Clusters or Continuum? A Validated Protocol for Distinguishing Discrete Structure From Gradients in Multi-Omics Data

Matevž Ogrinc
matevz.ogrinc@ijs.si
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Vida Groznik
vida.groznik@fri.uni-lj.si
University of Ljubljana, Faculty of Computer and Information Science
Ljubljana, Slovenia

Barbara Koroušić Seljak
barbara.korousic@ijs.si
Jožef Stefan Institute
University of Ljubljana, Medical Faculty
Ljubljana, Slovenia

Biljana Mileva Boshkoska
biljana.mileva@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

A common question in medical microbiome and multi-omics research is whether participants form distinct subgroups or vary along a continuum. A popular approach is to apply clustering to a UMAP embedding, but this can produce apparent groups that do not reflect discrete structure in the original data. In this work, we evaluate a more robust method in which HDBSCAN is applied directly in the original feature space across a broad range of hyperparameter values, with every candidate cluster assessed against multiple robustness criteria, including stability across parameter settings and reproducibility under a second metric or in a second data layer. To establish that the method is trustworthy, we first apply it to synthetic data with known structure. It recovers discrete groups when they are present (adjusted Rand index 0.92 to 0.97) but identifies no stable clusters in data generated along a continuous gradient, whereas clustering of UMAP embeddings retains spurious groups on that same gradient in 80–96% of configurations. We then apply the approach to a dietary-intervention cohort ($n \approx 97$) across five data layers. Four layers: microbiome species, wearables, biomarkers, and biomarker change turn out to be continuous, and their apparent clusters are artifacts. The fifth, bacterial functional pathways, contains a single validated group of six participants, whose functional signature (branched-chain amino-acid and rhamnose biosynthesis) and taxonomy both indicate domination by a single gut bacterium, *Prevotella copri* (51–79%). Previous studies have associated *P. copri* and its branched-chain amino-acid production with insulin resistance, but we do not observe this association. Instead, the relevant metabolic phenotype is associated with overall microbial diversity. The key finding of our study is methodological, where apparent clusters in low-dimensional embeddings should not be interpreted as biological subgroups unless they are reproducible in the original data space under a second metric or in a second readout of the same sample.

Keywords

Multi-omics, Microbiome, HDBSCAN, UMAP

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.skui.93

1 Introduction

Understanding whether a group of participants separates into distinct types or varies along a continuous gradient is a common question when analysing microbiome and other multi-omics data. If real subgroups exist, they can be used to define subtypes, guide stratification and motivate targeted follow-up studies. If the variation is continuous, dividing people into types imposes a structure that is not there. In the gut microbiome this question has been asked frequently and the community was originally described in terms of discrete enterotypes [1], yet later work argued that these apparent types are better understood as gradients [9] and the current view treats them as labels for regions of a continuous landscape rather than discrete categories [6]. Whether enterotypes are detected at all depends on the clustering method and the distance metric [10]. Deciding between the two is more difficult than it appears, because a clustering algorithm returns clusters whether or not any exist. The difficulty lies not in running the algorithm but in judging whether its output reflects real structure in the data. A common way is to project the data into two dimensions with UMAP [13], inspect the resulting plot and treat any visually separated regions as groups.

UMAP is a useful method, yet it can fragment a continuous gradient into groups and can impose structure on data that has none, with the outcome depending on parameter choices [5]. Clustering an embedding therefore risks reporting subgroups that are artifacts of the projection rather than features of the biology. In this work we develop a protocol for separating discrete structure from continuous variation, validate it on synthetic data with known ground truth, and apply it to five layers of a weight loss intervention cohort. Unlike Bulygin et al. [3], who reach the continuum conclusion by clustering non-linear embeddings, we cluster in the original distance space, extend the question beyond the microbiome, and report a positive case alongside the negative ones.

The remainder of the paper is organized as follows. Section 2 describes the cohort, the five data layers and the clustering protocol. Section 3 reports the validation on synthetic data and the analysis of the five layers. Section 4 interprets the findings and sets out the limits of the approach.

2 Methods

2.1 Data

For the application to a real cohort we utilized AI4FoodDB [17, 12], a publicly available multi-omics resource from a one-month

Table 1: The five data layers. Features are counted before and after the 10% prevalence filter, which applies to the compositional layers only.

Layer	n	Features	Representation
Microbiome species	96	726 → 246	CLR / rel. abundance
Wearables	88	43	standardised / min-max
Biomarkers	97	39	standardised / min-max
Biomarker change	90	39	standardised / min-max
Functional pathways	96	498 → 403	CLR / rel. abundance

dietary weight-loss intervention in 100 overweight and obese adults (mean age 50 years, mean BMI 30.7, 69% female and 31% male). The two groups defined in the original study differed only in the order of manual and digital data collection. Biological samples were collected at baseline (visit V1), and after the intervention (visit V3). The present work is a secondary analysis of the released data. Taxonomic and functional profiling of the stool metagenomes was carried out using the MetaBakery pipeline [15]. From this resource we assembled five baseline data layers, each reduced to a single feature vector per participant (Table 1). Two are derived from the same V1 stool metagenomes: relative species abundances, and community-level pathway abundances after removal of unmapped and per-species stratified entries. The others are physiological features from Fitbit Sense recordings of heart rate, sleep and activity; blood measures at V1, including insulin, HOMA-IR, a lipid panel, adiponectin and inflammatory markers; and the change in each blood measure between V1 and V3. Values reported below the detection limit were set to that limit.

2.2 Clustering and Evaluation

We cluster with HDBSCAN [4], a density-based method that does not require the number of clusters in advance, in four steps.

(1) Representation. Each metric is paired with a representation it is valid on. Microbiome profiles are compositional, so Euclidean-style distances on raw proportions are not interpretable [7]: the compositional layers use a centred log-ratio transform, with zeros replaced multiplicatively ($\delta = 10^{-6}$), for the Euclidean, Manhattan and cosine metrics, and relative abundance for Bray-Curtis, which is defined only on non-negative data. The real valued layers are median imputed, then standardised for the first three metrics and scaled to $[0, 1]$ for Bray-Curtis.

(2) Grid search. We evaluate all combinations of minimum cluster size {3, 5, 10, 15}, minimum samples {1, 2, 5}, selection method {excess of mass, leaf} and metric {Euclidean, Manhattan, cosine, Bray-Curtis}, giving 96 configurations per layer, each run on the full pairwise distance matrix with a fixed random seed and, for the compositional layers, a filter keeping features present in at least 10% of participants.

(3) Screening. A configuration is retained only if it yields at least two clusters, leaves fewer than half of the participants as noise, and reaches a silhouette score [18] of at least 0.25, computed on the clustered participants under the same metric. Both thresholds are deliberately permissive: 0.25 is the boundary below which no substantial cluster is considered found, and a solution leaving most of the cohort unassigned does not describe how it divides.

(4) Validation. A retained result must have a positive density based validation score (DBCV [14]), and its candidate cluster must hold at least 5% of the cohort, keep its members together at every adjacent grid point that returns a clustering, a parameter-space analogue of cluster stability [8], and, since the structure detected depends on the metric [10], group those members again under a second metric or in a second layer from the same samples. A candidate supported by one metric in one layer is not interpreted. Cluster persistence was also examined but did not separate retained from rejected candidates.

Comparator. The common alternative, clustering a UMAP embedding [13], is run for comparison. UMAP accepts a non-Euclidean metric for its neighbour graph but returns points in a Euclidean space, so clustering the embedding is always Euclidean and the metric no longer distinguishes configurations, leaving 24 per layer. We embed into eight dimensions rather than two, so the comparator is not limited by what a plot can show.

Synthetic validation. We validate the procedure on synthetic data matched to the dimensionality and sparsity of the real layers: for both the compositional and the real valued setting, one dataset with three planted clusters and one varying along a single gradient, over ten random seeds. We record the fraction of configurations retained and, on the clustered data, the adjusted Rand index (ARI) of each against the planted partition. The comparator is run on the same data.

For a subgroup that passes validation we compare it with the rest of the cohort on seven blood markers using the Mann-Whitney U test, correcting across the panel with the Benjamini-Hochberg procedure [2]. Because six participants give a low-powered test, we additionally relate the relevant microbiome quantities to the biomarkers across the whole cohort by partial Spearman correlation, adjusting for age and BMI.

3 Results

3.1 Finding Clusters and Rejecting Gradients

We repeated the entire procedure on the synthetic data ten times, generating the datasets with a different random seed each time. On data containing three planted clusters, the grid search retained 62% of the 96 parameter combinations in the compositional setting and 96% in the real-valued setting, and the retained configurations recovered the planted partition with a mean ARI of 0.92 (SD 0.14) and 0.97 (SD 0.10). On pure-gradient data with no group structure, it retained fewer than 1% in both cases. Running the same procedure on a UMAP embedding of those same gradients retained 96% and 80% of combinations (Figure 1).

On synthetic data, the protocol therefore finds structure when it is present and reports none when it is absent. The data, the algorithm, the parameter grid and the screening criteria are held constant. The only difference is whether distances are computed in the original space or in the embedding. Clustering the embedding did recover the planted clusters, retaining 97% and 100% of settings on that data (ARI 0.83 and 0.78), but could not tell the two cases apart.

3.2 Four of the Five Layers Are Continuous, Not Clustered

On the real cohort data, four of the five layers turn out to be continuous (Table 2). The microbiome species layer retained 14 parameter combinations, but all 14 used Bray-Curtis and none of the remaining three metrics reproduced them; each contains a *Prevotella*-rich cluster of 14 to 16 participants with 20% to 46% of the cohort as noise and a silhouette of at most 0.34. The biomarker layer retained no combinations, and the wearable layer only one weak solution under cosine (silhouette 0.26, 36% noise), counted

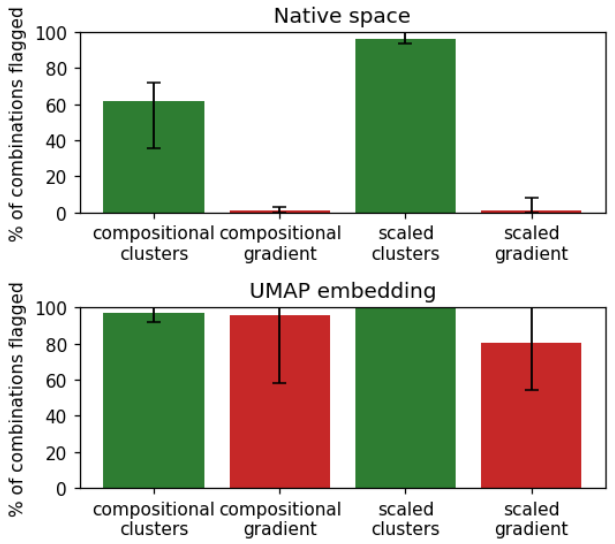


Figure 1: Validation on synthetic data. Bars show the percentage of parameter combinations retained by the screening criteria across ten random seeds for four datasets containing either planted clusters (compositional/scaled clusters) or a pure gradient (compositional/scaled gradient). Clustering in the original distance space (top) flags structure on the clustered datasets but almost never on the gradients, whereas the UMAP based pipeline (bottom) flags structure in both.

under both selection methods. In both layers the highest silhouette scores came from solutions that discarded 61% to 72% of participants as noise, which describes a few outliers rather than a split of the cohort. The biomarker change layer retained two combinations, both under the cosine metric and both with near-zero cluster validity. In all four layers, the highest silhouette scores came from the UMAP embedding rather than the original space, the same pattern observed on the synthetic gradients.

Table 2: Grid search results per layer in the original data space. Retained combinations are given out of 96. Silhouette and DBCV values are the best achieved by any combination.

Layer	Retained	Metrics	Silhouette	DBCV
Microbiome species	14	1	0.343	0.148
Wearables	2	1	0.347	0.125
Biomarkers	0	0	0.282	0.124
Biomarker change	2	1	0.402	0.123
Functional pathways	13	3	0.511	0.404

3.3 The One Validated Subgroup Is an Enterotype Without a Metabolic Phenotype

The functional pathway layer differed. Its 13 retained combinations span three metrics, yet the solutions are not the same. The four Bray-Curtis combinations separate the same six participants from the rest; the clearest, with a minimum cluster size of 5, a minimum samples value of 2 and excess-of-mass selection, left 5% of participants as noise and achieved a silhouette score of 0.51 and a DBCV of 0.40, the highest cluster validity observed in any layer. The six Manhattan combinations instead isolate the same 13 participants at every setting, who are not reproduced under cosine, do not differ from the rest in age, BMI, *P. copri* abundance or diversity (all $p > 0.25$), and are not interpreted further. All six subgroup members are assigned to noise in every Manhattan and cosine solution. The six therefore pass the validation criteria not across metrics but across readouts: they remain together at every adjacent grid point that returns a clustering, and they are the core of the *Prevotella*-rich cluster found under Bray-Curtis in the species layer, a second readout of the same metagenomes.

The six participants share a distinct functional signature. In each of them, dTDP-L-rhamnose biosynthesis is the most abundant pathway, and as a group they are enriched in branched-chain amino acid (BCAA) biosynthesis ($p \approx 10^{-9}$) and depleted in fatty acid and purine pathways. The depletions are informative, because a simple low-diversity artefact would inflate all detected pathways rather than shift them in a particular direction. The taxonomy of the same stool samples showed that all six are dominated by *Prevotella copri* at 51–79% of their community. This separates them clearly from the remainder of the cohort, in which no participant exceeds 46%, and they are the six most *P. copri*-rich samples in the dataset (Figure 2). They also show the usual enterotype contrast, with 83% *Prevotella* and 1.4% *Bacteroides* against 1% and 35% in the remainder, together with lower diversity (Shannon 1.6 against 3.0) and younger age (median 38 against 51 years, $p = 0.024$).

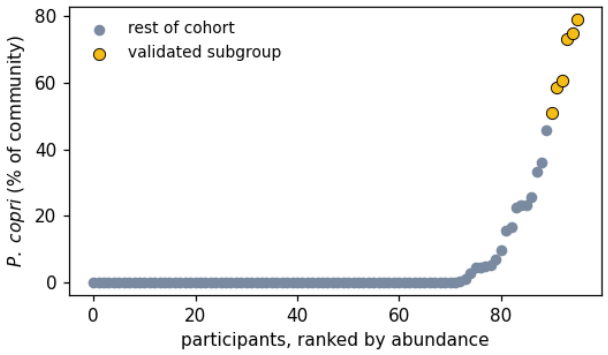


Figure 2: The six participants identified in the functional pathway layer are the *P. copri* dominated samples: relative abundance of *P. copri* across all participants, ranked. No other participant exceeds 46%.

Microbial BCAA biosynthesis is linked to insulin resistance, with *P. copri* identified as a driver [16], so we tested the six against the blood markers. They were nominally higher on HOMA-IR (2.25 against 1.74) and insulin, and lower on adiponectin and HDL. However, none of these differences remained statistically significant after correction across the seven-marker panel (minimum $q = 0.48$), nor after adjustment for age (minimum $q = 0.58$). Additionally neither *P. copri* abundance ($\rho = +0.09$, $p = 0.38$) nor the *Prevotella*-to-*Bacteroides* ratio ($\rho = +0.01$, $p = 0.96$) tracks HOMA-IR across the cohort. On the other hand Shannon diversity correlated with HOMA-IR at $\rho = -0.41$ ($p < 0.001$), and this held after adjustment for age and BMI ($\rho = -0.36$) and was

unchanged when the six were excluded. In this cohort, therefore, the metabolic signal follows overall diversity rather than enterotype.

4 Discussion

The protocol did two things. First, it rejected the apparent structure in four layers by showing that it came from a single metric, from discarding most of the cohort, or from the embedding alone. Because we had already verified that the protocol can detect clusters when they are present, those rejections are informative. Second, it found one subgroup, which is visible under a single metric and qualifies because the same participants group together in a second readout of the same samples.

The clearest methodological difference is between the two spaces. Clustering a UMAP embedding did recover the real clusters, and did so reliably. But it also reported clusters in data that had none, under most parameter settings and across every seed. The problem is therefore not that embeddings miss structure, but that they cannot indicate when none is present. This is important since the reason to cluster an omics cohort is usually to determine whether groups exist at all. When that is the question, distances should be computed in the original space. Embeddings remain useful for visualisation of the data and for recovering groups once their existence has been established. Working in the original space also keeps the choice of metric, which the embedding discards, and that comparison is informative even when it does not confirm. The subgroup appears only under Bray-Curtis, the metric sensitive to dominance of a single feature, and dissolves under the centred log-ratio representation, which removes that dominance. It qualified only because the same six reappeared in the species readout.

The biological result, however, explains why the cluster we identified required validation. The pathway that made the subgroup easy to detect is not the pathway that would make it meaningful. Rhamnose biosynthesis dominates these samples because one organism dominates the community, and Bray-Curtis weights abundant features heavily. BCAA biosynthesis is the interesting one, because it is the activity linking *P. copri* to insulin resistance in prior work. We do not, however, reproduce that link here. It is absent both in the underpowered comparison of the six and in the continuous test. Other work suggests the relationship is context-dependent, as it has been reported in Western-diet settings, whereas *Prevotella*-rich guts are also associated with fibre rich diets and, in some studies, with improved glucose handling [11].

Several limitations should be noted. The subgroup comprises six participants, so the metabolic comparison can detect only very large effects, and the continuous analysis is the more reliable of the two. The cohort itself is small, although the synthetic validation was generated to match the dimensionality and sparsity of the real layers. The simulations test two interpretable extremes, discrete clusters and a continuous gradient, and therefore do not cover intermediate structures such as overlapping clusters, heterogeneous densities, branching manifolds, or mixtures of discrete and continuous variation. The analysis is cross-sectional and dietary data were not available, so whether the enterotype reflects habitual fibre intake remains an open question. The gradient strength is itself a calibration choice, since a sufficiently strong gradient produces sparse extremes that HDBSCAN separates on legitimate grounds. Clustering in the original space also requires a full pairwise distance matrix, which becomes memory-bound above a few thousand samples.

Within these limits, the protocol is simple, reusable and easy to implement. Its main contribution is the principle that a clustering result becomes evidence only once it has been validated across parameters, across metrics or independent readouts, and against a method whose behaviour on known data has been quantified.

Acknowledgements

We thank Blaž Stres and Boštjan Murovec for processing the raw sequencing data and providing the taxonomic and functional profiles used in this work.

The research was partially supported by Slovenian Research and Innovation Agency under the research programmes Artificial intelligence and intelligent systems (grant no. P2-0209) and Computer Systems and Structures (grant No. P2-0098), the Gravity project Artificial Intelligence for Science (grant no. GC-0001), and the research project AI4Food (grant No. J7-70265). This work was also funded by the European Union under Grant Agreement No. 101287404 (HE NutriMind).

References

[1] Manimozhiyan Arumugam et al. 2011. Enterotypes of the human gut microbiome. *Nature*, 473, 7346, 174–180.

[2] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57, 1, 289–300.

[3] Ivan Bulygin, Vladislav Shatov, Anton Rykachevskiy, et al. 2023. Absence of enterotypes in the human gut microbiomes reanalyzed with non-linear dimensionality reduction methods. *PeerJ*, 11, e15838.

[4] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 160–172.

[5] Tara Chari and Lior Pachter. 2023. The specious art of single-cell genomics. *PLOS Computational Biology*, 19, 8, e1011288.

[6] Paul I Costea et al. 2018. Enterotypes in the landscape of gut microbial community composition. *Nature Microbiology*, 3, 1, 8–16.

[7] Gregory B Gloor, Jean M Macklaim, Vera Pawlowsky-Glahn, and Juan J Egozcue. 2017. Microbiome datasets are compositional: and this is not optional. *Frontiers in Microbiology*, 8, 2224.

[8] Christian Hennig. 2007. Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, 52, 1, 258–271.

[9] Dan Knights, Tonya L Ward, Christopher E McKinlay, Hannah Miller, Antonio Gonzalez, Daniel McDonald, and Rob Knight. 2014. Rethinking "enterotypes". *Cell Host & Microbe*, 16, 4, 433–437.

[10] Omry Koren, Dan Knights, Antonio Gonzalez, Levi Waldron, Nicola Segata, Rob Knight, Curtis Huttenhower, and Ruth E Ley. 2013. A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Computational Biology*, 9, 1, e1002863.

[11] Petia Kovatcheva-Datchary et al. 2015. Dietary fiber-induced improvement in glucose metabolism is associated with increased abundance of prevotella. *Cell Metabolism*, 22, 6, 971–982.

[12] Blanca Lacruz-Pleguezuelos et al. 2025. Ai4food, a feasibility study for the implementation of automated devices in the nutritional advice and follow up within a weight loss intervention. *Clinical Nutrition*, 48, 80–89.

[13] Leland McInnes, John Healy, and James Melville. 2018. Umap: uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

[14] Davoud Moulavi, Pablo A Jaskowiak, Ricardo JGB Campello, Arthur Zimek, and Jörg Sander. 2014. Density-based clustering validation. In *Proceedings of the 2014 SIAM international conference on data mining*. SIAM, 839–847.

[15] Boštjan Murovec, Leon Deutsch, Damjan Osredkar, and Blaž Stres. 2024. Metabakery: a singularity implementation of biobakery tools as a skeleton application for efficient hpc deconvolution of microbiome metagenomic sequencing data to machine learning ready information. *Frontiers in Microbiology*, 15, 1426465.

[16] Helle Krogh Pedersen et al. 2016. Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature*, 535, 7612, 376–381.

[17] Sergio Romero-Tapiador et al. 2023. Ai4fooddb: a database for personalized e-health nutrition and lifestyle through wearable devices and artificial intelligence. *Database: The Journal of Biological Databases and Curation*, 2023, baad049.

[18] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.