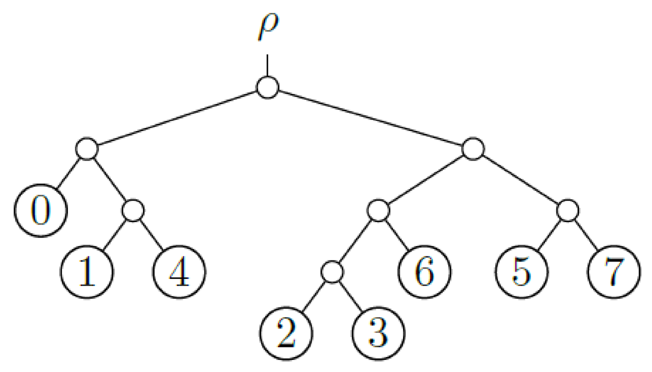
OLA encoding of binary trees

Erick Matsen *Fred Hutch Cancer Center*, Harry Richman† *Fred Hutch Cancer Center*, Cheng Zhang *Peking University*

How to encode a phylogenetic tree on a computer?

- Can I generate a random tree?
- Can I interpolate between trees?
- Can I estimate distance between trees?
- Can I infer a tree using a **deep neural network**?



$\leftrightarrow$ `((0,(1,4)),(((2,3),6),(5,7)));`

Figure 1. Current "standard" format: Newick string

OLA encoding

- New format for encoding trees:

$$\mathrm{OLA}_n : \left\{ \begin{matrix} \text{rooted binary trees on} \\ 0,1,2,\ldots,n-1 \end{matrix} \right\} \xrightarrow{\sim} \left\{ \begin{matrix} \text{vectors } (a_1, a_2, \ldots) \in \mathbb{Z}^{n-1} \\ \text{with } -i < a_i < i \end{matrix} \right\}$$

- Encoding space is **compact**, easy to sample.
- Easy **induced distance** from Hamming distance on $\mathbb{Z}^{n-1}$.
- Distance in encoding space is **somewhat-low distorsion** vs. SPR distance.
- Tree in Figure 1 is encoded as `(0,-1,2,1,-3,-3,5)`.

Related work

For all (unrooted) trees with internal & leaf nodes labeled:

- Prüfer code (1918) to vectors in $\mathbb{Z}^{n-2}$ with $0 < a_i \leq n$.

For rooted binary trees with leaves labeled:

- Diaconis and Holmes (2002) define bijection to perfect matchings on leaf set.
- Chauve, Colijn, and L. Zhang (2024) encoding by certain permutations of multiset $\{1, 1, 2, 2, \ldots, n, n\}$.
- Penn, Scheidwasser et al. (2024) "Phylo2vec" encoding by integer vectors, quadratic time complexity.

Example encoding: Ordered leaf attachment

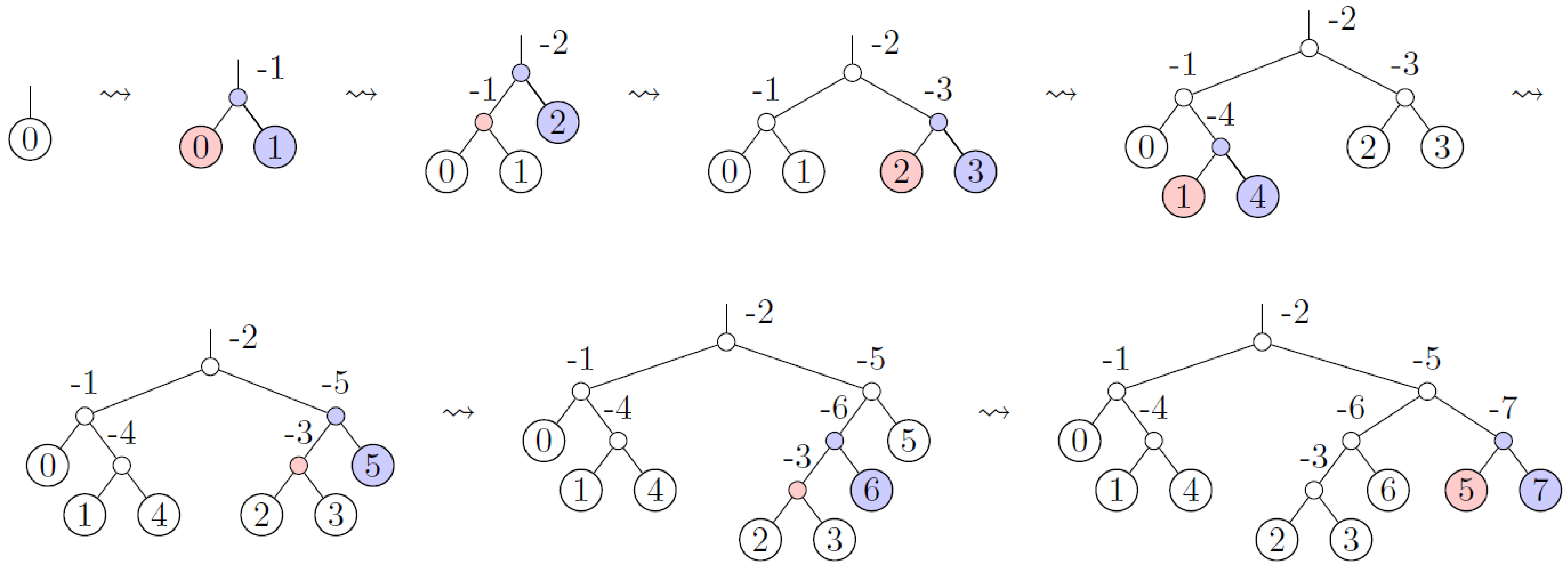


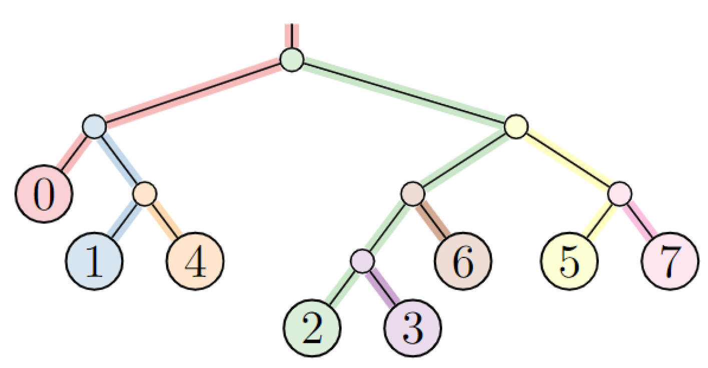
Figure 2. Constructing the tree with OLA code `(0,-1,2,1,-3,-3,5)`.

Results

Theorem (Matsen–R–Zhang, 2024)

OLA encoding and decoding takes **linear time**.

- Practical implementation uses "branch decomposition" for internal node labels



Theorem (Matsen–R–Zhang, 2024)

(a) If T' is a random **NNI neighbor** of T, then $\mathbb{E}(d_{\mathrm{OLA}}(T, T')) = O(1)$.
(b) If T' is a random **SPR neighbor** of T, then $\mathbb{E}(d_{\mathrm{OLA}}(T, T')) = O(\mathrm{height}(T))$.

OLA distance on trees

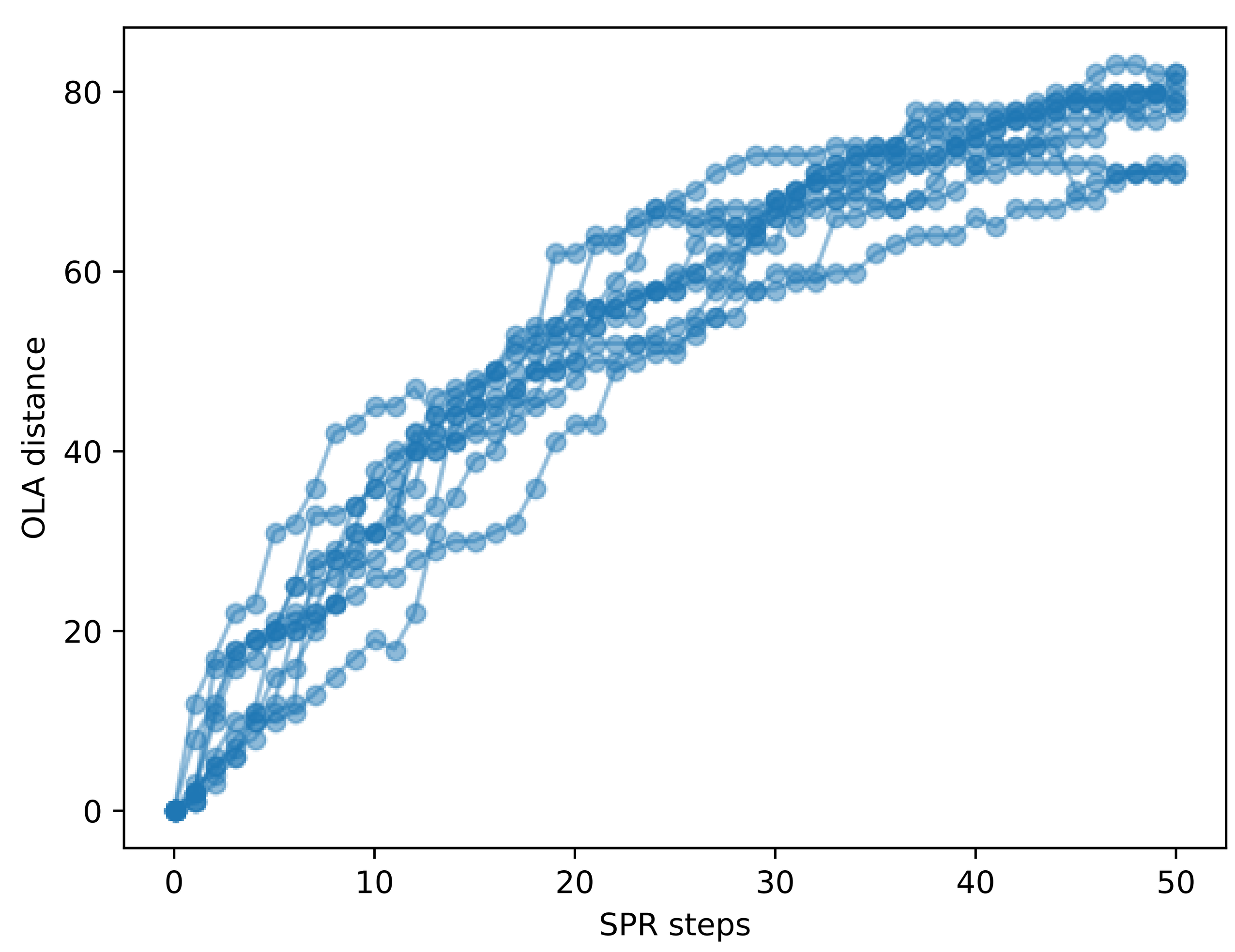


Figure 3. Experimental data from trees on 100 leaves.

https://harryrichman.github.io/
hrichman@alum.mit.edu