Microgastropod Taxon-Omics: Towards a Probabilistic and Automated Species-Discovery System

Torsten Hauffe[1], Jens Schauer[1,2], Thomas Wilke[1]

1 Systematic & Biodiversity Group, Justus Liebig University Gießen, Germany 2 Bioinformatic & Systems Biology, Justus Liebig University Gießen, Germany

Background

Current state of Taxonomy

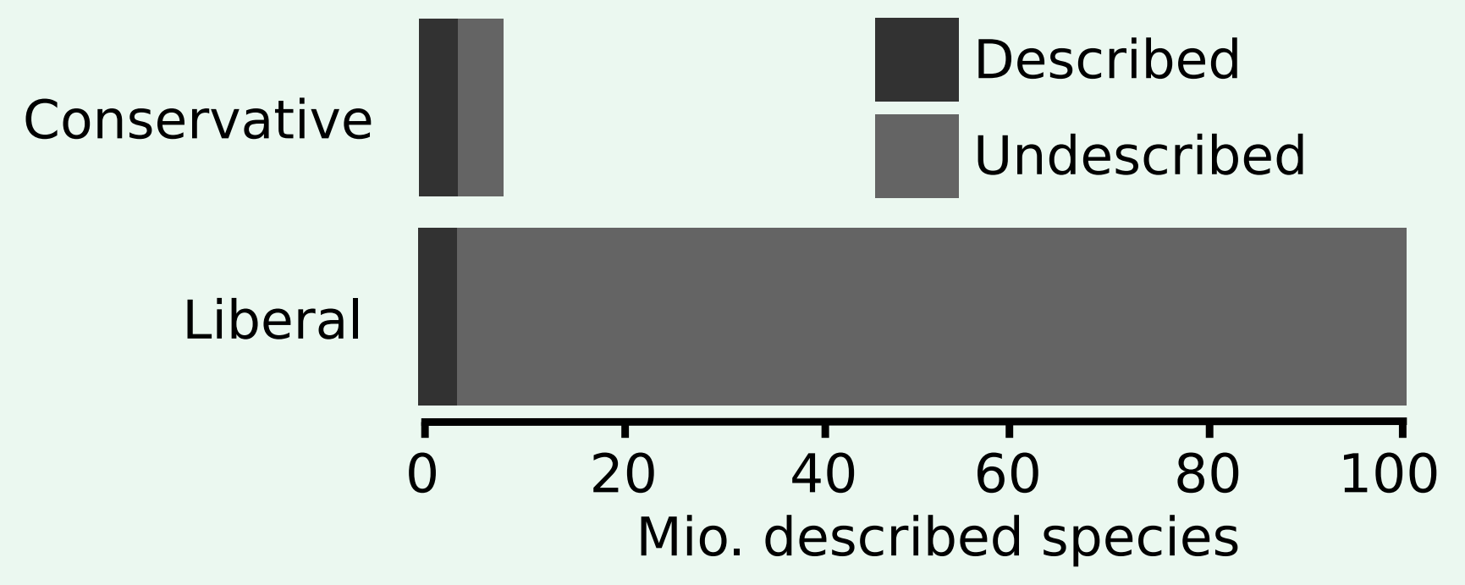


Fig. 1: Over the past 260 years, taxonomy yielded ca. 1.8 million validly described extant species[1]. However, a high number of species remains undescribed[2].

Reducing the share of undocumented species conflicts with the two main interests of taxonomy — quality and speed of species delimitation/description.

➤ We address both aspects by developing a probabilistic species-discovery system (proSDS) and applying it to a newly compiled reference database for microgastropods.

Microgastropods

The aquatic family Hydrobiidae (Stimpson, 1865) comprises > 900 described extant species, occurring mainly in springs and lakes of the northern hemisphere (Fig. 2)[3]. Their typically small (ca. 2 mm) and featureless shells, simple anatomy, and varying evolutionary rates (Fig. 3) require sophisticated integrative taxonomic approaches.

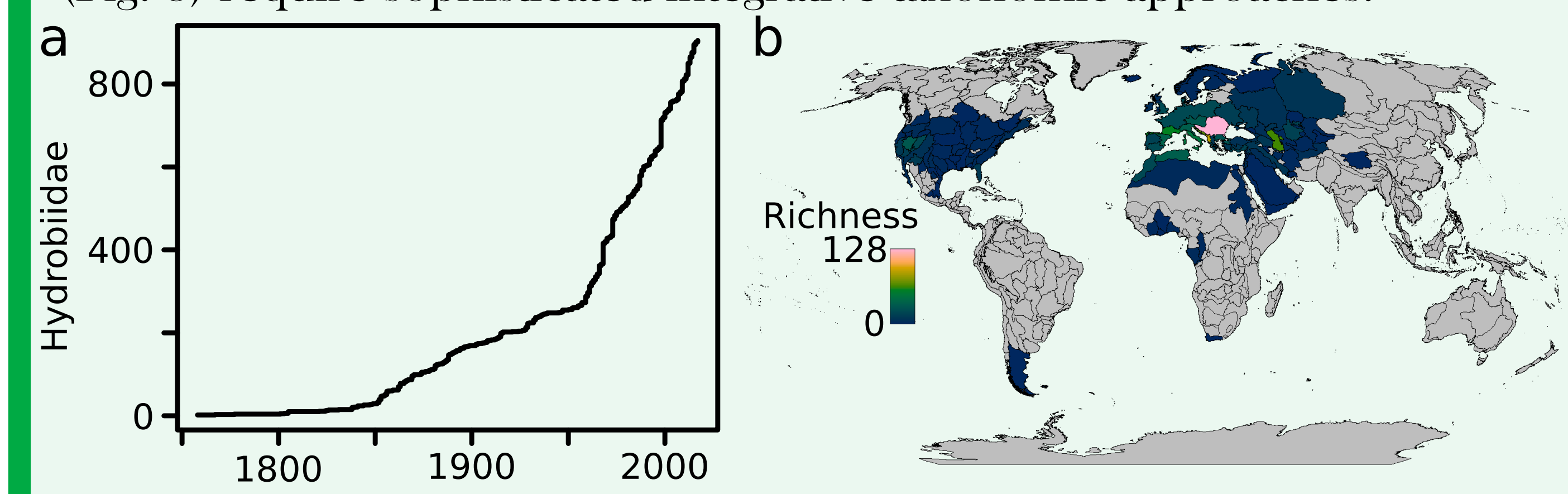


Fig. 2: (a) Described species through time. (b) Spatial distribution of species richness.

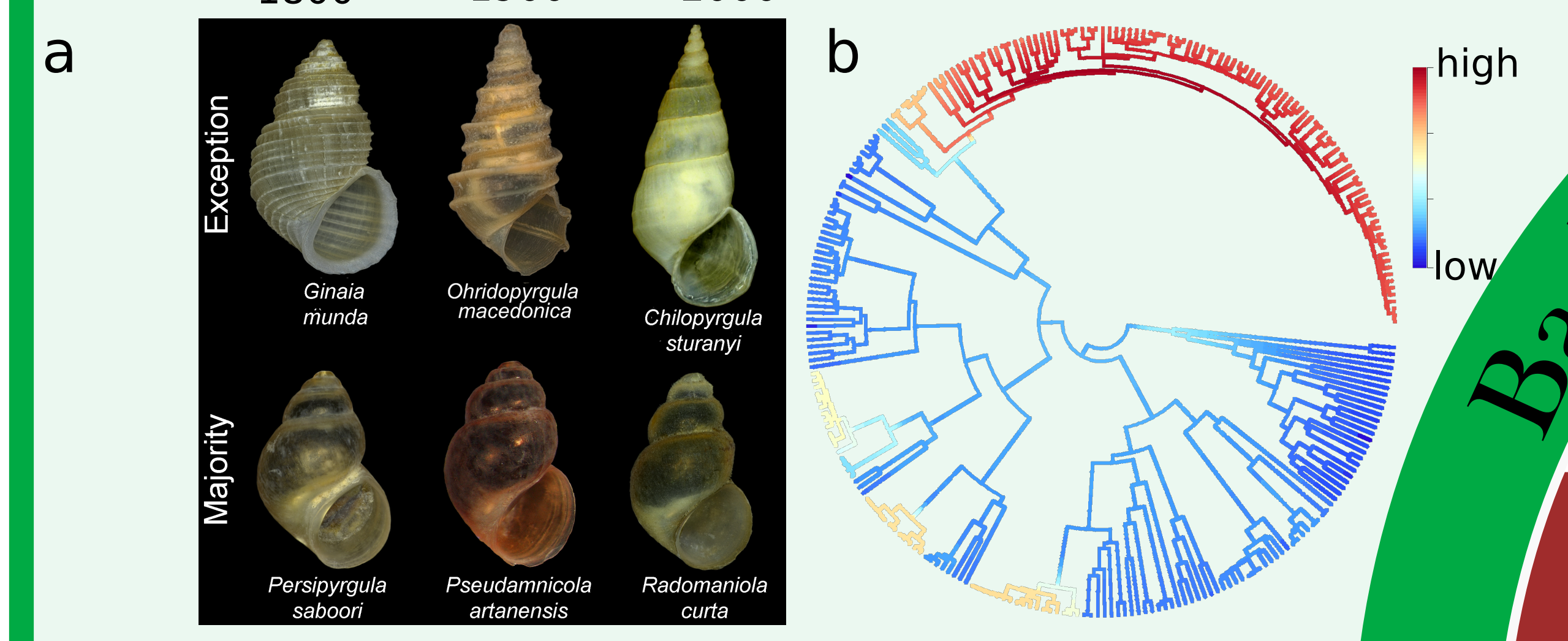
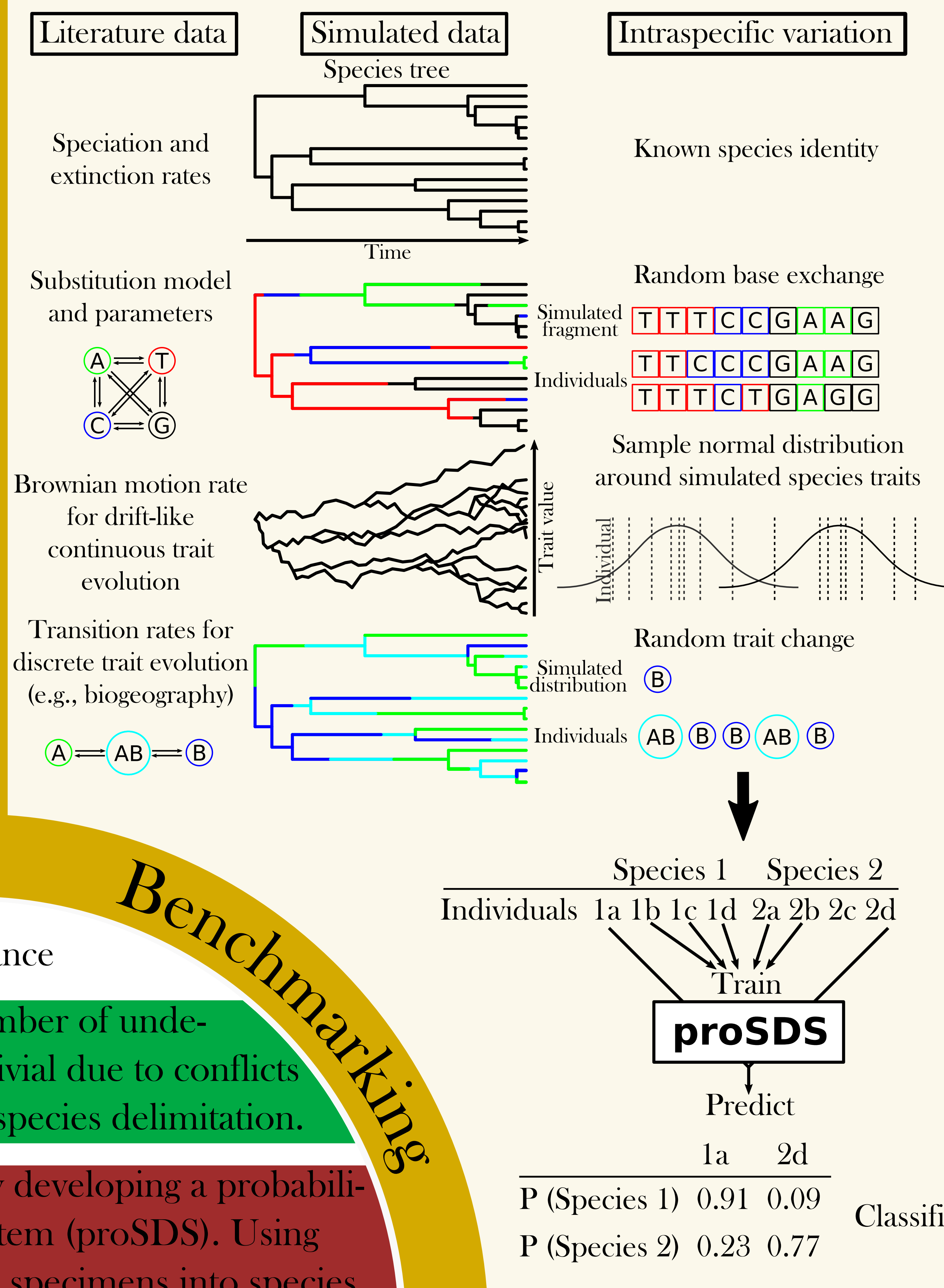


Fig. 3: (a) Examples of shell diversity. (b) Diversification rates.

Benchmarking

Performance evaluation of novel approaches is challenging because insufficient statistical robustness may result from an ill-defined method and/or inconsistencies in the input data (e.g., wrongly determined specimens). We therefore used simulated data of known properties and illustrated the use of proSDS on the two model microgastropod genera Pseudamnicola and Corrosella[4].



Literature data

Simulated data

Intraspecific variation

Species tree

Speciation and extinction rates

Time

Known species identity

Substitution model and parameters

Random base exchange

Simulated fragment T T T C C G A A G

Individuals T T C C C G A A G

T T T C T G A G G

Brownian motion rate for drift-like continuous trait evolution

Sample normal distribution around simulated species traits

Trait value

Individual

Transition rates for discrete trait evolution (e.g., biogeography)

Random trait change

Simulated distribution B

Individuals AB B B AB B

Species 1 Species 2

Individuals 1a 1b 1c 1d 2a 2b 2c 2d

Train

proSDS

Predict

	1a	2d
P (Species 1)	0.91	0.09
P (Species 2)	0.23	0.77

Classification performance?

Fig. 5: Simulation pipeline. Based on evolutionary parameters compiled from literature, we simulated sequence and morphological/ecological traits along species phylogenies and added population variability.

At a glance

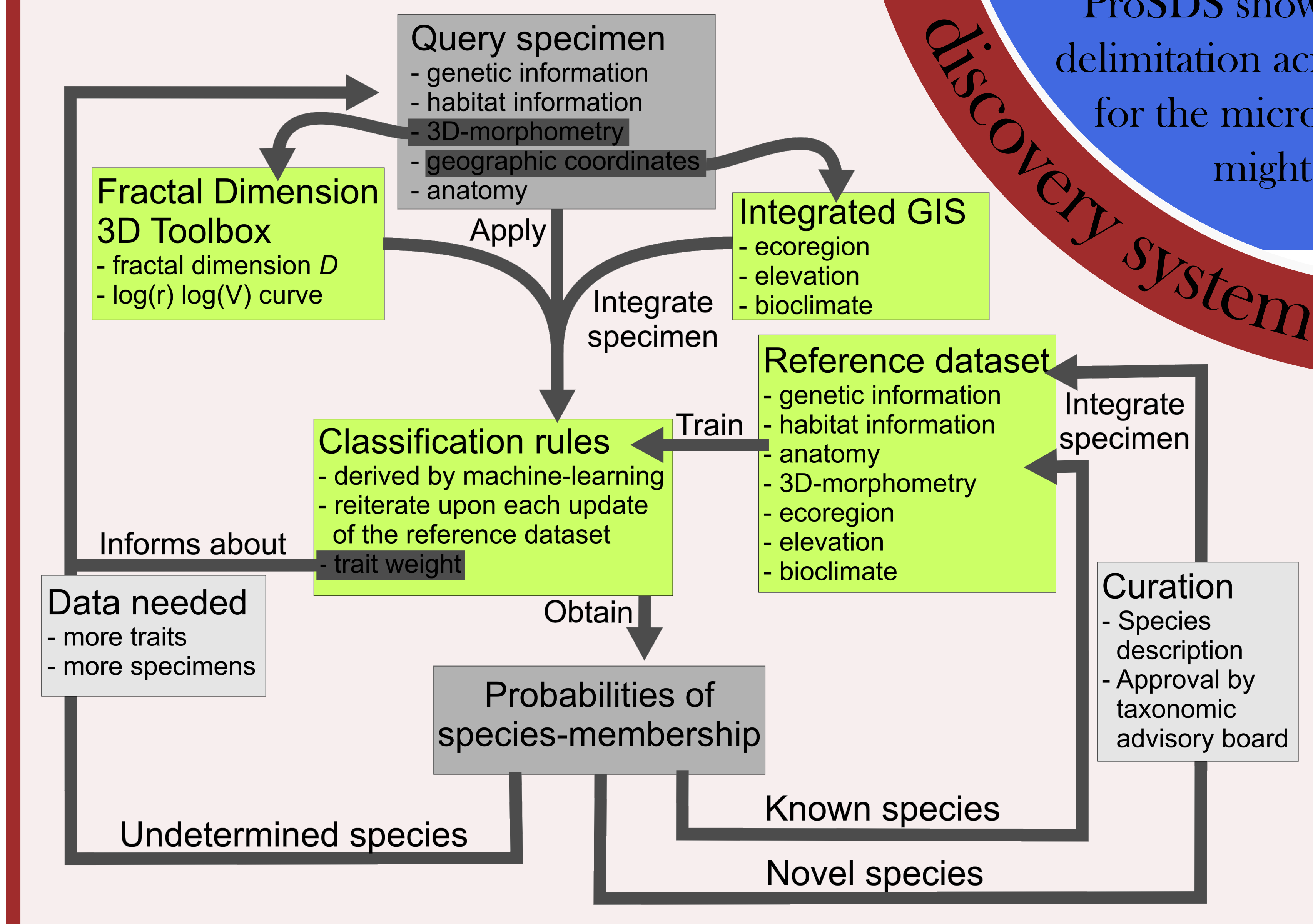
Reducing the number of undescribed species is not trivial due to conflicts in quality and speed of species delimitation.

We address both aspects by developing a probabilistic species-discovery system (proSDS). Using machine learning, it groups specimens into species based on genetic, morphological, and ecological data.

We evaluated the performance using simulated data and a reference database for microgastropods.

ProSDS showed high accuracy in species delimitation across simulation scenarios and for the microgastropods. Our approach might facilitate taxonomic decisions.

New discovery system



Query specimen
- genetic information
- habitat information
- 3D-morphometry
- geographic coordinates
- anatomy

Fractal Dimension 3D Toolbox
- fractal dimension *D*
- log(r) log(V) curve

Apply

Integrated GIS
- ecoregion
- elevation
- bioclimate

Integrate specimen

Reference dataset
- genetic information
- habitat information
- anatomy
- 3D-morphometry
- ecoregion
- elevation
- bioclimate

Integrate specimen

Train

Classification rules
- derived by machine-learning
- reiterate upon each update of the reference dataset
- trait weight

Informs about

Data needed
- more traits
- more specimens

Obtain

Probabilities of species-membership

Curation
- Species description
- Approval by taxonomic advisory board

Undetermined species

Known species

Novel species

Fig. 4: Workflow of the probabilistic species-delimitation system. Supervised machine learning derives rules for classifying specimens into species utilizing a taxonomist's curated reference dataset of genetic, morphological, and ecological traits. Querying the species identity of an unknown specimen by applying the classification rules results in a probability for belonging to a species not included in the reference dataset ('novel species'). Next, for each species in the reference dataset a probability for the query specimen to be a member of that species is obtained ('known species'). In case of throughout low probabilities per reference species, more data need to be collected ('undetermined species').

Results & Discussion

The benchmarking results suggest a generally good performance of proSDS in both real and simulated data, with high rates of correct species classification with fair probabilities over a wide range of intra/interspecific variation (Figs. 6, 7).

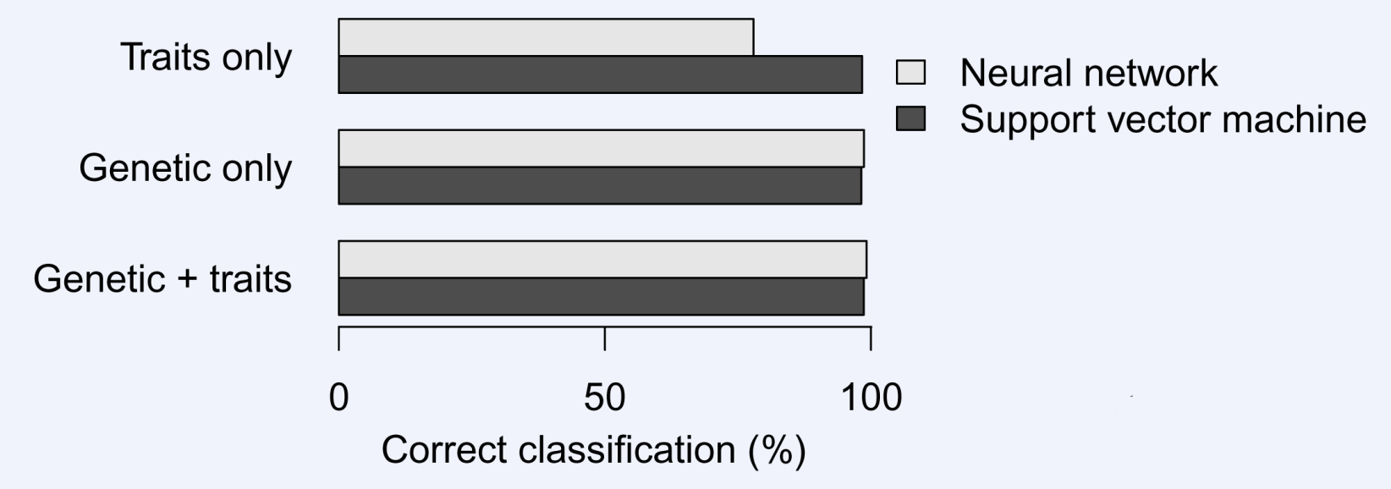


Fig. 6: Performance of proSDS exemplified on two micrograstropod genera. Ratio of correct species classification varies slightly by machine learning algorithm and utilized data types.

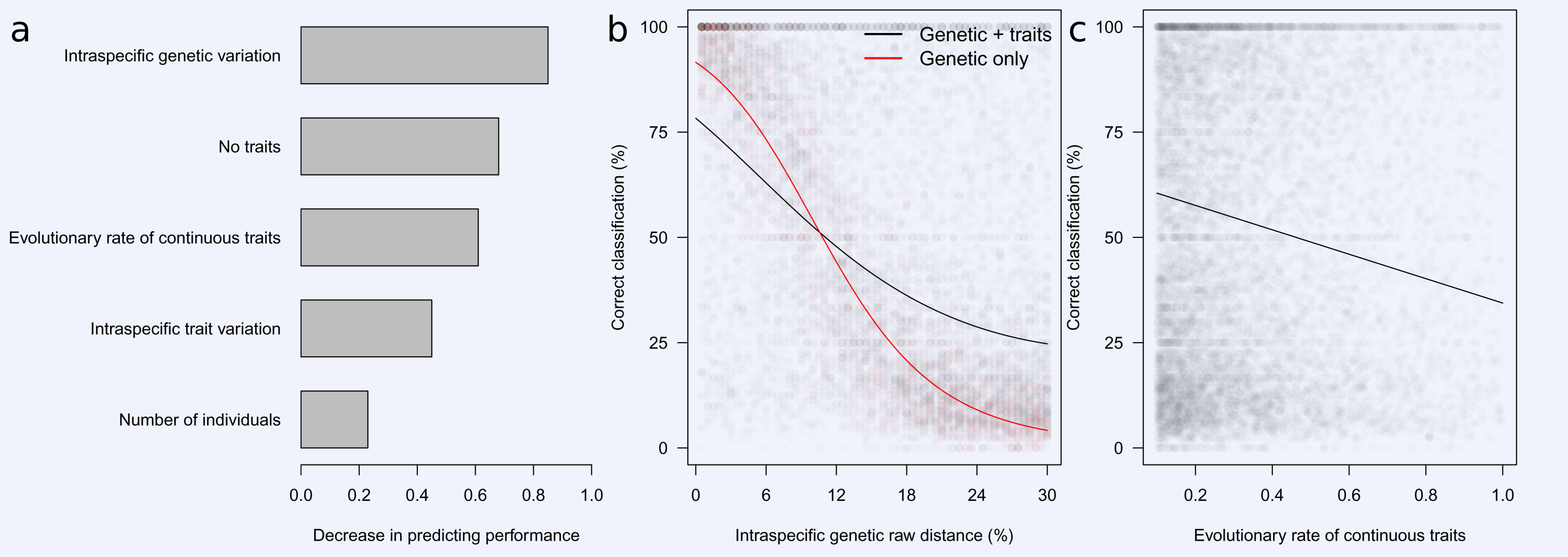


Fig. 7: Robustness and limits of proSDS inferred by regression analyses. (a) Decrease in explaining correct classification ratio upon omitting explanatory simulation features. (b) Decrease of correct classification power with increasing intraspecific variation is mitigated by morphological and ecological traits. (c) Rapid trait divergence limits performance.

Our simulations comprise simple evolutionary scenarios. We will benchmark proSDS against (i) convergent evolution[5] and (ii) speciation with gene flow and trait evolution along the resulting phylogenetic networks[6]; scenarios identified in our previous Taxon-Omics meetings as major obstacles to species delimitation.

In the long run, our approach might assist scientists in making taxonomic decisions by estimating the probability for a query specimen to belong to a known or novel species.

[1] Costello et al. Predicting total global species richness using rates of species description and estimates of taxonomic effort 2012
[2] Mora et al. Comment on 'Can we name earth's species before they go extinct? 2013
[3] Miller et al. Global species richness of hydrobiid snails determined by climate and evolutionary history 2018
[4] Delicado et al. Ecological opportunity may facilitate diversification in Palearctic freshwater organisms: a case study on hydrobiid gastropods 2018
[5] Mahler et al. Exceptional convergence on the macroevolutionary landscape of island lizard radiations 2013
[6] Bastide et al. Phylogenetic comparative methods on phylogenetic networks with reticulations 2018

Acknowledgments: We thank the Bioinformatic and Systems Biology group (Justus Liebig University, Germany) for access to high-performance computation and the DFG program SPP 1991 for funding.