From Easy to Hopeless - Predicting the Difficulty of Phylogenetic Analyses

Julia Haag[1*], Dimitri Höhler[1], Ben Bettisworth[1] and Alexandros Stamatakis[1,2]

[1]Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies, 69118 Heidelberg, Germany.

[2]Institute for Theoretical Informatics, Karlsruhe Insititute of Technology, 76131 Karlsruhe, Germany.

*Corresponding author(s). E-mail(s): julia.haag@h-its.org;

Abstract

Phylogenetic analyses under the Maximum Likelihood model are time and resource intensive. To adequately capture the vastness of tree space, one needs to infer multiple independent trees. On some datasets, multiple tree inferences converge to similar tree topologies, on others to multiple, topologically highly distinct yet statistically indistinguishable topologies. At present, no method exists to quantify and predict this behavior. We introduce a method to quantify the degree of difficulty for analyzing a dataset and present Pythia, a Random Forest Regressor that accurately predicts this difficulty. Pythia predicts the degree of difficulty of analyzing a dataset *prior* to initiating Maximum Likelihood based tree inferences. Pythia can be used to increase user awareness with respect to the amount of signal and uncertainty to be expected in phylogenetic analyses, and hence inform an appropriate (post-)analysis setup. Further, it can be used to select appropriate search algorithms for easy-, intermediate-, and hard-to-analyze datasets.

Keywords: Phylogenetics, Maximum Likelihood, Machine Learning, Random Forest Regression

1 Introduction

The goal of a phylogenetic inference is to find the phylogenetic tree that best explains the given biological sequence data. Since the number of possible tree topologies grows super-exponentially with the number of taxa, one cannot compute and score every possible tree topology. Instead, one deploys tree inference heuristics that explore the tree space to find a tree with a ‘good’ score, for example under the Maximum Likelihood (ML) criterion [1]. However, these heuristics, do not guarantee that the tree inference will converge to the *globally* optimal tree. Therefore, under ML, one typically infers multiple trees and subsequently summarizes the inferred, *locally* optimal trees via a consensus tree. One can observe that for some datasets, all individual, independent ML tree searches converge to topologically similar trees. This suggests that the likelihood surface of such datasets exhibits a single likelihood peak, yielding the dataset easy to analyze. For other datasets, one observes that the independent tree inferences converge to multiple topologically distinct, yet, with respect to their ML score, statistically indistinguishable, locally optimal trees. These datasets are hence difficult to analyze, and we say that they exhibit a rugged likelihood surface. This diverse behavior of phylogenetic tree searches has already been reported in several publications [2, 3, 6]. In general, the more tree inferences we perform, the better our understanding of the dataset's behavior and coverage of the respective tree space will be. However, under ML, inferring a single tree can already require multiple hours or even days of CPU time. In order to save time and resources, an optimal analysis setup will perform as few tree inferences as necessary. For easy-to-analyze datasets with a single likelihood peak, we require fewer and less involved tree search heuristics and bootstrap replicate searches to adequately sample the tree space and obtain stable support values, as opposed to difficult-to-analyze datasets with rugged likelihood surfaces. To the best of our knowledge, and despite anecdotal reports on the behavior of difficult datasets, there does not yet exist a *quantifiable* definition of dataset difficulty.

Here, we initially introduce such a difficulty quantification based on the result of 100 ML tree inferences per dataset. We then show that this quantification adequately represents the behavior of the ML searches on the dataset. Since executing 100 ML tree searches is computationally prohibitive in general, we train a Random Forest Regressor [23] that can predict the difficulty of a given Multiple Sequence Alignment (MSA) that is exclusively based on MSA attributes and some fast and thus substantially less expensive parsimony-based tree inferences [9, 10]. Our difficulty predictor Pythia predicts the difficulty of a dataset on a scale ranging between 0.0 (easy) to 1.0 (difficult). We attain a high prediction accuracy, with a mean absolute prediction error (MAE) of 0.09 and a mean absolute percentage error (MAPE) of 2.9 %. Computing the prediction features and predicting the difficulty is on average approximately five times faster than a single ML tree inference.

Using Pythia *prior* to the actual phylogenetic analyses allows for faster analyses, as the number of required tree inferences can be carefully considered before performing costly tree inferences. Note that the predicted difficulty does not directly predict the number of tree inferences required to sufficiently sample the tree space, as this number also depends on the implemented tree inference heuristic. We therefore suggest that Pythia should be included in ML phylogenetic analysis pipelines. The difficulty prediction we present here focuses on ML analyses. However, we also provide anecdotal evidence that the predicted difficulty correlates with the convergence diagnostics of Markov chain Monte Carlo (MCMC) based Bayesian phylogenetic inference. Besides informing the computational setup of phylogenetic analyses, Pythia can also potentially adjust user expectations regarding the stability and absolute values of bootstrap support as well as related support measures. For instance, the perhaps most common and recurrent user inquiry on the RAxML Google user support group concerns possible reasons for often unexpected and disappointingly low support values.

Pythia is available as open source software libraries in C and Python. Both libraries include the trained Random Forest Regressor and the computation of the required prediction features. The C library CPythia is an addition to the COre RAXml LIBrary (Coraxlib) [15] and is available at https://github.com/tschuelia/CPythia. Additionally, we provide PyPythia, a lightweight, stand-alone Python library, including a respective command line interface. PyPythia is available at https://github.com/tschuelia/PyPythia. Finally, by using the phylogenetic tree data that is being collected by our dynamically growing RAxML Grove [4] database, we regularly retrain Pythia and update the predictor in both libraries.

2 Results

2.1 Difficulty Prediction Accuracy

Our training data contains 3250 empirical MSAs obtained from TreeBASE [5]. We divide this training data into a *training set* (80 %) and a *test set* (20 %). The training set is used for training the predictor and the test set is exclusively used for evaluating the trained predictor. Pythia predicts the degree of difficulty on a scale between 0.0 to 1.0. A value of 1.0 indicates a difficult (hopeless) MSA with a rugged tree space. We expect such an MSA to exhibit multiple, statistically indistinguishable locally optimal yet topologically highly distinct trees. In contrast, we expect an MSA with a value of 0.0 to be easy to analyze by requiring only few independent tree searches. Pythia attains a mean absolute error (MAE) of 0.09. This corresponds to a mean average percentage error (MAPE) of 2.9 %. The mean squared error

(MSE) is 0.02 and the R^2 score is 0.79. Supplementary Figures S5a and S5b show the distribution of prediction errors for the training data. When analyzing the prediction error, we notice that Pythia tends to overestimate the difficulty of MSAs with a difficulty ≤ 0.3 and underestimate the difficulty for MSAs with a difficulty > 0.3 (Supplementary Figure S4). We suspect that this is caused by an uneven distribution of difficulties in the training data. Our training data contains substantially more 'easy' MSAs than difficult MSAs: for approximately 60 % of MSAs the assigned difficulty is ≤ 0.3 and only about 10 % have a difficulty ≥ 0.7 (Supplementary Figure S2). We expect this effect to cancel out in the long run as we regularly retrain Pythia.

As stated above, the predicted difficulty is not intended to predict the number of tree inferences required to sufficiently sample the tree space, as this number also depends on the implemented tree inference heuristic. However, to demonstrate the predictive power of Pythia, we analyze the impact of the prediction on the quality of tree inferences with RAxML-NG [7]. To this end, we compare the log-likelihoods obtained from 100 independent RAxML-NG tree searches ($LnLs_{100}$) to the log-likelihoods of $predicted\_difficulty \cdot 100$ tree searches ($LnLs_{\text{pred}}$) for all MSAs in our training data. We compare the respective best found log-likelihoods LnL^*_{100} and LnL^*_{pred}, as well as the average log-likelihoods $\overline{LnL}_{100}$ and $\overline{LnL}_{\text{pred}}$.

For 81 % of the MSAs, the best found log-likelihoods LnL^*_{100} and LnL^*_{pred} are identical. For the remaining 19 % of MSAs, LnL^*_{pred} is on average $\ll 0.01\,\%$ worse than LnL^*_{100}. The average log-likelihoods $\overline{LnL}_{100}$ and $\overline{LnL}_{\text{pred}}$ deviate on average by 0.01 % only.

2.2 Runtime of Feature Computation

Computing the selected set of prediction features takes on average 5 ± 31 s ($\mu \pm \sigma$) with a median runtime of 1 s. For our training data, this corresponds to a runtime of $21.5 \pm 88.6\,\%$ relative to the runtime for inferring a single ML tree using RAxML-NG. The median is 6.8 %. The high average compared to the median, and the large spread, are due to the fact that the runtime of computing the prediction features predominantly depends on the size of the MSA. The larger the MSA, the faster the feature computation is compared to a single ML tree inference. Supplementary Figure S3 depicts this correlation. For benchmarking the runtimes of the feature computation, we used the implementation in our Python library. When running a subsequent ML tree inference, the runtime overhead induced by the prediction can be amortized by passing the inferred maximum parsimony trees as starting trees to the ML inference tool (e.g, RAxML-NG). Instead of re-computing parsimony starting trees, the RAxML-NG simply initiates its tree searches on the provided parsimony starting trees.

2.3 Feature Importance

Pythia predicts how difficult an MSA will be to analyze based on eight features. Four of these are direct attributes of the MSA: the sites-over-taxa ratio, the patterns-over-taxa ratio, the percentage of gaps, and the percentage of invariant sites. Two features quantify the amount of information in the MSA: the Shannon entropy [8] and the Bollback multinomial [21]. Two additional features are based on the rapid parsimony tree inferences: we infer 100 parsimony trees via a randomized stepwise addition order procedure and compute their average pair-wise topological distances using the Robinson-Foulds distance metric (RF-Distance) [11], as well as the proportion of unique topologies in this set of 100 parsimony trees. In Section 4.5 we provide a more detailed description of the features as well as a rationale for selecting these features. Table 1 shows the prediction importances of the eight features upon which the difficulty prediction is based. We use the permutation importance [16] for computing feature importances. As the table shows, the difficulty prediction heavily relies on the average RF-Distance and the proportion of unique topologies among the inferred parsimony trees. This is expected, as our difficulty definition under ML reflects the ruggedness of the tree space.

2.4 MCMC Convergence Prediction

The features we use to predict the difficulty of an MSA are independent of the inference method used for the subsequent analyses. However, as we describe in Section 4.1, our difficulty quantification is based on 100 tree inferences using RAxML-NG which implements the ML method. Therefore,

Feature	Impurity importance
% Unique topologies parsimony trees	42.9 %
RF-Distance parsimony trees	33.2 %
Entropy	17.0 %
Patterns-over-taxa	13.6 %
% Gaps	2.5 %
Bollback	2.3 %
Sites-over-taxa	1.5 %
% Invariant	0.6 %

Table 1 Importance of the subset of features we use to train Pythia.

our predictions might be biased towards ML analyses and potentially not describing the ruggedness of the tree space in a model-independent manner. To assess if our predictions can be generalized, we anecdotally compare our difficulty prediction to convergence diagnostics of MCMC based phylogenetic analyses. For three DNA MSAs (D27 [17], D125 [19], and D354 [20]) we perform MCMC analysis using MrBayes [22]. We run four chains for 10 million generations using the general time reversible (GTR) model with four Γ rate categories to account for among site rate heterogeneity. MrBayes reports the average standard deviation of split frequencies (ASDSF) as a convergence diagnostic metric and suggests executing additional generations as long as the ASDSF is ≥ 0.01. D125 is an easy dataset with a presumable clear, single likelihood peak. As expected, the assigned difficulty is low ($\ll 0.1$) and MrBayes appears to converge: the ASDSF value drops below 0.01 after 150 000 generations and is $\ll 0.01$ after 10 million generations. D27 exhibits at least two distinct likelihood peaks, suggesting that the MSA is rather difficult to analyze [18]. The difficulty according to our definition is 0.45 and after 10 million generations MrBayes reports an ASDSF of 0.011, indicating that the MCMC did not converge to a single local optimum. D354 exhibits a rugged likelihood surface [20], so we expect a high difficulty and no convergence. The assigned difficulty for D354 is 0.6 and after 10 million generations the ASDSF is 0.007. According to MrBayes this suggests convergence and adding more generations should improve the ASDSF. However, we observe that the ASDSF did not improve during the last 2 million generations, and adding more generations did not further improve the ASDSF.

2.5 Bootstrap Support Values

As already mentioned, the perhaps most common question on the RAxML user support Google group is related to disappointingly low support values. Therefore, we anecdotally test the capability of Pythia to adjust user expectations regarding the stability of bootstrap support values and related support measures by means of three example MSAs. We use the same MSAs for the same reasons as for the anecdotal MCMC convergence prediction conducted above: D27, D125, and D354. For each MSA, we run RAxML-NG using its `--all` execution mode. This mode infers 20 ML trees for the MSA, infers 1000 bootstrap replicate trees, and draws support values on the tree with the highest log-likelihood (best-known tree). To anecdotally show the correlation between the difficulty prediction value and the bootstrap support values, we compute the average and standard deviation $\mu \pm \sigma$ of bootstrap support values on the respective best-known trees. As stated above, D125 is an easy dataset exhibiting a clear signal with an assigned difficulty $\ll 0.1$. This is reflected by the bootstrap support values: $\mu \pm \sigma = 97.64 \pm 8.38\,\%$. The assigned difficulty for D27 is 0.45 and RAxML-NG reports the bootstrap support values as $\mu \pm \sigma = 51.5 \pm 29.02\,\%$. Dataset D354 is the most difficulty among the three example MSAs with a predicted difficulty of 0.6. Hence, the bootstrap support values are the lowest among the three MSAs with $\mu \pm \sigma = 43.41 \pm 32.48\,\%$. Since computing bootstrap support values is time and resource intensive, we decided against computing the bootstrap support for all MSAs in our comprehensive training data. Instead, we only show anecdotal evidence for three representative MSAs in this paper.

3 Discussion

Predicting the difficulty of MSAs to gain a priori insights into the expected behavior of phylogenetic tree searches and the shape of the likelihood surface constitutes a vital step towards faster phylogenetic inference and a more targeted setup of the computational analyses and post-analyses.

Our difficulty prediction allows for careful consideration of the number of tree inference required to sufficiently sample tree space *prior* to ML analyses. Especially for easy MSAs, this has the potential to save valuable time and resources. In this paper, we presented a quantifiable definition of difficulty for MSAs and showed that this definition adequately represents the tree space of the dataset. Using this definition, we trained Pythia, a Random Forest Regressor, to predict the difficulty on a scale ranging between 0.0 to 1.0. We showed that Pythia achieves a high prediction accuracy. We further showed that the runtime to compute the prediction features is on average only approximately one fifth of the runtime required for inferring a single ML tree with RAxML-NG. The more taxa and sites the MSA has, the faster the feature computation is relative to a single ML tree inference, making Pythia especially valuable for phylogenetic analyses on MSAs with many sites and taxa. We conclude that predicting the difficulty of an MSA prior to any tree inference allows for faster analyses, informing user expectations regarding the stability of the inferred tree, and Pythia should be included in ML phylogenetic inference pipelines. Using our dynamically growing RAxML Grove database, we will perpetually enlarge and diversify our training data, and regularly retrain Pythia.

While the difficulty prediction is targeted towards ML phylogenetic analyses, we anecdotally showed that the predicted difficulty based on ML phylogenetic inference correlates with convergence diagnostics of MCMC methods. Investigating this correlation more systematically will be subject of future work.

Analogous, we anecdotally show that Pythia can also potentially inform about the stability and absolute values of bootstrap support. Confirming this observation on a broader set of MSAs remains future work.

Another avenue for future work is to implement a difficulty-aware tree inference heuristic. Depending on the difficulty of the MSA, we can, for example, apply different heuristic search strategies. For instance, on easy MSAs it might be sufficient to explore the tree space via a less thorough exploration strategy, that is, by only using Nearest-Neighbor-Interchange (NNI) moves. In comparison to Subtree Pruning and Regrafting (SPR) moves, this reduces the tree topology search complexity from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$ [27].

Further potential applications of Pythia include, for instance, the assembly of benchmark dataset collections for testing novel phylogenetic models and tools which cover a broad and representative difficulty range. Pythia can also serve as a criterion during the empirical dataset assembly process. For instance, additional sequences data can be added to yield the dataset easier to analyze.

4 Materials and Methods

We formulate the difficulty prediction challenge as a supervised regression task. The goal is to predict the difficulty on a scale ranging between 0.0 (easy) to 1.0 (difficult). We face two main challenges: (i) obtaining a sufficiently large set of MSAs to train Pythia on, ideally consisting of empirical MSAs, and (ii) obtaining ground-truth difficulties that represent the actual difficulty of the training data. In the following, we present how we obtain the training data and assign ground-truth difficulties. We further present our trained regression model, motivate how we select the features we use for difficulty prediction, and finally present our heuristic for regularly retraining the regression model.

4.1 Quantification of Difficulty

In order to train a reliable difficulty predictor, we need a reliable ground-truth label for each training datum. To obtain such labels, we require a quantifiable difficulty definition. To stringently quantify the difficulty of an MSA, we would have to explore the entire tree space. Since this is computationally not feasible, we need to rely on a heuristic definition. Our heuristic to quantify the difficulty is based on 100 ML tree inferences. In our analyses, we use RAxML-NG. First, we infer $N_{\text{all}} = 100$ ML trees and compute the average pairwise relative RF-Distance between all trees (RF_{all}), as well as the number of unique topologies among the 100 inferred trees (N^*_{all}). We determine the best tree among the 100 inferred trees according to the log-likelihood, and compare all trees to this best tree using statistical significance tests. We assign trees that are not significantly worse than the best tree to a so-called *plausible tree set*. In our analyses, we use the statistical significance tests as implemented in the IQ-TREE software package [12]. Due to the continuing debate about the most appropriate significance test for comparing phylogenetic trees, we use the approach suggested by Morel *et al.* [6]: we only include trees that pass *all* significance tests in the plausible tree set. We further refer to the number of trees in this plausible tree set as N_{pl}. We compute the average pairwise relative RF-Distances between trees in the plausible tree set (RF_{pl}), as well as the number of unique topologies (N^*_{pl}). Finally, we compute the difficulty of the dataset based on the following formula:

$$\begin{aligned} \text{difficulty} = \frac{1}{5} \cdot \Bigg[& RF_{\text{all}} + RF_{\text{pl}} && (1) \\ & + \frac{N^*_{\text{all}}}{N_{\text{all}}} + \frac{N^*_{\text{pl}}}{N_{\text{pl}}} && (2) \\ & + \left(1 - \frac{N_{\text{pl}}}{N_{\text{all}}}\right) \Bigg] && (3) \end{aligned}$$

The reasoning for expression (1) is that if the RF-Distance is high, the tree space consists of multiple distinct, locally optimal tree topologies which characterize a dataset that is difficult to analyze. With expression (2) the reasoning is that the tree surface becomes more rugged, the more distinct locally optimal tree topologies the tree inference yields, and the more tree topologies are not significantly different from the best tree. Finally, the rationale for expression (3) is that, the more tree inferences yield a plausible tree, the more informative the MSA will be about the underlying evolutionary process and the easier the MSA will be to analyze. Each term is a value between 0.0 and 1.0, leading to an average value between 0.0 and 1.0 that quantifies the overall difficulty.

For each MSA in our training data, we compute the difficulty according to this definition. To this end, we implement a training data generation pipeline that automatically performs all required tree inferences, statistical tests, and computes the difficulty label alongside the features required for training Pythia. We implement this pipeline using the Snakemake workflow management system [13] and Python 3. The pipeline code is available at https://github.com/tschuelia/difficulty-prediction-training-data. In Supplementary Section 5 we list the software versions we use in the described pipeline.

4.2 Training Data

We train Pythia using empirical MSAs obtained from TreeBASE [5]. To date, our training data consists of 3250 MSAs, of which 74% contain DNA data and 26% contain Amino Acid (AA) data. The training data includes partitioned and unpartitioned MSAs. We provide a detailed overview of the training data in Supplementary Section 1. Figure 1 depicts the workflow for training data generation. For each MSA, we compute the difficulty according to the above definition as ground-truth label for supervised training using the training data generation pipeline. We compute the corresponding prediction features using our Python library. The set of prediction features and the corresponding difficulty label form our training data. For training the regression model, we split this training data into two sets: a training set and a test set. The training set comprises 80% of the training data and the test set the remaining 20%. The test set is exclusively used for evaluating the predictive power of the difficulty predictor. To ensure an even distribution of difficulty labels in the training and test sets, we deploy stratified sampling. Stratified sampling splits all difficulty labels into disjoint subsets and draws random samples from each subset independently. In principle, using simulated data would allow us to increase the size of the training data. However, since simulating data that is comparable to empirical data constitutes a challenging task [4], we decided against using additional simulated data.

4.3 Label Validation

Due to the lack of absolute ground-truth labels, we need to rely on the inferred difficulty labels. The motivation of the difficulty prediction is to limit the number of tree inferences required to sufficiently sample the tree space and obtain a representative consensus tree. To verify the label assignment for each dataset, we conduct the following analysis. We compare the consensus tree obtained from the plausible tree set constructed from all 100 ML tree inferences (*baseline tree*) to the consensus of the plausible trees we obtain

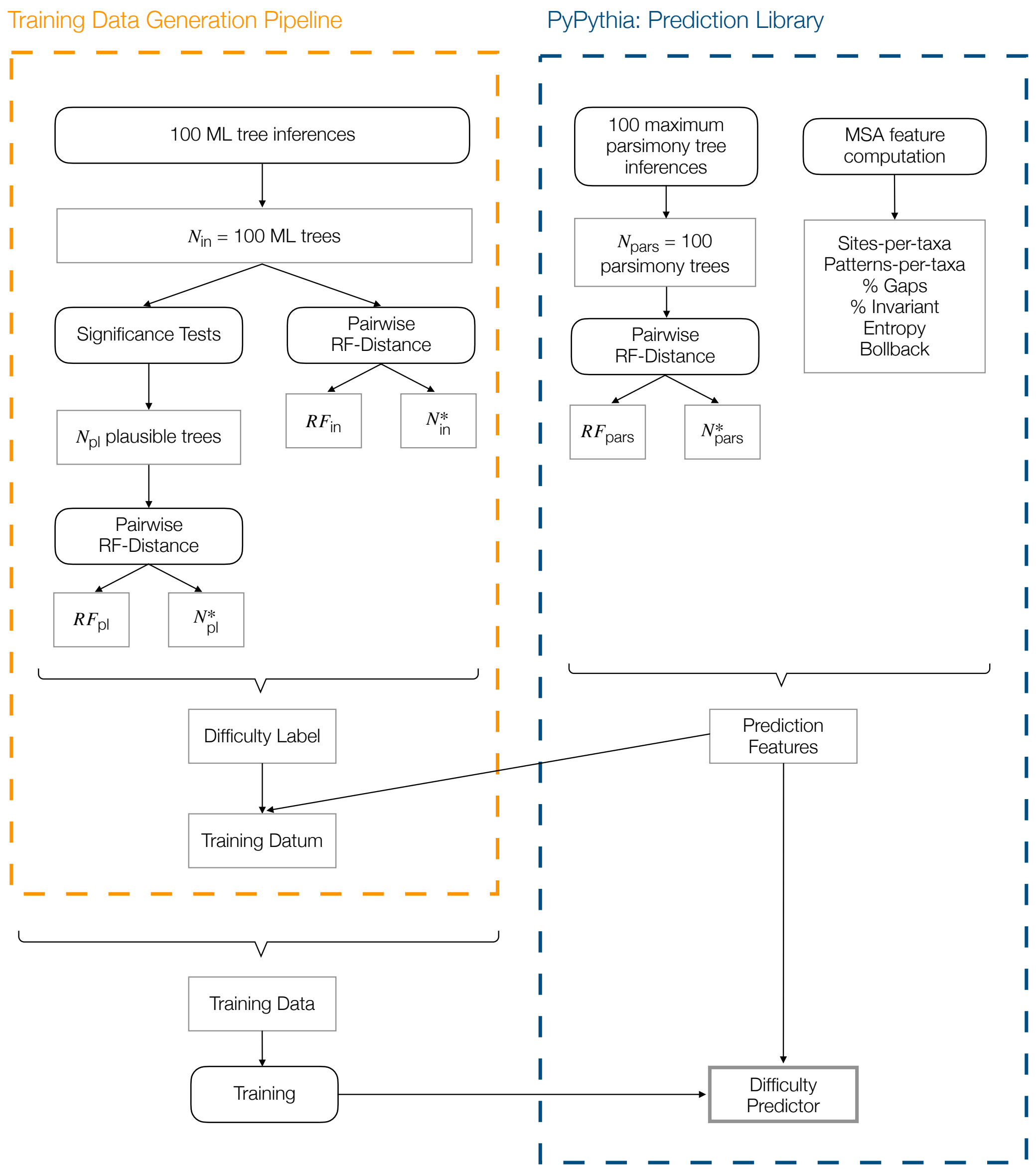

Fig. 1 Schematic depiction of the training data generation procedure. For each MSA, we compute the difficulty label based on our difficulty quantification using our training data generation pipeline (left dashed box). We further compute the prediction features using our Python prediction library PyPythia (right dashed box). Using the difficulty label and the corresponding prediction features for all MSAs in our training data, we train Pythia.

when inferring only 100 * *difficulty* trees (*prediction tree*). Note that for this analysis we use the *difficulty* we compute according to the above definition rather than using a predicted difficulty. We compare the topologies of the consensus trees using the RF-Distance. The RF-Distance between the *baseline tree* and the *prediction tree* is on average $9.6 \pm 15.8\,\%$. This noticeable topological difference suggests that either a) the difficulty labels do not sufficiently represent the tree search behavior of the dataset, or b) 100 tree inferences do not sufficiently sample the tree space. To determine the impact of b), we repeatedly sample 99 trees out of the 100 tree inferences and compute the consensus tree C_i of the respective plausible tree set. We then assess the average RF-Distance between all consensus trees C_i. For our training data, this RF-Distance is on average $8.1 \pm 14.5\,\%$. We conclude that mostly b) causes the high topological distances between the *baseline tree* and the *prediction tree*. In fact, a high RF-Distance between the consensus trees C_i for an MSA is correlated with its difficulty. The Spearman's rank correlation coefficient is 0.88 with a p-value of 0.0 ($\ll 10^{-300}$). Thus, the more difficult the MSA, the higher the topological distances between the consensus trees C_i will be. Our difficulty quantification accounts for the remaining $1.5 \pm 1.3\,\%$ topological differences. With this low RF-Distance, we conclude that our difficulty quantification is sufficiently accurate to capture the tree space complexity and the behavior of an MSA under ML based phylogenetic analysis.

4.4 Machine Learning and Evaluation

During our experiments, we trained distinct regression algorithms and compared their predictive power according to the R^2 score, the MSE, the MAE, and the MAPE. We divide the training data into two sets: a training set and a test set. We use the training set to train the prediction algorithms and the test set to evaluate the trained predictors on unseen data. We train multiple different regression models, namely Linear Regression, Lasso Regression [26], Random Forest Regression [23], Adaptive Boosting (AdaBoost) [24], and Support Vector Regression [25]. Random Forest Regression proves to be the most suitable Machine Learning algorithm for the task at hand, and outperforms all other tested regression models according to all our metrics. In Supplementary Information Section 3, we present the results for all trained regression models. Random Forest Regression is an ensemble method that averages over the predictions of multiple independently trained decision trees. To determine the optimal set of hyperparameters for the Random Forest Regression, we implemented a grid search that tests various combinations of hyperparameter values. For this grid search, we use an additional validation set, obtained by further subdividing the training set. We then perform hyperparameter optimization using this validation set. Our final difficulty predictor consists of 100 decision trees with a maximum depth of 10. To prevent overfitting, we set the minimum number of samples in a leaf node to 10 and the minimum number of samples required for a split to 20. Further, we train the individual decision trees on bootstrapped training data. We set the sample size for the bootstrapping to 75 % of the training data size. Note that this bootstrapping procedure samples the training data (features and corresponding label) and is not the phylogenetic bootstrap.

4.5 Feature Engineering

In our study, we analyze a plethora of distinct features based on the MSA, trees inferred under parsimony, and features based on a single ML tree inference using RAxML-NG. In order to decrease the runtime of our difficulty prediction, we analyze the runtime of computing each feature for all MSAs in our training data, as well as the importance of the feature for the prediction. Based on these results, we select the following feature subset:

- Sites-over-taxa ratio:

$$\frac{\#\text{ Sites}}{\#\text{ Taxa}} = \frac{\text{Number of alignment columns}}{\text{Number of taxa}}$$

- Patterns-over-taxa ratio:

$$\frac{\#\text{ Patterns}}{\#\text{ Taxa}} = \frac{\text{Number of unique sites}}{\text{Number of taxa}}$$

- % Invariant sites: Percentage of fully conserved sites.
- % Gaps: Proportion of gaps in the MSA.
- Entropy: Shannon Entropy [8] as average over all per-column/site entropies. See the supplementary information for a more detailed description.
- Bollback Multinomial: Multinomial test statistic according to Bollback [21]. See the supplementary information for a more detailed description.
- RF-Distance Parsimony Trees: RF-Distances between 100 trees inferred using parsimony.
- % Unique Topologies Parsimony Trees: Percentage of unique topologies among the 100 inferred parsimony trees.

In Supplementary Information Section 2 we present all features we analyzed in more detail, alongside the respective feature importance and runtime to justify the selection of features for this subset.

4.6 Retraining the Model

Using the anonymized MSAs that we continuously obtain during our RAxML Grove database updates, we perpetually enlarge the training data and retrain Pythia. Note that these MSAs are only available internally in RAxML Grove and are not publicly available. To limit the amount of resources required for retraining, we do not include every incoming, new MSA. We select MSAs based on a heuristic instead. At the time of writing, we select the set of new MSAs such that it diversifies the distribution of features in our training data. Algorithm 1 shows the heuristic for deciding whether to use a given MSA for retraining. For each feature f_i, we compute the respective histogram H_i on the training data using a predefined number of bins n_{bins}. Next, we compute the respective feature value for the given MSA and find the corresponding bin *hist_bin* in the histogram H_i. The goal is to attain an even distribution of features, that is, all histogram bins should have the same height $\bar{h}_i = 1/n_{\text{bins}}$. To quantify the deviation v_i from this even distribution, we divide this desired height $\bar{h}_i$ by the actual height h_i of *hist_bin*. The deviation v_i is negatively correlated to the number of samples in the corresponding histogram bin. For bins with fewer samples than the desired even distribution, the deviation is > 1. We sum the deviations v_i across all features. We use the given MSA for retraining if this sum is ≥ 14 or any of the deviations v_i is ≥ 4. The rationale for the first threshold is that in this case, on average, for each feature f_i the corresponding bin *hist_bin* has only half the desired height. The rationale for the second threshold is that in this case, one of the feature bins has only 1/4-th of the desired height.

For all MSAs we select, we compute the ground-truth label and prediction features as described in Section 4.2. Based on this enlarged training data, we retrain Pythia and automatically update the trained predictor in our Python and C libraries.

Algorithm 1 Heuristic for deciding whether to use a given MSA for retraining Pythia.

```
foreach feature $f_i$ do
    $H_i$ = histogram(training_data, $f_i$, $n_{\text{bins}}$)
    feat = compute_feature_value(MSA)
    $\bar{h}_i = 1/n_{\text{bins}}$
    hist_bin = find_bin_for_value($H_i$, feat)
    $h_i$ = height(hist_bin)
    $v_i = \bar{h}_i / h_i$
end
$V = \sum v_i$
analyze_msa = $V \geq 14$ or $max(v_i) \geq 4$
return analyze_msa
```

5 Code and Data availability

We provide Pythia as open source software libraries in C and Python. Both libraries include the trained Random Forest Regressor and the computation of the required prediction features. The C library CPythia is an addition to Coraxlib and is available at https://github.com/tschuelia/CPythia. Additionally, we provide PyPythia, a lightweight, stand-alone Python library, including a command line interface. PyPythia is available at https://github.com/tschuelia/PyPythia. The implemented pipeline to compute the prediction features and ground-truth difficulty labels for the training data is available at https://github.com/tschuelia/difficulty-prediction-training-data. This repository also contains the training data as parquet file.

6 Supplementary Information

Supplementary information is available online.

7 Acknowledgments and Funding

The authors gratefully acknowledge the support of the Klaus Tschira Foundation. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 764840.

References

[1] Yang, Z., Goldman, N. & Friday, A. Maximum Likelihood Trees from DNA Sequences: A Peculiar Statistical Estimation Problem. *Systematic Biology.* **44**, 384–399 (1995), https://doi.org/10.2307/2413599

[2] Stamatakis, A. Phylogenetic Search Algorithms for Maximum Likelihood. *Algorithms In Computational Molecular Biology: Techniques, Approaches And Applications.* 547–577 (2010), https://doi.org/10.1002/9780470892107.ch25

[3] Lakner, C., Mark, P., Huelsenbeck, J., Larget, B. & Ronquist, F. Efficiency of Markov Chain Monte Carlo Tree Proposals in Bayesian Phylogenetics. *Systematic Biology.* **57**, 86–103 (2008), https://doi.org/10.1080/10635150801886156

[4] Höhler, D., Pfeiffer, W., Ioannidis, V., Stockinger, H. & Stamatakis, A. RAxML Grove: an empirical phylogenetic tree database. *Bioinformatics.* **38**, 1741–1742 (2021), https://doi.org/10.1093/bioinformatics/btab863

[5] Piel, W., Donoghue, M., Sanderson, M. & Person, C. TreeBASE: A database of phylogenetic information. (2000)

[6] Morel, B., Barbera, P., Czech, L., Bettisworth, B., Hübner, L., Lutteropp, S., Serdari, D., Kostaki, E., Mamais, I., Kozlov, A., Pavlidis, P., Paraskevis, D. & Stamatakis, A. Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult. *Molecular Biology And Evolution.* **38**, 1777–1791 (2021), https://doi.org/10.1093/molbev/msaa314

[7] Kozlov, A., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics.* **35**, 4453–4455 (2019), https://doi.org/10.1093/bioinformatics/btz305

[8] Shannon, C. A mathematical theory of communication. *The Bell System Technical Journal.* **27**, 379–423 (1948), https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[9] Farris, J. Methods for Computing Wagner Trees. *Systematic Biology.* **19**, 83–92 (1970), https://doi.org/10.1093/sysbio/19.1.83

[10] Fitch, W. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Zoology.* **20**, 406–416 (1971), https://doi.org/10.2307/2412116

[11] Robinson, D. & Foulds, L. Comparison of phylogenetic trees. *Mathematical Biosciences.* **53**, 131–147 (1981), https://doi.org/10.1016/0025-5564(81)90043-2

[12] Minh, B., Schmidt, H., Chernomor, O., Schrempf, D., Woodhams, M., Haeseler, A. & Lanfear, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology And Evolution.* **37**, 1530–1534 (2020), https://doi.org/10.1093/molbev/msaa015

[13] Köster, J. & Rahmann, S. Snakemake – a scalable bioinformatics workflow engine. *Bioinformatics.* **28**, 2520–2522 (2012), https://doi.org/10.1093/bioinformatics/bts480

[14] Huelsenbeck, J. Performance of Phylogenetic Methods in Simulation. *Systematic Biology.* **44**, 17–48 (1995), https://doi.org/10.1093/sysbio/44.1.17

[15] COre RAXml LIBrary (Coraxlib), https://codeberg.org/Exelixis-Lab/coraxlib.

[16] Breiman, L. Random Forests. *Machine Learning.* **45**, 5–32 (2001), https://doi.org/10.1023/A:1010933404324

[17] Hedges, S., Moberg, K. & Maxson, L. Tetrapod phylogeny inferred from 18S and 28S ribosomal RNA sequences and a review of the evidence for amniote relationships. *Molecular Biology And Evolution.* **7**, 607–633 (1990), https://doi.org/10.1093/oxfordjournals.molbev.a040628

[18] Lakner, C., Mark, P., Huelsenbeck, J., Larget, B. & Ronquist, F. Efficiency of Markov Chain Monte Carlo Tree Proposals in Bayesian Phylogenetics. *Systematic Biology.* **57**, 86–103 (2008), https://doi.org/10.1080/10635150801886156

[19] Poulakakis, N. & A. Stamatakis Recapitulating the evolution of Afrotheria: 57 genes and rare genomic changes (RGCs) consolidate their history. *Systematics And Biodiversity*. **8**, 395–408 (2010), https://doi.org/10.1080/14772000.2010.484436

[20] Grimm, G., Renner, S., Stamatakis, A. & Hemleben, V. A Nuclear Ribosomal DNA Phylogeny of Acer Inferred with Maximum Likelihood, Splits Graphs, and Motif Analysis of 606 Sequences. *Evolutionary Bioinformatics*. **2** (2006), https://doi.org/10.1177/117693430600200014

[21] Bollback, J. Bayesian Model Adequacy and Choice in Phylogenetics. *Molecular Biology And Evolution*. **19**, 1171–1180 (2002), https://doi.org/10.1093/oxfordjournals.molbev.a004175

[22] Ronquist, F., Teslenko, M., Mark, P., Ayres, D., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. & Huelsenbeck, J. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology*. **61**, 539–542 (2012), https://doi.org/10.1093/sysbio/sys029

[23] Ho, T. Random decision forests. *Proceedings Of 3rd International Conference On Document Analysis And Recognition*. **1** 278–282 (1995), https://doi.org/10.1109/ICDAR.1995.598994

[24] Freund, Y. & Schapire, R. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal Of Computer And System Sciences*. **55**, 119–139 (1997), https://doi.org/10.1006/jcss.1997.1504

[25] Boser, B., Guyon, I. & Vapnik, V. A Training Algorithm for Optimal Margin Classifiers. (Association for Computing Machinery, 1992), https://doi.org/10.1145/130385.130401

[26] Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal Of The Royal Statistical Society. Series B (Methodological)*. **58**, 267–288 (1996)

[27] Heath, L. & Ramakrishnan, N. Problem Solving Handbook in Computational Biology and Bioinformatics. (Springer-Verlag, 2010), https://doi.org/10.1007/978-0-387-09760-2