

HELM on BigEarthNet

Benchmarking hierarchical multi-label
land-cover classification

Nikola Marić · Marjan Stoimchev · Dragi Kocev · Sašo Džeroski

Department of Knowledge Technologies, Jožef Stefan Institute

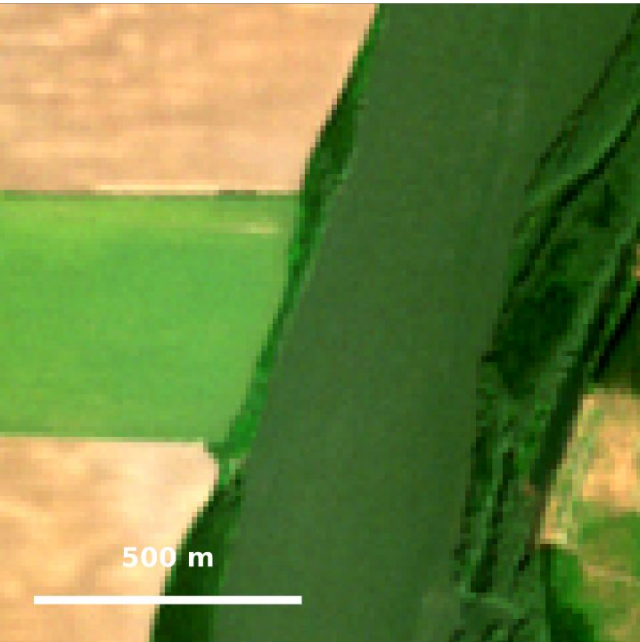
AutoML4EO Workshop at AutoML 2026 · Ljubljana, 1. 10. 2026



Land cover is hierarchical by design

5 → 15 → 44

CORINE Land Cover arranges its classes in nested levels



Sentinel-2 · 2018-04-30 · RGB

One patch: 5 leaf labels switch on 14 of 63 nodes

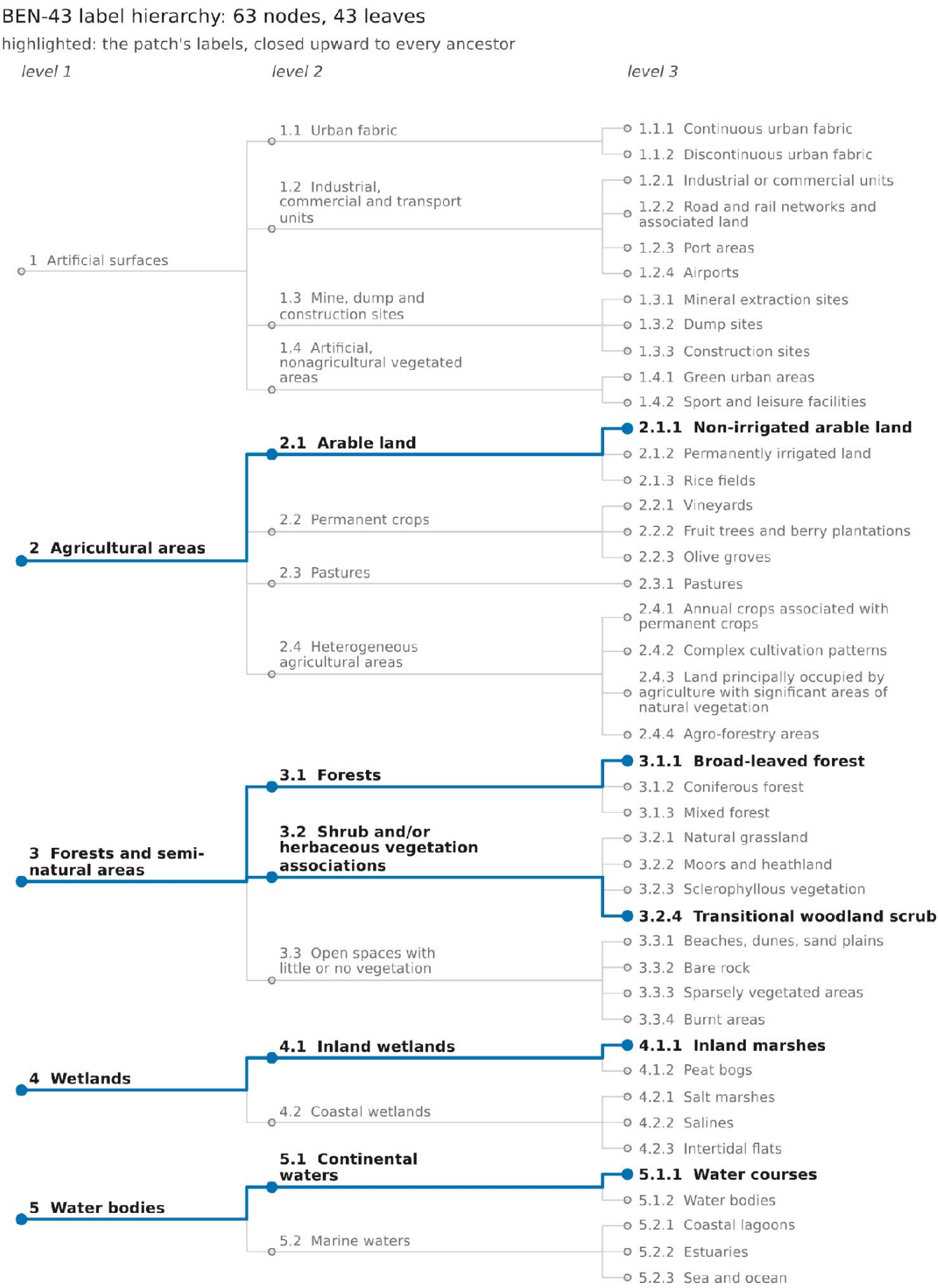
Multi-path: four of the five top-level branches at once

Every leaf implies its ancestors

- Broad-leaved forest
 - ⇒ Forests
 - ⇒ Forests and semi-natural areas

Other national land-cover systems are trees too (UK Environment Agency: 4 levels · China's land-use system: 2 levels)

Source · Feranec et al., 2016 · EEA / Copernicus CLC · Sumbul et al., 2019; 2021 · Ai et al., 2025



Mature in ML, scarce in Earth observation

Machine learning



HMLC is a long-standing ML task: gene function, text, images — decision trees (Vens et al., 2008), networks (HMCN 2018, C-HMCNN 2020), contrastive learning (2022)

Hierarchies help in low-label regimes

Earth observation



- **Earth observation labels are scarce**
- 2024 · first HMLC remote-sensing benchmarks: BigEarthNet-19/43-HMLC, UCM-HMLC
- 2025–26 · hierarchy-aware models appear: HieraRS · SAHC · MAPLE · HELM
- Flat BigEarthNet benchmarks are well established

“a lack of HMLC datasets and methods in the remote sensing (RS) domain”

Stoimchev et al., ECML PKDD 2024

“these hierarchies are frequently overlooked”

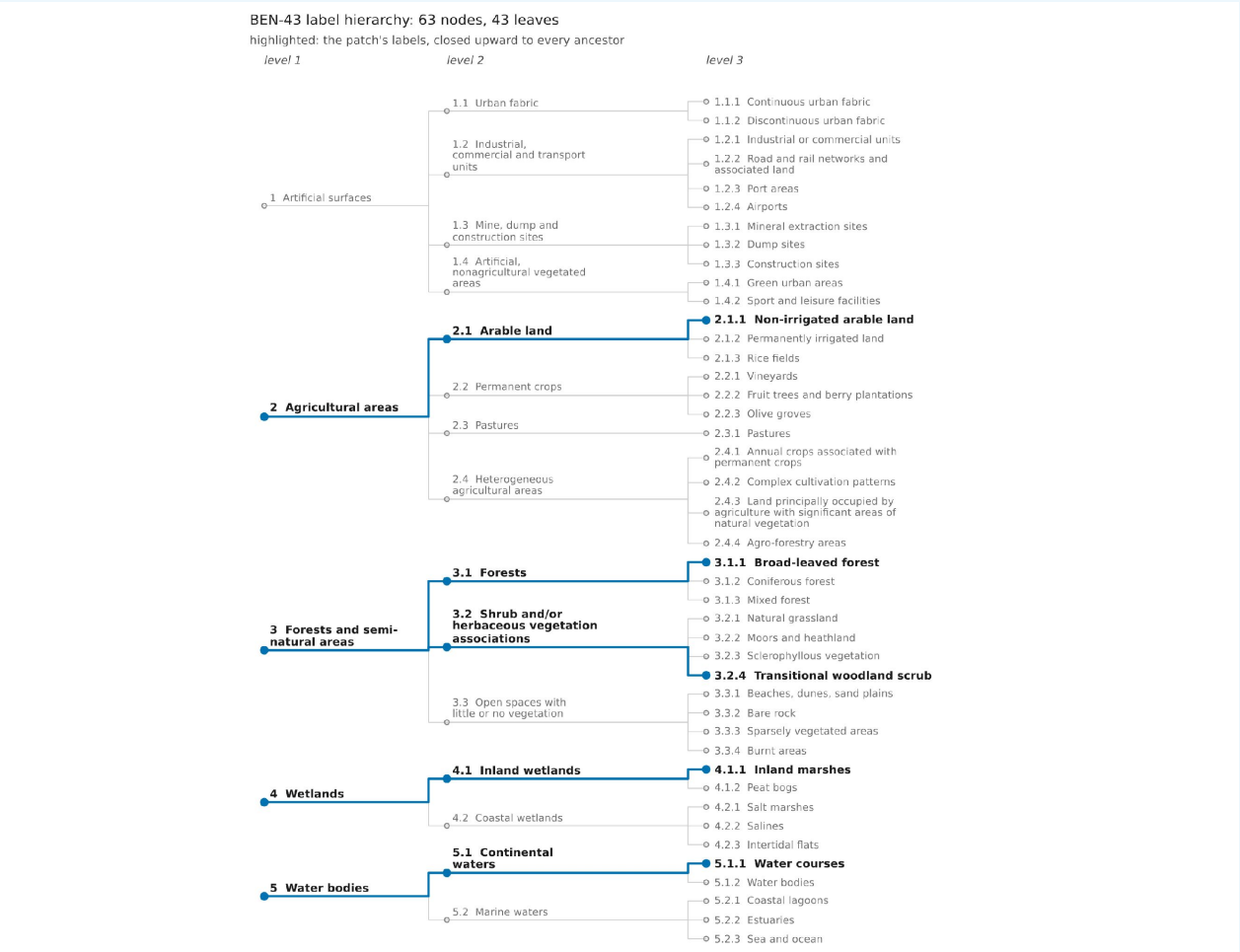
Weikmann, Perantoni & Bruzzone, 2025

Source · Vens et al., 2008 · Wehrmann et al., 2018 · Giunchiglia & Lukasiewicz, 2020 · Zhang et al., 2022 · Stoimchev et al., 2026a/b · Ai et al., 2025 · Weikmann et al., 2025 · Koloski et al., 2025 · Dimitrovski et al., 2023

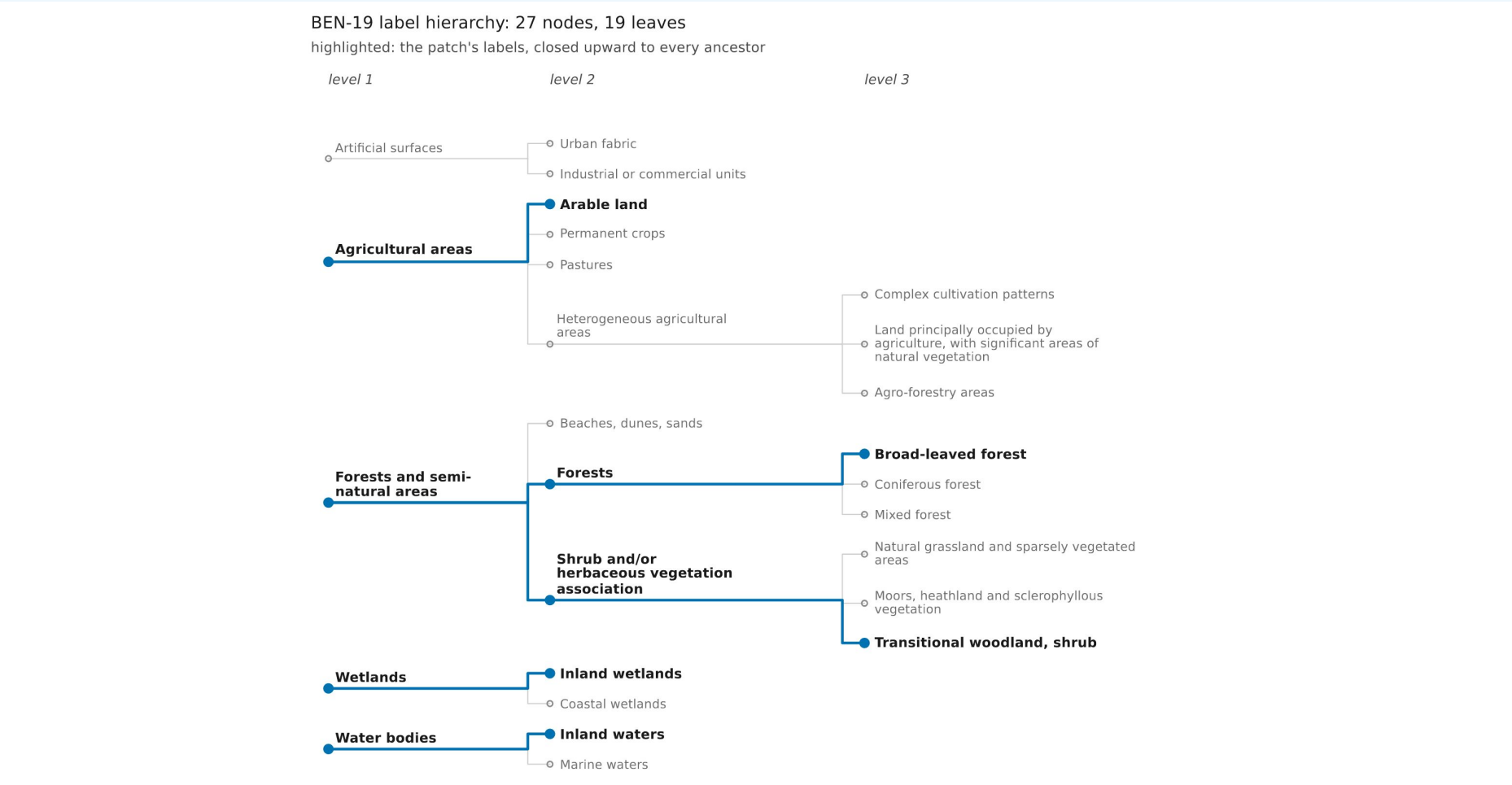
Our push: HELM on BigEarthNet

- BigEarthNet-S2 v1.0, official split: **519,284 patches** (269,695 / 123,723 / 125,866)
- Two hierarchies over the same images: **BEN-43** (63 nodes, uniform depth) · **BEN-19** (27 nodes, variable depth)

BEN-43 tree



BEN-19 tree



Evaluated models

One ViT-B/16 backbone, one output contract, one scoring path, 3 seeds

	HELM-SSL	+ unlabelled patches
	HELM-SL	+ BYOL branch
	HELM_g	+ GraphSAGE over the tree
	Depth-weighted · HMLC	hierarchy in the loss
	MLC	HELM's leaf-only variant
Leaf BCE	flat control	

HMLC = BCE on every node · HELM-SSL is meaningful once part of the archive is unlabelled

13 equivalence tests against HELM's code · 4 consistency corrections · hierarchical metrics

The obliteration study

k =

1	2	3	4	5		10	20	50	100	200	500	1000
---	---	---	---	---	--	----	----	----	-----	-----	-----	------

- k patches drawn per leaf label, $n = k \times \text{leaves}$
 - **BEN-43:** 43 $\rightarrow$ 43,000 patches · **BEN-19:** 19 $\rightarrow$ 19,000 · same seven models, same splits
 - Balanced until rare leaves run out: BEN-43 from k = 100 (rarest leaf capped at 235), BEN-19 from k = 500
-

Full results: leaf micro-F1

Mean of 3 seeds · k = patches per leaf label

BEN-43 · k =	1	2	3	4	5		10	20	50	100	200	500	1000	Mean
█ Leaf BCE	0.313	0.412	0.440	0.460	0.476		0.489	0.523	0.557	0.590	0.606	0.642	0.664	0.514
█ MLC	0.342	0.381	0.411	0.414	0.446		0.468	0.524	0.570	0.588	0.610	0.648	0.663	0.505
█ Depth-weighted	0.308	0.403	0.450	0.464	0.484		0.518	0.552	0.578	0.603	0.628	0.658	0.669	0.526
█ HMLC	0.335	0.401	0.410	0.431	0.450		0.486	0.524	0.573	0.588	0.617	0.649	0.670	0.511
█ HELM_g	0.335	0.403	0.428	0.441	0.453		0.495	0.544	0.584	0.603	0.622	0.659	0.674	0.520
█ HELM-SL	0.366	0.432	0.444	0.454	0.467		0.516	0.569	0.601	0.616	0.629	0.658	0.673	0.535
█ HELM-SSL	0.356	0.438	0.466	0.482	0.492		0.547	0.586	0.607	0.620	0.637	0.660	0.671	0.547

BEN-19 · k =	1	2	3	4	5		10	20	50	100	200	500	1000	Mean
█ Leaf BCE	0.364	0.414	0.452	0.478	0.495		0.532	0.549	0.563	0.574	0.602	0.623	0.678	0.527
█ MLC	0.384	0.416	0.442	0.454	0.448		0.487	0.532	0.557	0.580	0.605	0.635	0.684	0.519
█ Depth-weighted	0.364	0.420	0.460	0.484	0.510		0.534	0.570	0.587	0.601	0.617	0.638	0.686	0.539
█ HMLC	0.351	0.436	0.431	0.470	0.456		0.498	0.530	0.568	0.591	0.613	0.631	0.686	0.522
█ HELM_g	0.408	0.423	0.445	0.466	0.471		0.504	0.541	0.578	0.603	0.617	0.636	0.691	0.532
█ HELM-SL	0.372	0.431	0.451	0.472	0.487		0.519	0.555	0.597	0.616	0.628	0.645	0.695	0.539
█ HELM-SSL	0.405	0.442	0.476	0.492	0.501		0.538	0.576	0.611	0.621	0.638	0.649	0.697	0.554

HELM-SSL is best at 10 of 12 budgets on both hierarchies (mean 0.547 on BEN-43, 0.554 on BEN-19)

The gain is largest at k = 10–50: +0.050 to +0.063 over the flat Leaf BCE on BEN-43

At k = 1000 the models converge: all seven within 0.011 (BEN-43) and 0.019 (BEN-19)

Bold: best at each budget · shaded: per-seed range disjoint from every rival · light-blue rows: hierarchy-aware models

Full results: leaf macro-AUPRC

Mean of 3 seeds · k = patches per leaf label

BEN-43 · k =	1	2	3	4	5		10	20	50	100	200	500	1000	Mean
█ Leaf BCE	0.137	0.171	0.194	0.202	0.217		0.252	0.277	0.335	0.381	0.417	0.457	0.496	0.295
█ MLC	0.120	0.150	0.164	0.173	0.191		0.215	0.262	0.300	0.338	0.373	0.418	0.507	0.267
█ Depth-weighted	0.137	0.173	0.192	0.209	0.215		0.259	0.300	0.334	0.393	0.419	0.469	0.500	0.300
█ HMLC	0.122	0.156	0.158	0.178	0.198		0.219	0.264	0.318	0.358	0.383	0.475	0.503	0.278
█ HELM_g	0.109	0.140	0.156	0.165	0.187		0.200	0.248	0.301	0.335	0.376	0.451	0.495	0.264
█ HELM-SL	0.121	0.147	0.155	0.158	0.182		0.229	0.235	0.296	0.332	0.373	0.399	0.498	0.260
█ HELM-SSL	0.116	0.150	0.164	0.173	0.186		0.221	0.254	0.289	0.317	0.349	0.452	0.489	0.263

BEN-19 · k =	1	2	3	4	5		10	20	50	100	200	500	1000	Mean
█ Leaf BCE	0.287	0.340	0.373	0.398	0.410		0.445	0.448	0.489	0.507	0.552	0.598	0.624	0.456
█ MLC	0.235	0.282	0.309	0.316	0.328		0.365	0.413	0.465	0.471	0.494	0.555	0.576	0.401
█ Depth-weighted	0.283	0.336	0.376	0.394	0.406		0.418	0.460	0.527	0.505	0.559	0.629	0.647	0.462
█ HMLC	0.240	0.282	0.289	0.319	0.333		0.383	0.423	0.459	0.501	0.518	0.521	0.576	0.404
█ HELM_g	0.242	0.253	0.283	0.300	0.315		0.363	0.400	0.434	0.449	0.455	0.561	0.536	0.383
█ HELM-SL	0.229	0.261	0.271	0.280	0.314		0.348	0.422	0.415	0.438	0.498	0.523	0.614	0.384
█ HELM-SSL	0.245	0.274	0.313	0.321	0.339		0.384	0.405	0.439	0.459	0.513	0.513	0.578	0.399

Leaf BCE or Depth-weighted is best at 22 of 24 budgets (all 12 on BEN-19)

The three HELM variants have the three lowest means on both hierarchies (0.260–0.264 vs 0.295 for Leaf BCE on BEN-43)

Ranking orders the models differently from micro-F1: the winner depends on the metric

Bold: best at each budget · shaded: per-seed range disjoint from every rival · light-blue rows: hierarchy-aware models

Full results: internal-node F1 (leaf-propagate)

Mean of 3 seeds · k = patches per leaf label

BEN-43 · k =	1	2	3	4	5		10	20	50	100	200	500	1000	Mean
█ Leaf BCE	0.282	0.366	0.383	0.407	0.426		0.456	0.479	0.479	0.505	0.537	0.562	0.575	0.455
█ MLC	0.315	0.372	0.391	0.382	0.418		0.413	0.490	0.509	0.507	0.539	0.558	0.573	0.456
█ Depth-weighted	0.284	0.363	0.391	0.409	0.434		0.460	0.510	0.504	0.528	0.557	0.574	0.576	0.466
█ HMLC	0.311	0.368	0.380	0.409	0.423		0.449	0.471	0.498	0.530	0.539	0.580	0.590	0.462
█ HELM_g	0.328	0.359	0.373	0.412	0.405		0.450	0.494	0.520	0.530	0.551	0.582	0.591	0.466
█ HELM-SL	0.334	0.376	0.379	0.406	0.418		0.462	0.511	0.526	0.531	0.556	0.569	0.588	0.471
█ HELM-SSL	0.272	0.369	0.390	0.407	0.427		0.465	0.509	0.526	0.535	0.551	0.568	0.585	0.467

BEN-19 · k =	1	2	3	4	5		10	20	50	100	200	500	1000	Mean
█ Leaf BCE	0.433	0.487	0.533	0.551	0.571		0.625	0.649	0.669	0.678	0.692	0.712	0.733	0.611
█ MLC	0.489	0.542	0.559	0.567	0.564		0.581	0.638	0.667	0.684	0.698	0.716	0.737	0.620
█ Depth-weighted	0.445	0.503	0.536	0.564	0.594		0.635	0.666	0.686	0.696	0.711	0.712	0.738	0.624
█ HMLC	0.500	0.558	0.562	0.597	0.576		0.613	0.637	0.673	0.692	0.694	0.723	0.741	0.631
█ HELM_g	0.530	0.567	0.571	0.590	0.588		0.597	0.647	0.683	0.692	0.709	0.720	0.742	0.636
█ HELM-SL	0.514	0.558	0.575	0.583	0.608		0.635	0.651	0.691	0.705	0.710	0.723	0.746	0.642
█ HELM-SSL	0.484	0.570	0.584	0.598	0.606		0.631	0.668	0.698	0.703	0.709	0.721	0.747	0.643

HELM-SL and HELM-SSL have the highest means on both hierarchies; the flat Leaf BCE has the lowest

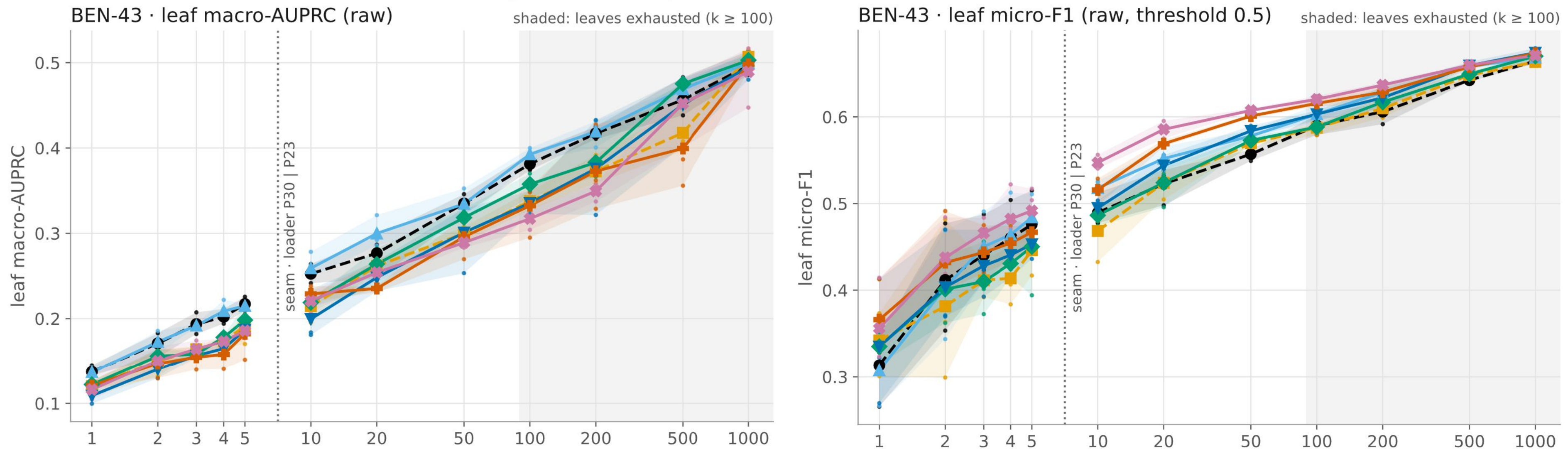
The margin over MLC peaks at k = 10: about +0.05 for both HELM variants on both hierarchies

At k = 1000 the spread shrinks to 0.018 (BEN-43) and 0.014 (BEN-19)

Bold: best at each budget · shaded: per-seed range disjoint from every rival · light-blue rows: hierarchy-aware models

The tree pays at the coarse levels — most when labels are scarce

Leaf BCE MLC Depth-weighted HMLC HELM_g HELM-SL HELM-SSL



Take-home

Land cover is a tree. Earth observation has barely started to use it.

WHAT THE HIERARCHY GIVES

- With few labels, HELM-SSL has the best leaf micro-F1 at 10 of 12 budgets on both hierarchies
- HELM-SSL at $k = 10$ matches Leaf BCE at $k \approx 38$: about 4× fewer labels
- Coarse levels: HELM-SL and HELM-SSL ahead of MLC at 10 of 14 budgets, behind at none
- Unlabelled patches add on top: +0.031 over HELM-SL at $k = 10$

LIMITATIONS AND CAVEATS

- Leaf ranking (macro-AUPRC): no HELM variant beats the flat Leaf BCE at any budget $k \geq 10$
- Gains shrink as labels grow; with 270k labelled patches the tree adds little
- RGB inputs only

OPEN QUESTIONS

- Why does the tree help thresholded leaves (micro-F1) but not leaf ranking (macro-AUPRC)?
- What is the right way to give a model the tree?
- Does the picture change with all spectral bands, other sensors, other nomenclatures?

Let's treat land cover as the hierarchy it is, and score our models on the whole tree, not just the leaves.

Thank you! Questions?

HELM on BigEarthNet: Benchmarking Hierarchical
Multi-Label Land-Cover Classification · AutoML 2026
Workshop Track



nikola.maric@ijs.si
kt.ijs.si



BACKUP

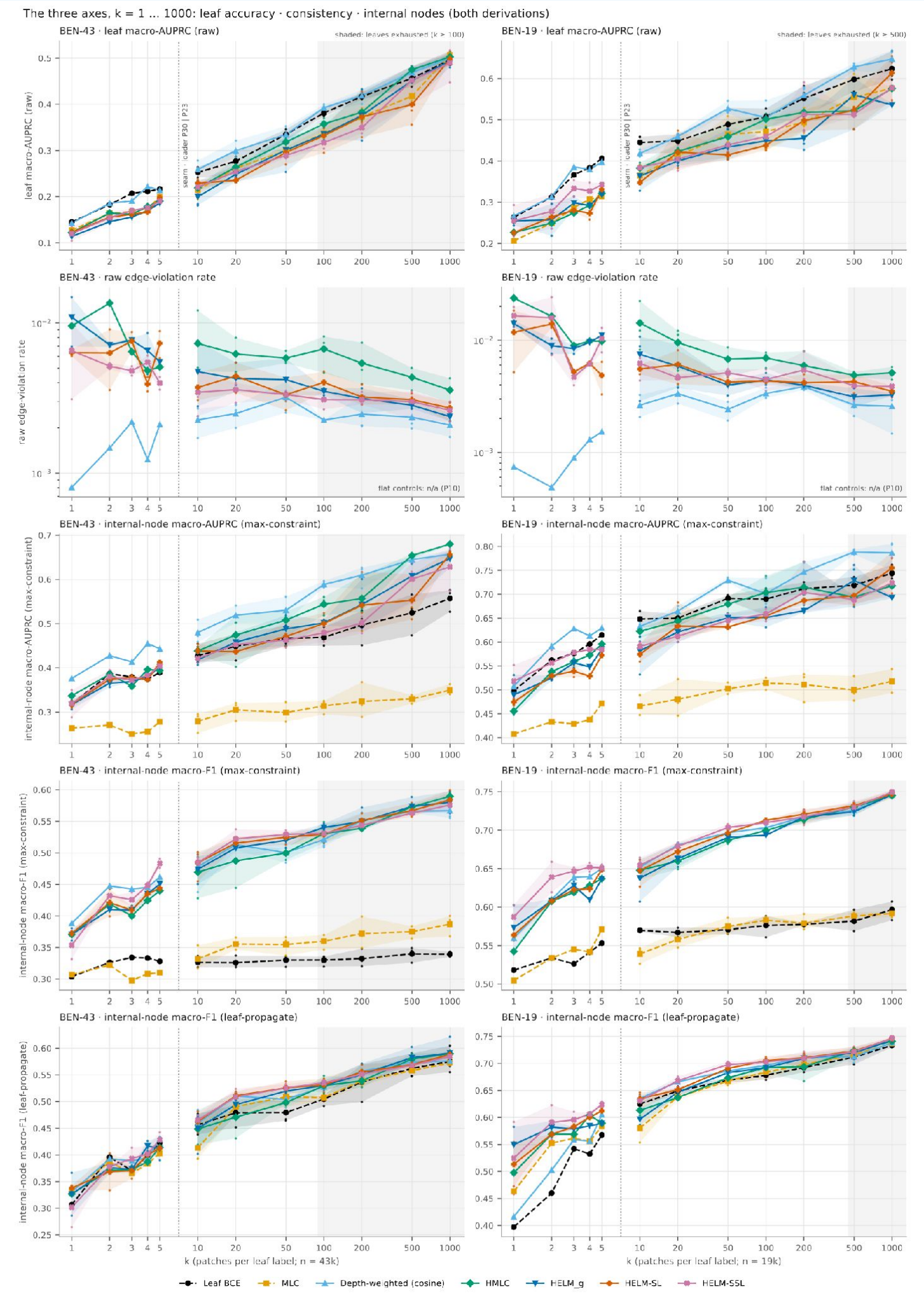
Backup slides

B1–B7 · figures, caveats, provenance

The three axes, all twelve rungs

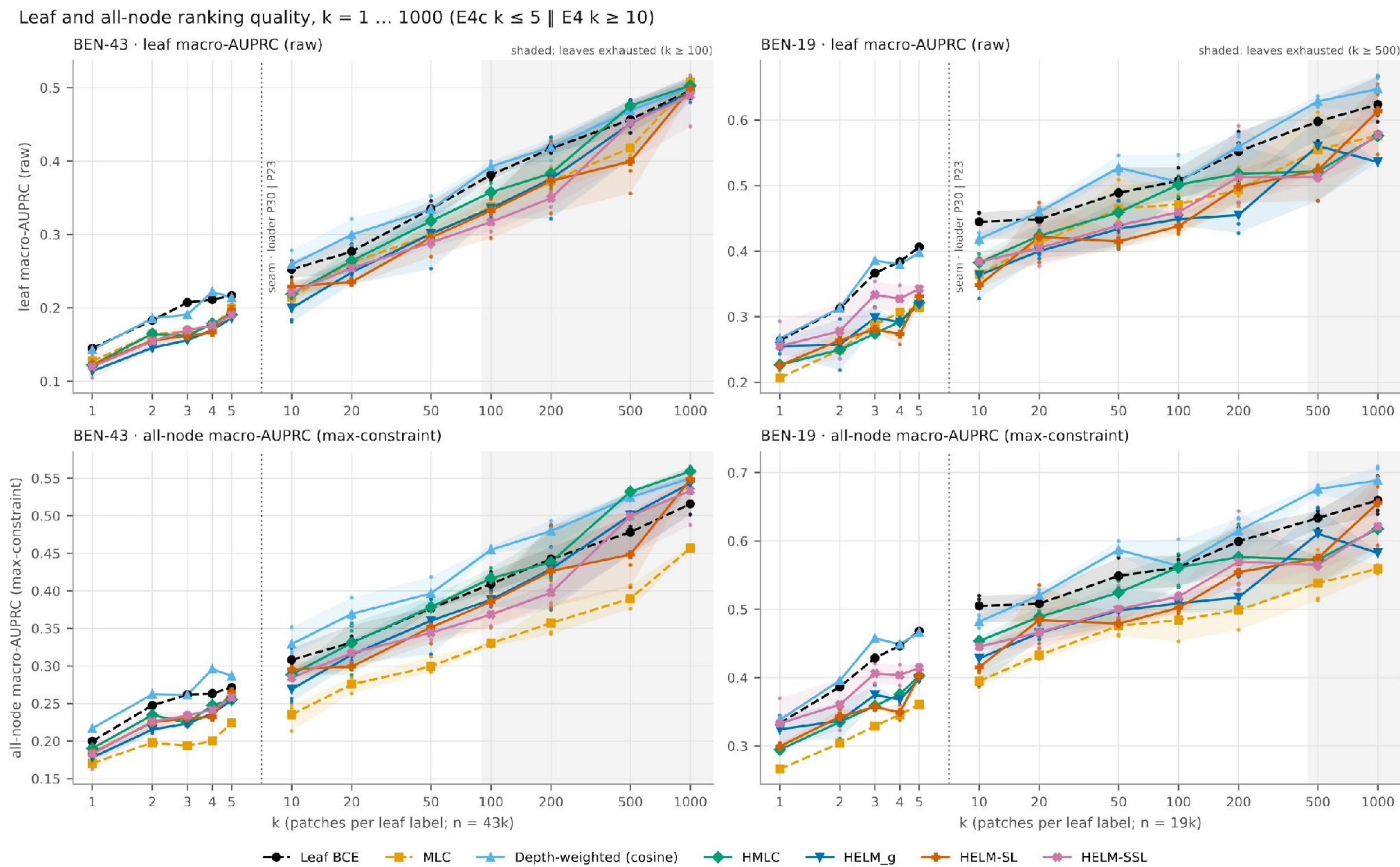
- Leaf accuracy
- Raw edge-violation rate (log scale; flat controls have no admissible value)
- Internal nodes under both derivations (max-constraint, leaf-propagate)

$k \leq 5$ points are preliminary; || marks the protocol seam between $k = 5$ and $k = 10$.



Full curves

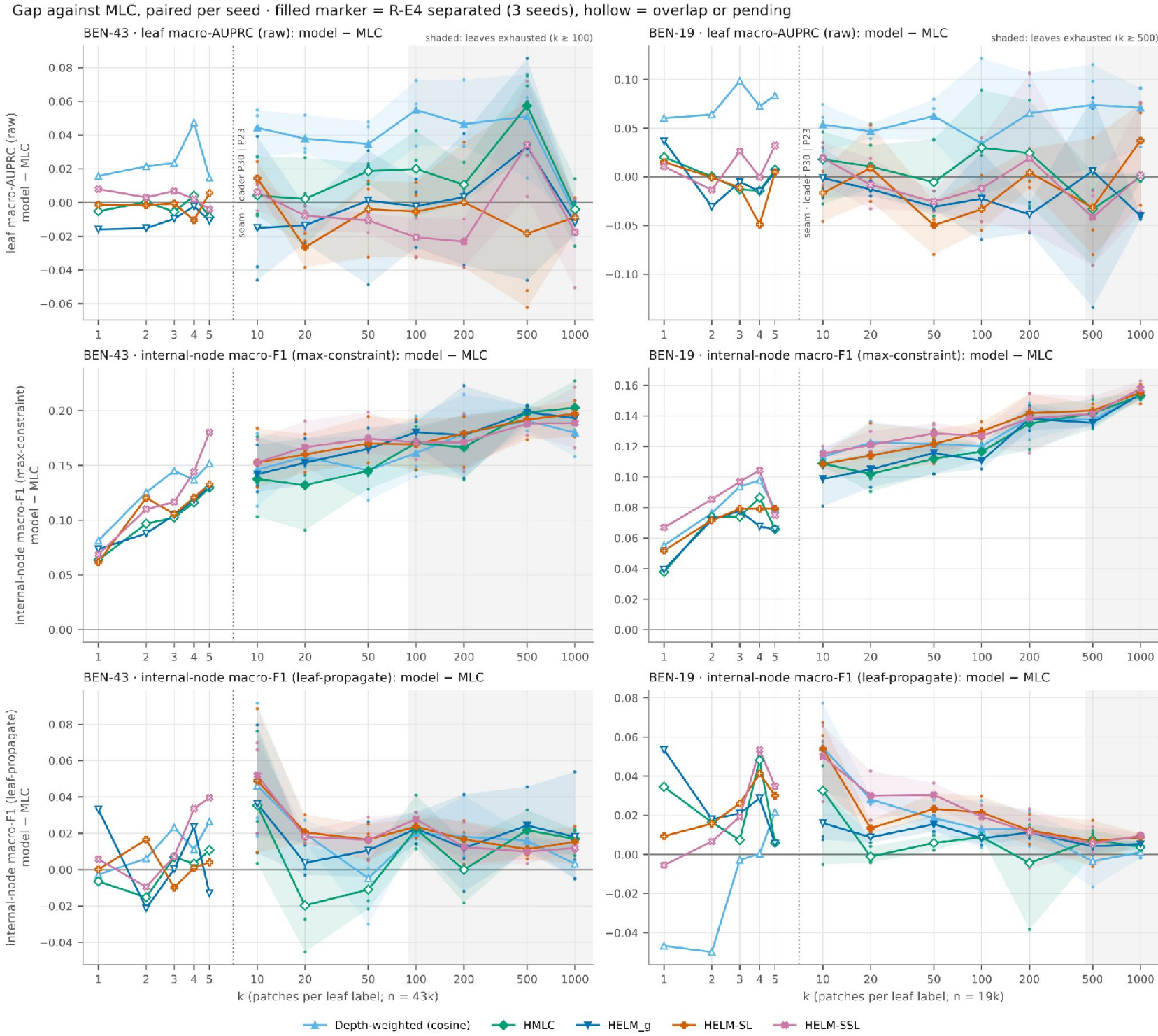
Leaf and all-node macro-AUPRC (max-constraint), both hierarchies · $k \leq 5$ preliminary



Full gap figure

- Rows: leaf macro-AUPRC · internal-node F1 (max-constraint) · internal-node F1 (leaf-propagate), each as model – MLC
- The middle row compares against the flat control's untrained internal head, so it overstates the gap
- Slide 10 uses the leaf-propagate row instead

fig_full_gap_vs_mlc



Consistency at full scale

- All-node models violate **0.0010–0.0027** of parent–child edges
- BEN-19 is harder to keep consistent than BEN-43 (HELM_g 0.0017 vs 0.0012), despite 22 edges vs 58

FIGURE

Paper p. 4, Figure 1 — fit, do not crop

Marić et al., 2026, Figure 1

Caveats, stated up front

- The seam: no claim about the $k = 5 \rightarrow 10$ step
- The F1 selection rule works against HELM-architecture models by $\sim 0.01\text{--}0.04$ val-F1 at small n — affects $k \leq 5$ comparisons with Leaf BCE and Depth-weighted
- The top of BEN-43 is imbalanced (max/min 56 at $k = 1000$; 18 leaves exhausted)
- Depth-weighted keeps its cosine schedule on the curve
- RGB inputs only
- 3 seeds: a separation rule is used, not significance tests

Separation counts, $k \geq 10$ (14 budgets per model)

Comparison	Result (of 14 budgets)
Leaf macro-AUPRC vs Leaf BCE	0 ahead for every model (HELM-SL 8 behind, HELM-SSL 7, HELM_g 7, HMLC 5)
Internal-node F1 (leaf-propagate) vs MLC	Ahead: HELM-SL 10, HELM-SSL 10, Depth-weighted 9, HELM_g 4, HMLC 3 — behind: none
Internal-node F1 (leaf-propagate) vs Leaf BCE	Ahead: Depth-weighted 7, HELM-SL 6, HELM-SSL 6, HELM_g 6, HMLC 3

Provenance

294/294

$k \geq 10$ grid complete

Learning-curve runs

98/210

$k \leq 5$ grid so far

61.9 GPU-h so far

114.2

GPU-h, paper grid

Four RTX 4090s

Preview commit 2b6593f, generated 2026-09-29 11:57 CEST