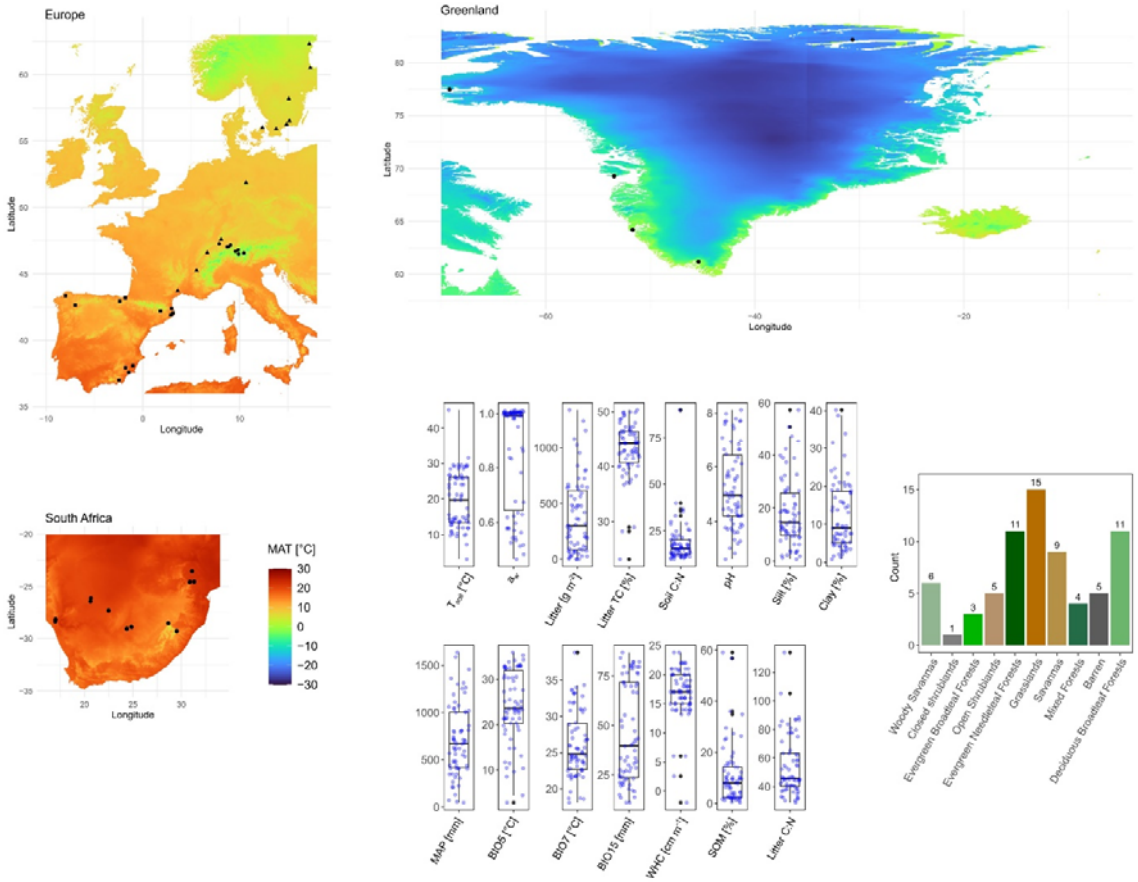

Figure S1: Overview of sites in Europe, Greenland and South Africa as well as distribution of climatic, vegetation and soil parameters across the dataset. MAT = mean annual temperature, a_w= water activity, MAP = mean annual precipitation, BIO5 = maximum temperature warmest month, BIO7 = annual temperature range, BIO15 = precipitation seasonality, WHC = water holding capacity, SOM = soil organic matter.

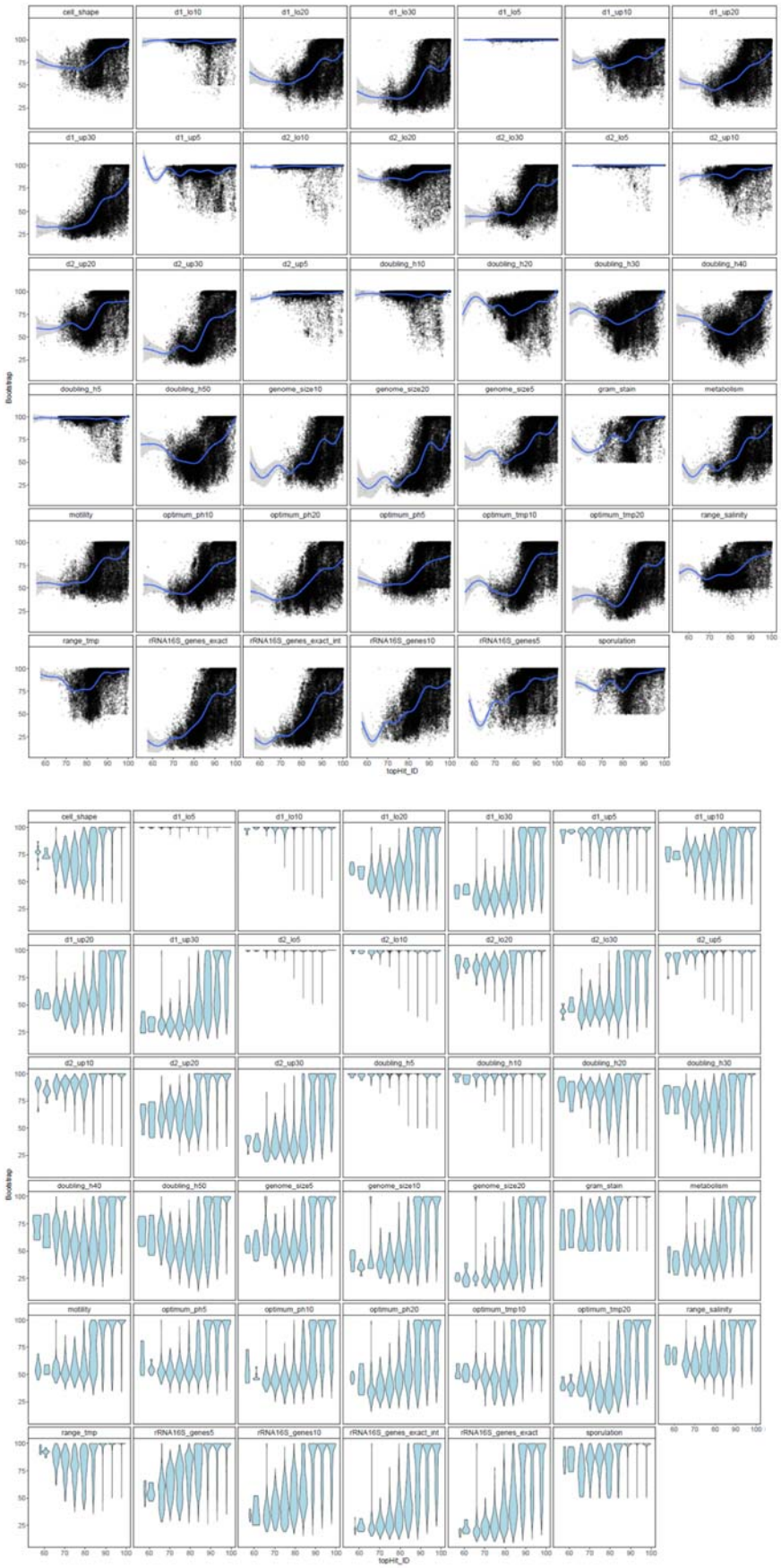

Figure S1 Bootstrap values as a function of the sequence identity with the top hit in the reference database as scatterplot (top) and as violin plot for 10 intervals of sequence identity (bottom). Intervals: [54.2,58.8] (58.8,63.4] (63.4,67.9] (67.9,72.5] (72.5,77.1] (77.1,81.7] (81.7,86.3] (86.3,90.8] (90.8,95.4] (95.4,100]

Supplementary methods

We used 16S rRNA gene amplicon sequencing data and metadata from a comprehensive soil dataset with 80 sites across Greenland, Europe and South Africa (Fig S1) to evaluate the performance of our workflow on real-world data. For evaluation of trait classifications, the full dataset of was used, for modelling trait – environment relationships, 10 sites from Greenland that were sampled to represent small-scale microclimatic heterogeneity were removed. We sampled five replicate soil cores (height: 10 cm, diameter: 10 cm) within an area of 20x20 meters. Soils were sieved through a 4-mm mesh and a subsample for DNA extraction was frozen at -20 °C. DNA was isolated from 0.25 g soil using DNeasy PowerSoil Pro Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. DNA concentrations were quantified using PicoGreen (Molecular Probes, Eugene, OR, USA) and the V3-V4 region of the prokaryotic16S rRNA gene was PCR amplified using the primers 341F and 801R as described previously (Frey et al., 2016). Amplicons were barcoded using the Fluidigm Access Array technology (Fluidigm) and paired-end sequenced on the Illumina MiSeq v3 platform (Illumina Inc., San Diego, CA, USA) at the Genome Quebec Innovation Center (Montreal, Canada). Raw sequences were processed using DADA2 (Callahan et al., 2016) implemented in Qiime2 (Bolyen et al., 2019). Primers were removed using cutadapt (Martin, 2011) with default settings and sequences were quality filtered and denoised with DADA2 (p-trunc-len-f = 270, p-trunc-len-r = 220, p-max-ee = 5).

Measured environmental variables in the dataset include pH, soil organic matter (SOM), total organic carbon (TOC), total carbon (TC) and nitrogen (TN), soil C:N, total litter, litter C, litter N, litter C:N, texture (sand, silt, clay), water activity (a_w) and in situ soil temperature at the time of sampling. pH was measured in a soil extract with 0.01 M $CaCl_2$ (extractant - soil ratio 2:1 v/w) using a pH meter. SOM was quantified through loss-on-ignition combusting the samples at 450 °C for 4 h (Davies, 1974). TOC content was determined upon HCl-fumigation with an elemental analyzer (Walthert et al., 2010). Soil and litter TC and TN were measured on ground material after drying at 60 °C using an elemental analyzer (NC-2500; CE Instruments, Wigan, United Kingdom). Soil texture was determined with the hydrometer method according to (Gee and Bauder, 1986). a_w indicating microbe available water (Daniel et al., 2004) was quantified using the aw meter LabSwift-aw (Novasina AG, Lachen, Switzerland).

In addition, we extracted further bioclimatic variables from the worldclim database (Fick and Hijmans, 2017) based on historical climate data between 1970 and 2000 at a resolution of 30 arc-seconds using the extract function in the *raster* package (Hijmans, 2022) with bipolar interpolation. These variables were: BIO1 (mean annual temperature), BIO5 (maximum temperature warmest month), BIO7 (temperature, annual range; maximum temperature of warmest month minus minimum temperature of coldest month), BIO15 (precipitation seasonality; ratio of the standard deviation of the monthly total precipitation to the mean monthly total precipitation). Similarly, we extracted the aridity index from the global aridity and PET database (http://www.cgiar-csi.org; (Zomer et al., 2008). Moreover, we extracted land cover classifications according the International Geosphere-Biosphere Programme classification using the MODIS product MCD12Q1_LC1 (Friedl et al., 2010) for the year 2020 which was downloaded using the *getModis* function in the R package *luna* (Ghosh et al., 2023) at 500 m spatial resolution. As the database is provided in sinusoidal projection, we projected the coordinates of our sites to the same projection to extract values using the function *spTransform* from the package *rgdal* (Bivand et al., 2023). We manually revised land cover classification based on photos from the sites. Finally, we extracted the water holding capacity from the ISRIC-WISE30sec data set (Batjes, 2016) provided at a resolution of 30 arc-seconds. We replaced the pixel identifier in the raster with the corresponding attribute using the *subs* function in the *raster* (Hijmans, 2022) package and subsequently extracted the values for our coordinates.

Missing values in the metadata were imputed based on principal components using the *estim_ncpPCA* and *imputePCA* function from the package *missMDA*.

References

Batjes, N.H., 2016. Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks. Geoderma 269, 61–68. doi:10.1016/j.geoderma.2016.01.034

Bivand, R., Keitt, T., Rowlingson, B., 2023. rgdal: Bindings for the "Geospatial" Data Abstraction Library.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K.B., Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hooft, J.J.J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J.G., 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37, 852–857. doi:10.1038/s41587-019-0209-9

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 13, 581–583. doi:10.1038/nmeth.3869

Daniel, R.M., Finney, J.L., Stoneham, M., Grant, W.D., 2004. Life at low water activity. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 359, 1249–1267. doi:10.1098/rstb.2004.1502

Davies, B.E., 1974. Loss-on-Ignition as an Estimate of Soil Organic Matter. Soil Science Society of America Journal 38, 150–151. doi:10.2136/sssaj1974.03615995003800010046x

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37, 4302–4315. doi:https://doi.org/10.1002/joc.5086

Frey, B., Rime, T., Phillips, M., Stierli, B., Hajdas, I., Widmer, F., Hartmann, M., 2016. Microbial diversity in European alpine permafrost and active layers. FEMS Microbiology Ecology 92, fiw018. doi:10.1093/femsec/fiw018

Friedl, M.A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., Huang, X., 2010. MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. Remote Sensing of Environment 114, 168–182. doi:10.1016/j.rse.2009.08.016

Gee, G.W., Bauder, J.W., 1986. Particle-size analysis, in: Methods of Soil Analysis, SSSA Book Series. American Society of Agronomy, Madison, pp. 383–411. doi:https://doi.org/10.2136/sssabookser5.1.2ed.c15

Ghosh, A., Mandel, A., Kenduiywo, B., Hijmans, R.J., 2023. luna: Tools for satellite remote sensing (Earth Observation) data processing.

Hijmans, R.J., 2022. raster: Geographic Data Analysis and Modeling. 2022. R package version 3.5-29. https://CRAN.R-project.org/package=raster.

Martin, M., 2011. CUTADAPT removes adapter sequences from high-throughput sequencing reads. EMBnet Journal 17, 10–12. doi:10.14806/ej.17.1.200

Walthert, L., Graf, U., Kammer, A., Luster, J., Pezzotta, D., Zimmermann, S., Hagedorn, F., 2010. Determination of organic and inorganic carbon, δ13C, and nitrogen in soils containing carbonates after acid fumigation with HCl. Journal of Plant Nutrition and Soil Science 173, 207–216. doi:10.1002/jpln.200900158

Zomer, R.J., Trabucco, A., Bossio, D.A., Verchot, L.V., 2008. Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. Agriculture, Ecosystems & Environment, International Agricultural Research and Climate Change: A Focus on Tropical Systems 126, 67–80. doi:10.1016/j.agee.2008.01.014