National Institute for Theory and Mathematics in Biology
Funded by the U.S. National Science Foundation and the Simons Foundation

Inferring microbiome community structure from biological function

Matthew Schmitt*, Kiseok Lee*, Vincenzo Vitelli, Seppe Kuehn

Bacteria are ubiquitous on earth and assemble into exceedingly complex communities. In natural environments, it is typical for thousands of species to coexist in a stable equilibrium. While modern experimental methods allow one to measure the abundances of all these individual species, distilling these lists of members into interpretable functional groups remains difficult. Here, we present a neural network architecture which performs this reduction automatically and can identify biologically relevant groups across a range of experimental datasets.

The connection of the identified groups to community function is a key element of our approach. Methods based only on genetic similarity or phylogeny may place species in separate groups even though they behave similarly when viewed through the lens of some functional output (for example, the production of some metabolite). To perform function-informed grouping we construct neural networks which simultaneously cluster species and then predict the functional measurement of interest.

Numerical approaches to clustering are difficult for two reasons. First, the space of potential groupings is exponentially large in the number of species. Second, clustering requires an assignment to a discrete set, which renders gradient-based optimization impossible. We overcome these issues by strategically introducing stochasticity and cluster smoothing which allow us to explore the space of clusters efficiently with gradient descent. In addition to outperforming other proposed approaches to the clustering problem, our approach benefits by inherently accounting for uncertainty.

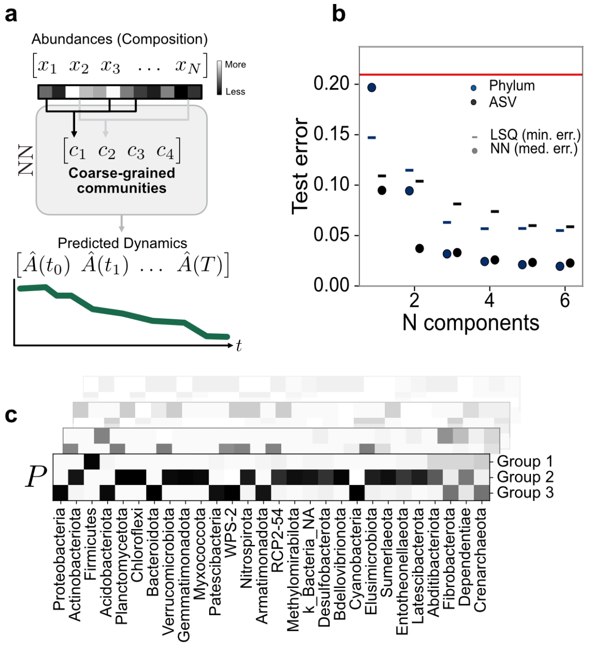

Figure: (a) We train neural networks to automatically identify groups of bacteria that are maximally predictive of metabolic function. (b) Neural networks outperform linear methods, suggesting non-linearity is essential. (c) Our approach reveals reproducible clusterings of species, accounting for uncertainty.