Supervised Projective Learning with Orthogonal Completeness

Tyler J Grear

Code 610.1 Intern

Global Modeling and Assimilation Office

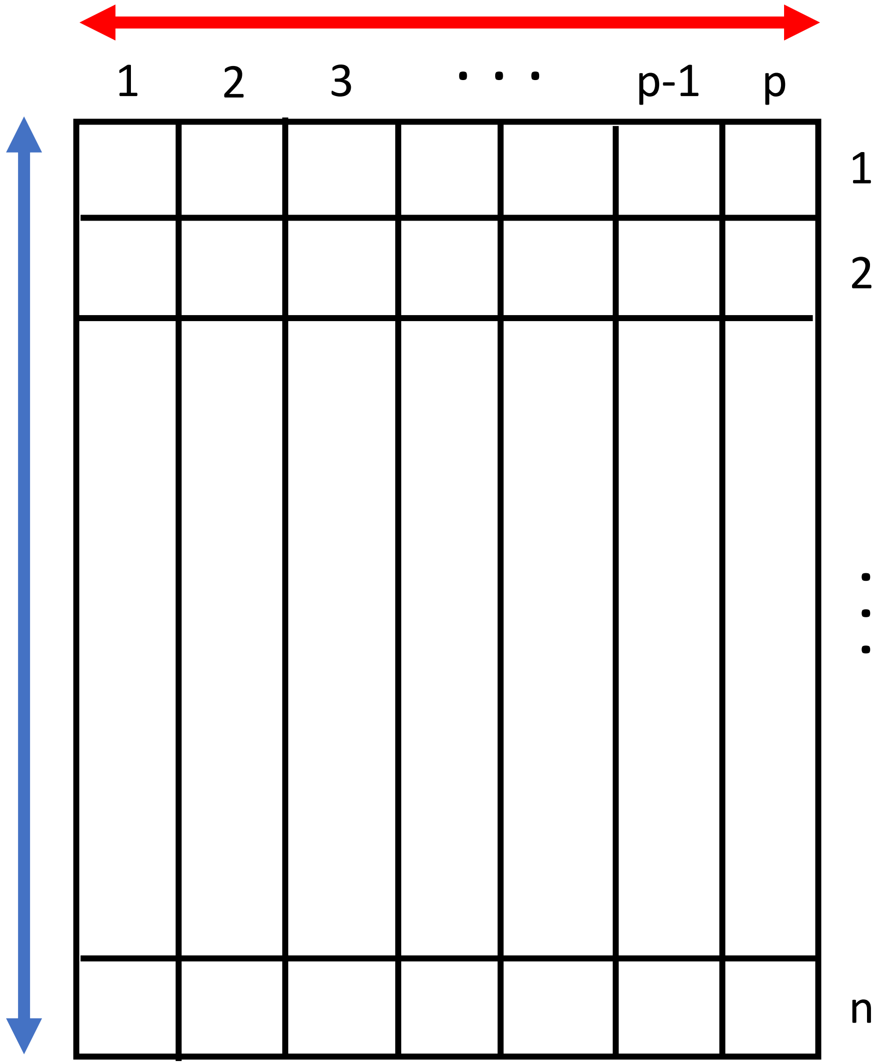
NASA Goddard Space Flight Center

7-23-2021

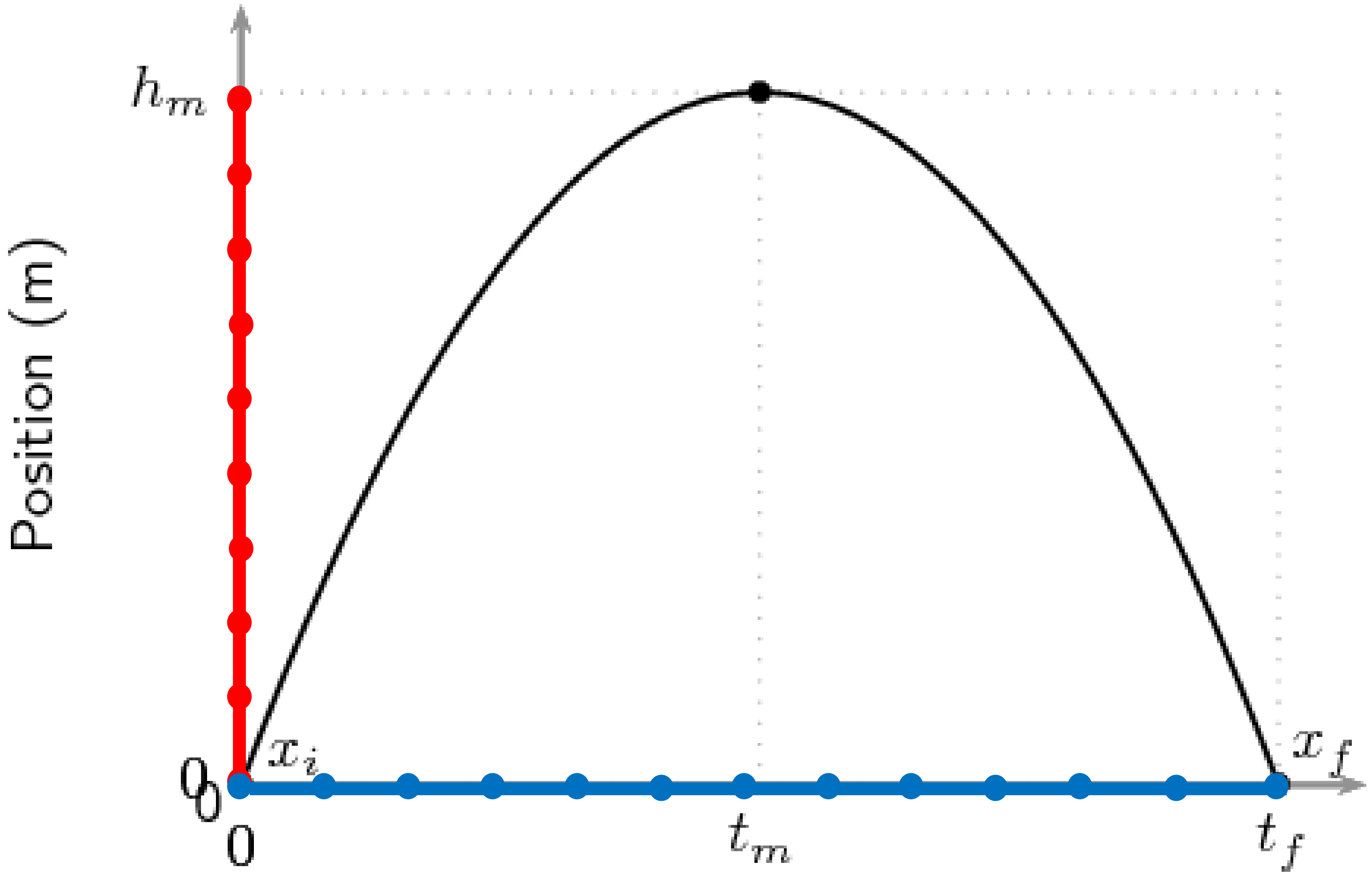
Dimension reduction (DR)

- Dimensions:
  - Features, independent variables, degrees of freedom (df), predictors, etc.

- Samples:
  - Observations, frames, measurements, time, etc.



Dimension reduction (DR): a projectile example



Supervised Projective Learning with Orthogonal Completeness

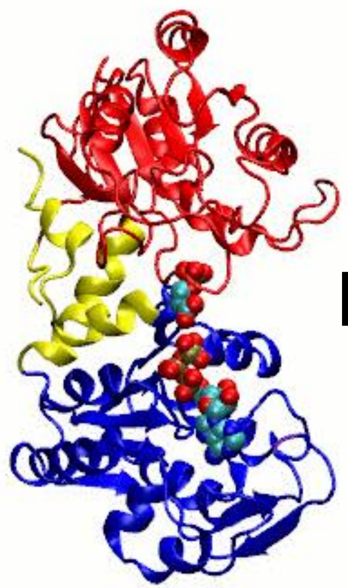
- **Supervised: class labels associated with data**
- **Projective: develop projection operators (linear operators)**
- **Learning: optimization formalized as a recurrent neural network (RNN)**
- **Orthogonal Completeness: procured basis sets are constrained to satisfy the completeness theorem such that:**

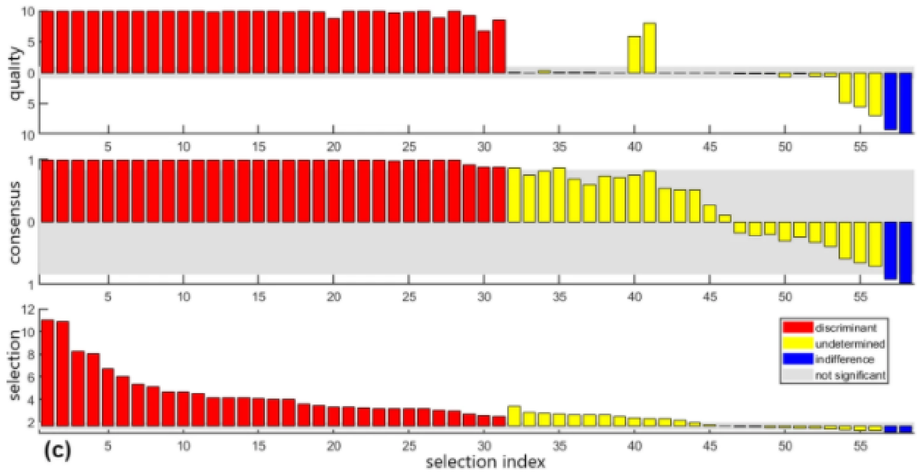
$$\langle j, s | i, s \rangle = \delta_{ji} \quad \text{and} \quad \sum_{i=1}^{p} |i, s\rangle\langle i, s| = I$$

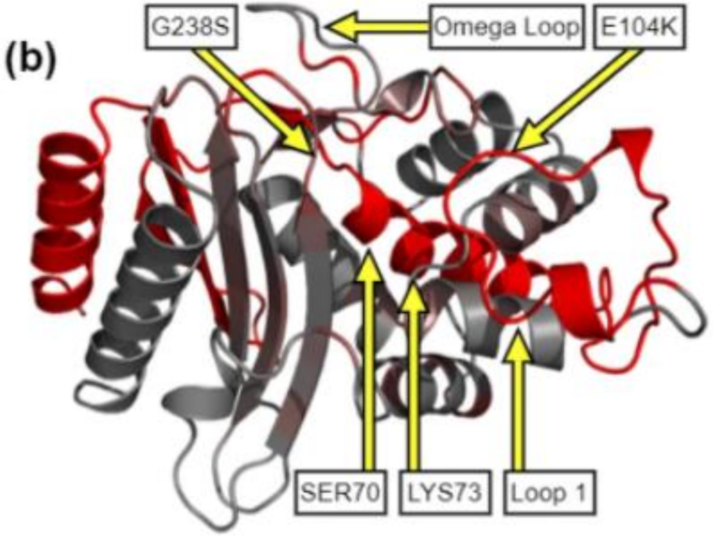
Molecular function recognition by supervised projection pursuit machine learning

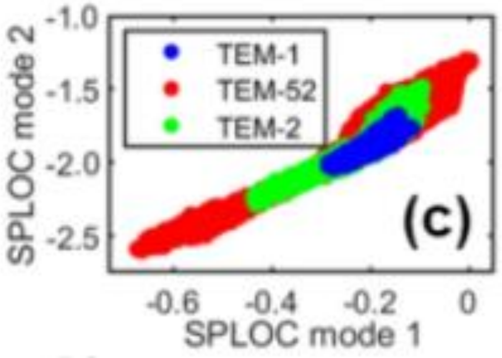
Tyler Grear, Chris Avery, John Patterson & Donald J. Jacobs

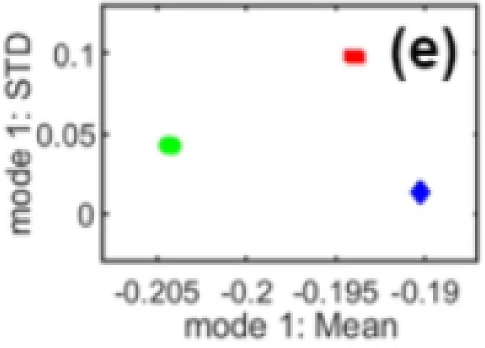
Scientific Reports 11, Article number: 4247 (2021) | Cite this article



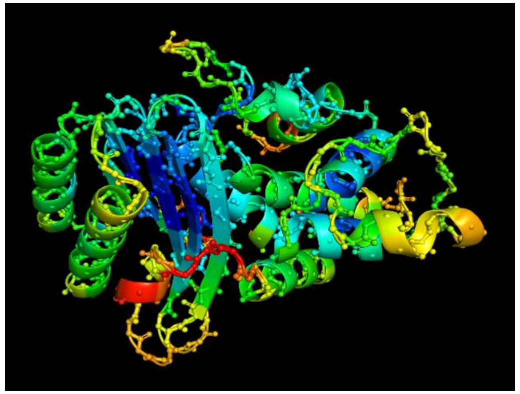








Example: Essential Dynamics (PCA)



=

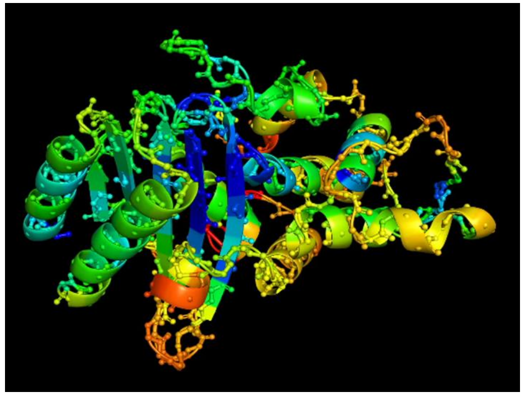
Empirical orthogonal function (EOFs)
Basis vectors

Each of these "modes" of motion are mapped to a single eigenvector from PCA.

We can map the PCA motions to the exact atoms taking part as shown by the coloring here.

+

+



What is projection pursuit?

Randomly generated unit vectors are iterated through a high-dimensional space while an objective function is optimized to identify interesting univariate projections.

PCA

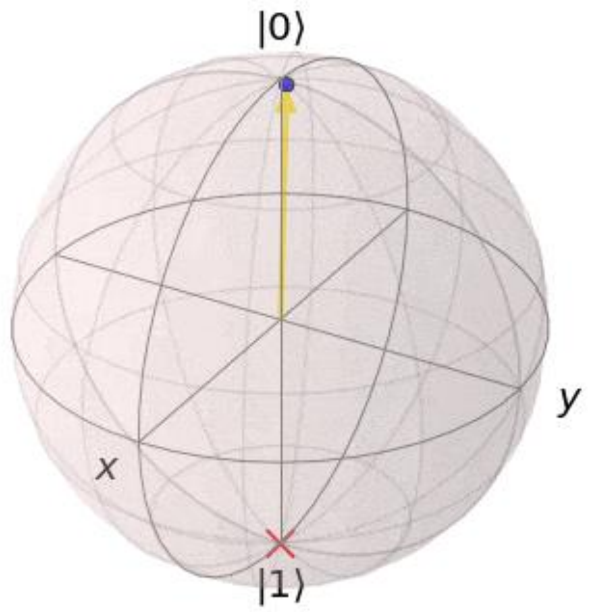
$$f(x) = max[var(\mathbf{Xr})]$$

PLS

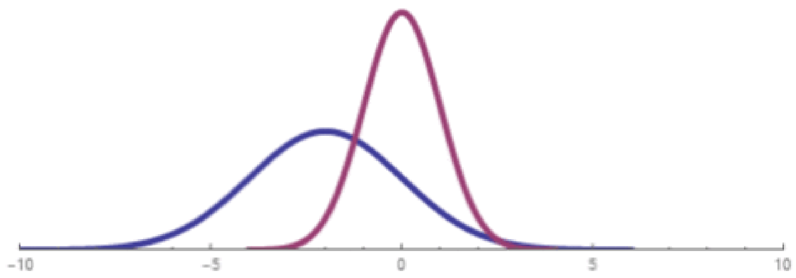
$$f(x) = max[cov(\mathbf{t}, \mathbf{u})]^2$$

ICA

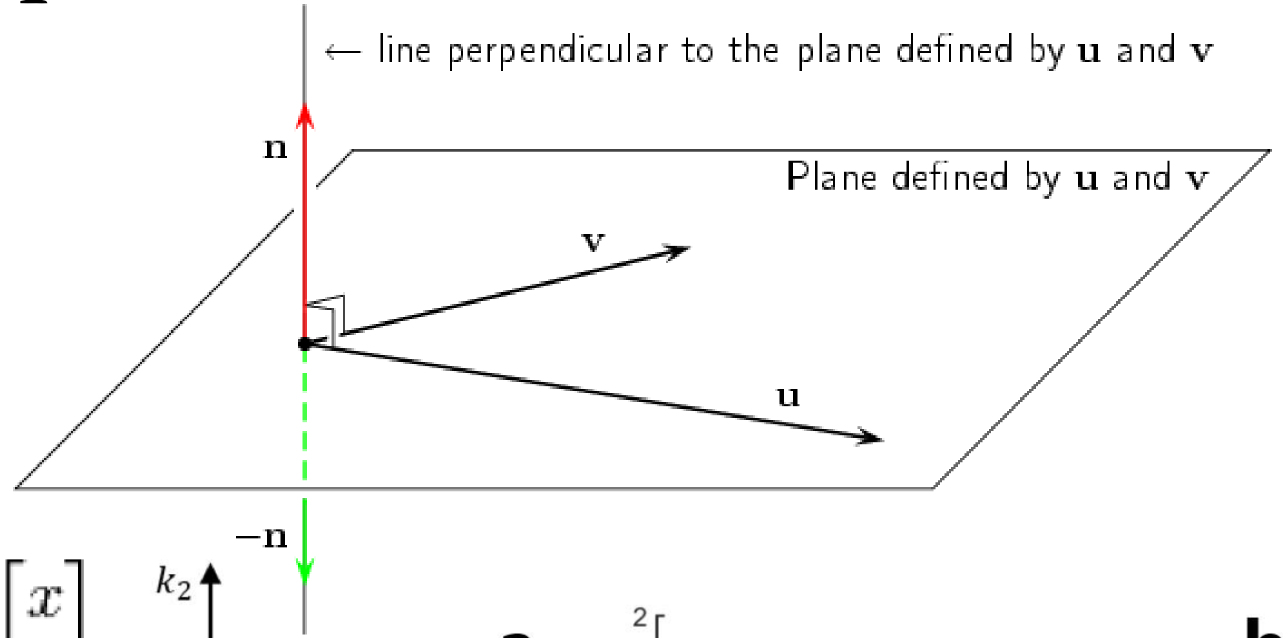
$$f(x) = max|kurt(x)|$$



Friedman and Tukey (1974)

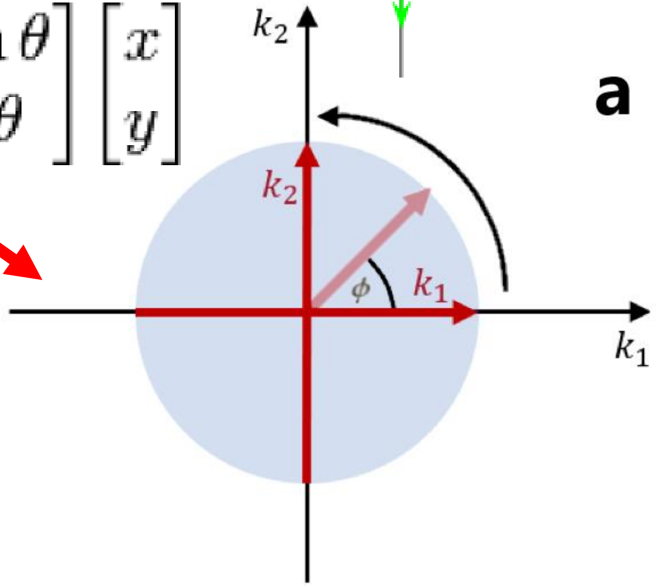


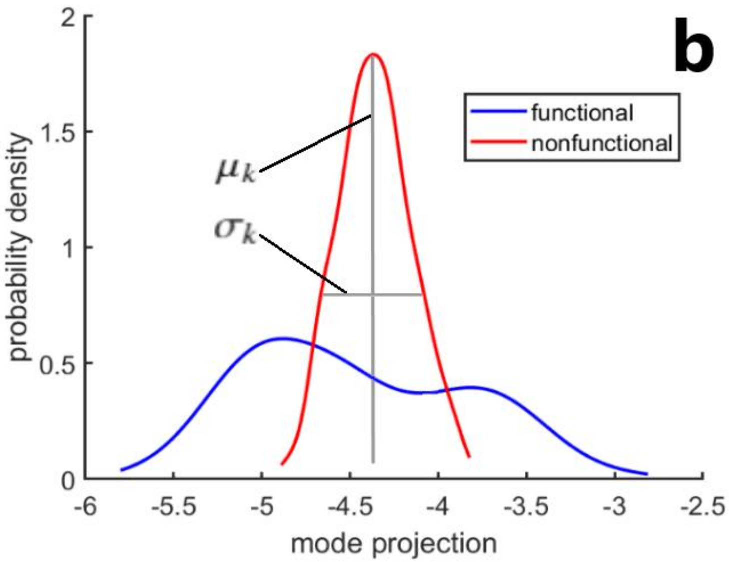
Projection pursuit in SPLOC



Rotation matrix

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$



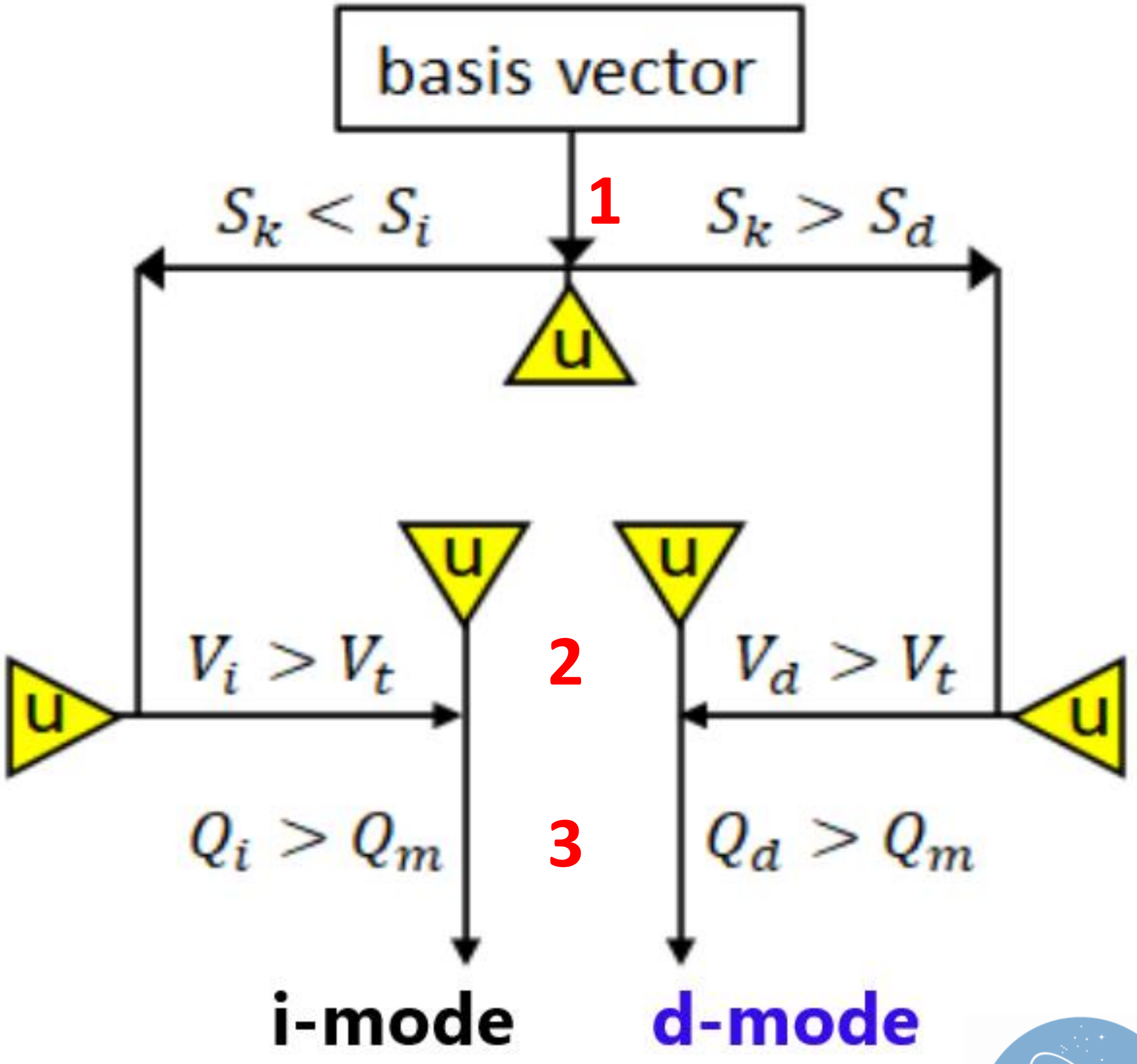


The decision triad

1) **Signal-to-noise**

2) **Statistical significance**

3) **Quality of clustering**



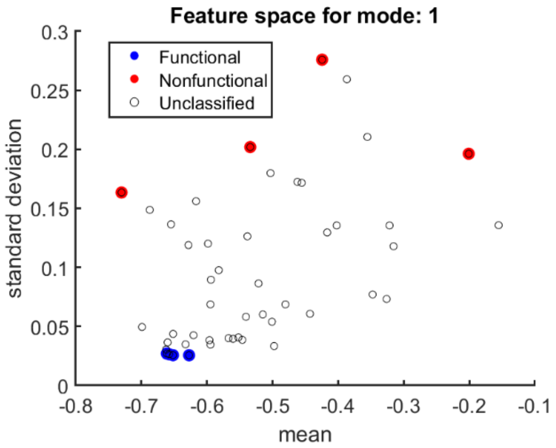
The decision triad

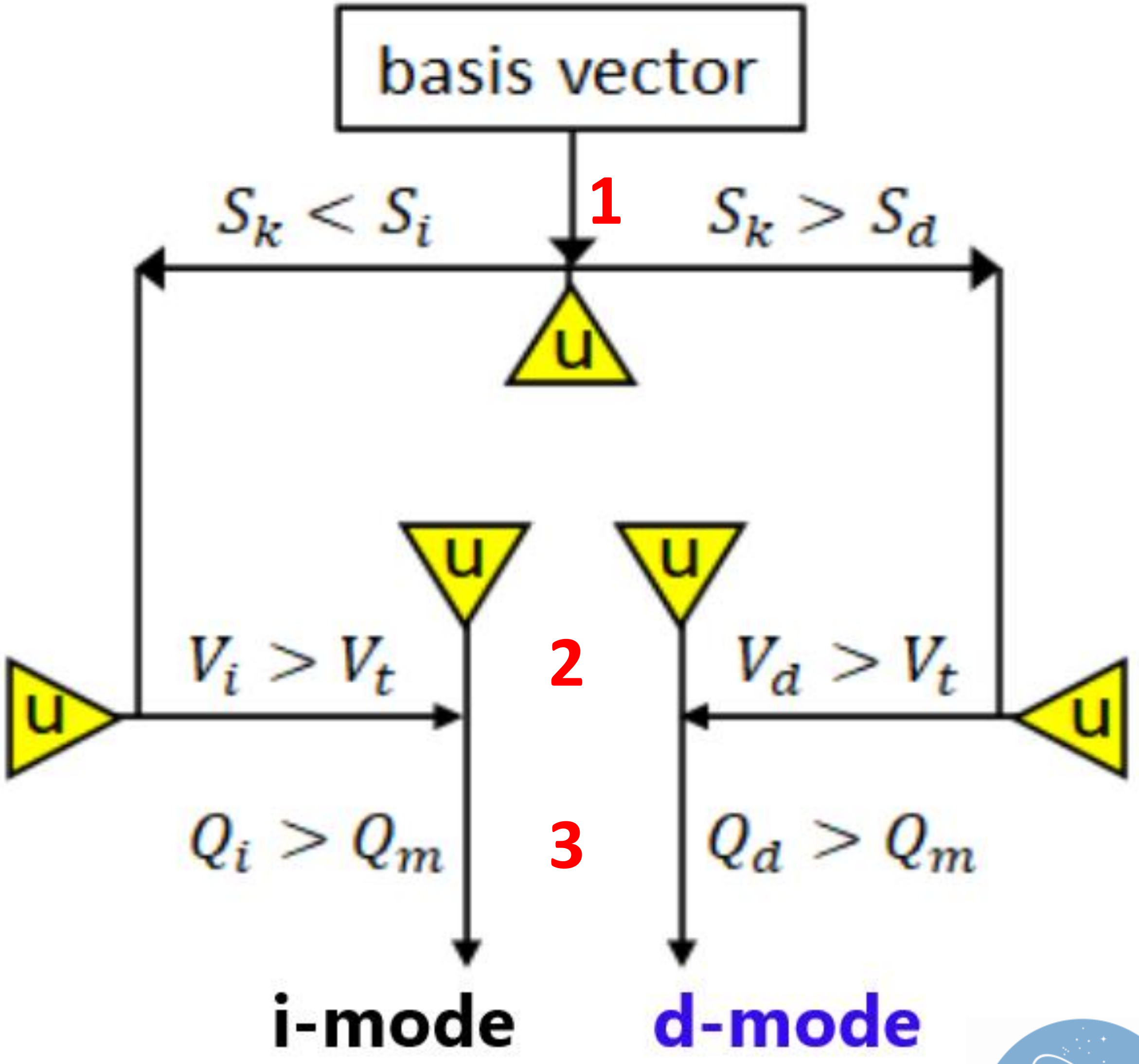
1) $$S_k(\alpha,\beta) = \begin{cases} \sqrt{sbn^2 + rex^2} + 1 & \text{when} > S_d \\ \sqrt{snr^2 + rex^2} + 1 & \text{when} < S_i \\ S_m & \text{otherwise} \end{cases}$$

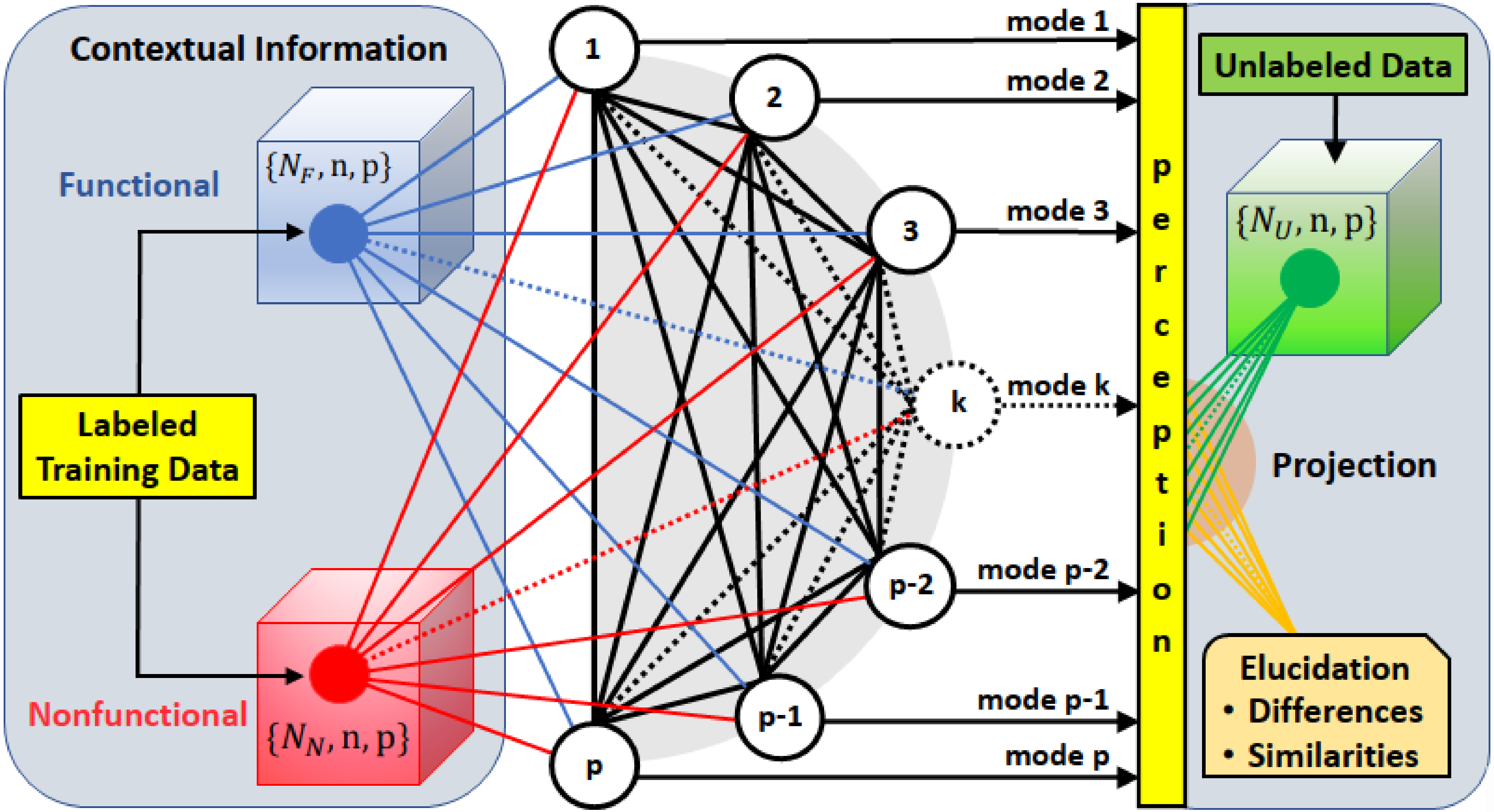
2) $$f_d(x) = [1 + \exp(16(x_d - x))]^{-p_d(x)}$$

$$f_i(x) = \left[1 - [1 + \exp(16(x_i - x))]^{-1}\right]^{p_i(x)}$$

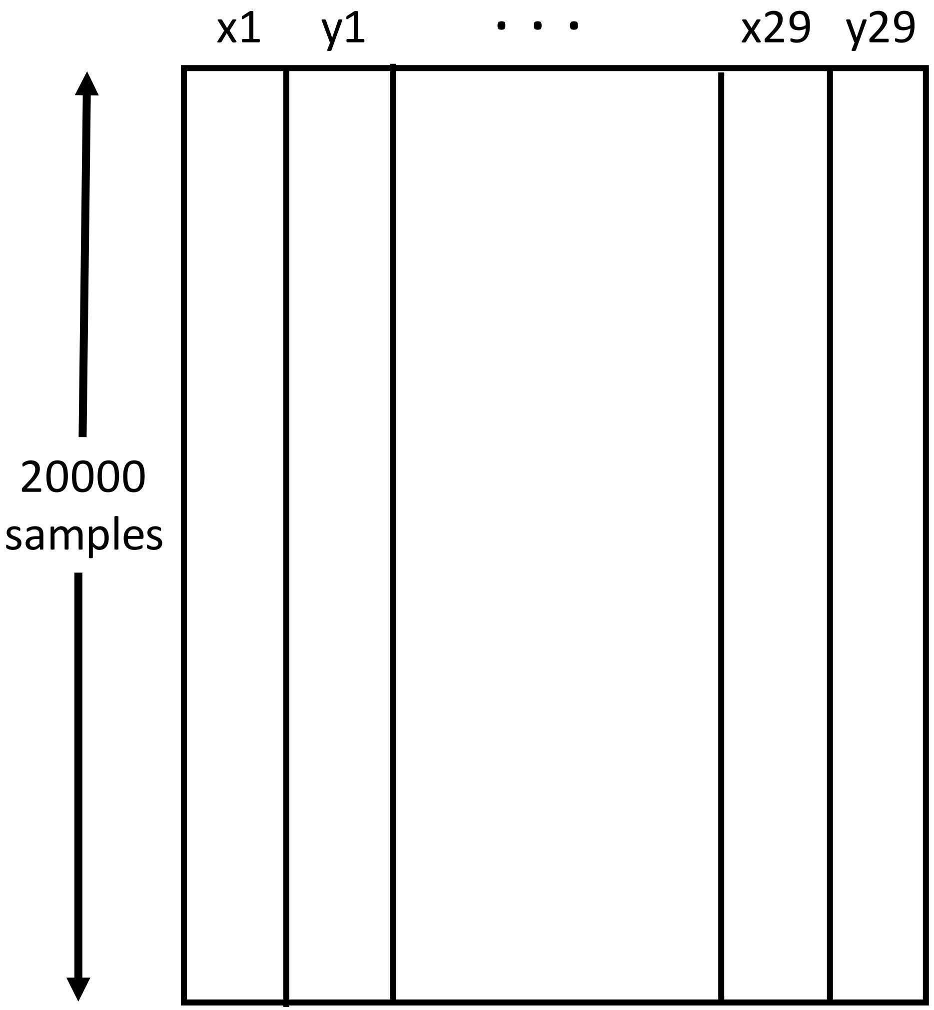
3)

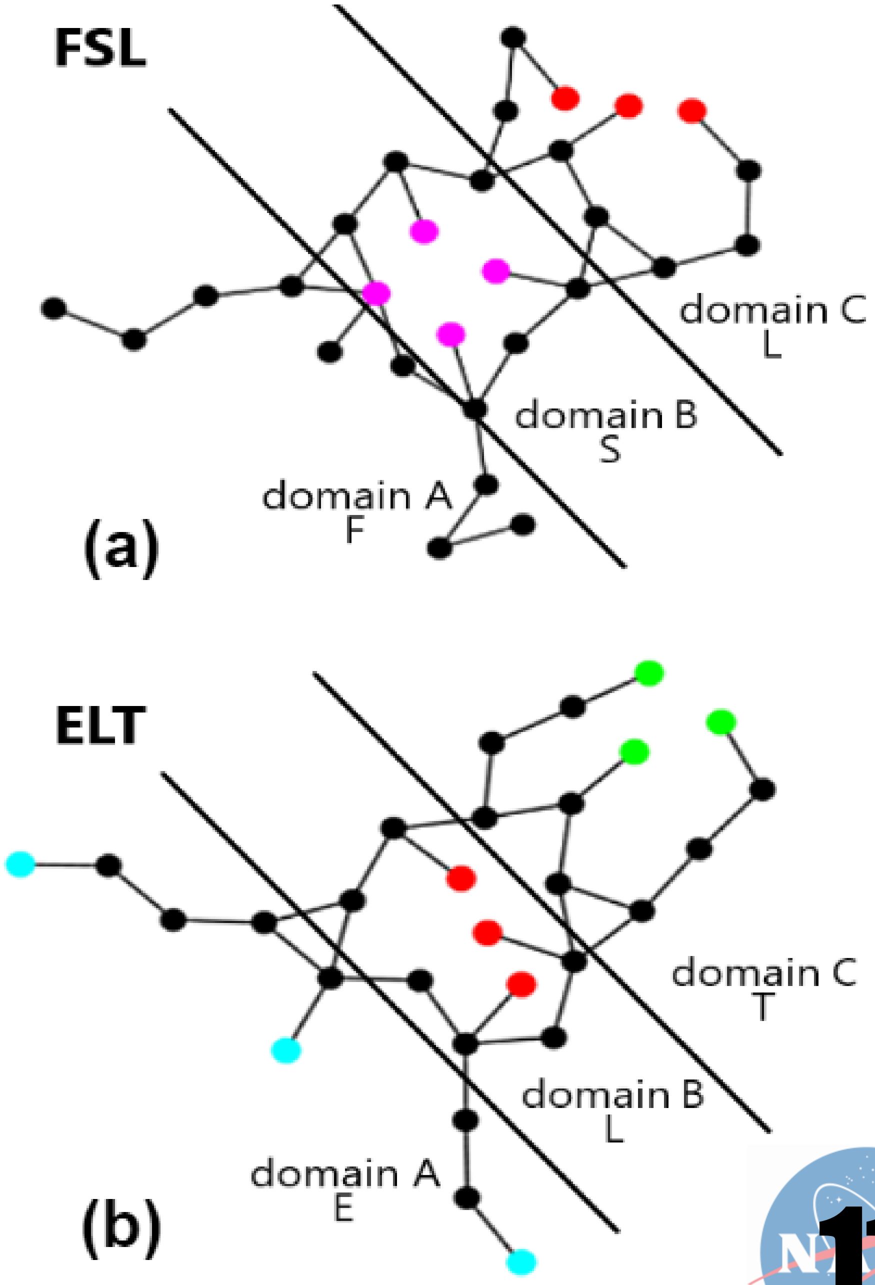





Contextual Information
Functional
$\{N_F, \mathrm{n}, \mathrm{p}\}$
Labeled
Training Data
Nonfunctional
$\{N_N, \mathrm{n}, \mathrm{p}\}$
1
2
3
k
p-2
p-1
p
mode 1
mode 2
mode 3
mode k
mode p-2
mode p-1
mode p
perception
Unlabeled Data
$\{N_U, \mathrm{n}, \mathrm{p}\}$
Projection
Elucidation
• Differences
• Similarities

The synthetic molecules





The training data

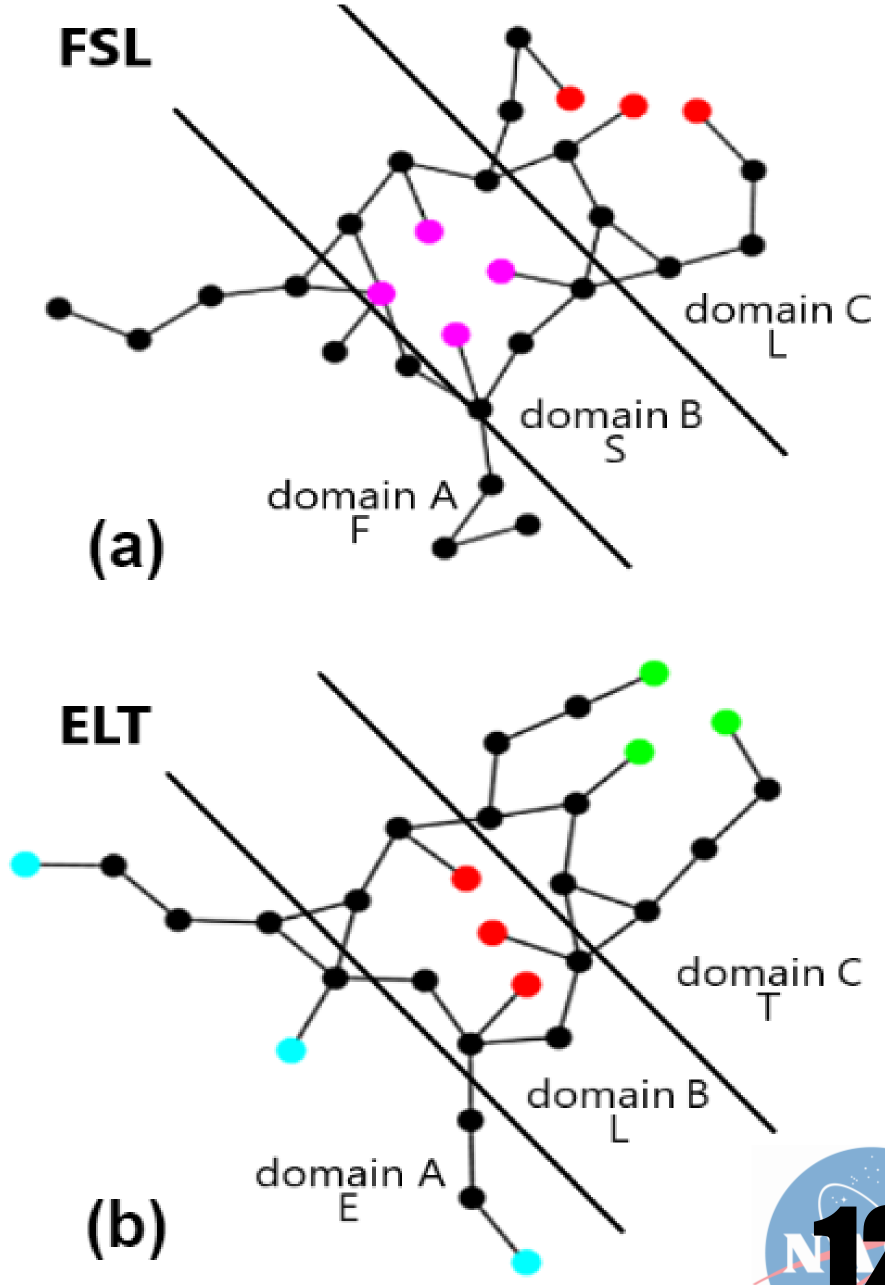
- Labeled functional:
  - {FLL,FLF}

Training set

- Labeled nonfunctional:
  - {FFF,FFL}

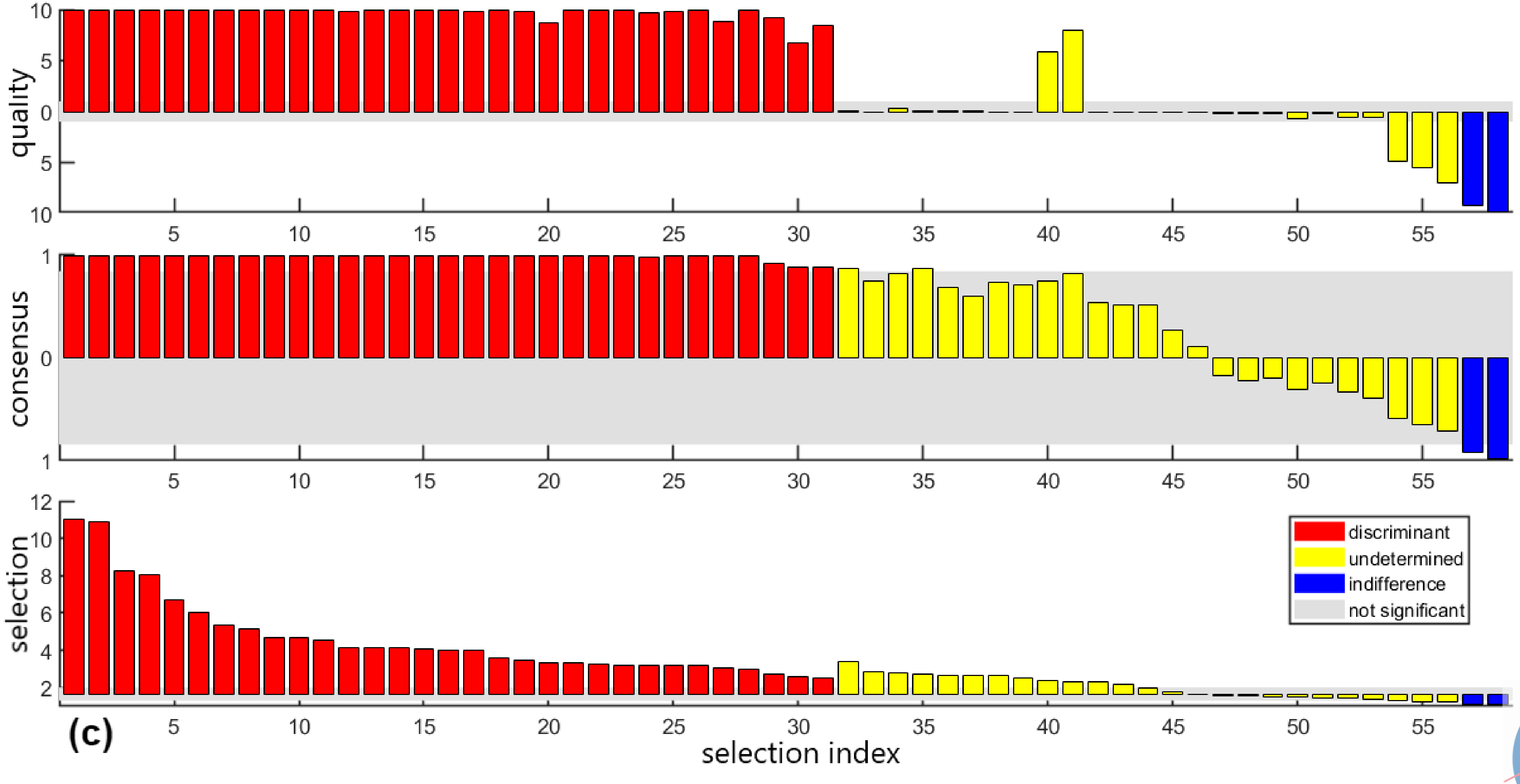
- Remaining 20/24 synthetic molecules are not labeled (unknown to machine).

Testing set

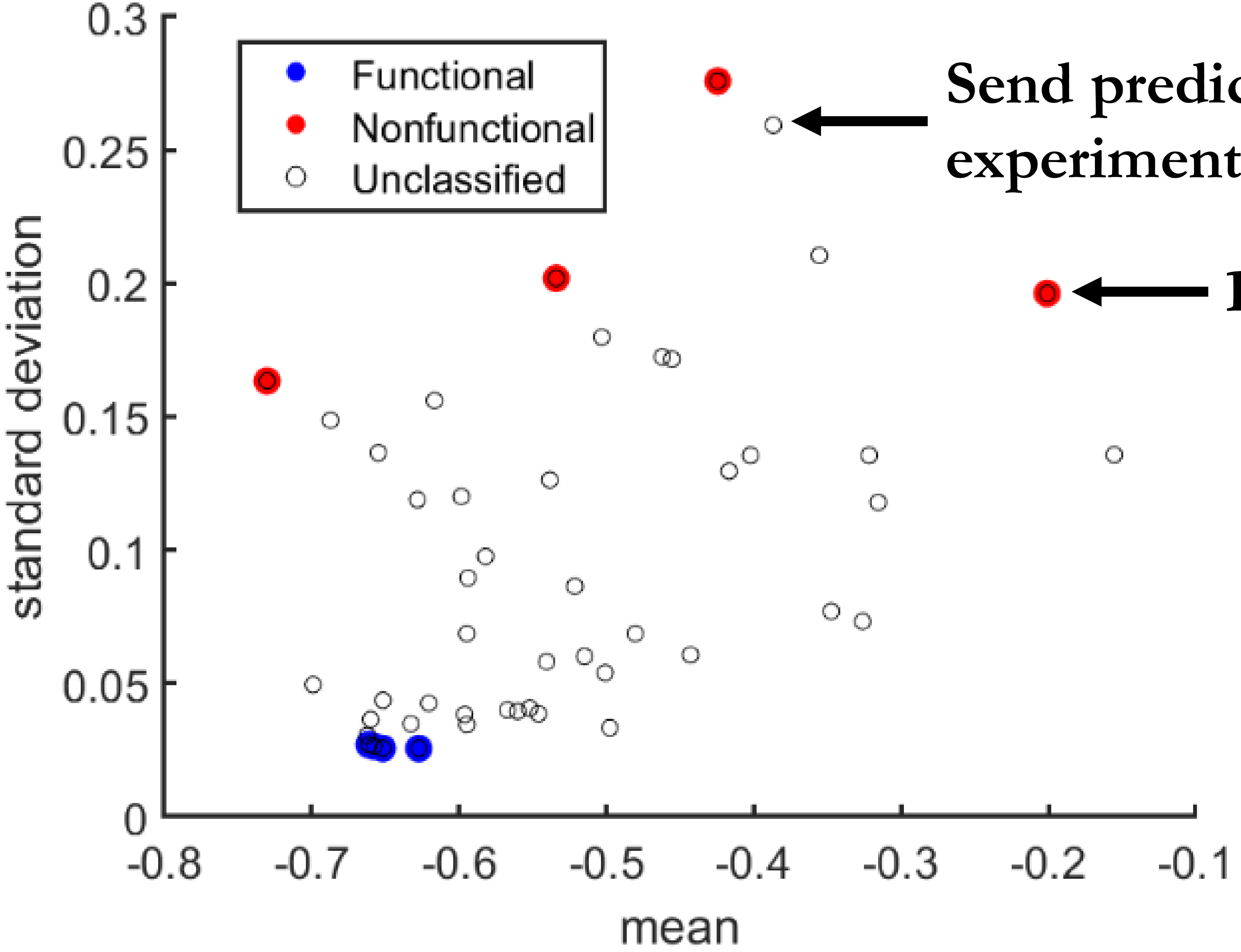


SPLOC basis vector spectrum

Decision triad



Mode feature space plane



Send prediction to experimentalists

Data packet

Thank you for your time,

Tyler J Grear

Code 610.1 intern
Global Modeling and Assimilation Office
NASA Goddard Space Flight Center
7-23-2021

https://github.com/BioMolecularPhysicsGroup-UNCC/MachineLearning