

# Phylogenetic Transfer Learning

Michael D. Catchen <sup>1,2</sup>

<sup>1</sup> McGill University   <sup>2</sup> Québec Centre for Biodiversity Sciences

## Correspondance to:

Michael D. Catchen — [michael.catchen@mcgill.ca](mailto:michael.catchen@mcgill.ca)

This work is released by its authors under a CC-BY 4.0 license



Last revision: *May 6, 2023*

## 1 What is this paper?

2 The goal of this paper is to introduce PTL to a broad audience, extend the framework from being specific to  
3 latent representation of networks (Strydom *et al.* 2022a, b) to arbitrary traits, and provide guidelines for when  
4 PTL is appropriate and how to validate its prediction using both simulated and empirical data

## 5 What is Phylogenetic Transfer Learning

6 The goal of PTL is to take a species pool for which a given trait is only *partially* observed, and impute the value  
7 of that trait for the rest of the species based on the evolutionary relatedness of each species.

## 8 Ancestral State Reconstruction

9 *Ancestral State Reconstruction* (ASR; also called ancestral character reconstruction, character estimation, or  
10 character mapping) is a core topic in phylogenetics (Joy *et al.* 2016). The goal is, given an estimate of a  
11 phylogeny  $\mathcal{P}$  and trait values  $T_i$  for each extant species in the phylogeny  $i$ , to estimate the value of that trait at  
12 some historical point in the phylogeny (typically at the node representing the MRCA of a clade of interest).

13 To do this, one assumes a statistical model of evolution (and often tries to infer the best among a set of candidate  
14 models). Depending on whether the trait of interest is discrete or continuous, the models are typically  
15 discrete-space Markov chains (where the transition matrix is a target of inference) or, in the continuous setting,  
16 either Brownian Motion or an Ornstein-Uhlenbeck (OU) process (where the parameters of the model are the  
17 target of inference). The former is used to model neutral evolution and the latter for when there is hypothesized  
18 selection on the trait.

19 The methodology used to fit models has followed the historical progression from maximum parsimony models  
20 (a naive approach that favors as few evolutionary changes as possible in discrete traits) to maximum likelihood  
21 estimation. Modern ASR typically uses Bayesian methods, which has the direct benefit of including uncertainty  
22 estimates in inferred ancestral states, and potentially uncertainty in the true topology of the phylogeny itself  
23 (Huelsenbeck & Bollback 2001).

24 Typically rate of trait evolution is learned as a single value across the whole phylogeny, and branch length  
25 enables ‘amount’ of evolution. Although in principle a hierarchical\* model could be used to infer both a global  
26 rate of evolution and rate values specific to each branch (\*sadly ‘hierarchical’ in this is sense different than

27 ‘hierarchical’ as it is used in Huelsenbeck & Bollback (2001), which refers to different tree models but a single  
28 set of parameters across all branches—this is because these models are constructed in the context of DNA  
29 substitution rates, which are assumed to be fixed).

## 30 **Phylogenetic Transfer Learning**

31 The core goal of Phylogenetic Transfer Learning (PTL) is to take a phylogeny  $\mathcal{P}$  where the species pool consists  
32 of two types of species: (1) species with trait observations, which we denote  $\mathcal{O}$  and call the *observed* species  
33 and (2) species *without* trait observations, which we denote  $\mathcal{U}$  and call *unobserved*, and produce predicted trait  
34 values for the unobserved species  $\mathcal{U}$ .

35 PTL does this by using ASR to infer a parameterized model of evolution and ancestral state of the partially  
36 observed trait, and then to simulate the parameterized evolutionary model forward from the ancestral node to  
37 get an estimate of trait values (with uncertainty) for each species in  $\mathcal{U}$ .

38 The *transfer learning* component in particular comes from the first use of this method to impute latent  
39 representations of species based on their position in food-webs (Strydom *et al.* 2022a, b), although the method  
40 is flexible enough to be applied to either latent or observed traits.

41 There are two possible models for PTL to be done in: (1) As in (Strydom *et al.* 2022a), the evolutionary  
42 model is inferred only from the observed trait values  $\mathcal{O}$ . (2) The evolutionary model from a trait for which there  
43 are observations available for the entire species pool. It may be the case that evolutionary dynamics inferred  
44 with auxiliary “global” traits available for every species (e.g. the raw sequences from which the tree is built)  
45 could improve imputation accuracy.

46 [Figure 1 about here.]

## 47 **Substance of the paper**

48 The main substance of this paper is to provide guidelines on when PTL is robust, and diagnostics to validate  
49 PTL estimates. This will be done in two parts: **(1)** using simulated phylogenies and trait values to compare  
50 efficacy of PTL imputation across known “ground-truth” evolutionary dynamics, and **(2)** using published ASR  
51 datasets, with known values for all extant species, to test imputation efficacy empirically.

## 52 **Simulated phylogenies**

53 First goal is to test predictive efficacy of PTL under various parameterizations of the “ground-truth”  
54 evolutionary dynamics, e.g.

- 55 • Rate of speciation
- 56 • Rate of evolution
- 57 • Trait dimensionality & correlation
- 58 • Frequency of “punctuations” (i.e. perturbations to trait value at a speciation event)

59 and second to compare efficacy based on different properties of the data, e.g.

- 60 • Proportion of species with trait values
- 61 • Predictive efficacy vs. distance to MRCA w/ data
- 62 • Is there a set of traits for all species to infer evolutionary dynamics?
  - 63 – How correlated are evolution between traits for all species vs. traits we want to impute

## 64 **Real data**

65 Thankfully there are lots of ACR studies out there with data for each extant species. So, in short, find as many  
66 ACR studies as we can, and do crossvalidation where we drop the trait values for ~20% of the species and  
67 impute them with PTL, and compare.

## 68 **Questions to address**

- 69 • What phylogenetic scales can PTL give robust predictions?
  - 70 – There is a clear upper limit (if the MRCA for set of species is too long ago, there will be massive
  - 71 variance in predicted state at the tip)
  - 72 – There is also a lower limit (if there is horizontal-gene-transfer and the phylogeny is not a reliable
  - 73 proxy for how traits are evolving)
  - 74 – Robustness to noise in trait measurements?
  - 75 – How does this vary w/ amount and different types of selection

- When is PTL overkill?
- Weighted average of neighbors by distance as alternative (ack. David Rolnick for idea)
- What diagnostics can we use to be confident a PTL imputed trait is statistically robust?

## Possible Applications

The core idea of PTL is to fill in gaps for data-sparse processes in ecology, so naturally the applications are going to be focused.

### Link prediction in networks

This is the inciting question for which the idea was conceived (Strydom *et al.* 2022a). Interactions are hard to sample (Catchen2020MisLin?). Not much to say here that isn't in (Strydom *et al.* 2022b).

### Forecasting species range shifts

Many projections of species range shifts under climate change are based on statistical associations between historical species observations and climatic conditions. The gap between so-called correlative vs. mechanistic species distribution models (Shabani *et al.* 2016) is of critical interest for forecasting species ranges, but robust mechanistic understandings about what and why climatic conditions limit where a species' range require detailed sampling and potentially experimental conditions (Lee-Yaw *et al.* 2016), which are difficult to scale. PTL could potentially alleviate this by giving good proxies of the physiological limitations of species ranges for more species.

### Connectivity and movement ecology

Reliable information about species movement is sparse, and PTL could fill this gap (Catchen *et al.* 2023).

### Model species and ecological monitoring

There are a lot of species on Earth. Monitoring them all would be hard. Can we use single species as a proxy for a larger group of species? Maybe a little. PTL can guide us on what good proxy species are.

## References

Catchen, M.D., Lin, M., Poisot, T., Rolnick, D. & Gonzalez, A. (2023). Improving ecological connectivity

assessments with transfer learning and function approximation.

Huelsenbeck, J.P. & Bollback, J.P. (2001). [Empirical and Hierarchical Bayesian Estimation of Ancestral States](#). *Systematic Biology*, 50, 351–366.

Joy, J.B., Liang, R.H., McCloskey, R.M., Nguyen, T. & Poon, A.F.Y. (2016). [Ancestral Reconstruction](#). *PLOS Computational Biology*, 12, e1004763.

Lee-Yaw, J.A., Kharouba, H.M., Bontrager, M., Mahony, C., Csergő, A.M., Noreen, A.M.E., *et al.* (2016). [A synthesis of transplant experiments and ecological niche models suggests that range limits are often niche limits](#). *Ecology Letters*, 19, 710–722.

Shabani, F., Kumar, L. & Ahmadi, M. (2016). [A comparison of absolute performance of different correlative and mechanistic species distribution models in an independent area](#). *Ecology and Evolution*, 6, 5973–5986.

Strydom, T., Bouskila, S., Banville, F., Barros, C., Caron, D., Farrell, M.J., *et al.* (2022a). [Food web reconstruction through phylogenetic transfer of low-rank network representation](#). *Methods in Ecology and Evolution*, 13, 2838–2849.

Strydom, T., Bouskila, S., Banville, F., Barros, C., Caron, D., Farrell, M.J., *et al.* (2022b). Graph embedding and transfer learning can help predict potential species interaction networks despite data limitations.

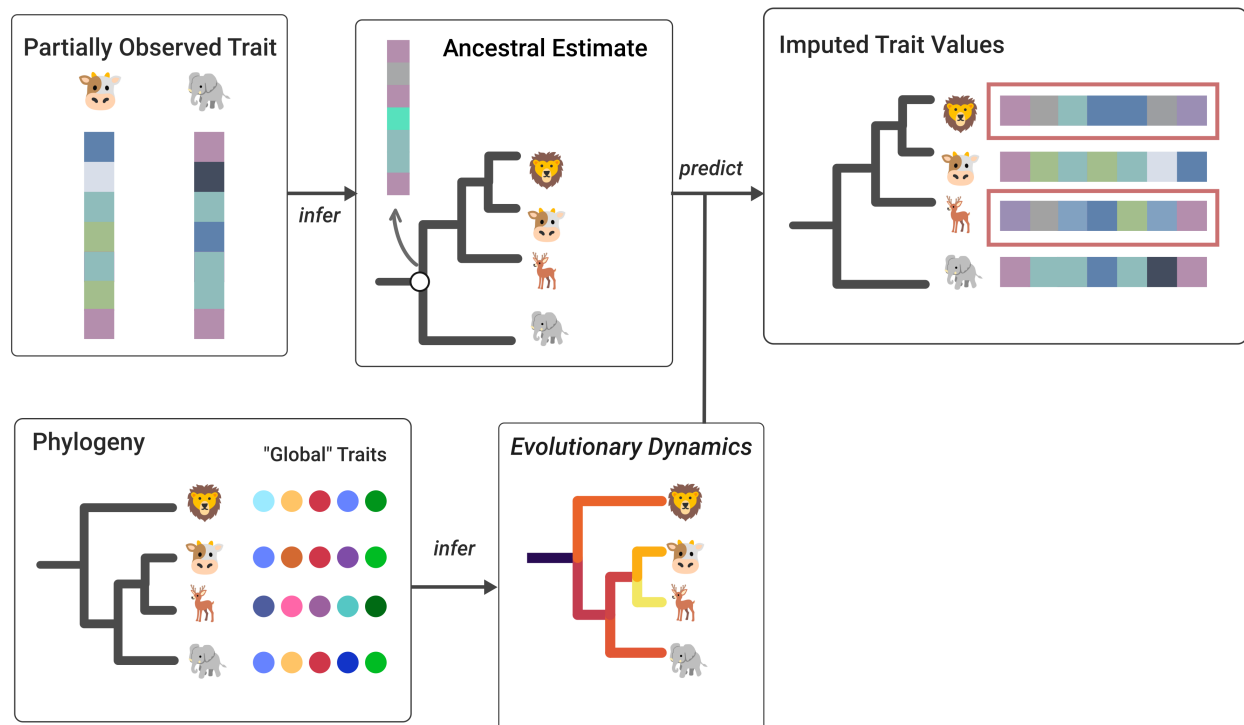


Figure 1: Conceptual visualization of PTL. Here, the trait of interest is partially observed only for cows and elephants. Using ASR, we can obtain an estimate of the ancestral state for the MRCA of the species considered here. Here we consider the second version of PTL described above, where there are “global” traits available for each species, which are used to infer the rate of evolutionary change. Using this parameterized model of evolutionary dynamics, we can then project forward to get imputed estimates of the trait value for the unobserved species, elk and lions.