https://doi.org/10.1038/s41592-025-02981-z

Scikit-bio: a fundamental Python library for biological omic data analysis

Modern biological studies are characterized by the involvement of various 'omic' data types that describe the totality of biological entities, such as genomics, transcriptomics, proteomics, metabolomics and metagenomics. They offer unprecedented insights into complex biological systems but also pose persistent analytical challenges, including high dimensionality (many more features than samples), sparsity (most features are zero) and compositionality (features are interdependent within a sample)[1,2]. Despite challenges, omic data also present a unique opportunity: the biological features are interconnected by knowledge-based, often tree-structured graphs, such as phylogenetic trees and functional classifications, that can be exploited to enhance analysis. These characteristics render generic data analysis methods inadequate for omic data.

To meet this need, we introduced scikit-bio, a Python library for bioinformatics that is oriented toward omic data analysis (Fig. 1). With more than 500 public-facing functions, classes and methods, scikit-bio provides a comprehensive suite of data structures and algorithms designed to address the fundamental analytical challenges and opportunities inherent to omics. Although scikit-bio supports basic sequence analysis, its true strength lies in the analysis of sample-by-feature tables

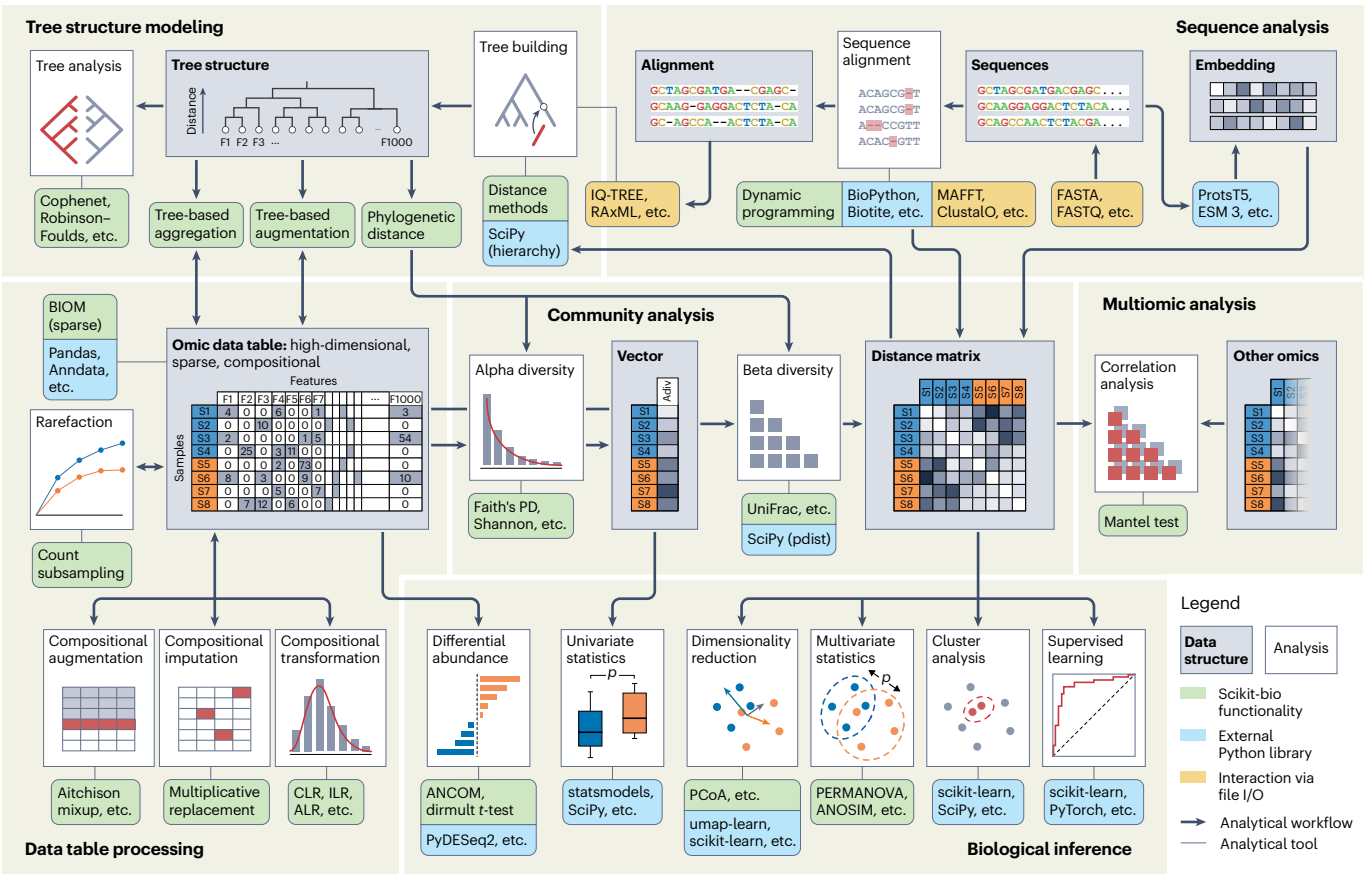

Fig. 1 | Scikit-bio's capacities for biological omic data analysis and its interactions with other tools in the Python scientific computing ecosystem. Squared-off boxes with illustrations represent data structures (light grey background) or analyses (white background). Round-cornered boxes represent specific analytical tools, categorized as follows: green, functionality implemented in scikit-bio; blue, functionality offered by external Python libraries that can be used in conjunction with scikit-bio within the Python framework—for example, a distance matrix generated by scikit-bio can be input into SciPy for hierarchical clustering, scikit-learn for *k*–nearest neighbors classification or umap-learn for UMAP embedding; yellow, functionality provided by non-Python programs that can interact with scikit-bio through file input and output (I/O). For example, scikit-bio can read a phylogenetic tree built by RAxML or a multiple sequence alignment generated by MAFFT.

and associated tree structures, the two data representations common to many omics studies. This unique focus distinguishes scikit-bio from general-purpose toolkits such as BioPython[3] and domain-specific packages such as ScanPy[4]. Scikit-bio also serves as a bridge between omics-based methods, many originally developed in R, and advanced machine learning frameworks commonly implemented in Python (Supplementary Table 1).

Scikit-bio addresses the sparsity challenge of omic data by integrating the Biological Observation Matrix (BIOM) format[5], a sparse matrix structure for tabular data. Developed and maintained by members of our team, BIOM has served as a standard in microbiome research. To address data compositionality, scikit-bio provides multiple transformation methods, such as center log-ratio (CLR) and a multiplicative replacement method for zero handling. It also implements compositionality-aware differential abundance tests, such as analysis of composition of microbiomes (ANCOM). Meanwhile, scikit-bio houses a variety of metrics to quantify biological diversity within and between samples. The latter category generates a distance matrix, which can then feed into multidimensional scaling techniques such as principal coordinates analysis (PCoA) for non-Euclidean data embedding, or multivariate statistical tests such as permutational multivariate analysis of variance (PERMANOVA) to assess omic-trait correlations. Moreover, scikit-bio implements the Mantel test for evaluating cross-omic correlations. The library also features functionalities for working with tree structures that model the biological relevance among features. Users can construct, traverse and manipulate trees, calculate the distances between features or groups, and integrate trees into community modeling (such as Faith's phylogenetic diversity (PD) and UniFrac metrics), thereby incorporating biological knowledge into the analysis of high-dimensional data that would otherwise be analytically intractable.

Many implementations are state of the art or unique in the Python ecosystem (Supplementary Table 1). They have been optimized for the efficient analysis of immense datasets in contemporary and emerging research (such as PCoA analyses of >100,000 samples[6]).

Scikit-bio integrates seamlessly with the Python scientific computing ecosystem. Its structured API adheres to modern styling and design principles. Many functionalities utilize widely adopted, efficient data structures, such as NumPy arrays, SciPy sparse matrices and Pandas dataframes, to enable users to apply scikit-bio functions to existing data without costly conversions. Outputs from scikit-bio analyses can be directly utilized by external libraries, such as statsmodels for statistical analysis, scikit-learn for machine learning and matplotlib for data visualization. Scikit-bio also features a flexible and extensible input/output system, making it interoperable with non-Python bioinformatics workflows.

Since its initial release in 2014, scikit-bio has supported numerous studies and tool developments. Notably, it has supported multiple essential functionalities of QIIME 1/2[7,8], a widely adopted microbiome data analysis platform. Though scikit-bio originated in microbiome research, with multiple core abstractions driven by that community, its design leverages common properties of omic data. As a stand-alone library, it has a modular architecture that lets users select and combine only the components they need, plugging them into domain-specific workflows to suit various omics analyses. Its adoption has extended into various fields beyond microbiome research, such as single-cell data analysis, structural variation analysis, metabolomics and epigenomics. The collective works leveraging scikit-bio have been cited tens of thousands of times (Supplementary Tables 2 and 3).

Our team adheres to industry-standard practices to ensure a high-quality software project, maintaining 98% unit-test coverage, with comprehensive doctests, and objective style enforcement. The online documentation details the mathematical background, biological applications and example code of functionalities, plus structured tutorials navigating through modules and workflows. Additionally, scikit-bio has served as the basis for a bioinformatics textbook[9]. These practices ease adoption and facilitate interdisciplinary collaboration.

The scikit-bio project thrives on a community-driven model, with contributions from more than 80 developers to date. This approach ensures scikit-bio's continuous evolution alongside advances in scientific computing. Our team is committed to improving and expanding scikit-bio's functionality to support up-to-date and broader research applications. We believe that this essential software library will continue to benefit the research community into the future.

Code availability

The scikit-bio project's official website is https://scikit.bio. The source code of scikit-bio is licensed under a BSD-3 licensed and hosted at the public GitHub repository https://github.com/scikit-bio/scikit-bio. Version 0.7.0 of the source code has been permanently archived at the Zenodo repository at https://zenodo.org/records/15988672. The scikit-bio software package is distributed free of charge via PyPI (https://pypi.org/project/scikit-bio/) and conda-forge (https://anaconda.org/conda-forge/scikit-bio). Comprehensive documentation of scikit-bio, including guidelines, API references and example usages, are available at the official website. A structured list of scikit-bio tutorials is available at https://github.com/scikit-bio/scikit-bio-tutorials.

Matthew Aton[1], Daniel McDonald[2], Jorge Cañardo Alastuey[3,15], Raeed Azom[1], Paarth Batra[1], Valentyn Bezshapkin[4], Evan Bolyen[5], Alexander Cagle[2], J. Gregory Caporaso[5], Justine W. Debelius[2,16], Kestrel Gorlick[5], Nirmitha Hamsanipally[1], Lars Hunger[6], Aryan Keluskar[1], Disen Liao[7], Yang Young Lu[7,8], Jose A. Navas-Molina[2,9,17], Anders Pitman[5,18], Jai Ram Rideout[5,19], Anton Sazonov[1], Bharath Sathappan[2], Karen Schwarzberg Lipson[5,20], Igor Sfiligoi[10], Chris Tapo[1,21], Yoshiki Vázquez-Baeza[11,22], Zijun Wu[1], Zhenjiang Zech Xu[2,23], Mingsong Sam Ye[12], Jianshu Zhao[2], Rob Knight[2,9,11,13] ✉, James T. Morton[6] ✉ & Qiyun Zhu[1,14] ✉

[1]Biodesign Center for Fundamental and Applied Microbiomics, Arizona State University, Tempe, AZ, USA. [2]Department of Pediatrics, University of California San Diego School of Medicine, La Jolla, CA, USA. [3]College of Engineering and Applied Science, University of Colorado Boulder, Boulder, CO, USA. [4]Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zurich, Zurich, Switzerland. [5]Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA. [6]Gutz Analytics, LLC, Boulder, CO, USA. [7]Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada. [8]Department of Biomedical Engineering and Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, WI, USA. [9]Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, USA. [10]San Diego Supercomputer Center, University of California San Diego, La Jolla, CA, USA. [11]Center for Microbiome Innovation, University

of California San Diego, La Jolla, CA, USA. [12]School of Business, Stevens Institute of Technology, Hoboken, NJ, USA. [13]Department of Bioengineering, University of California San Diego, La Jolla, California, USA. [14]School of Life Sciences, Arizona State University, Tempe, AZ, USA. [15]Present address: Amazon Robotics, North Reading, MA, USA. [16]Present address: Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. [17]Present address: Google LLC, Mountain View, CA, USA. [18]Present address: University of Utah, Salt Lake City, UT, USA. [19]Present address: One Codex, Inc., Wilmington, DE, USA. [20]Present address: Chasing Value Asset Management, Inc., Santa Monica, CA, USA. [21]Present address: Department of Mathematics, College of Engineering and Polymer Science, University of Akron, Akron, OH, USA. [22]Present address: BiomeSense Inc., Chicago, IL, USA. [23]Present address: State Key Laboratory of Food Science and Resources, Nanchang University, Nanchang, China. ✉e-mail: rknight@ucsd.edu; jamie@gutzanalytics.com; qiyun.zhu@asu.edu

Published online: 11 December 2025

References

1. Quinn, T. P. et al. *Gigascience* **8**, giz107 (2019).
2. Rahnenführer, J. et al. *BMC Med.* **21**, 182 (2023).
3. Cock, P. J. A. et al. *Bioinformatics* **25**, 1422–1423 (2009).
4. Wolf, F. A., Angerer, P. & Theis, F. J. *Genome Biol.* **19**, 15 (2018).
5. McDonald, D. et al. *Gigascience* **1**, 7 (2012).
6. McDonald, D. et al. *Nat. Methods* **15**, 847–848 (2018).
7. Bolyen, E. et al. *Nat. Biotechnol.* **37**, 852–857 (2019).
8. Caporaso, J. G. et al. *Nat. Methods* **7**, 335–336 (2010).
9. Bolyen, E. et al. *J. Open Source Educ.* **1**, 27 (2018).

Acknowledgements

We thank a large number of community contributors for their valuable inputs to the scikit-bio project. The development of scikit-bio is currently funded by the US Department of Energy, Office of Science, under award DE-SC0024320.

Author contributions

R.K. and J.G.C. initiated the scikit-bio project in 2013; J.G.C. led the project from 2013 to 2023; and Q.Z. has led the project since 2023. M.A., D.M., J.T.M. and Q.Z. comprise the current core development team of scikit-bio. M.A. is responsible for the maintenance of scikit-bio. R.K. provides advisory support. J.G.C. provides consulting. All authors have made significant contributions to the development of scikit-bio, including but not limited to code, documentation, test data and educational materials. Q.Z. and M.A. led the writing of the manuscript, with input from other authors. All authors reviewed and approved the final manuscript.

Competing interests

R.K. is a scientific advisory board member and consultant for BiomeSense, Inc., and has equity in and receives income from the company. He is a scientific advisory board member for and has equity in GenCirq. He is a consultant for and receives income from DayTwo. He has equity in and acts as a consultant for Cybele. He is a co-founder of Biota, Inc., and has equity in the company. He is a cofounder of Micronoma, and has equity in and is a scientific advisory board member for the company. The terms of these arrangements have been reviewed and approved by the University of California, San Diego, in accordance with its conflict of interest policies. D.M. is a consultant for BiomeSense, Inc., and has equity in and receives income from the company. The terms of these arrangements have been reviewed and approved by the University of California, San Diego, in accordance with its conflict of interest policies. Y.V.-B. is a full-time employee for BiomeSense, Inc., and has equity in and receives income from the company. J.R.R. is a full-time employee for One Codex, Inc., and has equity in and receives income from the company. J.G.C. and E.B. have equity in and receive income from Cymis Benefit Corp. J.C.A.'s contribution to this work was completed before he joined Amazon Robotics. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-025-02981-z.

Peer review information *Nature Methods* thanks Zewen Kelvin Tuong, who co-reviewed with Amos Choo, and Ralf Gommers and Tyler Reddy for their contributions to the peer review of this work.