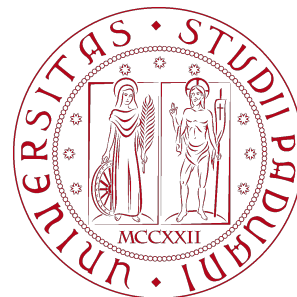
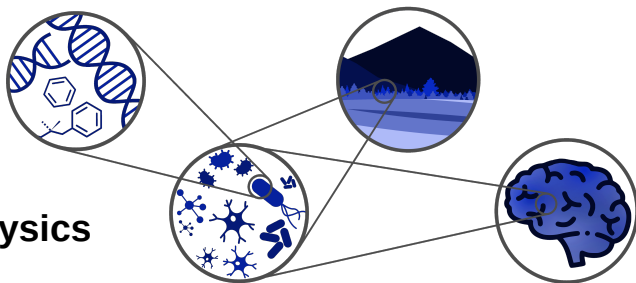


Archetypal Decomposition of Metagenomic Profiles for Cross-Study Diagnostic Classification

LIPh
Laboratory of
Interdisciplinary Physics



**UNIVERSITÀ
DEGLI STUDI
DI PADOVA**

Authors:

Marcello Seppi¹, Jacopo Pasqualini¹, Guido Zampieri², Samir Suweis¹

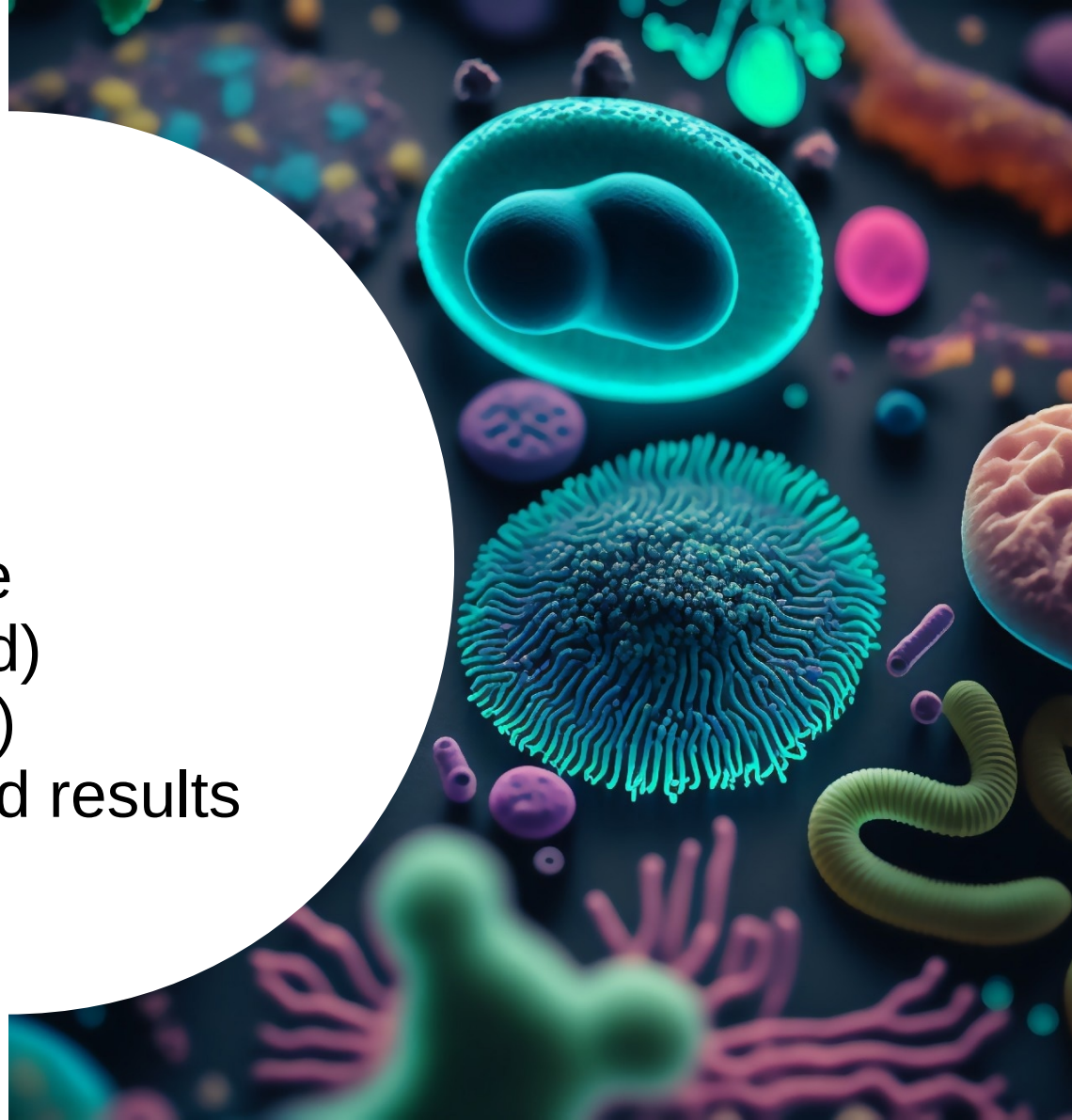
¹Dipartimento di Fisica e Astronomia Galileo Galilei, Università degli Studi di Padova

²Dipartimento di Biologia, Università degli Studi di Padova

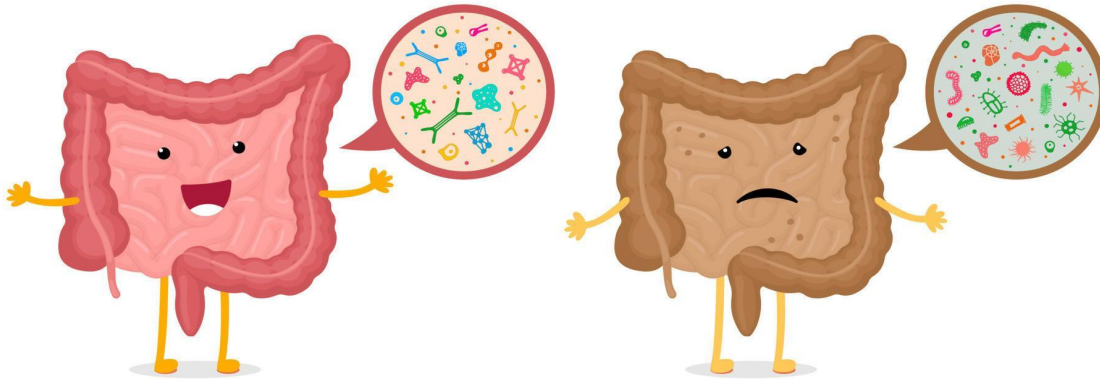
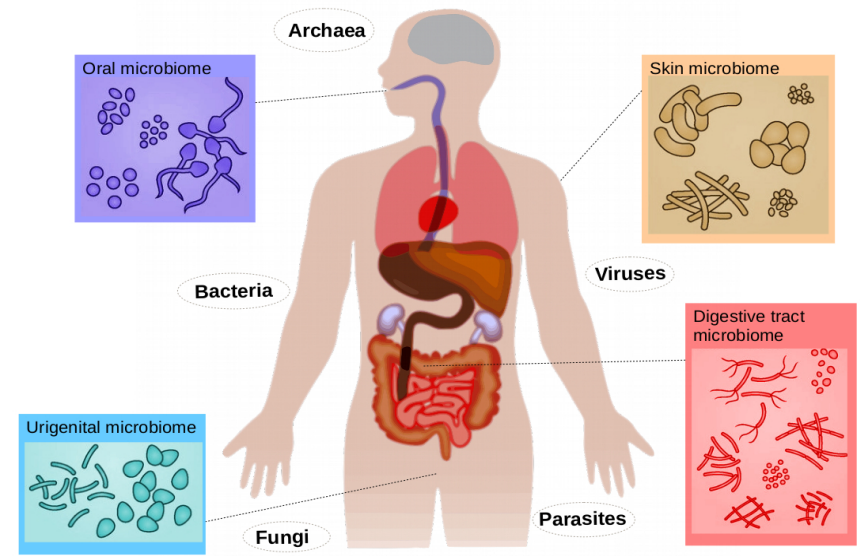
Marcello Seppi

Contents:

- Context
- Classification problem
- Microbiome data
- Challenges of data structure
- Archetypal Analysis (method)
- Archetypal Analysis (results)
- Classification framework and results



Human-body as ecosystem



Connected to health

Predict disease from gut microbiome composition

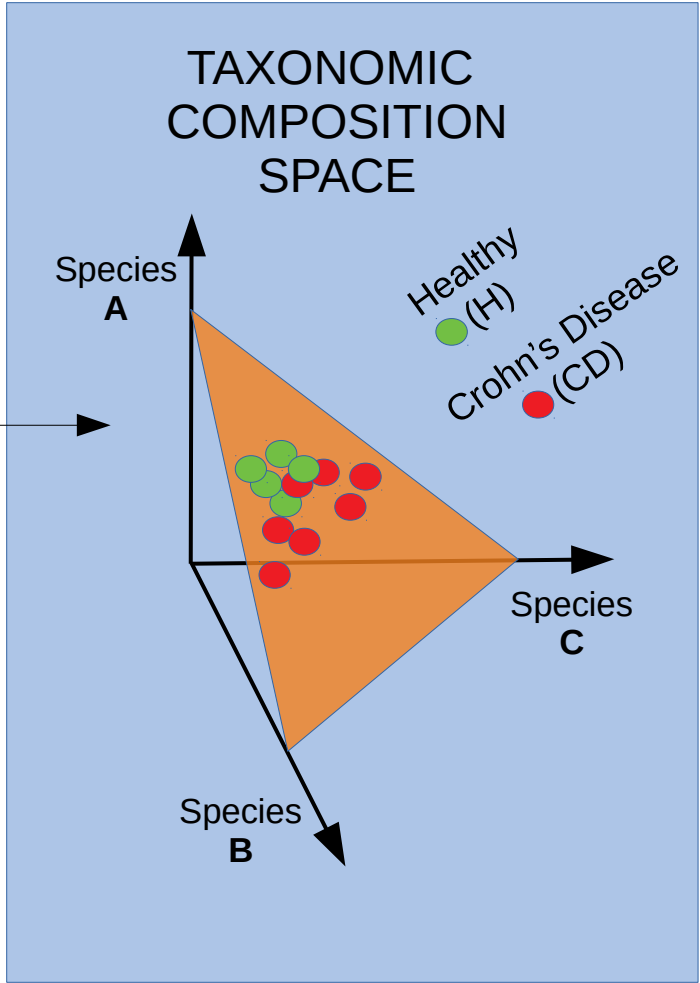
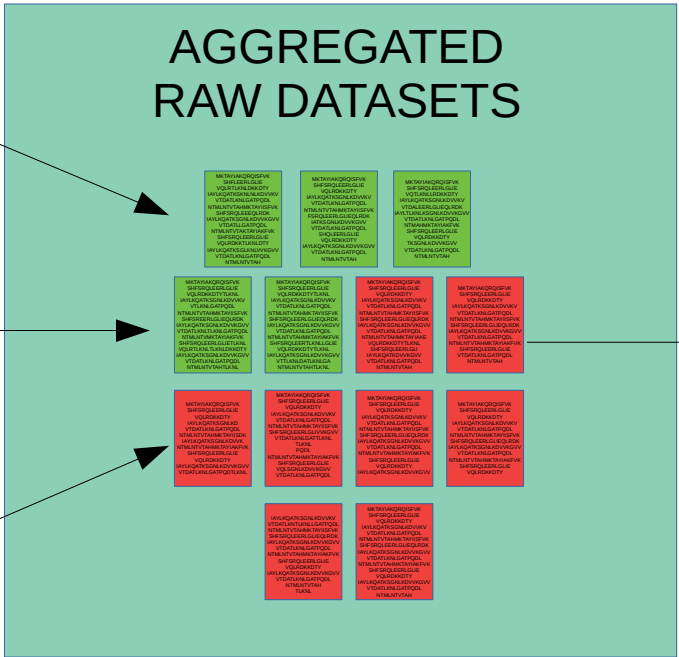
What we are attempting:

Healthy vs Crohn's Disease classification through gut microbiome data

Study 1
[he2019two]

Study 2
[ijaz2017distinct]

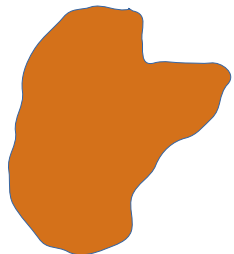
Study 3
[lloid2019multi]



			N_genomes		diagnosis		
			mean	std	total	H	CD
study	country	instrument					
he2019two	China	Illumina HiSeq 2000	105.295082	14.950435	122	55	67
ijaz2017distinct	Scotland	Illumina MiSeq	107.306122	13.153270	49	27	22
lloyd2019multi	USA	Illumina HiSeq 2000	106.315068	9.985930	73	25	48

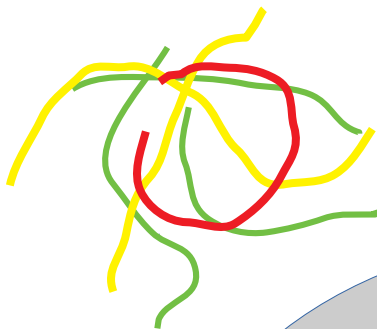
Measuring microbial composition:

BIOLOGICAL
SAMPLE



DNA
extraction

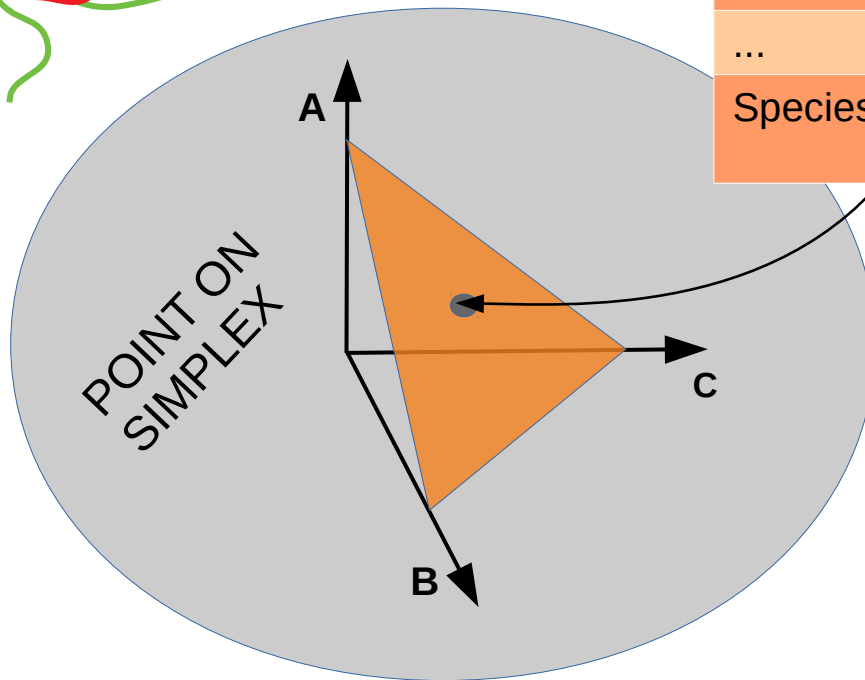
MICROBIAL DNA



Sequencing &
Classification

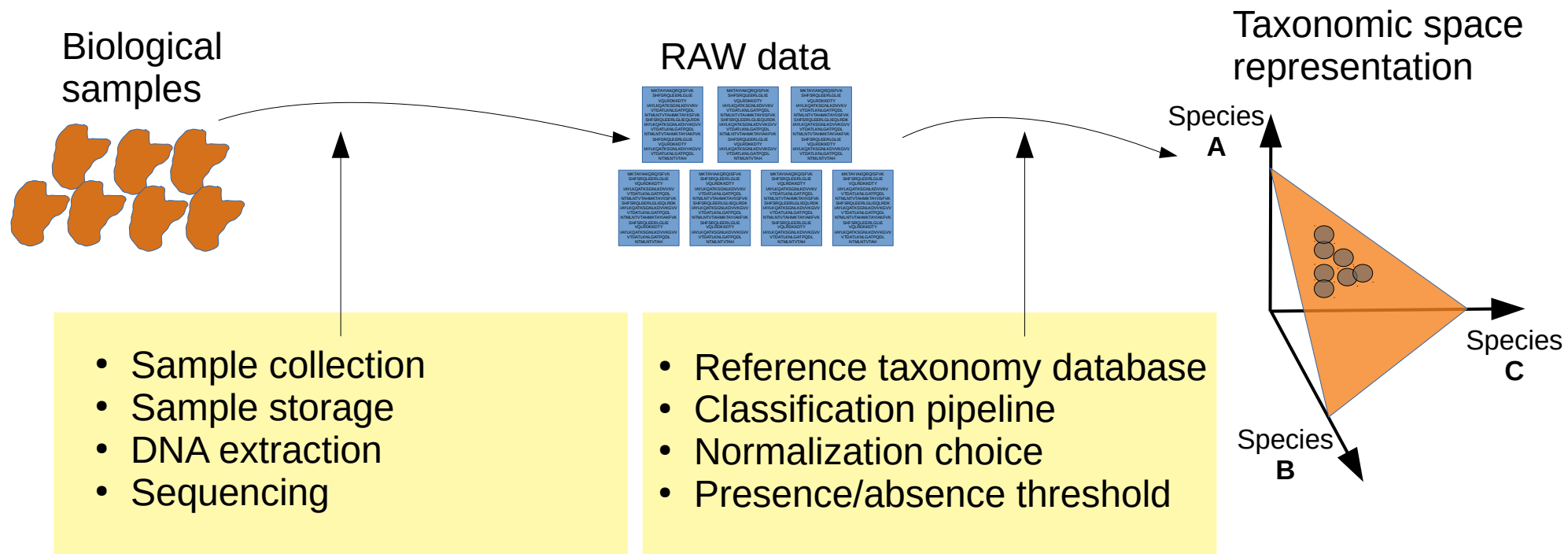
RELATIVE ABUNDANCE
PROFILE

<i>SPEC.</i>	<i>REL. AB.</i>
Species A	12%
Species B	15%
...	...
...	...
Species N	2%



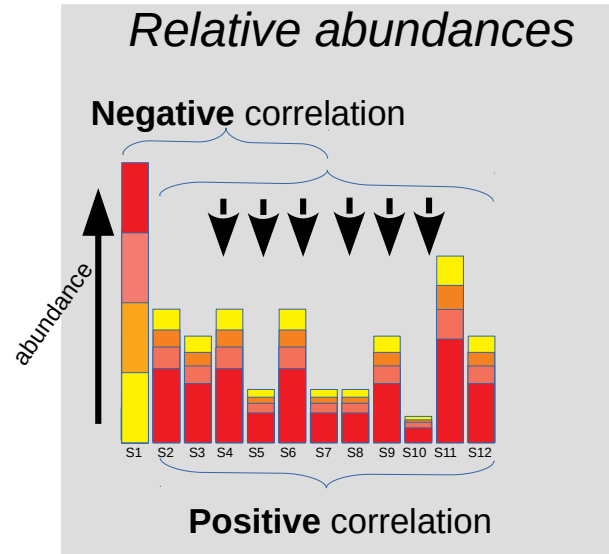
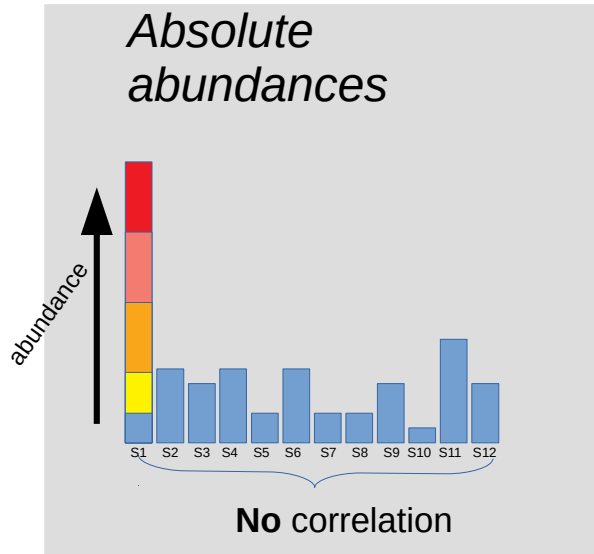
POINT ON
SIMPLEX

Each study comes with its own bias:



Challenging data environment:

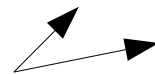
- Data dimensionality: ~1000 dimensions, ~100 samples;
- Compositional data: abundance comparison meaningful?



Dimensionality reduction through **Archetypal Analysis**

Convex combination:

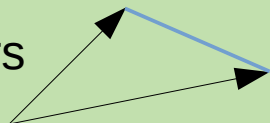
$$\vec{v} = a\vec{v}_1 + b\vec{v}_2$$



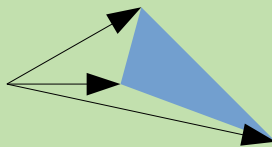
Kinds of Vector Combinations:

NAME	CONSTRAINT	SPAN	
1) Linear combination	→ no constraint	→ linear span	
2) Conical combination	→ coefficients ≥ 0 ($a > 0, b > 0$)	→ conical span	
3) Affine combination	→ coefficients sum = 1 ($a + b = 1$)	→ affine hull	
4) Convex combination	→ coefficients ≥ 0 & sum = 1 ($a > 0, b > 0, a + b = 1$)	→ convex hull	

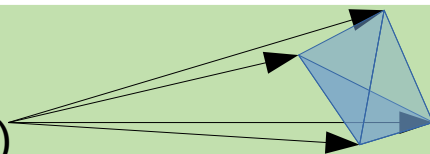
2 vectors
(line)



3 vectors
(triangle)



4 vectors
("tetrahedron")

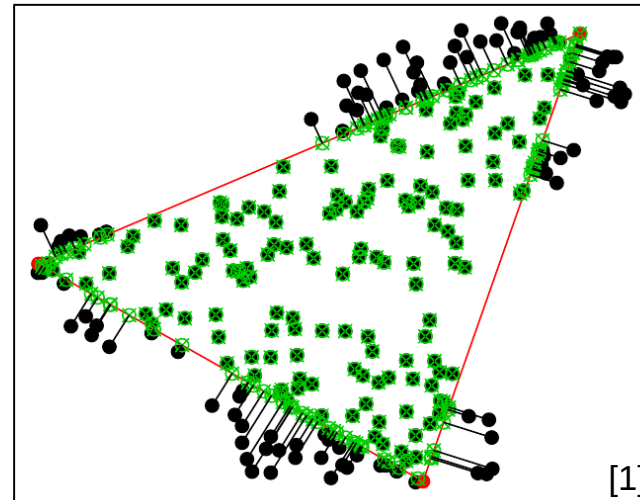
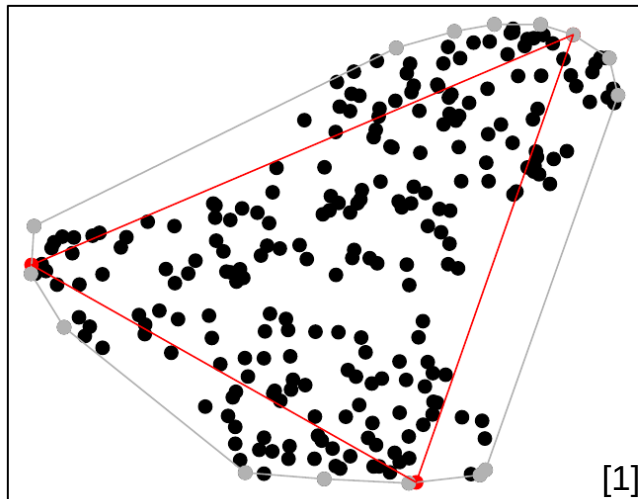


Archetypal Analysis:

2 Steps:

- 1) **Construct Archetypes** as convex combinations of data (B matrix)
- 2) **Reconstructed Data** as convex combination of Archetypes (A matrix)

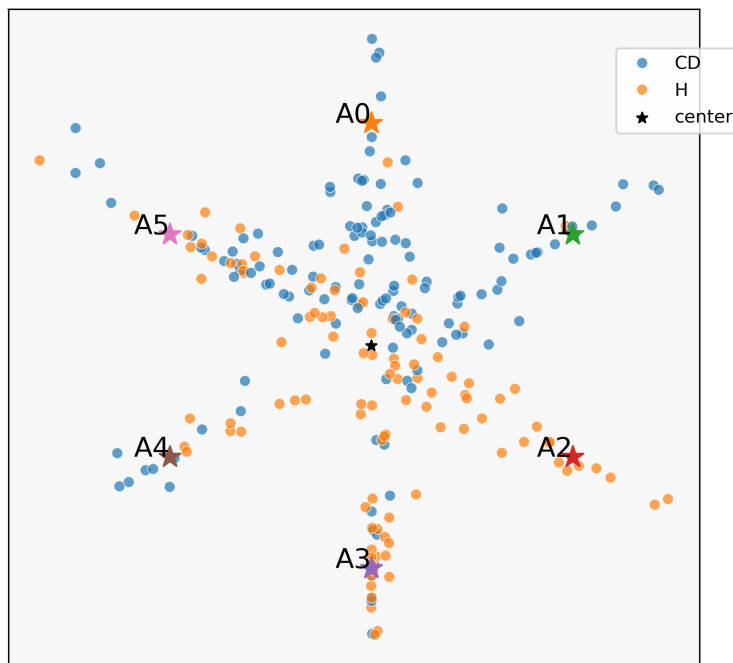
$$\begin{aligned}(A^*, B^*) &= \arg \min_{A, B} \text{RSS}(A, B|X) \\ &= \arg \min_{A, B} ||X - ABX||^2\end{aligned}$$



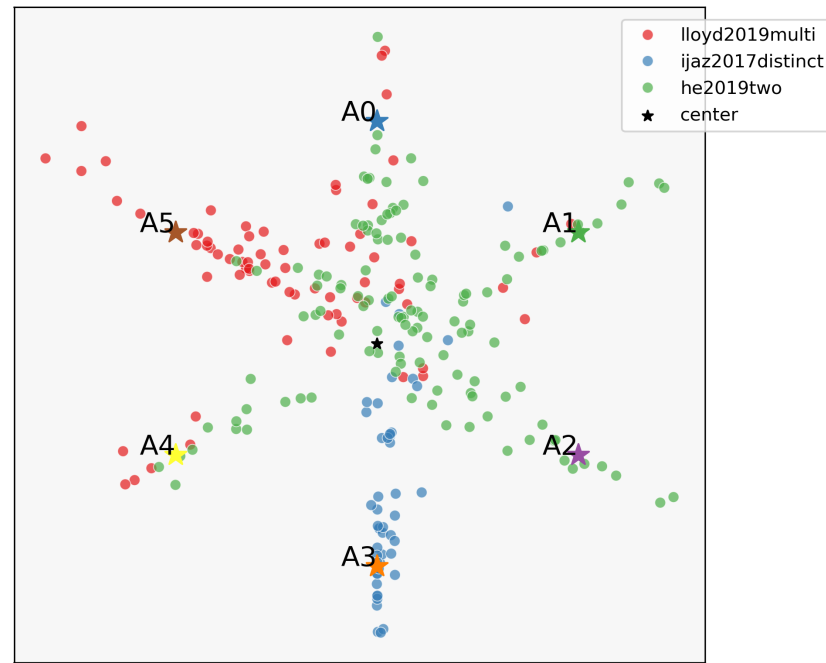
Archetypal representation of data:

Informative and **noninformative** archetypes:

Diagnosis:

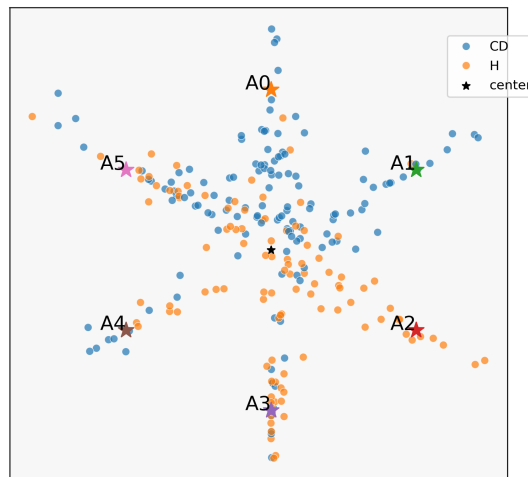


Study:

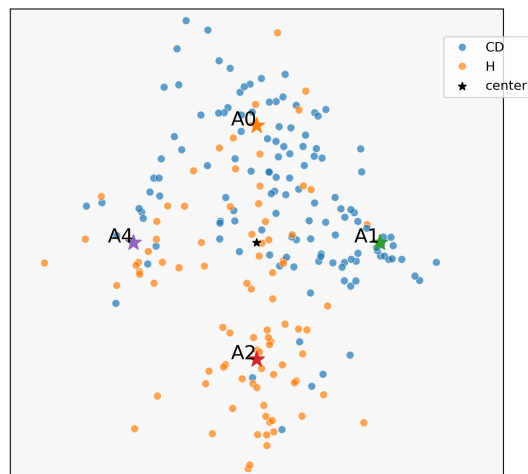


Mitigating study biases:

Diagnosis:

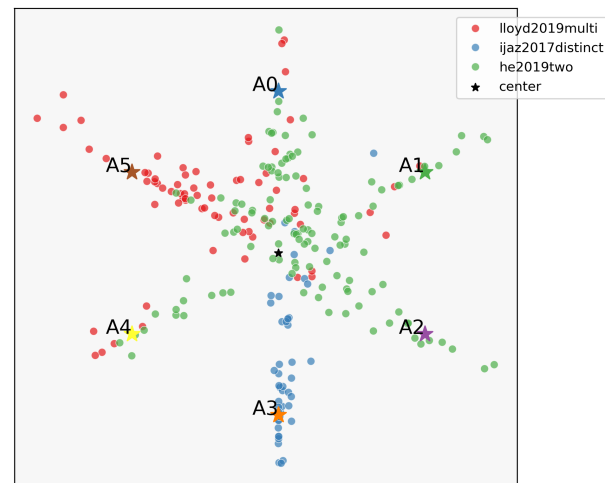


Diagnosis:

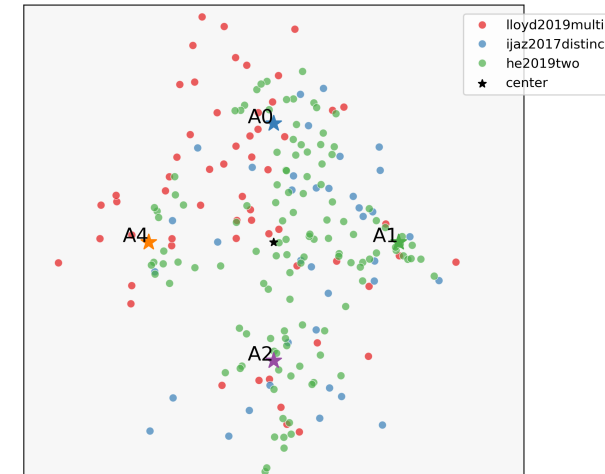


projection

Study:



Study:

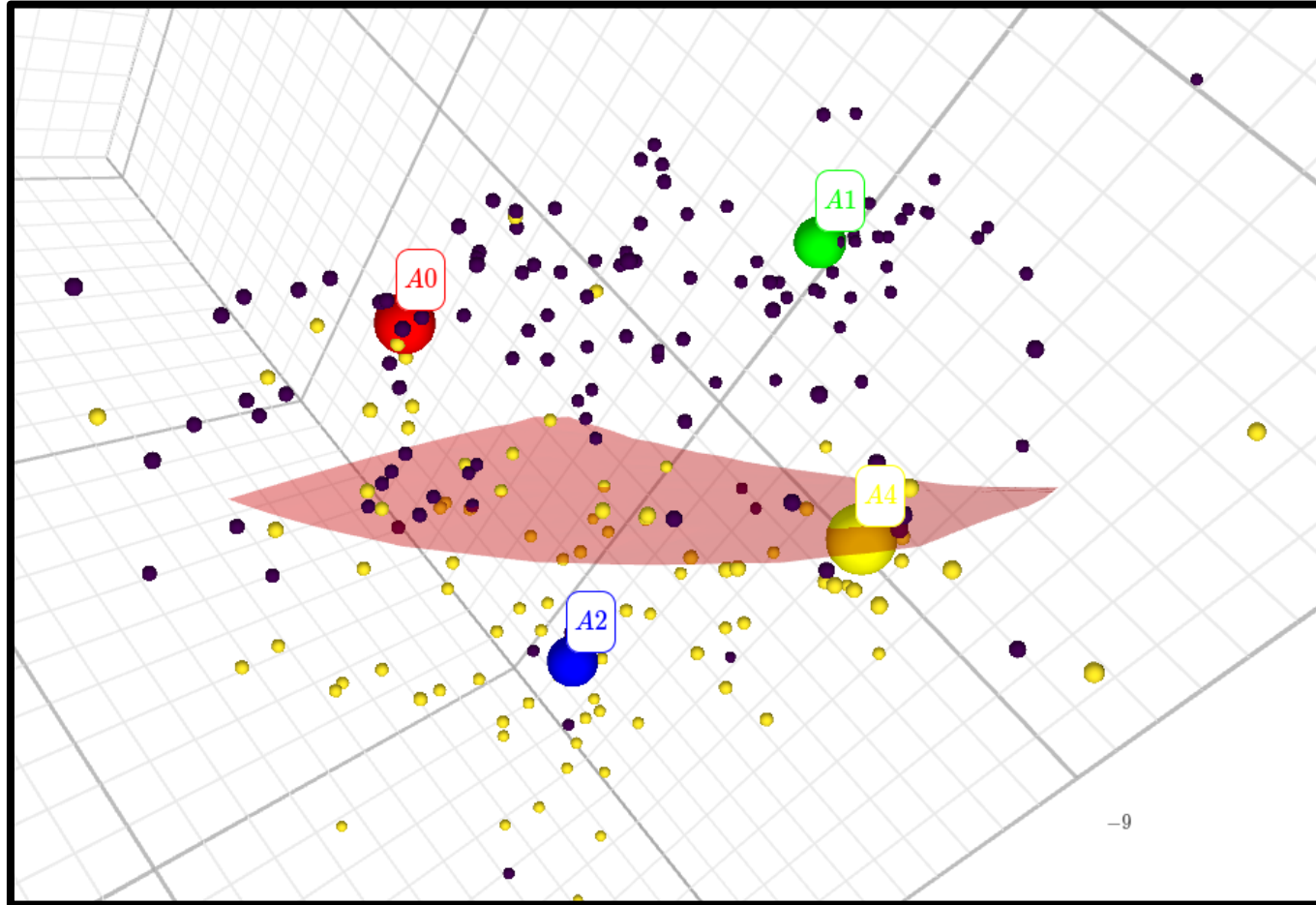


Remove dataset
specific information

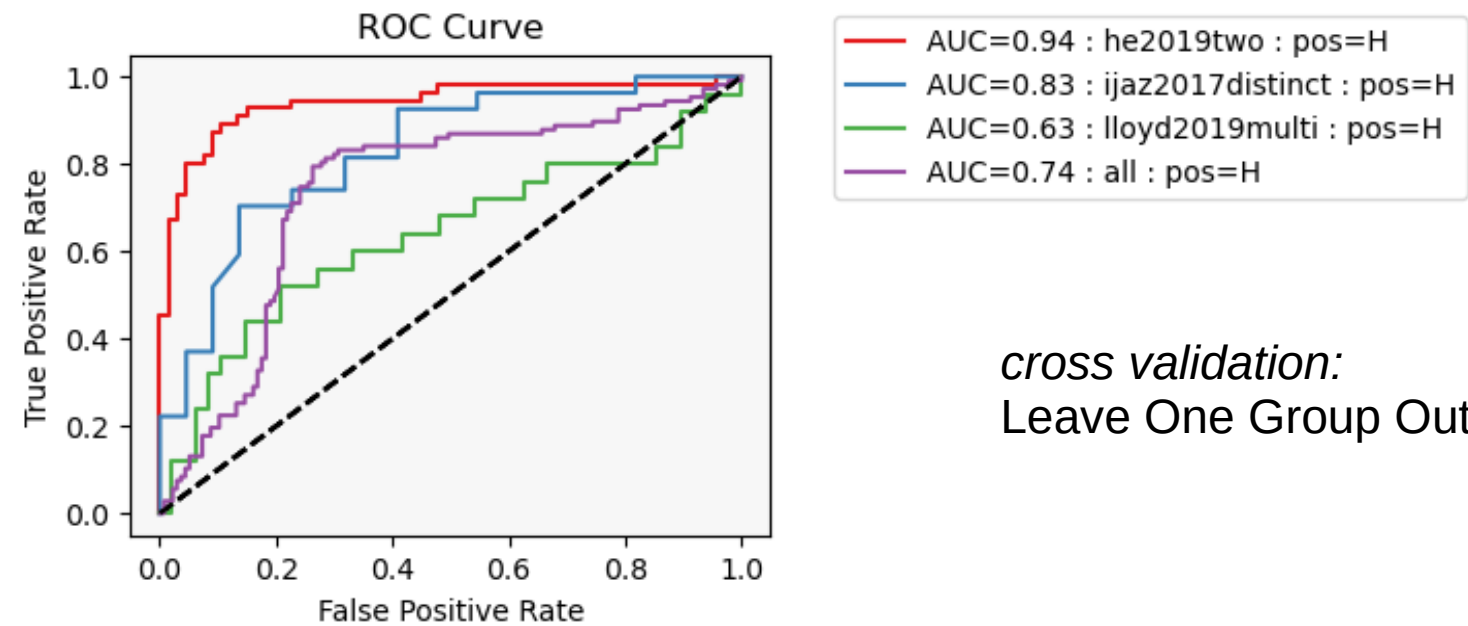
Keep diagnosis
related information

Classification with kernel SVM:

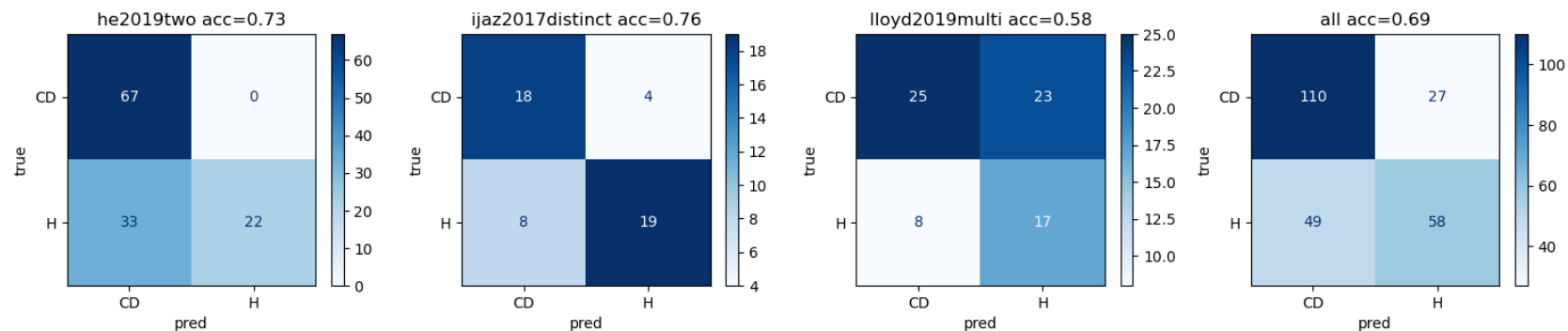
Decision boundary in archetypal representation space:



Preliminary Results of classification framework:

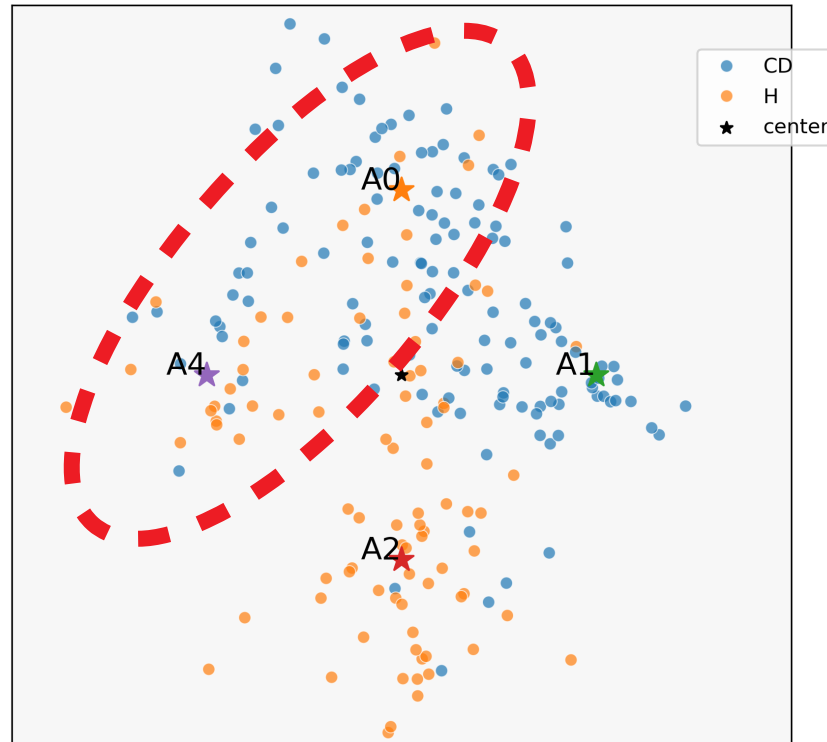


cross validation:
Leave One Group Out (LOGO)



Room for of improvement:

- Preselection of informative species
- Defining a non-discrimination region



Thank you!

