KBase
PREDICTIVE BIOLOGY

DOE Systems Biology Knowledgebase

Understanding Earth's Ecosystems Using Machine Learning

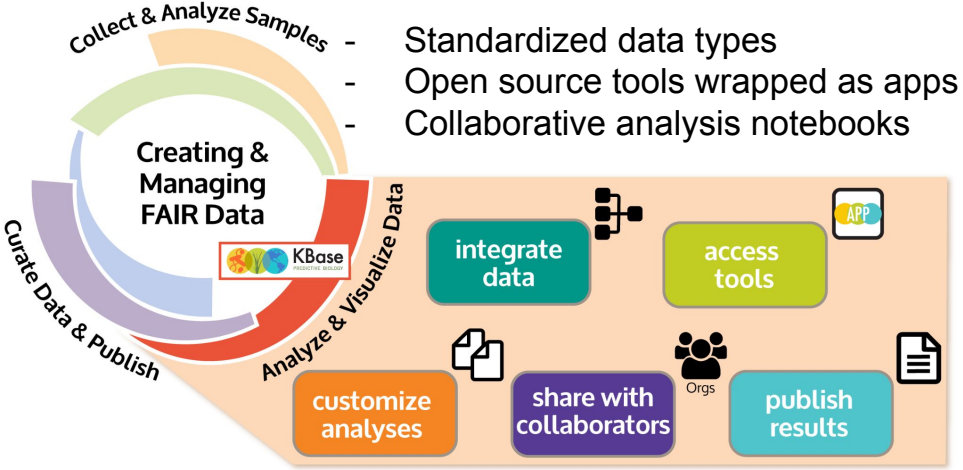
Marcin P. Joachimiak
Staff Researcher
Lawrence Berkeley National Laboratory

RuleML & KEG Seminar 8/1/23

U.S. DEPARTMENT OF ENERGY | Office of Science
Office of Biological and Environmental Research

INTEGRATION and MODELING *for* PREDICTIVE BIOLOGY

KBase: Maximize the Impact of Your Research



- Standardized data types
- Open source tools wrapped as apps
- Collaborative analysis notebooks

US Department of Energy KBase system:

- Empower scientists and teams.
- Collaboratively perform computational biology.
- Plants, microbes and their communities.

KBase **Knowledge Engine** – an internal R&D project to support prediction and inference on the system.

https://www.kbase.us

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

KBase in Numbers

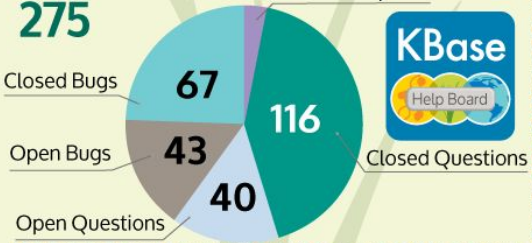
KBase Annual Report

THE NUMBERS
2022
January — December

15 Dec 22

Help Board tickets

Total **275**



Twitter

113 K	Impressions
84 K	Profile Visits
162	New Followers
1981	Followers

YouTube

11	Videos posted
15.4 K	Views
280	New Subscribers
811	Subscribers

kbase.us

28.4 K	New vistors
33 K	Visitors
90 K	Site visits

New User Accounts

Total	27319
in 2022	6188

User Narratives

Total	62.6 K
Growth	16.6 K

All User Data

Total	518 TB
Growth	161 TB

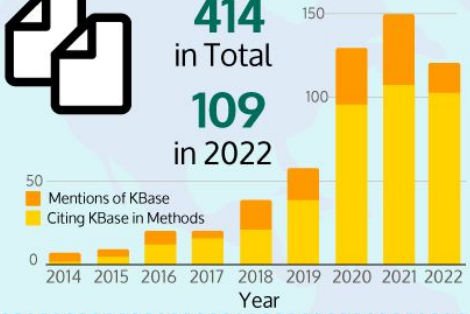
Public Data

Total	16.7 TB
Growth	0.6 TB

Publications using KBase

414 in Total

109 in 2022

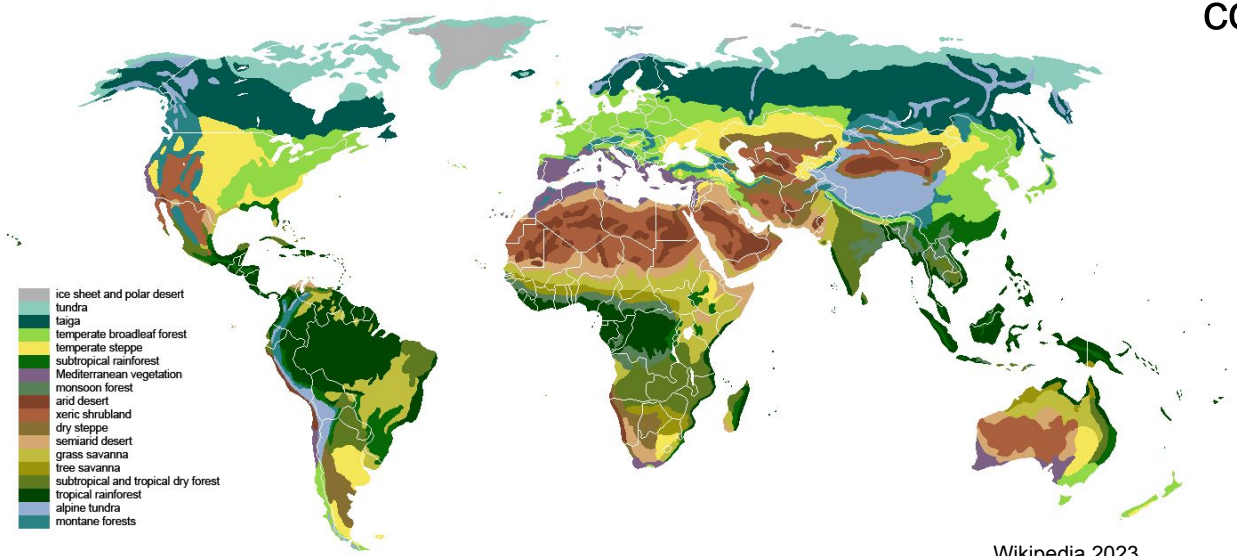


Top App Categories

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Exploring Earth's Ecosystems with Metagenomes



Wikipedia 2023

Our planet is made up of a network of interconnected biomes categorized into types.
However, our understanding is limited …

Metagenome samples are snapshots of biome communities

Analysis types

356039	amplicon
28873	assemblies
2039	metabarcoding
33827	metagenomes
2205	metatranscriptomics
2	long reads assemblies

Public data

4577	studies
346287	samples
444172	analyses
315252	genomes in 8 MAG catalogues

Or by selected biomes

Largest collection of assembled, annotated, standardized metagenomes.

www.ebi.ac.uk/metagenomics/

Associating Metagenomic Features With Ecosystems

Experimental

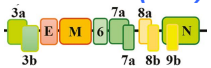
Physical samples → Reads → Assembled contigs

Expect to discover features adaptive for specific ecosystems:

O_2 = aerobic metabolism + oxidative stress functions

Protein domains

InterPro (IPR)



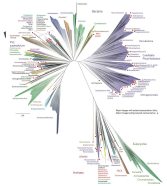
InterPro2GO

Function terms

Gene Ontology

Taxa

NCBI Taxonomy



?

?

?

?

Ecosystem

Environmental features

?

GOLD environmental classification

Blue = Standardization with ontological labels

→ Data associations

Latent associations

Context: shotgun metagenomics, EBI MGnify

KBase PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

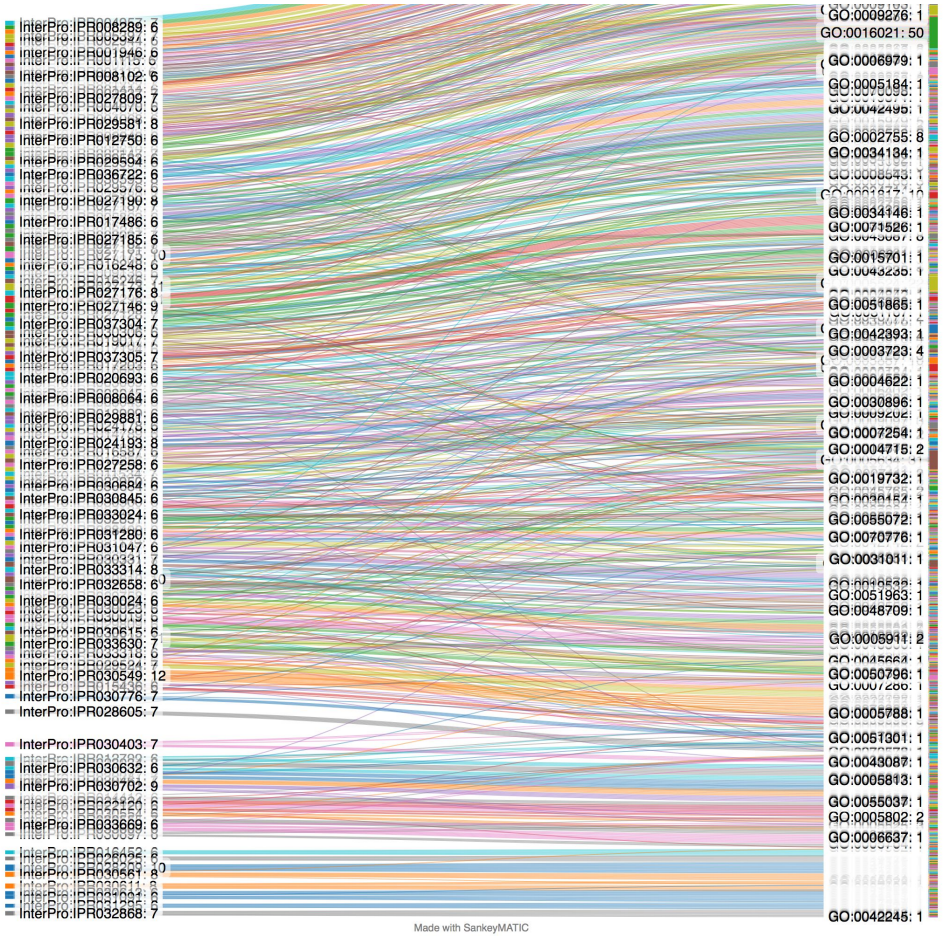
The InterPro domain to GO term mapping

Curated by InterPro.

34847 associations
15839 InterPro domains
6337 GO terms

29684 associations 1-to-many

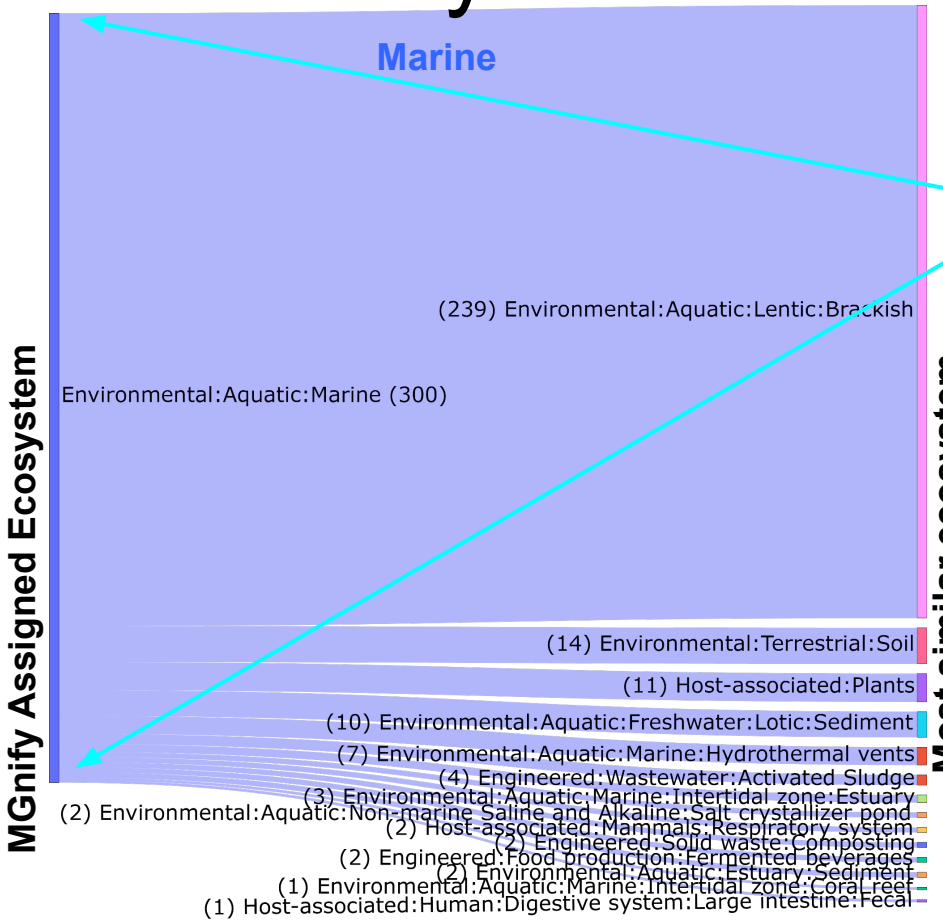
Note: Showing subset of associations with more than 5 mappings.

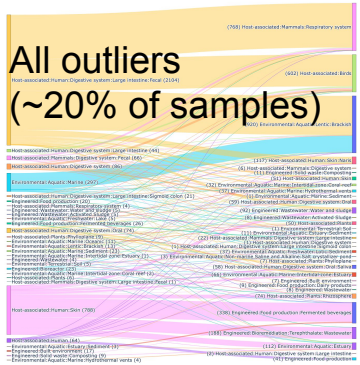


Limits of Ecosystem Labeling

Some samples are more similar to samples from another ecosystem than their original assignment.

300/686 *Environmental:Aquatic:Marine* samples more similar to another ecosystem





Cases
Classification degeneracy
Contamination
Mixture
Overly broad
Mislabeled

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

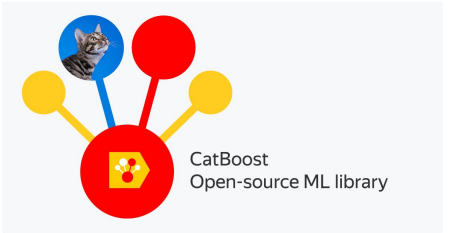
Machine Learning For Classification

Task

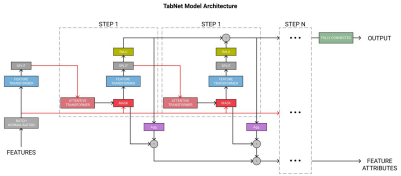
Predict sample ecosystem labels: use input data table with target classes to train models.

Machine Learning methods:

- Gradient boosted decision trees (GBDT) – CatBoost
  - Boosting means combining a learning algorithm in series to achieve a strong learner from many sequentially connected weak learners, in this case decision trees.
  - Each tree attempts to minimize the errors of the previous tree. Trees are weak learners but adding many trees in series and each focusing on the errors from the previous one make boosting a highly efficient and accurate model.

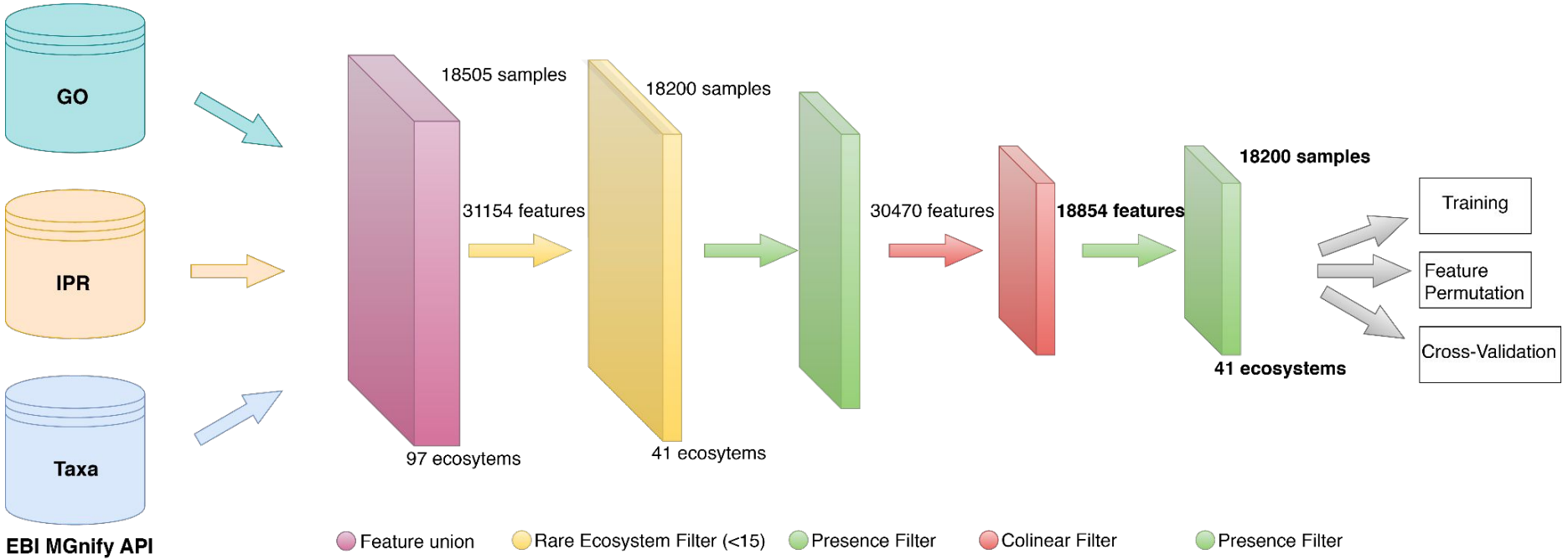


- TabNet - deep tabular learning (Google)
  - Attention mechanism to provide context for each position in input.
  - Uses sequential attention to choose which features to reason from at each decision step, enabling interpretability and more efficient learning as the learning capacity is used for the most salient features.
  - Can parallelize input processing e.g. sentence vs words.



KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

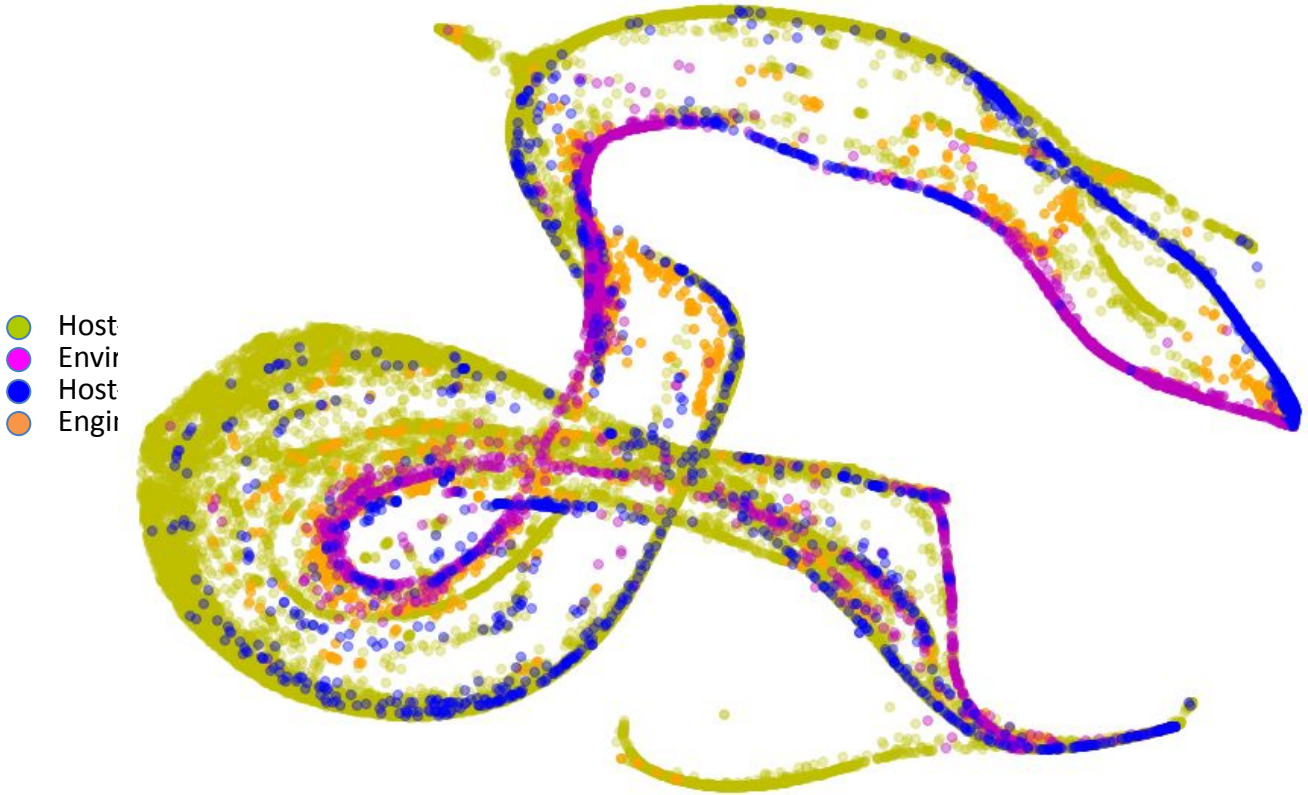
Data Preprocessing for Metagenome ML Pipeline



KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Overview of metagenome samples: GO term profiles

*tSNE of 31,939 MGnify samples representing 122 biomes. Function abundance profiles across **4,402 GO term annotations** from InterPro.*



~75% of samples are from human.

Model Evaluation Winner: CatBoost with All Feature Types

Models trained on different feature type combinations.

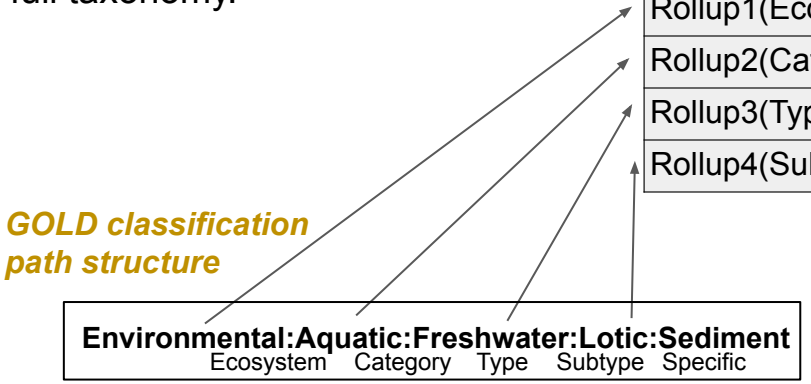
Model	Mean test accuracy
Cat: GO+IPR+Full taxa	0.89
Cat: IPR	0.89
TabNet: GO+IPR+Full Taxa	0.88
Cat: GO	0.87
Cat: Full taxonomy	0.84
Cat: Phylum taxa rollup	0.77

Hyperparameter optimization
&
10-fold cross-validation

We focus on the top full resolution model, which uses all three feature types: sequence domains, GO terms, full taxonomy.

Lower resolution models trained on environmental classification rollups.

Model	Mean test accuracy
Rollup1(Ecosystem)	0.99
Rollup2(Category)	0.99
Rollup3(Type)	0.97
Rollup4(Subtype)	0.95

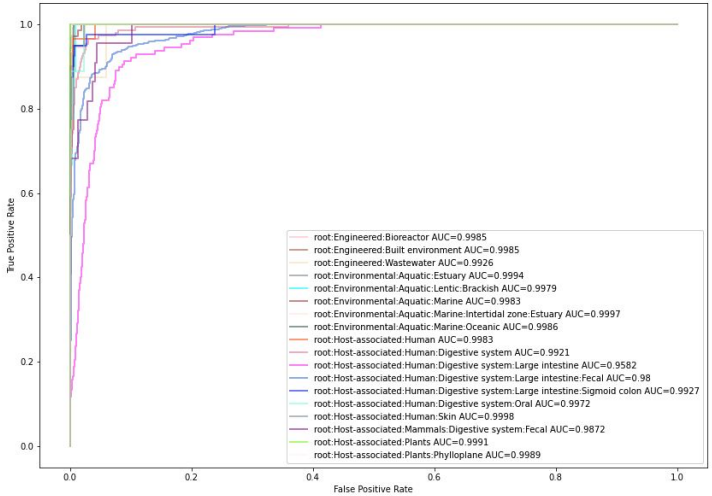


KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

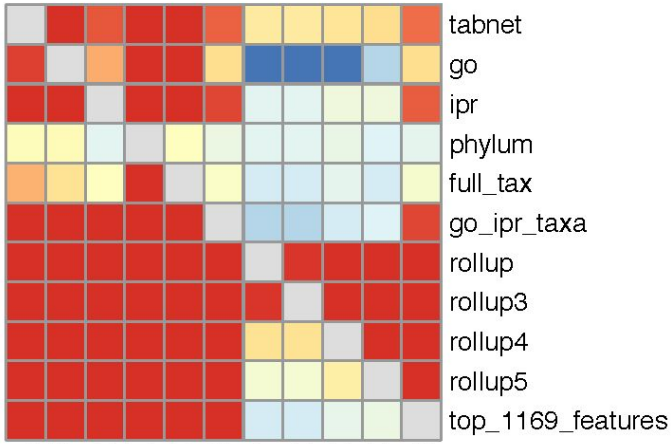
Model evaluation and selection

10-fold cross-validation

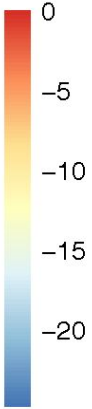
	Train Acc.	Train St. Dev.	Test Acc.	Test St. Dev.
CatBoost				
GO	0.997	0.001	0.870	0.004
IPR	1.000	0.000	0.887	0.006
GO+IPR	1.000	0.000	0.886	0.005
Phylum	0.829	0.002	0.770	0.008
Full taxonomy	0.959	0.001	0.836	0.006
GO+IPR+Full tax.	0.999	0.000	0.890	0.004
GO+ IPR +Full tax. top 148 features	1.000	0.000	0.884	0.004
GO+ IPR +Full tax. rollup3	1.000	0.000	0.986	0.001
GO+ IPR +Full tax. rollup4	1.000	0.000	0.972	0.002
GO+ IPR +Full tax. rollup5	1.000	0.000	0.949	0.003
TabNet				
GO+ IPR +Full tax.	0.984	0.012	0.874	0.011



Model paired cross-validation accuracy tests



Log_{10} p-value

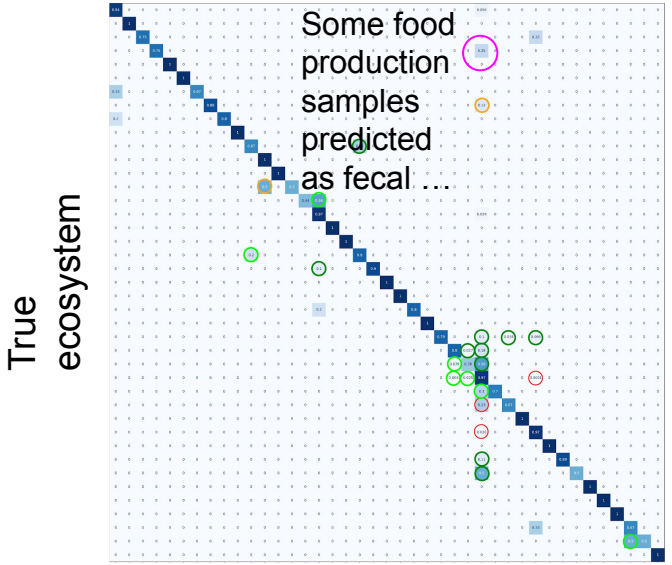


Environmental:Aquatic:Freshwater:Lotic:Sediment
Ecosystem Category Type Subtype Specific

Best full resolution models

KBase PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Model Errors Can Be Explained With Semantic Relations



Possible contamination/Mixture

Hypernym

Hyponym

Contextual synonym

Examples:

Engineered:Food production
Vs
Human:Digestive system:Large intestine:Fecal

Human:Digestive system:Large intestine
Vs
Human:Digestive system:Large intestine:Fecal

Human:Digestive system:Large intestine
Vs
Human:Digestive system

Engineered:Wastewater
Vs
Human:Digestive system:Large intestine:Fecal

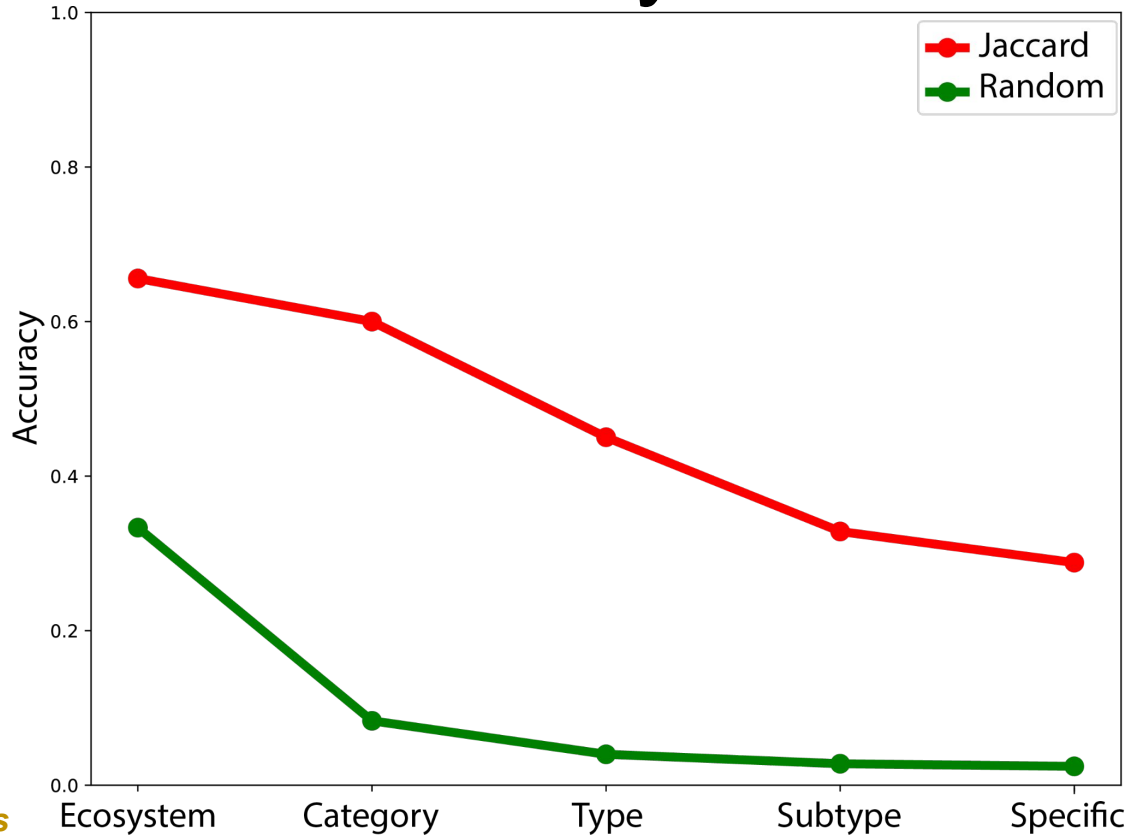
KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Semantic Similarity Reveals Model Generalizability

305 samples were from ‘rare’ ecosystems without enough samples for training.

We used these untrained samples to assess model generalizability.

Semantic similarity was significantly higher than expected by chance.

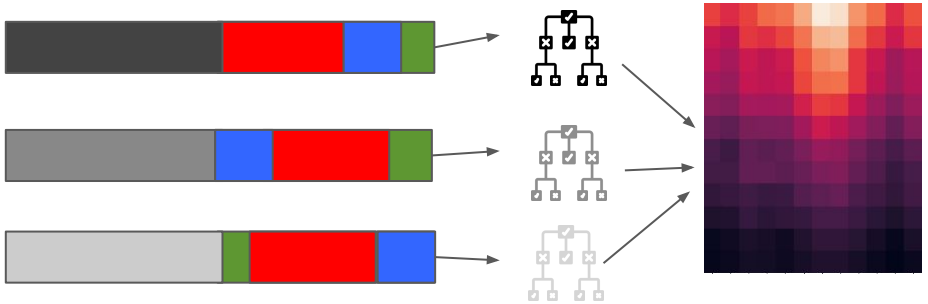


GOLD classification levels

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Important Features

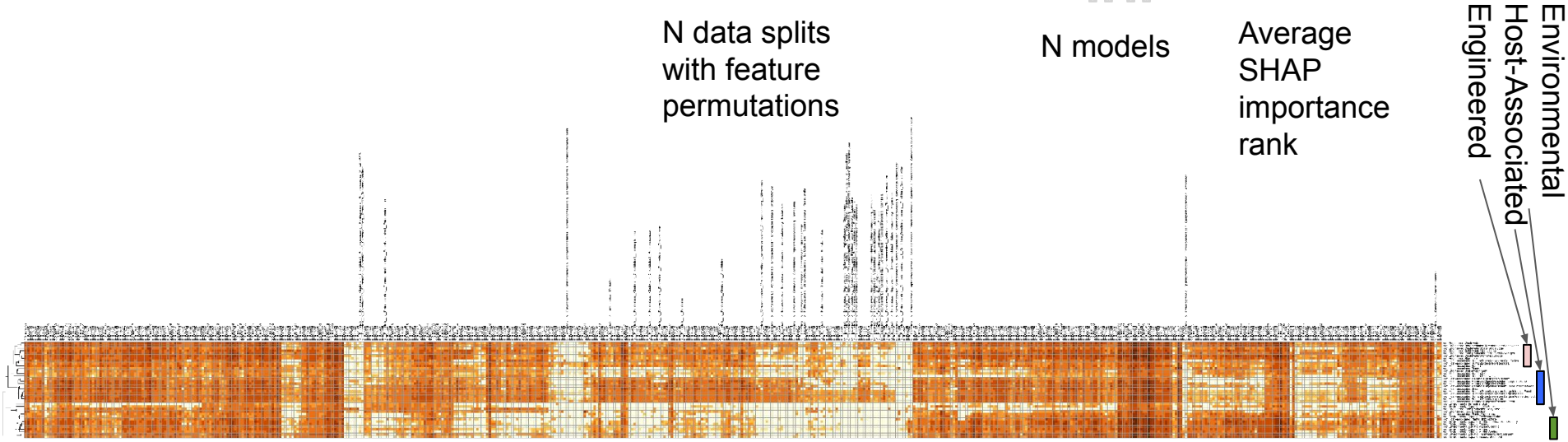
568 important features for overall ecosystem classification from feature permutation model analysis.



N data splits with feature permutations

N models

Average SHAP importance rank



Reveals data patterns that relate classification features to the ecosystem hierarchy.
Hierarchy is not used for training.

TabNet SHAP analysis requires > 1 TB RAM :(

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Feature Importance Values

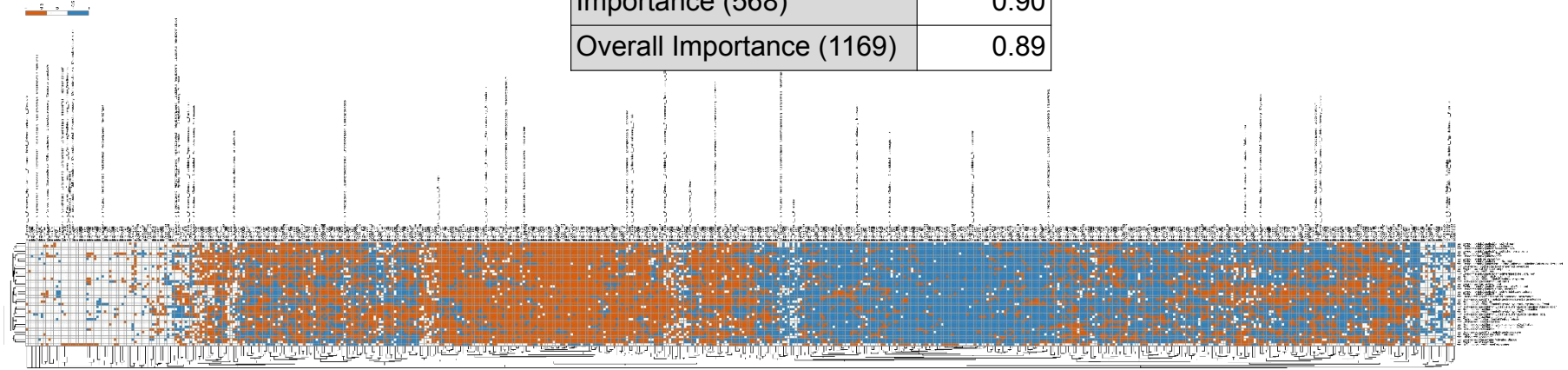
568 important features for overall ecosystem classification from feature permutation model analysis.

Feature importance models

Model	Mean test accuracy
Ecosystem-specific Importance (568)	0.90
Overall Importance (1169)	0.89

Positive importance - contributes information to class prediction

Negative importance - contributes to confusion for that class



Rare

Mixed +/-

Mostly + importance

Mostly - importance

Mixed +/-

KBase PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Important Feature Vignettes

Feature id	Feature label	Signal/chemical	Similar features	In ecosystem
IPR022387	Carbohydrate ABC transporter substrate-binding, CPR0540	**carbohydrates**		Skin, Marine, Digestive system, Brackish
IPR000060	Betaine/Carnitine/Choline Transporter (BCCT)	**quaternary nitrogen compounds**	*IPR018093* - BCCT transporter, conserved site	Respiratory system, Oceanic, Salt crystallizer pond, Skin:Naris
GO:0009584	Visible detection of light	**light/infrared**	*IPR001294* - Phytochrome, *GO:0018298* - obsolete protein-chromophore linkage, *IPR013654* - PAS fold-2, *IPR013515* - Phytochrome, central region	Large intestine, Fecal, Marine, Marine:Sediment
IPR017813	Mycothiol acetyltransferase	**mycothiol**	*IPR021678* - Protein of unknown function DUF3263	Digestive System, Large intestine, Fecal, Wastewater

Extreme importance

Biology/ecosystem pattern

KBase PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Why Would Phytochromes Be Present In Gut Microbes?

Phytochromes across all ecosystems

*Bacteria 3118 / **5***
*Eukaryota 1166 / **1***
Viridiplantae 36
Bamfordvirae (viruses) 2
Fungi 3
Archaea 3
Root
Expression vector
Uncultured organism
Unknown

Most likely a digestive ecosystem secondary biomarker due to diet or environment.

Phytochrome photoreceptors absorb far-red and near-infrared (NIR) light and regulate light responses in plants, fungi, and bacteria.

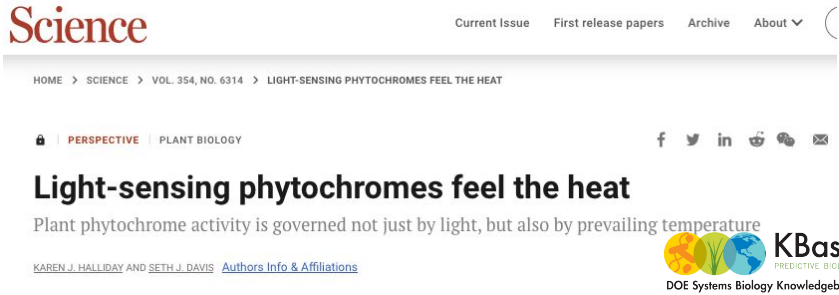
A subclass of bacterial phytochromes (BphPs) utilizes heme-derived biliverdin tetrapyrrole, which is ubiquitous in mammalian tissues, as a chromophore. Because biliverdin possesses the largest electron-conjugated chromophore system among linear tetrapyrroles, BphPs exhibit the most NIR-shifted spectra that reside within the NIR tissue transparency window.

doi: 10.1021/acs.chemrev.6b00700

WIKIPEDIA

The **near-infrared (NIR) window** (also known as optical window or therapeutic window) defines the range of wavelengths from 650 to 1350 nanometre (nm) where light has its maximum depth of penetration in tissue.

NIR window is primarily limited by the light absorption of blood at short wavelengths and water at long wavelengths. The technique using this window is called NIRS. Medical imaging techniques such as fluorescence image-guided surgery often make use of the NIR window to detect deep structures.



KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

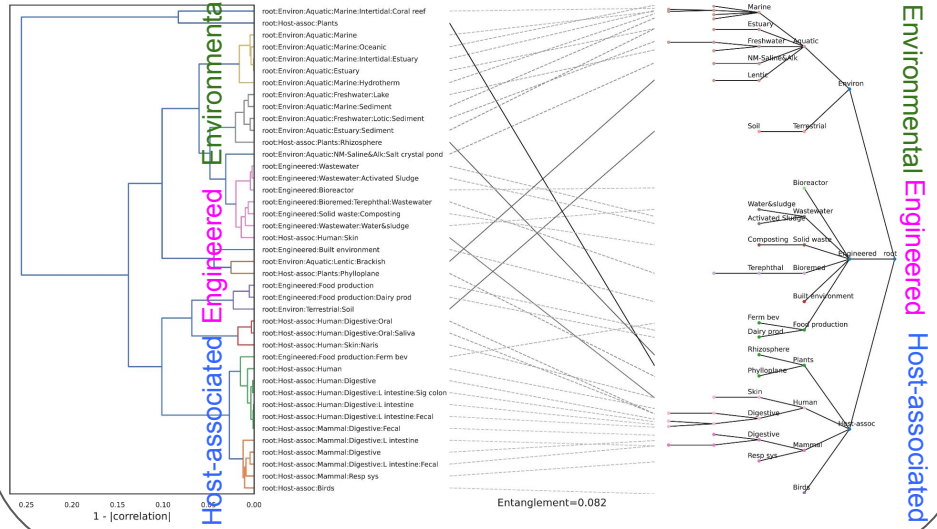
20% Of Features Represent All Features And Ecosystem Hierarchy

All features

Important + untrained similar

Important + untrained similar

GOLD classification hierarchy



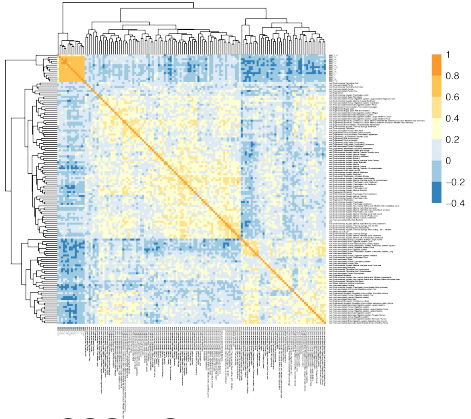
Important+similar features represent information about higher order relationships in the ecosystem classification.

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

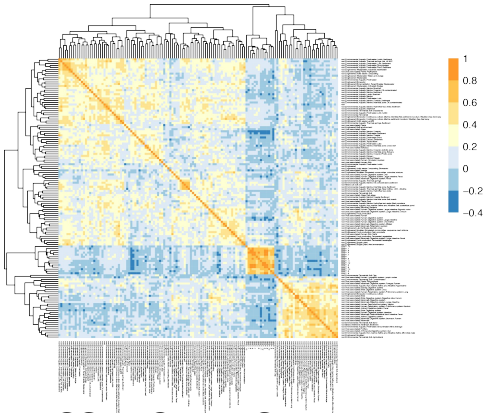
Feature selection matters more ...

Features restricted important **GO terms** from a CatBoost biome classification model.

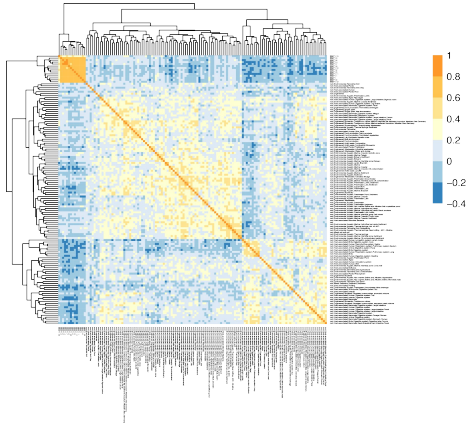
CSS - Pearson



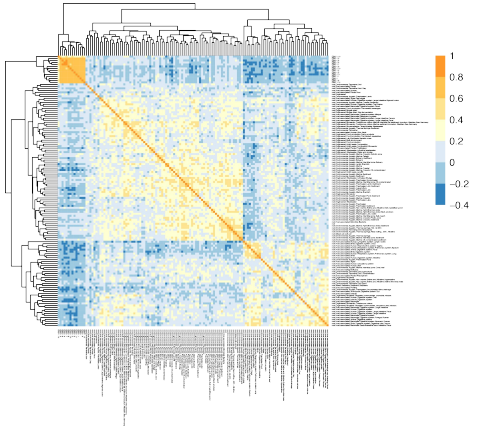
TSS + CLR - Pearson



CSS - Spearman



TSS + CLR - Spearman



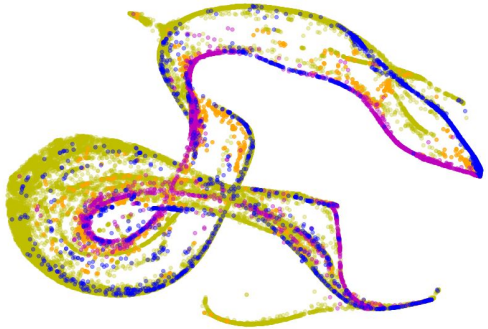
Important functions help resolve metagenome differences

Some host-associated samples still overlap with human samples, suggesting label improvements.

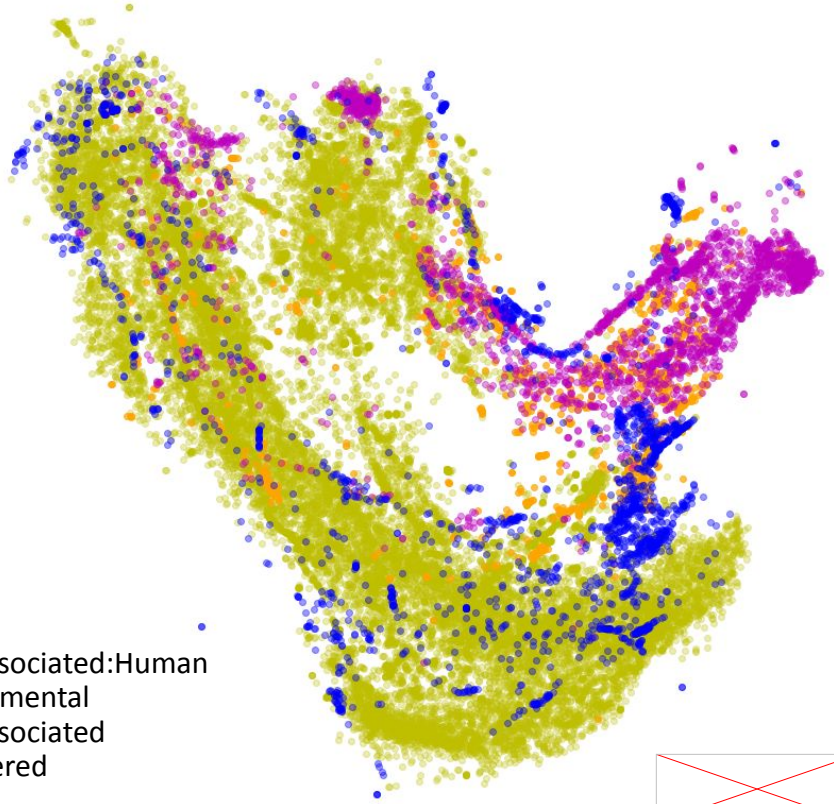
For example, most samples with the label:

Host-associated:Mammals:Digestive System:Large Intestine:Fecal

are predicted to be the ***Human*** equivalent biome.



All features (GO)



Important Feature Similarity Network

<u>568 ecosystem-specific important features</u>

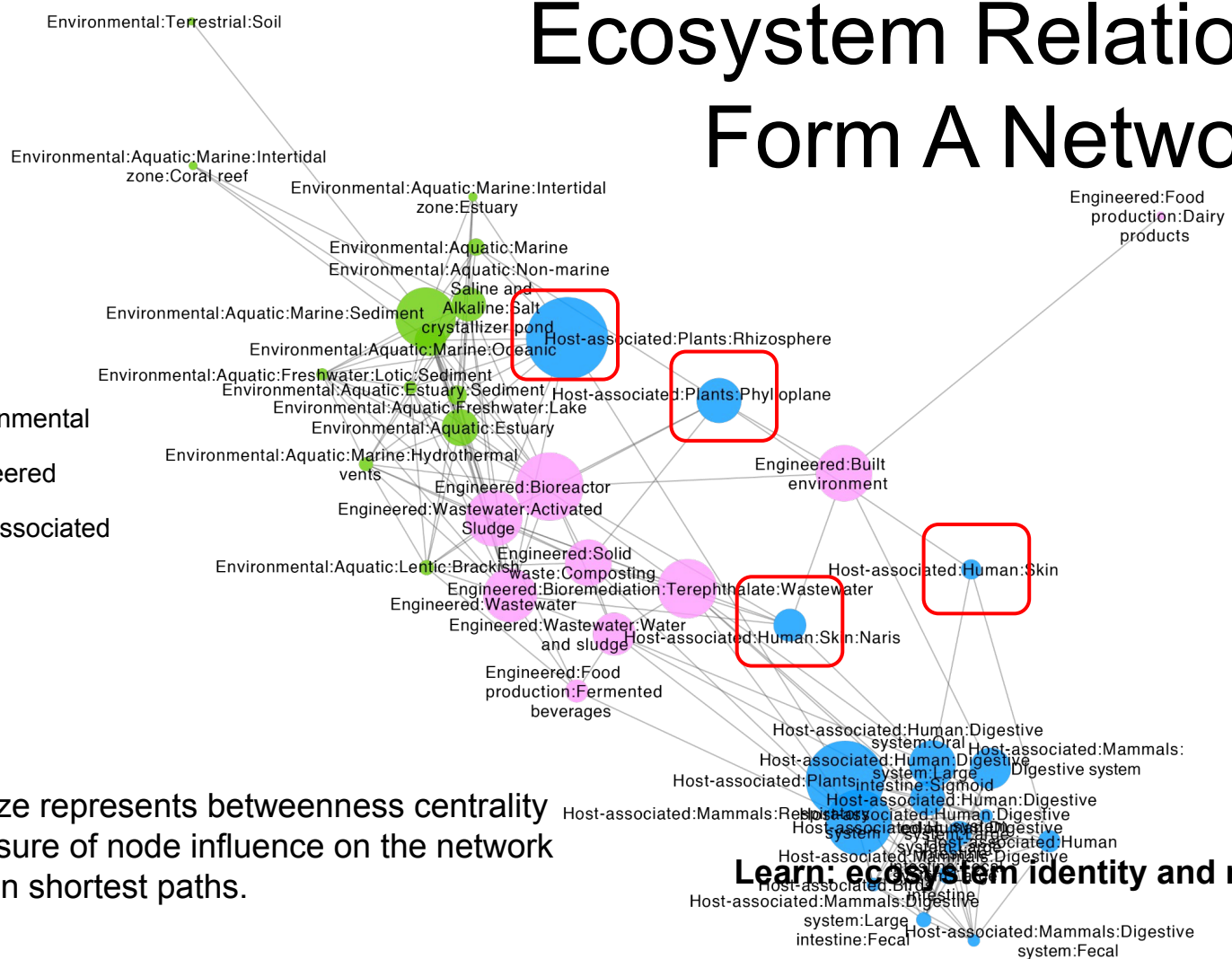
Taxa	36
GO	90
IPR	442
No IPR -> GO	340
DUF IPR	110

*DUF = Domain of Unknown Function

210 important features have untrained similar features.
127 important IPR domains have untrained similar GO feature (and no taxa).

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Ecosystem Relationships Form A Network



- Environmental
- Engineered
- Host-associated

Node size represents betweenness centrality - a measure of node influence on the network based on shortest paths.

Learn: ecosystem identity and relationship rules!

KBase PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

KG-MicroCult: Scientific Approach

Prediction tasks

1. Withhold known taxa-media pairings
2. Known examples of discovering defined media starting from rich media
3. Known examples of discovered growth conditions
4. Selection of 100 microbes important for LBL projects

Difficulty

Computational model training evaluation

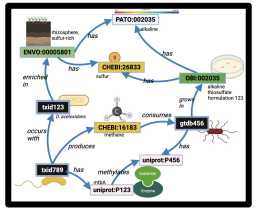
Experimental evaluation

- *Assess growth using HT optical density readout from plates.*
- *Physical replication and similar condition sampling for robust assessment.*
- *Attempt to validate growth condition discoveries with functional annotations and sample metadata.*

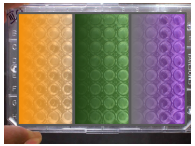
1. User or AI pick microbe(s) or conditions

4. while() - KG data update and reasoning

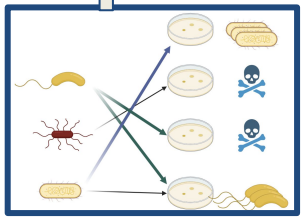
KG



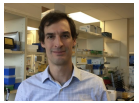
1b: Predict starting points



2. Design/walk growth parameters and ranges

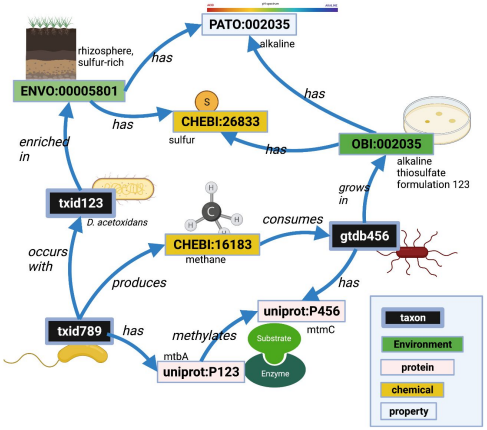


3. Run HT experiment & Record data



KG-MicroCult: Expected Outcomes

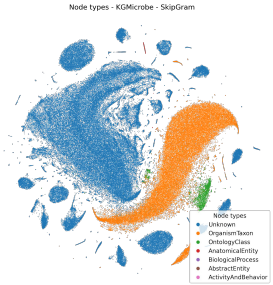
Microbial Culturing Knowledge Graph (KG-MicroCult)



AI/ML applications

AI/ML

Knowledge graph embeddings



Nature Comp. Sci. 2023

https://github.com › AnacletoLAB › grape

AI/ML

Predict growth parameters

Near term impact

Culturing recipes

Live cells

Predictive model and service

Integration of culturing experiments

Major impacts:

- Current and standardized microbial culturing data including reference knowledge and experimental results.
- Reproducibly, systematically expands culturing data and catalog of culturable microbes.
- **New model organisms, chassis for bioengineering, experimental hypothesis testing.**
- **Base R&D platform: bioremediation, biomanufacturing, drug discovery, agriculture, food production, communities for health and environment.**

KG-Microbe -- a reference microbial knowledge graph

Another KG based on the KG-Hub KG build template. Also our prototype for introducing Named Entity Recognition (NER) and more broadly Natural Language Processing (NLP) into the KG construction workflow.

- Bacteria (eubacteria) *Click on organism name to get more information.*
  - Acidobacteria
    - Acidobacteriia
      - Acidobacteriales
      - Bryobacterales
      - Candidatus Acidoferrales
      - unclassified Acidobacteriia
      - environmental samples
    - Blastocatellia
      - Blastocatellales
      - Chloracidobacterium
      - unclassified Blastocatellia
      - environmental samples
    - Holophagae
      - Acanthopleuribacterales
      - Holophagales
      - Thermotomaculales
      - unclassified Holophagae
      - environmental samples
    - Thermoanaerobaculia
      - Thermoanaerobaculales
      - unclassified Thermoanaerobaculia
    - Vicinamibacteria
      - Vicinamibacteraceae
      - unclassified Vicinamibacteria
    - Acidobacteria incertae sedis
      - Acidobacteria subdivision 13
      - Acidobacteria subdivision 2
      - Acidobacteria subdivision 22
      - Candidatus Guanabacteria
    - unclassified Acidobacteria
      - Acidobacteria bacterium
      - Acidobacteria bacterium 01QDG
      - Acidobacteria bacterium 01QDH_23S
      - Acidobacteria bacterium 13_1_20CM_2_55_15
      - Acidobacteria bacterium 13_1_20CM_2_57_8
      - Acidobacteria bacterium 13_1_20CM_2_59_6

Data Descriptor | Open Access | Published: 05 June 2020

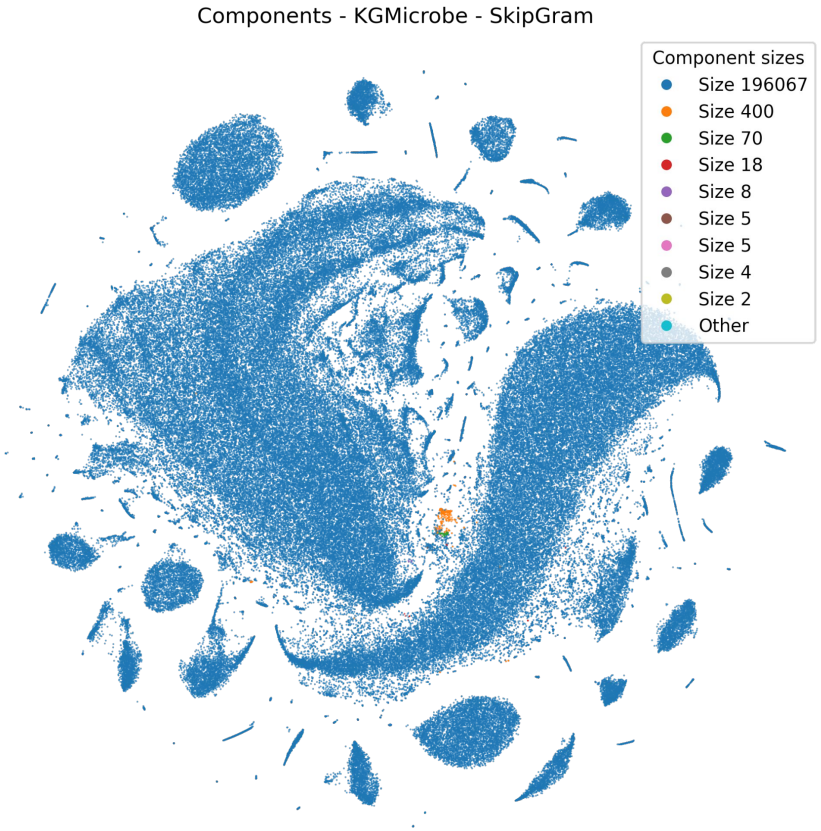
A synthesis of bacterial and archaeal phenotypic trait data

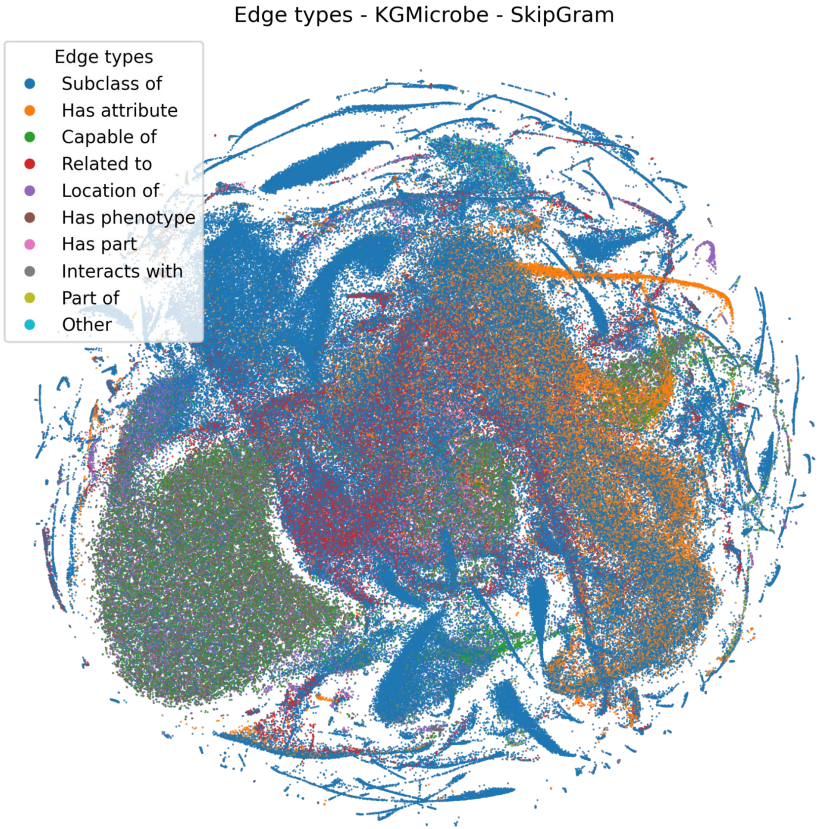
Joshua S. Madin ✉, Daniel A. Nielsen, [...] Mark Westoby

With Chris Mungall and Harshad Hegde

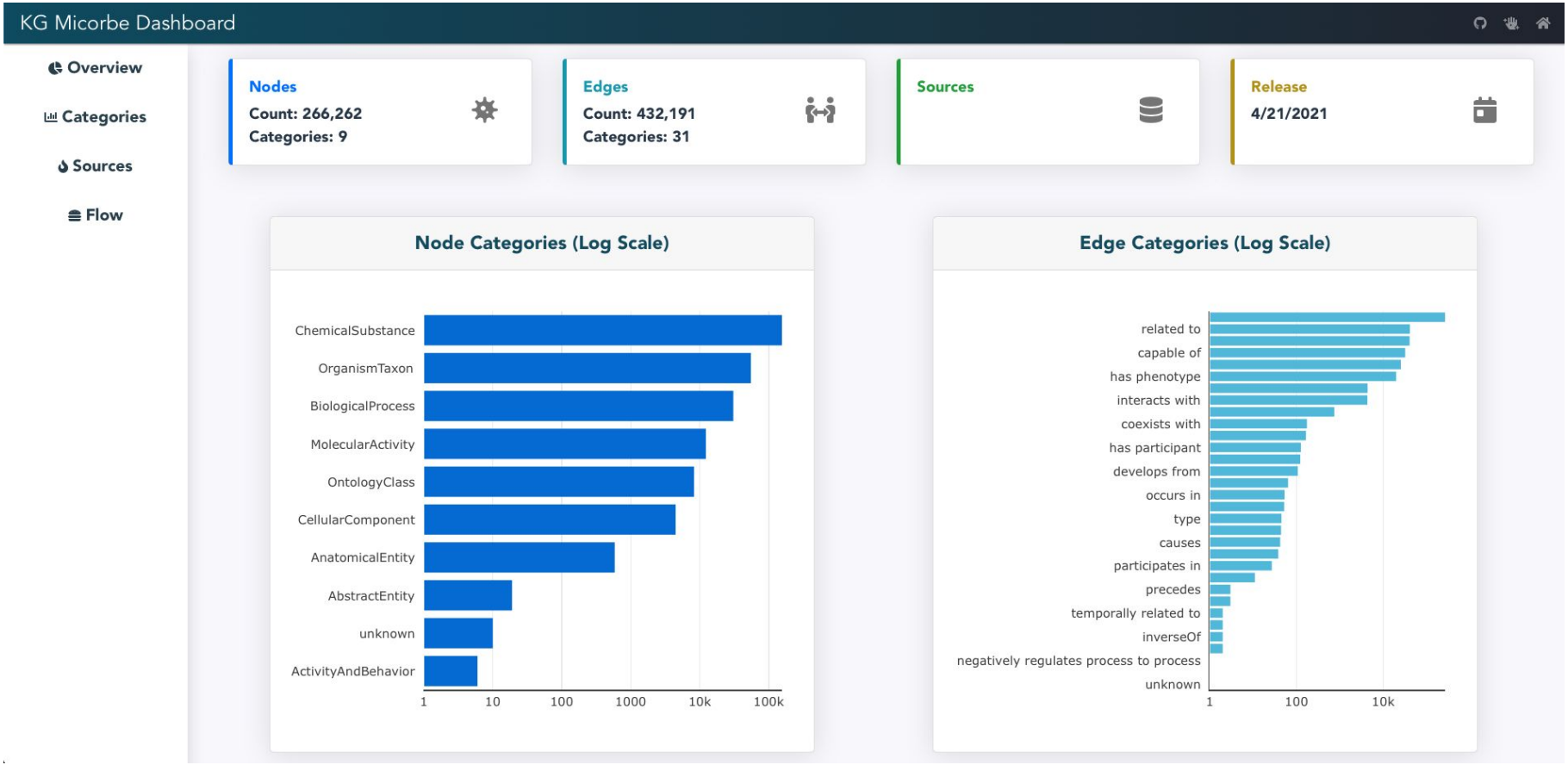
Readthedocs: https://knowledge-graph-hub.github.io/kg-microbe/index.html

KG-Microbe dimensionality reduction





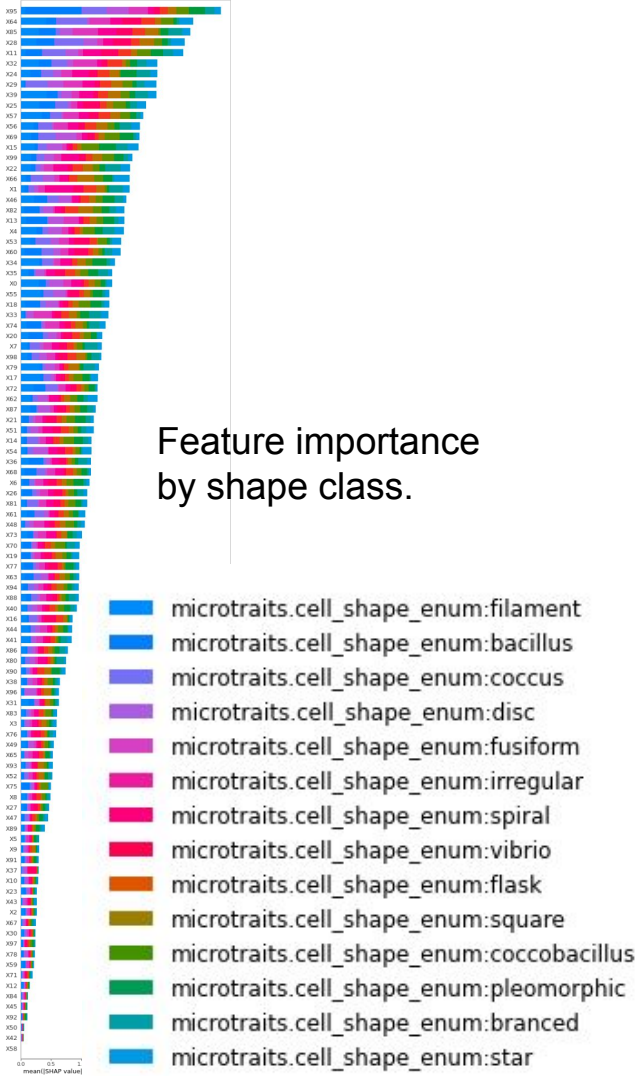
KG-Microbe graph stats dashboard



Predicting … microbial traits … from graph embeddings

Confusion matrix for **microbial shape** predictions

Feature importance by shape class.



Many prediction targets of interest: pathways, metabolism mode, growth substrates.

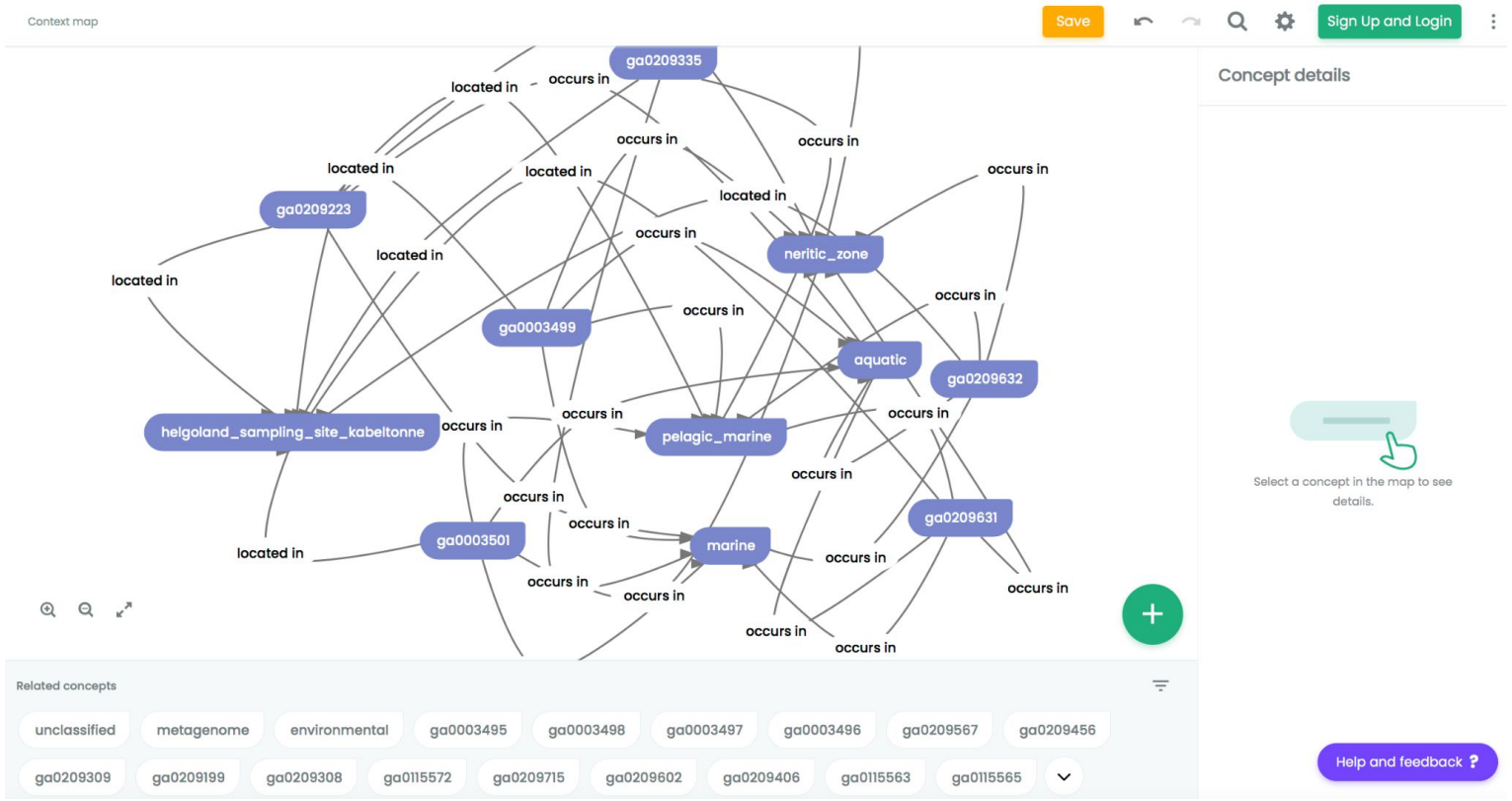
Unbiased performance evaluation of microbial shape predictions from graph embeddings

	precision	recall	f1-score	support
microtraits.cell_shape_enum:bacillus	0.99	0.98	0.98	2343
microtraits.cell_shape_enum:coccus	0.93	0.94	0.94	103
microtraits.cell_shape_enum:spiral	0.95	0.98	0.97	1123
microtraits.cell_shape_enum:vibrio	1.00	0.75	0.86	4
microtraits.cell_shape_enum:coccobacillus	0.99	0.98	0.98	96
microtraits.cell_shape_enum:filament	1.00	1.00	1.00	1
microtraits.cell_shape_enum:pleomorphic	1.00	1.00	1.00	3
microtraits.cell_shape_enum:disc	1.00	0.50	0.67	2
microtraits.cell_shape_enum:fusiform	1.00	1.00	1.00	11
microtraits.cell_shape_enum:flask	1.00	0.97	0.99	157
microtraits.cell_shape_enum:star	1.00	0.50	0.67	2
microtraits.cell_shape_enum:square	1.00	1.00	1.00	1
microtraits.cell_shape_enum:irregular	1.00	0.99	1.00	102
accuracy			0.98	3948
macro avg	0.99	0.89	0.93	3948
weighted avg	0.98	0.98	0.98	3948

As before, predictions using 80% graph embeddings from stratified 80/20 split of target edges.

Evaluation using 20% held out edges and edge embeddings from 80% graph embeddings.

ContextMinds: visualizing and interacting with KGs



https://contextminds.vse.cz/imgvr/

Marek Dudas and Tomas Kliegr, VSE Prague

Conclusions

- Machine learning can reduce feature noise and improve interpretability.
- ~20% of all features can retain ecosystem relationships and the environmental hierarchy.
- Important features are a data product and resource that can support biological discovery.
- Need more: high quality data and metadata, curation, and data and processing standardization, and memory.

Acknowledgements

The Department of Energy Systems Biology Knowledgebase (KBase) Team

Special thanks to:

The KBase Knowledge Engine team

- Ziming Yang
- Sean Jungbluth
- William Riehl
- Prachi Gupta
- Chris Neely
- Meghan Drake
- Shane Cannon
- Paramvir Dehal
- Adam Arkin

U.S. DEPARTMENT OF ENERGY
Office of Science

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase