USDA United States Department of Agriculture
Agricultural Research Service

SCINet Newsletter: April 2023

Research Spotlight | News | Training | Support | Connect

RESEARCH SPOTLIGHT

Research Spotlight: A Machine Learning Tool to Collapse Diet or Microbiome Data Using Taxonomic Structure

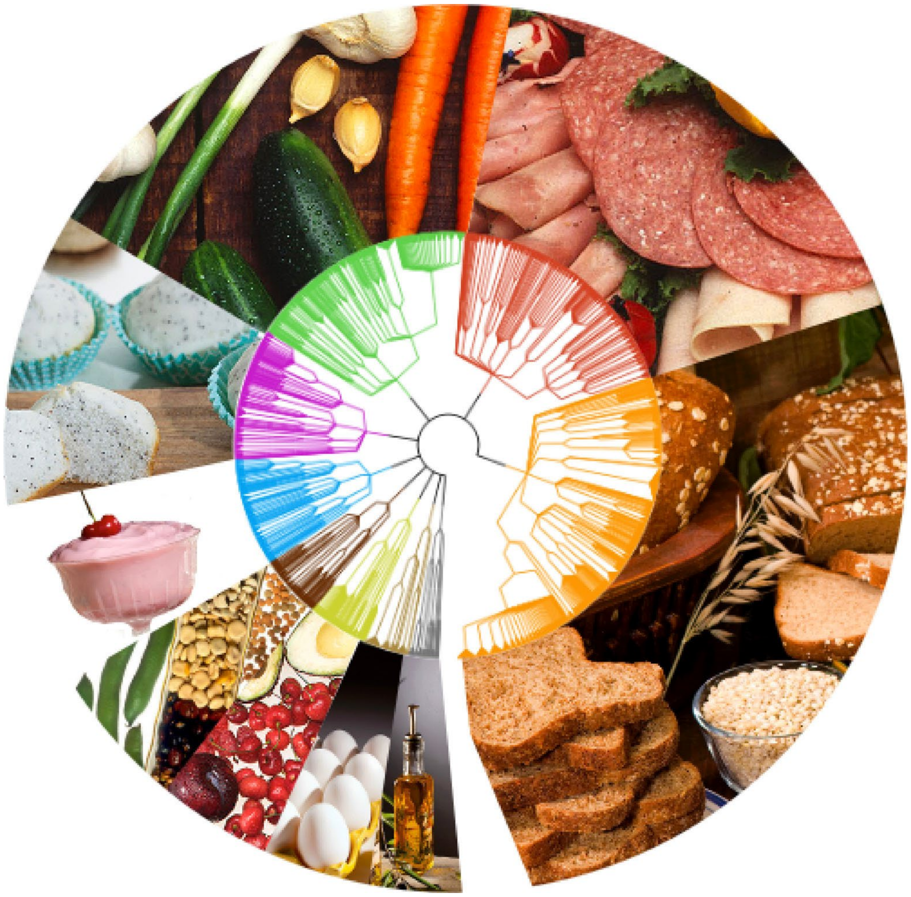


Figure 1. Taxonomic relationships of foods consumed in a USDA cohort

By Andrew Oliver, SCINet Fellow, Western Human Nutrition Research Center, Davis, CA

Researchers at the USDA-ARS Western Human Nutrition Research Center (WHNRC) are broadly interested in how diet impacts human health. Dr. Danielle Lemay, a Research Molecular Biologist in the Immunity and Disease Prevention Research Unit, together with SCINet Postdoctoral Fellow Dr. Andrew Oliver contribute to this research theme by leveraging machine learning to uncover the molecular underpinnings of diet's association with health. Of particular interest is the role of the human gut microbiota – a vast and diet-responsive community of microorganisms that live inside our digestive tract and contribute to health through production of beneficial metabolites and commensal interactions with our immune system.

One challenging aspect of investigating the diet-microbiome-health axis is how highly dimensional the data are. For example, a person may consume dozens of different foods in a week, which move through a digestive tract containing thousands of different microbial species, which collectively employ millions of genes to make or modify hundreds of metabolites found in fecal samples. These combinations are highly personalized; indeed, no two individuals share the same gut microbes.

To reduce the dimensionality of the problem, Dr. Oliver developed an algorithmic approach to leverage a property of microbiome data: microbial taxonomy. A common question in microbiome analyses is, “Which microbial species or genera co-vary with my treatment or physiological trait of interest?” This question can be problematic for several reasons; specifically, it 1) assumes the microbiome variability associated with a trait is conserved at the taxonomic level analyzed (e.g., the species or genus) and 2) forces an analysis that can handle hundreds to thousands of (likely) zero-inflated features. Dr. Oliver's program, called taxaHFE (for taxonomic hierarchical feature envineering), examines every taxonomic level of each feature and determines if it adds discriminatory value, collapsing features to higher taxonomic levels if they are redundant or uninformative. In a test of six published microbiome studies, applying taxaHFE resulted in an 88% reduction in the number of features. Moreover, this feature reduction came at no cost to the downstream predictive power of the microbiome in machine learning models. Indeed, machine learning models using data processed with taxaHFE performed better than models trained on any single taxonomic level.

Beyond the microbiome, taxaHFE can be applied to other data structures containing hierarchical features. Dr. Mary Kable, a colleague of Dr. Lemay and Dr. Oliver, has investigated diet-microbiome relationships by placing foods on a “food tree” (PMID: 34958387). Closely related foods exist more proximally on the food tree, much like a phylogeny of microorganisms. taxaHFE can be applied to this novel method for diet analysis, allowing nutritionists to investigate which foods or general food groups are associated with a particular health outcome.

To develop taxaHFE, and the many other questions Dr. Oliver and Dr. Lemay investigate, access to HPC environments is critical. taxaHFE runs hundreds of models to collapse high dimensional data and was designed to utilize parallel compute resources. Containerization of taxaHFE, using Singularity, and the high performance of SCINet resources such as Ceres enables access to this method across ARS.

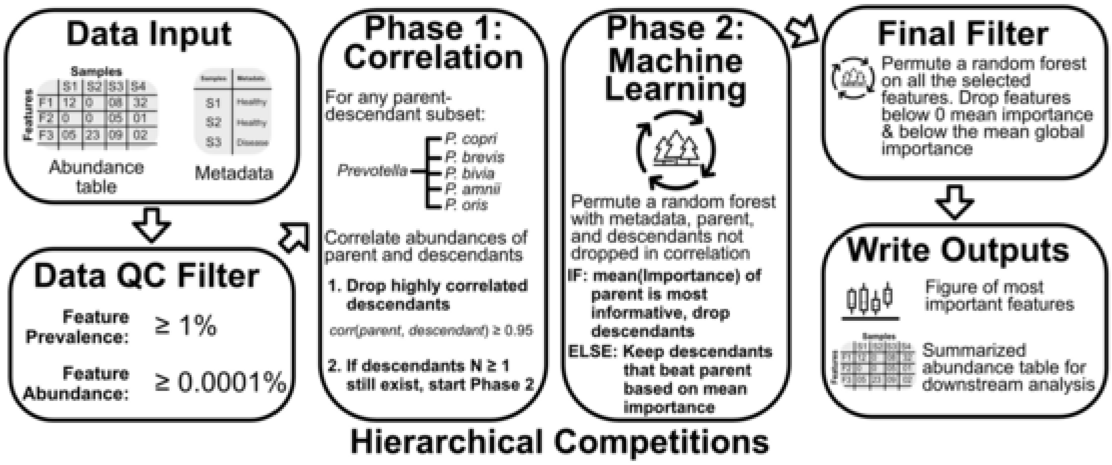

Data Input
Samples
Features
S1 S2 S3 S4
F1 12 0 08 32
F2 0 0 05 01
F3 05 23 09 02
Abundance table
Samples Metadata
S1 Healthy
S2 Healthy
S3 Disease
Metadata
Data QC Filter
Feature Prevalence: ≥ 1%
Feature Abundance: ≥ 0.0001%
Phase 1: Correlation
For any parent-descendant subset:
Prevotella
P. copri
P. brevis
P. bivia
P. amnii
P. oris
Correlate abundances of parent and descendants
1. Drop highly correlated descendants
corr(parent, descendant) ≥ 0.95
2. If descendants N ≥ 1 still exist, start Phase 2
Phase 2: Machine Learning
Permute a random forest with metadata, parent, and descendants not dropped in correlation
IF: mean(Importance) of parent is most informative, drop descendants
ELSE: Keep descendants that beat parent based on mean importance
Hierarchical Competitions
Final Filter
Permute a random forest on all the selected features. Drop features below 0 mean importance & below the mean global importance
Write Outputs
Figure of most important features
Summarized abundance table for downstream analysis

Figure 2: Outline of taxaHFE algorithm

SCINet and AI COE Fellows

Welcome Dr. Suzanne M. Fleishman, SCINet/AI-COE postdoctoral fellow! Dr. Fleishman is interested in applying data-intensive methods to clarify the relationship between belowground interactions and agroecosystem function. She received a MS in Horticulture (2018) and PhD Ecology (2023) from Penn State University, with projects examining inter-species competition in a grapevine-groundcover agroecosystem. Her research involved integrated approaches to characterizing root function, with a particular focus on root-associated microbiomes and root multi-omics (transcriptomic and metabolomic). As a SCINet Fellow working with Dr. Adam Rivers in the Genomic and Bioinformatic Research Unit, Dr. Fleishman is developing a web app that facilitates rapid search and analysis of metagenomes in NCBI's sequence read archive (SRA). Through her work, she hopes to support other researchers by providing accessible bioinformatic tools that streamline the analysis of large genomic datasets.

NEWS

Congratulations AI Innovation Fund Awardees

Five ARS scientists emerged from an incredibly competitive field to receive 2023 Artificial Intelligence Innovation Fund awards (in alphabetical order by lead PI's last name):

- **Lina Castano-Duque**: *Development of predictive models and a digital interface to manage mycotoxin outbreaks in the USA*
- **Kossi Nouwakpo**: *A Deep Learning model for irrigation methods mapping*
- **Christopher L. Owen** and Alex Konstantinov: *Using museum collections to develop a machine learning application to stop biological invasion*
- **Melissa Johnson**: *CoffeeMD: Development of machine learning to rapidly identify damage from coffee pests, diseases, and nutrient deficiencies*
- **Jana Lee** and Tim Warren: *Automated, image-based monitoring of spotted-wing* Drosophila *using deep learning*

We look forward to sharing the results from these exciting research projects with the SCINet community!

Congratulations SCINet/AI-COE Fellowship Mentors

22 ARS scientists were awarded funding to host SCINet/AI-COE postdoctoral fellows. This was another highly competitive application process. Congratulations to the following mentors (in alphabetical order by lead mentor's last name):

- **Rachel A Koch Bach**: *Determining the Evolutionary Potential of the Causal Agent of Coffee Leaf Rust Disease from Big Data*
- **Raman Bansal**: *Solving Big Data Challenges in California's Tree Nut Pests*
- **Paula Chen**, Norman Best, David Kang, Bethany Redel, Adam Rivers, and Jacob Washburn: *Developing A Cross-Kingdom Gene Editing Tool Kit For USDA-ARS Scientists*
- **Pat Clark**: *AI Modeling of Grazingland Animal Behavior*
- **Octavia Crompton** and Kyle Knipper: *Networking local measurements to management scales using machine learning*

- **Milton Drott**, Mitch Elmore, and Hye-Seon Kim: *Machine learning reveals RNAi molecular targets from transcriptomic architectures of mycotoxigenic fungi.*
- **Jay Evans** and Judy Chen: *Functional genomics and microbes for the Beenome100 pollinator project*
- **Clair Friedrichsen**, Katie Pisarello, and Dave Archer: *Food Security Agency in Alaska*
- **Alexander J . Hernandez**: *Spatiotemporal effects of restoration on soil moisture dynamics in semiarid ecosystems*
- **David Kang**: *Using AI to address large, complex datasets in microbiome based IPM*
- **Danielle G. Lemay**: *Using A.I. and Metatranscriptomics to Improve Predictions of Enzyme Function in the Gut Microbiome*
- **Brian Lovette** and Kathryn Bushley: *Identification of fungal proteins and chemicals targeting insect pests*
- **Steven B. Mirsky** and Chris Reberg-Horton: *Transforming ARS scientists use of synthetic image creation pipelines for computer vision and AI*
- **Adelumola Oladeine** and Zaid Abdo: *A supervised machine learning tool to predict parameters that are correlated with a reduction in antimicrobial resistance and probiotic administration*
- **Brenda Oppert**, Gerard Lazo, and Chris Mattison: *Engineering agricultural products using an AI/ML approach to study insect enzyme and substrate interactions*
- **Lindsey Perkin**: *Machine learning to distinguish pest from non-pest weevils*
- **Erin Scully**: *Machine Learning Approaches to Improve Functional Predictions of Chemosensory Genes in Stored Product Insects*
- **Edward Spinard**: *ASFV-Swine Reactome: Adapting Deep Learning to Predict Novel Protein Interactions*
- **Sheina Sim** and Scott Geib: *Implications of taxon-specific training sets on the accuracy of Google DeepVariant variant calling in non-model (non-human) systems for improved marker-assisted breeding and trait mapping.*
- **Kerri Steenwerth**, Amisha Poret-Peterson, Cristina Lazcano, and Jorge Rodgrigues: *Metagenomic analysis of winegrape soils to develop a unique approach to evaluate soil health outcomes*
- **Huihui Zhang** and David Barnard: *Improve Aerial Image-based Forest Fire Detection Using Deep Learning*
- **Xuesong Zhang**, Glenn E Moglen, and Kaiguang Zhao: *Robust explainable AI methods for understanding climate impacts on crop yields*

Graduate Student Internships Update

Thanks to the enthusiasm and interest of ARS mentors and student participants, we are busy onboarding 25 new AI-COE/SCINet graduate student interns for summer internships with ARS labs. These interns will work on a wide variety of research projects, all of which include significant artificial intelligence/machine learning or data science components. Two interns are already working with ARS in spring internships. Many thanks to the ARS scientists who volunteered to serve as internship mentors, and we also thank the universities who have partnered with us for this pilot internships program. We are planning an internships research symposium in the fall – we expect to announce details this summer!

Group Licensing for ASReml

Breeding Insight OnRamp Director Amanda Hulse-Kemp, SCINet/AI-COE fellow Keo Corak, and USDA Network Administrator Rob Butler are exploring group licensing options for the popular statistics software ASReml and/or ASReml-R.

You can fill out this survey to indicate your interest in joining an ASReml group purchase: https://forms.office.com/g/BiFSuzm9zW.

For questions about this survey, please contact Amanda Hulse-Kemp (amanda.hulse-kemp@usda.gov) or Keo Corak (keo.corak@usda.gov).

AI Training Images Workshop

The ARS AI Center of Excellence hosted a virtual workshop to help lay the foundations for building shared AI training image resources for ARS research. The workshop was held on February 28, 2023. Recordings from the Introductory Presentation, Lightning Talks, and AI Barriers and Needs Discussion session can be viewed here.

Invasion Genomics Symposium at Entomology Society of America Annual Meeting

SCINet Fellow Rebecca Clement is organizing a symposium on "Invasion Genomics" at the Entomology Society of America meeting this November. If you are using some sort of genomics (or other "omics") tool to study biological invasions or an invasive group, or know someone who might be interested in giving a 15-minute talk at this session, please contact Dr. Clement (rebecca.clement@usda.gov). Early career scientists and scientists from diverse gender/ethnic backgrounds are especially encouraged to inquire.

Help test upgraded Ceres nodes

During the week of June 19, all Ceres nodes will be upgraded from their current operating system, Enterprise Linux 7 (EL7), to Enterprise Linux 9 (EL9). We are working hard to make this transition as seamless as possible, but because this is a major upgrade with the possibility of impacting user workflows, we are asking for your help!

The VRSC has created an EL9 “sandbox” (small number of test nodes) running EL9 and the latest versions of SCINet’s software packages. We ask that you log in to the sandbox and test the software that you normally use prior to the full system upgrade. If you encounter any problems using your usual software and workflows, please let us know at scinet_vrsc@usda.gov. (Note: If you have installed software on your own, you may need to rebuild it under the new OS as well.)

To log in to the EL9 sandbox from the command line:

ssh login.dev.scinet.usda.gov

From there you can use the same commands you’d normally use on Ceres. E.g.:

module avail – to see available modules

module load <module_name> – to load a module

salloc – to submit an interactive job

sbatch <batch_script> – to submit a batch job

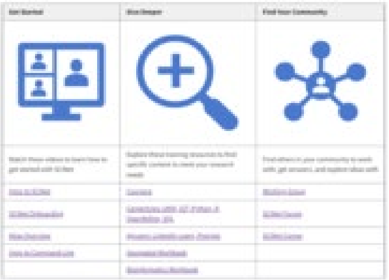
IMPORTANT: Although /project and /90daydata are available from the sandbox, the filesystem performance will be slow during this testing phase. Normal performance will be restored when the entire cluster is upgraded to EL9.

The EL9 sandbox is also available via Ceres Open OnDemand (OOD) (https://ceres-ood.scinet.usda.gov) both for shell access and for interactive apps (select the “ceres-dev” cluster when launching the job in OOD).

Please submit any questions you may have via email to scinet_vrsc@usda.gov.

TRAINING

Training Opportunities



Getting Started: With the expansive list of free training available online, finding the right training to meet your learning needs can be daunting. Take the first steps in getting started with the SCINet Introductory Learning Pathway. Learn about SCINet, how to sign up for an account, and what is possible when supported by SCINet infrastructure. Then dive in with hands-on tutorials available across multiple searchable platforms to find the information you need for just in time learning.

The Carpentries Workshops: We have one more self-hosted workshop on the calendar. We also are offering Data Carpentry's Genomics Workshop this summer.

Workshop	Dates	Time	Registration
Software Carpentry with R	May 9, 16	11am-6pm EDT	signup form
Data Carpentry (Genomics)	Aug 1, 2	12-5pm EDT	signup form

Protein Function and Phenotype Prediction Working Group: The Protein Function and Phenotype Prediction Working Group is organizing a workshop for June 27 and 29th from 1-5 pm EDT. The tentative agenda is:

June 27th: 1 pm-5 pm

1. How to log in to SCINet and transfer the files via Globus
2. How to run a Slurm script with GPU nodes for both “Alphafold-Ceres” and “Alphafold-Atlas”
3. How to visualize AlphaFold output (e.g., a pdb file) by using a public website (Cn3D) Windows Install (nih.gov) and VAST Search (nih.gov)

June 29th: 1 pm-5 pm

1. How to set up and use the Alphafold-CPU version in Ceres

2. How to create a Foldseek database and use the Foldseek search server with the proteome level
3. How to use ESM fold by using the API

Scientists can sign up here to register for the workshop.

Geospatial Research Working Group: The Geospatial Research Working Group has two upcoming training sessions. First, a tutorial on spatial interpolation (e.g., translating point data to gridded data) will be held on May 24 from 3-4pm EDT. The agenda will include an introduction to interpolation techniques as well as code examples. Next, a presentation and discussion on the Geospatial Data Act of 2018 and other geospatial data management topics will be held on June 22 from 1-2pm EDT. To sign up for either session, please use this form.

Courses by Mississippi State University: Mississippi State regularly offers Introduction to Atlas courses. Additionally, there are waiting lists available for several other courses, including an Intensive R course to help scientists with no R experience become familiar with the programming language and start performing statistical analyses in 4 days. Sign up to get notified when these courses are offered.

Coursera.org Courses: The SCINet Office and the AI-COE are excited to provide training opportunities through Coursera. Coursera licenses are available to ARS scientists and support staff for training focused on scientific computing, data science, artificial intelligence, and related topics. Successful completion of courses and specializations result in widely recognized certificates and credentials. Please visit the SCINet Coursera Training Page to request a license. Licenses will be assigned on a rolling basis and are active for three months. Users may be able to extend their licenses upon request.

Training opportunities are continuously being updated on the SCINet Upcoming Training webpage. For more information on any of the above trainings, registration questions or suggestions, please email SCINet-training@usda.gov.

SUPPORT

Getting Started with SCINet is as Easy as 1,2,3

In October 2022, we reached the milestone of having 2,000 registered SCINet users. If you do not already have a SCINet account, we hope you will consider joining the 2,000+ researchers who do. Follow the steps below to get your SCINet account.

1. Request a SCINet account to get started.

2. Read the SCINet FAQs covering general info, accounts/login, software, storage, data transfer, support/policy/O&M, parallel computing, and technical issues.

3. Register for a SCINet Forum account to connect to other users, ask questions, and learn how SCINet can enable your research.

P.S. Don't forget to complete your annual security training! This is required to maintain your account.

For technical assistance with your SCINet account, please email scinet_vrsc@usda.gov.

Support email addresses

All requests for help with user accounts, login problems, resource requests, or support for the Ceres HPC cluster should be sent to the SCINet Virtual Research Support Core (VRSC) at scinet_vrsc@usda.gov. Help requests specific to the Atlas HPC cluster should be sent to help-usda@hpc.msstate.edu.

Many emails are currently being sent to other SCINet email boxes. For the most expedient response to your support requests, be sure to send them to scinet_vrsc@usda.gov or to help-usda@hpc.msstate.edu for Atlas-specific requests.

SCINet User Tip

Each of SCINet's HPC clusters, Ceres and Atlas, provide two main locations for data and other file storage: /project and /90daydata. If you are not sure about which to use, consider these key points:

- Directories in /project have a fixed quota, which means that there is a limit to how much data they can contain. Files in these directories are retained indefinitely (that is, old files are not automatically removed over time).
- Directories in /90daydata do not have quotas. However, files in /90daydata that have not been used for more than 90 days are automatically removed to reclaim storage space.

/90daydata is ideal for storing files that might be very large but are only needed temporarily. This might include, for example, intermediate outputs from analysis workflows or staging large input datasets from external sources. /project is ideal for storing files that are *not* temporary, such as analysis code.

As a final note, neither /90daydata nor /project are backed up, so *neither* location should be considered long-term, permanent storage. SCINet's Juno storage system is the preferred location for permanent storage.

Do you have tips to share? Email them to SCINet-Newsletter@usda.gov to be included in future newsletters.

SCINet Corner: First Thursdays Each Month

SCINet Corner is a VRSC-moderated virtual space for people to share knowledge, discuss best practices, learn about new opportunities, and explore resources to support progress on their projects.

The next SCINet Corner will be held Thursday, June 8, 2023 at 1 pm EST. You can register for this and future SCINet Corners here.

Have a question that just can't wait? Want to see what other users are doing? Reach out to the ever-expanding SCINet Forum community for ideas, support, or just someone to bounce ideas off of at https://forum.scinet.usda.gov/.

CONNECT

The SCINet Team

Every newsletter highlights SCINet community members as a way to connect the ARS scientific computing community. To see all the SCINet community and review past newsletters, visit the Newsletter Archive.

Contribute

Do you use SCINet for your research? We would love to share your story! Email SCINet-Newsletter@usda.gov to contribute content, ask questions, or provide feedback on the SCINet newsletter or website.

SCINet Leadership Team

Brian Stucky, Acting Chief Science Information Officer
Rob Butler, SCINet Program Manager
Jeremy Edwards, Science Advisory Committee (SAC) Chair
Steve Kappes, Associate Administrator

Note: This newsletter is edited to comply with ARS editorial standards.

SCINet Website

Stay Connected with the USDA Agricultural Research Service
5601 Sunnyside Avenue, Beltsville, MD 20705