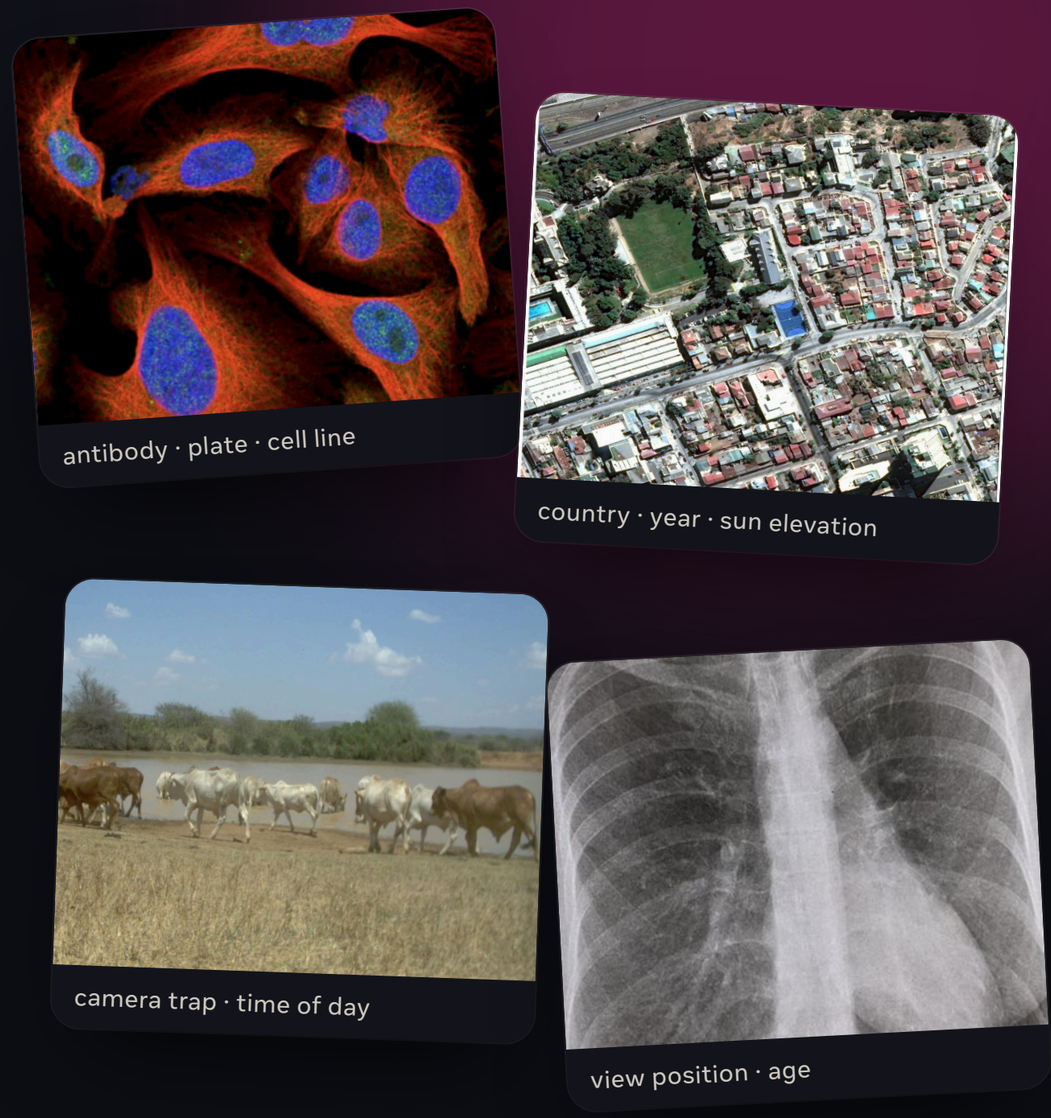
LABEL-FREE REPRESENTATION ADAPTATION

Who Needs Labels?

Adapting Vision Foundation Models With the Metadata You Already Have

Elouan Gardès[1,2] Seung Eun Yi[1] Kartik Ahuja[1] Théo Moutakanni[1]
Huy V. Vo[1] Piotr Bojanowski[1] Wolfgang M. Pernice[3*] Loïc Landrieu[2*]
Camille Couprie[1*]

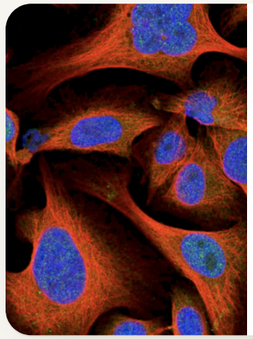
[1] Meta FAIR [2] LIGM, CNRS, Univ Gustave Eiffel, ENPC, IPP [3] Columbia University, New York
[*] Equal supervision



MOTIVATION

A single approach for very different scientific data

Goal: obtain the best representations across entirely different scientific fields, with a **unified** approach.



HPA

Fluorescence microscopy

••• plate ••• antibody ••• cell line



FMoW

Earth observation

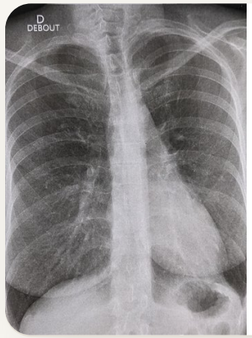
••• country ••• weather ↔ viewing angle



iWildCam

Wildlife monitoring

••• camera trap ↔ timestamp



MIMIC-CXR

Chest radiography†

••• view position ↔ age

These datasets have two things in common:

1 Domain shift

Sites, sensors, batches, regions and seasons differ between training and test data.

2 Lots of free metadata

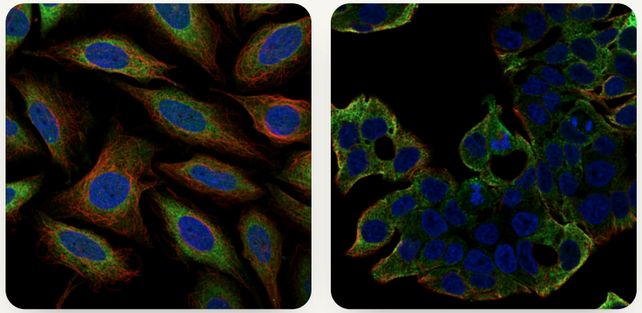
Recorded with every image, so it can serve as weak supervision at no labelling cost.

••• discrete ↔ continuous

† Illustration is a personal X-ray of one of the authors, not a MIMIC-CXR image.

MOTIVATION

Domain shifts in scientific datasets



Cell microscopy · HPA

- cell morphology
- batch effects (plate, reagent aliquot)

Wildlife monitoring · iWildCam

- background of each camera trap
- timestamp (night infrared vs. daylight)





Remote sensing · FMoW

- country, region
- timestamp (season, year)

MOTIVATION

Why not just fine-tune with labels?

Labels are scarce

Expert annotation in microscopy, radiology or ecology is slow and expensive.

Fine-tuning hurts OOD

The model fits the training distribution and its task, and loses the generality it had after pretraining.

Domain adaptation is still task-driven

UDA adds unlabelled target data, but still optimises a single label space.

43.4 fine-tuned vs **51.4** frozen

OOD macro F1 · iWildCam

On the out-of-distribution test set (camera traps never seen in training), supervised fine-tuning of DINOv3 ends up **below the frozen model** it started from.

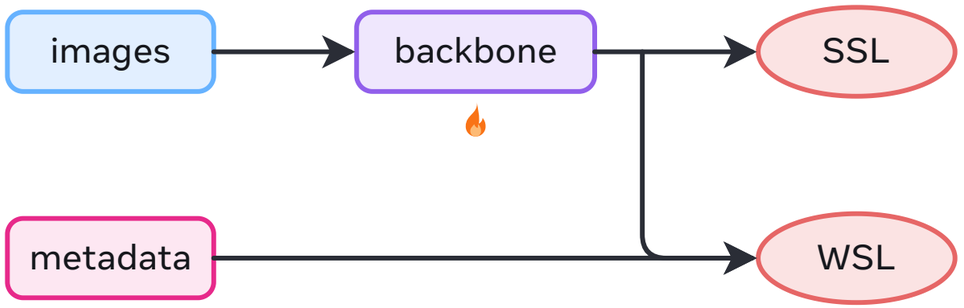
SETTING

Representation adaptation

Adapt a foundation model to a new domain **without task labels**, then evaluate the **frozen** representation with a lightweight probe.

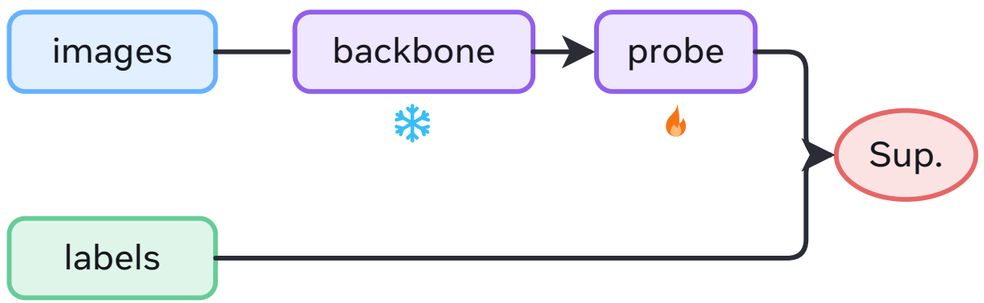
1 Adapt the representation

no labels



2 Lightweight probing

with labels



Backbone ViT-L, 306M params

Probe attentive, 2.7M params (linear <1M)

No metadata needed at inference

RELATED WORK

Limitations of previous metadata-guided methods

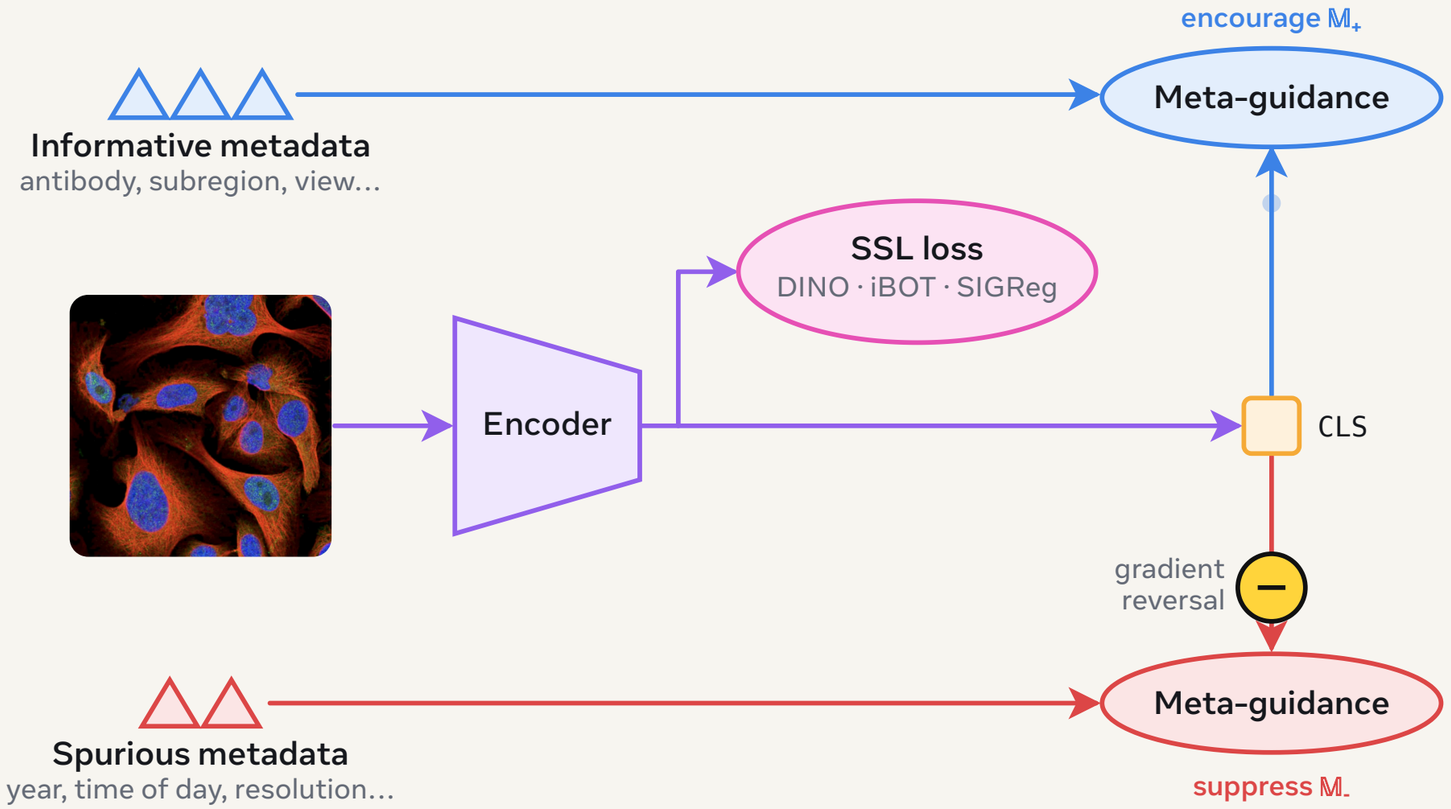
Previous works typically have at least one of the following limitations:

✕ Domain specificity	✓ Same objective and loss weights for all 4 domains
✕ Backbone architecture modifications	✓ Plain ViT, guidance only in training heads
✕ User constraints, e.g. metadata at inference	✓ Metadata only used during adaptation
✕ Specific metadata type (only continuous, or few classes)	✓ Discrete with 10^4+ values, and continuous
✕ Fixed “angle”: a factor is either spurious or informative	✓ Each factor can be encouraged or suppressed

So we build a **unified framework**: **FINO**

METHOD

FINO: FIne-tuning with NO labels



- **Self-supervision** keeps the representation general: DINO on the CLS token, iBOT on patches, SIGReg as regulariser.
- **Informative** factors $\mathbb{M}_+$ are *encouraged*: the embedding should be organised along them.
- **Spurious** factors $\mathbb{M}_-$ are *suppressed*: the guidance gradient is reversed before it reaches the encoder.

METHOD · METADATA GUIDANCE

Discrete metadata: prototype-based contrastive loss

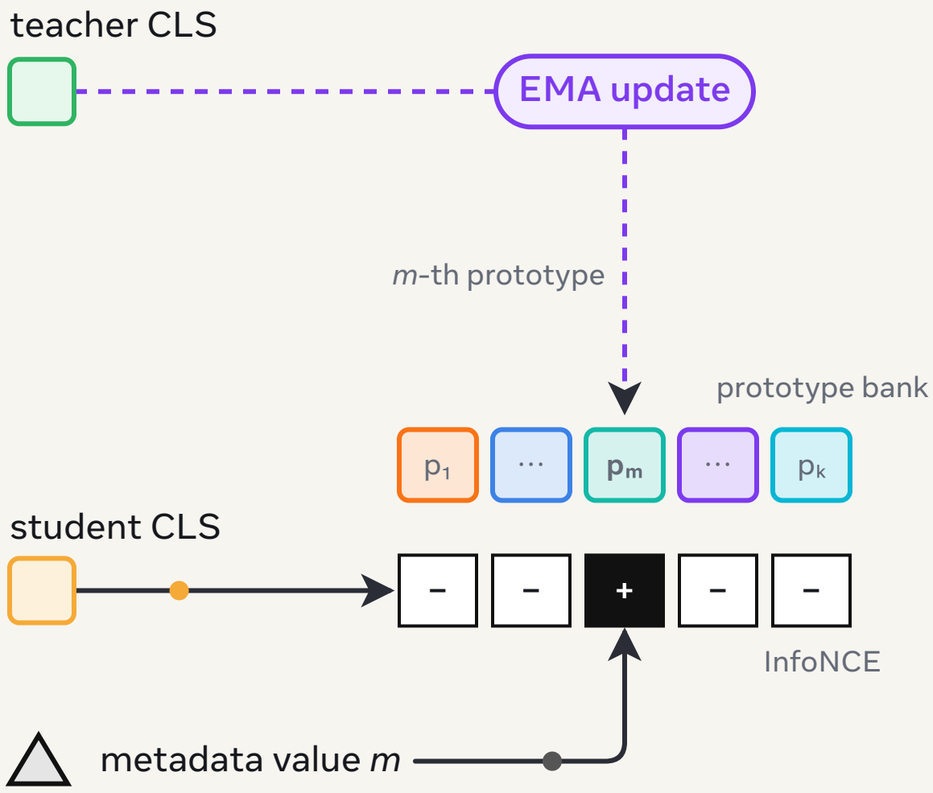
$$\mathcal{L}_{\text{meta}}^{(t)}\left(x\right) = -\log\frac{\exp\left(\left\langle \varphi\left(x\right), p_m^t \right\rangle / \tau\right)}{\sum_{n=1}^{M} \exp\left(\left\langle \varphi\left(x\right), p_n^t \right\rangle / \tau\right)}$$

InfoNCE between the student embedding and **one prototype per metadata value** (cosine similarity).

$$p_m^t \leftarrow \alpha\, p_m^t + \left(1-\alpha\right) \varphi_{\text{teacher}}\left(x\right)$$

The matched prototype is updated from the **teacher** embedding by EMA, then re-normalised.

No positive pairs needed in the batch, so it scales to the **14,678** antibodies of HPA.



METHOD · METADATA GUIDANCE

Continuous metadata and gradient reversal

Continuous metadata (timestamp, age, viewing angle…)

$$\mathcal{L}_{\text{meta}}^{(t)}\left(x\right) = \left\| g^{(t)}\left(\varphi\left(x\right)\right) - m \right\|_2^2$$

A small predictor head $g^{(t)}$ regresses the value from the embedding.

Balancing contrastive and regression losses

$$\tilde{n}_t \leftarrow \mu\,\tilde{n}_t + \left(1-\mu\right)\left\|\nabla_{\varphi(x)}\mathcal{L}_{\text{meta}}^{(t)}\right\|_2$$

$$s_t = \text{sg}\left(\frac{\overline{n}}{\tilde{n}_t}\right), \quad \overline{n} = \left(\prod_t \tilde{n}_t\right)^{1\,/\,|\mathcal{T}|}$$

Each branch is rescaled by s_t so its gradient norm at the CLS embedding matches the geometric mean of all branches (μ = 0.99). Nothing is learned.

Same loss, opposite sign

+ Informative $\mathbb{M}_+$

The encoder **minimises** the guidance loss, so features align with the factor.

− Spurious $\mathbb{M}_-$

The gradient is **reversed** before the encoder, so features stop carrying the factor.

The heads themselves (prototypes, $g^{(t)}$) always minimise the loss, as in DANN.

METHOD

Same objective and hyperparameters for all datasets

$$\mathcal{L} = \mathcal{L}_{\text{DINO}} + \mathcal{L}_{\text{iBOT}} + \lambda_{\text{SIGReg}}\mathcal{L}_{\text{SIGReg}} + \sum_{t \in \mathbb{M}_+ \cup \mathbb{M}_-} \lambda_{\text{meta}}^{(t)}\mathcal{L}_{\text{meta}}^{(t)}$$

self-supervision · regularisation (from LeJEPA) · metadata guidance

1 recipe

The same losses and weights on HPA, FMoW, iWildCam and MIMIC-CXR. Only the metadata fields change.

$\lambda_{\text{meta}} = 0.03$

One value for every metadata branch, discrete or continuous, on every dataset.

No sweep

No per-dataset loss-weight search, so the method can be applied to a new dataset as is.

Init DINOv3 ViT-L

Schedule heads trained on a frozen backbone, then full adaptation (shorter run on iWildCam)

EXPERIMENTS

Datasets and evaluation protocol

Same ViT-L backbone for every method. Metadata is selected on held-out validation data, never on test scores.

Dataset	Domain	Metric	Encouraged $\mathbb{M}_+$	Suppressed $\mathbb{M}_-$	Transfer target
HPA	Fluorescence microscopy	macro F1 (Kaggle private)	antibody	none	OpenCell
iWildCam	Wildlife camera traps	OOD macro F1	none	time of day	none
FMoW	Satellite imagery	OOD worst-group acc.	subregion	year	FLAIR-Hub
MIMIC-CXR	Chest X-ray	AUROC	view position	none	CheXpert

Baselines supervised (scratch, fine-tuning), DANN, CORAL, Group-DRO, frozen DINOv3, DINO SSL adaptation

RESULTS

Main results

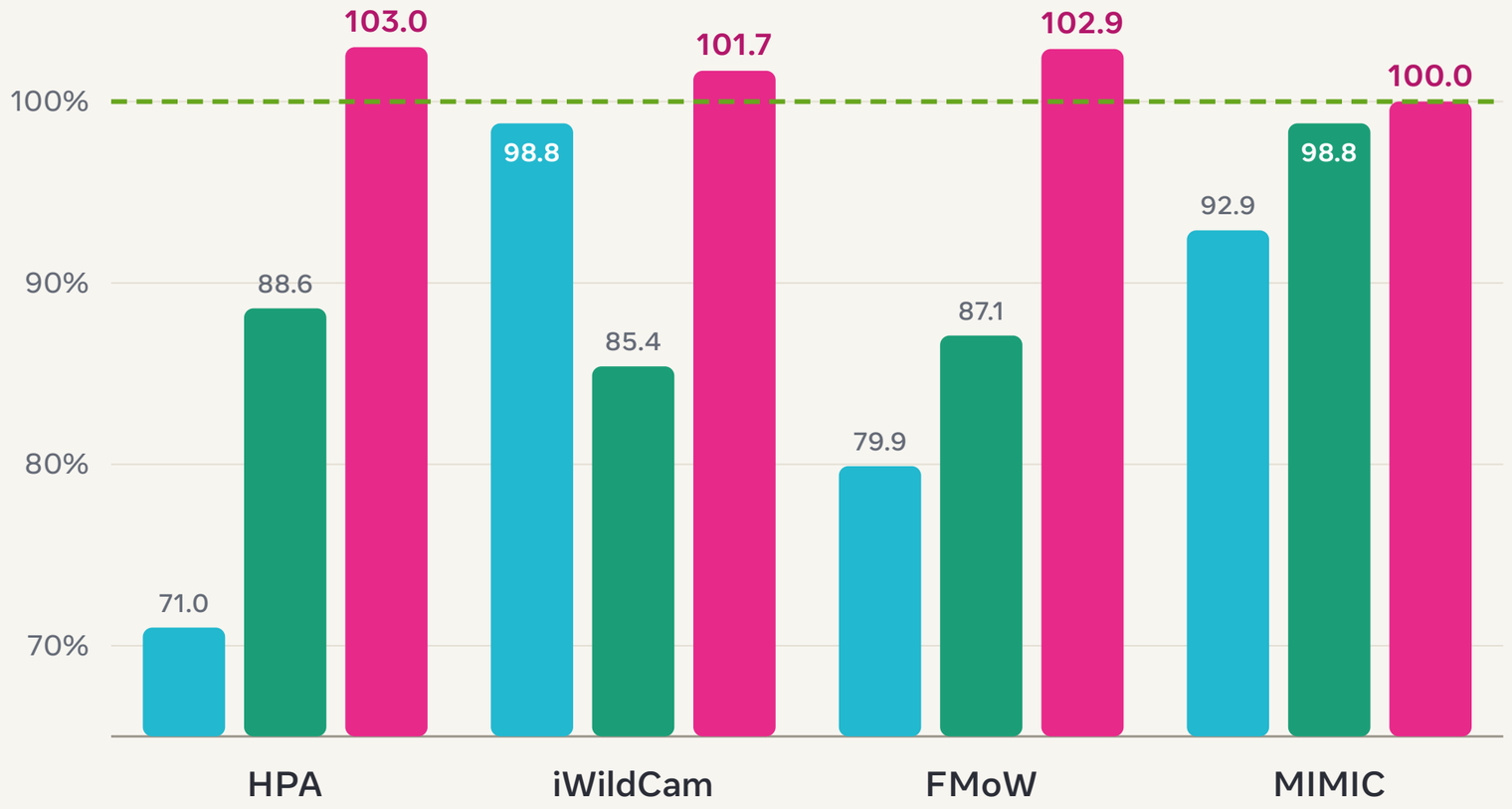
Method		Label supervision		HPA	iWildCam	FMoW	MIMIC
		Param.	*Backbone*	*F1*	*OOD F1*	*WGA*	*AUROC*
Published reference		*80–300M*	🔥/❄️	*59.4*	*52.0* ±0.4	*51.8* ±0.4	*81.7* ±0.1
Supervised training	From scratch	306M	🔥	44.7	10.1	24.3	76.8
	Fine-tuning	306M	🔥	52.6	43.4	40.5	80.7
Domain adapt. & general.	DANN	306M	🔥	50.2	38.5	46.2	n/a
	CORAL	306M	🔥	41.1	38.5	45.9	n/a
	Group-DRO	306M	🔥	52.6	26.1	46.4	80.0
Representation adaptation	Frozen DINOv3	2.7M	❄️	42.2	51.4	45.0	75.9
	DINO (SSL)	2.7M	❄️	60.0	50.0	48.4	81.1
	FINO (ours)	2.7M	❄️	**61.2***	**52.9** ±0.8	**53.3** ±0.8	**81.7** ±0.1

Mean ± std over 8 / 5 / 5 complete adaptation runs (iWildCam / FMoW / MIMIC). *Full-resolution HPA, single run; the low-resolution repeat gives 54.7 ± 0.7 over 10 runs. Published references use different protocols.

RESULTS

Comparison with published specialised methods

Performance relative to the published reference on each dataset (reference = 100%).



Frozen DINOv3

Supervised fine-tuning

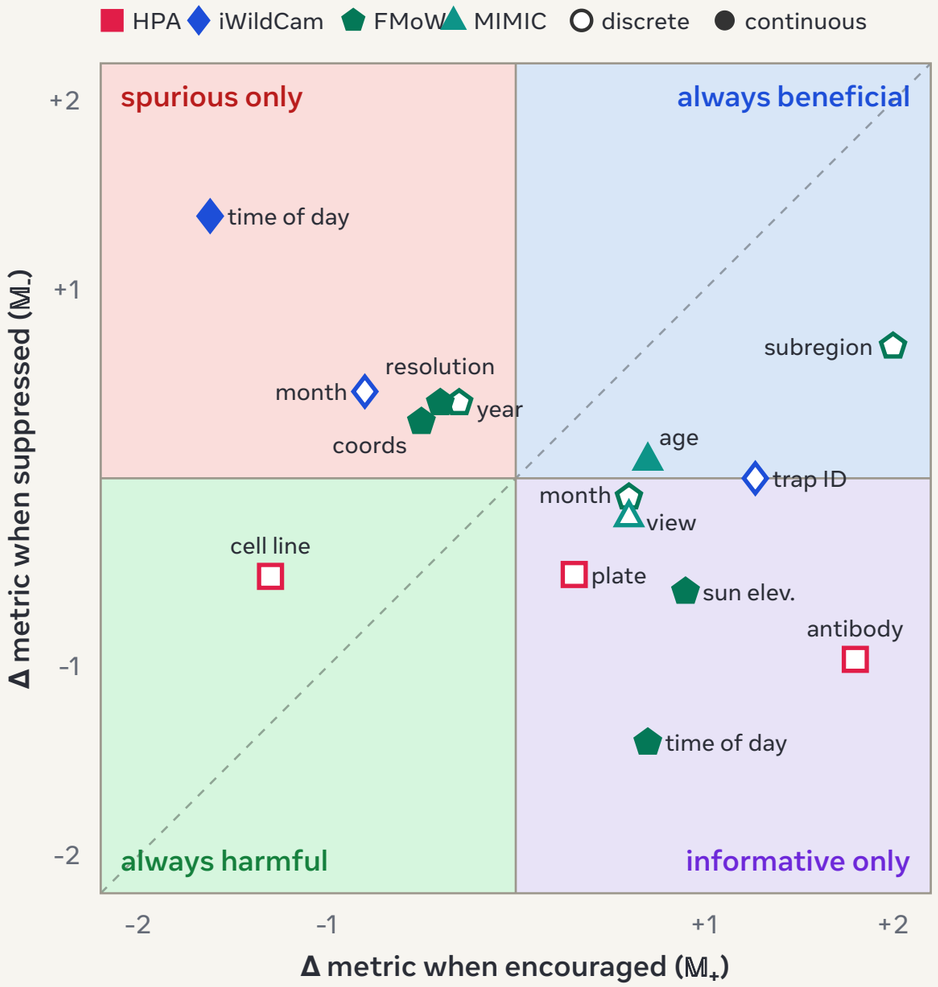
FINO

Published reference

References: Ouyang et al. 2019 (HPA), AutoFT 2024 (iWildCam, FMoW), Moutakanni et al. 2024 (MIMIC). Upstream data, supervision, resolution and inference differ between papers, so this is not a controlled comparison.

ANALYSIS

Which metadata helps?



Each of the **15 candidate factors** is added alone, either encouraged (*x*) or suppressed (*y*). We report the change in the target metric compared to no guidance.

- **Always beneficial**: useful as an auxiliary signal either way (FMoW subregion, MIMIC age).
- **Informative only / spurious only**: most factors, they help only with the right sign.
- **Always harmful**: only HPA cell line, which is entangled with antibody identity.

ANALYSIS

What if we guide with expert labels instead?

Same adaptation and evaluation. We only replace the metadata with the downstream **expert labels** (SupCon).

EXPECTATION Expert labels are the very task we evaluate on, so they should guide better than metadata.

Guidance during adaptation	FMoW OOD WGA	HPA OOD F1
no guidance (SSL adaptation)	48.4	53.0
expert labels	48.3 −0.1	54.5 +1.5
expert labels + metadata	51.0 +2.6	54.0 +1.0
metadata only (FINO)	**53.3** +4.9	**54.7** +1.7

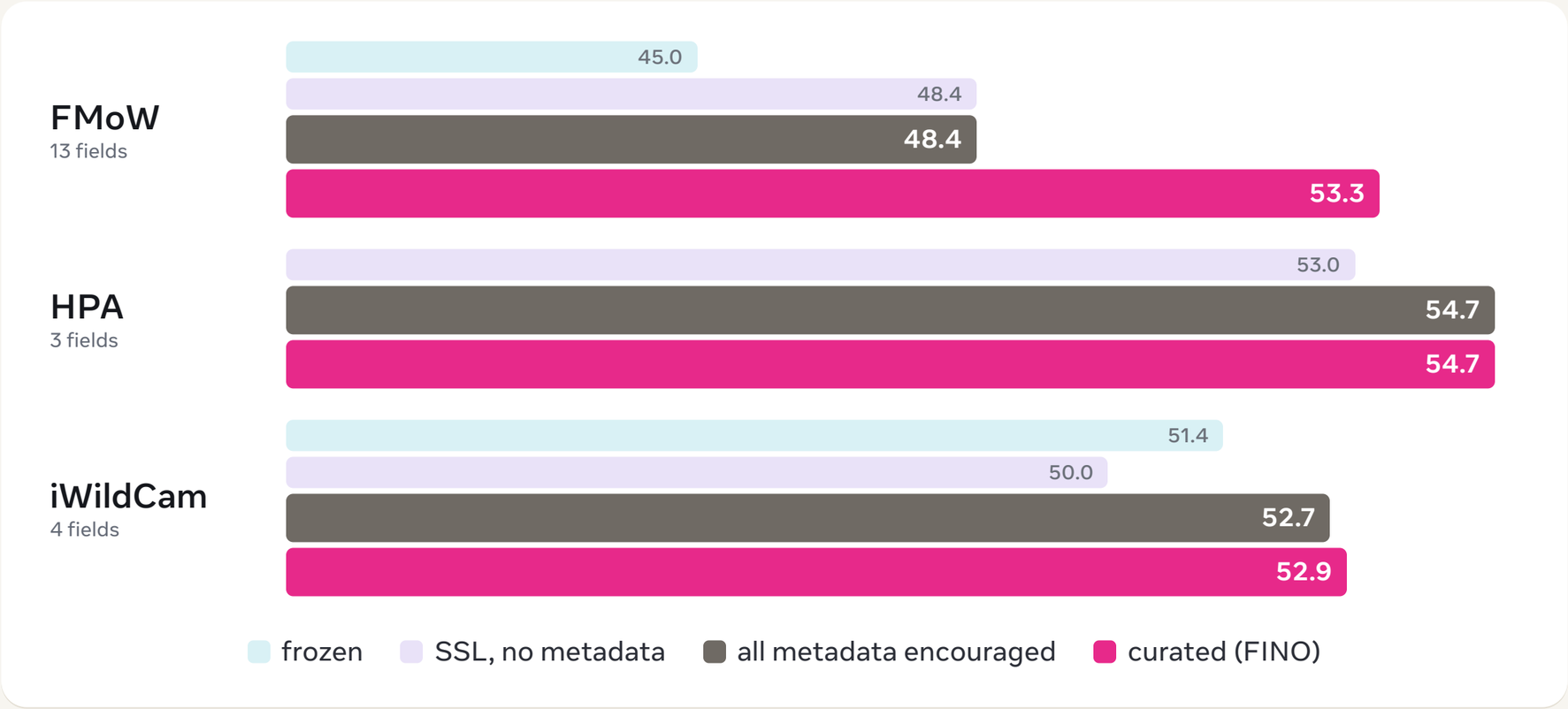
It does not: metadata alone is best and adding expert labels hurts.

Differences relative to SSL adaptation without guidance. HPA at 224^2 resolution.

ANALYSIS

Does the metadata need to be curated?

Stress test: encourage **all** available metadata fields, without any selection.



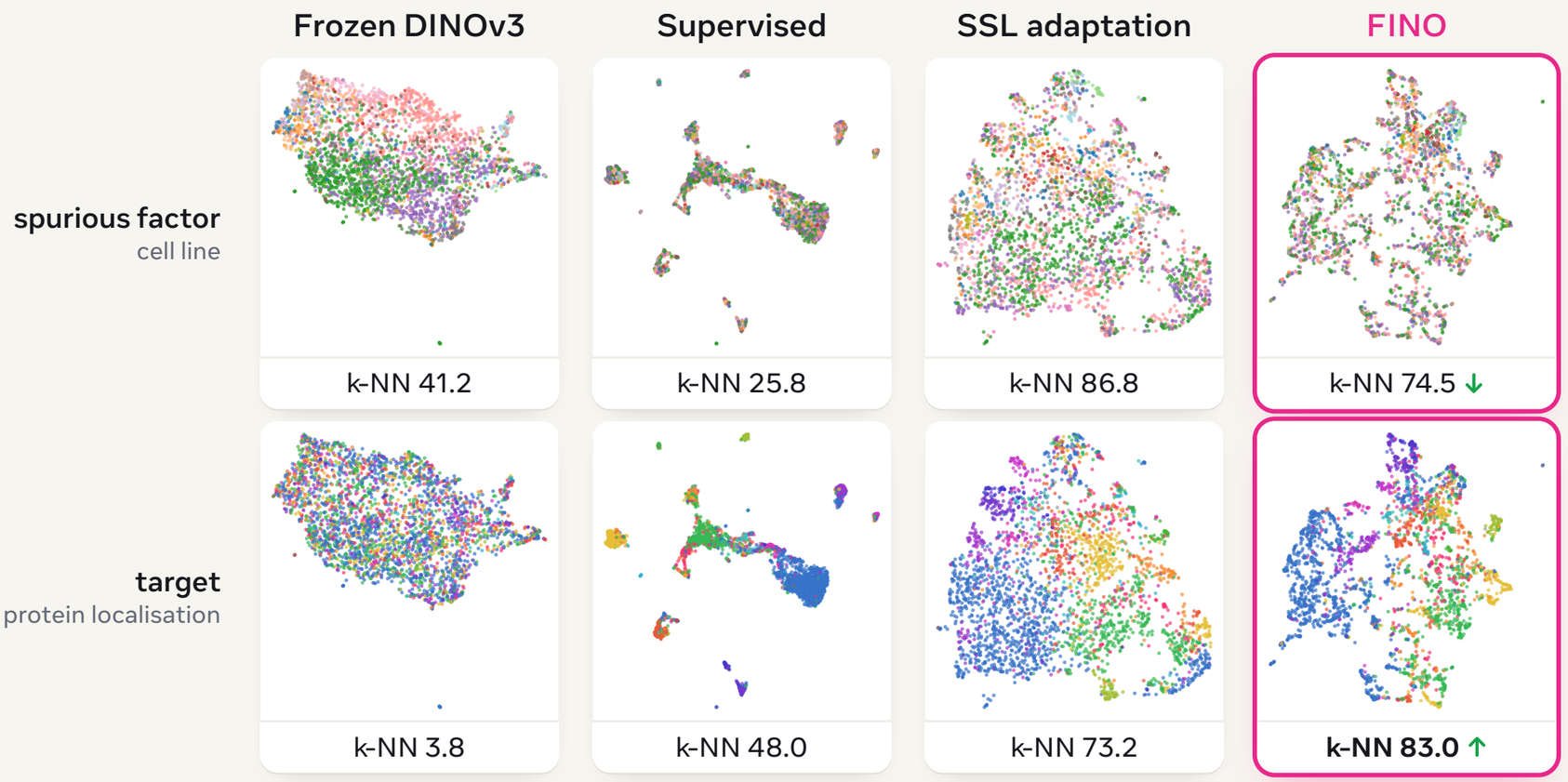
Encouraging everything is never worse than SSL alone. Curation matters most when many fields overlap (FMoW, +4.9).

HPA at 224^2 resolution; no low-resolution frozen baseline for HPA. Single-run stress tests.

ANALYSIS

Representation geometry on HPA

k-NN accuracy on the spurious factor should go **down**, and on the target **up**.

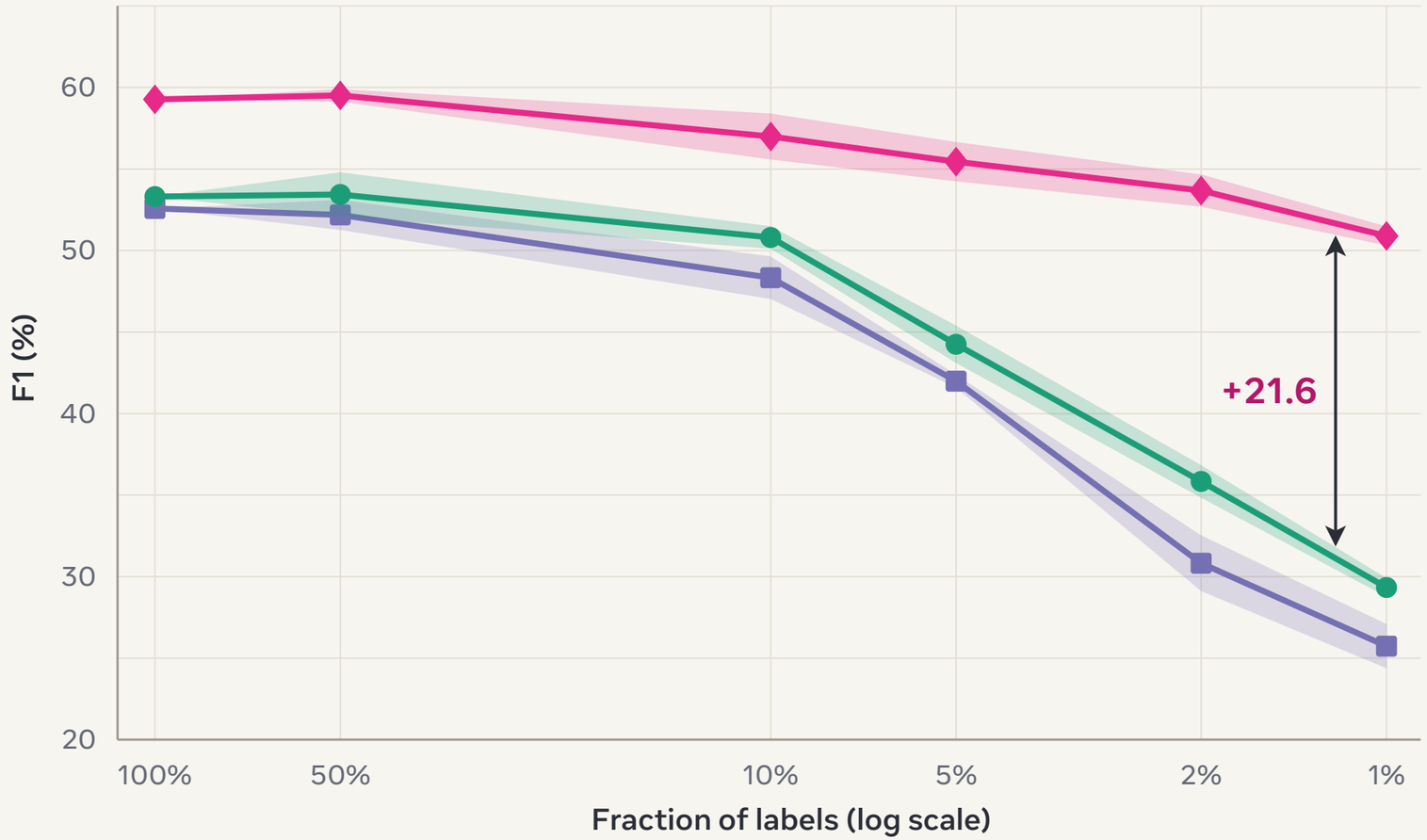


FINO is only guided by antibody, yet cell-line k-NN drops from 86.8 to 74.5.

RESULTS

Low-supervision regime

HPA F1 as a function of the fraction of the 94,270 labelled images used to train the probe.



FINO

Supervised fine-tuning

Group-DRO (cell-line groups)

50.9

FINO F1 with **1%** of the labels, vs. 29.3 for fine-tuning and 25.7 for Group-DRO

2% > 100%

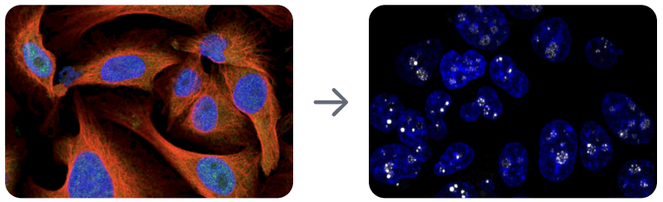
With 2% of the labels, FINO already beats supervised fine-tuning trained on all of them

Single fold; 5 stratified label samplings per point (± std shaded).

RESULTS

Transfer to other datasets

Adapt on one dataset, then freeze, probe and evaluate on a **different** dataset of the same domain, with different classes, channels or tasks.



HPA → OpenCell

100× ARI · clustering

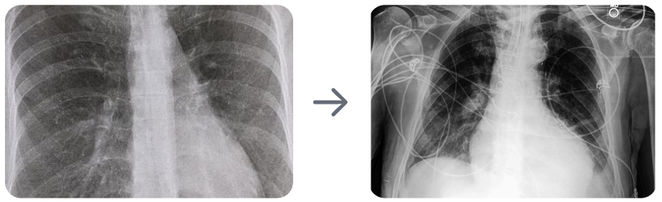
Frozen	46.6
Supervised	53.0
FINO	58.8



FMoW → FLAIR-Hub

mIoU · segmentation

Frozen	62.0
Supervised	61.9
FINO	62.3



MIMIC-CXR → CheXpert

AUROC · classification

Frozen	87.0
Supervised	87.3
FINO	88.0

ROBUSTNESS

Starting from SigLIP2 instead of DINOv3

Same pipeline and hyperparameters, initialised from a SigLIP2 ViT-L/16 (image-text contrastive).



ABLATION

Ablation study

HPA Kaggle private F1, low-resolution (224^2) single-fold setting.

metadata	regularisation	pretrained	discrete guide	F1	
✓	SIGReg	✓	proto	**54.7** ±0.7	ours
✗	SIGReg	✓	n/a	53.0	−1.7
✓	KoLeo	✓	proto	53.8	−0.9
✗	KoLeo	✓	n/a	51.9	−2.8
✓	SIGReg	✗	proto	51.4	−3.3
✓	SIGReg	✓	MLP	51.4	−3.3

First row: mean ± std over 10 complete runs; other rows are single runs.

- **Metadata and SIGReg add up**: −1.7 and −0.9 when removed alone, −2.8 together.
- **Pretraining still matters** after long adaptation to microscopy (−3.3 from scratch).
- **EMA prototypes work better than an MLP classifier** for high-cardinality metadata (−3.3).

CONCLUSION

Main points

01 **Metadata as supervision**

The metadata that comes with the data is enough to adapt a foundation model to a scientific domain, without task labels and without metadata at test time.

02 **Works with any metadata**

Prototypes for discrete fields with thousands of values, a regression head for continuous ones, gradient reversal for spurious factors.

03 **Better representations**

Outperforms supervised, domain adaptation and SSL baselines on 4 domains, holds up with 1% of labels and transfers to other datasets.

04 **The role of each factor matters**

Factors can be informative, spurious, redundant or entangled. A quick dependence check on the metadata table helps assign them.

Thank you!

Who Needs Labels? Adapting Vision Foundation Models With the Metadata You Already Have

End-to-end code, configs & splits provided

HPA · FMoW · iWildCam · MIMIC-CXR

Code

github.com/facebookresearch/dinov3/tree/FINO#metadata-guided-training

Paper

arxiv.org/abs/2606.05107

Meta

École des Ponts ParisTech

COLUMBIA

COLUMBIA UNIVERSITY IRVING MEDICAL CENTER