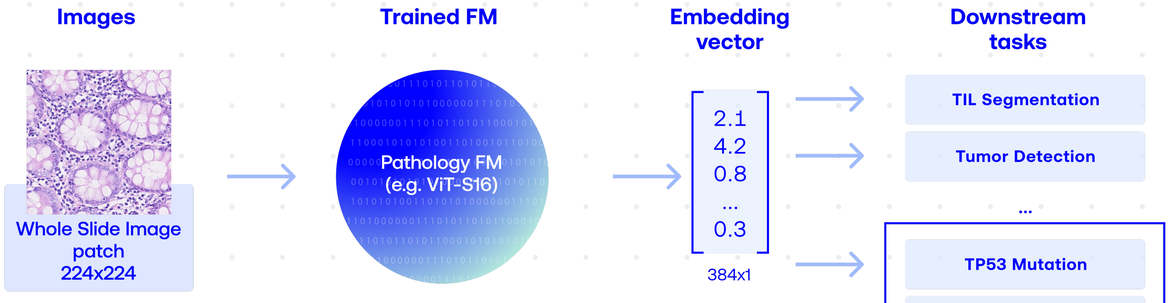
Enhancing Pathology Foundation Models with Transcriptomics

Edwin D. de Jong, Mikhail Karasikov, Marharyta Kurban, Moritz Platscher, Marie Stettler, Fei Tang

Interested in a collaboration?
Check out our website! ⟶

FM for clinical tasks

Pathology FMs are computer vision models trained with Self-Supervised Learning (SSL) on massive datasets of pathology images, designed to work across various tasks.



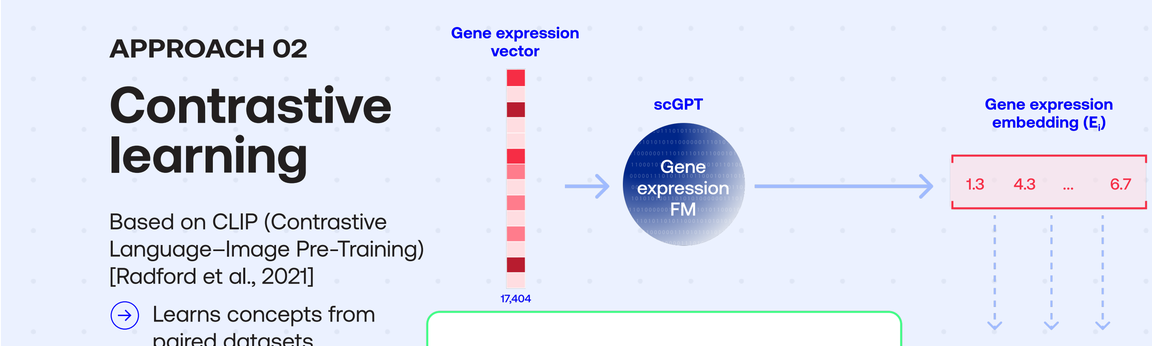
Observation: Pathology FMs don't perform as well as genomics-based models on some downstream tasks.

However: Sequencing is costly while pathology slides are generated nearly for each patient.

Idea: Enhance vision-only pathology FMs by fine-tuning them on transcriptomics data.

Training / fine-tuning FM

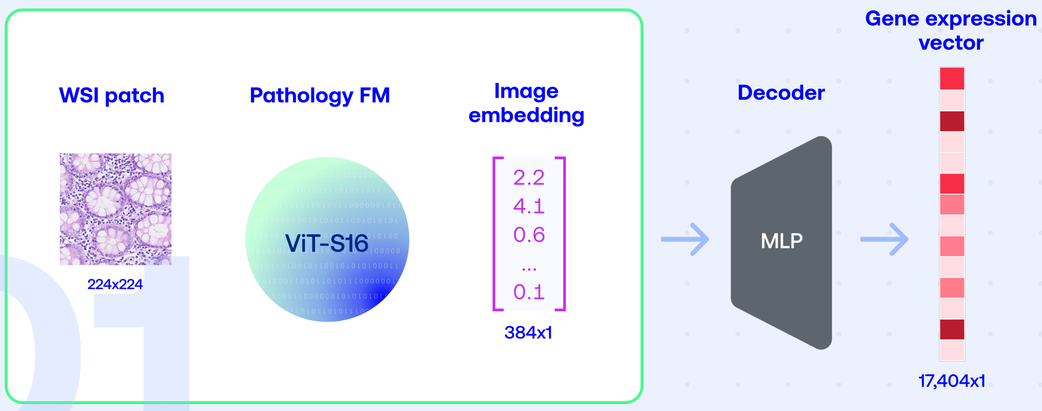
Enhancing FM with transcriptomics by fine-tuning on paired pathology image — gene expression data (e.g., from TCGA).



APPROACH 01

Direct prediction

Fine-tune the model to directly predict gene expression from the embeddings



APPROACH 02

Contrastive learning

Based on CLIP (Contrastive Language–Image Pre-Training) [Radford et al., 2021]

Learns concepts from paired datasets

We use gene expression instead of text as in CLIP

Results

Contrastive learning approach outperforms direct fine-tuning.

The highest improvement is achieved for TCGA downstream tasks.

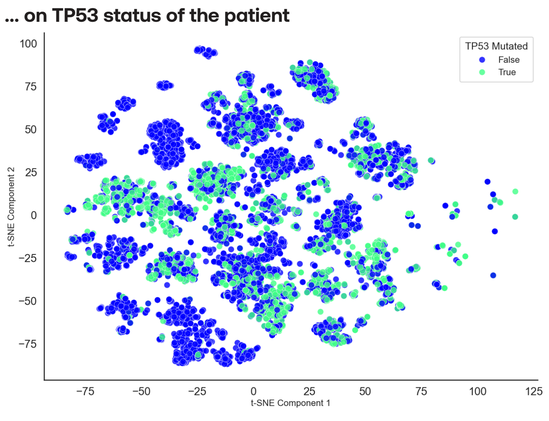
	Balanced accuracy									Average correlation		
	BACH	PCam Test	MHIST	CRC MSI Kather	TCGA Cancer	TCGA TP53	TCGA Breast	TCGA Kidney	TCGA Lung	TCGA Expression	Mean delta	
Base model	67.1%	89.0%	74.1%	66.7%	58.7%	66.6%	53.0%	87.7%	70.7%	41.6%	-	
Direct fine-tuning	-0.6%	-1.3%	-1.5%	0.4%	0.4%	1.3%	2.9%	1.3%	0.3%	2.1%	**0.5%**	Δ
Contrastive base	62.7%	86.5%	74.3%	66.7%	58.5%	67.1%	54.7%	86.7%	69.9%	40.9%	-	
Contrastive fine-tuning	1.8%	-0.5%	1.3%	-0.5%	1.1%	0.3%	0.0%	2.2%	1.2%	1.1%	**0.8%**	Δ

Table 1. The accuracy (in %) and its delta for the two fine-tuned models on various downstream tasks. The TCGA expression prediction is evaluated with the average Pearson correlation coefficient. All other tasks use balanced accuracy.

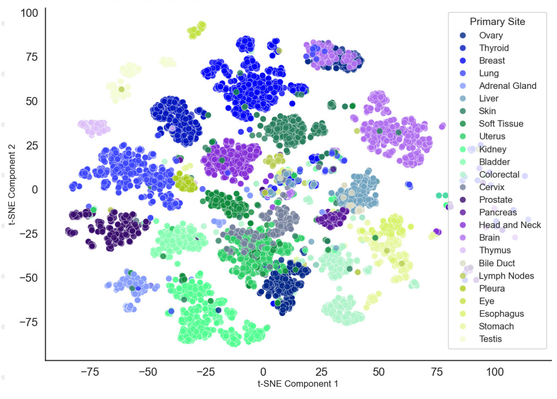
Text integration & future work

T-SNE plots of the TCGA-PRAD report embeddings (from OpenAI text-embedding-ada-002) indicate high predictive power of the text modality

... on TP53 status of the patient



... on initial tumor site



Many downstream targets are predictable not only from RNA-Seq and WSI but also from other modalities, e.g., text.

Future work:

Fine-tune the model on other datasets of linked modalities.

Final model will provide insights from various modalities using only images as input.

REFERENCES

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

kaiko.ai