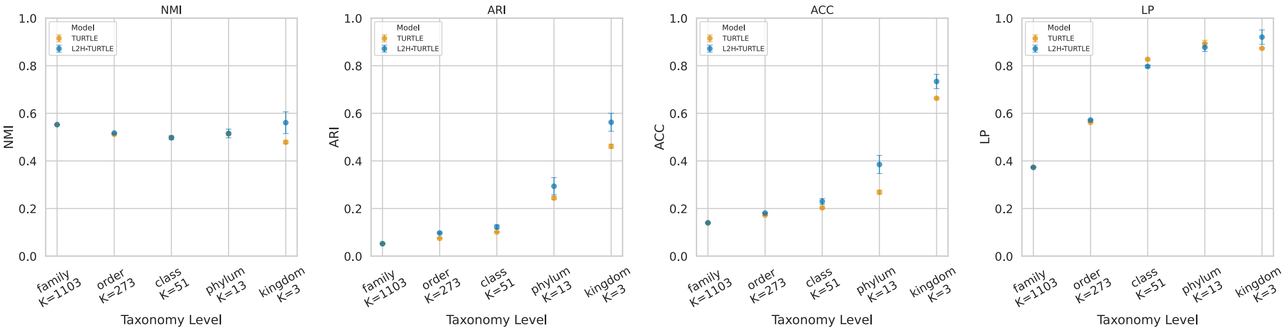

Model	Family				Order				Class				Phylum				Kingdom			
	nmi	ari	acc	lp	nmi	ari	acc	lp	nmi	ari	acc	lp	nmi	ari	acc	lp	nmi	ari	acc	lp
TURTLE	**0.552**	**0.052**	**0.140**	**0.373**	0.512	0.075	0.172	0.562	**0.498**	0.101	0.203	**0.827**	**0.514**	0.244	0.268	**0.893**	0.479	0.461	0.663	0.873
L2H-TURTLE	**0.552**	**0.052**	**0.140**	**0.373**	**0.517**	**0.097**	**0.181**	**0.572**	**0.497**	**0.123**	**0.229**	0.797	**0.515**	**0.294**	**0.385**	0.877	**0.561**	**0.562**	**0.734**	**0.920**

Figure: Experiment on the INaturalist21 dataset to test whether, with hiearchical datasets, our proposed approach can bring an advantage over flat clustering with the backbone model. To do so, we train TURTLE to model clusters at the *family* taxonomy level ($K_{family} = 1103$). Then we implement our L2H procedure on top, and use the produced hierarchy to make clustering predictions at more coarse taxonomy levels. For instance, K_{order} steps from the end of the procedure (see Algorithm 1), we have a K_{order}-clustering of the data points, and we test its performance on the test set (at the *order* taxonomy level). Then, we train instances of TURTLE at each coarse taxonomy level i.e. $K \in \{273, 51, 13, 3\}$ and report the corresponding test set performance for comparison. Notably training TURTLE at the fine-grained level, and using our L2H approach to construct a hierarchy and make predictions at more coarse levels, achieves better clustering performance that training separate TURTLE models at each taxonomy level. Note that this performance improvement comes also with much lower compute time, as only a single instance of the TURTLE model is trained (at the finest-grained taxonomy level). We report the results with means and standard deviations in the plots above, but also report mean values in the table below for better readability, bolding best results and results that are statistically indistinguishable from the best.