Evaluating strategies of phylogenetical analyses by the coherence of their results

Blaise Li

Journées de la SFS - Paris - 10/10/2012

Method selection

Frequent approach:

- A priori decision, based on a compromise between complexity of the operations and expected accuracy of the results.

Method selection

Frequent approach:

- A priori decision, based on a compromise between complexity of the operations and expected accuracy of the results.
- Use of a testing tool, like `jModelTest` (Posada, 2008), that makes likelihood calculations on trees obtained by fast methods.

Method selection

Frequent approach:

- A priori decision, based on a compromise between complexity of the operations and expected accuracy of the results.
- Use of a testing tool, like `jModelTest` (Posada, 2008), that makes likelihood calculations on trees obtained by fast methods.

My proposition:

- Apply various methods.

Method selection

Frequent approach:

- A priori decision, based on a compromise between complexity of the operations and expected accuracy of the results.
- Use of a testing tool, like `jModelTest` (Posada, 2008), that makes likelihood calculations on trees obtained by fast methods.

My proposition:

- Apply various methods.
- Look how coherent the results obtained are when a given method is used on several datasets.

Method selection

Frequent approach:

- A priori decision, based on a compromise between complexity of the operations and expected accuracy of the results.
- Use of a testing tool, like `jModelTest` (Posada, 2008), that makes likelihood calculations on trees obtained by fast methods.

My proposition:

- Apply various methods.
- Look how coherent the results obtained are when a given method is used on several datasets.
- Choose a posteriori the method that generated the most coherent results.

More than just method: Strategy of analysis

Only the results count, so the thing to select can be a whole analysis pipeline, including such things as:

More than just method: Strategy of analysis

Only the results count, so the thing to select can be a whole analysis pipeline, including such things as:

- Data selection and pre-processing (alignment, trimming, recoding, . . .)

More than just method: Strategy of analysis

Only the results count, so the thing to select can be a whole analysis pipeline, including such things as:

- Data selection and pre-processing (alignment, trimming, recoding, . . .)
- Model (sustitution rates, composition, heterogeneities, correlations, . . .)

More than just method: Strategy of analysis

Only the results count, so the thing to select can be a whole analysis pipeline, including such things as:

- Data selection and pre-processing (alignment, trimming, recoding, . . .)
- Model (sustitution rates, composition, heterogeneities, correlations, . . .)
- Method of inference (distance, parsimony, likelihood, . . .)

More than just method: Strategy of analysis

Only the results count, so the thing to select can be a whole analysis pipeline, including such things as:

- Data selection and pre-processing (alignment, trimming, recoding, . . .)
- Model (sustitution rates, composition, heterogeneities, correlations, . . .)
- Method of inference (distance, parsimony, likelihood, . . .) and it implementation (program, options, tunings, . . .)

More than just method: Strategy of analysis

Only the results count, so the thing to select can be a whole analysis pipeline, including such things as:

- Data selection and pre-processing (alignment, trimming, recoding, . . .)
- Model (sustitution rates, composition, heterogeneities, correlations, . . .)
- Method of inference (distance, parsimony, likelihood, . . .) and it implementation (program, options, tunings, . . .)
- Support evaluation (bootstrapping, . . .)

More than just method: Strategy of analysis

Only the results count, so the thing to select can be a whole analysis pipeline, including such things as:

- Data selection and pre-processing (alignment, trimming, recoding, . . .)
- Model (sustitution rates, composition, heterogeneities, correlations, . . .)
- Method of inference (distance, parsimony, likelihood, . . .) and it implementation (program, options, tunings, . . .)
- Support evaluation (bootstrapping, . . .)
- <Insert your favourite strategy item here>

Result coherence

Why?:

Result coherence

Why?:

- The most accurate strategies should extract more historical signal than the others.

Result coherence

Why?:

- The most accurate strategies should extract more historical signal than the others.
- So the different datasets should produce more similar trees whith these strategies.

Result coherence

Why?:

- The most accurate strategies should extract more historical signal than the others.
- So the different datasets should produce more similar trees whith these strategies.

Result coherence

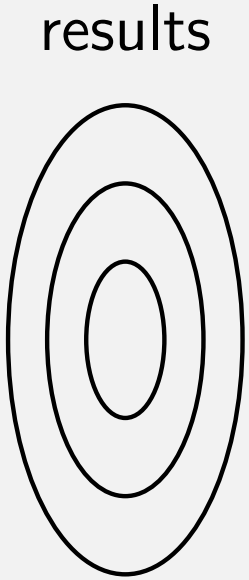
Why?:

- The most accurate strategies should extract more historical signal than the others.
- So the different datasets should produce more similar trees whith these strategies.

Result coherence

Why?:

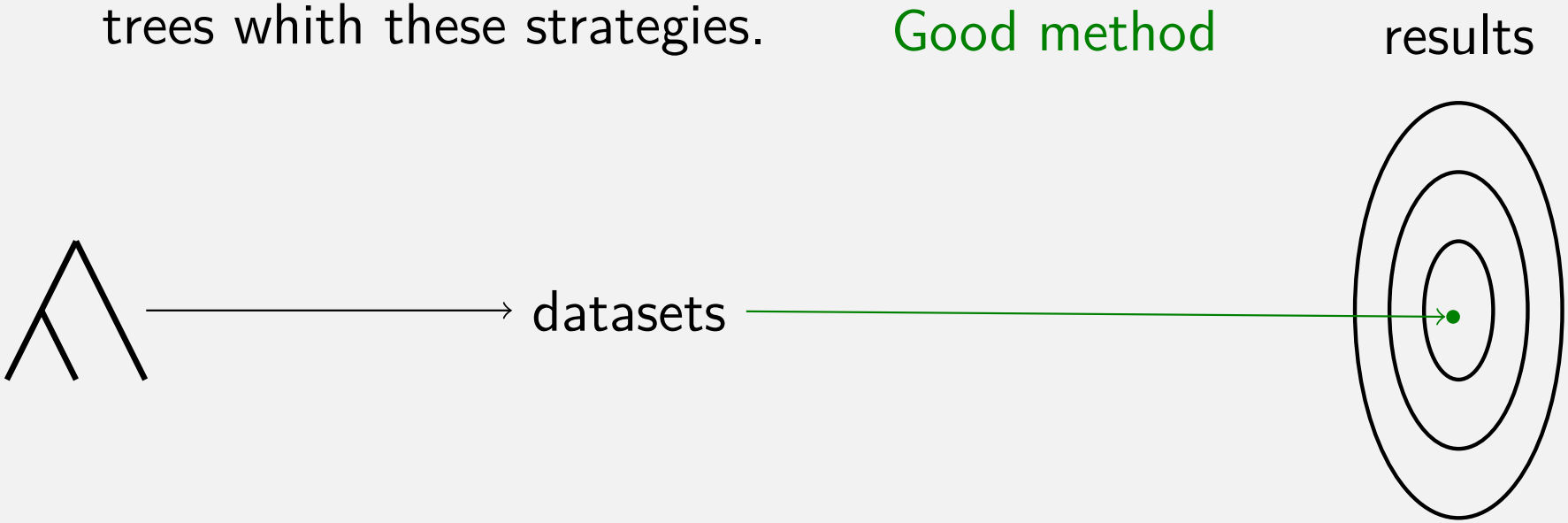
- The most accurate strategies should extract more historical signal than the others.
- So the different datasets should produce more similar trees whith these strategies.



Result coherence

Why?:

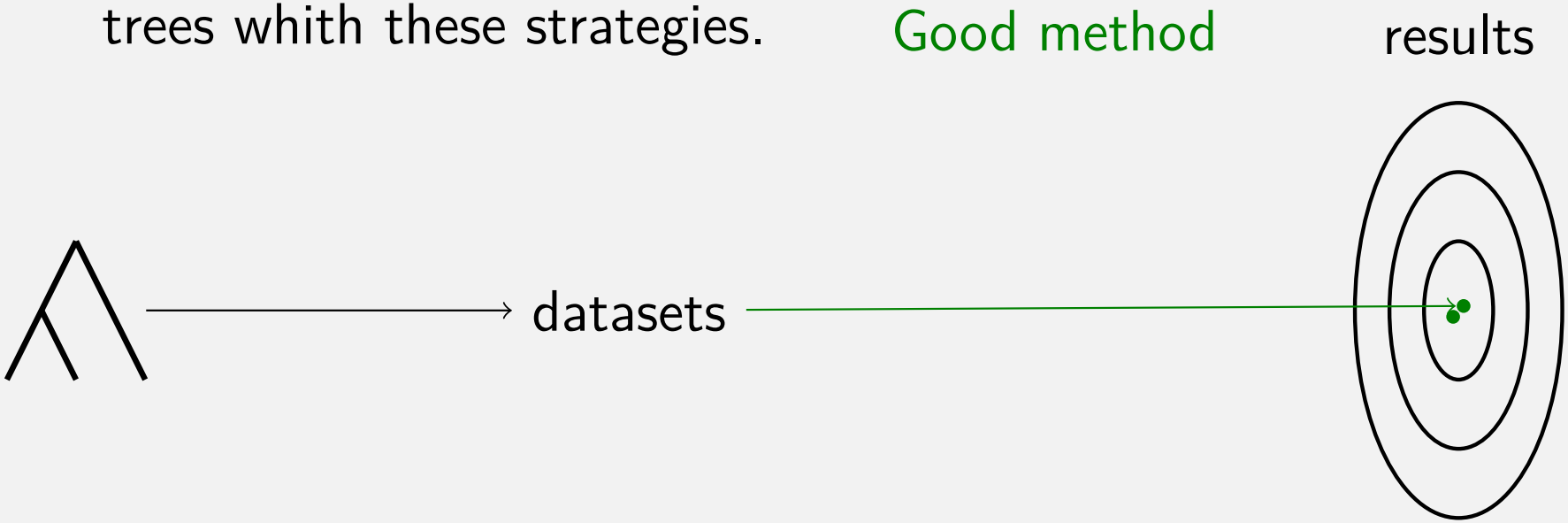
- The most accurate strategies should extract more historical signal than the others.
- So the different datasets should produce more similar trees whith these strategies.



Result coherence

Why?:

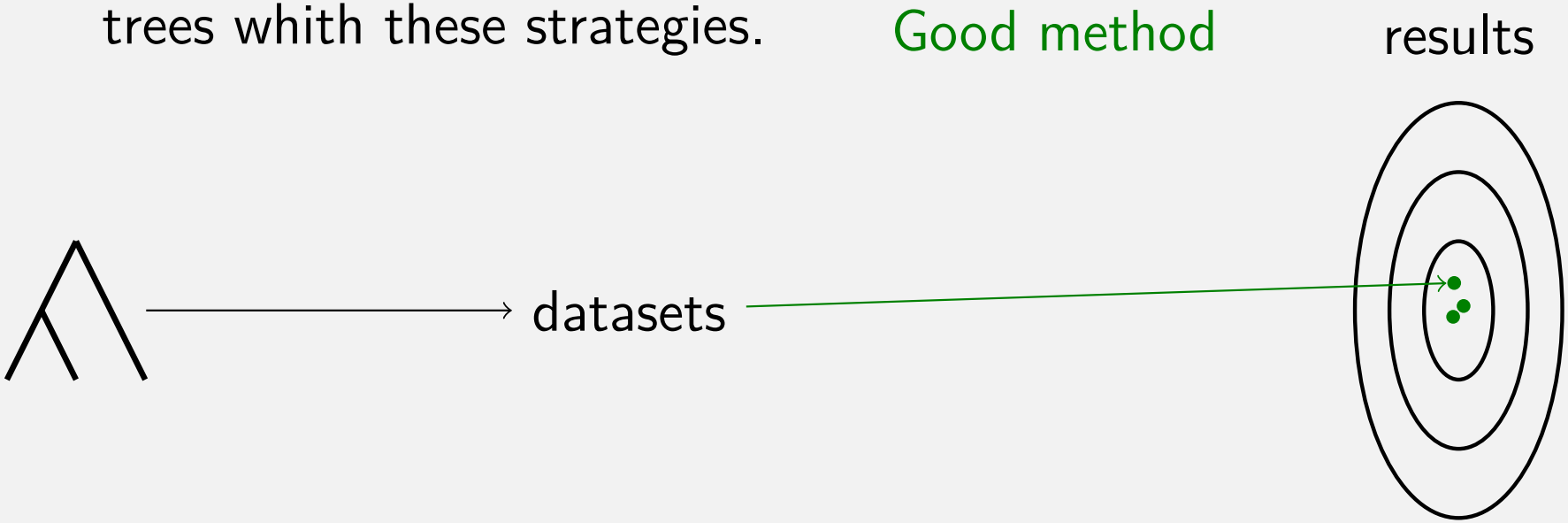
- The most accurate strategies should extract more historical signal than the others.
- So the different datasets should produce more similar trees whith these strategies.



Result coherence

Why?:

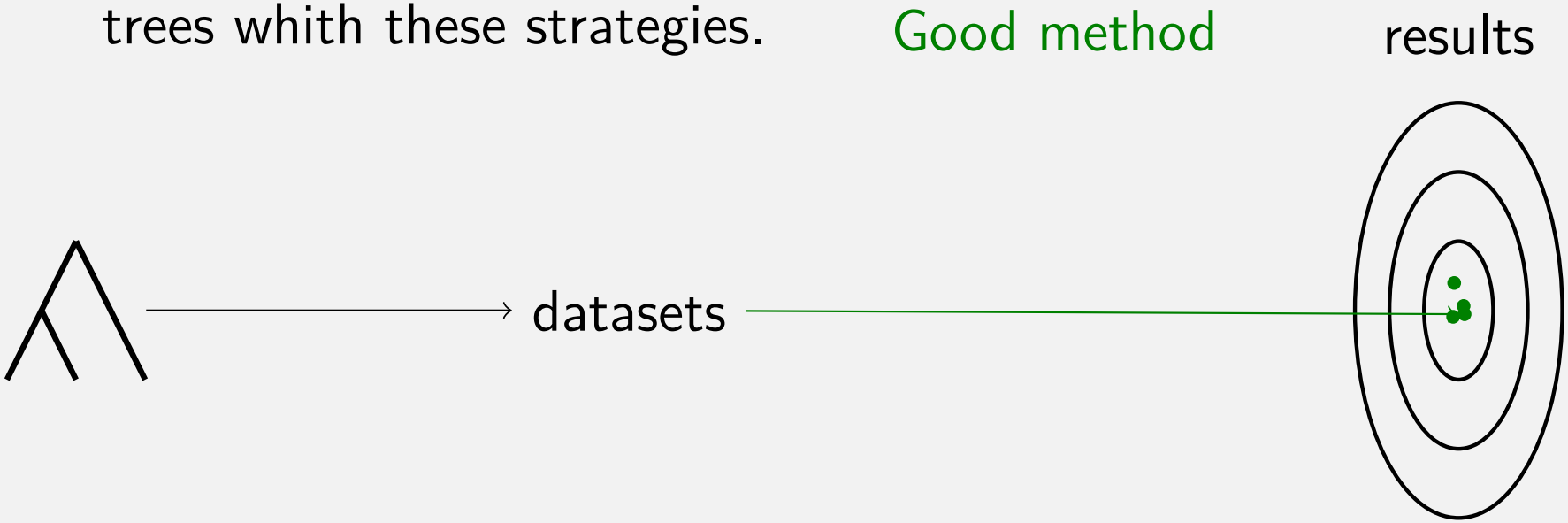
- The most accurate strategies should extract more historical signal than the others.
- So the different datasets should produce more similar trees whith these strategies.



Result coherence

Why?:

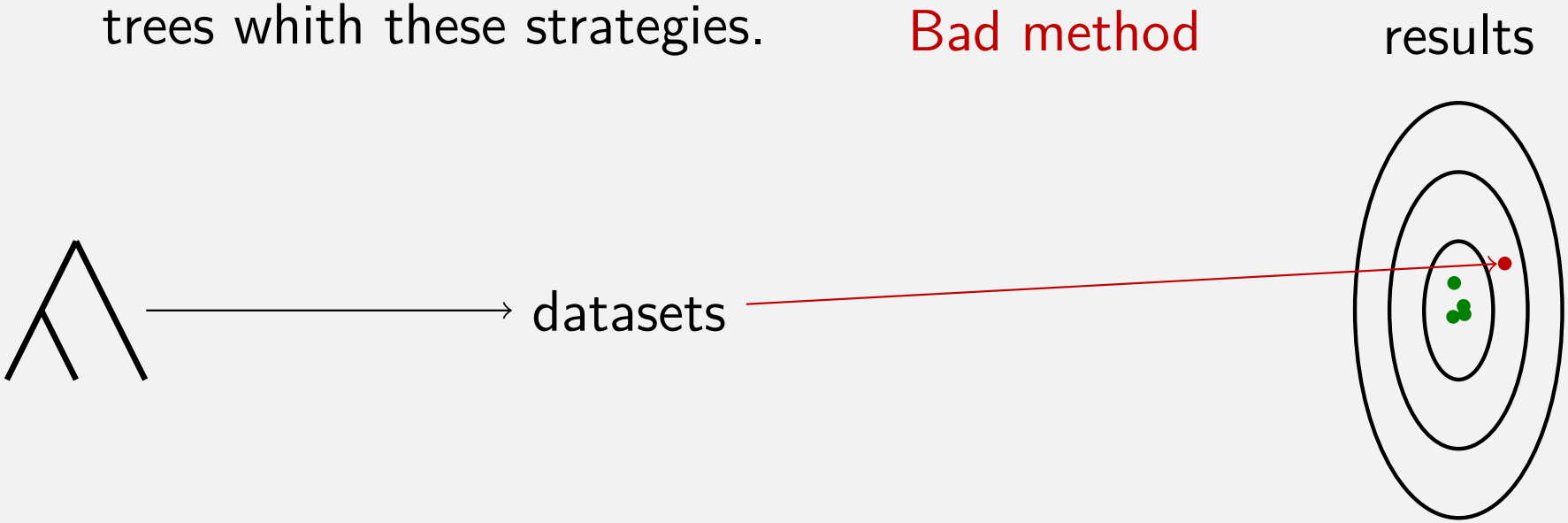
- The most accurate strategies should extract more historical signal than the others.
- So the different datasets should produce more similar trees whith these strategies.



Result coherence

Why?:

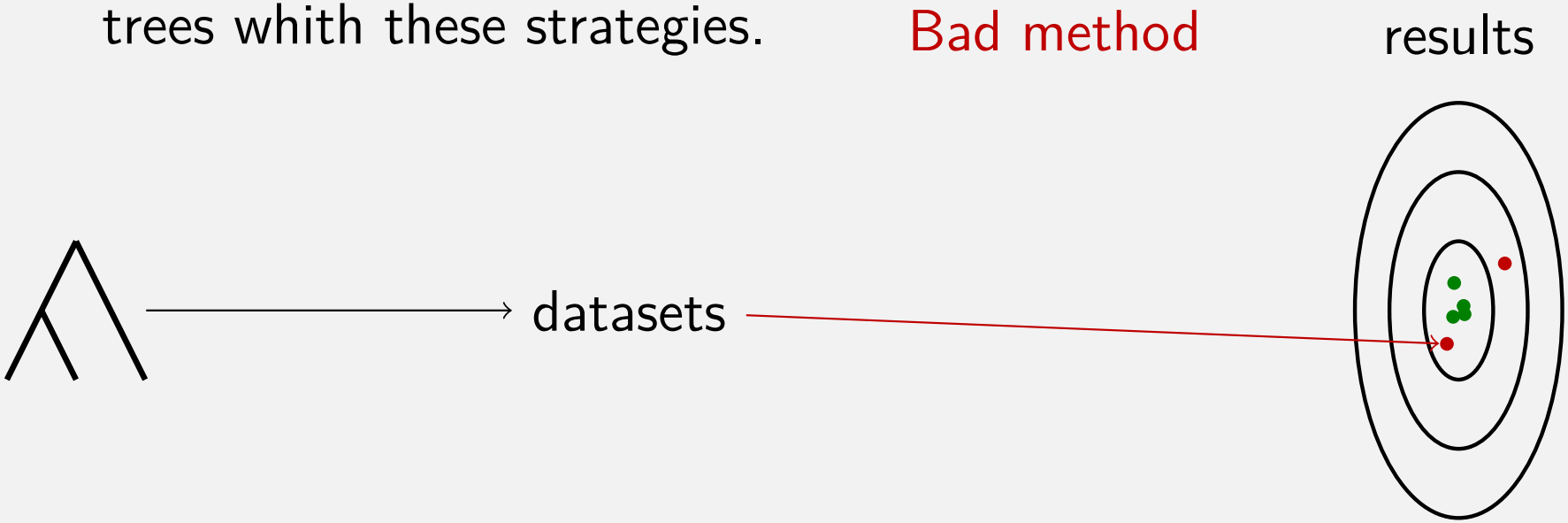
- The most accurate strategies should extract more historical signal than the others.
- So the different datasets should produce more similar trees whith these strategies.



Result coherence

Why?:

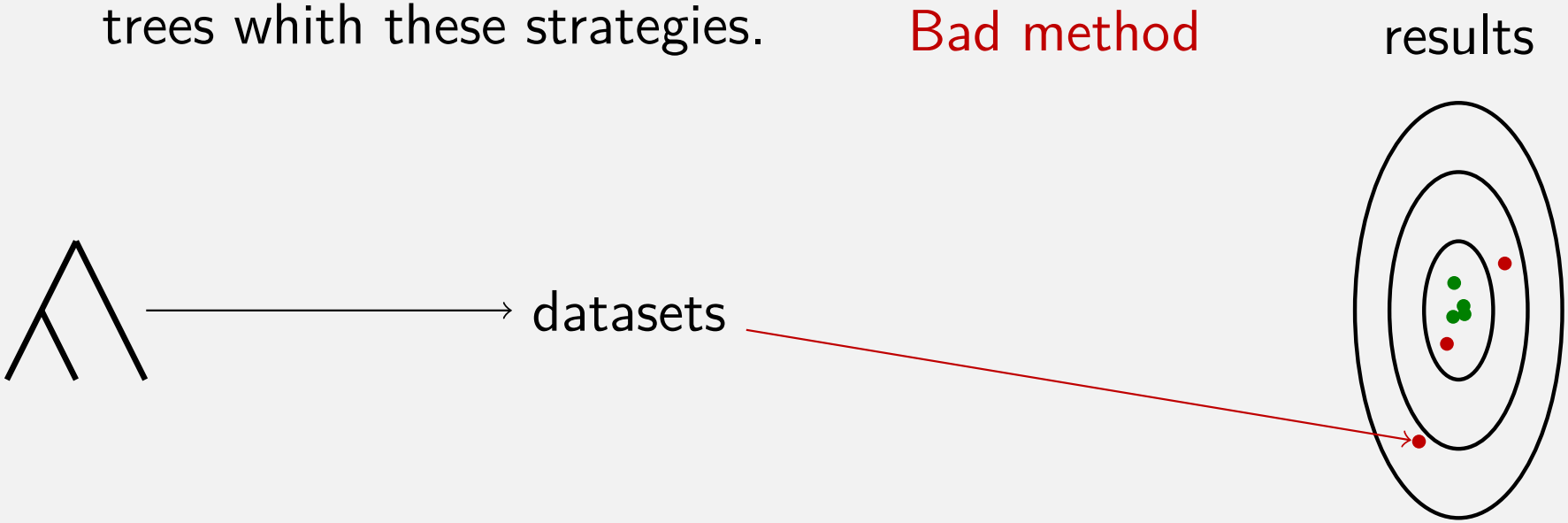
- The most accurate strategies should extract more historical signal than the others.
- So the different datasets should produce more similar trees whith these strategies.



Result coherence

Why?:

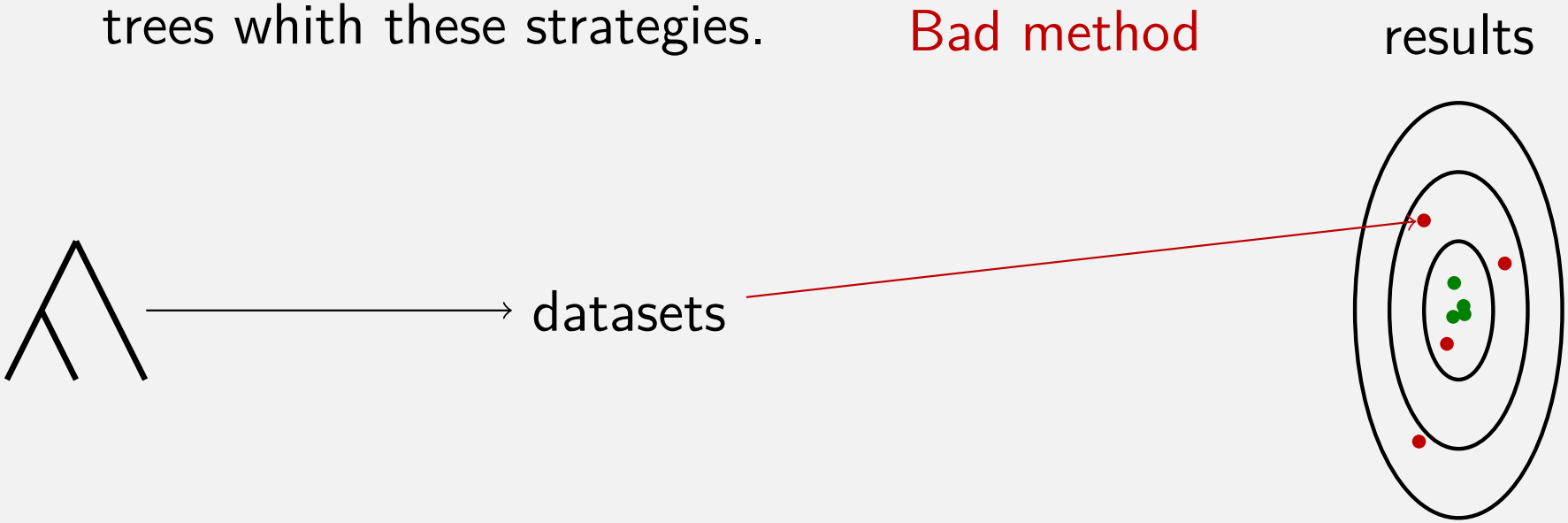
- The most accurate strategies should extract more historical signal than the others.
- So the different datasets should produce more similar trees whith these strategies.



Result coherence

Why?:

- The most accurate strategies should extract more historical signal than the others.
- So the different datasets should produce more similar trees whith these strategies.



Result coherence

But:

Result coherence

But:

- Accurate strategies will produce consistently divergent results if the datasets evolved under different histories (false negative).

Result coherence

But:

- Accurate strategies will produce consistently divergent results if the datasets evolved under different histories (false negative).

Result coherence

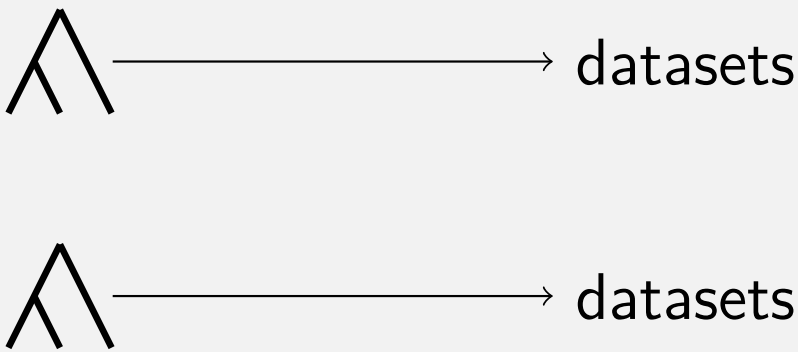
But:

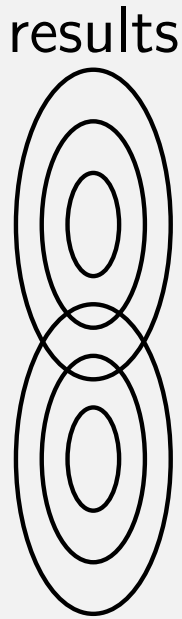
- Accurate strategies will produce consistently divergent results if the datasets evolved under different histories (false negative).

Result coherence

But:

- Accurate strategies will produce consistently divergent results if the datasets evolved under different histories (false negative).





Result coherence

But:

- Accurate strategies will produce consistently divergent results if the datasets evolved under different histories (false negative).

Result coherence

But:

- Accurate strategies will produce consistently divergent results if the datasets evolved under different histories (false negative).

Result coherence

But:

- Accurate strategies will produce consistently divergent results if the datasets evolved under different histories (false negative).

Result coherence

But:

- Accurate strategies will produce consistently divergent results if the datasets evolved under different histories (false negative).

Result coherence

But:

- Accurate strategies will produce consistently divergent results if the datasets evolved under different histories (false negative).

Result coherence

But:

- Accurate strategies will produce consistently divergent results if the datasets evolved under different histories (false negative).

Result coherence

But:

- Accurate strategies will produce consistently divergent results if the datasets evolved under different histories (false negative).

Result coherence

But:

- Accurate strategies will produce consistently divergent results if the datasets evolved under different histories (false negative).

Result coherence

But:

- Strategies prone to reconstruction errors will produce wrong results in a consistent manner if the datasets share the error-inducing characteristics (false positive).

Result coherence

But:

- Strategies prone to reconstruction errors will produce wrong results in a consistent manner if the datasets share the error-inducing characteristics (false positive).

Result coherence

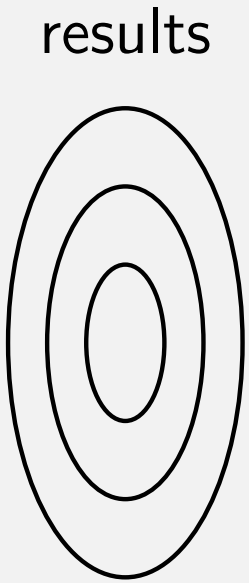
But:

- Strategies prone to reconstruction errors will produce wrong results in a consistent manner if the datasets share the error-inducing characteristics (false positive).

Result coherence

But:

- Strategies prone to reconstruction errors will produce wrong results in a consistent manner if the datasets share the error-inducing characteristics (false positive).



Result coherence

But:

- Strategies prone to reconstruction errors will produce wrong results in a consistent manner if the datasets share the error-inducing characteristics (false positive).

Result coherence

But:

- Strategies prone to reconstruction errors will produce wrong results in a consistent manner if the datasets share the error-inducing characteristics (false positive).

Result coherence

But:

- Strategies prone to reconstruction errors will produce wrong results in a consistent manner if the datasets share the error-inducing characteristics (false positive).

Result coherence

But:

- Strategies prone to reconstruction errors will produce wrong results in a consistent manner if the datasets share the error-inducing characteristics (false positive).

Data

- 42 cyanobacteria and plastids
- 73 protein-coding genes

Data

- 42 cyanobacteria and plastids
- 73 protein-coding genes
- 4 sets of congruent markers according to `concaterpillar` (Leigh et al., 2008)

Data

- 42 cyanobacteria and plastids
- 73 protein-coding genes
- 4 sets of congruent markers according to `concaterpillar` (Leigh et al., 2008)

Due to the internals of `concaterpillar` the analyses of the 4 sets should yield results with some degrees of incoherence, at least for standard maximum likelihood under a $\mathrm{GTR} + \mathrm{I} + \Gamma$ model.

Strategies

- Maximum likelihood bootstrap (200 pseudo-replicates):
  - `RAxML` (Stamatakis, 2006), $\mathrm{GTR} + \mathrm{I} + \Gamma$

Strategies

- Maximum likelihood bootstrap (200 pseudo-replicates):
  - `RAxML` (Stamatakis, 2006), $\mathrm{GTR} + \mathrm{I} + \Gamma$
  - recodings to eliminate signal associated with synonymous substitutions

Strategies

- Maximum likelihood bootstrap (200 pseudo-replicates):
  - `RAxML` (Stamatakis, 2006), $\mathrm{GTR} + \mathrm{I} + \Gamma$
  - recodings to eliminate signal associated with synonymous substitutions
- Bayesian MCMC inference:

Strategies

- Maximum likelihood bootstrap (200 pseudo-replicates):
  - `RAxML` (Stamatakis, 2006), $\mathrm{GTR} + \mathrm{I} + \Gamma$
  - recodings to eliminate signal associated with synonymous substitutions
- Bayesian MCMC inference:
  - `Phylobayes` (Lartillot and Philippe, 2004), $\mathrm{GTR} + \mathrm{I} + \Gamma + \mathrm{CAT}$ (site-wise composition heterogeneity)

Strategies

- Maximum likelihood bootstrap (200 pseudo-replicates):
  - `RAxML` (Stamatakis, 2006), $\mathrm{GTR} + \mathrm{I} + \Gamma$
  - recodings to eliminate signal associated with synonymous substitutions
- Bayesian MCMC inference:
  - `Phylobayes` (Lartillot and Philippe, 2004), $\mathrm{GTR} + \mathrm{I} + \Gamma + \mathrm{CAT}$ (site-wise composition heterogeneity)
  - `P4` (Foster, 2004), $\mathrm{GTR} + \mathrm{I} + \Gamma + \mathrm{NDCH}$ (clade-wise composition heterogeneity)

Strategies

- Maximum likelihood bootstrap (200 pseudo-replicates):
  - `RAxML` (Stamatakis, 2006), $\mathrm{GTR} + \mathrm{I} + \Gamma$
  - recodings to eliminate signal associated with synonymous substitutions
- Bayesian MCMC inference:
  - `Phylobayes` (Lartillot and Philippe, 2004), $\mathrm{GTR} + \mathrm{I} + \Gamma + \mathrm{CAT}$ (site-wise composition heterogeneity)
  - `P4` (Foster, 2004), $\mathrm{GTR} + \mathrm{I} + \Gamma + \mathrm{NDCH}$ (clade-wise composition heterogeneity)
- A priori less accurate strategies (using `Phylip`, Felsenstein, 2005):

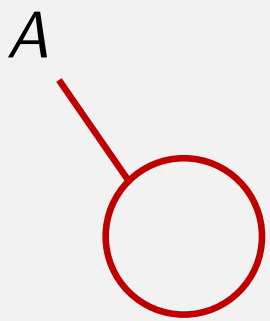
Strategies

- Maximum likelihood bootstrap (200 pseudo-replicates):
  - `RAxML` (Stamatakis, 2006), $\mathrm{GTR} + \mathrm{I} + \Gamma$
  - recodings to eliminate signal associated with synonymous substitutions
- Bayesian MCMC inference:
  - `Phylobayes` (Lartillot and Philippe, 2004), $\mathrm{GTR} + \mathrm{I} + \Gamma + \mathrm{CAT}$ (site-wise composition heterogeneity)
  - `P4` (Foster, 2004), $\mathrm{GTR} + \mathrm{I} + \Gamma + \mathrm{NDCH}$ (clade-wise composition heterogeneity)
- A priori less accurate strategies (using `Phylip`, Felsenstein, 2005):
  - parsimony bootstrap (200 pseudo-replicates)

Strategies

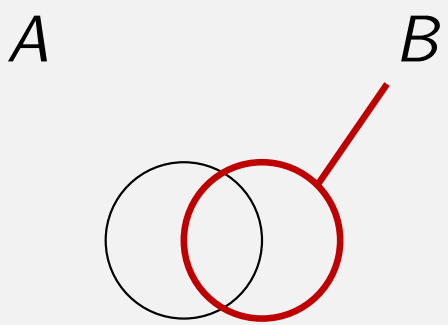
- Maximum likelihood bootstrap (200 pseudo-replicates):
  - `RAxML` (Stamatakis, 2006), $\mathrm{GTR} + \mathrm{I} + \Gamma$
  - recodings to eliminate signal associated with synonymous substitutions
- Bayesian MCMC inference:
  - `Phylobayes` (Lartillot and Philippe, 2004), $\mathrm{GTR} + \mathrm{I} + \Gamma + \mathrm{CAT}$ (site-wise composition heterogeneity)
  - `P4` (Foster, 2004), $\mathrm{GTR} + \mathrm{I} + \Gamma + \mathrm{NDCH}$ (clade-wise composition heterogeneity)
- A priori less accurate strategies (using `Phylip`, Felsenstein, 2005):
  - parsimony bootstrap (200 pseudo-replicates)
  - distance bootstrap (Jukes-Cantor and LogDet, 200 pseudo-replicates)

Robinson-Foulds distance (a.k.a symmetric difference)



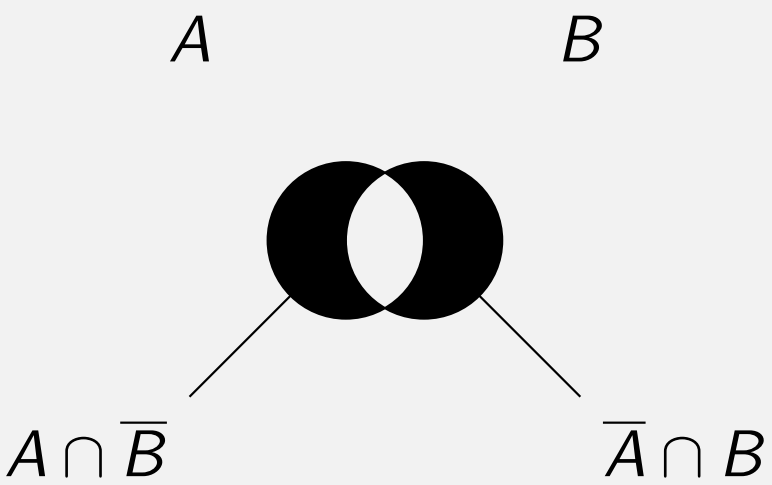
A: bipartitions defined by the branches of one tree

Robinson-Foulds distance (a.k.a symmetric difference)



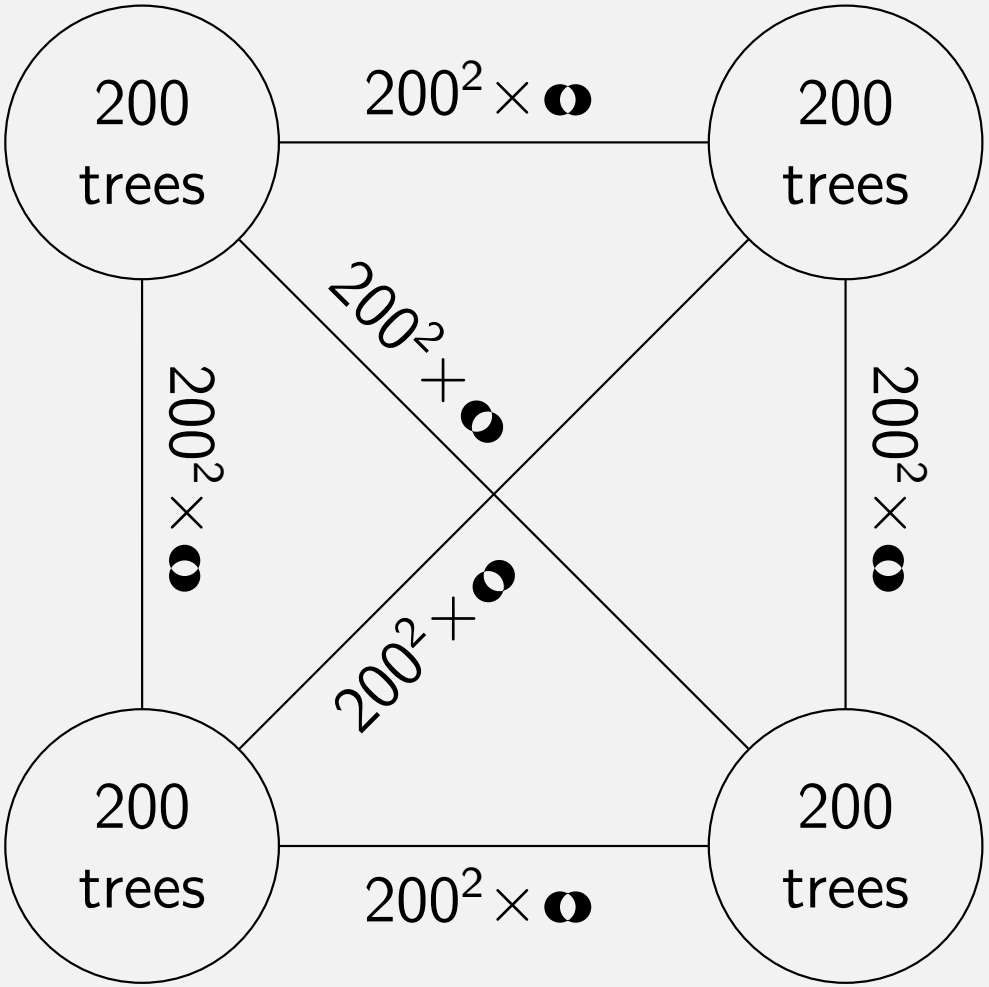
A: bipartitions defined by the branches of one tree
B: bipartitions defined by the branches of the other tree

Robinson-Foulds distance (a.k.a symmetric difference)



A: bipartitions defined by the branches of one tree
B: bipartitions defined by the branches of the other tree
Distance between the trees:
$RF = |A \cap \overline{B}| + |\overline{A} \cap B|$

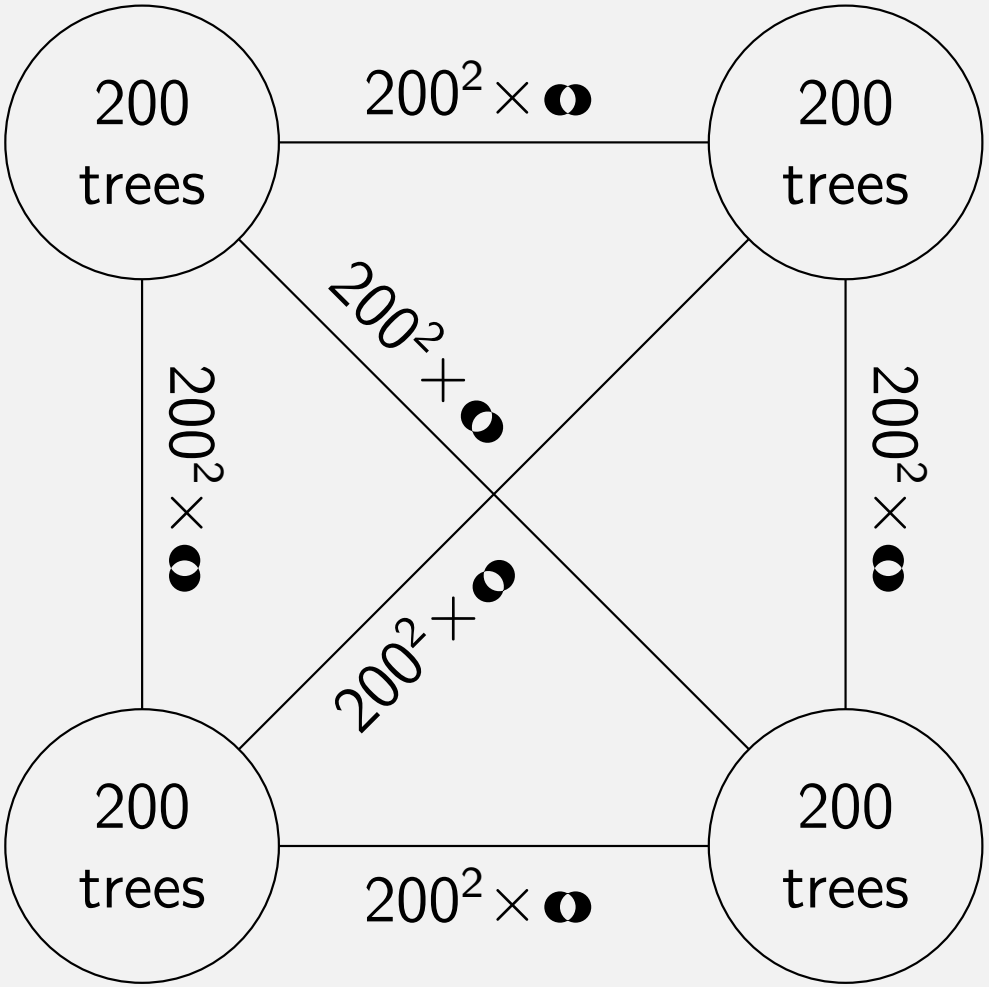
Coherence measure



Full distribution:

- $200^2 \times 6$ Robinson-Foulds distances

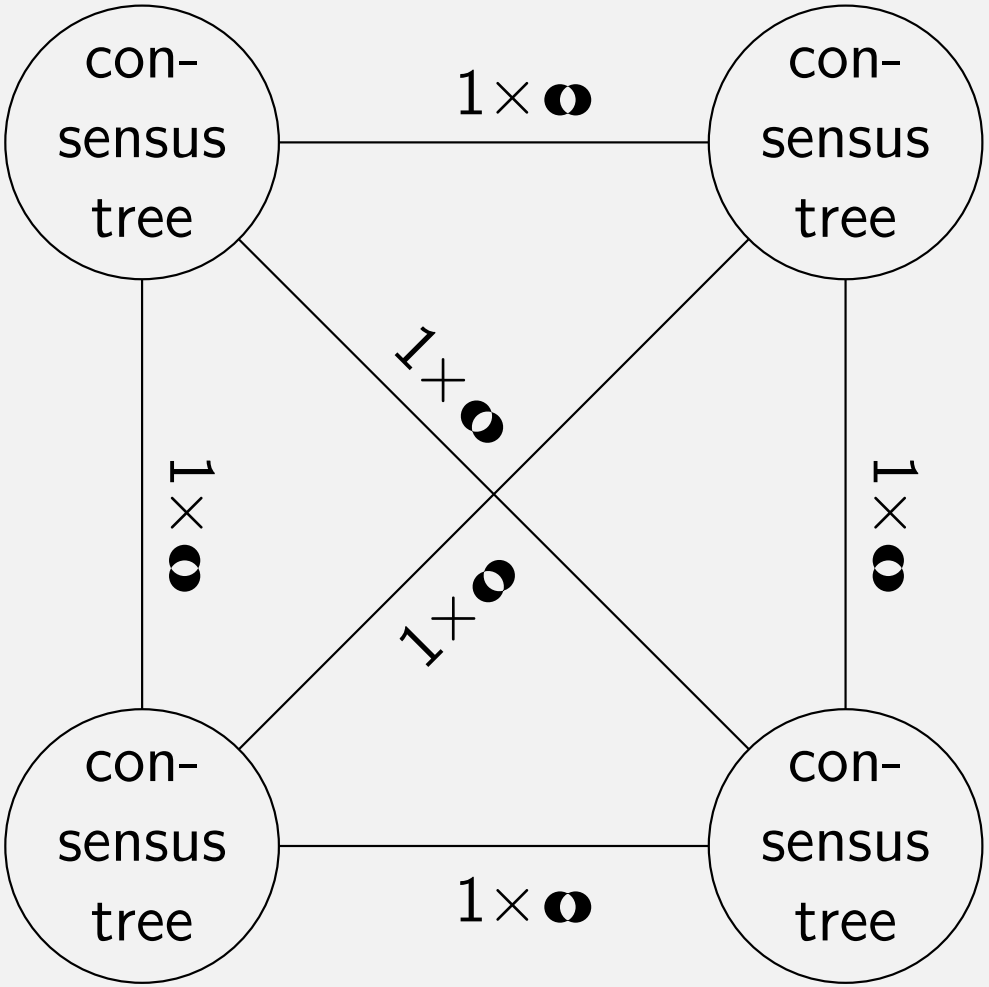
Coherence measure



Full distribution:

- $200^2 \times 6$ Robinson-Foulds distances
- Coherence assessed using the distribution and average of these distances

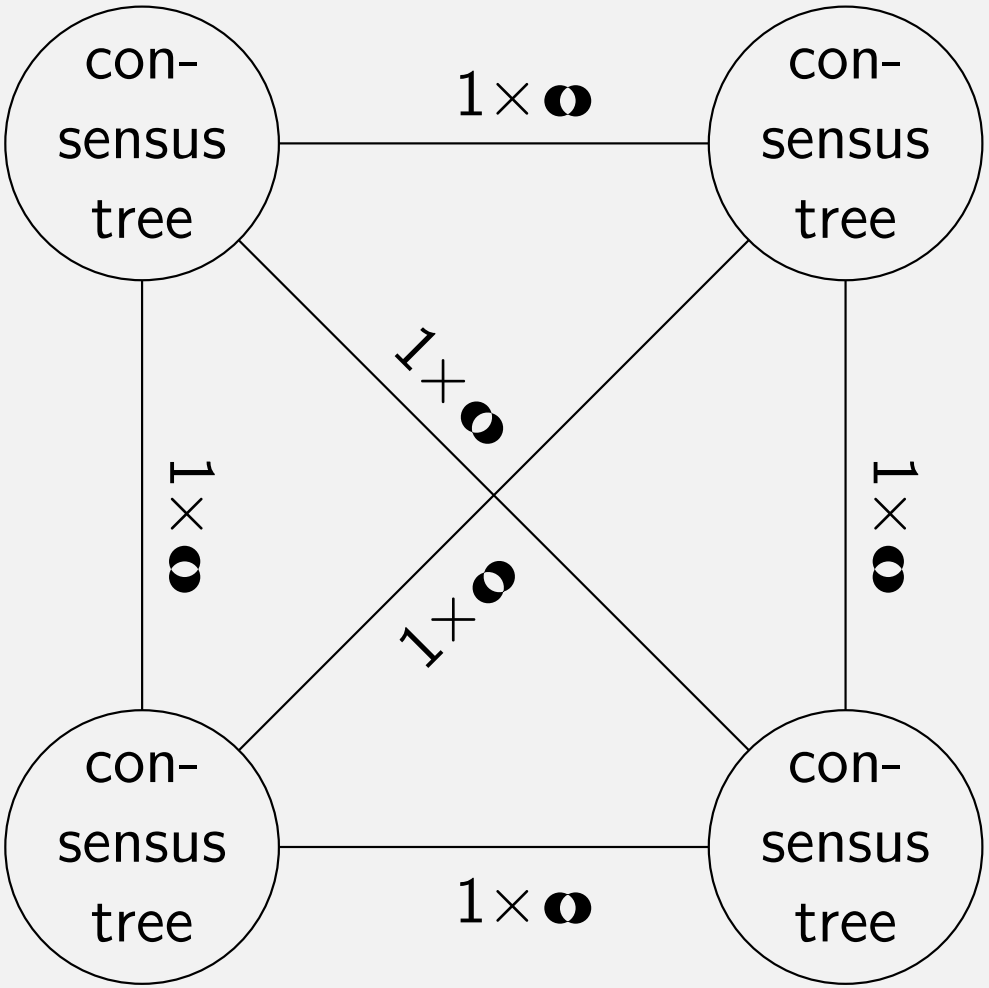
Coherence measure



Between consensus:

- 6 Robinson-Foulds distances

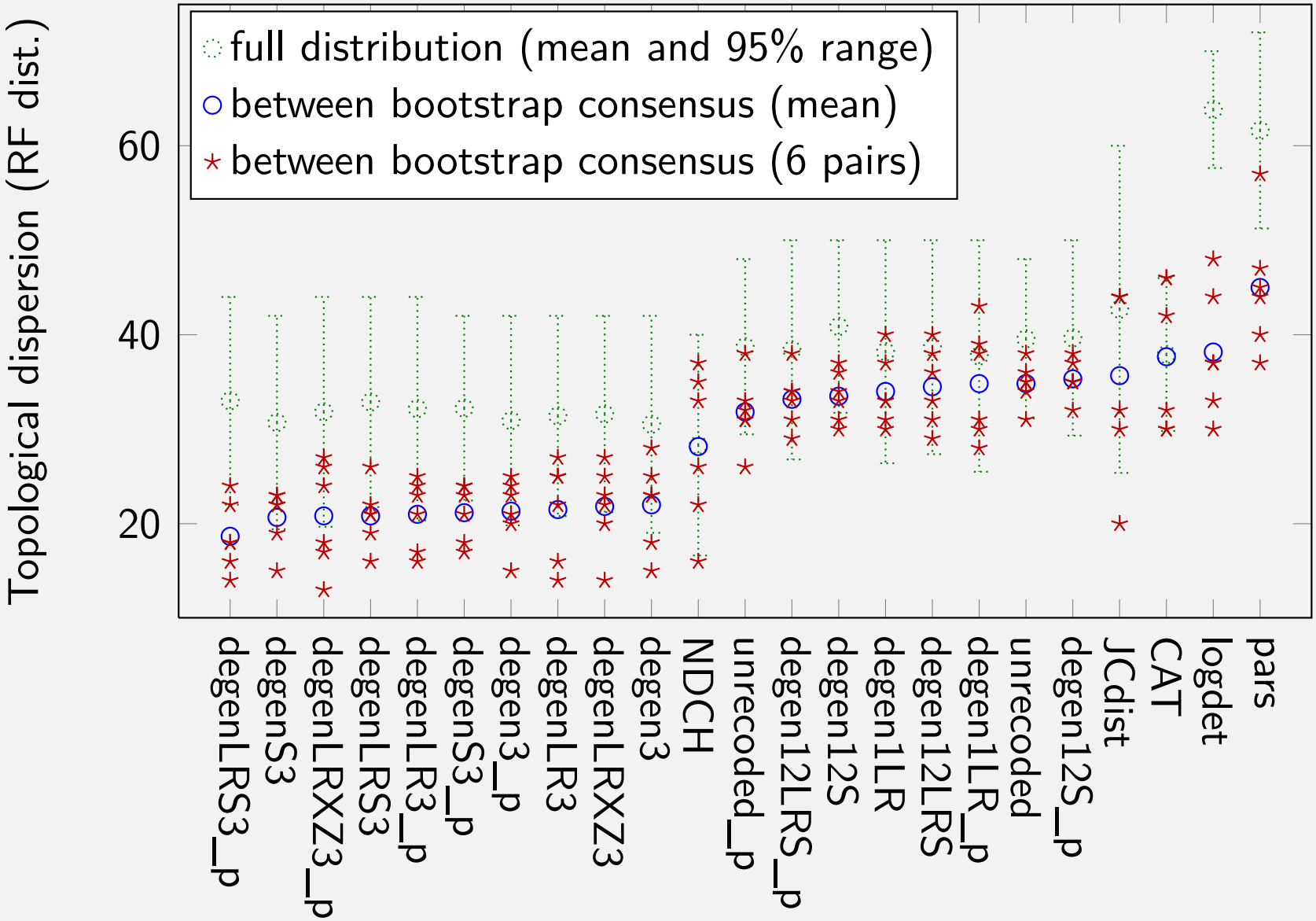
Coherence measure



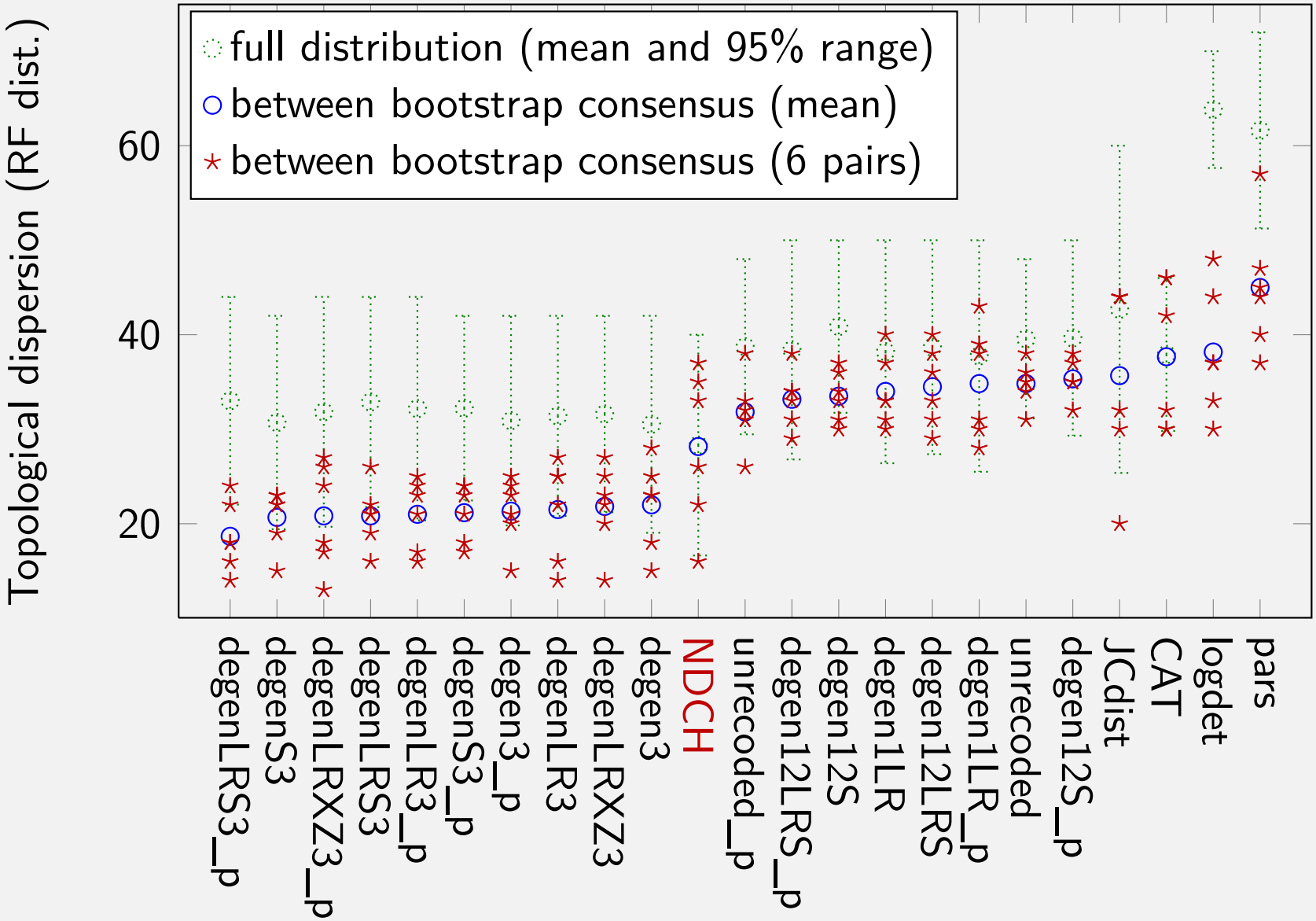
Between consensus:

- 6 Robinson-Foulds distances
- Coherence assessed using these distances and their average

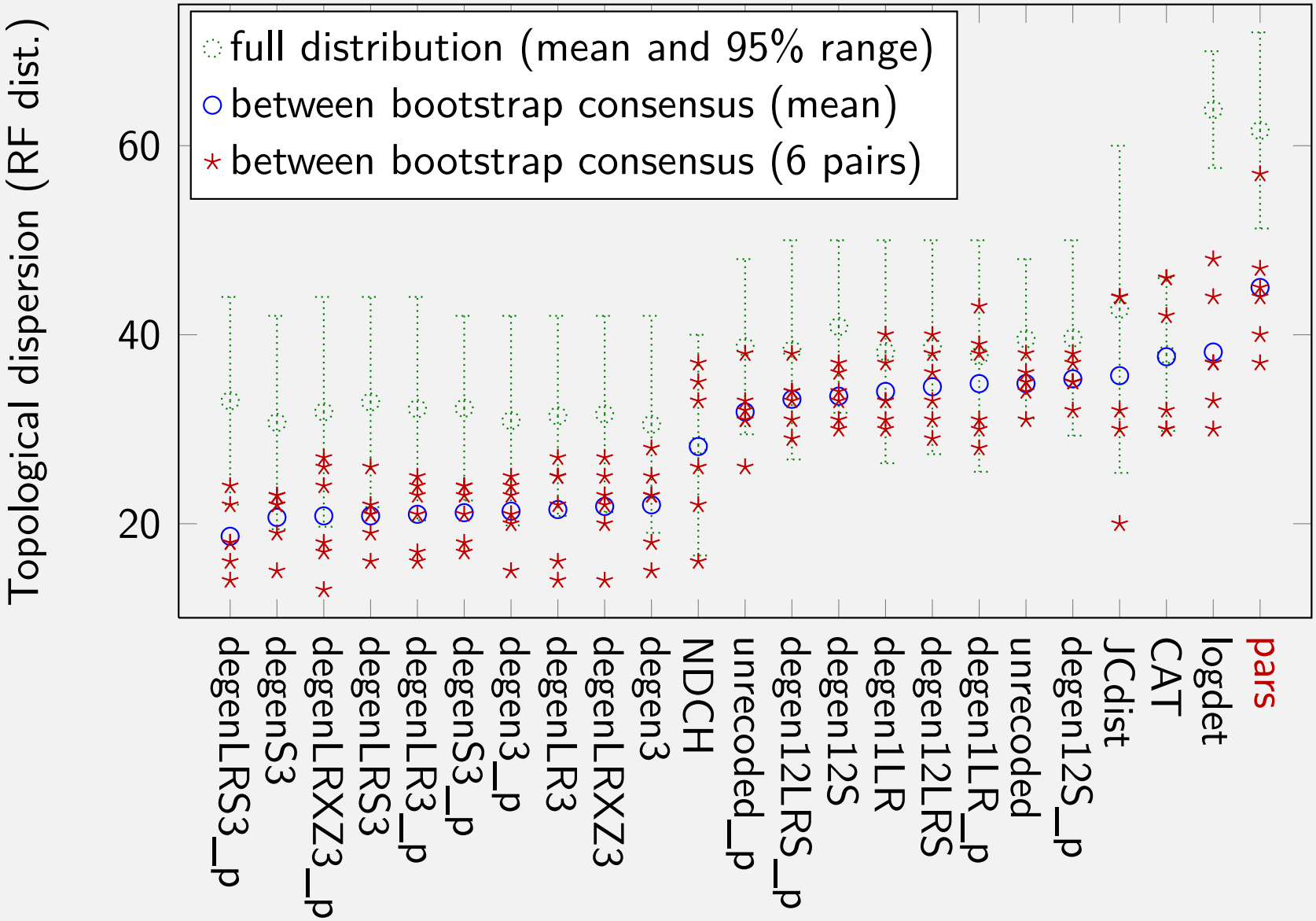
Results


full distribution (mean and 95% range)
between bootstrap consensus (mean)
between bootstrap consensus (6 pairs)
Topological dispersion (RF dist.)
20
40
60
degenLRS3_p
degenS3
degenLRXZ3_p
degenLRS3
degenLR3_p
degenS3_p
degen3_p
degenLR3
degenLRXZ3
degen3
NDCH
unrecoded_p
degen12LRS_p
degen12S
degen1LR
degen12LRS
degen1LR_p
unrecoded
degen12S_p
JCdist
CAT
logdet
pars

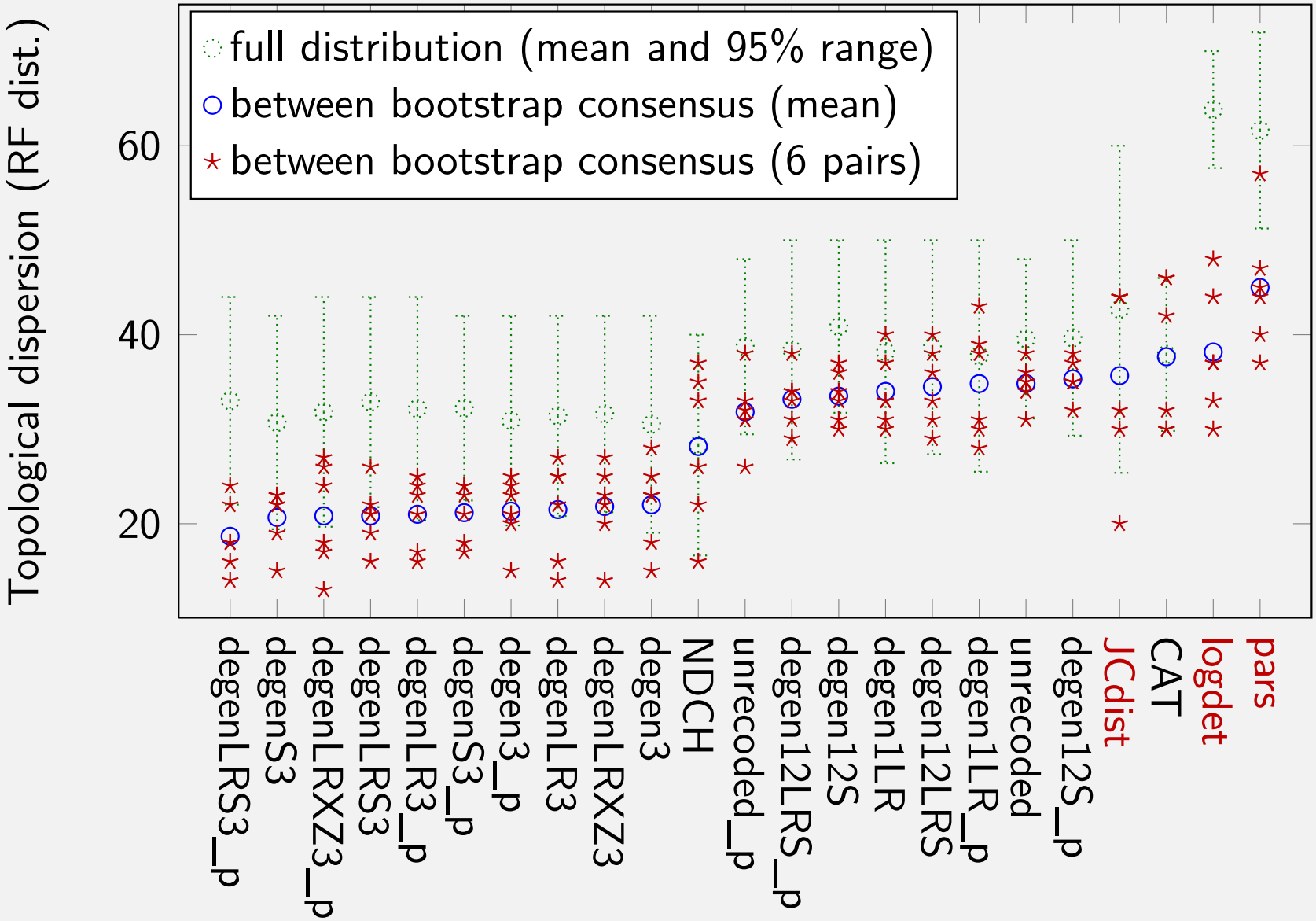
Results


full distribution (mean and 95% range)
between bootstrap consensus (mean)
between bootstrap consensus (6 pairs)
Topological dispersion (RF dist.)
20
40
60
degenLRS3_p
degenS3
degenLRXZ3_p
degenLRS3
degenLR3_p
degenS3_p
degen3_p
degenLR3
degenLRXZ3
degen3
NDCH
unrecoded_p
degen12LRS_p
degen12S
degen1LR
degen12LRS
degen1LR_p
unrecoded
degen12S_p
JCdist
CAT
logdet
pars

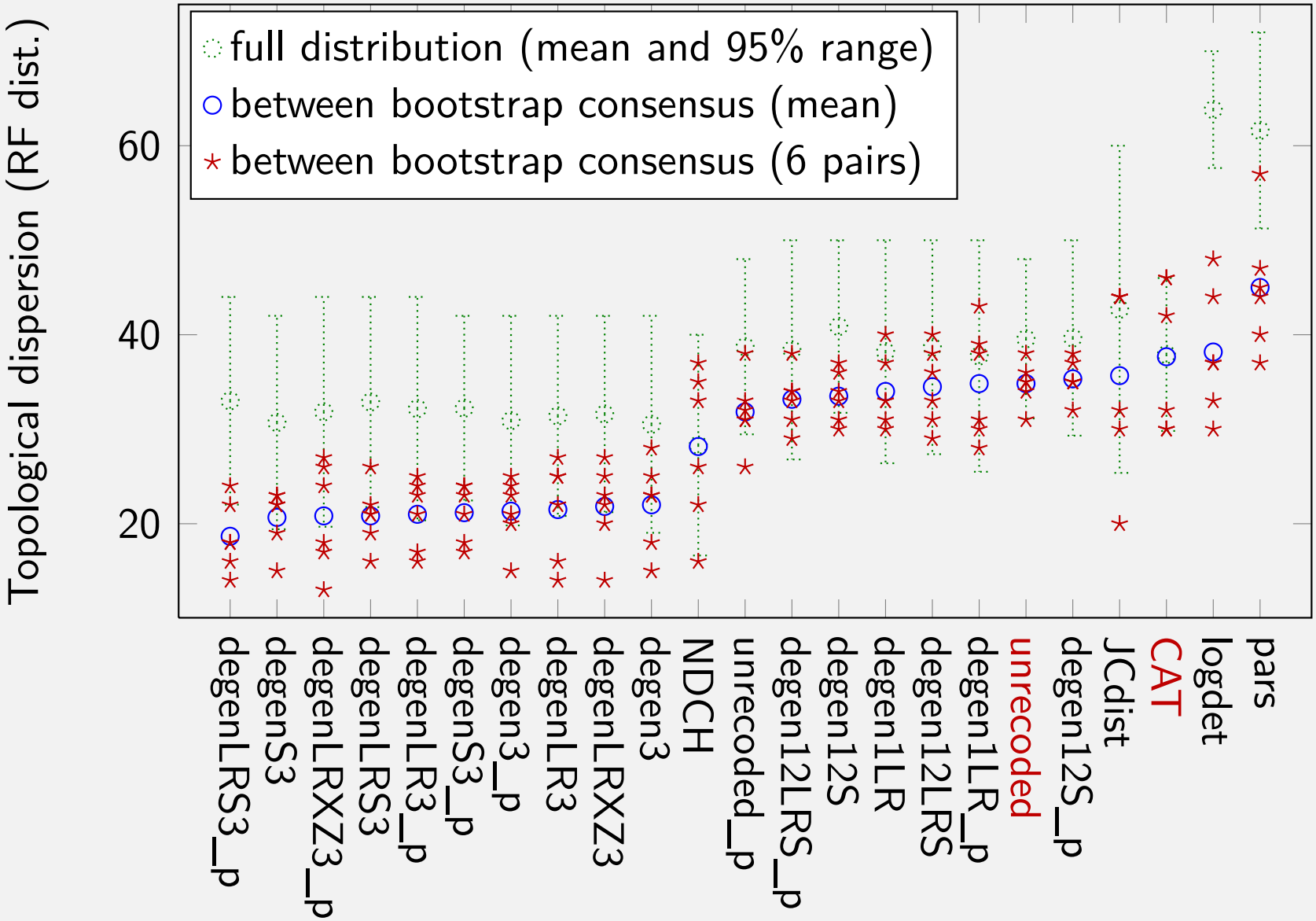
Results


full distribution (mean and 95% range)
between bootstrap consensus (mean)
between bootstrap consensus (6 pairs)
Topological dispersion (RF dist.)
20
40
60
degenLRS3_p
degenS3
degenLRXZ3_p
degenLRS3
degenLR3_p
degenS3_p
degen3_p
degenLR3
degenLRXZ3
degen3
NDCH
unrecoded_p
degen12LRS_p
degen12S
degen1LR
degen12LRS
degen1LR_p
unrecoded
degen12S_p
JCdist
CAT
logdet
pars

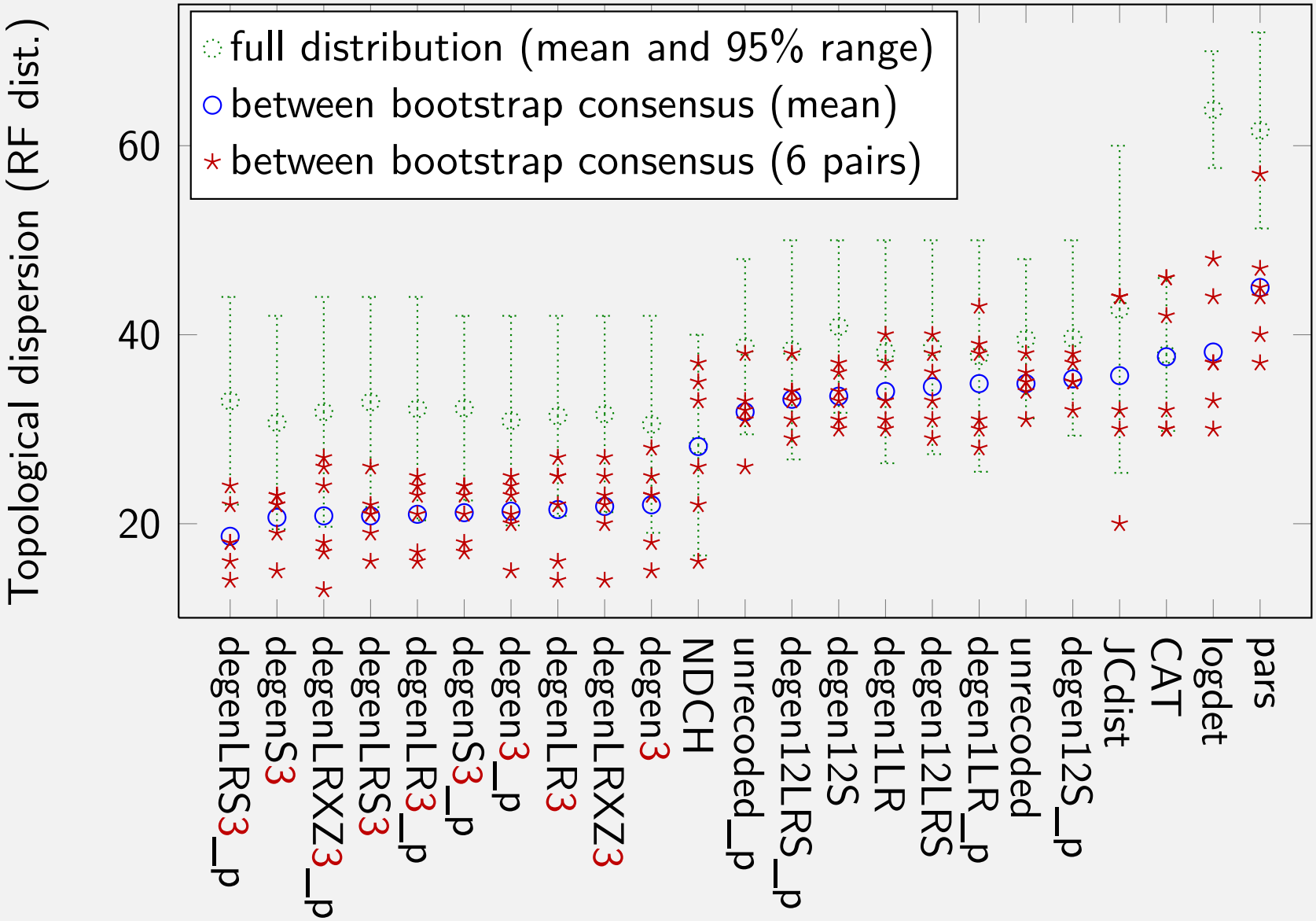
Results


full distribution (mean and 95% range)
between bootstrap consensus (mean)
between bootstrap consensus (6 pairs)
Topological dispersion (RF dist.)
20
40
60
degenLRS3_p
degenS3
degenLRXZ3_p
degenLRS3
degenLR3_p
degenS3_p
degen3_p
degenLR3
degenLRXZ3
degen3
NDCH
unrecoded_p
degen12LRS_p
degen12S
degen1LR
degen12LRS
degen1LR_p
unrecoded
degen12S_p
JCdist
CAT
logdet
pars

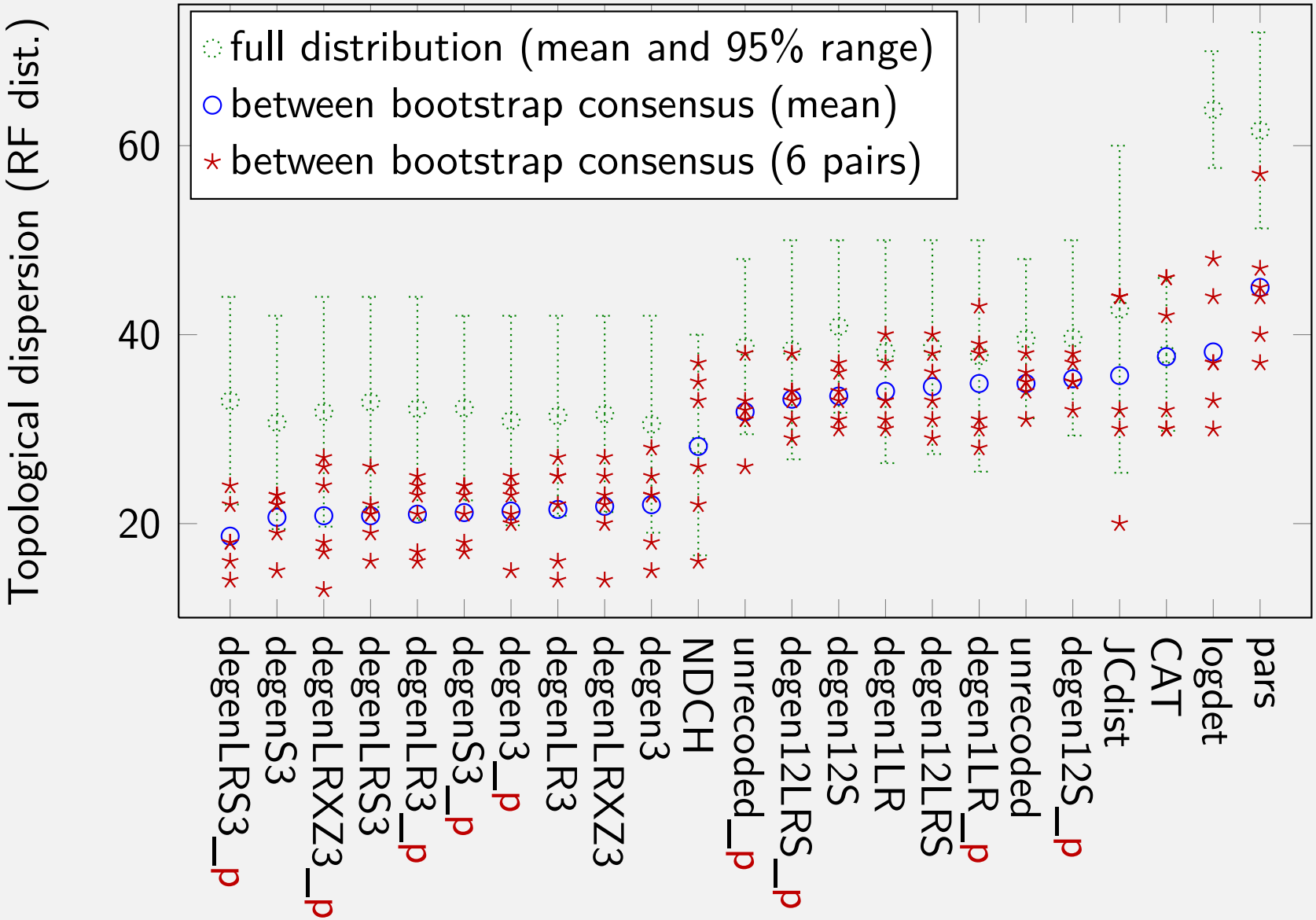
Results


full distribution (mean and 95% range)
between bootstrap consensus (mean)
between bootstrap consensus (6 pairs)
Topological dispersion (RF dist.)
20
40
60
degenLRS3_p
degenS3
degenLRXZ3_p
degenLRS3
degenLR3_p
degenS3_p
degen3_p
degenLR3
degenLRXZ3
degen3
NDCH
unrecoded_p
degen12LRS_p
degen12S
degen1LR
degen12LRS
degen1LR_p
unrecoded
degen12S_p
JCdist
CAT
logdet
pars

Results


full distribution (mean and 95% range)
between bootstrap consensus (mean)
between bootstrap consensus (6 pairs)
Topological dispersion (RF dist.)
20
40
60
degenLRS3_p
degenS3
degenLRXZ3_p
degenLRS3
degenLR3_p
degenS3_p
degen3_p
degenLR3
degenLRXZ3
degen3
NDCH
unrecoded_p
degen12LRS_p
degen12S
degen1LR
degen12LRS
degen1LR_p
unrecoded
degen12S_p
JCdist
CAT
logdet
pars

Results


full distribution (mean and 95% range)
between bootstrap consensus (mean)
between bootstrap consensus (6 pairs)
Topological dispersion (RF dist.)
20
40
60
degenLRS3_p
degenS3
degenLRXZ3_p
degenLRS3
degenLR3_p
degenS3_p
degen3_p
degenLR3
degenLRXZ3
degen3
NDCH
unrecoded_p
degen12LRS_p
degen12S
degen1LR
degen12LRS
degen1LR_p
unrecoded
degen12S_p
JCdist
CAT
logdet
pars

The conclusions

- To some extent, coherence seems correlated to phylogenetic accuracy.

The conclusions

- To some extent, coherence seems correlated to phylogenetic accuracy.
- But the discriminative power of the Robinson-Foulds based coherence measure is low

The conclusions

- To some extent, coherence seems correlated to phylogenetic accuracy.
- But the discriminative power of the Robinson-Foulds based coherence measure is low and this measure may be subject to biases related to the degree of resolution that a strategy of analysis typically produces.

The conclusions

- To some extent, coherence seems correlated to phylogenetic accuracy.
- But the discriminative power of the Robinson-Foulds based coherence measure is low and this measure may be subject to biases related to the degree of resolution that a strategy of analysis typically produces. (Is this a bug or a feature?)

The conclusions

- To some extent, coherence seems correlated to phylogenetic accuracy.
- But the discriminative power of the Robinson-Foulds based coherence measure is low and this measure may be subject to biases related to the degree of resolution that a strategy of analysis typically produces. (Is this a bug or a feature?)
- Could better measures of coherence be designed?

The conclusions

- To some extent, coherence seems correlated to phylogenetic accuracy.
- But the discriminative power of the Robinson-Foulds based coherence measure is low and this measure may be subject to biases related to the degree of resolution that a strategy of analysis typically produces. (Is this a bug or a feature?)
- Could better measures of coherence be designed?
- Including in the pannel a priori poorly performing analysis strategies may help to detect false positives.

Thanks for your attention

- This work was supported by a Fundação para a Ciência e a Tecnologia (FCT, Portugal) grant to Cymon J. Cox, Centro de Ciencias do Mar (CCMAR) - CIMAR-Lab. Assoc., (PTDC/BIA-BCM/099565/2008).

- I'm currently looking for a job, so if you think I can be useful in your lab, feel free to contact me.
- Contact: `blaise.li@normalesup.org`

This work relied heavily on the `Python` programming language.

This presentation was made using the excellent `beamer` and `TikZ/pdf` LaTeX packages.