Nonlinear models based on leaf architecture traits explain the variability of mesophyll conductance across plant species

Milad Rahimi-Majd[1,2], Alistair Leverett[3], Johannes Kromdijk[3,4], and Zoran Nikoloski[* 1,2]

[1]Bioinformatics Department, Institute of Biochemistry and Biology, University of Potsdam, 14476 Potsdam, Germany
[2]Systems Biology and Mathematical Modeling Group, Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam, Germany
[3]Department of Plant Sciences, University of Cambridge, Cambridge, Cambridgeshire, CB23EA, UK
[4]Carl R Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, IL61801, Illinois, USA

*Contact: nikoloski@mpimp-golm.mpg.de
Short title: Nonlinear models of mesophyll conductance

Abstract

Mesophyll conductance (g_m) describes the efficiency with which CO_2 moves from substomatal cavities to chloroplasts. Despite the stipulated importance of leaf architecture in affecting g_m, there remains a considerable ambiguity about how and whether anatomy influences g_m. This is, in part, because studies exploring the relationship between leaf architecture and g_m have often relied on simple linear or exponential models to identify correlations. Here, we employed non-linear machine learning models to more comprehensively assess the relationship between ten leaf architecture traits and g_m. These models achieved excellent predictability of g_m, which depended on the leaf architecture traits considered as predictors. Dissection of the importance of leaf architecture traits in the models indicated that cell wall thickness and chloroplast area exposed to internal airspace have a large impact on interspecific variation in g_m. Additionally, other leaf architecture traits, such as: leaf thickness, leaf density, and chloroplast thickness emerged as important predictors of g_m. We found significant differences in the predictability between models trained on different plant functional types (PFTs): those trained on woody species could predict g_m by anatomical traits on other woody PFTs, ferns, and C_3 herbaceous plants, whereas the converse did not hold in general. By moving beyond simple linear and exponential models, our analyses demonstrated that a larger suite of leaf architecture traits drive differences in g_m than has been previously acknowledged. These findings pave the way for modulating g_m by strategies that modify its leaf architecture determinants.

keywords: mesophyll conductance, leaf architecture traits, plant functional types, machine learning, non-linear regression models, impurity-based feature importance

1 Introduction

Mesophyll conductance, g_m, is a numerical measure of the rate of diffusion of CO_2 from the substomatal cavities to RuBisCO, the site of carboxylation in the chloroplasts. An increase in mesophyll conductance is thus expected to elevate the rate at which RuBisCO can fix CO_2, thereby decreasing the water and nitrogen costs for carbon acquisition and fixation. Therefore, understanding factors controlling g_m is considered important for increasing the availability of CO_2 at RuBisCO's site of carboxylation, with expected concomitant improvement in the rate of photosynthesis (Zhu et al., 2010).

Relatively few leaf anatomical traits have been linked to interspecific variation in g_m. Existing evidence has indicated that cell wall thickness, T_{cw}, and surface area of chloroplasts exposed to the intercellular airspaces per unit leaf area, S_c, are important determinants of g_m, as these traits negatively and positively correlate with g_m, respectively (Clemente-Moreno et al., 2019; Carriquí et al., 2020; Veromann-Jürgenson et al., 2020; Tosens et al., 2016; Veromann-Jürgenson et al., 2017 ; Gago et al., 2019 and references therein). However, there is still considerable ambiguity

regarding the extent to which T_{cw} and S_c affect g_m, as their predictive power can be weak or even nonsignificant. For example, (Xiong, 2023) found that neither of these anatomical traits correlated with g_m in C_3 crops, using simple linear regression models. Furthermore, other studies presenting regression analyses on data collected from the literature (Flexas et al., 2021; Knauer et al., 2022a) have generally yielded weak models based on the two aforementioned anatomical traits for different plant functional types (PFTs). Whilst it is true that some exceptional cases have shown very high predictive power, these are based on only very few data points (e.g. seven), and thus the generalizability of these models remains unexplored (Peguero-Pina et al., 2017; Carriquí et al., 2020). In addition, it remains unclear if other leaf architecture traits, besides T_{cw} and S_c, contribute to explaining variance of g_m.

The ambiguity surrounding the importance of anatomy is perhaps not surprising if one considers that g_m is a composite parameter that integrates the effects of multiple factors, including: cell wall, plasma membrane (via its permeability, affected by aquaporins), cytosol, chloroplast envelope and stroma (Evans, 2021). This problem is further exacerbated by the differences in g_m values obtained by different measuring approaches. As leaf development will often be governed by allometric scaling rules (John et al., 2013), and anatomical traits may have antagonistic and/or complex impacts on g_m, it is likely that simple models based on one or two explanatory variables may be insufficient to robustly capture the relationships between anatomy and g_m. However, to date, the majority of models have applied this approach, describing the relationship between anatomy and g_m have been based on single- and two-variable linear or exponential relationships.

Advances in machine learning approaches provide one suitable means to obtain data-driven insights in the determinants of g_m. Modern machine learning approaches can capture non-linear relationships, and comparisons of models built using different plant functional types (PFTs) can test the generalizability of the resulting models. Here, we used machine learning techniques to address four questions: (1) Can machine-learning approaches be used to improve the predictive power of models describing the relationship between anatomy and g_m? (2) Do S_c and T_{cw} emerge as important determinants of g_m when several leaf architecture traits are used as inputs into non-linear models? (3) Can these non-linear models identify other leaf architecture traits (besides S_c and T_{cw}) influence g_m? (4) Do the best fitting models vary between different PFTs, and are they generalizable?

To address these questions, we make use of the largest compendium of g_m values along with leaf cell architecture traits published to date, measured over different PFTs and species. These data allow us to also investigate and fully address the extent of generalizability of the developed non-linear models between different PFTs. Lastly, we show how exhaustive consideration of different combinations of predictors can help in characterizing the role of leaf cell architecture in the control of g_m, and, thereby, photosynthesis.

2 Results

2.1 Predictive performance of the random forest models within PFTs

To identify and analyze the relationships between leaf architecture traits and g_m, we used a recently published comprehensive data set (Knauer et al., 2022b) providing measurements of diverse leaf traits on the same set of plants. This is currently the largest available data set for g_m, collecting measurements from 563 peer-reviewed studies over 617 species partitioned into 13 major PFTs. To train random forest (RF) models, we then constructed all possible combinations of traits for the *global data set* (consisting of all PFTs) and for each of the individual PFTs, respectively. Further, we considered only those combinations with at least 50 samples, with no missing data, allowing us to avoid data imputation that may bias the findings (see section Data and preprocessing, for details).

Some of the random forest (RF) models (see section The model), assessed by cross-validation on the global data set, revealed excellent relationships between different combinations of leaf architecture traits and g_m across all PFTs and species (Fig. 1). The performance of the models, i.e. predictability, was assessed by the adjusted coefficient of determination, R^2_{adj}, that controls for the number of predictors, and the Pearson correlation coefficient, r, between the predicted and measured g_m values. We note that R^2_{adj} assesses the quantitative agreement, while r captures the qualitative agreement between the measured and predicted g_m values.

The model based on the combination of five anatomical traits, namely, T_{cw}, S_c, T_{leaf}, T_{chl}, and D_{leaf} (model 1 in Fig. 1), showed both quantitatively and qualitatively the best predictability ($R^2_{adj} = 0.63$ and $r = 0.90$). Combinations involving some of these five traits were included as predictors in seven of the ten models ranked high with respect to their predictability (i.e., models $2-5$, 7, 8, and 10 in Fig. 1). Furthermore, the model that considered T_{cw} and S_c (model 14 in Fig. 1), the model that considered T_{cw}, T_{leaf}, and D_{leaf} (model 3 in Fig. 1), and the one based on the combination of T_{cw}, S_c, T_{chl}, and D_{leaf} (model 4 in Fig. 1) were the best-performing among those trained on two to four anatomical traits as predictors.

The best-performing model ($R^2_{adj} = 0.55$ and $r = 0.89$) based on a combination of six traits included: T_{cw}, S_c, T_{leaf}, T_{chl}, D_{leaf}, and S_m, while the best-performing model ($R^2_{adj} = 0.49$ and $r = 0.88$) on seven traits included: ***LMA***, T_{mes}, T_{cw}, T_{chl}, S_m, S_c, and T_{leaf}. Interestingly, the best-performing model ($R^2_{adj} = 0.22$ and $r = 0.85$) with eight traits, namely: ***LMA***, T_{mes}, T_{cw}, T_{cyt}, T_{chl}, S_m, S_c, and T_{leaf} ($R^2_{adj} = 0.22$ and $r = 0.85$), was considerably weaker in comparison to the top performing models with fewer traits as predictors.

This raised the question of why the introduction of additional predictors did not result in a further increase in model performance. The considerably smaller number of models on six or more traits in comparison to the number of models based on five traits (i.e. 80 models with six traits, 19

models with seven, one model with eight, and no models with nine or ten traits in comparison to 157 models on five traits) seemed as a plausible explanation (see section Data and preprocessing). To address the concern about the data limitations for models that include more than five traits as predictors, we then considered a different data-splitting approach. To this end, we used a smaller value (of 25 rather than 50) for the minimum number of data points required for a combination of traits, and performed the same model training and cross-validation. In this robustness analysis, we did not use R^2_{adj} as a performance measure since the number of traits and test data points was equal or close to each other for some trait combinations, resulting in infinity or out-of-bound values for R^2_{adj}. Interestingly, the traits contributing to the best model with respect to the r values were the same as in the previous analysis; in addition, other models with at most six traits again displayed high performance scores (Fig. S1). Therefore, the robustness analysis indicated that the models on a larger number of traits as predictors did not outperform the best model with five traits, identified based on the stricter data consideration.

Other factors that can contribute to a poor performance of a RF model include (multi)collinearity of predictors and presence of irrelevant predictors. In other words, adding irrelevant and highly correlated predictors is not expected to improve model performance and may also have an opposite effect on model performance due to the increasing model uncertainty and complexity (Kuhn et al., 2013). Indeed, we found pairwise correlations between different traits, indicating their collinearity (Fig. 2a). For example, T_{leaf} and T_{mes}, T_{leaf} and *LMA*, D_{leaf} and *LMA* as well as T_{mes} and *LMA* represent trait pairs showing strong, moderate, weak, and no correlations, respectively (Fig. S2). Thus, (multi)collinearity of the predictors can explain the negative effect of increasing number of predictors on the performance of RF models.

Feature selection is a common strategy to resolve the problem of (multi)collinearity among the predictors. This is performed either by preselecting the predictors according to defined criteria (filter methods) or by iteratively identifying the predictors that maximize the performance of the target model (wrapper methods) (Kuhn et al., 2013). Both approaches aim to remove non-informative and highly correlated traits and reach an optimal subset with respect to different criteria (e.g. minimum number of predictors retained). In our setting, having only ten predictors allowed us to investigate all possible combinations of predictors along with the respective models. As a result, we did not rely on selection of features since we performed exhaustive training of models of each of these combinations of predictors. We expected the five traits appearing in our best-performing model (model 1 in Fig. 1) to be the best representatives for the rest of the traits and have no high correlations with each other. Indeed, as expected, we found that all pairs of predictors in the best-performing model show weak correlations, except for D_{leaf} and T_{cw} that exhibit moderate correlations (Fig. 2b). In addition, removing D_{leaf} from the set of predictors resulted in the second best-performing model ($R^2_{adj} = 0.61$, model 2 in Fig. 1). This demonstrates that D_{leaf}, despite the

moderate correlations to the other predictors, still contributes to explaining some of the variance in g_m. Therefore, our modeling strategy contributed to understanding how (multi)colinearity between leaf architecture traits impacts on the predictability of the resulting models.

To better assess the last claim, we observed that a trait can be excluded from a model either because it is correlated to another predictor present in the model or because it does not contribute to explaining variance in g_m. Inspection of models for all possible combinations of traits as predictors allows us to assess the reasons for not considering a trait in a predictive model for g_m. Having more models with a positive R^2_{adj}, with different combination of traits, indicate one or both of the two possibilities: *i*) more traits significantly contribute to explaining variance in g_m, and *ii*) correlated traits also contribute to explanation of variance in g_m, due to lack of high correlation between each other. In this way, the number of models with positive R^2_{adj} and the distribution of their predictability values provide a general view in assessing the relationship between leaf anatomy and g_m. The information provided by this metric can also be more robust than the information obtained by the best-performing model, considering the possible overfitting due to the number of data points and biological differences between species appearing in each model (i.e. trait combination).

Following this logic, we summarized the information about the performance of the models on the global data set and on data of individual PFTs using the distribution of predictability scores (Fig. 3). In addition to the global data set (as discussed above), the cross-validation over the data of eight individual PFTs showed several models with non-negative R^2_{adj}. For instance, woody evergreens, woody evergreen angiosperms, gymnosperms, evergreen gymnosperms, $C_3 - C_4$ herbaceous, woody angiosperms, C_3 herbaceous, and extended ferns were the PFTs with at least one model with positive R^2_{adj}. Excluding C_3 herbaceous, the best models of the mentioned PFTs showed weak to moderate R^2_{adj} scores alongside high values for r between the measured and predicted values of g_m in the test set (Table 1). These results provide further, strong evidence for the effect of leaf anatomy on g_m within PFTs, in agreement with what has already been presented in the literature (e.g., Knauer et al., 2022a).

However, these results raise the question of why there are such differences in the predictive performance scores between the global data set and the PFTs, as well as between the PFTs themselves. To address this issue, we aimed to further examine the clear statistical differences in the data of the different PFTs. We observed that data from individual PFTs contained fewer data points compared to the global data set, which, as mentioned above, can negatively affect the performance of the models when trained on the data from the individual PFTs. Moreover, the data sets of each of the considered PFTs omit some traits, resulting in the consideration of only a fraction of the possible combinations with the ten traits as predictors (Table S2).

To investigate the effect of missing traits and combinations, we focused on the most special case, that of C_3 herbaceous PFT, which showed very poorly performing models ($R^2_{adj} <= 0.04$)

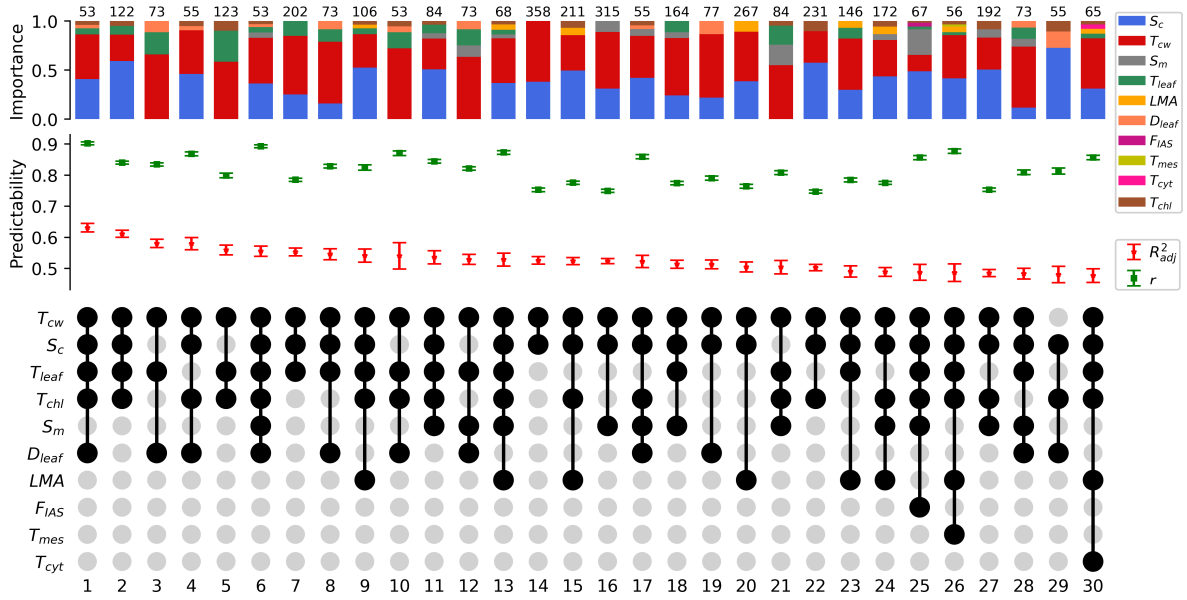

Figure 1: Predictive performance of random forest models using different combinations of leaf anatomical traits. The UpSet plot shows the predictability evaluation of top 30 models for g_m, based on average R^2_{adj}, consisting of ten anatomical traits *LMA*, T_{mes}, F_{IAS}, T_{cw}, T_{cyt}, T_{chl}, S_m, S_c, T_{leaf}, and D_{leaf} over all available species and PFTs of Knauer et al. (2022b) data set. The lower panel shows the intersection of traits contributing to the training model. The middle panel indicates the average R^2_{adj} and r between the measured and predicted values of g_m in the test set. The error bars show the standard errors of the predictability measures. The upper panel shows the average Gini importance of the corresponding traits at each combination of the traits. The number of data points in each model is provided above the importance bars. For all models, the average predictability scores were achieved by the RF model in 150 executions, with 70% randomly chosen data elements used for the training set and the remaining 30% used for the test set.

across the 49 trained models. To this end, we applied the same analyses to the data set of a recently published paper (Xiong, 2023), providing the g_m data for eight of our anatomical traits across ten C_3 crops. This yielded several models with considerably higher predictive performance for C_3 herbaceous plants, i.e., moderate R^2_{adj} and high r values, where at least one of the traits involved in each of the top 30 models was missing in our C_3 herbaceous data set (Fig. S3). The observed effect of missing traits and combinations in this particular case, along with strong correlation values in all PFTs, suggests that increasing the available data for individual PFTs may improve model performance.

2.2 Predictive performance of the RF models between PFTs

Our other approach to assess the relationships between leaf structural traits and g_m was cross-prediction over PFTs, i.e., prediction of the g_m values in one PFT by the RF models trained on the data set of another PFT. This approach allowed us to assess if and to what extent the models are

Table 1: Best-performing models for g_m in cross-validation within different PFTs.

PFT	R^2_{adj}	r	n
Global data set	0.63	0.90	599
Woody evergreens	0.42	0.75	72
Woody evergreen ang.	0.3	0.74	63
$C_3 - C_4$ herbaceous	0.37	0.69	49
Evergreen gym.	0.3	0.41	31
Extended ferns	0.22	0.78	7
Woody angiosperms	0.21	0.68	72
C_3 herbaceous	0.04	0.51	49

Cross-validation predictability scores of the model with the highest R^2_{adj} for global data set and each of seven individual PFTs with at least one model with positive R^2_{adj}. The number of trained models, n, for each data set is also given in the table.

generalizable, i.e. their performance remains good on unseen data sets. This modeling strategy also sharply decreases the probability that a pair of data samples from an identical study is split such that one lies in the training set and the other on the test set, given the majority of the studies in the data set provide measured g_m for a few species from one PFT. As a result, this strategy overcomes the bias in the models due to possible systematic errors in measurements of different studies. Finally, by following this strategy, we aimed to investigate if the same relationship, captured in a RF model, holds across PFTs.

First, we developed RF models in a setting where the data of one PFT was considered as the test set and the remaining data as the training set. This resulted in numerous models with moderate to strong R^2_{adj} and r values (Fig. 4a). The prediction of g_m on woody angiosperms, including both evergreen and deciduous species, showed the largest number of RF models with a positive R^2_{adj} set as well as the model with the highest predictability across all scenarios (Fig. 4a and Table 2). The two subgroups of these species, woody evergreen angiosperms and woody deciduous angiosperms, also showed several models with moderate R^2_{adj} and strong r values. A special case was the scenario with C_3 annual herbaceous as the test set, which showed a weak performance for one model and a negative R^2_{adj} for the rest of 347 trained models. This finding distinctly contrasts the scenario in which C_3 perennial herbaceous was considered as a test set, showing several models with moderate to strong predictability scores on only 83 trained models. In addition, the union of these PFTs, i.e., C_3 herbaceous and $C_3 - C_4$ herbaceous, as the test sets also showed predictability scores with performances between these two cases. The prediction of g_m on evergreen gymnosperms and woody evergreens as test sets showed 28 and four models with a positive R^2_{adj}, respectively. Finally, the scenarios with the (extended) ferns as the test sets also showed ten models with non-negative R^2_{adj}, with the best models showing moderate to strong

predictability scores. The data on the remaining PFTs either did not result in a model with non-negative R^2_{adj} or did not contain sufficient points to apply the same prediction scenarios (Tabel S2).

In the next step, we investigated prediction scenarios in which the different pairs of non-overlapping PFTs were considered as the training and test sets for the RF models, respectively. Among all the possible pairs of PFTs, 121 scenarios had at least one combination of traits with sufficient training and test data points, with 31 of them resulting in at least one model with positive R^2_{adj} values (Fig. 4b and Fig. S4). Different groups of woody plants, i.e., woody (evergreen/deciduous) angiosperms and evergreen gymnosperms, C_3 (annual/perennial) herbaceous plants, and (extended) fern plants were included in the training and the test sets of all 31 scenarios.

The RF models trained on data from selected woody species and tested on the other woody species, C_3 herbaceous plants, and ferns were generally the best-performing (Fig. 4b, Fig. S4 and Table 2).

Woody angiosperm species represented a special case, since: (*i*) the model trained on the global set (excluding this PFT) resulted in the best-performing model when tested on this PFT (Table 2) and (*ii*) the model trained on data from this PFT predicted g_m with the best performance on data from C_3 perennial herbaceous species (Table 2). In addition, models trained on data from woody plants resulted in 40 models with positive R^2_{adj} when tested on data from (extended) ferns (Fig. 4b and S4). However, the models trained on (extended) ferns could only predict g_m on C_3 ($-C_4$) herbaceous and woody plants in 8 and 2 models, respectively. On the other hand, the models trained on C_3 herbaceous plants could only predict the g_m from the other C_3 herbaceous plants (15 models) and woody deciduous plants (2 models). The scenarios with the C_3 perennial herbaceous as the test sets showed several models with positive R^2_{adj}, including the one with the best predictability scores. However, this was not the case for the C_3 annual herbaceous. This result was in line with the significant difference between the predictability of the models tested on C_3 annual herbaceous and C_3 perennial herbaceous, both of which were trained on the rest of the global data set (Table 2).

2.3 Importance of anatomical traits in predicting g_m

The relative importance of the traits contributing to the RF models is of particular interest when interpreting the nonlinear relationships between anatomical traits and g_m.

Previous works investigating the relationships between g_m and leaf architecture generally considered and investigated models with one or two traits. They then identified the traits with and without significant regression scores as important and unimportant, respectively (e.g., Knauer et al., 2022a, Flexas et al., 2021, Xiong, 2023). However, here we follow a different approach: we developed a model for each possible combination of ten traits, available in our data set, and

computed the performance of the RF models in predicting g_m by anatomical traits. Hence, we required different strategies to identify which traits were more important in explaining g_m. To this end, we considered three different aspects to evaluate the importance of the traits: *i*) the contribution of a trait in the given model, *ii*) the relative importance of a contributing trait in the given RF model, and *iii*) the overall impact of a trait in a set of models in terms of its contribution and relative importance, taking into account the performance of the models containing the trait.

The traits contributing to the optimal model can initially be considered as the most important in explaining g_m. However, the contribution of a trait still does not provide details about its share in predicting the g_m in the optimal model. Therefore, we considered the average impurity-based Gini importance of each trait across different runs of the RF model as its relative importance (Fig. 5). The first surprising result was the major share of importance of one or two traits included in each model. Further, in most models, one trait accounted for at least 50% of the Gini importance and another trait accounted for most of the remaining portion. Interestingly, S_c and T_{cw} were among the important traits in the majority of models. This is in line with several previous works that recognized these traits as the two essential anatomical traits to explain the variation of g_m across PFTs and species (see Section Introduction). In addition, each of the other eight traits contributed to at least one of the best-performing models. This provides evidence that the ten investigated anatomical traits contribute to explaining the variance in g_m across plant species.

Next, we investigated the contribution and importance of the traits in other RF models. Similar to the best-performing, the remaining models with a non-negative performance showed a large proportion of Gini importance for only one or two traits (e.g., the upper panel of Fig. 1, S1). To summarize the importance of each trait over all the models with positive R^2_{adj} we used two total importance measures IMP_C and IMP_G (see Section Measures of predictor importance in RF models). Interestingly, except in three cases (i.e., the scenario trained on the C_3 annual herbaceous and tested on C_3 perennial herbaceous plants along with the scenarios trained on C_3 and $C_3 - C_4$ herbaceous and tested on woody deciduous angiosperms), one or both of S_c and T_{cw} were again the most important traits based on IMP_G in all the scenarios (Fig. 3, 4a, 4b, and S4). However, the ordering of traits based on importance values assessed by IMP_C were different: in ten cases, neither S_c nor T_{cw} were found to be among the top two important traits, and the importance share of the most important traits small compared to the results obtained by IMP_G. Excluding S_c and T_{cw}, again, all the remaining eight traits showed a considerable contribution of total importance, at least in one of the prediction scenarios, particularly when using IMP_C. As special cases, D_{leaf} and S_m were the most important traits in two scenarios, and *LMA* and f_{IAS} were the most important in one scenario in terms of both measures of total importance.

In summary, our findings indicated that the ten considered anatomical traits are important in explaining g_m in different prediction scenarios, based on considering the best-performing models

or all the models with a positive R^2_{adj}. We also showed that the relative importance of a few of the traits in explaining g_m is considerably higher than the others. Meanwhile, there are still uncertainties in ranking the importance of the traits due to data limitations. More specifically, except for two prediction scenarios (i.e., cross-validation over the global data set and the scenario with woody angiosperms as the test set and the rest of the global data set as the training set), one or more traits were missing in other scenarios. Therefore, we avoid ranking the contribution of the traits and only highlight the main trends, such as the major importance of the two traits, namely, S_c and T_{cw}.

3 Discussion

Our study aimed to address the relationship between leaf architecture traits and g_m, thus helping assess the suitability of modulating leaf anatomy as a way towards engineer g_m. Several studies have already investigated and attempted to find significant empirical relationships between leaf architecture traits and variability of g_m across plant species. The models reported in the existing studies mainly suggested that two anatomical traits, T_{cw} and S_c, can explain a small proportion of the variability in g_m, as assessed by weak to moderate R^2; these models were developed often using a limited number of data points (see section Introduction). In addition, these modelling efforts generally failed to find a significant relationship between leaf structural traits (e.g., *LMA*, D_{leaf}, T_{leaf}, and T_{mas}) and variation of g_m across PFTs and species (Knauer et al., 2022a). As a result, the existing models tend to not generalize well on unseen data. Further, the existing models are rooted in different linear and nonlinear regression approaches. For instance, different studies have used linear (Carriquí et al., 2020), exponential (Tosens et al., 2016), logarithmic (Tomás et al., 2013; Veromann-Jürgenson et al., 2017, 2020), and power-law (Flexas et al., 2021; Knauer et al., 2022a) models to fit the g_m based on T_{cw}.

However, comprehensive models that consider the majority of measured leaf architecture traits as predictors have not yet been carefully investigated and compared. Here, we ask if non-linear machine-learning models, with more than two leaf architecture traits as predictors, can be used to improve the predictive power of models describing the relationship between anatomy and g_m. Interestingly, the RF model built based on data for the two anatomical traits, T_{cw} and S_c, found moderately correlated with g_m, demonstrated that increases in any of these traits does not necessarily lead to an increase in g_m (see the rugged surface on Fig. S5). This was a further motivation to employ multivariate nonlinear models that consider other anatomical and structural traits describing different parts of leaf architecture. In this regard, we created an RF model for each possible combination of ten leaf architecture traits on the available data across 34 distinct prediction scenarios, representing the global data set, different PFTs, and combinations thereof.

Following these strategies, we identified several models, considering both anatomical and structural traits, with strong predictability scores for different prediction scenarios. Particularly, we found that the model trained on all the PFTs can predict g_m based on three anatomical traits (i.e., T_{CW}, S_C, and T_{chl}) and two structural traits (i.e., D_{leaf}, and T_{leaf}) over unseen data with $R^2_{adj} = 0.63$ and $r = 0.9$. This evidence reliably indicates that the leaf architecture is a primary determinant of the variation of g_m within and between PFTs. Furthermore, these findings suggest that a comprehensive analysis of both leaf structure and anatomy is necessary to explain the variation of g_m across species. On the other hand, our analysis indicated that in addition to the best-performing model for each scenario, other models based on different combinations of the traits should also be taken into account. This can result in an exhaustive understanding of different aspects of the effect of leaf architecture on g_m, considering weak to strong (but not perfect) correlations between anatomical and structural traits.

The aim of this study was to also provide robust and generalizable models allowing to assess the extent to which different parts of the leaf architecture associate to g_m across PFTs and species. To this end, we also examined whether and how the models trained on one or more PFTs can predict g_m from other, unseen PFTs. This resulted in the identification of the most robust models tested on completely unseen species. Moreover, this strategy can uncover similarities and differences in the association of g_m with leaf architecture across different PFTs. Our results yielded strong predictability for several models built based on this idea. For instance, the models trained on the global data set, with no overlap with the test sets, could predict g_m on woody angiosperms, C_3 herbaceous, and (extended) ferns with an $R^2_{adj} > 0.5$. On the other hand, among the models trained and tested on individual PFTs, the ones either trained or tested on woody plants generally showed higher performances. The models trained on data from these plants could predict g_m on other woody plants, C_3 herbaceous plants, and ferns. However, the models trained on data from ferns and C_3 annual herbaceous generalized to a much smaller degree to other PFTs.

Interestingly, our analysis of the data from Xiong (2023) found that only two of the 30 best-performing models contained both T_{cw} and S_c as, with these traits making up only a small fraction of the Gini importance scores (Fig. S3). This outcome varies considerably from the analysis based on the Knauer et al. (2022b) data set including all PFTs (Fig. 1). Furthermore, the observation that T_{cw} seems less important in the crop species studied by Xiong (2023) is in stark contrast with a published comparison of anatomy across 15 species, spanning multiple PFTs. Tomás et al. (2013) showed that the slope between g_m (standardised by S_c) and T_{cw} was much steeper within herbaceous C_3 species, than for evergreen trees, suggesting that T_{cw} plays a larger role in determining g_m within the C_3 herbaceous annual leaves. The importance of T_{cw} in determining g_m has also been difficult to assess from experimental studies. For example, work on tobacco found that the reduction in g_m coinciding with leaf age was strongly correlated with an increase in T_{cw} (Clarke

et al., 2021). However, knocking down cell wall mixed-linkage glucan production in rice plants resulted in lower g_m, alongside concurrent reductions to T_{cw} (Ellsworth et al., 2018). As a result of these contrasting observations, it remains unclear if, and to what extent, T_{cw} is influencing g_m within C_3 herbaceous annuals.

One open question from this study is why models to describe C_3 annual plants underperformed, compared to other PFTs. One possible explanation is that this is an artifact, caused by the averaging of g_m values derived from different experimental methods. Knauer et al. (2022a) showed that linear regressions between g_m and V_{cmax} fit the data considerably better when separate models were built depending on the method used to estimate g_m (i.e. isotope, fluorescence or curve fitting). Whilst V_{cmax} bears no importance for our analysis, this indicates that averaging g_m values may not always yield the most reliable results. Estimations of g_m rely on several assumptions (e.g., fractionation factors, the photorespiratory compensation point, methods chosen to estimate respiration). As such, it is conceivable that combining independent estimations of g_m may have introduced unforeseen errors into the dataset that may interfere with model construction. Given that there is a bias towards research on C_3 annual species (which the majority of the world's staple crop species belong), a greater number of measurements have been recorded, per species, for this PFT. Consequently, within the data set collated by Knauer et al. (2022b) C_3 annual herbaceous species had 538 measurements for 52 species, whereas the ratio of measurements to species was < 2.5 for all other PFTs. This remains to be tested, but it may also explain why models could be built to describe the relationship between anatomy and g_m based on data from Xiong (2023), as these were derived from a single source and were not subject to the same averaging.

4 Conclusions

By using well-established machine learning approach, that of random forest, we demonstrated that one can obtain models based on leaf architecture traits that achieve excellent predictability of g_m. In addition, we showed that these models are generalizable, particularly if trained with data from specific PFTs. We also presented a systematic approach for determining the importance of anatomical and structural traits based on the Gini importance of traits in best-performing models and two total importance measures that consider all models with a positive R^2_{adj} in each prediction scenario. Using the systematic approach, we found that not only T_{cw} and S_c are two critical traits in explaining the variation of g_m across plant species, but the remaining eight structural and anatomical traits considered play a role in explaining g_m. In future work, our approach can also be used for the exact ranking of the importance of the traits by increasing the data availability or considering natural variability within species.

5 Methods

5.1 Data and preprocessing

To identify and analyze the relationships between anatomical traits and g_m, we used a recently published comprehensive data set (Knauer et al., 2022b) providing leaf structural, anatomical, biochemical, and physiological traits measured on the same set of plants. This is currently the largest available data for g_m, which collected measurements from 563 peer-reviewed studies over 617 species partitioned to 13 major PFTs, namely: evergreen gymnosperms, deciduous gymnosperms, woody evergreen angiosperms, woody deciduous angiosperms, semi-deciduous angiosperms, CAM plants, ferns, fern allies, mosses, C_3 perennial herbaceous, C_3 annual herbaceous, C_4 annual herbaceous, and C_4 perennial herbaceous.

Since most of the individual PFTs do not contain enough data points to train a model for many of the possible combinations of the traits as predictors, we also formed five more groups from the union of the above PFTs. This was performed according to the shared functional characteristics among the PFTs. The groups involved the following: woody evergreens (union of woody evergreen angiosperms and evergreen gymnosperms), woody angiosperms (union of woody evergreen angiosperms, woody deciduous angiosperms, and semi-deciduous angiosperms), extended ferns (union of ferns and fern allies), C_3 herbaceous (union of C_3 perennial herbaceous and C_3 annual herbaceous), and ($C_3 - C_4$) herbaceous (union of C_4 annual herbaceous, C_4 perennial herbaceous, C_3 perennial herbaceous, and C_3 annual herbaceous). This strategy allowed us to not only increase the data available for model training, but also to compare the findings for different groups and their subgroups. This modeling strategy also facilitated the investigation of whether or not the combination of data from PFTs increase the generalizability of the models.

In our analyses, we used g_m values standardized to temperature of $25^{\circ}C$ and atmopheric pressure of 1 bar ($10^5 Pa$), as provided in the data set (see Knauer et al., 2022a). The data set contains information about all the published methods for estimating g_m in each study. Except for a few cases, all collected measurements were based on one of three methods: isotope (Evans et al., 1986; Caemmerer and Evans, 1991; Lloyd et al., 1992; Scartazza et al., 1998; Tazoe et al., 2009, 2011; Evans and Von Caemmerer, 2013; Mizokami et al., 2015), fluorescence (Harley et al., 1992; Loreto et al., 1992; Epron et al., 1995; Maxwell et al., 1997; Bernacchi et al., 2002; Yin and Struik, 2009; Yin et al., 2009), and curve fitting (Ethier and Livingston, 2004; Ethier et al., 2006; Sharkey et al., 2007; Gu et al., 2010; Sharkey, 2015). To have only one value for each individual experiment, we aggregated the repeated data by calculating per-species g_m as the average of its values measured with the different methods. After aggregation, we used all the remaining data with no additional filters.

The data sets includes measurements for 31 anatomical traits. However, in addition to the two

frequently reported traits (T_{cw} and S_c), we selected eight other anatomical and structural traits that have received the most attention in the literature in terms of published data (Table S1): Leaf dry mass per area (*LMA*), leaf density (D_{leaf}), leaf thickness (T_{leaf}), mesophyll thickness (T_{mes}), cytosol thickness (T_{cyt}), chloroplast thickness (T_{chl}), surface area of mesophyll cells exposed to the intercellular airspaces per unit leaf area (S_m), and fraction of intercellular airspaces in leaf mesophyll (F_{IAS}). In this way, we kept all the data samples with a value for standardized g_m and *at least one* of the mentioned anatomical traits, resulting in 882 data samples from 453 species and all the mentioned PFTs. The number of data samples and species for individual PFTs and groups are provided in Table S2.

To investigate the performance of selected models, we used the data set from Xiong (2023) consisting of eight anatomical and structural traits T_{mes}, F_{IAS}, T_{cw}, T_{cyt}, T_{chl}, S_m, S_c, and T_{leaf} measured for ten C_3 crops. The measurements of g_m in this data set were obtained using online carbon isotope discrimination and chlorophyll fluorescence methods, and we used the g_m provided by the second method in our analyses.

To perform the model training based on data for each PFT, we then constructed all possible combinations of traits for the *global data set* (consisting of all PFTs) and for each of the individual PFTs, respectively. In each combination, we removed the data samples with a missing value in one or more traits. We then kept only the combinations with at least 50 data samples, with no missing data, and ignored the rest. This strategy allowed us to avoid data imputation, that may bias the findings given that the measurements are made across different plant species. Future studies may consider investigating the effect of bias by relying on recently proposed imputation techniques (Ellington et al., 2015; Scherer and Emslander, 2023; Lee and Beretvas, 2023). This resulted in 599 combinations for the set of all PFTs, each containing from one to ten anatomical traits as independent variables and the corresponding standardized g_m as a response variable (Table S2). We also ensured that the training set was larger than the test set, to achieve generalizable models.

5.2 The model

The random forest (RF) model in a regression setting (Breiman, 2001) was used to predict g_m by the anatomical traits, used as predictors. To achieve a robust result for the prediction scenarios within PFTs, for each combination of predictors we performed the training in a Monte-Carlo cross-validation setting (Smyth, 1996), by running 150 independent executions with 70% randomly chosen data points for the training set and the remaining 30% used for the test set. We also run the RF model 150 times for prediction scenarios between PFTs, with fixed training and test sets, to capture the effect of different random seeds controlling the bootstrapping and feature sampling in the trees (Raste et al., 2022). The training models and splitting data were implemented using the Python package Scikit-learn (Pedregosa et al., 2011). The source code ensuring reproducibility of

our analyses is available on GitHub: github.com/MRahimiMajd/leaf_gm_architecture.

The predictive performance of the models was assessed quantitatively and qualitatively by using the coefficient of determination (R^2_{adj}) and Pearson correlation (r), respectively. The coefficient of determination (R^2) is used as a quantitative measure of how much variance in g_m is explained by the anatomical traits, employed as predictors. However, our models have different numbers of independent variables (i.e., anatomical traits as predictors). To capture the effect of this difference on the performance of models, we relied on the R^2_{adj}, which adjusts the R^2 value based on the number of predictors (Hocking, 1976).

5.3 Measures of predictor importance in RF models

To assess the relative importance of a trait contributing to a RF model, we used the impurity-based feature importance (Gini importance). Since the RF model is an ensemble of decision trees (obtained by node splitting), Gini importance measures the total reduction of the impurity of the RF model attributed to that feature, averaged over all trees in the ensemble (Pedregosa et al., 2011). For a single run or an ensemble of runs, we ensure that the (average) values of the relative importance of the contributing traits always sum up to one, as explained in the following.

For the case where we have several models with different combinations of traits as predictors, we are interested in obtaining a total importance for each trait across all these models. In our analyses, the number of models is given by the number of possible combinations of traits. However, poorly performing models do not provide any information about trait importance. Thus, by excluding these models, based on a threshold for the measure of performance, the number of models that include a given trait can simply be used as a measure of total importance for the trait. While seemingly sound, this measure does not discriminate between models of weak, moderate, and strong performance. To address this issue, we also employ the quality of the regression and the Gini importance of the traits in each model to define two total importance measures: the total contribution importance (IMP_C) and the total Gini importance (IMP_G). More specifically, IMP_C of a trait is defined as the average of R^2_{adj} values of all models with positive R^2_{adj} including the trait. This measure captures the contribution of the traits in the models weighted by the performance of the models. The IMP_G follows the same steps, but the R^2_{adj} values for each model are also multiplied by the Gini importance of the traits before averaging these values. This total importance measure captures more detail about the impact of the traits on achieving models of good performance while considering the importance of features in the RF model. Having the average values for each trait, we normalize them such that the sum of the importance values of all the traits equals to one. In our analyses, we set the threshold at which a model contributes to the total importance measures as $R^2_{adj} = 0$. This threshold indicates that the model explains the variance of g_m better than the average of its values (Chicco et al., 2021).

6 Data availability

All used data along with the Python functions used in our analyses are available on the following URL: github.com/MRahimiMajd/leaf_gm_architecture. The Knauer et al. (2022b) data set is also available using the link: https://doi.org/10.6084/m9.figshare.19681410.v1.

7 Funding

All authors acknowledge the funding support by the NovoNordisk Foundation, Data Science Initiative, project DIRECTION (Grant NNF 21OC0068884, to Z.N. and J.K.).

8 Author contributions

M.R.-M. performed analyses, interpreted the results, and wrote the paper. A.L. interpreted results and commented on drafts of the paper. J.K. interpreted results and commented on the drafts of the paper. Z.N. conceptualized the study, interpreted results, wrote the paper. All authors contributed to finalizing the manuscript.

9 ORCID

Milad Rahimi-Majd https://orcid.org/0009-0009-1217-8563
Alistair Leverett https://orcid.org/0000-0002-7064-1917
Johannes Kromdijk https://orcid.org/0000-0003-4423-4100
Zoran Nikoloski https://orcid.org/0000-0003-2671-6763

10 Acknowledgments

The authors would like to thank Dr. Dongliang Xiong for kindly sharing the data on anatomical and structural leaf architecture traits from C_3 crops.

11 Declaration of interest

The authors declare no competing interests.

References

Carl J Bernacchi, Archie R Portis, Hiromi Nakano, Susanne Von Caemmerer, and Stephen P Long. Temperature response of mesophyll conductance. implications for the determination of rubisco enzyme kinetics and for limitations to photosynthesis in vivo. *Plant physiology*, 130(4):1992–1998, 2002.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

SV Caemmerer and John R Evans. Determination of the average partial pressure of co2 in chloroplasts from leaves of several c3 plants. *Functional Plant Biology*, 18(3):287–305, 1991.

Marc Carriquí, Miquel Nadal, María J Clemente-Moreno, Jorge Gago, Eva Miedes, and Jaume Flexas. Cell wall composition strongly influences mesophyll conductance in gymnosperms. *The Plant Journal*, 103(4):1372–1385, 2020.

Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, 7:e623, 2021.

Victoria C Clarke, Florence R Danila, and Susanne von Caemmerer. Co2 diffusion in tobacco: a link between mesophyll conductance and leaf anatomy. *Interface Focus*, 11(2):20200040, 2021.

María José Clemente-Moreno, Jorge Gago, Pedro Díaz-Vivancos, Agustina Bernal, Eva Miedes, Panagiota Bresta, Georgios Liakopoulos, Alisdair R Fernie, José Antonio Hernández, and Jaume Flexas. The apoplastic antioxidant system and altered cell wall dynamics influence mesophyll conductance and the rate of photosynthesis. *The Plant Journal*, 99(6):1031–1046, 2019.

E Hance Ellington, Guillaume Bastille-Rousseau, Cayla Austin, Kristen N Landolt, Bruce A Pond, Erin E Rees, Nicholas Robar, and Dennis L Murray. Using multiple imputation to estimate missing data in meta-regression. *Methods in Ecology and Evolution*, 6(2):153–163, 2015.

Patrícia V Ellsworth, Patrick Z Ellsworth, Nuria K Koteyeva, and Asaph B Cousins. Cell wall properties in oryza sativa influence mesophyll co 2 conductance. *New Phytologist*, 219(1):66–76, 2018.

D Epron, D Godard, G Cornic, and B Genty. Limitation of net co2 assimilation rate by internal resistances to co2 transfer in the leaves of two tree species (fagus sylvatica l. and castanea sativa mill.). *Plant, Cell & Environment*, 18(1):43–51, 1995.

GJ Ethier and NJ Livingston. On the need to incorporate sensitivity to co2 transfer conductance into the farquhar–von caemmerer–berry leaf photosynthesis model. *Plant, Cell & Environment*, 27(2):137–153, 2004.

GJ Ethier, NJ Livingston, DL Harrison, TA Black, and JA Moran. Low stomatal and internal conductance to co2 versus rubisco deactivation as determinants of the photosynthetic decline of ageing evergreen leaves. *Plant, Cell & Environment*, 29(12):2168–2184, 2006.

John R Evans. Mesophyll conductance: walls, membranes and spatial complexity. *New Phytologist*, 229(4):1864–1876, 2021.

John R Evans and Susanne Von Caemmerer. Temperature response of carbon isotope discrimination and mesophyll conductance in tobacco. *Plant, Cell & Environment*, 36(4):745–756, 2013.

JR Evans, TD Sharkey, JA Berry, and GD Farquhar. Carbon isotope discrimination measured concurrently with gas exchange to investigate co2 diffusion in leaves of higher plants. *Functional Plant Biology*, 13(2):281–292, 1986.

Jaume Flexas, María J Clemente-Moreno, Josefina Bota, Tim J Brodribb, Jorge Gago, Yusuke Mizokami, Miquel Nadal, Alicia V Perera-Castro, Margalida Roig-Oliver, Daisuke Sugiura, et al. Cell wall thickness and composition are involved in photosynthetic limitation. *Journal of Experimental Botany*, 72(11):3971–3986, 2021.

Jorge Gago, Marc Carriquí, Miquel Nadal, María José Clemente-Moreno, Rafael Eduardo Coopman, Alisdair Robert Fernie, and Jaume Flexas. Photosynthesis optimized across land plant phylogeny. *Trends in Plant Science*, 24(10):947–958, 2019.

Lianhong Gu, Stephen G Pallardy, Kevin Tu, Beverly E Law, and Stan D Wullschleger. Reliable estimation of biochemical parameters from c3 leaf photosynthesis–intercellular carbon dioxide response curves. *Plant, Cell & Environment*, 33(11):1852–1874, 2010.

Peter C Harley, Francesco Loreto, Giorgio Di Marco, and Thomas D Sharkey. Theoretical considerations when estimating the mesophyll conductance to co2 flux by analysis of the response of photosynthesis to co2. *Plant physiology*, 98(4):1429–1436, 1992.

Ronald R Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, pages 1–49, 1976.

Grace P John, Christine Scoffoni, and Lawren Sack. Allometry of cells and tissues within leaves. *American Journal of Botany*, 100(10):1936–1948, 2013.

Jürgen Knauer, Matthias Cuntz, John R Evans, Ülo Niinemets, Tiina Tosens, Linda-Liisa Veromann-Jürgenson, Christiane Werner, and Sönke Zaehle. Contrasting anatomical and biochemical controls on mesophyll conductance across plant functional types. *New Phytologist*, 236(2):357–368, 2022a.

Jürgen Knauer, Matthias Cuntz, John R. Evans, Ülo Niinemets, Tiina Tosens, Linda-Liisa Veromann-Jürgenson, Christiane Werner, and Sönke Zaehle. A global dataset of mesophyll conductance measurements and accompanying leaf traits. 7 2022b. doi: 10.6084/m9.figshare.19681410.v1. URL `https://figshare.com/articles/dataset/A_global_dataset_of_mesophyll_conductance_measurements_and_accompanying_leaf_traits/19681410`.

Max Kuhn, Kjell Johnson, et al. *Applied predictive modeling*, volume 26. Springer, 2013.

Jihyun Lee and S Natasha Beretvas. Comparing methods for handling missing covariates in meta-regression. *Research Synthesis Methods*, 14(1):117–136, 2023.

J Lloyd, JP Syvertsen, PE Kriedemann, and GD Farquhar. Low conductances for co2 diffusion from stomata to the sites of carboxylation in leaves of woody species. *Plant, Cell & Environment*, 15(8):873–899, 1992.

Francesco Loreto, Peter C Harley, Giorgio Di Marco, and Thomas D Sharkey. Estimation of mesophyll conductance to co2 flux by three different methods. *Plant physiology*, 98(4):1437–1443, 1992.

Kate Maxwell, Susanne von Caemmerer, and John R Evans. Is a low internal conductance to co2 diffusion a consequence of succulence in plants with crassulacean acid metabolism? *Functional Plant Biology*, 24(6):777–786, 1997.

Yusuke Mizokami, KO Noguchi, Mikiko Kojima, Hitoshi Sakakibara, and Ichiro Terashima. Mesophyll conductance decreases in the wild type but not in an aba-deficient mutant (aba1) of n icotiana plumbaginifolia under drought conditions. *Plant, Cell & Environment*, 38(3):388–398, 2015.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

José Javier Peguero-Pina, Sergio Sisó, Jaume Flexas, Jeroni Galmés, Ana García-Nogales, Ülo Niinemets, Domingo Sancho-Knapik, Miguel Ángel Saz, and Eustaquio Gil-Pelegrín. Cell-level

anatomical characteristics explain high mesophyll conductance and photosynthetic capacity in sclerophyllous mediterranean oaks. *New Phytologist*, 214(2):585–596, 2017.

Soham Raste, Rahul Singh, Joel Vaughan, and Vijayan N Nair. Quantifying inherent randomness in machine learning algorithms. *arXiv preprint arXiv:2206.12353*, 2022.

A Scartazza, M Lauteri, MC Guido, and E Brugnoli. Carbon isotope discrimination in leaf and stem sugars, water-use efficiency and mesophyll conductance during different developmental stages in rice subjected to drought. *Functional Plant Biology*, 25(4):489–498, 1998.

Ronny Scherer and Valentin Emslander. Quality assessment in meta-analysis (quama). 2023.

Thomas D Sharkey. What gas exchange data can tell us about photosynthesis. *Plant, Cell & Environment*, 39(6):1161–1163, 2015.

Thomas D Sharkey, Carl J Bernacchi, Graham D Farquhar, and Eric L Singsaas. Fitting photosynthetic carbon dioxide response curves for c3 leaves. *Plant, cell & environment*, 30(9): 1035–1040, 2007.

Padhraic Smyth. Clustering using monte carlo cross-validation. In *Kdd*, volume 1, pages 26–133, 1996.

Youshi Tazoe, Susanne Von Caemmerer, Murray R Badger, and John R Evans. Light and co2 do not affect the mesophyll conductance to co2 diffusion in wheat leaves. *Journal of Experimental Botany*, 60(8):2291–2301, 2009.

Youshi Tazoe, Susanne Von Caemmerer, Gonzalo M Estavillo, and John R Evans. Using tunable diode laser spectroscopy to measure carbon isotope discrimination and mesophyll conductance to co2 diffusion dynamically at different co2 concentrations. *Plant, Cell & Environment*, 34(4): 580–591, 2011.

Magdalena Tomás, Jaume Flexas, Lucian Copolovici, Jeroni Galmés, Lea Hallik, Hipólito Medrano, Miquel Ribas-Carbó, Tiina Tosens, Vivian Vislap, and Ülo Niinemets. Importance of leaf anatomy in determining mesophyll diffusion conductance to co2 across species: quantitative limitations and scaling up by models. *Journal of experimental botany*, 64(8):2269–2281, 2013.

Tiina Tosens, Keisuke Nishida, Jorge Gago, Rafael Eduardo Coopman, Hernán Marino Cabrera, Marc Carriquí, Lauri Laanisto, Loreto Morales, Miquel Nadal, Roke Rojas, et al. The photosynthetic capacity in 35 ferns and fern allies: mesophyll co 2 diffusion as a key trait. *New phytologist*, 209(4):1576–1590, 2016.

Linda-Liisa Veromann-Jürgenson, Tiina Tosens, Lauri Laanisto, and Ülo Niinemets. Extremely thick cell walls and low mesophyll conductance: welcome to the world of ancient living! *Journal of Experimental Botany*, 68(7):1639–1653, 2017.

Linda-Liisa Veromann-Jürgenson, Timothy J Brodribb, Ülo Niinemets, and Tiina Tosens. Variability in the chloroplast area lining the intercellular airspace and cell walls drives mesophyll conductance in gymnosperms. *Journal of Experimental Botany*, 71(16):4958–4971, 2020.

Dongliang Xiong. Leaf anatomy does not explain the large variability of mesophyll conductance across c3 crop species. *The Plant Journal*, 113(5):1035–1048, 2023.

Xinyou Yin and Paul C Struik. Theoretical reconsiderations when estimating the mesophyll conductance to co2 diffusion in leaves of c3 plants by analysis of combined gas exchange and chlorophyll fluorescence measurements. *Plant, Cell & Environment*, 32(11):1513–1524, 2009.

Xinyou Yin, Paul C Struik, Pascual Romero, Jeremy Harbinson, Jochem B Evers, Peter EL Van Der Putten, and JAN Vos. Using combined measurements of gas exchange and chlorophyll fluorescence to estimate parameters of a biochemical c3 photosynthesis model: a critical appraisal and a new integrated approach applied to leaves in a wheat (triticum aestivum) canopy. *Plant, cell & environment*, 32(5):448–464, 2009.

Xin-Guang Zhu, Stephen P Long, and Donald R Ort. Improving photosynthetic efficiency for greater yield. *Annual review of plant biology*, 61:235–261, 2010.

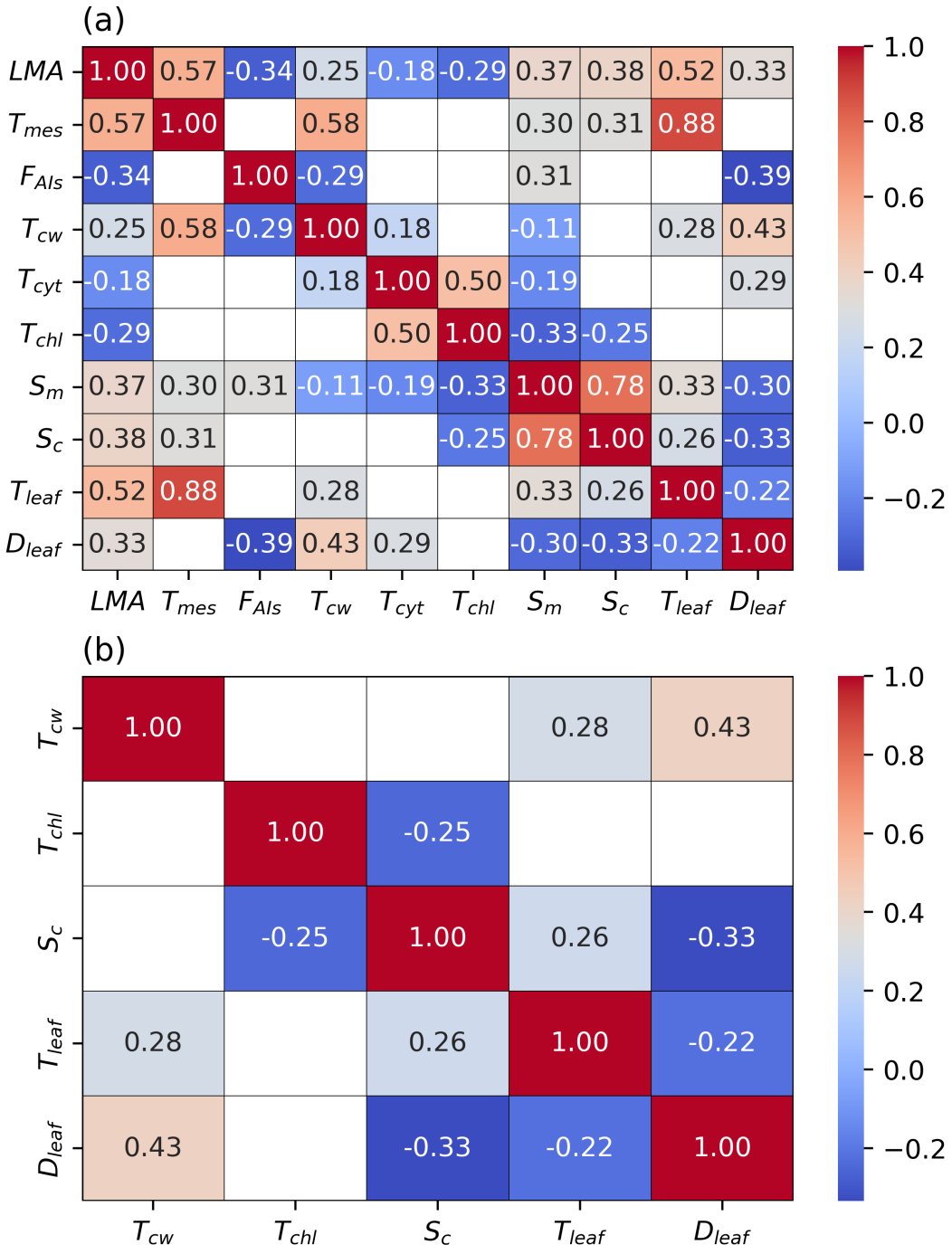

Figure 2: Correlation structure of anatomical traits. Pearson correlation between a) all anatomical traits and b) the anatomical traits of the best-performing model for cross-validation on the global data set. The correlation was calculated on the available data for each trait pair, regardless of other traits. Missing entries in matrices indicate that the p-value of the correlation was larger than our considered significance level 0.05.

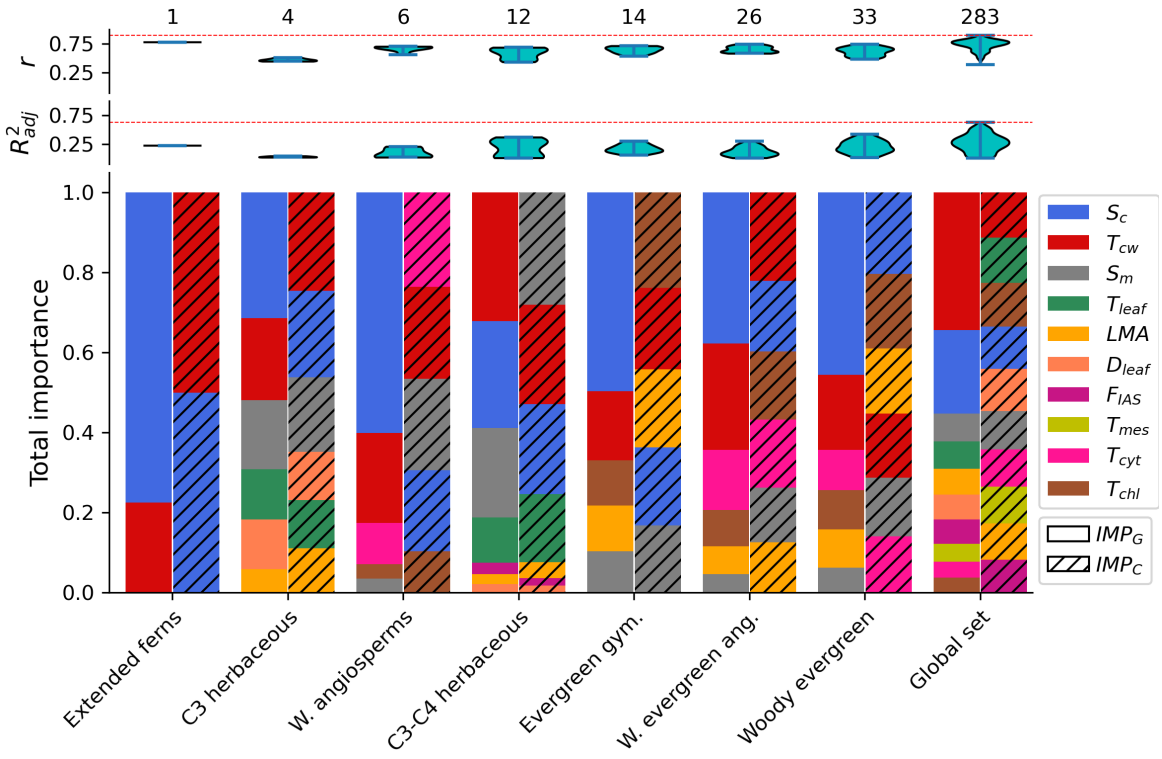

Figure 3: The predictive performance and total feature importance ratios of the trained models within PFTs. Evaluation of the performance of models in different prediction scenarios based on cross-validation within different PFTs. The upper panels show the violin plots of the average predictability scores (r and R^2_{adj}) of the models for the global data set and each PFT. The number of models with positive R^2_{adj} in each scenario is given above the violin plots. The dashed red lines indicate the maximum predictability scores across all models. The bar charts in the lower panel show the total importance measures, IMP_C and IMP_G, of the contributing traits in the models of the global data set and individual PFTs. The position of traits in the bar charts has been sorted from top to bottom based on their total importance. For all the cases, the average predictability scores were achieved by RF model in 150 executions, with 70% randomly chosen data elements used for the training set and the remaining 30% used for the test set.

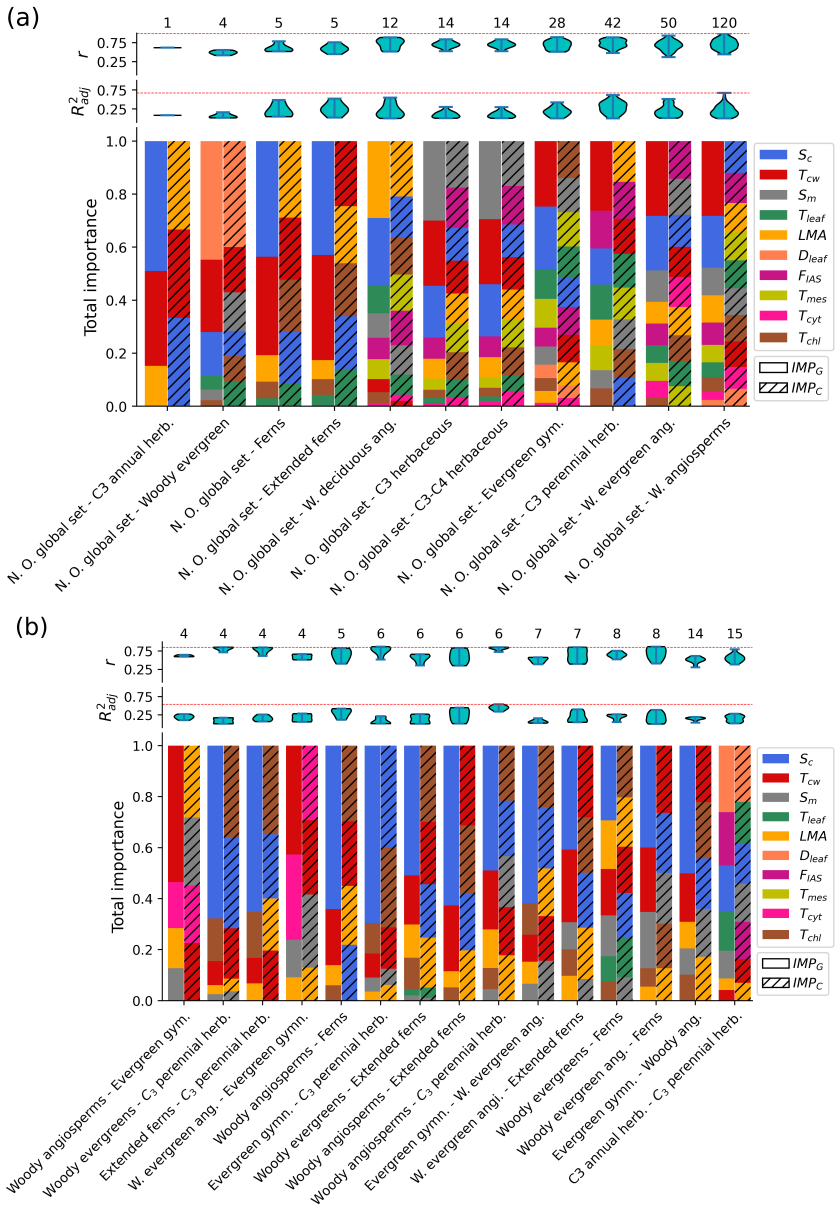

Figure 4: The predictive performance and total feature importance ratios of the trained models between PFTS. The evaluation of the performance of the models in different prediction scenarios with the test sets of individual PFTs and the training sets of: a) non-overlapping global data set with corresponding test sets and b) other individual PFTs. Figure b) contains the 15 of the 31 scenarios with at least one model with a positive R^2_{adj}. The rest of the scenarios are illustrated in Fig. S4. The upper panels show the violin plots of the average predictability scores (r and R^2_{adj}) of the models for each prediction scenario. The number of models with positive R^2_{adj} in each scenario is given above the violin plots. The dashed red lines indicate the maximum of predictability scores across all the models. The bar charts in the lower panels show the total importance measures, IMP_C and IMP_G, of the contributing traits in the models of each scenario. The position of traits in the bar charts has been sorted from top to bottom based on their total importance. For all models, the average predictability scores and importance ratios were achieved by RF model in 50 executions, with the fixed training and test sets.

Table 2: Best-performing models for g_m between PFTs.

Prediction scenario	R^2_{adj}	r	n
Non-overlapping global set - Woody angiosperms	0.67	0.84	371
Non-overlapping global set - C_3 perennial herbaceous	0.62	0.85	83
Non-overlapping global set - Woody deciduous ang.	0.55	0.89	210
Non-overlapping global set - Extended ferns	0.52	0.76	57
Non-overlapping global set - Woody evergreen ang.	0.51	0.94	387
Non-overlapping global set - Ferns	0.49	0.79	55
Non-overlapping global set - Evergreen gym.	0.42	0.84	442
Non-overlapping global set - C_3 herbaceous	0.30	0.81	441
Non-overlapping global set - $C_3 - C_4$ herbaceous	0.30	0.81	436
Non-overlapping global set - Woody evergreens	0.16	0.56	171
Non-overlapping global set - C_3 annual herbaceous	0.09	0.62	347
Woody angiosperms - C_3 perennial herbaceous	0.54	0.86	40
Woody angiosperms - Extended ferns	0.45	0.79	30
Woody angiosperms - Ferns	0.42	0.83	30
Woody evergreen angiosperms - Extended ferns	0.40	0.85	23
Woody evergreen angiosperms - Ferns	0.38	0.86	23
C_3 annual herbaceous - C_3 perennial herbaceous	0.28	0.63	37
Woody evergreen angiosperms - Evergreen gymnosperms	0.28	0.67	35
Woody evergreens - Ferns	0.26	0.66	47
Woody evergreens - Extended ferns	0.27	0.66	43
Woody angiosperms - evergreen gymnosperms	0.27	0.65	55
Extended ferns - C_3 perennial herbaceous	0.26	0.84	16
Evergreen gymnosperms - C_3 perennial herbaceous	0.21	0.87	23
Evergreen gymnosperms - Woody angiosperms	0.19	0.52	29
Woody evergreens - C_3 perennial herbaceous	0.17	0.85	65
Evergreen gymnosperms - Woody evergreen angiosperms	0.16	0.48	30

Predictability scores of the models with the highest R^2_{adj} for 9 prediction scenarios between the non-overlapping global data set and different PFTs (upper side) and 15 prediction scenarios between different PFTs (lower side), with at least one model with a positive R^2_{adj}. The number of trained models for each scenario is also given in the table by n.

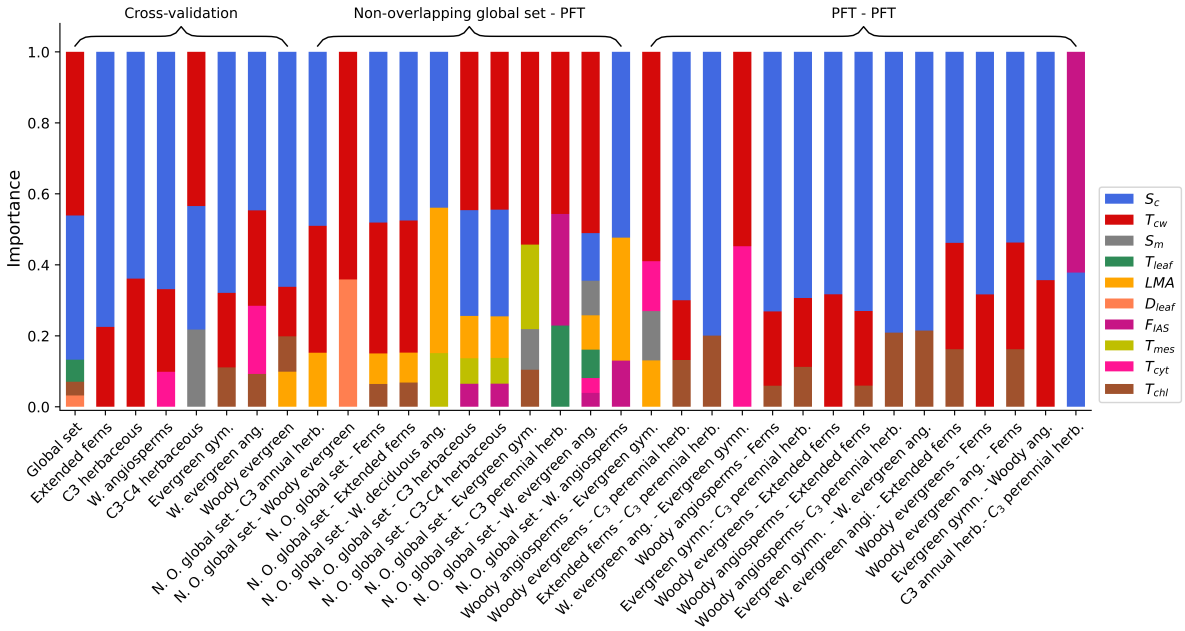

Figure 5: The importance of the traits in the best models of different prediction scenarios. The average Gini importance of the contributing traits on the optimal models of different prediction scenarios, based on R^2_{adj}. The position of traits in the bar charts has been sorted from top to bottom based on their relative importance. The prediction scenarios were classified into three groups based on the data splitting methods: *i*) cross-validation scenarios, *ii*) scenarios with the individual PFTs as the test sets and the non-overlapping part of the global data set with them as the training set, and *iii*) scenarios in which the training and test sets are non-overlapping individual PFTs.

Supplemental information

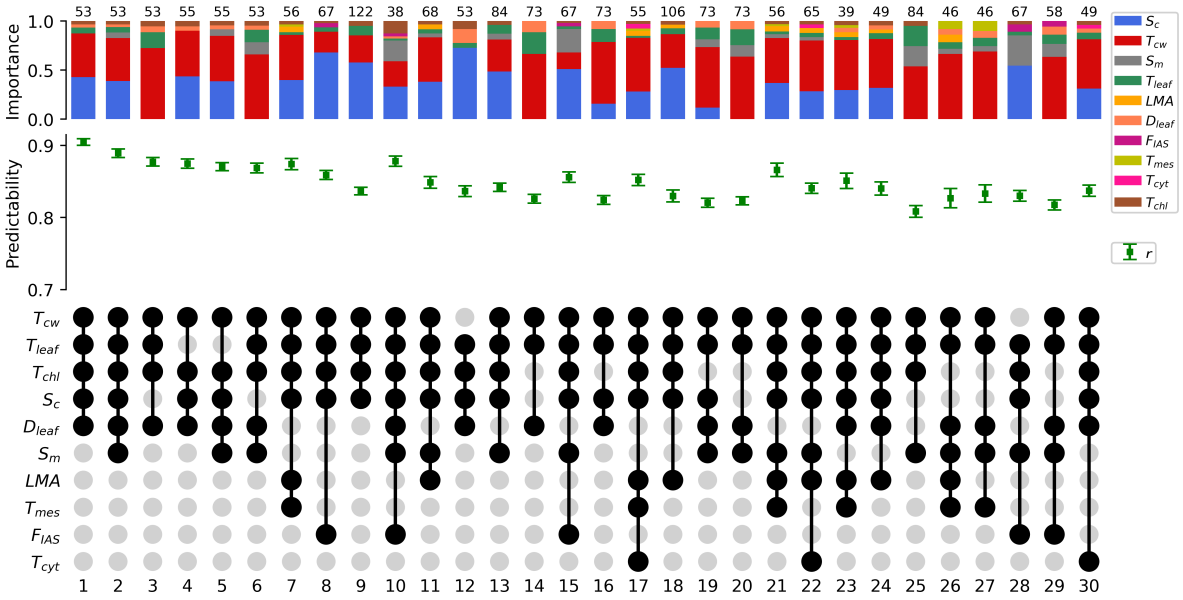

Figure S1: Predictive performance of random forest models, with at least 25 data points, using different combinations of leaf anatomical traits. The UpSet plot shows the predictability evaluation of top 30 models, with 25 data points or more, for g_m, based on average r, consisting of ten anatomical traits *LMA*, T_{mes}, F_{IAS}, T_{cw}, T_{cyt}, T_{chl}, S_m, S_c, T_{leaf}, and D_{leaf} over all available species and PFTs of Knauer et al. (2022b) data set. The lower panel shows the intersection of traits contributing to the training model. The middle panel indicates the average r between the measured and predicted values of g_m in the test set. The error bars show the standard errors of the predictability measures. The upper panel shows the average Gini importance of the corresponding traits at each combination of the traits. The number of data points in each model is provided above the importance bars. For all models, the average predictability scores were achieved by the RF model in 150 executions, with 70% randomly chosen data elements used for the training set and the remaining 30% used for the test set.

Table S1: Available data for anatomical traits.

	Trait	No. of data points
1	Leaf dry mass per area (gm^{-2})	712
2	Surface area of chloroplasts exposed to the intercellular airspaces per unit leaf area (m^2m^{-2})	408
3	Surface area of mesophyll cells exposed to the intercellular airspaces per unit leaf area (m^2m^{-2})	377
4	Cell wall thickness (μm)	362
5	Leaf thickness (μm)	346
6	Chloroplast thickness (μm)	237
7	Fraction of intercellular airspaces in leaf mesophyll	225
8	Mesophyll thickness (μm)	191
9	Leaf density ($g\ cm^{-3}$)	170
10	Cytosol thickness (μm)	131
11	Stomatal density abaxial (mm^{-2})	111
12	Stomatal density adaxial (mm^{-2})	84
13	Chloroplast length (μm)	79
14	Palisade mesophyll thikness (μm)	71
15	Spongy mesophyll thickness (μm)	71
16	Stomatal density (mm^{-2})	52
17	Single stomatal area on abaxial leaf side (μm^2)	39
18	Single stomatal area (μm^2)	33
19	Leaf width (mm)	33
20	Epidermis thickness on adaxial leaf side (μm)	31
21	Single stomatal area on adaxial leaf side (μm^2)	31
22	Stomatal length on abaxial leaf side (μm)	31
23	Stomatal length on adaxial leaf side (μm)	30
24	Epidermis thickness on abaxial leaf side (μm)	30
25	Interveinal distance (μm)	26
26	Stomatal length (μm)	24
27	Stomatal width (μm)	22
28	Stomatal index ($\%$)	21
29	Chloroplast surface area (μm^2)	16
30	Chloroplast width (μm)	14
31	Surface area of bundle sheath cells per unit leaf area (m^2m^{-2})	7

The number of data points in the data set of Knauer et al. (2022a) for all measured anatomical traits. Due to the size of the available data points, we used the first ten traits in our analyses.

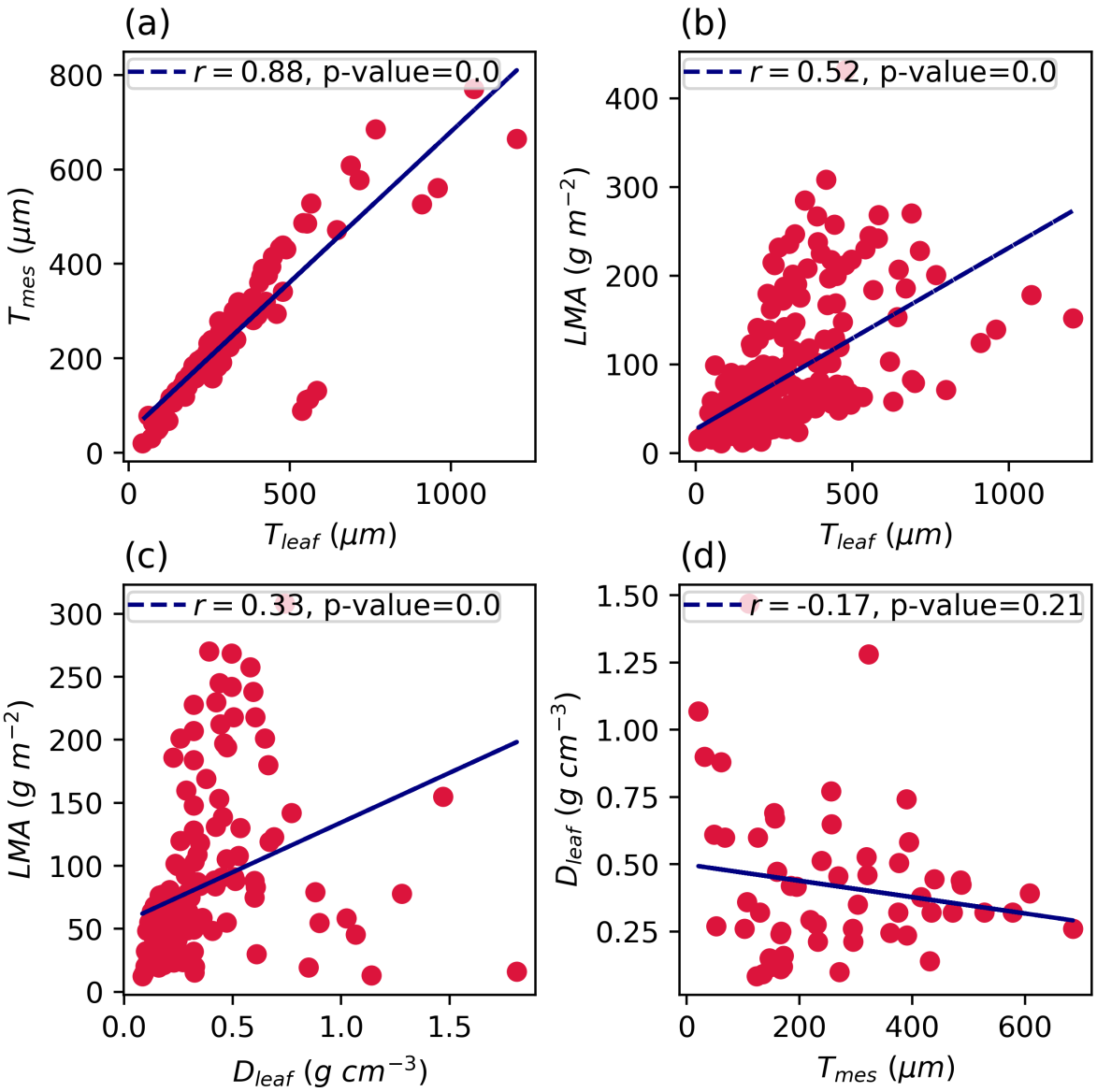

Figure S2: Linear regression between different anatomical traits. Linear regression between the pairs of anatomical traits: a) T_{leaf} and T_{mes}, b) T_{leaf} and *LMA*, c) D_{leaf} and *LMA*, and d) T_{mes} and *LMA*.

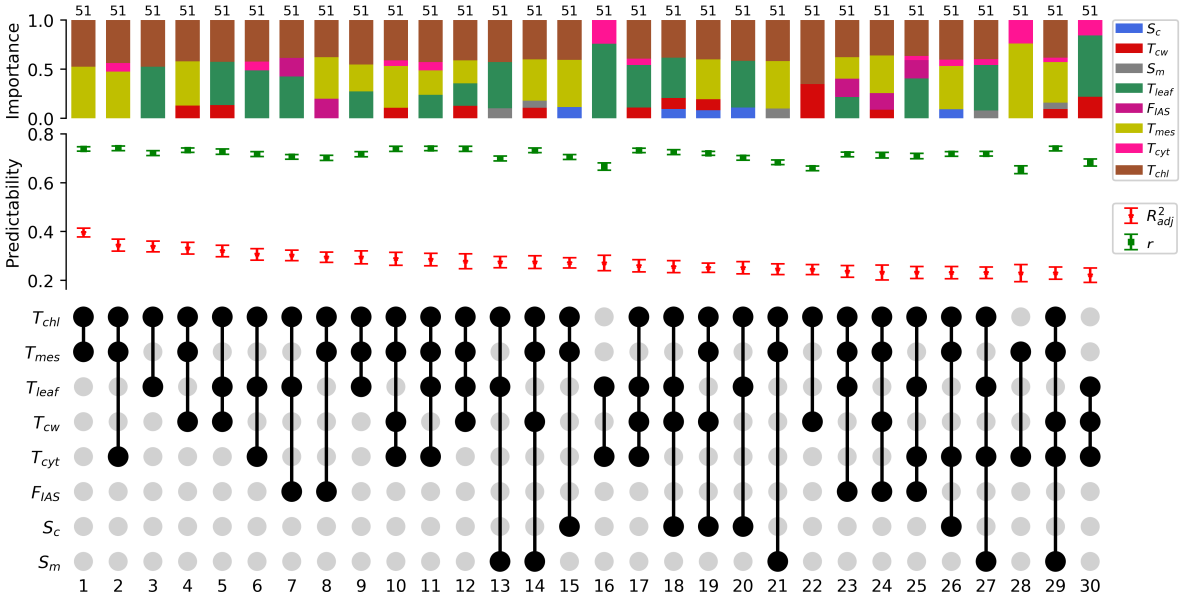

Figure S3: Predictive performance of random forest models using different combinations of leaf anatomical traits. The UpSet plot shows the predictability evaluation of top 30 models for g_m, based on average R^2_{adj}, consisting of eight anatomical and structural traits T_{mes}, F_{IAS}, T_{cw}, T_{cyt}, T_{chl}, S_m, S_c, and T_{leaf} over all available species of Xiong (2023) data set. The lower panel shows the intersection of traits contributing to the training model. The middle panel indicates the average R^2_{adj} and r between the measured and predicted values of g_m in the test set. The error bars show the standard errors of the predictability measures. The upper panel shows the average Gini importance of the corresponding traits at each combination of the traits. The number of data points in each model is provided above the importance bars. For all models, the average predictability scores were achieved by the RF model in 150 executions, with 70% randomly chosen data elements used for the training set and the remaining 30% used for the test set. The data set consists of ten C_3 crops, glycine max, oryza sativa, arundo donax, helianthus annuus, triticum aestivum, gossypium hirsutum, beta vulgaris, astragalus sinicus, lycopersicon esculentum, and solanum tuberosum, where our C_3 data set contain the first 6 species.

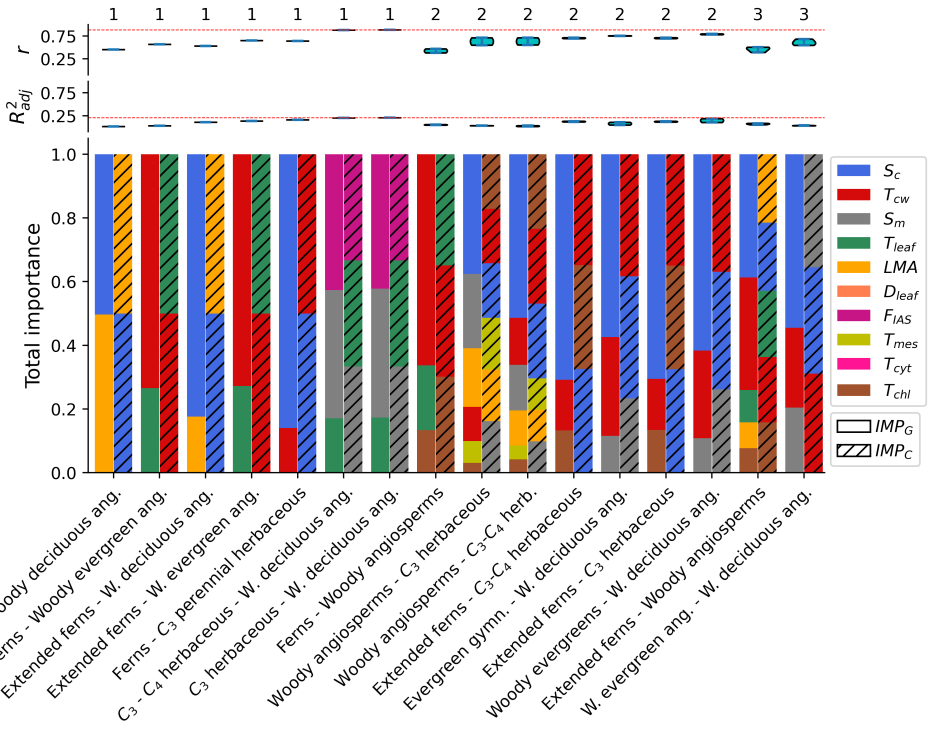

Figure S4: The predictive performance and total feature importance ratios of the trained models between PFTS. The evaluation of the performance of the models in 16 prediction scenarios with the test sets of individual PFTs and the training sets of other individual PFTs. The figure shows the results for the rest of the scenarios shown in Fig. 4b. The upper panels show the violin plots of the average predictability scores (r and R^2_{adj}) of the models for each prediction scenario. The number of models with positive R^2_{adj} in each scenario is given above the violin plots. The dashed red lines indicate the maximum of predictability scores across all the models. The bar charts in the lower panel show the total importance measures, IMP_C and IMP_G, of the contributing traits in the models of each scenario. The position of traits in the bar charts has been sorted from top to bottom based on their total importance. For all models, the average predictability scores and importance ratios were achieved by RF model in 50 executions, with the fixed training and test sets.

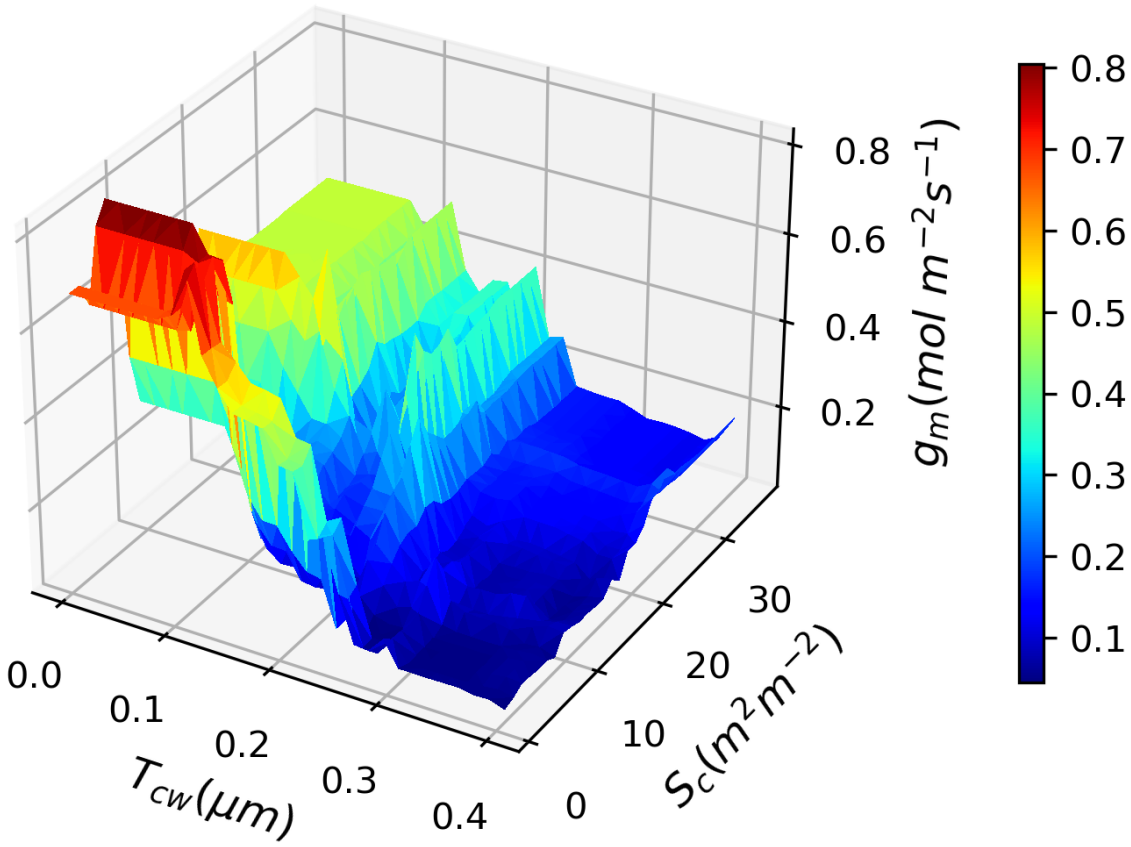

Figure S5: Variation of g_m based on T_{cw} and S_c The surface plot of g_m based on a trained model on T_{cw} and S_c. The model was built by cross-validation on the global data set with 70% randomly chosen data elements used for the training set and the remaining 30% used for the test set. After training the model, 30 equally spaced data points were determined for each of the predictors, and then g_m values were predicted for all 900 pairs of T_{cw} and S_c data points. The ranges of both predictors were based on their minimum and maximum in the test set, except for the maximum of T_{cw} where we reduced it to 0.4 to indicate the fluctuations of g_m.

Table S2: The statistics of the data for global data set and individual PFTs.

Plant functional type	No. of data samples	No. of species	No. of $n >= 50$
global data set	882	453	599
$C_3 - C_4$ herbaceous	382	116	49
C_3 herbaceous	354	93	49
Woody evergreens	302	214	72
Woody angiosperms	287	185	72
C_3 annual herbaceous	255	35	22
Woody evergreen angiosperms	176	122	63
Evergreen gymnosperms	126	92	31
Woody deciduous angiosperms	108	60	1
C_3 perennial herbaceous	99	58	2
Extended ferns	64	44	7
Ferns	58	39	1
C_4 annual herbaceous	17	12	0
C_4 perennial herbaceous	11	11	0
Deciduous gymnosperms	10	4	0
Mosses	10	10	0
Fern allies	6	5	0
Semi-deciduous angiosperms	3	3	0
CAM plants	3	2	0

The number of data samples with a value for g_m and at least one of the anatomical traits for the global data set and different PFTs are provided in the first column. The second column shows the number of species contributing to each set. The number of combinations with at least 50 data points in each set is given in the following column.