

# Lecture 10

Maximum likelihood  
Botany/Plant Path 563

- **Previous class check-up:**
  - We studied the different models of evolution
- **Learning Objectives:** At the end of today's session, you will be able to
  - Explain how the likelihood of a tree is computed
  - Explain the steps in maximum likelihood phylogenetic inference
- **Pre-class work**
  - Read HAL 1.2 and canvas quiz

# Phylogenetic inference

Step 1: Choose the criterion to use:  
distances, parsimony, likelihood

Step 2: Search the space of trees  
until you find the optimum

# Phylogenetic inference

Step 1: Choose the criterion to use:  
~~distances, parsimony, likelihood~~

Step 2: Search the space of trees  
until you find the optimum



**You know how to  
calculate the likelihood  
for a given tree**

# Maximum likelihood

1. Choose a substitution model

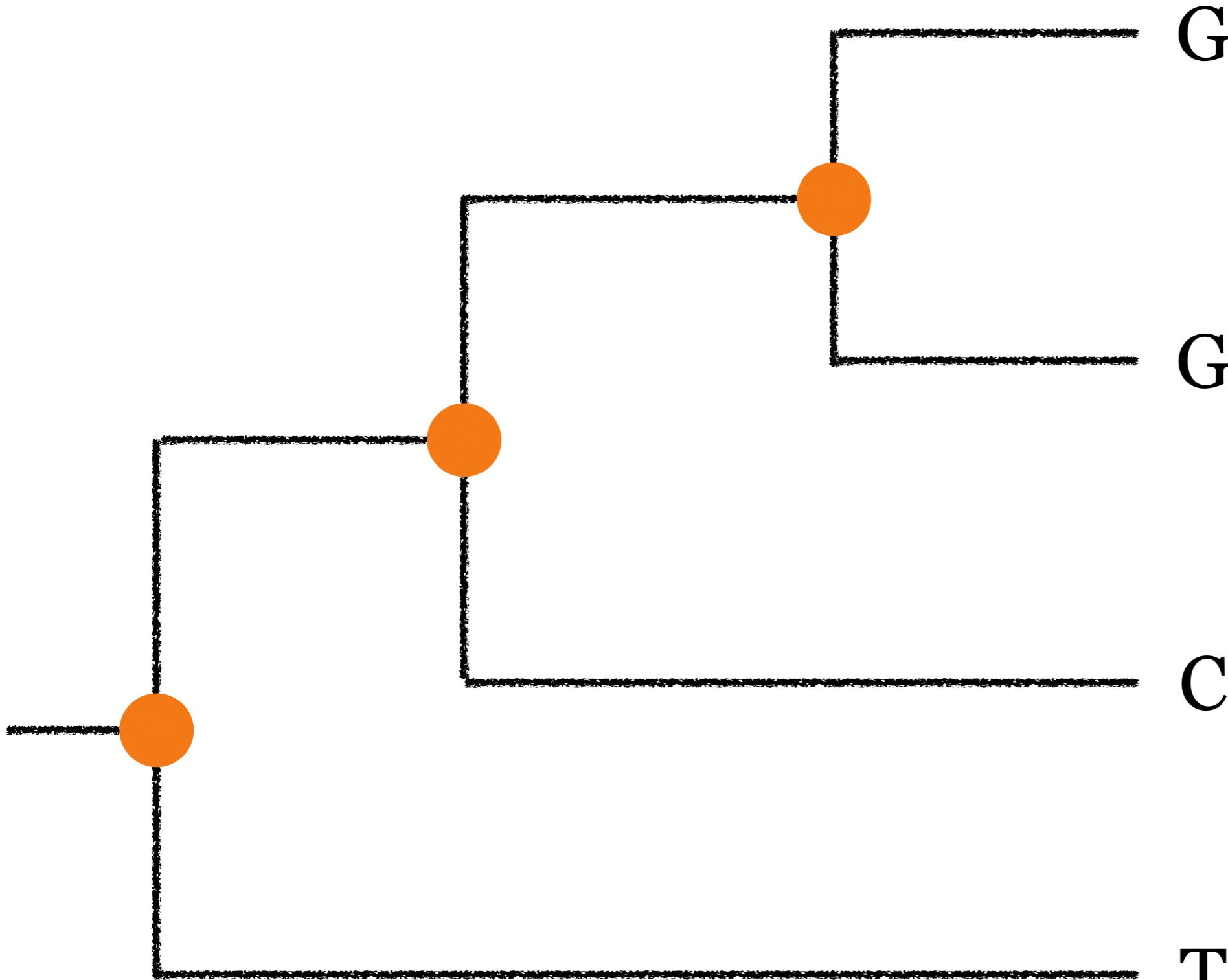
$$\mathbf{P}(t) = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & & & & \\ C & & & & \\ G & & & & \\ T & & & & \end{array} = e^{\mathbf{Q}\mu t}$$

2. For a given tree, calculate the likelihood given the data and the substitution model

$$\mathcal{L}_Q(\text{tree}) \mid \begin{array}{l} \text{AAGTCTAG} \\ \text{AAGTCTAG} \\ \text{AACTCTAG} \\ \text{AATTCTAG} \end{array}$$

3. Search the space of trees using the tree moves (NNI, SPR, TBR) until you find the maximum likelihood tree

# Calculate the likelihood for this tree

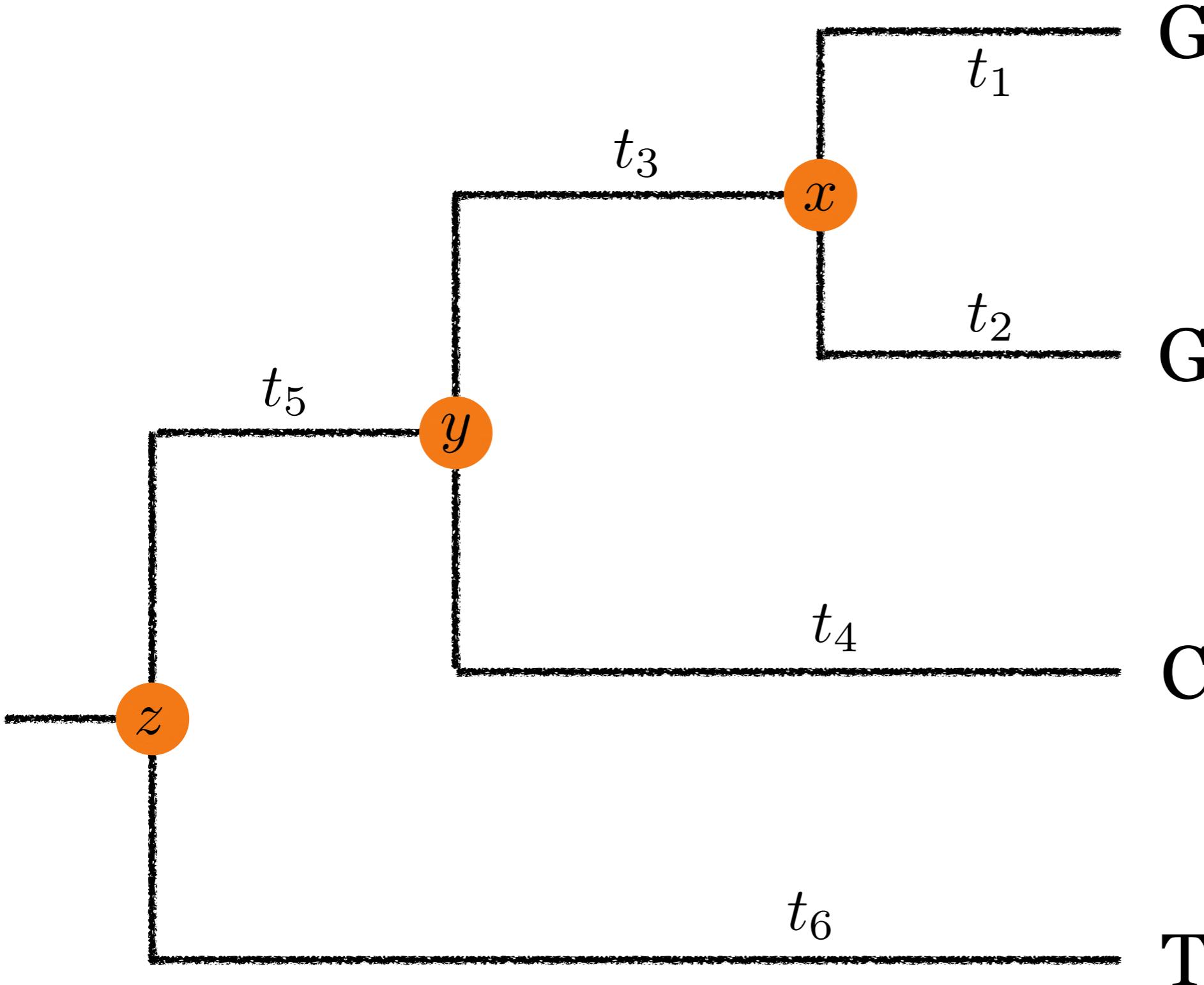


**Assumption 1:** The mutation process is the same at every branch of the tree

**Assumption 2:** We assume sites evolve independently

**Assumption 3:** All sites evolve the same

# Calculate the likelihood for this tree



Depends on parameters:

$Q$

You choose which form  
(each model has its own parameters)

$\mathbf{t} = (t_1, \dots, t_6)$

Branch lengths

$x, y, z$

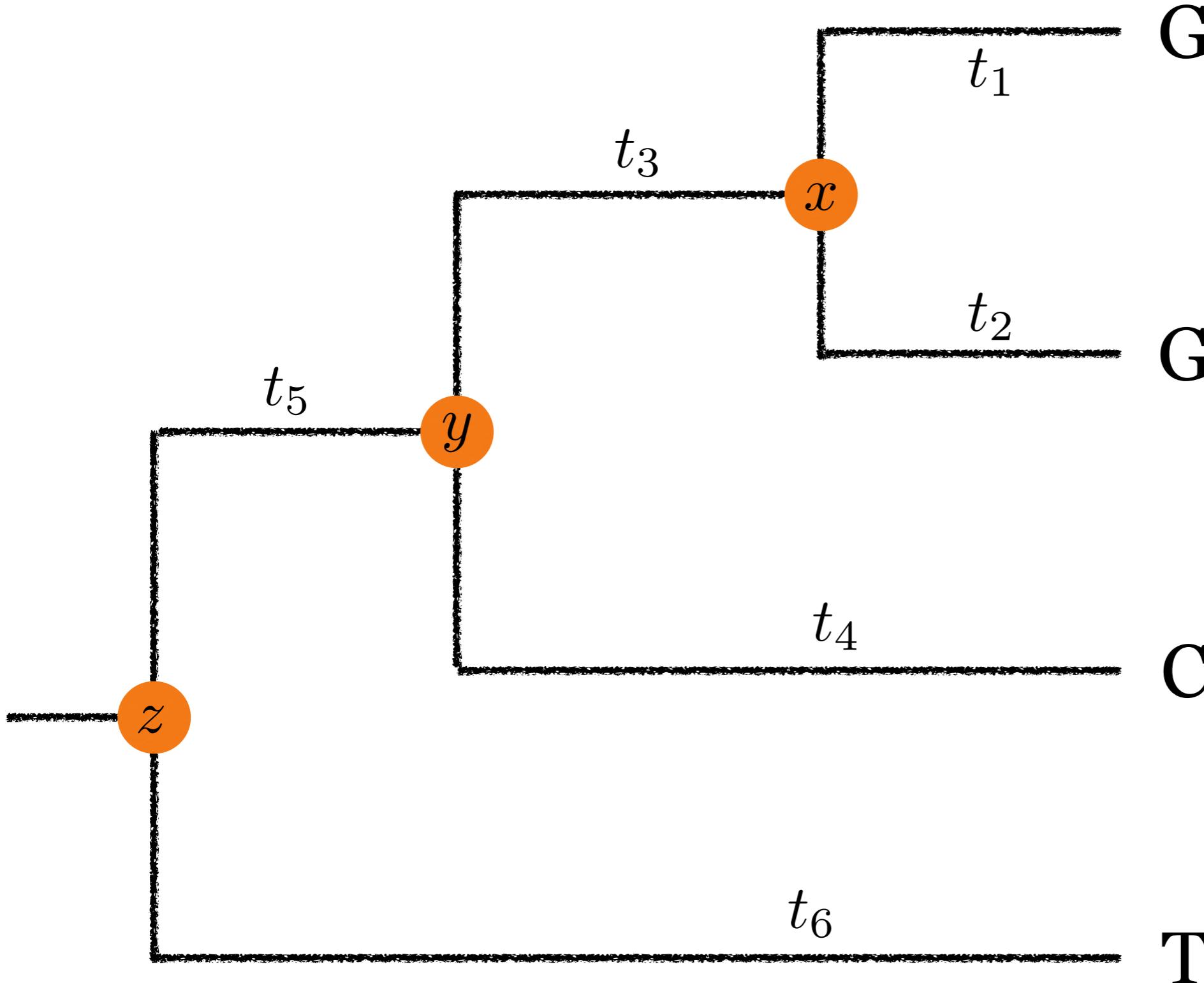
Ancestral states

**Assumption 1:** The mutation process is the same at every branch of the tree

**Assumption 2:** We assume sites evolve independently

**Assumption 3:** All sites evolve the same

# Calculate the likelihood for this tree



Depends on parameters:

**Q** You choose which form (each model has its own parameters)

**t** = (t<sub>1</sub>, ..., t<sub>6</sub>)

Branch lengths

**x, y, z** Ancestral states

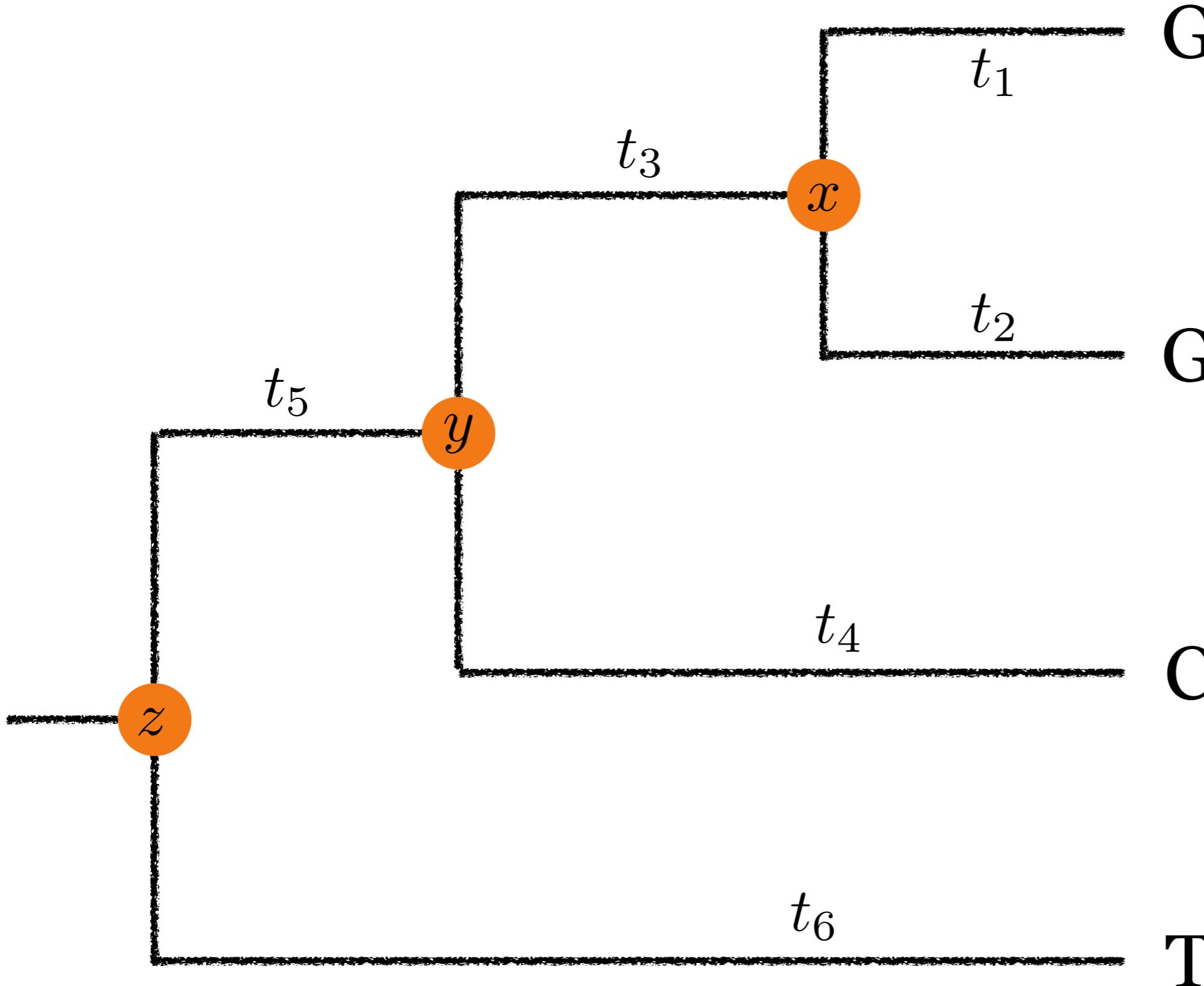
$$\mathcal{L}(T, Q, t, x, y, z)$$

**Assumption 1:** The mutation process is the same at every branch of the tree

**Assumption 2:** We assume sites evolve independently

**Assumption 3:** All sites evolve the same

# Calculate the likelihood for this tree



Depends on parameters:

$\mathbf{Q}$  You choose which form (each model has its own parameters)

$\mathbf{t} = (t_1, \dots, t_6)$   
Branch lengths

$x, y, z$  Ancestral states

$$\mathcal{L}(T, \mathbf{Q}, \mathbf{t}, x, y, z)$$

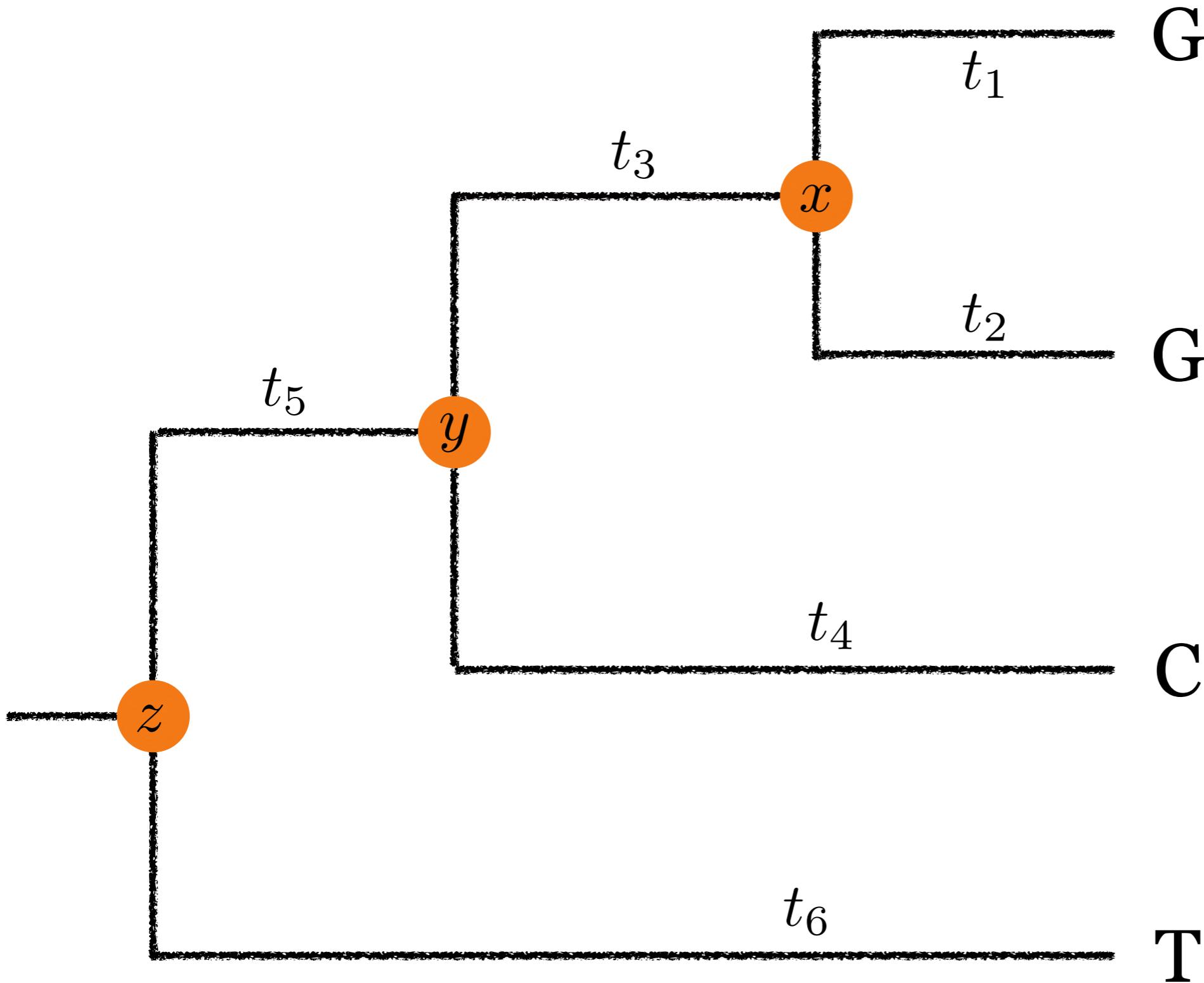
$$C = P(GGCT|T, \mathbf{Q}, \mathbf{t}, x, y, z)$$

**Assumption 1:** The mutation process is the same at every branch of the tree

**Assumption 2:** We assume sites evolve independently

**Assumption 3:** All sites evolve the same

# Calculate the likelihood for this tree

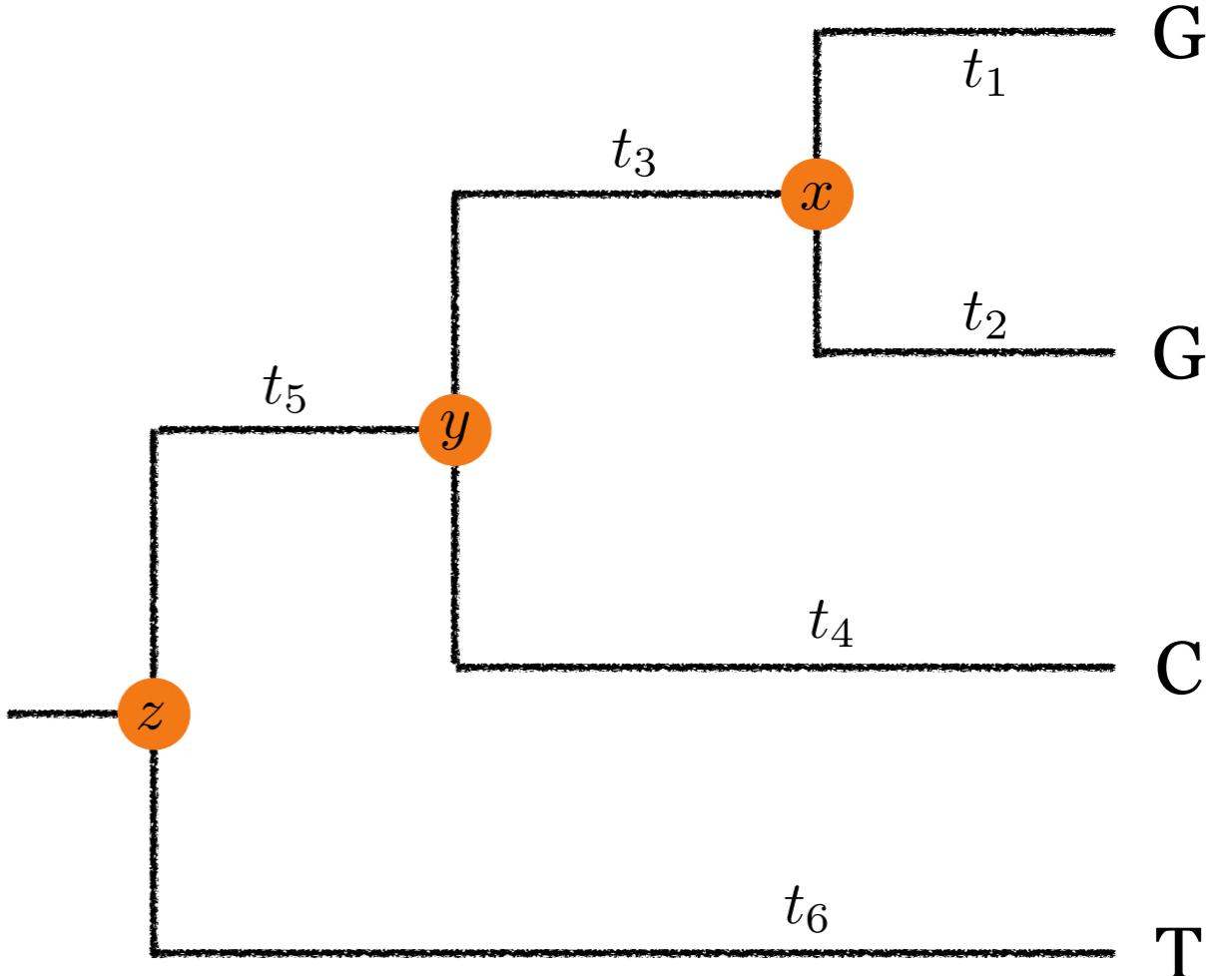


**Assumption 1:** The mutation process is the same at every branch of the tree

**Assumption 2:** We assume sites evolve independently

**Assumption 3:** All sites evolve the same

# Calculate the likelihood for this tree



**Assumption 1:** The mutation process is the same at every branch of the tree

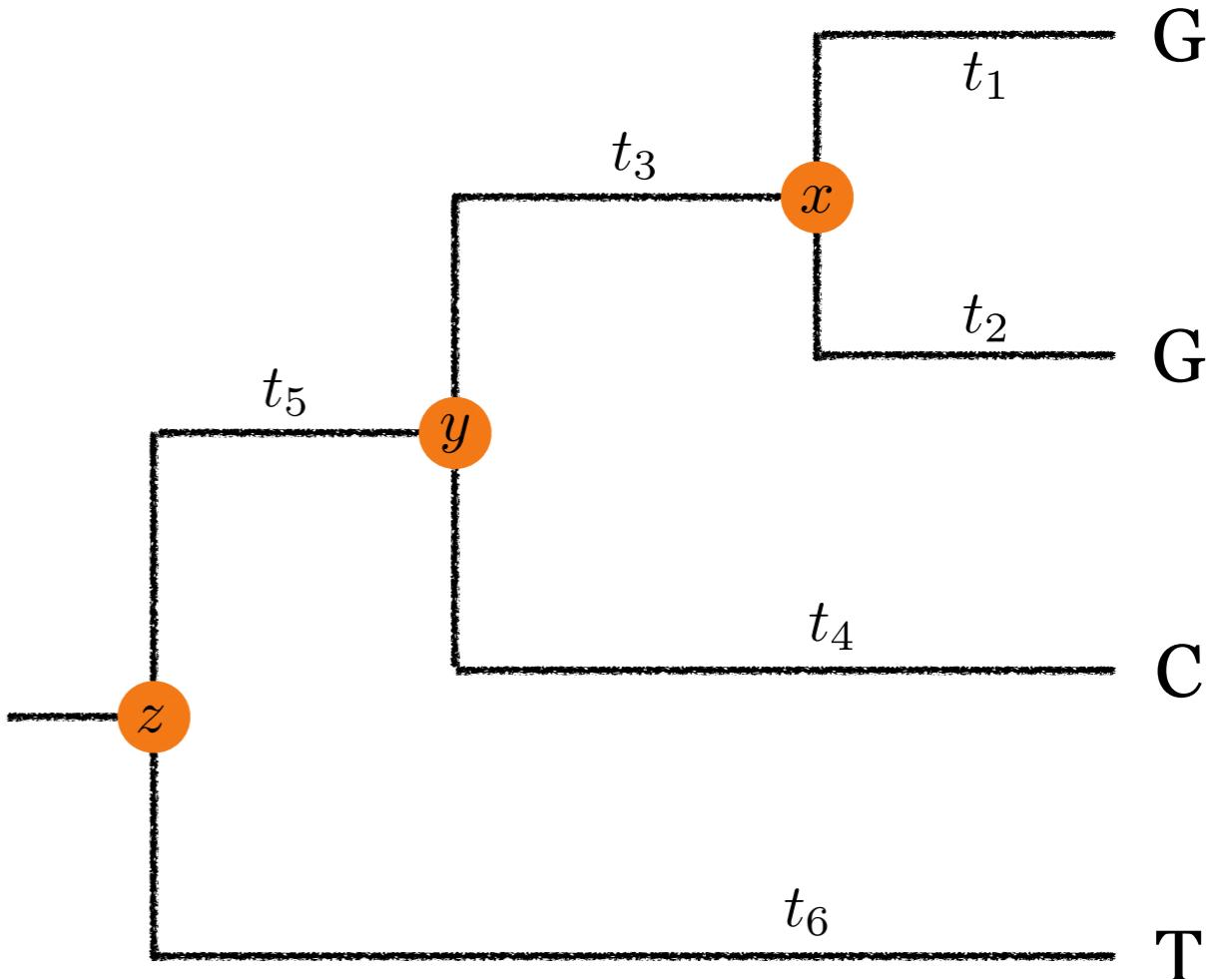
**Assumption 2:** We assume sites evolve independently

**Assumption 3:** All sites evolve the same

# Calculate the likelihood for this tree

$$\mathbf{P}(t) = e^{\mathbf{Q}\mu t}$$

$$\mathbf{Q} = \begin{bmatrix} A & C & G & T \\ * & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & * & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & * & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & * \end{bmatrix}$$



$$L = \sum_z \sum_y \sum_x \pi(z) P_{t_6}(z, T) P_{t_5}(z, y) P_{t_4}(y, C) P_{t_3}(y, x) P_{t_2}(x, G) P_{t_1}(x, G)$$

**Assumption 1:** The mutation process is the same at every branch of the tree

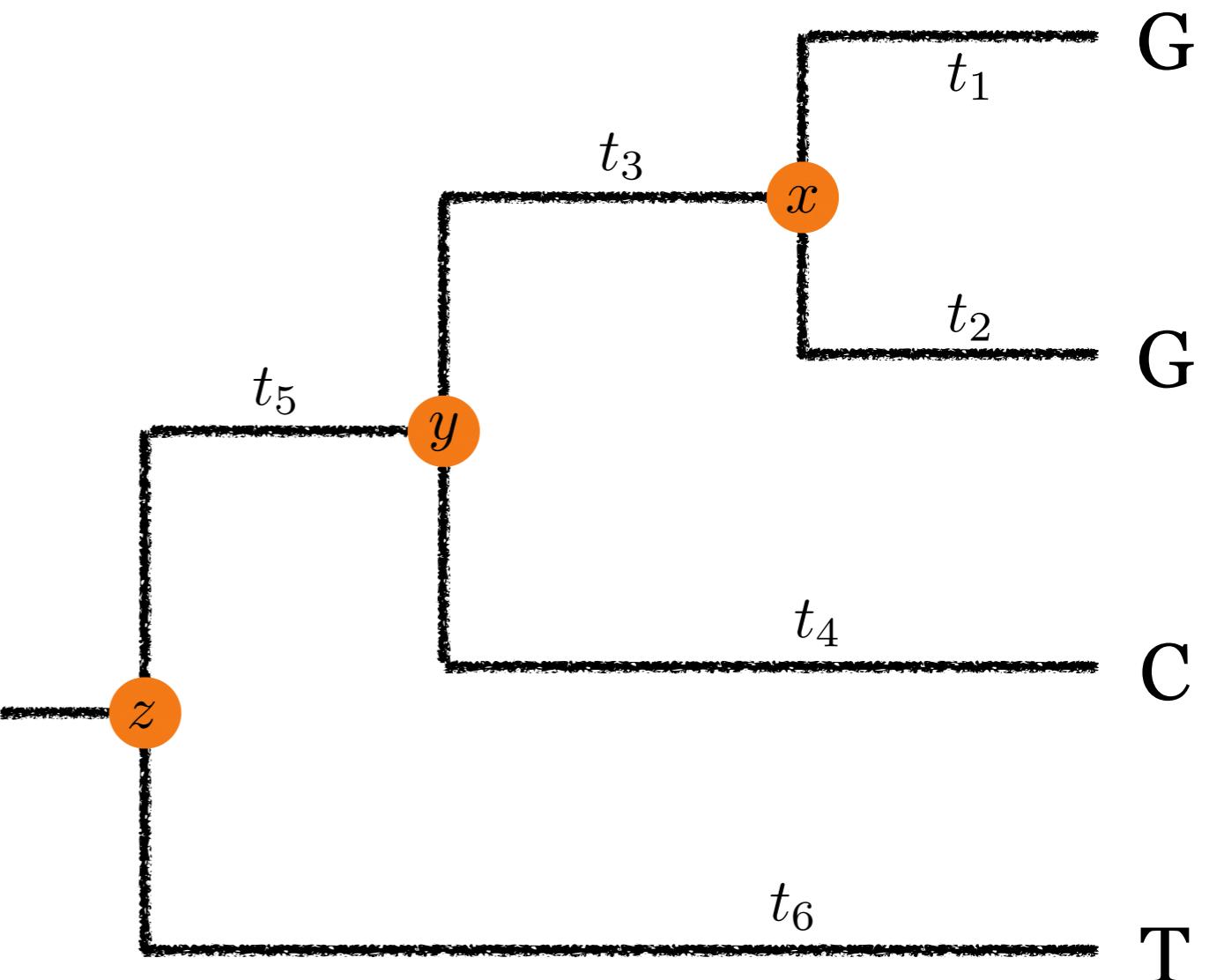
**Assumption 2:** We assume sites evolve independently

**Assumption 3:** All sites evolve the same

# Calculate the likelihood for this tree

$$\mathbf{P}(t) = e^{\mathbf{Q}\mu t}$$

$$\mathbf{Q} = \begin{bmatrix} A & C & G & T \\ * & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & * & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & * & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & * \end{bmatrix}$$



$$L = \sum_z \sum_y \sum_x \pi(z) P_{t_6}(z, T) P_{t_5}(z, y) P_{t_4}(y, C) P_{t_3}(y, x) P_{t_2}(x, G) P_{t_1}(x, G)$$

Where do the assumptions play a role?

**Assumption 1:** The mutation process is the same at every branch of the tree

**Assumption 2:** We assume sites evolve independently

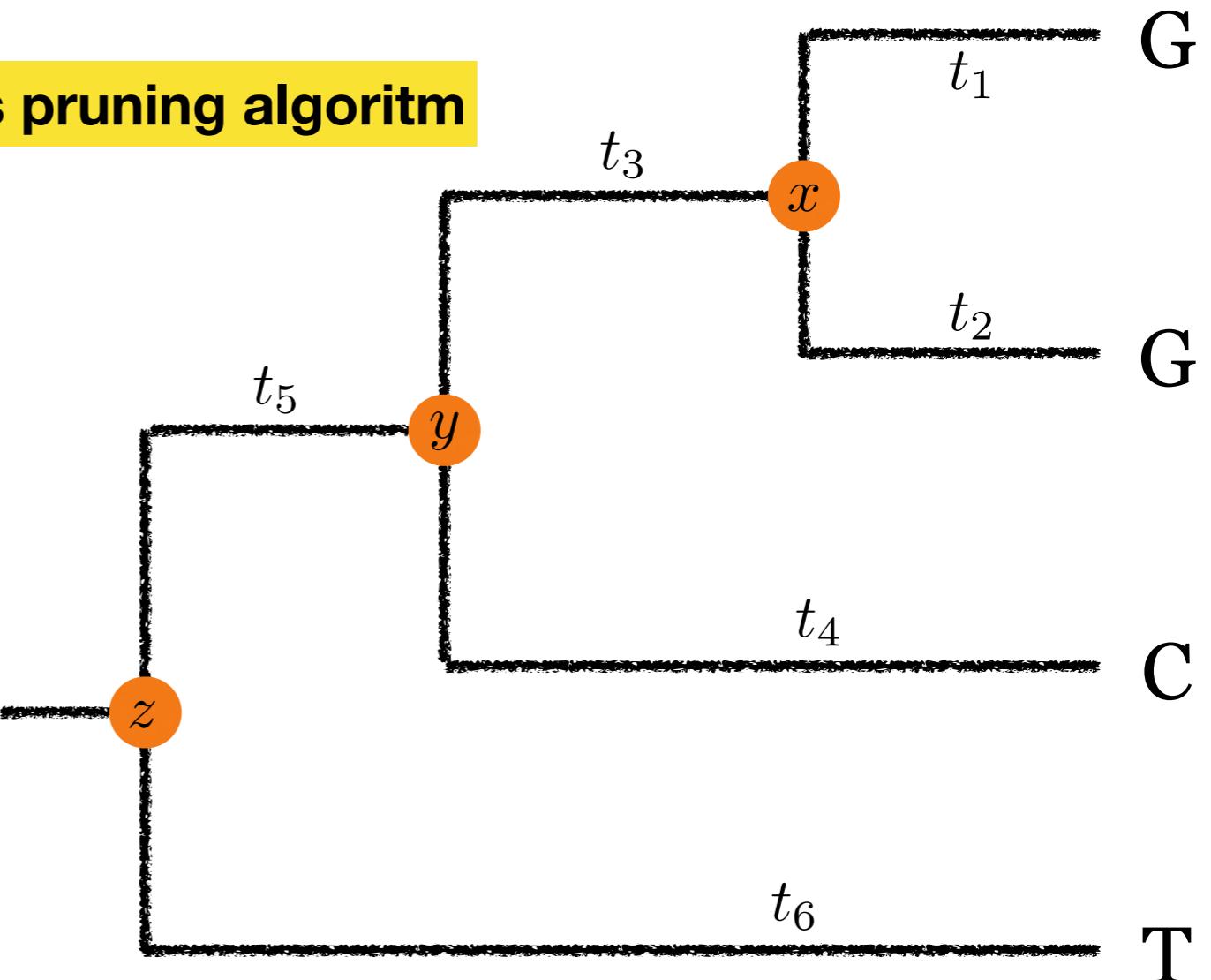
**Assumption 3:** All sites evolve the same

# Calculate the likelihood for this tree

## Felsenstein's pruning algorithm

$$\mathbf{P}(t) = e^{\mathbf{Q}\mu t}$$

$$\mathbf{Q} = \begin{bmatrix} A & C & G & T \\ * & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & * & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & * & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & * \end{bmatrix}$$



$$L = \sum_z \sum_y \sum_x \pi(z) P_{t_6}(z, T) P_{t_5}(z, y) P_{t_4}(y, C) P_{t_3}(y, x) P_{t_2}(x, G) P_{t_1}(x, G)$$

Where do the assumptions play a role?

**Assumption 1:** The mutation process is the same at every branch of the tree

**Assumption 2:** We assume sites evolve independently

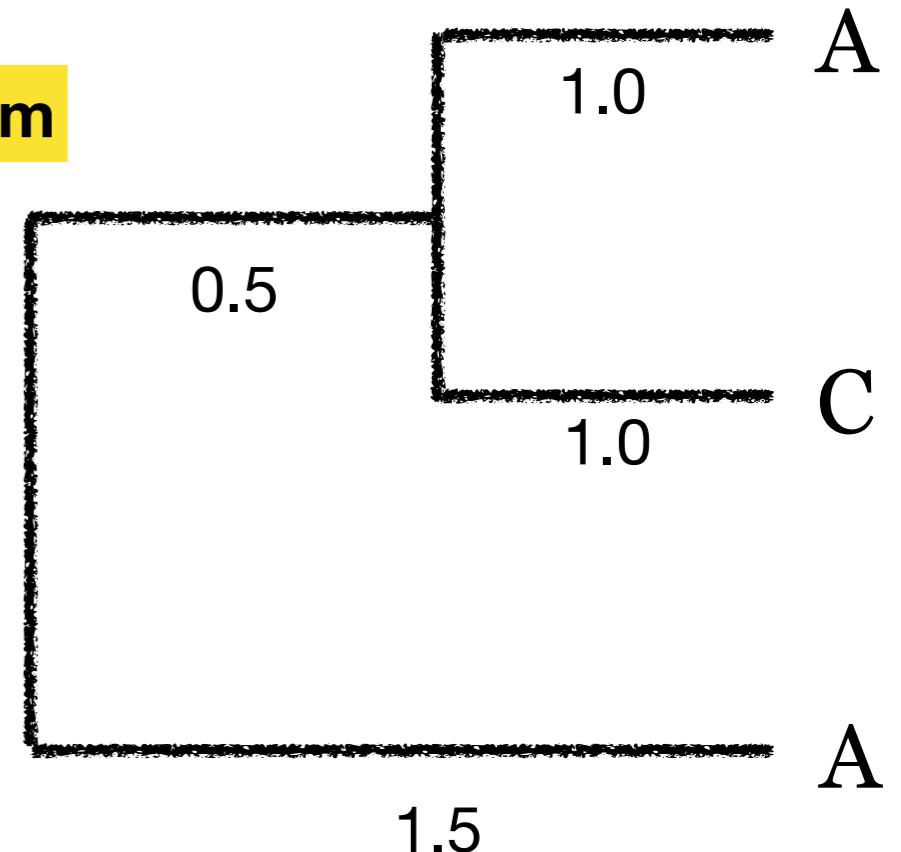
**Assumption 3:** All sites evolve the same

# Calculate the likelihood for this tree

Felsenstein's pruning algorithm

$$\mathbf{P}(t) = e^{\mathbf{Q}\mu t}$$

$$\mathbf{Q} = \begin{bmatrix} A & C & G & T \\ -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{bmatrix} \begin{matrix} A \\ C \\ G \\ T \end{matrix}$$



$$L_p(i) = \left( \sum_x P(x|i, t_L) L_L(x) \right) \left( \sum_x P(x|i, t_R) L_R(x) \right)$$

Full example in [video on YouTube](#)

# Maximum likelihood

1. Choose a substitution model

$$\mathbf{P}(t) = \begin{array}{c|cccc} & & A & C & G & T \\ \hline A & & & & & \\ C & & & & & \\ G & & & & & \\ T & & & & & \end{array} = e^{\mathbf{Q}\mu t}$$

2. For a given tree, calculate the likelihood given the data and the substitution model

$$\mathcal{L}_Q(\text{tree} | \text{AAGTCTAG, AAGTCTAG, AACTCTAG, AATTCTAG})$$

3. Search the space of trees using the tree moves (NNI, SPR, TBR) until you find the maximum likelihood tree

# Maximum likelihood

1. Choose a substitution model

$$\mathbf{P}(t) = \begin{array}{|c|c|c|c|} \hline & A & C & G & T \\ \hline A & & & & \\ \hline C & & & & \\ \hline G & & & & \\ \hline T & & & & \\ \hline \end{array} = e^{\mathbf{Q}\mu t}$$

2. For a given tree, calculate the likelihood given the data and the substitution model

**Depends on parameters:**

$$\mathcal{L}_Q(\text{tree} | \text{AAGTCTAG, AAGTCTAG, AACTCTAG, AATTCTAG})$$

$Q$   
You choose which form  
(each model has its own parameters)

$t = (t_1, \dots, t_6)$   
Branch lengths

$x, y, z$   
Ancestral states

3. Search the space of trees using the tree moves (NNI, SPR, TBR) until you find the maximum likelihood tree

# Maximum likelihood

1. Choose a substitution model

$$\mathbf{P}(t) = \begin{array}{|c|c|c|c|} \hline & A & C & G & T \\ \hline A & & & & \\ \hline C & & & & \\ \hline G & & & & \\ \hline T & & & & \\ \hline \end{array} = e^{\mathbf{Q}\mu t}$$

2. For a given tree, calculate the likelihood given the data and the substitution model

**Depends on parameters:**

$$\mathcal{L}_Q(\text{tree} | \text{AAGTCTAG, AAGTCTAG, AACTCTAG, AATTCTAG})$$

$\mathbf{Q}$

You choose which form  
(each model has its own parameters)

$t = (t_1, \dots, t_6)$

Branch lengths

$x, y, z$  Ancestral states

Average across them

3. Search the space of trees using the tree moves (NNI, SPR, TBR) until you find the maximum likelihood tree

# Maximum likelihood

1. Choose a substitution model

$$P(t) = \begin{array}{|c|c|c|c|} \hline & A & C & G & T \\ \hline A & & & & \\ \hline C & & & & \\ \hline G & & & & \\ \hline T & & & & \\ \hline \end{array} = e^{Q\mu t}$$

2. For a given tree, calculate the likelihood given the data and the substitution model

$$\mathcal{L}_Q(\text{tree} | \text{AAGTCTAG, AAGTCTAG, AACTCTAG, AATTCTAG})$$

**Depends on parameters:**

**Q**

You choose which form  
(each model has its own parameters)

**t** =  $(t_1, \dots, t_6)$

Branch lengths

~~x, y, z~~ Ancestral states

Average across them

3. Search the space of trees using the tree moves (NNI, SPR, TBR) until you find the maximum likelihood tree

→ At each proposed tree,  
we maximize Q and t

Need to optimize

# Maximum likelihood

1. Choose a substitution model

$$\mathbf{P}(t) = \begin{array}{|c|c|c|c|} \hline & A & C & G & T \\ \hline A & & & & \\ \hline C & & & & \\ \hline G & & & & \\ \hline T & & & & \\ \hline \end{array} = e^{\mathbf{Q}\mu t}$$

2. For a given tree, calculate the likelihood given the data and the substitution model

$$\mathcal{L}_Q(\text{tree} | \text{AAGTCTAG, AAGTCTAG, AACTCTAG, AATTCTAG})$$

3. Search the space of trees using the tree moves (NNI, SPR, TBR) until you find the maximum likelihood tree

**Depends on parameters:**

**Q**: You choose which form (each model has its own parameters)

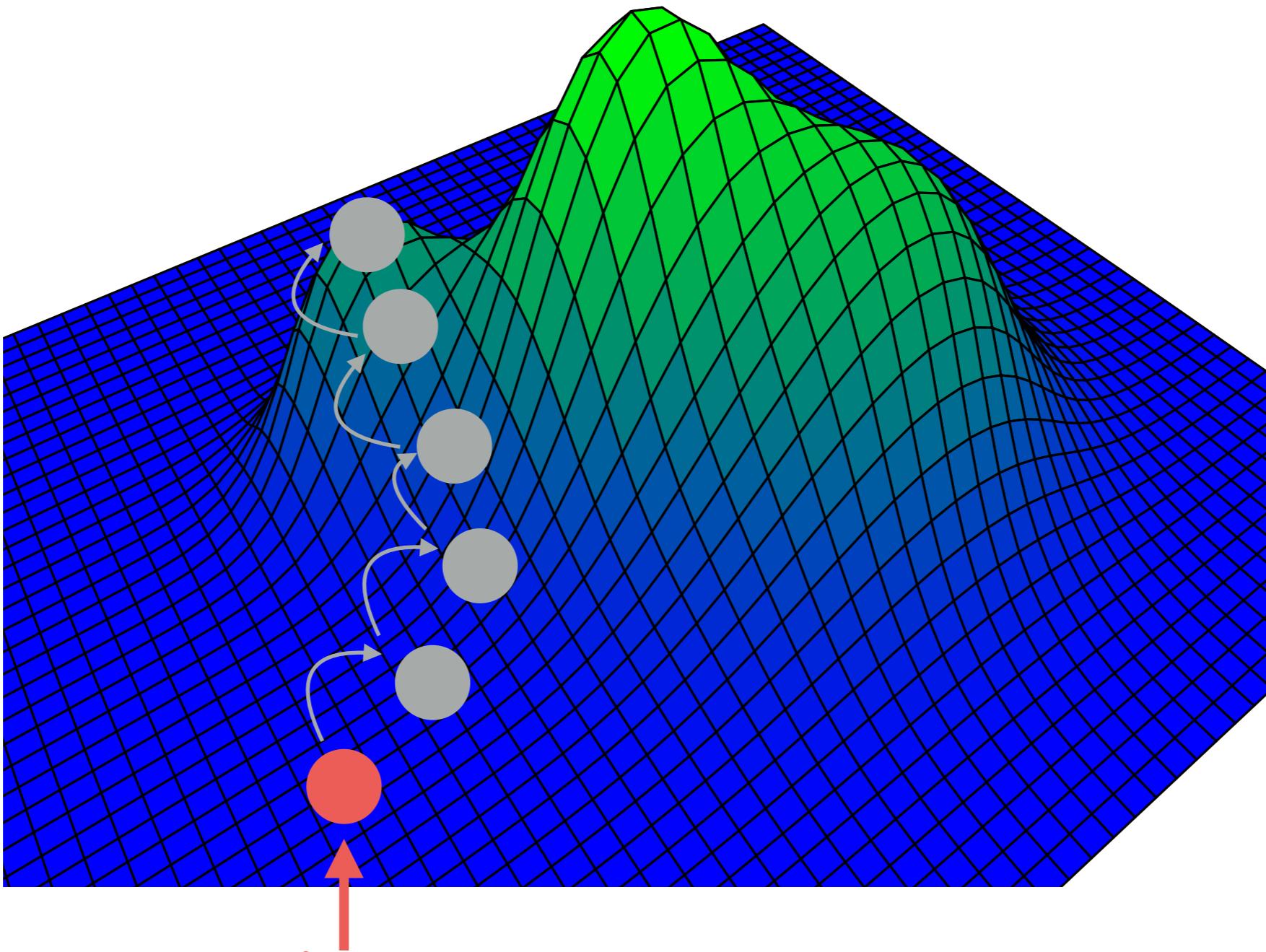
**t** =  $(t_1, \dots, t_6)$ : Branch lengths

**x, y, z**: Ancestral states

Average across them

Need to optimize

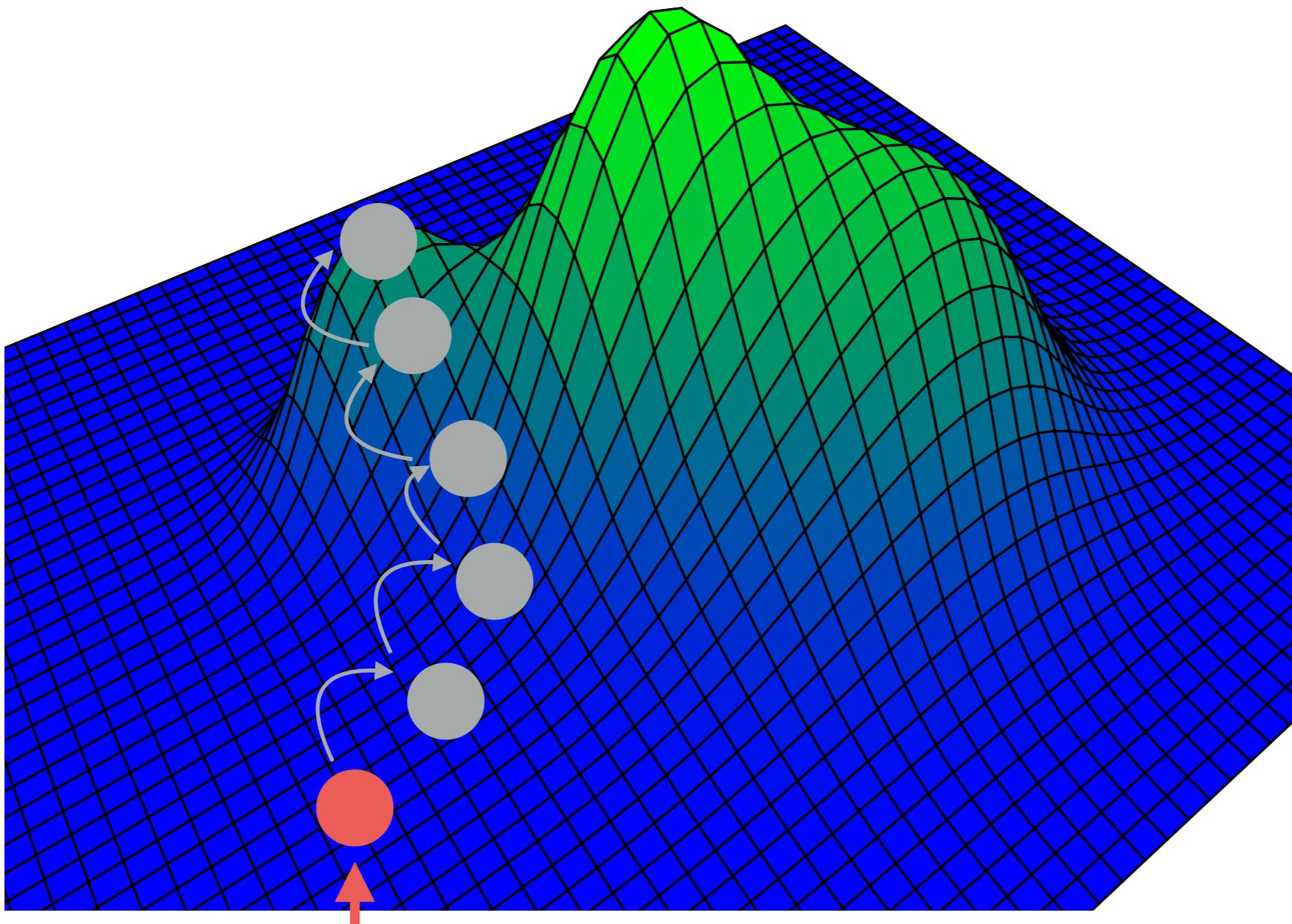
# Traverse tree space: finding the MLE



Starting tree

Nearest Neighbor Interchange (NNI)  
Subtree Pruning and Regrafting (SPR)  
Tree Bisection and Reconnection (TBR)

# Traverse tree space: finding the MLE

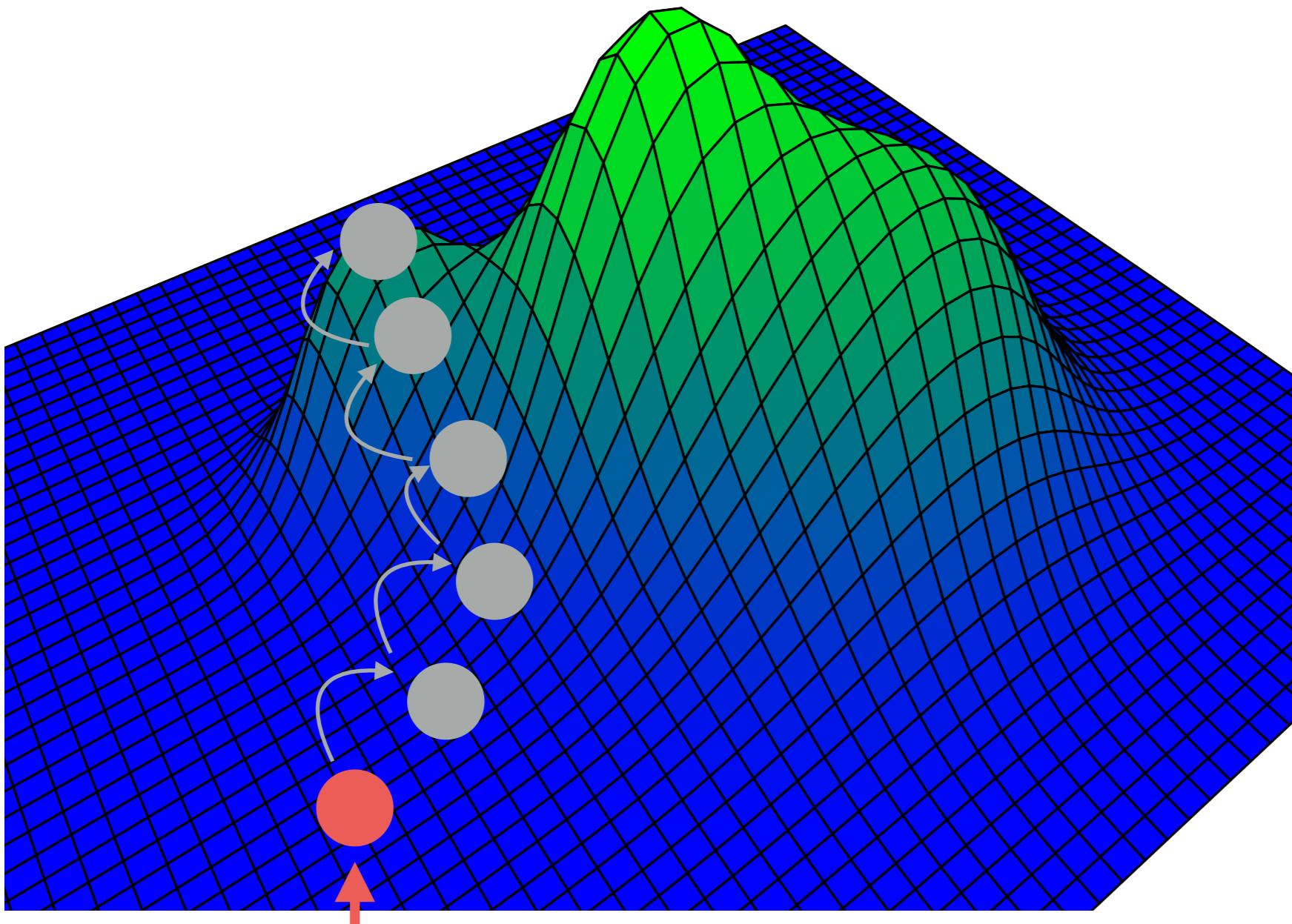


Starting tree

## Four things affecting performance:

- ▶ Starting tree
- ▶ Model chosen
- ▶ Data
- ▶ Convergence

# Traverse tree space: finding the MLE

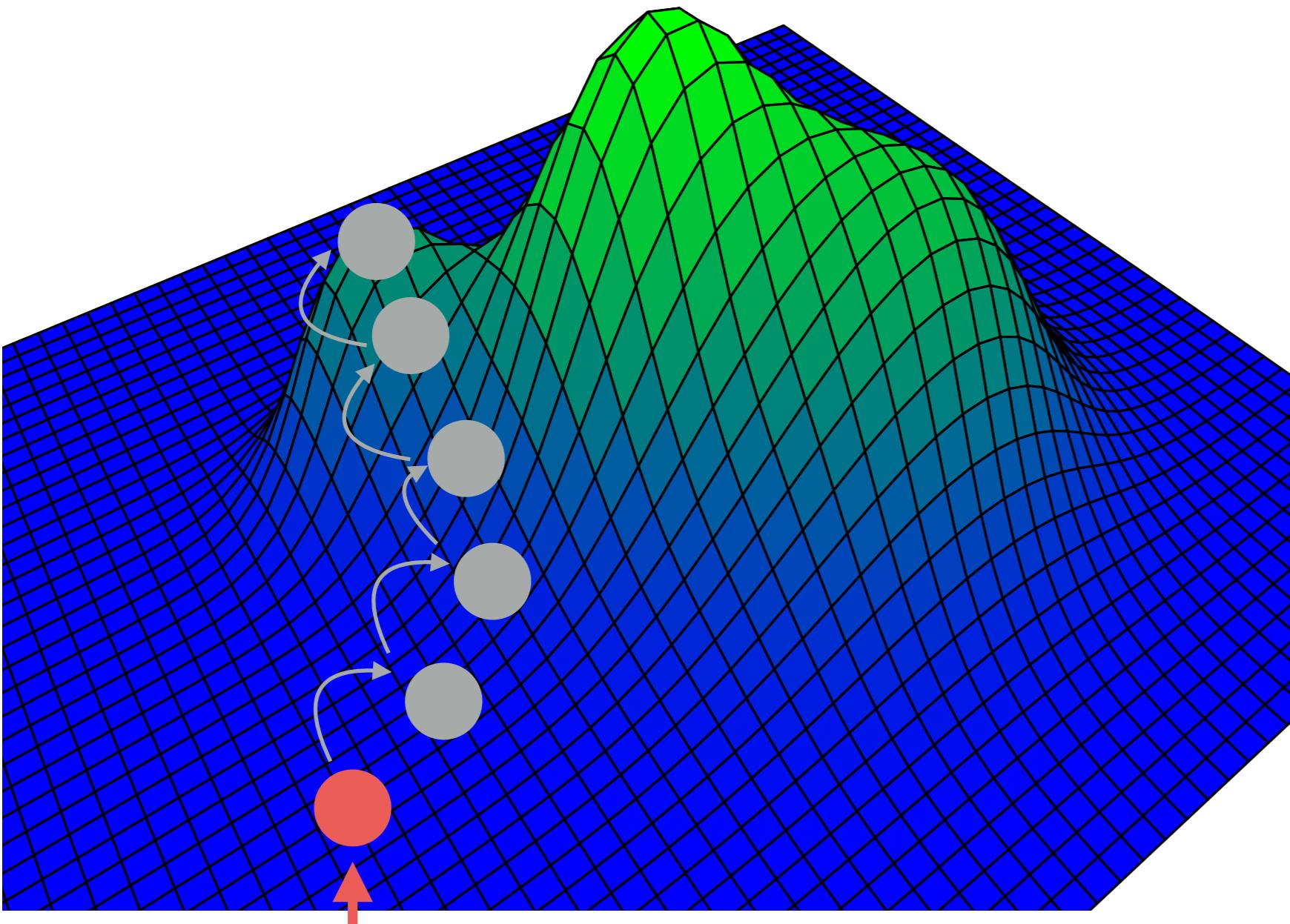


Starting tree

## Four things affecting performance:

- ▶ Starting tree
  - ▶ Affects optimization
  - ▶ Get stuck on poor likelihood region
  - ▶ Bad starting tree?
    - ▶ Best case: slows down
    - ▶ Worst case: suboptimal tree
- ▶ Model chosen
- ▶ Data
- ▶ Convergence

# Traverse tree space: finding the MLE

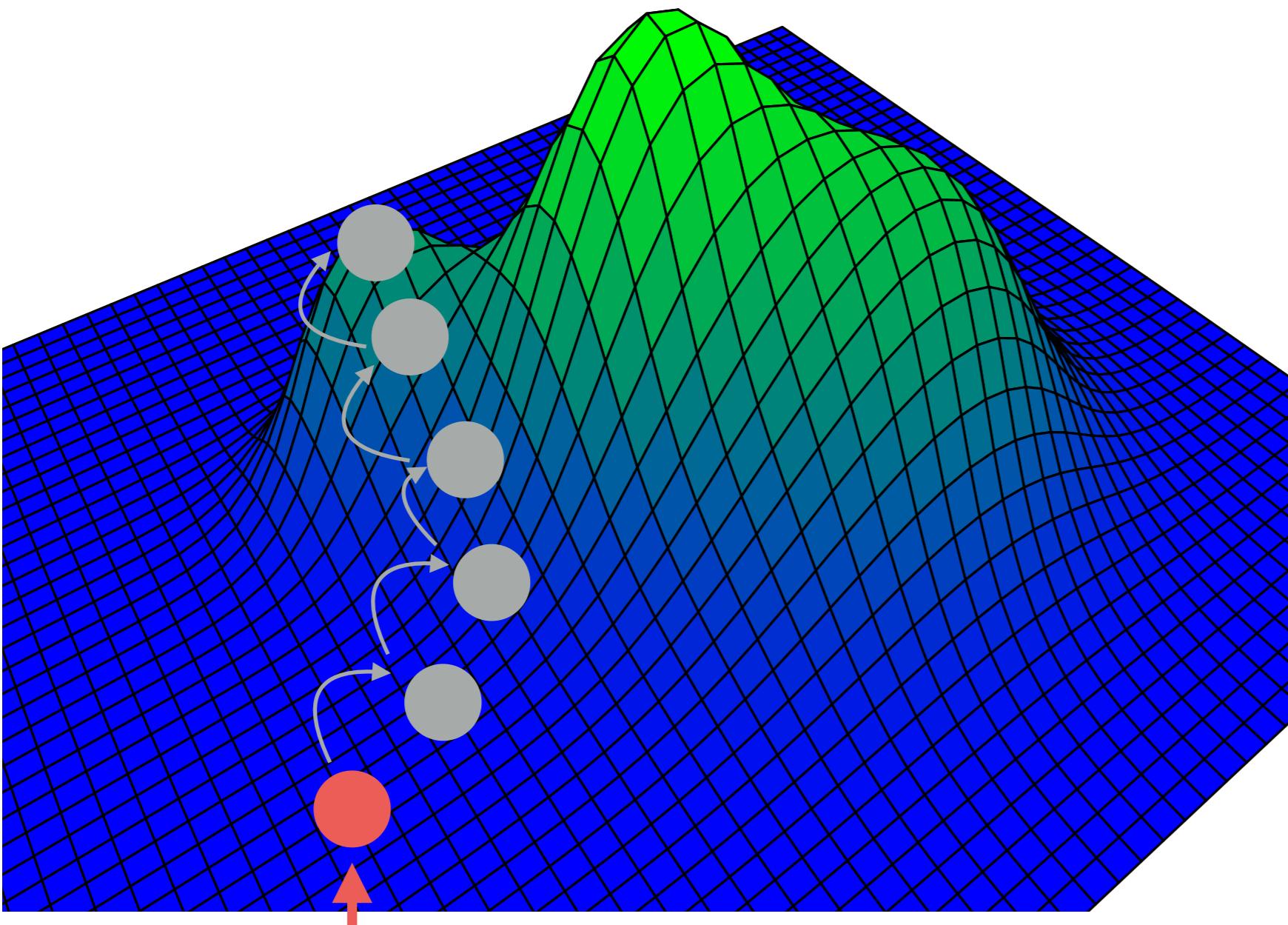


Starting tree

## Four things affecting performance:

- ▶ Starting tree
  - ▶ Affects optimization
  - ▶ Get stuck on poor likelihood region
  - ▶ Bad starting tree?
    - ▶ Best case: slows down
    - ▶ Worst case: suboptimal tree Model chosen
  - ▶ Affects shape of the surface we optimize
  - ▶ You might be optimizing the wrong function
  - ▶ Identifiability
- ▶ Data
- ▶ Convergence

# Traverse tree space: finding the MLE



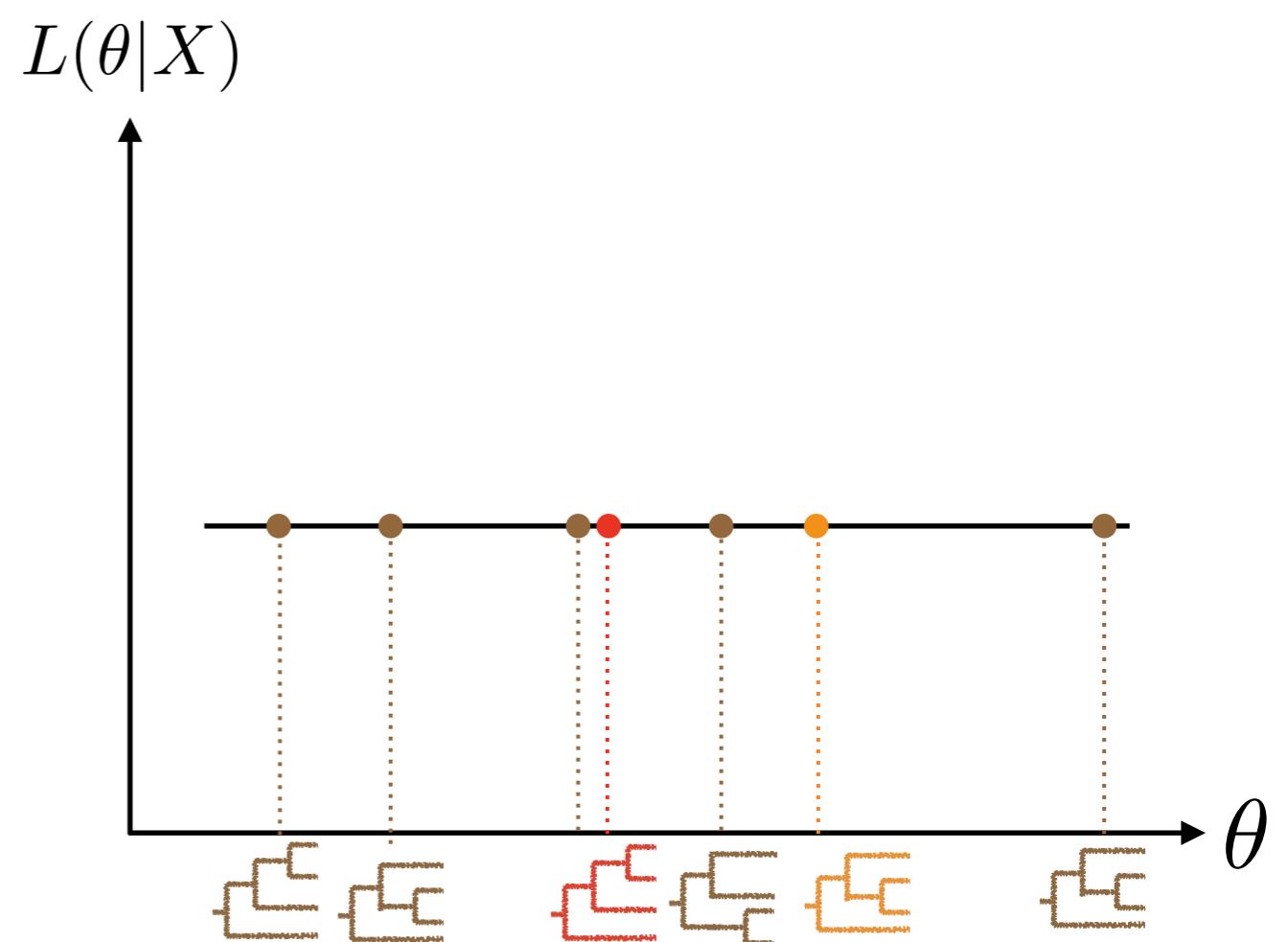
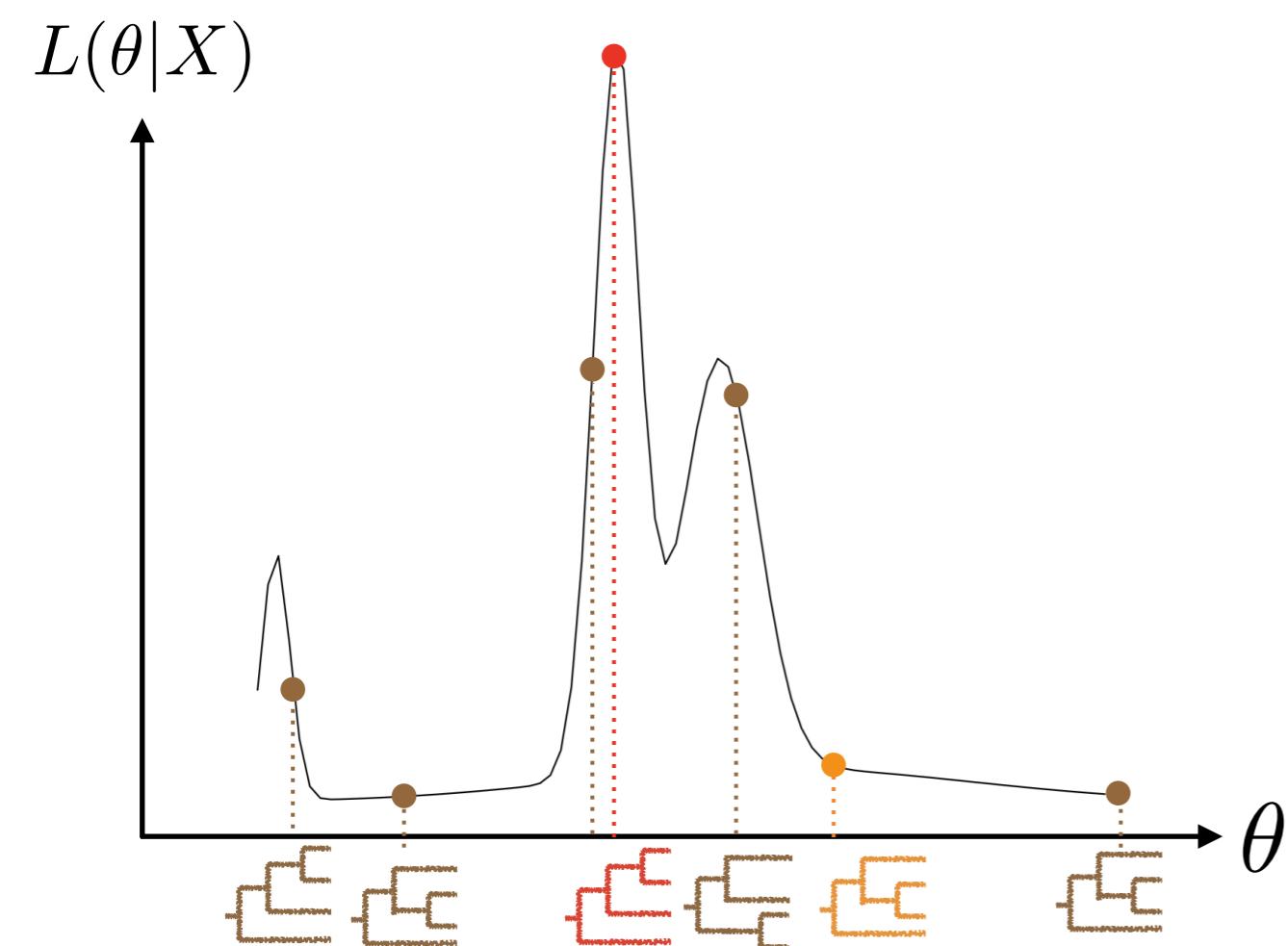
Starting tree

## Four things affecting performance:

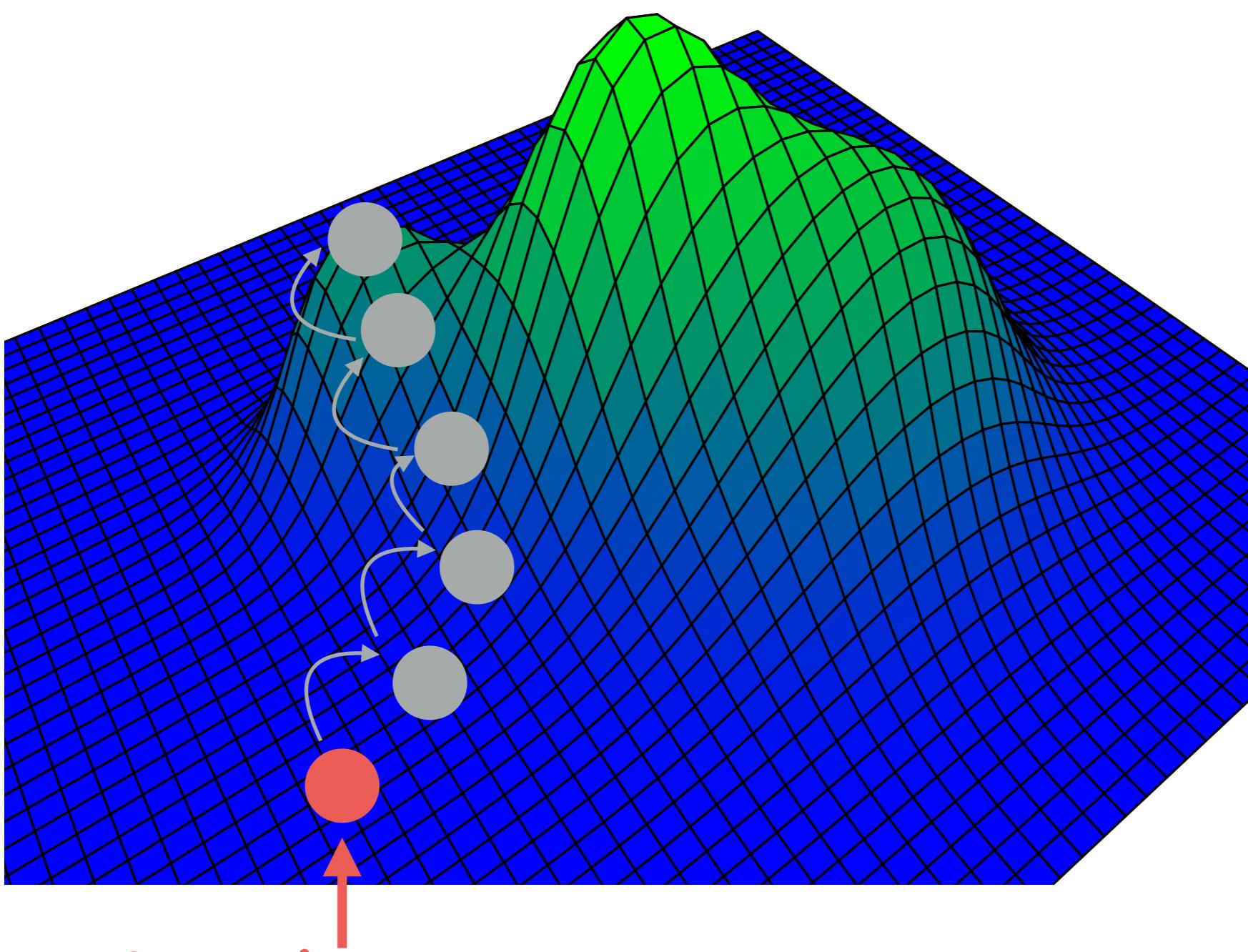
- ▶ Starting tree
  - ▶ Affects optimization
  - ▶ Get stuck on poor likelihood region
  - ▶ Bad starting tree?
    - ▶ Best case: slows down
    - ▶ Worst case: suboptimal tree
- ▶ Model chosen
  - ▶ Affects shape of the surface we optimize
  - ▶ You might be optimizing the wrong function
  - ▶ **Identifiability**
- ▶ Data
- ▶ Convergence

HAL 1.2 "Rough likelihood surface: when analyzing datasets with comparatively few sites and a large number of taxa. The key challenge with such datasets is that 100 distinct ML searches are likely to yield 100 topologically substantially different, but statistically indistinguishable trees."

# Identifiability



# Traverse tree space: finding the MLE

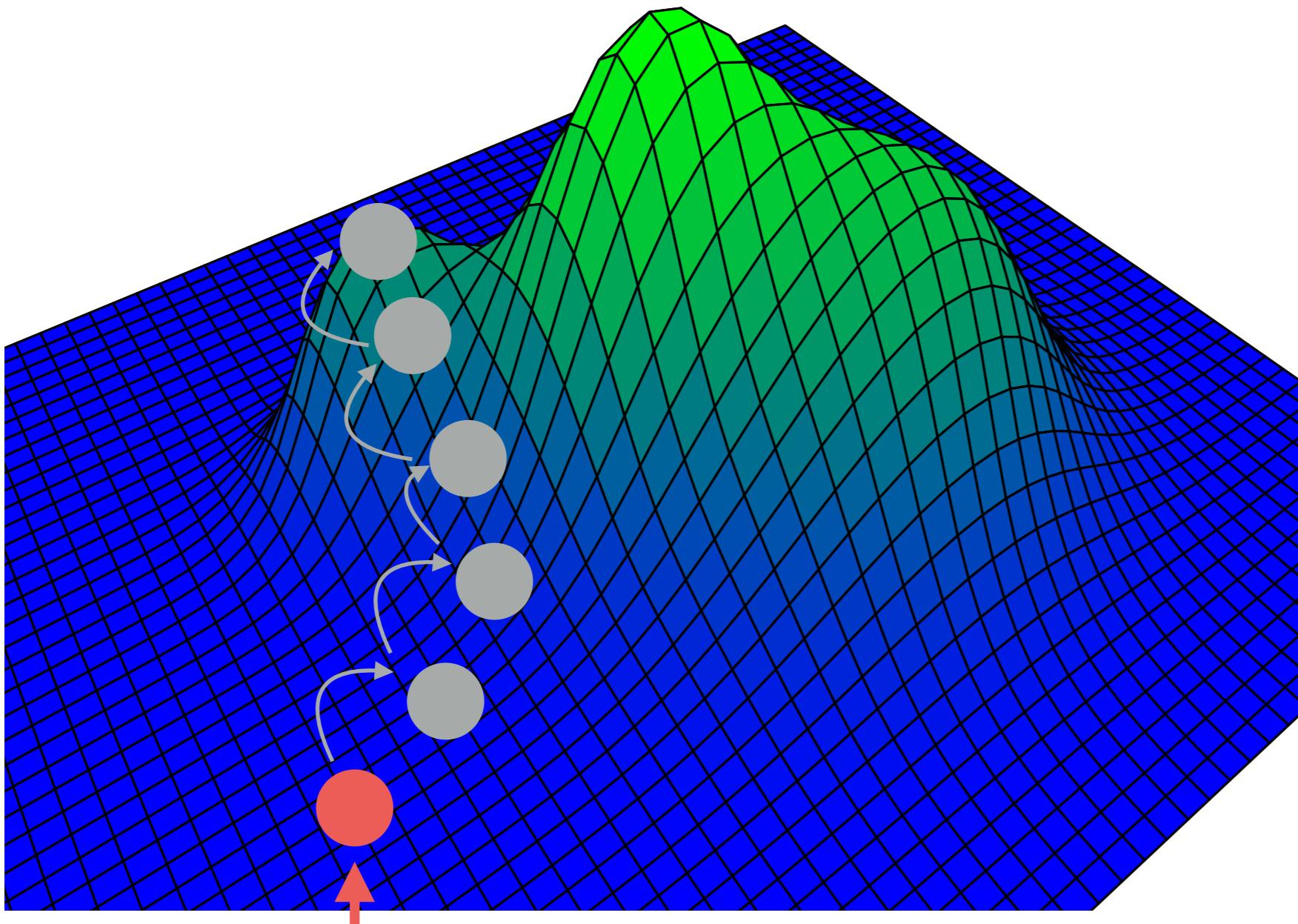


Starting tree

## Four things affecting performance:

- ▶ Starting tree
  - ▶ Affects optimization
  - ▶ Get stuck on poor likelihood region
  - ▶ Bad starting tree?
    - ▶ Best case: slows down
    - ▶ Worst case: suboptimal tree
- ▶ Model chosen
  - ▶ Affects shape of the surface we optimize
  - ▶ You might be optimizing the wrong function
  - ▶ Identifiability
- ▶ Data
  - ▶ Lack of signal (sample size or poorly chosen region)
  - ▶ Difference between data and information
  - ▶ Identifiability
- ▶ Convergence

# Traverse tree space: finding the MLE



## Four things affecting performance:

- ▶ Starting tree
  - ▶ Affects optimization
  - ▶ Get stuck on poor likelihood region
  - ▶ Bad starting tree?
    - ▶ Best case: slows down
    - ▶ Worst case: suboptimal tree
- ▶ Model chosen
  - ▶ Affects shape of the surface we optimize
  - ▶ You might be optimizing the wrong function
  - ▶ Identifiability
- ▶ Data
  - ▶ Lack of signal (sample size or poorly chosen region)
  - ▶ Difference between data and information
  - ▶ Identifiability
- ▶ Convergence
  - ▶ When do you stop the traversal of tree space?
  - ▶ Affects optimization

# Statistical Consistency

- Maximum likelihood (and Bayesian), neighbor joining, ME OLS are all statistically consistent methods
- UPGMA and maximum parsimony are not statistically consistent methods

# For next class:

- We will go over IQ-Tree and RAxML
- Each student is assigned to one software and has to read two papers for that software (software papers are short: two papers are fewer than 9 pages combined)
- Focus on the 4 things affecting performance: starting tree, model chosen, data, convergence
- We will have a class discussion followed by installing and using the software