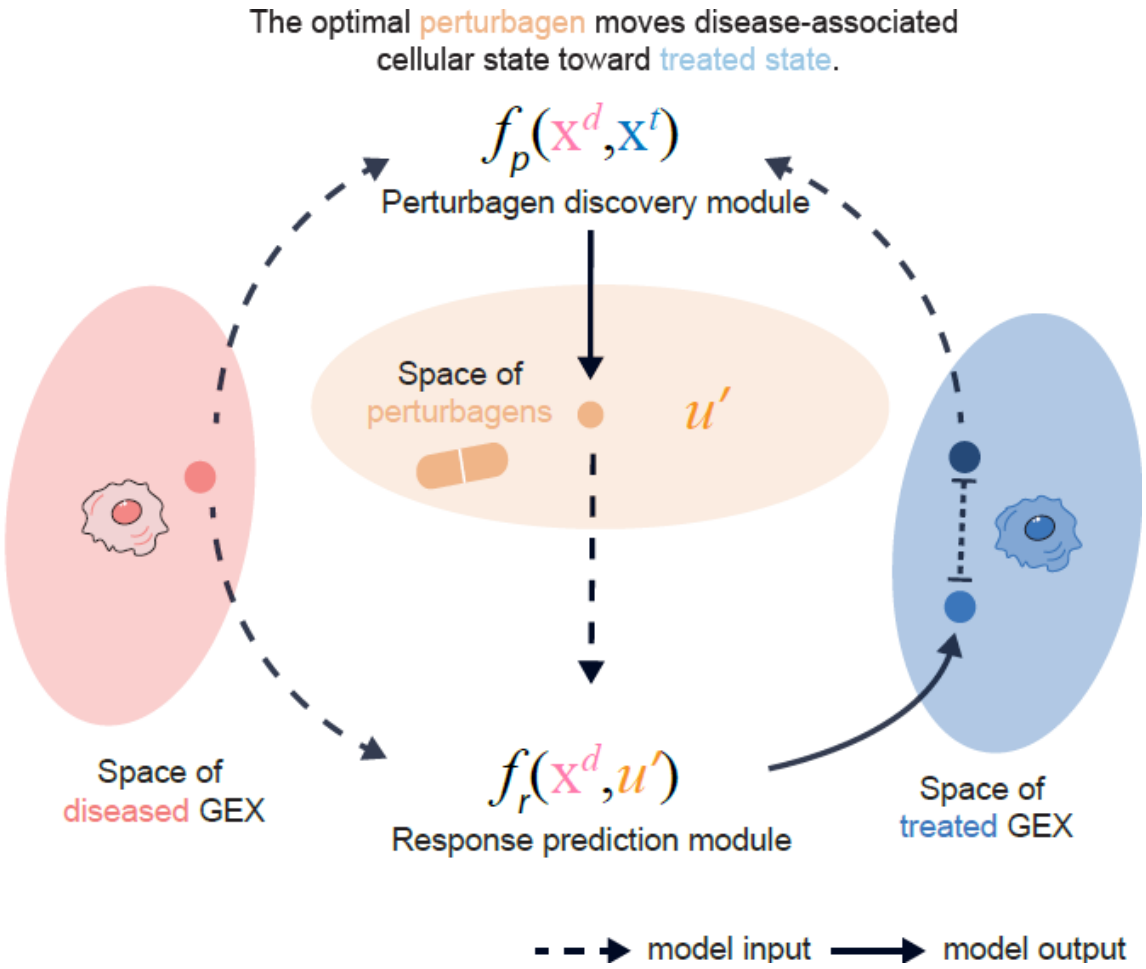
Perturbagen discovery



Response prediction



Optimization



Optimization

The perturbagen discovery module f_p is optimized using a cycle loss, ensuring that the response to the predicted intervention set $\mathcal{U}'$ closely matches the desired treated state. In addition, we provide a supervisory signal for predicting $\mathcal{U}'$ in the form of cross-entropy loss (CE). So, the total loss is defined as (with f_r frozen):

$$\mathcal{L}_{f_p} = \mathrm{CE}(\mathbf{x}^t, f_r(\mathbf{x}^d, f_p(\mathbf{x}^d, \mathbf{x}^t)) + \mathrm{CE}(\mathcal{U}', f_p(\mathbf{x}^d, \mathbf{x}^t)).$$

The response prediction module f_r is optimized using cross-entropy (CE) loss on known triplets of disease intervention $< \mathbf{x}^h, \mathcal{U}, \mathbf{x}^d >$ and treatment intervention $< \mathbf{x}^d, \mathcal{U}', \mathbf{x}^t >$ defined as:

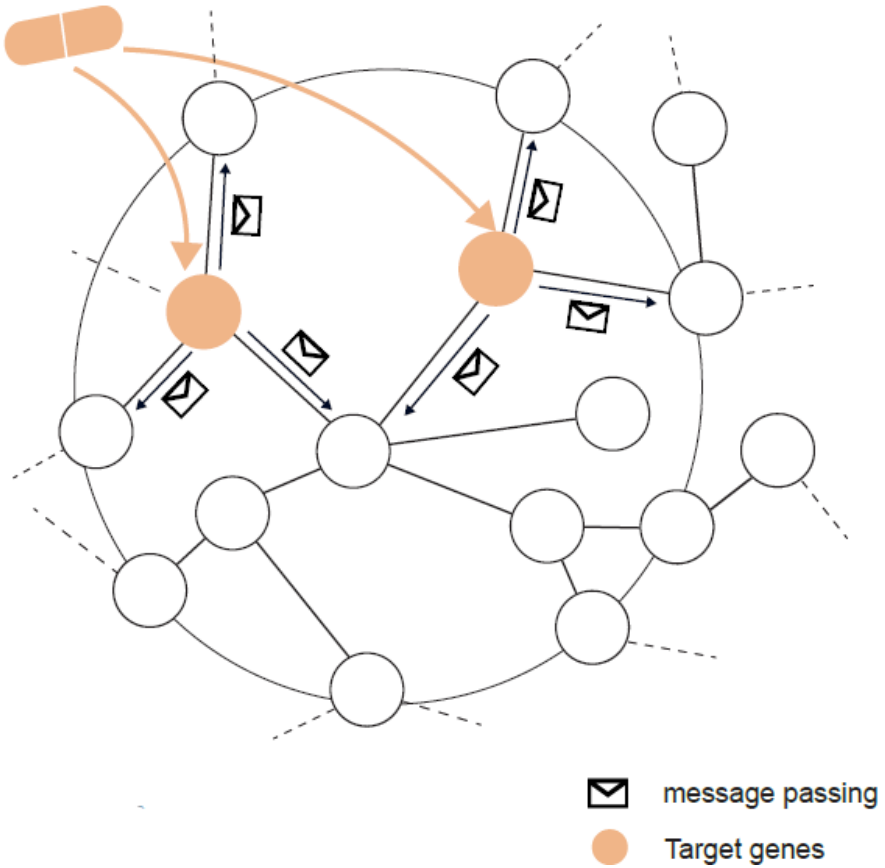
$$\mathcal{L}_{f_r} = \mathrm{CE}(\mathbf{x}^d, f_r(\mathbf{x}^h), \mathcal{U}) + \mathrm{CE}(\mathbf{x}^t, f_r(\mathbf{x}^d), \mathcal{U}').$$

Operating on GNNs

1. **Perturbagen discovery:** Given a triplet $< \mathbf{x}^d, \mathcal{U}', \mathbf{x}^t >$, we propose a neural model operating on graph $G^{\mathcal{U}'}$ with node features $\mathbf{x}^d$ and $\mathbf{x}^t$ that predicts a ranking for each node, where the top P ranked nodes should be predicted as the nodes in $\mathcal{U}'$.
2. **Response prediction:** For each node, we use the binary perturbation flag to assign a d-dimensional learnable embedding. To embed the gene expression value, we calculate thresholds using quantiles to assign the gene expression value into one of the B bins and use the bin index to assign a d-dimensional learnable embedding. To increase our model's representation power, we concatenate a d-dimensional positional embedding.

For each node $i \in \mathcal{V}$, an embedding $\mathbf{z}_i$ is computed using a GNN operating on the attributes of node's neighbors.

Message passing



Both f_p and f_r follow the standard message-passing framework, where node representations are updated by aggregating the information from neighbors in the graph.

References

1. Gonzalez, G., Lin, X., Herath, I., Veselkov, K., Bronstein, M. and Zitnik, M., Combinatorial prediction of therapeutic perturbations using causally-inspired neural networks. Nature Biomedical Engineering, In press, 2025.