

Do Pathology Foundation Models Learn Biology? Uncertainty-Based Evaluation on AML Cytomorphology

Aadil Bhat

`aadil.bhat@uok.edu.in`

University of Kashmir

Sajad Kawa

University of Kashmir

Mohammad Munzir

University of Kashmir

Article

Keywords:

Posted Date: August 31st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10474276/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Do Pathology Foundation Models Learn Biology?

Uncertainty-Based Evaluation on AML Cytomorphology

Aadil Rashid Bhat^{1,*}, Sajad Ahmad Kawa¹, Muhammad Munzir Ahanger¹

¹Department of Computer Sciences, University of Kashmir, Hazratbal, Srinagar 190006, Jammu & Kash

*Correspondence: aadil.bhat@uok.edu.in

Abstract

Pathology foundation models are increasingly deployed as feature extractors for medical image analysis, yet their advantage over general-purpose vision models in unsupervised settings has not been directly tested. We held the clustering pipeline fixed and varied only the backbone across three pretraining regimes on the AML-Cytomorphology_LMU dataset (17,447 single-cell images, 15 leukaemic cell types): ResNet50 (supervised, ImageNet), DINOv2-large (self-supervised, natural images), and Virchow2 (self-supervised, 3.1 million pathology whole-slide images). Per-cell uncertainty scores served as a probe of biological structure. Virchow2 uncertainty differed substantially across cell types (Kruskal-Wallis $H = 178.2$, $p = 1.5 \times 10^{-30}$), with transitional myeloid precursors showing higher uncertainty than mature stable types ($\bar{\Omega} = 0.384$ vs. 0.313 , $p = 4.3 \times 10^{-15}$). DINOv2 and ResNet50 showed no such pattern. DINOv2 ranked first on all internal clustering metrics (Silhouette=0.522) yet last on biological recovery (ARI=0.033); Virchow2 showed the inverse ordering. Across five seeds, Virchow2 clustering accuracy fluctuated widely (ARI= 0.016 ± 0.006) while its uncertainty structure remained stable, identifying uncertainty as the more reliable signal of representation quality.

Introduction

Acute myeloid leukaemia arises when myeloid progenitor cells fail to complete normal differentiation, accumulating instead as immature blasts in the marrow and peripheral blood[1]. Morphological examination of stained smears remains central to diagnosis, but the task is demanding: a single case can take an experienced haematologist close to an hour, and inter-observer disagreement rates of 15–30% have been reported even among specialists[2]. Automated analysis is an obvious target for reducing this workload, yet most deep learning approaches have relied on large annotated collections that are expensive to assemble[3, 4]. Annotation-free methods sidestep this requirement by grouping cells according to learned visual similarity, but without labels during training there is no guarantee that the groupings produced correspond to anything biologically coherent.

The AML-Cytomorphology_LMU dataset[5] provides a useful testbed for this question. Its 17,447 single-cell images span 15 expert-defined morphological categories; because the labels exist but need not be used during training, they can serve as a post-hoc check on whether unsupervised groupings track real cell biology. Pathology-specific foundation models such as Virchow2[6], UNI[7], and CONCH[8] have shown strong results on supervised pathology benchmarks[9], but whether that advantage carries over to unsupervised settings is an open question. A large general-purpose model like DINOv2[10], trained on natural images rather than slides, might perform comparably if scale matters more than domain.

To isolate the effect of the backbone, we fixed the clustering pipeline (FuzzyGNN-Clust) and tested three extractors: ResNet50[11] as a supervised convolutional baseline, DINOv2-large[10] as a large self-supervised general-purpose model, and Virchow2[6] as a pathology-domain model. Alongside conventional clustering metrics, we computed per-cell uncertainty scores and asked whether they reflect known myeloid differentiation structure — a test applicable in genuinely unsupervised scenarios where no labels are available at inference time.

Virchow2 is the only backbone whose uncertainty aligns with differentiation biology across five independent runs ($H = 178.2$, $p = 1.5 \times 10^{-30}$; transitional vs. stable: $p = 4.3 \times 10^{-15}$).

DINOv2 and ResNet50 produce flat uncertainty profiles across all 15 cell types. DINOv2 ranks first on every internal clustering metric yet last on biological recovery — a complete inversion with direct implications for how backbone representations are evaluated in medical imaging.

Results

Clustering Performance

Table 1 summarises clustering performance for all three backbones. All permutation tests yield $p \approx 0$, confirming that clustering exceeds chance for every backbone.

Table 1: Clustering performance on AML-Cytomorphology_LMU ($N = 1,577$). ARI and NMI computed post-hoc against 15 expert labels. Best values in **bold**.

Backbone	Pretraining	ARI $\uparrow$	NMI $\uparrow$	Silhouette $\uparrow$	DBI $\downarrow$	CHS $\uparrow$
ResNet50	ImageNet (supervised)	0.041	0.130	0.380	0.954	1,335
DINOv2	ImageNet (self-sup.)	0.033	0.106	0.522	0.657	5,195
Virchow2	3.1M pathology WSIs	0.113	0.270	0.263	1.459	615

Virchow2 achieves ARI=0.113 at $c = 15$, a $2.8\times$ improvement over ResNet50 and $3.4\times$ over DINOv2. Figure 1 shows the rank reversal between internal and external metrics.

Metric Dissociation

The ranking of backbones by internal metrics is the exact inverse of their ranking by external metrics:

$$\text{Internal: } \text{DINOv2} > \text{ResNet50} > \text{Virchow2} \quad (1)$$

$$\text{External: } \text{Virchow2} > \text{ResNet50} > \text{DINOv2} \quad (2)$$

Table 2 shows the dominant cell types in each cluster at silhouette-optimal $c = 2$. DINOv2 groups lymphocytes, segmented neutrophils, and eosinophils together — cells that share

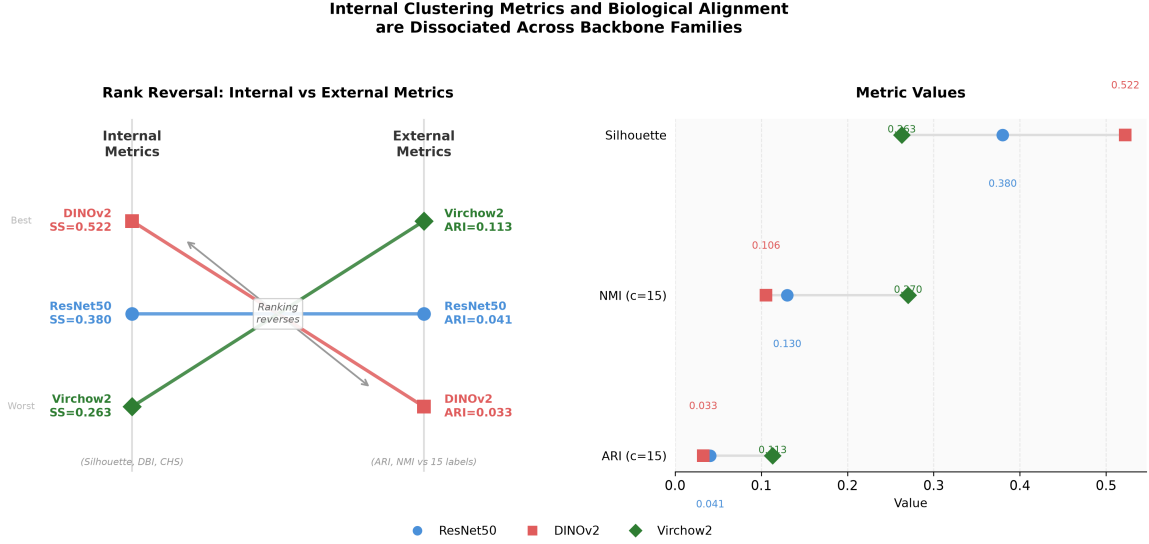


Figure 1: **Rank reversal between internal and external clustering metrics.** Left: slope graph showing how backbone rankings flip between internal and external evaluation. Right: metric values per backbone. DINOv2 leads on all internal metrics; Virchow2 leads on both external metrics.

compact nuclear morphology but belong to entirely different lineages. Virchow2 Cluster 1 shows partial monocytic and myeloid coherence (MOB, MYO, MON), indicating some lineage awareness in its representations.

Table 2: **Cluster composition at silhouette-optimal $c = 2$.** Top 3 cell types with counts shown.

Backbone	Cluster 0	Cluster 1	Coherence
ResNet50	EBO(101), MYB(82), MOB(81)	EOS(104), MYO(100), NGS(91)	Mixed
DINOv2	LYT(118), NGS(105), EOS(101)	MMZ(111), MOB(105), MYB(101)	Mixed
Virchow2	EBO(123), NGS(118), LYT(105)	MOB(133), MYO(100), MON(99)	Partial

Uncertainty Structure

Figure 2 shows per-cell-type uncertainty for all three backbones. Table 3 reports Kruskal-Wallis statistics.

Transitional cell types (MYB, PMB, PMO, MYO, MMZ, MOB) carry significantly higher Virchow2 uncertainty ($\bar{\Omega} = 0.384$) than stable mature types (NGS, NGB, LYT, EOS, BAS;

Table 3: **Kruskal-Wallis uncertainty structure test results.**

Backbone	KW H	KW p	Ω range	Structure
ResNet50	38.8	3.9×10^{-4}	0.32–0.46	Weak
DINOv2	31.0	5.6×10^{-3}	0.25–0.39	Weakest
Virchow2 (1 seed)	167.1	2.7×10^{-28}	0.30–0.60	Strong
Virchow2 (5 seeds)	178.2	1.5×10^{-30}	–	Confirmed

$\bar{\Omega} = 0.313$), $p = 4.3 \times 10^{-15}$.

UMAP Visualisation

Figure 3 shows UMAP projections comparing ground-truth cell types with predicted clusters for all three backbones.

Multi-Seed Reproducibility

Table 4 reports Virchow2 results across five seeds. ARI varies considerably (0.016 ± 0.006), reflecting sensitivity to GAT initialisation. The uncertainty structure (H ranging 167–178) is stable throughout.

Table 4: **Virchow2 multi-seed results** (seeds: 42, 123, 256, 512, 1024) at silhouette-optimal $c = 2$.

Metric	Mean	Std
ARI	0.016	0.006
NMI	0.042	0.012
Silhouette	0.583	0.049

Discussion

Uncertainty as a Differentiation Readout

Myeloblasts, promyelocytes, myelocytes, metamyelocytes, and monoblasts occupy transitional positions in the myeloid hierarchy: each shares morphological features with the stage

immediately before and after it. For a model trained on pathology material, these cells are genuinely ambiguous — not because the model is poorly trained, but because the cells themselves sit at biological boundaries. Virchow2 uncertainty reflects this: $\bar{\Omega}$ is consistently higher for transitional types than for morphologically stable mature cells across all five seeds. ResNet50 and DINOv2 show no such gradient. Given that DINOv2 is the larger model by parameter count, the difference cannot be attributed to capacity alone. The more parsimonious explanation is that exposure to millions of pathology slides during pretraining instilled sensitivity to cytological features that natural-image training does not capture. We treat this as a hypothesis rather than a conclusion, since confirming it would require targeted probing experiments beyond the scope of this study.

Why DINOv2 Clusters Look Good but Are Not

With 307 million parameters and a self-distillation objective applied to 142 million photographs, DINOv2 is a formidable general-purpose encoder. Its Silhouette score of 0.522 and Davies-Bouldin index of 0.657 are the best in this comparison by a wide margin. The clusters are tight and well-separated in embedding space. The problem is what they contain: lymphocytes, segmented neutrophils, and eosinophils grouped together on the basis of compact nuclear morphology, a visual property that cuts across lineage boundaries. Giemsa staining produces colour and texture patterns — nuclear-to-cytoplasmic ratio, granule density, chromatin texture — that carry diagnostic weight in haematology but bear no resemblance to the features that dominate natural image datasets. DINOv2’s representations appear to organise cells by the wrong visual axes for this domain.

The Metric Inversion Problem

The complete reversal of rankings between internal and external metrics is not a marginal effect. DINOv2 is first on Silhouette, Davies-Bouldin, and Calinski-Harabasz; it is last on ARI and NMI. Virchow2 is last on all three internal metrics and first on both external ones.

A practitioner who benchmarks backbone representations using only label-free metrics — as is common when labels are unavailable — would select the wrong model for AML analysis. Any domain where visual compactness and biological coherence diverge will produce the same trap.

Single-Seed Results Are Unreliable

Virchow2 with seed 42 gave ARI=0.113. Across five seeds the mean drops to 0.016 with a standard deviation of 0.006. The gap between the best single run and the average is large enough to change the interpretation of the experiment entirely. The uncertainty structure, by contrast, gave Kruskal-Wallis H between 167 and 178 across all five seeds — a range of roughly 6% around a stable signal.

Limitations

Graph memory constraints forced a stratified subsample of 1,577 images rather than the full 17,447; results on the complete dataset may differ. Only three backbones were tested. A broader comparison including UNI, CONCH, and other recent pathology models[9] would clarify whether the pattern observed here generalises across the foundation model landscape. The findings are specific to AML cytomorphology; other haematological conditions with different differentiation hierarchies may not show the same uncertainty structure. The mechanism linking pathology pretraining to structured uncertainty remains unconfirmed.

Conclusion

Three findings emerge from this study. First, Virchow2 uncertainty tracks myeloid differentiation boundaries in a way that neither DINOv2 nor ResNet50 does, and this pattern holds across five independent runs. Second, the backbone that produces the most geometrically compact clusters (DINOv2) recovers the least biological structure, while the backbone with

the loosest internal geometry (Virchow2) recovers the most — a complete inversion that makes internal-only benchmarking unreliable for this type of task. Third, Virchow2 clustering accuracy is highly seed-dependent whereas its uncertainty structure is not, which has practical consequences for how results from single experimental runs should be interpreted.

Domain-specific pretraining leaves a detectable imprint on how a model distributes uncertainty across biologically distinct cell types, and this imprint is more informative than standard clustering scores when the goal is biological coherence rather than geometric tidiness.

Funding

This research received no specific funding.

Data Availability

The AML-Cytomorphology_LMU dataset is publicly available from the Cancer Imaging Archive at <https://doi.org/10.7937/tcia.2019.dc64i46r>. No new data were generated in this study.

Author Contributions

A.R.B.: Conceptualisation, Methodology, Software, Formal Analysis, Investigation, Writing – Original Draft, Writing – Review & Editing, Visualisation. S.A.K.: Writing – Review & Editing, Validation. M.M.A.: Writing – Review & Editing, Validation.

Competing Interests

The authors declare no competing interests.

Acknowledgements

The authors thank the creators of the AML-Cytomorphology_LMU dataset. Portions of this manuscript were drafted with assistance from Amazon Q Developer; all scientific content and conclusions are the sole responsibility of the authors.

References

- [1] D.A. Arber et al. The 2016 revision to the WHO classification of myeloid neoplasms and acute leukemia. *Blood*, 127(20):2391–2405, 2016. doi: 10.1182/blood-2016-03-643544.
- [2] J.M. Bennett et al. Proposals for the classification of the acute leukaemias: French-American-British (FAB) co-operative group. *Br. J. Haematol.*, 33(4):451–458, 1976. doi: 10.1111/j.1365-2141.1976.tb03563.x.
- [3] C. Matek, S. Schwarz, K. Spiekermann, and C. Marr. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. *Nat. Mach. Intell.*, 1(11):538–544, 2019. doi: 10.1038/s42256-019-0101-9.
- [4] J. Park, S. Kim, and H. Lee. Deep learning model for differentiating acute myeloid and lymphoblastic leukemia in peripheral blood cell images. *Sci. Rep.*, 14:8234, 2024. doi: 10.1038/s41598-024-58234-x.
- [5] C. Matek, S. Schwarz, K. Spiekermann, and C. Marr. AML-Cytomorphology_LMU dataset. Cancer Imaging Archive, 2019.
- [6] E. Zimmermann et al. Virchow2: Scaling self-supervised mixed magnification models in pathology, 2024.
- [7] R.J. Chen et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.*, 30:850–862, 2024. doi: 10.1038/s41591-024-02857-3.

- [8] M.Y. Lu et al. A visual-language foundation model for computational pathology. *Nat. Med.*, 30:863–874, 2024. doi: 10.1038/s41591-024-02856-4.
- [9] P. Neidlinger et al. Benchmarking foundation models as feature extractors for weakly supervised computational pathology. *Nat. Commun.*, 2025.
- [10] M. Oquab et al. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024.
- [11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 770–778, 2016.
- [12] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proc. 33rd Int. Conf. Mach. Learn. (ICML)*, pages 1050–1059, 2016.
- [13] P. Veličković et al. Graph attention networks. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.
- [14] J.C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.
- [15] P.J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20:53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.
- [16] D.L. Davies and D.W. Bouldin. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-1(2):224–227, 1979. doi: 10.1109/TPAMI.1979.4766909.
- [17] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Commun. Statist.*, 3(1):1–27, 1974. doi: 10.1080/03610927408827101.
- [18] W.M. Rand. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.*, 66(336):846–850, 1971. doi: 10.2307/2284239.

[19] A. Strehl and J. Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617, 2002.

Methods

Dataset

The AML-Cytomorphology_LMU collection[5] comprises 17,447 Giemsa-stained peripheral blood cell images from 100 AML patients, spanning 15 morphological categories: MYB, PMB, PMO, MYO, MMZ, NGB, NGS, MOB, MON, EOS, BAS, LYT, LYA, EBO, and KSC. All images were captured at 100 $\times$ oil-immersion magnification. Cell-type labels were withheld from every stage of the computational pipeline and used only to evaluate outputs after the fact. A stratified random subsample of 1,577 images was drawn (approximately 105 cells per class) due to graph memory constraints.

Backbone Architectures

ResNet50[11] (25M parameters, supervised ImageNet-1K): the 2048-dimensional final pooling vector was projected to 512 dimensions with L2 normalisation.

DINOv2-large[10] (307M parameters, self-supervised on 142M natural photographs): the 1024-dimensional CLS token was projected to 512 dimensions.

Virchow2[6] (632M parameters, ViT-H/14, self-supervised on 3.1M pathology whole-slide images): the 1280-dimensional CLS token was projected to 512 dimensions.

All backbones were frozen during experiments; gradient flow was restricted to the projection head.

FuzzyGNN-Clust Pipeline

Features were extracted with dropout rate 0.3. Monte Carlo dropout[12] with $T = 20$ forward passes provided per-sample epistemic uncertainty u_i^{epi} . A symmetric k -nearest neighbour graph ($k = 10$, cosine similarity) was constructed, then processed by a 2-layer Graph Attention Network[13] (8 heads, ELU activation) trained with a self-supervised reconstruction objective to produce 128-D embeddings Z . Fuzzy c-means[14] ($m = 2$, k -means++ initialisation, 10 restarts) was applied to Z , with the optimal cluster count c chosen by silhouette maximisation over $c \in \{2, \dots, 19\}$. A unified uncertainty score combined epistemic uncertainty and fuzzy membership entropy:

$$\Omega_i = \beta \tilde{u}_i^{\text{epi}} + (1 - \beta) \tilde{H}_i^{\text{fuzz}}, \quad \beta = 0.5 \quad (3)$$

Evaluation

Internal metrics — Silhouette[15], Davies-Bouldin[16], Calinski-Harabasz[17] — assessed cluster geometry without reference to ground truth. External metrics — ARI[18] and NMI[19] — compared predicted assignments against the 15 expert labels post-hoc. A permutation test (1,000 shuffles) verified that clustering exceeded chance for each backbone. Two Kruskal-Wallis tests examined uncertainty: one across all 15 cell types, and one comparing transitional precursors (MYB, PMB, PMO, MYO, MMZ, MOB) against stable mature types (NGS, NGB, LYT, EOS, BAS). Virchow2 experiments were repeated across five random seeds (42, 123, 256, 512, 1024).

Ethical Statement

This study used only publicly available, fully anonymised data. No patient consent or ethical approval was required.

Cell-Type Uncertainty by Backbone
Virchow2 shows strongly structured, cell-type-specific Ω

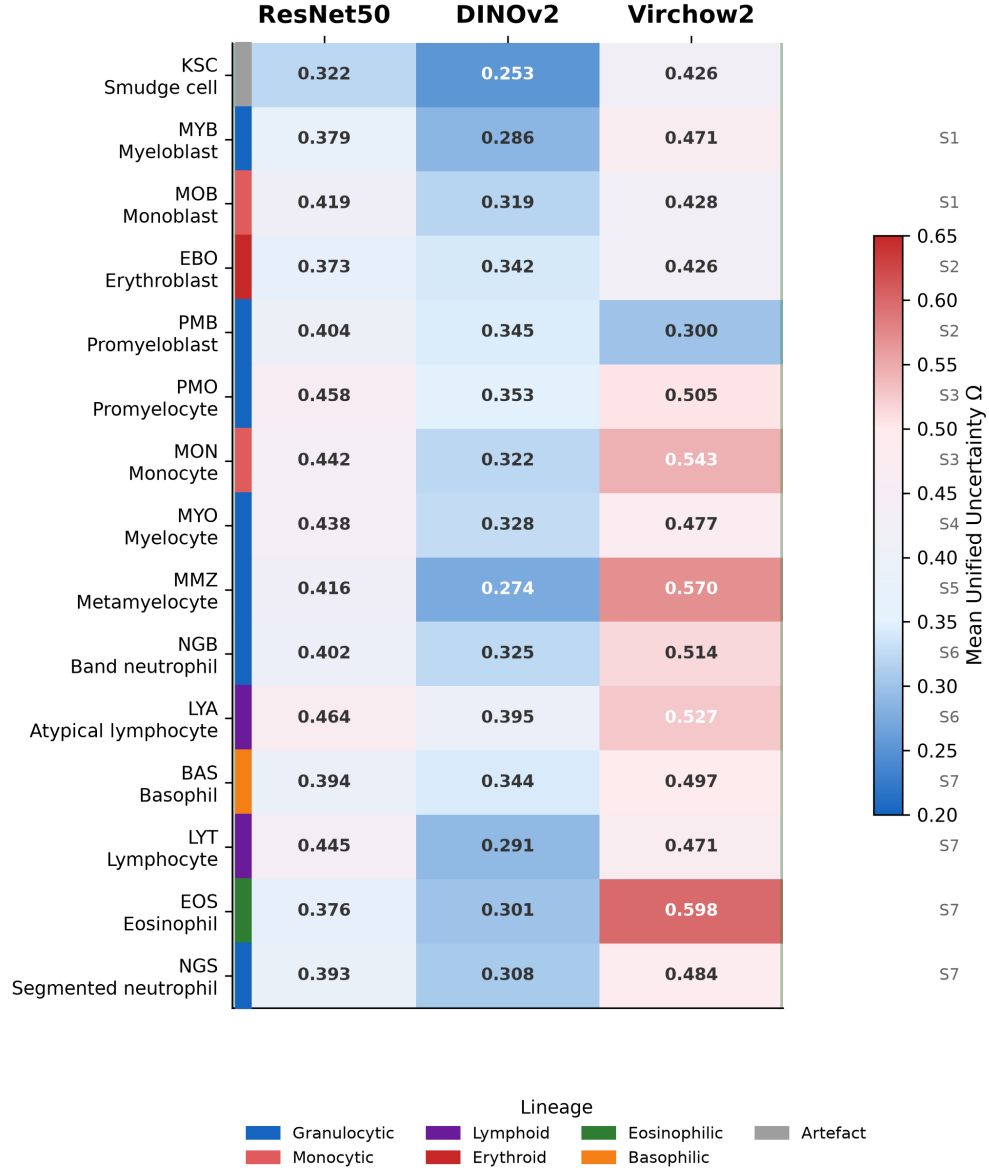


Figure 2: **Cell-type uncertainty heatmap.** Rows ordered by differentiation stage. Virchow2 shows strongly differentiated uncertainty ($H = 178.2$, $p = 1.5 \times 10^{-30}$). DINOv2 and ResNet50 are nearly uniform across cell types.

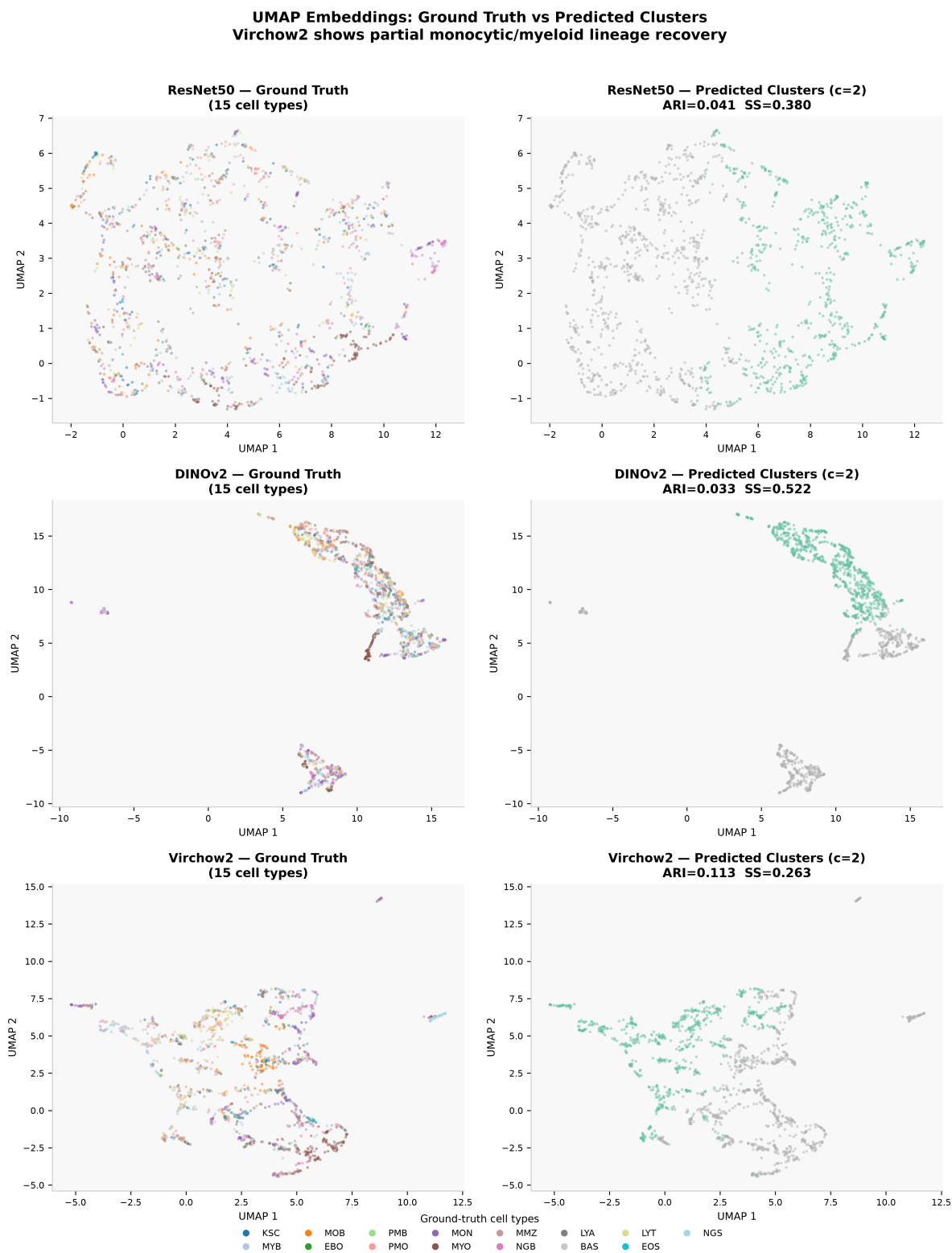


Figure 3: **Six-panel UMAP comparison.** Left column: ground-truth cell types (15 classes). Right column: predicted clusters at silhouette-optimal c . Virchow2 Cluster 1 shows partial monocytic and myeloid coherence.