Prior Knowledge Injection into Deep Learning Models Predicting Gene Expression from Whole Slide Images

Max Hallemeesch, Marija Pizurica, Paloma Rabaey, Olivier Gevaert, Thomas Demeester, Kathleen Marchal

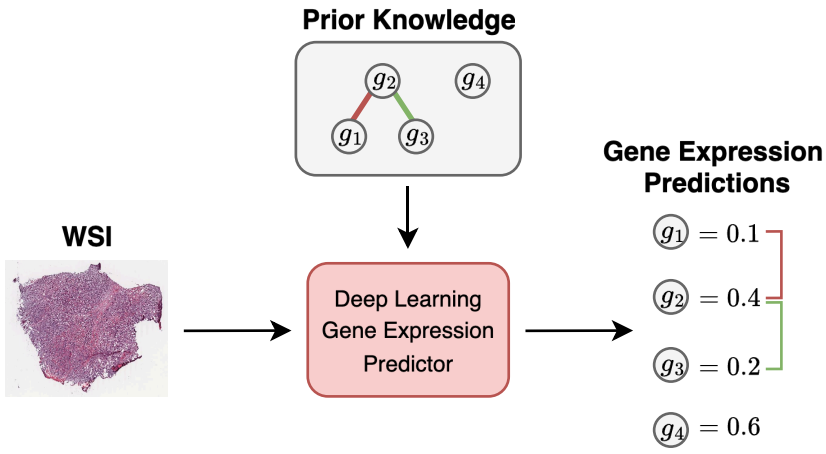
Problem Statement

Recent advances in Deep Learning allow to predict molecular information from morphological features within Whole Slide Images (WSIs). While promising, **current methods lack the robustness** to fully replace direct sequencing.

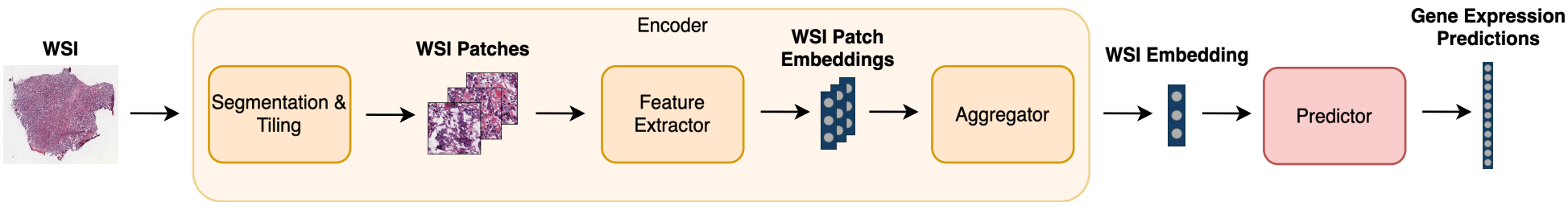
Goal

Here we aim to improve existing methods by introducing a **model-agnostic framework that allows to inject prior knowledge on gene-gene interactions** into Deep Learning architectures.

High Level Approach

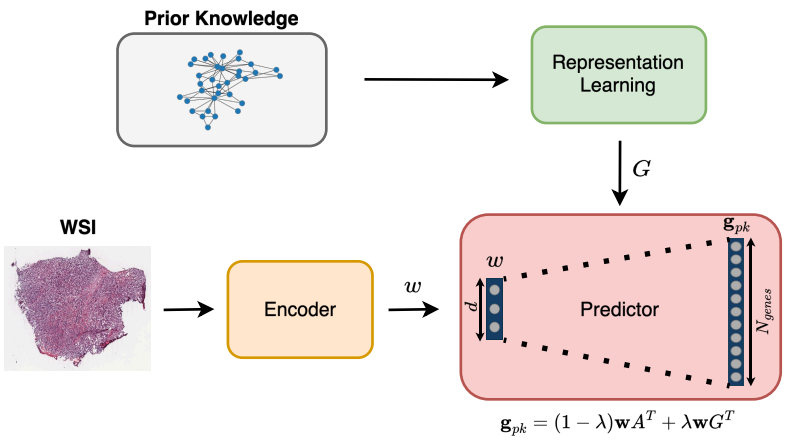


General Workflow Gene Expression Prediction



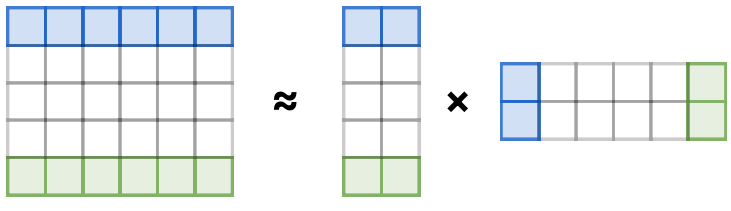
General workflow of predicting gene expression from Whole Slide Images. First, the WSI is processed by an encoder, which extracts patches, their corresponding features, and aggregates them into a single WSI embedding. Then, a predictor transforms the embedding into expression predictions for 25,761 genes. We evaluate our model-agnostic framework by considering two feature extractors (CTrans (ctr) and UNI) and three aggregators (MLP, Transformer (tf) and SummaryMixing (smx)).

Framework Architecture



Overview of the framework. First, we transform prior knowledge (gene-gene interaction network) into gene embeddings *G* using a representation learning technique. The Encoder transforms the WSI into an embedding **w**. We then inject the gene embeddings into the Predictor by linearly transforming **w** into gene predictions $\mathbf{g}_{pk}$ using a weighted sum of the linear predictor layer *A*, and the gene embeddings *G*. Hyperparameter λ controls the effect of prior knowledge.

Representation Learning



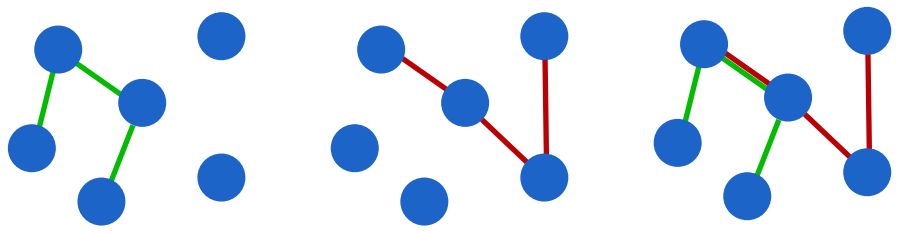
Each prior knowledge co-expression graph can be represented by an adjacency matrix. We compress this high-dimensional information into low-dimensional gene embeddings by employing Nonnegative Matrix Factorization (NMF), which minimizes the Frobenius Norm.

/in/maxhallemeesch

max@hallemeesch.net

github.com/maxhallemeesch/PRALINE

Prior Knowledge Sources



Internal
Train set

External
Independent set

Combined
Train set + Independent set

In our research, we exploit gene co-expression as prior knowledge. We define a pair of genes (*i*, *j*) to be co-expressed if the Pearson correlation R_{ij} between their respective expression levels (p_i, p_j) exceeds a threshold. We consider three sources of prior knowledge: internal (TCGA-BRCA), external (Cancer Gene Neighborhood of MsigDB) and combined.

Evaluation

TCGA	No PK	External	Internal	Combined
ctr_mlp	21,233	22,225 (0.9) ↑992	22,278 (0.9) ↑1,045	22,160 (0.9) ↑927
ctr_tf	19,155	21,647 (0.9) ↑2,429	20,618 (0.2) ↑1,463	20,548 (0.8) ↑1,393
ctr_smx	20,945	22,564 (0.9) ↑1,619	21,451 (0.5) ↑506	21,944 (0.5) ↑999
uni_mlp	21,721	22,666 (0.8) ↑945	22,802 (0.8) ↑1,081	**23,214** (0.9) ↑1,493
uni_tf	22,124	22,461 (0.2) ↑337	22,645 (0.1) ↑521	22,597 (0.1) ↑473
uni_smx	**22,997**	**23,578** (0.5) ↑581	**23,732** (0.1) ↑735	**23,162** (0.9) ↑165

CPTAC	No PK	External	Internal	Combined
ctr_mlp	16,936	16,313 ↓623	**18,116** ↑1,180	**17,363** ↑427
ctr_tf	15,677	15,146 ↓531	14,983 ↓694	14,386 ↓1,291
ctr_smx	15,714	16,682 ↑968	15,731 ↑17	15,753 ↑39
uni_mlp	15,560	16,106 ↑546	16,045 ↑485	15,784 ↑224
uni_tf	14,705	15,648 ↑763	15,469 ↑764	15,400 ↑695
uni_smx	16,952	**17,280** ↑328	**17,091** ↑139	**16,981** ↑29

We evaluated the number of significantly predicted genes across 18 experiments, including three sources of prior knowledge and six deep learning architectures, on both TCGA-BRCA (upper table) and CPTAC-BRCA (lower table). Across 14 experiments we observed an increase in the number of significant genes on both TCGA and CPTAC, demonstrating an enhanced generalization performance.