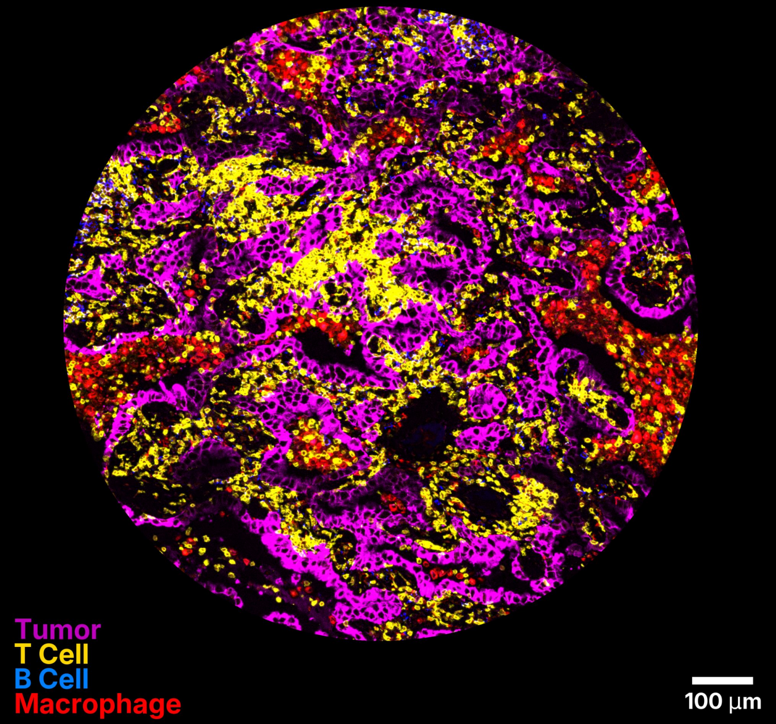
May 2025
CS25 | Stanford

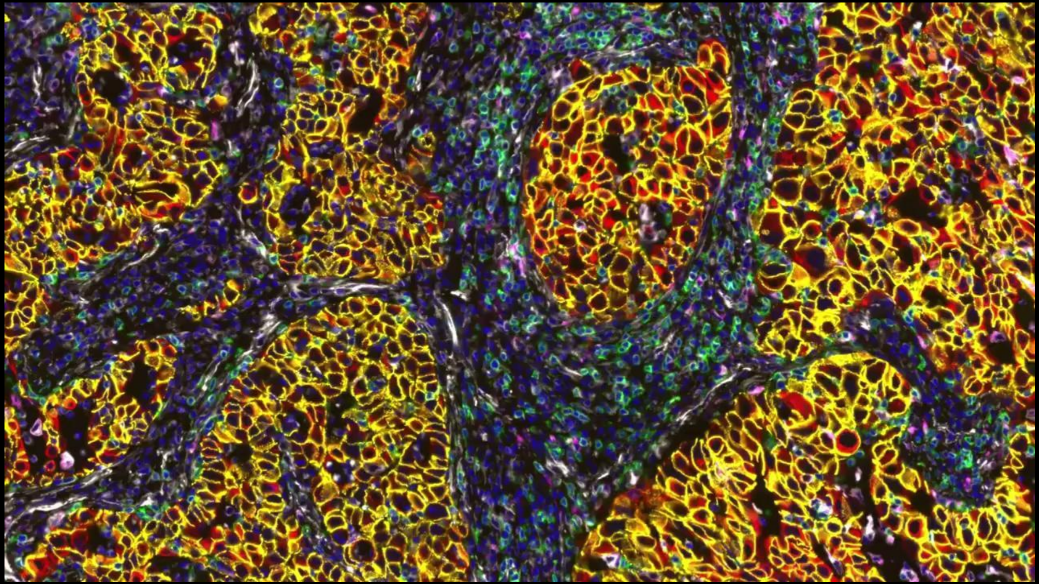
Transformers vs. Cancer

Eshed Margalit, PhD

NOETIK



Today's topics



1 | Multimodal Model Madness

2 | Cracking Cancer con Context

3 | Futuristic figures + Follow-ups

What I assume about you:

- you’re interested in research on novel transformer architectures and training tasks
- you’re curious about “real-world” applications of transformers, including those beyond LLMs
- you’re familiar with the basics of transformers and ML
- you are not familiar with cancer immunology, but think curing cancer would be neat

What you should know about me

- background in computational neuroscience, computer vision, and visual cortex @ Stanford

- broadly interested in understanding how complex biological systems are assembled, how they function, and how they break

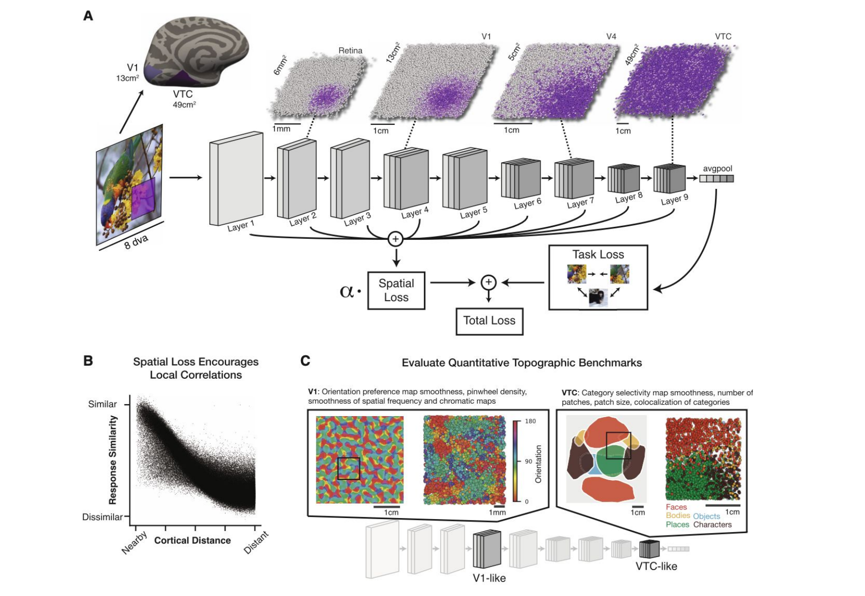


Figure 1. Constructing a unified model of the functional and spatial constraints of the ventral visual cortex

NOETIK

ML @ Noetik

Daniel Bear

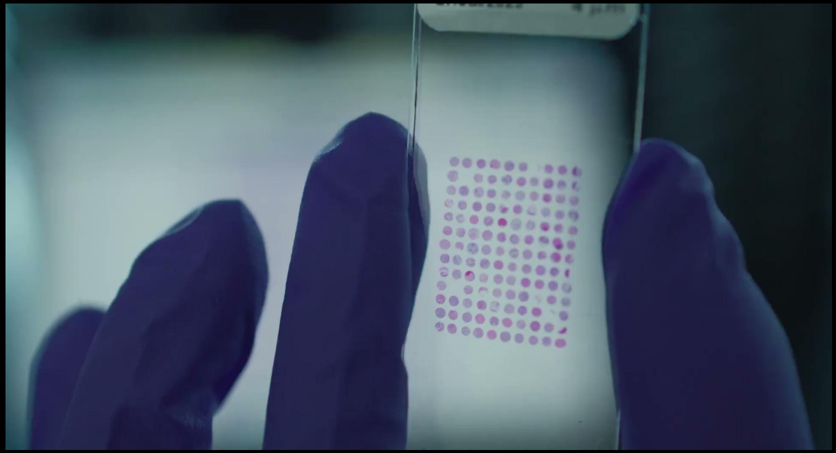
Jake Schmidt

Michela Meister

Ryan Huang

Yubin Xie

Eshed Margalit



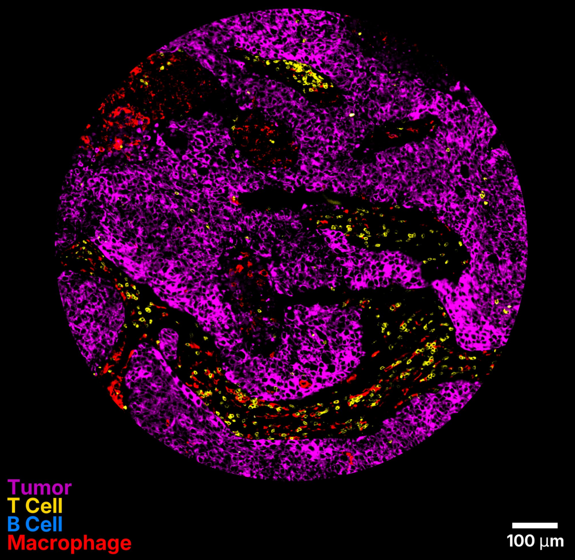
NOETIK

How the next 60 minutes are probably going to go

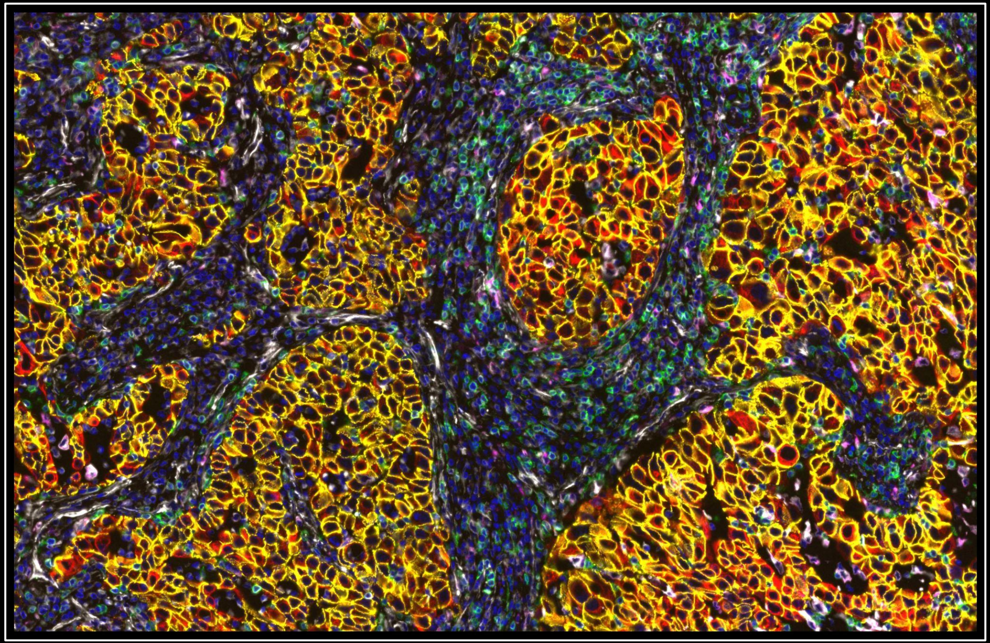
- I'm going to try to convince you that 1) there's a lot of very exciting and creative work to be done with multimodal transformers, and 2) that cancer biology is a fantastic place to do basic ML research

- Ok and for bonus points: 3) that we're making meaningful progress in understanding cancer biology @ Noetik

- Interruptions for clarifying questions strongly encouraged, but please hold larger/philosophical questions for the end



Today's topics



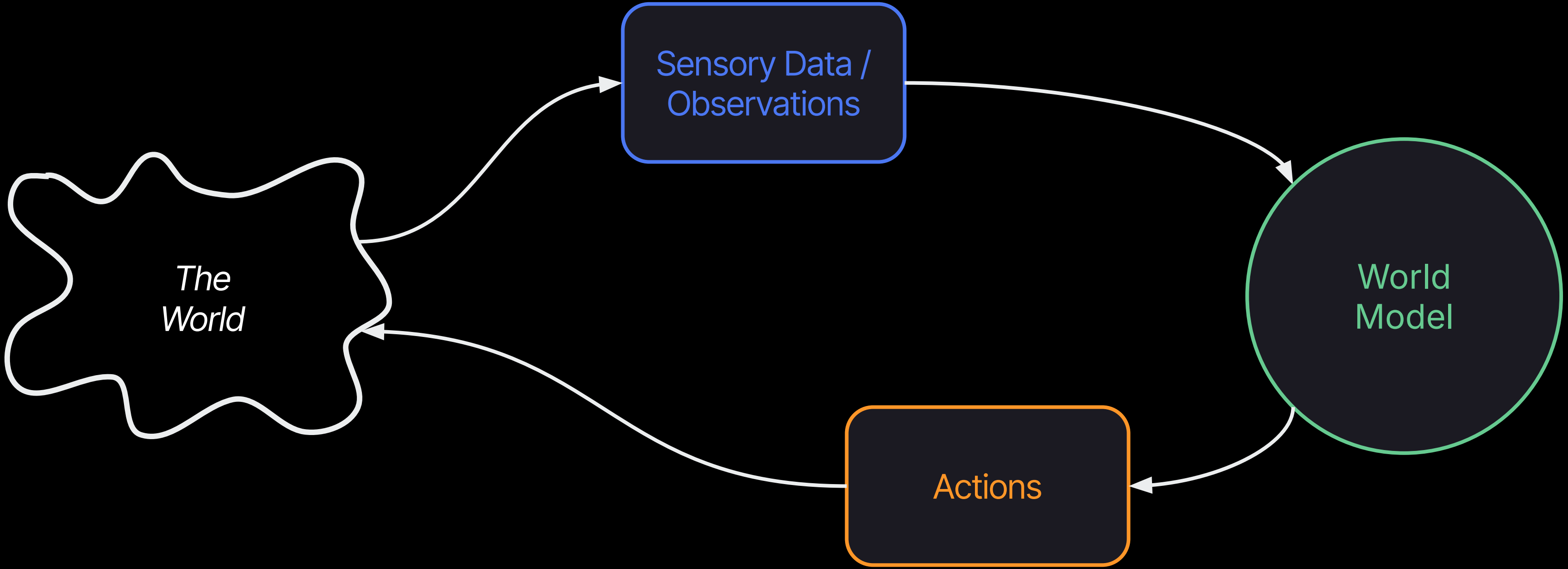
1 | Multimodal Model Madness

2 | Cracking Cancer con Context

3 | Futuristic figures + Follow-ups

NOETIK

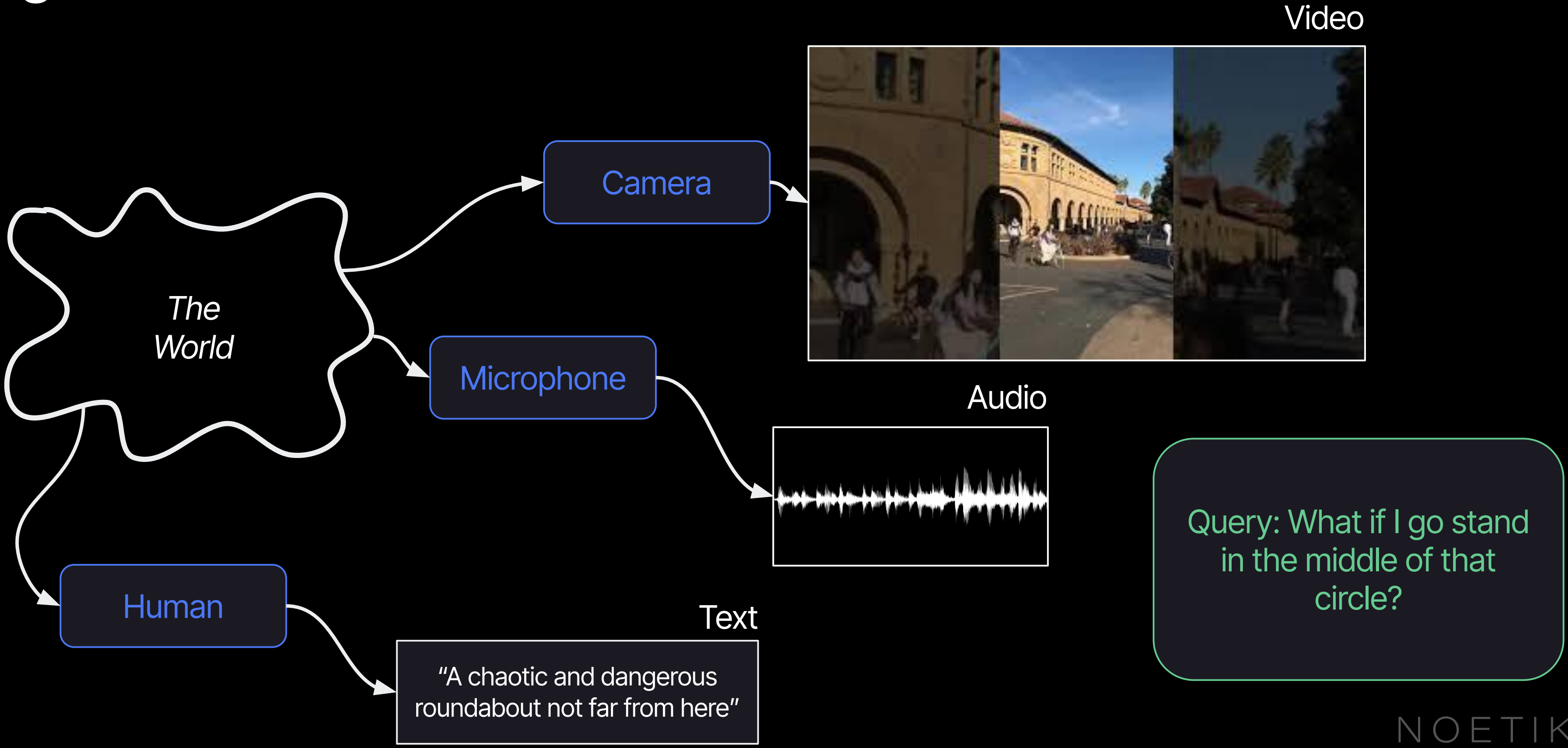
A unifying goal in AI research is to build 'world models'



Operational definition: a world model is a system that can simulate the future state of the world conditioned on existing state and actions

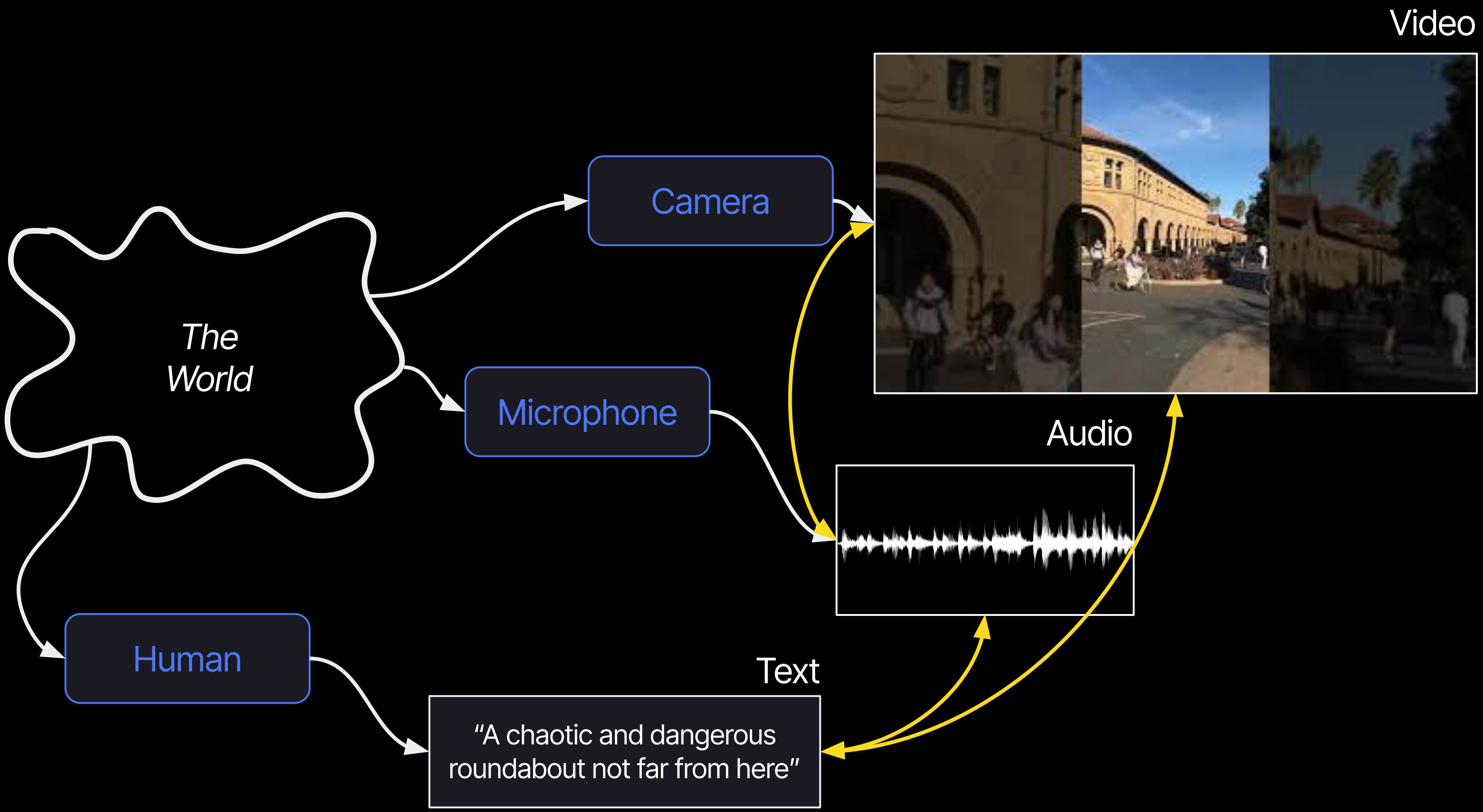
NOETIK

The world is perceived in multiple modalities, and the best agents will reason about all of them



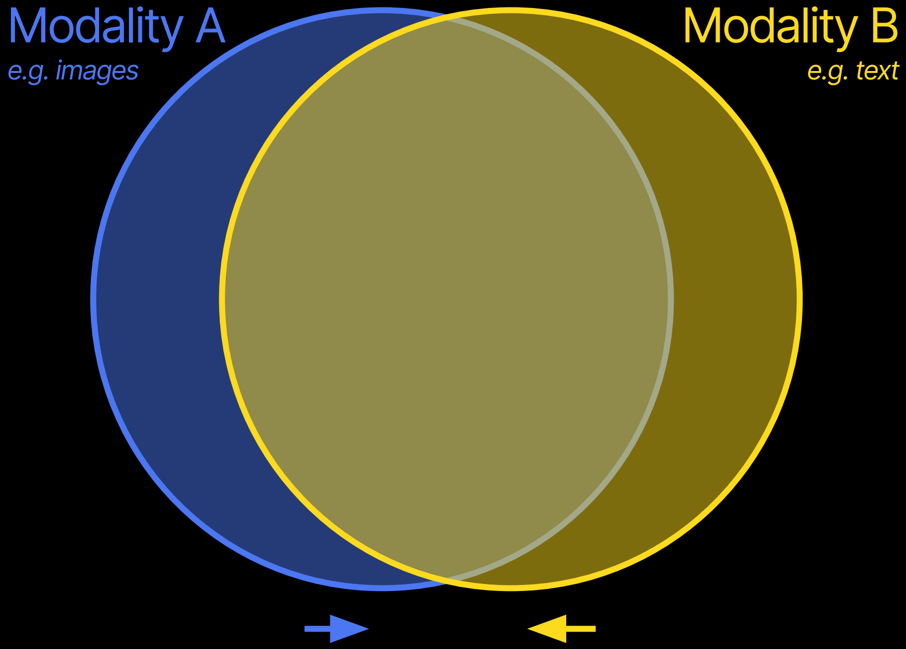
NOETIK

Multimodal learning refers to the fusion of, or translation between, different modalities



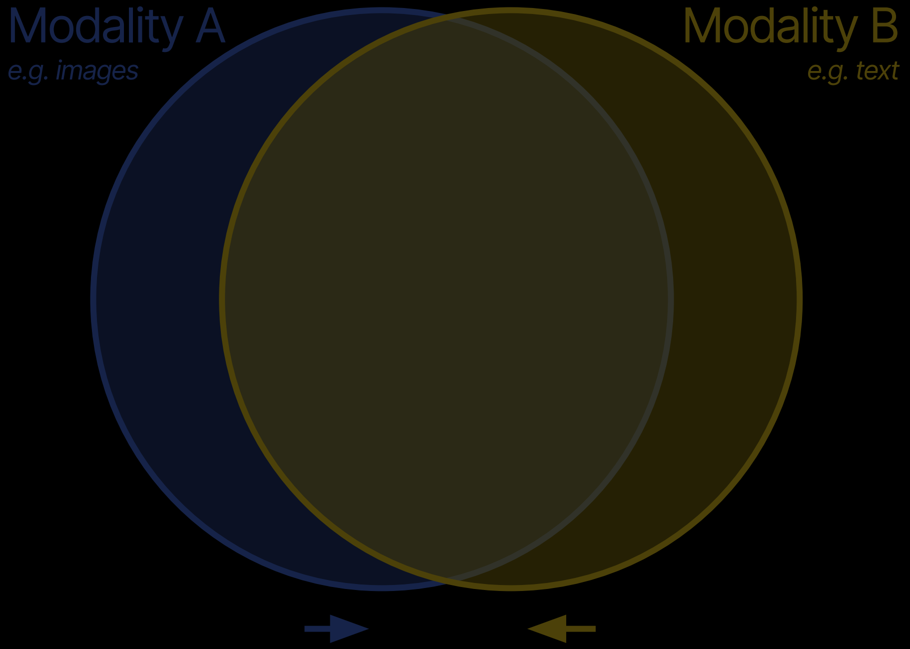
NOETIK

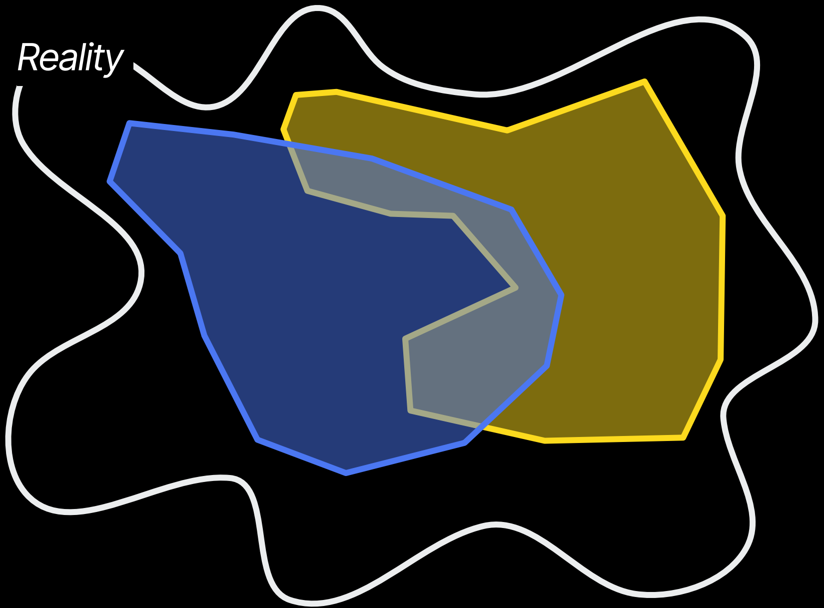
Multimodality as *translation*



NOETIK

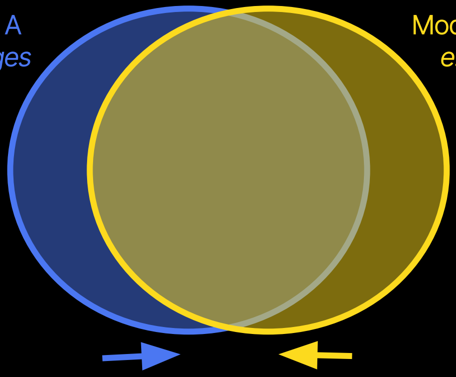
Multimodality as *translation* vs. *disambiguation*





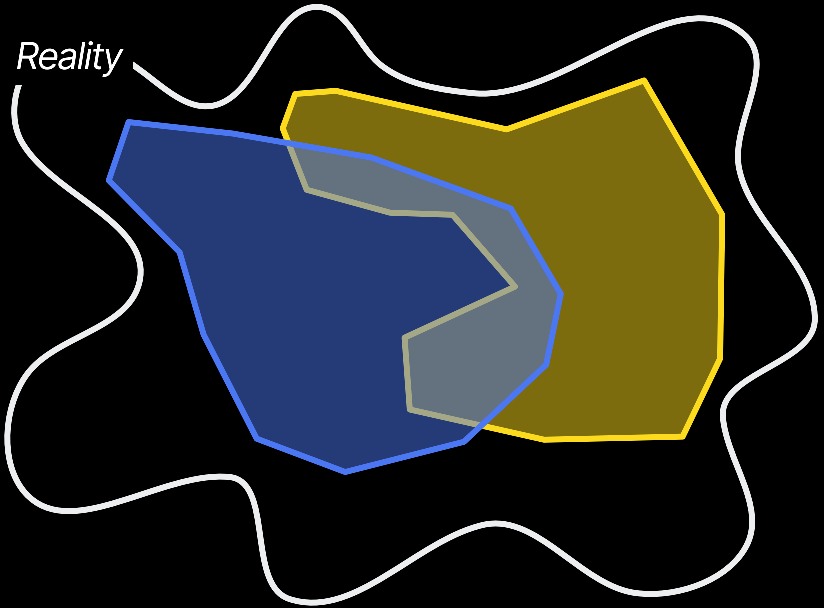
Multimodality as *translation*

Emphasis on unified representation of samples across modalities: all of the image content should reflect all of the text content



Multimodality as *disambiguation*

There can be information in one modality that does not exist in the others, and we want to learn about the underlying world by combining both.



NOETIK

Multimodality as *disambiguation*

Example: you approach a building and see everybody running out of it.

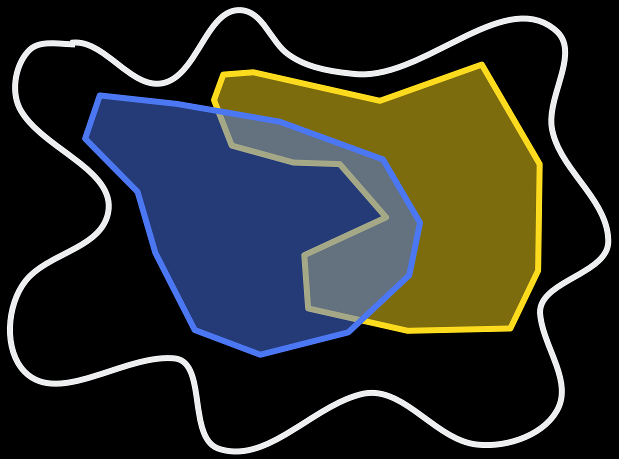
Scenario 1: you hear a fire alarm going off inside

Scenario 2: you hear an announcement behind you that there's free boba for the first 10 students to claim it

How does multimodal learning actually work?

<u>A brief and very incomplete tour of ideas:</u>

1. Learning joint embedding spaces
2. Concatenation of inputs
3. Cross-attention
4. Concatenation of tokens
5. Layernorm context

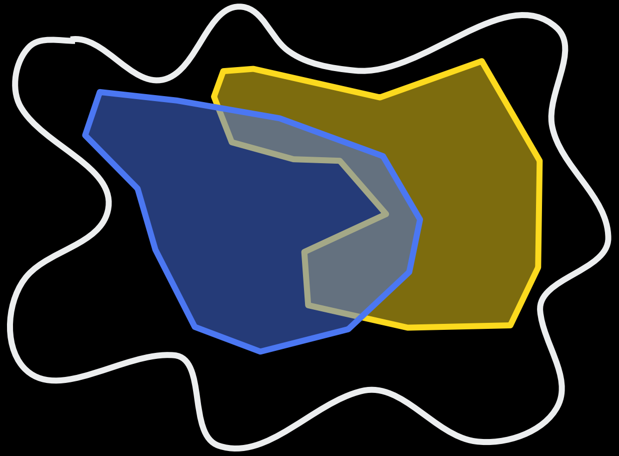


NOETIK

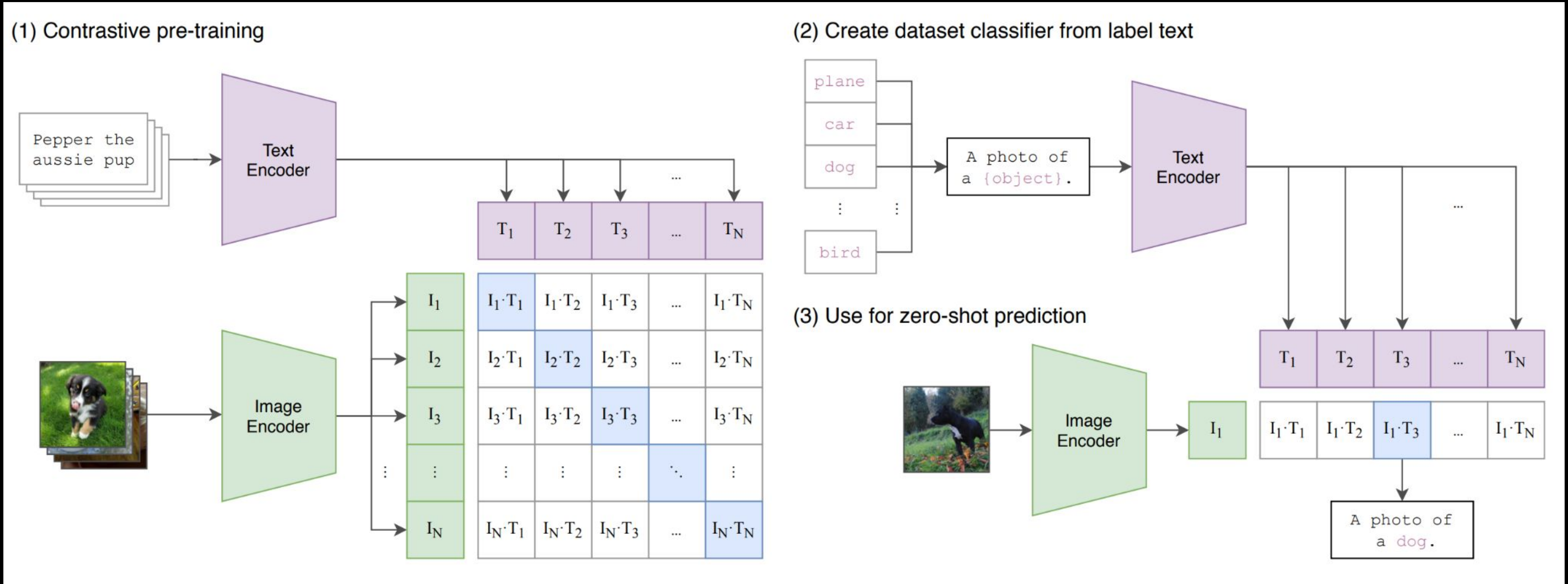
How does multimodal learning actually work?

A brief and very incomplete tour of ideas:

1. Learning joint embedding spaces (late fusion)
2. Concatenation of inputs (super-early fusion)
3. Cross-attention (mid-fusion?)
4. Concatenation of tokens (early-ish fusion?)
5. Layernorm context (early-ish fusion?)

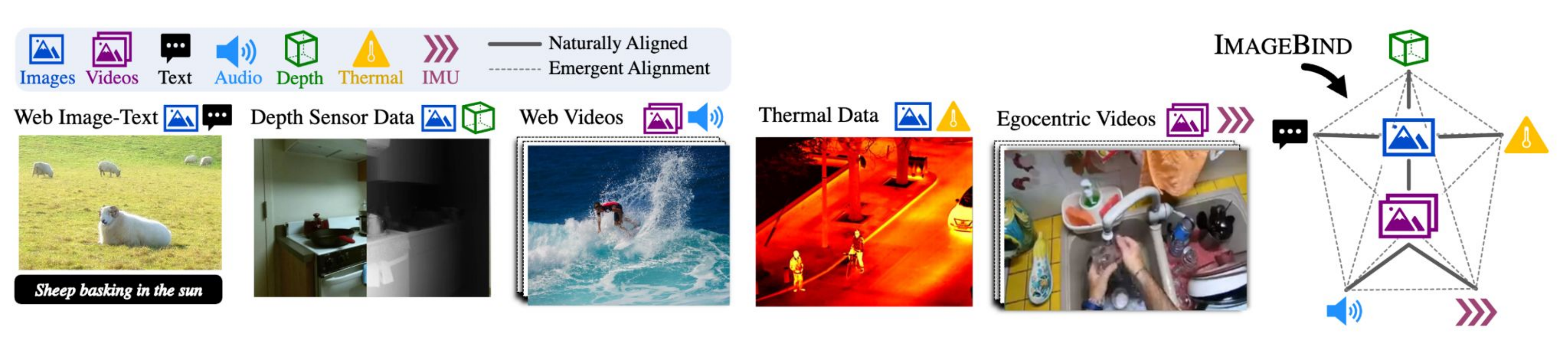


Option 1: encourage similar embeddings from different modalities



CLIP: Radford et al., 2021

Option 1: encourage similar embeddings from different modalities



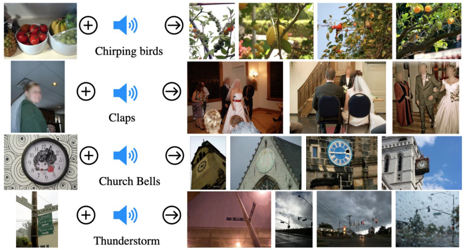


Figure 4. **Embedding space arithmetic** where we add image and audio embeddings, and use them for image retrieval. The composed embeddings naturally capture semantics from different modalities. Embeddings from an image of fruits + the sound of birds retrieves images of birds surrounded by fruits.

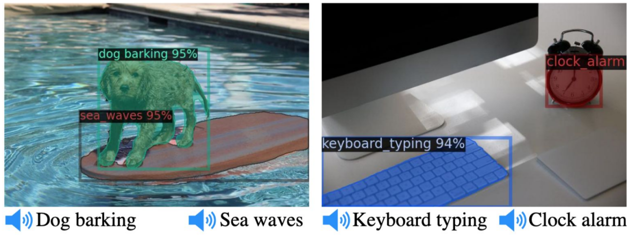


Figure 5. Object detection with audio queries. Simply replacing Detic [86]'s CLIP-based 'class' embeddings with our audio embeddings leads to an object detector promptable with audio. This requires no re-training of any model.

ImageBind: Girdhar et al., 2023

All encoders are transformers

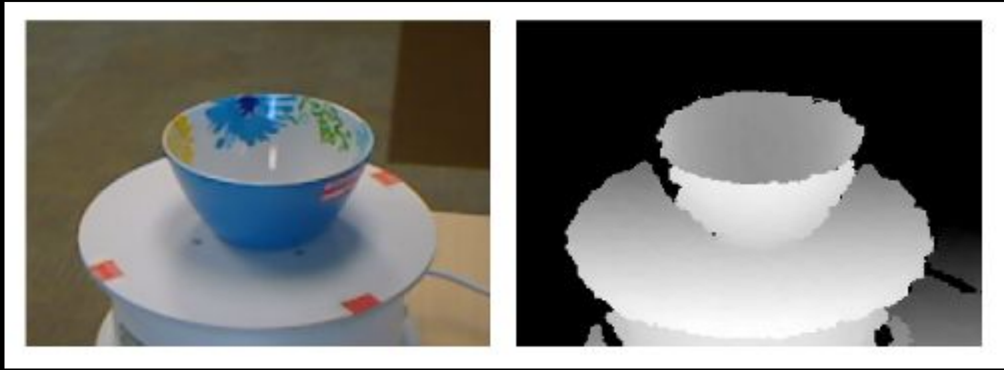
Spiritually similar to CLIP, but with many contrastive pairs

NOETIK

Option 2: concatenate inputs directly

- Not very common
- Requires that modalities have some shared dimensions (e.g. spatial dims)
- Sort of “extremely early” fusion

RGB (3-D) Depth (1-D)



RGB-D dataset: https://rgbd-dataset.cs.washington.edu/

Option 3: cross-attention

Lu et al., 2019

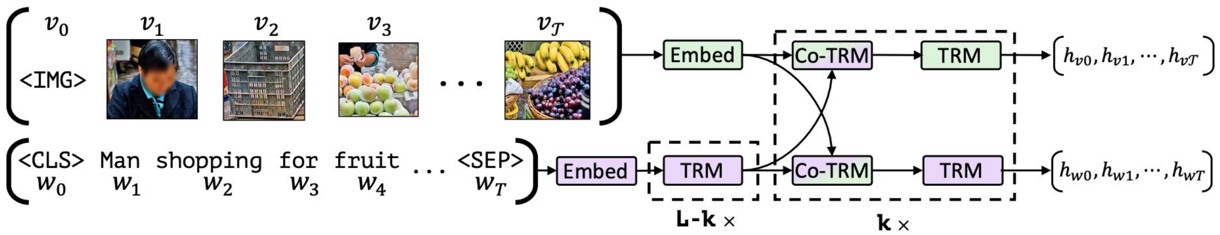
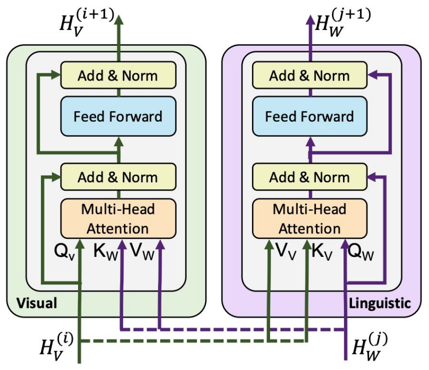
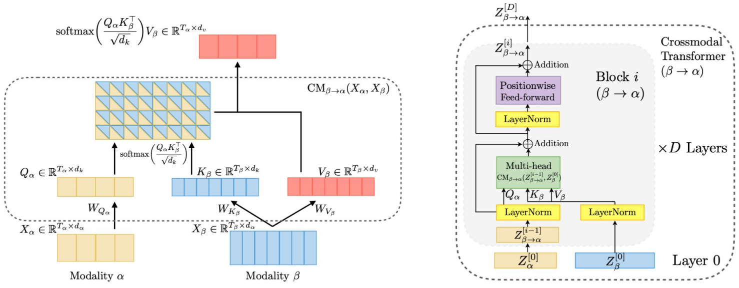


Figure 1: Our ViLBERT model consists of two parallel streams for visual (green) and linguistic (purple) processing that interact through novel co-attentional transformer layers. This structure allows for variable depths for each modality and enables sparse interaction through co-attention. Dashed boxes with multiplier subscripts denote repeated blocks of layers.



(b) Our co-attention transformer layer

MulT: Tsai et al., 2019



(a) Crossmodal attention $\text{CM}_{\beta\rightarrow\alpha}(X_\alpha, X_\beta)$ between sequences X_α, X_β from distinct modalities.

(b) A crossmodal transformer is a deep stacking of several crossmodal attention blocks.

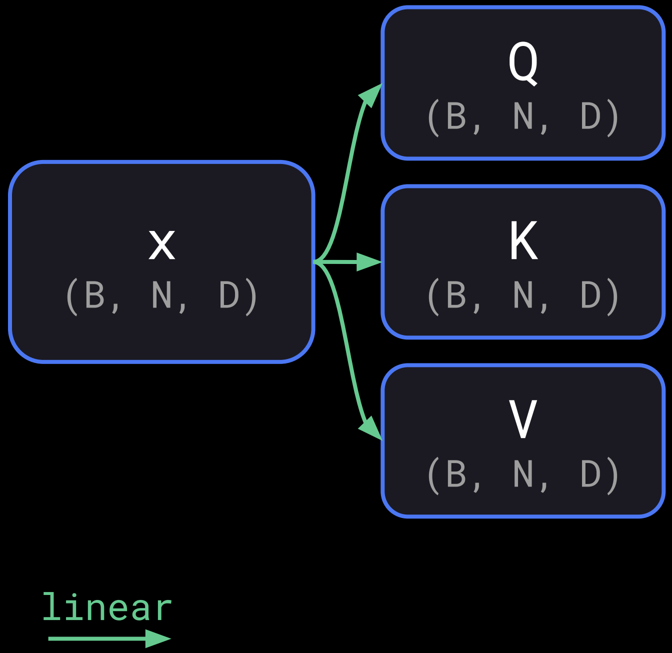
Figure 3: Architectural elements of a crossmodal transformer between two time-series from modality α and β.

NOETIK

Reminder: single-modality self-attention

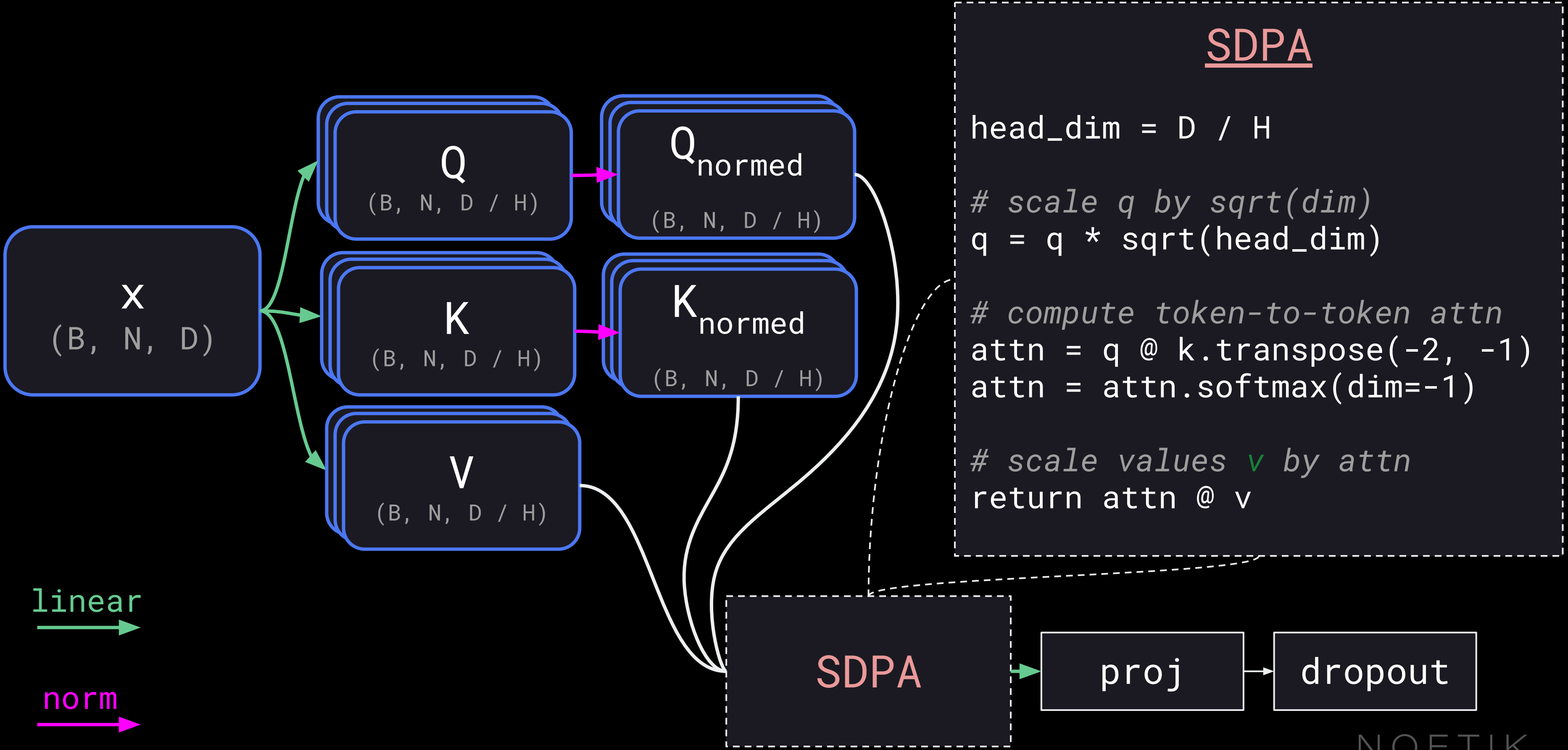
X
(B, N, D)

Reminder: single-modality self-attention

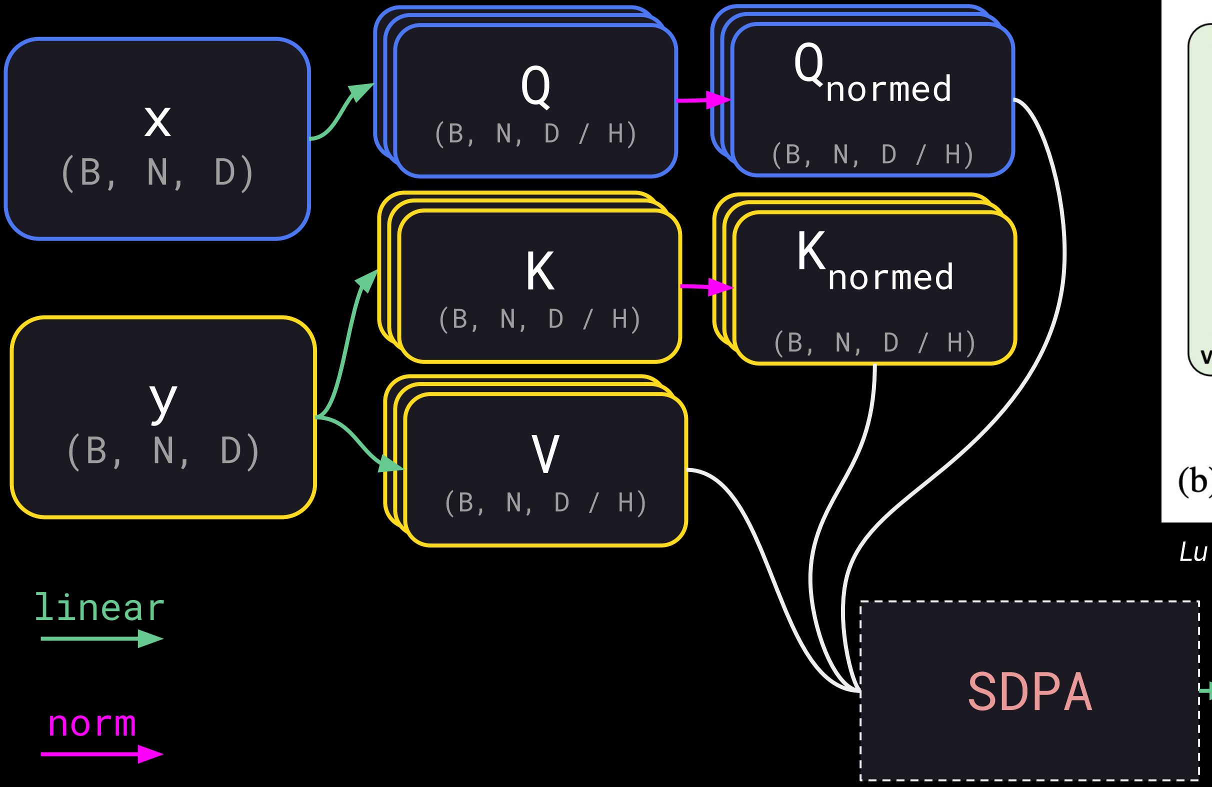


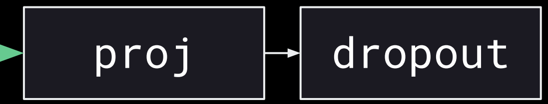
NOETIK

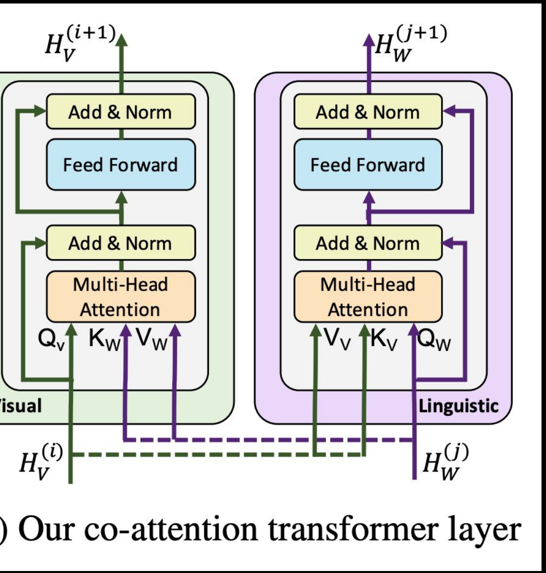
Reminder: single-modality multi-head self-attention



Cross-attention: queries from one stream, keys/values from the other







(b) Our co-attention transformer layer

Lu et al., 2019

NOETIK

Option 4: add modality information as a bonus/CLS token

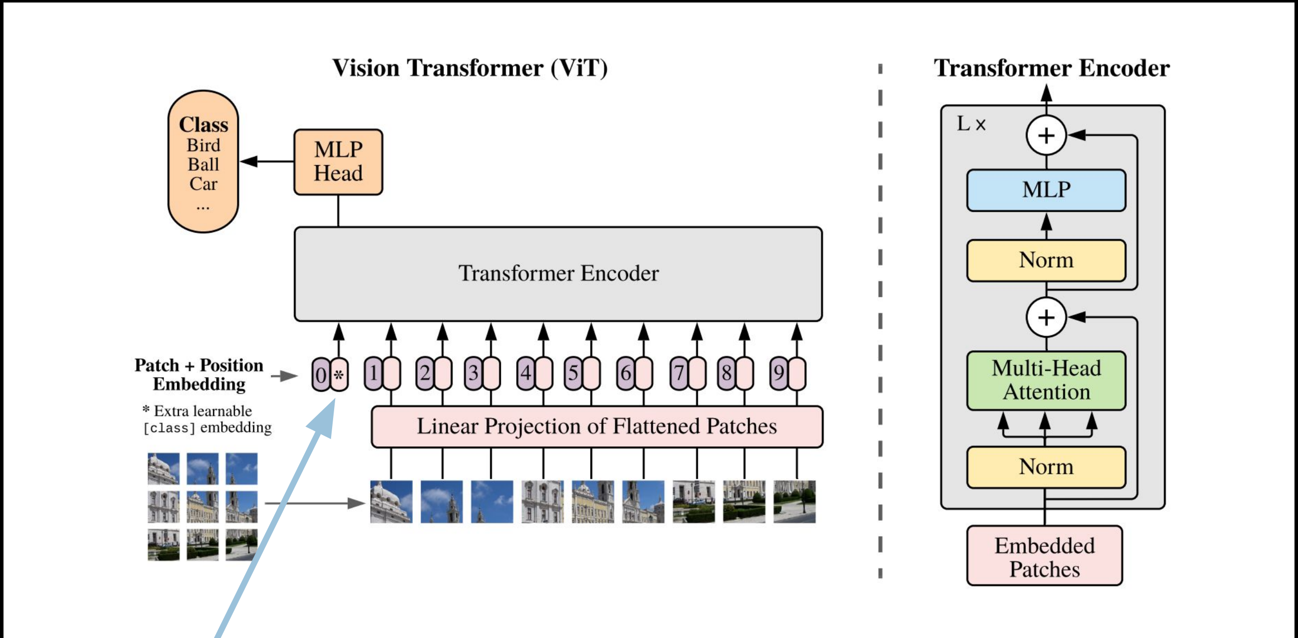


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).

ViT: Dosovitskiy et al., 2021

I'll show an example of how this might work for an input (instead of discrete label) later – for now just imagine you have some other NN encoder pushing input from another modality into a single token

NOETIK

Option 4: add modality information as a bonus/CLS token

Image credit: OpenAI (https://openai.com/index/dall-e/)

DALL-E: Ramesh et al., 2021

Concatenate text tokens and image tokens (encoded with a discrete VAE) into a single stream

Also uses CLIP to rank possible generated images

NOETIK

Option 5: adaptive layernorm

DiT: Peebles and Xie, 2023

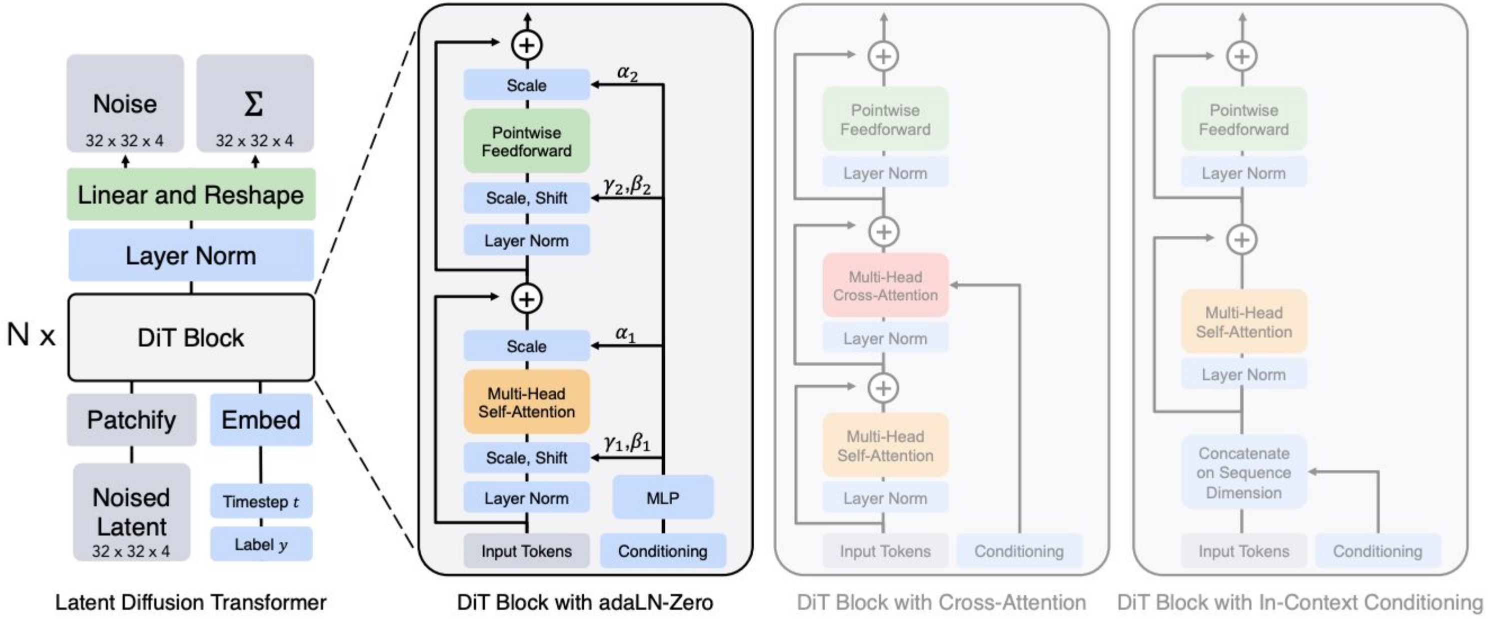


Figure 3: **The Diffusion Transformer (DiT) architecture.** *Left:* We train conditional latent DiT models. The input latent is decomposed into patches and processed by several DiT blocks. *Right:* Details of our DiT blocks. We experiment with variants of standard transformer blocks that incorporate conditioning via adaptive layer norm, cross-attention and extra input tokens. Adaptive layer norm works best.

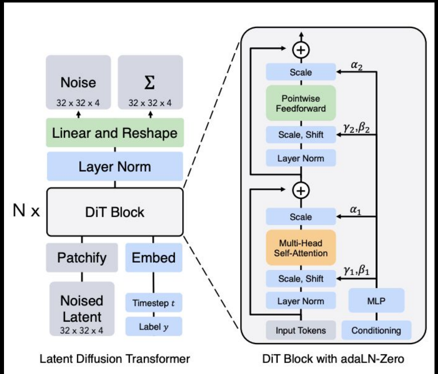
What? With LayerNorn? Yeah, with LayerNorm.

```python
# from https://github.com/facebookresearch/DiT/blob/main/models.py

def modulate(x, shift, scale):
    return x * (1 + scale.unsqueeze(1)) + shift.unsqueeze(1)


class DiTBlock(nn.Module):
    def __init__(self):
        ...
        self.adaLN_modulation = nn.Sequential(
            nn.SiLU(),
            nn.Linear(hidden_size, 6 * hidden_size, bias=True)
        )
        ...

    def forward(self, x, c):
        shift_msa, scale_msa, gate_msa, shift_mlp, scale_mlp, gate_mlp = self.adaLN_modulation(c).chunk(6, dim=1)
        x = x + gate_msa.unsqueeze(1) * self.attn(modulate(self.norm1(x), shift_msa, scale_msa))
        x = x + gate_mlp.unsqueeze(1) * self.mlp(modulate(self.norm2(x), shift_mlp, scale_mlp))
        return x
```

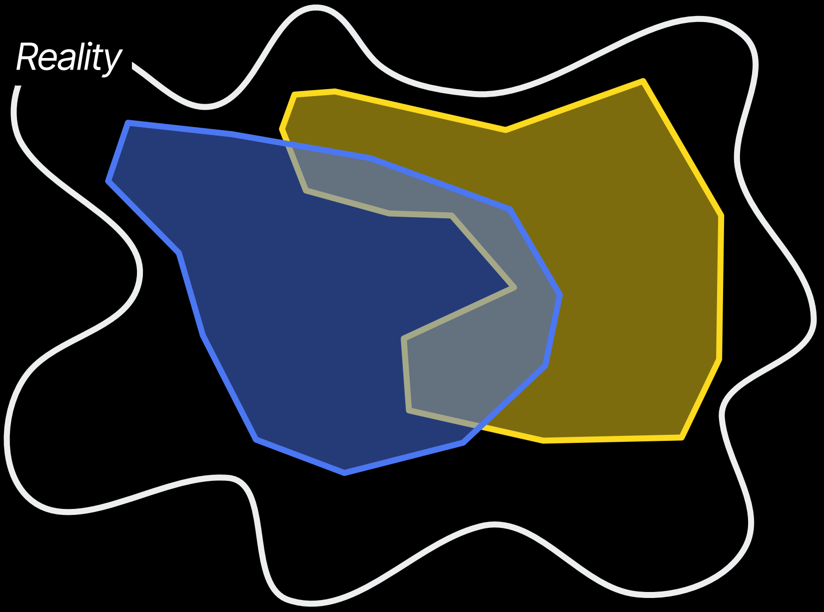


DiT: Peebles and Xie, 2023

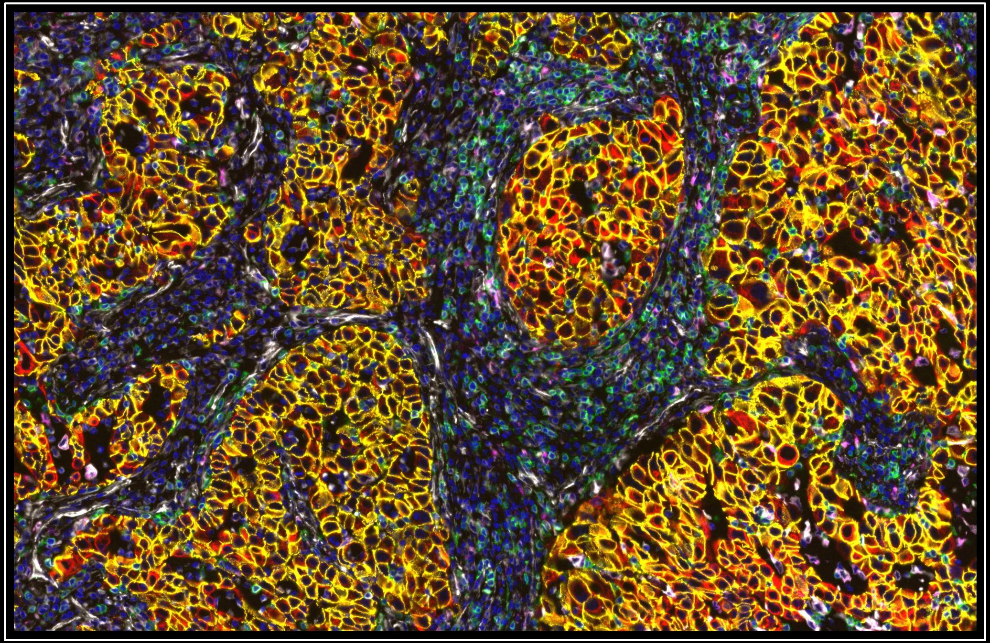
So: lots of options for learning across modalities, especially when everything is just token soup

A brief and very incomplete tour of ideas:

1. Learning joint embedding spaces
2. Concatenation of inputs (early fusion)
3. Cross-attention
4. Concatenation of tokens
5. Layernorm context



Today's topics



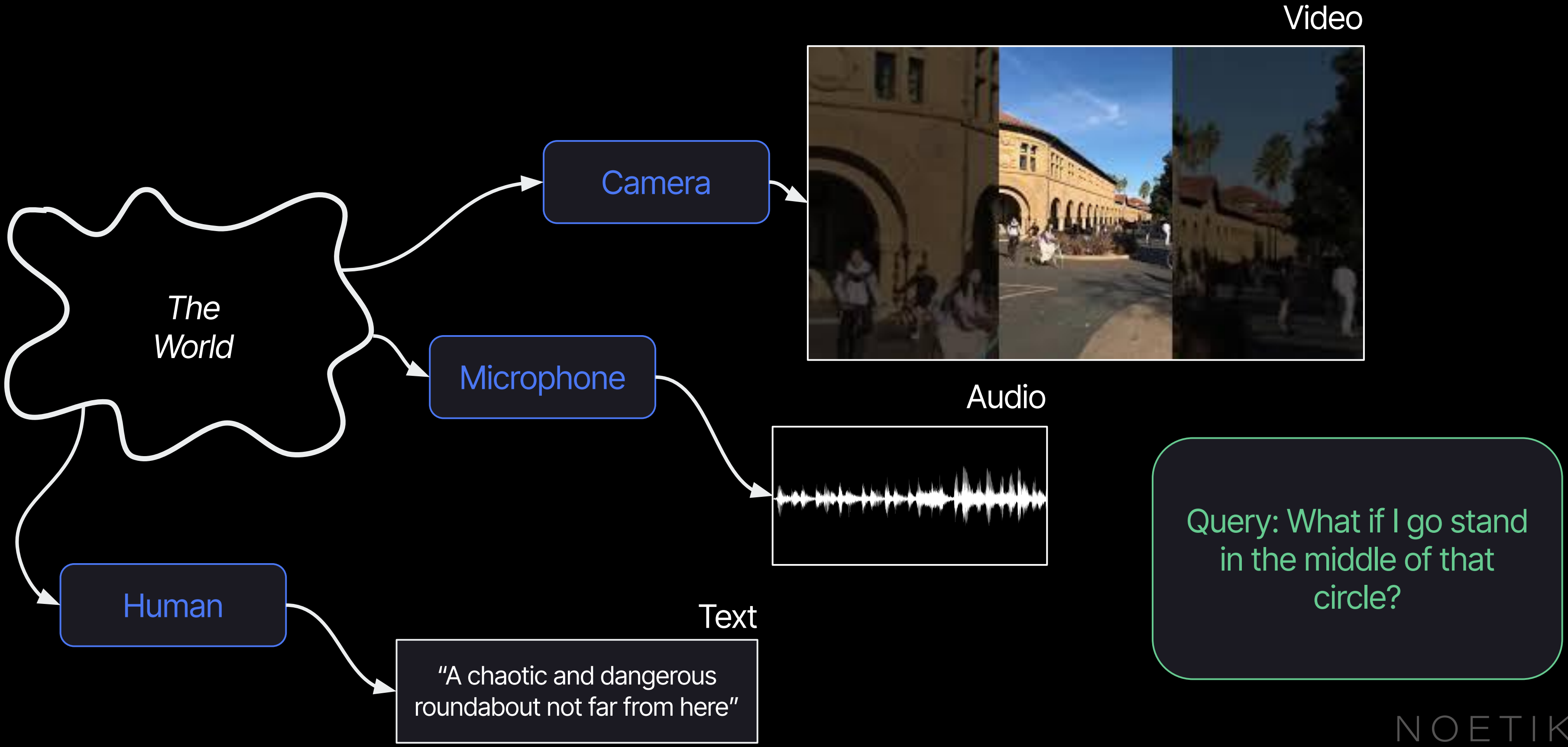
1 | Multimodal Model Madness

2 | Cracking Cancer con Context

3 | Futuristic figures + Follow-ups

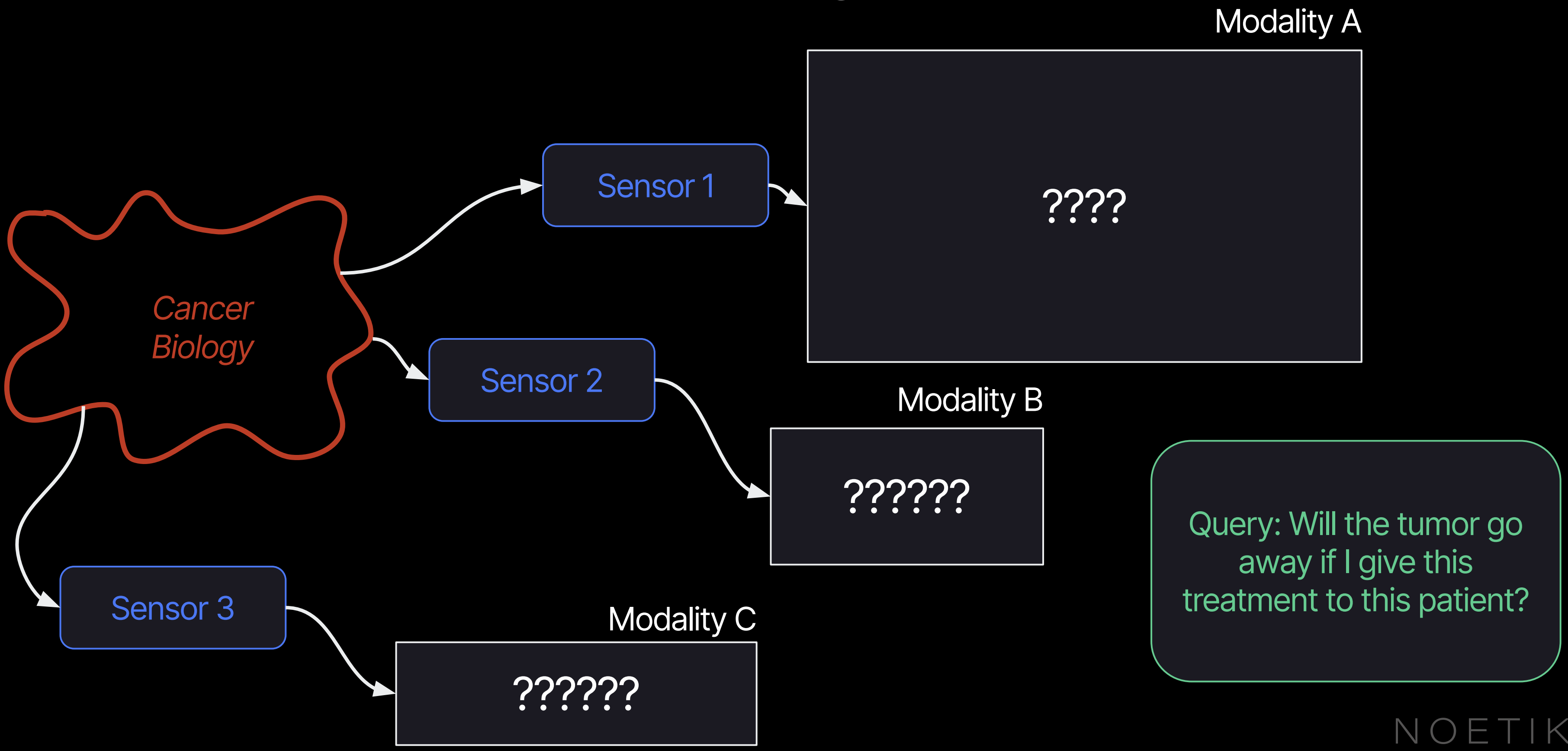
NOETIK

Models of the macroscopic world



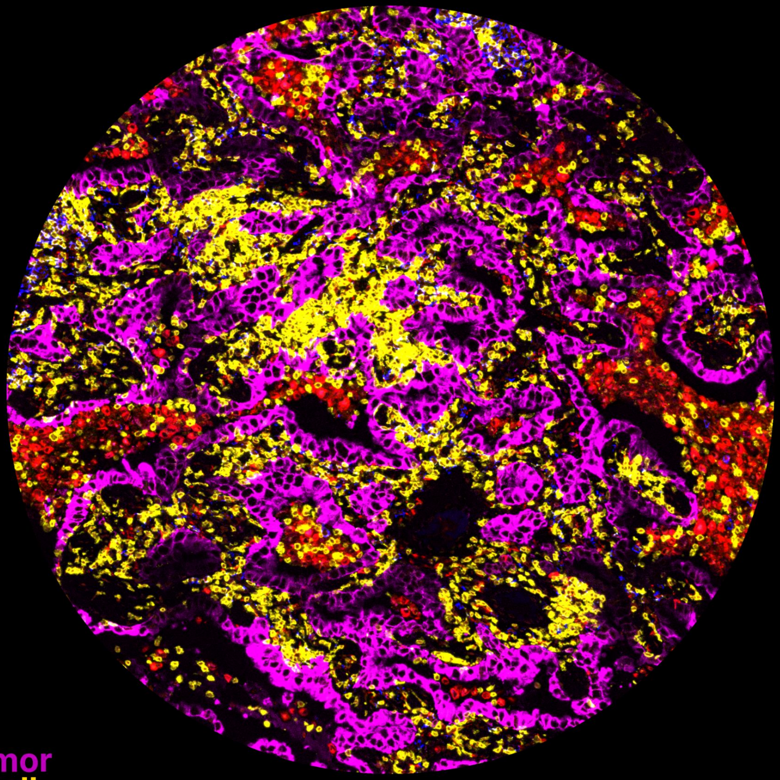
NOETIK

A world model for tumor biology



NOETIK

Cancer immunology in one slide (sorry, immunologists 😬)



- The immune system can detect and destroy cancer, but tumors evolve to hide or suppress immune responses.

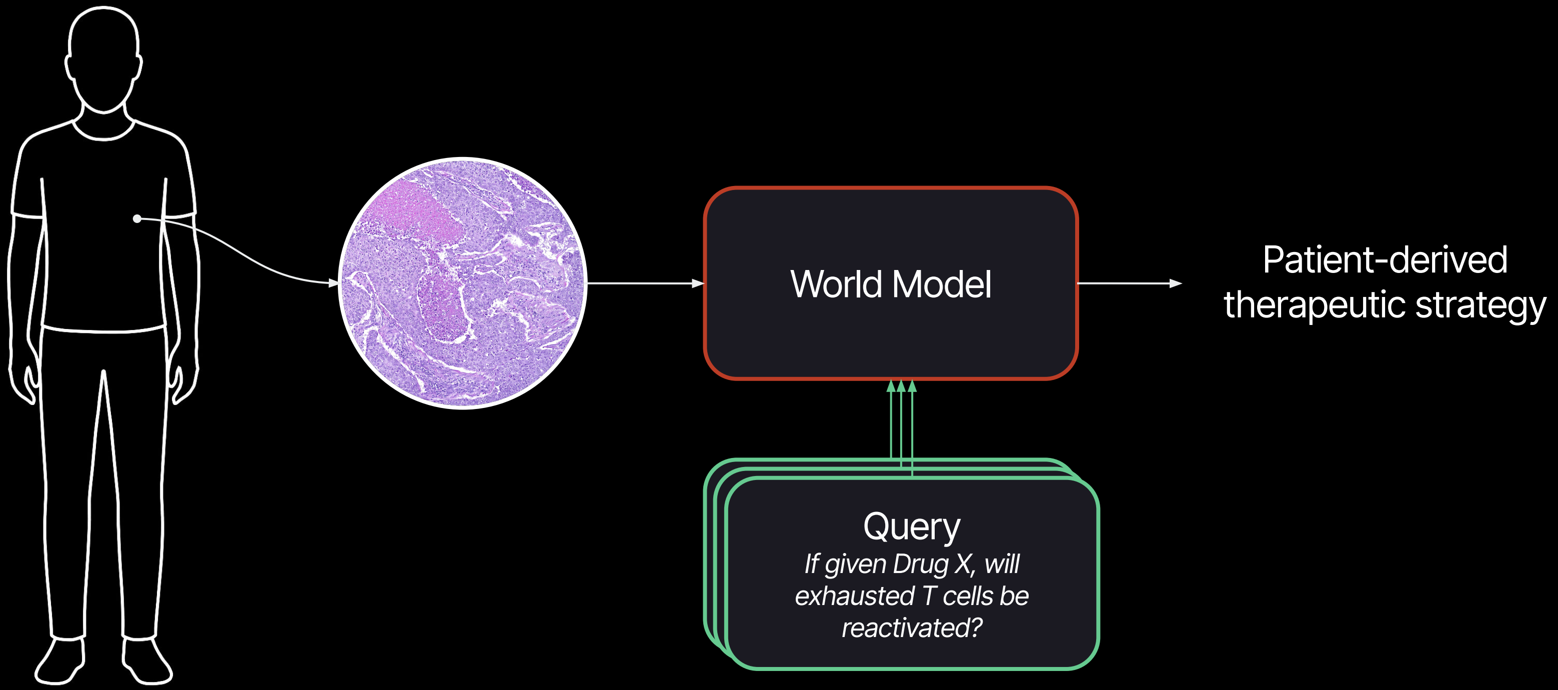
- Immunotherapy boosts or reactivates immune cells to target and kill cancer cells more effectively.

- We need both 1) new drugs and 2) better ways to target the right drug to the right patient. So, we need a model of the tumor-immune world that lets us run realistic simulations.

Query: Will the tumor go away if I give this treatment to this patient?

NOETIK

In an ideal world, we just have a simulator of tissue-level biology of any patient that comes into the clinic



Noetik's huge (and growing) multimodal dataset of cancer biology

1042
Human Lung tumor specimens

1800
Slides processed

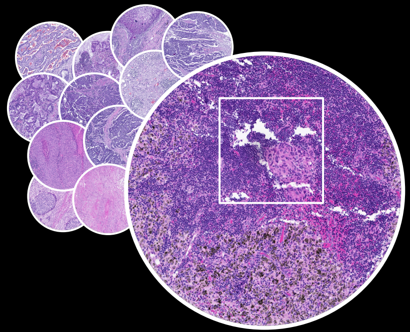
1.5
Petabyes of multimodal spatial data generated

40
Million cells of spatial transcriptomics (> 2 percent of all CosMX data)

Noetik is continuously building a massive multimodal dataset of cancer biology

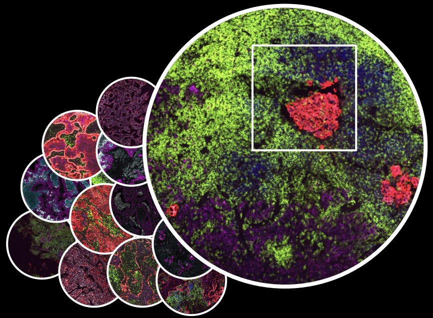
H&E (haematoxylin and eosin)

- Cheap and easy to acquire; ubiquitous
- Highlights gross morphology
- Most similar to RGB images in other ML/CV contexts



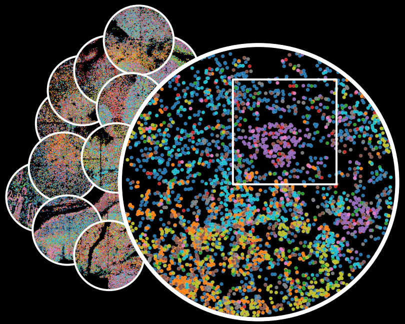
Protein

- 16-plex immunofluorescence panel highlighting tumor and immune markers

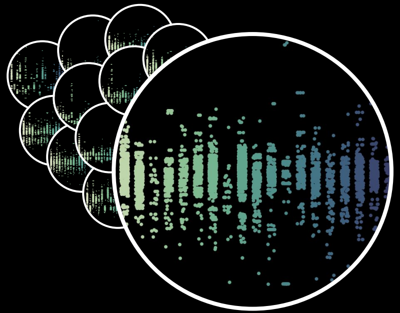


Spatial Transcriptomics

- 1000-plex measurement of RNA
- Perfectly aligned to H&E and Protein
- Richest and most complicated



Genetic Sequencing

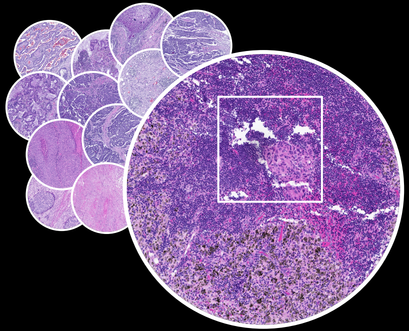


- Identify mutations in key genes

Noetik is continuously building a massive multimodal dataset of cancer biology

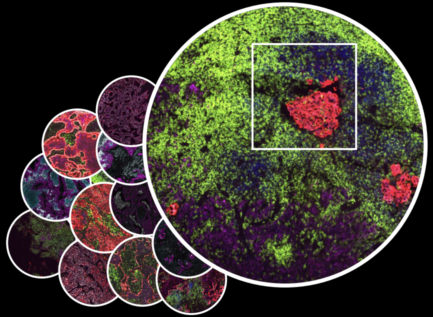
H&E (haematoxylin and eosin)

- Cheap and easy to acquire; ubiquitous
- Highlights gross morphology
- Most similar to RGB images in other ML/CV contexts



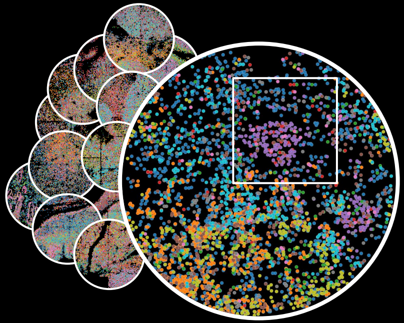
Protein

- 16-plex immunofluorescence panel highlighting tumor and immune markers

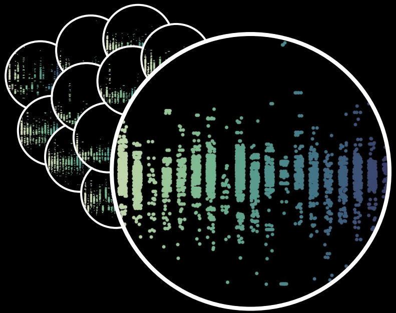


Spatial Transcriptomics

- 1000-plex measurement of RNA
- Perfectly aligned to H&E and Protein
- Richest and most complicated

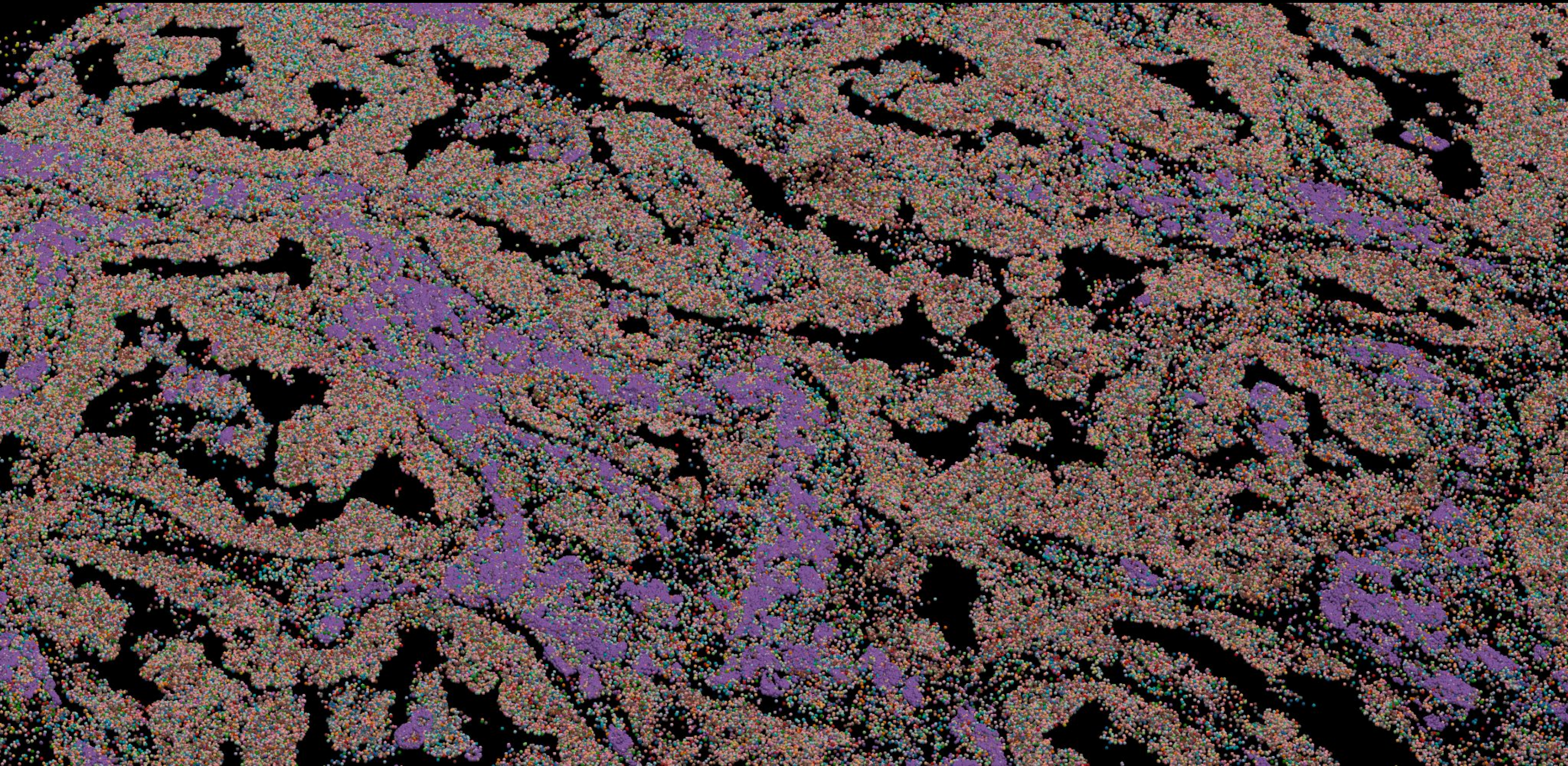


Genetic Sequencing

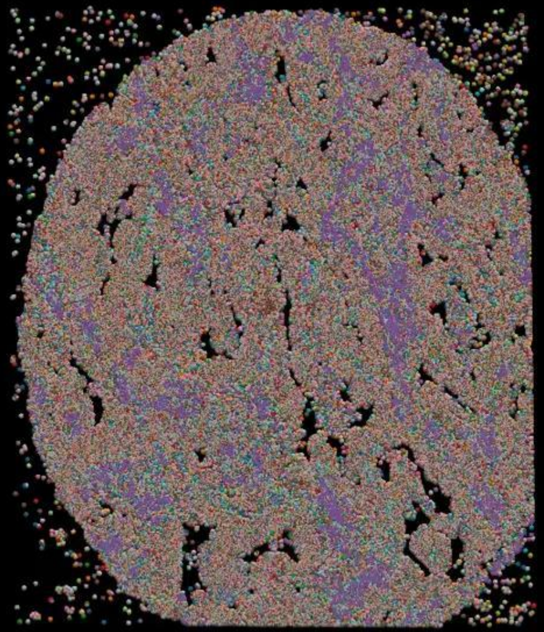


- Identify mutations in key genes

Spatial transcriptomics data are incredibly rich and complex



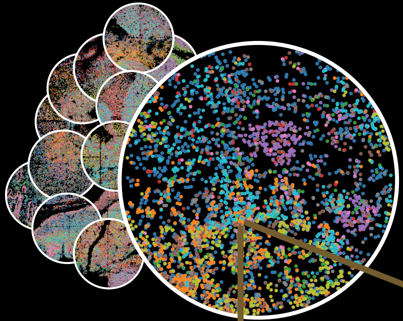
Spatial transcriptomics data are incredibly rich and complex



NOETIK

Spatial transcriptomics data are incredibly rich and complex

Multiple "cores" per patient

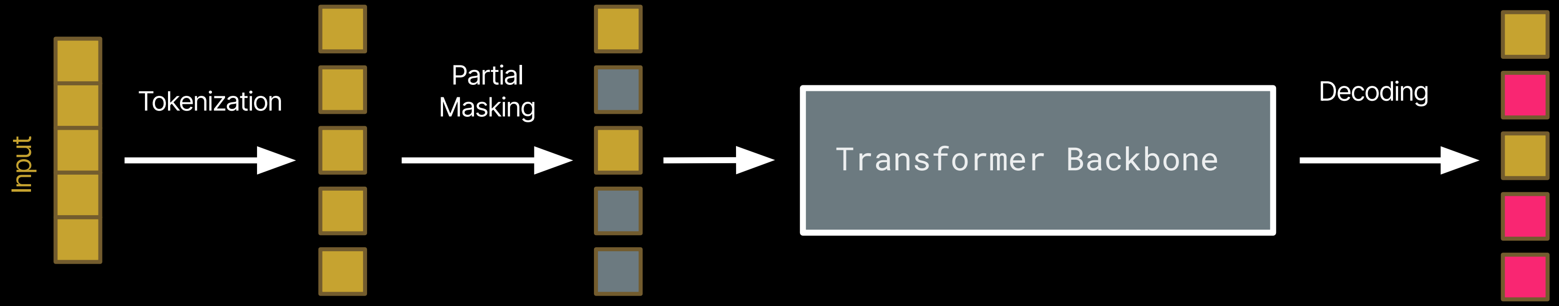


Thousands of cells per core

Thousands of genes per cell, but sparse

Gene	Reads
CD3E	6
CD20	0
KRT19	0
CD8A	7
...	...

Masked autoencoding is a flexible and powerful framework for learning world models

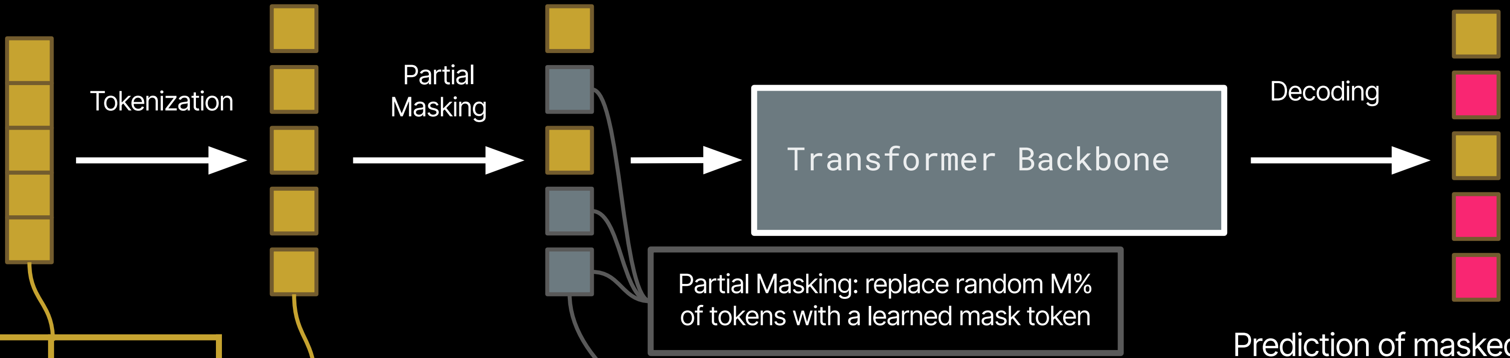


Prediction of masked tokens: loss is computed over masked tokens only

See also Counterfactual World Modeling by Bear et al., 2023 (https://arxiv.org/abs/2306.01828)

NOETIK

A model that predicts masked gene counts...



Prediction of masked tokens: loss is computed over masked tokens only

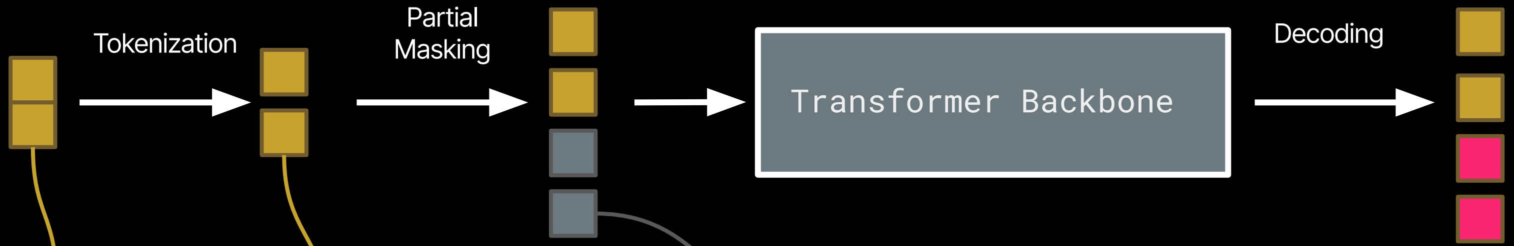
Gene	Reads
CD3E	6
CD20	0
KRT19	0
CD8A	7
...	...

idx	token
0	t_0
1	t_1
2	t_2
3	t_3
4	t_4

idx	token
0	t_0
1	[MASK]
2	t_2
3	[MASK]
4	[MASK]

NOETIK

At inference time: can provide a "prompt" and mask out the rest



Predict all masked tokens

Gene	Reads
CD3E	3
KRT19	0

idx	token
0	t_0
1	t_1

idx	token
0	t_0
1	t_1
2	[MASK]
3	[MASK]

Prompt: "You are a T cell and not a tumor"

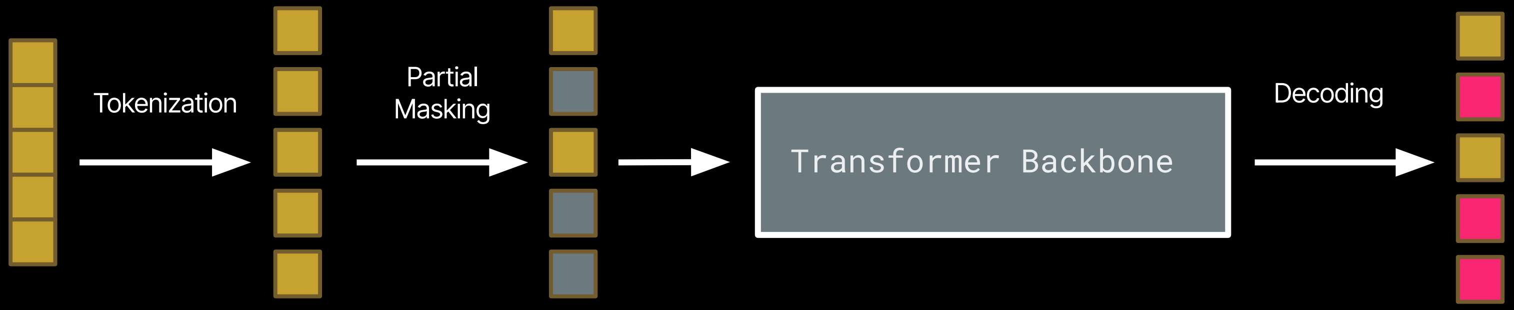
Note: the only information available to the predictor is the prompt (and of course, knowledge of the entire dataset via training)

Latent space of this model useful for identifying cell type and state.

Can also run "counterfactuals", e.g. by asking what rest of genes look like if CD3E is 3x as high.

NOETIK

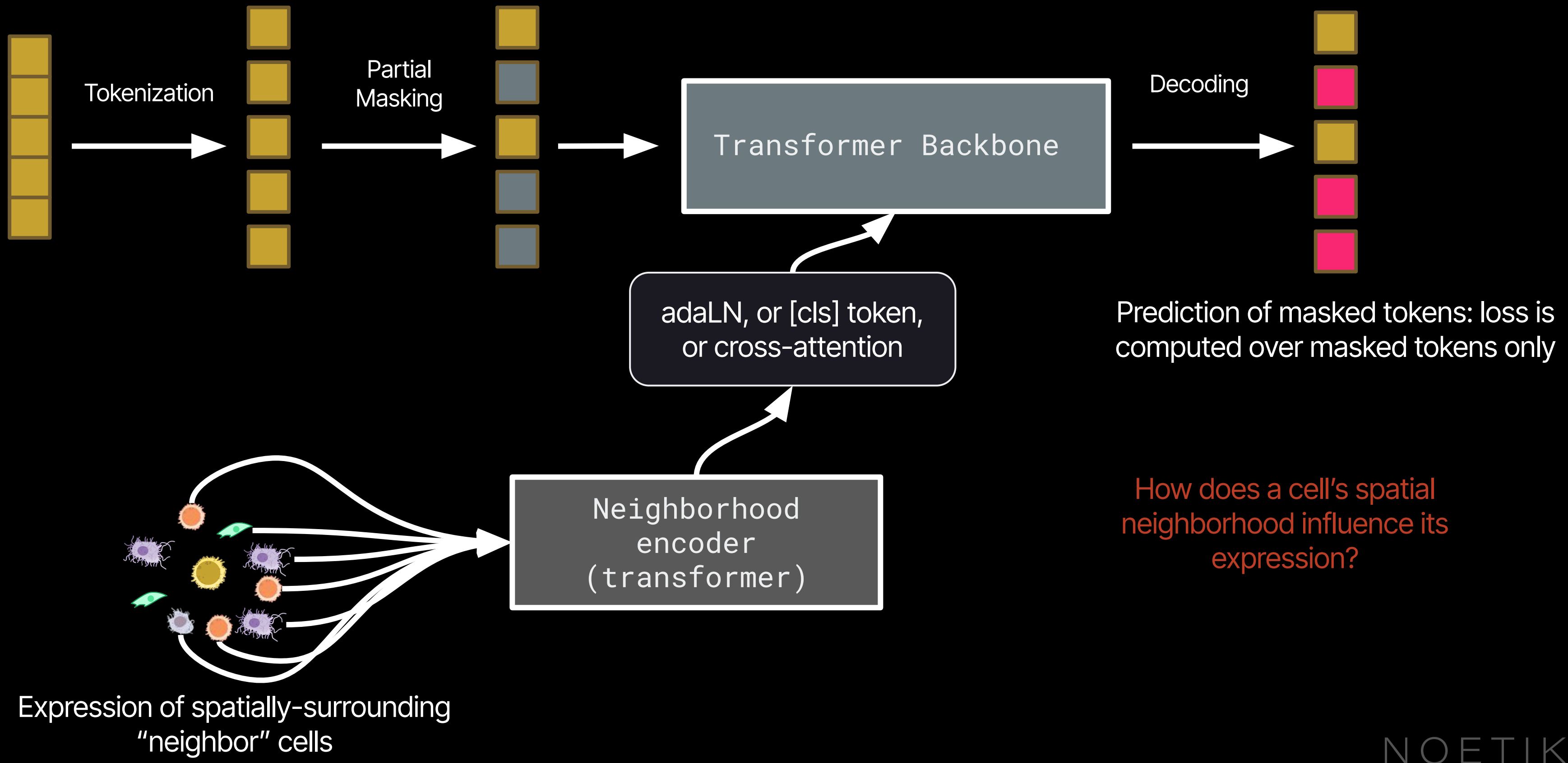
Not quite "multimodality" but similar: virtual cells embedded in spatial neighborhoods



Prediction of masked tokens: loss is computed over masked tokens only

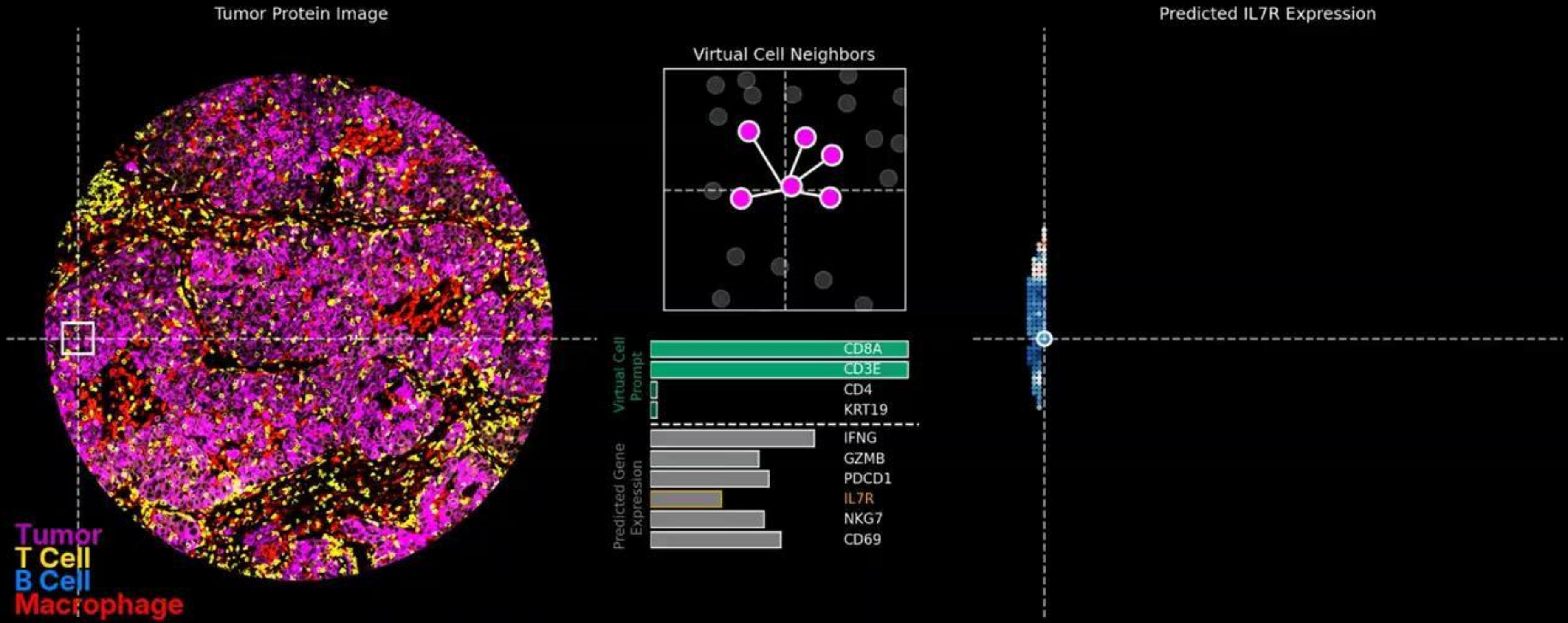
NOETIK

Not quite "multimodality" but similar: virtual cells embedded in spatial neighborhoods



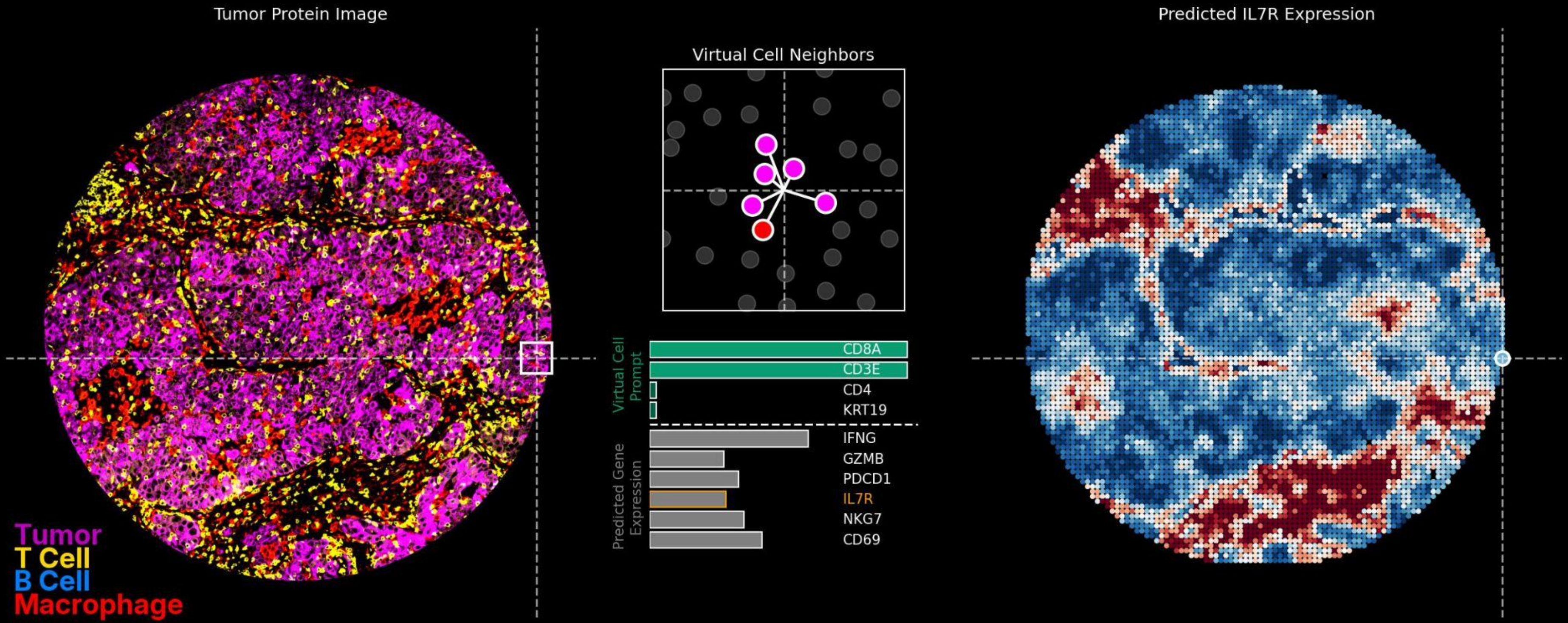
How does a cell's spatial neighborhood influence its expression?

Not quite "multimodality" but similar: virtual cells embedded in spatial neighborhoods



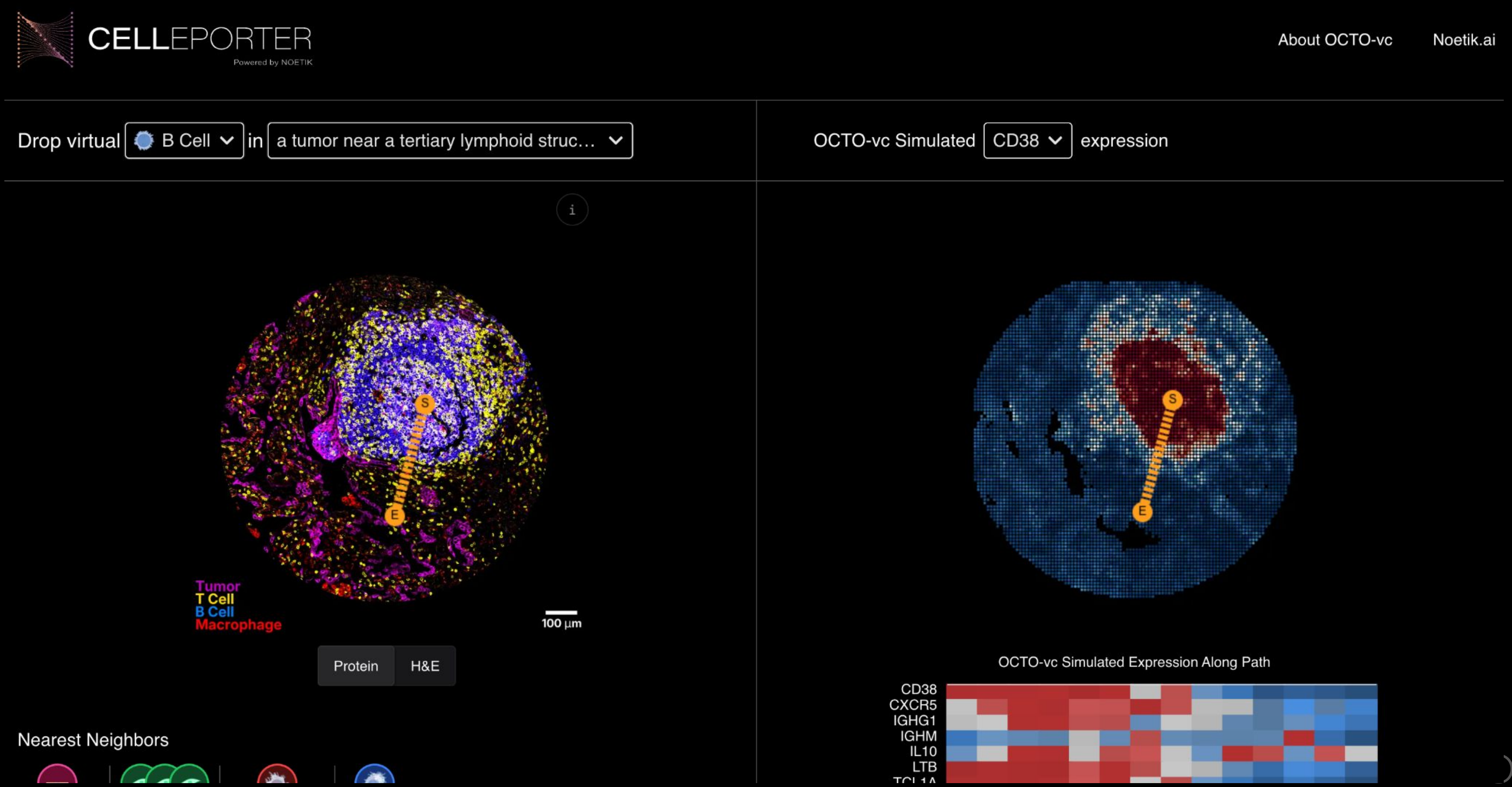
NOETIK

Not quite "multimodality" but similar: virtual cells embedded in spatial neighborhoods



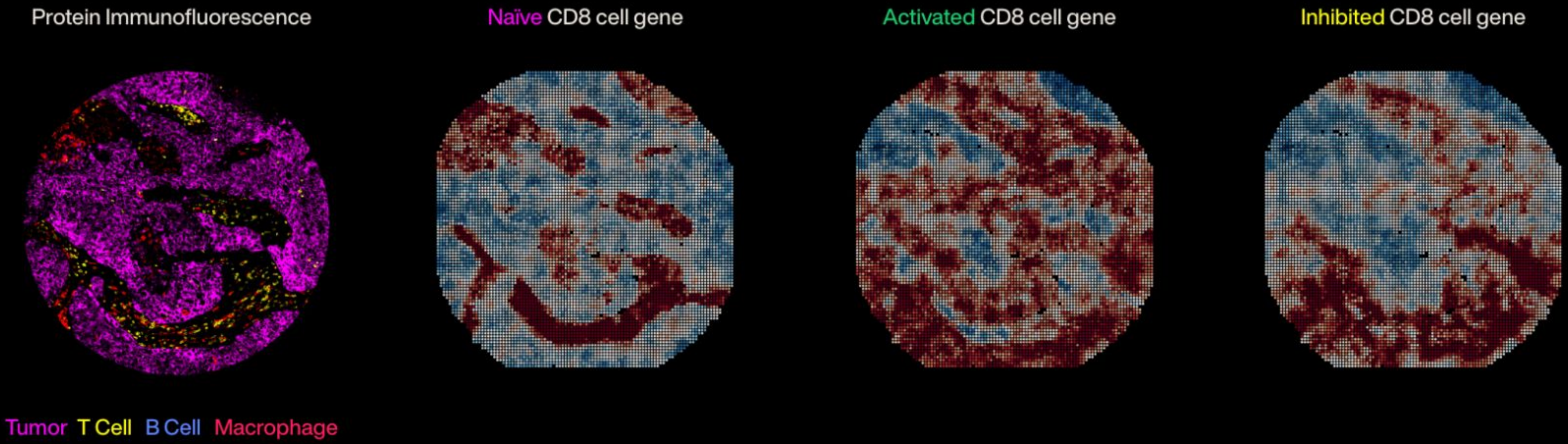
NOETIK

Explore for yourself at celleporter.ai



Virtual cell predictions depend on 1) prompt and 2) spatial context

Same prompt, different context, different predicted genes

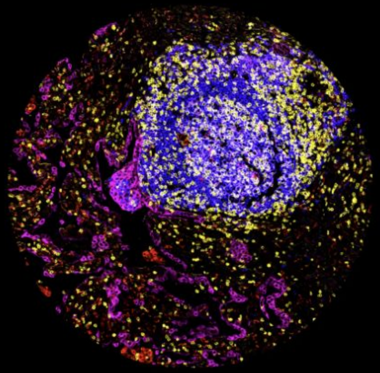


NOETIK

Virtual cell predictions depend on 1) prompt and 2) spatial context

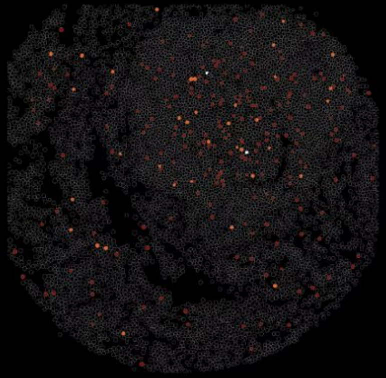
Different prompt, different context

Protein Immunofluorescence

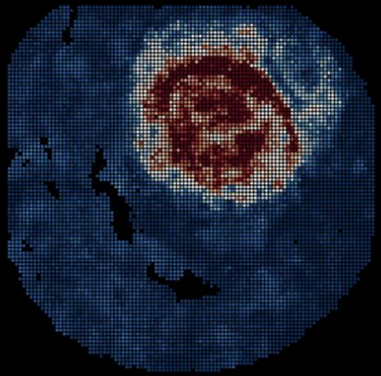


Tumor T Cell B Cell Macrophage

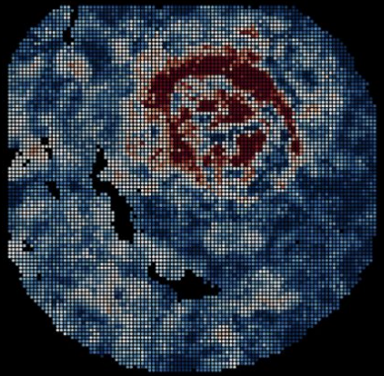
True PD1 (all cells)



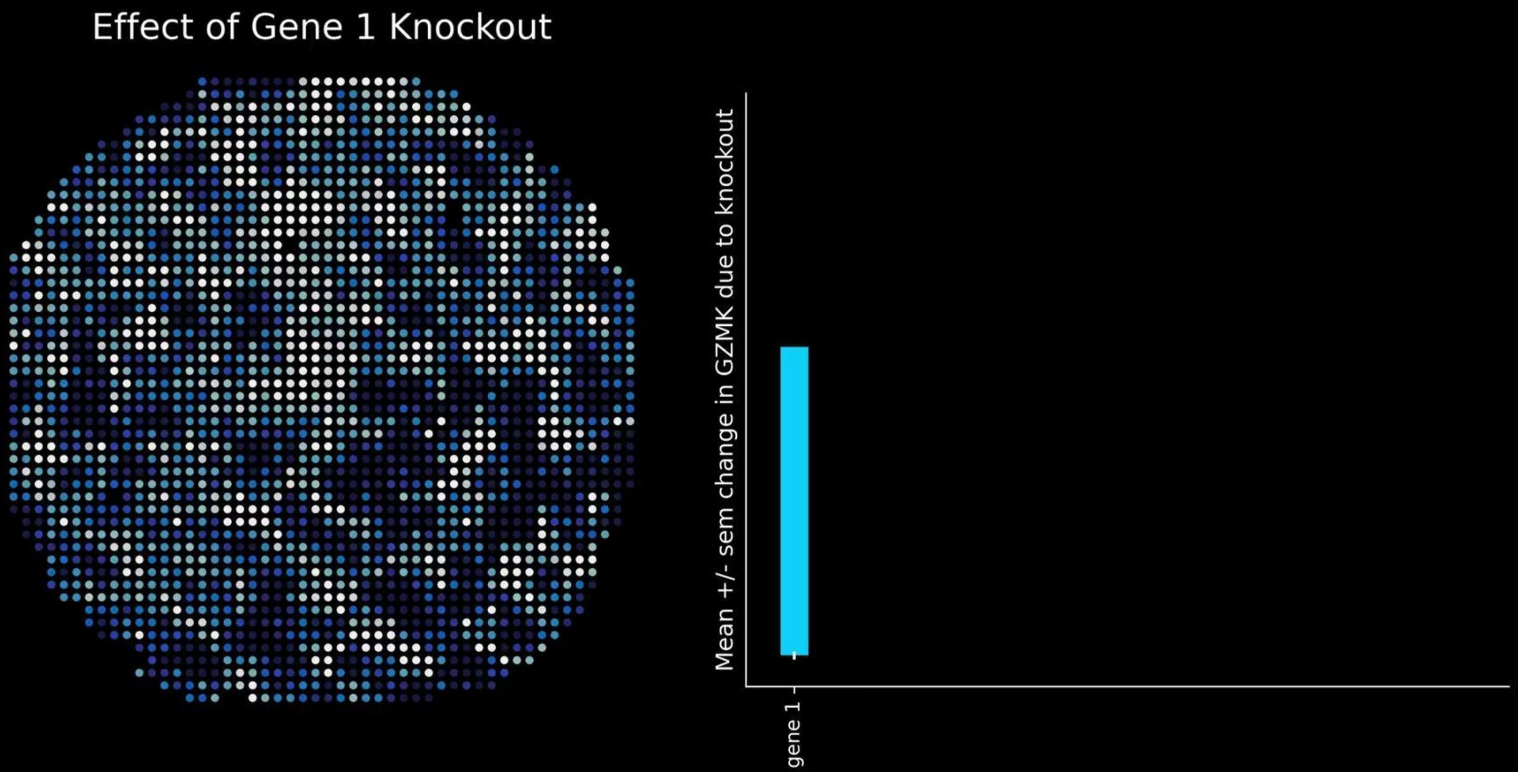
Simulated PD1 (Virtual CD4 T Cells)



Simulated PD1 (Virtual CD8 T Cells)

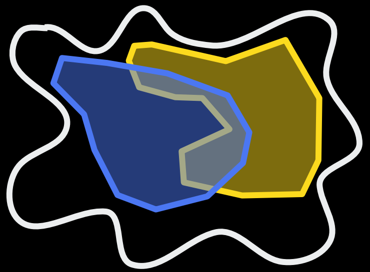


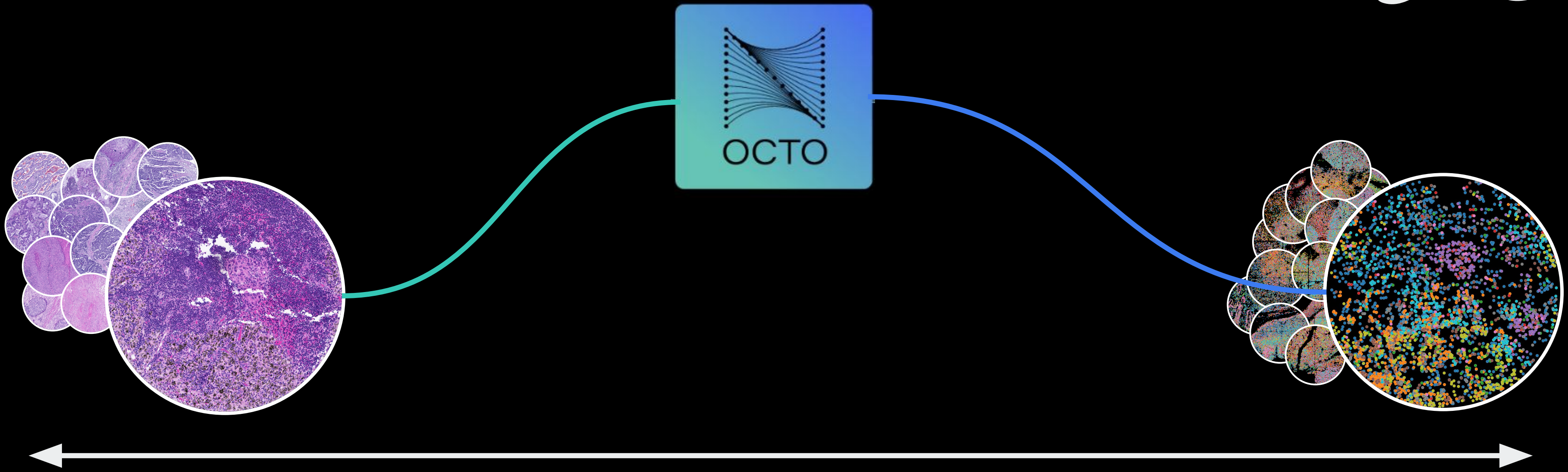
Using virtual cell predictions to run counterfactual simulations



ETIK

Multimodal learning to impute data

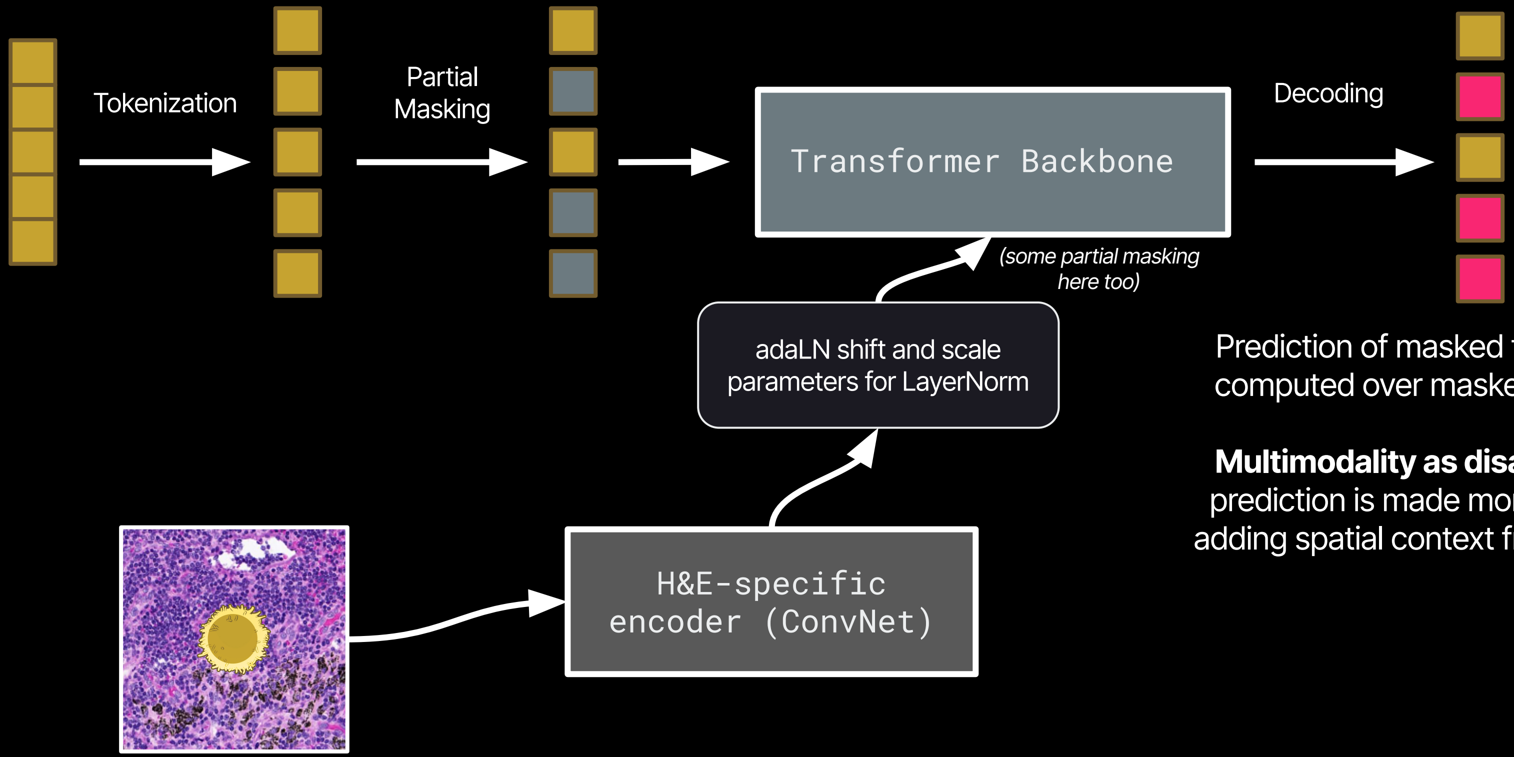




Cheap, ubiquitous

Expensive, rare

A multi-modal model that predicts masked gene counts… conditioned on aligned H&E images

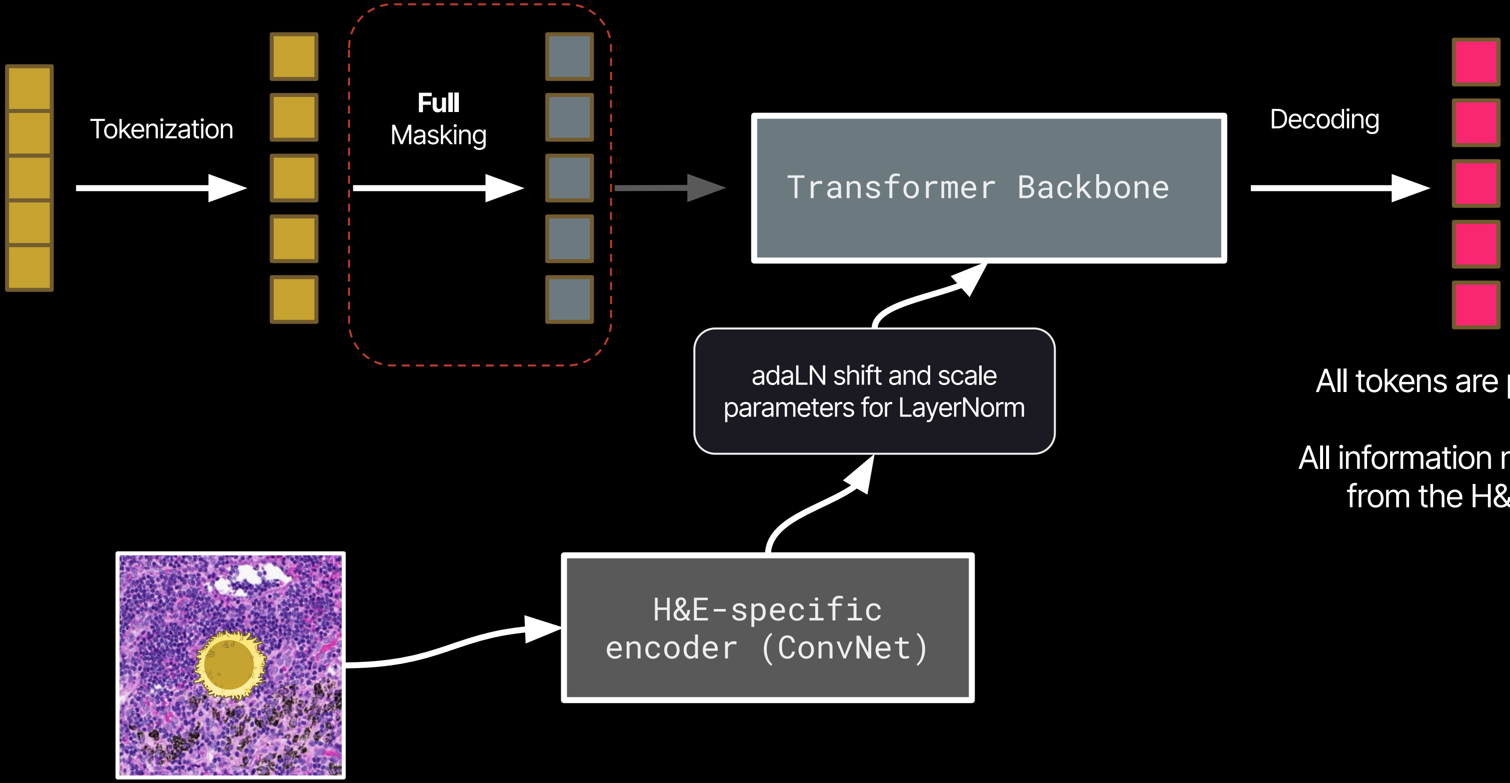


H&E image centered on the coordinates of this particular cell

Prediction of masked tokens: loss is computed over masked tokens only

Multimodality as disambiguation: prediction is made more accurate by adding spatial context from H&E image

You can use this model for “translation” between modalities



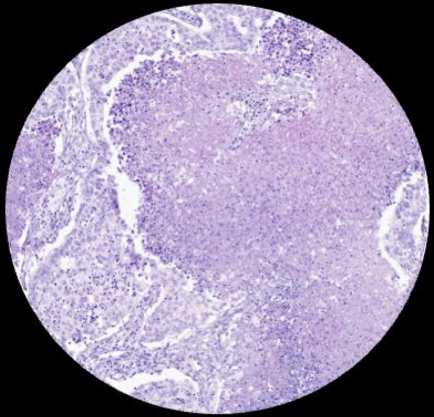
H&E image centered on the coordinates of this particular cell

All tokens are predicted

All information must come from the H&E input

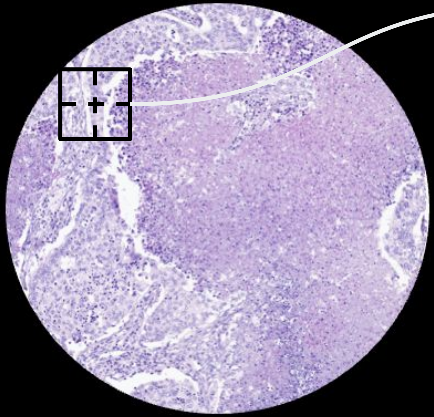
Model accurately predicts expression of genes from H&E alone

H&E



Model accurately predicts expression of genes from H&E alone

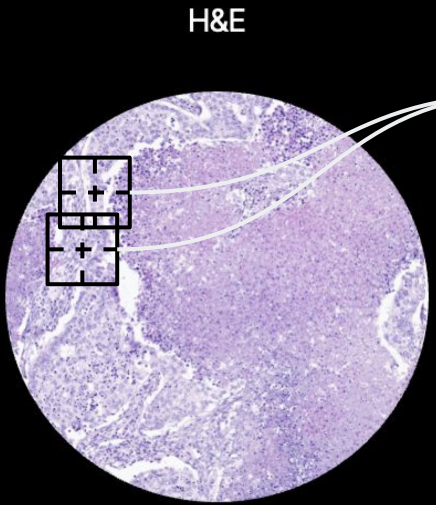
H&E



Inference with all genes masked: prediction depends only on H&E input

Pred KRT19 = 0.1

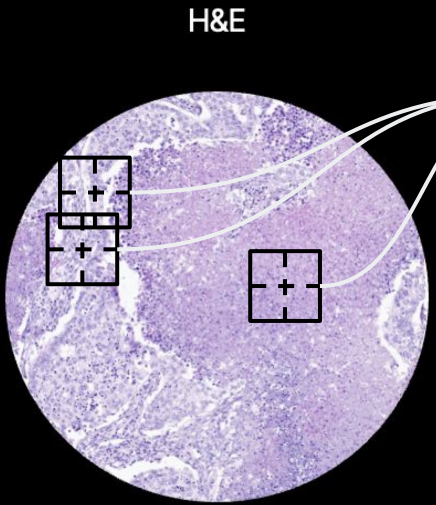
Model accurately predicts expression of genes from H&E alone



Inference with all genes masked: prediction depends only on H&E input

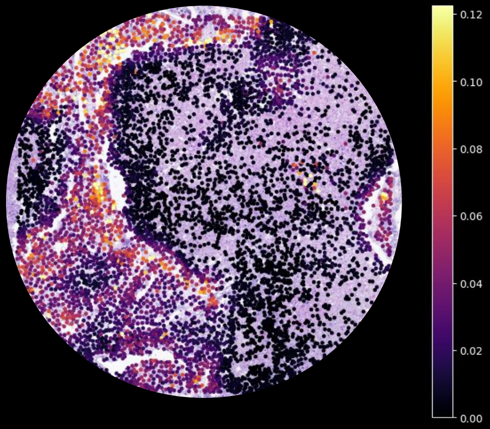
Pred KRT19 = 0.09

Model accurately predicts expression of genes from H&E alone



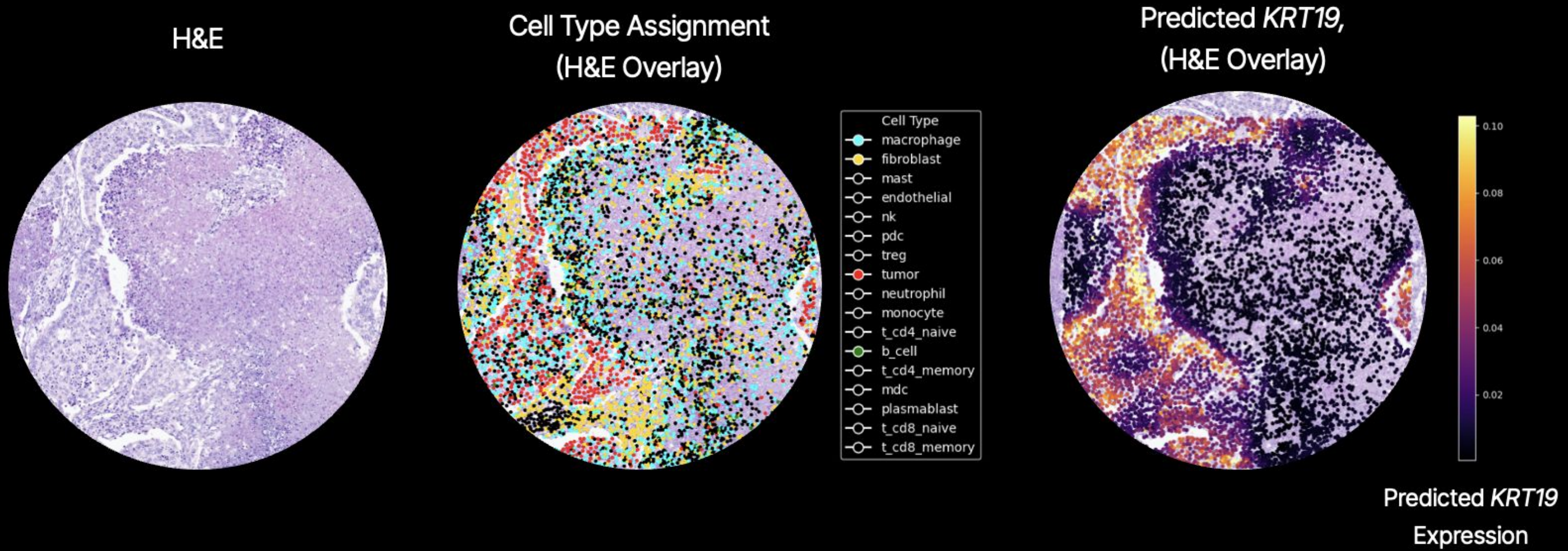
Inference with all genes masked: prediction depends only on H&E input

Pred KRT19 = 0.001

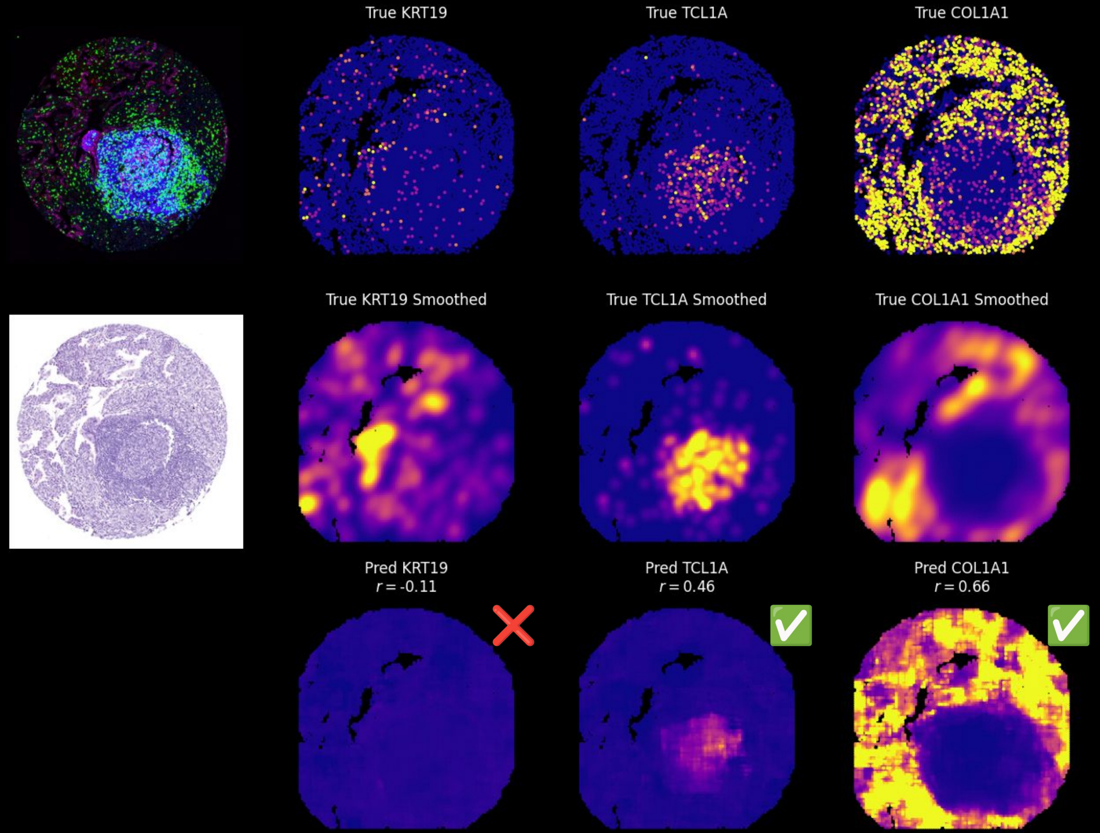


NOETIK

Model accurately predicts expression of genes from H&E alone



Imputation is accurate, moreso for some genes than others



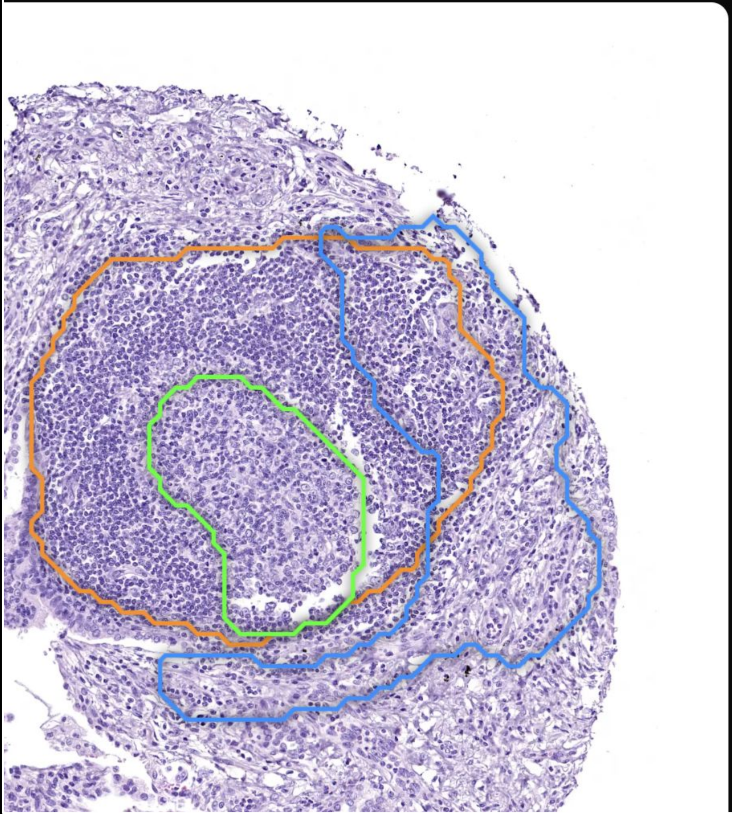
Real transcript detections

Interpolated transcript detections

Model prediction

NOETIK

Aside: this capability lets us combine model predictions with LLMs to build some pretty cool tools!



Cytotoxic T cell activation

Top genes like CD8A, NKG7, and CCL5 are signature markers of cytotoxic T lymphocytes. Downregulated keratins and epithelial markers suggest a shift away from epithelial lineage toward immune activity.

Top 5: CD2, NKG7, CCL5. Bottom 5: KRT19, ENO1, KRT18, KRT8, LGALS3BP

MHC II antigen presentation

B cells or antigen-presenting cells

Immediate early response cells

Activated B cells

Top genes like MS4A1 (CD20), CD19, and TNFRSF13B are B cell-specific activation markers. Bottom genes include ribosomal and immature lymphoid markers, suggesting a shift to a mature, activated B cell phenotype.

Top 5: HBB, CD19, MS4A1. Bottom 5: RPL34, PTPRC, ITGAX, RPL21, TCL1A

T cell–rich immune response

Mature plasma cells

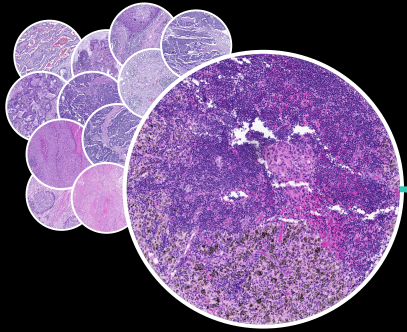
Stress-responding epithelial cells

Activated T cells with B cell interaction

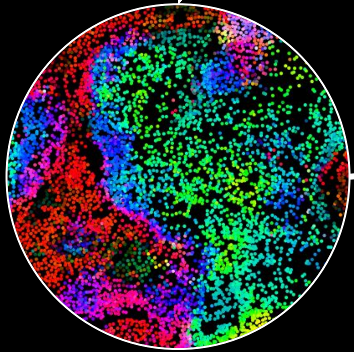
CD2, CD69, and JUNB are T cell activation markers, while IGKC, IGHG1/2 reflect interaction with B cells or expression in dual-phenotype cells. The downregulation of MHC I and immunoglobulin genes implies a complex interplay of immune states.

NOETIK

A system to translate easy-to-acquire data into rich patient representations that surface therapeutic hypotheses

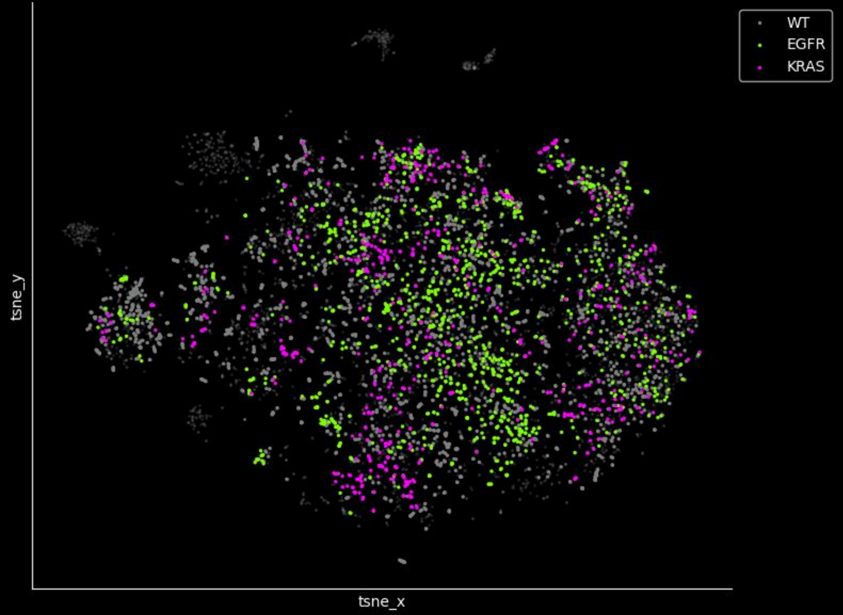


OCTO



1000-dimensional prediction of gene expression at thousands of locations inside each sample (viz: mapping of first 3 PCs to RGB space)

Tumor Genetics

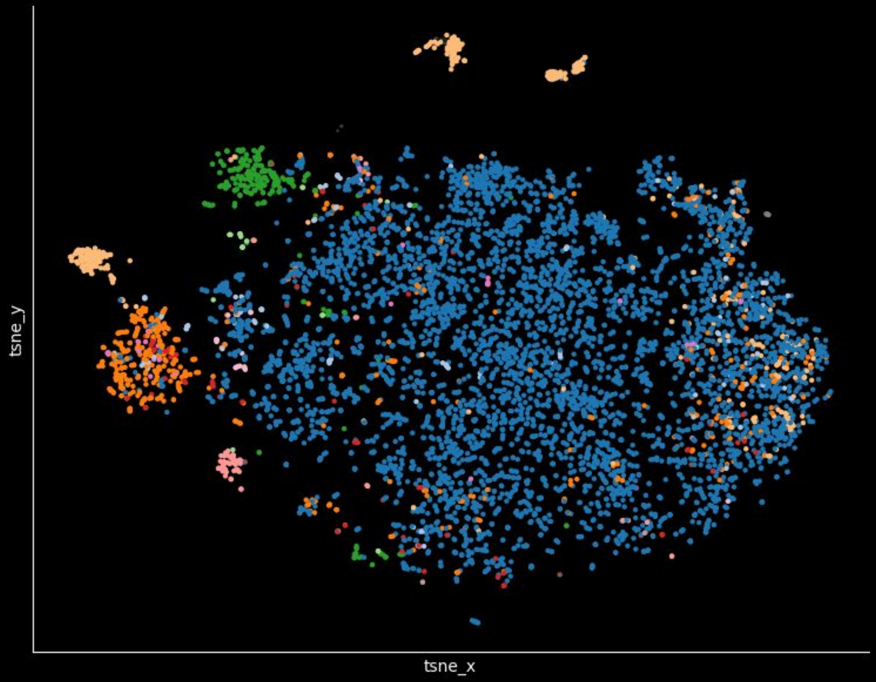


t-SNE dimensionality reduction on predicted gene expression

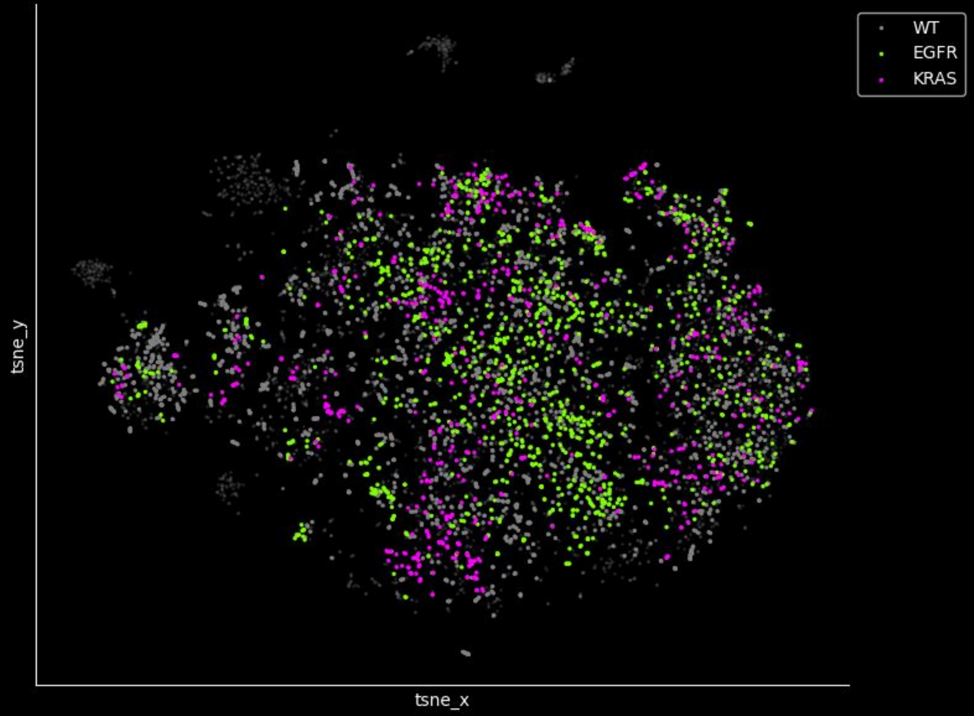
Each dot is one sample from one patient

Embedding spaces produced by billions of simulations recover known biology

Histology



Tumor Genetics

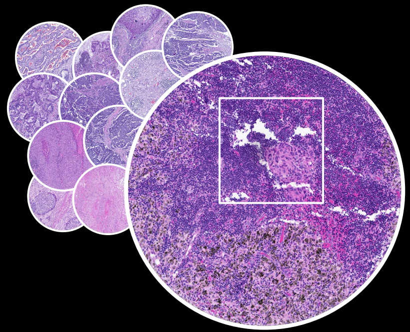


NOETIK

Noetik is continuously building a massive multimodal dataset of cancer biology

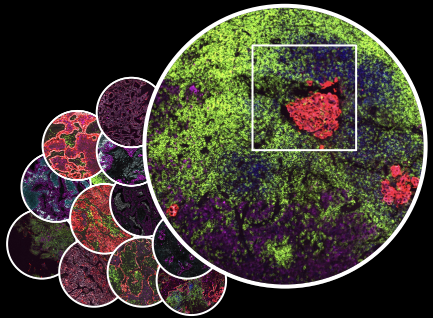
H&E (haematoxylin and eosin)

- Cheap and easy to acquire; ubiquitous
- Highlights gross morphology
- Most similar to RGB images in other ML/CV contexts



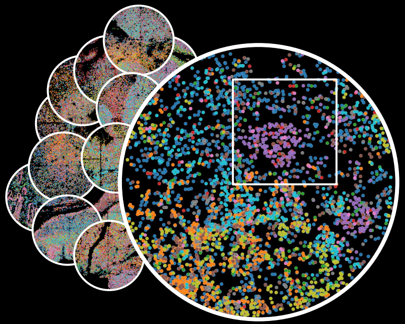
Protein

- 16-plex immunofluorescence panel highlighting tumor and immune markers

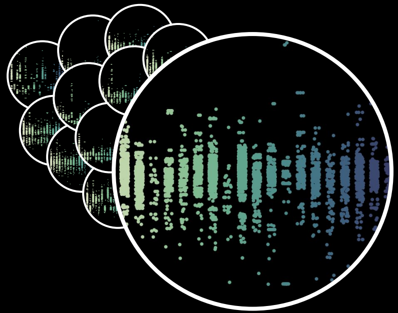


Spatial Transcriptomics

- 1000-plex measurement of RNA expression
- Perfectly aligned to H&E and Protein
- Richest and most complicated



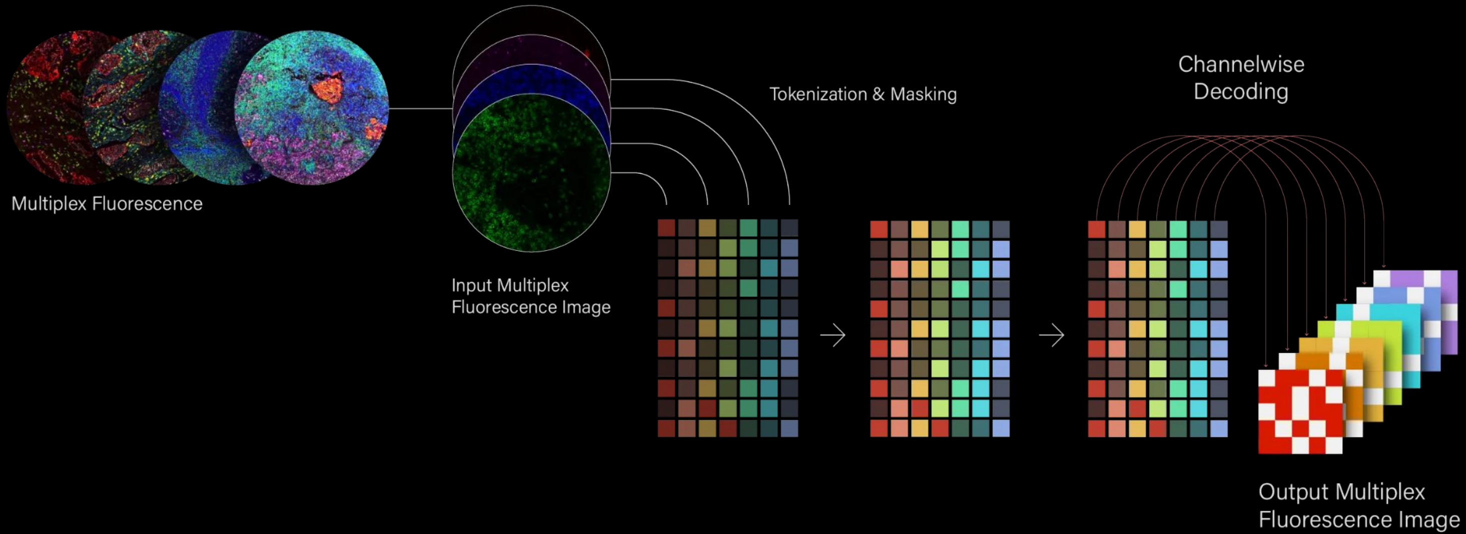
Genetic Sequencing



- Identify mutations in key genes

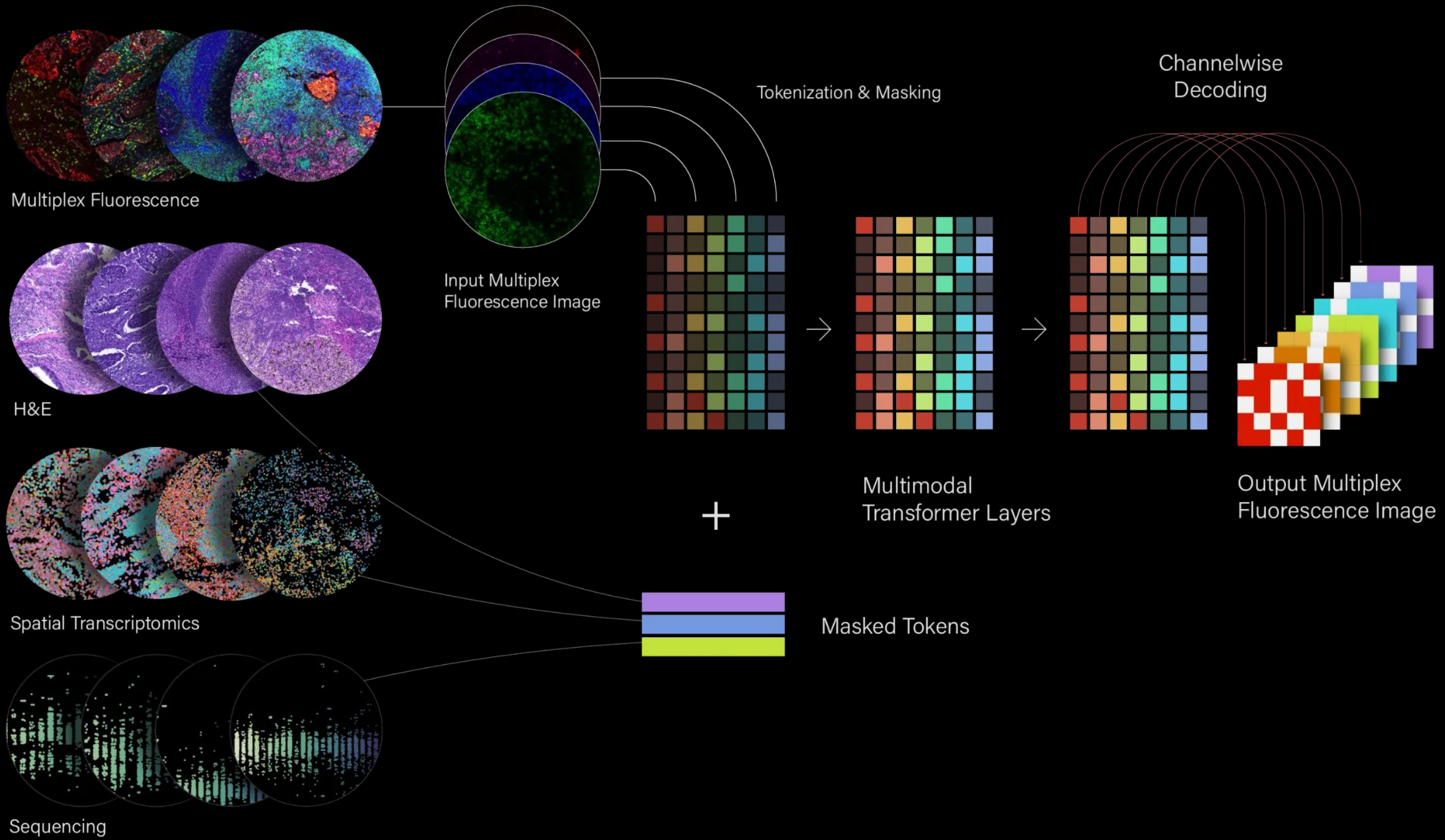
Unimodal transformers for cancer biology

Predicting fluorescence image as target



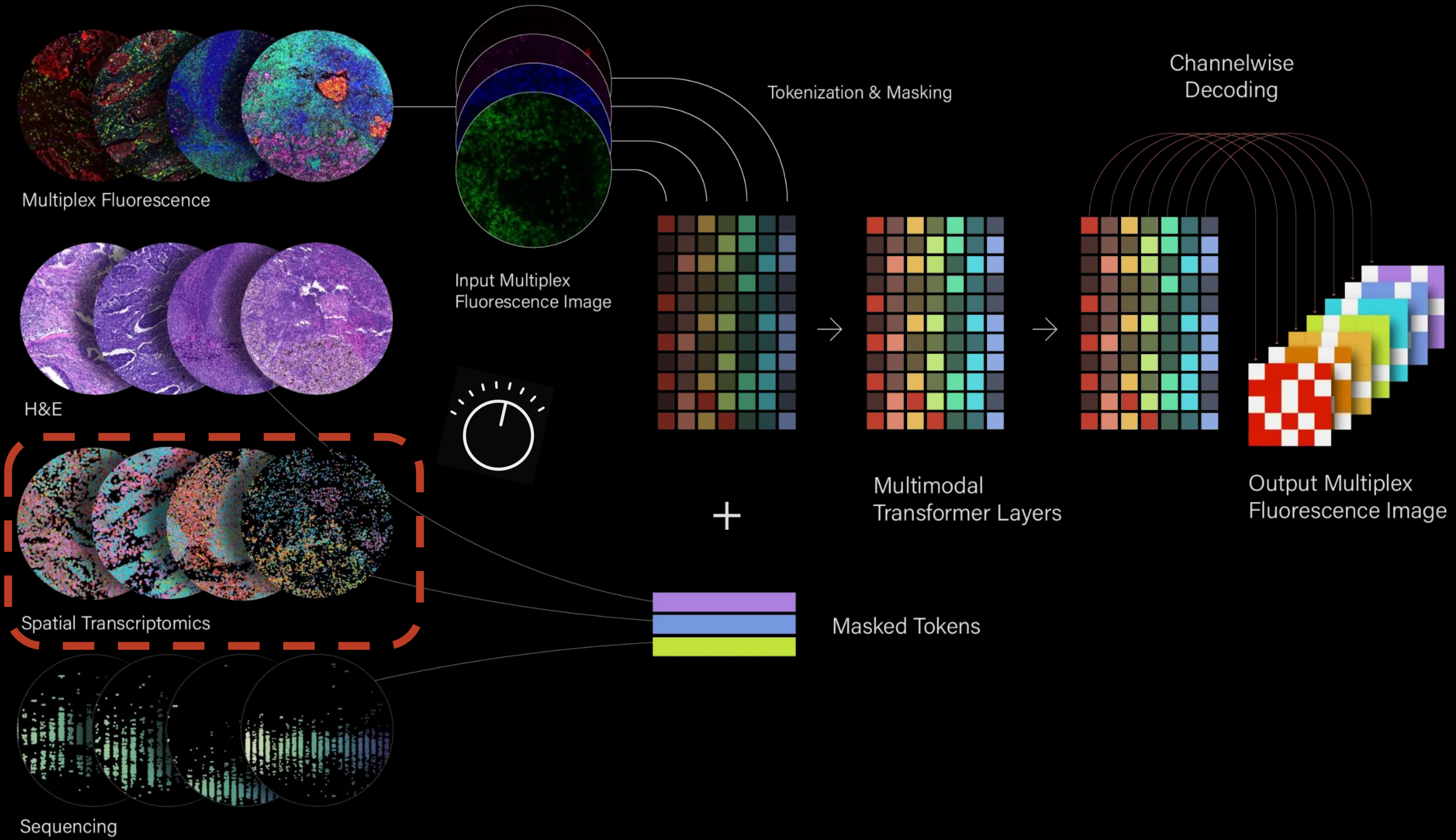
Multimodal transformers for cancer biology

Predicting fluorescence image as target



NOETIK

Multimodal counterfactual simulations: how would prediction change if one of the input modalities changed?



NOETIK

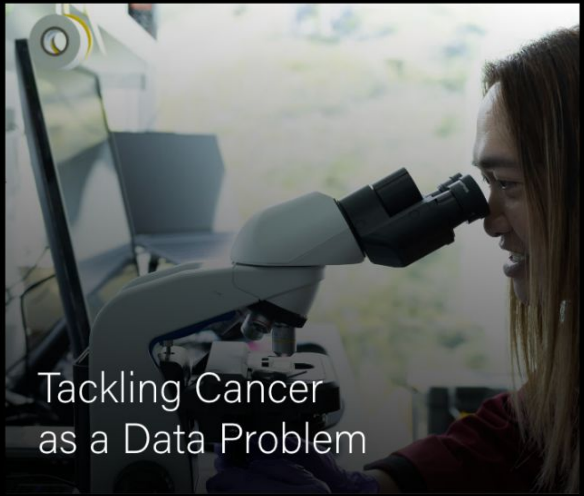
Multimodal counterfactual simulations: how would prediction change if one of the input modalities changed?

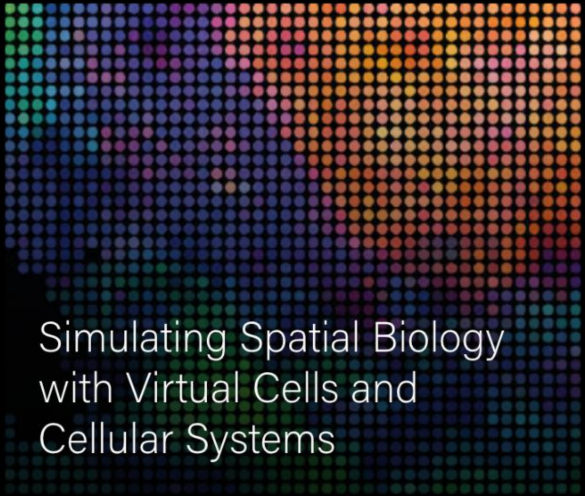
NOETIK

Multimodal counterfactual simulations: how would prediction change if one of the input modalities changed?

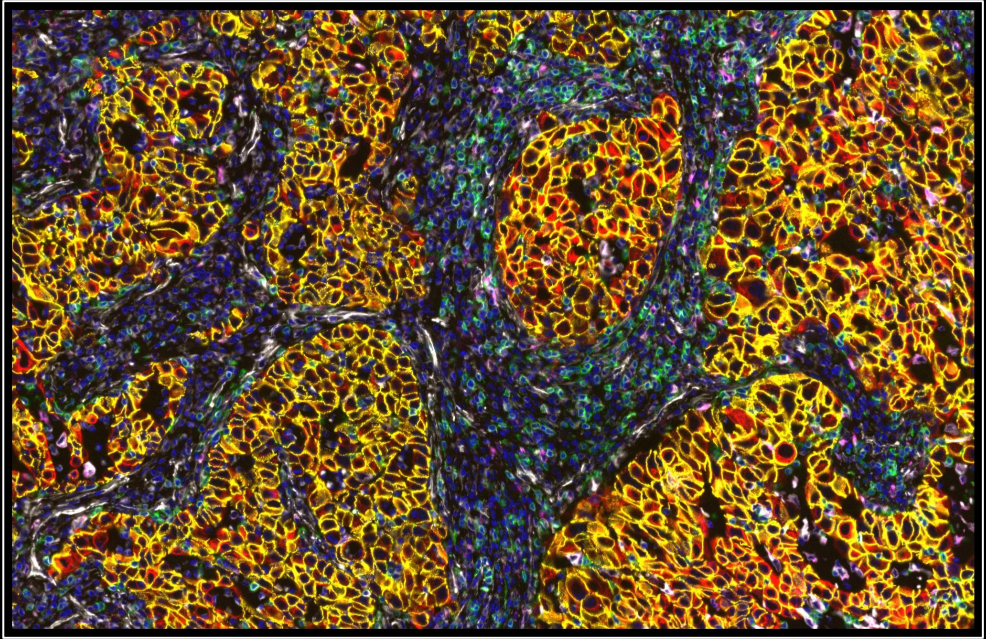
For more: https://www.noetik.ai/research

Research





Today's topics



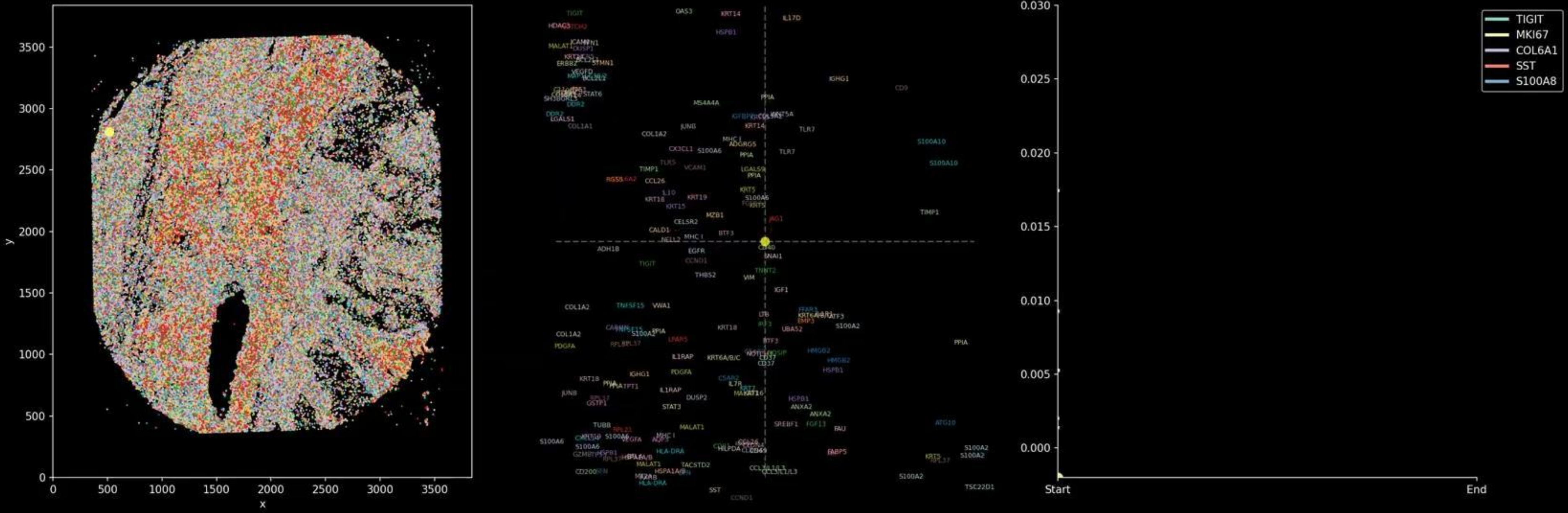
1 | Multimodal Model Madness

2 | Cracking Cancer con Context

3 | Futuristic figures + Follow-ups

NOETIK

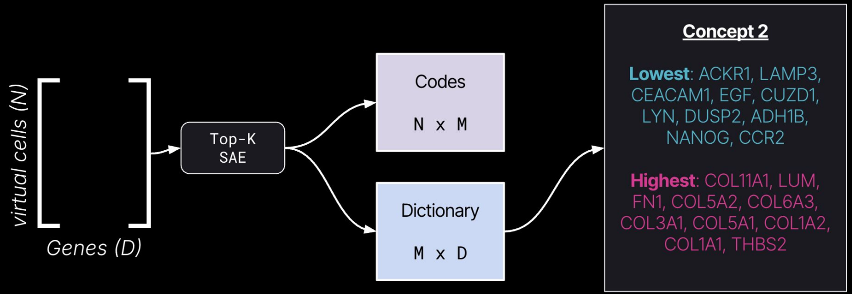
In progress: training on a ton of raw RNA transcript data

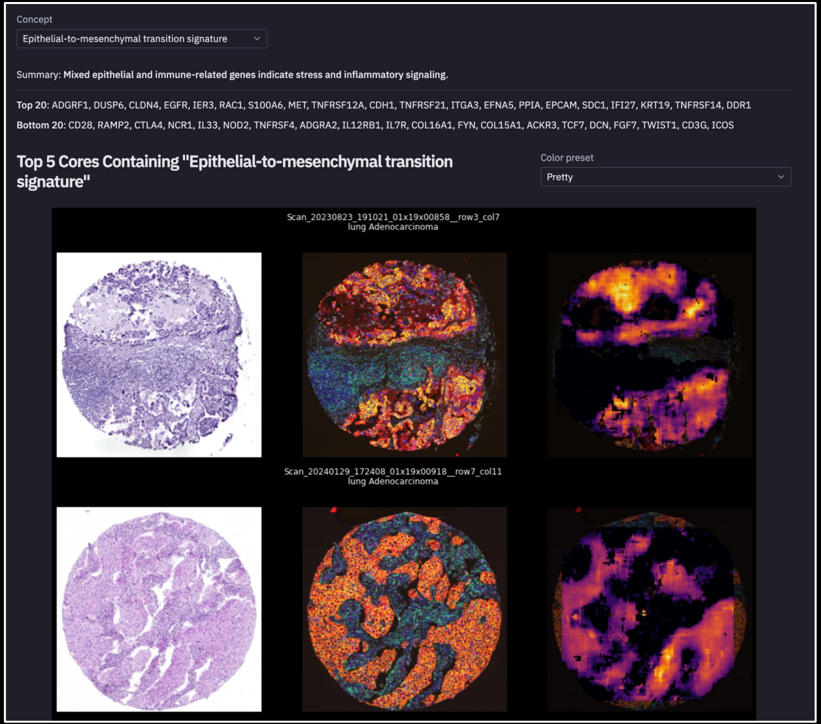


NOETIK

On the biology of a large multimodal model for biology

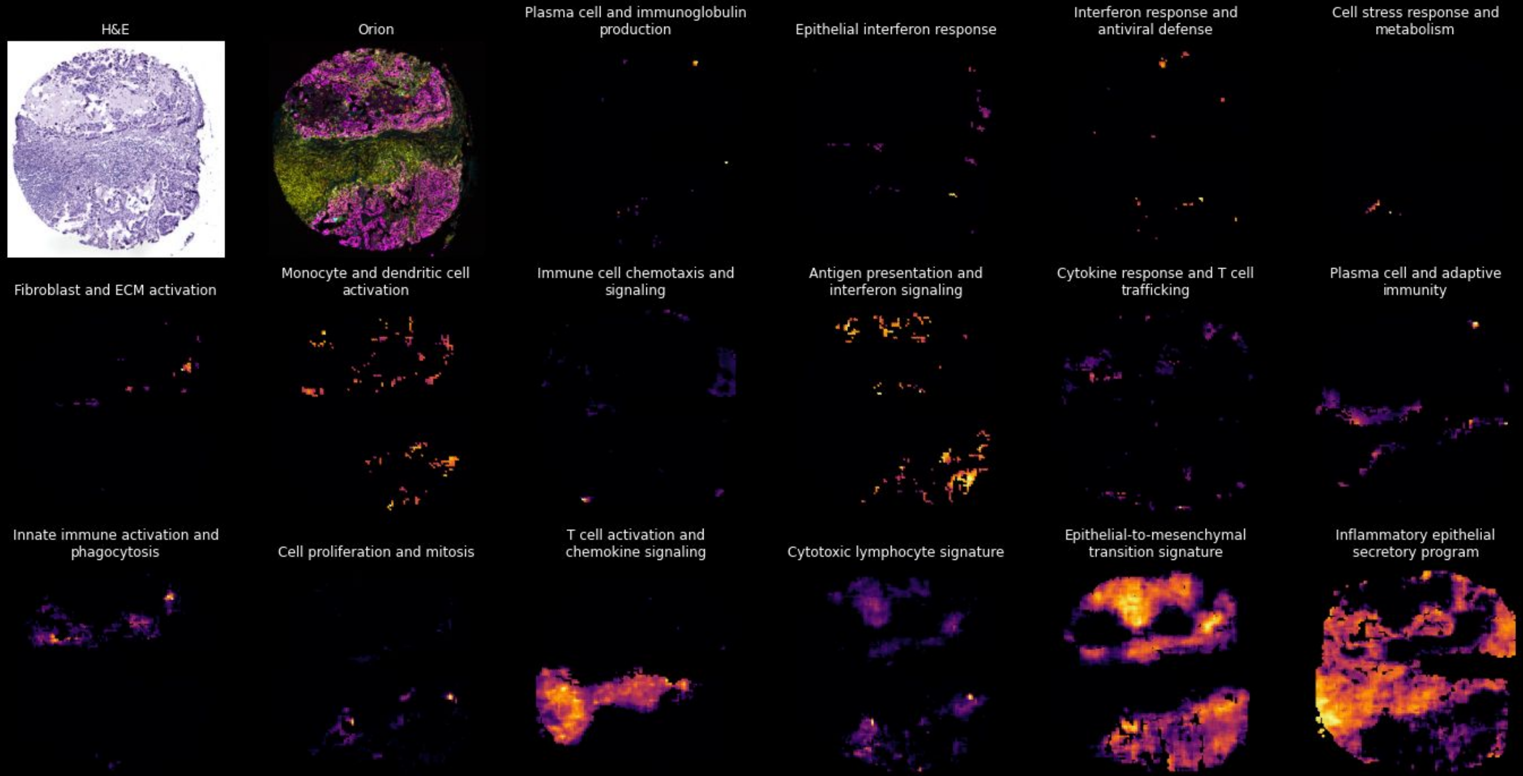
May 13	On the Biology of a Large Language Model **[In-Person]** Speaker: Josh Batson (**Anthropic**)





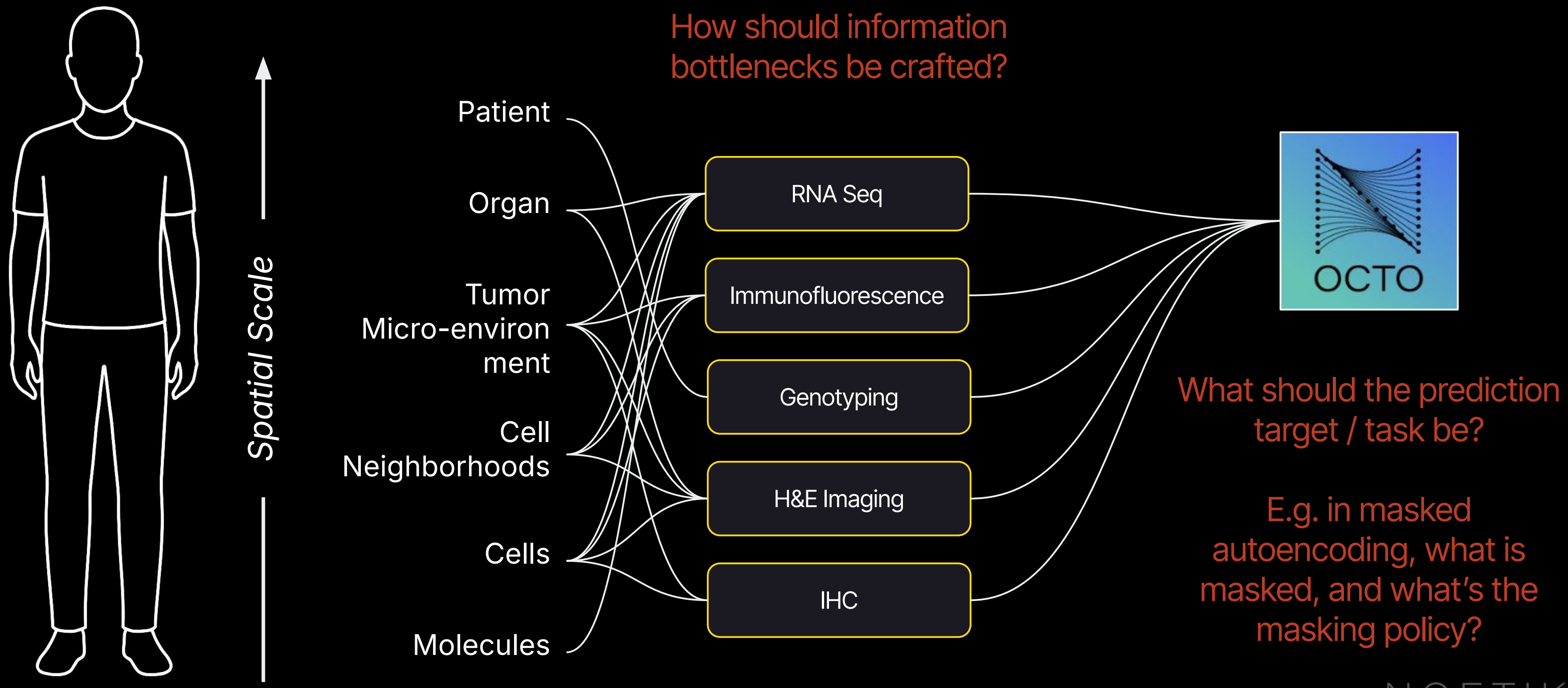
NOETIK

On the biology of a large multimodal model for biology



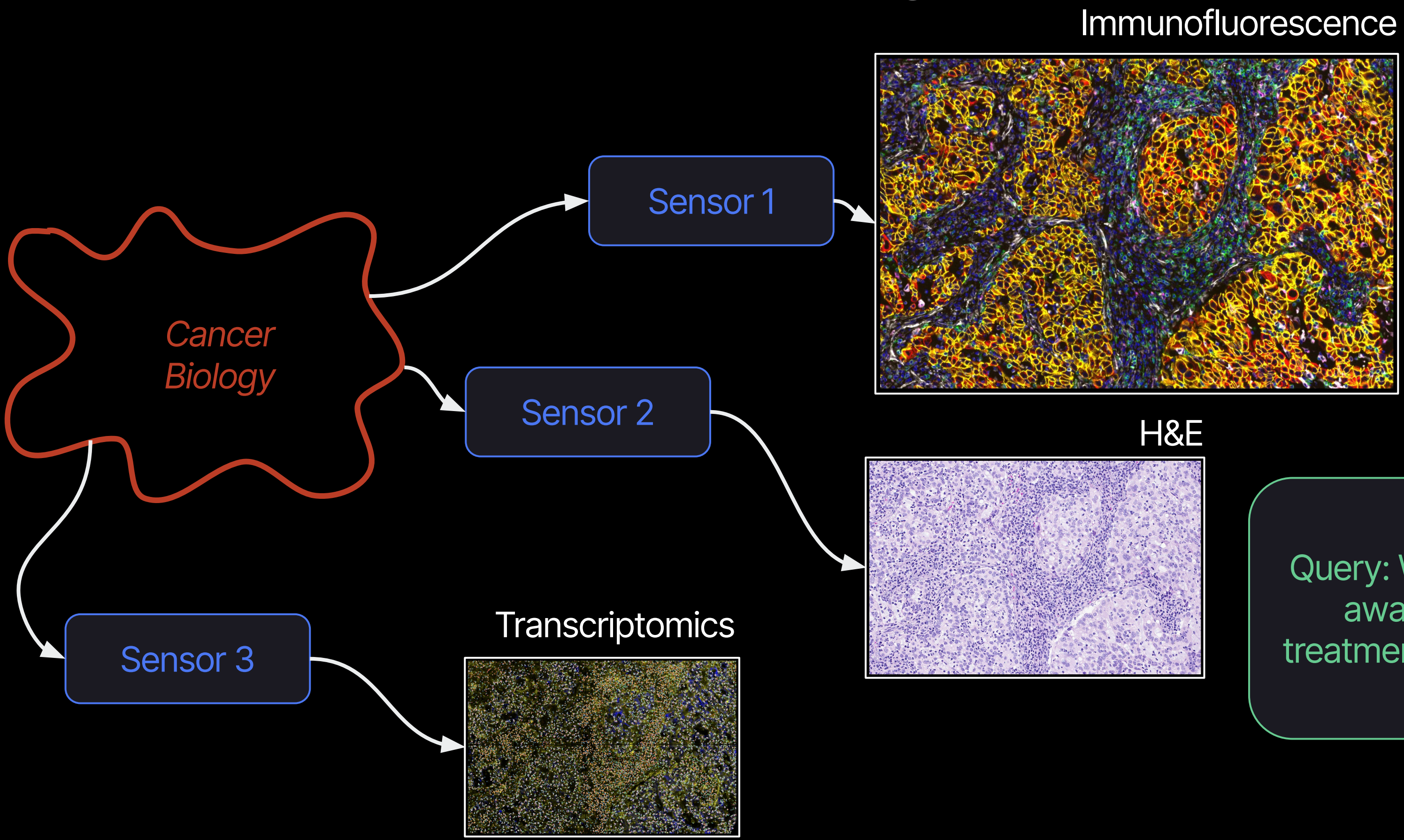
NOETIK

Toward massive multimodal transformers for cancer biology



NOETIK

A world model for tumor biology



Query: Will the tumor go away if I give this treatment to this patient?

NOETIK

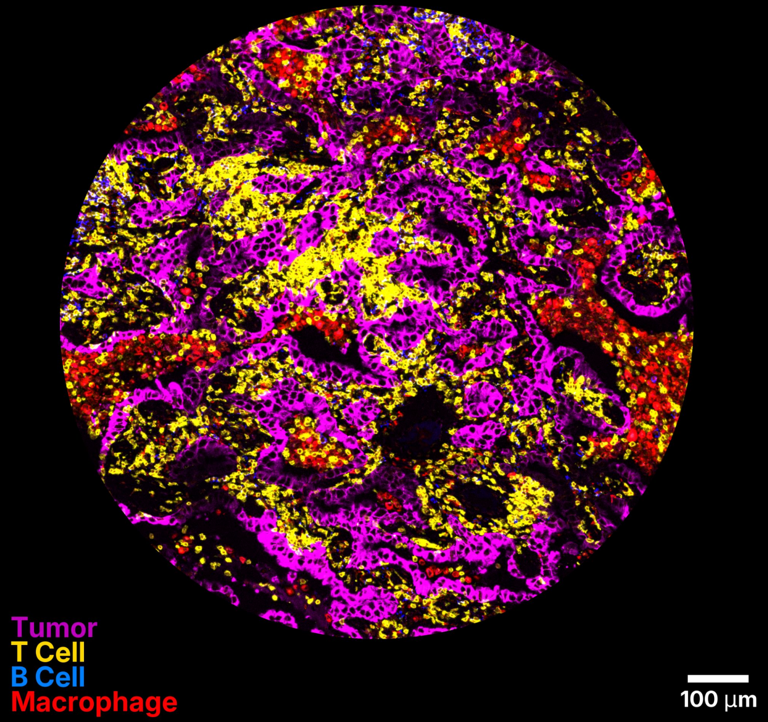
Thank you!

noetik.ai

eshed.margalit@noetik.ai

eshedmargalit.com

eshedmargalit



NOETIK