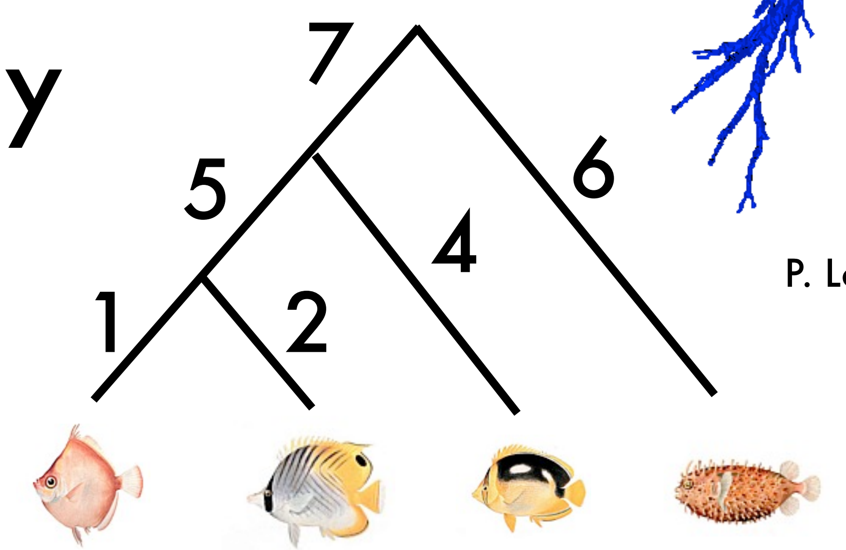
Statistics in BHV Tree Space

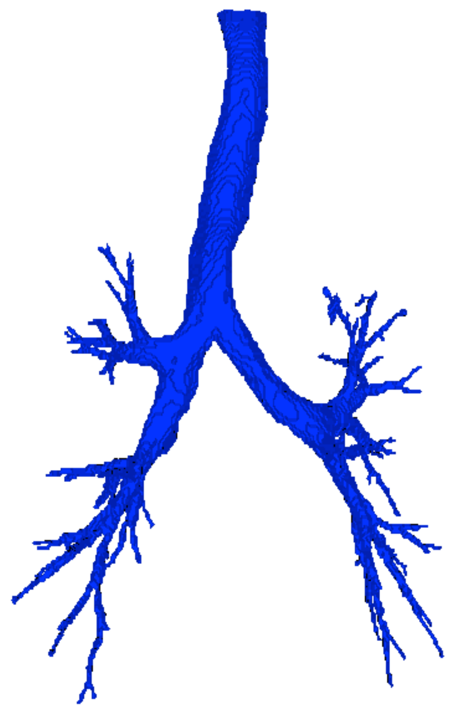
Megan Owen
Lehman College, CUNY
Joint with Daniel Brown (U. Waterloo)

BIRS Workshop on Mathematical Approaches to Evolutionary Trees and Networks
Feb. 13, 2017

Overarching Goal

- many examples of tree-shaped data (phylogenies, anatomical trees, etc.)
- parameters:
  - tree shape = tree topology
  - edge lengths
- not Euclidean data!



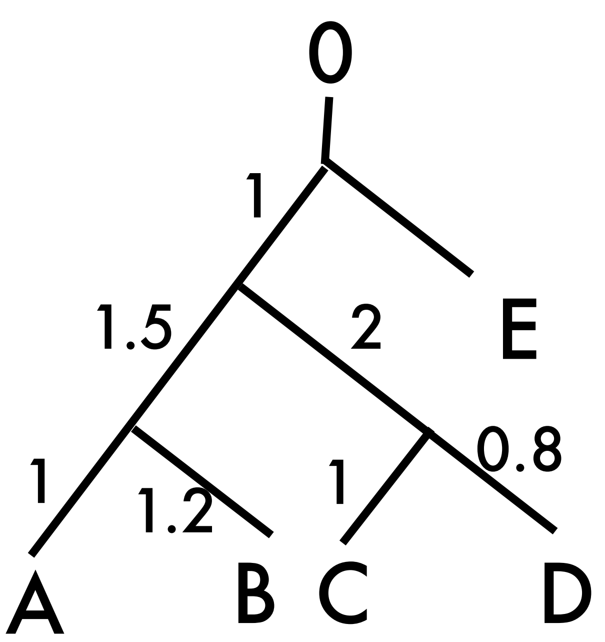


P. Lo et al. EXACT'09

Goal: develop methods for statistical analysis (i.e. mean, PCA) in a space of metric trees analogous to those for Euclidean space

Tree Space Framework

- constructed by Billera, Holmes, and Vogtmann (2001)
- tree space $\mathbb{T}_n$ = set of all trees with *n* leaves and branch lengths
- includes degenerate trees (non-binary)



Tree Space

- represent each tree as a vector
- coordinates = splits

Tree Space

- not all sets of splits form a tree

  ⇒ not all vectors are possible

  ⇒ not a Euclidean space

A|BCD0 B|ACD0 ... D|ABC0 AB|CD0 AC|BD0 ... ABC|D0 ABD|C0 AB0|CD

(2, 4, 2, 3, 2, 2, 0, 0, 0, 0, 0, ...)

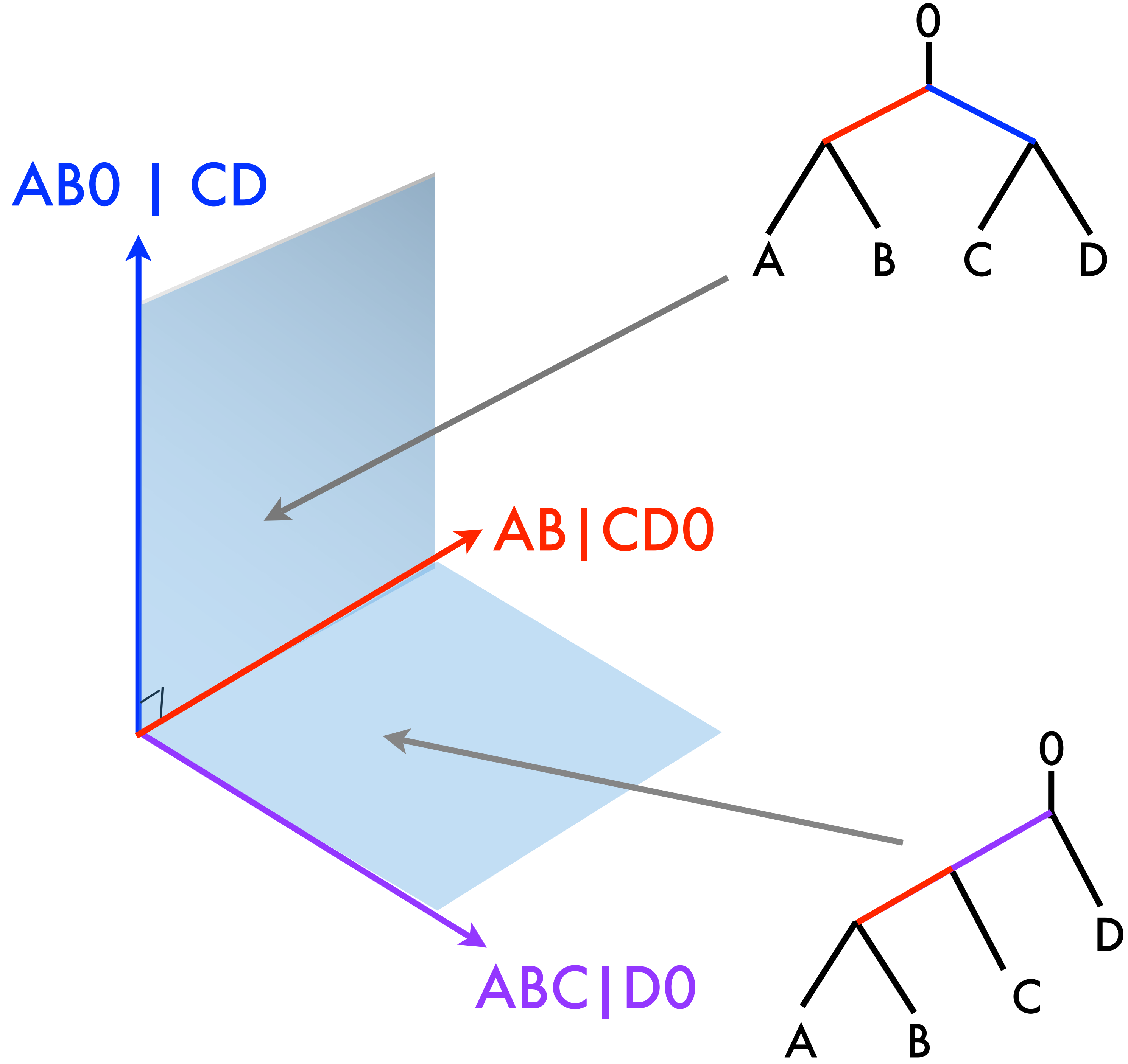
Tree Space

- not all sets of splits form a tree

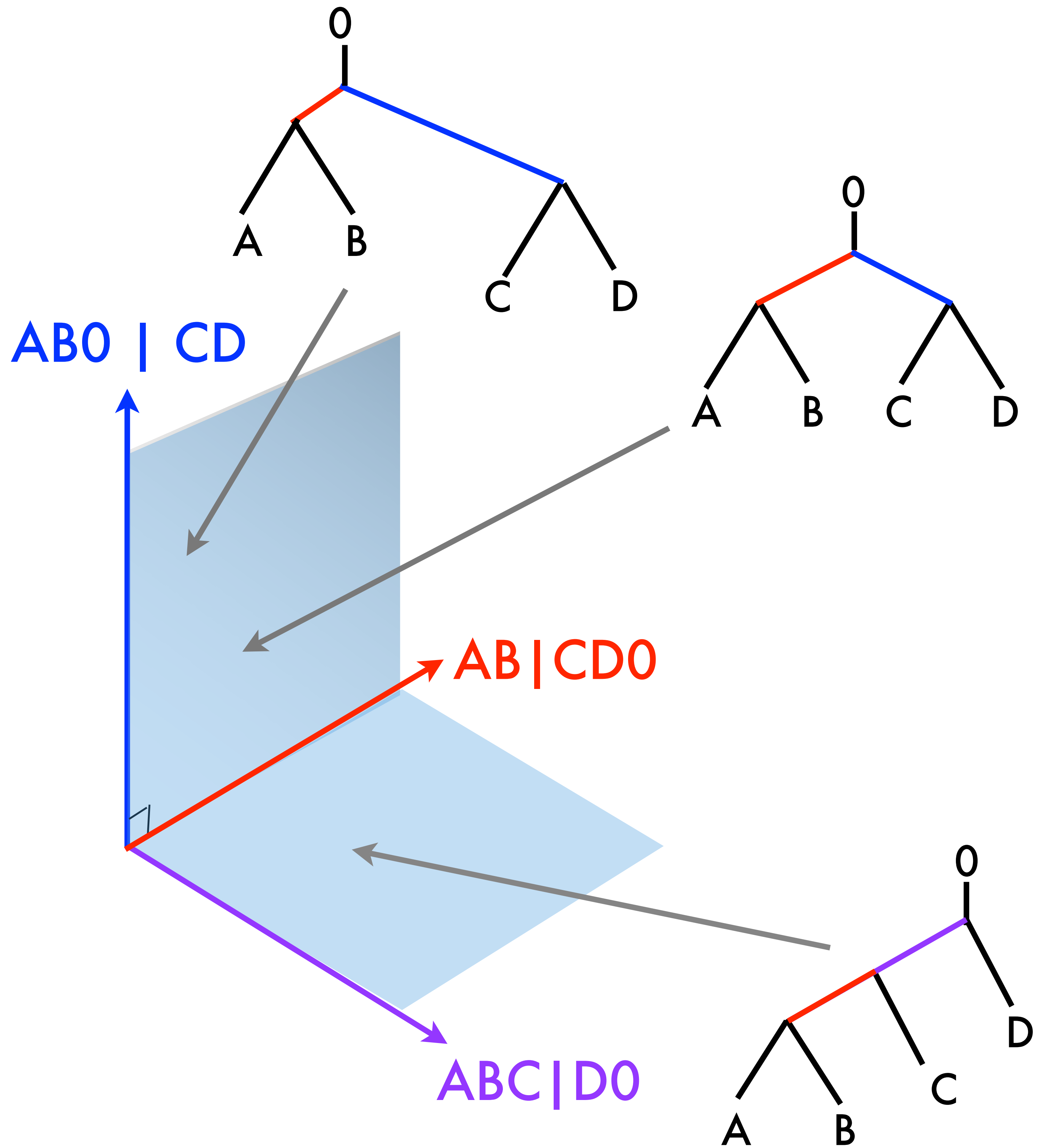
  ⇒ not all vectors are possible

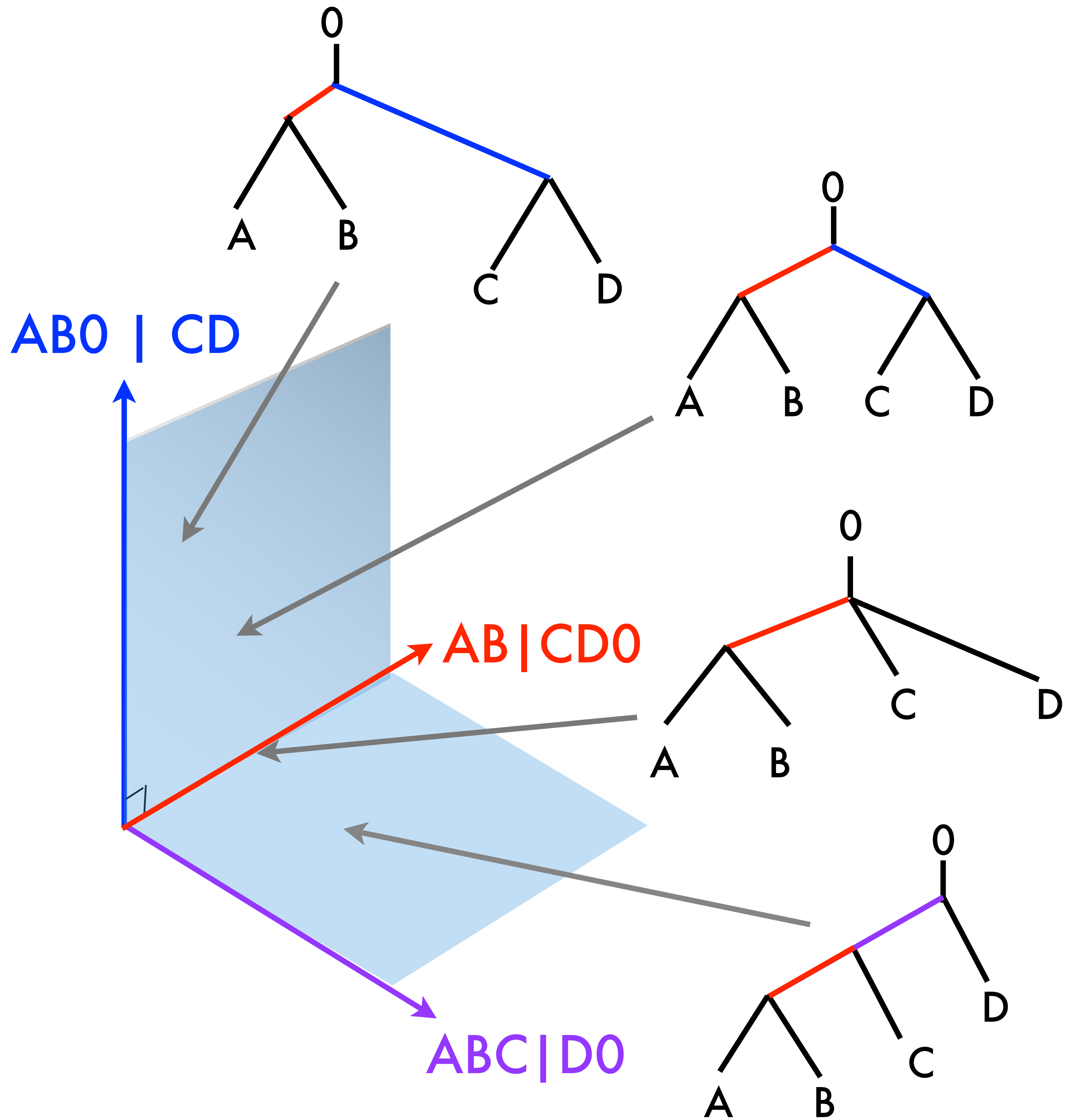
  ⇒ not a Euclidean space

$$\begin{array}{ccccccccccc} A|BCD0 & B|ACD0 & \ldots & D|ABC0 & AB|CD0 & AC|BD0 & & \ldots & ABC|D0 & ABD|C0 & AB0|CD \\ (2, & 4, & 2, & 3, & 2, & 2, & 0, & 0, & 0, & 0, & 0, \ldots) \end{array}$$

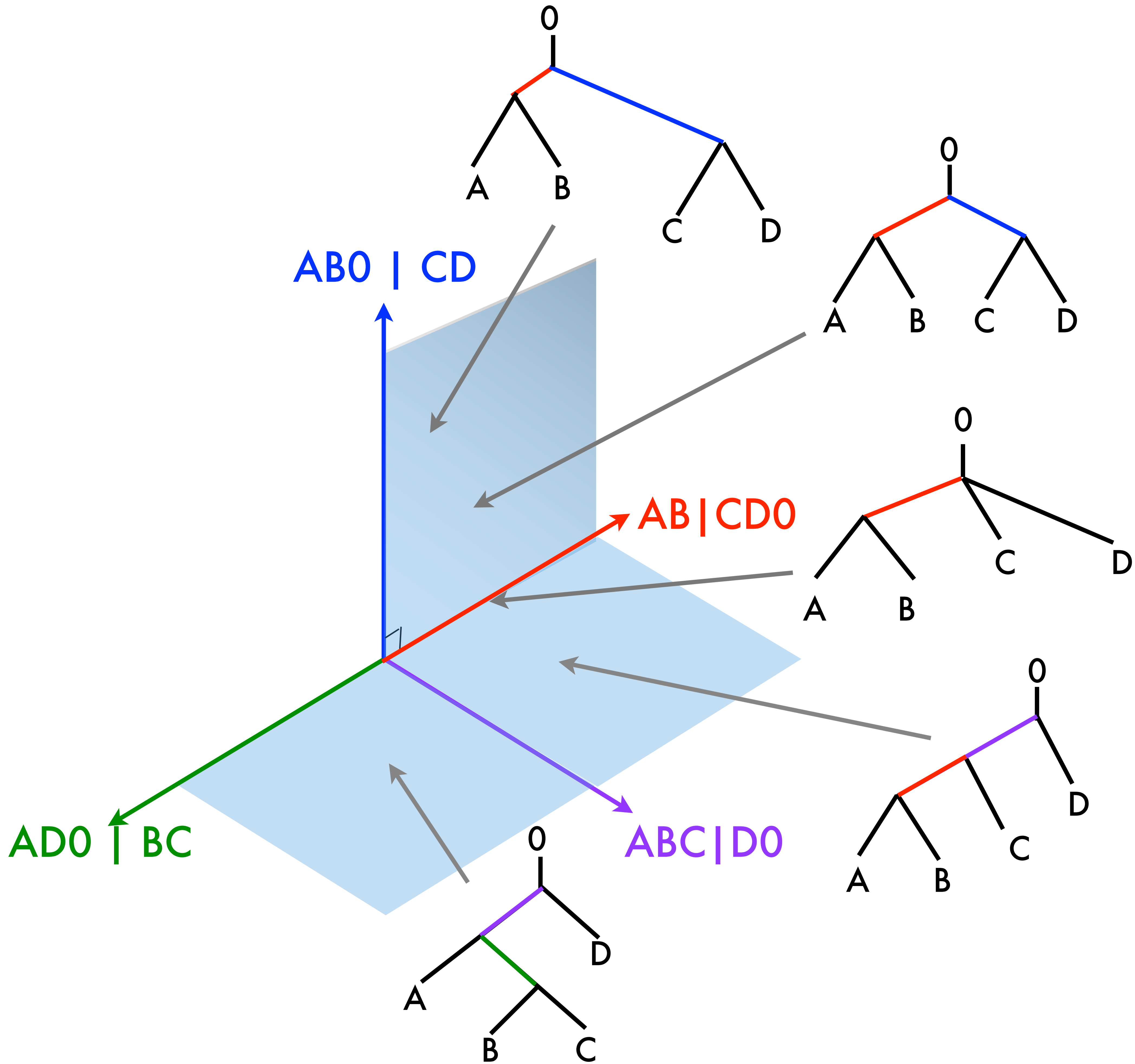


AB0 | CD
AB|CD0
ABC|D0
0
A
B
C
D
0
A
B
C
D

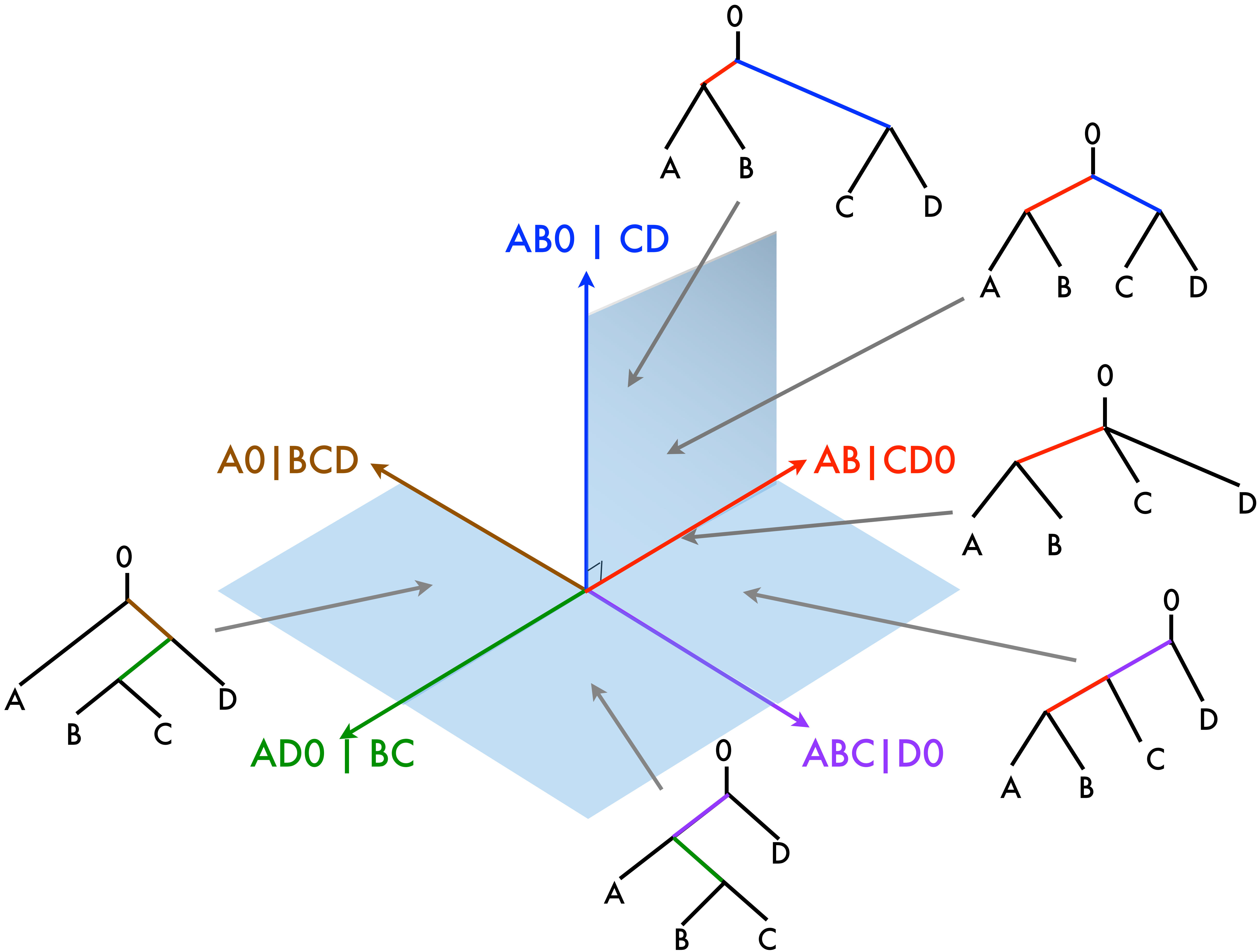

0
A
B
C
D
AB0 | CD
0
A
B
C
D
AB|CD0
0
D
C
A
B
ABC|D0

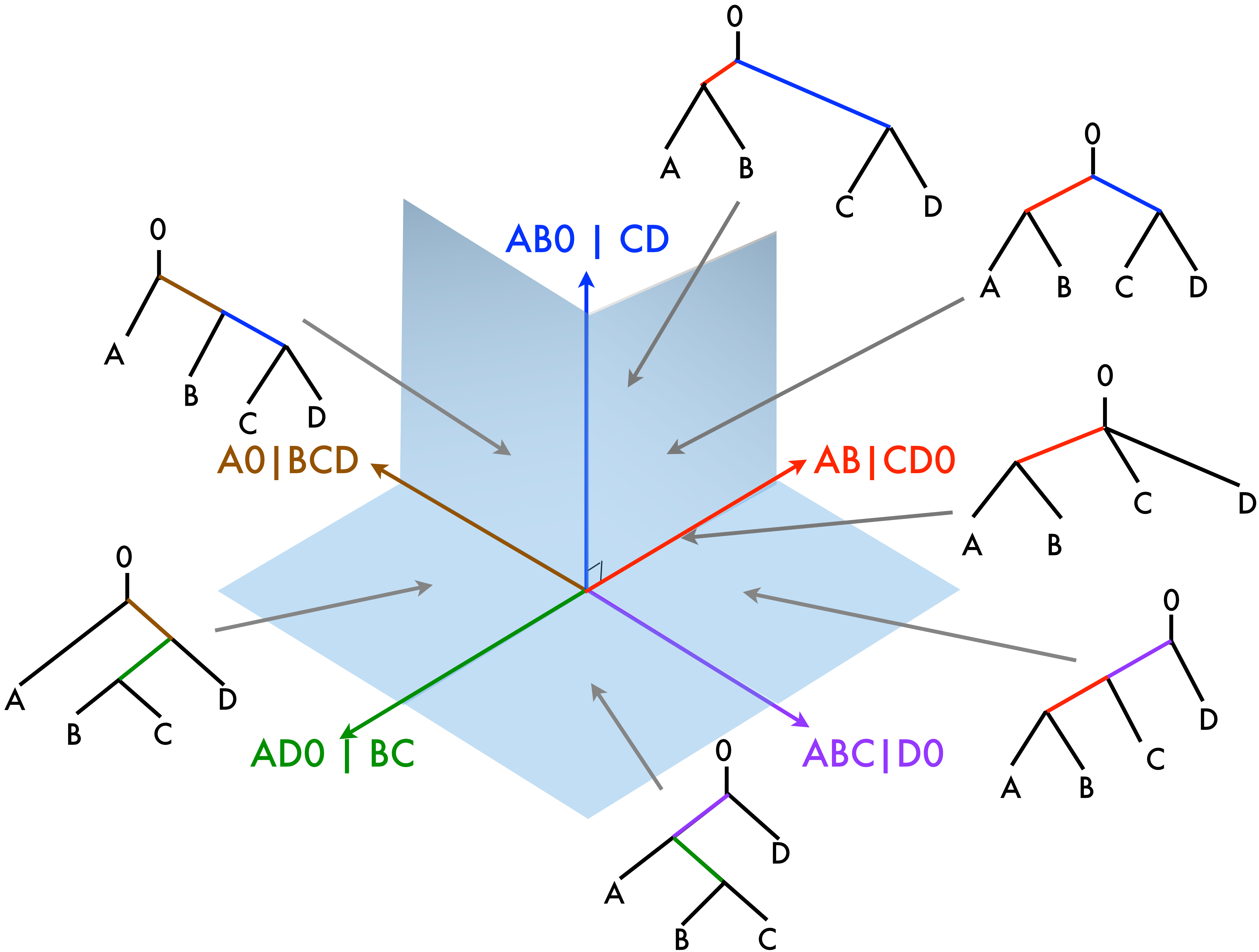


AB0 | CD
AB|CD0
ABC|D0
0
A
B
C
D

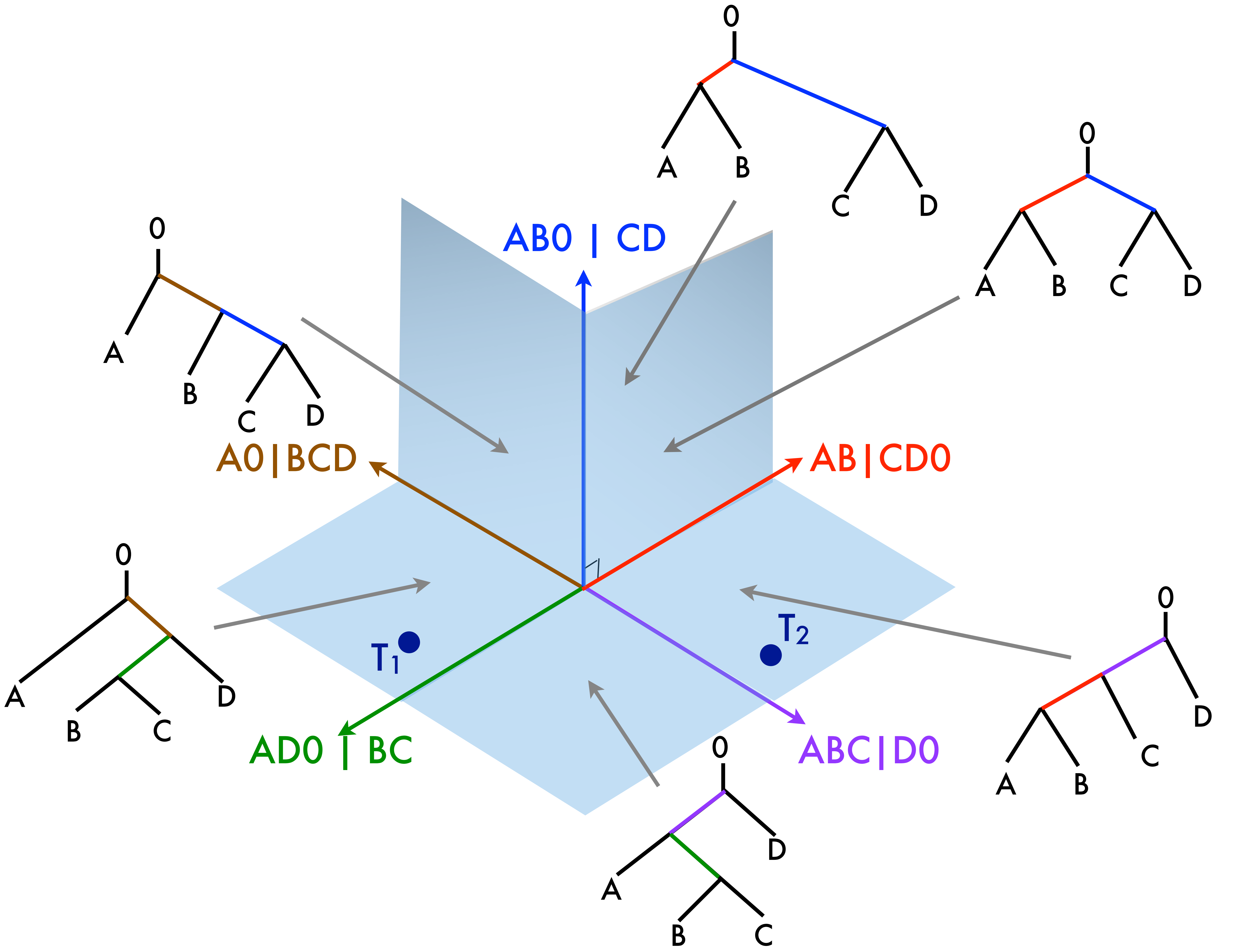


AB0 | CD
AB|CD0
AD0 | BC
ABC|D0
0
A
B
C
D

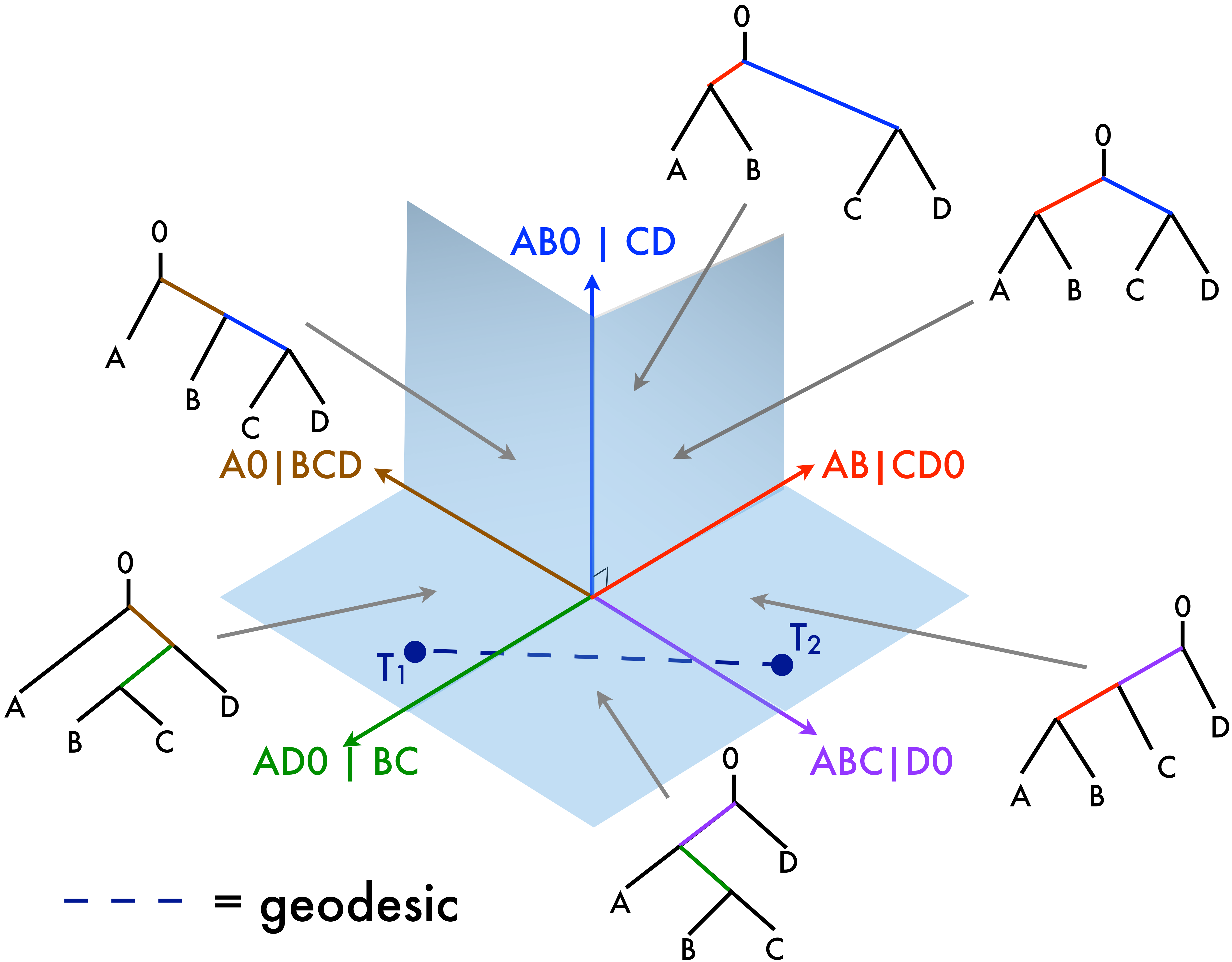

AB0 | CD
A0|BCD
AB|CD0
AD0 | BC
ABC|D0
0
A
B
C
D



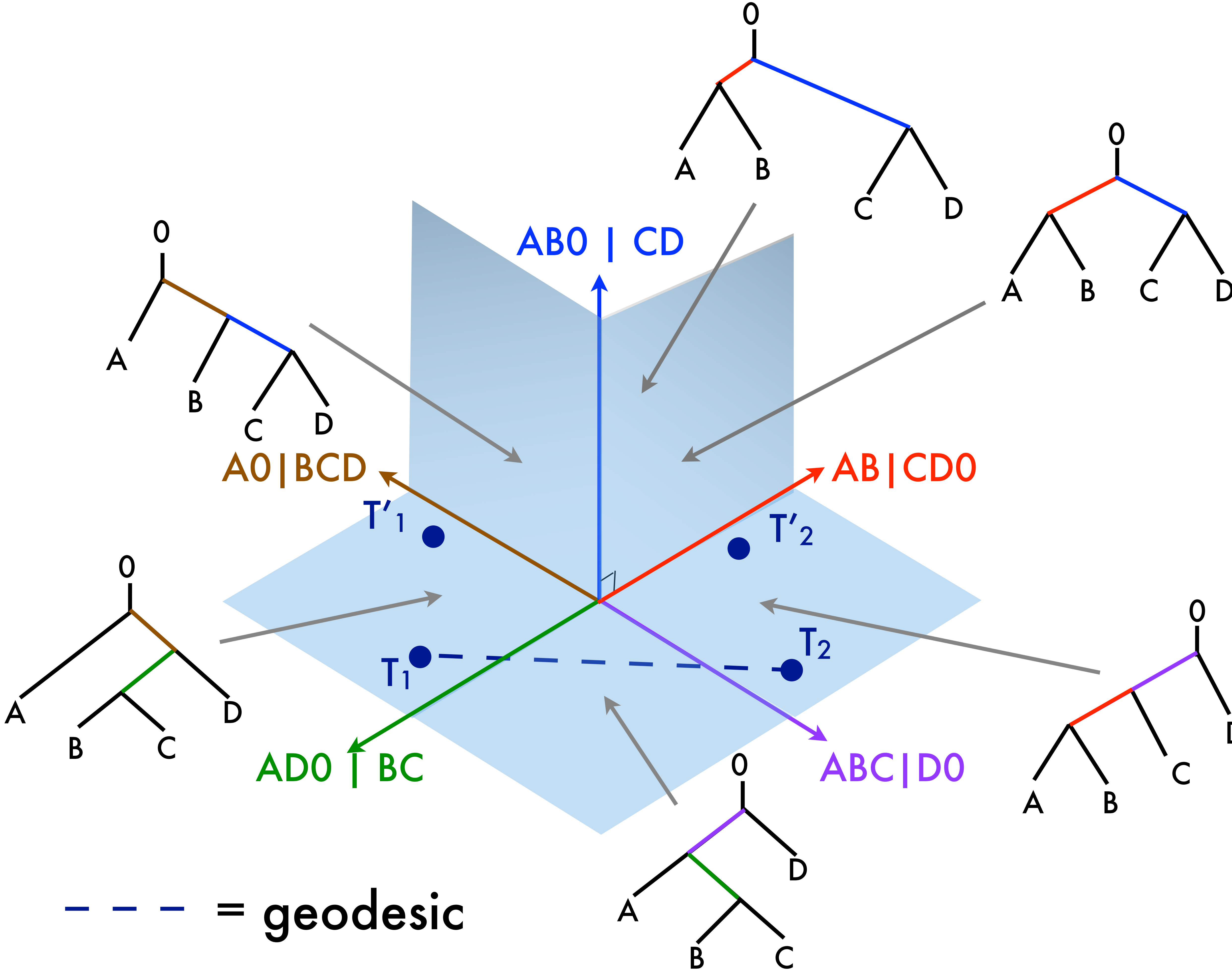
AB0 | CD
AB | CD0
A0 | BCD
AD0 | BC
ABC | D0
0
A
B
C
D

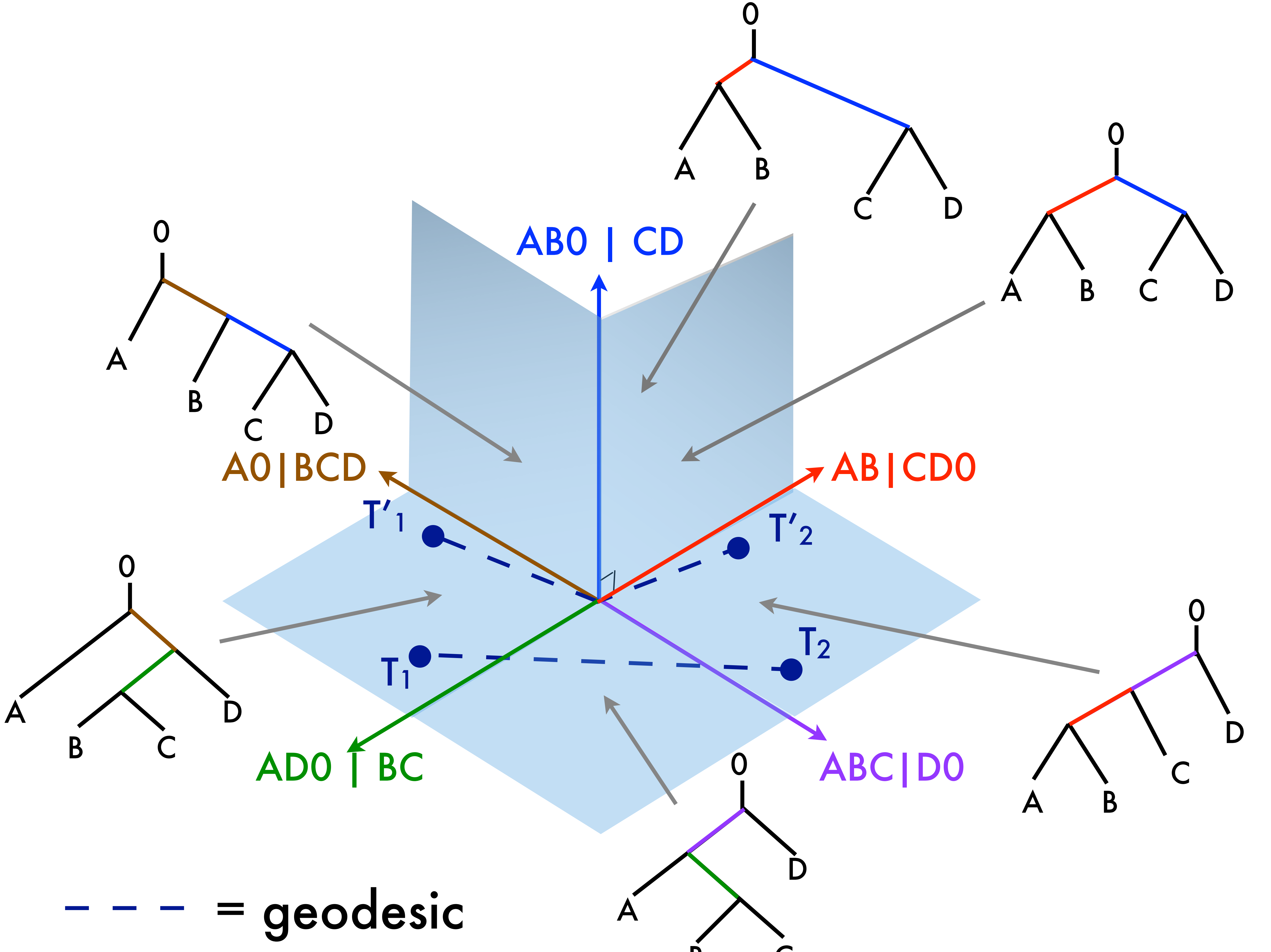


AB0 | CD
AB|CD0
A0|BCD
AD0 | BC
ABC|D0
T1
T2
0
A
B
C
D



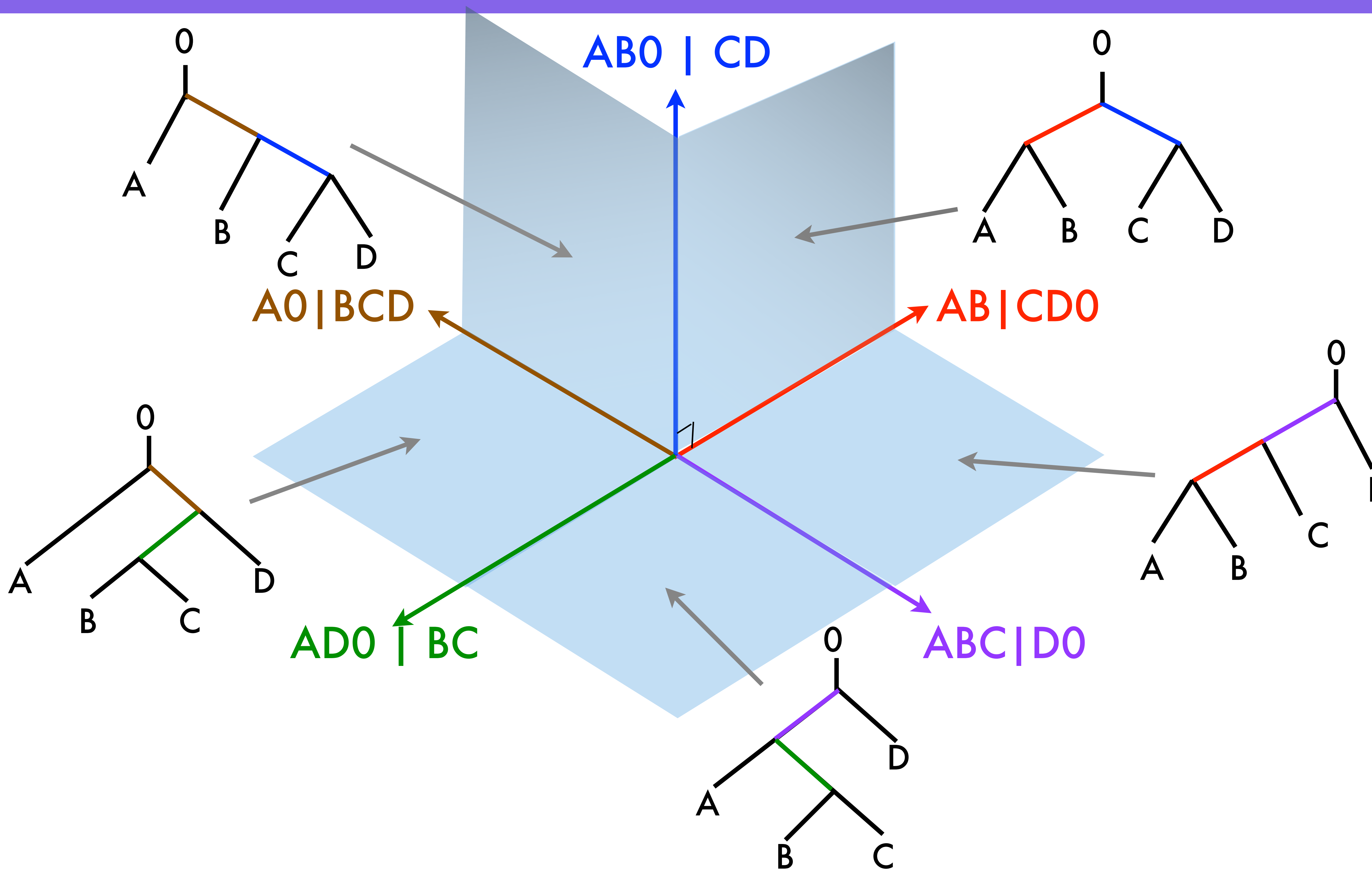
AB0 | CD
AB|CD0
A0|BCD
AD0 | BC
ABC|D0
T1
T2
0
A
B
C
D
= geodesic


AB0 | CD
A0|BCD
AB|CD0
AD0 | BC
ABC|D0
T'1
T'2
T1
T2
0
A
B
C
D
= geodesic

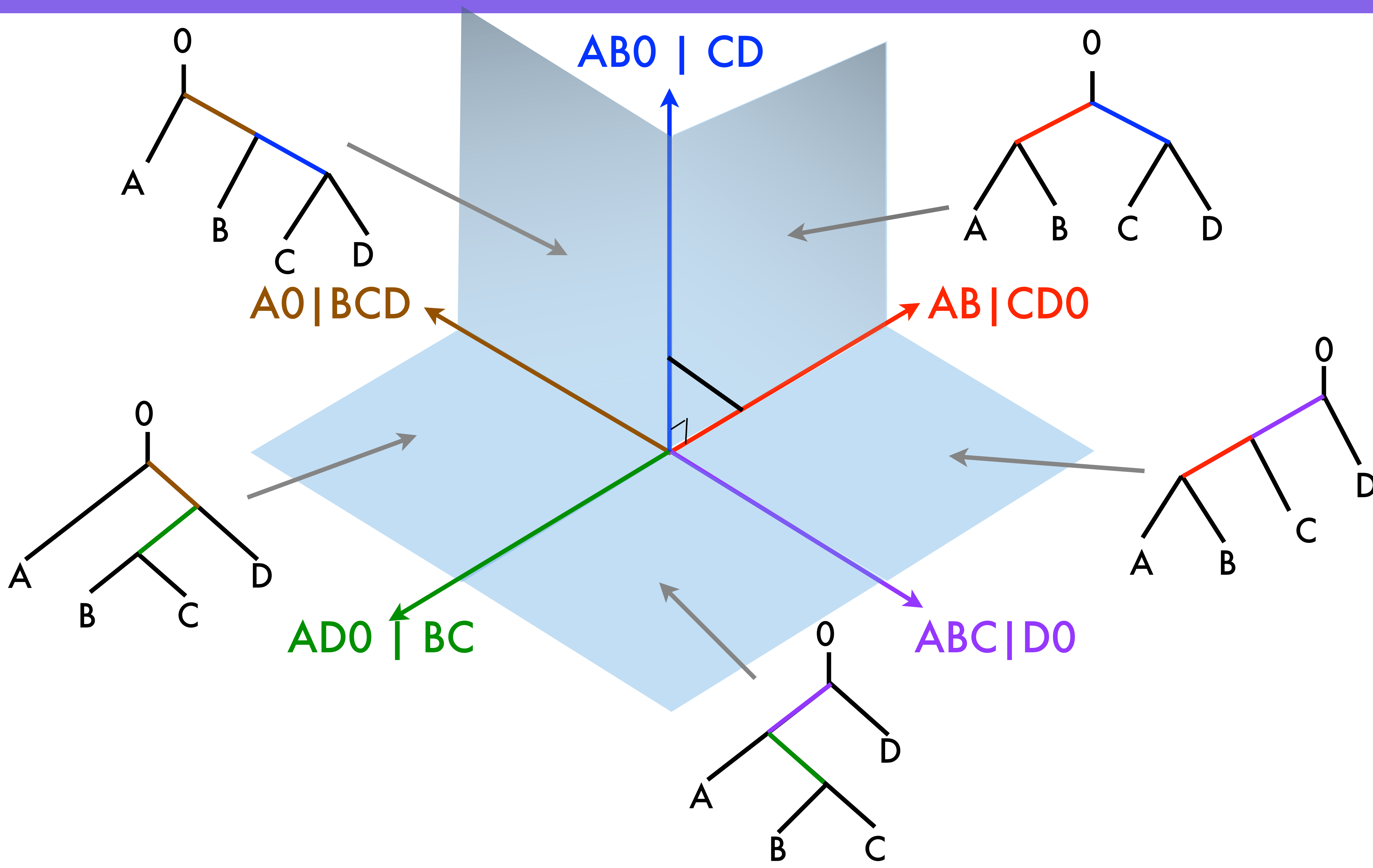


AB0 | CD
A0|BCD
AB|CD0
AD0 | BC
ABC|D0
T'1
T'2
T1
T2
0
A
B
C
D
= geodesic

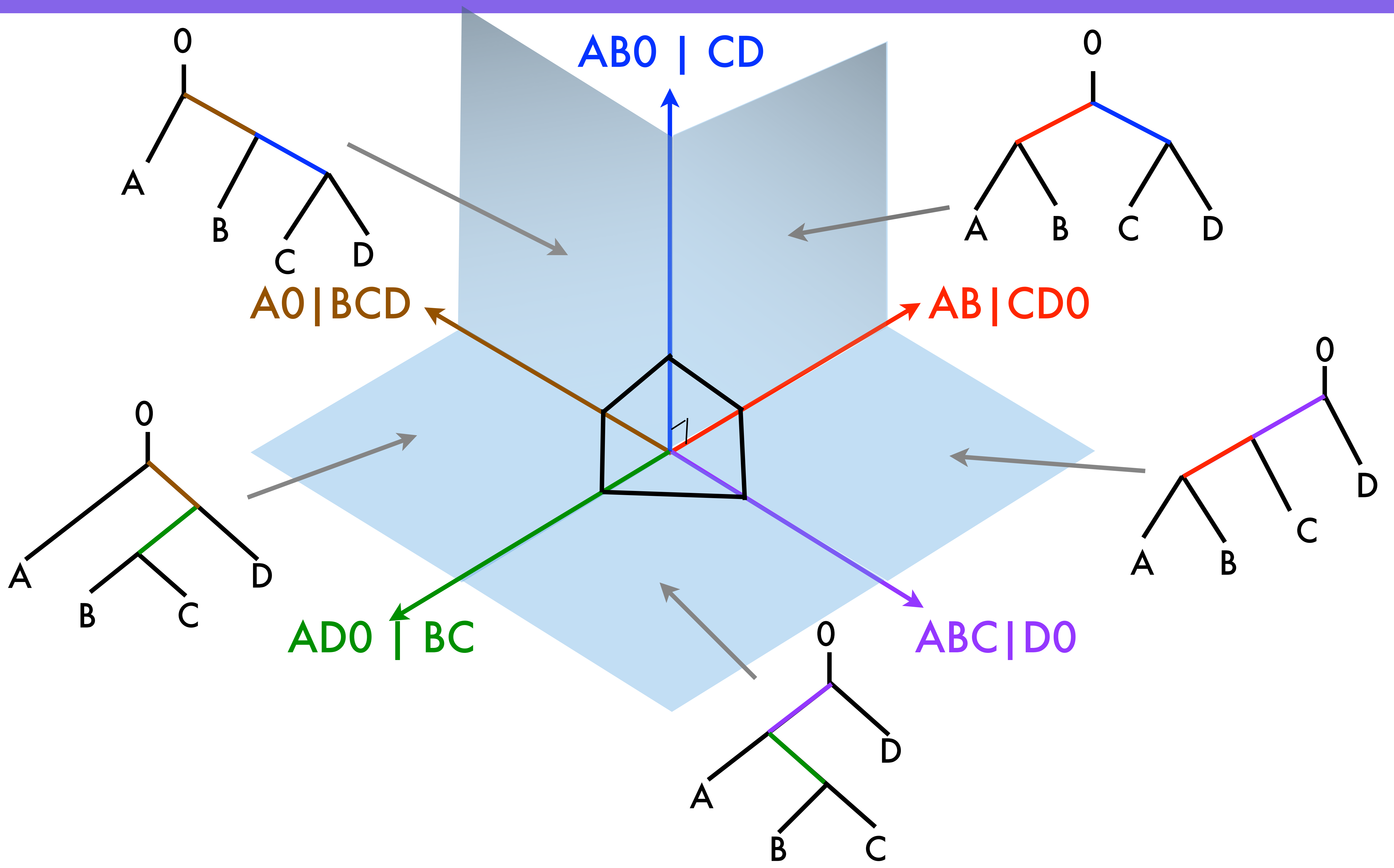
Structure of $\mathbb{T}_4$



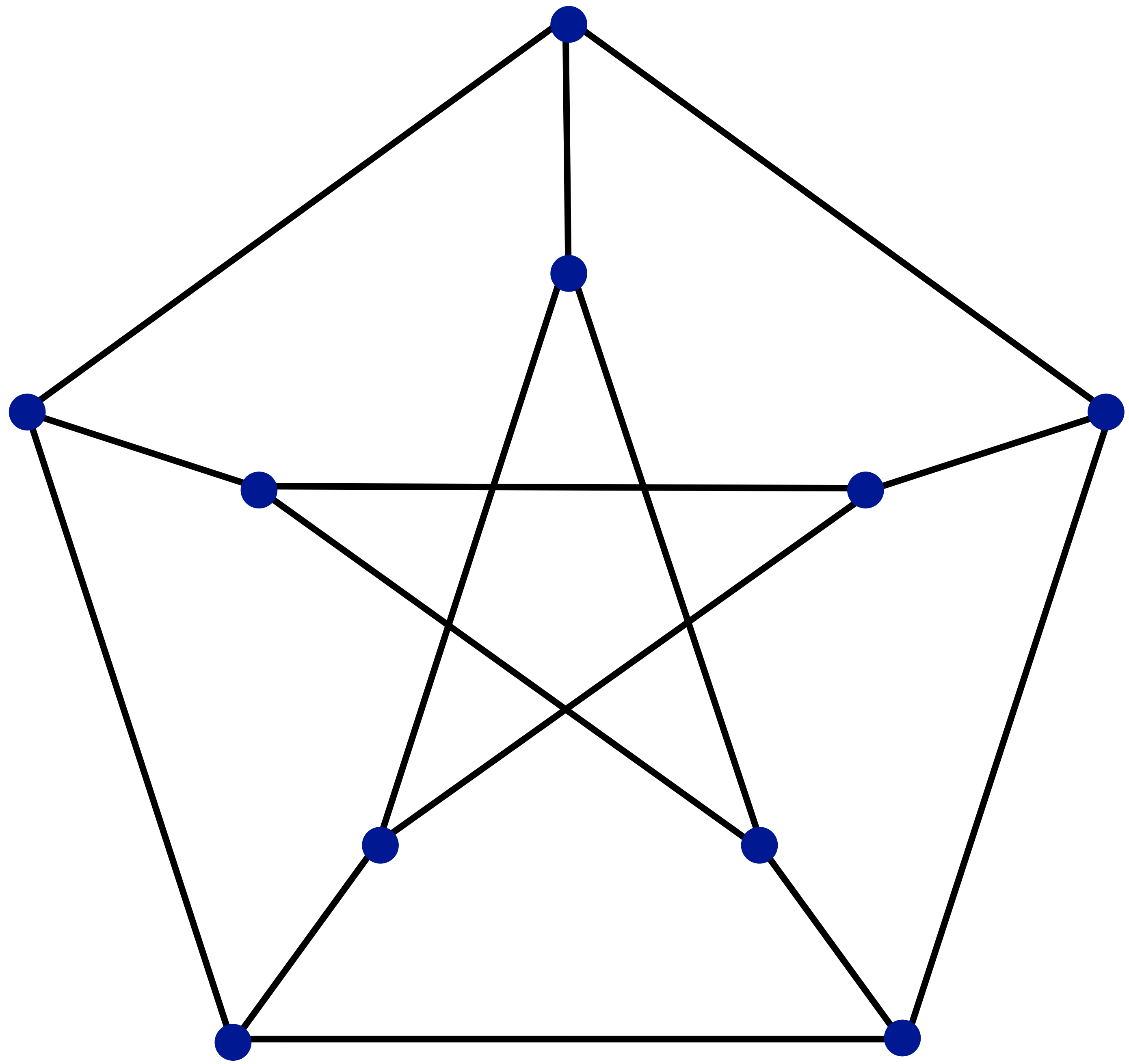
Structure of $\mathbb{T}_4$



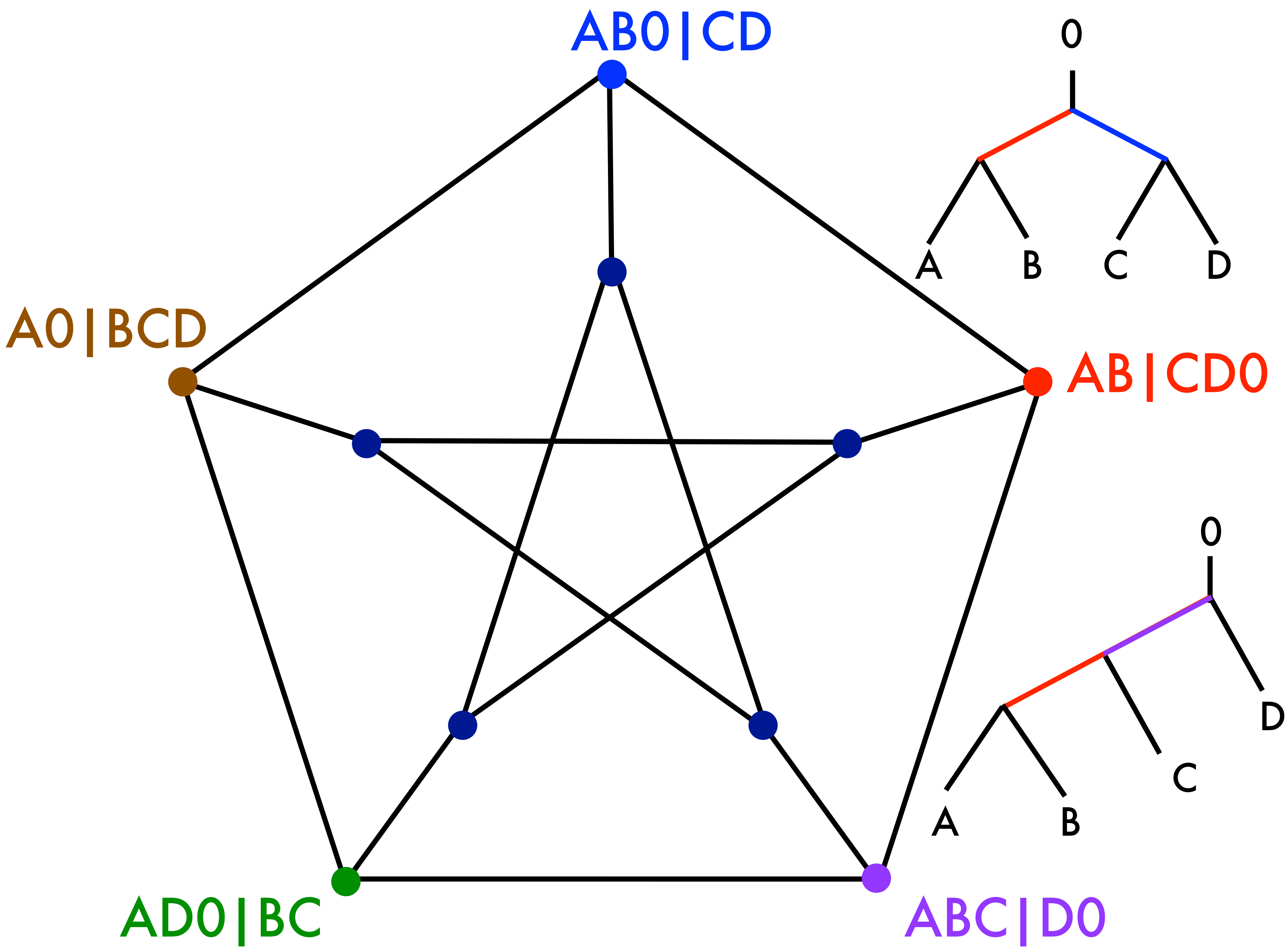
Structure of $\mathbb{T}_4$



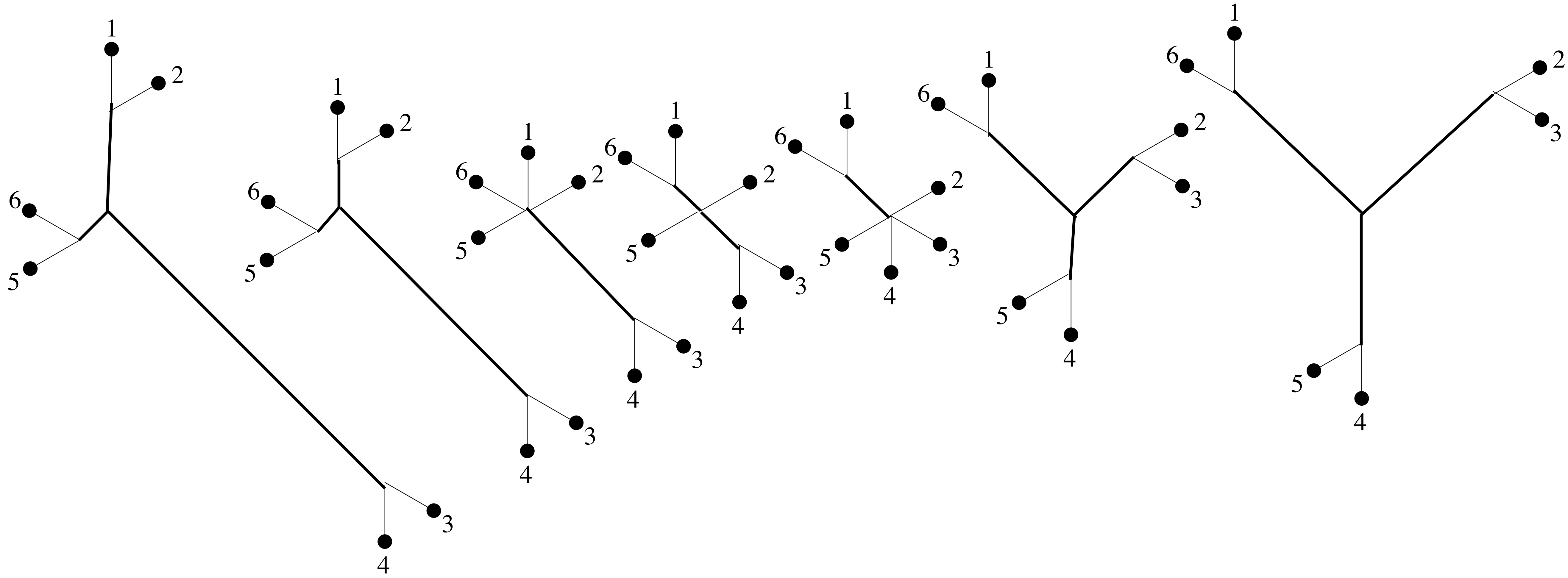
Structure of $\mathbb{T}_4$



Structure of $\mathbb{T}_4$



Geodesics



Tree Space Properties

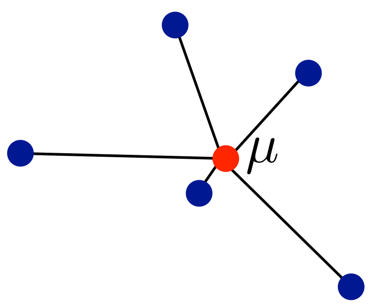
Theorem (Billera, Holmes,Vogtmann, 2001):
Tree space has global non-positive curvature.

$\Rightarrow$ unique geodesics (shortest paths)

$\Rightarrow$ well-defined mid-point tree

- BHV or geodesic distance = length of shortest path between two trees T_1 and T_2
- polynomial time algorithm to compute geodesic distance (O. and Provan, 2011)

Mean and Variance



- weighted set X in tree space:
    - Fréchet mean(X) = centre of mass

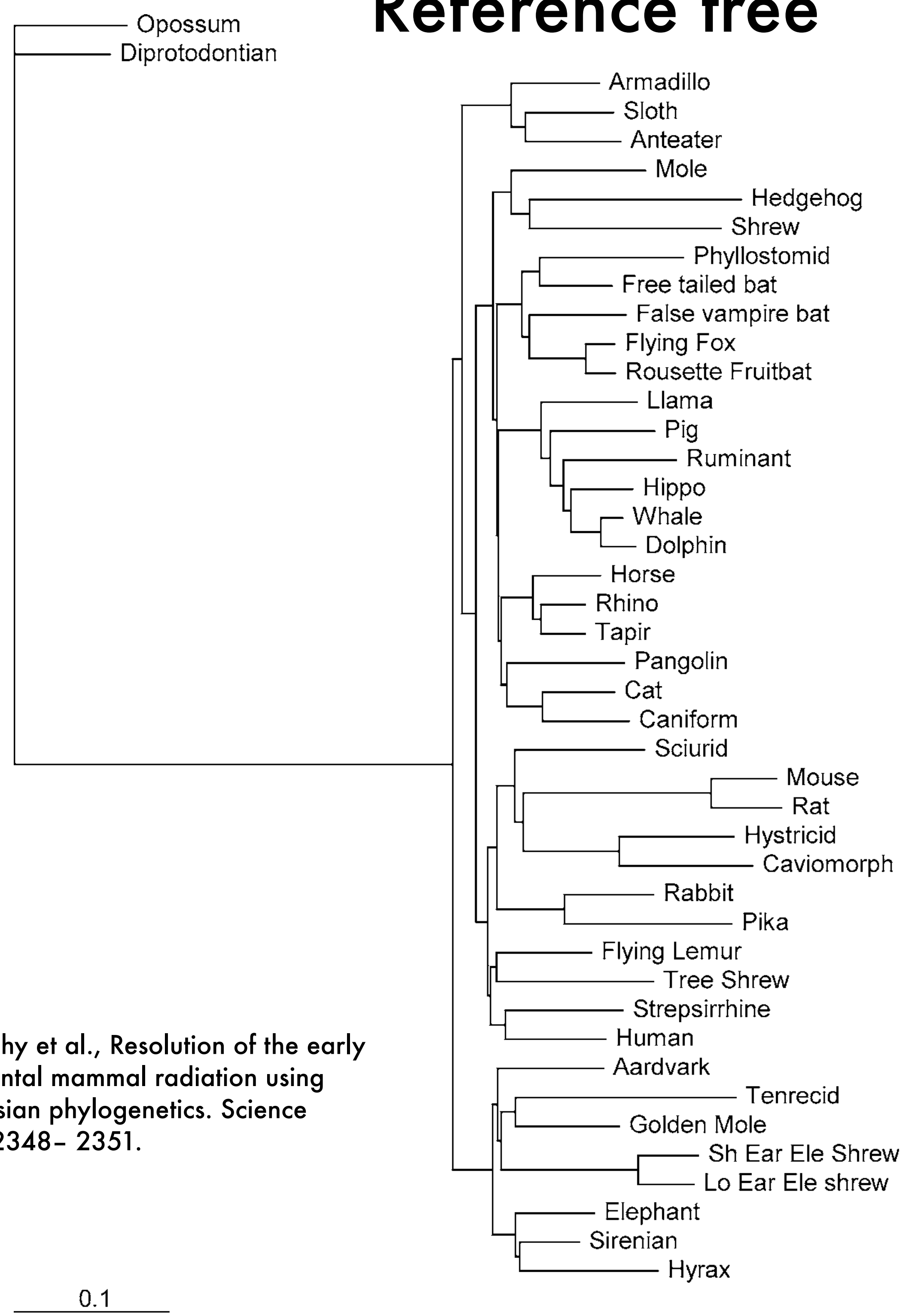
$$= \underset{\mu}{\operatorname{argmin}} \sum_{x \in X} p(x) d(x, \mu)^2$$

(tree minimizing sum of square BHV distances)

- variance(X) = $\sum_{x \in X} p(x) d(x, \mu)^2$
- computable by algorithm based on Law of Large Numbers (Sturm 2003; Miller, O, Provan 2015; Bačák 2014)

Experimental Results

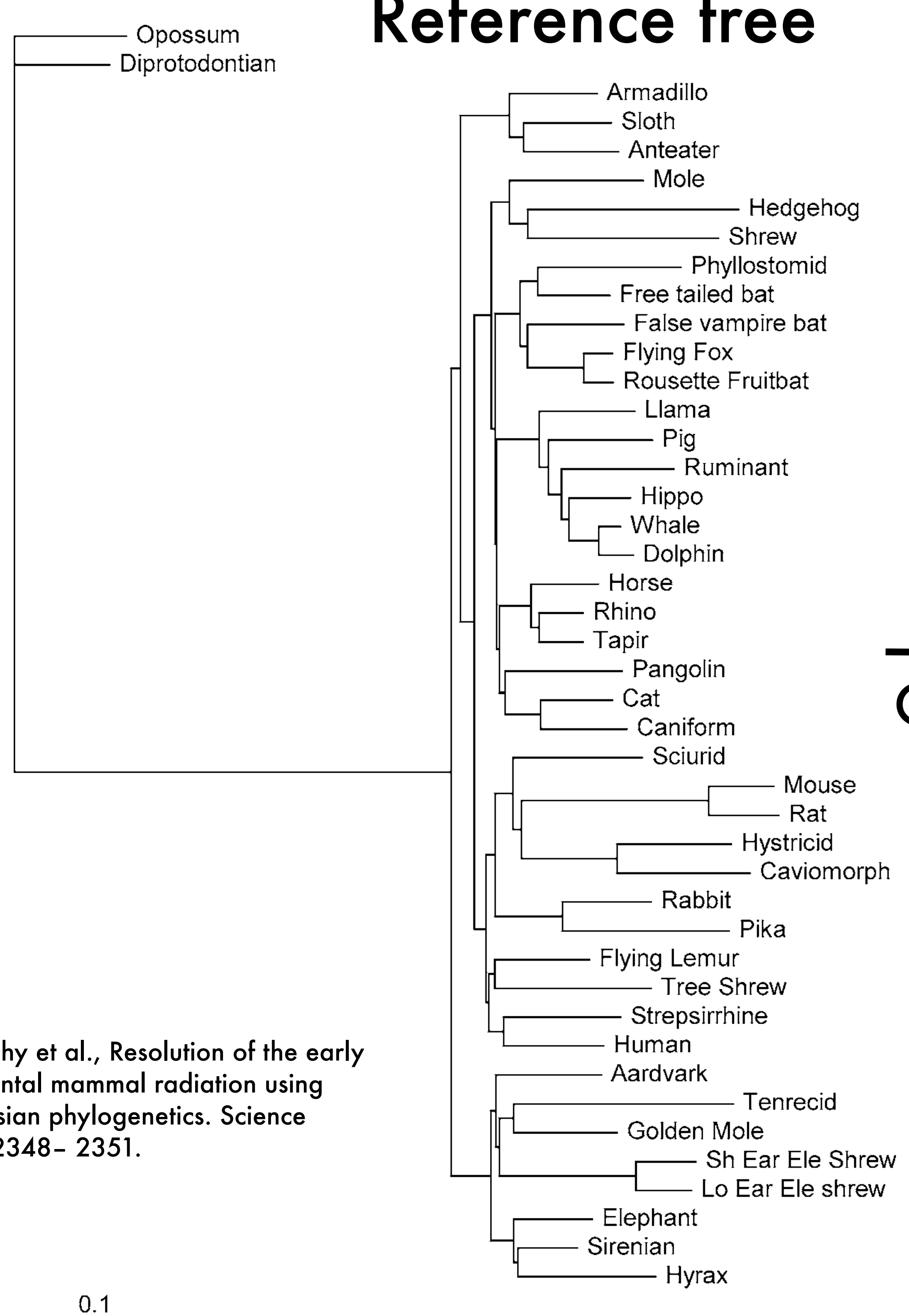
Reference tree



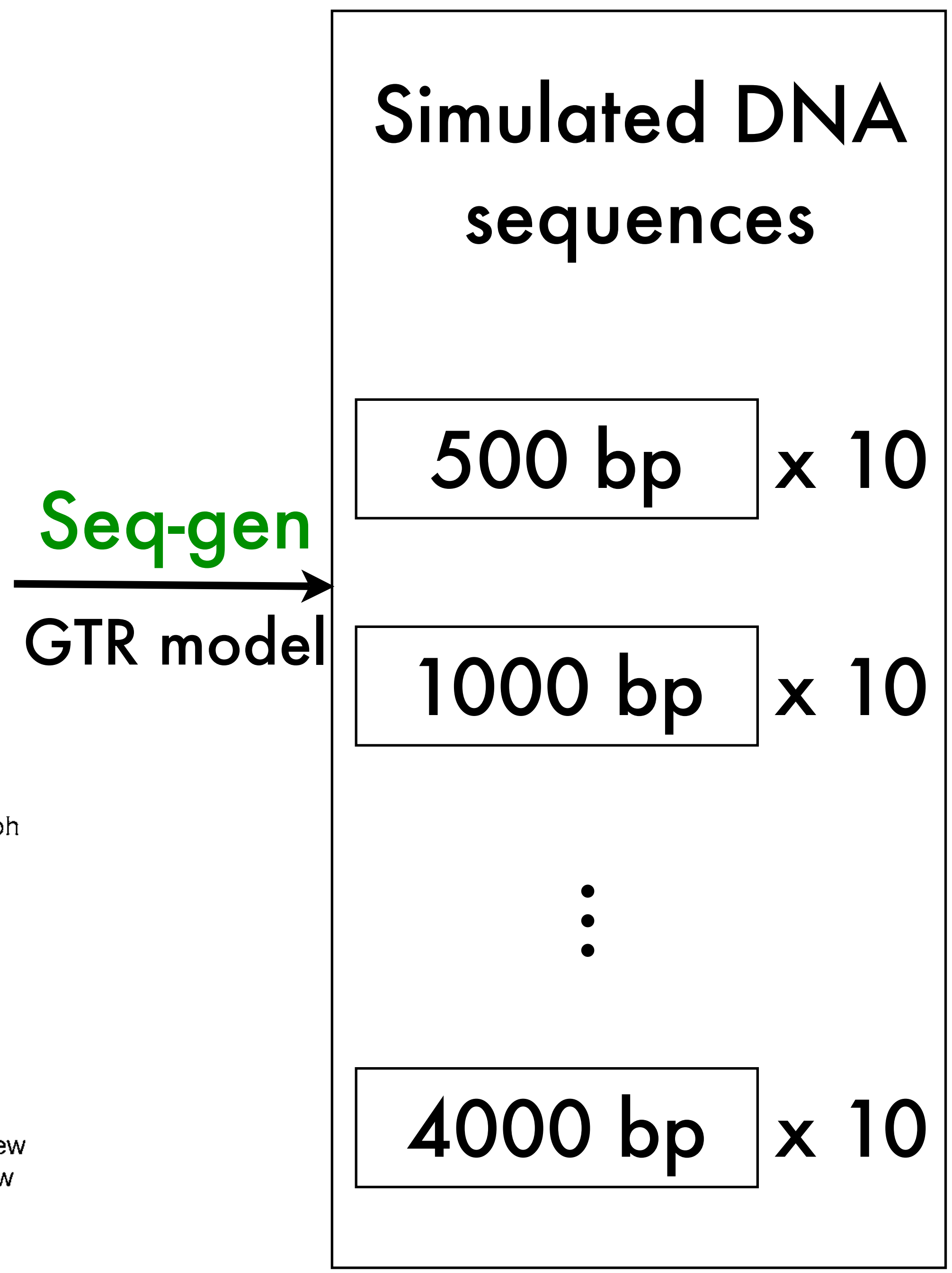
Murphy et al., Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294:2348– 2351.

Experimental Results

Reference tree



Murphy et al., Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294:2348– 2351.

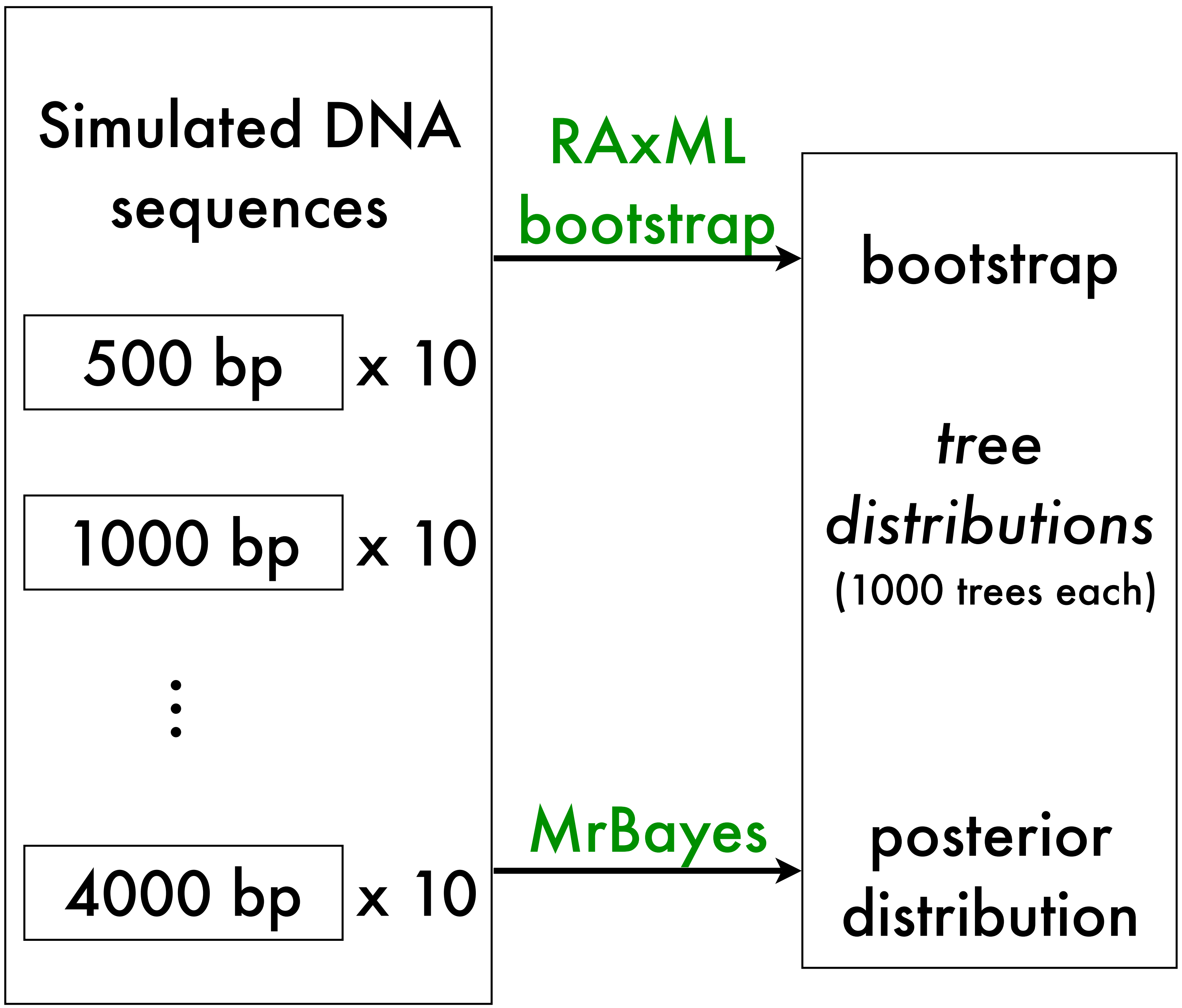


Experimental Results

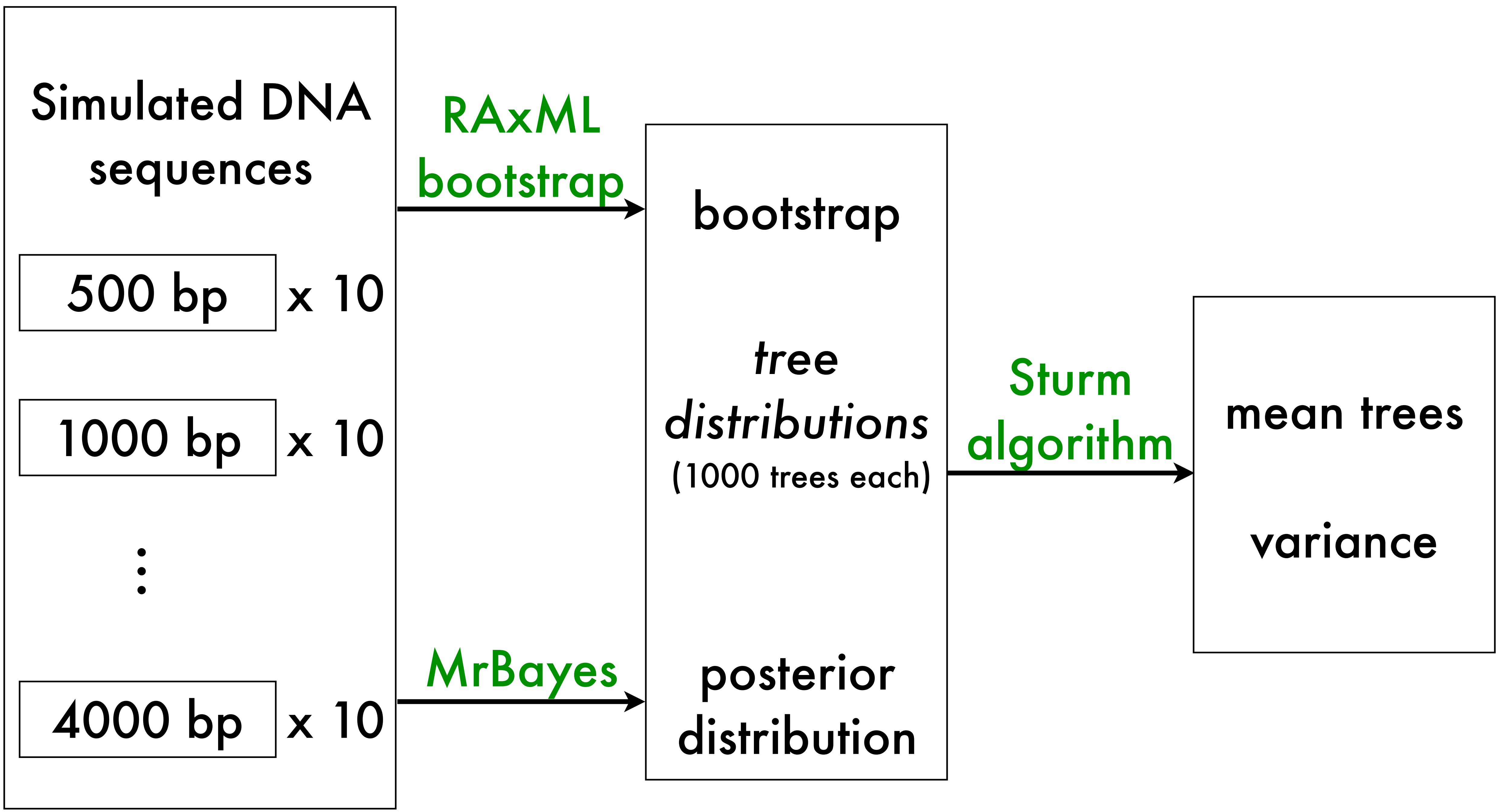
Simulated DNA sequences

500 bp	x 10
1000 bp	x 10
⋮	
4000 bp	x 10

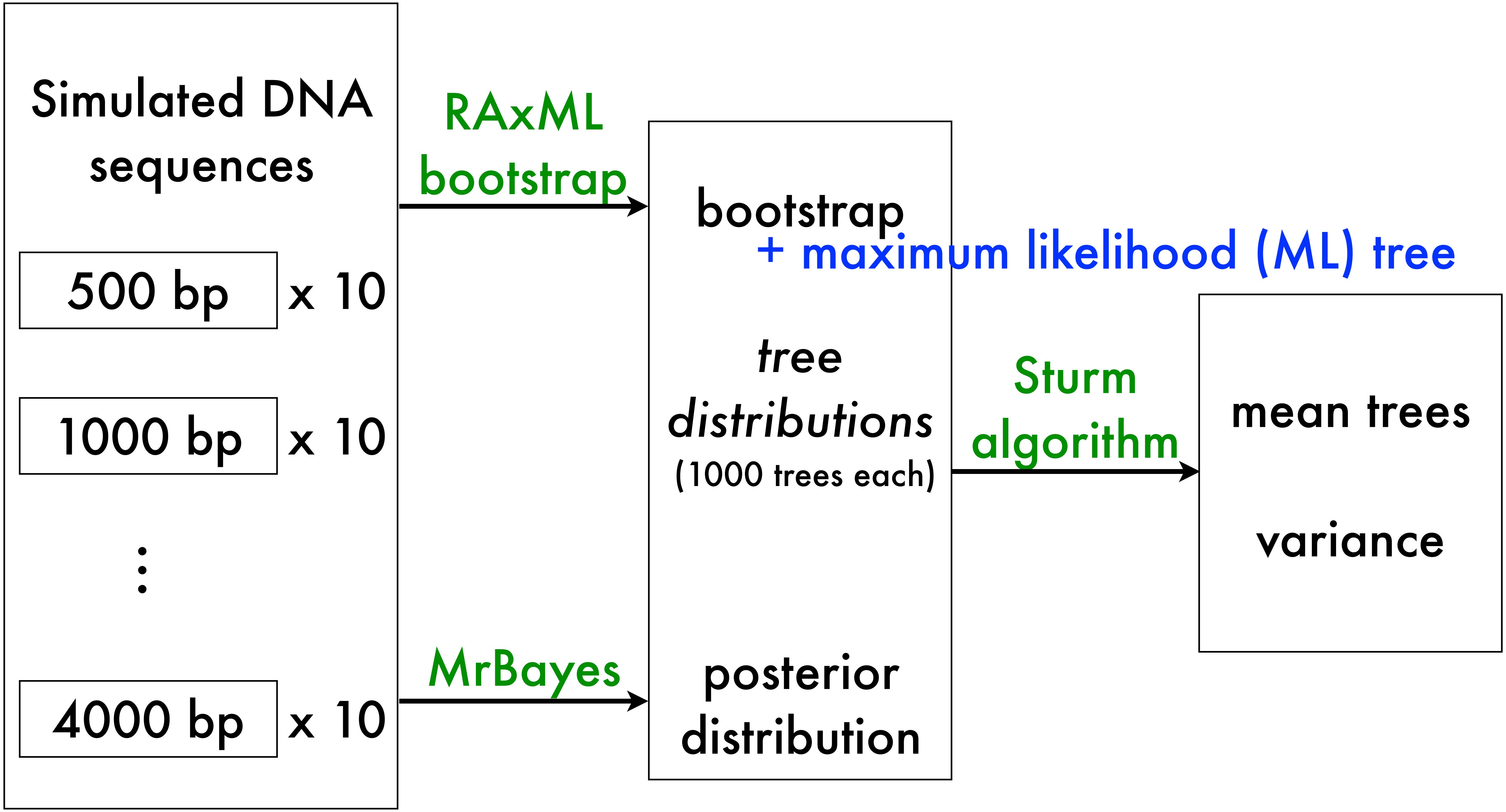
Experimental Results



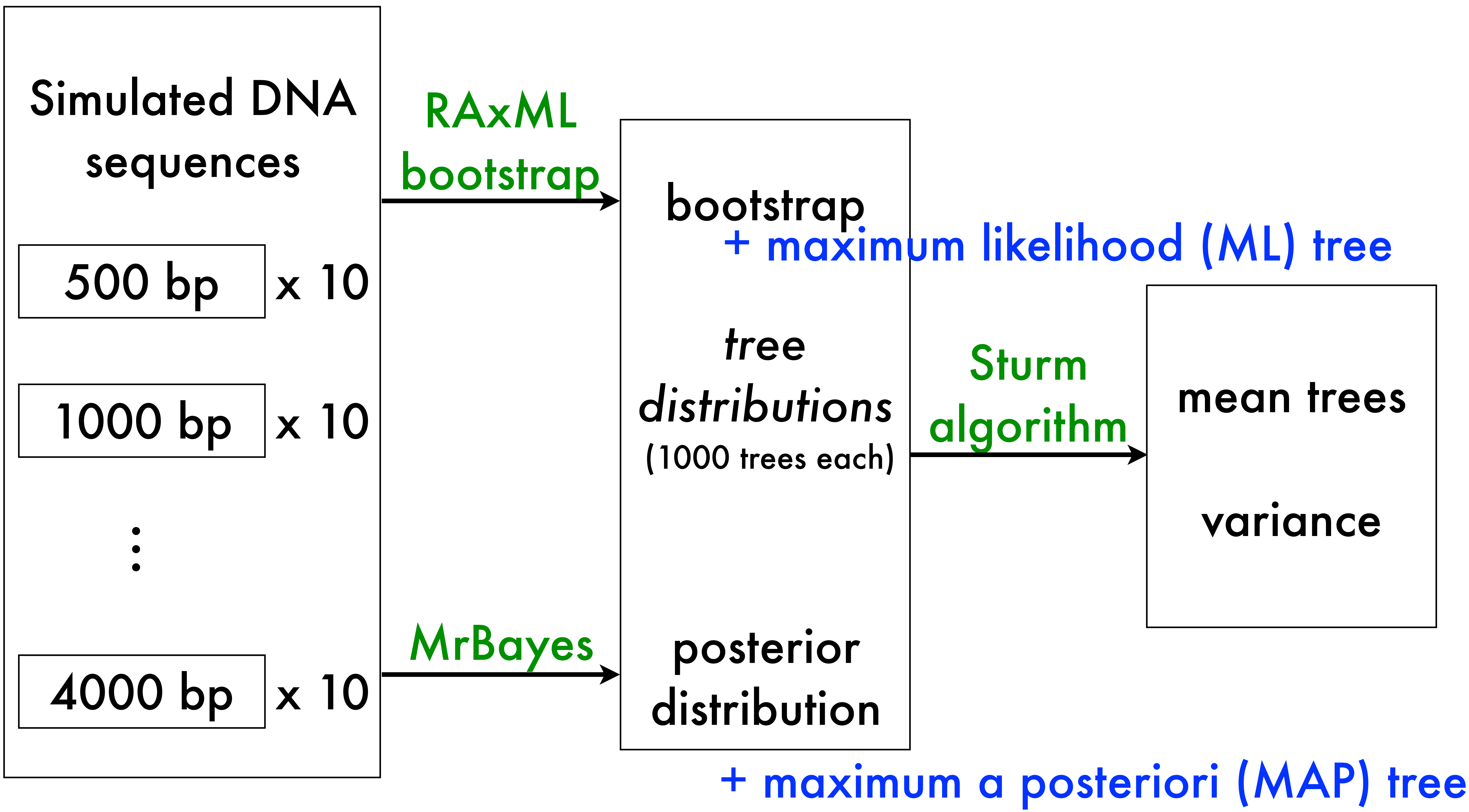
Experimental Results



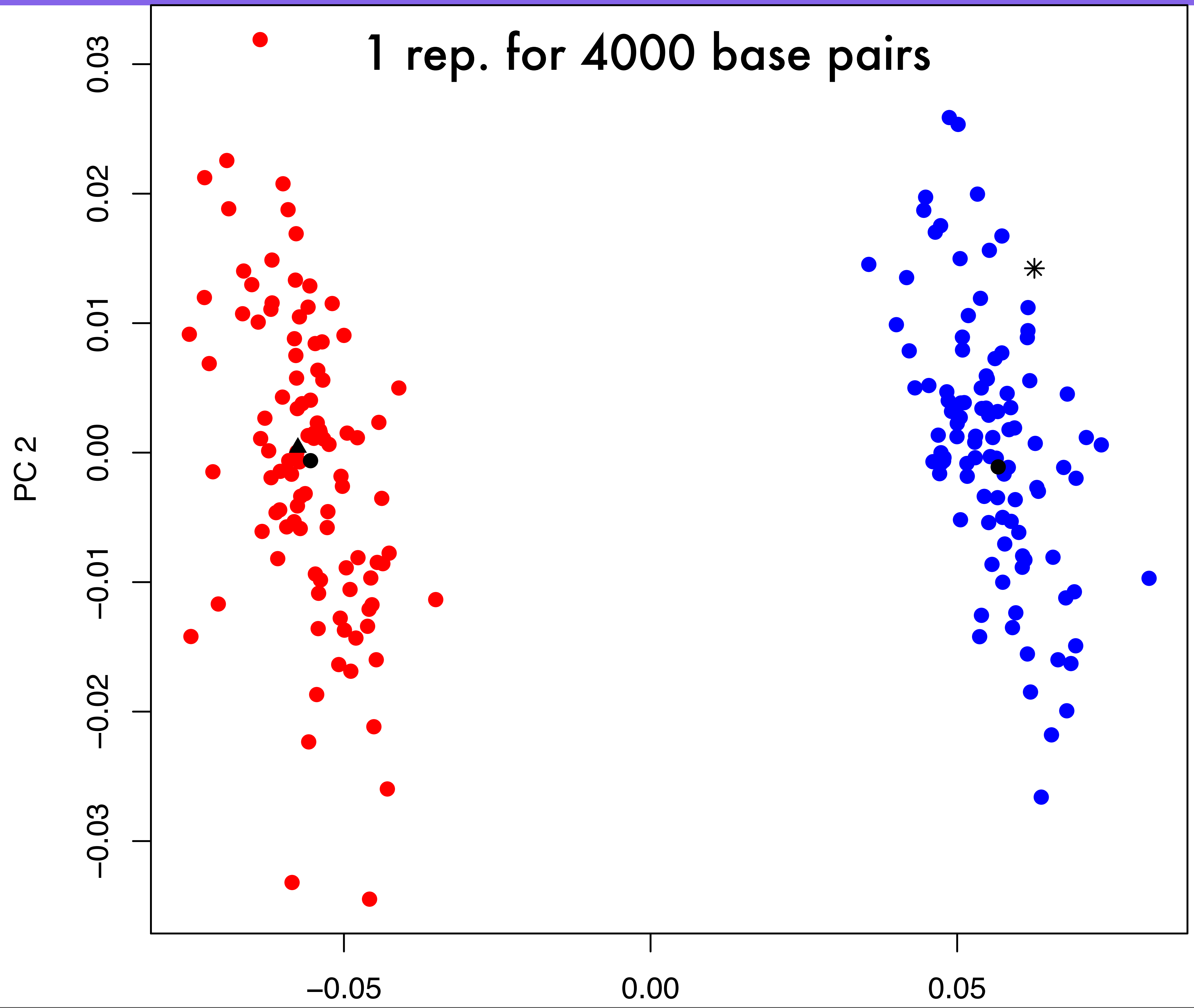
Experimental Results



Experimental Results

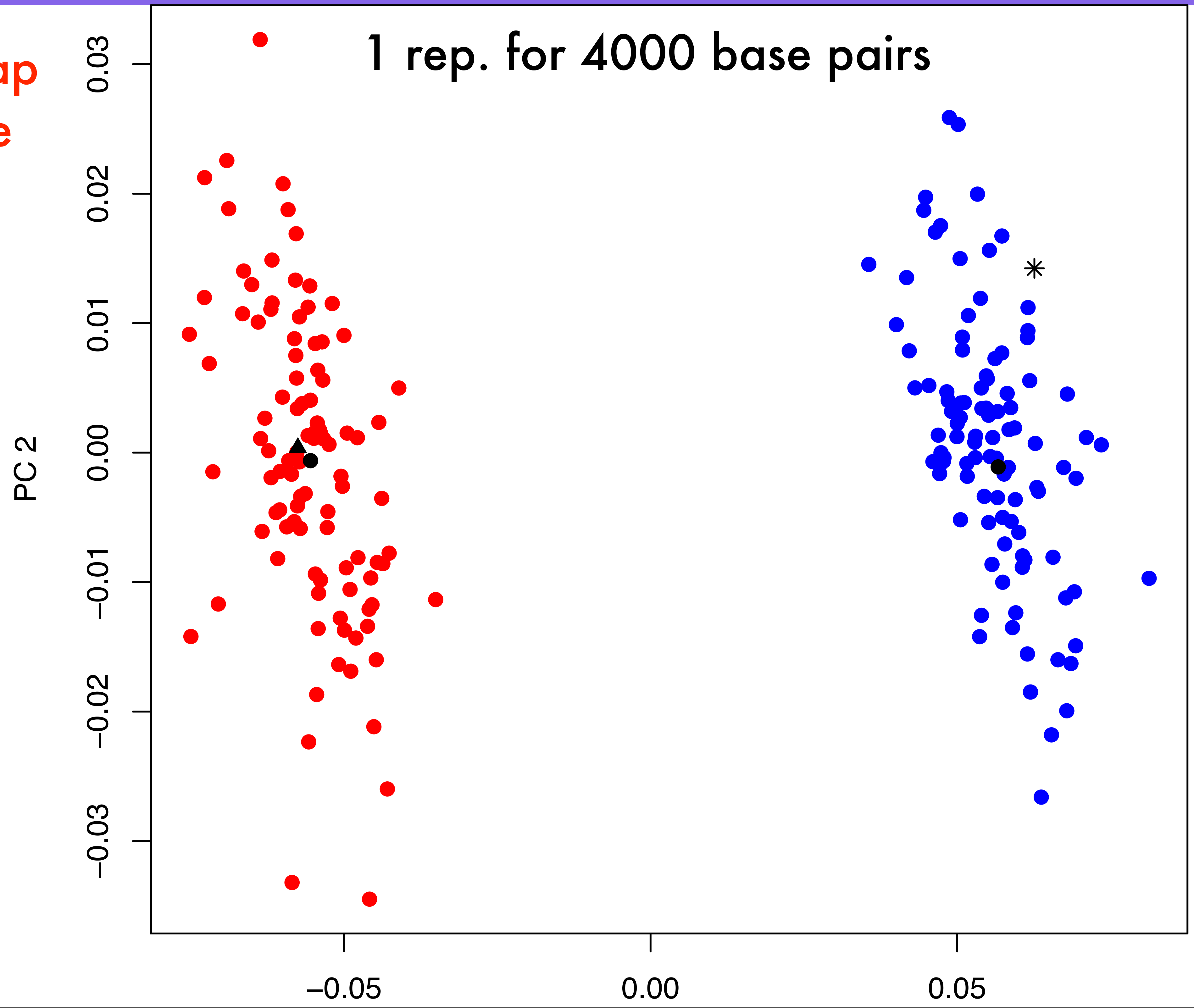


Visualization vis MDS

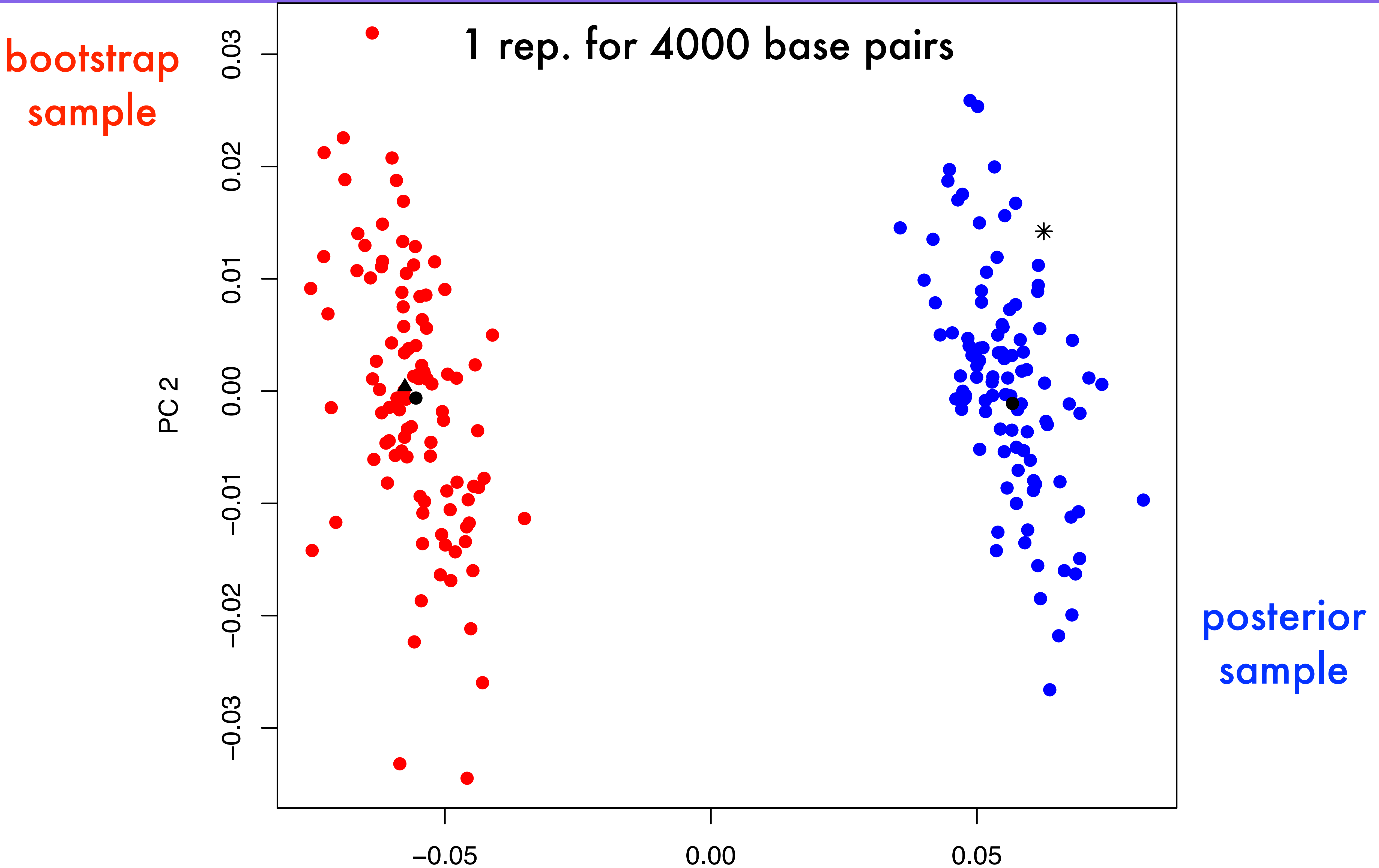


Visualization vis MDS

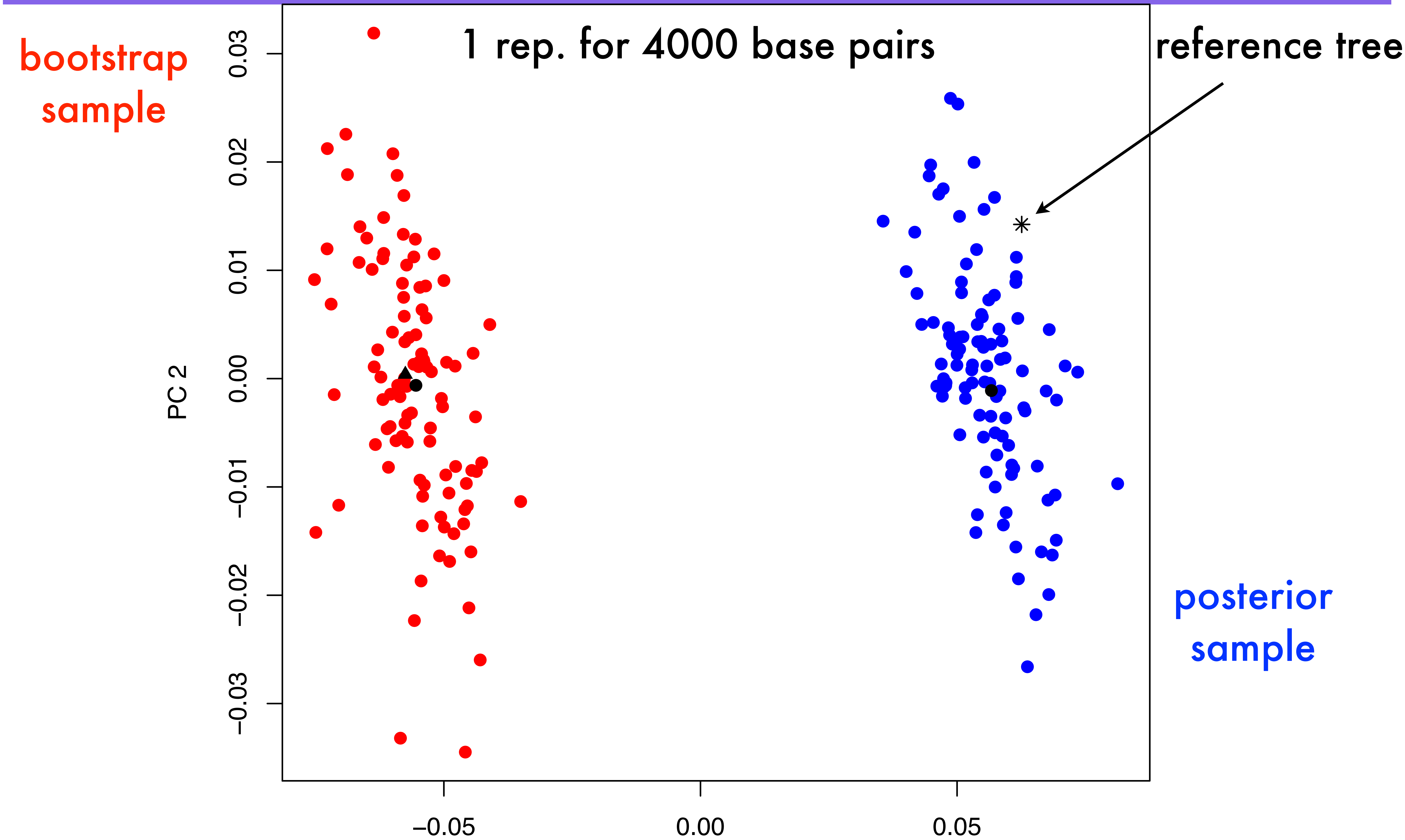
bootstrap sample



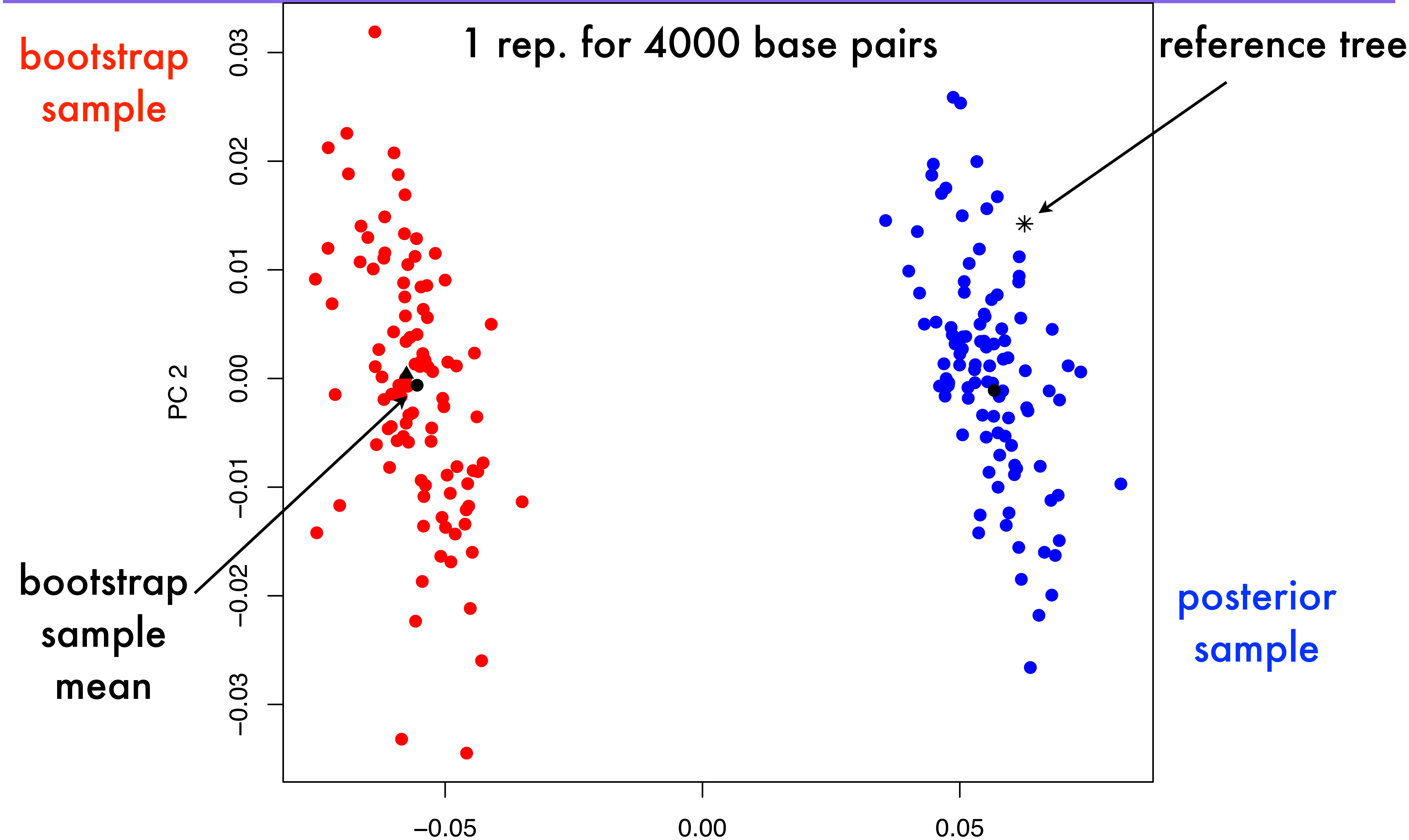
Visualization vis MDS



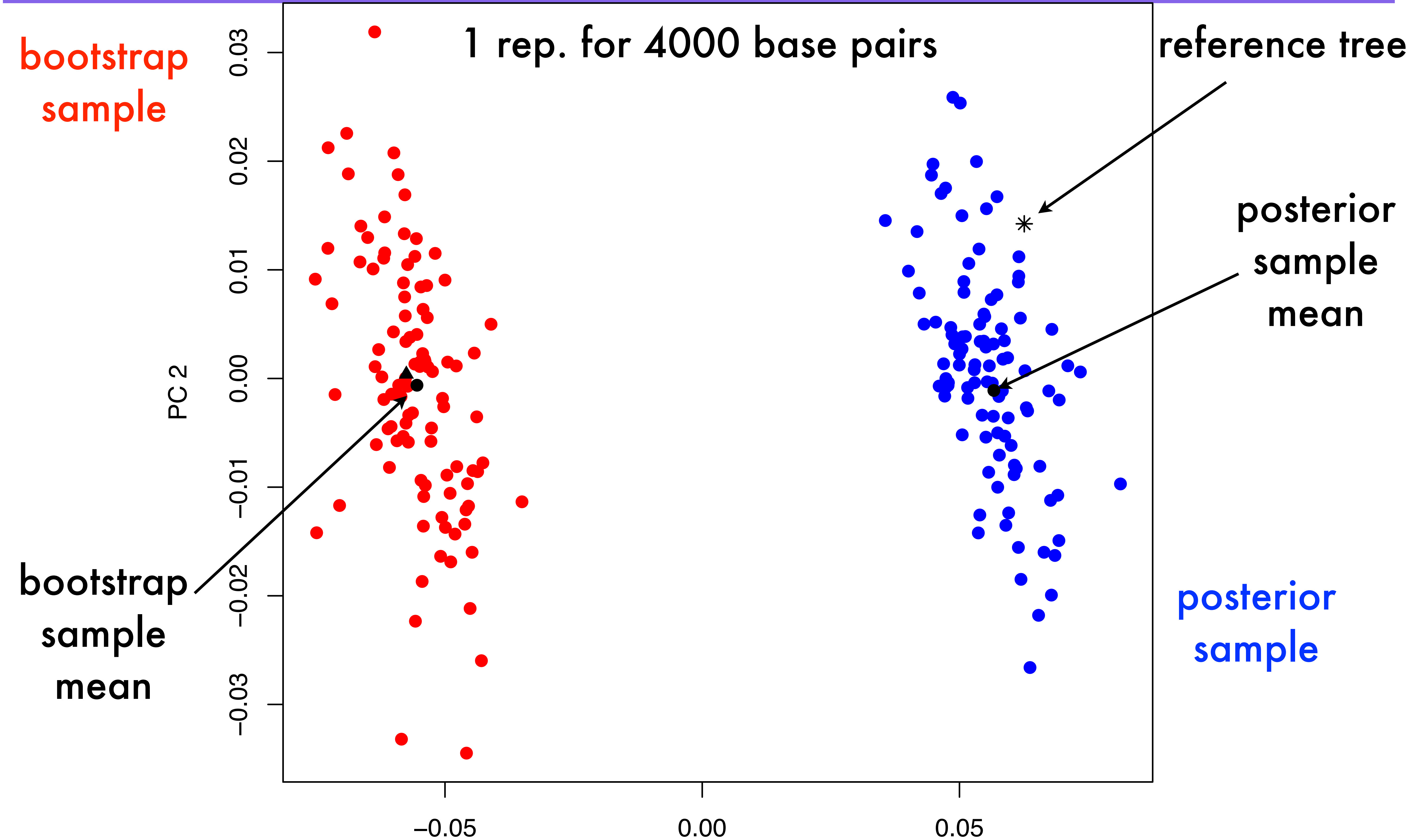
Visualization vis MDS



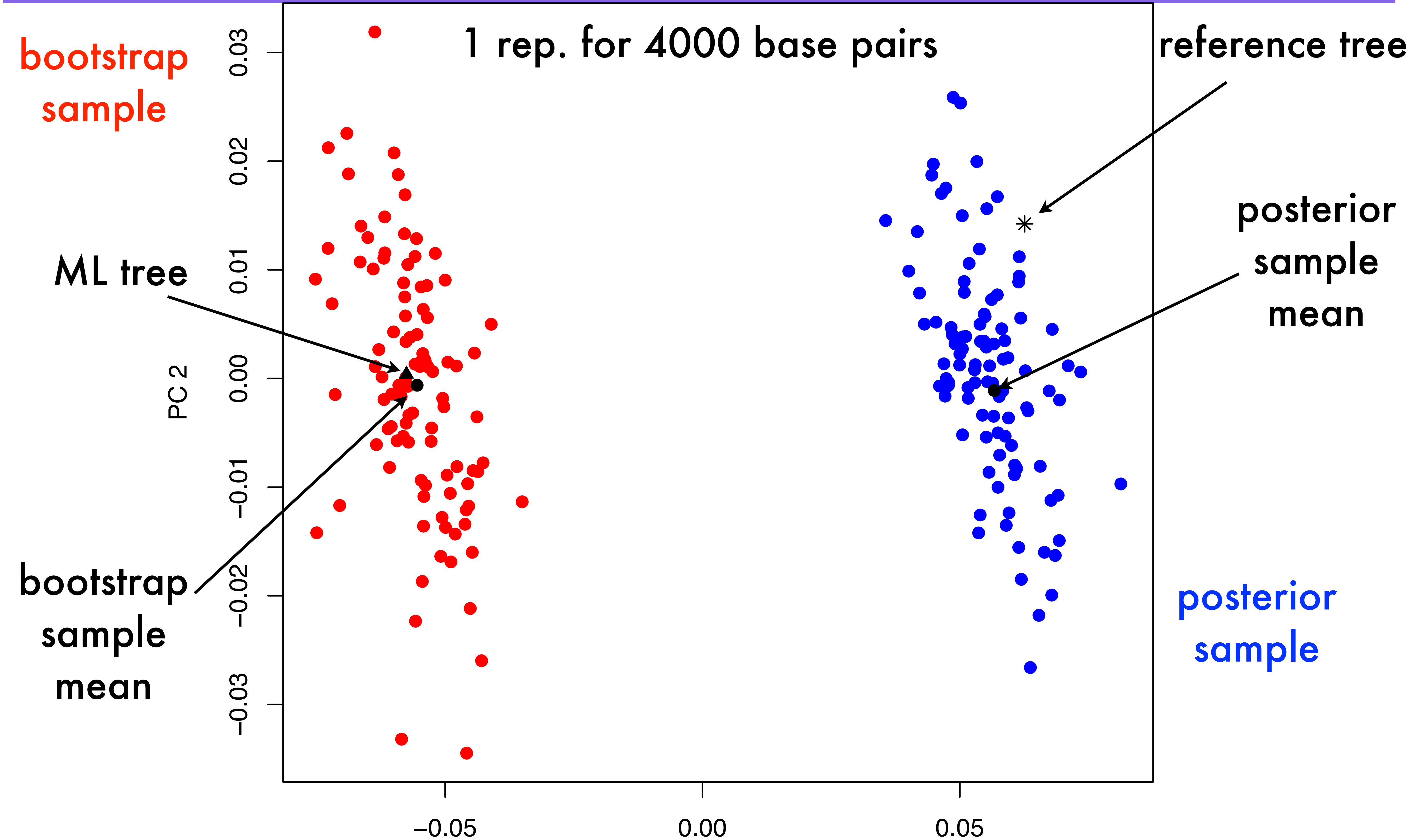
Visualization vis MDS

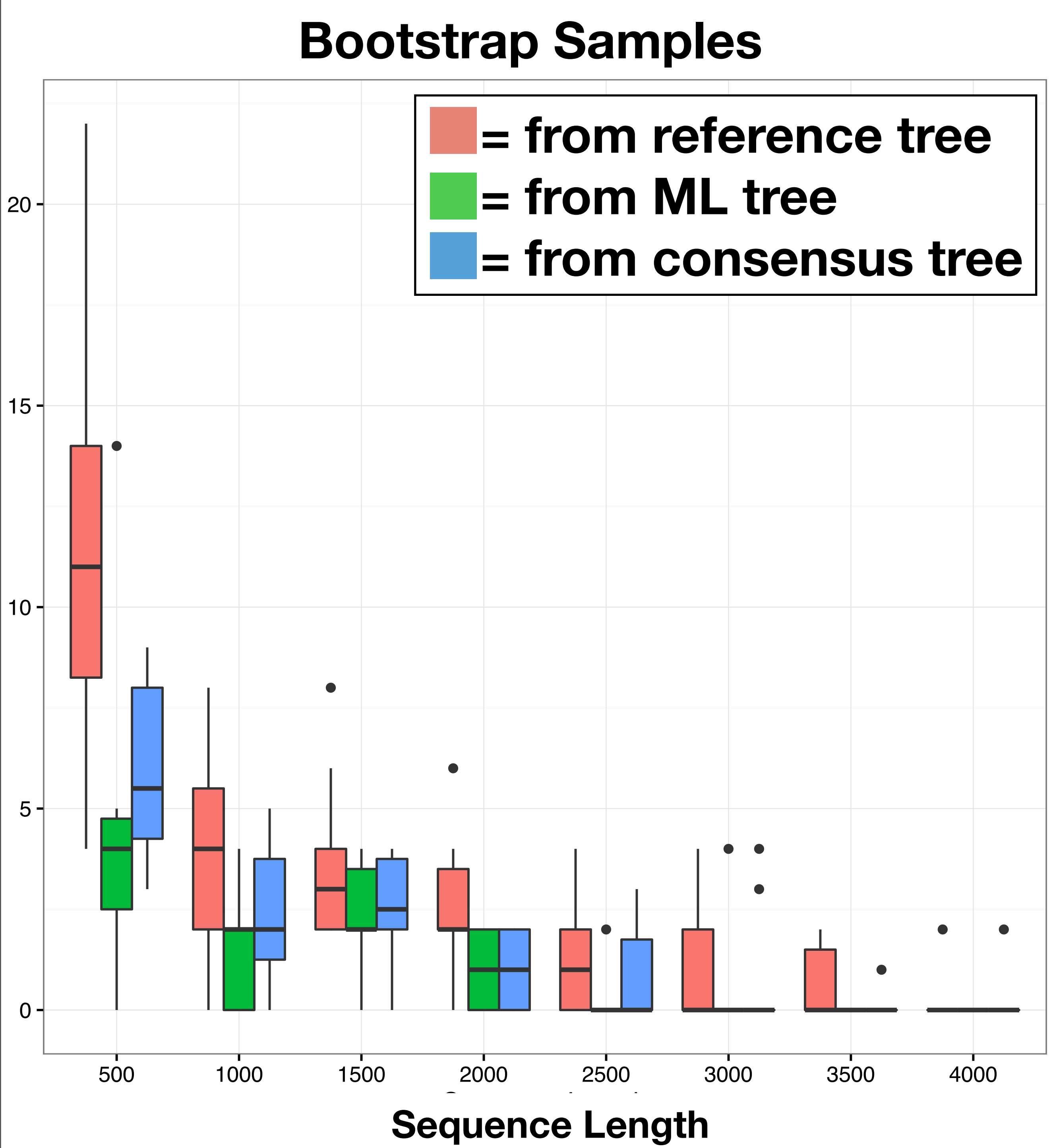


Visualization vis MDS

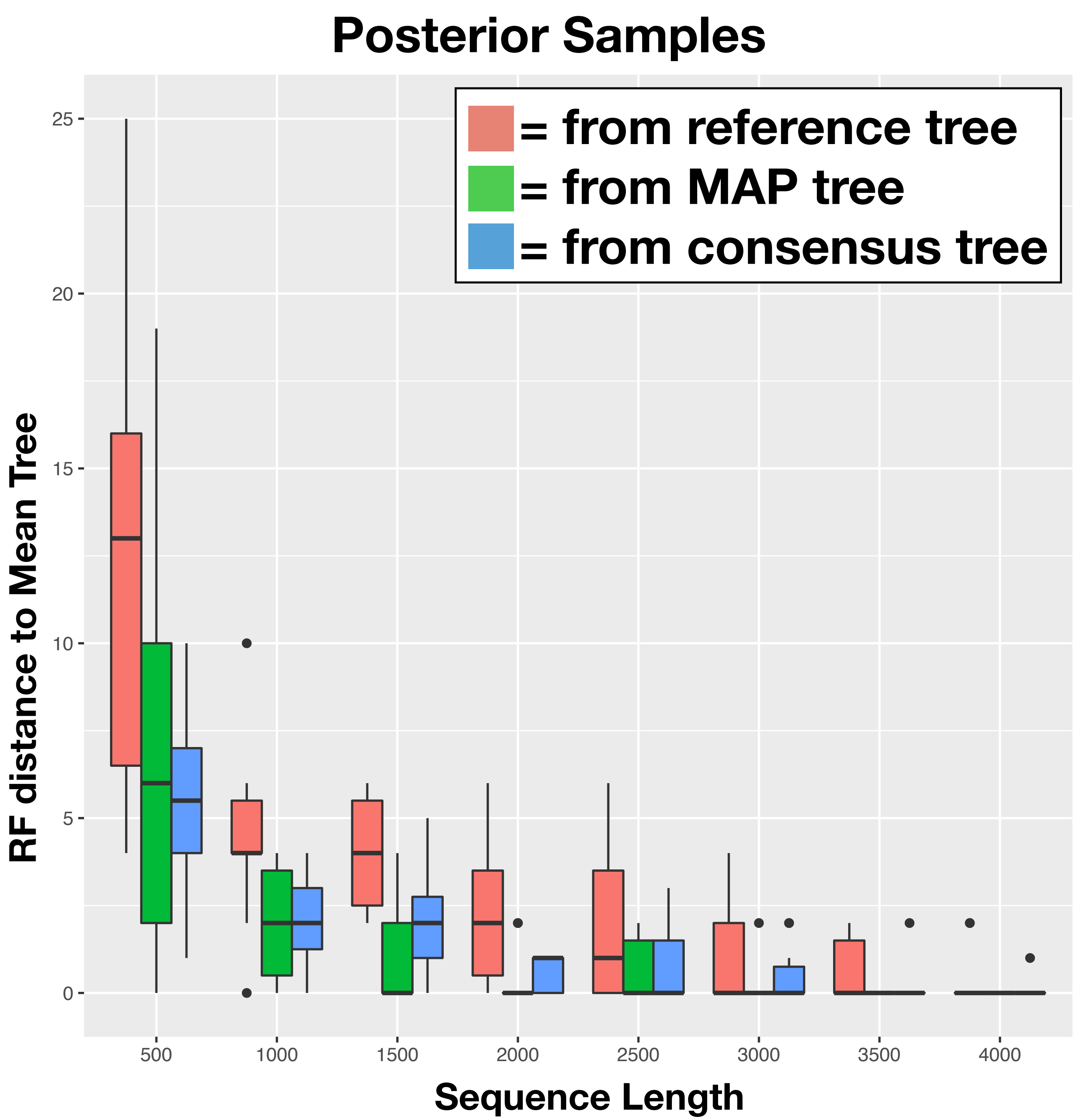


Visualization vis MDS

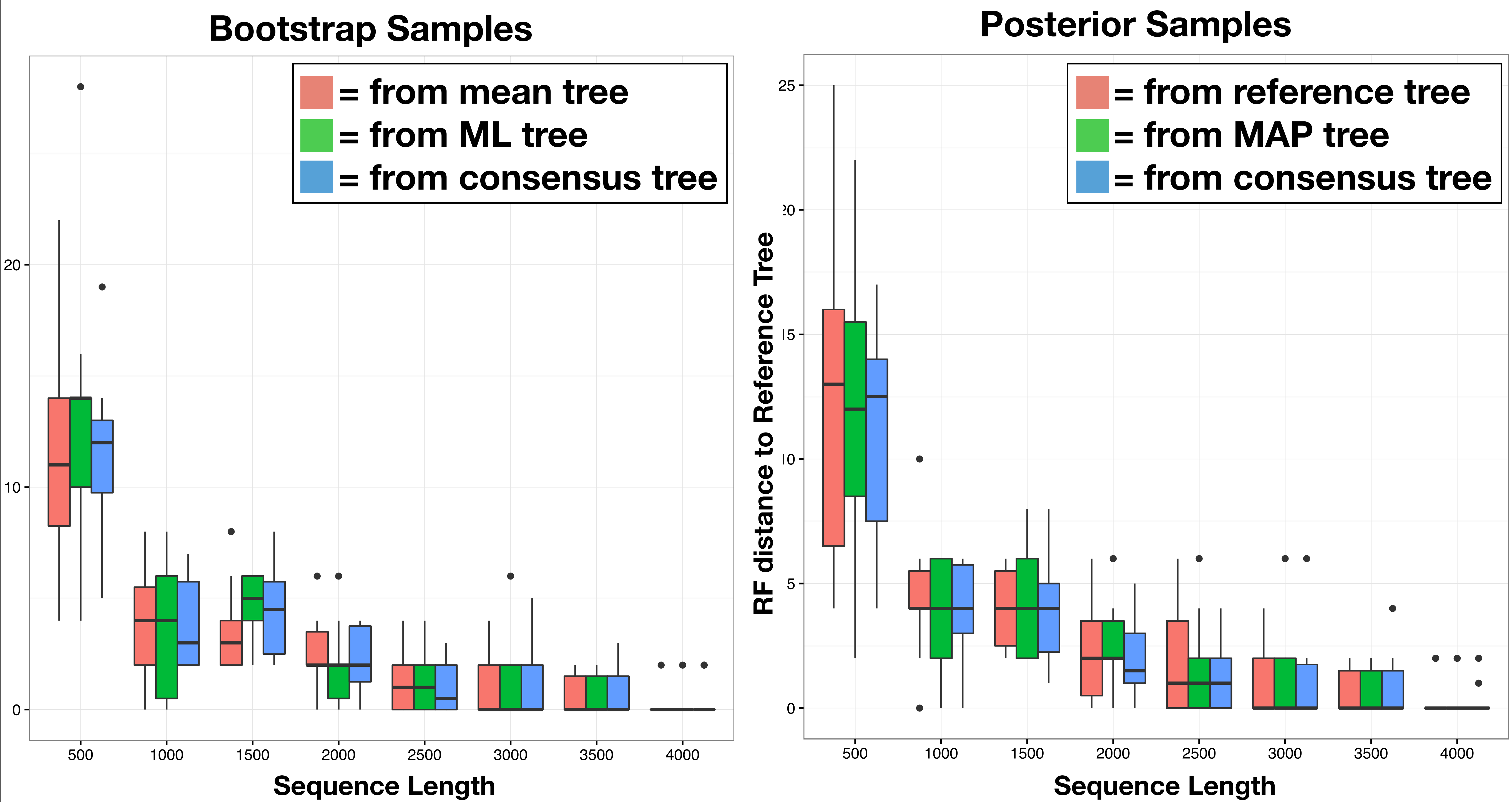




Bootstrap Samples
= from reference tree
= from ML tree
= from consensus tree
0
5
10
15
20
500
1000
1500
2000
2500
3000
3500
4000
Sequence Length



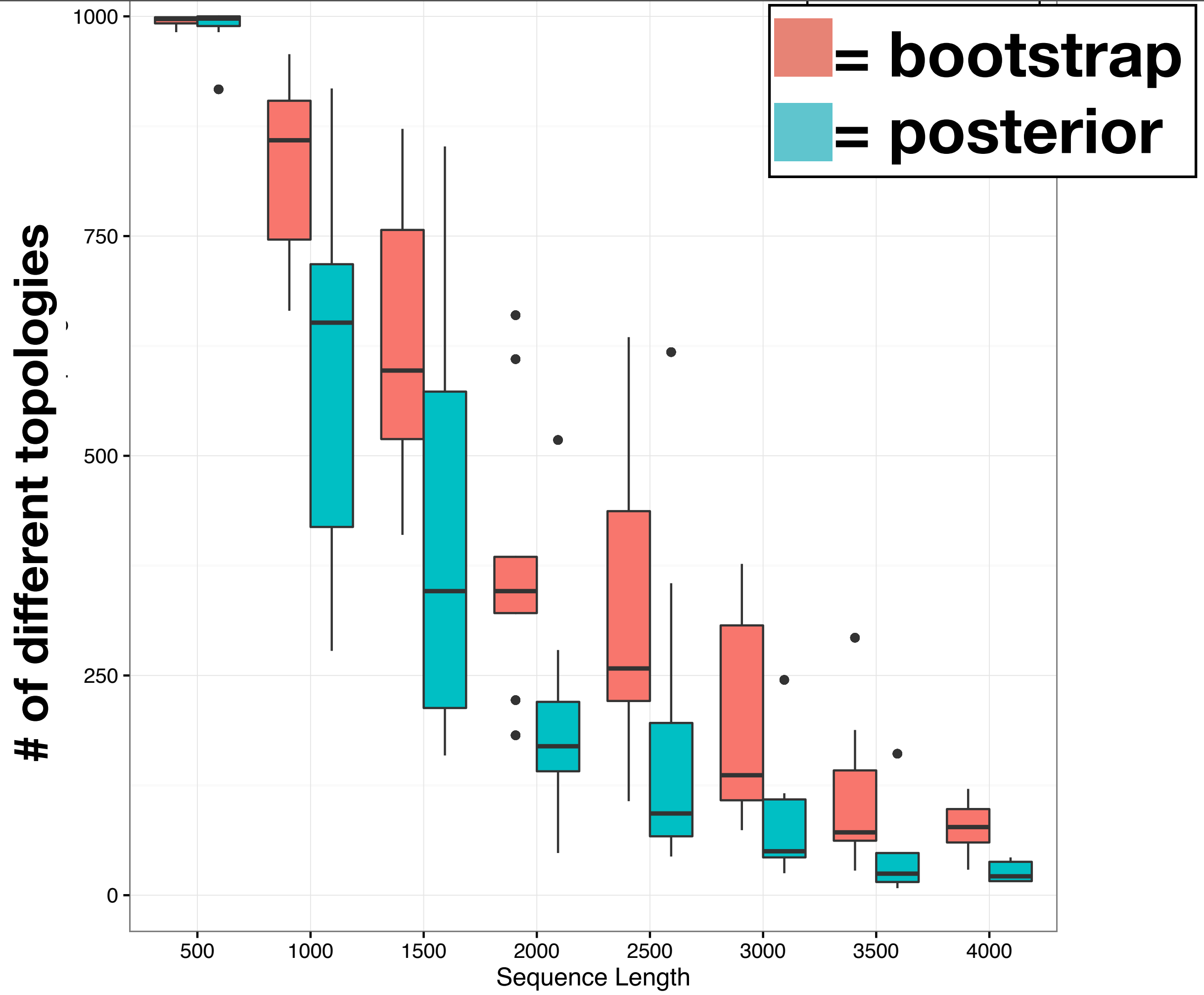
Posterior Samples
= from reference tree
= from MAP tree
= from consensus tree
RF distance to Mean Tree
0
5
10
15
20
25
500
1000
1500
2000
2500
3000
3500
4000
Sequence Length

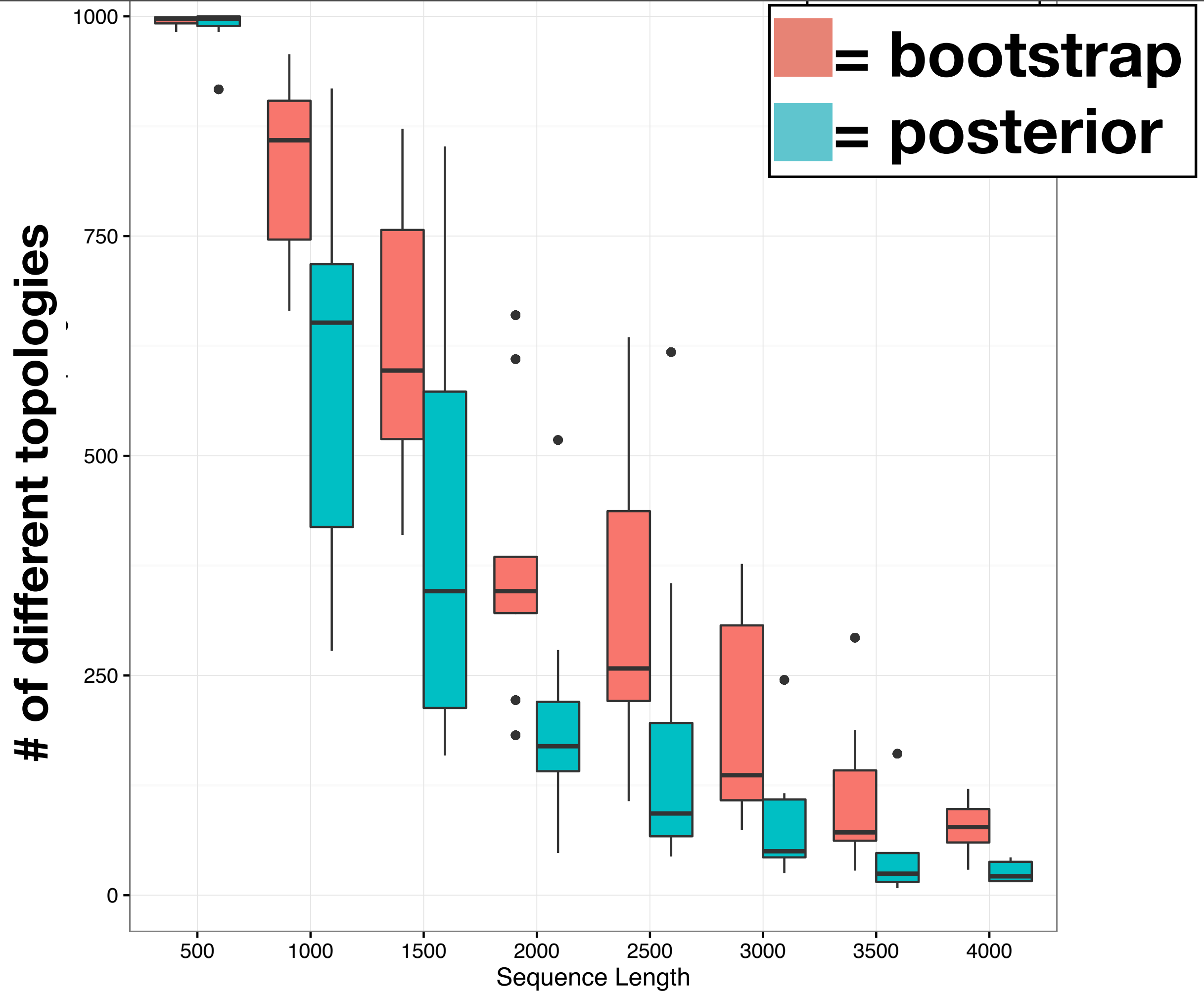

Bootstrap Samples
= from mean tree
= from ML tree
= from consensus tree
0
10
20
500
1000
1500
2000
2500
3000
3500
4000
Sequence Length
Posterior Samples
= from reference tree
= from MAP tree
= from consensus tree
RF distance to Reference Tree
0
5
25
500
1000
1500
2000
2500
3000
3500
4000
Sequence Length

Measures of Variance

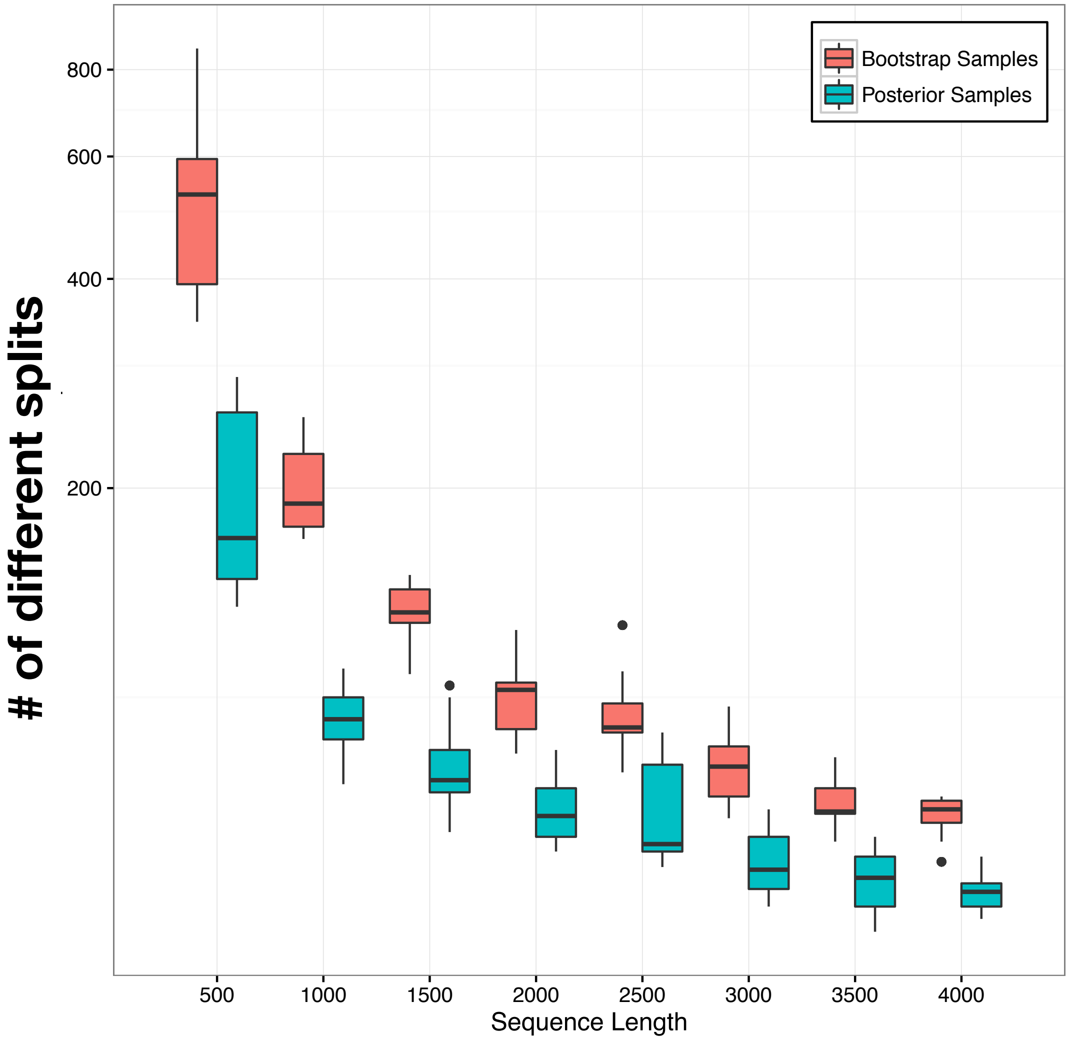
- # of different topologies in sample
- # of different splits in sample
- sum of squared distances between trees

$$\sum_{T,T' \in \mathcal{T}} d(T,T')^2$$

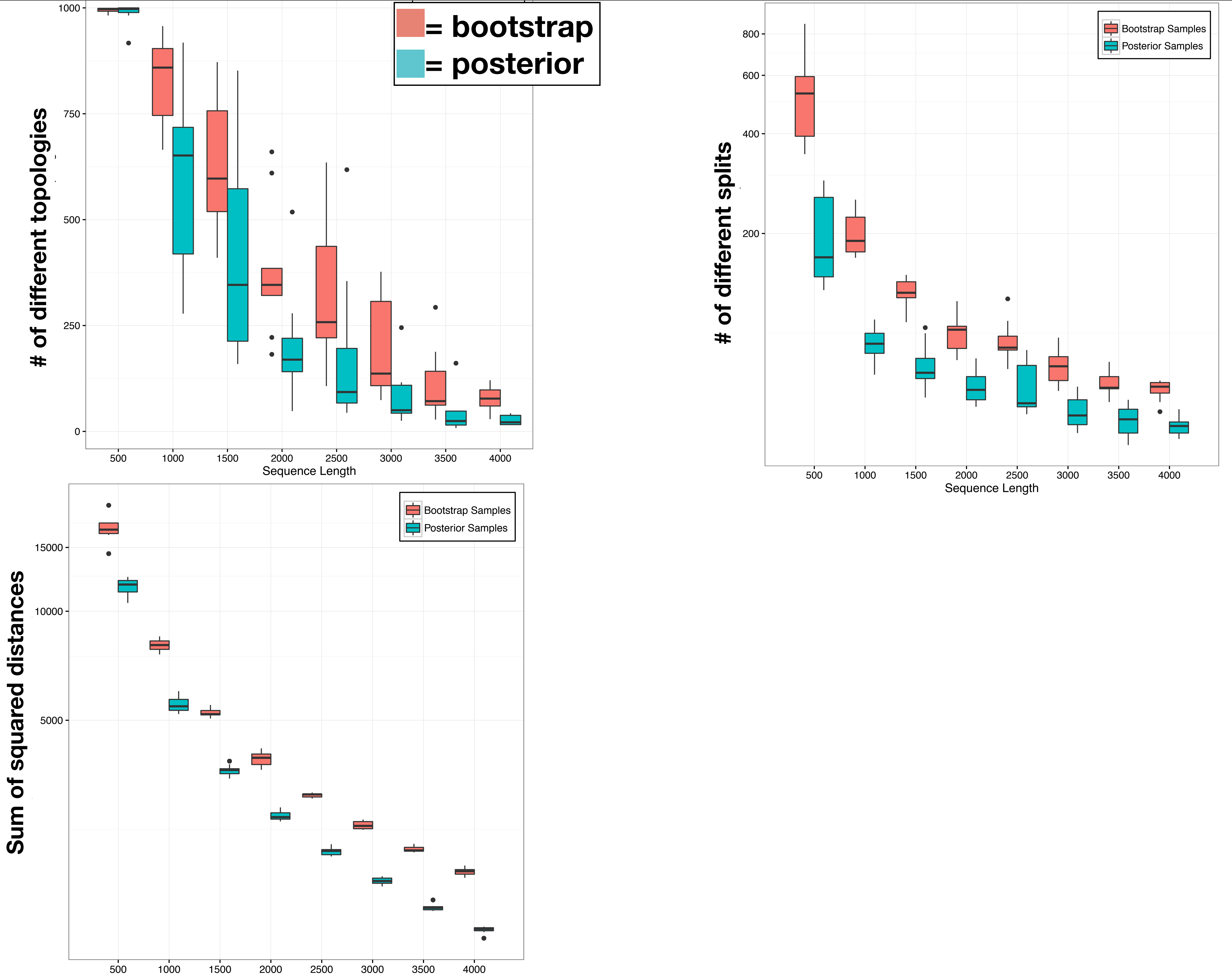

= bootstrap
= posterior
# of different topologies
1000
750
500
250
0
500
1000
1500
2000
2500
3000
3500
4000
Sequence Length



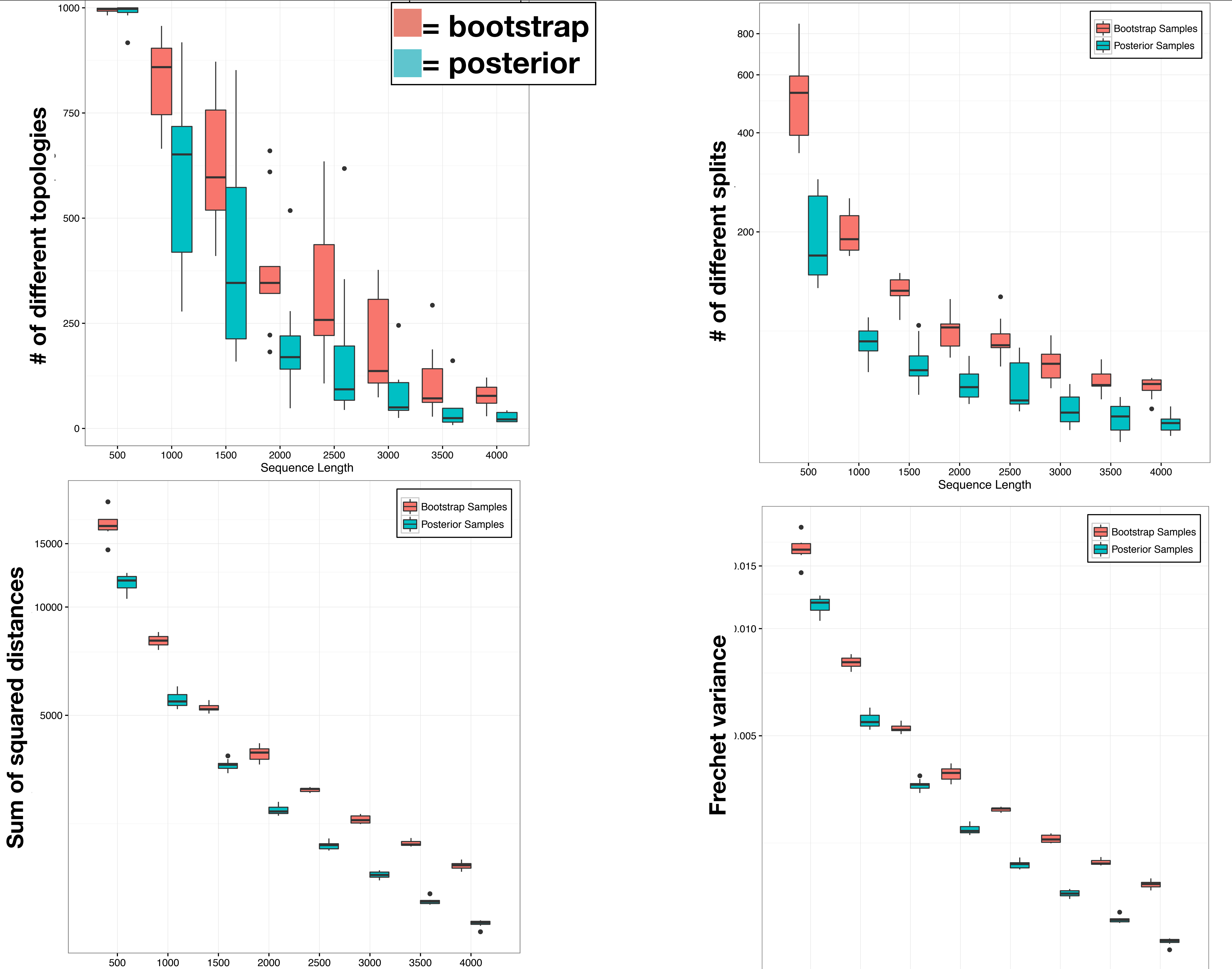
= bootstrap
= posterior
# of different topologies
Sequence Length



Bootstrap Samples
Posterior Samples
# of different splits
Sequence Length

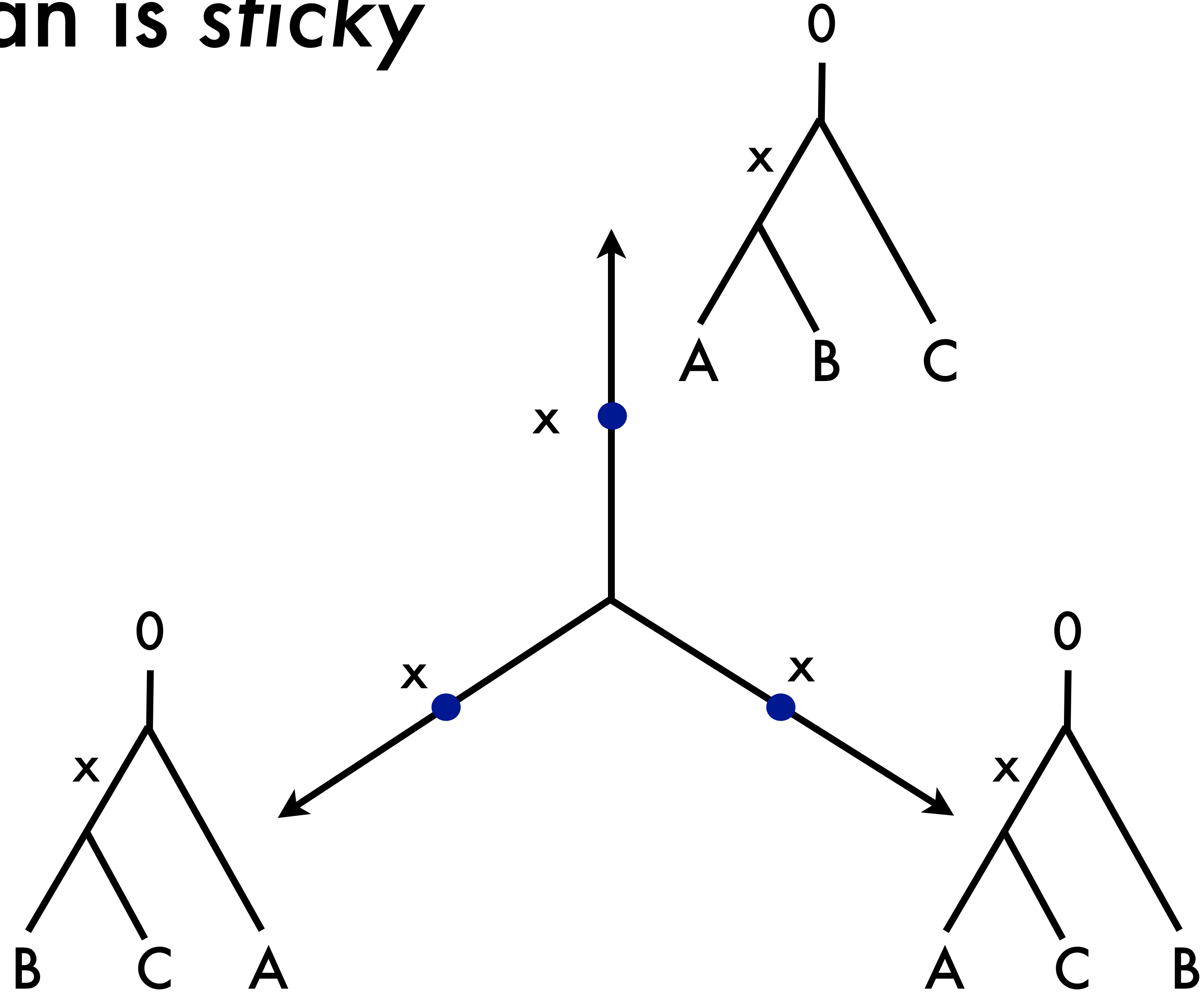


= bootstrap
= posterior
# of different topologies
1000
750
500
250
0
500
1000
1500
2000
2500
3000
3500
4000
Sequence Length
Bootstrap Samples
Posterior Samples
# of different splits
800
600
400
200
500
1000
1500
2000
2500
3000
3500
4000
Sequence Length
Bootstrap Samples
Posterior Samples
Sum of squared distances
15000
10000
5000
500
1000
1500
2000
2500
3000
3500
4000


= bootstrap
= posterior
# of different topologies
1000
750
500
250
0
500
1000
1500
2000
2500
3000
3500
4000
Sequence Length
Bootstrap Samples
Posterior Samples
# of different splits
800
600
400
200
500
1000
1500
2000
2500
3000
3500
4000
Sequence Length
Bootstrap Samples
Posterior Samples
Sum of squared distances
15000
10000
5000
500
1000
1500
2000
2500
3000
3500
4000
Bootstrap Samples
Posterior Samples
Frechet variance
0.015
0.010
0.005

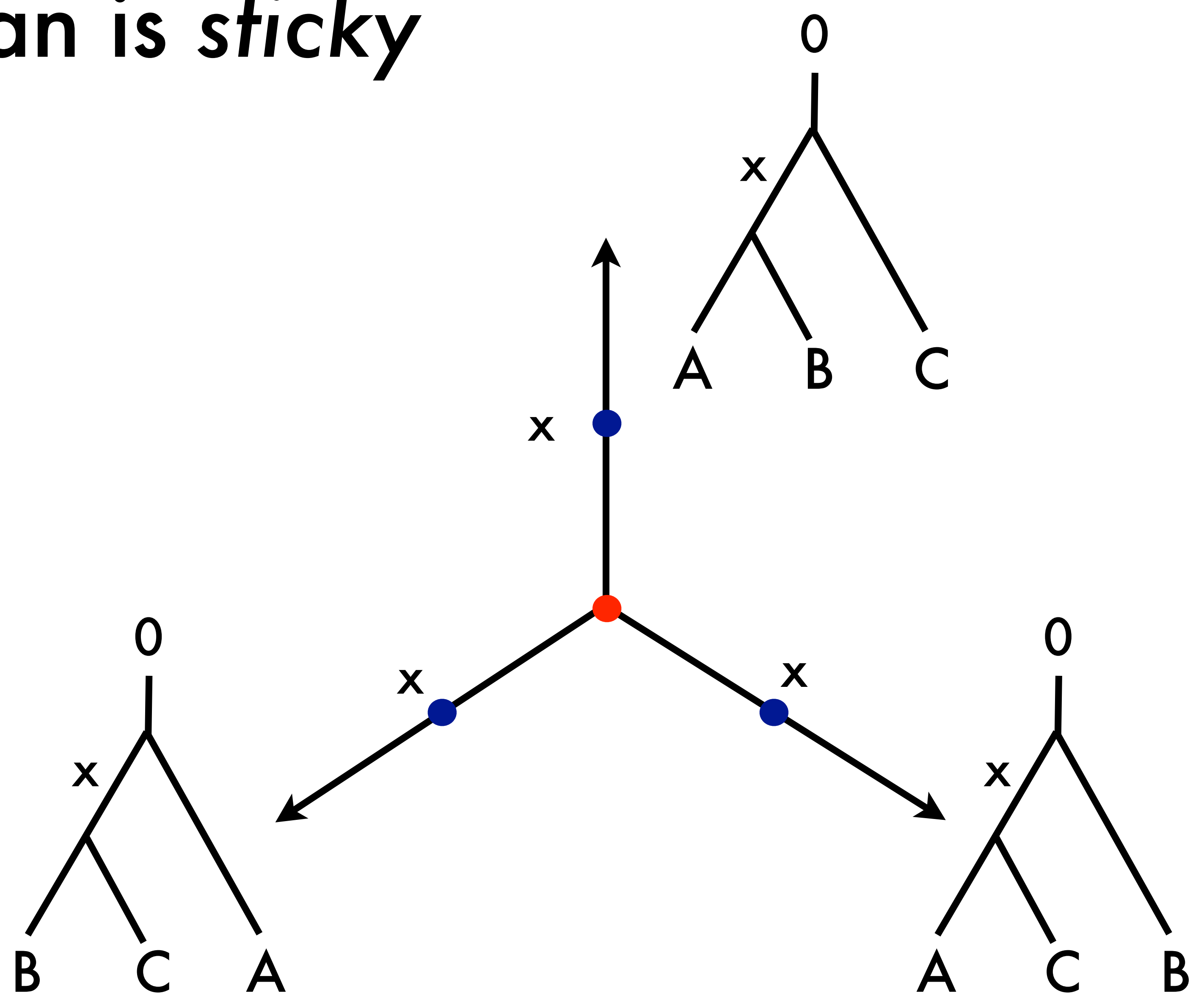
Caveat

- Mean is *sticky*



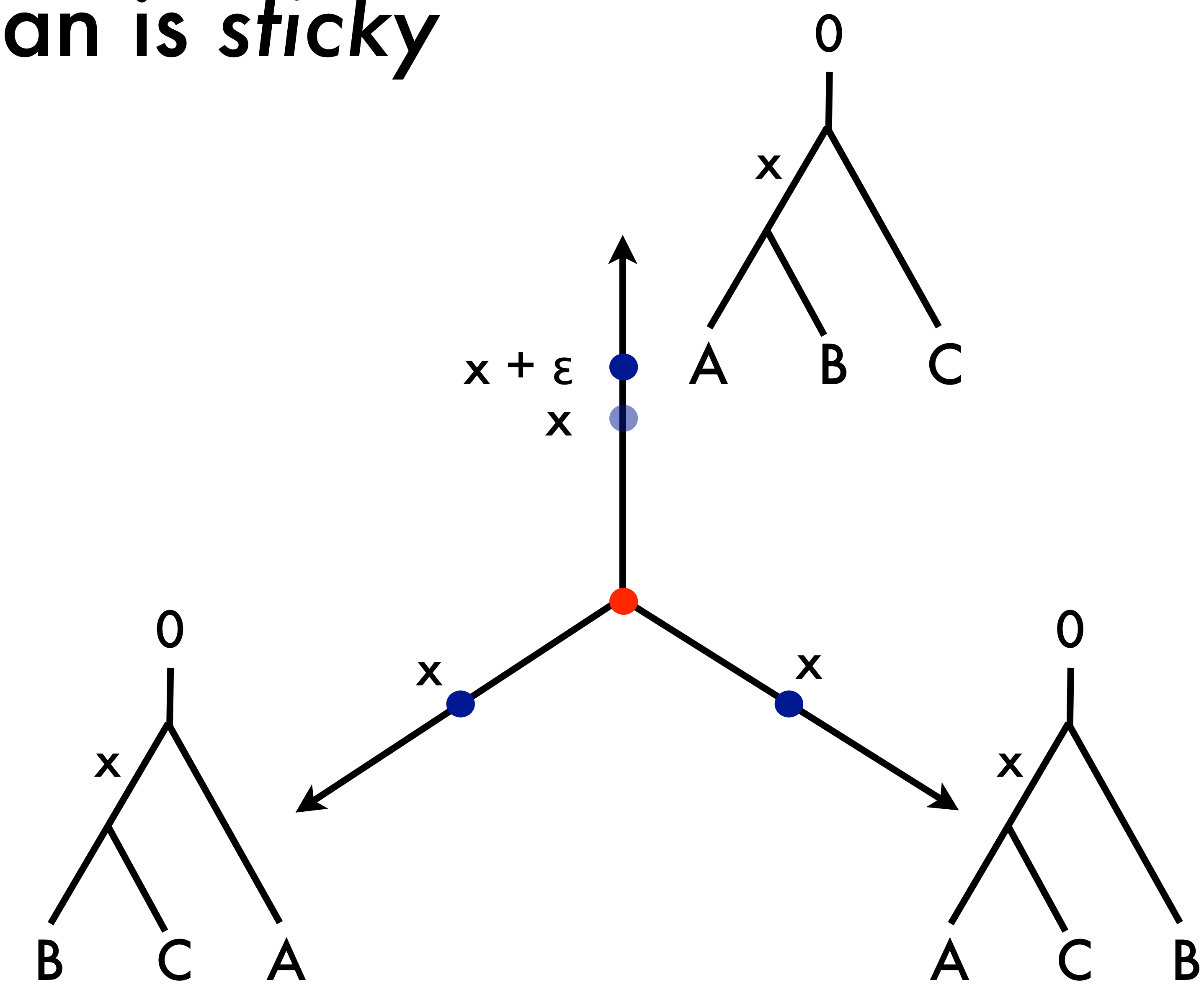
Caveat

- Mean is *sticky*

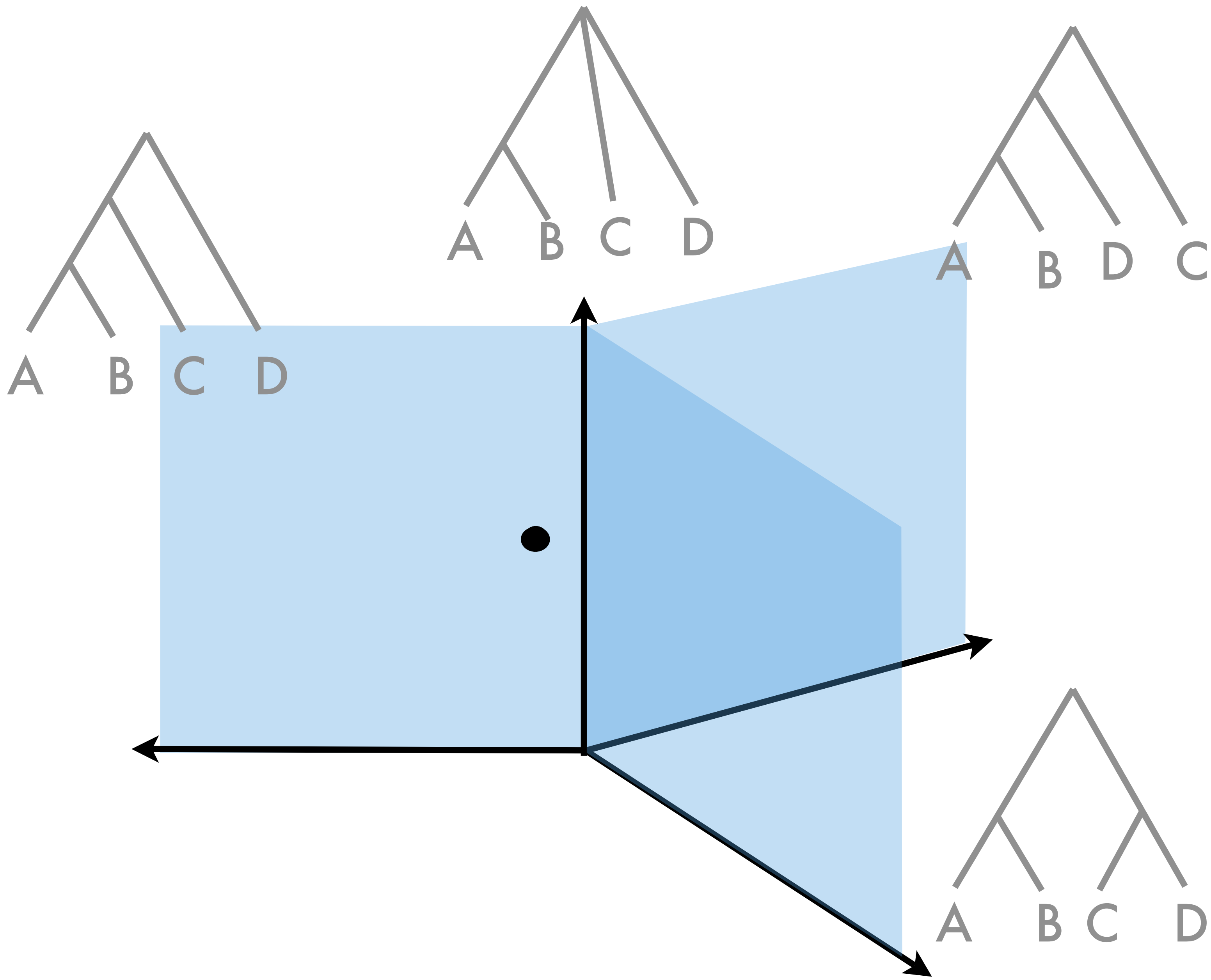


Caveat

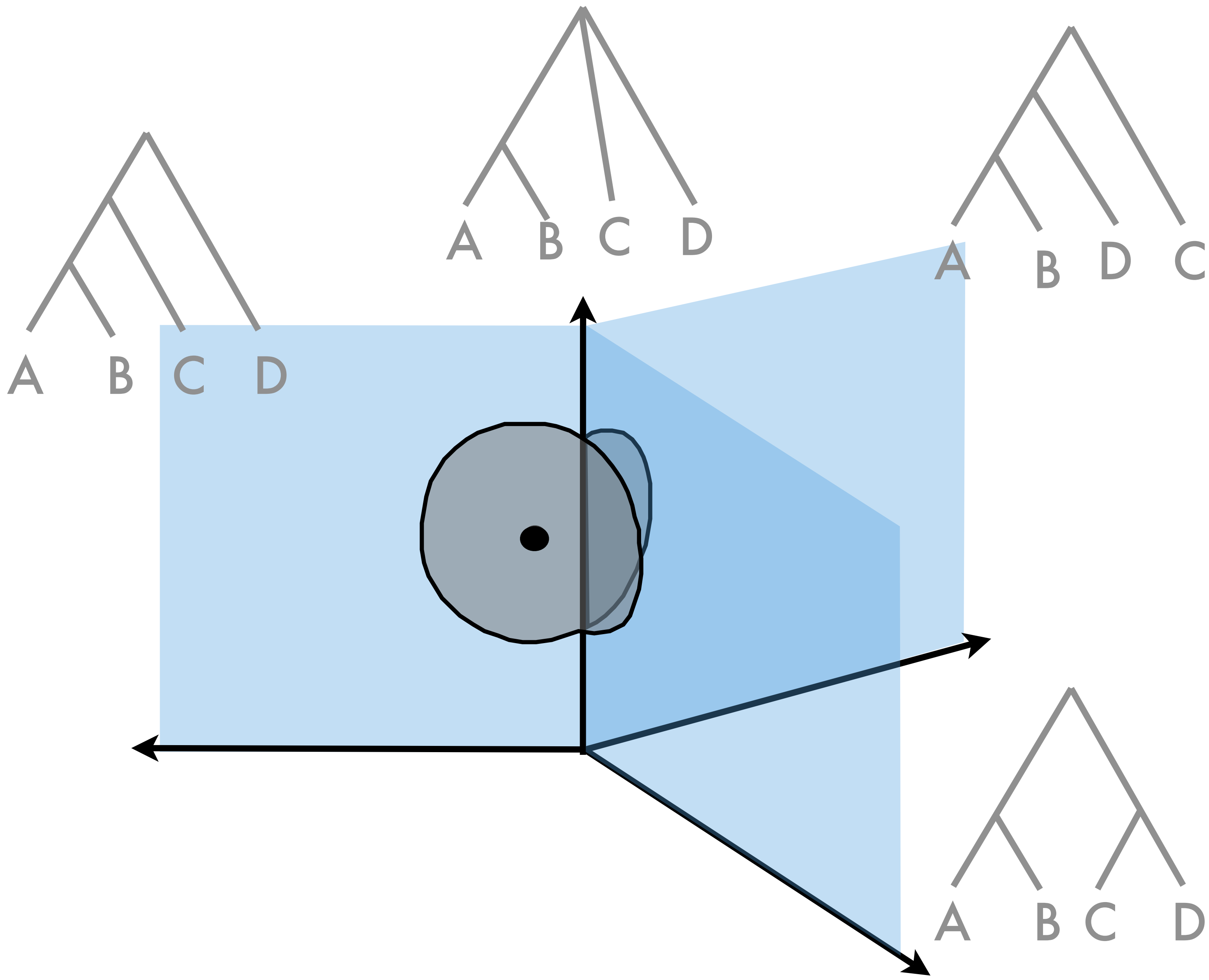
- Mean is *sticky*



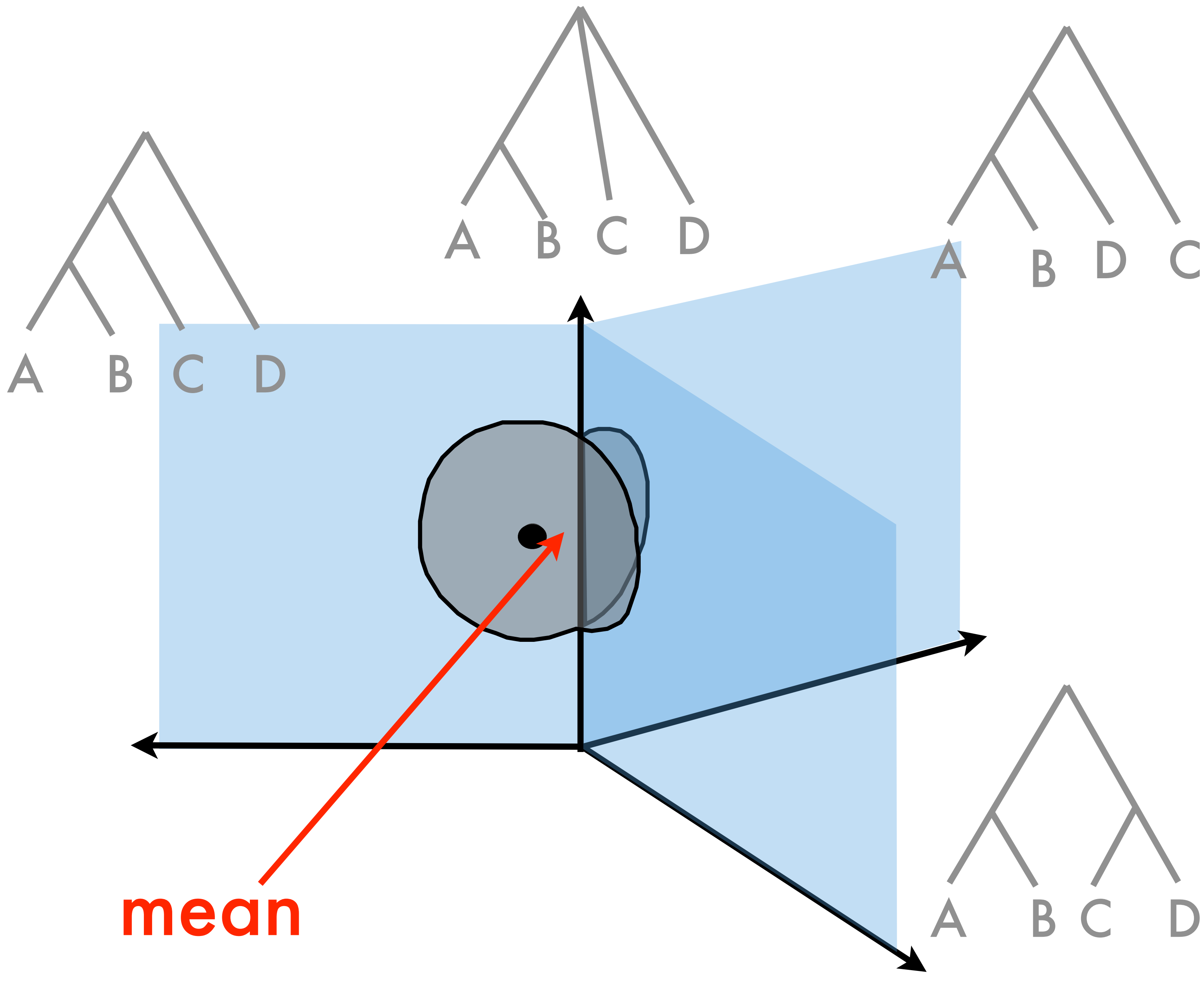
Caveat



Caveat



Caveat



Other Statistics

- Central Limit Theorem on BHV tree space:
  - special cases: Hotz, O., et al. 2012; Barden, Le, O., 2013, 2014; Huckemann et al. 2015
- Principal Components Analysis (PCA): (Nye 2011, 2014; Feragen, O. et al. 2013; Nye et al. 2016)
- confidence regions: Willis 2016
- multiple techniques: Chakerian and Holmes 2012, Zairis et al. 2016
- and more...

Thank You

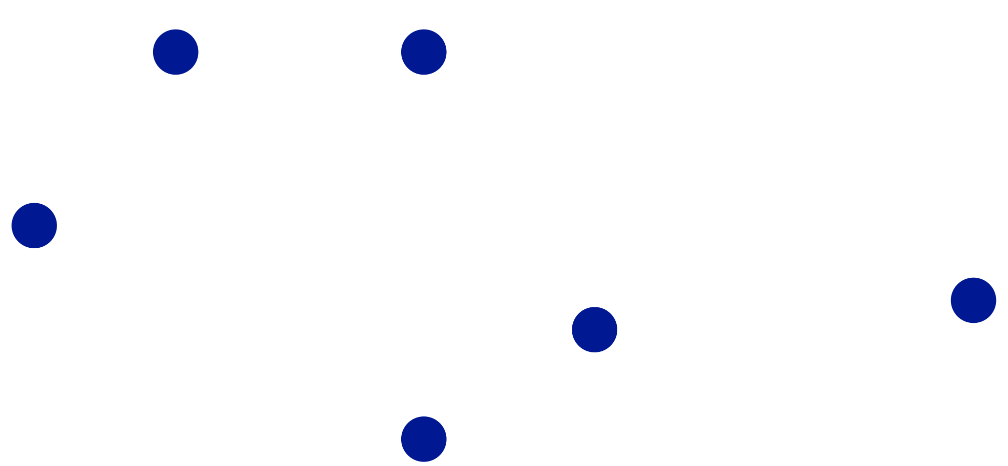
- funding: SIMONS FOUNDATION

- webpage: http://comet.lehman.cuny.edu/owen

Computing Mean

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

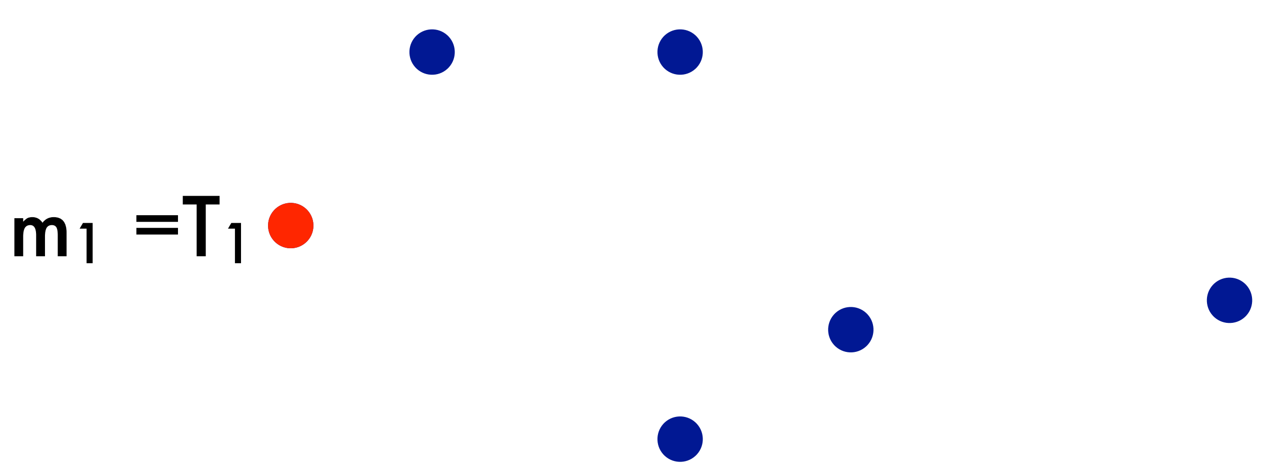
- $m_1 = T_1$
- i^{th} iteration :
  - randomly choose tree T_i from tree set with replacement
  - $m_i = \frac{1}{i}$ (geodesic from m_{i-1} to T_i)



Computing Mean

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

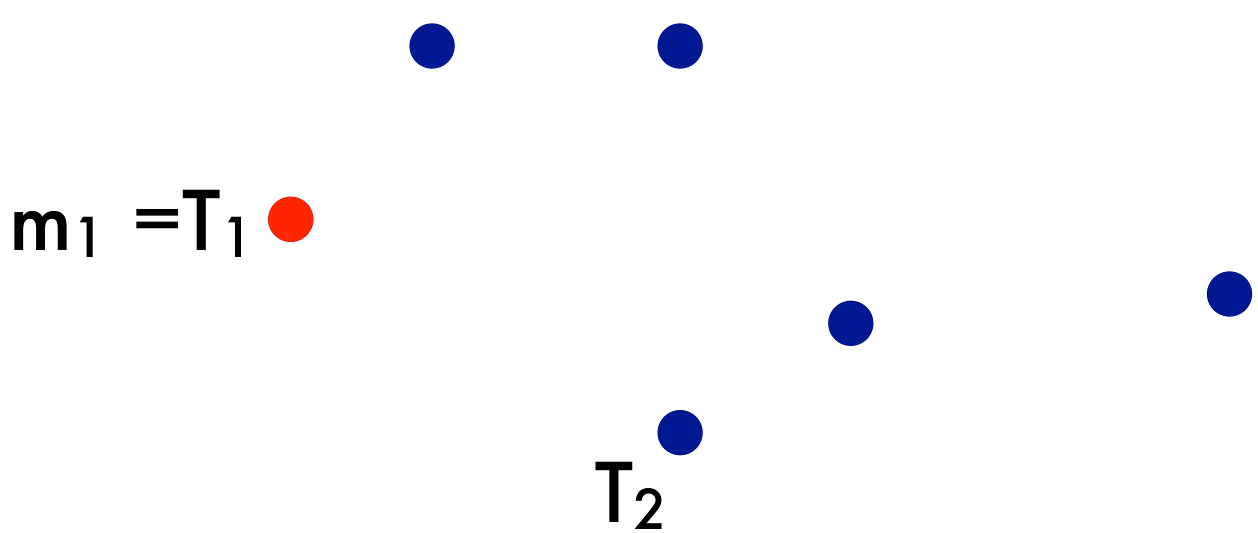
- $m_1 = T_1$
- i^{th} iteration :
  - randomly choose tree T_i from tree set with replacement
  - $m_i = \frac{1}{i}$ (geodesic from m_{i-1} to T_i)



Computing Mean

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

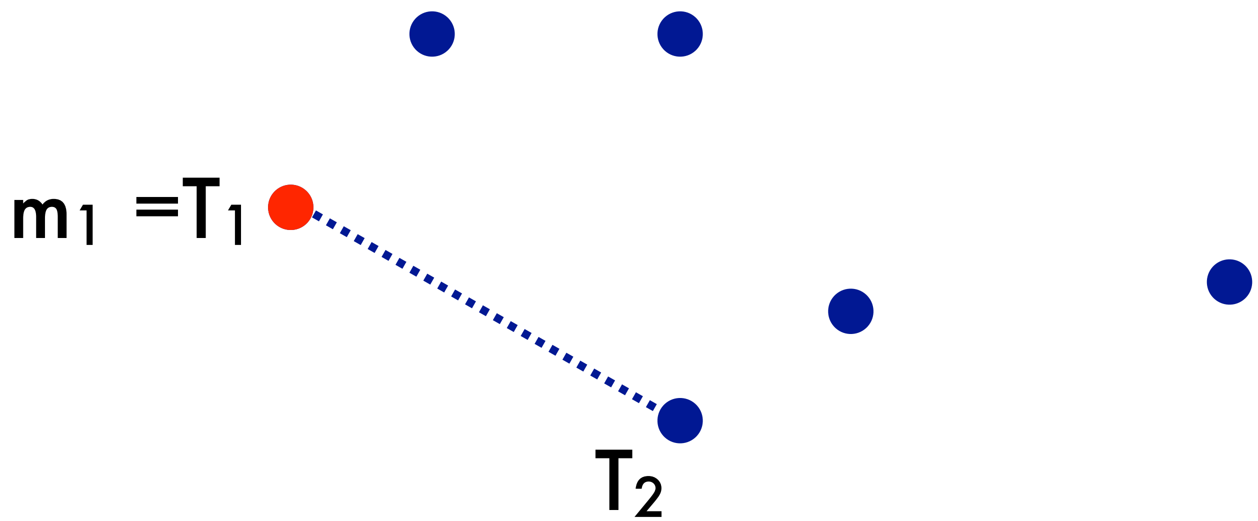
- $m_1 = T_1$
- i^{th} iteration :
  - randomly choose tree T_i from tree set with replacement
  - $m_i = \frac{1}{i}$ (geodesic from m_{i-1} to T_i)



Computing Mean

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

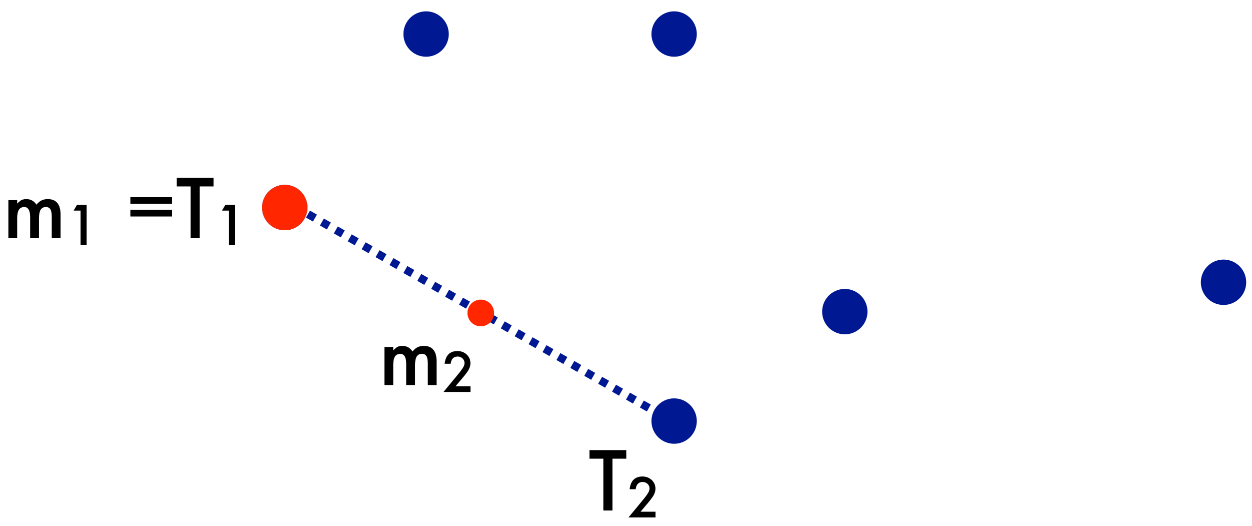
- $m_1 = T_1$
- i^{th} iteration :
  - randomly choose tree T_i from tree set with replacement
  - $m_i = \frac{1}{i}$ (geodesic from m_{i-1} to T_i)



Computing Mean

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

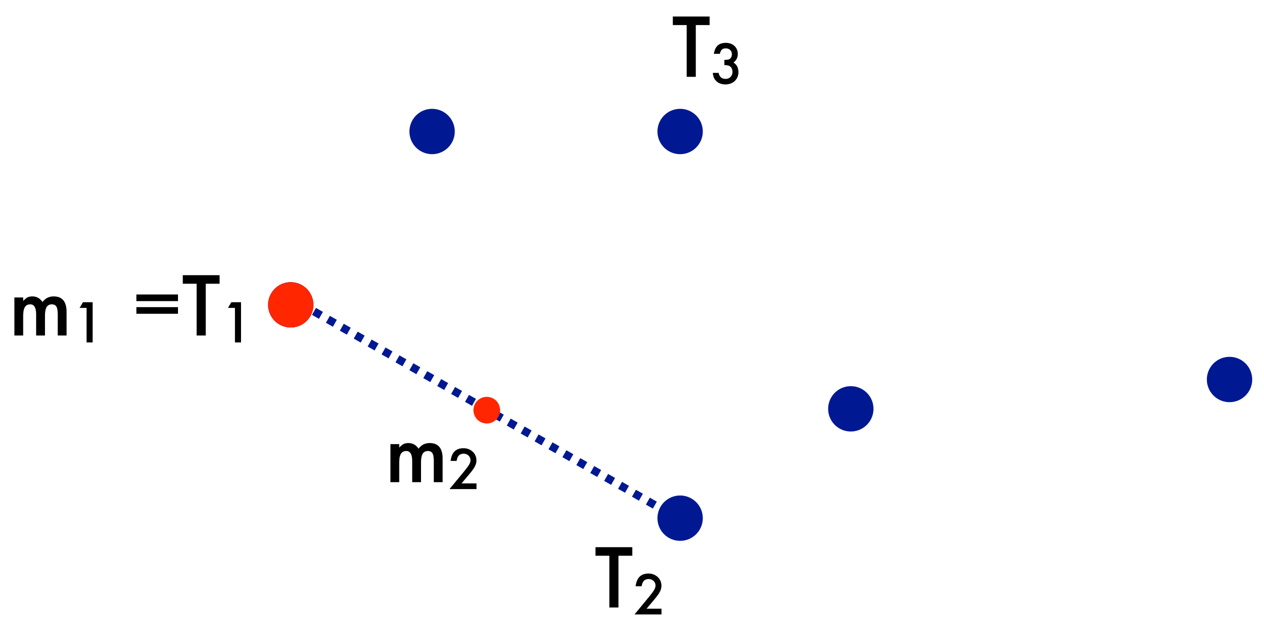
- $m_1 = T_1$
- i^{th} iteration :
  - randomly choose tree T_i from tree set with replacement
  - $m_i = \frac{1}{i}$ (geodesic from m_{i-1} to T_i)



Computing Mean

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

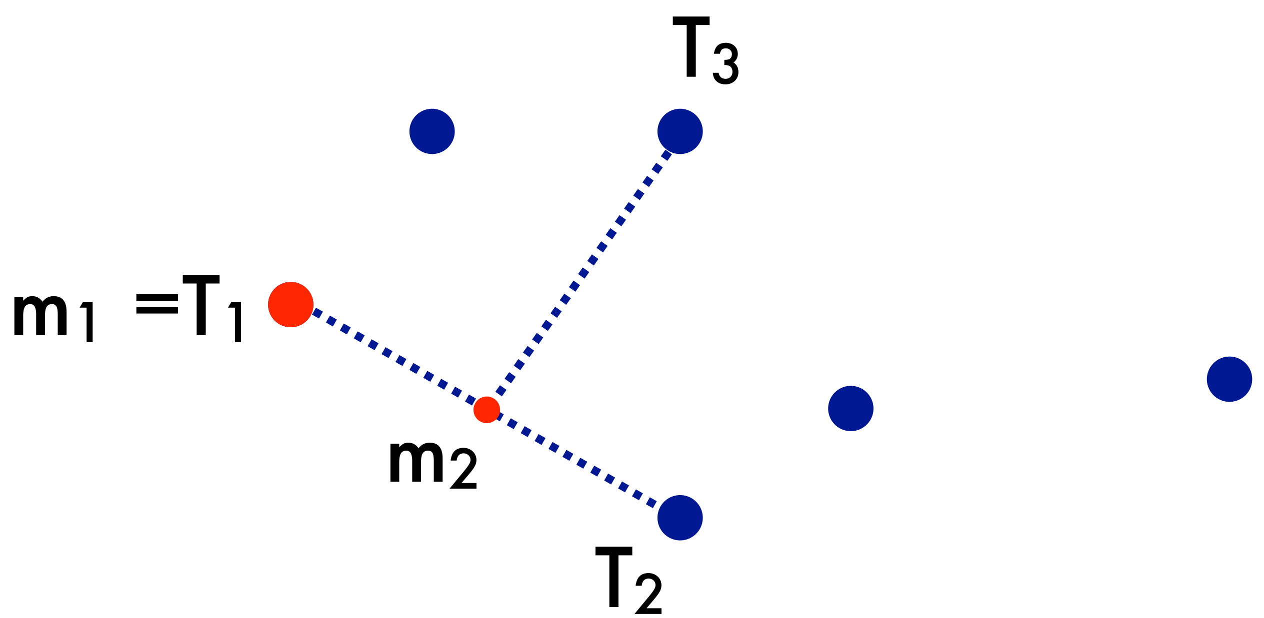
- $m_1 = T_1$
- i^{th} iteration :
  - randomly choose tree T_i from tree set with replacement
  - $m_i = \frac{1}{i}$ (geodesic from m_{i-1} to T_i)



Computing Mean

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

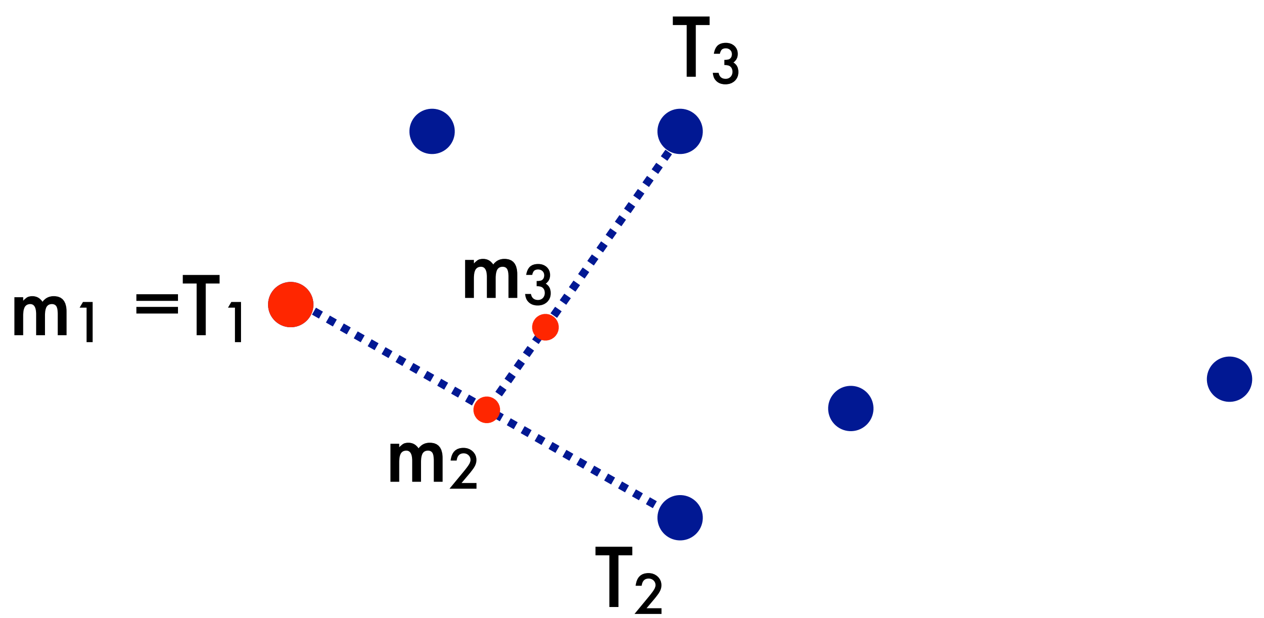
- $m_1 = T_1$
- i^{th} iteration :
  - randomly choose tree T_i from tree set with replacement
  - $m_i = \frac{1}{i}$ (geodesic from m_{i-1} to T_i)



Computing Mean

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

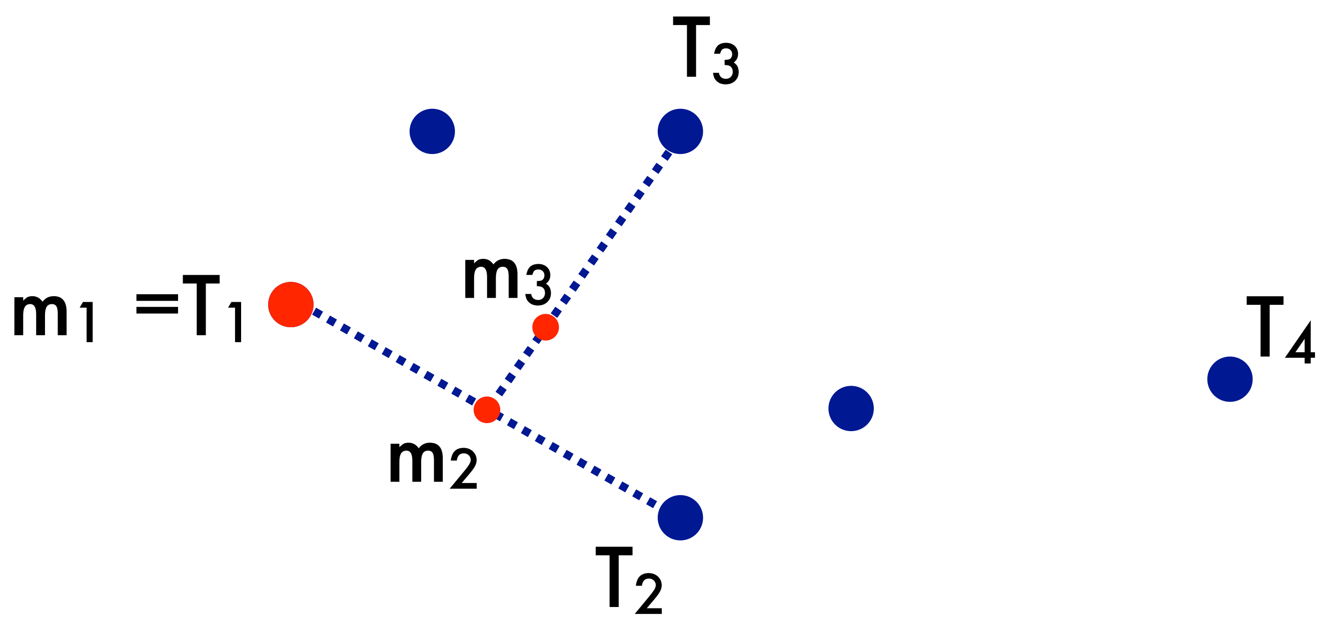
- $m_1 = T_1$
- i^{th} iteration :
  - randomly choose tree T_i from tree set with replacement
  - $m_i = \frac{1}{i}$ (geodesic from m_{i-1} to T_i)



Computing Mean

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

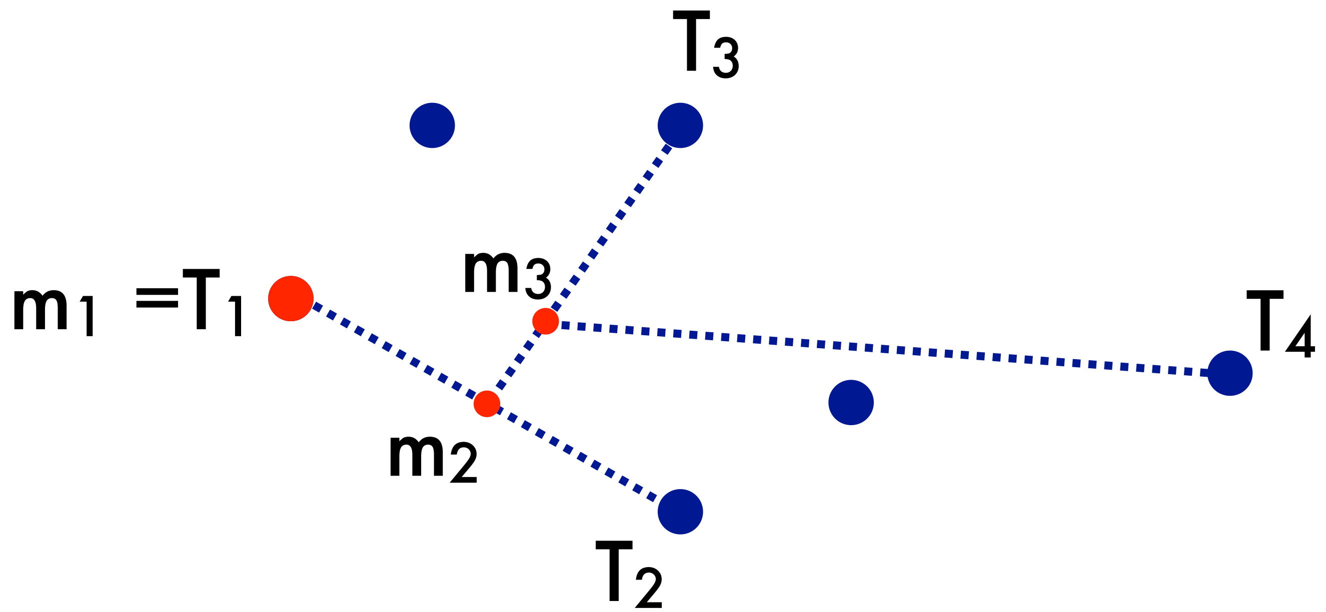
- $m_1 = T_1$
- i^{th} iteration :
  - randomly choose tree T_i from tree set with replacement
  - $m_i = \frac{1}{i}$ (geodesic from m_{i-1} to T_i)



Computing Mean

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

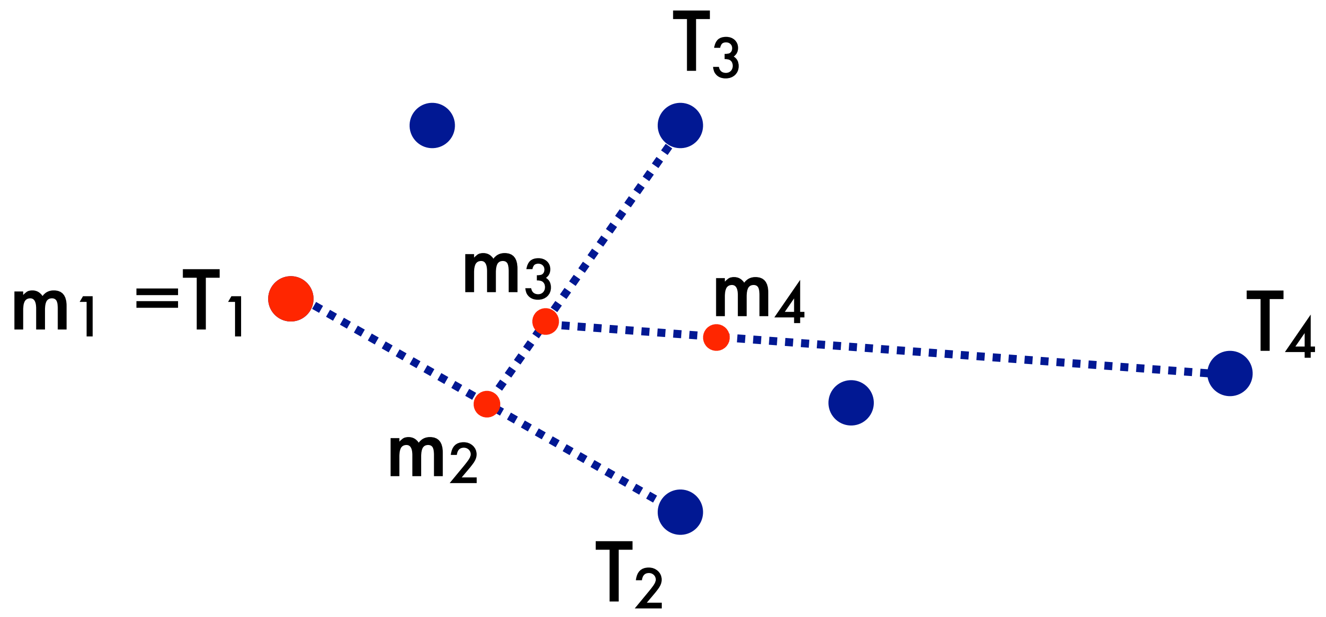
- $m_1 = T_1$
- i^{th} iteration :
  - randomly choose tree T_i from tree set with replacement
  - $m_i = \frac{1}{i}$ (geodesic from m_{i-1} to T_i)



Computing Mean

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

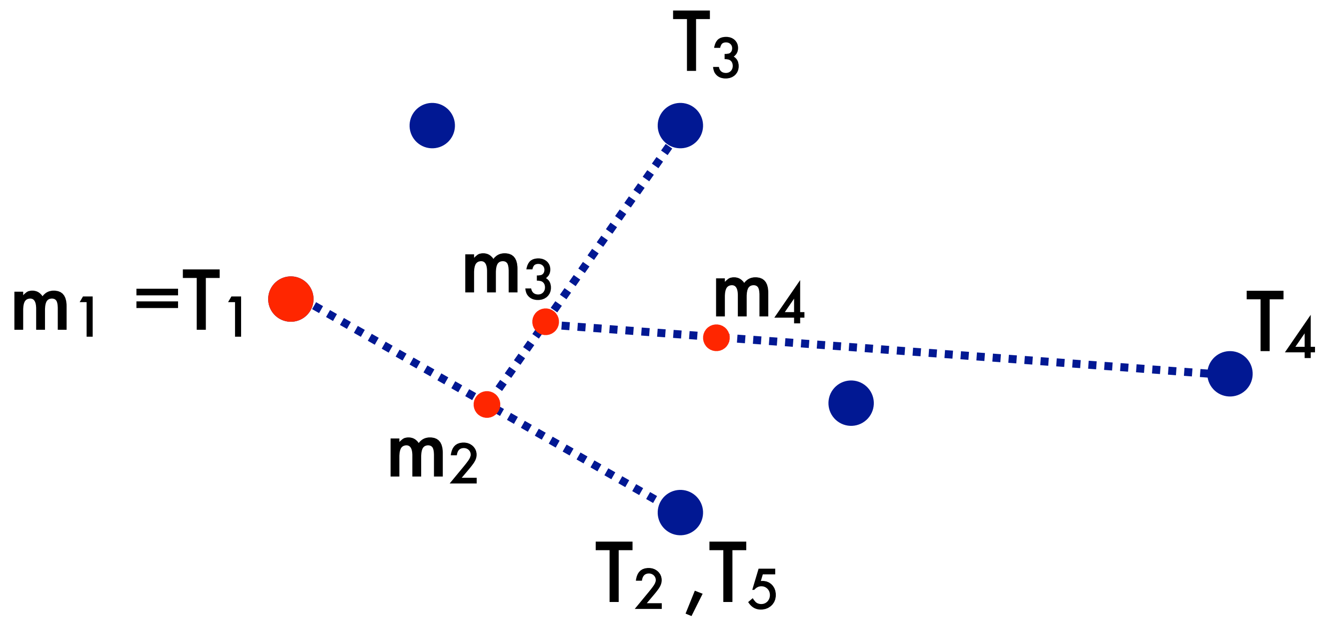
- $m_1 = T_1$
- i^{th} iteration :
  - randomly choose tree T_i from tree set with replacement
  - $m_i = \frac{1}{i}$ (geodesic from m_{i-1} to T_i)



Computing Mean

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

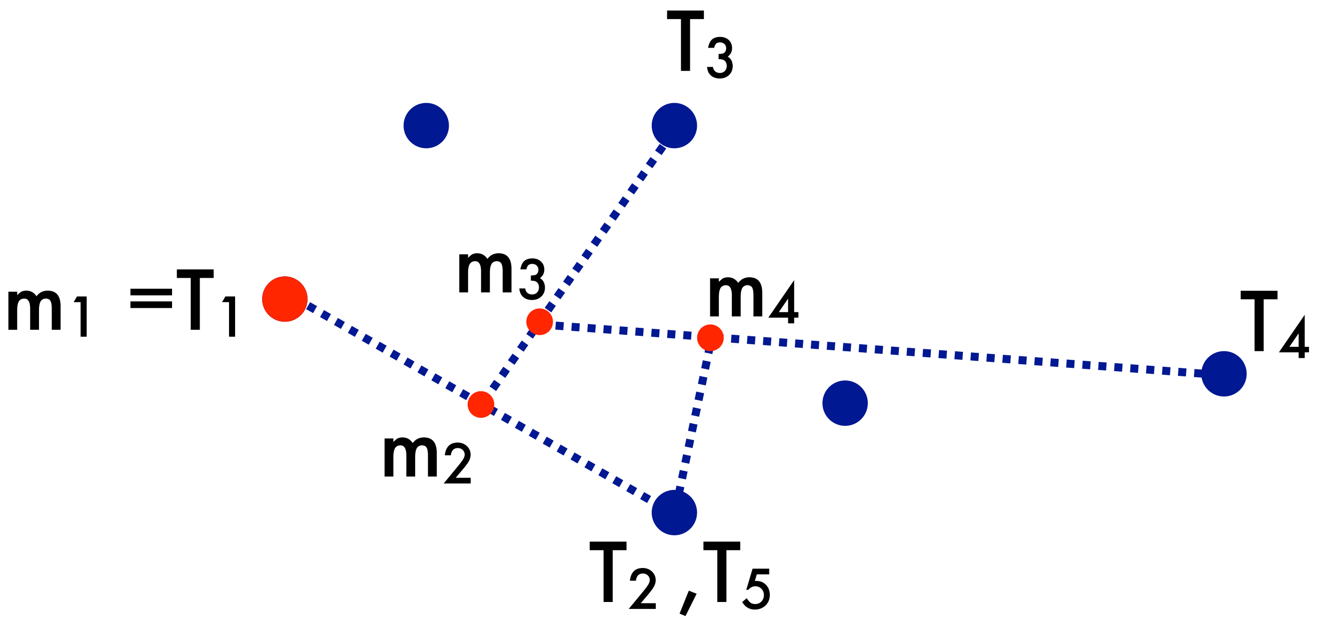
- $m_1 = T_1$
- i^{th} iteration :
  - randomly choose tree T_i from tree set with replacement
  - $m_i = \frac{1}{i}$ (geodesic from m_{i-1} to T_i)



Computing Mean

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

- $m_1 = T_1$
- i^{th} iteration :
  - randomly choose tree T_i from tree set with replacement
  - $m_i = \frac{1}{i}$ (geodesic from m_{i-1} to T_i)



Computing Mean

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

- $m_1 = T_1$
- i^{th} iteration :
  - randomly choose tree T_i from tree set with replacement
  - $m_i = \frac{1}{i}$ (geodesic from m_{i-1} to T_i)