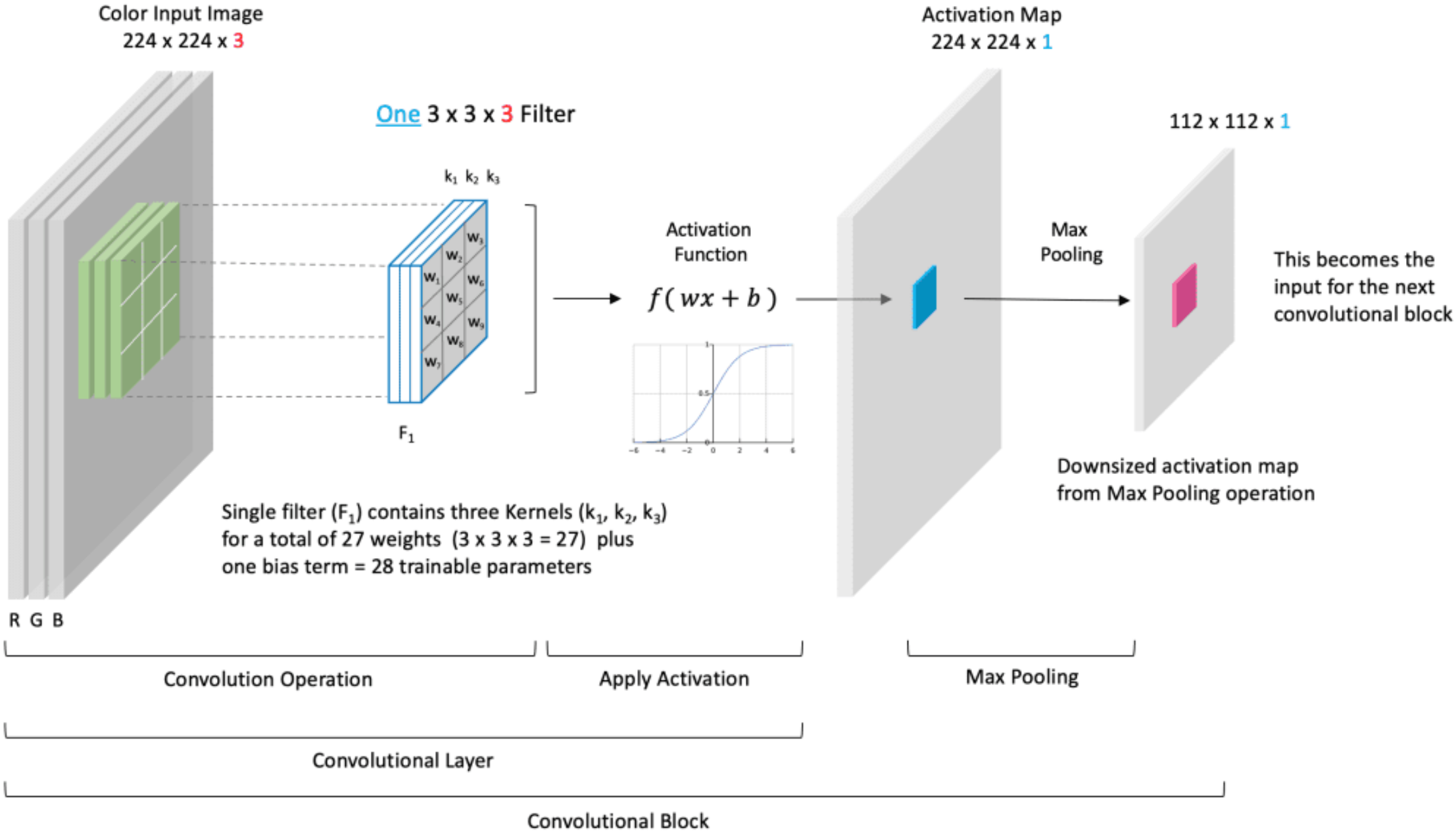
Photo by Dan Nelson, http://10000thingsofthepnw.com/2021/12/20/ocypus-olens-devils-coach-horse/

Roberta Hunt, José L. Reyes-Hernández, Josh Jenkins Shaw, Alexey Solodovnikov, Kim Steenstrup Pedersen. 2025. "Integrating Deep Learning Derived Morphological Traits and Molecular Data for Total-Evidence Phylogenetics." *Systematic Biology* 74(3): 453-468 https://doi.org/10.1093/sysbio/syae072

Class focus area: Machine learning for traits

EEB603: Brian O'Meara

All quotes and images from the above paper unless otherwise noted



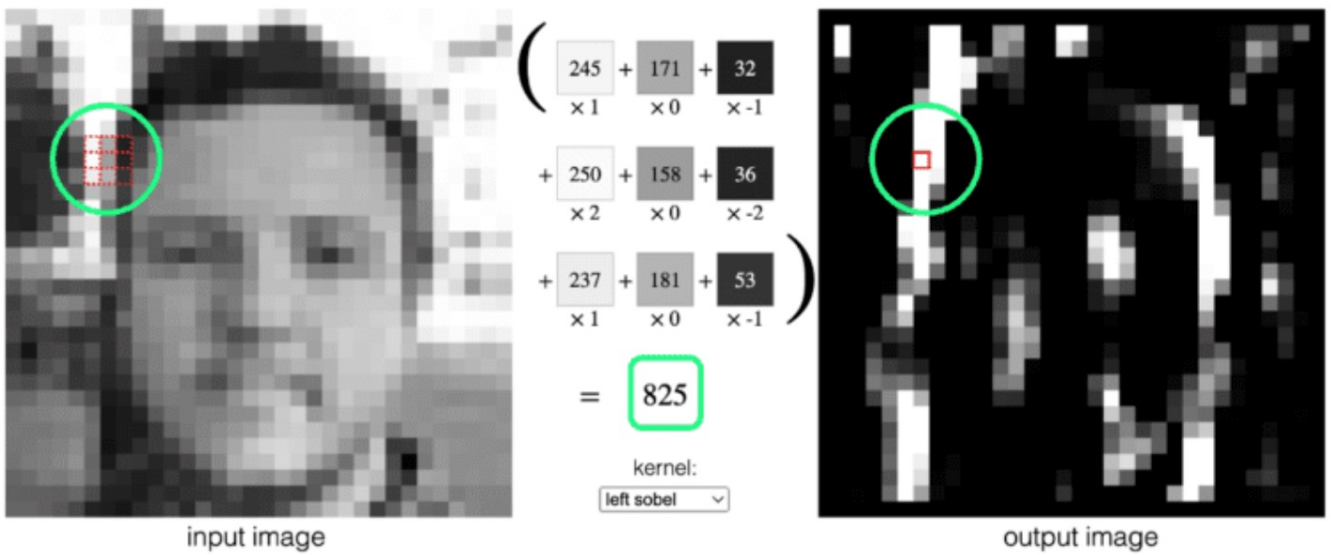
From (and a really good explanation at): https://learnopencv.com/understanding-convolutional-neural-networks-cnn/

Sobel Kernel Example

Here we show a concrete example of how a **Sobel Kernel** detects vertical edges. Recall the convolution operation defined above is the weighted sum of the kernel values with the corresponding input values. Since the Sobel kernel has positive values in the left column, zeros in the center column, and negative values in the right column, as the kernel is moved from left to right across the image, the convolution operation is a numerical approximation of the derivative in the horizontal direction, and therefore, the output produced detects vertical edges. This is an example of how specific kernels can detect various structures in images like edges. Other kernels can be used to detect horizontal lines/edges or diagonal lines/edges. In CNN, this concept is generalized. Since the kernel weights are learned during the training process, CNNs can therefore learn to detect many types of features that support image classification.

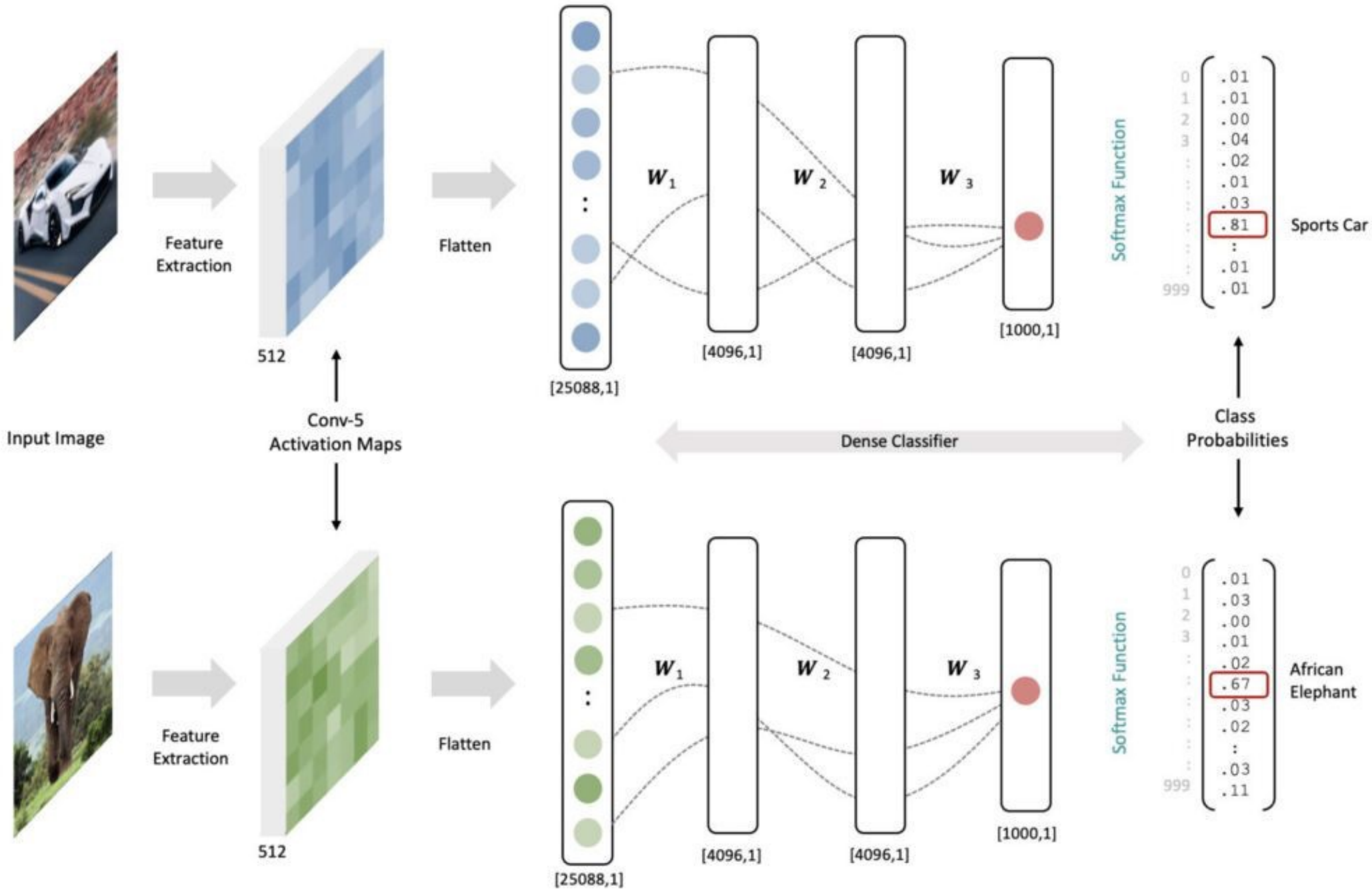
Sobel Kernel
(3 x 3)

1	0	-1
2	0	-2
1	0	-1

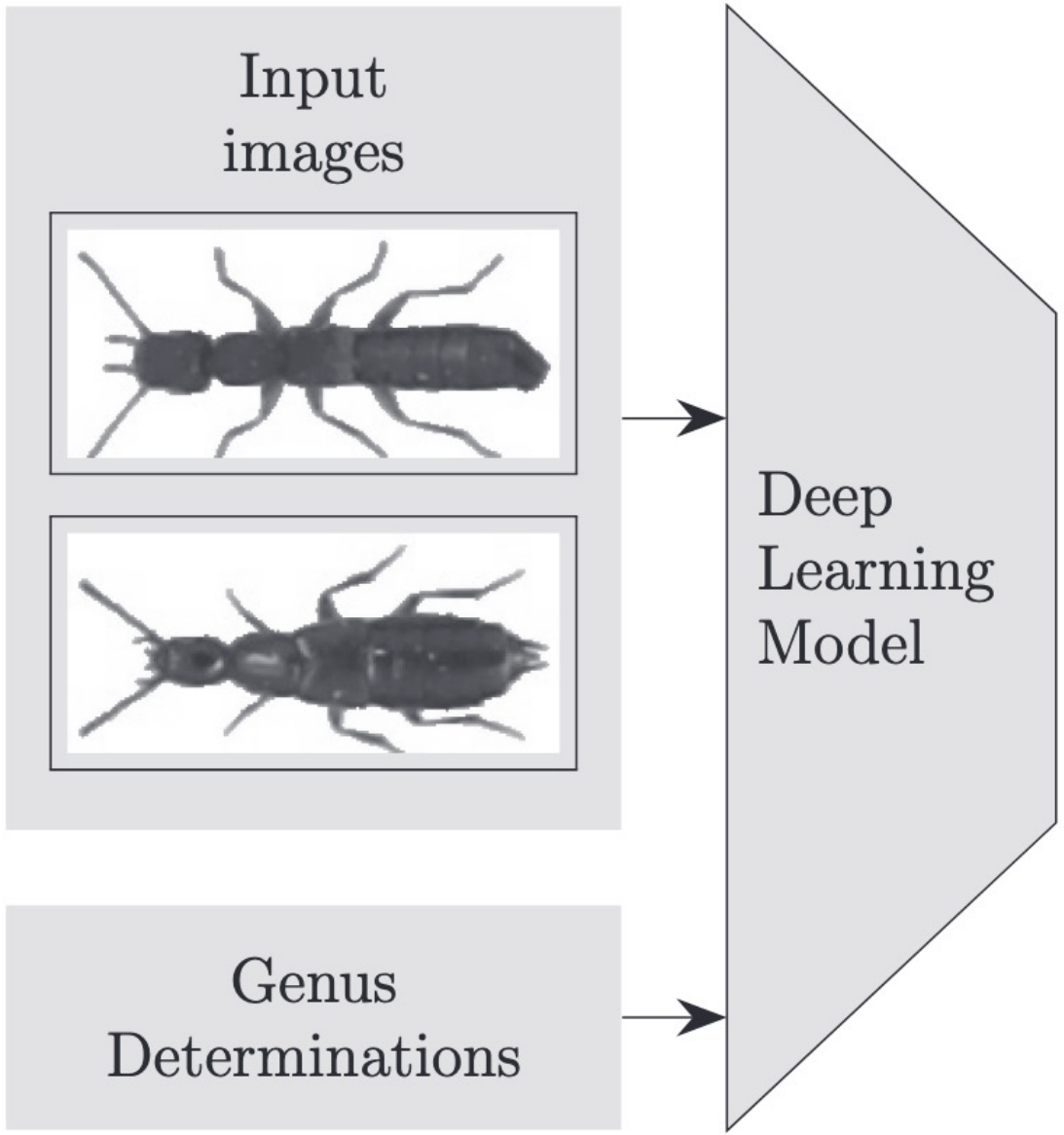


Source: https://setosa.io/ev/image-kernels/

From (and a really good explanation at): https://learnopencv.com/understanding-convolutional-neural-networks-cnn/



From (and a really good explanation at): https://learnopencv.com/understanding-convolutional-neural-networks-cnn/



Average Continuous
Traits Per Species

genus	trait0	...	trait128
g. 1	0.124	...	0.124
g. 2	1.453	...	-4.792
...	...	...	...
g. 44	-0.569	...	2.223

FIGURE 1. Pipeline for generation of continuous morphological traits per genus. A deep metric learning model was trained on d
ages. It used the genus determinations to pull specimens from the same genus closer together in space. A vector of 128 traits was
from the model for each image. The average vector for each genus was then calculated.

Combined Continuous and Molecular Traits

genus	trait0	...	trait128	bp0	...	bp4553
g. 1	0.124	...	0.124	G	...	A
g. 2	1.453	...	-4.792	T	...	A
...	...	...	...	...	...	...

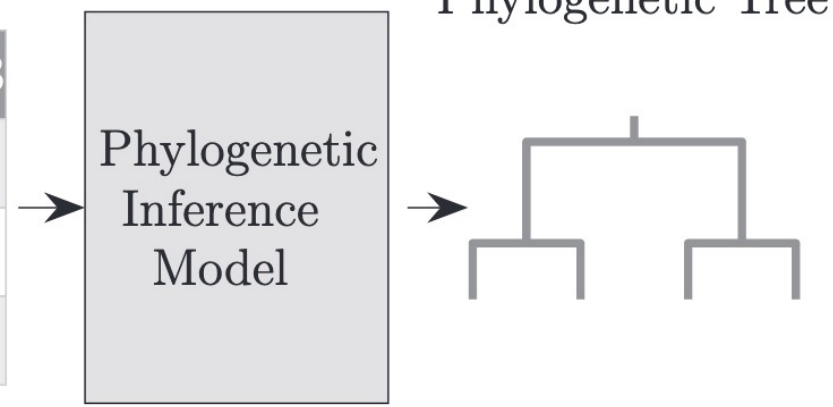


FIGURE 2. Pipeline of phylogenetic tree generation. First the continuous traits given by the deep learning model were combined with the molecular data. Both of these were then fed into the phylogenetic inference model which generates a tree.

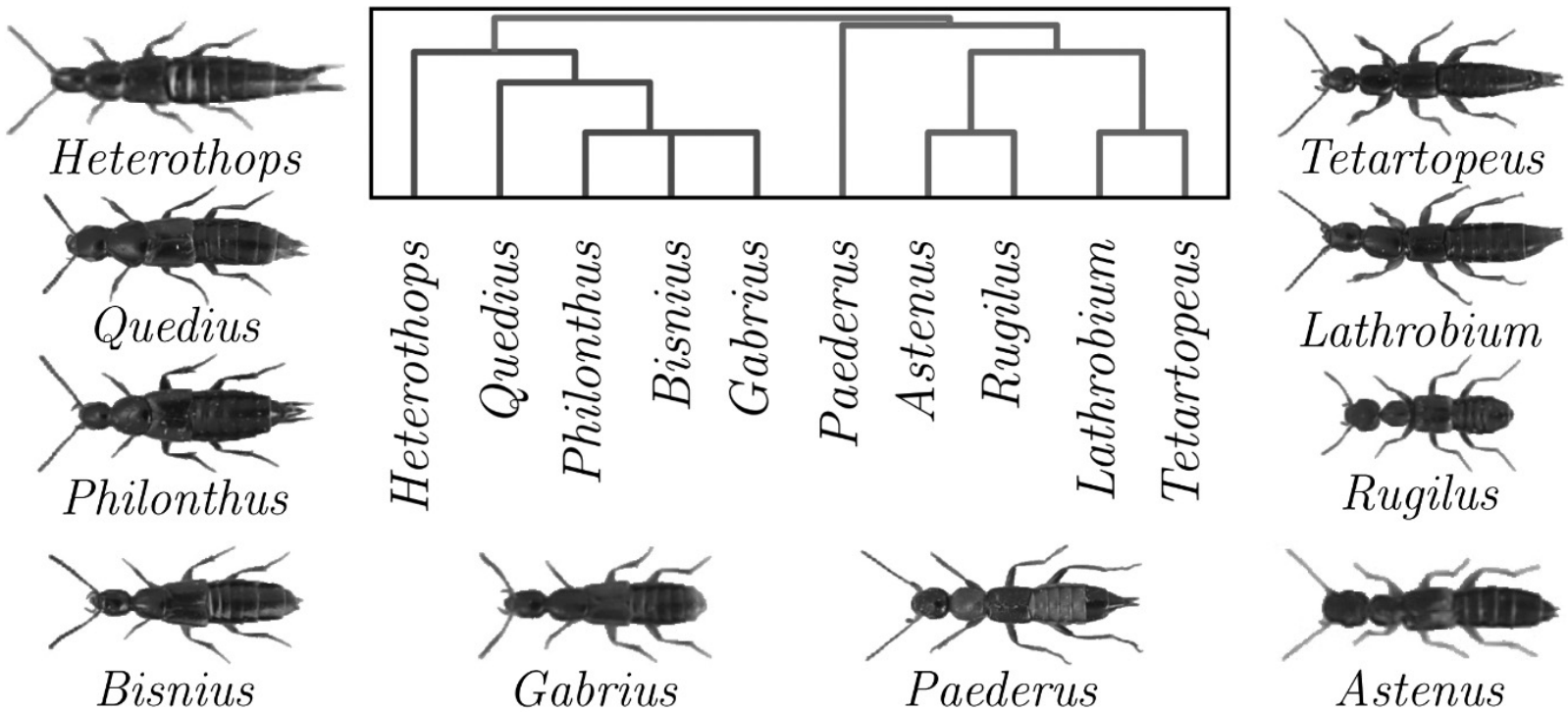


FIGURE 3. Subset of the reference phylogeny from the Rove-Tree-11 dataset, for the 10 genera with the most images in the dataset. Each leaf represents a genus. Example specimens from each of the genera are shown in black and white for reference. Reproduced from Hunt and Pedersen (2022) with permission.

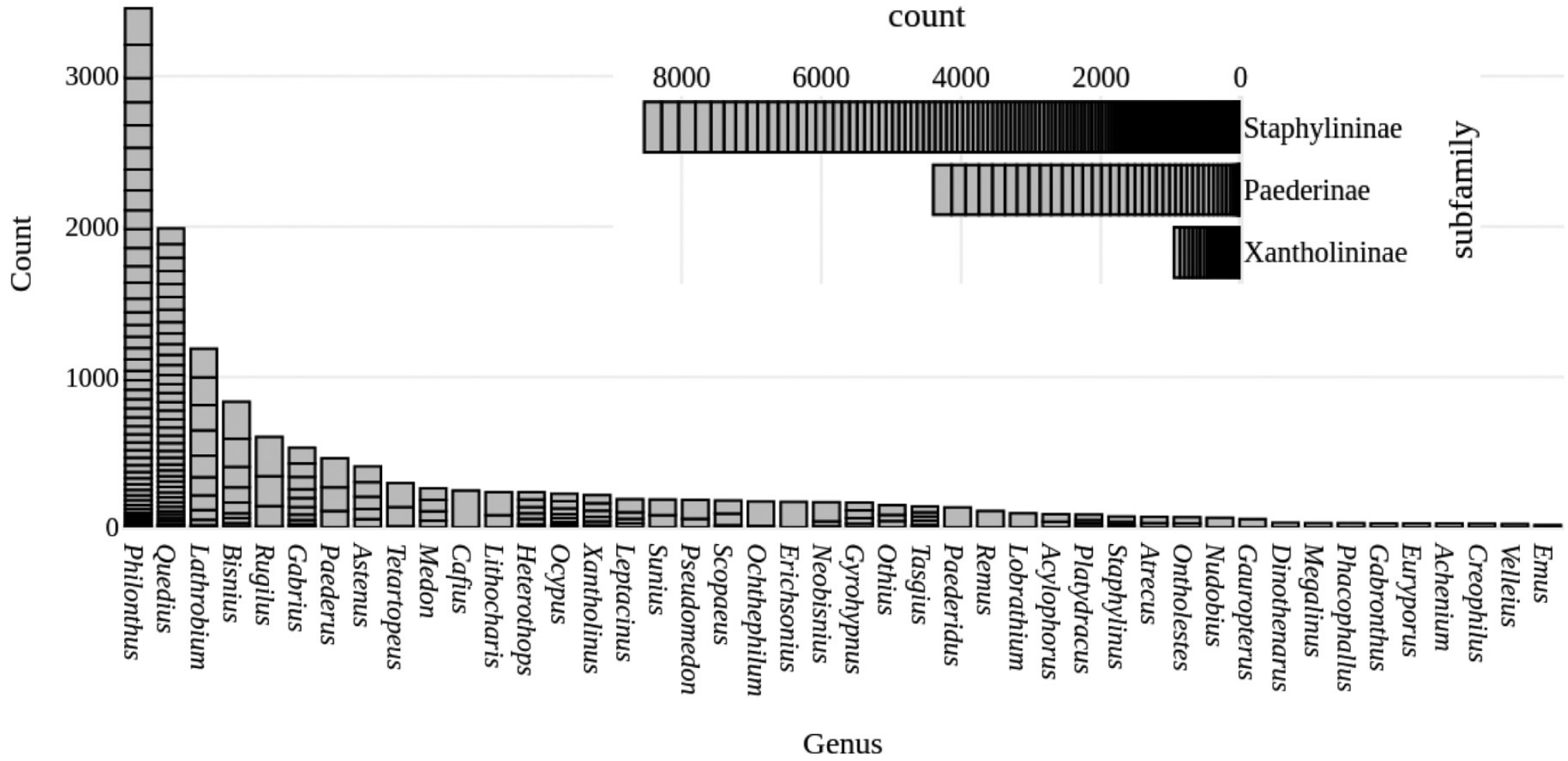


FIGURE 4. Distribution of specimens per genus (bottom left) and per subfamily (top right). Each slice in the stacked bar chart represents a different species within that genus. Reproduced from Hunt and Pedersen (2022) with permission.

TABLE 1. Tree Inference Results. Each row is the average of 5 runs. As a baseline for the scores, five randomly generated trees of this size gave a nAS^a of 0.702 ± 0.022 and a nRF^b score of 0.993 ± 0.007. Results are reported with 95% confidence intervals, using a student's t-distribution. Best results in bold. Results within confidence interval of the best score are underlined.

		nAS			nRF		
Dataset Split	Loss Function		Average	Median		Average	Median
Clade	Arcface	0.595 ± 0.031	0.627 ± 0.022	0.625	0.913 ± 0.028	0.975 ± 0.017	0.969
Clade	Contrastive		0.629 ± 0.017	0.637		0.967 ± 0.005	0.969
Clade	Margin		0.615 ± 0.045	0.616		0.975 ± 0.017	0.970
Clade	Multisim.		$\mathbf{0.506 \pm 0.034}$	0.512		0.877 ± 0.043	0.897
Clade	Proxy		0.638 ± 0.009	0.634		0.987 ± 0.022	1.000
Clade	Triplet		0.560 ± 0.045	0.563		0.901 ± 0.065	0.903
Stratified	Arcface	0.596 ± 0.020	0.692 ± 0.042	0.708	0.947 ± 0.019	0.992 ± 0.022	1.000
Stratified	Contrastive		0.531 ± 0.024	0.538		$\mathbf{0.829 \pm 0.026}$	**0.815**
Stratified	Margin		0.596 ± 0.036	0.597		0.928 ± 0.045	0.925
Stratified	Multisim.		0.524 ± 0.076	**0.505**		0.845 ± 0.057	0.825
Stratified	Proxy		0.699 ± 0.009	0.699		1.000 ± 0.000	1.000
Stratified	Triplet		0.530 ± 0.067	0.532		0.886 ± 0.062	0.900

a normalized Align Score
b normalized Robinson-Foulds

TABLE 2. Phylogenetic signal quantification using Abouheif's Cmean. Each row is the average of the 128 traits of all 5 runs. Results reported with 95% confidence intervals, using a student's t-distribution. Best results in bold. Results within confidence interval of the best score are underlined

		Average Cmean		Maximum Cmean		P value < 0.05 with Cmean		
Dataset split	Loss function	Cmean	P value	Cmean	P value	> 0.3	> 0.5	> 0.7
Clade	Arcface	0.107 ± 0.010	0.219 ± 0.019	0.480	0.001	9%	0%	0.0%
Clade	Contrastive	$\mathbf{0.288 \pm 0.014}$	0.067 ± 0.011	**0.794**	0.001	**46%**	**14%**	**0.6%**
Clade	Margin	0.188 ± 0.012	0.118 ± 0.015	0.610	0.001	21%	2%	0.0%
Clade	Multisim.	0.208 ± 0.013	0.118 ± 0.016	0.658	0.001	28%	4%	0.0%
Clade	Proxy	0.101 ± 0.010	0.226 ± 0.020	0.519	0.001	6%	0%	0.0%
Clade	Triplet	0.244 ± 0.012	0.076 ± 0.011	0.683	0.001	35%	5%	0.0%
Stratified	Arcface	-0.028 ± 0.007	0.511 ± 0.022	0.241	0.008	0%	0%	0.0%
Stratified	Contrastive	0.281 ± 0.012	0.050 ± 0.009	0.706	0.001	44%	8%	0.2%
Stratified	Margin	0.164 ± 0.011	0.138 ± 0.015	0.578	0.001	16%	1%	0.0%
Stratified	Multisim.	0.128 ± 0.010	0.179 ± 0.018	0.517	0.001	10%	0%	0.0%
Stratified	Proxy	-0.025 ± 0.008	0.498 ± 0.023	0.273	0.007	0%	0%	0.0%
Stratified	Triplet	0.232 ± 0.011	0.076 ± 0.012	0.645	0.001	31%	4%	0.0%

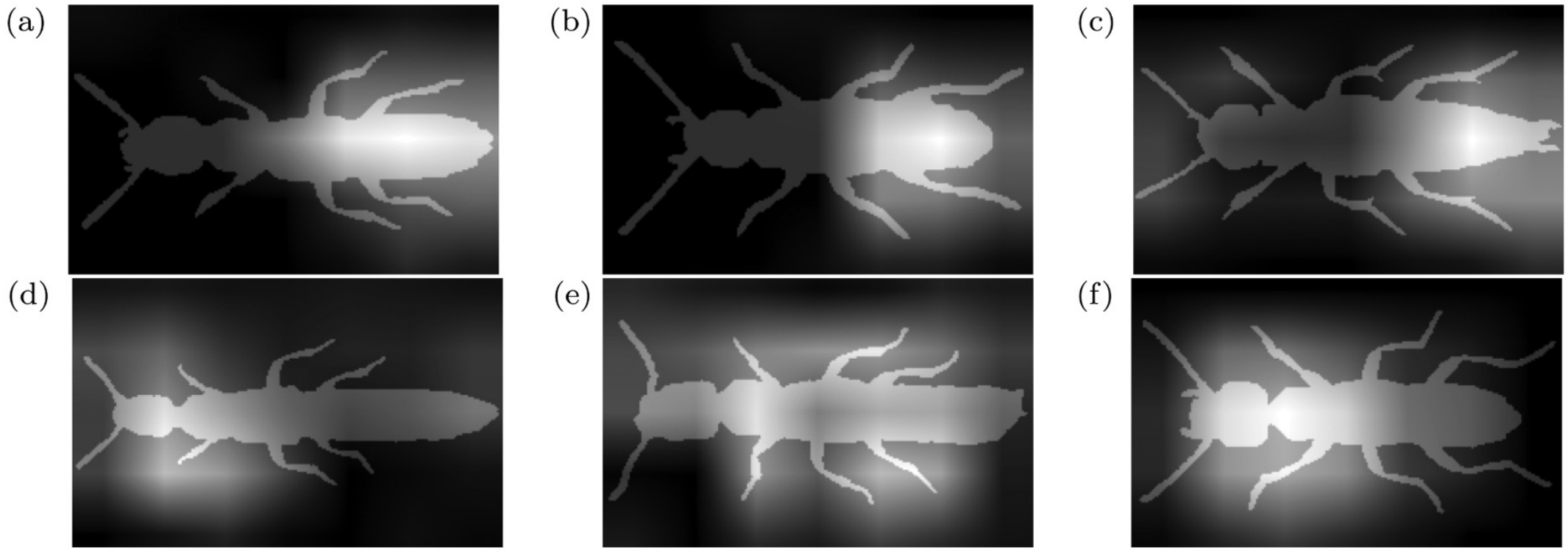


FIGURE 5. Examples of gradcam saliency maps (trait 30 from the stratified dataset with triplet loss, seed 4). Saliency maps are shown superimposed on the mask of the beetle with brighter pixel values indicating higher influence on the latent variable. Saliencies in the top row are from genera (a) *Atrecus*, (b) *Lithocharis* and (c) *Ontholestes*, these show the model focusing on the abdomen for this trait. And (d) *Gyrohypnus*, (e) *Nudobius* and (f) *Rugilus* demonstrate some counter examples for this trait. In the case of *Rugilus* (f) the model focuses instead on the neck region which is distinctively small in the *Rugilus* genus.

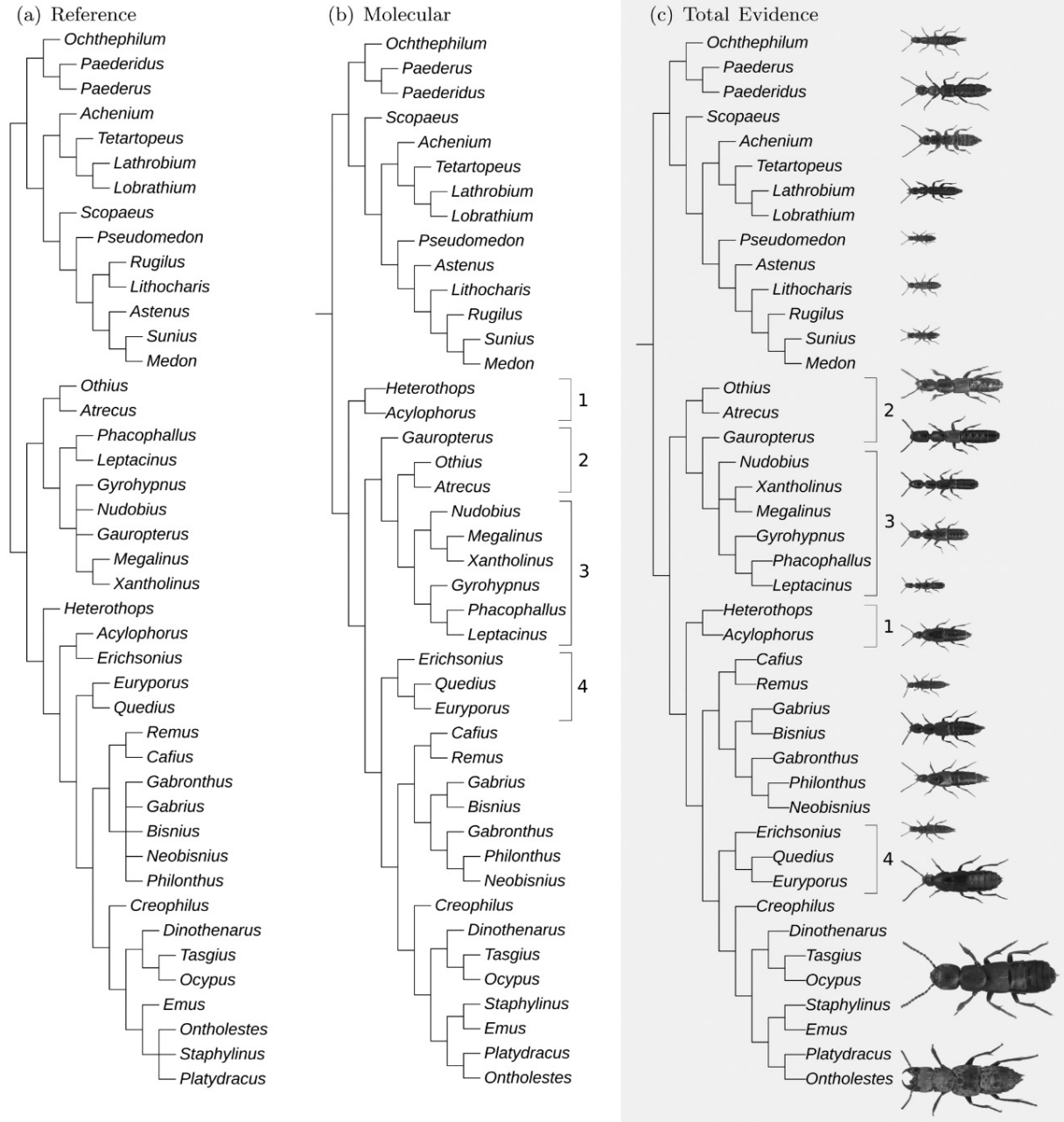


FIGURE 6. Comparison of a) reference phylogeny, b) best molecular-only tree and c) best total evidence tree. Differences between best molecular-only and best total evidence tree highlighted by indicating the controversial groups 1,2,3,4 on both trees. Plots produced in part using iTOL (Letunic and Bork, 2021)

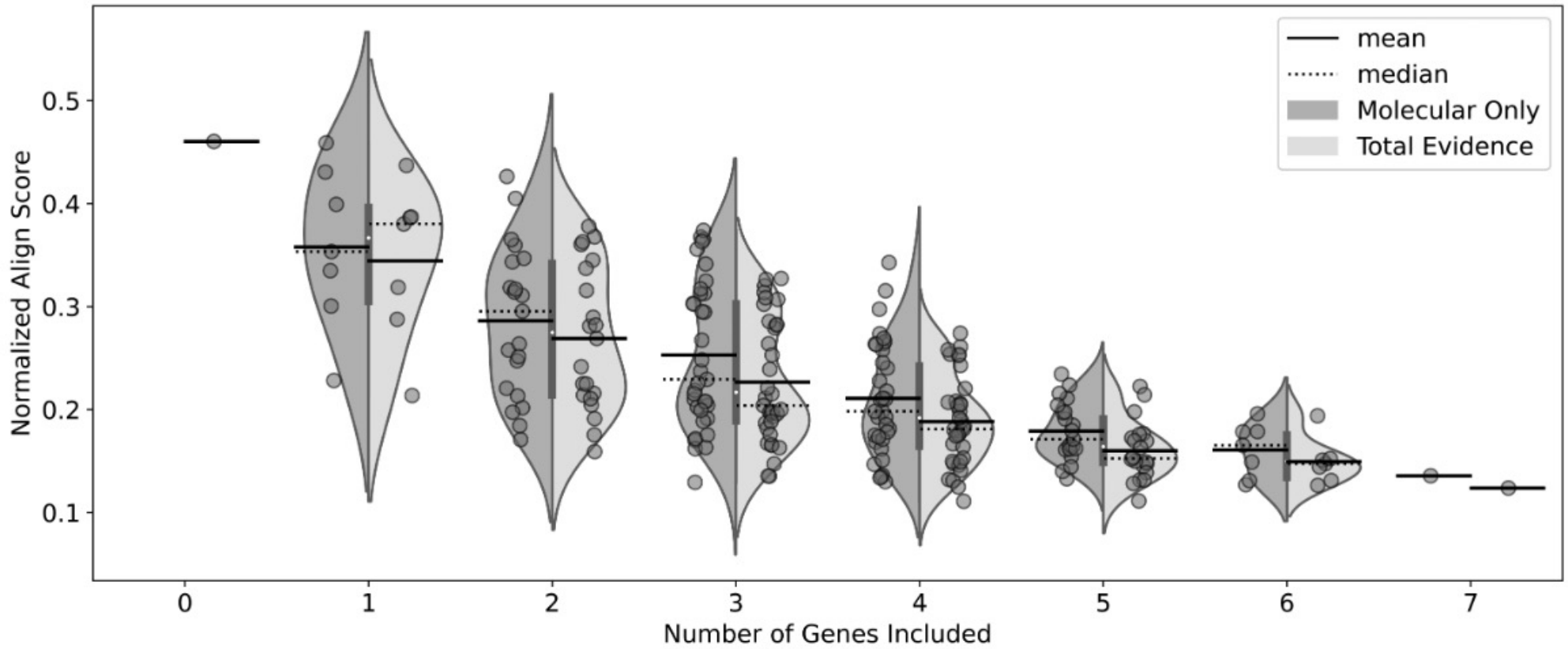


FIGURE 7. Combined Violin plot/Strip chart of the effect of including different gene combinations on the normalised Align Score. In each column the results to the left (darker) are for molecular-only ablations, and to the right (lighter) are for total evidence ablations. Each point represents an individual result.

TABLE 3. Results of total evidence analysis – combining molecular data and deep learning derived morphological traits. Confidence intervals are 95% level based on 5 runs. As a baseline comparison for the scores, five randomly generated trees of this size would give a nAS of 0.702 ± 0.022 and a nRF score of 0.993 ± 0.007. Best results in bold. Results within confidence interval of the best score are underlined

			nAS		nRF	
Traits	Dataset split	Loss function	Average	Median	Average	Median
Molecular Only	-	-	0.141 ± 0.017	0.147	0.382 ± 0.040	0.405
Total Evidence	Clade	Arcface	0.139 ± 0.021	0.138	0.337 ± 0.024	0.333
Total Evidence	Clade	Contrastive	$\mathbf{0.119 \pm 0.014}$	0.119	0.311 ± 0.033	0.315
Total Evidence	Clade	Margin	0.127 ± 0.017	0.127	0.319 ± 0.025	0.315
Total Evidence	Clade	Multisim	0.135 ± 0.024	0.136	0.316 ± 0.045	**0.306**
Total Evidence	Clade	Proxy	0.127 ± 0.013	0.126	0.324 ± 0.041	0.315
Total Evidence	Clade	Triplet	0.121 ± 0.023	**0.118**	$\mathbf{0.309 \pm 0.031}$	0.324
Total Evidence	Stratified	Arcface	0.166 ± 0.037	0.160	0.383 ± 0.028	0.389
Total Evidence	Stratified	Contrastive	0.141 ± 0.032	0.139	0.313 ± 0.019	**0.306**
Total Evidence	Stratified	Margin	0.162 ± 0.058	0.143	0.324 ± 0.011	0.324
Total Evidence	Stratified	Multisim	0.135 ± 0.016	0.133	0.337 ± 0.026	0.333
Total Evidence	Stratified	Proxy	0.155 ± 0.028	0.149	0.365 ± 0.021	0.361
Total Evidence	Stratified	Triplet	0.121 ± 0.012	0.119	0.310 ± 0.017	0.315

0.147 ± 0.011 0.128 ± 0.006 0.339 ± 0.012 0.320 ± 0.010

TABLE 4. Best subset of genes[c] given the number of genes. Best results in bold

	Molecular only			Total evidence		
No. genes	Best gene combination	nAS	nRF	Best gene combination	nAS	nRF
0	-	-	-	-	0.461	0.846
1	28S	0.228	0.433	28S	0.214	0.420
2	28S, ArgK	0.171	0.362	28S, ArgK	0.159	0.408
3	28S, COI, Wg	0.129	**0.275**	28S, ArgK, topo	0.135	0.286
4	28S, COI, topo, Wg	0.130	0.314	28S, cadB, COI, Wg	**0.111**	0.278
5	28S, ArgK, cadB, COI, Wg	0.133	0.432	28S, cadB, COI, topo, Wg	**0.111**	0.288
6	28S, ArgK, cadA, cadB, COI, topo	0.127	0.342	28S, ArgK, cadB, COI, topo, Wg	0.126	0.324
7	All	0.136	0.378	All	0.124	0.315

Genes used in this analysis: nuclear ribosomal 28S (28S), arginine kinase (ArgK), carbamoyl- phosphate synthetase (cadA and cadC), mitochondrial protein-encoding COI (COI) topoisomerase I (topo) and wingless (Wg)

Data Availability

The data underlying this article are available at http://doi.org/10.17894/ucph.39619bba-4569-4415-9f25-d6a0ff64f0e3 for the Rove-Tree-11 dataset and in the article's dryad repository (https://doi.org/10.5061/dryad.9cnp5hqqq) for the further molecular data and associated genbank accession numbers, example inference code, all generated trees, and stratified dataset split. All trained model runs and extracted trait matrices are available in the following erda repository https://erda.ku.dk/archives/440063cabdb1789ad82f31366c926b4e/published-archive.html. The reference tree, best molecular tree and best total-evidence tree can be found on TreeBASE at http://purl.org/phylo/treebase/phylows/study/TB2:S31300?x-access-code=397cc12bd8047bf52b312b4743f23e2b&format=html. The code used in this analysis is available on github https://github.com/robertahunt/Revisiting_Deep_Metric_Learning_PyTorch, commit a6654453c3b7785a17511255e02c468c53fe6f5d, forked from Roth et al. (2020).

Quantitative morphological characteristics, extracted through the application of deep learning techniques applied to images of pinned insect specimens produced for mass collections digitization purposes, have been demonstrated to possess phylogenetic relevance. These traits, when integrated into molecular phylogenies, have the potential to augment the phylogenetic framework in a comprehensive total-evidence based approach, offering the possibility of incorporating species lacking molecular data into phylogenetic trees. However, the improvement of phylogenetic reconstructions by the inclusion of such morphological data derived through deep learning methodologies remains minimal. While this approach has shown promise, scaling up its implementation is still not feasible for 2 reasons. First, the phylogenetic signal of the deep learning derived traits, at least in our dataset, was not strong enough to justify the additional effort in gathering the data. Second, the effort required for our image-based model testing, even though likely lesser than an effort by the expert to assemble a traditional morphological phylogenetic matrix, is still significant.

- “Homology”
- “Habitus”
- Cmeans vs Pagel lambda?
- Loss functions
- Other traits, like calls?
- How long would it take a skilled taxonomist in this field to take all the measurements and make a tree that way?
- Some say if you build a tree from morphological characters, you can't use that tree to test the evolution of those characters (because it would be circular). Is that true?
- Is it possible to incorporate continuous trait information in a maximum likelihood framework?
- How realistic are these morphotyping approaches for species with high plasticity?
- How do we scale deep-learning-based morphotyping to megaflora/megafauna that outscale standard imaging/camera resolutions?
- What are the assumptions of total evidence approach or what total evidence approach means?
- Wouldn't it be essential to have a taxonomist or group expert curating the data, such as collection names that will later be used by the deep learning, and also reviewing the iNaturalist names before using them for these analyses? or is not so relevant?
- Does this show more evidence of need to digitize collections?