Lecture 10

Maximum likelihood

Botany 563 — Spring 2022

- **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

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

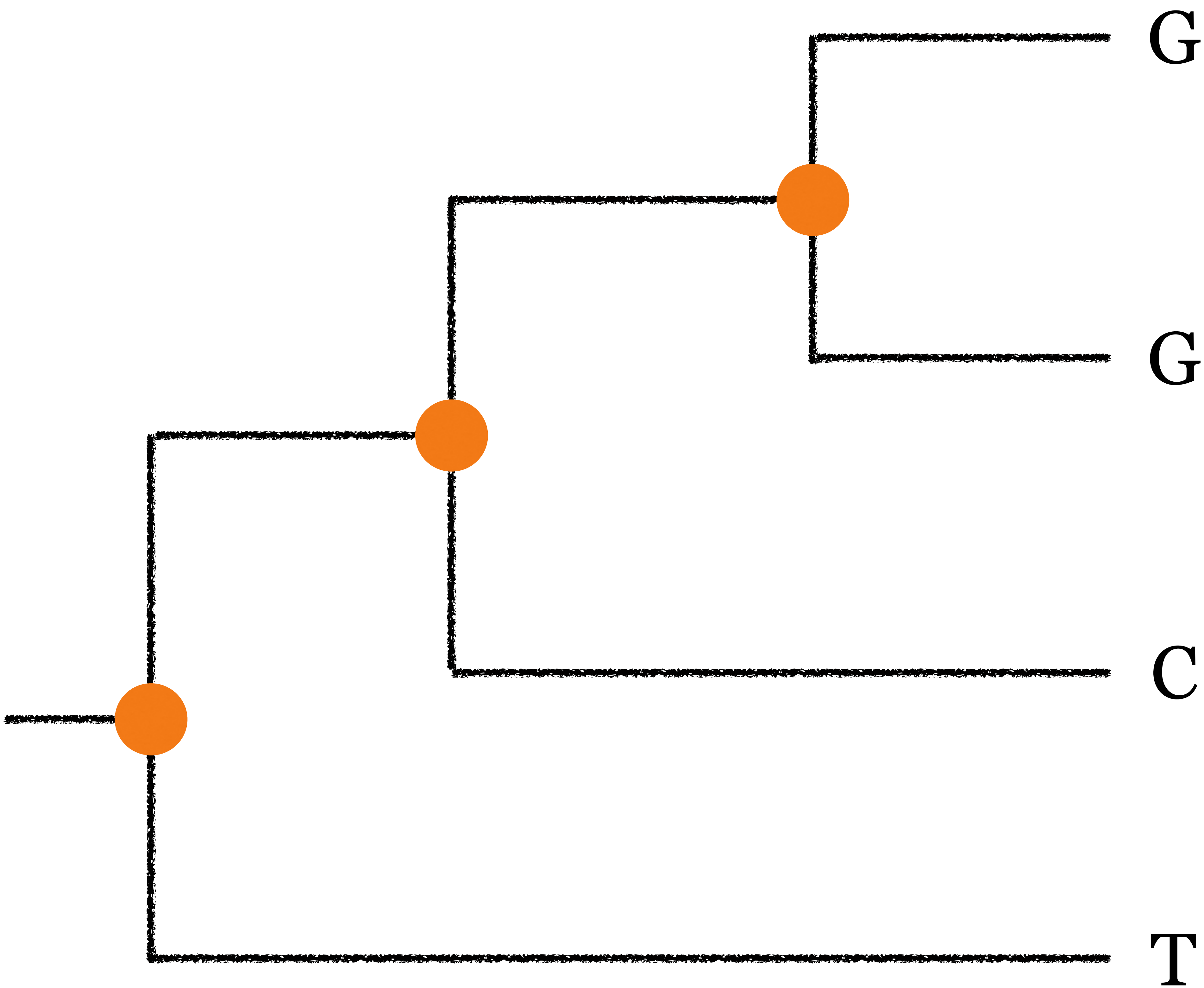
2. For a given tree, calculate the likelihood given the data and the substitution model

$$\mathcal{L}_{\mathbf{Q}}\Bigg($$

$$\Bigg| \begin{array}{c} \text{AAGTCTAG} \\ \text{AAGTCTAG} \\ \text{AACTCTAG} \\ \text{AATTCTAG} \end{array} \Bigg)$$

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

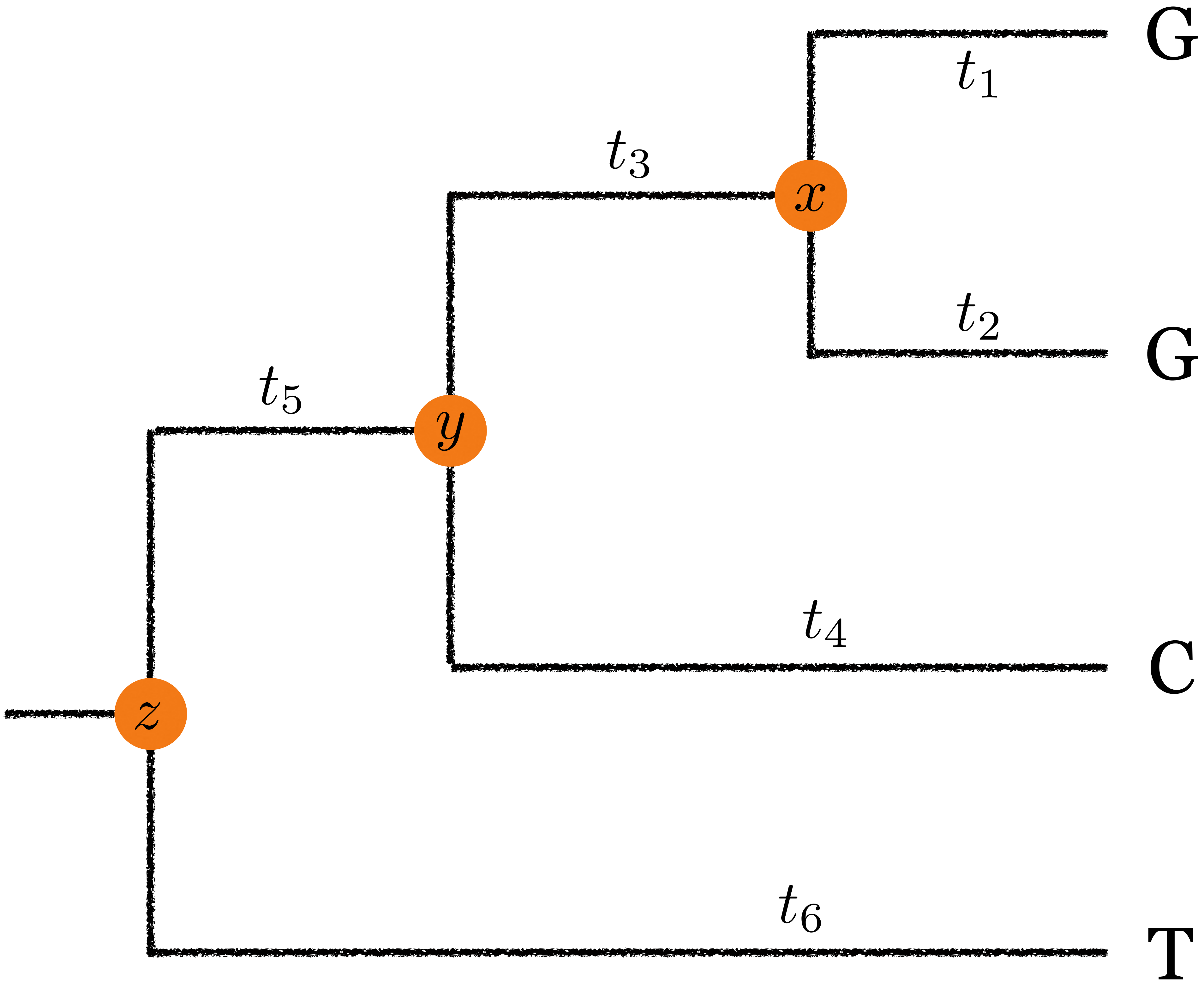


Assumption 1: The mutation process is the same at every branch of the tree

Assumption 2: We assume sites evolve independently

Assumption 3: All sites evolve the same

Calculate the likelihood for this tree



Depends on parameters:

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

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

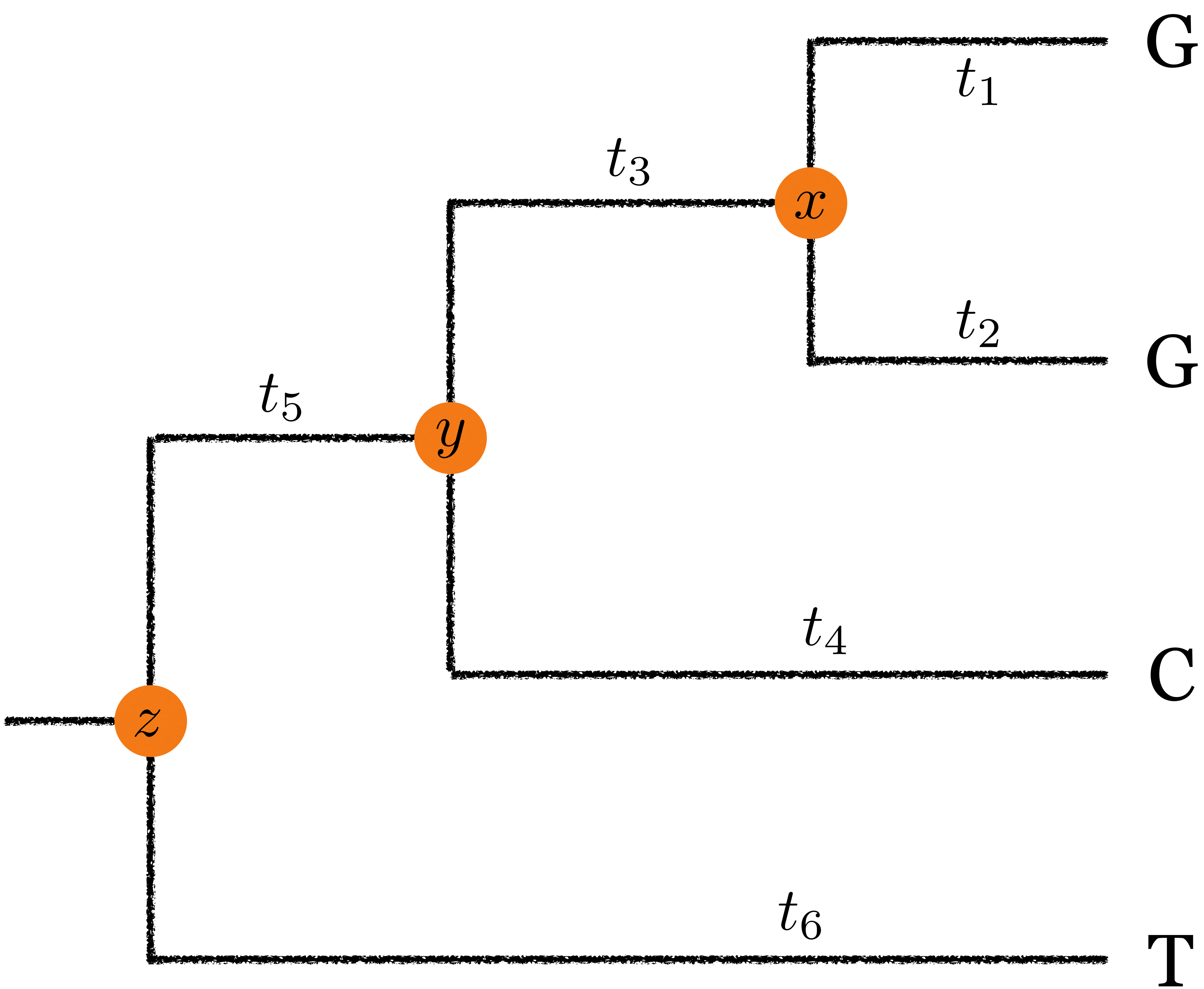
x, y, z — Ancestral states

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:

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

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

x, y, z Ancestral states

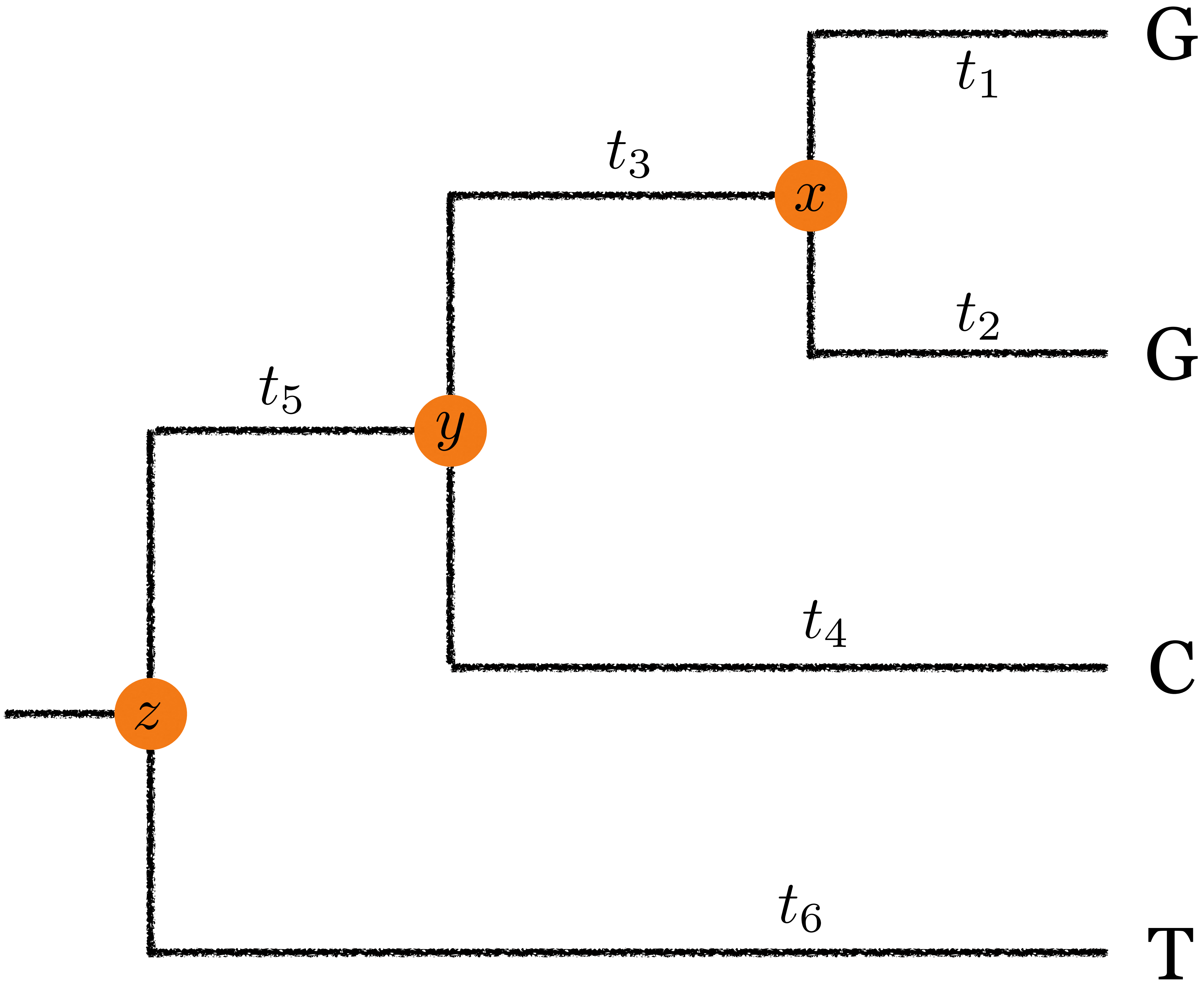
$$\mathcal{L}(T, \mathbf{Q}, \mathbf{t}, x, y, z)$$

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



Depends on parameters:

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

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

x, y, z Ancestral states

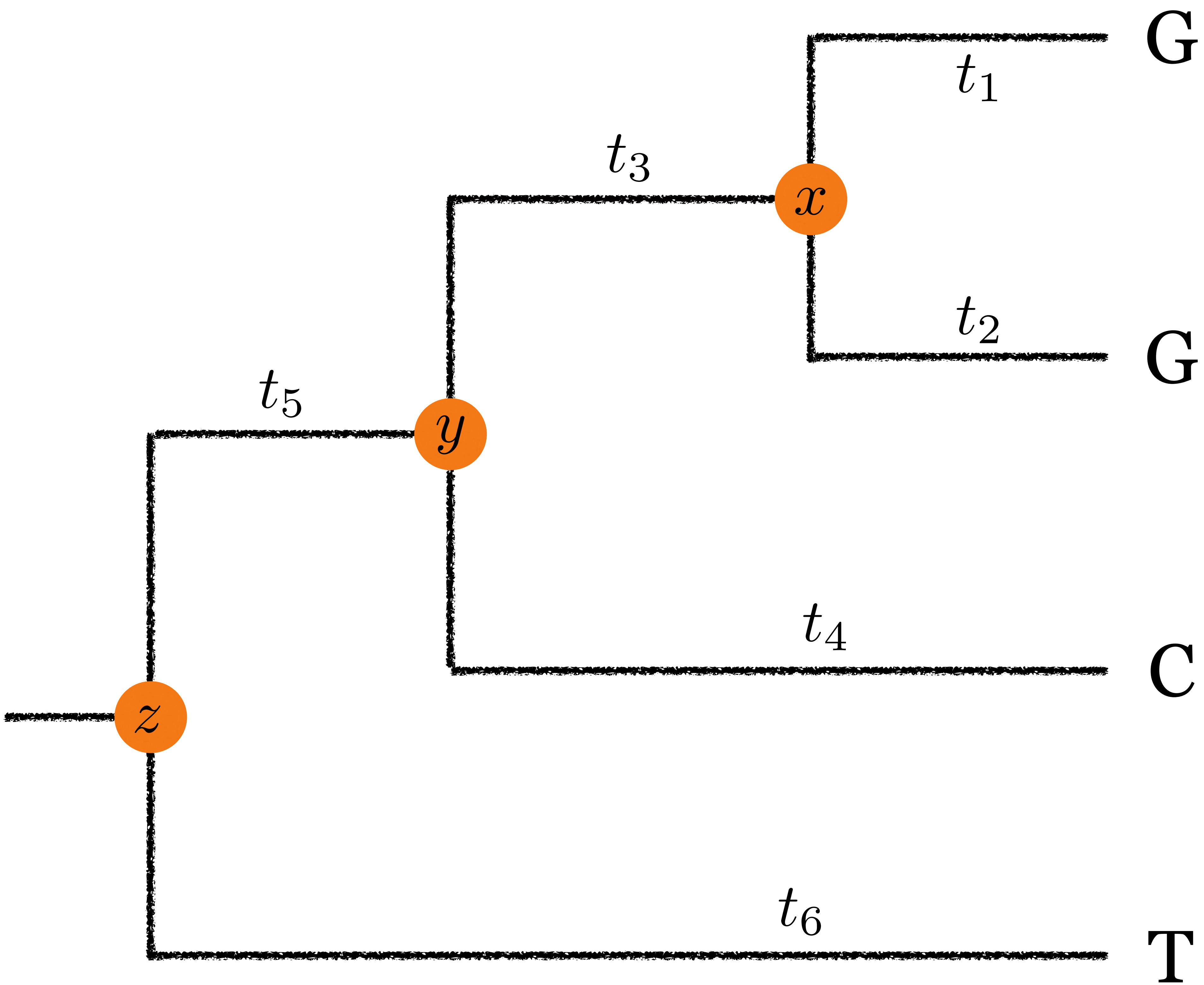
$$\mathcal{L}(T, \mathbf{Q}, \mathbf{t}, x, y, z)$$
$$= P(GGCT | T, \mathbf{Q}, \mathbf{t}, x, y, z)$$

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

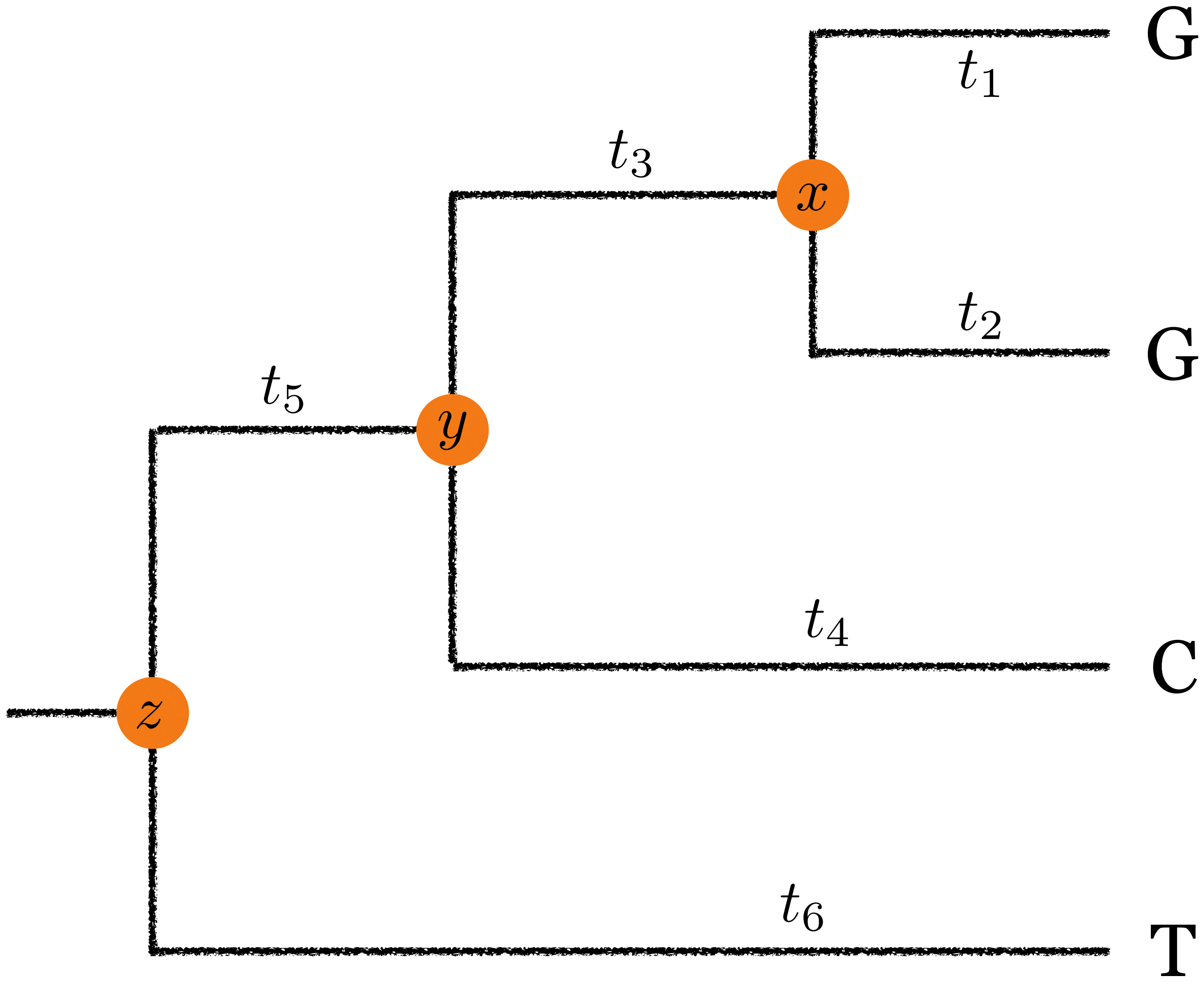


Assumption 1: The mutation process is the same at every branch of the tree

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

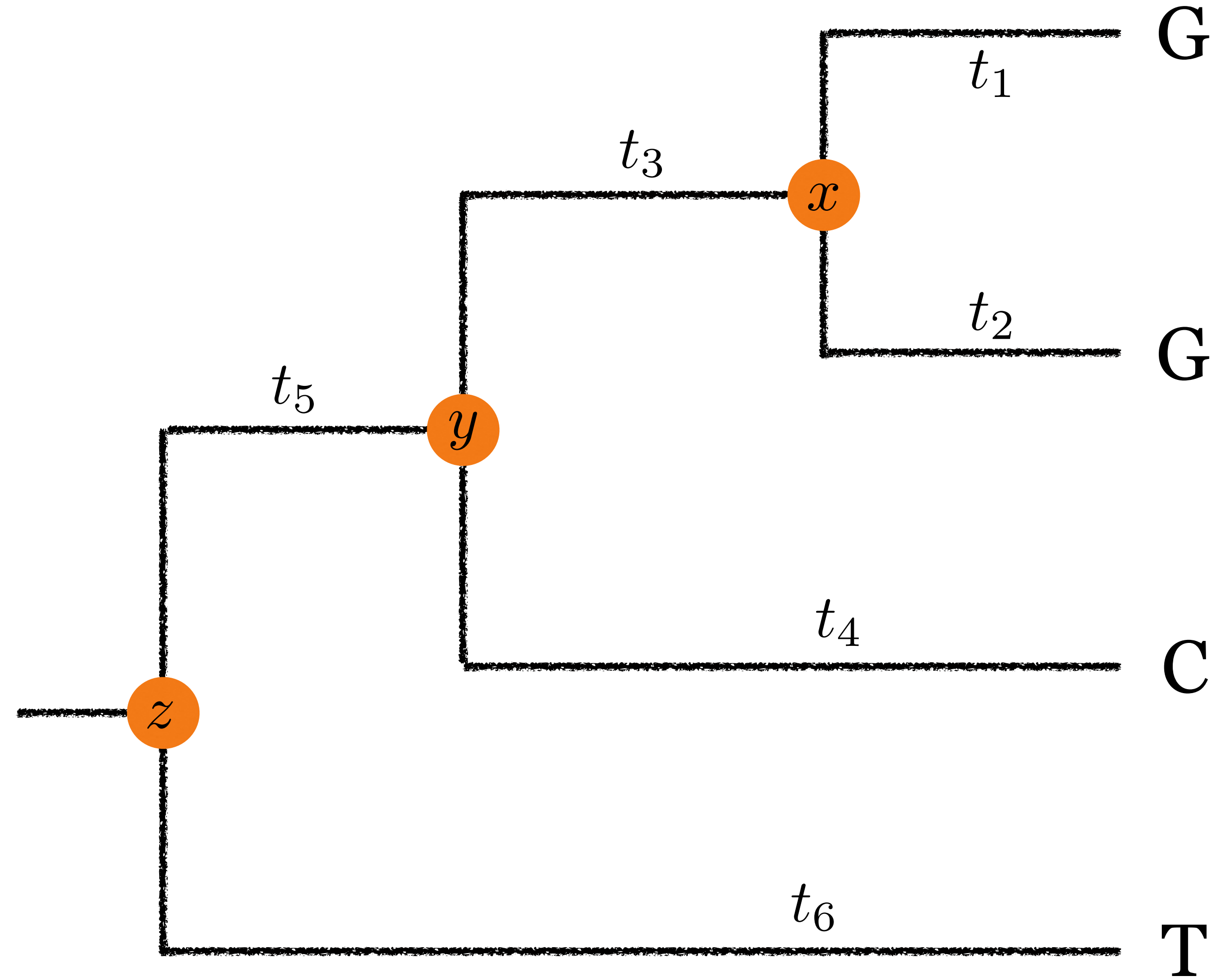
Assumption 2: We assume sites evolve independently

Assumption 3: All sites evolve the same

Calculate the likelihood for this tree

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

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



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

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

Assumption 2: We assume sites evolve independently

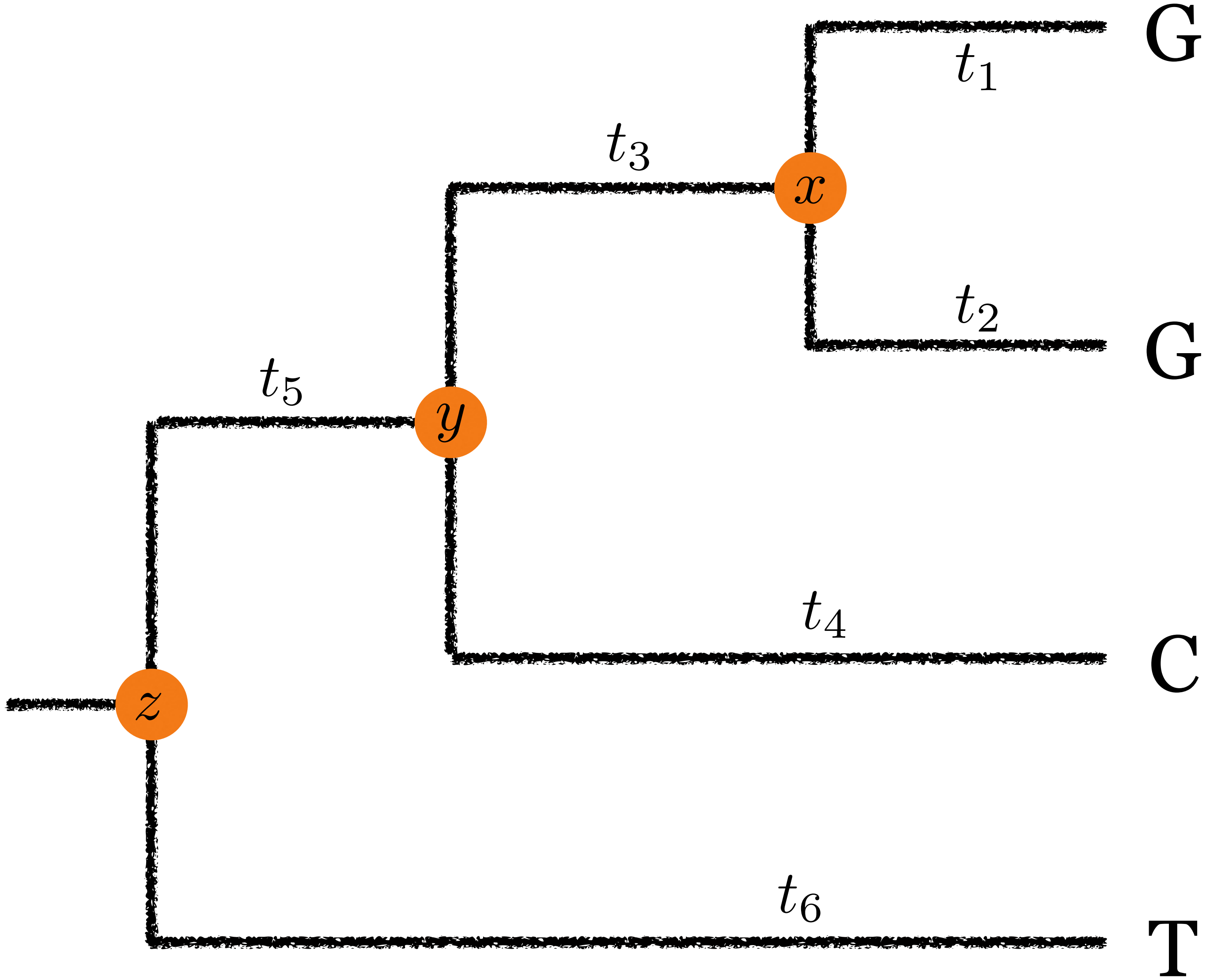
Assumption 3: All sites evolve the same

Calculate the likelihood for this tree

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

$$\mathbf{Q} = \begin{array}{c} \\ \\ \\ \\ \\ \end{array}\begin{array}{cccc} A & C & G & T \\ \end{array}$$

$$\mathbf{Q} = \begin{array}{ccccc} A & C & G & T & \\ \left[\begin{array}{c} * \\ a\pi_A \\ b\pi_A \\ c\pi_A \end{array}\right. & \begin{array}{c} a\pi_C \\ * \\ d\pi_C \\ e\pi_C \end{array} & \begin{array}{c} b\pi_G \\ d\pi_G \\ * \\ f\pi_G \end{array} & \left.\begin{array}{c} c\pi_T \\ e\pi_T \\ f\pi_T \\ * \end{array}\right] & \begin{array}{c} A \\ C \\ G \\ T \end{array} \end{array}$$



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

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

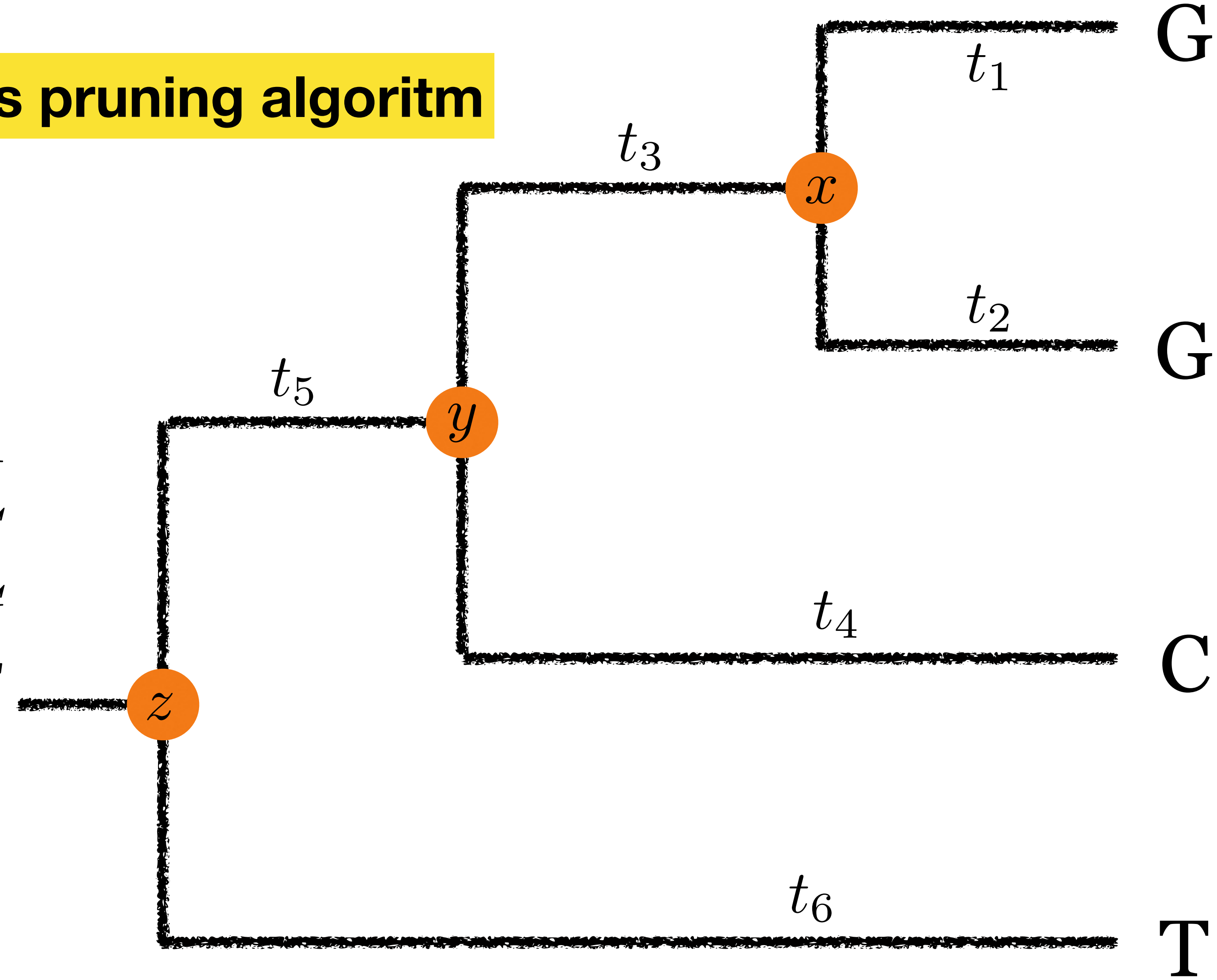
Calculate the likelihood for this tree

Felsenstein's pruning algoritm

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

$$\mathbf{Q} = \begin{array}{c} \\ \\ \\ \\ \end{array}\begin{array}{cccc} A & C & G & T \end{array}$$

$$\mathbf{Q} = \begin{array}{cc} \begin{array}{cccc} \quad A \quad & \quad C \quad & \quad G \quad & \quad T \quad \end{array} & \\ \left[\begin{array}{cccc} * & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & * & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & * & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & * \end{array}\right] & \begin{array}{c} A \\ C \\ G \\ T \end{array} \end{array}$$



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

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

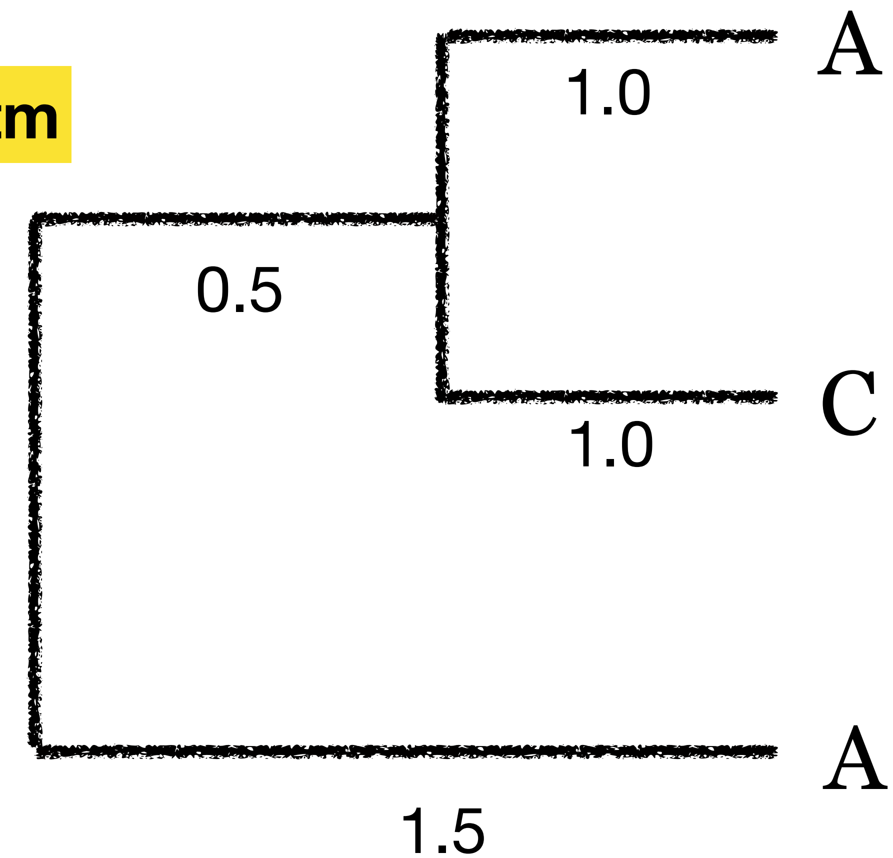
Assumption 3: All sites evolve the same

Calculate the likelihood for this tree

Felsenstein's pruning algoritm

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

$$\begin{array}{cc} & \begin{array}{cccc} A & C & G & T \end{array} \\ \mathbf{Q} = & \begin{bmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{bmatrix} \begin{array}{c} A \\ C \\ G \\ T \end{array} \end{array}$$



Full example in video on canvas

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

$$\mathcal{L}_{\mathbf{Q}}\Bigg($$  $$\Bigg| \begin{array}{l} \text{AAGTCTAG} \\ \text{AAGTCTAG} \\ \text{AACTCTAG} \\ \text{AATTCTAG} \end{array} \Bigg)$$

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

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

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

$\mathscr{L}_{\mathbf{Q}}\Big($  $\Big|$ AAGTCTAG AAGTCTAG AACTCTAG AATTCTAG $\Big)$

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

Depends on parameters:

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

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

x, y, z Ancestral states

Maximum likelihood

1. Choose a substitution model

$$\mathbf{P}(t) =$$

	A	C	G	T
A				
C				
G				
T				

$$= e^{\mathbf{Q}\mu t}$$

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

$\mathcal{L}_{\mathbf{Q}}($  $|$ 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:

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

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

Branch lengths

$\cancel{x, y, z}$ ~~Ancestral states~~

Average across them

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

$\mathcal{L}_{\mathbf{Q}}\Big($  $\Big|$ AAGTCTAG AAGTCTAG AACTCTAG AATTCTAG $\Big)$

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

Depends on parameters:

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

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

Branch lengths

Need to optimize

$\cancel{x, y, z}$ ~~Ancestral states~~

Average across them

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

$$\mathcal{L}_{\mathbf{Q}}\Bigg($$  $$\Bigg| \begin{matrix} \text{AAGTCTAG} \\ \text{AAGTCTAG} \\ \text{AACTCTAG} \\ \text{AATTCTAG} \end{matrix} \Bigg)$$

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

Depends on parameters:

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

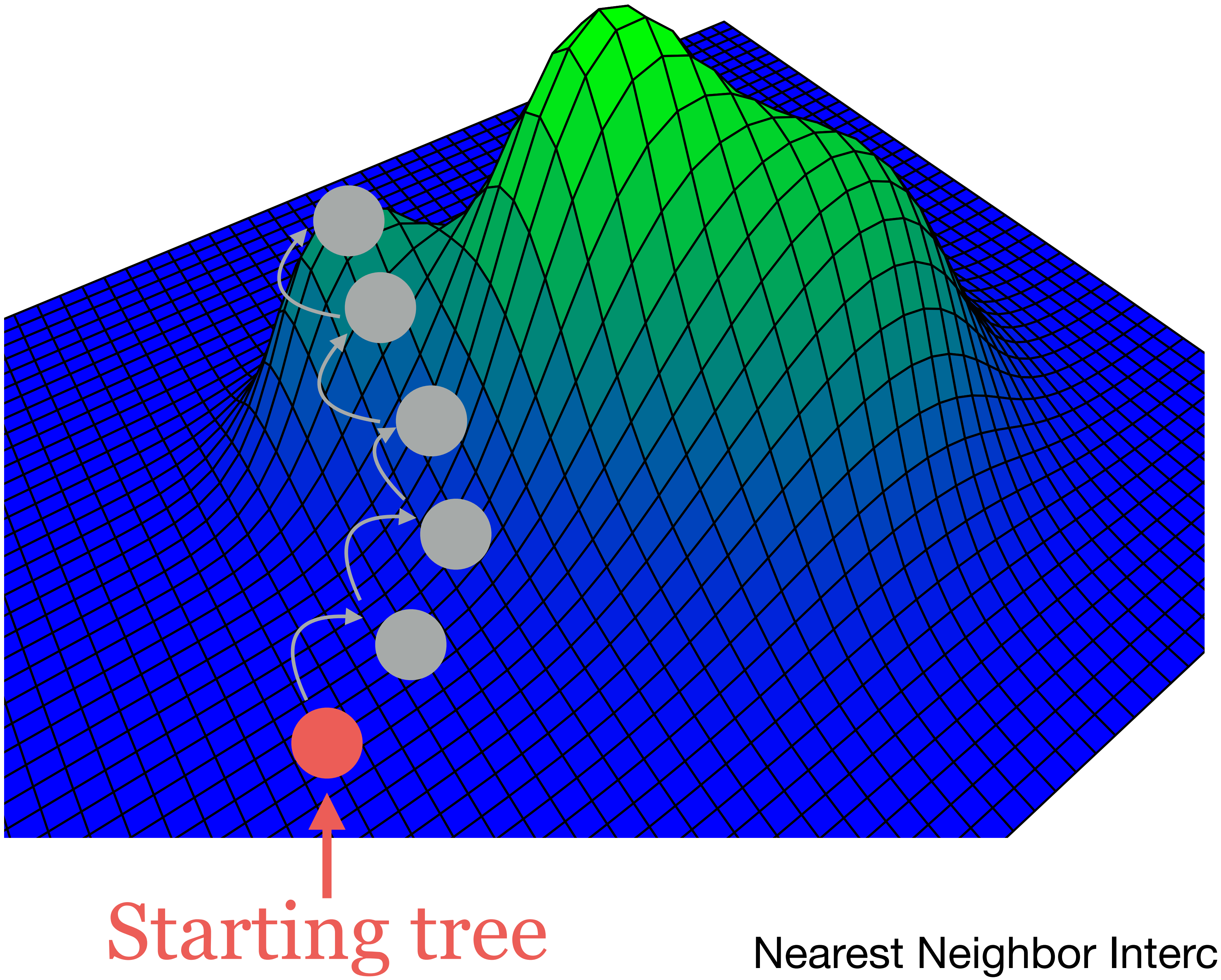
$\mathbf{t} = (t_1, \ldots, t_6)$ — Branch lengths

$\cancel{x, y, z}$ ~~Ancestral states~~

Average across them

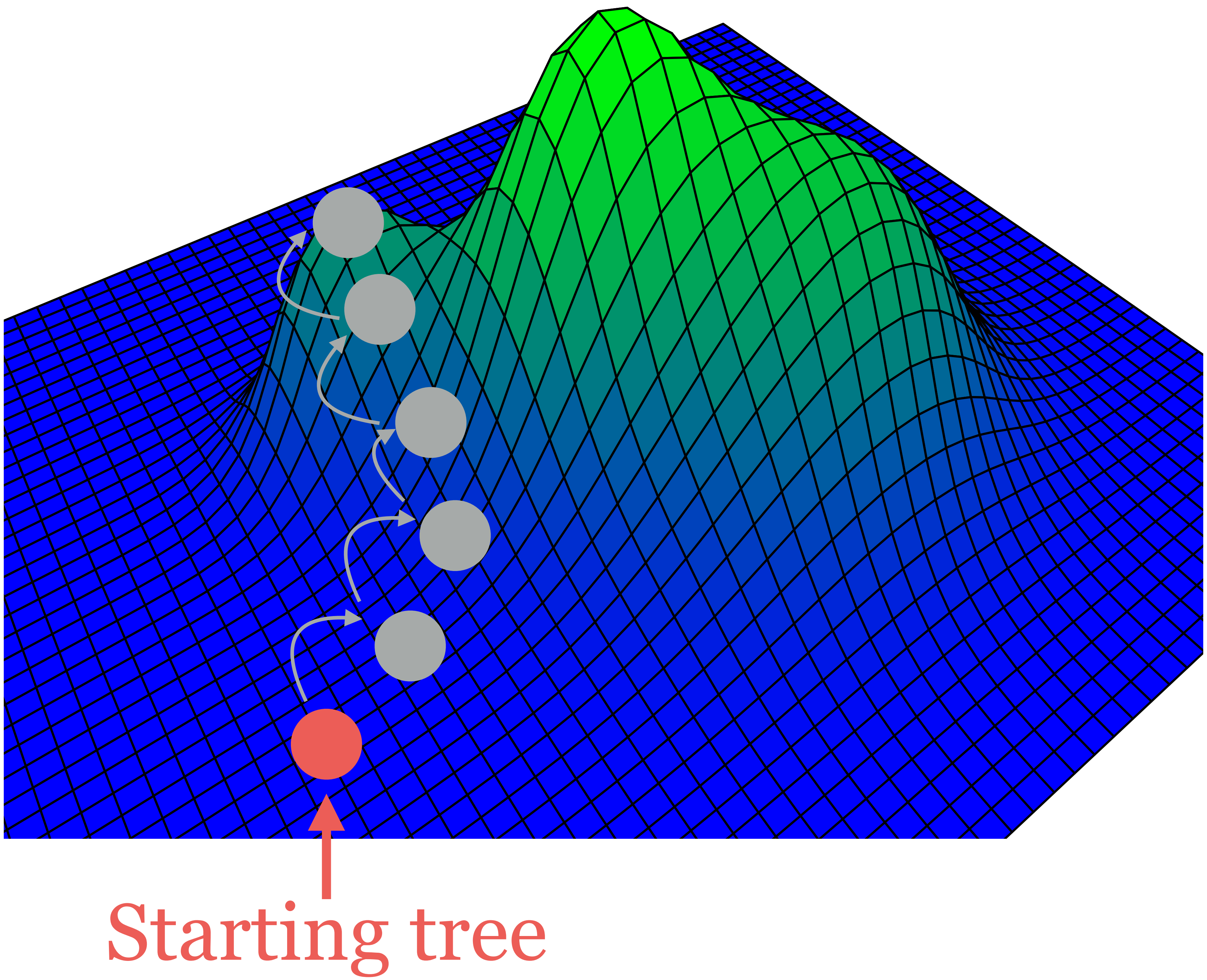
Need to optimize

Traverse tree space: finding the MLE



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

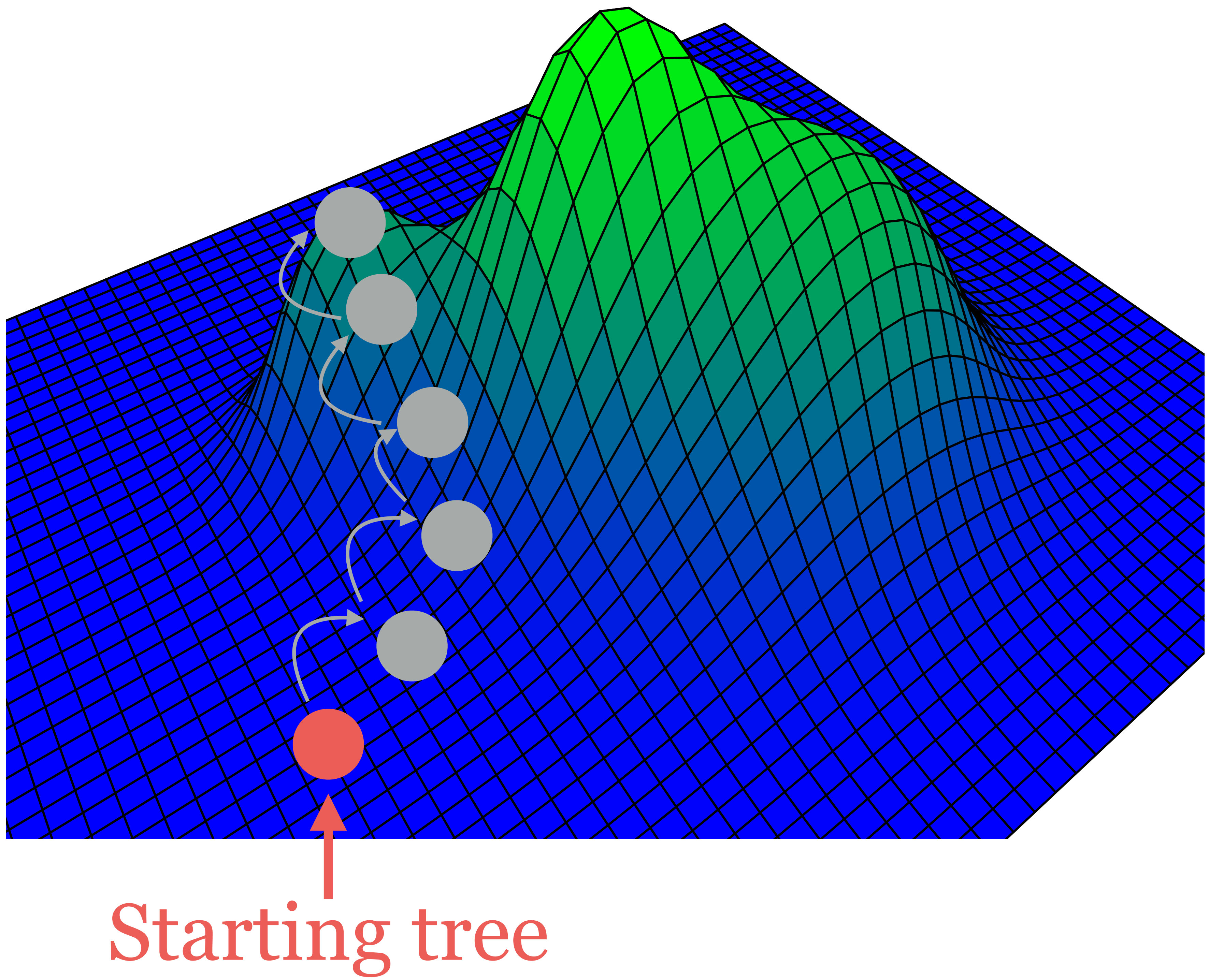
Traverse tree space: finding the MLE



Four things affecting performance:

- Starting tree
- Model chosen
- Data
- Convergence

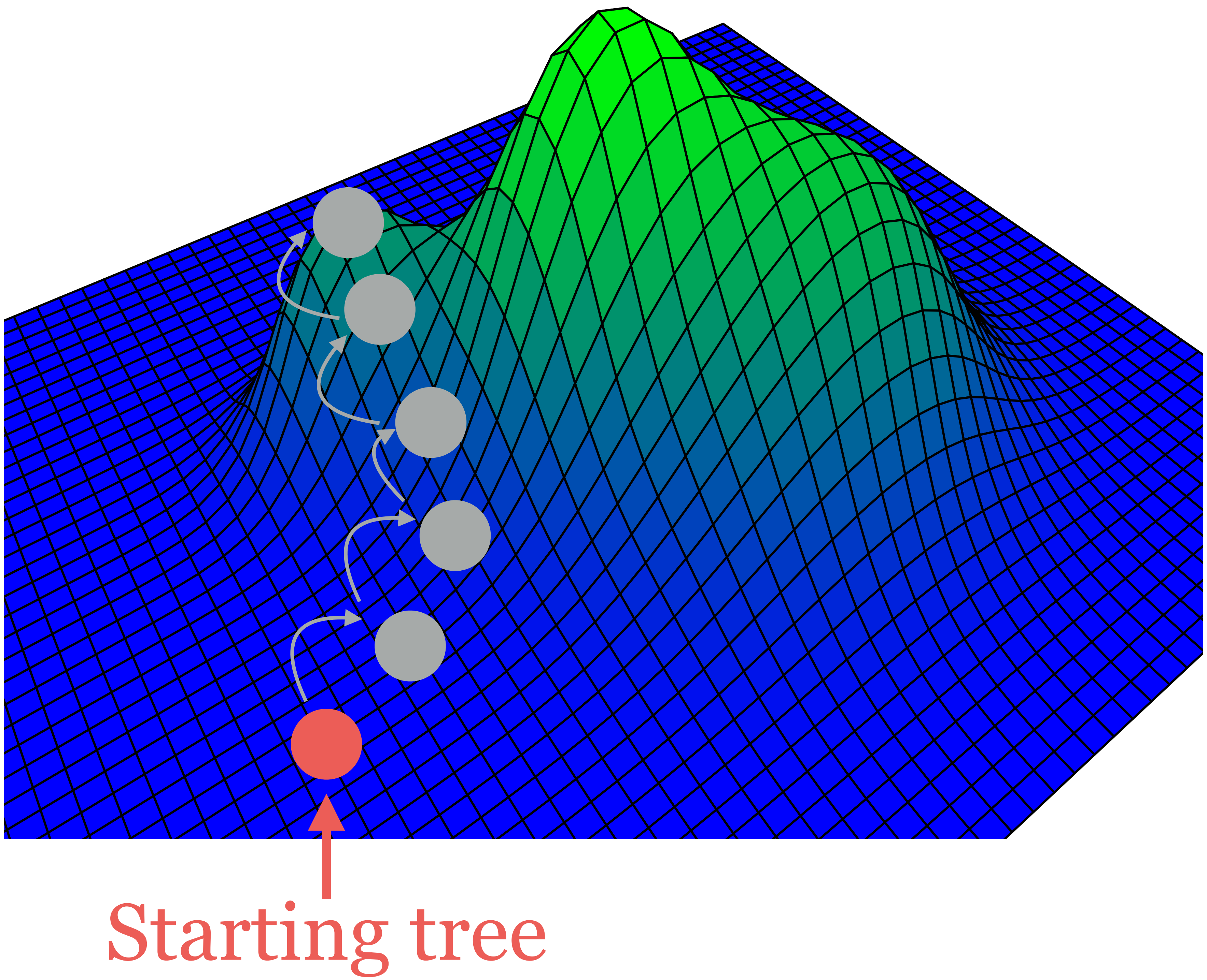
Traverse tree space: finding the MLE



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
- Data
- Convergence

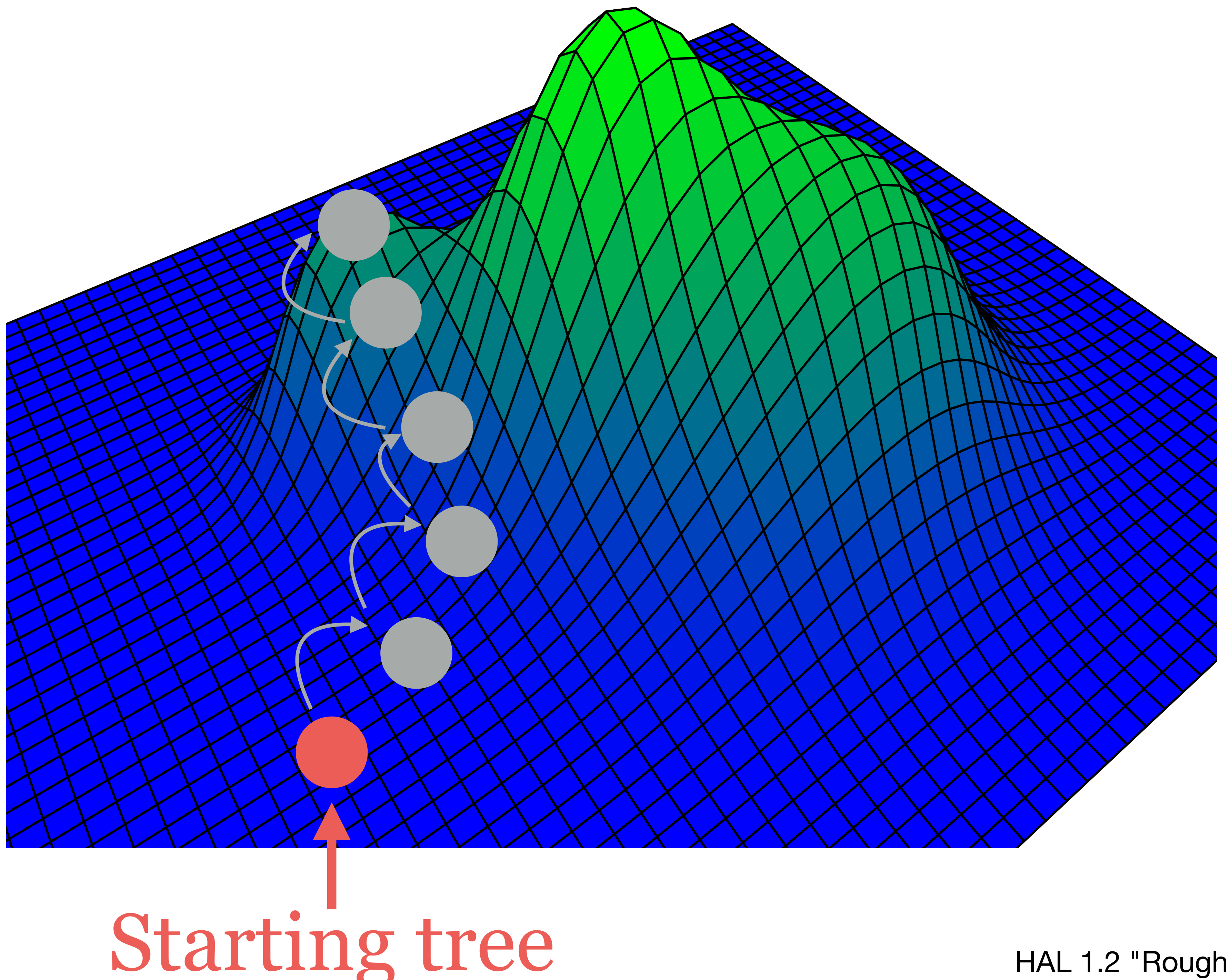
Traverse tree space: finding the MLE



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
- Convergence

Traverse tree space: finding the MLE

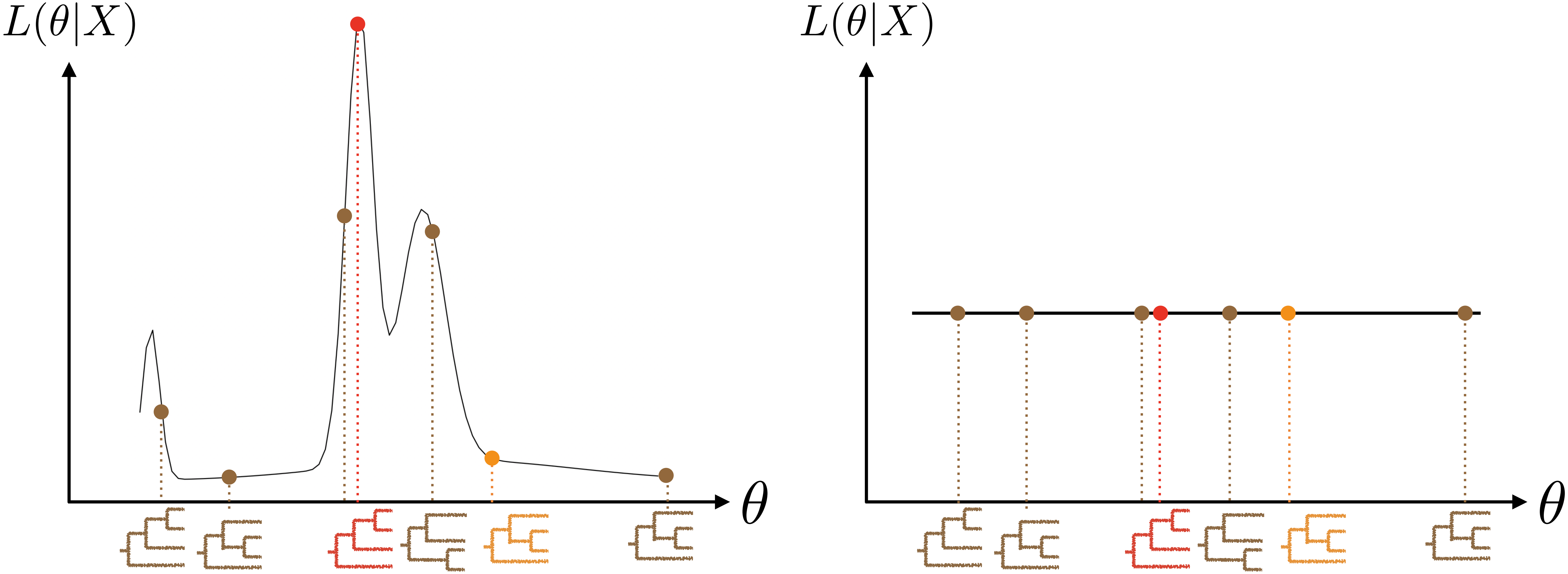


Four things affecting performance:

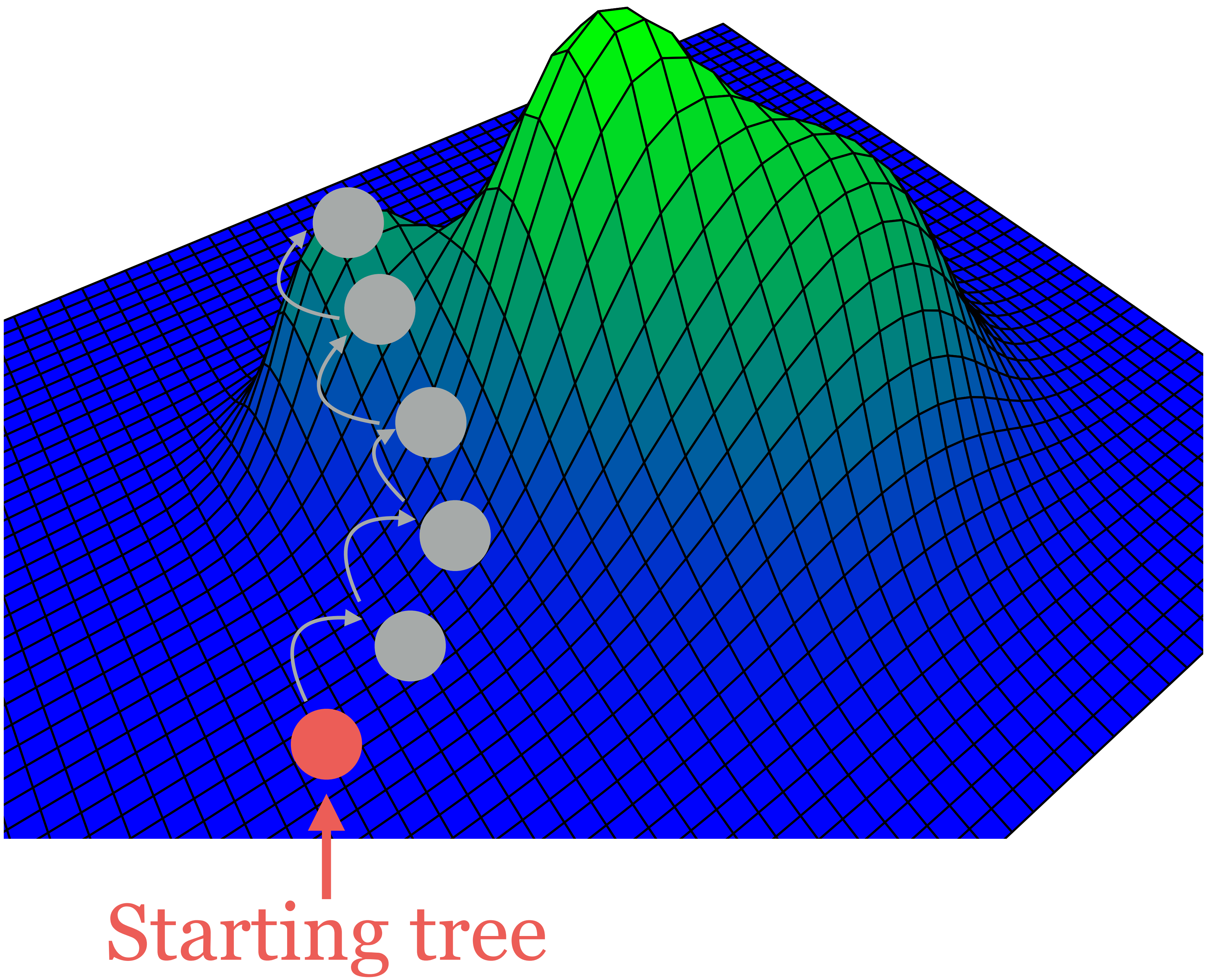
- Starting tree
  - Affects optimization
  - Get stuck on poor likelihood region
  - Bad starting tree?
    - Best case: slows down
    - Worst case: suboptimal tree
- Model chosen
  - Affects shape of the surface we optimize
  - You might be optimizing the wrong function
  - Identifiability
- Data
-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



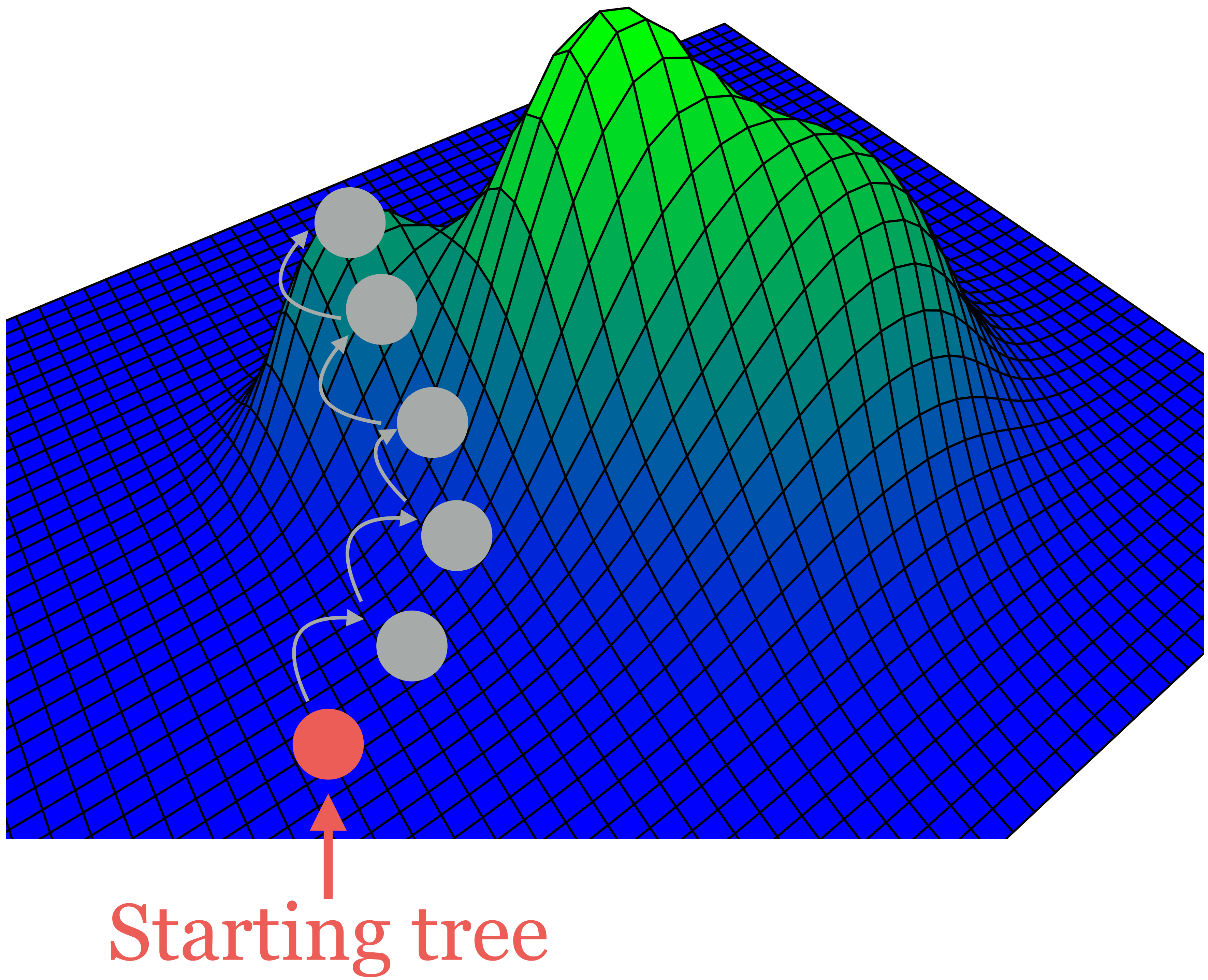
Traverse tree space: finding the MLE



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