**Mikko Impiö***    **Jenni Raitoharju***[†]

\* Finnish Environment Institute Syke, Quality of Information
[†] University of Jyväskylä, Faculty of Information Technology

Sample

Images from **inlier** and **outlier** classes

Outlier images are **unknown**

Unique DNA barcodes
$c_1$ GTAAGCG
$c_2$ AGCTCAG
$c_3$ CGATTAAC

**Outlier** classes are **known**

CNN

OOD detection methods used to order samples based on score $m$

$[p_1, p_2, p_3]$ $m_1$ $c_2$
$[p_1, p_2, p_3]$ $m_2$ $c_2$
$[p_1, p_2, p_3]$ $m_3$ $c_1$

Calculate DNA distances and rank classes based on it

$d_{32} < d_{31}$

Highest score class $c$ is chosen as candidate prediction and compared to the true outlier

**Proposed re-ordering methods**

**DNA ordering**
Order first by most proximate DNA class $c$, then by $m$

**DNA quantile ordering**
Order first by top and bottom blocks, then by DNA inside the blocks

ROC and PR curves are calculated

## Takeaways

- DNA-based re-ordering improved OOD detection compared to all baselines.

- The results demonstrate potential for the use of DNA data as supplementary information for computer vision systems.

- Taxa with high DNA barcode similarity are often confused for each other by a neural network.

# Improving Taxonomic Image-based Out-of-distribution Detection With DNA Barcodes

Outlier detection for biological samples can be improved if there is access to a bulk DNA sequencing that reveals outlier classes.

Outlier class images can be found by re-ordering outlier candidates by their DNA distance to the known outlier DNA sequence.

## Background

- Reliable out-of-distribution (OOD) detection if important for image-based species identification system.

## Example use case

- Samples are imaged and identified using a classifier with a fixed class distribution.
- **DNA metabarcoding** (no individual assignment) is done to the bulk sample.
- The DNA data **reveals an outlier class.**
- We want to find the images that represent this class.

## Method

- **Images are classified** using a neural network.
- Standard OOD scoring methods like Maximum Sigmoid Probability [1], MaxLogit [2] and Energy score [3] can be used for **initial ranking** of outlier candidates.
- DNA barcode proximity to outlier class is used for **re-ranking** candidates.
- **Quantile method:** Re-ranking is done separately for the first $q$ and last $1 - q$ samples.

## Data

- FinBenthic2 dataset: 460 009 images of 39 taxa [4]
- Prototypical DNA barcodes collected from BOLD database

## DNA distances

- We assume each taxa class has a representative DNA sequence that can be used for distance calculations.
- Kimura 2-parameter distance (K80) [5] is used.

## Results

- The proposed quantile re-ordering method improves OOD detection compared to all baselines.
- The method works the best with genus-level outliers.
- Taxa with close DNA proximity are often visually similar and mistaken for each other.

| OOD scoring | Re-ordering method | AUPRC ↑ | AUROC ↑ | FPR@95 ↓ |
|---|---|---|---|---|
| Entropy | OOD baseline | 0.317 (±0.22) | 0.808 (±0.10) | 0.501 (±0.17) |
| | DNA | 0.444 (±0.26) | 0.806 (±0.18) | 0.467 (±0.30) |
| | DNA quantile | **0.469 (±0.24)** | **0.864 (±0.06)** | **0.390 (±0.13)** |
| MSP | OOD baseline | 0.302 (±0.21) | 0.804 (±0.09) | 0.503 (±0.17) |
| | DNA | 0.443 (±0.26) | 0.806 (±0.18) | 0.467 (±0.30) |
| | DNA quantile | **0.468 (±0.24)** | **0.863 (±0.06)** | **0.395 (±0.14)** |
| MaxLogit | OOD baseline | 0.319 (±0.24) | 0.801 (±0.13) | 0.493 (±0.24) |
| | DNA | 0.447 (±0.26) | 0.806 (±0.18) | 0.468 (±0.30) |
| | DNA quantile | **0.455 (±0.22)** | **0.856 (±0.06)** | **0.337 (±0.08)** |
| Energy | OOD baseline | 0.317 (±0.23) | 0.800 (±0.13) | 0.493 (±0.24) |
| | DNA | 0.447 (±0.26) | 0.806 (±0.18) | 0.468 (±0.30) |
| | DNA quantile | **0.454 (±0.22)** | **0.855 (±0.06)** | **0.335 (±0.09)** |
| Ratio (logit) | OOD baseline | 0.253 (±0.17) | 0.762 (±0.10) | 0.605 (±0.13) |
| | DNA | 0.435 (±0.26) | 0.804 (±0.18) | 0.469 (±0.30) |
| | DNA quantile | **0.436 (±0.24)** | **0.833 (±0.08)** | 0.490 (±0.17) |
| Ratio (softmax) | OOD baseline | 0.282 (±0.19) | 0.798 (±0.09) | 0.511 (±0.16) |
| | DNA | 0.442 (±0.26) | 0.806 (±0.18) | 0.467 (±0.30) |
| | DNA quantile | **0.464 (±0.24)** | **0.861 (±0.07)** | 0.408 (±0.14) |

Re-ordering methods compared to OOD baseline ranking for different OOD scoring methods. Values are averages over 39 models trained for 39 different outlier classes.



$r=-0.38, p=2.62e-30$
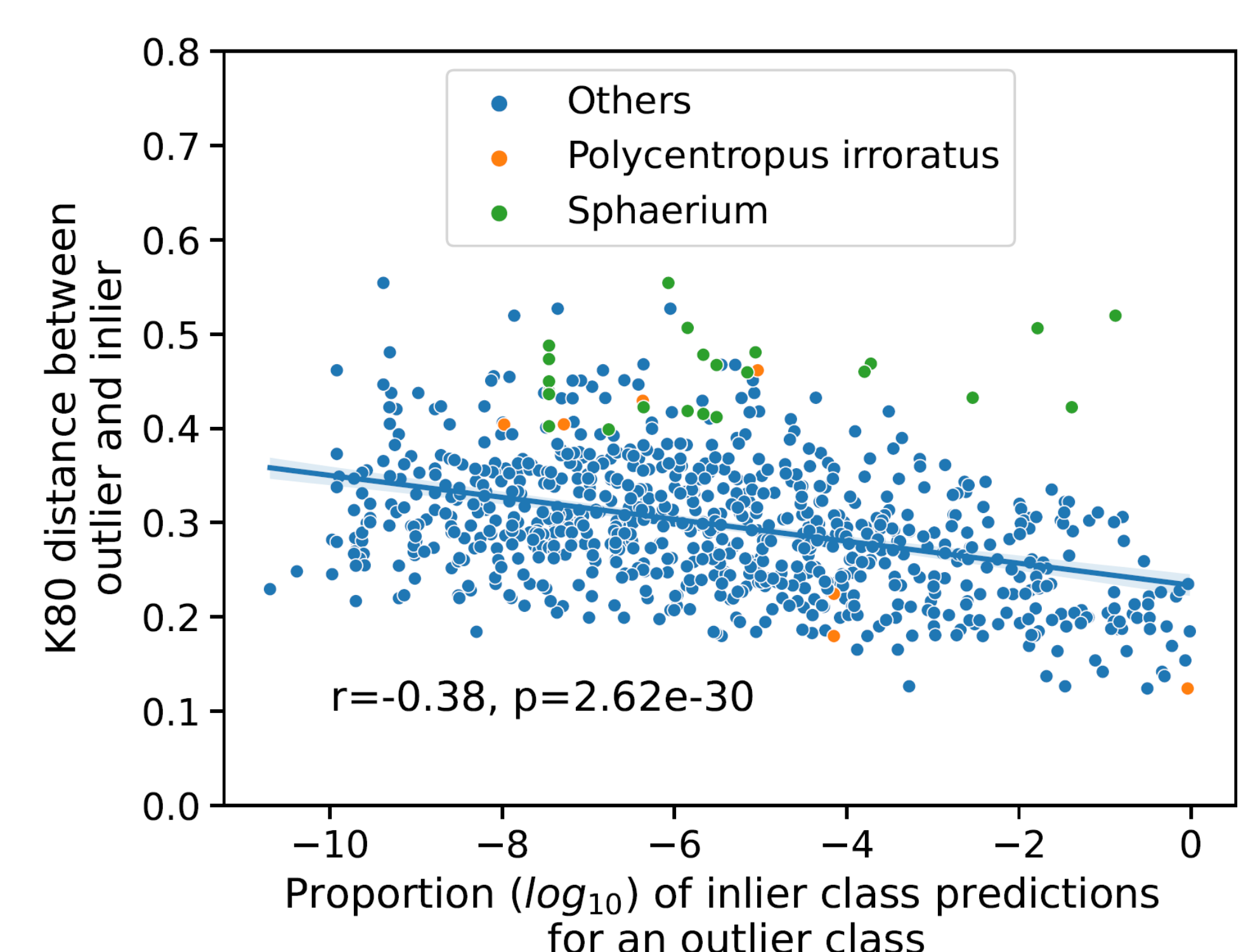
Comparison of the proportion of (false) inlier predictions and DNA-based distance between classes for outlier-inlier pairs. Outliers where re-ordering works the best (Polycentropus irroratus) and worst (Sphaerium) are highlighted.

[1] D. Hendrycks and K. Gimpel, "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks," in International Conference on Learning Representations, 2017

[2] D. Hendrycks, et al., "Scaling Out-of-Distribution Detection for Real- World Settings," in International Conference on Machine Learning, 2022.

[3] W. Liu, X.Wang, J. Owens, and Y. Li, "Energy-based Out-of-distribution Detection," in Advances in Neural Information Processing Systems, vol. 33, 2020.

[4] J. Ärje et al., "Human experts vs. machines in taxa recognition," Signal Processing: Image Communication, vol. 87, 2020.

[5] M. Kimura, "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences," Journal of Molecular Evolution, vol. 16, no. 2, 1980.

**https://github.com/mikkoim/dnaimg-ood**