



OPEN

Deep learning and citizen science enable automated plant trait predictions from photographs

Christopher Schiller¹, Sebastian Schmidlein¹, Coline Boonman², Alvaro Moreno-Martínez³ & Teja Kattenborn⁴✉

Plant functional traits ('traits') are essential for assessing biodiversity and ecosystem processes, but cumbersome to measure. To facilitate trait measurements, we test if traits can be predicted through visible morphological features by coupling heterogeneous photographs from citizen science (iNaturalist) with trait observations (TRY database) through Convolutional Neural Networks (CNN). Our results show that image features suffice to predict several traits representing the main axes of plant functioning. The accuracy is enhanced when using CNN ensembles and incorporating prior knowledge on trait plasticity and climate. Our results suggest that these models generalise across growth forms, taxa and biomes around the globe. We highlight the applicability of this approach by producing global trait maps that reflect known macroecological patterns. These findings demonstrate the potential of Big Data derived from professional and citizen science in concert with CNN as powerful tools for an efficient and automated assessment of Earth's plant functional diversity.

Global change driven by global warming, land-cover conversion and landscape fragmentation imposes a threat on global biodiversity¹. The loss of biodiversity inevitably leads to a loss of ecosystem functioning and processes¹, which are essential to human well-being². Ecosystem functioning, in turn, can be assessed using plant functional traits (hereafter: 'traits'), since it results from the trait composition of the species that compose a plant community³. Therefore, the impacts of environmental changes on ecosystem functioning can most simply be assessed on species level using traits^{4–7}. These traits can, for instance, be related to plant size, e.g. leaf area or growth height, or tissue constituents, e.g. leaf nitrogen concentration or stem specific density. Such traits, though, have in common a high measurement effort. Thus, an effective trait measurement tool would greatly facilitate rapid ecosystem monitoring¹, which is urgently needed in light of expected trait shifts around the globe⁸.

Convolutional Neural Networks (CNN), a deep learning-oriented computer vision technique, are evolving as promising tools for ecological research^{9,10}, e.g. in plant community identification on unmanned aerial vehicle (UAV) images¹¹ or species identification by harnessing plant photograph databases¹². Particularly, the fundamental innovation of CNN is their efficiency in target-oriented learning of image features from raw input data. Hence, CNN may also enable to infer trait expressions from plant photographs by means of directly related morphological plant features or covariance with indirect causal links¹³ among visible and non-visible plant features. For instance, features such as the shape and thickness of plant leaves might be indicative of traits such as leaf nitrogen concentration and therefore be predictable by CNN.

Currently, this approach is hampered by the lack of a dataset comprising plant photographs with matching trait measurements that cover the global plant functional spectrum. In recent years, however, global data initiatives and citizen science projects have emerged as a strongly growing data source, which builds upon the collective effort of the scientific community^{14,15}. Consequently, we employ a weakly supervised learning approach that combines the independent data sources of (1) the iNaturalist database providing a worldwide record of millions of plant photographs including taxa and geolocations¹⁶ and (2) the TRY database containing more than 11 million trait measurements across more than 270,000 taxa⁷.

With this setup, we test if it is possible to infer plant traits from simple RGB photographs using CNN (setup 1, Fig. 1). As intra-specific trait variability can be substantial along environmental gradients^{17–19}, we used species-specific trait distributions rather than mean trait values in the CNN training in a second step (setup 2).

¹Institute of Geography and Geoecology, Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany. ²Department of Environmental Science, Institute for Water and Wetland Research, Radboud University, Nijmegen, The Netherlands. ³Image Processing Laboratory (IPL), Universitat de València, Valencia, Spain. ⁴Remote Sensing Center for Earth System Research, Leipzig University & Helmholtz Centre for Environmental Research (UFZ), Leipzig, Germany. ✉email: teja.kattenborn@uni-leipzig.de

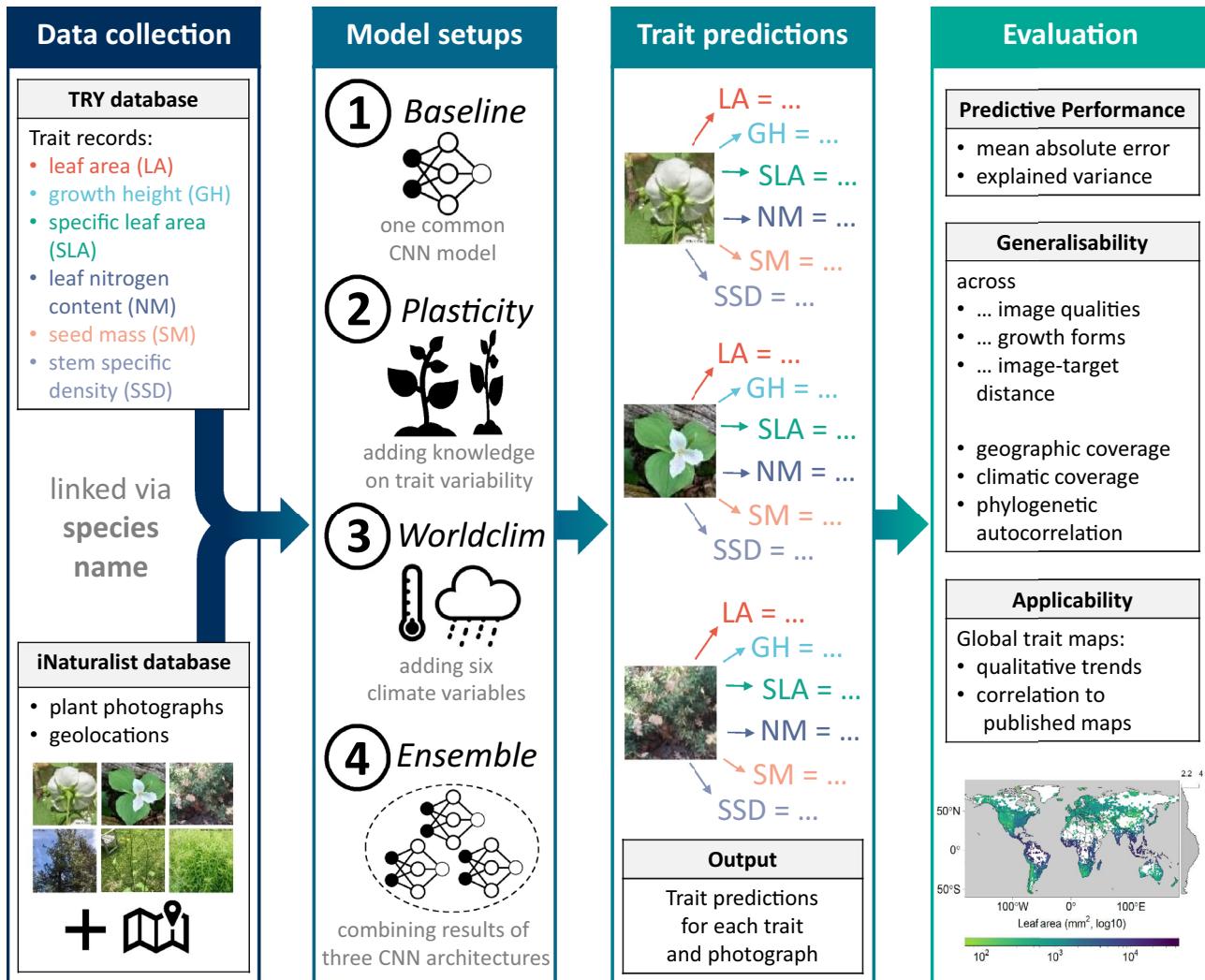


Figure 1. Conceptual diagram of the analyses, from data collection to evaluation. Data collection included linking plant functional trait records from the TRY database and plant photographs from the iNaturalist database via species names. The Convolutional Neural Networks were trained in four setups, which enabled trait predictions. The evaluation included analyses of predictive performance, generalisability and applicability. CNN, Convolutional Neural Networks.

Moreover, contextual cues might foster CNN predictive performance¹⁰. Consequently, bioclimatic variables were incorporated into the CNN due to their link to trait expressions (setup 3)^{8,20,21}. Fourthly, we assessed the effect on predictive performance of an ensemble approach combining different CNN models (setup 4)²².

In addition to a statistical evaluation on independent observations, we assessed the plausibility of the trained CNN by global trait maps derived by applying each trait model on approx. 185,000 iNaturalist observations that were not included in the training process. This highlighted both a potential future application in light of an exponential growth of the iNaturalist database providing new plant photographs day by day, and the practical value of the presented approach. Therefore, we consider this study an important pioneering work to pave the way for rapid and efficient trait assessments.

Results

We built datasets with image-trait couples for leaf area (LA), growth height (GH), specific leaf area (SLA), leaf nitrogen concentration (LNC), seed mass (SM) and stem specific density (SSD; for details see Supplementary Table 1). Some of these traits, such as LA and GH, could be readily visible by the computer vision models, whereas the others might be indirectly visible on account of the strong covariances between the traits investigated in this study¹³. Moreover, traits such as SLA and LNC might be predictable in photographs on account of the thickness, color intensity or shape of the plant's leaves. We restricted the number of images per species to prevent the model from learning species-specific trait expressions. Simultaneously, we maximized the number of species in the dataset to achieve a wide range of trait expressions and a sufficiently large dataset. We applied a stratified

| Trait | Max. images per species | N _{species} | N _{woody} | N _{non-woody} | N _{TA} | N _{training} | N _{validation} | N _{test} | N _{total} |
|-------|-------------------------|----------------------|--------------------|------------------------|-----------------|-----------------------|-------------------------|-------------------|--------------------|
| LA | 8 | 1361 | 3937 | 6096 | 7773 | 7216 | 1804 | 1013 | 10,033 |
| GH | 2 | 8161 | 6862 | 8897 | 11,040 | 11,348 | 2836 | 1575 | 15,759 |
| SLA | 3 | 4615 | 7099 | 6041 | 9351 | 9461 | 2365 | 1314 | 13,140 |
| LNC | 3 | 4339 | 7103 | 5261 | 8254 | 8903 | 2225 | 1236 | 12,364 |
| SM | 1 | 9725 | 3903 | 5822 | 5404 | 7003 | 1750 | 972 | 9725 |
| SSD | 5 | 2455 | 10078 | 685 | 7176 | 7750 | 1937 | 1076 | 10,763 |

Table 1. Summary of plant functional trait datasets. Information on the maximum number of images per species, total number of species (N_{species}), number of images with woody (N_{woody}) and non-woody species ($N_{\text{non-woody}}$), number of observations with non-zero standard deviations for Plasticity setup (N_{TA}), number of images in training (N_{training}), validation ($N_{\text{validation}}$) and test (N_{test}) dataset as well as total number of images (N_{total}) in each trait dataset. LA, leaf area; GH, growth height; SLA, specific leaf area; LNC, leaf nitrogen concentration; SM, seed mass; SSD, stem specific density.

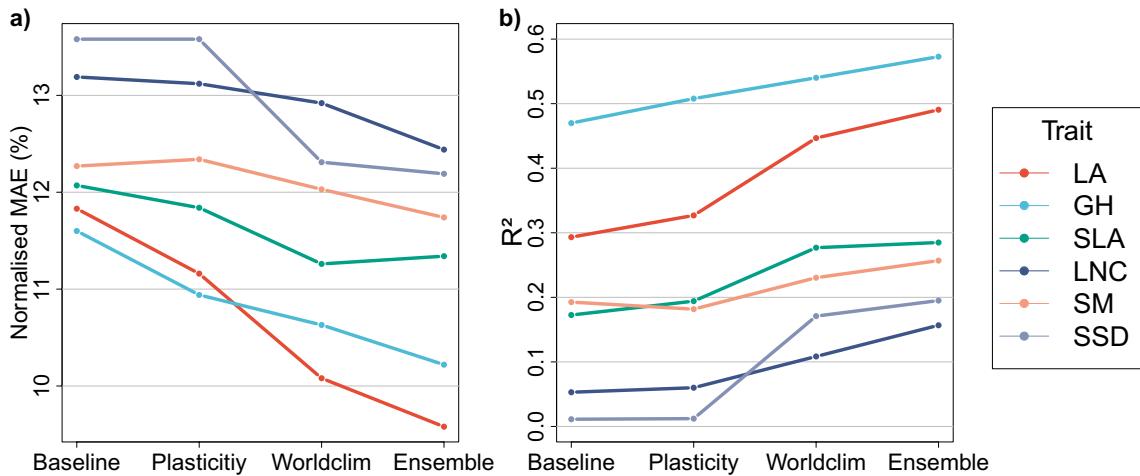


Figure 2. Results of the four model setups. Normalised mean absolute error (a) and R^2 (b) for the basic CNN model setup ('Baseline'), the setup including trait variability ('Plasticity'), the setup including bioclimatic data ('Worldclim') and the ensemble setup ('Ensemble'). Mean absolute error was normalised by the respective test dataset's range to enable a comparison between the predictive performance of the six traits. R^2 shows the explained variance of the linear fit of test predictions versus test targets.²³ LA, leaf area; GH, growth height; SLA, specific leaf area; LNC, leaf nitrogen concentration; SM, seed mass; SSD, stem specific density.

sampling design (each species being a stratum) allowing up to eight images per species to acquire a sufficient amount of training data while balancing the taxonomic evenness (Table 1).

Implementation of trait plasticity, bioclimatic data and ensembles. The prediction of the basic CNN ('Baseline' setup) yielded normalised mean absolute errors (NMAE) between 13.6% (SSD) and 11.6% (GH; Fig. 2, Supplementary Table 2). The explained variance of the linear fit did not exceed 5% for SSD and LNC, but reached up to 47% for GH. Accounting for the intra-specific variability of traits by providing a distribution of trait values instead of a single mean value for each species ('Plasticity' setup) generally improved predictive performance with the exception of SM, increasing the explained variance by up to 3.79%-points (GH). Adding knowledge on the local climate of a photograph's location ('Worldclim' setup) generally increased the predictive performance. Its effect on R^2 was higher on LA, SLA and SSD (between approx. +8 and +16%-points) than on GH, LNC and SM (between approx. +3 and +5%-points). Furthermore, the 'Ensemble' setup, in which we averaged the predictions of three common CNN architectures, generally increased R^2 , yielding a rise of more than 4%-points in explained variance for LA and LNC.

The results of a threefold cross-validation based on the ensemble models revealed that the traits characterizing leaf form or habitus - GH and LA - demonstrated the lowest NMAE (9.91% and 9.92%; for prediction errors in original unit see Supplementary Table 3) and highest R^2 (.58 and .45; Fig. 3), whereas the traits related to tissue constituents, LNC and SSD, ranked lowest (11.71% and 11.98%; .16 and .2).

Model performance vs. data heterogeneity. The assessment of the predictive performance by visual interpretation of 200 independent images per trait regarding the distance of the photographer to the target species ('image-target distance') showed that images with the longest distance attained 7.6%, whereas most images showed the shortest distance (63.5%; Fig. 4). Additionally, the vast majority of the images (92.9%) expressed a

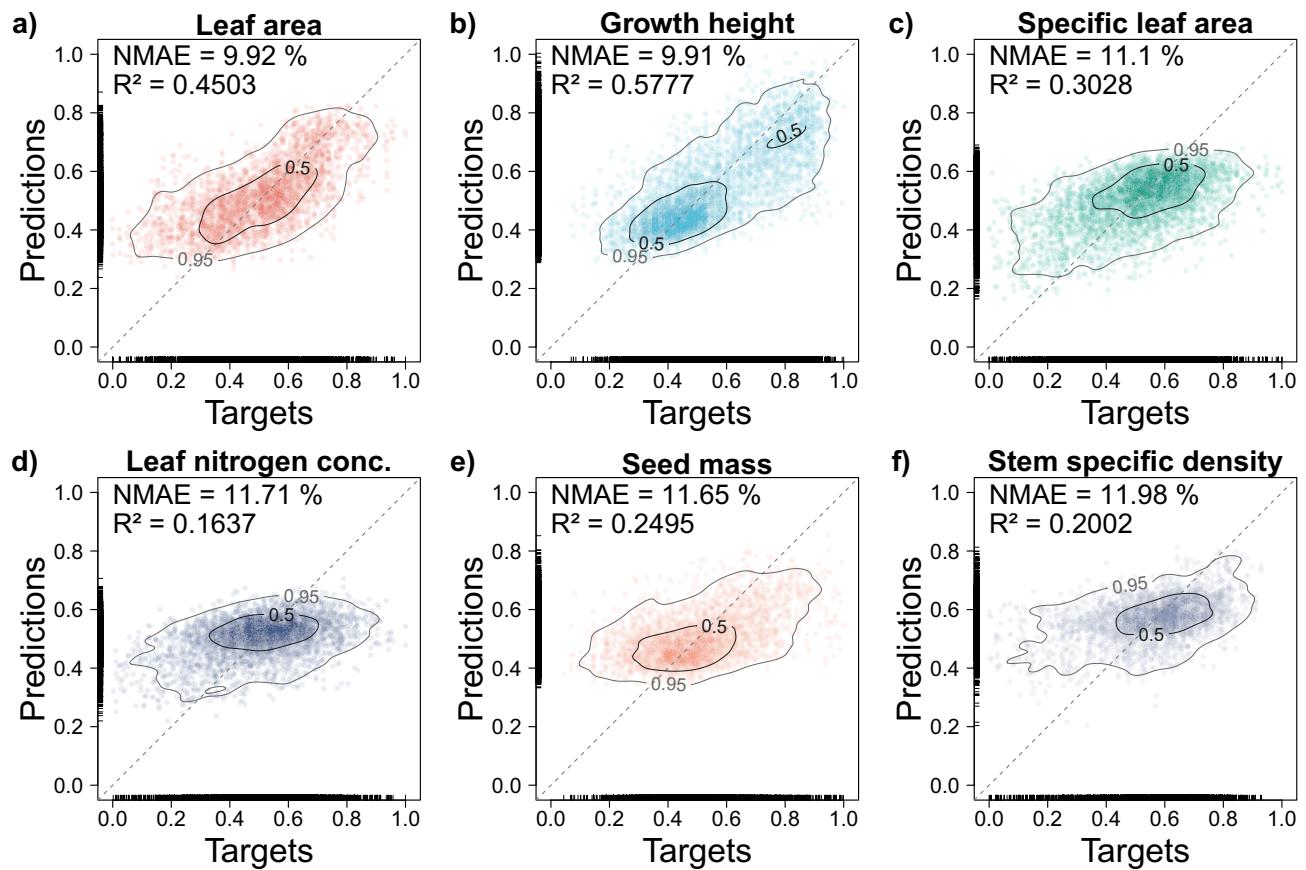


Figure 3. Distributions of predictions and targets of 3-fold cross-validation. **a-f**, All target-reference pairs ($N = 3 \times N_{\text{test}}$) of Ensemble setup of 3-fold cross-validation for the six plant functional traits leaf area (a), growth height (b), specific leaf area (c), leaf nitrogen concentration (d), seed mass (e) and stem specific density (f). Dashed grey lines indicate the one-to-one line for reference. Contour lines indicate the bivariate occurrence probabilities (50% and 95%) computed by kernel density estimation using R²³ package 'ks' (version 1.11.7)²⁴.

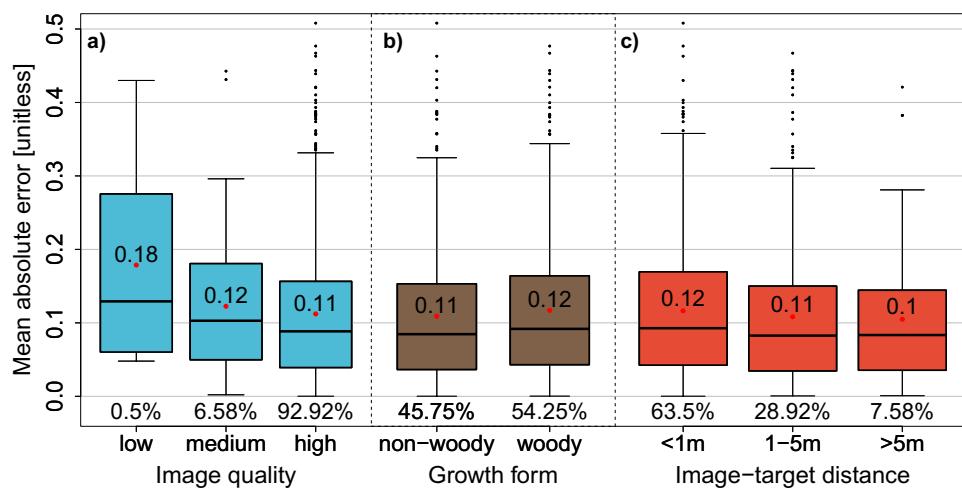


Figure 4. Model evaluation results. **(a-c)** Mean absolute errors across different image qualities (a), growth forms (b) and image-target distances (c) concerning 200 images per plant functional trait ($N = 1200$ images). None of the images of the evaluation dataset were used in the training phase. The percentage of images falling in a category is indicated below each box plot. Red dots and the number linked to them are the mean absolute errors for each category. None of the groups differ significantly on $p < .05$ (Supplementary Table 4). ²³

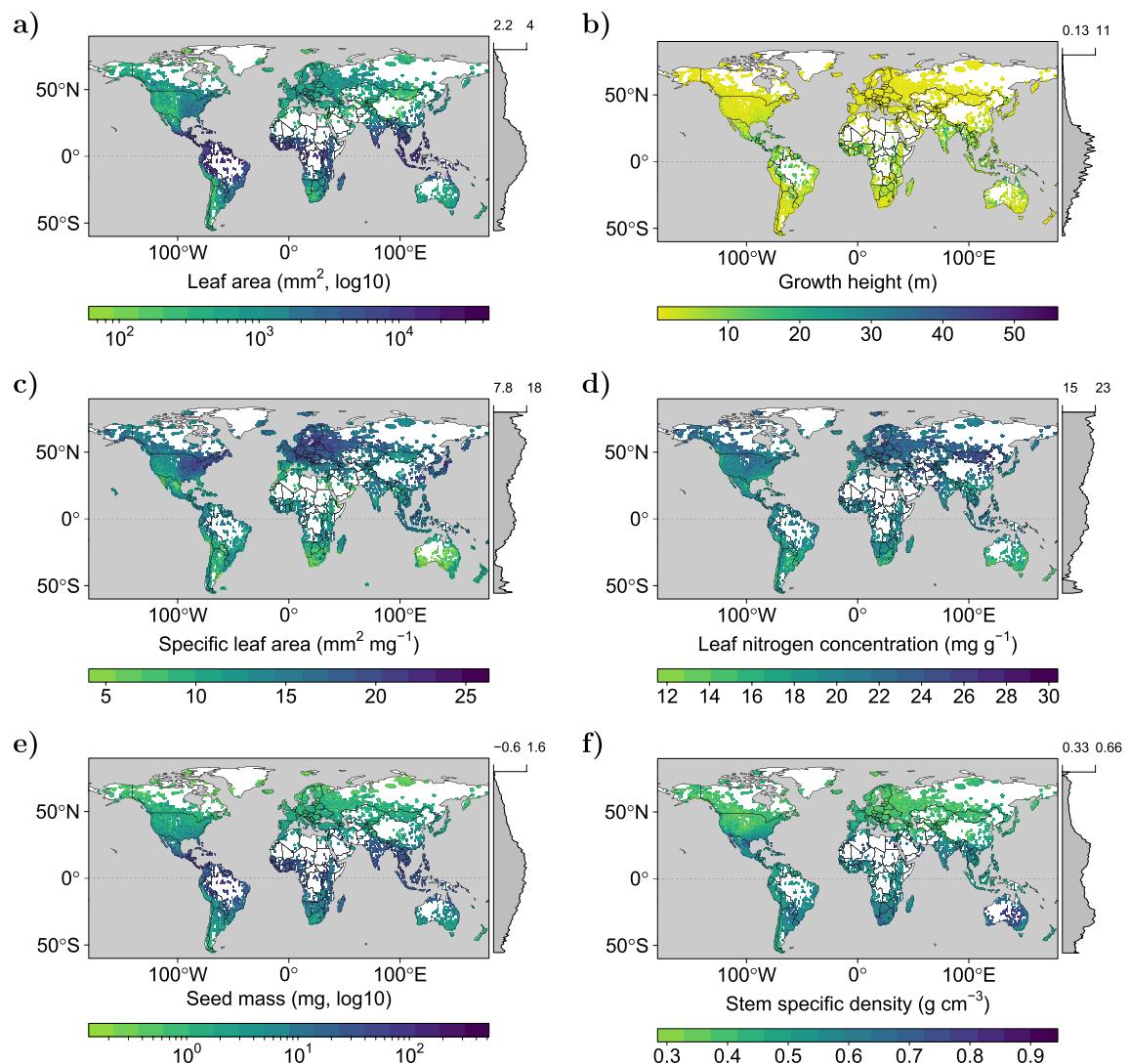


Figure 5. Global trait distribution maps. Global maps of mean plant functional traits produced by inverse-distance weighted interpolation on ensemble predictions for leaf area (a), growth height (b), specific leaf area (c), leaf nitrogen concentration (d), seed mass (e), stem specific density (f) including latitudinal distribution. Values of leaf area and seed mass were \log_{10} -transformed for improved visualisation.

high image quality, meaning that all plant organs appearing in the image were clearly recognisable. The assessment of the growth form (woody vs. non-woody) of the images' target species revealed that the share of woody and non-woody species was approximately equal. None of the groups in Fig. 4 differed significantly on $p < .05$ (Supplementary Table 4).

Global trait distribution maps and their validation. Global trait distribution maps (GTDM) derived from model predictions on more than 185,000 independent plant photographs, which were not used for model training, expressed a unimodal latitudinal distribution peaking around the equator for LA, GH and SM (Fig. 5). SLA and LNC expressed their highest values around northern temperate and polar zones as well as the equator. SSD showed a bimodal distribution with peaks in the subtropics. All trait expressions were similar along the equator. A gradient from high to low LA and SLA was found from the east of North America towards the west, while an opposite trend was observed for LNC and SSD. A quantitative comparison with published GTDM revealed significant correlations with Pearson's $r > .5$ concerning GH²⁶, SM⁶ and SSD^{6,26}, whereas insignificant or small correlations were found between all of the four GTDM concerning LNC^{26–28} (Supplementary Fig. 1). For SLA, our GTDM correlated significantly to ref.^{6,26} ($r > .5$), whereas correlations were negative or insignificant between the former and ref.^{27,28}.

Discussion

Our results suggest that certain plant functional traits can be retrieved from simple RGB photographs. The key for this trait retrieval are deep learning-based Convolutional Neural Networks in concert with Big Data derived from open and citizen science projects. Although these models are subject to some noise, there is a wealth of applications for this approach, such as global trait maps, monitoring of trait shifts in time and the identification of large-scale ecological gradients. This way, the problem of limited data that still impedes us to picture global gradients⁷ could be alleviated by harnessing the exponentially growing iNaturalist database¹⁶. The performance of the CNN models across traits varied strongly, but revealed a clear trend: As expected, the more a trait referred to morphological features, the more accurate the predictions were. The models of the Baseline setup explained a substantial amount of the variance for LA and GH, whereas traits that are partly related to morphological features, SLA and SM, show moderate R^2 values. The predictions of LNC and SSD explain almost none of the variance, suggesting that tissue constituents are not directly expressed or related to visible features. It also indicates that the strong covariance among these traits¹³ does not suffice in supporting their predictions from photographs. If the RGB images do not contain relevant information, the model will minimise the prediction errors by the regression-to-the-mean bias seen in Fig. 3 (especially lower panels).

The value of informing the model on the known trait variability through an augmentation of the target values (Plasticity setup) depended on the results of the Baseline setup of the trait. That is, the better the predictive performance of the Baseline setup, the more the trait seemed to profit from the Plasticity setup, rendering it ineffective for SSD, LNC and SM (Fig. 3). Refraining to cling to species mean values by considering within-species trait variation has been applied before using conventional methods²⁶, but to our knowledge has never been tried in CNN models, yet. We expected that providing a distribution of trait values rather than a single mean for each species can convey to the CNN that different trait realisations can be expected from the same species. Obviously, this idea can only work if a distribution rather than a single value is available for each species. The SM dataset, for instance, contained only one image per species (Table 1). In this case, the Plasticity setup reduced the predictive performance compared to the Baseline setup, possibly by increasing the discrepancy to the true trait value. Since the traits with more accurate predictions profited most from the Plasticity setup, we assume that it supports the model in learning to predict the trait expressions themselves rather than extracting them indirectly through taxa-specific morphological features. Given that we restricted the number of images per species to a maximum of 8, while successful deep learning-based plant species identification usually requires thousands of images^{12,22}, it seems very unlikely that the models inferred traits from species-specific plant features visible in the imagery. The latter was underpinned by our finding that the predictions of most traits are void of phylogenetic autocorrelation (Supplementary Information 1 and Supplementary Table 5), indicating that taxonomic relationships were insignificant for the trait predictions. The absence of phylogenetic autocorrelation of the prediction errors underlines that the models did not learn species-specific features for most traits, as this would imply similar trait predictions for related species.

On the contrary, the SSD model predictions express a phylogenetic signal (Supplementary Information 1 and Supplementary Table 5). Trait expressions are generally clustered under similar climatic conditions^{29–31}. Simultaneously, climatic conditions constrain the geographic distribution of species and growth forms^{29–32}. The SSD dataset is biased towards woody species (Table 1), which confines it to a smaller taxonomic range. Hence, the phylogenetic signal of SSD might result from its phylogenetic clustering and predominant dependence on bioclimatic information rather than on RGB imagery (Fig. 2).

Nevertheless, the benefit of including climate information on temperature, precipitation and their seasonality^{8,20,21,26} on predicting trait expressions was confirmed for all traits in this study, which underlines the value of contextual constraints in CNN models¹⁰ (see below for a discussion of the relevance of climate vs. image data). This also highlights the general flexibility of deep learning frameworks in adapting to variable input data from different scales and sensors¹⁰, which makes them a promising tool for ecological research. Our results particularly revealed this effect for SSD, SLA and LA, whereas it was smaller for GH, LNC and SM (Fig. 2). For the latter traits, other physical constraints such as disturbance^{33,34}, seasonal variation^{35,36} and soil conditions^{6,26,28} come into consideration. As the focus of the Worldclim setup was to show that contextual cues can improve the trait retrieval from photographs rather than identifying the best set of auxiliary data, we confined the analysis to the most promising^{20,26} data source (WorldClim³⁷).

In the Worldclim setup, a single model accumulated knowledge about the trait learning task. Combined predictions of different CNN models, however, have shown to surpass the predictive performance of single CNN, e.g. in plant species identification tasks²². Each of the CNN models is prone to literally ‘look’ at different aspects of the learning task by focusing on different image features. Previous research also showed increased model performance in a trait prediction task in case of ensembles of regression and machine learning models²⁶. Accordingly, and as demonstrated by our results, an ensemble approach seems promising to further enhance predictive performance of CNN models concerning trait prediction.

The predictive performance of these Ensembles has shown to be reproducible with different sets of training images (cp. Figs. 2, 3, Supplementary Fig. 2). In our heterogeneous dataset, model performance was not affected by different growth forms, image qualities and image-target distances (Fig. 4). Different growth forms and plant functional types show their own characteristic trait spectrum¹³. Possibly, contextual cues within the image might have supported the CNN in inferring the plant functional type of a species, e.g. by a long-distance image being indicative of a tree species. Yet, since the majority of the images only shows single plant organs on close-up photographs (Fig. 4), we assume that the trait predictions are not confounded by the identification of growth forms. Furthermore, the absence of a phylogenetic signal in the prediction errors for most traits highlights the model’s ability to generalise by extracting trait information independent of taxonomic relationships, meaning that the

models (except for SSD) did not learn species-specific mean trait expressions (see Supplementary Information 1 and Supplementary Table 5).

Additionally, we disclosed the high generalisability of these results by investigating the datasets' underlying distributions both spatially (Supplementary Fig. 3) and across biomes (Supplementary Fig. 4). Although some regions such as Central Europe and North America show higher data coverage, the datasets used for this study contain data across all biomes and regions on Earth. Therefore and despite this clustering, we expect the models to be applicable for all biomes around the globe. This was highlighted by an additional analysis showing that the predictive performance of the models is reasonably constant across biomes (Supplementary Fig. 5). As suggested by refs.^{38,39}, we tolerated a certain spatial bias in favor of larger datasets. Although the SSD dataset predominantly contained woody species, neither of the six datasets expressed an exclusion of either growth form (Table 1, Fig. 4).

The application of our models to global gradients of traits revealed that our GTDM indeed cover macroecological patterns and trends known from other publications: The latitudinal distributions could roughly be confirmed for GH²⁶, LNC^{26,27} and SM^{6,8} (Fig. 5). Predicted trends for maximum leaf size hint at the applicability of our GTDM of LA⁴⁰. The trait gradients for North America were confirmed for SLA^{6,8,26–28}, SM^{6,8}, LNC²⁷ and SSD⁶ alike. Although based on different input data and modelling methods, the major global latitudinal gradients found in previous studies could be reproduced by our GTDM, which indicates the plausibility of the latter^{6,8,26,27}.

We further validated the GTDM quantitatively by means of correlations with other GTDM. Regarding SSD, the detected high correlations might be due to method similarity, as our GTDM product of SSD primarily builds upon climate data (see above), just as ref.^{6,26}. For GH, SLA and SM, however, the high correlations are unlikely to result from climate data exclusively, as the explained variances of the RGB imagery (R^2 of Plasticity setup) are higher than the additional contributions of the Worldclim setup (approx. 94%, 70% and 79% share of imagery on total explained variance, respectively; Supplementary Table 2). We decoupled the GTDM products from bioclimatic information in an additional analysis (Supplementary Fig. 6). Remarkably, the macroecological patterns could roughly be reproduced when the GTDM were based exclusively on RGB imagery, which shows that the bioclimatic information merely serves to smooth the macroecological trait patterns for most of the traits.

Despite of all GTDM being at least partly build upon climate data and using trait data from the same source (TRY database), some GTDM of SLA and all GTDM of LNC vary strongly in their correlations (Supplementary Fig. 1). On the one hand, this might indicate that LNC varies at a different scale, e.g. on account of its seasonal and within-species variation^{35,36}. On the other hand, other GTDM products are based on mean trait values weighted by abundances of plant functional types^{27,28} rather than single trait predictions, which might explain negative or non-significant correlations as well.

Hence, a potential pitfall of the presented approach is that it is prone to express an observation bias, e.g. by citizen scientists only taking pictures of the most striking species. The sampling design underlying the GTDM does not account for plant community composition, meaning that we cannot tell if plant photographs at a certain location represent the actual community structure. Since many images contain more than one individual plant and different species, the CNN model predictions, however, might be based on more than one species, thereby partly resembling trait expressions of the community. The representativeness of trait data for plant communities, though, remains an ubiquitous problem of global trait maps, including those fully based on trait data from the TRY database⁷, since every available dataset is far from representing the actual plant community composition⁷. Hence, at present our GTDM have to be considered a plausibility check of the model predictions rather than an application-ready trait map product, not least because the sampling of images might not be representative of the respective plant community.

Nevertheless, our results indicate that exploiting a Big Data approach is viable to reveal macroecological trait patterns, maybe because the most striking species of an ecosystem are likely to suffice in describing its functional footprint⁵. Since the strong growth of the iNaturalist database leads to a steadily increasing geographic coverage, the representativeness of these data is likely to grow as well. A recent study investigating the records of FloraIncognita¹², a citizen-science and deep learning-based application for identifying plant species from photographs, suggested that such crowd-sourced data can reproduce primary dimensions in plant community composition⁴¹. This underlines the future potential of harnessing citizen science databases for identifying these patterns. Here, we demonstrated the practical value and applicability of the CNN models by producing GTDM that were able to reproduce known macroecological trait patterns while displaying one anticipated application of this method. Additionally, in these GTDM we bypass the issue of spatial error analysis that is challenging for most GTDM products²⁶ by obtaining a potentially arbitrary number of observations in light of the strongly increasing number of observations in iNaturalist, almost rendering an extrapolation obsolete. Our GTDM are based on individual trait measurements rather than estimated on behalf of a small set of covariates, which is typical for climate-based GTDM²⁶. Since plant traits vary strongly within species^{17–19}, these measurements express a high practical relevance. As the iNaturalist plant photograph database is witnessing an exponential growth of data inputs, the potential of exploiting this data source for plant trait predictions is growing rapidly. It is worth mentioning that this approach also led to the first publication of a GTDM of mean LA (available for download under <https://doi.org/10.6084/m9.figshare.13312040>), since former publications were limited to modelling upper limits of LA based on climatic constraints⁴⁰.

Future studies building on our work, which benefit from the ever-growing data accumulation of both the iNaturalist and TRY database, might not face restrictions of dataset size as we did in our study. This might allow for more representative samples in future studies, e.g. enabling to stratify training data by species while simultaneously balancing the trait distribution. This might support a reduction of the regression-to-the-mean bias seen in all of the results (Fig. 3) by avoiding to overrepresent common trait expressions. Another possible approach would be to select only species with particularly low variability for model training, since it decreases the chance of incorporating images showing plants with an extreme trait expression that differ strongly from the chosen

mean trait values from TRY. By that, we might be able to derive more reliable and accurate predictions in the context of weakly supervised learning by reducing noise in the training data.

Although weakly supervised learning approaches generally have shown to be an effective way of compensating a shortage of individually labeled data^{42,43}, an image dataset including in-field trait measurements under natural conditions representing the global trait spectrum would be necessary for a conclusive validation. In our study, it even remains unclear to what extent the trait values actually refer to the individual plant shown in an image, particularly as the images sometimes show more than one individual plant and more than one species (Supplementary Fig. 7). This may hinder the model from predicting a trait value corresponding to the dominant species in the image (but might also partly resemble the community composition, see above). Although we attempted to compensate the lack of a dataset that enables a conclusive validation by eliminating possible biases concerning image settings (Fig. 4), growth forms (Fig. 4), phylogenetic autocorrelation (Supplementary Information 1, Supplementary Table 5), predictions based only on climate data (Supplementary Fig. 6), predictive performance across biomes (Supplementary Fig. 5), a training dataset subject to limited geographic or climatic coverage (Supplementary Figs. 3, 4) and effects of a specific set of training data (Supplementary Fig. 2), we cannot conclusively prove that the model predictions are based on causal relationships. Our model results suggest that the trait predictions reflect the feature space of natural trait expressions (Fig. 3), but an in-depth analysis of the image features the models learned for inferring the respective traits will be necessary to rule out any possible biases in future studies. An explicit analysis might involve investigating which plant organs are relevant for the trait predictions by means of feature attribution techniques and could ultimately provide clear evidence. This may not only enable to build trust in such artificial intelligence (AI) models, but also to generate new knowledge from them in order to deepen our understanding of plant morphology and trait covariance.

Nevertheless, this study can only be considered a pioneering work testing the feasibility of the approach, as application-ready models require a conclusive and explicit validation. A dataset enabling this has to incorporate image-trait pairs measured and photographed on the same individuals. One possible solution would be to generate a database of plant traits including respective photographs, which then can serve as a benchmark for future studies.

Conclusion and outlook

Following the urgently needed transition of ecology towards a data-sharing scientific discipline^{15,44} and the call for integration of powerful datasets in ecology⁴⁴, we built upon this revolution of Big Data provided by professional and citizen science alike. Therefore, we exploited the potential of Convolutional Neural Networks⁴⁴ to produce generic models generalising across all biomes and regions around the globe to extract a set of plant functional traits from simple RGB photographs. Thus, the burden of limited geographical coverage of trait databases might be lifted by the worldwide coverage of the iNaturalist plant photograph database. The traits referring to the models with the highest predictive performance cover the primary axes of plant form and function¹³, which are plant and organ size (GH, LA) as well as the leaf economics spectrum (SLA). Despite of the disputable model performance for some traits, our results highlight the potential of this approach to facilitate non-invasive, cost-efficient and automated assessments of functional gradients in future real-world applications. In order to achieve application-ready models, however, a conclusive validation is mandatory and has to incorporate image-trait pairs derived from the same individual plants. Once this is done, future applications of this approach might be next generation global trait maps^{33,34} as input for modelling and monitoring ecosystem processes and biochemical cycles, trait monitoring on time series data from PhenoCam images⁴⁵ on local scales, or assessments of ecosystem functions on landscape-scale through high resolution imagery from drones¹¹. Each of those applications might empower researchers to further close the gap between actual and intended spatio-temporal coverage, which ecology has been falling short of for decades⁴⁶.

Methods

Data acquisition and preparation. We acquired trait records of the six plant functional traits leaf area (LA), growth height (GH), specific leaf area (SLA), leaf nitrogen concentration (LNC), seed mass (SM) and stem specific density (SSD) from the TRY database⁷ (see Supplementary Table 1 for details). These traits explain most of the global trait variation in plants¹³ since they directly relate to plant's nutrient economics and competitive abilities^{3,4,13}. Accordingly, these traits are among those with the highest data coverage in the TRY database⁷. The traits LA and GH could be readily visible by computer vision, whereas traits such as SLA and LNC might be predictable through visible plant features such as thickness, color intensity or shape of leaves. Moreover there is a strong covariance between all of these traits, for instance a high GH usually implying a high SSD^{3,13}. We utilised the standardized values given in the TRY database, which have been converted to uniform units. Observations with a difference to the trait mean value of greater than 4 were removed as recommended in the TRY release notes, since these observations are considered outliers. Next, the mean and standard deviation was computed for the six plant traits for each species over all observations.

Plant photographs together with their geolocation were downloaded from the iNaturalist database (Research-grade Observations) via the Global Biodiversity Information Facility (GBIF)⁴⁷. The observations containing presumed wrong species names and geospatial issues such as missing coordinates and a coordinate uncertainty of more than 100 km were removed from analysis. Next, the geolocations from the photographs' metadata were used to extract bioclimatic variables from the Worldclim database with a resolution of $0^{\circ} 10' 0''$ ³⁷. We chose the following bioclimatic variables: Annual Mean Temperature (BIO1), Temperature Seasonality (BIO4), Temperature Annual Range (BIO7), Annual Precipitation (BIO12) and Precipitation Seasonality (BIO15). Additionally, we computed the Precipitation Annual Range (BIO13–14) by subtracting Precipitation of Wettest Month (BIO13) by Precipitation of Driest Month (BIO14). BIO1 and BIO12 are known predictors of plant traits^{6,17,20,21}. Similarly,

climate variables referring to range and seasonality have shown to coincide with plant traits^{8,26}. Therefore, we selected the other four climate variables (BIO4, BIO7, BIO13–14, BIO15) to characterize annual variation of climatic conditions of the respective site. Photographs that could not be linked to climate data, e.g. because of geolocations off the land surface, were removed from the analysis.

Sampling and pre-processing. Based on the species names, we linked the trait observations obtained from the TRY database (species-specific mean and standard deviation) with the plant photographs (iNaturalist) already linked to bioclimatic variables (Worldclim). For the purpose of balancing the dataset, i.e. avoiding over-represented species, while maximizing the overall dataset size, we included at least one but at maximum eight observations per species (Table 1). We sampled a wide range of species rather than focusing on an equal distribution of trait expressions, since we wanted to prevent the model from learning species-specific trait expressions. Therefore, the number of images per species was kept as low as possible, and the number of species in the dataset as high as possible, while attempting to achieve a sufficiently large dataset of at least 10,000 images. Next, we extracted a random sample of 10% of the dataset of each trait before model training. This ‘test dataset’ was not involved in the training process and exclusively served for the independent evaluation of the trained models. The remaining data was split into ‘training dataset’ and ‘validation dataset’ by a ratio of 4:1 (Table 1). The training dataset was employed to train the weights of the CNN model, whereas the validation dataset indicated the training progress after each full training cycle (‘epoch’).

We clipped the images to be quadratic by removing the spare margins and down-sampled the resulting image to a resolution of 512×512 pixels. Further pre-processing included \log_{10} -transformation of the reference trait data (‘targets’) due to skewed distributions and removing outliers exceeding three standard deviations above or below the mean. Afterwards, we normalised all targets (training, validation and test datasets) by the minimum and maximum values of the training dataset according to

$$\text{target}_{\text{norm}} = \frac{\text{target} - \min_{\text{train}}}{\max_{\text{train}} - \min_{\text{train}}}, \quad (1)$$

target denoting the \log_{10} -transformed target value, and \min_{train} and \max_{train} being the minimum and maximum of the \log_{10} -transformed training dataset. Note that the minimum and maximum values used for normalising the targets were derived from the training dataset exclusively, preventing a leakage of information of the validation and test datasets to the training process. As a result, the final target values ranged exactly (for validation and test datasets: approximately) between 0 and 1. The same normalisation scheme was undergone for the bioclimatic variables.

Convolutional neural network setups. CNN, a sub-group of deep learning models for image analysis, are designed to harness the spatial context of image pixels for a specific modelling problem. The basic structure of CNNs includes a cascade of convolutions, i.e. optimizable filter operations, for extracting activation maps and down-sampling (‘pooling’) operations, which enable to perform these operations at multiple spatial scales. Hence, the feature maps derived this way can contain both low- (e.g. edges) and high-level features (e.g. leaf shapes) of the image. Depending on the target value, the CNN learns which features are relevant and aggregates this information to a specific prediction in the last layer. In general, CNNs are computationally efficient and do not require a manual feature design process, making them readily adaptable to many image interpretation tasks. Former studies revealed their applicability in vegetation science tasks such as plant species identification using plant images^{12,22,48} and identification of plant communities utilising imagery from unmanned aerial vehicles (UAV)¹¹. In view of our research objectives, we tested the potential of CNN for plant trait retrieval with four different setups (Fig. 1):

- (1) As a baseline, a state-of-the-art CNN architecture called Inception-Resnet-v2⁴⁹ was used (‘Baseline’ in Fig. 2, setup 1 in Fig. 1).
- (2) In order to test if prior knowledge on within-species variability can improve the weakly supervised learning process, ‘Plasticity’ (or ‘target augmentation’, TA) was implemented with exactly the same model configuration as in setup 1 (‘Plasticity’ in Fig. 2, setup 2 in Fig. 1). Plasticity was realised by harnessing the standard deviations obtained from TRY database regarding every trait and species (for information on data availability, see Table 1). The target values were altered within a Gaussian distribution truncated by one standard deviation of the trait values using R package ‘truncnorm’ (version 1.0–8)⁵⁰, thereby leaving a small deviation from the mean value more likely than a large one. To avoid negative as well as overly large values, the targets were clipped to the range between 0 and 1 after normalisation.
- (3) As a contextual constraint, bioclimatic data (see above) was fed into the CNN with the same model and data configuration including Plasticity in a mixed data approach (‘Worldclim’ in Fig. 2, setup 3 in Fig. 1).
- (4) We tested an Ensemble approach, in which two more state-of-the-art model architectures, namely Xception⁵¹ and MobileNetV2, the latter with halved number of trainable parameters⁵², were trained on the configuration of setup 3, and their predictions were subsequently averaged (‘Ensemble’ in Fig. 2, setup 4 in Fig. 1). These models differed strongly in their number of trainable weights, resulting in a different depth. The final model performance was assessed using a 3-fold cross-validation, with three different training, validation and test splits. To enable a comparison of model performance across traits, the resulting mean absolute error (MAE) was normalised by division over the range of the target values of the respective test dataset (‘normalised mean absolute error’, NMAE).

Training process and hyperparameters. In order to build upon a pre-existing knowledge base, we employed ‘transfer learning’ by using pre-trained layer weights (the storage of the model’s knowledge) from a classification task on a dataset on www.image-net.org³⁸ for all CNN models used in this study. The regressor following the basic CNN consisted of a global average pooling layer followed by two dense layers with 512 and 1 output units. The latter forces the CNN to output exactly one prediction (trait) value. In case of the mixed data model (setups (3) and (4)), the CNN consisted of parallel branches to incorporate the different input data types. The branch processing the bioclimatic data consisted of three dense layers with 64, 32 and 4 output units, and the last layer of the CNN regressor contained 4 output units. After concatenating the two branches (image and bioclimatic branch), the regressor contained four dense layers with 8, 8, 4 and 1 output units. The last dense layer of each branch and the final layer of the model were linearly activated, whereas for all other dense layers, a ‘relu’ activation function was utilised. The latter enables the model to use non-linear separation boundaries of the feature space. We determined this configuration by its model performance in preliminary runs.

We increased the robustness and transferability of the model predictions by means of ‘image augmentation’ (also: ‘data augmentation’), which works independent of the choice of the specific CNN architecture⁵³. The image augmentation procedure serves to inflate the amount of training data, while simultaneously assisting the CNN to learn spatial features independent of the data acquisition, e.g. camera settings. Therefore, the images of the training data were subjected to horizontal and vertical flipping as well as adjusting the contrast, saturation and brightness (between factors of .9 and 1.1 each) of the images. After these adjustments, the pixel values were clipped to a range between 0 and 1 in order to prevent the value range from being enlarged.

For the training process, a batch size of 20 images and an RMSprop optimiser with a learning rate of 0.001 and a learning rate decay of 0.0001 was used. The chosen loss function was mean squared error, while the prediction accuracy was quantified by the MAE of the respective dataset. The MAE of the validation dataset was computed after each epoch. Models were trained until the validation MAE did not further improve compared to the preceding epochs and diverged from the training MAE (‘overfit’). The trained model was then applied to the test dataset.

All CNN were implemented using the Keras API version (2.3.0.0)⁵⁴ and the TensorFlow backend (version 2.2.0)⁵⁵ in R (version 3.6.3)²³. Model training was undergone on a workstation with two CUDA-compatible NVIDIA GPUs (GeForce RTX 2080 Ti, CUDA version 11.0).

Evaluation. We tested the robustness of the CNN models in view of the heterogeneous dataset on 200 images per trait retained before training. Therefore, we extracted three criteria for each image: 1) We allocated the growth form (woody vs. non-woody) to the image using information from the TRY database. 2) The first author assessed the image quality in three categories (low, medium and high) as well as 3) the distance of the photographer to the target species in the image (‘image-target distance’, categories: < 1m, 1–5m and > 5m) by visual interpretation (see Supplementary Information 2 for details and Supplementary Fig. 7 for example images). We predicted the respective trait values for these images using the trained Ensemble models and tested for significant differences of the MAEs across the three criteria above.

Global trait distribution maps. We acquired further images including geolocations from the iNaturalist database via GBIF (see above) with a stratified sampling design, attempting to get the most even distribution possible by countries on Earth. Pre-processing, including bioclimatic data, was undergone as described above. We removed duplicates with the training and validation datasets of the respective trait in order to ensure that all of the images were unknown to the CNN model to avoid artefacts from the training process. This resulted in 188,156 (LA), 188,318 (GH), 186,873 (SLA), 187,115 (LNC), 188,523 (SM) and 185,688 (SSD) images. Next, we employed the trained Ensemble models to predict each of the six traits. Afterwards, we re-transformed the predictions to yield the traits in their original unit by reconverting Eq. (1). Then, we applied an inverse-distance weighting interpolation for each trait on all of the predictions using their geolocations. The result was averaged for each cell of a grid with a resolution of 0° 30' 0" that was superimposed on the world map in WGS84 coordinate reference system using the R package ‘raster’ (version 3.3-13)⁵⁶. To minimize uncertainties associated with extensive extrapolations²⁶, we masked the interpolation output with a buffer of 100 km around each observation. Additionally, we excluded all grid cells that did not fall within the Earth’s landmass. For the same grid cells, the .9 and .1 quantile was computed and their difference was mapped as the quantile range for reference (see Supplementary Fig. 8). All trait maps were produced with R package ‘rasterVis’ (version 0.49)²⁵.

Quantitative plausibility check of global trait distribution maps. We obtained the GTDM from ref.^{6,26–28} in order to check the plausibility of our GTDM. All maps had the same coordinate reference system (WGS84), but were resampled to the same grid as our maps by bilinear interpolation using R package ‘raster’ (version 3.3-13)⁵⁶. The Pearson correlation coefficients and their significance value was computed for all available GTDM combinations and plotted using the R package ‘corrplot’ (version 0.84)⁵⁷.

Data availability

The image data (Fig. 5 and Supplementary Fig. 8) and all CNN models of the Ensemble setup supporting the findings of this study are available in ‘figshare’ with the identifier <https://doi.org/10.6084/m9.figshare.13312040>. Additionally, the raw data tables containing the download links for the plant images as well as the mean trait values that were the basis for further processing are available on figshare (<https://doi.org/10.6084/m9.figshare.14410379>). The raw image dataset can be obtained from iNaturalist database via <https://doi.org/10.15468/ab3s5x>⁴⁷. Raw trait data are available upon request from the TRY database (<https://www.try-db.org/>)⁷.

Code availability

The code supporting this manuscript is available online at https://github.com/ChrSchiller/cnn_traits.

Received: 29 January 2021; Accepted: 30 June 2021

Published online: 12 August 2021

References

- Cardinale, B. J. *et al.* Biodiversity loss and its impact on humanity. *Nature* **486**, 59–67. <https://doi.org/10.1038/nature11148> (2012).
- Díaz, S., Fargione, J., Chapin, F. S. & Tilman, D. Biodiversity loss threatens human well-being. *PLOS Biol.* **4**, e277. <https://doi.org/10.1371/journal.pbio.0040277> (2006).
- Bruelheide, H. *et al.* Global trait-environment relationships of plant communities. *Nat. Ecol. Evol.* **2**, 1906–1917 (2018).
- Lavorel, S. & Garnier, É. Predicting changes in community composition and ecosystem functioning from plant traits: Revisiting the holy grail. *Funct. Ecol.* **16**, 545–556 (2002).
- Garnier, Eric *et al.* Plant functional markers capture ecosystem properties during secondary succession. *Ecology* **85**, 2630–2637. <https://doi.org/10.1890/03-0799> (2004).
- van Bodegom, P. M., Douma, J. C. & Verheyen, L. M. A fully traits-based approach to modeling global vegetation distribution. *Proc. Natl. Acad. Sci. USA* **111**, 13733–13738. <https://doi.org/10.1073/pnas.1304551110> (2014).
- Kattge, J. *et al.* TRY plant trait database—enhanced coverage and open access. *Glob. Chang. Biol.* **26**, 119–188. <https://doi.org/10.1111/gcb.14904> (2020).
- Madani, N. *et al.* Future global productivity will be affected by plant trait response to climate. *Sci. Rep.* **8**, 2870. <https://doi.org/10.1038/s41598-018-21172-9> (2018).
- Brodrick, P. G., Davies, A. B. & Asner, G. P. Uncovering ecological patterns with convolutional neural networks. *Trends Ecol. Evol.* **34**, 734–745. <https://doi.org/10.1016/j.tree.2019.03.006> (2019).
- Reichstein, M. *et al.* Deep learning and process understanding for data-driven earth system science. *Nature* **566**, 195–204. <https://doi.org/10.1038/s41586-019-0912-1> (2019).
- Kattenborn, T., Eichel, J. & Fassnacht, F. E. Convolutional neural networks enable efficient, accurate and fine-grained segmentation of plant species and communities from high-resolution UAV imagery. *Sci. Rep.* **9**, 17656. <https://doi.org/10.1038/s41598-019-53797-9> (2019).
- Wäldchen, J. & Mäder, P. Machine learning for image based species identification. *Methods Ecol. Evol.* **9**, 2216–2225. <https://doi.org/10.1111/2041-210X.13075> (2018).
- Díaz, S. *et al.* The global spectrum of plant form and function. *Nature* **529**(7585), 167–171 (2016).
- Dickinson, J. L. *et al.* The current state of citizen science as a tool for ecological research and public engagement. *Front. Ecol. Environ.* **10**, 291–297. <https://doi.org/10.1890/110236> (2012).
- Hampton, S. E. *et al.* Big data and the future of ecology. *Front. Ecol. Environ.* **11**, 156–162. <https://doi.org/10.1890/120103> (2013).
- Boone, M. E. & Basille, M. Using iNaturalist to contribute your nature observations to science. *EDIS* **2019**, 5 (2019).
- Jin, T., Liu, G., Fu, B., Ding, X. & Yang, L. Assessing adaptability of planted trees using leaf traits: A case study with Robinia pseudoacacia L. in the loess plateau, China. *Chin. Geogr. Sci.* **21**, 290–303. <https://doi.org/10.1007/s11769-011-0470-4> (2011).
- Chave, J. The problem of pattern and scale in ecology: What have we learned in 20 years?. *Ecol. Lett.* **16**, 4–16. <https://doi.org/10.1111/ele.12048> (2013).
- Albert, C. H. *et al.* Intraspecific functional variability: Extent, structure and sources of variation. *J. Ecol.* **98**, 604–613 (2010).
- Gong, H., Cui, Q. & Gao, J. Latitudinal, soil and climate effects on key leaf traits in northeastern China. *Glob. Ecol. Conserv.* **22**, e00904. <https://doi.org/10.1016/j.gecco.2020.e00904> (2020).
- Yang, Y. *et al.* A novel approach for modelling vegetation distributions and analysing vegetation sensitivity through trait-climate relationships in China. *Sci. Rep.* **6**, 24110. <https://doi.org/10.1038/srep24110> (2016).
- Goëau, H., Bonnet, P., & Joly, A. Plant identification based on noisy web data: the amazing performance of deep learning (LifeCLEF 2017). In: CLEF: Conference and Labs of the Evaluation Forum, No. 1866 (2021).
- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria (2020).
- Duong, T. *et al.* ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *J. Stat. Softw.* **21**, 1–16 (2007).
- Perpiñán, O. & Hijmans, R. *rasterVis*. R package version 0.50.3 (2020).
- Boonman, C. C. F. *et al.* Assessing the reliability of predicted plant trait distributions at the global scale. *Glob. Ecol. Biogeogr.* **29**, 1034–1051. <https://doi.org/10.1111/geb.13086> (2020).
- Moreno-Martínez, Á. *et al.* A methodology to derive global maps of leaf traits using remote sensing and climate data. *Remote Sens. Environ.* **218**, 69–88. <https://doi.org/10.1016/j.rse.2018.09.006> (2018).
- Butler, E. E. *et al.* Mapping local and global variability in plant trait distributions. *Proc. Natl. Acad. Sci. USA* **114**, E10937–E10946. <https://doi.org/10.1073/pnas.1708984114> (2017).
- Baraloto, C. *et al.* Using functional traits and phylogenetic trees to examine the assembly of tropical tree communities. *J. Ecol.* **100**, 690–701 (2012).
- Eiserhardt, W. L., Borchsenius, F., Plum, C. M., Ordóñez, A. & Svenning, J.-C. Climate-driven extinctions shape the phylogenetic structure of temperate tree floras. *Ecol. Lett.* **18**, 263–272 (2015).
- Li, D., Ives, A. R. & Waller, D. M. Can functional traits account for phylogenetic signal in community composition?. *New Phytol.* **214**, 607–618 (2017).
- Yang, J. *et al.* Functional traits of tree species with phylogenetic signal co-vary with environmental niches in two large forest dynamics plots. *J. Plant Ecol.* **7**, 115–125 (2014).
- Yang, Y., Zhu, Q., Peng, C., Wang, H. & Chen, H. From plant functional types to plant functional traits: A new paradigm in modelling global vegetation dynamics. *Prog. Phys. Geogr.* **39**, 514–535. <https://doi.org/10.1177/030913315582018> (2015).
- Van Bodegom, P. *et al.* Going beyond limitations of plant functional types when predicting global ecosystem-atmosphere fluxes: Exploring the merits of traits-based approaches. *Glob. Ecol. Biogeogr.* **21**, 625–636 (2012).
- DeJong, T. & Doyle, J. Seasonal relationships between leaf nitrogen content (photosynthetic capacity) and leaf canopy light exposure in peach (*Prunus persica*). *Plant Cell Environ.* **8**, 701–706 (1985).
- Han, Q., Kawasaki, T., Nakano, T. & Chiba, Y. Spatial and seasonal variability of temperature responses of biochemical photosynthesis parameters and leaf nitrogen content within a *Pinus densiflora* crown. *Tree Physiol.* **24**, 737–744 (2004).
- Fick, S. E. & Hijmans, R. J. Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315. <https://doi.org/10.1002/joc.5086> (2017).
- Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition (pp. 248–255). IEEE (2009).
- Rolnick, D., Veit, A., Belongie, S. & Shavit, N. Deep learning is robust to massive label noise. arXiv preprint. [arXiv:1705.10694](https://arxiv.org/abs/1705.10694) (2017).
- Wright, I. J. *et al.* Global climatic drivers of leaf size. *Science* **357**, 917–921. <https://doi.org/10.1126/science.aal4760> (2017).

41. Mahecha, M. D. *et al.* Crowd-sourced plant occurrence data provide a reliable description of macroecological gradients. *Ecography* **44**, 1–12 (2021).
42. Zhou, Z.-H. A brief introduction to weakly supervised learning. *Natl. Sci. Rev.* **5**, 44–53. <https://doi.org/10.1093/nsr/nwx106> (2018).
43. Cao, Y. *et al.* Predicting pathogenicity of missense variants with weakly supervised regression. *Hum. Mutat.* **40**, 1579–1592 (2019).
44. Wüest, R. O. *et al.* Macroecology in the age of big data—where to go from here?. *J. Biogeogr.* **47**, 1–12 (2020).
45. Seyednasrollah, B. *et al.* Tracking vegetation phenology across diverse biomes using version 2.0 of the phenocam dataset. *Sci. Data* **6**, 1–11 (2019).
46. Estes, L. *et al.* The spatial and temporal domains of modern ecology. *Nat. Ecol. Evol.* **2**, 819–826. <https://doi.org/10.1038/s41559-018-0524-4> (2018).
47. Ueda, K.-I. iNaturalist Research-grade Observations. Accessed via GBIF.org on 2020-06-27, <https://doi.org/10.15468/ab3s5x> (2020).
48. Joly, A. *et al.* A look inside the pl@ntnet experience. *Multimed. Syst.* **22**, 751–766 (2016).
49. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. Preprint at <http://arxiv.org/pdf/1602.07261v2.pdf> (2016).
50. Mersmann, O., Trautmann, H., Steuer, D., Bornkamp, B. *truncnorm: Truncated Normal Distribution*. R package version 1.0-8 (2018).
51. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 1251–1258 (2017).
52. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 4510–4520 (2018).
53. Hernández-García, A. & König, P. Further advantages of data augmentation on convolutional neural networks. *ICANN* 95–103, (2018).
54. Allaire, J. & Chollet, F. *keras: R Interface to 'Keras'*. R package version 2.3.0.0 (2020).
55. Allaire, J. & Tang, Y. *tensorflow: R Interface to 'TensorFlow'*. R package version 2.2.0.9000 (2020).
56. Hijmans, R. J. *raster: Geographic Data Analysis and Modeling*. R package version 3.4–5 (2020).
57. Wei, T. & Simko, V. *corrplot: Visualization of a Correlation Matrix*. R package version 0.84 (2017).

Author contributions

T.K. conceived the study and assisted in conducting the experiments. C.S. planned and conducted the experiments, performed the analyses, and wrote the manuscript. T.K. and S.S. assisted in the analysis and in writing the manuscript. C.B. and A.M.-M. contributed data and helped to interpret the results. All authors provided edits and reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. We acknowledge support from Leipzig University for OpenAccess Publishing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-95616-0>.

Correspondence and requests for materials should be addressed to T.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021