SI Appendix : **Spatial and evolutionary predictability of phytochemical diversity**

Emmanuel Defossez, Camille Pitteloud, Patrice Descombes, Gaétan Glauser, Pierre-Marie Allard, Tom W. N. Walker, Pilar Fernandez-Conradi, Jean-Luc Wolfender, Loïc Pellissier, Sergio Rasmann

This file includes:

Supplementary methods for species distribution models (SDMs)

We predicted the potential distribution of 403 plant species at a 93 m spatial grid resolution in Switzerland by using species distribution models (SDMs (1)). We used valid (expert-based) and precise (< 100 m) occurrences from the National Data and Information Center of the Swiss Flora (Infoflora.ch), which were disaggregated at 300 m to reduce spatial autocorrelation in model residuals and to reduce the potential effect of observations sampled on similar populations. SDMs were performed only for species presenting at least 150 occurrences (403 species out of the 416 species). We selected 8,000 pseudo-absences in Switzerland with a higher frequency in regions presenting higher number of occurrences to account for spatial bias in sampling intensities. The density layer was generated by calculating the number of occurrences into a grid of 5 km resolution. We finally added 2,000 pseudo-absences selected following a stratified sampling as a background to represent all possible combination of environmental conditions of the predictors used in the species modelling. We built SDMs by relating occurrence and pseudo-absence observations to ten environmental variables: 1) annual mean temperature (Tave), 2) annual precipitation sum (Precip), 3) annual sum of solar radiation (Srad), 4) topographic position index (i.e. topographic orientation, TPI), 5) topographic roughness index (TRI), 6) topographic wetness index (TWI), 7) inner forest density (Forest height Q25), 8) mean normalized difference vegetation index (NDVI), 9) soil moisture (EIV-F), and 10) soil pH (EIV-R). Tave, Precip, and Srad (direct and diffuse potential solar radiation) layers were retrieved from (2). TPI and TRI were calculated using the "terrain" function from the raster package (3) in R 3.5.1 (4) and TWI with SAGA (5), by using a 93 m resolution DEM (6). Inner forest density (Forest height Q25) was measured by analysing swisstopo and cantonal LiDAR data acquired during multiple seasons between 2000 and 2014 with a density of 0.5–35 points/m2 (7, 8). We calculated the 25th height-percentile of returns above 40 cm as a measure of inner forest density (8), respectively, at a spatial resolution of 25 m, and aggregated them at 93 m using bilinear interpolation (8). NDVI was calculated by using Landsat data (https://espa.cr.usgs.gov/) collected for the years 2007-2015 from 1 July to 15 September at 30 m resolution. Soil moisture (EIV-F) and soil pH (EIV- R) were averaged by grid cell, and modelled at 93 m spatial resolution with Random Forest algorithm and 16 predictors representing meso-climate, land use, topography and geology (see (9) for further details). Because SDM predictions have the tendency to vary among different statistical techniques (10, 11), we implemented an ensemble approach (12) by averaging the results of five statistical algorithms: generalized linear models (13), gradient boosting machines (14), generalized additive models (15, 16), random forests (17) and maximum entropy models (18, 19). We gave equal weights to presences and pseudo-absences in the calibration of the model and ran 5 iterations of the models with different sets of pseudo-absences (20, 21). For model evaluation, we calibrated the model on 80% of the data and evaluated them on the remaining 20% (random split-sampling). We used the True Skill Statistics (TSS)(22), which evaluates the ability of the model to discriminate presences from absences, defined by the following equation: Sensitivity + Specificity - 1. We used the threshold maximising the TSS for converting model probabilities to simulated presences and absences, and

calculated the rate of true positive (sensitivity) and true negative (specificity) using functions implemented in the *PresenceAbsence* package (23). TSS scales between -1 and 1, with 0 indicating random predictions. Models are considered to have reliable predictive performances with TSS values > 0.40 (i.e. excellent TSS > 0.75; good 0.40 < TSS < 0.75; poor TSS < 0.40) (24). Finally, metrics of predictive performance were averaged across replicates and across algorithms for every species. We considered only specie presenting TSS values >= 0.4 for further analyses (i.e. 391 species). All models were used to predict the potential distribution of plant species in Switzerland. We generated an Ensemble projection, by averaging single model projections across replicates and across algorithms of the model considering all the data. The Ensemble projection was finally converted into simulated presences and absences using the threshold maximising the TSS evaluated on the Ensemble. Additionally, for each plant species we extracted the mean, the median and quantiles (0.95-0.05) for all the environmental variables.

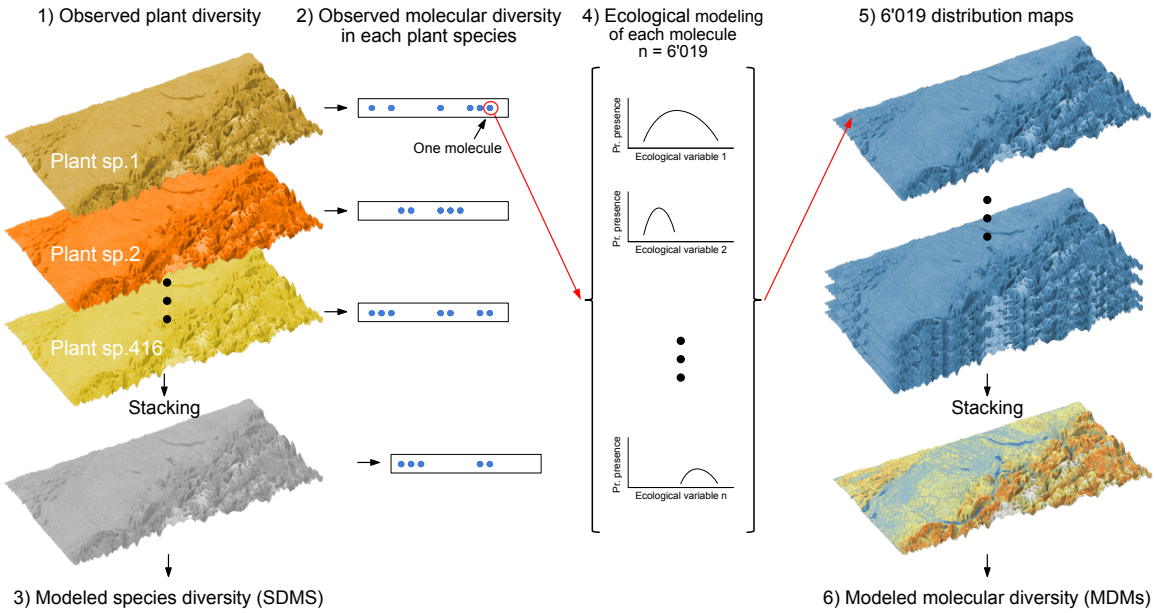

Figure S1. Scheme illustrating the methodological framework for developing molecular distribution models (MDMs) across the landscape. The workflow is summarized as follows: 1) exhaustively sampling species along ecological transects that cover the gamut of regional plant growth geographic boundaries, 2) build plant species distribution models (SDMs), 3) assess species-level phytochemical composition, 4) combine phytochemical information for each species with species distribution models (SDMs) for extracting climatic and topographical variables associated with each unique molecule observed across all species, 5) build distribution maps for each molecules (MDMs), and 6) stack them for predicting phytochemical diversity across the landscape.

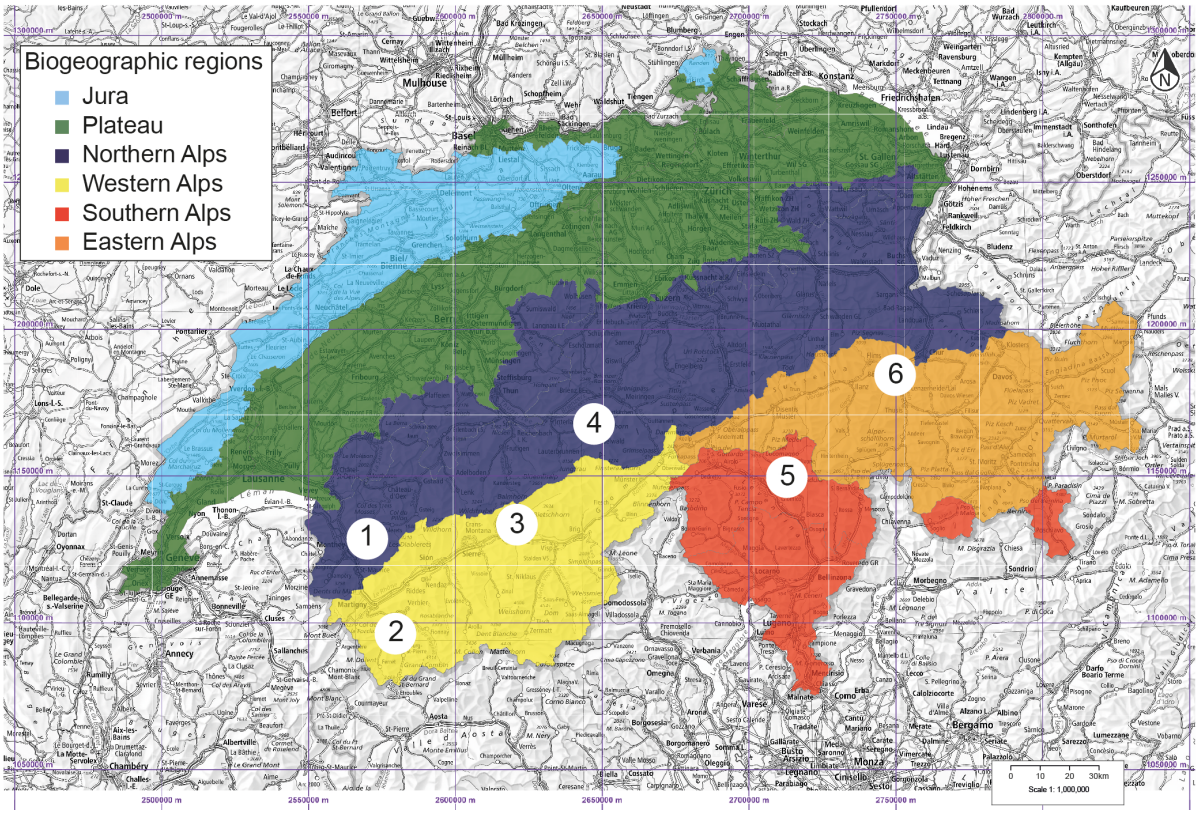

Figure S2. Sampling map of vegetation transects across Switzerland. Each dot represents a transect in the Swiss landscape where vegetation was inventoried and sampled from the bottom of the valley up to about 2000 m, for a total of 38 vegetation plots. The map of Switzerland is color-coded based on the different biogeographic regions as described in the legend. Figure modified from swisstopo (www.geo.admin.ch).

Figure S3. Molecular clustering based on spectral similarities. The clusters were generated by a molecular networking process (cosine >0,7). Each sub-network corresponding to specific compounds family. Each point represents the raw fragmented spectrum. Out of the 416 species and 69 families of plants, we aligned about 40’000 spectra, out of which, we were able to cluster >6000 molecular families of compounds and around 10’000 unique unclustered compounds. Green colour shows the annotated molecular families, while black shows the undetermined molecular families. MS/MS data and molecular network can be accessed through GNPS in the webpage : https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=a5a98ddc777a42d29ac33c3faca41a11 .

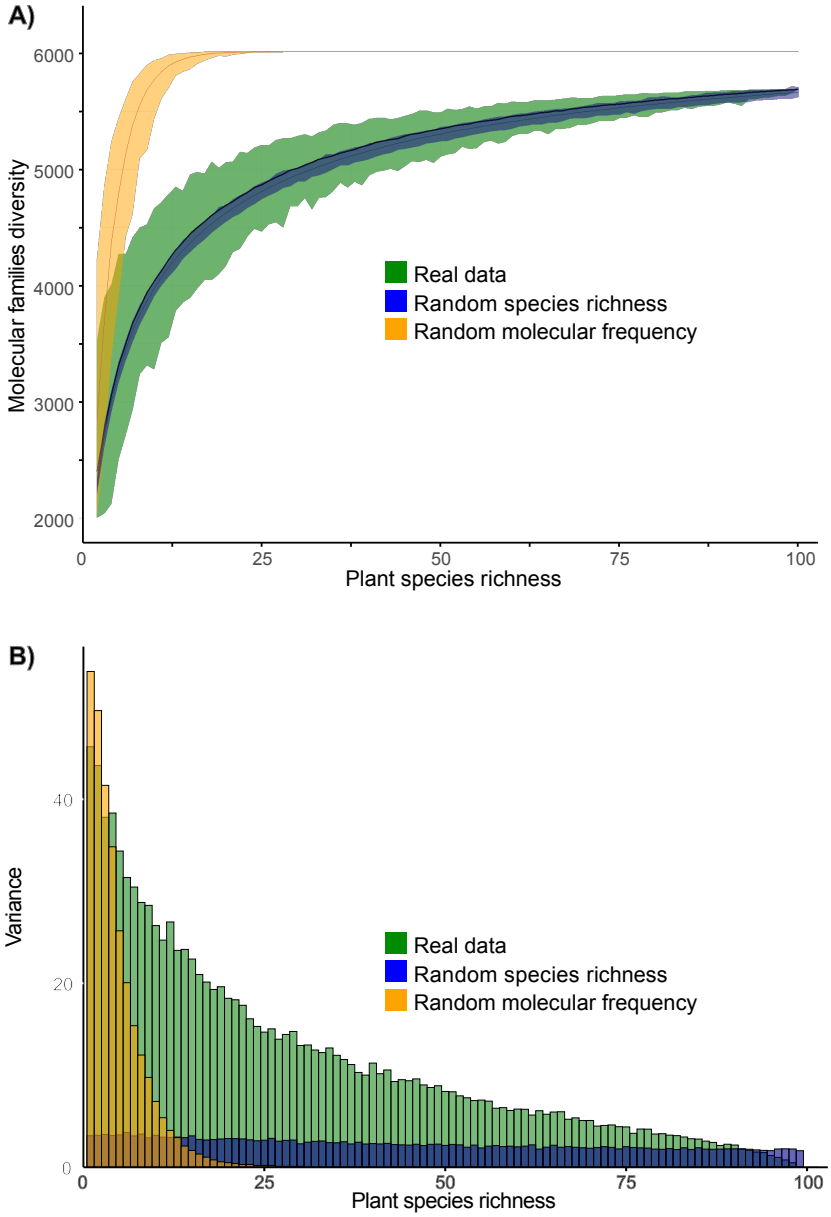

Figure S4. Sensitivity analyses for interpreting the correlation between plant species richness and molecular family diversity. Sensitivity of such correlation (A) was tested by building randomization-based null models, first by randomly varying molecular identity within species (blue ribbon), or by randomly varying species diversity (yellow ribbon). Panel (B) shows the error associated with each line. By comparing the null models with the real data value (green ribbon) we show that phytochemical diversity based on plant species richness is dependent on the specific structure the molecular diversity and frequency of molecules across plant species.

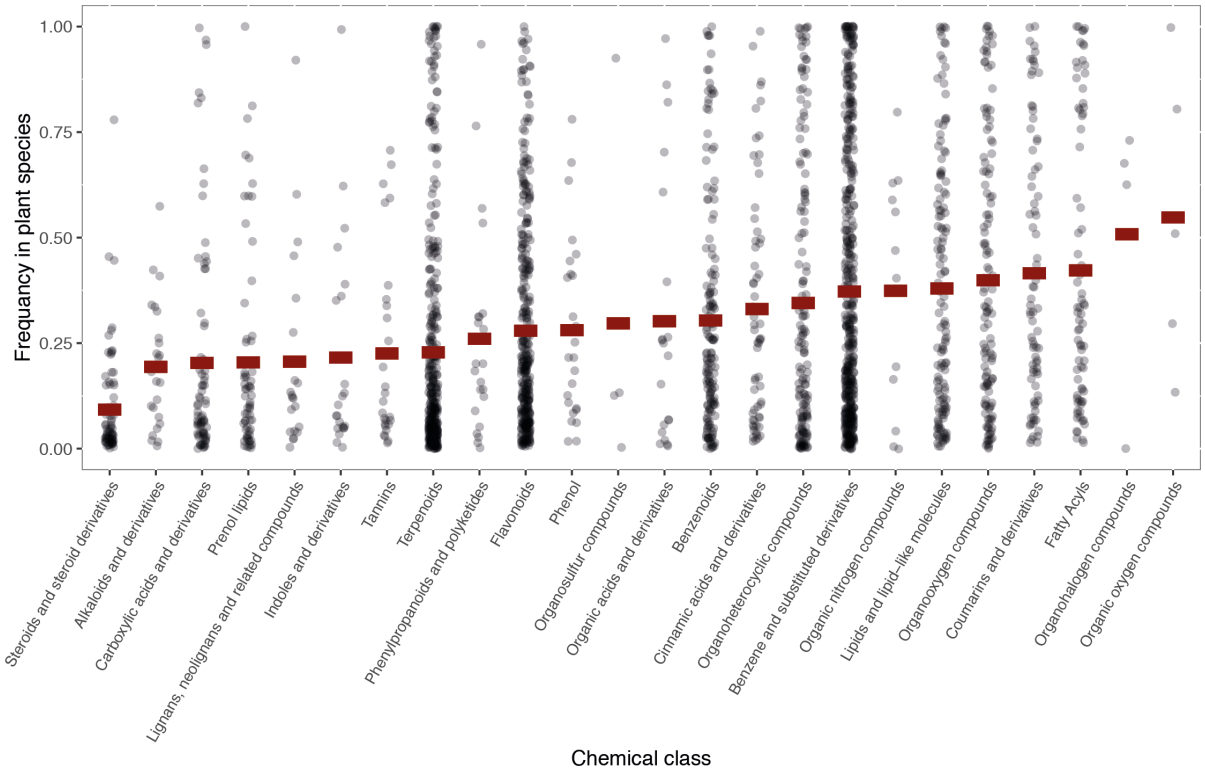

Figure S5. Frequency distribution of chemical classes of annotated molecular families found across 416 plant species About 40% of all molecular families were assigned to a known compound class or infraclass.

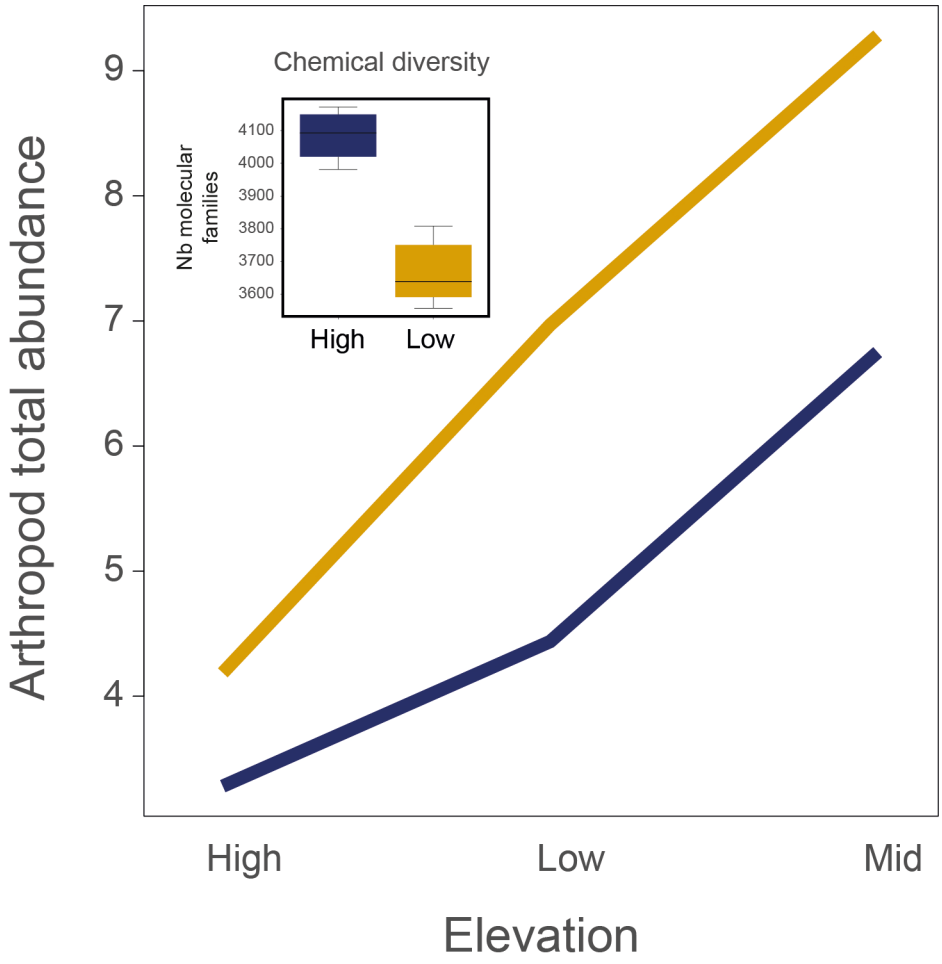

Figure S6. Field manipulation of molecular diversity and arthropods' abundance. The experiment consisted in building plant communities of high or low molecular families diversity by randomly selecting ten plant species per community from a pool of 50 species (inset: communities of high molecular families diversity were in average 11% richer compared to communities of low molecular family diversity; ANOVA: $F_{1,190} = 1220$, $p < 0.001$). Five pairs of high and low chemical diversity communities were placed at low (430, and 900 m above sea level (asl)), mid (1520, and 1470 m asl) and at high (1875 and 1950 m asl) elevation along the transects 1 and 5 as shown in Fig. S1, respectively. Pairs of plots were separated in average by 30 cm. All arthropods in each plant community were sampled by sucking through a fine-meshed net with inverted leaf blower (Hilti, AG, Switzerland) four times during the growing season (10.07.2019, 29.07.2019, 22.08.2019, 15.09.2019) and counted. Overall, insect abundance was 1.41 times higher in low molecular families diversity communities, independently of elevation (mixed effect linear model with transect as random factor: chemical diversity treatment effect: $F_{1,185} = 5.21$, $p = 0.023$; elevation effect: $F_{2,185} = 5.78$, $p = 0.004$, and treatment by elevation interaction: $F_{2,185} = 5.21$, $p = 0.33$, $p = 0.714$).

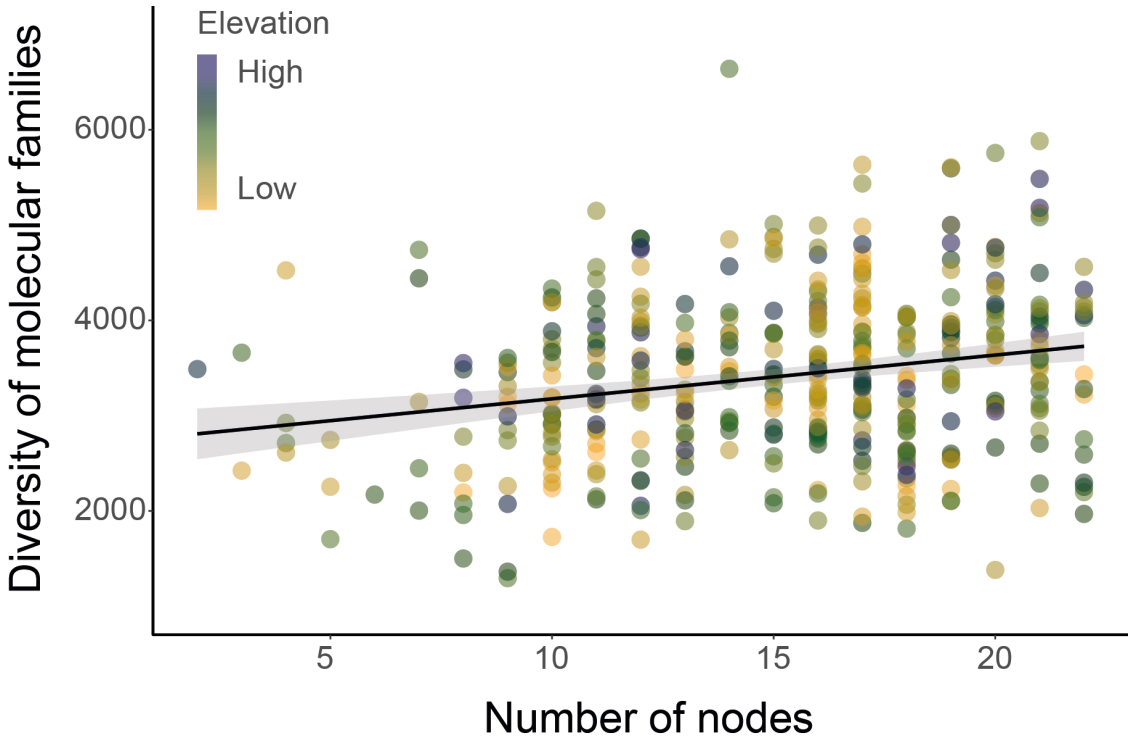

Figure S7. Regression analysis between the richness of molecules and the number of intervening nodes in the phylogeny for each plant species surveyed. (Pearson correlation: N = 416, r = 0.5, *P* < 0.001, estimate = 20).

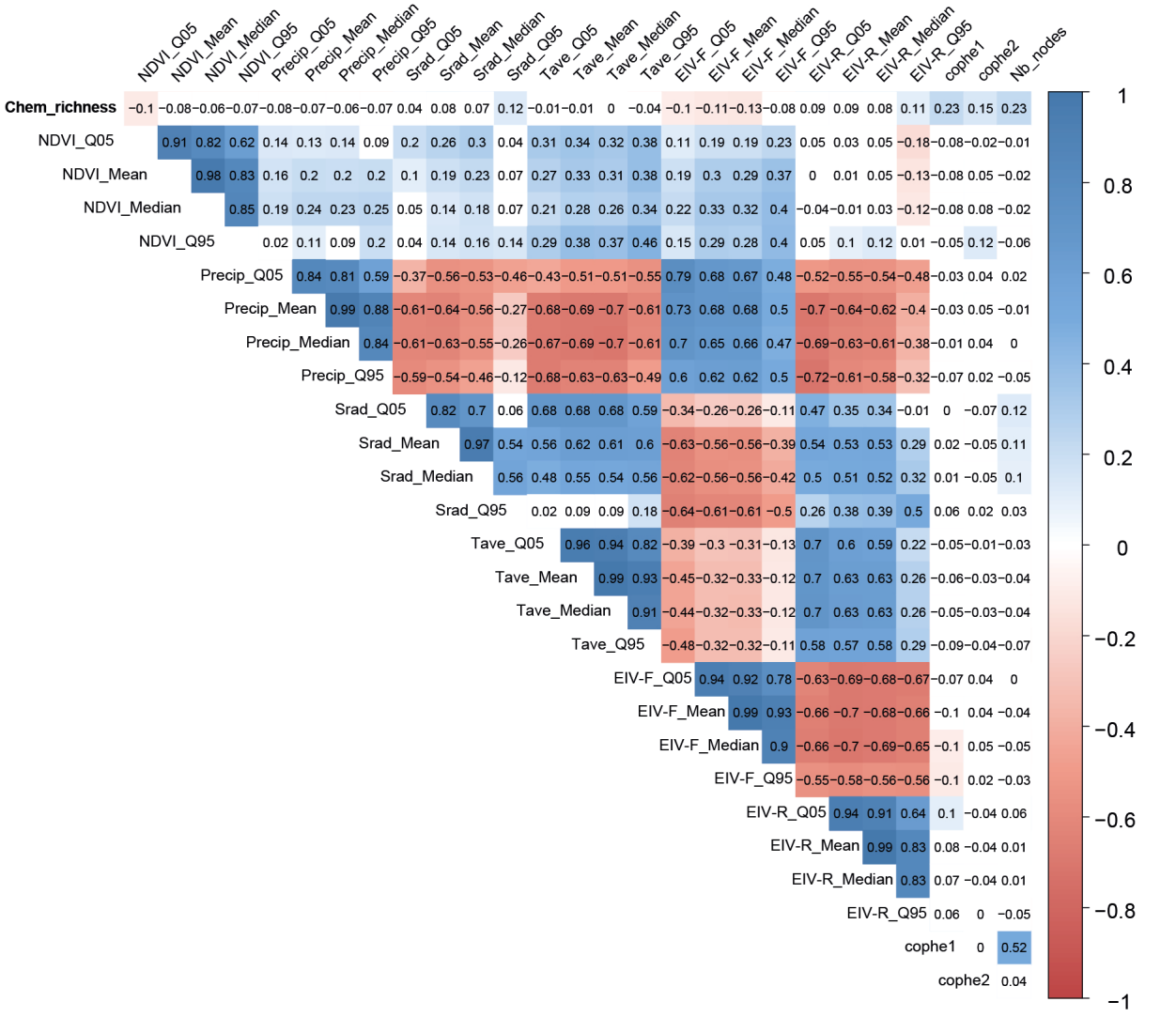

Figure S8. Correlation plot among all predictive variables, including: Phytochemical diversity (Chem_richness), annual mean temperature (Tave), annual precipitation sum (Precip), annual sum of solar radiation (Srad), mean normalized difference vegetation index (NDVI), soil humidity (EIV-F), and soil pH (EIV-R), cophenetic distance 1 and 2, and number of intervening nodes (Nn). Pairwise significant correlations are highlighted with colours (blue = positive, and red = negative correlations).

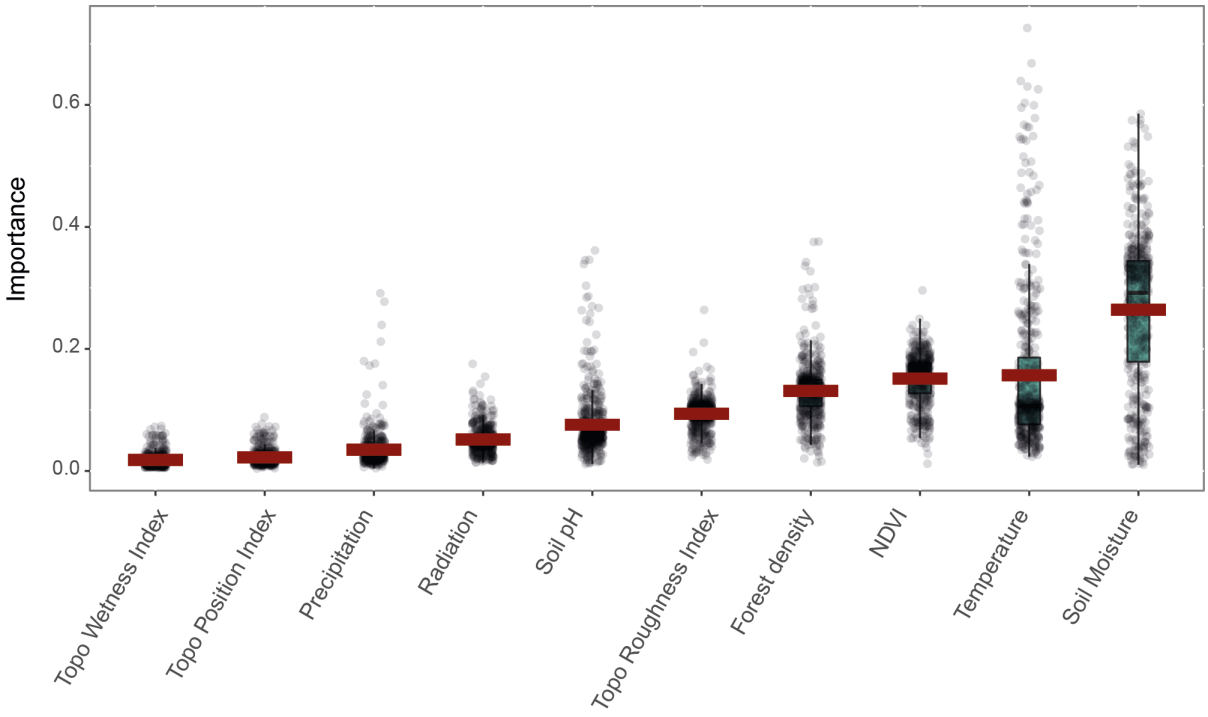

Figure S9. Variable importance for all the environmental predictors used to build the spatial distribution of molecular families richness across the landscape.

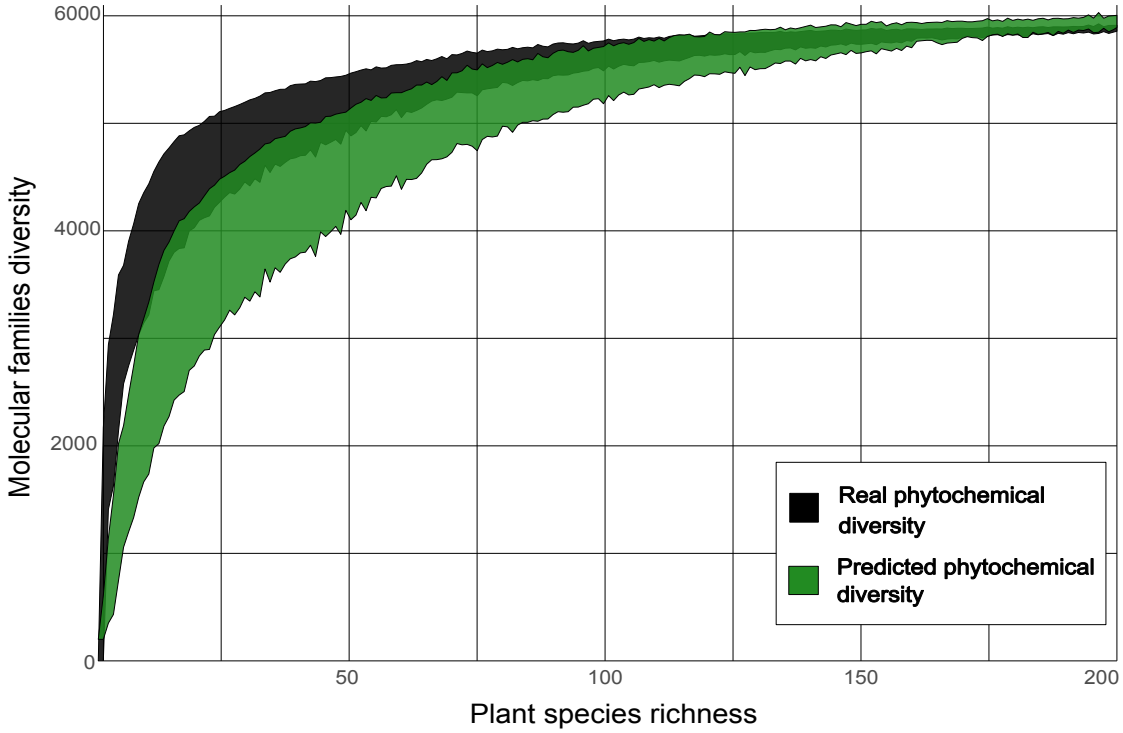

Figure S10. Correlation between plant species richness and molecular families diversity. Black ribbon corresponds to the correlation between SDMs-based species richness and real phytochemical composition of each plant species (i.e. calculated from chromatographic analyses of field-sampled plants). The green ribbon shows the correlation between SDMs-based species richness and the MDMs phytochemical predictions (i.e. predicted phytochemical diversity). The observed slight underestimation of phytochemical diversity based on the MDMs is an artifact based on model parameters calibration, and not a true difference.

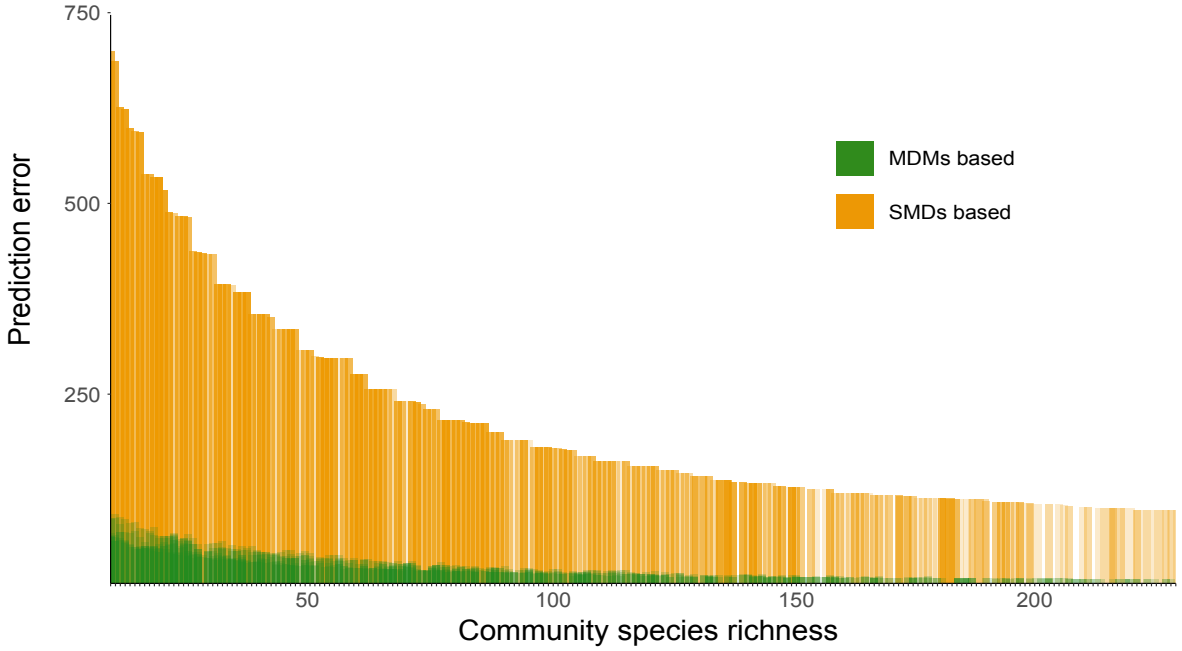

Figure S11. Testing the accuracy of MDMs predictions of phytochemical diversity across plant species assemblages. Shown are prediction errors derived from a model aiming at predicting phytochemical diversity from predicted SMDs species richness (yellow bars), or from MDMs models (green bars). Species assemblages were obtained by randomly sampling 50'000 locations. Overall, the MDMs model increase accuracy by up to 12-fold, and this effect is particularly important for communities with less than 100 species.

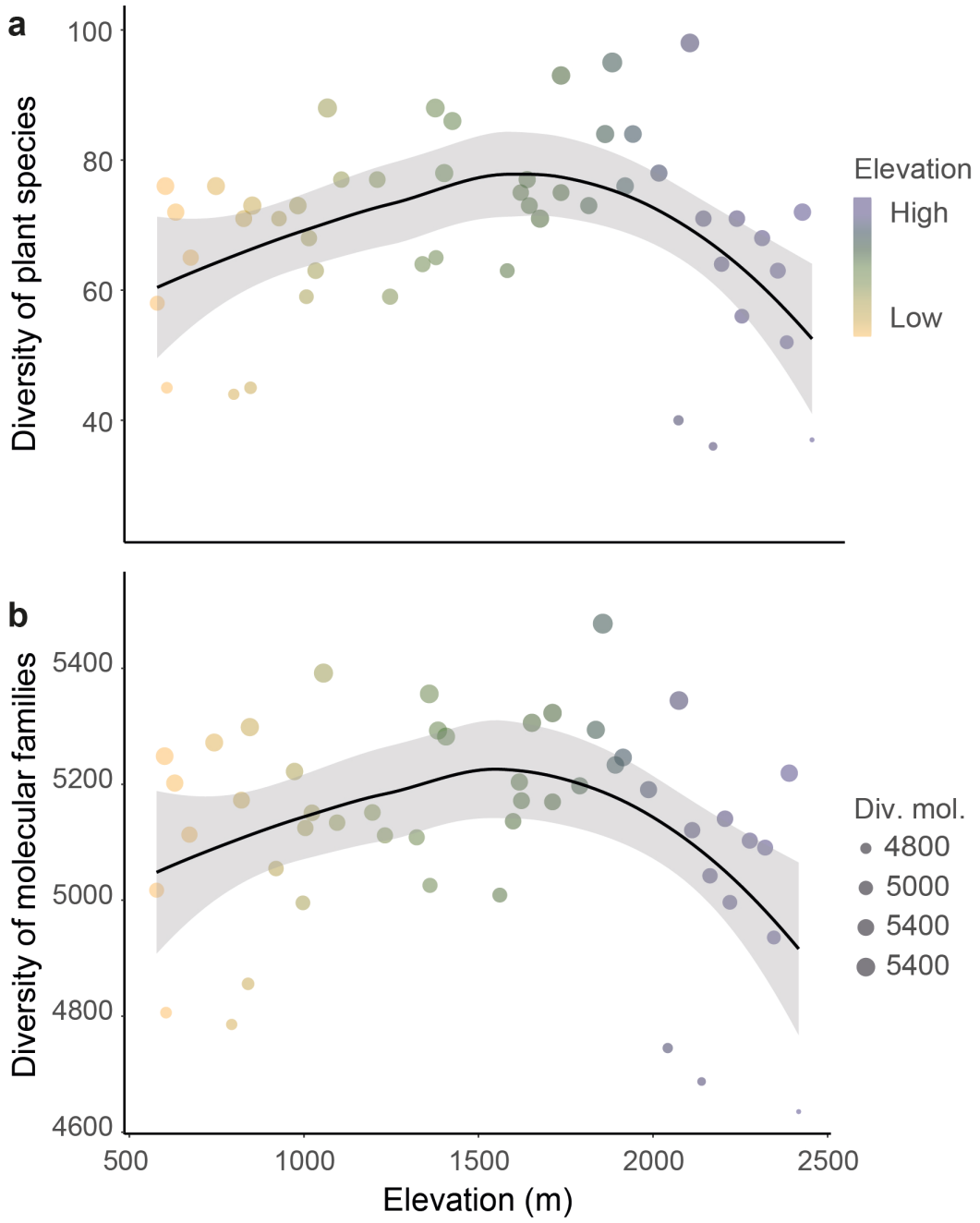

Figure S12. Community-level species a) and molecular b) diversity. Each dot represents a community of plants sampled along the six elevation transects in the Alps. The size of the dots is scaled according to the diversity of unique molecules found across all species in each community. Second-degree regression curves are the best fit for the diversity distributions along elevation (for species diversity; elevation: N = 48; estimate = 0.08, t = 4.18, $p < 0.0001$, and elevation^2 : estimate= -2.5e-05, t = -4.09, $p < 0.001$. For chemical diversity; elevation: estimate = 7.133e-02, t = 3.444, $p < 0.0001$, and elevation^2 : estimate= -2.403e-05, t = -3.488, $p < 0.001$).

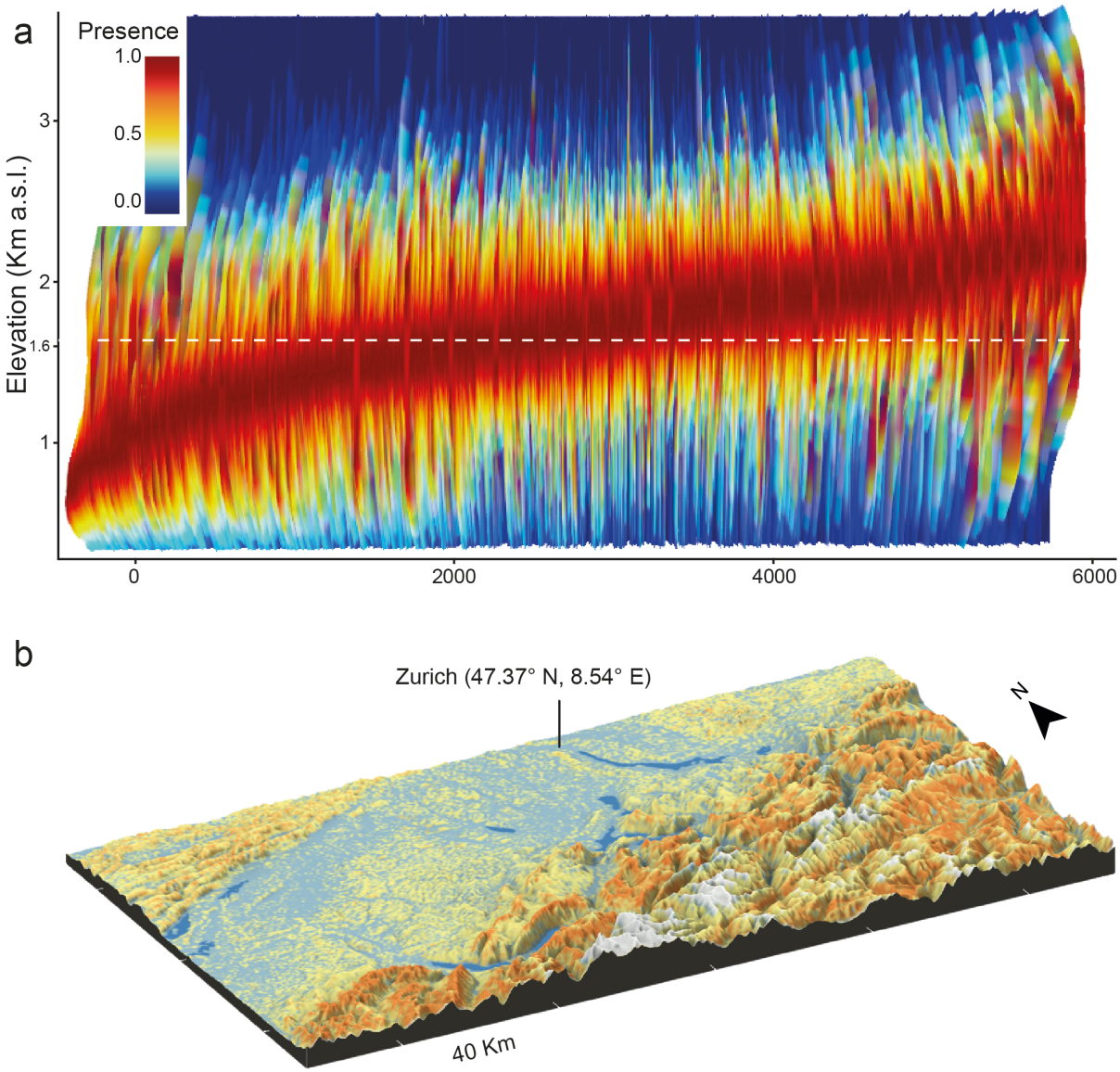

Figure S13. Elevational distribution of endemic molecular families. (a) Optimal distribution of the endemism of 6'012 molecular families found in the Alps along elevation (km a.s.l.). The distribution of the molecular families along elevation is based on the niche of each species retrieved from spatial modeling occurrence data in Switzerland. Red colors indicate maximal probability of occurrence, and blue colors indicate absence. The dotted white line represents the average value of presence across all molecular families. Molecular endemism was estimated by filtering the quantile 0.8 of molecular families elevational range for probability of presence corresponding to 0.5. (b) Spatial representation of molecular endemism in Switzerland. The map is color-coded as the described for panel (a), except for the white layer of snow on the top of the mountains and the lakes.

References

1. A. Guisan, N. E. Zimmermann, Predictive habitat distribution models in ecology. *Ecol. Model.* **135**, 147-186 (2000).
2. N. Zimmermann, F. Kienast, Predictive mapping of alpine grasslands in Switzerland: species versus community approach. *Journal of Vegetation Science* **10**, 469-482 (1999).
3. R. J. Hijmans (2019) raster: Geographic Data Analysis and Modeling.
4. R Development Core Team, R: A language and environment for statistical computing.
5. O. Conrad *et al.*, System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geoscientific Model Development* **8**, 1991-2007 (2015).
6. N. Robinson, J. Regetz, R. P. Guralnick, EarthEnv-DEM90: A nearly-global, void-free, multi-scale smoothed, 90m digital elevation model from fused ASTER and SRTM data. *ISPRS Journal of Photogrammetry and Remote Sensing* **87**, 57-67 (2014).
7. R. Artuso, S. Bovet, A. Streilein, Practical methods for the verification of countrywide terrain and surface models. *International Archives of Photogrammetry and Remote Sensing* **XXXIV, Part 3/W13** (2003).
8. R. O. Wüest, A. Bergamini, K. Bollmann, A. Baltensweiler, LiDAR data as a proxy for light availability improve distribution modelling of woody species. *For. Ecol. Manage.* **456**, 117644 (2020).
9. P. Descombes *et al.*, Spatial modelling of ecological indicator values at high resolution improves predictions of plant distributions in complex landscapes. *in review* (2020).
10. J. Elith *et al.*, Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29**, 129-151 (2006).
11. W. Thuiller, L. Brotons, M. B. Araújo, S. Lavorel, Effects of restricting environmental range of data to project current and future species distributions. *Ecography* **27**, 165-172 (2004).
12. M. B. Araújo, M. New, Ensemble forecasting of species distributions. *Trends Ecol. Evol.* **22**, 42-47 (2007).
13. P. McCullagh, Generalized linear models. *European Journal of Operational Research* **16**, 285-292 (1984).
14. G. Ridgeway, The state of boosting. *Computing Science and Statistics* **31**, 172–181 (1999).
15. T. Hastie, R. Tibshirani, Generalized Additive Models: Some Applications. *Journal of the American Statistical Association* **82**, 371-386 (1987).
16. A. Guisan, T. C. Edwards, T. Hastie, Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecol. Model.* **157**, 89-100 (2002).
17. L. Breiman, Random Forests. *Machine Learning* **45**, 5-32 (2001).
18. J. Elith *et al.*, A statistical explanation of MaxEnt for ecologists. *Divers. Distrib.* **17**, 43-57 (2011).
19. S. J. Phillips, R. P. Anderson, R. E. Schapire, Maximum entropy modeling of species geographic distributions. *Ecol. Model.* **190**, 231-259 (2006).
20. M. S. Wisz, A. Guisan, Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. *BMC Ecol.* **9**, 8 (2009).

21. M. Barbet-Massin, F. Jiguet, C. H. Albert, W. Thuiller, Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution* **3**, 327-338 (2012).
22. O. Allouche, A. Tsoar, R. Kadmon, Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *J. Appl. Ecol.* **43**, 1223-1232 (2006).
23. E. A. Freeman, G. Moisen, PresenceAbsence: An R Package for Presence-Absence Model Analysis. *Journal of Statistical Software* **23**, 1-31 (2008).
24. J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data. *Biometrics* **33**, 159-174 (1977).