# Rove-Tree-11: The not-so-Wild Rover A hierarchically structured image dataset for deep metric learning research

Roberta Hunt[1] and Kim Steenstrup Pedersen[1,2]

[1] Department of Computer Science,
University of Copenhagen, Universitetsparken 1, 2100, Copenhagen, Denmark
{r.hunt,kimstp}@di.ku.dk
[2] Natural History Museum of Denmark, Øster Voldgade 5 - 7, 1350, Copenhagen,
Denmark kimstp@snm.ku.dk

**Abstract.** We present a new dataset of images of pinned insects from museum collections along with a ground truth phylogeny (a graph representing the relative evolutionary distance between species). The images include segmentations, and can be used for clustering and deep hierarchical metric learning. As far as we know, this is the first dataset released specifically for generating phylogenetic trees. We provide several benchmarks for deep metric learning using a selection of state-of-the-art methods.

**Keywords:** Phylogeny · Dataset · Tree · Hierarchy · Hierarchical Dataset · Rove · Staphylinidae · Phylogenetic Tree

## 1 Introduction

A phylogeny is a fundamental knowledge frame which hypothesizes how different species relate to each other [11]. A fully annotated phylogeny, i.e. a tree of life anchored in time scale, placed in the geographic context, and with a multitude of organismal traits mapped along the tree branches is an important tool in biology. It explains biodiversity changes over millennia or geological epochs, traces organismal movements in space and evolution of their properties, models populations response to climate change, navigates new species discovery and advises classification and taxonomy. An example phylogeny from our dataset is shown in fig. 1 along with some example images from the most abundant species in the dataset.

Traditionally biologists generate phylogenies [9,10] using genetic data or morphological features (relating to the shape or development of the organism, for example the head shape, or the pattern of the veins on the wings). Despite genetic data dominating phylogenetic research in recent years, morphological features extracted by visual inspection of specimens are still of use. Fossils, for example, contain no genetic data, but morphological features on the fossils can be used to relate them to existing biodiversity [26]. Occasionally morphological and genetic

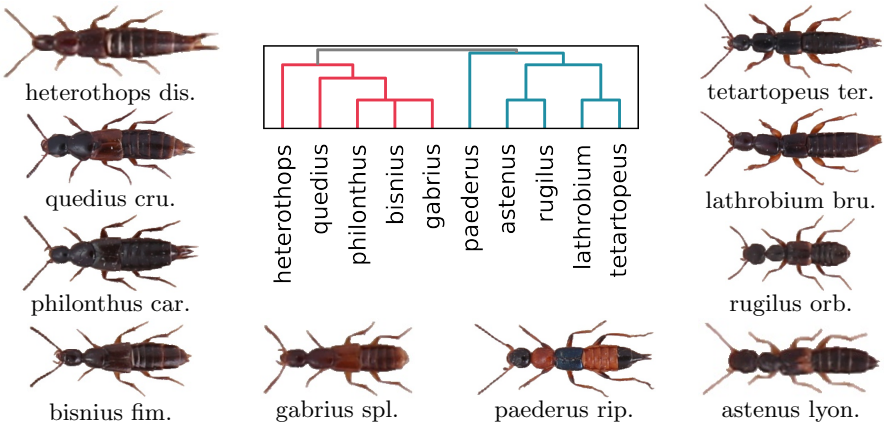


**Fig. 1.** Subset of phylogeny from the Rove-Tree-11 dataset, for the 10 genera with the most images in the dataset. Each leaf represents a genus. Genera which are closer together on the tree are more closely related, and nodes in the tree represent common ancestors. Nodes with more than two branches are considered not yet fully resolved. Many phylogenetic trees include estimations of time representing when the speciation event occurred (when the common ancestor split into two species). These dates are usually based on fossil evidence. This dated information is unfortunately not currently available for our ground truth tree. Example specimens from each genera are shown for reference.

data are even combined to generate a so called 'total-evidence' phylogeny [34]. Morphological features are also of importance for species/specimens which lack good quality genetic data. Much of phylogenetic research on insects is done from museum specimens captured many years ago. Often the DNA of such specimens has degraded and is no longer of use. Genetic extraction is also expensive, time consuming, and a destructive process which can require completely destroying the specimen, particularly in the case of small insects.

However, the traditional process of generating morphological features is slow, meticulous and introduces some aspects of subjectivity by the researcher performing the analysis. Typically a phylogenetic researcher would generate a matrix of discrete traits (although the use of continuous traits has recently been explored [35]) which they hypothesize are of use in distinguishing the species and are evolutionary important. With thousands of new species of insects discovered each year [1], it is difficult for phylogeneticists to keep up.

**Deep metric learning** [38,22] is a proven technique to generate informative embedding matrices from images, and we posit that it can be used to generate morphological embeddings which more objectively represent the morphological features of a specimen. In this dataset we are unfortunately only looking at one view of the insect, in our case, the dorsal view (the back), whereas biologists would ideally examine and compare all external and internal features of the insect. However, we hypothesize that this can be offset by the model's ability

to learn minute details. Our intention is that these methods could eventually be improved and used as a tool for biologists to inform their decision making process. Additionally, many natural history museums worldwide [8,20] are digitizing their collections, including in many cases, taking images of millions of museum insects. The Natural History Museum of Denmark (NHMD) alone estimates they have over 3.5 million pinned and dried insect specimens spanning 100,000 described species [32] and is in the process of digitizing their collections [31]. The importance of such digitization efforts have been studied from a biology research perspective [17,36]. Thus, given the increased data availability, we predict that phylogenetic generation from images will become a growing field of research within computer vision and related areas of artificial intelligence.

Despite the rapidly growing availability of images of pinned insects from natural history museums, the ongoing push from the biological community to generate phylogenies, and the increasing ability of deep learning to learn complex shapes and relationships, few publicly available datasets exist targeting the generation of phylogenies from images using deep learning techniques. There are several reasons for this, as we will explore in more detail in sec. 2.1, when we compare with existing datasets. In brief, although the number of image analysis datasets is steadily growing, often the graphs which are included in the datasets are subjectively resolved (such as [5]) or the groupings they provide are too coarse-grained (such as [12]) or, particularly for biological datasets, the images are natural photos taken in the wild, meaning they are from various viewpoints and often obscured (such as [40],[41]). This makes it difficult for the model to learn which distinct morphological features are more related to those from others species. Typical morphology based phylogenies are generated from careful inspection and comparison of features, meaning we expect direct comparison to be very important for this task.

In this paper we present 'Rove-Tree-11', a dataset of 13,887 segmented dorsal images of rove beetles along with a ground truth phylogeny down to genus level[‡]. The species-level phylogeny is not included, because this level of information is not yet readily available. Our intention with releasing this data is that it can further research on deep hierarchical metric learning and computer vision solutions for building morphological phylogenies on interesting biological groups, leveraging the current digitization-wave that is gripping natural history museums worldwide.

The **contributions** of this paper are:

1. The release of a new hierarchically structured image dataset including segmentations and ground truth genus-level phylogeny
2. We provide baseline results on this dataset for the tasks of classification, clustering, and for predicting phylogenetic trees.

[‡] to genus-level means that each species within a genus is considered unresolved, or equally likely to be related to any other species within that genus.

## 2 Related Work

### 2.1 Comparison with Existing Datasets

Hierarchically structured data is often found in computer vision related tasks. Examples include cognitive synonym relations between object categories such as clothing items[27] and is especially found in tasks concerning nature. However, current datasets which present a ground truth hierarchical grouping of the data are not intended for morphological phylogenetic research, and therefore poorly suited to the task.

There are several natural history related image datasets which do, or could easily be adapted to, include a taxonomy (ie IP102 [44], CUB-200-2011 [41], iNaturalist [40], Mammal dataset [12], PlantCLEF 2021 [13] and ImageNet [6]). With the exception of PlantCLEF, these are however all 'in the wild' images and identification has typically been done by non-experts with the naked eye. The phylogenies are also usually superficial - including only a few levels, and typically based only on the current taxonomy, which is not fine-grained and not necessarily representative of the state of the art phylogenetic tree, as taxonomies have a longer review process[§]. In the case of PlantCLEF the majority of the training images are of herbaria sheets, and therefore not 'in-the-wild', however only a shallow taxonomy is provided with the PlantCLEF dataset. In the case of IP102, the hierarchical tree is grouped by the plant the insect parasitizes, and is not related to ancestral traits at all. With the exception of CUB-200-2011, iNaturalist and PlantCLEF, the species are also easily identified by a layman/amateur by the images alone, which is not necessarily the case in our dataset, where many of the identifications traditionally require a microscope or dissection. It is also often the case that the taxonomy is not properly updated until years after the phylogeny has been altered, particularly in the case of entomology where new species are discovered regularly, so using the most recent taxonomy may not actually represent the state-of-the-art knowledge of the evolution of the species. In the case of iNaturalist, the dataset does include a tree with the same number of levels as Rove-Tree-11, however, this depth begins from kingdom-level, whereas ours begins from family level (four taxonomic ranks lower on the taxonomic hierarchy), and represents the most recent phylogeny.

Additionally there are non-biological hierarchical datasets, such as DeepFashion [27], for which others have created their own hierarchy [5]. This hierarchy is however based on loose groupings of clothing items which are highly subjective. For example, the top-level groupings are: top, bottom, onepiece, outer and special, where special includes fashion items such as kaftan, robe, and onesie, which might morphologically be more related to coats, which are in the 'outer' cate-

[§] the taxonomy represents how the organism is classified - ie which class, order, family the organism belongs to, and is a non binary tree. The phylogeny represents how related different species are together, and would ideally be a binary tree. In an ideal world the taxonomy would be a congruent to the phylogeny, but in reality they tend to diverge as taxonomic revisions take longer

**Table 1.** Comparison of dataset properties. The table indicates number of images and categories, tree depth and whether or not the images are 'in the wild'. Tree depth is calculated as the maximum number of levels in the tree. For example, with iNaturalist this is 7 (corresponding to: kingdom, phylum, class, order, family, genus and species)

| Dataset | No. Images | No. Cat. | Tree Depth | Wild? | year |
|---|---|---|---|---|---|
| Rove-Tree-11 | 13,887 | 215 | 11 | No | 2022 |
| ImageNet [6] | 14,197,122 | 21,841 | 2 | yes | 2018 |
| IP102 [44] | 75,222 | 102 | 4 | yes | 2019 |
| CUB-200-2011 [41] | 11,788 | 200 | 4 [2] | yes | 2011 |
| Cars196 [24] | 16,185 | 196 | 1 | yes | 2013 |
| INaturalist 2021 [40] | 857,877 | 5,089 | 7 | yes | 2021 |
| PlantCLEF 2021 [13] | 330,772 | 997 | 3 | mixed | 2021 |
| DeepFashion [27] | 800,000 | 50 | 4 ¶ | yes | 2016 |

gory. This kind of subjective hierarchy can be useful in other applications, but not particularly for research on generating relationships based on morphology.

The Rove-Tree-11 dataset on the other hand is a well-curated museum collection, where the identification has been done by experts, often using a microscope, and the ground truth phylogeny is as up to date as possible. Additionally, because the images are of museum collections and not 'in-the-wild', the specimen is always fully visible, and the dataset has been curated to include only whole dorsal images. Whether dorsal-view images are sufficient to generate a phylogeny remains to be seen. Typically biologists would use features from all over the body, including ventral and sometimes internal organs. We hypothesize that dorsal view may be sufficient given the ability of deep learning models to learn patterns which are difficult for the human eye to distinguish. Additionally results from our classification experiments shown in table 3 suggest that distinguishing features can be learnt from the images, supporting our belief that phylogenies may be learnt from this dataset.

### 2.2 Related Methodologies

**Classification** Classification is one of the most developed fields in computer vision and deep learning, with numerous new state of the art architectures and methods discovered each year. However, there are some architectures which have gained widespread usage in recent years, which we will use to give baselines for this dataset. In particular, we will compare classification results using ResNet [16] and EfficientNet B0 [39]. ResNet is a series of models, introduced in 2015, which uses residual convolution blocks. EfficientNet was introduced in 2019 and is known for achieving high accuracies with few parameters. Classification is not the main focus of this dataset, but we provide classification results for comparison with similar datasets.

¶ hierarchy presented in [5]

**Fig. 2.** Example image of museum unit tray from Stage 1 of image processing.

**Deep Metric Learning** The goal of deep metric learning (DML) is to learn an embedding of the data which represents the dataset and distances between datapoints meaningfully. This could be through clustering related data together, or through creating independence and interpretability in the variables. Recent research into deep metric learning can be split into three groups [38]. **Ranking-based** methods attempt to pull instances from the same class (positive examples) closer together in the embedding space, and typically push examples from other classes further away (eg, [15] [43]). **Classification-based methods**, such as ArcFace[7], work by modifying the discriminative classification task. Finally **Proxy Based** methods, such as Proxy NCA [29] compare each sample with a learned distribution for each class.

In this paper we demonstrate results for this dataset using seven deep metric learning methods; Five ranking-based losses: margin loss [43], triplet loss [43], contrastive loss [15], multisimilarity loss [42], lifted loss [45], one classification-based loss: arcface loss [7] and one proxy-based loss: proxynca [29]. With many state of the art methods and variations on these, choosing which to use is difficult. We chose these firstly because they are all used in [38] as benchmarks, making our results directly comparable. Of the 23 described in [38], we focus on seven which represented some of the better results and show a variety of methods. For a detailed description of each loss we refer the reader to [38].

During training DML models are typically evaluated not just on the loss, but also on a number of clustering metrics. In our case, to do this the dataset is evalu-

ated using nearest neighbors Recall@1 (R1) and Normalized Mutual Information (NMI) after clustering using the k-means algorithm [28]. NMI is presented in our main results, and R1 in the supplemental material. NMI is a symmetric quantity measuring the overlap between clusters. A NMI of 1 indicates that the clusters are the same. Recall@1 is a measure of the % of results with a nearest neighbour in the same class. Both are described in further detail in [38].

**Generating a Phylogeny from Embeddings** In order to use this dataset for deep phylogenetic generation, we need methods to generate binary graphs from embedding spaces. We could treat this as a classification problem, however, with only one graph to generate, this dataset is not large enough to perform direct graph generation. Instead, the graph can be generated indirectly from the embedding space and compared with the ground truth. This is analogous to how biologists would traditionally generate phylogenetic trees for small datasets using morphological matrices. Biologists use maximum parsimony or bayesian methods [10] to find the best-fitting tree based on discrete characters (either morphological or genetic). However, the use of continuous characters in improving phylogeny generation has been recently explored [35]. Therefore if we assume our embedding space represents morphological features and is a morphological space, this could similarly be used to generate a phylogeny using the same continuous trait bayesian phylogenetic inference methods. We use RevBayes[19], a popular bayesian inference package to complete the analysis. Similar methods have been used to generate phylogenetic trees [23].

**Phylogenetic Comparison** The main purpose of this dataset is to allow exploration of methods for generating phylogenetic trees based on morphology. To do this, we need methods for comparing phylogenies. There are many standard methods of doing this in biology, a thorough comparison of them is provided in [25]. In brief, the metrics can be split into those which do and do not compare branch lengths. As branch lengths (i.e. evolutionary time) are not yet available in our ground truth phylogeny, we will focus on those which do not include branch lengths, called topology-only comparison methods. The most widely used of these is called the Robinson-Foulds (RF) metric, introduced in 1981 [37]. The RF metric defines the dissimilarity between two trees as the number of operations that would be required to turn one tree into another*. However, it has some notable disadvantages, including that apparently similar trees can have a disproportionately high RF score.

One of the more recently introduced metrics is called the Align Score [33]. The Align Score works in two stages. In the first stage, a 1:1 mapping of edges from each tree ($T_1$ and $T_2$) is assigned. This is done by calculating a similarity score $s(i, j)$ between the edges, i and j in $T_1$ and $T_2$ respectively, based on how similarly they partition the tree. More concretely, in tree $T_1$, edge $i$ will partition the tree into two disjoint subsets $P_{i0}$ and $P_{i1}$. The similarity scores can then by

* it is, however, different from the edit distance popular in computer science

**Fig. 3.** Examples of specimen images before (above) and after (below) segmentation and rotation adjustment.

computed as:

$$s(i,j) = 1 - \max(\min(a_{00}, a_{11}), min(a_{01}, a_{10})) \tag{1}$$

where $a_{rs}$ is the intersection over the union of the partitions:

$$a_{rs} = |\frac{P_{ir} \cap P_{js}}{P_{ir} \cup P_{js}}| \tag{2}$$

The munkres algorithm is then used to find the edge $j = f(i)$ that minimizes the assignment problem, and then the group with the minimum pairs are summed as follows to calculate the total align score for the two trees:

$$\sum_{i \in T_1} s(i, f(i)) \tag{3}$$

Unlike the RF score, for each set of partitions the align score calculates the similarity, $s(i,j)$, as a continuous variable instead of a binary value. That said, it has the disadvantage that the value is not normalized - a larger tree will likely have a larger align score, making the result difficult to interpret. Despite this, we choose to use it as it is a more accurate representation of the topological similarity between two trees[25].

## 3 An Overview of Rove-Tree-11

### 3.1 Image Collection

The images in the dataset were collected and prepared in 4 stages [14]:

**Stage 1: Unit Tray Image Collection** Rove-Tree-11 was collected by taking overview images of 619 unit trays from the entomology collection at Natural History Museum of Denmark, see fig. 2. A Canon EOS R5 mounted on a camera stand with a macro lens was used to take images of 5760 × 3840 pixels (px) resolution. Since the camera height and focus were kept fixed, the images can be related to physical distance as approx. 400 px per cm. Artificial lighting was used to minimize lighting variance over the images.

**Table 2.** Species-level classification results on segmented and unsegmented images. We can see that using segmentations drastically reduces the accuracy, indicating that the model is learning from the background and not the morphology of the beetle, as desired. Top-1 and Top-5 represent accuracies. Uncertainties represent 95% confidence intervals.

| Model | Dataset | Top-1 | Top-5 |
|---|---|---|---|
| ResNet-18[16] | segmented | 90.9 ± 1.2 | 99.1 ± 1.2 |
| ResNet-18[16] | unsegmented | 99.1 ± 0.3 | 99.9 ± 0.3 |

**Stage 2: Bounding Box Identification and Sorting** After image capture, bounding boxes for the individual specimens were then manually annotated using Inselect [18]. Images of 19,722 individual specimens were then sorted. Only dorsal views (views from the 'back' of the beetle) where the specimen was largely intact and limbs were mostly visible were included, resulting in images of 13,887 specimens in final dataset. See fig. 3 for examples of bounding boxes around specimens. Estimates of body rotation were also annotated in 45 degree increments which allows for coarse correction of the orientation of the crops.

**Stage 3: Segmentation** Segmentations were then generated through an iterative process. First 200 images were manually segmented. Then U-Net was trained on these 200 images and was used to generate predictions for the rest of the images. 3000 of these segmentations were considered good enough. U-Net was then retrained with these images, then rerun and new segmentations produced. The final segmentations were then manually corrected. Examples of segmentation masks and final segmented specimens can be seen in fig. 3 and fig. 4. The dataset is released with both the original crops and the segmentation masks, however, as we show in table 2, the segmentations are extremely important for phylogenetic analysis, as the background of the image is highly correlated with the species. This is because many of the same species were collected at the same time in the same place by the same person, meaning whether the specimen was glued to a card, the age and color of the card, could be correlated with the species, despite being unrelated to the phylogeny. The segmentations are not perfect. In particular they cut off some of the finer hairs on the body; It could therefore be the case that the segmentations are removing vital information which the model can use to complete classification. We consider this unlikely and suspect the model is instead learning from the backgrounds.

**Stage 4: Rotation Adjustment** Rotations were corrected by finding the principal axis of inertia of the segmentation masks, (see [21] for details). Since all the beetles are more or less oval shaped, the minimal axis of rotation of their masks tends to line up well with their heads and tails. Using this we further standardized the rotations of the segmentations. This process is shown in fig. 4.

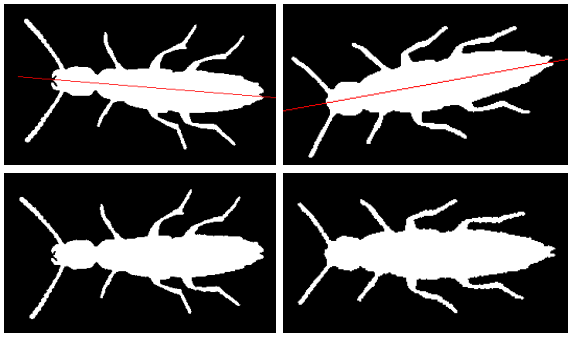

**Fig. 4.** Illustration of rotation adjustment algorithm. Example original masks (top) and rotated masks (bottom). The red line represents the principal axis of inertia found.

### 3.2 Preparation of phylogeny

A current genus-level phylogeny of the closely related subfamilies Staphylininae, Xantholininae and Paederinae is provided for the sample of genera used in our analysis. The full phylogeny is visualized in fig. 1 our supplementary material. A subset is shown in fig. 1. This phylogeny represents the current state of knowledge as it was pieced together from the most relevant recently published phylogenetic analyses, such as [47] for sister-group relationships among all three subfamilies and the backbone topology of Xantholininae and Staphylininae, [3] for the subtribe Staphylinina, [4] for the subtribe Philonthina and [46] for the subfamily Paederinae. Below genus-level the phylogeny is considered unresolved as we were unable to find species-level phylogenies for the 215 species included in Rove-Tree-11. A newick file of the phylogeny is provided with the dataset.

### 3.3 Dataset Statistics

In total, 13,887 images of beetles from the family Staphylinidae, commonly known as rove beetles, are included from 215 species - spanning 44 genera, 9 tribes and 3 subfamilies. Example images are shown in fig. 1.

The distribution of the dataset per genus is shown in fig. 7. A species-level distribution is provided in the supplementary material. From this we can see that the dataset is not evenly distributed, with the species with the highest number of specimens having 261 examples and the lowest having 2 with the genus Philonthus accounting for 24.8% of the dataset. This is due to the number of specimens the museum had in the unit trays that were accessed and imaged at the time, although the curators also includes samples of species which were easily distinguishable from each other, and examples which were hard and can only usually be determined by genital extraction by experts (ie Lathrobium geminum and Lathrobium elongatum. Examples from these two species are shown in fig. 5 to demonstate the difficulty of the task). The distribution of image sizes in the dataset is shown in fig. 6. The majority of the images (82%) are under $500 \times 250$ pixels.

**Fig. 5.** Example images of Lathrobium geminum (top) and Lathrobium elongatum (bottom) from the dataset. Typically even experts need to dissect the specimen to complete the determination between these two species.

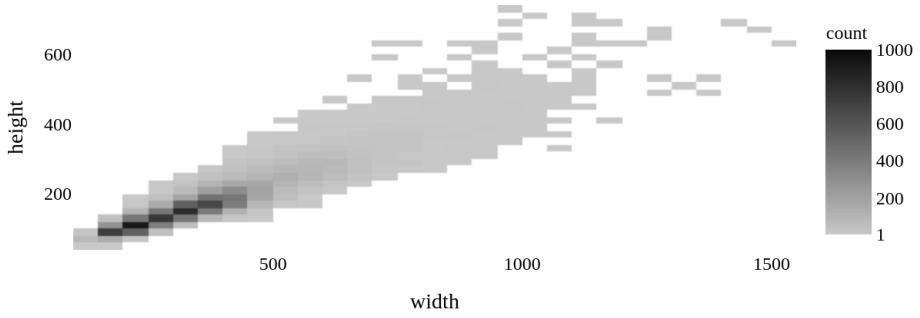


**Fig. 6.** Distribution of image sizes included in the dataset. The majority (82%) of images are under $500 \times 250$ pixels.

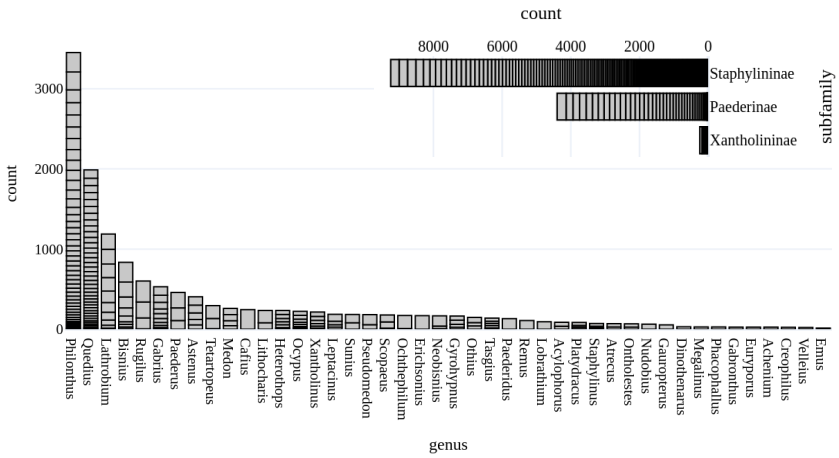


**Fig. 7.** Distribution of specimens per genus (bottom left) and per subfamily (top right). Each slice in the stacked bar chart represents a different species within that genus. Subfamily distribution is included as it is used to generate the validations and test sets for the clustering results in sec. 4.2. A full species level distribution is shown in the supplemental material.

**Table 3.** Classification results using deep learning architectures. Top-1 and Top-5 represent accuracies. Uncertainties represent 95% confidence intervals based on 3 runs.

| Model | Params | Species | | Genus | |
|---|---|---|---|---|---|
| | | Top-1 | Top-5 | Top-1 | Top-5 |
| ResNet-18 [16] | 11.4 M | 90.9±1.2 | 99.2±1.2 | 98.9±0.3 | 100±0.3 |
| ResNet-50 [16] | 23.9 M | 89.4±1.4 | 99.2±1.4 | 98.2±0.4 | 100±0.4 |
| EfficientNet B0 [39] | 5.3 M | 91.9±1.8 | 99.3±1.8 | 99.1±0.2 | 100±0.2 |

## 4 Evaluation

Here we evaluate the dataset by performing benchmark experiments. As stated previously, the main purpose of this dataset is for deep metric learning on hierarchical phylogenetic relationships, so this is also the focus of the benchmarks, although we also provide benchmarks for the classification and clustering tasks. The same augmentations were applied to the dataset as for CUB200 and Cars176 and as in [38], with the exception that the RandomHorizontalFlip was changed to a RandomVerticalFlip, as this makes more sense for the Rove-Tree-11 dataset. Gradient accumulation was also used in some cases due to memory constraints on the available clusters. The details of which experiments this was applied to are provided in the codebase.

### 4.1 Classification

Results from classification experiments are provided in table 3. For these experiments the official pytorch implementations were used with default parameters: categorical cross entropy loss with an initial learning rate of 0.1, momentum of 0.9, weight decay of $1e-4$ and SGD optimizer. Training details are released with the code for this dataset. The only alterations from the defaults were to reduce the batch-size to 32 due to memory constraints and to alter the data augmentations, detailed in the code. A species-stratified train/val/test split of 70/15/15 was used. The split is provided with the code.

As shown in table 3, the models are able to achieve a top-1 species-level accuracy of 92% with no hyperparameter tuning, and a top-1 genus level of almost 100%. These results suggest that although this dataset could be used for classification tasks and might be useful as such for biologists, classification of this dataset is not particularly difficult, and this dataset is probably not ideal as a benchmark for classification in deep learning.

### 4.2 Clustering and Phylogenetic Results

In table 4, we present benchmark results of applying state of the art methods for deep metric learning to the Rove-Tree-11 dataset and comparing phylogenies generated using phylogenetic bayesian methods on the embedding space to the

ground truth phylogeny as described in sec. 2. A more complete table showing R1 scores and Cars176 results, is provided for reference in the supplementary material (table 1). The 'Random' row represents the align score of a randomly generated tree with the 9 genera leaves included in the test set, against the ground truth tree based on 5 random initializations. Since the align score is not normalized, this random baseline is useful to gauge our results and represents an upper bound our models should achieve. Following best practice, as described in [30], the dataset was split into three groups for training, validation and testing. To properly test the ability of the model to generalize, the groups were split at subfamily level, so the train, validation and test sets should be as phylogenetically distinct as possible, in the sense that they belong to different parts of the phylogenetic tree. This results in 8534 training images from the subfamily Staphylininae, 4399 validation images from the subfamily Paederinae and 954 test images from the subfamily Xantholininae.

All results on Rove-Tree-11 were generated using implementations used in [38], modified to calculate the align score. A forked codebase is provided as a submodule in the github repository.

Based on the clustering results in table 4, we see that Rove-Tree-11 has similar NMI scores to CUB200, suggesting this dataset has a similar clustering difficulty to CUB200 and may be appropriate as a clustering benchmark. As with CUB200, the best models on Rove-Tree-11 are Triplet [43] and Multisimilarity [42]. We can also see that the align score results somewhat correspond with the NMI, with the best results being achieved with Triplet Loss. We can also see that the best test set align score of 4.0 is a marked improvement to the random align score baseline of 6.6, but still significantly far away from a perfect align score of 0, suggesting there is room for improvement. We find it surprising that the align score of the best model on the CUB200 dataset shows a 60% improvement to the random score, while on Rove-Tree-11 the improvement is only 40% on the test set and 51% on the validation set. This suggests that either CUB200 is an easier dataset to generate phylogenies from, or could be an artifact of the align score on trees of different depths (CUB200 has a depth of 4, while Rove-Tree-11 has a depth of 11). It is surprising that it could be an easier dataset, given that the images are in-the-wild, but this could also be due to phylogenetically close birds having similar backgrounds in the images (water-faring birds might typically have ocean backgrounds, for example, and be more closely phylogenetically related). The phylogenetic tree produced by the best model is provided in the supplementary material along with the ground truth tree for visual inspection.

## 5 Conclusions

In this paper we present Rove-Tree-11, a novel dataset of segmented images of and research-grade classifications of rove beetles for researching methods for generating phylogenies from images. We provide an eleven-level fine-grained ground

**Table 4.** Benchmark clustering and Align-Score results on Rove-Tree-11 dataset. 'Random' represents the average align score of 5 randomly generated trees. This gives us a metric to compare our results with. A perfect align score would be 0. 95% confidence errors are provided based on 5 runs.

| | CUB200 | | Rove-Tree-11 | | | |
|---|---|---|---|---|---|---|
| | Test | | Validation | | Test | |
| Loss | NMI | Align | NMI | Align | NMI | Align |
| Random | - | 21.9±0.2 | - | 15.8±0.9 | - | 6.6±0.5 |
| Triplet | 64.8±0.5 | 9.9±0.9 | 68.9±0.4 | **7.8±1.1** | 66.3±0.3 | 4.1±0.5 |
| Margin | 60.7±0.3 | 10.6±1.2 | 68.0±0.7 | 8.2±0.7 | 65.9±0.5 | 4.2±0.7 |
| Lifted | 34.8±3.0 | 15.9±2.0 | 55.0±0.6 | 10.5±0.7 | 56.0±1.1 | 4.9±0.8 |
| Constrast. | 59.0±1.0 | 11.0±1.2 | 66.7±0.5 | 8.5±1.0 | 65.4±0.5 | 4.5±0.6 |
| Multisim. | **68.2±0.3** | **8.6±0.8** | **70.7±0.2** | 8.2±0.4 | **67.3±0.5** | **4.0±0.5** |
| ProxyNCA | 66.8±0.4 | 9.8±0.8 | 67.5±0.7 | 9.0±0.8 | 65.5±0.3 | 4.2±0.4 |
| Arcface | 67.5±0.4 | 9.8±0.8 | 66.9±0.9 | 8.5±0.4 | 64.8±0.5 | 4.1±0.4 |

truth phylogeny for the 44 (train, validation and test) genera included in this dataset.

We start by demonstrating the importance of the provided segmentations as the model can learn from the background. We show benchmark results on this dataset for classification, deep metric learning methods and tree alignment. We further demonstrate that this dataset shows similar clustering results to the CUB200 dataset suggesting it may be appropriate as an alternative clustering benchmark. Finally, we demonstrate how this dataset can be used to generate and compare phylogenies based on the align score, and show that while it is possible to generate such trees, there is plenty of room for improvement and we hope this will be a growing field of research. Code and data are available (code: https://github.com/robertahunt/Rove-Tree-11, data: http://doi.org/10.17894/ucph.39619bba-4569-4415-9f25-d6a0ff64f0e3).

**Ethical Concerns** Models similar to those described, if applied to images of faces, could be used to generate family trees for humans. This could result in public images being used to infer familial relationships which could have a negative societal impact. The authors strongly discourage this form of misuse of the proposed methods.

**Acknowledgements** Many people have been involved in this project. First we would like to thank David Gutschenreiter, Søren Bech and André Fastrup who took the photos of the unit trays and completed the initial segmentations of the images as part of their theses. Next, we would like to thank Alexey Solodovnikov of the Natural History Museum of Denmark for providing the specimens, the ground truth phylogeny and guidance for all things entomological. Also, thanks Francois Lauze and the entire Phylorama team for their input the project.

## References


1. Bakalar, N.: Nicholas. The New York Times (2014), https://www.nytimes.com/2014/05/27/science/welcoming-the-newly-discovered.html
2. Bameri, F., Pourreza, H.R., Taherinia, A.H., Aliabadian, M., Mortezapour, H.R., Abdilzadeh, R.: TMTCPT: The tree method based on the taxonomic categorization and the phylogenetic tree for fine-grained categorization. Biosystems **195**, 104137 (jul 2020). https://doi.org/10.1016/j.biosystems.2020.104137, https://doi.org/10.1016%2Fj.biosystems.2020.104137
3. Brunke, A., Smetana, A.: A new genus of staphylinina and a review of major lineages (staphylinidae: Staphylininae: Staphylinini). Systematics and Biodiversity **17**, 745–758 (11 2019). https://doi.org/10.1080/14772000.2019.1691082
4. Chani-Posse, M.R., Brunke, A.J., Chatzimanolis, S., Schillhammer, H., Solodovnikov, A.: Phylogeny of the hyper-diverse rove beetle subtribe philonthina with implications for classification of the tribe staphylinini (coleoptera: Staphylinidae). Cladistics **34**(1), 1–40 (2018). https://doi.org/https://doi.org/10.1111/cla.12188
5. Cho, H., Ahn, C., Min Yoo, K., Seol, J., Lee, S.g.: Leveraging class hierarchy in fashion classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops (Oct 2019)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Deng, J., Guo, J., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. CoRR **abs/1801.07698** (2018), http://arxiv.org/abs/1801.07698
8. DiSSCo: Distributed system of scientific collections. https://www.dissco.eu/ (July 2022)
9. Felsenstein, J.: Evolutionary trees from dna sequences: A maximum likelihood approach. Journal of Molecular Evolution **17**, 368–376 (1981)
10. Felsenstein, J.: Statistical inference of phylogenies. Journal of the Royal Statistical Society: Series A (General) **146**(3), 246–262 (May 1983). https://doi.org/https://doi.org/10.2307/2981654
11. Felsenstein, J.: Inferring phylogenies. Sinauer associates, Sunderland, MA (2003)
12. Fink, M., Ullman, S.: From aardvark to zorro: A benchmark for mammal image classification. Int. J. Comput. Vision **77**(1–3), 143–156 (may 2008). https://doi.org/10.1007/s11263-007-0066-8, https://doi.org/10.1007/s11263-007-0066-8
13. Goëau, H., Bonnet, P., Joly, A.: Overview of plantclef 2021: cross-domain plant identification. In: Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum. vol. 2936, pp. 1422–1436 (2021)
14. Gutschenreiter, D., Bech, S.: Deep-learning methods on taxonomic beetle data Automated segmentation and classification of beetles on genus and species level. Master's thesis, University of Copenhagen (2021)
15. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 1735–1742 (2006). https://doi.org/10.1109/CVPR.2006.100
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)

17. Hedrick, B.P., Heberling, J.M., Meineke, E.K., Turner, K.G., Gassa, C.J., Park, D.S., Kennedy, J., Clarke, J.A., Cook, J.A., Blackburn, D.C., Edwards, S.V., Davis, C.C.: Digitization and the future of natural history collections. BioScience **70**(3), 243–251 (February 2020). https://doi.org/https://doi.org/10.1093/biosci/biz163
18. Hudson, L.N., Blagoderov, V., Heaton, A., Holtzhausen, P., Livermore, L., Price, B.W., van der Walt, S., Smith, V.S.: Inselect: Automating the digitization of natural history collections. PLOS ONE **10**(11), 1–15 (11 2015). https://doi.org/10.1371/journal.pone.0143402, https://doi.org/10.1371/journal.pone.0143402
19. Höhna, L., Heath, B., Lartillot, M., Huelsenbeck, R.: Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Systematic Biology **65**, 726–736 (2016)
20. iDigBio: Integrated digitized biocollections. https://www.idigbio.org/ (July 2022)
21. J. Peraire, S.W.: Lecture notes from mit course 16.07 dynamics, fall 2008. l26 - 3d rigid body dynamics: The inertia tensor (2008), https://ocw.mit.edu/courses/16-07-dynamics-fall-2009/dd277ec654440f4c2b5b07d6c286c3fd_MIT16_07F09_Lec26.pdf
22. KAYA, M., BİLGE, H.S.: Deep metric learning: A survey. Symmetry **11**(9) (2019). https://doi.org/10.3390/sym11091066, https://www.mdpi.com/2073-8994/11/9/1066
23. Kiel, S.: Assessing bivalve phylogeny using deep learning and computer vision approaches. bioRxiv (2021). https://doi.org/10.1101/2021.04.08.438943, https://www.biorxiv.org/content/early/2021/04/09/2021.04.08.438943
24. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13). Sydney, Australia (2013)
25. Kuhner, M.K., Yamato, J.: Practical Performance of Tree Comparison Metrics. Systematic Biology **64**(2), 205–214 (12 2014). https://doi.org/10.1093/sysbio/syu085
26. Lee, M., Palci, A.: Morphological phylogenetics in the genomic age. Current Biology **25**(19), R922–R929 (2015). https://doi.org/https://doi.org/10.1016/j.cub.2015.07.009, https://www.sciencedirect.com/science/article/pii/S096098221500812X
27. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
28. MacQueen, J.: Classification and analysis of multivariate observations. In: 5th Berkeley Symp. Math. Statist. Probability. pp. 281–297 (1967)
29. Movshovitz-Attias, Y., Toshev, A., Leung, T.K., Ioffe, S., Singh, S.: No fuss distance metric learning using proxies. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 360–368 (2017). https://doi.org/10.1109/ICCV.2017.47
30. Musgrave, K., Belongie, S., Lim, S.N.: A metric learning reality check. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 681–699. Springer International Publishing, Cham (2020)
31. Natural History Museum of Denmark: Digital nature: Giant grant makes the natural history collections of denmark accessible to everyone. Newsletter (2021)
32. Natural History Museum of Denmark: Entomology - Dry and Wet Collections. Homepage (2022)

33. Nye, T., Lio, P., Gilks, W.: A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. Bioinformatics (Oxford, England) **22**, 117–9 (02 2006). https://doi.org/10.1093/bioinformatics/bti720
34. Orlov, I., Leschen, R.A., Żyła, D., Solodovnikov, A.: Total-evidence backbone phylogeny of aleocharinae (coleoptera: Staphylinidae). Cladistics **37**(4), 343–374 (2021). https://doi.org/https://doi.org/10.1111/cla.12444
35. Parins-Fukuchi, C.: Use of Continuous Traits Can Improve Morphological Phylogenetics. Systematic Biology **67**(2), 328–339 (09 2017). https://doi.org/10.1093/sysbio/syx072, https://doi.org/10.1093/sysbio/syx072
36. Popov, D., Roychoudhury, P., Hardy, H., Livermore, L., Norris, K.: The value of digitising natural history collections. Research Ideas and Outcomes **7**, e78844 (December 2021). https://doi.org/https://doi.org/10.3897/rio.7.e78844
37. Robinson, D., Foulds, L.: Comparison of phylogenetic trees. Mathematical Biosciences **53**(1), 131–147 (1981). https://doi.org/https://doi.org/10.1016/0025-5564(81)90043-2, https://www.sciencedirect.com/science/article/pii/0025556481900432
38. Roth, K., Milbich, T., Sinha, S., Gupta, P., Ommer, B., Cohen, J.P.: Revisiting training strategies and generalization performance in deep metric learning (2020)
39. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. PMLR (09–15 Jun 2019), https://proceedings.mlr.press/v97/tan19a.html
40. Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., Mac Aodha, O.: Benchmarking representation learning for natural world image collections. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12884–12893 (2021)
41. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
42. Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5017–5025 (2019). https://doi.org/10.1109/CVPR.2019.00516
43. Wu, C.Y., Manmatha, R., Smola, A.J., Krähenbühl, P.: Sampling matters in deep embedding learning. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2859–2867 (2017). https://doi.org/10.1109/ICCV.2017.309
44. Wu, X., Zhan, C., Lai, Y., Cheng, M.M., Yang, J.: Ip102: A large-scale benchmark dataset for insect pest recognition. In: IEEE CVPR. pp. 8787–8796 (2019)
45. Yuan, Y., Chen, W., Yang, Y., Wang, Z.: In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation (2019)
46. Żyła, D., Bogri, A., Hansen, A., Jenkins Shaw, J., Kypke, J., Solodovnikov, A.: A new termitophilous genus of paederinae rove beetles (coleoptera, staphylinidae) from the neotropics and its phylogenetic position. Neotropical Entomology (02 2022). https://doi.org/10.1007/s13744-022-00946-x
47. Żyła, D., Solodovnikov, A.: Multilocus phylogeny defines a new classification of staphylininae (coleoptera, staphylinidae), a rove beetle group with high lineage diversity. Systematic Entomology **45**(1), 114–127 (2020). https://doi.org/https://doi.org/10.1111/syen.12382