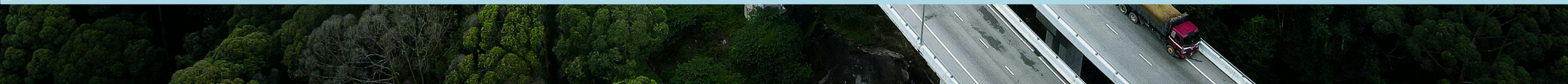
Highways of Horizontal Gene Transfer for Modeling Large Gene Flow Events Under the UndatedDTL Model

Noah A. Wahl, Benoit Morel, Oliver Schick, Alexandros Stamatakis,
Tom A. Williams, Gergely J. Szöllősi (TBF)

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

FORTH
INSTITUTE OF COMPUTER SCIENCE



Gene Tree Reconciliation and Highways

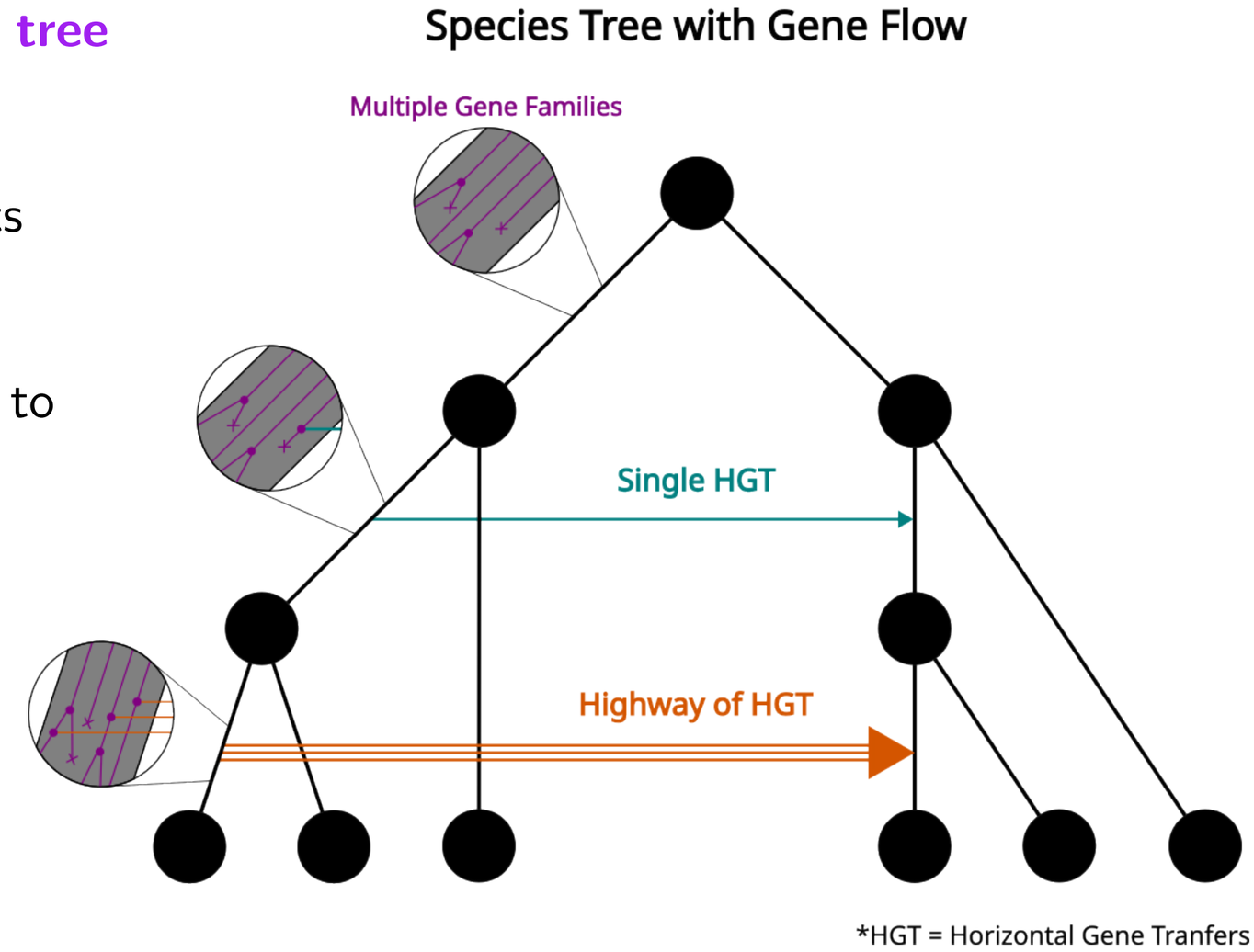
- Species and gene tree topologies often disagree
- Given a species tree and a collection of **gene tree samples** of gene families
- We try to infer a series of gene origination, **duplication, loss and transfer** (DTL) events under maximum likelihood
- The **UndatedDTL** (multi-)model uses branch-wise DTL and origination parameters to sample these events

FORTH INSTITUTE OF COMPUTER SCIENCE
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ UNIVERSITY OF CRETE

Gene Tree Reconciliation and Highways

- Species and gene tree topologies often disagree
- Given a species tree and a collection of **gene tree samples** of gene families
- We try to infer a series of gene origination, **duplication, loss and transfer** (DTL) events under maximum likelihood
- The **UndatedDTL** (multi-)model uses branch-wise DTL and origination parameters to sample these events

- Highways of HGT are pairs of species tree branches that have an **elevated** rate of gene exchange [1]
- Often an indicator of symbiont genes being transferred to the host genome



[1] Beiko, Robert G., Timothy J. Harlow, and Mark A. Ragan. "Highways of gene sharing in prokaryotes." Proceedings of the National Academy of Sciences 102.40 (2005)

Motivation

- **Horizontal gene transfer** (HGT) makes a lot of things more challenging
- The transfer rate needs to consider **all possible transfer targets**
- Global or family-wise DTL rates can be inferred even for large trees
- Branch-wise rates for small trees
- Ideal case: **donor** and **recipient** rates for all species tree branches
- Instead: can we just identify **highways** to deal with the most extreme outliers?

FORTH INSTITUTE OF COMPUTER SCIENCE
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ UNIVERSITY OF CRETE

Motivation

- **Horizontal gene transfer** (HGT) makes a lot of things more challenging
- The transfer rate needs to consider **all possible transfer targets**
- Global or family-wise DTL rates can be inferred even for large trees
- Branch-wise rates for small trees
- Ideal case: **donor** and **recipient** rates for all species tree branches
- Instead: can we just identify **highways** to deal with the most extreme outliers?

- Extract that pair as a separate parameter $H_{e,d}$

Extending the UndatedDTL Model in AleRax [2] with Highways

Collection of gene families with a distribution of gene trees per family
1 species tree
D, T, L, **H** branch-wise parameters

↓

Gene tree distributions get turned into clade splits and conditional clade probabilities (CCPs) [3,4]

↓

Double recursion over species tree branches (e) and clade splits ($\gamma', \gamma'' \mid \gamma$) in the gene trees to compute conditional likelihoods $\Pi(e, \gamma)$

[2] Morel, Benoit, et al. "AleRax: a tool for gene and species tree co-estimation and reconciliation under a probabilistic model of gene duplication, transfer, and loss." Bioinformatics 40.4 (2024)
[3] Höhna, Sebastian, and Alexei J. Drummond. "Guided tree topology proposals for Bayesian phylogenetic inference." Systematic biology 61.1 (2012)
[4] Larget, Bret. "The estimation of tree posterior probabilities using conditional clade probability distributions." Systematic biology 62.4 (2013)

FORTH INSTITUTE OF COMPUTER SCIENCE
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ UNIVERSITY OF CRETE

Extending the UndatedDTL Model in AleRax [2] with Highways

Collection of gene families with a distribution of gene trees per family
1 species tree
D, T, L, **H** branch-wise parameters

↓

Gene tree distributions get turned into clade splits and conditional clade probabilities (CCPs) [3,4]

↓

Double recursion over species tree branches (e) and clade splits ($\gamma', \gamma'' \mid \gamma$) in the gene trees to compute conditional likelihoods $\Pi(e, \gamma)$

UndatedDTL:

$$\begin{aligned}\Pi(e,\gamma) = S &+ SL \\ &+ D + DL \\ &+ T + TL\end{aligned}$$

→

UndatedDTL + **H**:

$$\begin{aligned}\Pi(e,\gamma) = S &+ SL \\ &+ D + DL \\ &+ T + TL \\ &+ H + HL\end{aligned}$$

[2] Morel, Benoit, et al. "AleRax: a tool for gene and species tree co-estimation and reconciliation under a probabilistic model of gene duplication, transfer, and loss." Bioinformatics 40.4 (2024)
[3] Höhna, Sebastian, and Alexei J. Drummond. "Guided tree topology proposals for Bayesian phylogenetic inference." Systematic biology 61.1 (2012)
[4] Larget, Bret. "The estimation of tree posterior probabilities using conditional clade probability distributions." Systematic biology 62.4 (2013)

FORTH INSTITUTE OF COMPUTER SCIENCE
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ UNIVERSITY OF CRETE

Extending the UndatedDTL Model in AleRax [2] with Highways

Collection of gene families with a distribution of gene trees per family
1 species tree
D, T, L, **H** branch-wise parameters

↓

Gene tree distributions get turned into clade splits and conditional clade probabilities (CCPs) [3,4]

↓

Double recursion over species tree branches (e) and clade splits ($\gamma', \gamma'' \mid \gamma$) in the gene trees to compute conditional likelihoods $\Pi(e, \gamma)$

UndatedDTL:

$$\begin{aligned}\Pi(e,\gamma) = S + SL \\ + D + DL \\ + T + TL\end{aligned}$$

→

Easy, right?

UndatedDTL + **H**:

$$\begin{aligned}\Pi(e,\gamma) = S + SL \\ + D + DL \\ + T + TL \\ + H + HL\end{aligned}$$

[2] Morel, Benoit, et al. "AleRax: a tool for gene and species tree co-estimation and reconciliation under a probabilistic model of gene duplication, transfer, and loss." Bioinformatics 40.4 (2024)
[3] Höhna, Sebastian, and Alexei J. Drummond. "Guided tree topology proposals for Bayesian phylogenetic inference." Systematic biology 61.1 (2012)
[4] Larget, Bret. "The estimation of tree posterior probabilities using conditional clade probability distributions." Systematic biology 62.4 (2013)

FORTH INSTITUTE OF COMPUTER SCIENCE
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ UNIVERSITY OF CRETE

Overfitting

- The UndatedDTL model can already have many parameters (branch-wise, per-family)
- But those are **user-regulated**
- Each higway introduces 1 additional parameter
- Highway parameters are chosen **during optimization**
- Too many (unnecessary) parameters slows down the optimizer significantly
- Need good heuristics with penalties for adding too many parameters and avoid **overfitting**

Overfitting

- The UndatedDTL model can already have many parameters (branch-wise, per-family)
- But those are **user-regulated**
- Each higway introduces 1 additional parameter
- Highway parameters are chosen **during optimization**
- Too many (unnecessary) parameters slows down the optimizer significantly
- Need good heuristics with penalties for adding too many parameters and avoid **overfitting**

- Easy case: user provides highway candidates to be **validated**
- Have at most this many highway parameters

Overfitting

- The UndatedDTL model can already have many parameters (branch-wise, per-family)
- But those are **user-regulated**
- Each higway introduces 1 additional parameter
- Highway parameters are chosen **during optimization**
- Too many (unnecessary) parameters slows down the optimizer significantly
- Need good heuristics with penalties for adding too many parameters and avoid **overfitting**

- Easy case: user provides highway candidates to be **validated**
- Have at most this many highway parameters

This requires a lot of domain knowledge and good hypotheses!

FORTH INSTITUTE OF COMPUTER SCIENCE
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ UNIVERSITY OF CRETE

Highway Inference Heuristics

A good highway inference heuristic needs to achive 2 things:

1. Find candidate node pairs with an exceptional number of transfers between them
2. Limit and test candidates quickly and discard ones with insufficient likelihood improvements

Highway Inference Heuristics

A good highway inference heuristic needs to achive 2 things:

1. Find candidate node pairs with an exceptional number of transfers between them
2. Limit and test candidates quickly and discard ones with insufficient likelihood improvements

Optimize DTL rates

↓

Get candidates with consistent transfers from 100 reconciliations

FORTH INSTITUTE OF COMPUTER SCIENCE
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ UNIVERSITY OF CRETE

Highway Inference Heuristics

A good highway inference heuristic needs to achive 2 things:

1. Find candidate node pairs with an exceptional number of transfers between them
2. Limit and test candidates quickly and discard ones with insufficient likelihood improvements

Optimize DTL rates

↓

Get candidates with consistent transfers from 100 reconciliations

↓

Discard by potential highway rate and donor-recipient distance

↓

Test for positive slope at a minimal highway rate

↓

Optimize highway rate and discard via BIC

FORTH INSTITUTE OF COMPUTER SCIENCE
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ UNIVERSITY OF CRETE

Highway Inference Heuristics

A good highway inference heuristic needs to achive 2 things:

1. Find candidate node pairs with an exceptional number of transfers between them
2. Limit and test candidates quickly and discard ones with insufficient likelihood improvements

Optimize DTL rates

↓

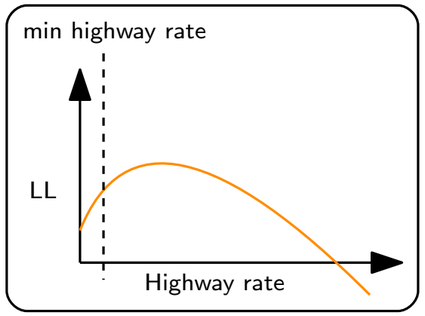
Get candidates with consistent transfers from 100 reconciliations

↓

Discard by potential highway rate and donor-recipient distance

↓

Test for positive slope at a minimal highway rate →



↓

Optimize highway rate and discard via BIC

FORTH INSTITUTE OF COMPUTER SCIENCE
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ UNIVERSITY OF CRETE

Simulation Pipeline

Species Tree	#
Cyanobacteria	36
Bacteria-28	38
Psuedomonadota-56	31

×

D	T	L
0.01	0.08	0.085
0.02	0.16	0.17
0.03	0.24	0.25

×

Highway Set	#
Set 1	2
Set 2	3
Set 3	5

×

Strength
0.1 x T
1.0 x T
10.0 x T

Parameters:
D, T, L, Highway rates and Species Trees

Simulation of Gene Families under UndatedDTL + Highways

Single Simulated Gene Family

bppseqgen

Model	LG08+G4+I
α	0.1
Inv	0.1
#Sites	300

Sequences

PhyloBayes

Model	CAT+P+G1
#Samples	100
Sample Freq	10
Burnin	10

Gene Tree Samples

Distribution of topologies through sequence simulation (bppseqgen) and tree inference (PhyloBayes)

Distribution of topologies through NNI moves

Γ-Distribution of random #NNI moves

α	θ
2.0	0.5
4.0	0.5
5.0	0.75

Verification of Candidates with False Positives

Reconciliation with AleRax to re-infer parameters

Inference Heuristic Based on Transfers

FORTH INSTITUTE OF COMPUTER SCIENCE
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ UNIVERSITY OF CRETE

Simulation Pipeline

focus on these results

Species Tree	#
Cyanobacteria	36
Bacteria-28	38
Psuedomonadota-56	31

×

D	T	L
0.01	0.08	0.085
0.02	0.16	0.17
0.03	0.24	0.25

×

Highway Set	#
Set 1	2
Set 2	3
Set 3	5

×

Strength
0.1 x T
1.0 x T
10.0 x T

Parameters:
D, T, L, Highway rates and Species Trees

Simulation of Gene Families under UndatedDTL + Highways

Single Simulated Gene Family

bppseqgen

Model	LG08+G4+I
α	0.1
Inv	0.1
#Sites	300

Sequences

PhyloBayes

Model	CAT+P+G1
#Samples	100
Sample Freq	10
Burnin	10

Gene Tree Samples

Distribution of topologies through sequence simulation (bppseqgen) and tree inference (PhyloBayes)

Distribution of topologies through NNI moves

Γ-Distribution of random #NNI moves

α	θ
2.0	0.5
4.0	0.5
5.0	0.75

Reconciliation with AleRax to re-infer parameters

Verification of Candidates with False Positives

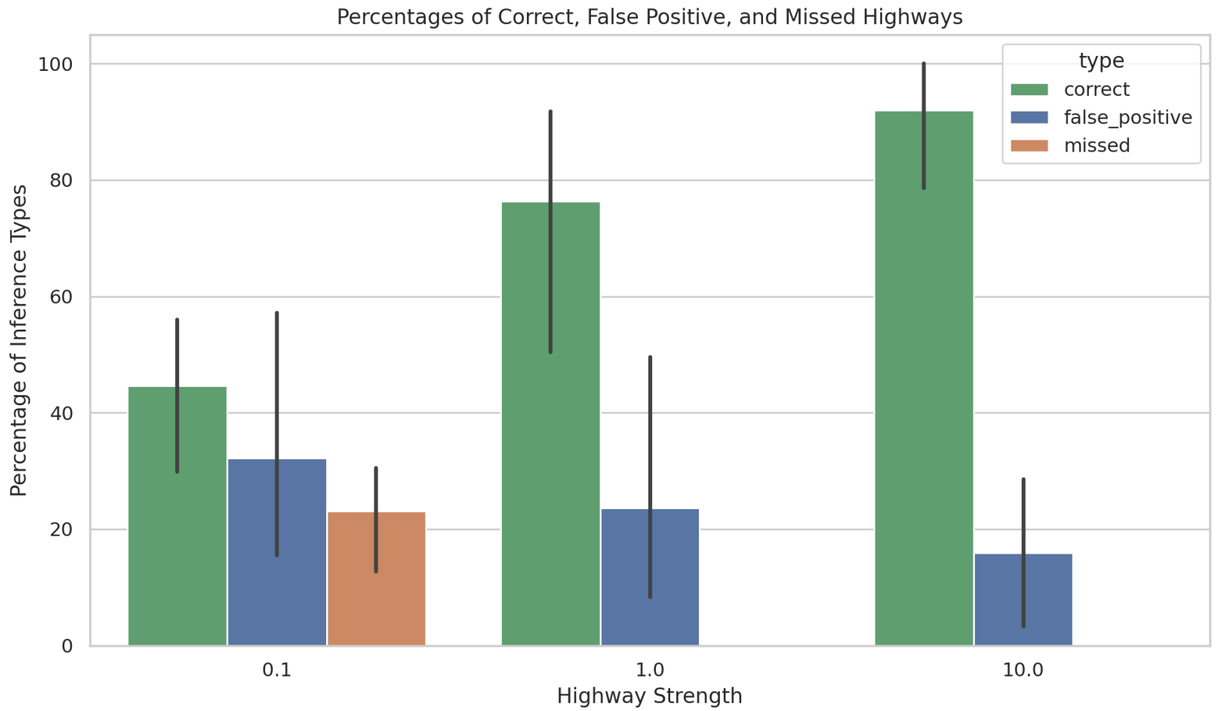
Inference Heuristic Based on Transfers

FORTH INSTITUTE OF COMPUTER SCIENCE
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ UNIVERSITY OF CRETE

Results (Transfer Heuristic)

BIC: $\alpha \times \Delta LL > log(N)$

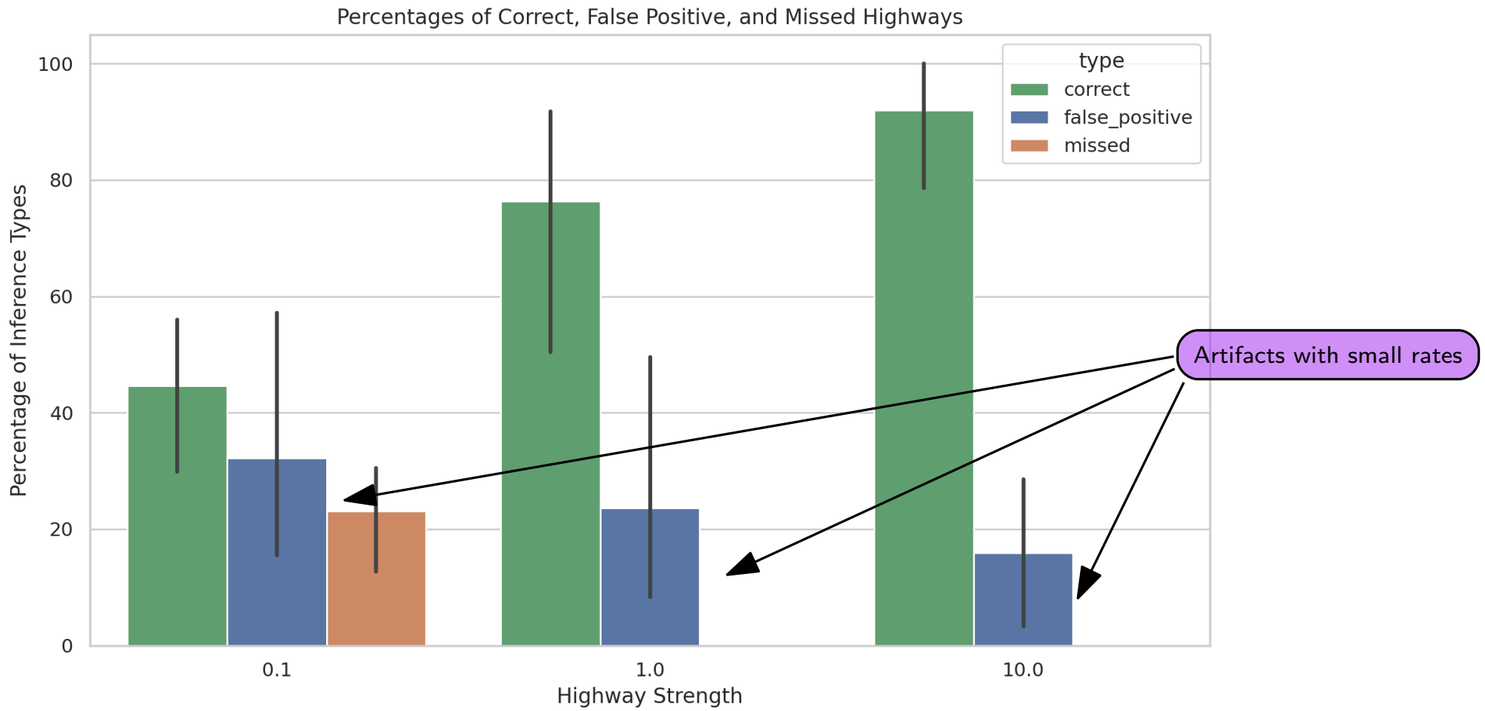
Weak pruning (distance, LL improvement, BIC $\alpha = 2$)



Results (Transfer Heuristic)

BIC: $\alpha \times \Delta LL > log(N)$

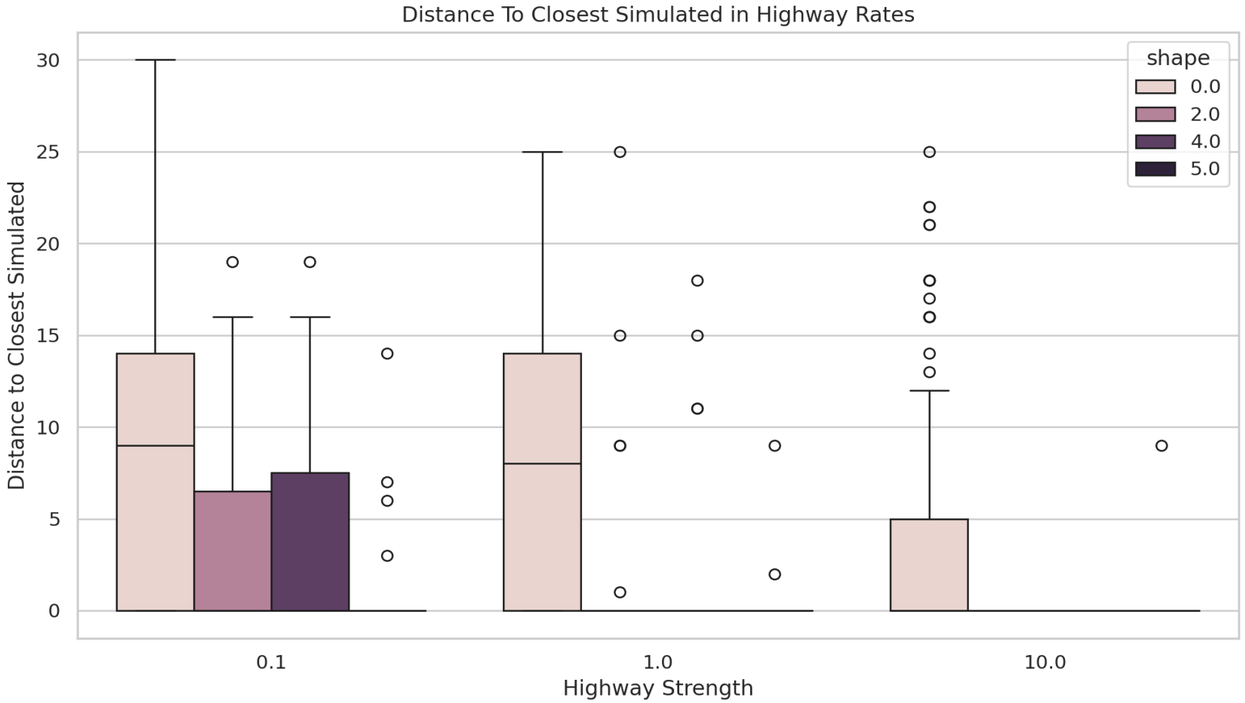
Weak pruning (distance, LL improvement, BIC $\alpha = 2$)



Results (Transfer Heuristic)

BIC: $\alpha \times \Delta LL > log(N)$

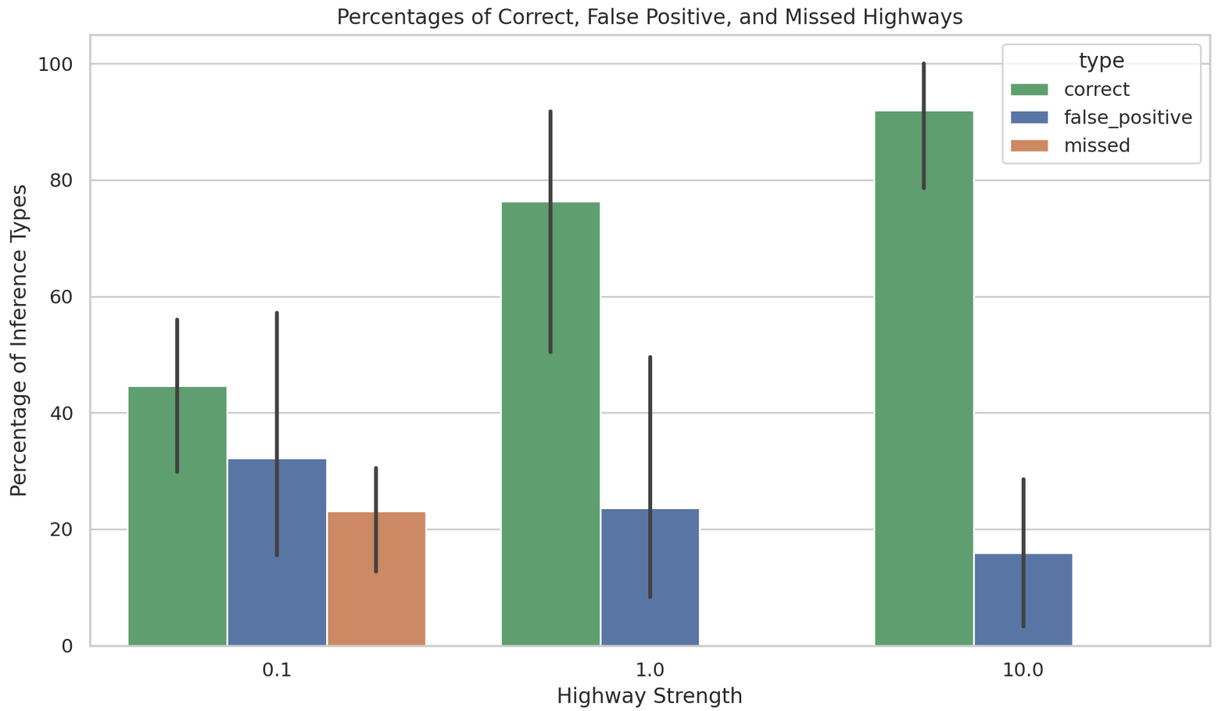
Weak pruning (distance, LL improvement, BIC $\alpha = 2$)



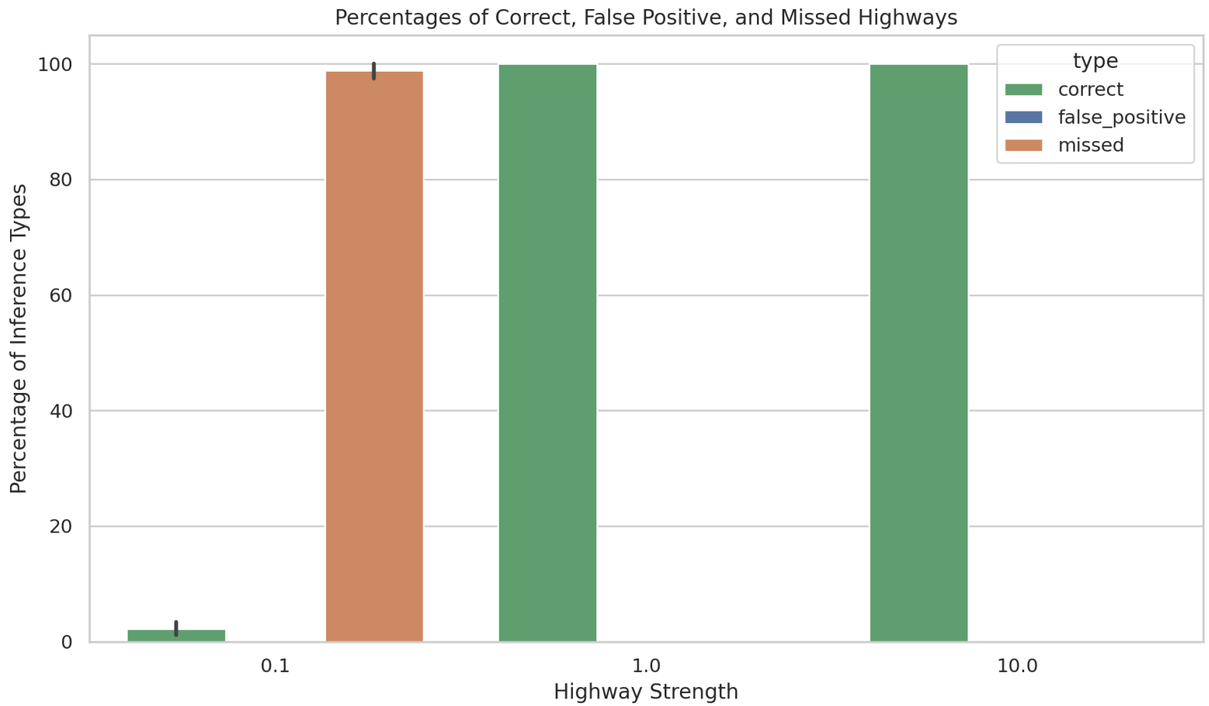
Results (Transfer Heuristic)

BIC: $\alpha \times \Delta LL > log(N)$

Weak pruning (distance, LL improvement, BIC $\alpha = 2$)



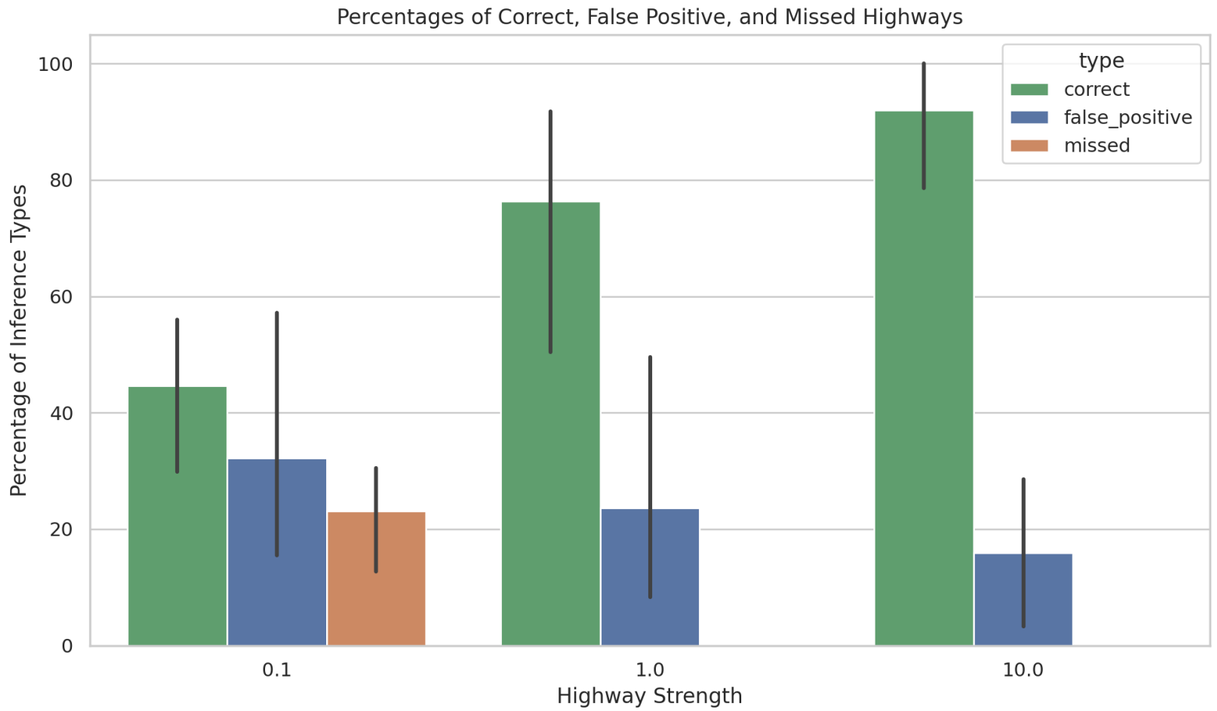
Strong pruning (distance, **potential highway rate**, LL improvement, BIC $\alpha = 1$)



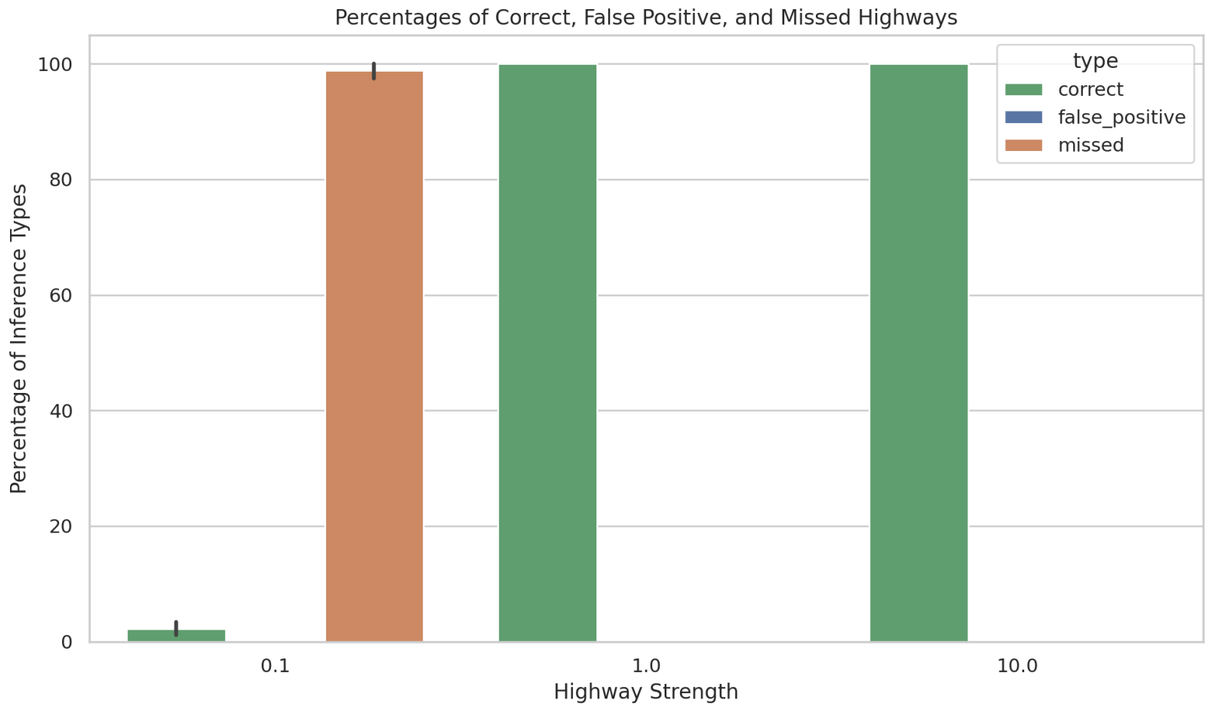
Results (Transfer Heuristic)

BIC: $\alpha \times \Delta LL > log(N)$

Weak pruning (distance, LL improvement, BIC $\alpha = 2$)



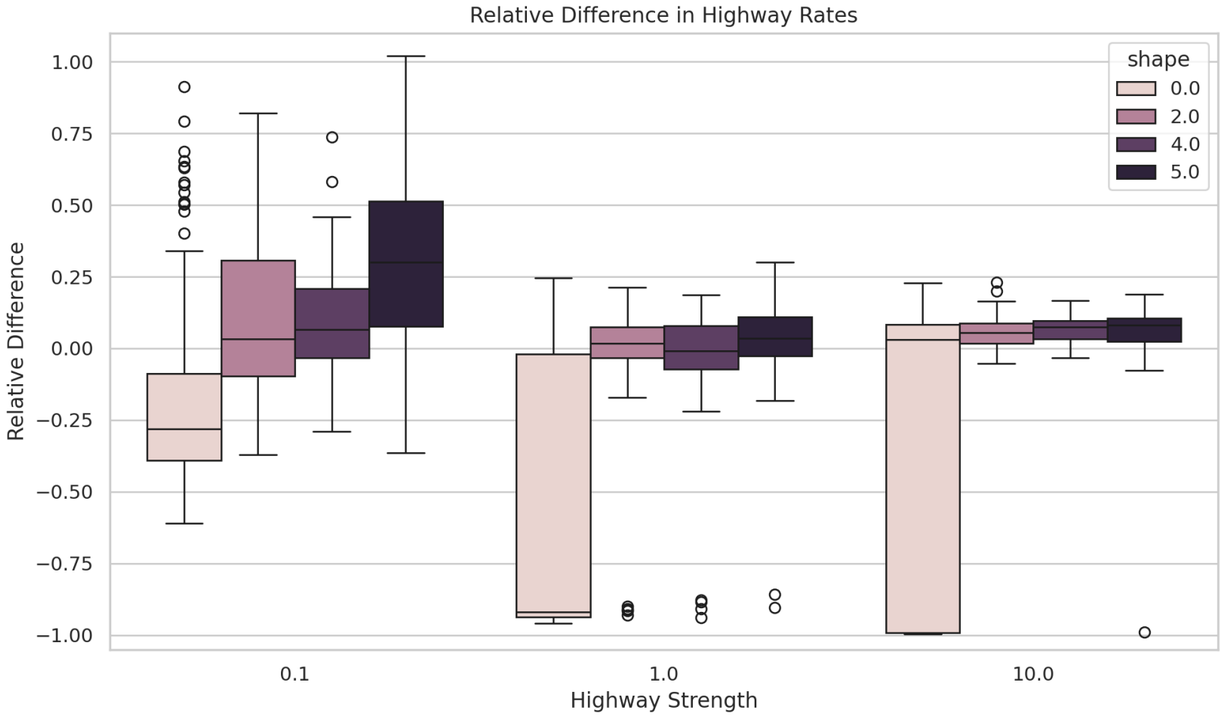
Strong pruning (distance, **potential highway rate**, LL improvement, BIC $\alpha = 1$)



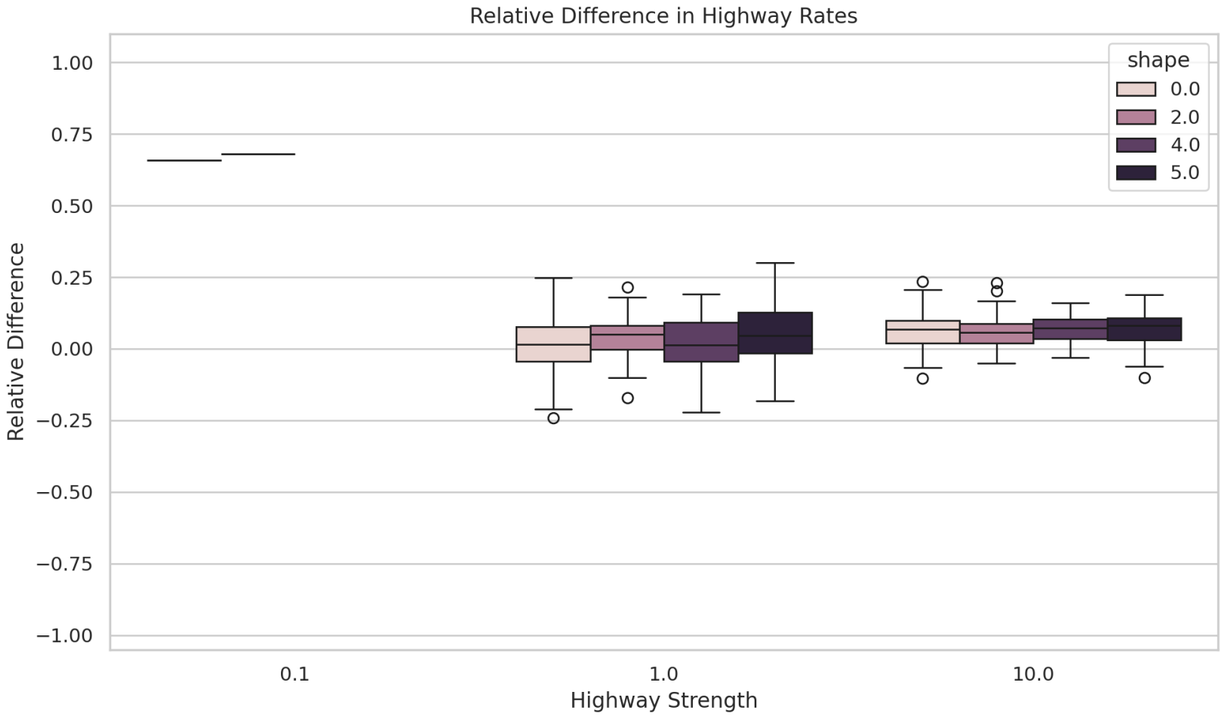
Depending on desired **sensitivity** to highway signals and/or runtime constraints heurstic may be adjusted

Results (Transfer Heuristic cont.)

Weak pruning (distance, LL improvement, BIC $\alpha = 2$)



Strong pruning (distance, **potential highway rate**, LL improvement, BIC $\alpha = 1$)



Limitations

- Missing metric of **realism** of simulated gene trees and their distributions
- Many other possibilities to model highways
- Empirical data is a lot more messy
- Empirical tests show many highways with varying strength, **how many should we capture**?
- Big **additional runtime** for the parameter optimization
- Some **numerical issues** remain

FORTH INSTITUTE OF COMPUTER SCIENCE
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ UNIVERSITY OF CRETE

Future Work

- (More) empirical tests
- Different ways of simulating gene trees and highways
- Different **penalties/optimization functions** to avoid overfitting
- Joint optimization of DTL + highway parameters
- **Faster inference** of highway rates that are "good enough"
- Investigating the impact of **ghost lineages** on highways

Appendix: Likelihood Math

$$\Pi_{e,\gamma} = p_e^S \sum_{\gamma',\gamma''|\gamma} p(\gamma',\gamma''|\gamma)(\Pi_{f,\gamma'}\Pi_{g,\gamma''} + \Pi_{f,\gamma''}\Pi_{g,\gamma'}) \quad (S)$$

$$+ p_e^S \ (\Pi_{f,\gamma}E_g + \Pi_{g,\gamma}E_f) \quad (SL)$$

$$+ p_e^D \sum_{\gamma',\gamma''|\gamma} p(\gamma',\gamma''|\gamma)\Pi_{e,\gamma'}\Pi_{e,\gamma''} \quad (D)$$

$$+ 2p_e^D \ \Pi_{e,\gamma}E_e \quad (DL)$$

$$+ p_e^T \sum_{\gamma',\gamma''|\gamma} p(\gamma',\gamma''|\gamma)(\Pi_{e,\gamma'}\bar{\Pi}_{e,\gamma''} + \Pi_{e,\gamma''}\bar{\Pi}_{e,\gamma'}) \quad (T)$$

$$+ p_e^T \ (\Pi_{e,\gamma}\bar{E}_e + \bar{\Pi}_{e,\gamma}E_e) \quad (TL)$$

$$+ \sum_{d\in H_e} p_{e,d}^H \sum_{\gamma',\gamma''|\gamma} p(\gamma',\gamma''|\gamma)(\Pi_{e,\gamma'}\Pi_{d,\gamma''} + \Pi_{e,\gamma''}\Pi_{d,\gamma'}) \quad (H)$$

$$+ \sum_{d\in H_e} p_{e,d}^H \ (\Pi_{e,\gamma}E_d + \Pi_{d,\gamma}E_e) \quad (HL)$$

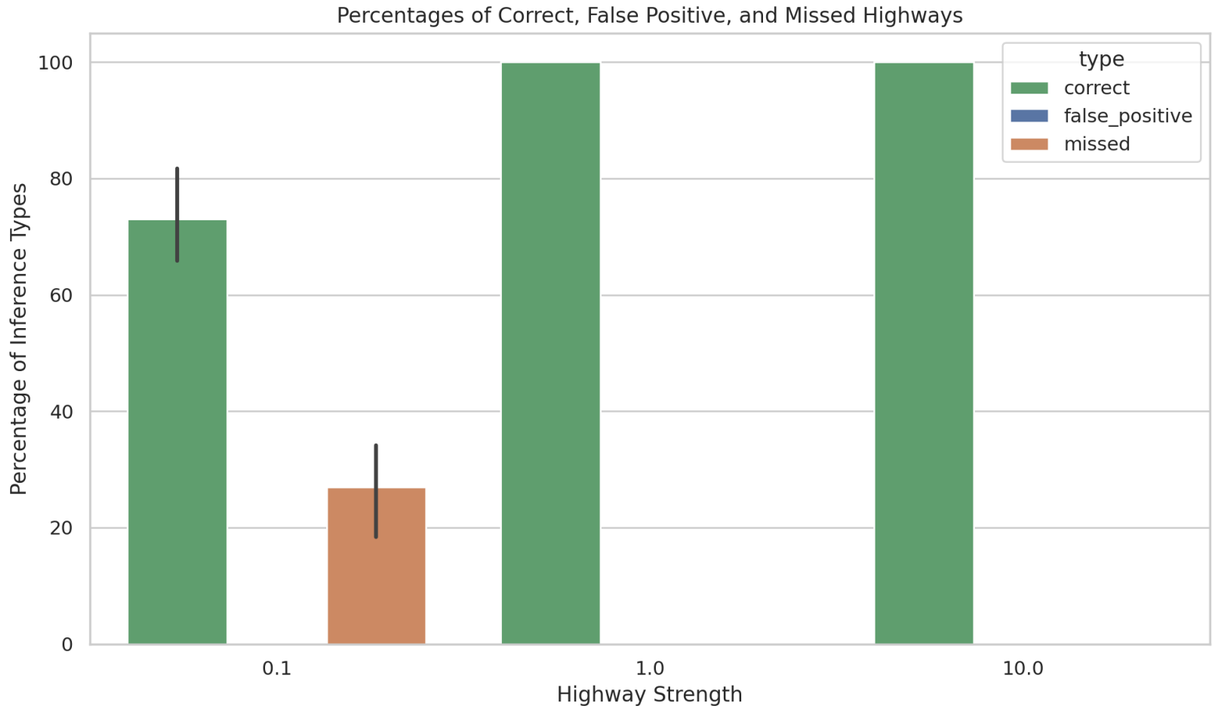
Appendix: Likelihood Math (cont.)

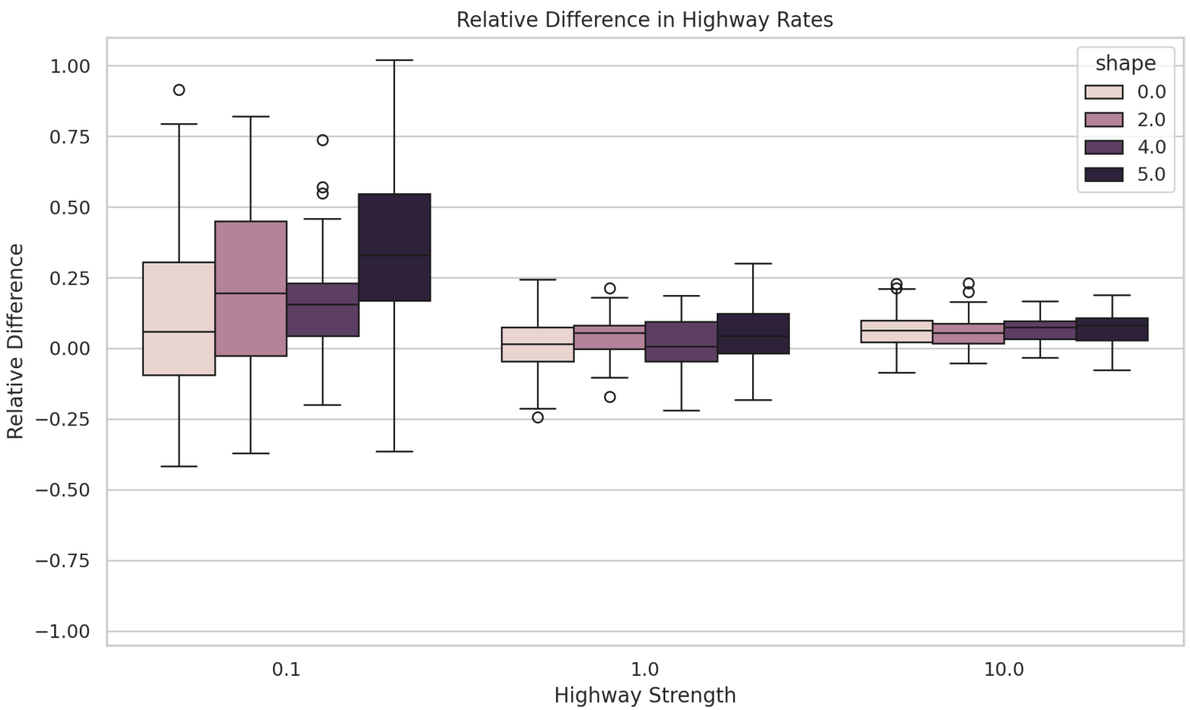
$$E_e = p_e^L + p_e^S \left(E_f E_g\right) + p_e^D \left(E_e^2\right) + p_e^T \left(E_e \bar{E}_e\right) + \sum_{d \in H_e} p_{e,d}^H \left(E_e E_d\right)$$

For a **gene alignment** A and a **species tree** S:

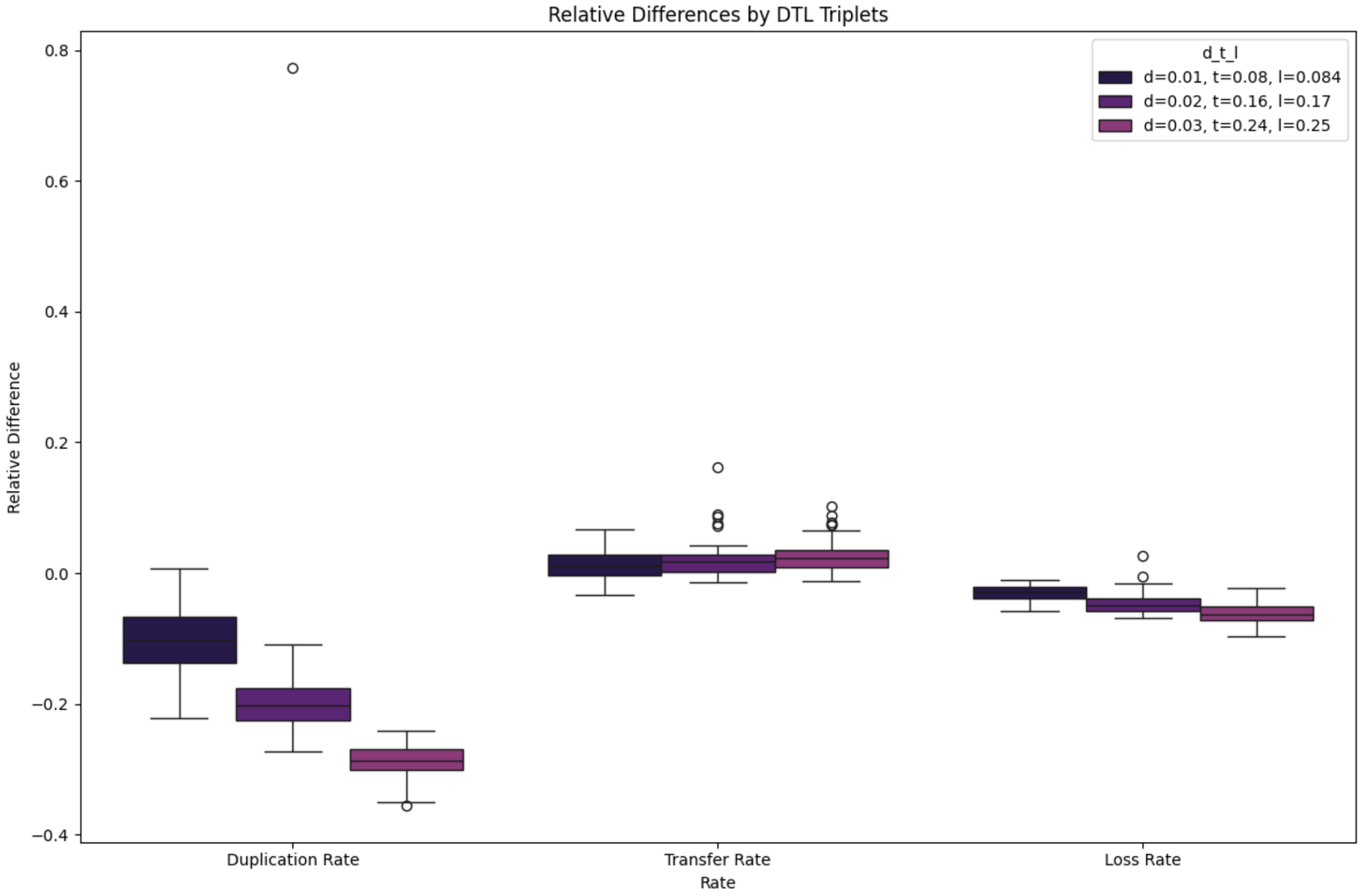
$$P(A|S) = \frac{\sum_e p_e^O \Pi_{e,\Gamma}}{\sum_e p_e^O (1 - E_e)}$$

Appendix: Results (Highway Validation)

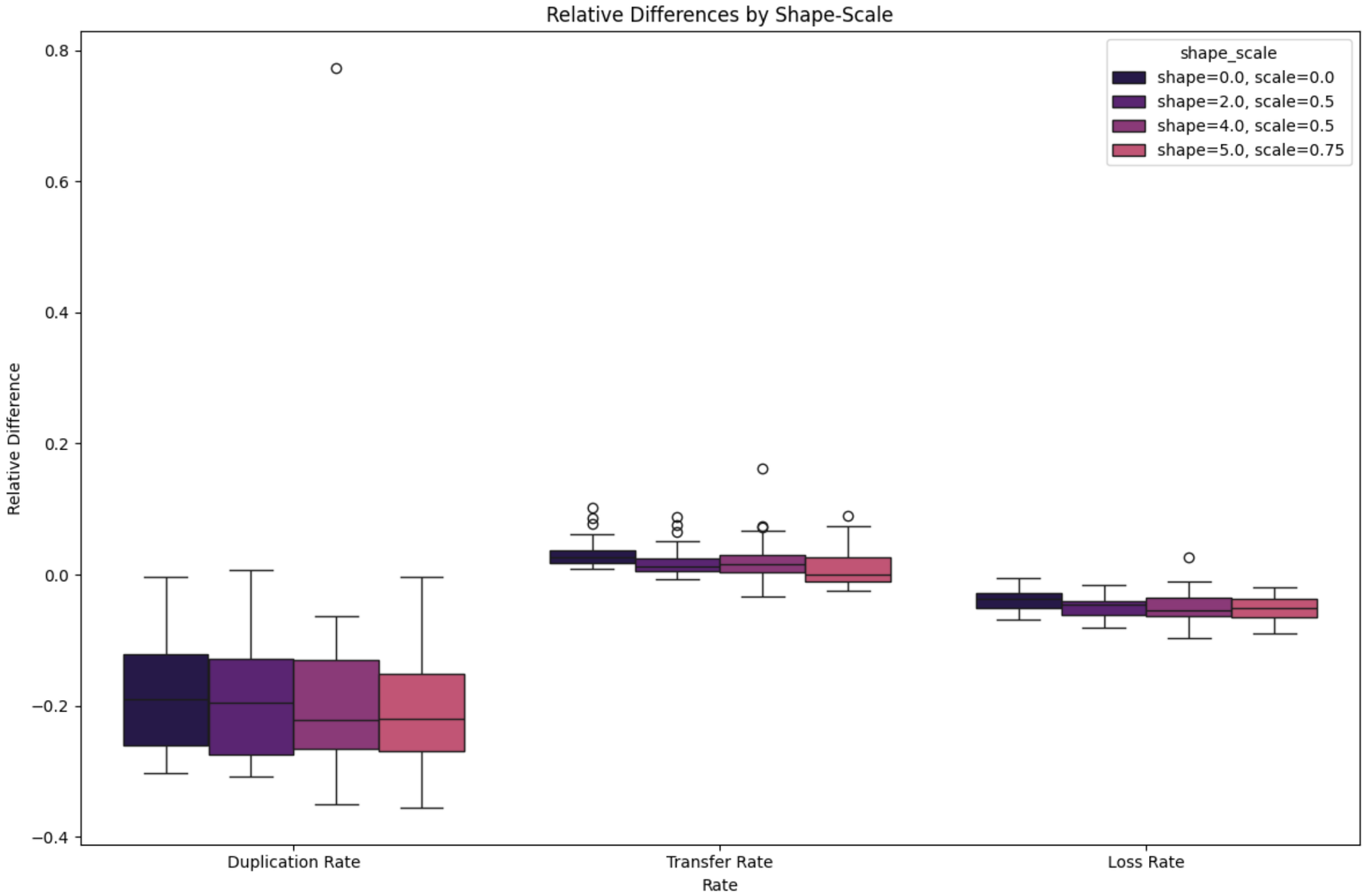




Appendix: Results (DTL)



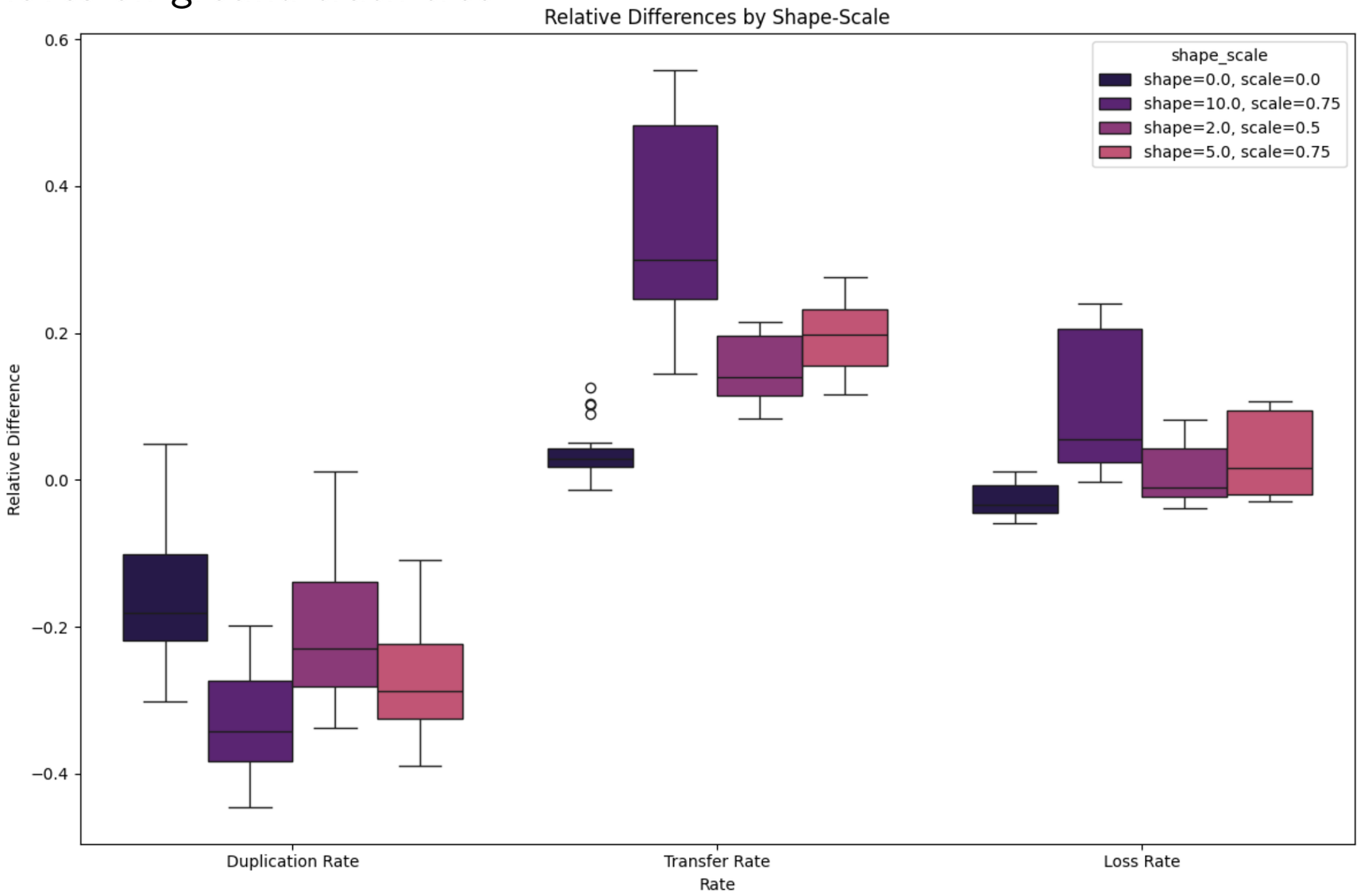
Appendix: Results (DTL cont.)



FORTH INSTITUTE OF COMPUTER SCIENCE
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ UNIVERSITY OF CRETE

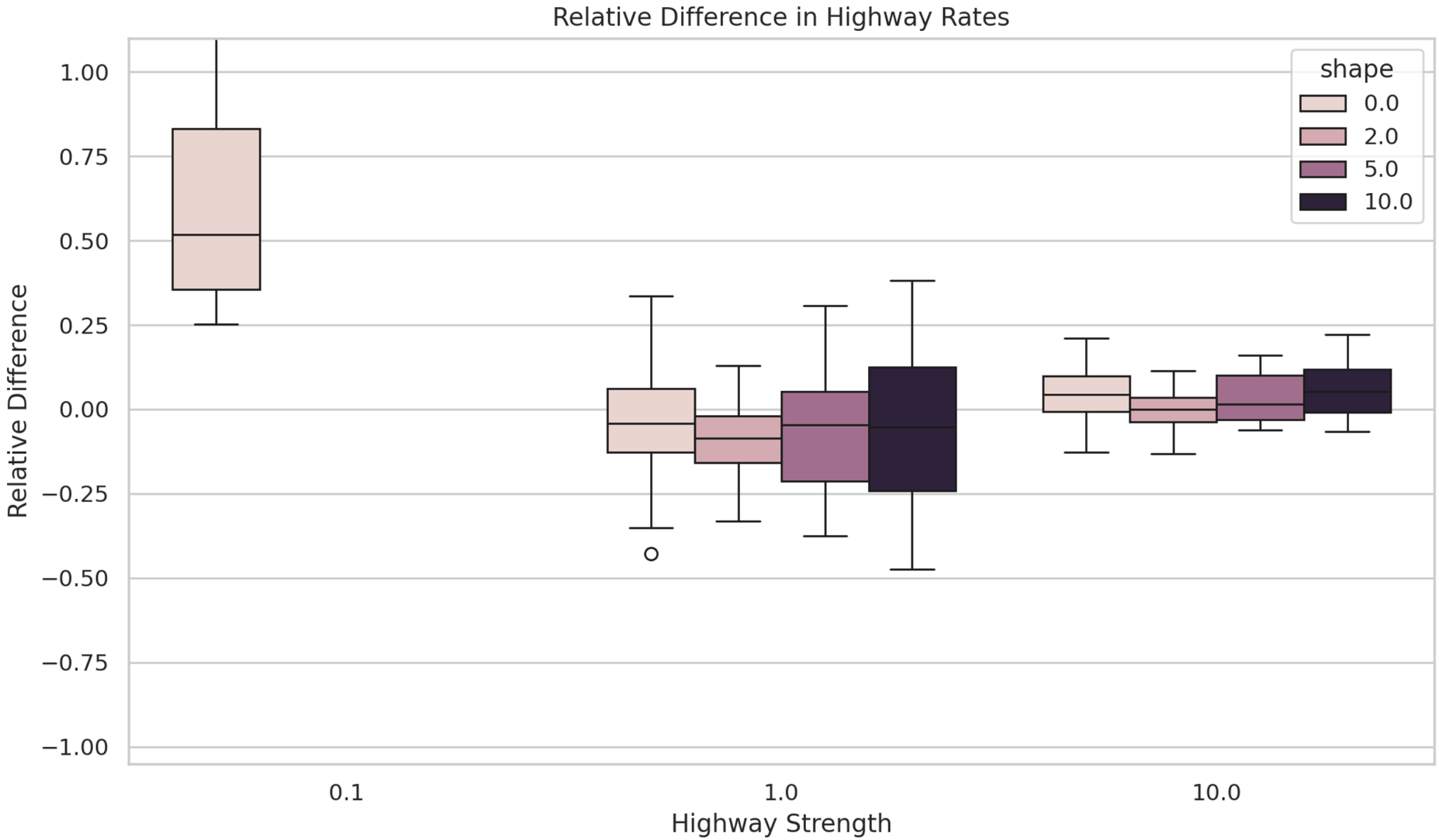
Appendix: Pushing Error and Noise

Additional NNI moves on ground truth tree

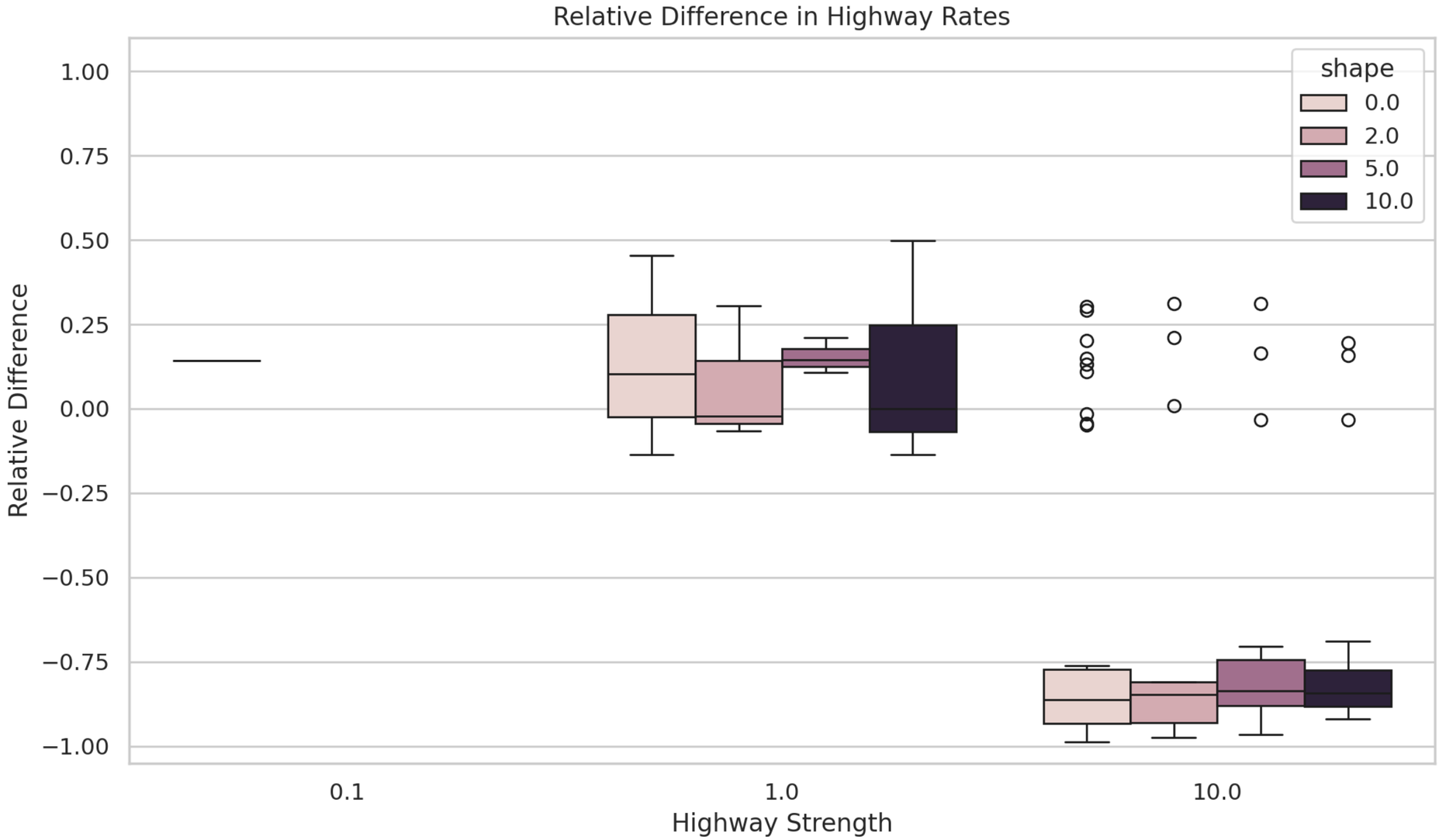


D, T, L = 0.1, 0.3, 0.4

Appendix: Pushing Error and Noise



Appendix: High Rates and Error



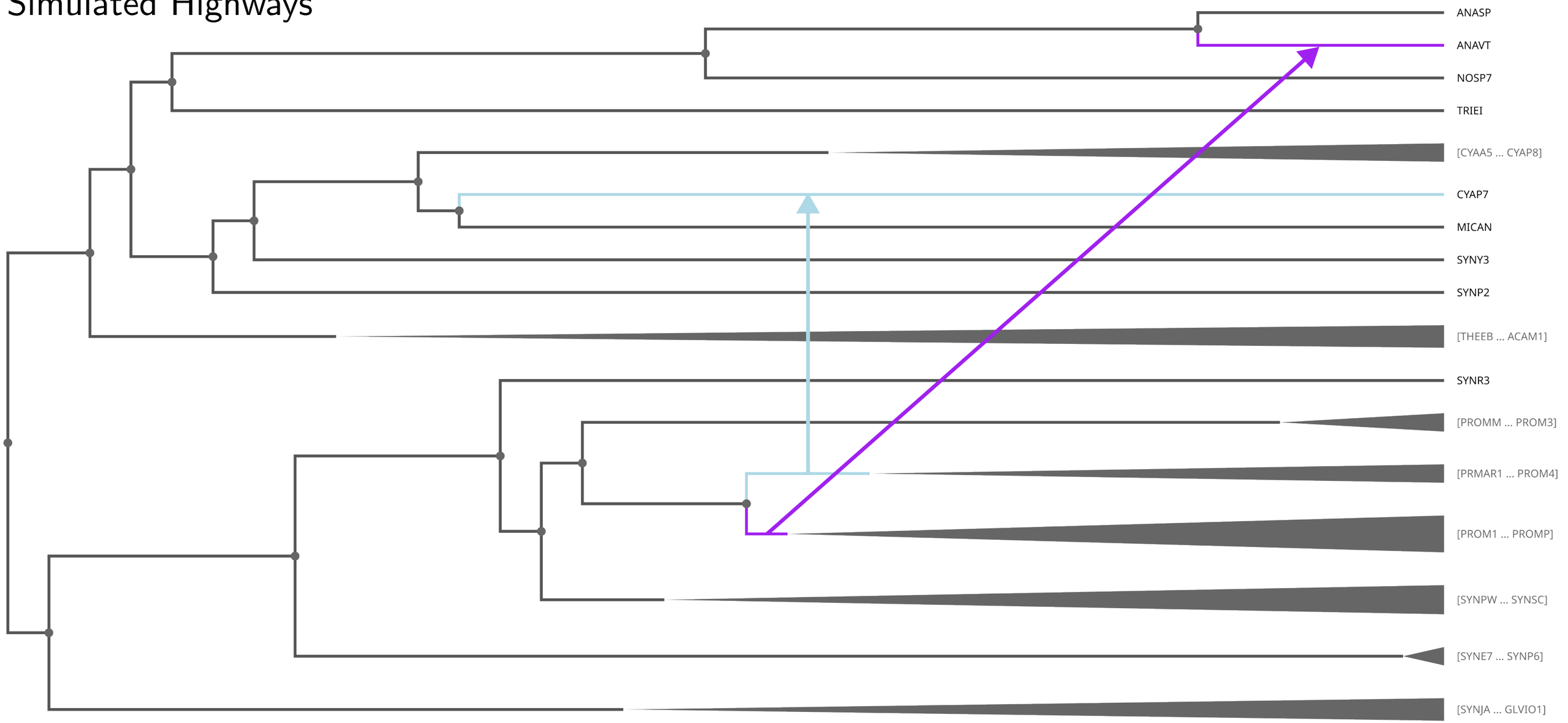
Appendix: Results (Complex Simulations)

Change in DTL inference error: $(\Delta D = 0.145, \Delta T = 0.157, \Delta L = -0.016)$

FORTH INSTITUTE OF COMPUTER SCIENCE
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ UNIVERSITY OF CRETE

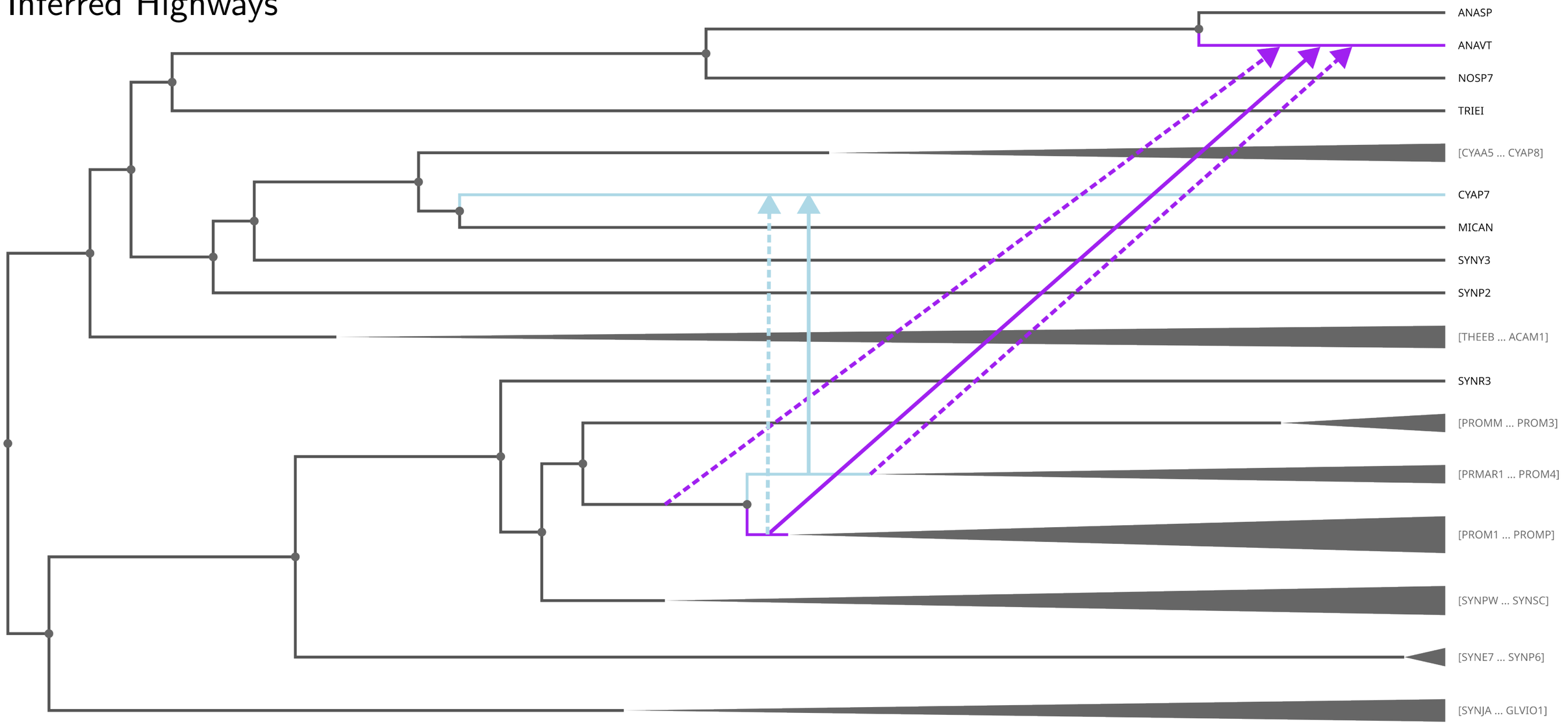
Appendix: Results (Complex Simulations)

Simulated Highways



Appendix: Results (Complex Simulations)

Inferred Highways



Appendix: Results (Complex Simulations)

Inferred Highways