**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.

Next, we subsampled the datasets with replacement, stratified by biome, to create 100 bootstrap samples. Using the best performing RF model hyperparameters, we then created 100 global prediction images, which were consequently reduced to coefficient of variation (standard deviation divided by mean), and lower 5% and upper 95% confidence intervals. To identify environmental conditions and associated geographic areas that fall outside the space represented by the training data, we first transformed the data into Principal Component (PC) space. Next, we selected the first XX axes that collectively explained 90% of variation. Then, for each of the one-to-one (bivariate) combinations, we assessed whether pixel values fell within or outside the sampled space. Representing the proportion of bivariate combinations as a percentage, we used this as a metric for the degree of extrapolation. We then combined this information with a map of geographic distance to sampling locations, to create a global spatial assessment of representativeness of our datasets.

We then performed buffered leave-one-out cross-validation tests (Roberts et al. 2017) to assess whether there was any effect of spatial autocorrelation on model performance. As expected, the predictive power declines with increasing buffer sizes, given that nearby locations provide the local environmental information that inform the RF model. Yet, we observe a stabilization of coefficient of determination R2 around ±0.2 and ±0.15 for EcM and AM ,respectively, from buffer sizes larger than 500 km. To transform this information into a spatial product, we projected the coefficient of determination R2 as a function the distance to nearest sampling location. This map is in agreement with the map of extrapolation, described above.

We then assessed whether our models exhibited any residual spatial autocorrelation. To do so, we first performed a Moran’s I test on model residuals. Whereas we observe no significant SAC for AM (Moran’s I = 0.008, p = 0.6184), the EcM model did show positive SAC (Moran’s I = 0.067, p < 0.001).

We performed several types of additional validation to address concerns of spatial autocorrelation and geographic bias in sampling locations. First,

1. Incorporation of Moran Eigenvector Maps as model covariates
2. Spatial blocking of cross validation
3. Spatially-buffered leave-one-out cross-validation

*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.