212028

2025.01.22 - Algorithm

Graph Types and Processing

1. Whole Graph (Cell Graph)

- Base graph $\mathcal{G}_{\text{whole}}$ that represents the unperturbed cell
- Single instance (batch size = 1) since it never changes
- Contains complete set of genes, metabolites, and their interactions
- Serves as reference point for measuring perturbation effects
- Data structure matches cell_graph format:

```
dataset.cell_graph

HeteroData(
  gene={
    num_nodes=6607,
    node_ids=[6607],
    x=[6607, 64],
  },
  metabolite={
    num_nodes=2534,
    node_ids=[2534],
  },
  (gene, physical_interaction, gene)={
    edge_index=[2, 144211],
    num_edges=144211,
  },
  (gene, regulatory_interaction, gene)={
    edge_index=[2, 16095],
    num_edges=16095,
  },
  (metabolite, reaction-genes, metabolite)={
    hyperedge_index=[2, 20960],
    stoichiometry=[20960],
    num_edges=4881,
    reaction_to_genes=dict(len=4881),
    reaction_to_genes_indices=dict(len=4881),
  }
)
```

2. Intact Graphs (Perturbed Instances)

- Collection of perturbed instances $\{\mathcal{G}_{\text{intact}}^{(i)}\}_{i=1}^{b}$ where b is batch size
- Each graph is derived from whole graph but with specific perturbations
- Processed in batches during training
- Contains additional perturbation-related data:

```
dataset[40]

HeteroData(
  gene={
    node_ids=[6605],
    num_nodes=6605,
    ids_pert=[2],
    cell_graph_idx_pert=[2],
    x=[6605, 64],
    x_pert=[2, 64],
    gene_interaction=[1],
    gene_interaction_p_value=[1],
    fitness=[1],
    fitness_std=[1],
  },
  metabolite={
    num_nodes=2534,
    node_ids=[2534],
  },
  (gene, physical_interaction, gene)={
    edge_index=[2, 144199],
    num_edges=144199,
  },
  (gene, regulatory_interaction, gene)={
    edge_index=[2, 16089],
    num_edges=16089,
  },
  (metabolite, reaction-genes, metabolite)={
    hyperedge_index=[2, 20939],
    stoichiometry=[20939],
    reaction_to_genes=dict(len=4881),
    reaction_to_genes_indices=dict(len=4881),
    num_edges=4875,
  }
)
```

Processing Flow

1. Whole Graph Processing:
   - Single pass through base cell graph
   - Outputs used as reference and for querying perturbed embeddings
2. Intact Graph Processing:
   - Batch processing of perturbed instances
   - Each instance compared against whole graph for fitness calculation
   - Perturbation effects measured relative to whole graph state

1. Gene-Gene Interaction Multigraph Let $\mathcal{G}_g = (\mathcal{V}_g, \mathcal{E}_g, \phi)$ represent the gene-gene interaction multigraph where:

- $\mathcal{V}_g$ is the set of gene vertices with $|\mathcal{V}_g| = n_g$ vertices
- $\mathcal{E}_g = \mathcal{E}_p \cup \mathcal{E}_r$ is the multiset of edges where:
  - $\mathcal{E}_p$ is the set of physical interaction edges
  - $\mathcal{E}_r$ is the set of regulatory interaction edges
- $\phi : \mathcal{E}_g \to \{\text{physical}, \text{regulatory}\}$ is the edge type mapping
- $X_g \in \mathbb{R}^{n_g \times d}$ is the gene feature matrix where d is the feature dimension

2. Metabolic Hypergraph Let $\mathcal{H}_m = (\mathcal{V}_m, \mathcal{E}_r, I_{m \to r}, I_{r \to g}, S)$ represent the metabolic hypergraph where:

- $\mathcal{V}_m$ is the set of metabolite vertices with $|\mathcal{V}_m| = n_m$ vertices
- $\mathcal{E}_r$ is the set of reaction hyperedges with $|\mathcal{E}_r| = n_r$ edges
- $I_{m \to r} \in \{0,1\}^{n_m \times n_r}$ is the metabolite-to-reaction incidence matrix
- $I_{r \to g} \in \{0,1\}^{n_r \times n_g}$ is the reaction-to-gene incidence matrix
- $S \in \mathbb{R}^{n_r}$ contains the stoichiometric coefficients
- $E_m \in \mathbb{R}^{n_m \times h}$ is the metabolite embedding lookup table

3. Label Data Structures For each batch of size b:

- $y_{\text{fitness}} \in \mathbb{R}^b$ (fitness ratio labels)

- $y_{\text{gene\_interaction}} \in \mathbb{R}^b$ (gene interaction labels)
- $P \in \mathbb{N}^p$ (perturbed gene indices for each sample)

Forward Pass Architecture

Base Forward Function Takes a graph $\mathcal{G}$ and outputs latent embeddings Z and pooled representation z

$\text{forward}(\mathcal{G}) \to (Z, z):$

1. Preprocessing:
   - $H_g = \text{MLP}(X_g) \in \mathbb{R}^{n_g \times h}$, where $n_g = 6607$ (gene nodes)
   - $H_r = \text{SAB}(H_g, I_{r\to g}) \in \mathbb{R}^{n_r \times h}$, where $n_r = 4881$ (reactions)
2. Parallel Processing:

   Gene Path:

   - $Z_g = \text{HeteroGNN}(H_g, \mathcal{E}_g) \in \mathbb{R}^{n_g \times h}$

   Metabolic Path:

   - $Z_m = \text{StoichiometricHypergraphConv}(E_m, H_r, \mathcal{E}_r, S) \in \mathbb{R}^{n_m \times h}$, where $n_m = 2534$ (metabolites)
   - $Z_r = \text{SAB}(Z_m, I_{m\to r}) \in \mathbb{R}^{n_r \times h}$
   - $Z_{mg} = \text{SAB}(Z_r, I_{r\to g}) \in \mathbb{R}^{n_g \times h}$
3. Integration:
   - $Z = \text{MLP}([Z_g \| Z_{mg}]) \in \mathbb{R}^{n_g \times h}$
   - $z = \text{ISAB}(Z) \in \mathbb{R}^h$

   Return: (Z, z)

Model Workflow

1. Process Whole Graph:
   - $(Z_W, z_W) = \text{forward}(\mathcal{G}_{\text{whole}})$
   - $Z_W \in \mathbb{R}^{n_g \times h}, z_W \in \mathbb{R}^h$
2. Process Intact Graph:
   - $(Z_I, z_I) = \text{forward}(\mathcal{G}_{\text{intact}})$
   - $Z_I \in \mathbb{R}^{n_g \times h}, z_I \in \mathbb{R}^h$
3. Query Perturbed Set:
   - Let $P \in \mathbb{N}^p$ be indices of perturbed genes from `ids_pert`, in example $p = 2$ (perturbed genes)
   - $Z_P = Z_W[P] \in \mathbb{R}^{p \times h}$
   - $z_P = \text{SAB}(Z_P) \in \mathbb{R}^h$

Prediction Heads

1. Growth and Fitness Calculation:
   - $\text{growth}_W = \text{MLP}_{\text{growth}}(z_W) \in \mathbb{R}^1$
   - $\text{growth}_I = \text{MLP}_{\text{growth}}(z_I) \in \mathbb{R}^1$
   - $\hat{y}_{\text{fitness}} = \text{growth}_I / \text{growth}_W \in \mathbb{R}^1$
2. Gene Interaction:
   - $\hat{y}_{\text{gene\_interaction}} = \text{MLP}_{\text{interaction}}(z_P) \in \mathbb{R}^1$

For a batch of size b: $\hat{Y} = [\hat{y}_{\text{fitness}} \| \hat{y}_{\text{gene\_interaction}}] \in \mathbb{R}^{2\times b}$

Loss Computation The total loss with weighting:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}}(Y, \hat{Y}) + \lambda_1 \mathcal{L}_{\text{dist}}(Y, \hat{Y}) + \lambda_2 \mathcal{L}_{\text{SupCR}}(z_P, z_I, Y) + \lambda_3 \mathcal{L}_{\text{cell}}(z_W, z_P, z_I)$$

Where:

- $Y, \hat{Y} \in \mathbb{R}^{2\times b}$ (ground truth and predictions)

- $z_P, z_I, z_W \in \mathbb{R}^h$ (latent representations)
- $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}^+$ (loss weights)