

# Visualizing Geophylogenies – Internal and External Labeling with Phylogenetic Tree Constraints

Jonathan Klawitter  



University of Auckland, New Zealand

Felix Klesen 

Universität Würzburg, Germany

Joris Y. Scholl

Ruhr-Universität Bochum, Germany

Thomas C. van Dijk  

Ruhr-Universität Bochum, Germany

Alexander Zaft

Universität Würzburg, Germany

---

## Abstract

A *geophylogeny* is a phylogenetic tree where each leaf (biological taxon) has an associated geographic location (site). To clearly visualize a geophylogeny, the tree is typically represented as a crossing-free drawing next to a map. The correspondence between the taxa and the sites is either shown with matching labels on the map (internal labeling) or with *leaders* that connect each site to the corresponding leaf of the tree (external labeling). In both cases, a good order of the leaves is paramount for understanding the association between sites and taxa. We define several quality measures for internal labeling and give an efficient algorithm for optimizing them. In contrast, minimizing the number of leader crossings in an external labeling is NP-hard. We show nonetheless that optimal solutions can be found in a matter of seconds on realistic instances using integer linear programming. Finally, we provide several efficient heuristic algorithms and experimentally show them to be near optimal on real-world and synthetic instances.

**2012 ACM Subject Classification** Human-centered computing → Geographic visualization; Applied computing → Biological networks; Theory of computation → Discrete optimization

**Keywords and phrases** geophylogeny, boundary labeling, external labeling, algorithms

**Digital Object Identifier** 10.4230/LIPIcs.GIScience.2023.5

**Related Version** *Full Version*: <https://arxiv.org/abs/2306.17348> [16]

**Supplementary Material** *Software (Source Code)*: <https://www.github.com/joklawitter/geophylo>

**Funding** *Jonathan Klawitter*: Beyond Prediction Data Science Research Programme (MBIE grant UOAX1932).

*Thomas C. van Dijk*: DFG grant Di2161/2-1.

## 1 Introduction

A *phylogeny* describes the evolutionary history and relationships of a set of taxa such as species, populations, or individual organisms [25]. It is one of the main tasks in phylogenetics to infer a phylogeny for some given data and a particular model. Most often, a phylogeny is modelled and visualized with a *rooted binary phylogenetic tree*  $T$ , that is, a rooted binary tree  $T$  where the leaves are bijectively labeled with a set of  $n$  taxa. For example, the phylogenetic tree in Figure 1a shows the evolutionary species tree of the five present-day kiwi (*Apteryx*) species. The tree is conventionally drawn with all edges directed downwards to the leaves and



© Jonathan Klawitter, Felix Klesen, Joris Y. Scholl, Thomas C. van Dijk, and Alexander Zaft; licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

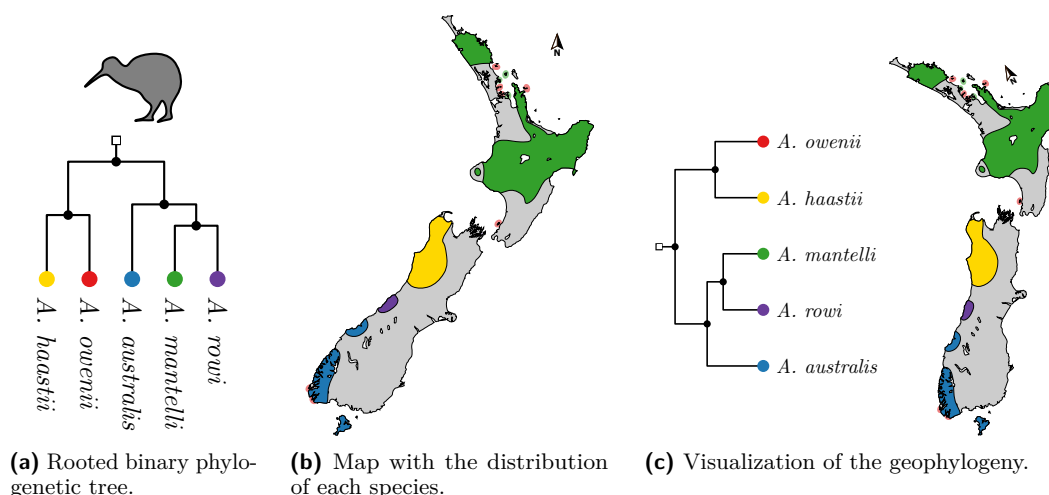
Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 5; pp. 5:1–5:16

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

## 5:2 Visualizing Geophylogenies



■ **Figure 1** To visualize Weir et al.’s geophylogeny of the five present-day kiwi species [27], we combine the phylogenetic tree (a) with the distribution map (b) into a single figure (c).

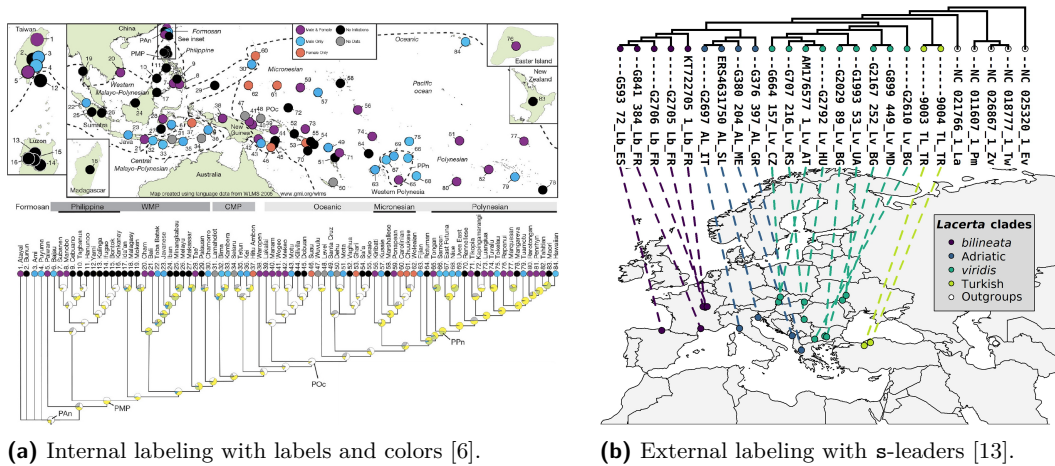
without crossings (*downward planar*). There exist several other models for phylogenies such as the more general phylogenetic networks and unrooted phylogenetic trees; here we only consider rooted binary phylogenetic trees and refer to them simply as phylogenetic trees.

In the field of phylogeography, geographic data is used in addition to genetic data. We may thus have spatial data associated with each taxon such as the distribution range of each species or the sampling site of each voucher specimen used in a phylogenetic analysis. For example, Figure 1b shows the distributions of the kiwi species from Figure 1a. We speak of a *geophylogeny* (or *phylogeographic tree*) if we have a phylogenetic tree  $T$ , a map region  $R$ , and a set  $P$  of features on  $R$  that correspond one-to-one with the taxa in  $T$ ; see Figure 1c for a geophylogeny of the kiwi species. In this paper, we focus on the case where  $P$  is a set of points, called *sites*.

### Visualizing Geophylogenies

When visualizing a geophylogeny, we may want to display its tree and its map together in order to show the connections (or the non-connections) between the leaves and the sites. For example, we may want to show that the taxa of a certain subtree are confined to a particular region of the map or that they are widely scattered. In the literature, we mainly find three types of drawings of geophylogenies. In a *side-by-side* drawing, the tree is drawn planar directly next to the map. To show the correspondences between the taxa and their sites, the sites are either labeled or color coded (as in Figure 2a and Figure 1c, respectively), or the sites are connected with *leaders* to the leaves of the tree (as in Figure 2b). We call this *internal labeling* and *external labeling*, respectively. There also exist *overlay* illustrations where the phylogenetic tree is drawn onto the map in 2D or 3D with the leaves positioned at the sites [15, 29], but for brevity we omit further discussion of this style.

Drawing a geophylogeny involves various subtasks, such as choosing an orientation for the map, a position for the tree, and the placement of the labels. Several existing tools support drawing geophylogenies but we suspect that in practice many drawings are made “by hand”. The tools **GenGIS** by Parks et al. [23, 22], a tool by Page [20], and the R-package **phytools** by Revell [24] can generate side-by-side drawings with external labeling. The former two try to minimize leader crossings by testing random leaf orders and by rotating



■ **Figure 2** Side-by-side drawings of geophylogenies from the literature.

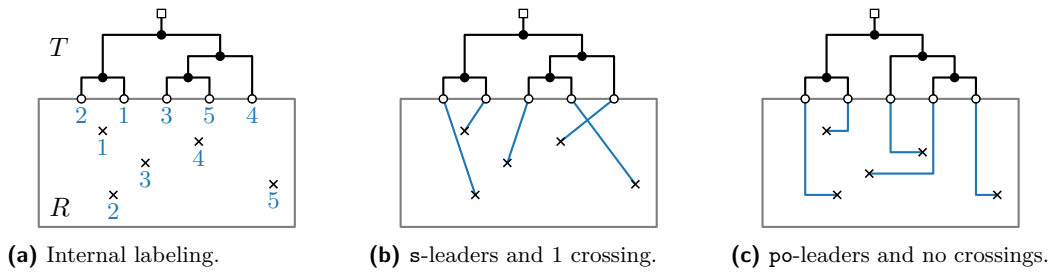
the phylogenetic tree around the map; Revell uses a greedy algorithm to minimize leader crossings. The R package `phylogeo` by Charlop-Powers and Brady [8] uses internal labeling via colors. Unfortunately, none of the articles describing these tools formally defines a quality measure being optimized or studies the underlying combinatorial optimization problem from an algorithmic perspective. In this paper, we introduce a simple combinatorial definition for side-by-side drawings of geophylogenies and propose several quality measures.

### Labeling Geophylogenies

Following standard map-labeling terminology, *internal labeling* places the labels inside or in the direct vicinity of a feature; *external labeling* [5] places the labels in the margin next to the map and a label is connected to the corresponding feature with a *leader*. An *s-leader* is drawn using a single (straight) line segment as in Figures 2b and 3b. Alternatively, a *po-leader* (for: parallel, orthogonal) consists of a horizontal segment at the site and a vertical segment at the leaf; see Figure 3c. In the literature, we have only encountered *s-leaders* in geophylogeny drawings, but argue below that *po-leaders* should be considered. In a user study on external labeling, Barth, Gemsa, Niedermann, and Nöllenburg [1] showed that *s-leaders* perform well when users are asked to associate sites with their labels and vice versa, but that *po-leaders* (and “diagonal, orthogonal” *do-leaders*) are among the aesthetic preferences.

For internal labeling, a common optimization approach is to place the most labels possible such that none overlap; see Neyer [18] for a survey on this topic. Existing algorithms can be applied to label the sites in a geophylogeny drawing and it is geometrically straight-forward to place the labels for the leaves of  $T$ . However, a map reader must also be aided in associating the sites on the map with the leaves at the border based on these labels (and potentially colors). Consider the drawing in Figure 1c, which uses color-based internal labeling: the three kiwi species *A. australis*, *A. rowi*, and *A. mantelli* occur in this order from South to North. When using internal labeling, we would thus prefer, if possible, to have the three species in this order in the tree as well – as opposed to their order in Figure 1a.

External labeling styles conventionally forbid crossing the leaders as such crossings could be visually confusing (cf. Figure 2b). Often the total length of leaders is minimized given this constraint; see the survey by Bekos, Niedermann, and Nöllenburg [5]. If one allows a many-to-one correspondence between sites and labels, the literature typically seeks a drawing



■ **Figure 3** We place  $T$  above  $R$  and use either internal or external labeling to show the mapping between  $P$  and  $L(T)$ . Figures (b) and (c) minimize the number of crossings for their leader type. Note the difference in embedding of  $T$  and that not all permutations of leaves are possible.

that minimizes the number of crossings between the leaders, and this is NP-hard [17]. The problem remains NP-hard even when leaders can share segments, so-called hyper-leaders [2]. Even though our drawings of geophylogenies have a one-to-one correspondence, the planarity constraint on the tree restricts which leaf orders are possible and it is not always possible to have crossing-free leaders in a geophylogeny. In order to obtain a drawing with low visual complexity, our task is thus to find a leaf order that minimizes the number of leader crossings.

## Results and Contribution

We formalize several graph visualization problems in the context of drawing geophylogenies. We propose quality measures for drawings with internal labeling and show that optimal solutions can be computed in quadratic time (Section 3). For external labeling (Section 4), we prove that although crossing minimization of s- and po-leaders is NP-hard in general, it is possible to check in polynomial time if a crossing-free drawing exists and to solve a certain class of instances efficiently in practice. Furthermore, we introduce an integer linear program (ILP) and several heuristics for crossing minimisation. We evaluate these solutions on synthetic and real-world examples and find that the ILP can solve realistic instances optimally in a matter of seconds and that the heuristics, which run in a fraction of a second, are often (near-)optimal as well (Section 5). We close the paper with a discussion and open problems; in particular, we point out further similarities between problems with geophylogeny drawings and with external labeling.

A longer version of this paper containing all proofs is available on arXiv [16]. Furthermore, implementations of the algorithms and the experiments are available online at [github.com/joklawitter/geophylo](https://github.com/joklawitter/geophylo).

## 2 Definitions and Notation

For a phylogenetic tree  $T$ , let  $V(T)$  be its vertex set,  $E(T)$  its edge set,  $L(T)$  its leaves, and  $I(T)$  its internal vertices. As size of an instance we let  $n = |L(T)|$  be the number of leaves. Let  $T(v)$  be the subtree rooted at  $v$  and  $n(v) = |L(T(v))|$ .

A *map*  $R$  is an axis-aligned rectangle and a *site* is a point on  $R$ . A *geophylogeny*  $G$  consists of a phylogenetic tree  $T$ , a map  $R$ , a set of points  $P$  on  $R$ , as well as a 1-to-1 mapping between  $L(T) = \{\ell_1, \dots, \ell_n\}$  and  $P = \{p_1, \dots, p_n\}$  so that without loss of generality the mapping is given by the indices.

We define a *drawing*  $\Gamma$  of  $G$  as consisting of drawings of  $R$  and  $T$  in the plane with the following properties (see Figure 3). We assume that  $T$  is always drawn at a fixed position above  $R$  such that the leaves of  $T$  lie at evenly spaced *positions* on the upper boundary

of  $R$ . Furthermore, we require that  $T$  is drawn *downward planar*, that is, all edges of  $T$  point downwards from the root towards the leaves, and no edges of  $T$  cross. (In our examples we draw  $T$  as a “rectangular cladogram”, but the exact drawing style is irrelevant given downward planarity.) The points of  $P$  are marked on  $R$  and the drawing uses either internal labeling as in Figure 3a or external labeling with s- or po-leaders as in Figures 3b and 3c. For drawings with external labeling, we use  $s_i$  to denote the leader that connects  $\ell_i$  and  $p_i$ .

Since the tree is drawn without crossings and the sites have fixed locations, the only combinatorial freedom in the drawing  $\Gamma$  is the embedding of  $T$ , i.e. which child is to the left and which is to the right. Furthermore, since we fixed the relative positions of the map and the leaves, note that there is also no “non-combinatorial” freedom. Hence, an embedding of  $T$  corresponds one-to-one with a left-to-right order of  $L(T)$  and we call this the *leaf order*  $\pi$  of  $\Gamma$ . For example, if a leaf  $\ell_i$  is at position 4 in  $\Gamma$ , then  $\pi(\ell_i) = 4$ . Further, let  $x(v)$  denote the x-coordinate of a site or leaf  $v$  of  $T$  in  $\Gamma$ .

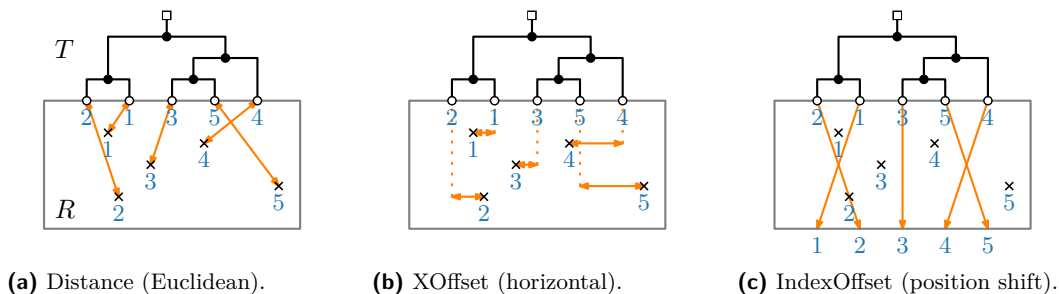
### 3 Geophylogenies with Internal Labeling

A good order of the leaves is crucial for internal labeling, since it can help the reader associate between  $L(T)$  and  $P$ . It is in general not obvious how to determine which leaf order is best for this purpose; we propose three quality measures and a general class of measures that subsume them. Any measure in this class can be efficiently optimized by the algorithm described below. In practice one can easily try several quality measures and pick whichever suits the particular drawing; a user study of practical readability could also be fruitful.

#### 3.1 Quality Measures

When visually searching for the site  $p_i$  corresponding to a leaf  $\ell_i$  (or the opposite direction), it seems beneficial if  $\ell_i$  and  $p_i$  are close together. Our first quality measure, *Distance*, sums the Euclidean distances over all pairs  $(p_i, \ell_i)$ . Since the tree organizes the leaves from left to right, it might be better to consider only the horizontal distances, i.e.  $\sum_{i=1}^n |x(p_i) - x(\ell_i)|$ , which we call *XOffset*. Finally, instead of the geometric offset, *IndexOffset* considers how much the leaf order permutes the geographic left-to-right order of the sites. Assuming without loss of generality that the sites are indexed from left to right, we sum how many places each leaf  $\ell_i$  is away from leaf position  $i$ , i.e.  $\sum_{i=1}^n |\pi(\ell_i) - i|$ . See Figure 4.

These measures have in common that they sum over some “quality” of the leaves, where the quality of a leaf depends only on its own position and that of the sites (but not the other leaves). We call such quality measures *leaf additive*. Unfortunately not all sensible quality measures are leaf additive (such as for example the number of inversions in  $\pi$ ).



■ **Figure 4** Orange arrows indicate what the three quality measures for internal labeling consider.

### 3.2 Algorithm for Leaf-Additive Quality Measures

Let  $f: L(T) \times \{1, \dots, n\} \rightarrow \mathbb{R}$  be a quality measure for placing one particular leaf at a particular position; the location of the sites is constant for a given instance, so we do not consider it an argument of  $f$ . This uniquely defines a leaf additive objective function on drawings by summing over the leaves; assume w.l.o.g. that we want to minimize this sum.

Now we naturally lift  $f$  to inner vertices of  $T$  by taking the sum over leaves in the subtree rooted at that vertex – in the best embedding of that subtree. More concretely, note that any drawing places the leaves of any subtree at consecutive positions and they take up a fixed width regardless of the embedding. Let  $F(v, i)$  be the minimum, taken over all embeddings of  $T(v)$  and assuming the leftmost leaf is placed at position  $i$ , of the sum of quality of the leaves of  $T(v)$ . Then by definition the optimal objective value for the entire instance is  $F(w, 1)$ , where  $w$  is the root of  $T$ .

► **Theorem 1.** *Let  $G$  be a geophylogeny on  $n$  taxa and let  $f$  be a leaf additive objective function. A drawing that minimizes (or maximizes)  $f$  can be computed in  $\mathcal{O}(n^2)$  time.*

**Proof.** For a vertex  $v$  with children  $x$  and  $y$ , we observe the following equality, since the embedding has only two ways of ordering them and those subtrees are then independent.

$$F(v, i) = \min\{ F(x, i) + F(y, i + n(x)), F(y, i) + F(x, i + n(y)) \} \quad (1)$$

Using dynamic programming on  $F$  allows us to calculate  $F(w, 1)$  in  $\mathcal{O}(n^2)$  time and space, since there are  $2n$  vertices,  $n$  possible leaf positions, and Equation (1) can be evaluated in constant time by precomputing all  $n(v)$ . The optimal embedding of  $T$  can be traced back through the dynamic programming table in the same runtime. ◀

Note that we can still define leaf additive quality measures when  $P$  contains regions (rather than just points) as in Figure 1. For example, instead of considering the distance between  $\ell_i$  and  $p_i$ , we could consider the smallest distance between  $\ell_i$  and any point in the region  $p_i$ .

With the above algorithm, we can restrict leaves and subtrees to be in a certain position or a range of positions, simply by marking all other positions as prohibitively expensive in  $F$ ; the rotation of an inner vertex can also be fixed by considering only the corresponding term of Equation (1). This can be used if there is a conventional order for some taxa or to ensure that an outgroup-taxon is placed at the leftmost or rightmost position. Furthermore, this enables an interactive editing experience where a designer can inspect the initial optimized drawing and receive re-optimized versions based on their feedback – for example “put the leaves for the sea lions only where there is water on the edge of the map”. (This is leaf additive.)

## 4 Geophylogenies with External Labeling

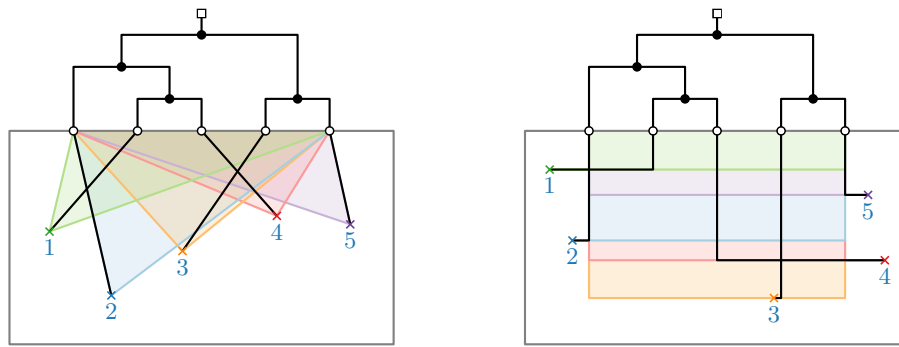
For external labeling, the optimization goal is to embed the tree such that the number of crossings between leaders is minimized. Unless otherwise stated, we use **s**-leaders.

For brevity, we omit proofs<sup>1</sup> of following foundational complexity results and move on to more practical algorithms.

---

<sup>1</sup> Proofs are available in the long version of this paper [16].





(a) A geometry-free instance for **s**-leaders: no site lies inside the **s**-area of another site.

(b) A geometry-free instance for **po**-leaders: no site lies inside the **po**-area of another site.

■ **Figure 5** In a geometry-free instance the leaf order  $\pi$  fully determines if any two leaders cross.

► **Proposition 2.** *Given a geophylogeny  $G$  and an integer  $k$ , it is NP-hard to decide, for both **s**- and **po**-leaders, if  $G$  admits a drawing with external labels and at most  $k$  leader crossings.*

► **Proposition 3.** *Given a geophylogeny  $G$  on  $n$  taxa, it can be decided in  $\mathcal{O}(n^6)$  time, for both **s**- and **po**-leaders, whether  $G$  admits a drawing with external labels and no leader crossings.*

### 4.1 Geometric Structure and Geometry-Free Instances

We start by making some observations about the structure of geophylogeny drawings. This leads to an  $\mathcal{O}(n \log n)$ -time algorithm for crossing minimization on a particular class of “geometry-free” instances and forms the basis for our ILP.

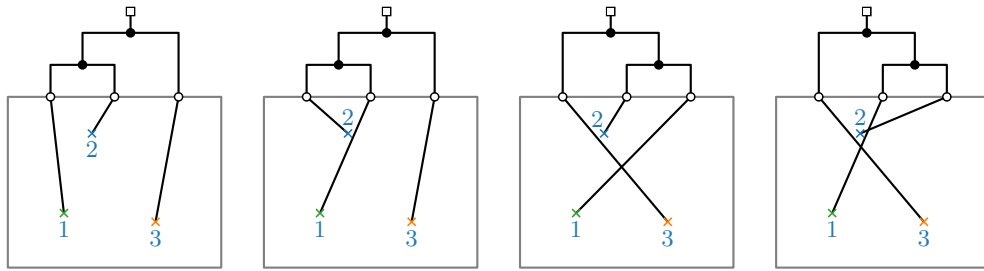
Let  $B$  be the line segment between leaf position 1 (left) and leaf position  $n$  (right); let the **s**-area of a site  $p_i$  be the triangle spanned by  $p_i$  and  $B$ . Note that the leader  $s_i$  lies within this triangle in any drawing. Now consider two sites  $p_i$  and  $p_j$  that lie outside each other’s **s**-area. Independently of the embedding of the tree,  $s_i$  always passes  $p_j$  on the same side: see Figure 5 where, for example,  $s_2$  passes left of  $p_4$  in any drawing. As a result, if  $p_i$  lies left of  $p_j$ , then  $s_i$  and  $s_j$  cross if and only if the leaf  $\ell_i$  is positioned right of the leaf  $\ell_j$  (cf. Figure 5). The case where  $p_i$  is right of  $p_j$  is flipped. We call such a pair  $(p_i, p_j)$  *geometry free* since purely the *order* of the corresponding leaves suffices to recognize if their leaders cross: the precise geometry of the leaf positions is irrelevant.

Conversely, consider a site  $p_k$  that lies inside the **s**-area of  $p_i$ . Whether the leaders  $s_i$  and  $s_k$  cross depends on the placement of the leaves  $\ell_i$  and  $\ell_k$  in a more complicated way than just their relative order:  $s_i$  might pass left or right of  $p_k$ . In this case, we call  $p_i$  *undecided* with respect to  $p_k$ . See Figure 6, where  $p_1$  is undecided with respect to  $p_2$ .

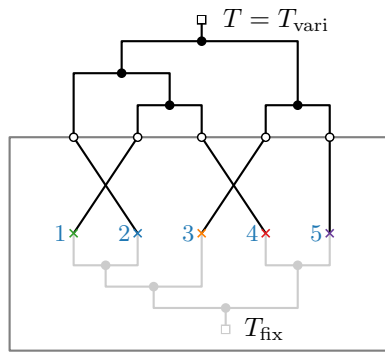
We call a geophylogeny *geometry free* if all pairs of sites are geometry free. Such instances are not entirely implausible: for example, researchers may have taken their samples along a coastline, a river, or a valley, in which case the sites may lie relatively close to a line. Orienting the map such that this line is horizontal could result in a geometry-free instance. Furthermore, unless two sites share an x-coordinate, increasing the vertical distance between the map and the tree eventually results in a geometry-free drawing for **s**-leaders; however, the required distance might be impractically large.

Concerning **po**-leaders, we can analogously define the **po**-area of a site (see Figure 5b).

► **Theorem 4.** *Given a geometry-free geophylogeny  $G$  on  $n$  taxa, a drawing with the minimum number of leader crossings can be found in  $\mathcal{O}(n \log n)$  time, for both **s**- and **po**-leaders.*



■ **Figure 6** Drawings of the same geophylogeny with different leaf orders. Whether  $s_1$  and  $s_2$  cross depends on the position of  $\ell_1$  and  $\ell_2$ , whereas  $s_1$  and  $s_3$  cross if and only if  $\ell_3$  is left of  $\ell_1$ . We call the pair  $(p_1, p_2)$  *undecided* and the pair  $(p_1, p_3)$  *geometry-free*.



■ **Figure 7** A geometry-free geophylogeny and a one-sided tanglegram  $(T_{\text{fix}}, T_{\text{vari}})$  that have the same combinatorics (in terms of leader crossings) as the two geometry-free instances in Figure 5.

**Proof.** We transform  $G$  into a so-called *one-sided tanglegram*  $(T_{\text{fix}}, T_{\text{vari}})$  that is equivalent in terms of crossings; see Figure 7. In a *tanglegram* [11] two phylogenetic trees on the same taxa are drawn planar opposite each other and the matching taxa are connected with straight line segments; the goal is to find leaf orders that minimize the number of crossings. In a one-sided tanglegram, the leaf order for one tree is given and fixed.

We take the sites  $P$  as the leaves of  $T_{\text{fix}}$  and embed the tree so that the points are ordered from left to right; the topology of  $T_{\text{fix}}$  is arbitrary. As the tree  $T_{\text{vari}}$  with variable embedding, we take the phylogenetic tree  $T$ . Since  $G$  is geometry-free, the crossings in the tanglegram correspond one-to-one with those in the geophylogeny drawing with the same embedding.

The number of crossings of  $(T_{\text{fix}}, T_{\text{vari}})$  can be minimized in  $\mathcal{O}(n \log n)$  time using an algorithm of Fernau et al. [11]: the resulting leaf order for  $T_{\text{vari}}$  then also minimizes the number of leader crossings in  $\Gamma$ . ◀

## 4.2 Optimal Drawings with Integer Linear Programming

For the following ILP, we consider an arbitrary embedding of the tree as *neutral* and describe all embeddings in terms of which internal vertices of  $T$  are rotated with respect to this neutral embedding, i.e. for which internal vertices to swap the left-to-right order of their two children. For two sites  $p_i$  and  $p_j$ , we use  $p_i \prec p_j$  to denote that  $\ell_i$  is left of  $\ell_j$  in the neutral embedding. Let  $U$  be the set of undecided pairs, that is, all ordered pairs  $(p, q)$  where  $q$  lies inside the  $\mathbf{s}$ -area of  $p$ ; note that these are ordered pairs.



### Variables and Objective Function

$\rho_i \in \{0, 1\} \forall i \in I(T)$ . Do we rotate internal vertex  $i$  (1) or keep its neutral embedding (0)?

Note that rotating the lowest common ancestor of  $\ell_i$  and  $\ell_j$  is the only way to flip their order, so for convenience we write  $\rho_{ij}$  to mean  $\rho_{\text{lca}(i,j)}$ .

$d_{pq} \in \{0, 1\} \forall (p, q) \in U$ . For each undecided pair  $(p, q)$ : should  $p$ 's leader pass to the left (0) or to the right (1) of site  $q$ ? (This is well-defined since the pair is undecided.)

$\chi_{pq} \in \{0, 1\} \forall p, q \in P, p < q$ . For each set of two sites: are the leaders of  $p$  and  $q$  *allowed* to cross? There is no requirement that noncrossing pairs have  $\chi_{pq} = 0$ , but that will be the case in an optimal solution.

To minimise the number of crossings, minimize the sum over all  $\chi_{pq}$ .

### Constraints

We handle geometry-free pairs and undecided pairs separately.

Consider a geometry-free pair of sites: if the leaders cross in the neutral embedding, we must either allow this, or rotate the lowest common ancestor. Conversely, if they do not cross neutrally, yet we rotate the lowest common ancestor, then we must allow their leaders to cross. Call these sets of pairs  $F_{\text{rotate}}$  and  $F_{\text{keep}}$  respectively, for how to prevent the crossing.

$$\chi_{ij} + \rho_{ij} \geq 1 \quad \forall (i, j) \in F_{\text{rotate}}; \quad \chi_{ij} - \rho_{ij} \geq 0 \quad \forall (i, j) \in F_{\text{keep}} \quad (2)$$

For undecided pairs  $(p, q)$ , a three-way case distinction on  $[p \prec q]$ ,  $\rho_{pq}$  and  $d_{pq}$  reveals the following geometry: pairs with  $p \prec q$  have crossing leaders if and only if  $\rho_{pq} + d_{pq} = 1$ ; pairs with  $p \succ q$  have crossing leaders if and only if  $\rho_{pq} + d_{pq} \neq 1$ . Recall that we do not force  $\chi$  to be zero if there is no intersection, only that it is 1 if there *is* an intersection; we implement these conditions in the ILP as follows. Let  $U_{\text{left}} \subseteq U$  be the undecided pairs with  $p \prec q$ .

$$\rho_{pq} - d_{pq} \leq \chi_{pq} \quad \forall (p, q) \in U_{\text{left}}; \quad d_{pq} - \rho_{pq} \leq \chi_{pq} \quad \forall (p, q) \in U_{\text{left}} \quad (3)$$

Conversely, let  $U_{\text{right}} \subseteq U$  be the undecided pairs with  $p \succ q$ .

$$\rho_{pq} + d_{pq} - 1 \leq \chi_{pq} \quad \forall (p, q) \in U_{\text{right}}; \quad 1 - \rho_{pq} - d_{pq} \leq \chi_{pq} \quad \forall (p, q) \in U_{\text{right}} \quad (4)$$

Finally, we must ensure that each leader  $s_i$  respects the  $d$  variables: the  $\mathbf{s}$ -leader from  $p_i$  to  $\ell_i$  must pass by each other site in the  $\mathbf{s}$ -area on the correct side. This does not affect geometry-free pairs, but we must constrain the leaf placement for undecided pairs.

Observe that the  $\rho$  variables together fix the leaf order, since they fix the embedding of  $T$ . Let  $L_i(\rho)$  be the function that gives the x-coordinate of  $\ell_i$  given the  $\rho$  variables. Note that  $L_i$  is linear in each of the  $\rho$  variables: rotating an ancestor of  $\ell_i$  shifts its leaf location by a particular constant, and rotating a non-ancestor does not affect it.

For an undecided pair  $(p_i, p_j)$ , let  $x^*(i, j)$  be the x-coordinate of where the ray from  $p_i$  through  $p_j$  intersects the top of the map and note that this is a constant. If  $d_{ij} = 0$ , then  $\ell_i$  must be to the left of this intersection; if  $d_{ij} = 1$ , it must be to the right. We model this in the ILP with two constraints and the *big-M method*, where we can set  $M = n$ .

$$L_i(\rho) - d_{ij}M \leq x^*(i, j), \quad L_i(\rho) + (1 - d_{ij})M \geq x^*(i, j); \quad \forall (p_i, p_j) \in U \quad (5)$$

The number of variables and constraints in the ILP are both quadratic in  $n$ .

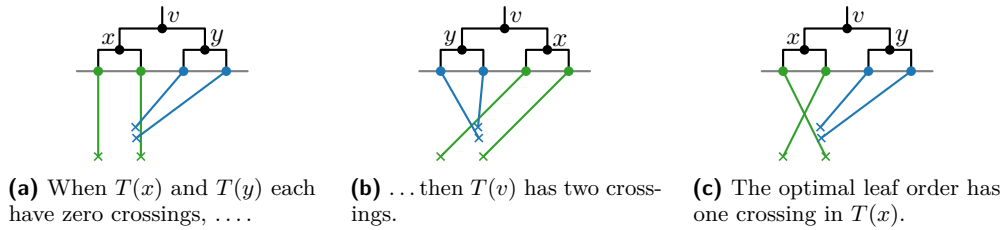
### 4.3 Heuristics

Since the ILP from the previous section can be slow in the worst case and requires advanced solver software, we now suggest a number of heuristics.

**Bottom-Up.** First, we use a dynamic program similar to the one in Section 3 and commit to an embedding for each subtree while going up the tree. At this point we note that counting the number of crossings is not a leaf additive objective function in the sense of Section 3. However, Equation (1) does enable us to introduce an additional cost based on where an entire subtree is placed and where its sibling subtree is placed – just not minimized over the embedding of these subtrees. More precisely, for an inner vertex  $v$  of  $T$  with children  $x$  and  $y$ , let  $C(x, y, i)$  be the number of crossings between  $T(x)$  and  $T(y)$  when placed starting at position  $i$  and  $i + n(x)$  respectively; this can be computed in  $\mathcal{O}(n(v)^2)$  time. Note that this ignores any crossings with leaders from other subtrees. With base case  $H(\ell, i) = 0$  for every leaf  $\ell$ , we use

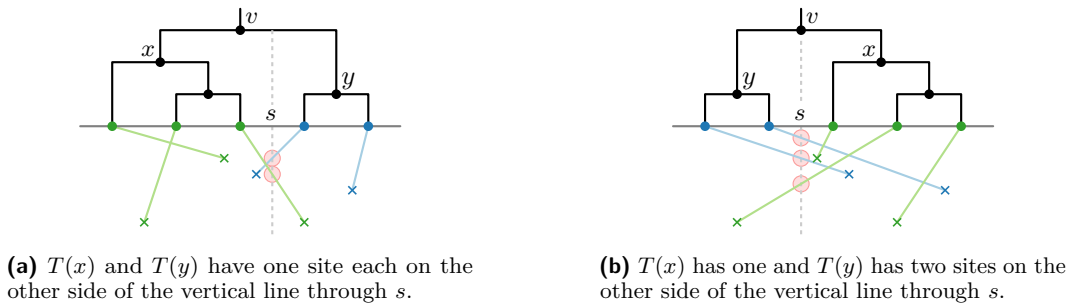
$$H(v, i) = \min\{ H(x, i) + H(y, i + n(x)) + C(x, y, i), H(y, i) + H(x, i + n(y)) + C(y, x, i) \}$$

to pick a rotation of  $T(v)$ . Since this can be evaluated in  $\mathcal{O}(n^2)$  time, the heuristic runs in  $\mathcal{O}(n^4)$  time. In the example in Figure 8 this does not minimize the total number of crossings.



■ **Figure 8** The bottom-up heuristic is not always optimal.

**Top-Down.** The second heuristic traverses  $T$  from top to bottom (i.e. in pre-order) and chooses a rotation for each inner vertex  $v$  based on how many leaders would cross the vertical line between the two subtrees of  $v$ ; see Figure 9. More precisely, suppose that  $T(v)$  has its leftmost leaf at position  $i$  based on the rotations of the vertices above  $v$ . For  $x$  and  $y$  the children of  $v$ , consider the rotation of  $v$  where  $T(x)$  is placed starting at position  $i$  and  $T(y)$  is placed starting at position  $i + n(x)$ . Let  $s$  be the x-coordinate in the middle between the last leaf of  $T(x)$  and the first leaf  $T(y)$ . We compute the number of leaders of  $T(v)$  that cross the vertical line at  $s$  and for the reverse rotation of  $v$ ; the smaller result is chosen and the rotation fixed. This procedure considers each site at most  $\mathcal{O}(n)$  times and thus runs in  $\mathcal{O}(n^2)$  time.



■ **Figure 9** The top-down heuristic tries both rotations of  $v$  and here would pick (a).

**Leaf-Additive Dynamic Programming.** Thirdly, we could optimize any of the quality measures for interior labeling (Section 3). These measures produce generally sensible leaf orders in quadratic time and we may expect the number of leader crossings to be low.

**Greedy (Hill Climbing).** Finally, we consider a hill climbing algorithm that, starting from some leaf order, greedily performs rotations that improve the number of crossings. This could start from a random leaf order, a hand-made one, or from any of the other heuristics. Evaluating a rotation can be done in  $\mathcal{O}(n^2)$  time and thus one round through all vertices runs in  $\mathcal{O}(n^3)$  time.

## 5 Experimental Evaluation

This section is based on our implementation of the ILP and the heuristics. The code is available online at [github.com/joklawitter/geophylo](https://github.com/joklawitter/geophylo), and data from the corresponding authors upon request.

### 5.1 Test Data

We use three procedures to generate random instances. For each type and with 10 to 100 taxa (in increments of 5), we generated 10 instances; we call these the *synthetic instances*. We stop at 100 since geophylogeny drawings with more taxa are rarely well-readable.

**Uniform.** Place  $n$  sites on the map uniformly at random. Generate the phylogenetic tree by repeating this merging procedure. Pick an unmerged site or a merged subtree uniformly at random, then pick a second with probability distributed by inverse distance to the first, and merge them; as position of a subtree, we take the median coordinate on both axis.

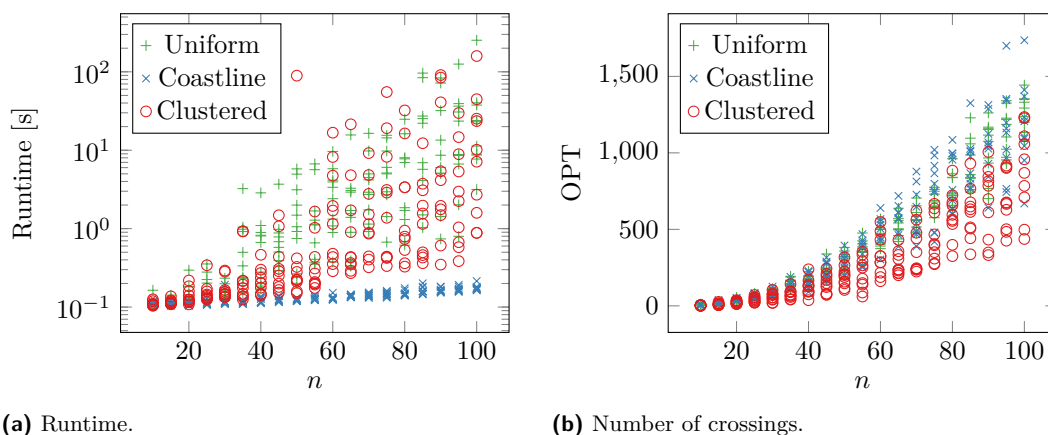
**Coastline.** Initially place all sites equidistantly on a horizontal line, then slightly perturb the x-coordinates. Next, starting at the central site and going outwards, change the y-coordinate of each site randomly (up to 1.5 times the horizontal distance) from the y-coordinate of the previous site. Construct the tree as before.

**Clustered.** These instances group multiple taxa into clusters. First a uniformly random number of sites between three and ten is allocated for a cluster and its center is placed at a uniformly random point on the map. Then for each cluster, we place sites randomly in a disk around the center with size proportional to the cluster size. Construct  $T$  as before, but first for each cluster separately and only then for the whole instance.

In addition, we consider three real world instances derived from published drawings. **Fish** is a 14-taxon geophylogeny by Williams and Johnson [28]. **Lizards** is 20-taxon geophylogeny by Jauss et al. [13], where the sites are mostly horizontally dispersed (see Figure 2b). **Frogs** is a 64-taxon geophylogeny by Ellepola et al. [10], where the sites are rather randomly dispersed on the map; the published drawing with s-leaders has over 680 leader crossings.

### 5.2 Experimental Results

**The ILP is fairly quick.** Our implementation uses a Python script to generate the ILP instance and Gurobi 10 to solve it; we ran the experiments on a 10-core Apple M1 Max processor. As expected, we observe that the runtime is exponential in  $n$ , but only moderately so (Figure 10). Instances with up to about 50 taxa can usually be solved optimally within a second, but for Clustered and Uniform instances the ILP starts to get slow at about 100 taxa. We note that geophylogenies with over 100 taxa should probably not be drawn with external labeling: for example, the Frogs instance can be drawn optimally by the ILP in



■ **Figure 10** Computing optimal drawings with the ILP.

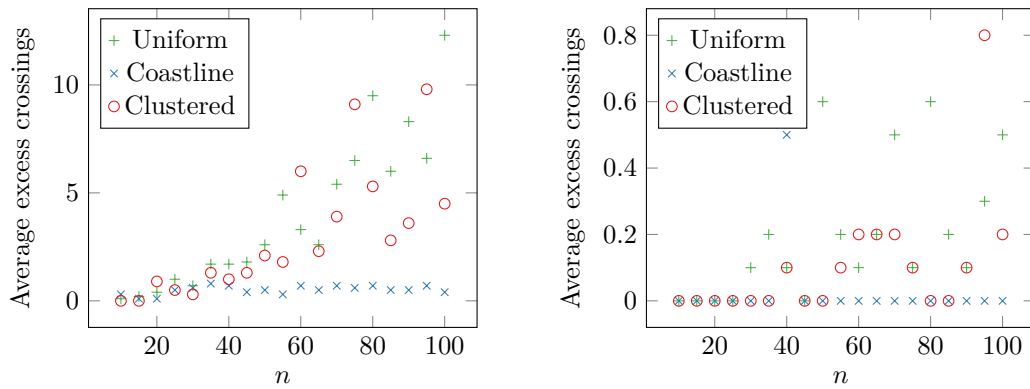
about 0.5 s, but even though this improves the number of crossings from the published 680 to the optimal 609, the drawing is so messy as to be unreadable (Figure 12b). We further observe that Coastline instances are solved trivially fast, since with fewer undecided pairs the ILP is smaller and presumably easier to solve.

**The synthetic instances have a superlinear number of crossings.** The Clustered instances can be drawn with significantly fewer crossings than Uniform: this matches our expectation, as by construction there is more correlation between the phylogenetic tree and the geography of the sites. More surprisingly we find that the Coastline instances require many crossings. We may have made them too noisy, but this does warn of the generally quadratic growth in number of crossings, which makes external labeling unsuitable for large geophylogenies unless the geographic correlation is exceptionally good.

**The heuristics run instantly and Greedy is often optimal.** The heuristics are implemented in single-threaded Java code. Bottom-Up, Top-Down and Leaf-Additive all run instantly, and even the Greedy hill climber runs in a fraction of a second. Of the first three heuristics, Bottom-Up consistently achieves the best results for both *s*- and *po*-leaders. Comparing the best solution by these heuristics with the optimal drawing (Figure 11), we observe that the number crossings in excess of the optimum increases with the number of taxa, in particular for Uniform and Clustered instances; Coastline instances are always drawn close to optimally by at least one heuristic. The Greedy hill climber often improves this to an optimal solution.

**For the number of crossings, *po*-leaders are promising.** In addition to *s*-leaders, our implementation of the heuristics can handle *po*-leaders. (The ILP cannot.) Our heuristics require on average only about 73% as many crossings when using *po*-leaders compared to *s*-leaders (55% for Coastline instances); the Lizard example in Figure 2b requires 11 *s*-leader crossings but only 2 *po*-leader crossings. We therefore propose that *po*-leaders deserve more attention from the phylogenetic community.

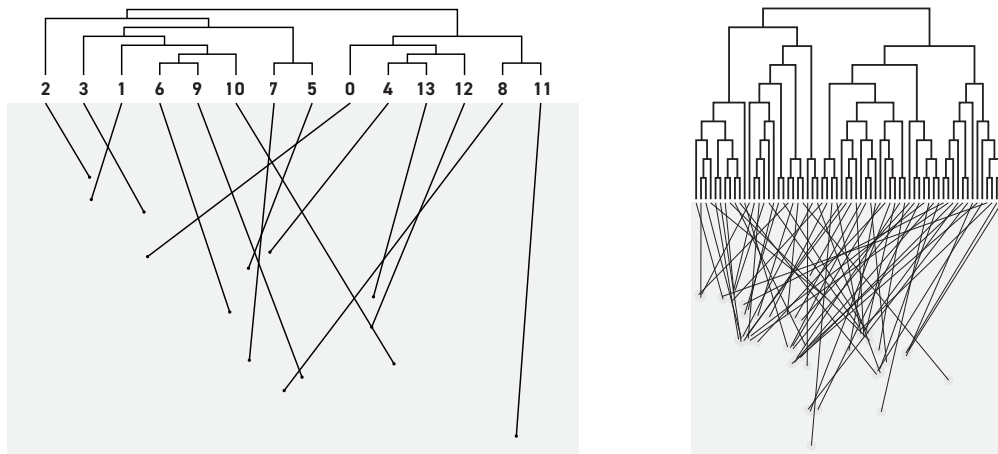
**Algorithmic recommendations.** Our results show that the ILP is a good choice for geophylogeny drawings with external labeling. If no solver is at hand or it is technically challenging to set up (for example when making an app that runs locally in a user’s web browser), then the heuristics offer an effective and efficient alternative, especially Bottom-Up and Greedy.



(a) Best heuristic without Greedy.

(b) Best after Greedy postprocessing.

■ **Figure 11** Number of crossings made by the best heuristic minus the number of crossings in the optimal drawing, averaged over 10 random instances per value of  $n$ .



(a) Drawing of **Fish** with 17 crossings.

(b) Drawing of **Frogs** with 609 crossings.

■ **Figure 12** Crossing-optimal drawings of Fish and Frogs with **s**-leaders.

For the Fish instance, for example, we found that the drawing with **s**-leaders and 17 crossings in Figure 12a is a good alternative to the internal labeling used in the published drawing [28]. However, for instances without a clear structure or with many crossings, it might be better to use internal labeling. Alternatively, the tree could be split like Tobler et al. [26], such that different subtrees are each shown with the map in separate drawings.

## 6 Discussion and Open Problems

In this paper, we have shown that drawings of geophylogenies can be approached theoretically and practically as a problem of algorithmic map labeling. We formally defined a drawing style for geophylogenies that uses either internal labeling with text or colors, or that uses external labeling with **s**/**po**-leaders. This allowed us to define optimization problems that can be tackled algorithmically. For drawings with internal labeling, we introduced a class of quality measures that can be optimized efficiently and even interactively constrained. In practice, designers can thus try different quality measures, pick their favorite, and make further

adjustments easily even for large instances. For external labeling, minimizing the number of leader crossings is NP-hard in general, but we provide multiple algorithmic approaches to solve this problem and demonstrated experimentally that they perform well in practice.

Even though we have provided a solid base of results, we feel the algorithmic study of geophylogeny drawings holds further promise by varying, for example, the type of leader used, the objective function, the composition of the drawing, or the nature of the phylogeny and the map. We finish this paper with several suggestions for future work.

One might consider *do*- and *pd*-leaders, which use a diagonal segment and can be aesthetically pleasing. We expect that some of our results (such as the NP-hardness of crossing minimization and the effectiveness of the heuristics) should hold for these leaders. The boundary labeling literature [5] studies even further types, such as *opo* and Bézier, and these might be more challenging to adapt.

For external labeling we have only considered the total number of crossings. If different colors are used for the leaders of different clades or if the drawing can be explored with an interactive tool, one might want to minimize the number of crossings within each clade (or a particular clade). Furthermore, one might optimize crossing angles. While we provided heuristics to minimize leader crossings, the development of approximation algorithms, which exist for other labeling problems [17, 3], could also be of interest.

Our model of a geophylogeny drawing can be expanded. One might allow the orientation of the map to be freely rotated, the extent of the map to be changed, or the leaves to be placed non-equidistantly. Optimizing over these additional freedoms poses new algorithmic challenges. Straying further from our model, some drawings in the literature have a circular tree around the map [21, 14]. (This is similar to contour labeling in the context of map labeling [19].) Also recall that Figure 1 has area features. Our quality measures for internal labeling are easily adapted to handle this, but (as is the case with general boundary labeling [4]) area features provide additional algorithmic challenges for external labeling. The literature contains many drawings where multiple taxa correspond to the same feature on the map [7], where we might want to look to many-to-one boundary labeling [17, 2]. Furthermore, one can consider non-binary phylogenetic trees and phylogenetic networks.

Lastly, we note that side-by-side drawings can also be used for a phylogenetic tree together with a diagram other than a map: Chen et al. [9] combine it with a scatter plot; Gehring et al. [12] even combine three things (phylogenetic tree, haplotype network, and map).

---

## References

- 1 Lukas Barth, Andreas Gemsa, Benjamin Niedermann, and Martin Nöllenburg. On the readability of leaders in boundary labeling. *Information Visualization*, 18(1), 2019. doi:10.1177/1473871618799500.
- 2 Michael A. Bekos, Sabine Cornelsen, Martin Fink, Seok-Hee Hong, Michael Kaufmann, Martin Nöllenburg, Ignaz Rutter, and Antonios Symvonis. Many-to-one boundary labeling with backbones. *Journal of Graph Algorithms and Applications*, 19(3):779–816, 2015. doi:10.7155/jgaa.00379.
- 3 Michael A. Bekos, Michael Kaufmann, Dimitrios Papadopoulos, and Antonios Symvonis. Combining traditional map labeling with boundary labeling. In Ivana Cerná, Tibor Gyimóthy, Juraj Hromkovic, Keith G. Jeffery, Rastislav Královic, Marko Vukolic, and Stefan Wolf, editors, *SOFSEM 2011*, volume 6543 of *LNCS*, pages 111–122. Springer, 2011. doi:10.1007/978-3-642-18381-2\_9.
- 4 Michael A. Bekos, Michael Kaufmann, Katerina Potika, and Antonios Symvonis. Area-feature boundary labeling. *The Computer Journal*, 53(6):827–841, 2010. doi:10.1093/comjnl/bxp087.

- 5 Michael A. Bekos, Benjamin Niedermann, and Martin Nöllenburg. External labeling techniques: A taxonomy and survey. *Computer Graphics Forum*, 38(3):833–860, 2019. doi:10.1111/cgf.13729.
- 6 R Alexander Bentley, William R Moritz, Damian J Ruck, and Michael J O'Brien. Evolution of initiation rites during the austronesian dispersal. *Science Progress*, 104(3):00368504211031364, 2021. doi:10.1177/00368504211031364.
- 7 Tyler K Chafin, Marlis R Douglas, Whitney JB Anthonysamy, Brian K Sullivan, James M Walker, James E Cordes, and Michael E Douglas. Taxonomic hypotheses and the biogeography of speciation in the tiger whiptail complex. *Frontiers*, 13(2), 2021. doi:10.21425/F5FBG49120.
- 8 Zachary Charlop-Powers and Sean F. Brady. phylogeo: an R package for geographic analysis and visualization of microbiome data. *Bioinformatics*, 31(17):2909–2911, 2015. doi:10.1093/bioinformatics/btv269.
- 9 Yi Chen, Lei Zhao, Huajing Teng, Chengmin Shi, Quansheng Liu, Jianxu Zhang, and Yaohua Zhang. Population genomics reveal rapid genetic differentiation in a recently invasive population of *rattus norvegicus*. *Frontiers in Zoology*, 18(1):6, 2021. doi:10.1186/s12983-021-00387-z.
- 10 Gajaba Ellepola, Jayampathi Herath, Kelum Manamendra-Arachchi, Nayana Wijayathilaka, Gayani Senevirathne, Rohan Pethiyagoda, and Madhava Meegaskumbura. Molecular species delimitation of shrub frogs of the genus *pseudophilautus* (anura, rhacophoridae). *PLOS ONE*, 16(10):1–17, 2021. doi:10.1371/journal.pone.0258594.
- 11 Henning Fernau, Michael Kaufmann, and Mathias Poths. Comparing trees via crossing minimization. *Journal of Computer and System Sciences*, 76(7):593–608, 2010. doi:10.1016/j.jcss.2009.10.014.
- 12 Philip-Sebastian Gehring, Maciej Pabijan, Jasmin E. Randrianirina, Frank Glaw, and Miguel Vences. The influence of riverine barriers on phylogeographic patterns of malagasy reed frogs (*heterixalus*). *Molecular Phylogenetics and Evolution*, 64(3):618–632, 2012. doi:10.1016/j.ympev.2012.05.018.
- 13 Robin-Tobias Jauss, Nadiné Solf, Sree Rohit Raj Kolora, Stefan Schaffer, Ronny Wolf, Klaus Henle, Uwe Fritz, and Martin Schlegel. Mitogenome evolution in the *lacerta viridis* complex (lacertidae, squamata) reveals phylogeny of diverging clades. *Systematics and Biodiversity*, 19(7):682–692, 2021. doi:10.1080/14772000.2021.1912205.
- 14 Monika Karmin, Rodrigo Flores, Lauri Saag, Georgi Hudjashov, Nicolas Brucato, Chelzie Crenna-Darusallam, Maximilian Larena, Phillip L Endicott, Mattias Jakobsson, J Stephen Lansing, Herawati Sudoyo, Matthew Leavesley, Mait Metspalu, François-Xavier Ricaut, and Murray P Cox. Episodes of Diversification and Isolation in Island Southeast Asian and Near Oceanian Male Lineages. *Molecular Biology and Evolution*, 39(3), 2022. doi:10.1093/molbev/msac045.
- 15 David M. Kidd and Xianhua Liu. geophylobuilder 1.0: an arcgis extension for creating “geophylogenies”. *Molecular Ecology Resources*, 8(1):88–91, 2008. doi:10.1111/j.1471-8286.2007.01925.x.
- 16 Jonathan Klawitter, Felix Klesen, Joris Y. Scholl, Thomas C. van Dijk, and Alexander Zaft. Visualizing geophylogenies – internal and external labeling with phylogenetic tree constraints. *CoRR*, abs/2306.17348, 2023. arXiv:2306.17348.
- 17 Chun-Cheng Lin, Hao-Jen Kao, and Hsu-Chun Yen. Many-to-one boundary labeling. *Journal of Graph Algorithms and Applications*, 12(3):319–356, 2008. doi:10.7155/jgaa.00169.
- 18 Gabriele Neyer. Map labeling with application to graph drawing. In Michael Kaufmann and Dorothea Wagner, editors, *Drawing Graphs: Methods and Models*, pages 247–273. Springer, 2001. doi:10.1007/3-540-44969-8\_10.
- 19 Benjamin Niedermann, Martin Nöllenburg, and Ignaz Rutter. Radial contour labeling with straight leaders. In Daniel Weiskopf, Yingcai Wu, and Tim Dwyer, editors, *IEEE Pacific Visualization Symposium*, pages 295–304. IEEE Computer Society, 2017. doi:10.1109/PACIFICVIS.2017.8031608.



- 20 Roderic Page. Visualising geophylogenies in web maps using geojson. *PLOS Currents*, 7, 2015. doi:10.1371/currents.tol.8f3c6526c49b136b98ec28e00b570a1e.
- 21 Da Pan, Boyang Shi, Shiyu Du, Tianyu Gu, Ruxiao Wang, Yuhui Xing, Zhan Zhang, Jiajia Chen, Neil Cumberlidge, and Hongying Sun. Mitogenome phylogeny reveals Indochina Peninsula origin and spatiotemporal diversification of freshwater crabs (Potamidae: Potamiscinae) in China. *Cladistics*, 38(1):1–12, 2022. doi:10.1111/c1a.12475.
- 22 Donovan H. Parks, Timothy Mankowski, Somayeh Zangoeei, Michael S. Porter, David G. Armanini, Donald J. Baird, Morgan G. I. Langille, and Robert G. Beiko. GenGIS 2: geospatial analysis of traditional and genetic biodiversity, with new gradient algorithms and an extensible plugin framework. *PLoS ONE*, 8(7):1–10, 2013. doi:10.1371/journal.pone.0069885.
- 23 Donovan H. Parks, Michael Porter, Sylvia Churcher, Suwen Wang, Christian Blouin, Jacqueline Whalley, Stephen Brooks, and Robert G. Beiko. GenGIS: A geospatial information system for genomic data. *Genome Research*, 19(10):1896–1904, 2009. doi:10.1101/gr.095612.109.
- 24 Liam J. Revell. phytools: an r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223, 2012. doi:10.1111/j.2041-210X.2011.00169.x.
- 25 Mike Steel. *Phylogeny: Discrete and Random Processes in Evolution*. Society for Industrial and Applied Mathematics, 2016. doi:10.1137/1.9781611974485.
- 26 Ray Tobler, Adam Rohrlach, Julien Soubrier, Pere Bover, Bastien Llamas, Jonathan Tuke, Nigel Bean, Ali Abdullah-Highfold, Shane Agius, Amy O’Donoghue, Isabel O’Loughlin, Peter Sutton, Fran Zilio, Keryn Walshe, Alan N. Williams, Chris S M Turney, Matthew Williams, Stephen M Richards, Robert J Mitchell, Emma Kowal, John R Stephen, Lesley Williams, Wolfgang Haak, and Alan Cooper. Aboriginal mitogenomes reveal 50,000 years of regionalism in australia. *Nature*, 544(7649):180–184, 2017. doi:10.1038/nature21416.
- 27 Jason T. Weir, Oliver Haddrath, Hugh A. Robertson, Rogan M. Colbourne, and Allan J. Baker. Explosive ice age diversification of kiwi. *Proceedings of the National Academy of Sciences*, 113(38):E5580–E5587, 2016. doi:10.1073/pnas.1603795113.
- 28 Trevor J. Williams and Jerald B. Johnson. History predicts contemporary community diversity within a biogeographic province of freshwater fish. *Journal of Biogeography*, 49(5):809–821, 2022. doi:10.1111/jbi.14316.
- 29 Xuhua Xia. Pgt: Visualizing temporal and spatial biogeographic patterns. *Global Ecology and Biogeography*, 28(8):1195–1199, 2019. doi:10.1111/geb.12914.