**Methods**

*Fungal data and mycorrhizal richness estimates*

Fungal occurrence records were generated from data-mining of published ITS and SSU sequencing studies (see [Větrovský](https://pubmed.ncbi.nlm.nih.gov/?term=V%C4%9Btrovsk%C3%BD%20T%5BAuthor%5D) et al. 2020 for details). Briefly, raw sequences and metadata from 210 ITS studies were processed through an established bioinformatic pipeline that incorporates X, Y, Z tools. The resulting OTU table with taxonomy assignments was compared against the FungalTraits database (Põlme et al. 2020) for subsetting ectomycorrhizal fungi. A similar approach was used for 45 SSU studies to create a VTX table of AMF fungi using the MaarjAM database (Öpik et al. 2010)…

We used Hill numbers (q=0) to measure the effective number of mycorrhizal species (OTUs or VTs) in each sample (Chao et al. 2014). This approach creates a sequencing depth-based rarefaction and extrapolation sampling curve (i.e., a sample-specific species accumulation curve), with diversity estimates and 95% confidence intervals calculated at curve asymptotes (Hsieh et al. 2016). Rarefied estimates of mycorrhizal richness allow for a more robust comparison of mycorrhizal patterns across studies of multiple sequencing technologies and primer sets.

We removed outliers in rarefied mycorrhizal richness prior to spatial modeling. Two Australian studies with ITS samples from Desert and Mediterranean biomes had unusually high ECM OTU richness levels and standard deviations (two orders of magnitude beyond other Australian ecoregions; Bisset et al. 2016, Yan et al. 2018), and have been previously marked as potentially inaccurate based on a recent database comparison (Tedersoo et al. 2021). All samples from these two studies were removed. We then filtered samples by biome by removing rarefied richness values that were more than five times the interquartile range higher than the biome-level median estimate. We only filtered values at the highest end of the distribution to avoid potentially removing ‘true’ estimates of low or zero mycorrhizal richness at a given location. See Supplemental Table # and Supplemental Figure # for a summary of outliers removed per biome.

*Geospatial modeling and validation*

Spatial predictions of mycorrhizal richness were created with a random forest (RF) modeling approach (modified from van den Hoogen et al. 2019). We first sampled a stack of global environmental covariate layers at each of the locations within the dataset. These layers included macroclimatic, soil texture and physiochemical information, vegetation, radiation and topographic indices and anthropogenic variables. Details of all layers, including descriptions, units, and source information, are described in Supplementary Table 3. In short, variables describing soil structure and physiochemical properties were obtained from SoilGrids (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0169748), limited to the upper 5 cm of soil. Climate information (i.e., mean annual temperature, annual precipitation, monthly maximum temperature, precipitation seasonality) was obtained from CHELSA (https://www.nature.com/articles/sdata2017122). Spectral vegetation indices (i.e., MODIS NPP product xxx, averaged annually) from the Google Earth Engine Data Catalog ([https://developers.google.com/earth-engine/datasets](https://developers.google.com/earth-engine/datasets/catalog/modis)). Landcover and topographic information were obtained from EarthEnv (<https://onlinelibrary.wiley.com/doi/10.1111/geb.12182>, https://onlinelibrary.wiley.com/doi/10.1111/geb.12365,). The Potential Evapotranspiration (PET) layer was obtained from CGIAR ([https://doi.org/10.6084/m9.figshare.7504448.v3](https://doi.org/10.6084/m9.figshare.7504448.v3" \t "_blank)). Anthropogenic information (Human Development Percentage) was obtained from [https://doi.org/10.1111/geb.12182](https://doi.org/10.1111/geb.12182" \t "_blank). Resolve Ecoregion classifications were used to categorize sampling locations into biome (https://doi.org/10.1093/biosci/bix014). The final set of predictors included a set of xx variables. All spatial covariate layers were reprojected and resampled to a unified pixel grid in EPSG:4326 (WGS84) at 30 arc-sec resolution (approximately 1 km2 at the equator). Areas covered by permanent snow or ice (e.g. the Greenland ice cap, glaciated mountain ranges, identified using SoilGrids, 76) were excluded from the analyses. Antarctic areas were excluded from analysis due to the limited coverage of several covariate layers in the region.

o harmonize the data across the different experimental set-ups of the original studies, we incuded four types of project-specific variables. To create the predictions, we harmonized these project-specific variables to the most commonly used levels; i.e., sequening platform: Illumnina, target\_gene: ITS2, sample type: soil, primer set: ITS3 – ITS4. Prior to modelling, all project-specific variables transformed from categorical to binary variables (i.e., one-hot encoding).

To create the RF training datasets, only distinct observations were used; when multiple samples fell within the same 1km2 pixel, we retained only those with unique rarefied richness values . To deal with the zero-inflated data structure of the EcM dataset, we adopted a two-step approach. First, a binary RF classification model was created to separate positive occurrence data (richness > 0) from samples with undetected ectomycorrhizal communities (richness = 0). We then trained a regression RF model on the subset of positive occurrence samples with a log(x+1) transformation of rarefied richness. Results of the regression model were multiplied by the binary classification model to create a combined prediction. For AM we only trained a regression model as there were no zeros in this dataset. For both AM and EcM, the final predictions are an ensemble (mean) of the top 10 best performing models based on coefficient of determination R2 with random 10-fold cross validation. The mean cross-validated R2 values were 0.71 and 0.61, for AM and EcM, respectively.

To generate a spatial understanding of our predictive accuracy, we created 100 bootstrap samples through subsampling the datasets with replacement, using biome-based stratification. Leveraging the optimal hyperparameters of the RF model, we generated 100 global prediction images. These were subsequently simplified into the coefficient of variation (derived by dividing the standard deviation by the mean), and the lower 5% and upper 95% confidence intervals. In order to locate environmental conditions and corresponding geographic regions that are not represented in the training data, we first converted the data into the Principal Component (PC) space. We then selected the initial XX axes that cumulatively accounted for 90% of the variation. For each of the one-to-one (bivariate) combinations, we evaluated whether pixel values fell within or outside the sampled space. We then expressed the proportion of bivariate combinations as a percentage and used this as a measure for the extent of extrapolation. Finally, we combined this extrapolation data with a map indicating geographic distance to sampling locations, thereby producing a global spatial assessment of the representativeness of our datasets.

We then assessed whether our models and/or data exhibited any spatial autocorrelation. To do so, we first create semivariogram plots usising the R package gstat, on both the rarefied richness values and model residuals. Here we observe, for AM, autocorrelation in the raw data up to ca. 600 km. For EcM the pattern is less clear, and we don’t observe obvious autocorrlation patterns. In the model residuals, however, there is no clear spatial autocorrelation visible from the semivariograms. We then performed a Moran’s I test on model residuals, aggregated by 1-km2 pixel. Whereas we observe no significant SAC for AM (*I* = 0.008, p = 0.6184), the EcM model did show small, but significant SAC (*I* = 0.067, p < 0.001). This led us to conclude that the models capture most variation, yet there might be some spatial processes that are not explained for AM.

As we observe spatial autocorrelation in the observations for AM, we set out to validate our models using spatial blocked cross-validation (rather than random fold assignment; which in this case might result in an overly optimistic R2 value).. First, we assess the appropriate block size calculating the Wasserstein distance between the distribution of our sampled data in environmental space and the prediction space. We find that hexagon shaped spatial blocks with a diameter of 600 km (AM) and 800 km (EcM), fit the prediction space the best, with coefficient of determination R2 values are 0.17 and 0.19, for AM and EcM, respectively. As we expect the model performance to decrease with distance to training locations, we next performed spatially buffered leave-one-out cross-validation, a computationally expensive approach where a separate model is trained for every location in the dataset, leaving out locations within a range of buffer sizes (Fig x). As expected, the R2 values at buffer sizes >500 km are in line with the spatial blocked R2 values. To transform these findings into a visual representation, we plotted the coefficient of determination, R2, against the distance to the nearest sampling location, creating a spatial product. The resulting map was consistent with our previously described extrapolation map.

A picture containing screenshot, diagram, plot, text

Description automatically generated

A picture containing diagram, map, screenshot, line

Description automatically generated

A map of the world

Description automatically generated

*Variable importance, latitude trends, and protected area analysis*

[Feature importance and SHAP analysis description]

To measure how mycorrhizal richness varies with latitude, richness values were averaged at every 0.2 latitude degrees within one degree-wide longitude bands (Zhu et al. 2012). Mycorrhizal hotspots were defined by setting a cutoff at the 95th percentile of predicted richness values. These hotspots were then overlaid with the World Database of Protected Areas (Bingham et al. 2019) before calculating the extent to which each hotspot currently falls within a protected area at the biome level.