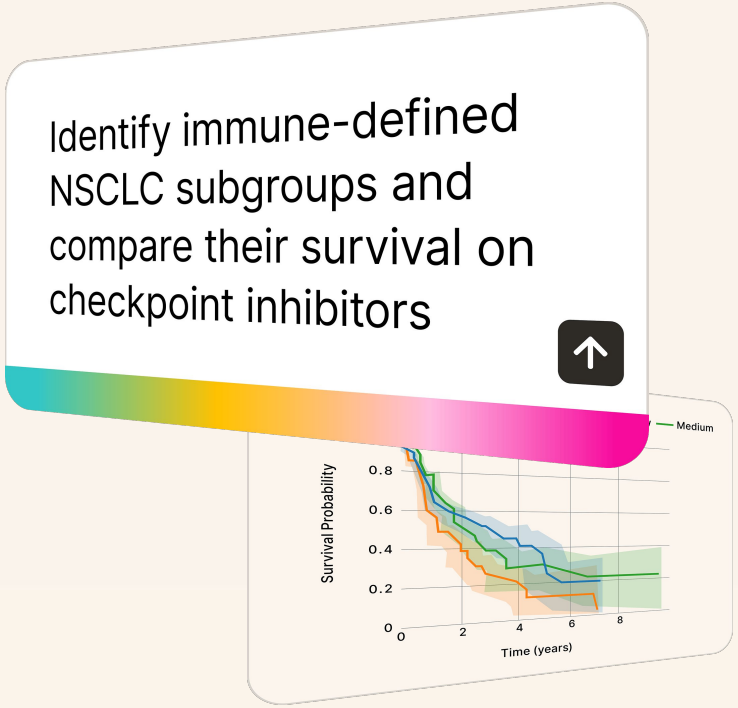
PRO

L’approche agentique en santé

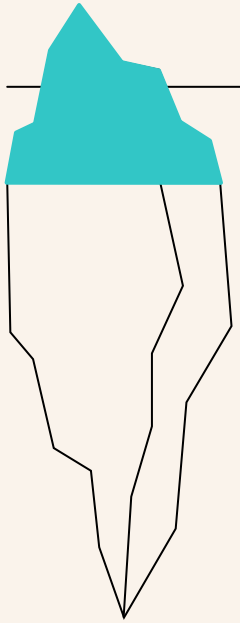
Combiner LLMs, modèles d’IA spécialisés et données de patients pour accélérer la recherche de nouveaux traitements.

↳ Yacine Bareche (Senior Product Manager)



The traditional approach for Target ID in R&D has major challenges

We need better targets, better drugs and better clinical trials



Drug failure

Targets don't convert to drugs:

- Pre-clinical models lack translatability to human biology.
- Biology is complexe & multimodal.
- Intra-tumoral heterogeneity is not well captured.

Drugs responses vary from patient to patient:

- Inter-patient heterogeneity is not well captured.
- Lack understanding of drug mechanism of action.
- Lack understanding of TME.

Clinical trials are slow and expensive:

- Recruitment is slow due to strict inclusion criteria.
- Non-data driven definition of inclusion criteria & Covariate selection makes it hard to detect treatment signal from the noise.

$2.7

Billion

Median cost of bringing an oncology/immunology drug to market

>13

Years

To bring a drug to market

10%

Success rate

High attrition rate at ph 1, 2 & 3 results in only 10% of drugs reaching the market.

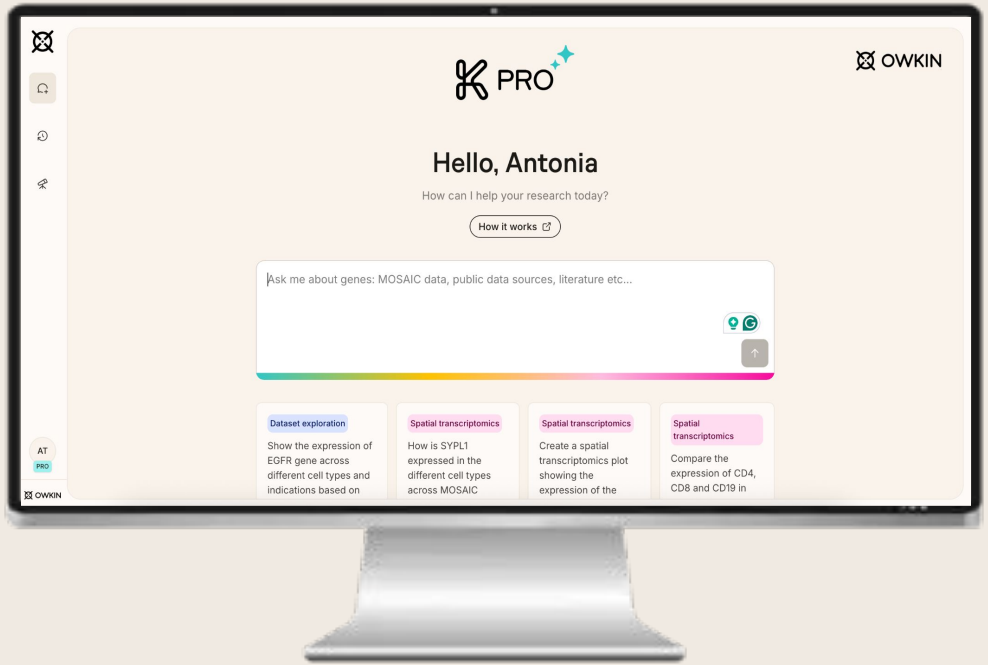
10 years of AI for biology & multimodal patient data packaged into **One AI Scientist** that gives answers in minutes rather than days / months

PRO

All-in-one interface

Data-rich

Superior visualization



- → 100 skills for biology R&D incl. biological reasoning
- → Enriched from 1.2M datasets everyday and 10M Q&A
- → Learns from multiple feedback loops: patient data, lab experiments, and real-world outcomes.

At Owkin we invest in 4 differentiating ingredients to build one AI Scientist

1

Powered & validated by multimodal patient data

- → 104 hospital patient data network w/ latest SoC and expert KOLs
- → Leading multimodal data assets like MOSAIC combined with public data
- → Ability to integrate user's data

2

Orchestration of expert AI models at scale

- → Cutting-edge AI tools & models specialized on biology
- → Orchestrated by one versatile coding agent composing tools
- → Parallel campaigns exploring multiple approaches

3

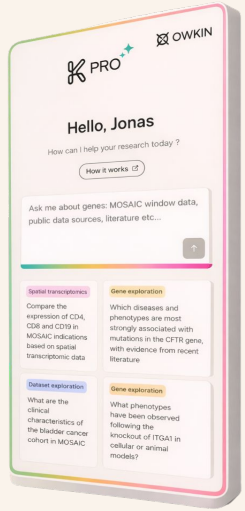
Guided by expert-grade AI skills

- → Owkin-certified skills powered by 10 yrs of biomedical AI research
- → KOL-inspired skills capturing the reasoning of leading domain experts
- → Your own custom workflows built on top

4

Deployable everywhere

- → Deployed securely where client data lives, alongside client tools
- → Integrates into existing workflows and connect with your tools
- → Modular architecture built for interoperability



Start with the right data. Public and proprietary. All in one place.

Breadth and depth of data

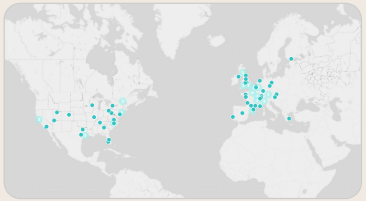
1.2M Patients data with extensive characterization

104 Hospitals from tier 1 academic institutions & labs (50% of the top 20 top academic hospitals)

Data generated, enriched and curated with associated KOL and recent cohorts capturing latest SoC

Multimodality by default (bulk, single cell, spatial RNA, pathology, longitudinal follow ups etc)

Growing network



Dynamic data network enriched everyday in onco, I&I, cardiometabolic, Neuro

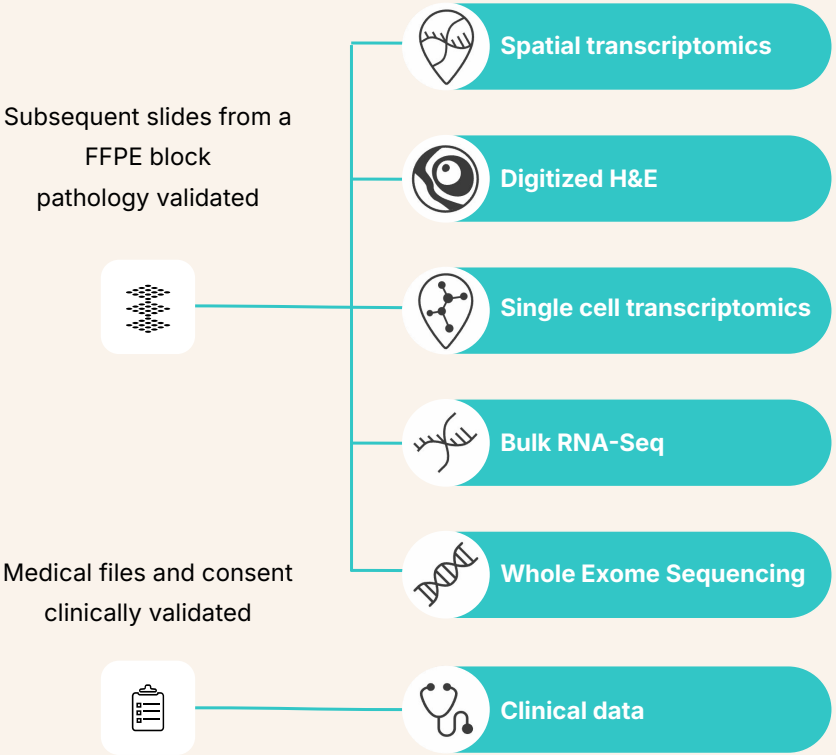
HOPE
MIIDAS
MINDS
CORE

- → **Depth where it matters**. MOSAIC delivers the largest spatially-resolved multimodal cohort in oncology
- → **Breadth when you need it**. K Pro catalogue unifies public and Owkin proprietary cohorts in one interface
- → **Bring your own data**. Connect your internal cohorts. Your data joins the product safely
- → **A validation hub**. AI insights validated by real patients, connecting discovery to clinical reality

A flagship data asset: MOSAIC, a large spatially resolved dataset with 6 data modalities per sample across 11 cancer indications in a centralized platform

- → **11 cancer types** NSCLC, Ovarian, Bladder, Mesothelioma, Glioblastoma, Breast, DLBCL, HNSCC, Pancreas, CRC, Gastric
- → **2,500+ patients** already included
- → Consistent and rigorous data generation: uniform sample processing, centralised NGS, rigorous QC steps at every stage
- → Partners:

OWKIN · GUSTAVE ROUSSY CANCER CAMPUS GRAND PARIS · Centre hospitalier universitaire vaudois · University of Pittsburgh · Uniklinikum Erlangen · CHARITÉ UNIVERSITÄTSMEDIZIN BERLIN



K Pro's data catalogue grows as our network expands and new consortia data is generated

Available	In progress	Mapped

HOPE

HAEMATOLOGY & ONCOLOGY PARTNER NETWORK

Owkin's unique data asset to support optimal patient stratification and enable treatment in Oncology. It includes the MOSAIC dataset.

MIIDAS

MULTIMODAL IMMUNOLOGY & INFLAMATION DATA ATLAS

The first spatial atlas of multimodal data in immunology and inflammation to support optimal patient stratification and enable treatment discovery.

CORE

CARDIO & METABOLIC OUTCOMES RESEARCH NETWORK

Patient data across the cardiovascular-kidney-metabolic syndrome spectrum, from early metabolic risk factors through advanced clinical disease stages.

MINDS

MULTIMODAL INTEGRATION OF ALZHEIMERS NEUROLOGY DATASETS

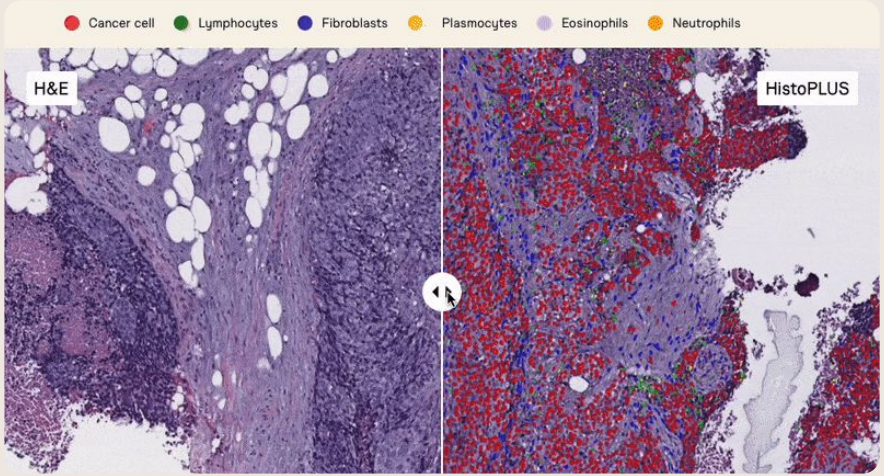
Comprehensive mapping of CNS data availability from partner centers with a focus on Alzheimer's.

Raw data is not enough for LLMs:

1. K Pro is connected to AI tools to transform data into quantifiable biology

AI cell detection tool

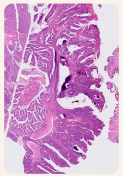
AI based digital pathology tool for cell detection & segmentation (incl. TLS/TILs) . Also available in Claude for Life Sciences



Adjadj et al. arxiv 2025

AI spatial prediction tool

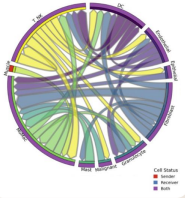
AI prediction of spatial gene expression from H&E



Schmauch et al. Nat Commun 2025

AI cell-cell communication tool

AI models to predict cell-cell interactions from spatial data



Herpin et al. biorxiv 2025

More tools incl.

Infer cell types, predict mutations etc.

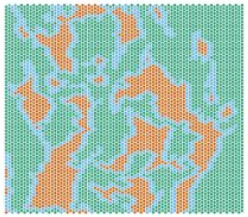
Raw data is not enough: We use it to train highly specialized mid sized LLMs for superior performance

Owkin Zero

K Pro's biological reasoning LLM purpose-built for biological discovery

Trained on ~10M curated Q&A pairs

Sourced from proprietary & public data. Focus on perturbations to ground the model in relevant signals




```
QUESTION: "Which gene is downregulated
in tumor islets versus stroma in
Bladder urothelial carcinoma?"
OPTIONS: {"A": "AFAP1", "B": "BPIFA1"}
ANSWER: "B"
```

Lightweight model + Post-training with RL

Starting from a lightweight 32B model + reinforcement learning with GRPO

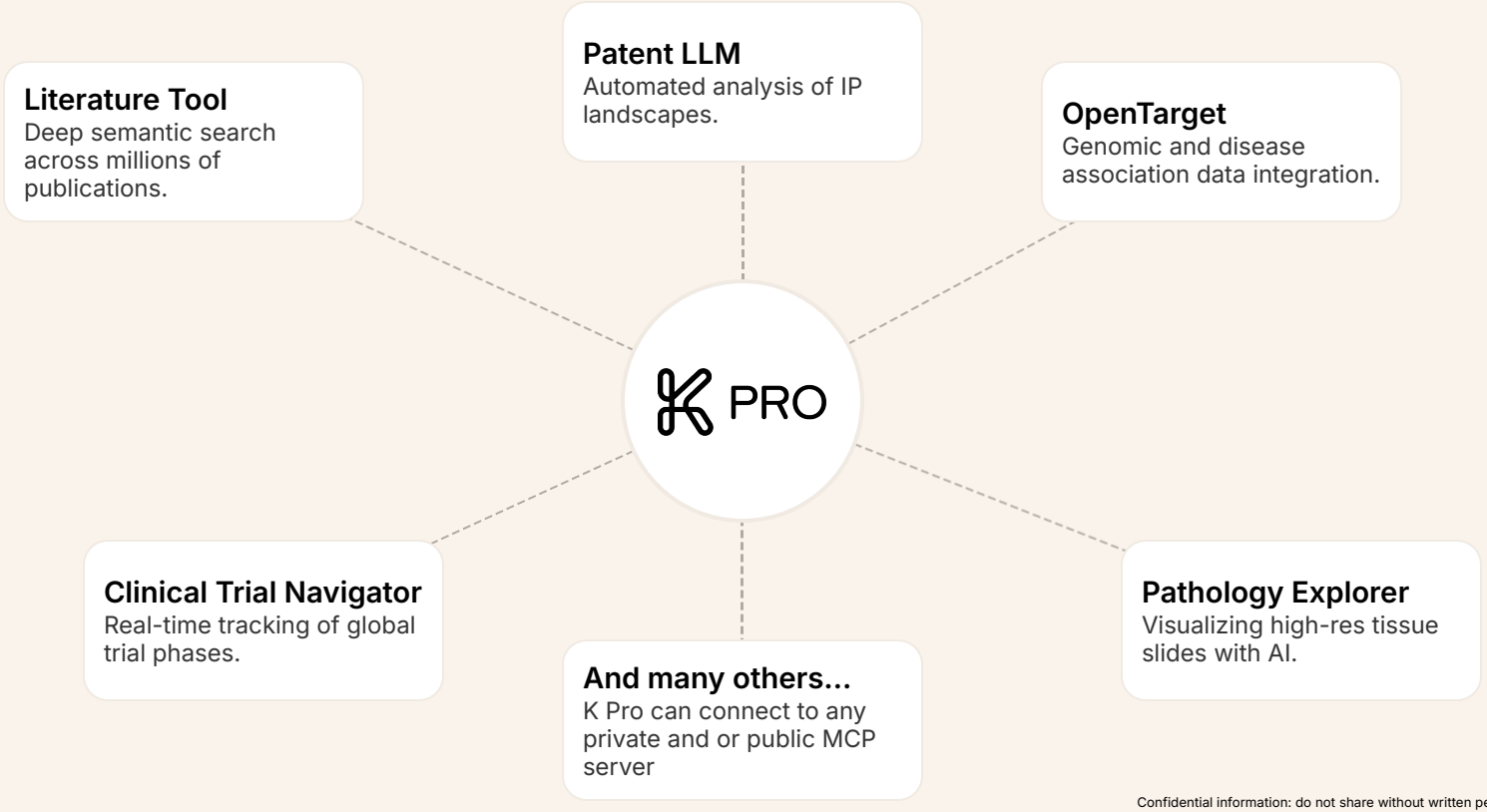
Emerging capabilities in biological reasoning and prediction of perturbation outcome

Sourced from proprietary & public data. Focus on perturbations to ground the model in relevant signals

Bigaud et al. arxiv 2025

Expert AI models

K Pro Connectivity: Integrated Research Tools



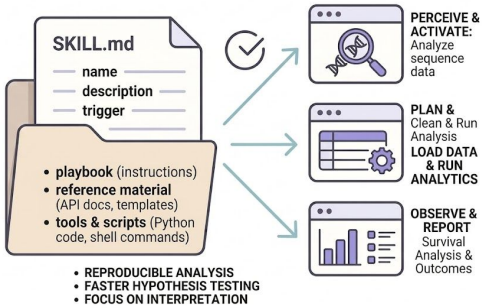
From a toolbox to capabilities with hundreds of expert skills

Encapsulated Logic

A skill is a self-contained, executable workflow that provides an LLM with specific domain capabilities beyond text generation.

It transforms a "reasoner" into a "doer" by providing a structured interface to external tools and data.

How Skill can be used in Oncology Research



Key Characteristics

- **Deterministic:** Reliable output for given inputs.
- **Verifiable:** Actions can be audited and validated.
- **Modular:** Skills can be chained into complex agents.

From a toolbox to capabilities with hundreds of expert skills

SKILL ID CARD

Name	Target Characterization Report
Purpose	Produces a comprehensive multi-omics characterization of a gene target
Status	Owkin Certified
Workflow	*Module 1* — Bulk mRNA *Module 2*— Single Cell RNA *Module 3* — Spatial RNA *Module 4* — Biomarker *Module 5* — Competition *Module X* — ...

Owkin-certified skills

Validated scientific workflows, built from 10 years of Owkin projects

OWKIN

KOL-derived skills

Leading domain experts' methods and reasoning encoded as executable workflows

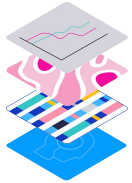
Your own skills

Your internal best practices and standards, on top of existing skills

An ADC toolbox powered by Owkin-certified skills

Deep & Augmented Data

mosaic
an owkin initiative



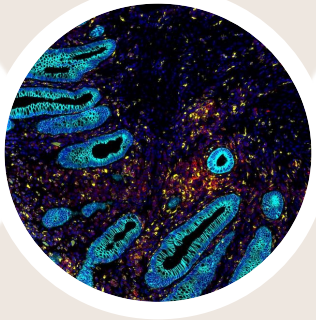
ADC Skills

Name	adc-target-id adc-subtyping, etc...
Purpose	Given input dataset and subtype labels perform....
Status	Owkin Certified

Smarter Decisions

Improved POS & differentiated positioning

1. Efficacy: target expression and heterogeneity in tumor cells
2. Safety: target expression in healthy tissues
3. Subgroups: target expression and heterogeneity in patient population

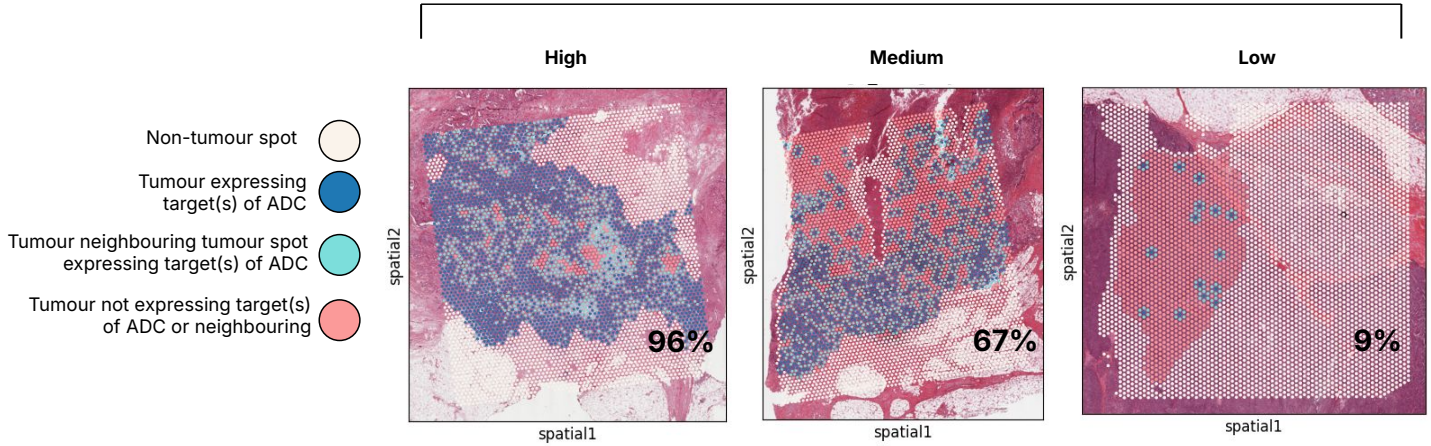


4. Bystander effect: spatial coverage
5. Bispecific ADC: target pair optimization
6. AI pathology on IHC

Skill: bystander-effect MOSAIC-derived biological features to assess the bystander effect and tumour coverage for targets of interest

What is the proportion of tumour spots either expressing the gene or that is a direct neighbour of a tumour spot expressing the gene? How are they distributed?

Example: ERBB2 spatial tumour coverage in Breast



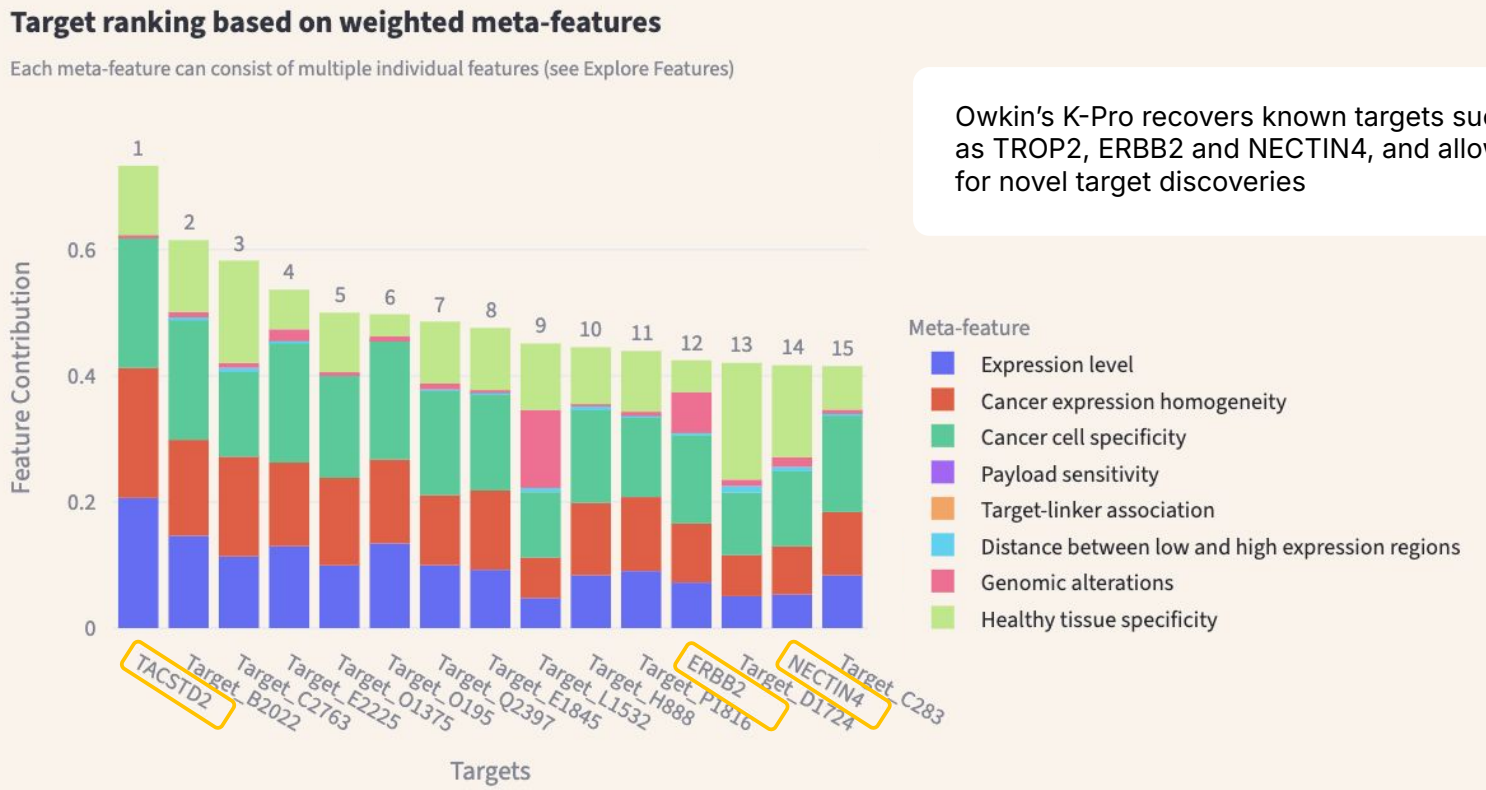
Hypothesis: A high coverage of tumour region spots by expression of the target either directly or via neighbouring spots indicates a greater portion of the tumour that is likely to be exposed to payload and therefore respond to ADC.

Skill: bystander-effect Tumour coverage heterogeneity across patients and targets

Gene-level Distribution

Patient-level Distribution

Skill: monospecific-adc-target-ID for integrated results in Bladder cancer

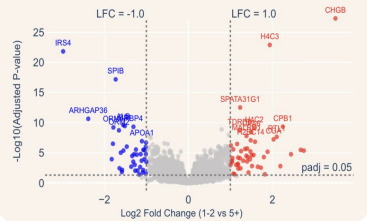


Owkin's K-Pro recovers known targets such as TROP2, ERBB2 and NECTIN4, and allows for novel target discoveries

What can you do with K Pro to improve each step of your pipeline?

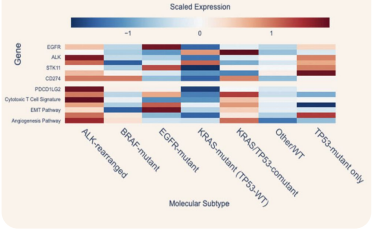
Prioritize targets

Investigate differential gene expression between breast cancer patients with different survival outcomes - 1-2 years vs. 5+ years survival - from TCGA data.



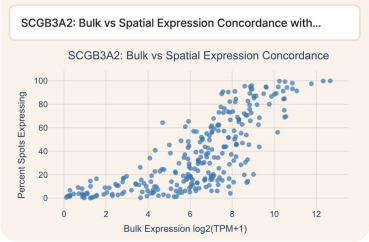
Characterize subgroups

Identify NSCLC molecular subtypes by mutations and expression. Characterize immune infiltration, pathways, and survival. Which offers the best targeted therapy opportunity?



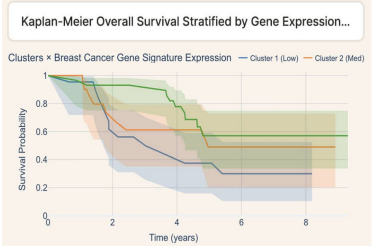
Validate biomarkers

Identify a potential biomarker in bulk data and validate its significance across single-cell and spatial modalities.



Quantify clinical impact

Perform unsupervised clustering of breast cancer patients based on gene expression and show survival differences between clusters



But also productize your **multimodal reports**, analyze **clinical competition**, inform **portfolio strategy** etc

All the data and decision-grade insights informing your drug discovery and development

Thank you

www.owkin.com

yacine.bareche@owkin.com

@OWKINscience

@owkin

Boston, London, Paris, Basel

MOSAIC

MOSAIC samples - breakdown by modality and cancer type

Samples	***Modalities***					
Indications	Clinical data	Spatial transcriptomics	Single cell RNA-seq	Bulk RNAseq	WES	Digitized H&E
NSCLC	533	468	414	312	349	504
Ovarian	473	394	411	387	414	451
Bladder	300	234	284	247	285	292
Mesothelioma	97	83	94	87	95	97
GBM	315	279	298	273	305	307
Breast	386	281	249	174	202	337
DLBCL	189	160	149	136	155	185
HNSCC	158	132	102	88	108	147
Pancreas	145	126	108	109	118	137
CRC	100	54	59	69	75	66
GC	20	20	12	12	12	20
Total	**2716**	**2331**	**2180**	**1894**	**2118**	**2543**

2.7 K
patients in the Study

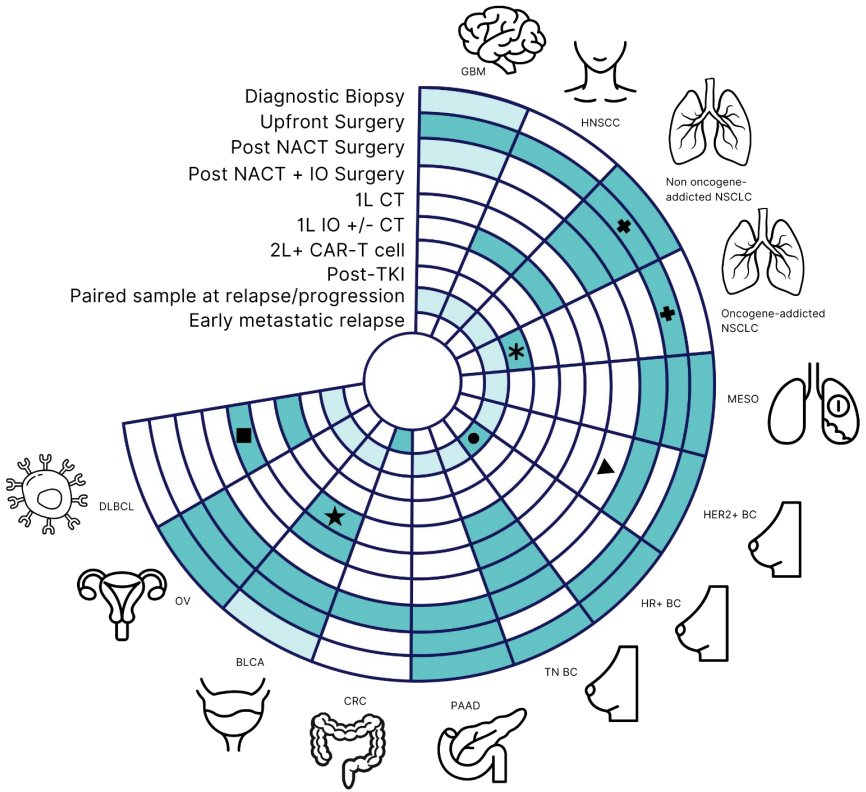
15% patients have multiple samples

~80% of samples are pre-treatment

~10% of samples are post-treatment

~10% of samples are relapse / recurrence

MOSAIC multimodal and longitudinal cohorts span across various indications



6 data modalities for every patient

Consecutive slices
From FFPE block
Pathology validated

- Spatial transcriptomics
- Digitized H&E
- Single cell RNA-seq
- Bulk RNA-seq
- WES

Consent
Medical data
Clinically validated

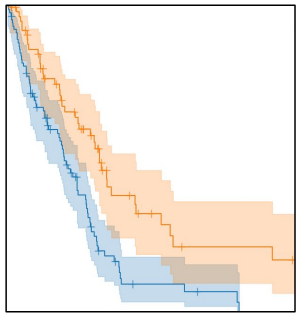
- Clinical data

- Main samples
- Optional samples
- ★ *Pembrolizumab combined with Enfortumab Vedotin (ADC)
- ▲ *Neoadjuvant chemotherapy in combination with HER2-targeted therapy
- ● *Samples from recurrence within 10 years
- ◆ *R-CHOP or 'R-CHOP-Like' regimen
- ✚ *Stage IA-II
- ✶ Paired pre-TKI samples were optional but highly encouraged

Resolution of MOSAIC Samples and Technology

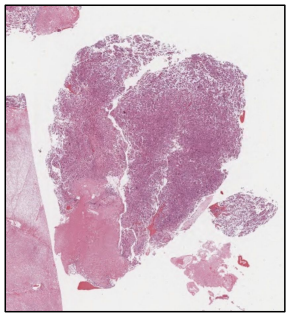
eCRF Patient Clinical information

- Precise clinical data with baseline clinical, pathological, and molecular information, treatments, evolution and outcomes



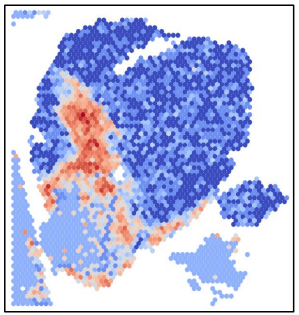
Whole Slide Images H&E

- H&E microscopic images, generated with Leica Aperio CS2 scanner with 40X magnification.



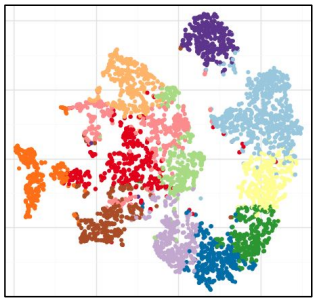
Spatial Transcriptomics

- 10x Visium SD v2 with Cytassist with total capture area of 6.5 × 6.5 mm
- 5,000 spots per sample, corresponding to about 1-10 cells per spot.



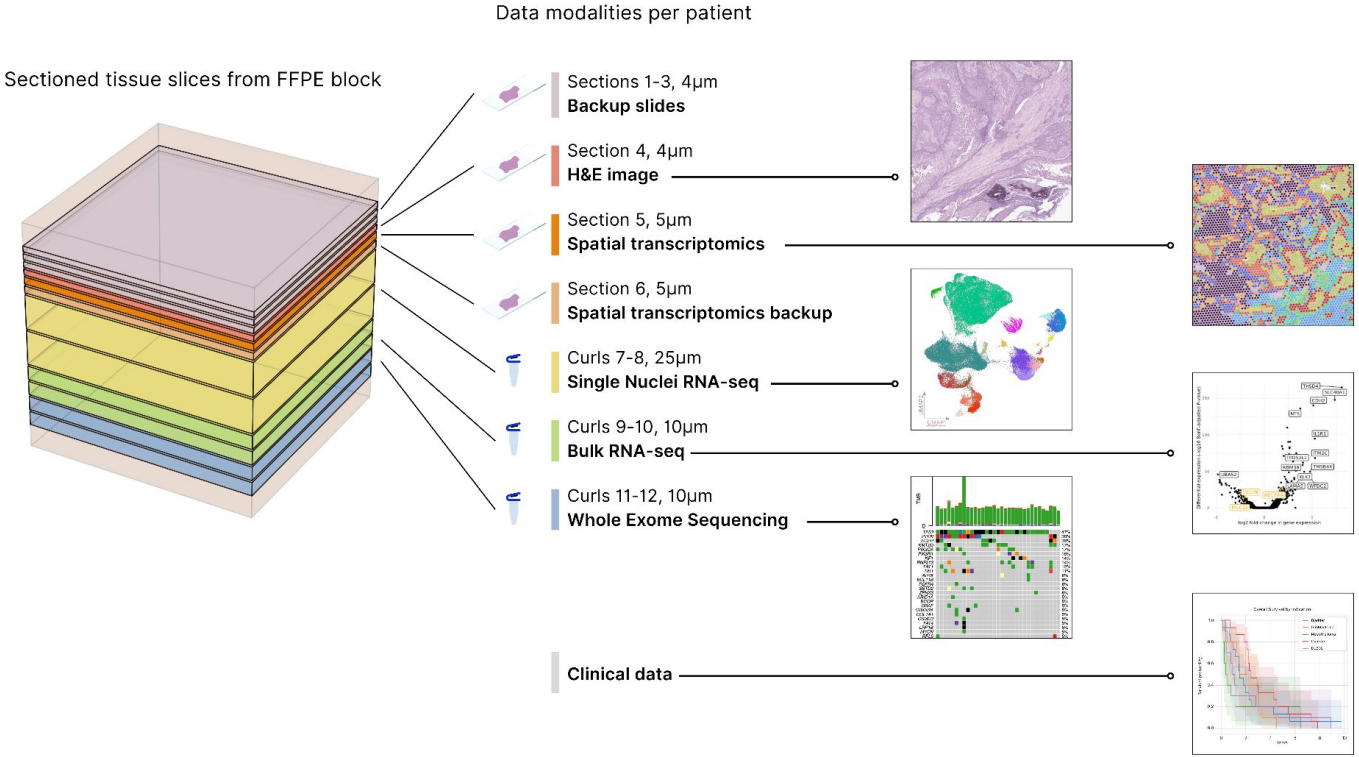
Single Nuclei Transcriptomics

- snRNAseq adapted from the Chromium FLEX protocol from 10X Genomics.
- 8000 cells per samples with 10,000 reads per cell.



MOSAIC dataset also contains matched bulk RNA and Whole Exome Sequencing

Schematic Representation of Data Modalities



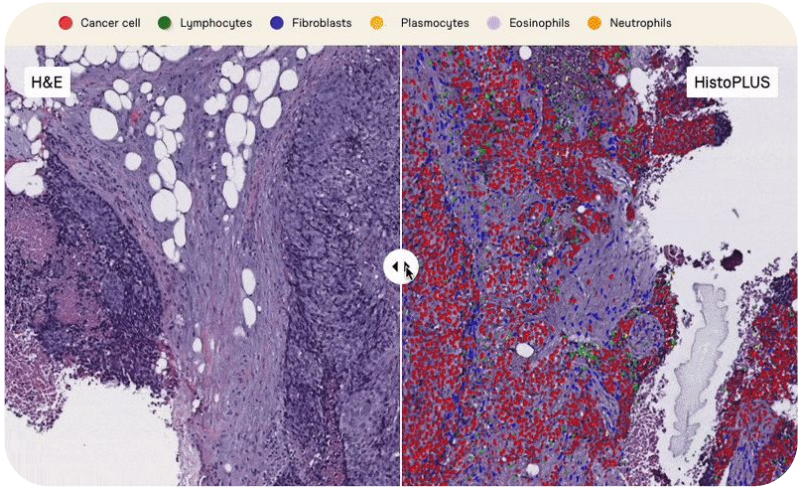
Expert AI Models

Augment H&E with proprietary cell segmentation and classification models

Adjadj et al. arXiv 2025

Histomics

AI based digital pathology tool for cell detection (including TILs and TLS) & segmentation



13 **cell types** covered, including understudied immune populations such as neutrophils & eosinophils.

5 **cancers** and leverage transfer learning technologies to maximize efficiencies

24% **better F1 classification of cells and 5% better detection** using 5 times less parameters

200k **consensus annotations** done by 10 pathologists

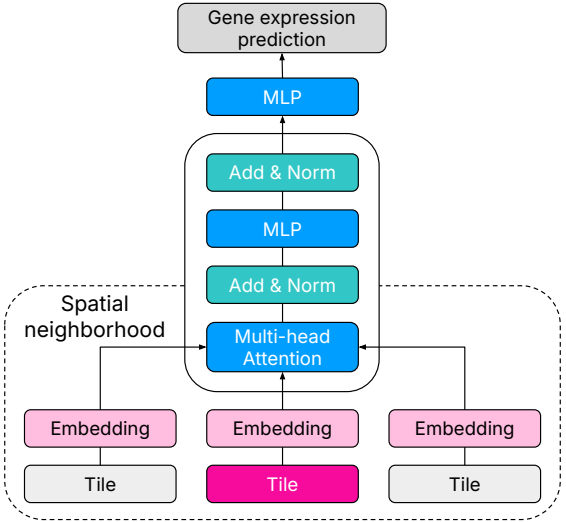
Augment H&E with proprietary cell models on spatial expression prediction

Schmauch et al. arXiv 2024

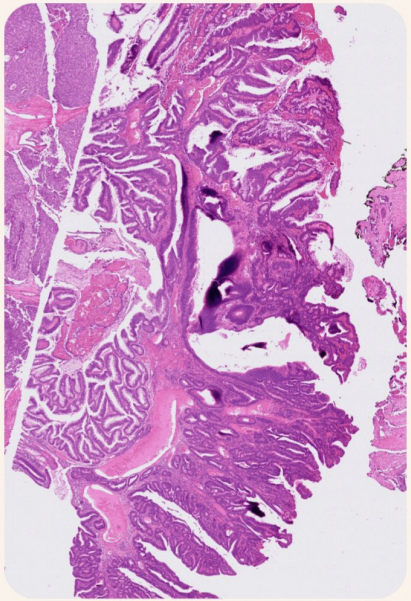
Spatial prediction

Predict gene expression in each spot of a spatial transcriptomics cohort using associate tile from H&E

Feature extractor	Training data	**HEST Average** (Pearson coefficient)
Baseline iBOT	FFCD	0.246
H0	FFCD	0.286
H0-mini	FFCD	0.344
H0-mini	**MOSAIC**	**0.381**



Near single-cell-resolution prediction through model distillation



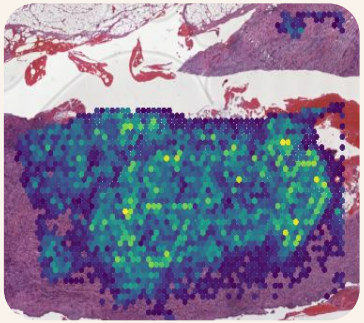
H&E-stained tissue

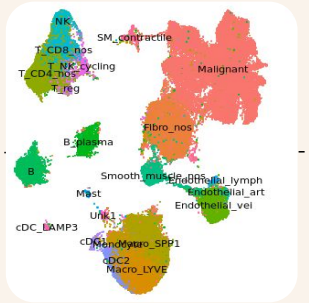
26

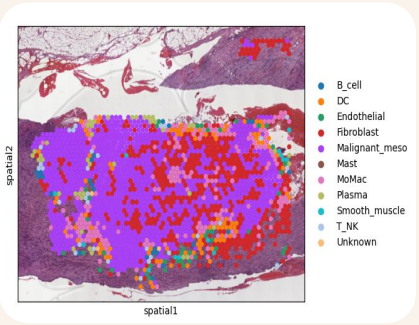
Augment molecular data with deconvolution algorithms (1)

↳ Increase resolution of Visium leveraging paired modalities (HE + ScRNA + Spatial)

Specific TME question?

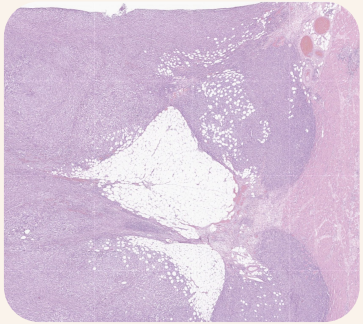




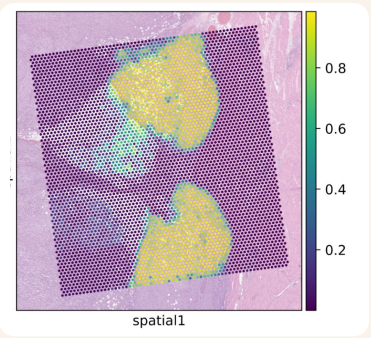


Spot-level deconvolution of cell types

Different tumor areas?



Learn cell signatures from single-cell RNAseq on paired samples within the same cohort



Spatialization of of specific tumour transcriptomic cluster (clone?)

27

Joint probabilistic modeling of pseudobulk and single-cell transcriptomics enables accurate estimation of cell type composition

Simon Grouard [*1] Khalil Ouardini [*1] Yann Rodriguez [1] Jean-Philippe Vert [1] Almudena Espin-Perez [1]

ICML International Conference On Machine Learning 2025

OwkinZero: a biological reasoning model designed to accelerate drug discovery

OWKIN Zero

- → Smaller fine tuned models are able to outperform frontier models on specialist tasks
- → Fine tuning on many contexts is better than one
- → Generalization to unseen tasks is still difficult