Generative continuous time model reveals epistatic signatures in protein evolution

Andrea Pagnani, **Pierre Barrat-Charlaix**

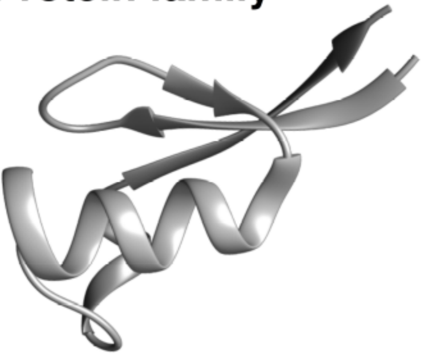
DISAT, Politecnico di Torino
CQSB, Sorbonne Université

Politecnico
di Torino
1859

SORBONNE
UNIVERSITÉ

Statistical modeling of protein sequences

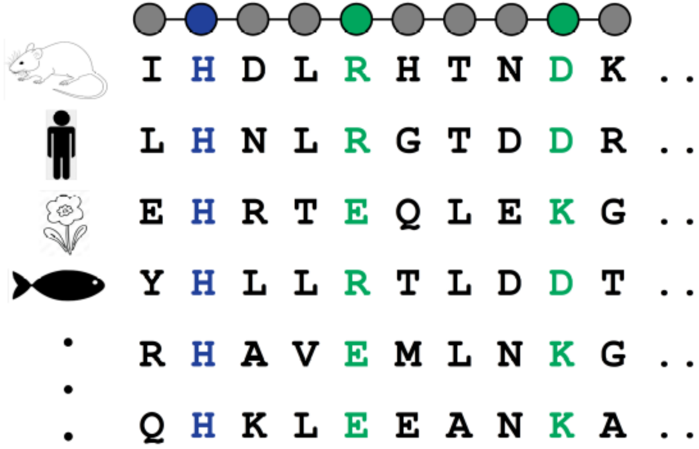
Protein family



In different species:
~ same structure
~ same function

Evolutionary constraints

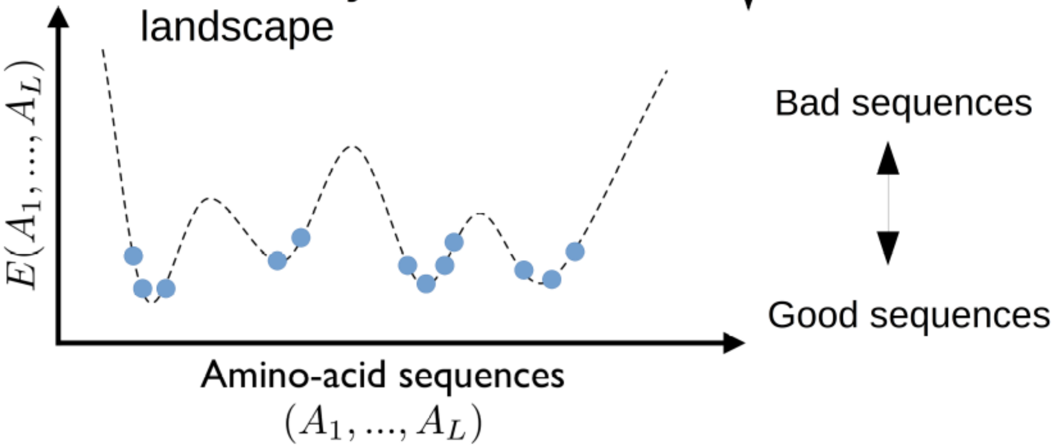
Multiple Sequence Alignment (MSA)



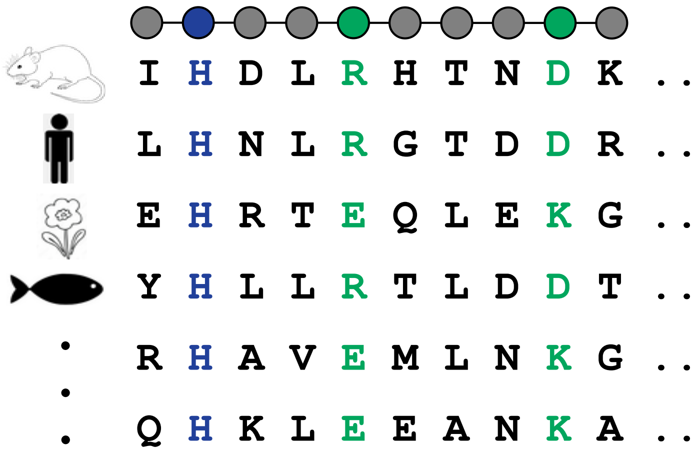
I H D L R H T N D K . .
L H N L R G T D D R . .
E H R T E Q L E K G . .
Y H L L R T L D D T . .
R H A V E M L N K G . .
Q H K L E E A N K A . .

Learning

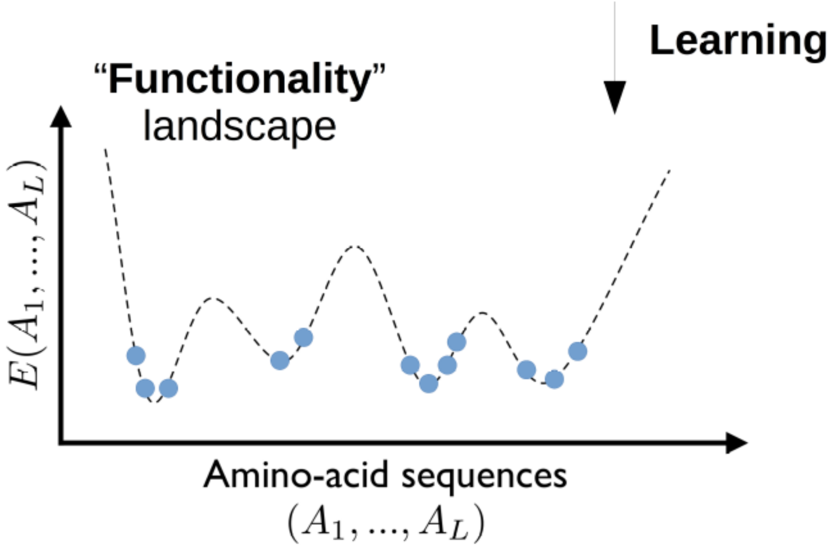
“**Functionality**” landscape



Statistical modeling of protein sequences



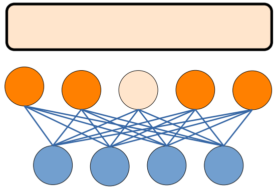
```
I H D L R H T N D K ..
L H N L R G T D D R ..
E H R T E Q L E K G ..
Y H L L R T L D D T ..
R H A V E M L N K G ..
Q H K L E E A N K A ..
```



Potts model

$$P(a_1,\ldots,a_N) = \frac{1}{Z}\exp\left(\sum_{i,j=1}^{L} \overset{\text{Couplings}}{J_{ij}(a_i,a_j)} + \sum_{i=1}^{L} \overset{\text{Fields}}{h_i(a_i)}\right)$$

Also: Deep models, RBMs, ...



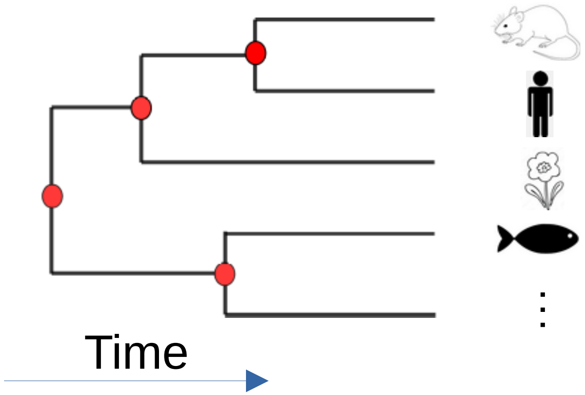
Generative model → **Design functional synthetic proteins!**

Key ingredient: **Epistasis**

→ Columns of the MSA are **not** independent

Modeling evolution?

Evolutionary history



Extant sequences

I **H** D L **R** H T N **D** K ..

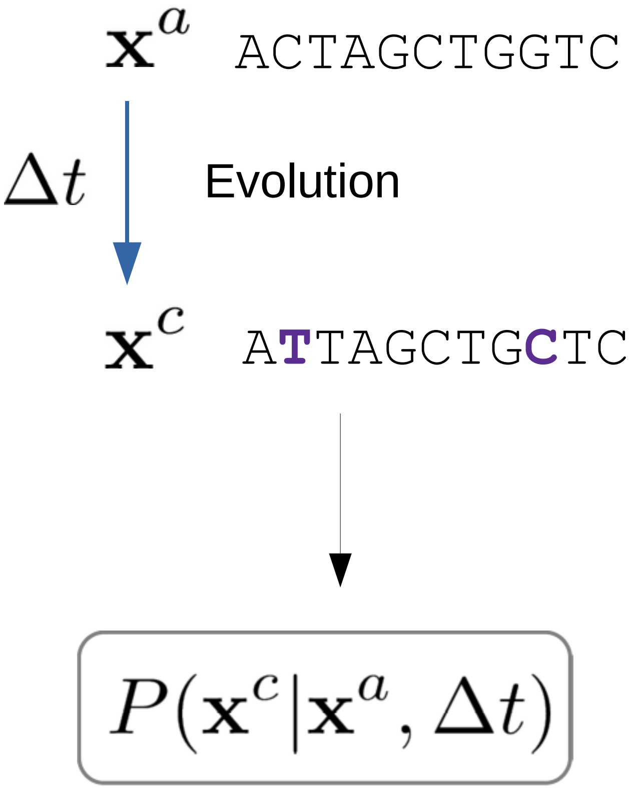
L **H** N L **R** G T D **D** R ..

E **H** R T **E** Q L E **K** G ..

Y **H** L L **R** T L D **D** T ..

R **H** A V **E** M L N **K** G ..

Sequence evolution model



$\mathbf{x}^a$ ACTAGCTGGTC

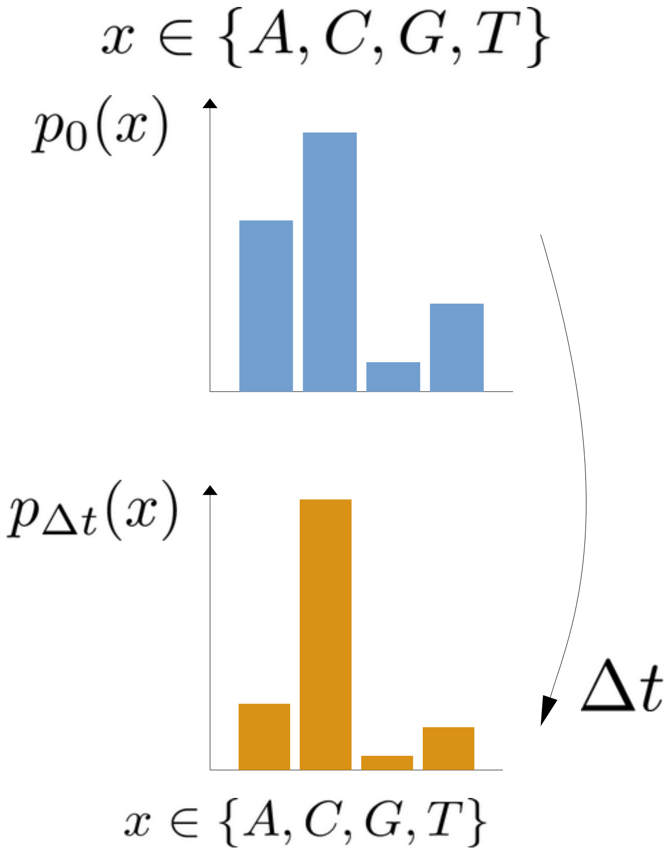
Δt Evolution

$\mathbf{x}^c$ A**T**TAGCTG**C**TC

$$P(\mathbf{x}^c|\mathbf{x}^a, \Delta t)$$

Sequence evolution models: state of the art

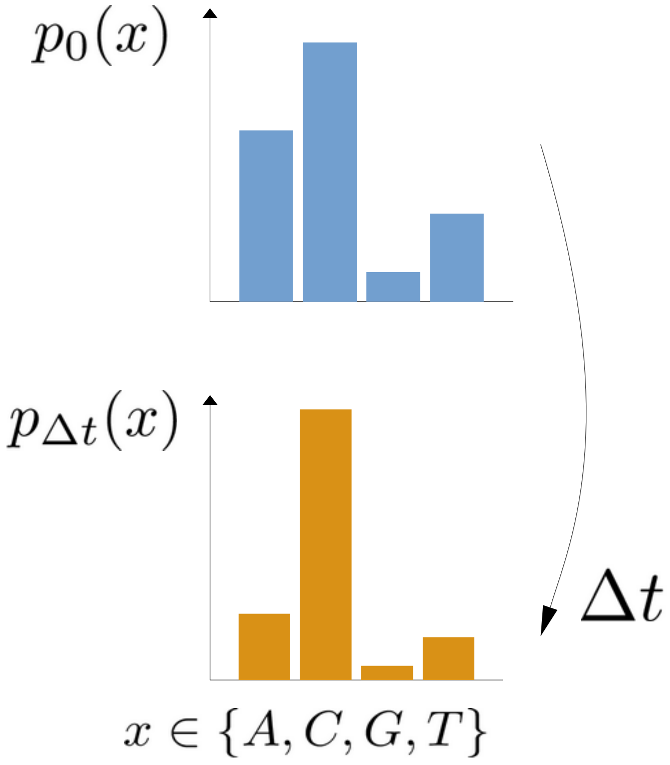
Focus on **one** sequence position

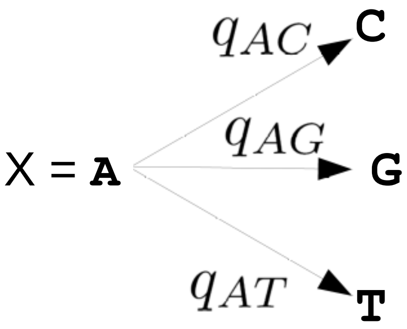


Sequence evolution models: state of the art

Focus on **one** sequence position

$$x \in \{A, C, G, T\}$$





Transition rate matrix

$$Q = \begin{pmatrix} -q_A & q_{CA} & q_{GA} & q_{TA} \\ q_{AC} & -q_C & q_{GC} & q_{TC} \\ q_{AG} & q_{CG} & -q_G & q_{TG} \\ q_{AT} & q_{CT} & q_{GT} & -q_T \end{pmatrix}$$

Continuous time Markov chain

$$\frac{\mathrm{d}p}{\mathrm{d}t} = p \cdot Q$$

$$p_{\Delta t} = p_0 e^{Q \Delta t}$$

- Nucleotides: 4x4 matrix
- Amino acids+gap: 21x21 matrix

$$q_x = \sum_{y \neq x} q_{xy}$$

Sequence evolution models

Transition rate matrix

$$Q = \begin{pmatrix} -q_A & q_{CA} & q_{GA} & q_{TA} \\ q_{AC} & -q_C & q_{GC} & q_{TC} \\ q_{AG} & q_{CG} & -q_G & q_{TG} \\ q_{AT} & q_{CT} & q_{GT} & -q_T \end{pmatrix}$$

Continuous time Markov chain

$$p_{\Delta t} = p_0 e^{Q \Delta t}$$

Each position independent

Transition probability between **sequences**

$\mathbf{x}$ ACTAGCTGGTC

$\Delta t \downarrow$

$\mathbf{y}$ A**T**TAGCTG**C**TC

$$P(\mathbf{y}|\mathbf{x}, \Delta t) = \prod_{i=1}^{L} \left(e^{\mu_i \Delta t Q}\right)_{x_i y_i}$$

Explicit **evolutionary rates** $\{\mu_1, \mu_2, \ldots\}$

Scaling of time $\langle \mu_i \rangle = 1$

$\Delta t = 1$ → ~ One substitution per site on average

Sequence evolution models

Transition rate matrix

$$Q = \begin{pmatrix} -q_A & q_{CA} & q_{GA} & q_{TA} \\ q_{AC} & -q_C & q_{GC} & q_{TC} \\ q_{AG} & q_{CG} & -q_G & q_{TG} \\ q_{AT} & q_{CT} & q_{GT} & -q_T \end{pmatrix}$$

Continuous time Markov chain

$$p_{\Delta t} = p_0 e^{Q\Delta t}$$

Each position independent

Transition probability between **sequences**

$$\begin{tikzcd} \mathbf{x}\ \texttt{ACTAGCTGGTC} \arrow[d, "\Delta t"'] \\ \mathbf{y}\ \texttt{A\textbf{T}TAGCTG\textbf{C}TC} \end{tikzcd}$$

$$P(\mathbf{y}|\mathbf{x}, \Delta t) = \prod_{i=1}^{L} \left(e^{\mu_i \Delta t Q}\right)_{x_i y_i}$$

Explicit **evolutionary rates** $\{\mu_1, \mu_2, \ldots\}$

Scaling of time $\langle \mu_i \rangle = 1$

$\Delta t = 1$ → ~ One substitution per site on average

Phylogenetic reconstruction

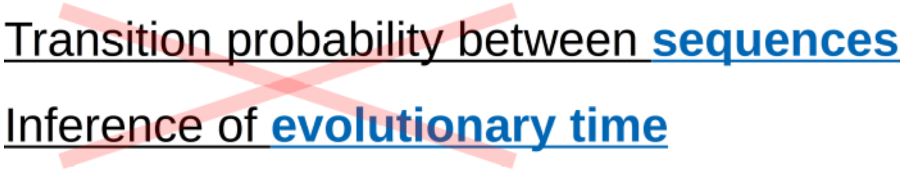
Likelihood of tree given sequences

Inference of **evolutionary time** $\frac{\mathrm{d}}{\mathrm{d}t} P(\mathbf{y}|\mathbf{x}, t) = \ldots$

Drawback: no epistasis, wrong fitness landscape → irrealistic

Evolution in complex landscape

Evolutionary model with **epistasis**?



~~Transition probability between **sequences**~~

~~Inference of **evolutionary time**~~

→ **no phylogenetic inference...**

… but **forward simulation** possible!

Is this useful?

- Study **effects of epistasis on evolution**
- **Robustness** of **state of the art models** to epistasis
- Use to **train machine learning models**
  ~ likelihood free phylogenetic inference

→

Goal

Design evolutionary model:
- realistic fitness landscape
- generative

Evolution in complex landscape

de la Paz *et. al.* 2020

di Bari *et. al.* 2024

Potts model

$$P(a_1,\ldots,a_N) = \frac{1}{Z}\exp\left(\sum_{i,j=1}^{L} J_{ij}(a_i,a_j) + \sum_{i=1}^{L} h_i(a_i)\right)$$

→ Good fitness landscape!

Discrete time **Markov chain** — Gibbs sampling

- Pick random position *i*
- Compute distribution at *i* conditioned on context → $P(a|\{a_j\}_{j\neq i})$
- Sample $a_i(t+1)$

Evolution in complex landscape

de la Paz *et. al.* 2020

di Bari *et. al.* 2024

Potts model

$$P(a_1, \ldots, a_N) = \frac{1}{Z} \exp\left(\sum_{i,j=1}^{L} J_{ij}(a_i, a_j) + \sum_{i=1}^{L} h_i(a_i) \right)$$

→ Good fitness landscape!

Discrete time **Markov chain** — Gibbs sampling

- Pick random position *i*
- Compute distribution at *i* conditioned on context → $P(a|\{a_j\}_{j \neq i})$
- Sample $a_i(t+1)$

Because of discrete time

~~Explicit evolutionary rates~~

~~Scaling of time~~

~~Compare with state of the art models~~

→

Goal

Design evolutionary model:

- realistic fitness landscape
- generative
- **continuous time**

Continuous time Potts model

Exponential model

Couplings or independent sites

$$P^{eq}(\mathbf{a}) \propto \exp\left(-E(\mathbf{a})\right)$$

Transition probability

$$P(\mathbf{b}|\mathbf{a}, t) = \left(e^{t\mathcal{Q}/\Omega}\right)_{\mathbf{ab}}$$

Detailed balance

$$P^{eq}(\mathbf{a})\mathcal{Q}_{\mathbf{ab}} = P^{eq}(\mathbf{b})\mathcal{Q}_{\mathbf{ba}}$$

→

Glauber dynamics

$$\mathcal{Q}_{\mathbf{ab}} = \left[1 + e^{E(\mathbf{b})-E(\mathbf{a})}\right]^{-1}$$

Continuous time Potts model

Exponential model

Couplings or independent sites

$$P^{eq}(\mathbf{a}) \propto \exp\left(-E(\mathbf{a})\right)$$

Transition probability

$$P(\mathbf{b}|\mathbf{a}, t) = \left(e^{t\mathcal{Q}/\Omega}\right)_{\mathbf{ab}}$$

Detailed balance

$$P^{eq}(\mathbf{a})\mathcal{Q}_{\mathbf{ab}} = P^{eq}(\mathbf{b})\mathcal{Q}_{\mathbf{ba}}$$

→

Glauber dynamics

$$\mathcal{Q}_{\mathbf{ab}} = \left[1 + e^{E(\mathbf{b})-E(\mathbf{a})}\right]^{-1}$$

Large matrix

$$\mathcal{Q} \rightarrow 21^L \times 21^L$$

→

Single mutations

$\mathcal{Q}_{\mathbf{ab}} = 0$ if **a** and **b** differ by more than one mut.

One row: *L x 21* non-zero rates

Continuous time Potts model: simulation?

→ Gillespie algorithm!

- Start from sequence **a**
- Compute **substitution rate**

$$R(\mathbf{a}) = \sum_{\mathbf{b}\neq\mathbf{a}} Q_{\mathbf{ab}}$$

- Sample **time to next substitution**

$$t \sim \mathrm{Exp}\left(R(\mathbf{a})\right)$$

Average waiting time $\langle t \rangle = R(\mathbf{a})^{-1}$

- Sample **next substitution**

For **b** one mutation away from **a**

$$p_{\mathbf{b}} = \frac{Q_{\mathbf{ab}}}{R(\mathbf{a})}$$

- **No rejection**
~ Gibbs sampling
- Complexity

$$\mathcal{O}(L \times 21)$$

Continuous time Potts model: scaling of time

$$R(\mathbf{a}) = \sum_{\mathbf{b} \neq \mathbf{a}} Q_{\mathbf{ab}}$$

Substitution rate when in state **a**

Convention in evolutionary model: $\langle R \rangle = L$ $\Delta t = 1$ ➞ One substitution per site

→ interpretation of **branch lengths**!

Scaling factor: $\Omega = \frac{1}{L} \left\langle \sum_{\mathbf{b} \neq \mathbf{a}} Q_{\mathbf{ab}} \right\rangle_{P^{eq}(\mathbf{a})}$

Compute **once** per model

What is the effect of epistasis on evolution?

Potts model

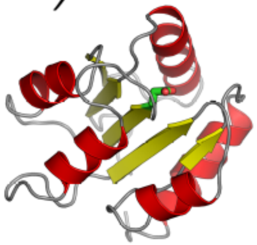
$$P^{eq}(\mathbf{a}) \propto \exp\left(\sum_{i<j} \overset{\text{Couplings}}{J_{ij}(a_i, a_j)} + \sum_{i} \overset{\text{Fields}}{h_i(a_i)}\right)$$

Profile model

$$P^{eq}(\mathbf{a}) \propto \exp\left(\sum_{i} h'_i(a_i)\right)$$

→ **Same single-site distribution**

Protein family
PF00072: Response regulator domain



What is the effect of epistasis on evolution?

Potts model

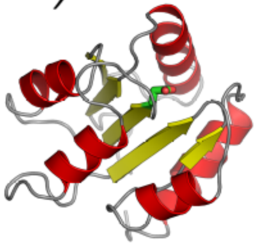
$$P^{eq}(\mathbf{a}) \propto \exp\left(\sum_{i<j} \overset{\text{Couplings}}{J_{ij}(a_i, a_j)} + \sum_i \overset{\text{Fields}}{h_i(a_i)}\right)$$

Profile model

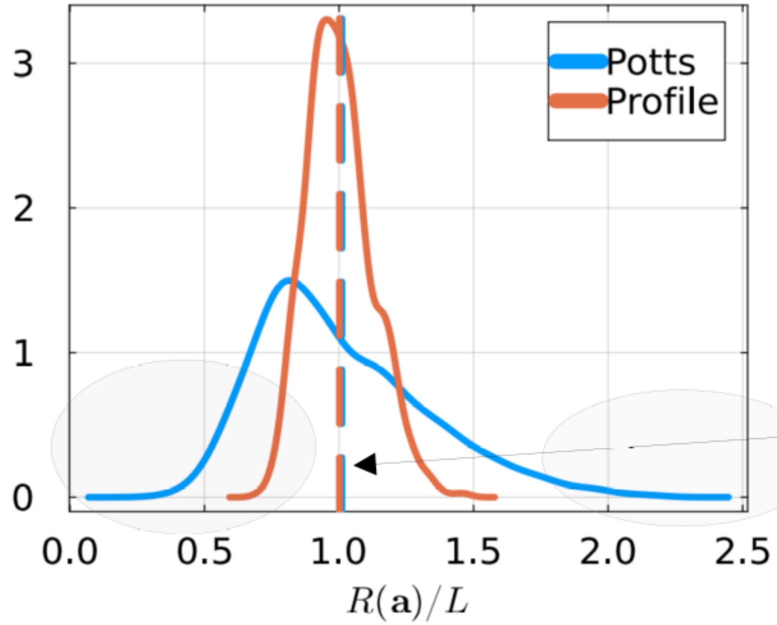
$$P^{eq}(\mathbf{a}) \propto \exp\left(\sum_i h'_i(a_i)\right)$$

→ **Same single-site distribution**

Protein family
PF00072: Response regulator domain



Natural sequences



Rates of evolution at sequence level

$$R(\mathbf{a}) = \sum_{\mathbf{b} \neq \mathbf{a}} Q_{\mathbf{ab}}$$

Sequences with high/low rates

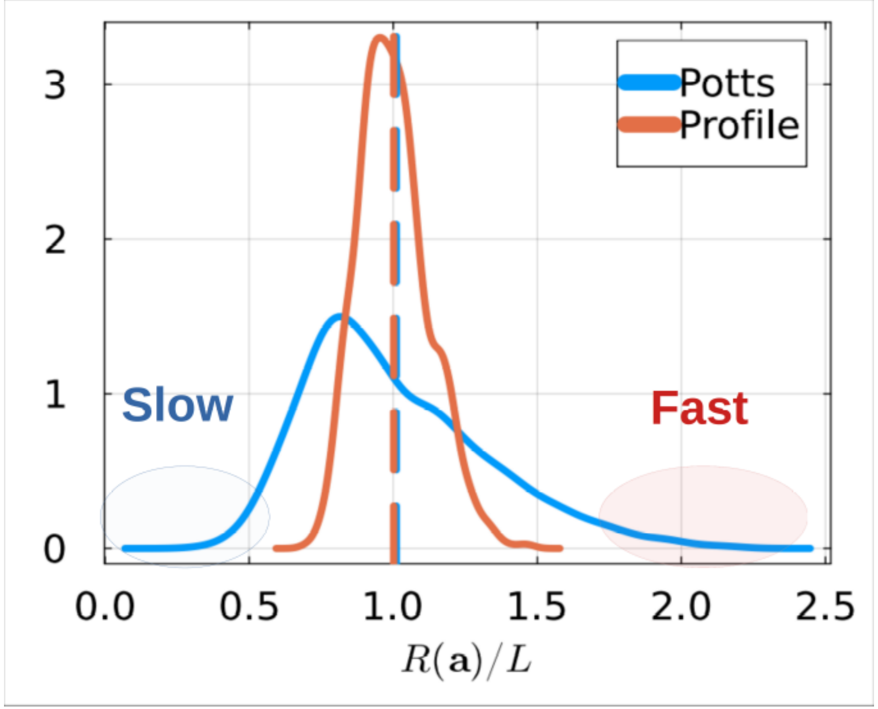
What is the effect of epistasis on evolution?

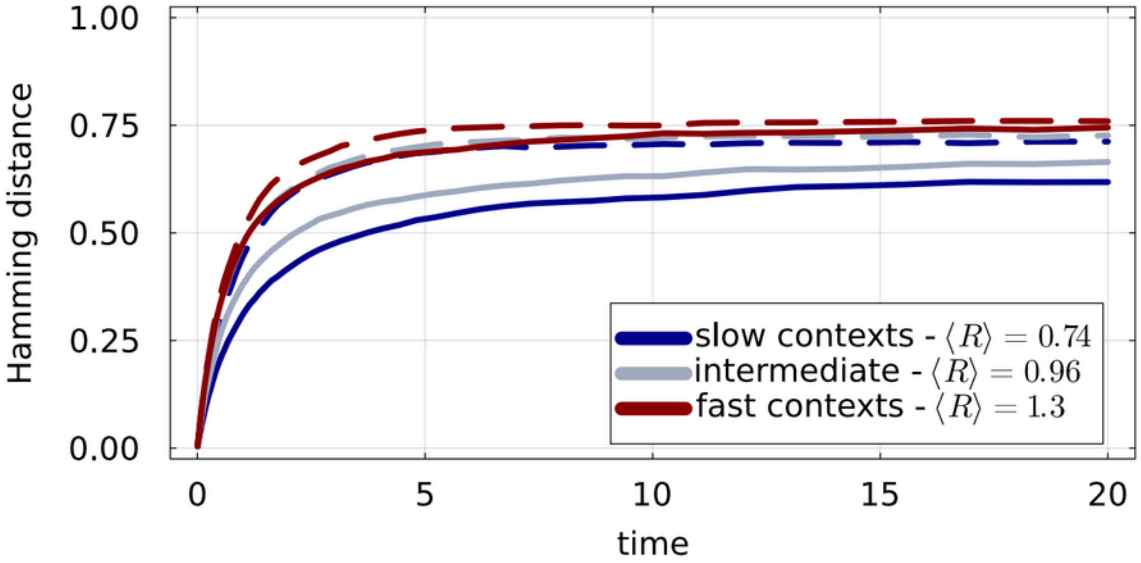
Potts model

$$P^{eq}(\mathbf{a}) \propto \exp\left(\sum_{i<j} \overset{\text{Couplings}}{J_{ij}(a_i,a_j)} + \sum_i \overset{\text{Fields}}{h_i(a_i)}\right)$$

Profile model

$$P^{eq}(\mathbf{a}) \propto \exp\left(\sum_i h'_i(a_i)\right)$$





→ **Slowing down of evolution**

Site specific rates

Average rate of substitution at site *i*

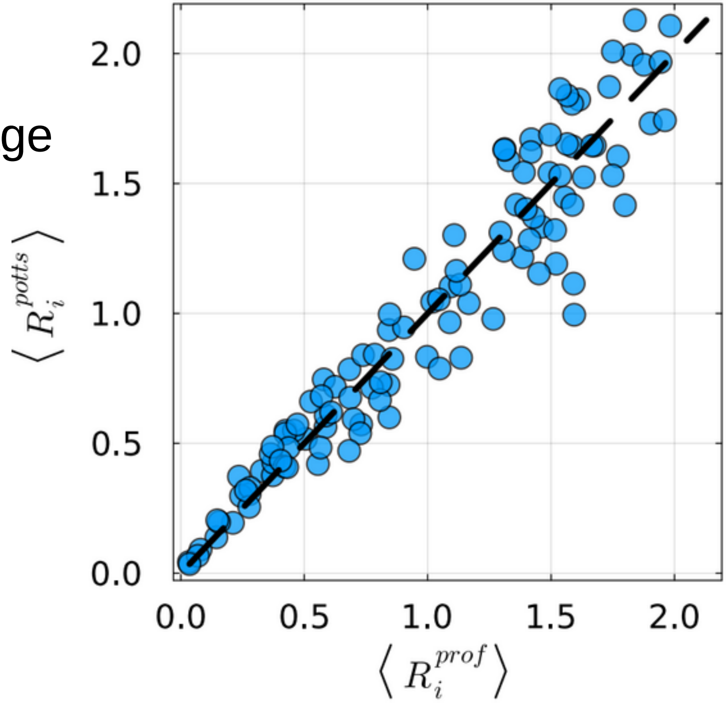
$$R_i = \left\langle \sum_{b \in \mathcal{N}_i(\mathbf{a})} Q_{\mathbf{ab}} \right\rangle_{\mathbf{a}}$$

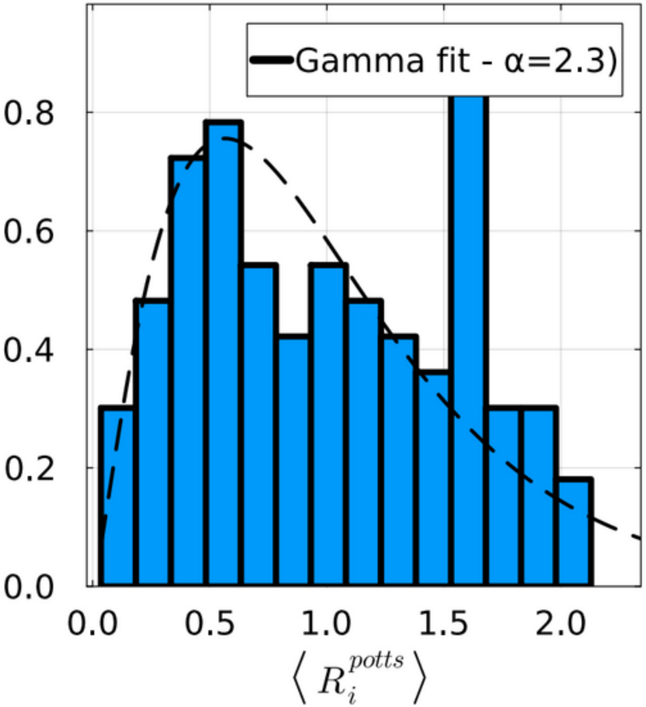
$\mathcal{N}_i(\mathbf{a})$ ~ sequences that differ from ***a*** at site ***i***

Common assumption in evolutionary models
Site-specific rates are **Gamma distributed**

Yang 1994
Rzhetsky & Nei 1993

No influence of epistasis on average site rates





The importance of context

Average substitution rate at *i*

$$R_i = \left\langle \sum_{b \in \mathcal{N}_i(\mathbf{a})} \mathcal{Q}_{\mathbf{ab}} \right\rangle_{\mathbf{a}}$$

Context-specific substitution rate at *i*

Context $\mathbf{a}_{\setminus i} = (a_1 \dots a_{i-1} \star a_{i+1} \dots a_L)$

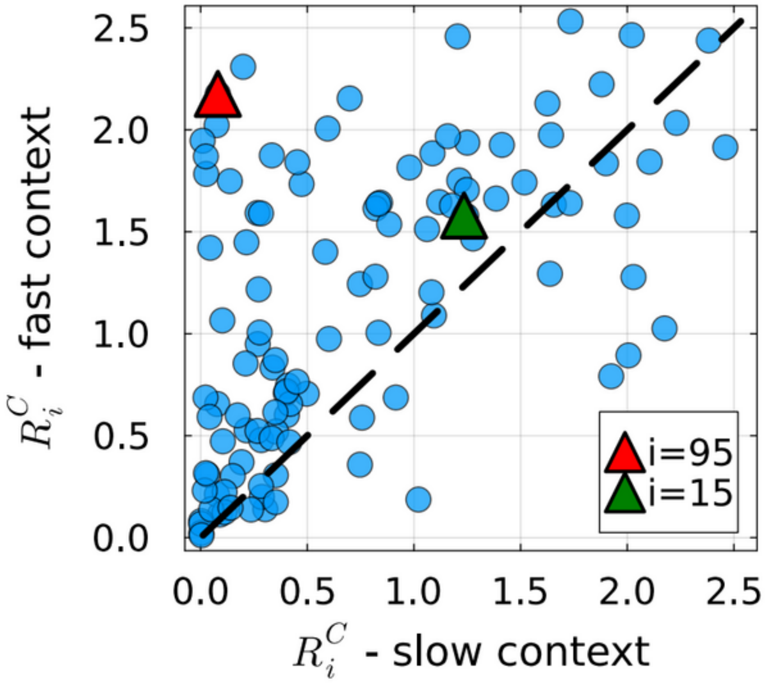
$$R_i^C(\mathbf{a}_{\setminus i}) = \sum_a P_i(a|a_{\setminus i}) \sum_{\mathbf{b} \in \mathcal{N}_i(\mathbf{a}_{\setminus i})} \mathcal{Q}_{\mathbf{ab}}$$

Average waiting time to next substitution at *i* in **context**

The importance of context

Average substitution rate at *i*

$$R_i = \left\langle \sum_{b\in\mathcal{N}_i(\mathbf{a})} \mathcal{Q}_{\mathbf{ab}} \right\rangle_{\mathbf{a}}$$

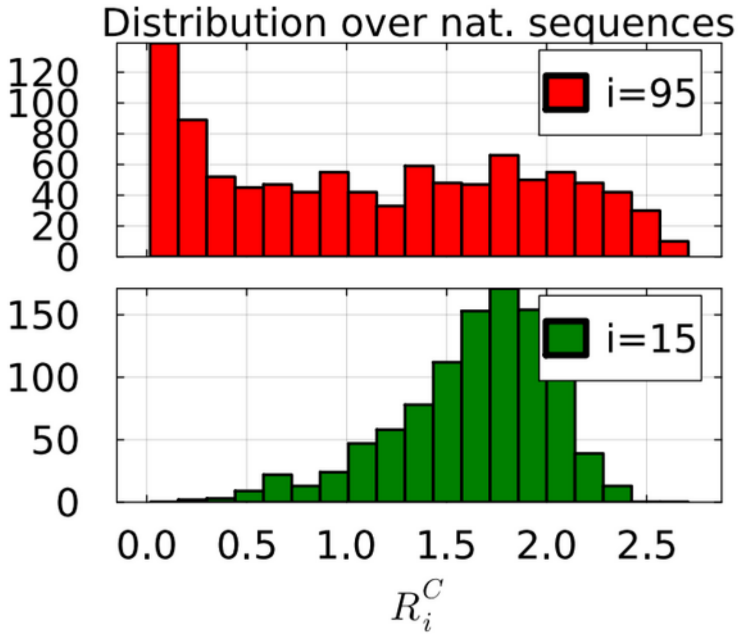


Context-specific substitution rate at *i*

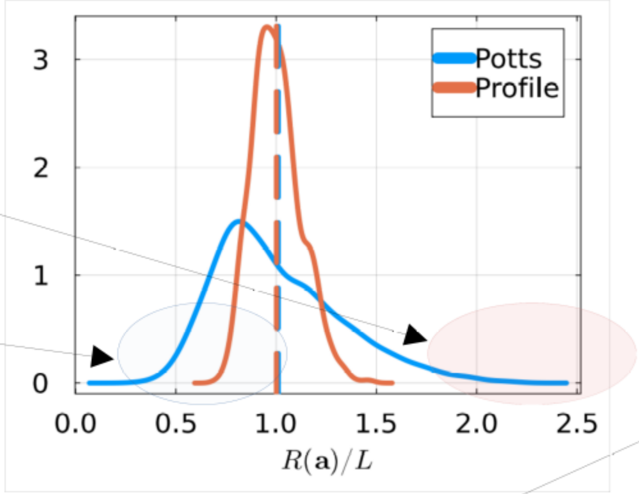
Context $\mathbf{a}_{\setminus i} = (a_1 \dots a_{i-1} \star a_{i+1} \dots a_L)$

$$R_i^C(\mathbf{a}_{\setminus i}) = \sum_a P_i(a|a_{\setminus i}) \sum_{\mathbf{b}\in\mathcal{N}_i(\mathbf{a}_{\setminus i})} \mathcal{Q}_{\mathbf{ab}}$$

Average waiting time to next substitution at *i* in **context**



The importance of context

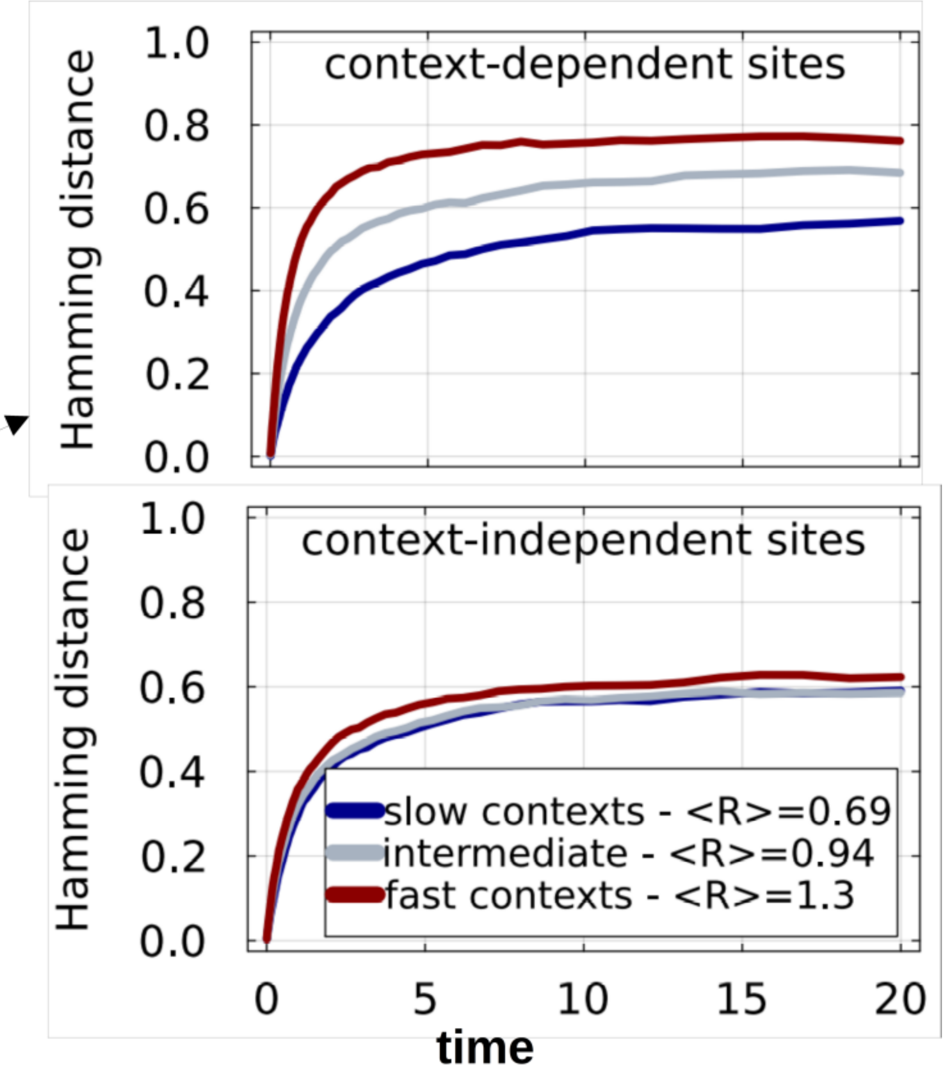


Fast

Slow

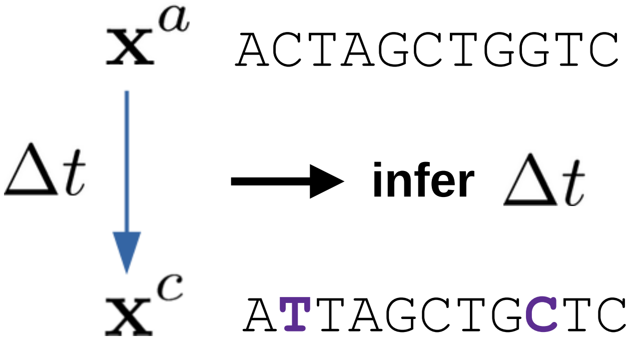
Strong epistasis

~ independent evolution



Reconstructing evolutionary time

Inferring evolutionary distance



$\mathbf{x}^a$ ACTAGCTGGTC

Δt → **infer** Δt

$\mathbf{x}^c$ A**T**TAGCTG**C**TC

Simulate data:

- pick ancestral sequence
- pick Δt
- simulate with **Potts** or **profile**

→ $\{\mathbf{x}^a, \mathbf{x}^c, \Delta t\}$

Infer time with **profile**

$$P(\mathbf{y}|\mathbf{x}, \Delta t) = \prod_{i=1}^{L} \left(e^{\mu_i \Delta t Q}\right)_{x_i y_i}$$

→ Most likely value of Δt

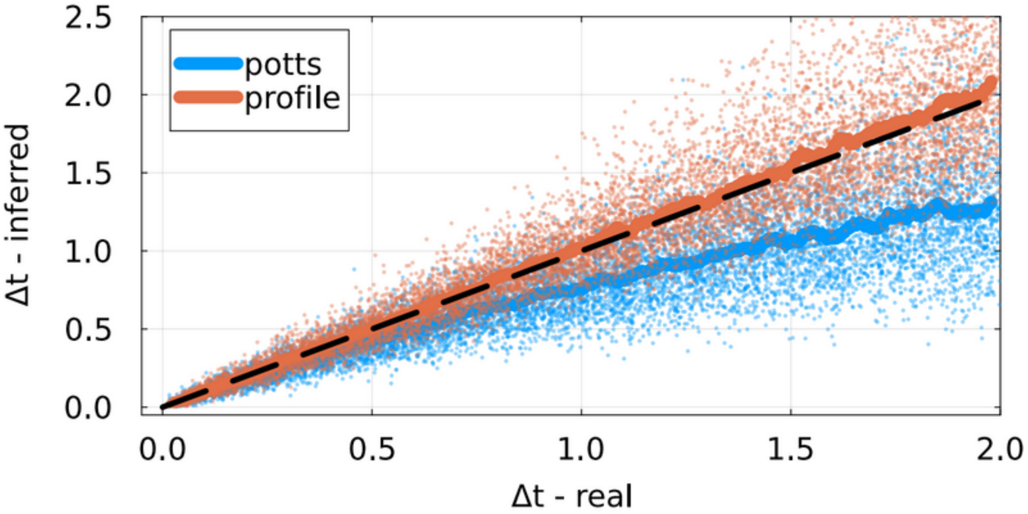
Reconstructing evolutionary time

Simulate data $\{\mathbf{x}^a, \mathbf{x}^c, \Delta t\}$

Infer time with **profile**

$$P(\mathbf{y}|\mathbf{x}, \Delta t) = \prod_{i=1}^{L} \left(e^{\mu_i \Delta t Q}\right)_{x_i y_i}$$

Underestimation of time



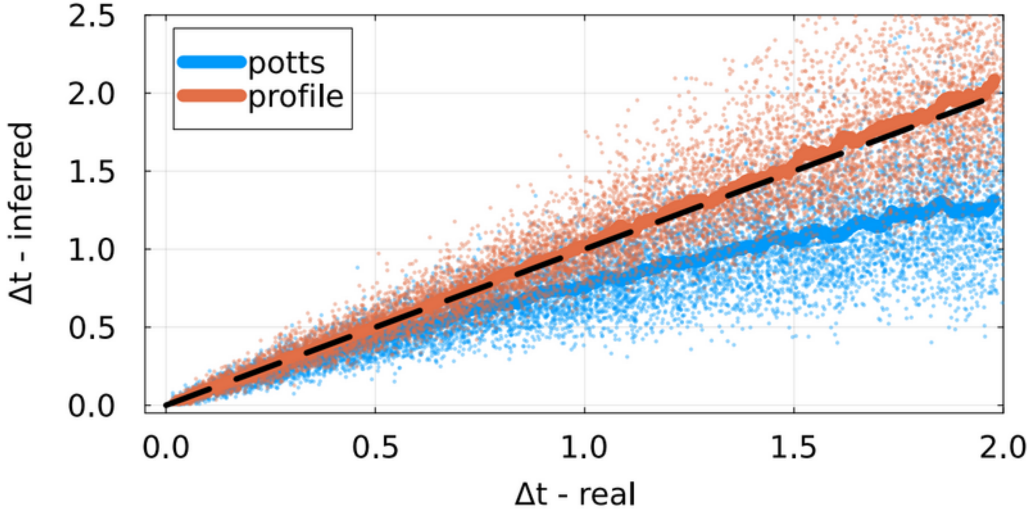
Reconstructing evolutionary time

Simulate data $\{\mathbf{x}^a, \mathbf{x}^c, \Delta t\}$

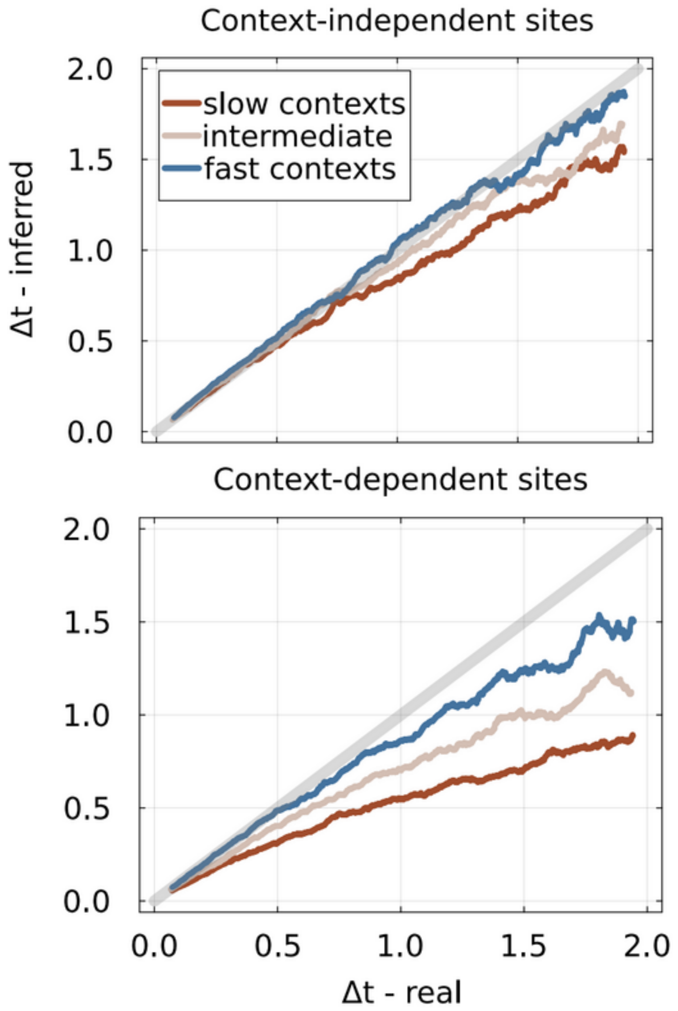
Infer time with **profile**

$$P(\mathbf{y}|\mathbf{x}, \Delta t) = \prod_{i=1}^{L} \left(e^{\mu_i \Delta t Q}\right)_{x_i y_i}$$

Underestimation of time



Caused by context dependent sites



Summary

Evolution with a continuous time generative model

- Informed by fitness landscape
- Comparable to state of the art methods (timescale)
- Direct access to evolutionary rates

Influence of epistasis on evolution

- Slows down evolution
- No effect on **average rates**
- Strong dependence on **context**

Influence of epistasis phylogenetic inference

Systematic underestimation of time!

Authors
Andrea Pagnani (PoliTo)
P.B.C

Thank you for listening