
HELM on BigEarthNet: Benchmarking Hierarchical Multi-Label Land-Cover Classification

Nikola Marić^{1,2} Marjan Stoimchev¹ Dragi Kocev¹ Sašo Džeroski¹

¹Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

²Jožef Stefan International Postgraduate School, Ljubljana, Slovenia

Abstract Hierarchical multi-label classification (HMLC) of satellite images is usually studied on purpose-built hierarchies, because hierarchy-annotated archives are scarce. BigEarthNet is an exception: its 43 leaf classes are CORINE Land Cover level-3 classes and its 19-class nomenclature is published with the archive. We describe `ben-hmlc-bench`, a benchmark that puts six flat and hierarchy-aware models through one backbone, one output contract and one scoring path on 519,284 official-split patches under two hierarchies, BEN-43 and BEN-19, scored with hierarchical metrics and four hierarchy-consistency corrections. Its flagship experiment extends the four supervised variants of HELM (Stoimchev et al., 2026a) to BigEarthNet under HELM’s own protocol, with equivalence tests against HELM’s code. Inside HELM’s family only the graph term raises leaf accuracy, and HELM’s full supervised model is indistinguishable from its graph variant; both beat our flat control somewhere on micro-F1, never on AUPRC, and depth-weighting’s lead over it is a schedule difference that does not survive matching. The corrections make the flat control hierarchy-consistent without improving a model already trained on every node, and the smaller of the two hierarchies is the harder one to keep consistent. All runs are fully supervised.

1 Introduction

Land cover is described by nomenclatures that are hierarchical by design: CORINE Land Cover (CLC) arranges its classes in nested levels (Feranec et al., 2016), and a Sentinel-2 patch usually carries several classes, the setting of hierarchical multi-label classification (HMLC). A prediction that marks a patch *Coniferous forest* but not *Forests* violates the hierarchy, which is why hierarchical metrics and hierarchy-consistency corrections belong in the evaluation. HELM, for Hierarchical and Explicit Label Modeling (Stoimchev et al., 2026a), reports gains over flat multi-label baselines on eight datasets that it also equipped with hierarchies.

Two limits temper what such results say about Earth observation. First, HELM’s hierarchies were built by mapping classes onto CLC or a clinical taxonomy, with the remaining placements queried from an LLM and checked by hand (Stoimchev et al., 2026a). Second, HELM’s largest dataset, PadChest, holds 121,230 images (Stoimchev et al., 2026a). BigEarthNet (Sumbul et al., 2019) removes both limits: its 43 leaf classes already are CLC level-3 classes (BEN-43) and its 19-class nomenclature (BEN-19) is published with the archive (Sumbul et al., 2021), while its official split covers 519,284 patches, 4.3× PadChest. We use the BEN-43 and BEN-19 hierarchies assembled for HMLC on land cover in earlier work (Stoimchev et al., 2026b). Flat multi-label benchmarks on BigEarthNet are well established (Dimitrovski et al., 2023), but none of them scores hierarchy consistency.

We describe `ben-hmlc-bench`; its flagship experiment is HELM on BigEarthNet. The six compared models are our flat control (Leaf BCE), HELM’s leaf-only variant (MLC), Depth-weighted BCE over all nodes (Stoimchev et al., 2026b), and HELM’s HMLC, HELM_g and HELM_{SL} . All six share one backbone (ViT-B/16), one output contract, one scoring path, the same labeled subset and the same three seeds. Inputs are RGB, as in HELM.

Table 1: The two BigEarthNet label hierarchies, over the same images. Patches: the official v1.0 train/validation/test split. Depth is 0-indexed (roots = 0); leaf depths are listed as count@depth.

Hierarchy	Patches	Nodes	Leaves	Roots	Edges	Leaf depth
BEN-43	519,284	63	43	5	58	uniform
BEN-19	519,284	27	19	5	22	variable (10@1, 9@2)

2 Related work

Hierarchy-aware image classifiers either translate the hierarchy into the network architecture (Noor and Robles-Kelly, 2024) or embed it in the objective, through an output layer that enforces parent-child consistency (Giunchiglia and Lukasiewicz, 2020) or a contrastive objective defined over the hierarchy (Zhang et al., 2022). A third line models the label set as a graph over label co-occurrence, not the hierarchy (Chen et al., 2019). HELM is where these lines meet: it equips eight public image datasets with hierarchies built for the study, none of them at BigEarthNet’s scale (Stoimchev et al., 2026a). HELM is the point of departure for the benchmark described here.

BigEarthNet annotates Sentinel-2 patches with CORINE Land Cover level-3 classes (Sumbul et al., 2019), a nomenclature nested by construction (Feranec et al., 2016): BEN-43’s leaves are taxonomy classes, not a mapping made for an experiment. A coarser nomenclature published with the archive supplies the label set behind BEN-19 (Sumbul et al., 2021). A refined release, reBEN, provides corrected patches and new splits (Clasen et al., 2025); the label tables and official split used here are the original archive’s. Both hierarchies were assembled from these nomenclatures and released as the first HMLC benchmarks in remote sensing (Stoimchev et al., 2026b), with that work’s all-node objective, `hmlc_net`, reported ahead of a flat multi-label baseline. Hierarchical evaluation gives partial credit for near misses (hierarchical precision, recall and F1 under ancestor closure (Kiritchenko et al., 2005)), and hierarchy violations can be penalised during training (Wehrmann et al., 2018). We add HELM’s variants, put flat and hierarchy-aware models through one protocol on both hierarchies, and report hierarchy consistency.

3 Benchmark and protocol

We use BigEarthNet-S2 v1.0 (Sumbul et al., 2019), bands B04/B03/B02 resized to 224 px. The official split gives 269,695/123,723/125,866 train/validation/test patches (519,284). Its lists exclude seasonal-snow, cloud- or shadow-covered, and no-19-class-target patches; we merge duplicated Pastures and Water bodies label-table columns (Stoimchev et al., 2024b,a). BEN-43’s leaves are 43 of the CORINE Land Cover level-3 classes present in BigEarthNet (Feranec et al., 2016); BEN-19’s are the published 19-class nomenclature (Sumbul et al., 2021), grouped as in its label table (Stoimchev et al., 2024a). They differ in shape (Table 1) and exist in four formats: JSON, edge list, CLUS+ paths (Vens et al., 2008), HELM dataset config.

Six models differ in how the hierarchy enters training. Leaf BCE, our flat control, is a timm ViT-B/16 (Dosovitskiy et al., 2021) with one multi-label head, trained on the leaf columns alone. Depth-weighted trains that backbone under `hmlc_net`’s objective (Stoimchev et al., 2026b): BCE over every node, weighted $w_c = w_0^{\text{depth}}$ with $w_0 = 0.75$. It keeps `hmlc_net`’s cosine schedule where the others use HELM’s OneCycle, so it is trained under both, and every number we report for it names the schedule. The other four instantiate HELM’s own classes: MLC, its leaf-only variant, puts the same leaf loss on a per-label-token ViT whose token bank is sized to the leaves; HMLC puts unweighted BCE on every node; HELM_g adds a two-layer GraphSAGE (Hamilton et al., 2017) over the hierarchy edges to the node logits; and HELM-SL, HELM’s full supervised model, adds its BYOL branch on two views of each labeled patch. Fairness is enforced: all six share one backbone (ViT-B/16, HELM’s own checkpoint, stochastic depth off), one (B, n_{nodes}) output contract, one

Table 2: Supervised test results on the official split, raw scores: leaf AUPRC (macro) and leaf micro-F1, mean $\pm$ sample std over 3 seeds. No cell is bold: no column extremum’s 3-seed range is disjoint from every rival’s.

Model	BEN-43		BEN-19	
	AUPRC $\uparrow$	F1 $\uparrow$	AUPRC $\uparrow$	F1 $\uparrow$
Leaf BCE	0.646 $\pm$ 0.005	0.774 $\pm$ 0.001	0.807 $\pm$ 0.001	0.785 $\pm$ 0.002
MLC	0.626 $\pm$ 0.005	0.771 $\pm$ 0.001	0.793 $\pm$ 0.003	0.782 $\pm$ 0.002
Depth-weighted (cosine)	0.657 $\pm$ 0.006	0.779 $\pm$ 0.001	0.804 $\pm$ 0.006	0.788 $\pm$ 0.002
HMLC	0.632 $\pm$ 0.001	0.773 $\pm$ 0.001	0.793 $\pm$ 0.002	0.782 $\pm$ 0.001
HELM _g	0.639 $\pm$ 0.004	0.777 $\pm$ 0.002	0.804 $\pm$ 0.001	0.789 $\pm$ 0.0002
HELM-SL	0.638 $\pm$ 0.013	0.777 $\pm$ 0.001	0.802 $\pm$ 0.005	0.789 $\pm$ 0.001

scoring path, one labeled subset, one optimiser (AdamW, base LR 1e-4; gradient clipping 1.0)) and HELM’s weak augmentation (flips, blur), each pinned by tests. No model starts from a pretrained positional encoding: HELM’s ViT declares its own, so the two timm ViTs are aligned with HELM.

The supervised protocol is identical to HELM’s (Stoimchev et al., 2026a). Thirteen equivalence tests pin hierarchy machinery, label extension, head arithmetic and 16 metrics to HELM’s own functions. Headline metric is HELM’s leaf AUPRC on raw scores, macro-averaged; we report leaf micro-F1. We add hierarchical precision/recall/F1 (hP/hR/hF) under ancestor closure (Kiritchenko et al., 2005) and a per-edge violation rate, at leaf, all-node and per-level scopes; thresholded metrics use a 0.5 threshold. Every model is scored under four hierarchy-consistency corrections (raw, max-constraint (Giunchiglia and Lukasiewicz, 2020), min-constraint, leaf-propagate), reported separately. Each cell is three seeds; early stopping on validation leaf micro-F1 ended every run at 11–20 epochs (mean 14.4). At batch 64, validation is once per epoch (4,213 steps) for every model. One GPU per run; the 36-run grid and it took 114.2 GPU-hours on four RTX 4090s.

4 Results

Table 2 reports leaf AUPRC and micro-F1 on the official test split, three seeds per cell as mean $\pm$ sample std. We count one model as above another only where their three-seed ranges are disjoint, and call them comparable otherwise.

Inside HELM’s own family, which shares one schedule, one initialisation and one codebase, only the graph term buys leaf accuracy. HELM_g reaches 0.639 $\pm$ 0.004 leaf AUPRC against MLC’s 0.626 $\pm$ 0.005 on BEN-43, and 0.804 $\pm$ 0.001 against 0.793 $\pm$ 0.003 on BEN-19. Training on every node without the graph does not: HMLC is comparable to MLC (0.632 $\pm$ 0.001, 0.793 $\pm$ 0.002). Stacking HELM’s BYOL branch on top of the graph then buys nothing we can measure, since HELM-SL is indistinguishable from HELM_g in all four headline cells (0.638 $\pm$ 0.013 and 0.777 $\pm$ 0.001 on BEN-43, 0.802 $\pm$ 0.005 and 0.789 $\pm$ 0.001 on BEN-19), though it does clear HMLC in three of them. Its BEN-43 AUPRC spread, 0.013 against HELM_g’s 0.004, is the widest in the table.

That ordering does not survive contact with the flat control. None of the four HELM variants is above Leaf BCE on AUPRC on either hierarchy. On micro-F1 two of them are, HELM_g on BEN-19 and HELM-SL on both, and after leaf-propagate the same two are above Leaf BCE on BEN-19 hF (Figure 1). Which model comes out ahead therefore depends on the metric being read.

Depth-weighted alone keeps its upstream cosine schedule where the others use OneCycle, so its row contrasts two published recipes. Re-trained on OneCycle over the same three seeds, its BEN-43 leaf AUPRC drops from 0.657 $\pm$ 0.006 to 0.642 $\pm$ 0.008. Matched that way it beats Leaf BCE nowhere: comparable on BEN-43, below on both BEN-19 metrics (0.799 $\pm$ 0.006 against 0.807 $\pm$ 0.001 AUPRC, 0.782 $\pm$ 0.0002 against 0.785 $\pm$ 0.002 micro-F1).

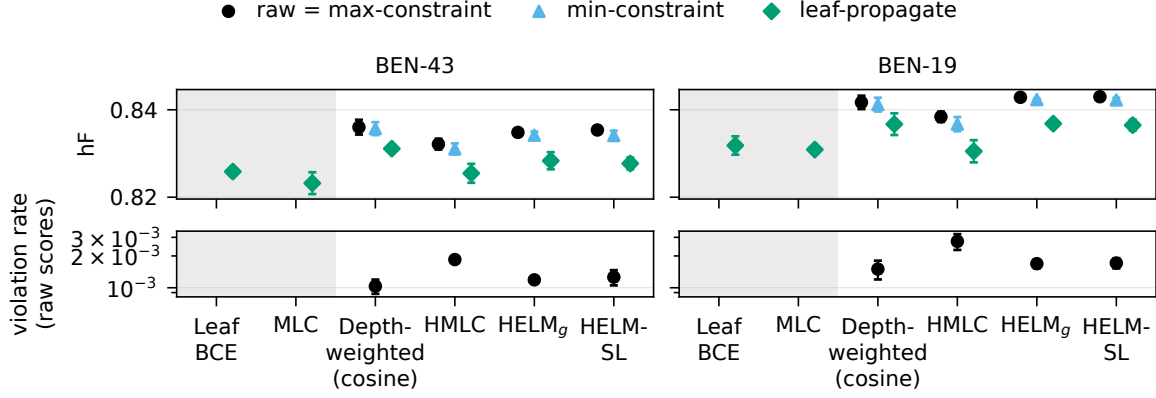


Figure 1: Hierarchy consistency, test split. On hF, max-constraint equals raw by construction. Shaded: Leaf BCE, MLC, scored after leaf-propagate only. Bars: sample std, $n = 3$.

A model already trained on every node leaves the corrections little to repair (Figure 1). Raw per-edge violation rates for the all-node models run from 0.0010 ± 0.0002 to 0.0027 ± 0.0005 , and the largest of the four lowers rather than raises what it touches: leaf-propagate takes HELM_g’s BEN-43 hF from 0.835 ± 0.001 to 0.828 ± 0.002 , while every correction but raw zeroes the violation rate by construction. The corrections earn their place on the flat models instead. Leaf BCE and MLC train nothing above the leaves, so leaf-propagate is the only one that gives them hierarchical scores; scored that way, Leaf BCE reaches 0.826 ± 0.001 on BEN-43, comparable to HELM_g and above MLC’s 0.823 ± 0.002 , and 0.832 ± 0.002 on BEN-19, below HELM_g’s 0.837 ± 0.001 and HELM-SL’s 0.836 ± 0.001 . The corrections make a flat model usable on a hierarchy; they do not improve one already trained on it.

The smaller of the two taxonomies is then the harder one to keep consistent, on the same images. The all-node models violate a larger fraction of parent–child pairs on BEN-19 than on BEN-43 – HELM_g, 0.0017 ± 0.0002 against 0.0012 ± 0.00001 – although BEN-19 has 22 edges to BEN-43’s 58. Fewer edges are therefore not easier edges, and one hierarchy alone would have told half of this.

5 Conclusion and further work

ben-hmlc-bench puts six flat and hierarchy-aware models through one protocol on two BigEarthNet hierarchies, scored with hierarchical metrics and four consistency corrections. The corrections do their work on the flat model, making it usable on a hierarchy without improving one already trained on it. The hierarchy itself does not pay: inside HELM’s family only the graph term buys leaf accuracy, no HELM variant clears the flat control on AUPRC though two do on micro-F1, and depth-weighting’s apparent lead is a schedule difference that does not survive matching. The archive is the likely reason: half a million labelled patches leave little for a hierarchy to add, and like multi-label learning over binary relevance, or semi-supervision, it earns its keep when labels are scarce and overfitting binds. Further work is a learning curve testing this, at n times the number of labels for n of 10, 20, 50, 100, 200, 500 and 1000, all labels equally represented, on the same splits.

Acknowledgements. N.M. acknowledges the support of the Slovenian Research and Innovation Agency (ARIS) through the Young Researcher programme (grant no. PR-13689). This work was further supported by ARIS through the projects Geospatial Information Technologies for Resilient and Sustainable Society, GeoAI (GC-0006), and Artificial Intelligence for Science, AI4Sci (GC-0001).

References

- Chen, Z.-M., Wei, X.-S., Wang, P., and Guo, Y. (2019). Multi-label image recognition with graph convolutional networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5172–5181. IEEE.
- Clasen, K. N., Hackel, L., Burgert, T., Sumbul, G., Demir, B., and Markl, V. (2025). reBEN: Refined BigEarthNet dataset for remote sensing image analysis. In *IGARSS 2025 – 2025 IEEE International Geoscience and Remote Sensing Symposium*, pages 1264–1268. IEEE.
- Dimitrovski, I., Kitanovski, I., Koccev, D., and Simidjievski, N. (2023). Current trends in deep learning for Earth Observation: An open-source benchmark arena for image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197:18–35.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.
- Feranec, J., Hazeu, G., Kosztra, B., and Arnold, S. (2016). CORINE land cover nomenclature. In *European Landscape Dynamics*, pages 15–26. CRC Press, Boca Raton, FL.
- Giunchiglia, E. and Lukasiewicz, T. (2020). Coherent hierarchical multi-label classification networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9662–9673. Curran Associates, Inc.
- Hamilton, W. L., Ying, R., and Leskovec, J. (2017). Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, pages 1024–1034. Curran Associates, Inc.
- Kiritchenko, S., Matwin, S., and Famili, A. F. (2005). Functional annotation of genes using hierarchical text categorization. In *BioLINK SIG: Linking Literature, Information and Knowledge for Biology, a joint meeting of the ISMB BioLINK Special Interest Group on Text Data Mining and the ACL Workshop on Linking Biological Literature, Ontologies and Databases*, Detroit, Michigan, USA.
- Noor, K. T. and Robles-Kelly, A. (2024). H-CapsNet: A capsule network for hierarchical image classification. *Pattern Recognition*, 147:110135.
- Stoimchev, M., Koloski, B., Levatić, J., Koccev, D., and Džeroski, S. (2026a). Fully- and semi-supervised hierarchical multi-label image classification with graph learning. *Information Sciences*, 754:123641.
- Stoimchev, M., Levatić, J., Koccev, D., and Džeroski, S. (2024a). BigEarthNet-19-HMLC. Hugging Face dataset, <https://huggingface.co/datasets/marjandl/BigEarthNet-19-HMLC>. Accessed 14 September 2026.
- Stoimchev, M., Levatić, J., Koccev, D., and Džeroski, S. (2024b). BigEarthNet-43-HMLC. Hugging Face dataset, <https://huggingface.co/datasets/marjandl/BigEarthNet-43-HMLC>. Accessed 14 September 2026.
- Stoimchev, M., Levatić, J., Koccev, D., and Džeroski, S. (2026b). Hierarchical multi-label classification of land use/land cover in remote sensing images. In Cerrato, M., Kalinauskaitė, D., Lukoševičius, M., Pechenizkiy, M., and Štutienė, K., editors, *Machine Learning and Principles and Practice of Knowledge Discovery in Databases. ECML PKDD 2024*, volume 2560 of *Communications in Computer and Information Science*, pages 169–179, Cham. Springer Nature Switzerland.

- Sumbul, G., Charfuelan, M., Demir, B., and Markl, V. (2019). BigEarthNet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019 – 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904. IEEE.
- Sumbul, G., de Wall, A., Kreuziger, T., Marcelino, F., Costa, H., Benevides, P., Caetano, M., Demir, B., and Markl, V. (2021). BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval. *IEEE Geoscience and Remote Sensing Magazine*, 9(3):174–180.
- Vens, C., Struyf, J., Schietgat, L., Džeroski, S., and Blockeel, H. (2008). Decision trees for hierarchical multi-label classification. *Machine Learning*, 73(2):185–214.
- Wehrmann, J., Cerri, R., and Barros, R. (2018). Hierarchical multi-label classification networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5075–5084. PMLR.
- Zhang, S., Xu, R., Xiong, C., and Ramaiah, C. (2022). Use all the labels: A hierarchical multi-label contrastive learning framework. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16639–16648. IEEE.

Submission Checklist

1. For all authors...

- (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes] See the abstract and Section 1. The three contributions map to claims recorded in the project’s claims ledger; the results sentence reports where our flat control is beaten and where it is not, rather than generalising either.
- (b) Did you describe the limitations of your work? [Yes] Section 3 lists the protocol decisions where we depart from HELM and names the two where HELM’s released code contradicts its paper; Section 4 states that one model family carries a different learning-rate schedule, that the leaf-only models have no internal-node scores and are reported under leaf-propagation only, and that separations are claimed only where 3-seed ranges are disjoint. Section 5 names what is not done.
- (c) Did you discuss any potential negative societal impacts of your work? [No] Land-cover classification from public Sentinel-2 imagery at 10 m ground sampling distance; we identified no negative societal impact specific to this work and did not manufacture a discussion.
- (d) Did you read the ethics review guidelines and ensure that your paper conforms to them? (see <https://2022.automl.cc/ethics-accessibility/>) [Yes]

2. If you ran experiments...

- (a) Did you use the same evaluation protocol for all methods being compared (e.g., same benchmarks, data (sub)sets, available resources, etc.)? [Yes] One backbone, one output contract, one scoring path, one labeled subset, one set of seeds, and identical data splits for every compared model; the deliberate exceptions are enumerated in Section 3 and in the protocol document.
- (b) Did you specify all the necessary details of your evaluation (e.g., data splits, pre-processing, search spaces, hyperparameter tuning details and results, etc.)? [Yes] Section 3 and Table 1: official v1.0 splits and their sizes, both label hierarchies, input bands and resolution, optimiser (AdamW, base LR 1e-4, gradient clipping 1.0), batch size, early-stopping criterion, patience and the resulting epoch range, decision threshold, weak augmentation, and the four hierarchy-consistency corrections. No hyperparameter search was run; values follow the two upstream papers except where Section 3 names a numbered decision. The epoch cap, the learning-rate schedule and the remaining optimiser literals are not in the paper; the repository is not public at submission time, so we state that rather than pointing to a document a reader cannot obtain.
- (c) Did you repeat your experiments (e.g., across multiple random seeds or splits) to account for the impact of randomness in your methods or data? [Yes] Every cell is three seeds (0, 1, 42).
- (d) Did you report the uncertainty of your results (e.g., the standard error across random seeds or splits)? [Yes] Mean $\pm$ sample standard deviation over the three seeds, in every table and figure.
- (e) Did you report the statistical significance of your results? [No] Three seeds cannot support a significance test, and no significance-testing code exists in the benchmark. We instead report the sample standard deviation and claim a separation only where the 3-seed ranges of the two cells are disjoint and the paired per-seed difference is sign-consistent. We state this convention rather than implying more.

- (f) Did you use enough repetitions, datasets, and/or benchmarks to support your claims? **[Yes]** For the claims actually made. Every comparative statement in Section 4 is restricted to pairs whose 3-seed ranges are disjoint; where they overlap, the text says the models are indistinguishable rather than ranking them.
 - (g) Did you compare performance over time and describe how you selected the maximum runtime? **[Yes]** Every run is capped at 100 epochs and stopped early on validation leaf micro-F1 with patience 5, so the runtime is set by the stopping rule rather than chosen; no run reached the cap.
 - (h) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **[Yes]** Section 3 reports the GPU type and the measured total over every run the paper reports: the 36-run grid and the six-run schedule control, 114.2 GPU-hours on four RTX 4090s.
 - (i) Did you run ablation studies to assess the impact of different components of your approach? **[Yes]** The benchmark is built around them: two hierarchies of different shape on the same images, six model configurations spanning flat and hierarchy-aware training, and four hierarchy-consistency corrections applied to every model through one scoring path.
3. With respect to the code used to obtain your results...
- (a) Did you include the code, data, and instructions needed to reproduce the main experimental results, including all dependencies (e.g., `requirements.txt` with explicit versions), random seeds, an instructive README with installation instructions, and execution commands (either in the supplemental material or as a URL)? **[No]** The repository is not public at the time of submission, so we make no availability claim in the paper. It carries an MIT licence and we intend to release it.
 - (b) Did you include a minimal example to replicate results on a small subset of the experiments or on toy data? **[No]** As above.
 - (c) Did you ensure sufficient code quality and documentation so that someone else can execute and understand your code? **[No]** As above. The code is documented and tested internally, but nothing is available to a reader at submission time and we do not claim otherwise.
 - (d) Did you include the raw results of running your experiments with the given code, data, and instructions? **[No]** As above.
 - (e) Did you include the code, additional data, and instructions needed to generate the figures and tables in your paper based on the raw results? **[No]** As above. Every table and figure in this paper is generated by script from the run outputs and is never hand-edited, but those scripts are not yet available to a reader.
4. If you used existing assets (e.g., code, data, models)...
- (a) Did you cite the creators of used assets? **[Yes]** BigEarthNet, the CORINE Land Cover nomenclature, the published 19-class nomenclature, HELM, hmlc_net, and the backbone are all cited where used.
 - (b) Did you discuss whether and how consent was obtained from people whose data you're using/curating if the license requires it? **[N/A]** The data is public satellite imagery; no personal data is involved.
 - (c) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **[N/A]** BigEarthNet contains Sentinel-2 satellite imagery at

10 m ground sampling distance, in 1.2 km patches, and no personally identifiable information or offensive content.

5. If you created/released new assets (e.g., code, data, models)...
 - (a) Did you mention the license of the new assets (e.g., as part of your code submission)? [\[Yes\]](#)
The benchmark carries an MIT licence in the repository. One of the two upstream codebases we build on states no licence, which we note rather than assume.
 - (b) Did you include the new assets either in the supplemental material or as a URL (to, e.g., GitHub or Hugging Face)? [\[No\]](#) The repository is not public at the time of submission.
6. If you used crowdsourcing or conducted research with human subjects...
 - (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [\[N/A\]](#)
 - (b) Did you describe any potential participant risks, with links to institutional review board (IRB) approvals, if applicable? [\[N/A\]](#)
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [\[N/A\]](#)
7. If you included theoretical results...
 - (a) Did you state the full set of assumptions of all theoretical results? [\[N/A\]](#) The paper contains no theoretical results.
 - (b) Did you include complete proofs of all theoretical results? [\[N/A\]](#) The paper contains no theoretical results.