Functional data analysis tools for autonomous experimentation

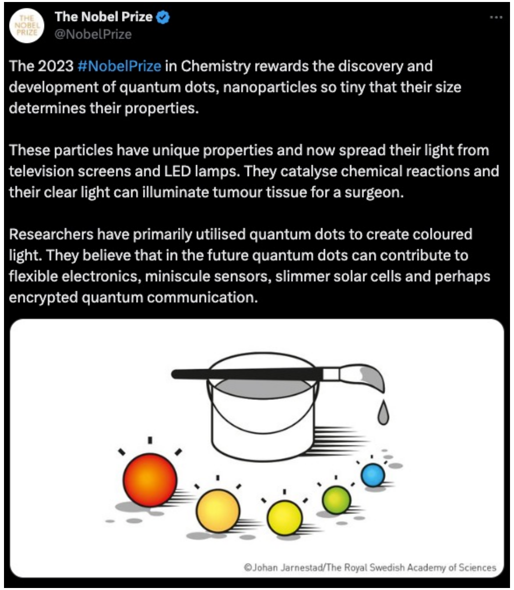
UW Data science Seminar

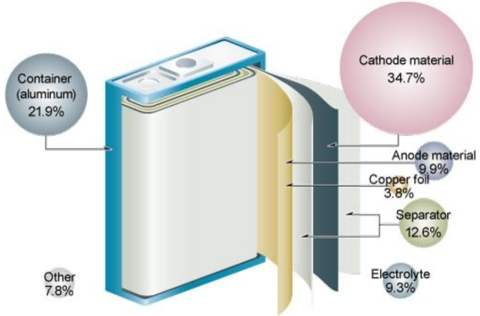
Kiran Vaddi [1,2], Huat T Chiang [1], Kacper Lachowski [1], Karen Li [1] , Lilo D Pozzo [1,2]

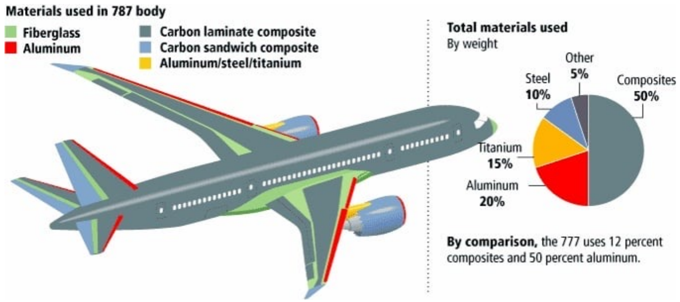
[1] Department of Chemical Engineering
[2] eScience Institute

Material solutions for modern challenges

- New and better materials can fundamentally change our future

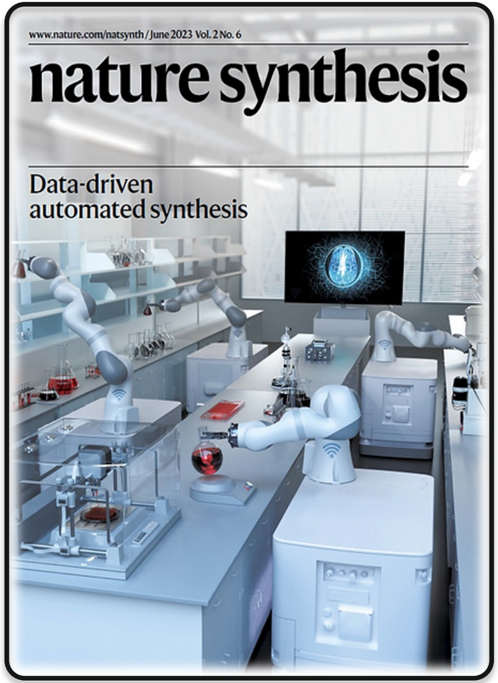


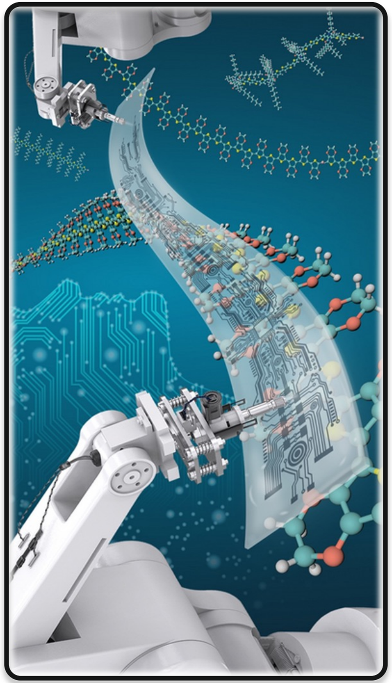




Accelerating design & discovery of materials

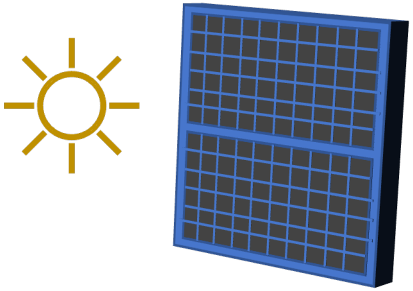
- ➢ Improve materials timeline from decades to a few years





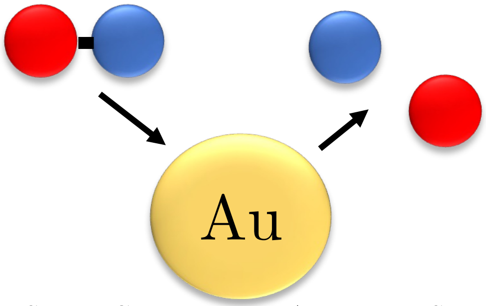
Nanomaterials in everyday life...

High-Efficiency Solar Cells



Omrani, M., Keshavarzi, R., Abdi-Jalebi, M. *et al.* *Sci Rep* 12, 5367 (2022)

Catalysis with high activity and selectivity

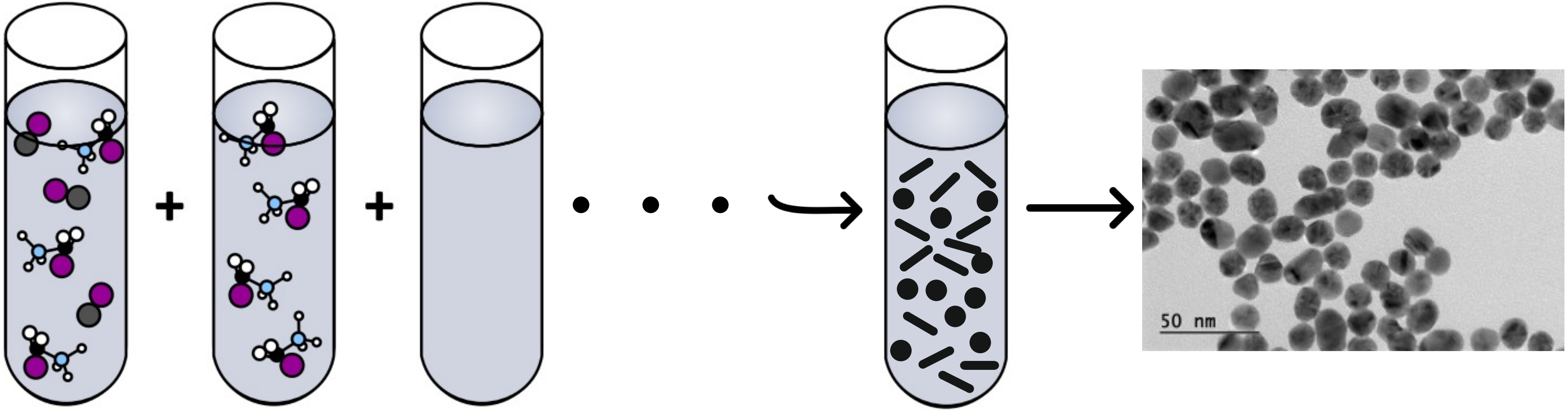


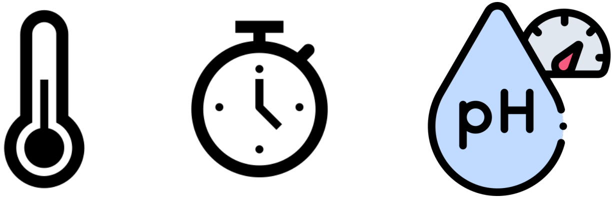
Steven Chavez, Umar Aslam, and Suljo Linic, *ACS Energy Letters* 2018

High-Performance Batteries

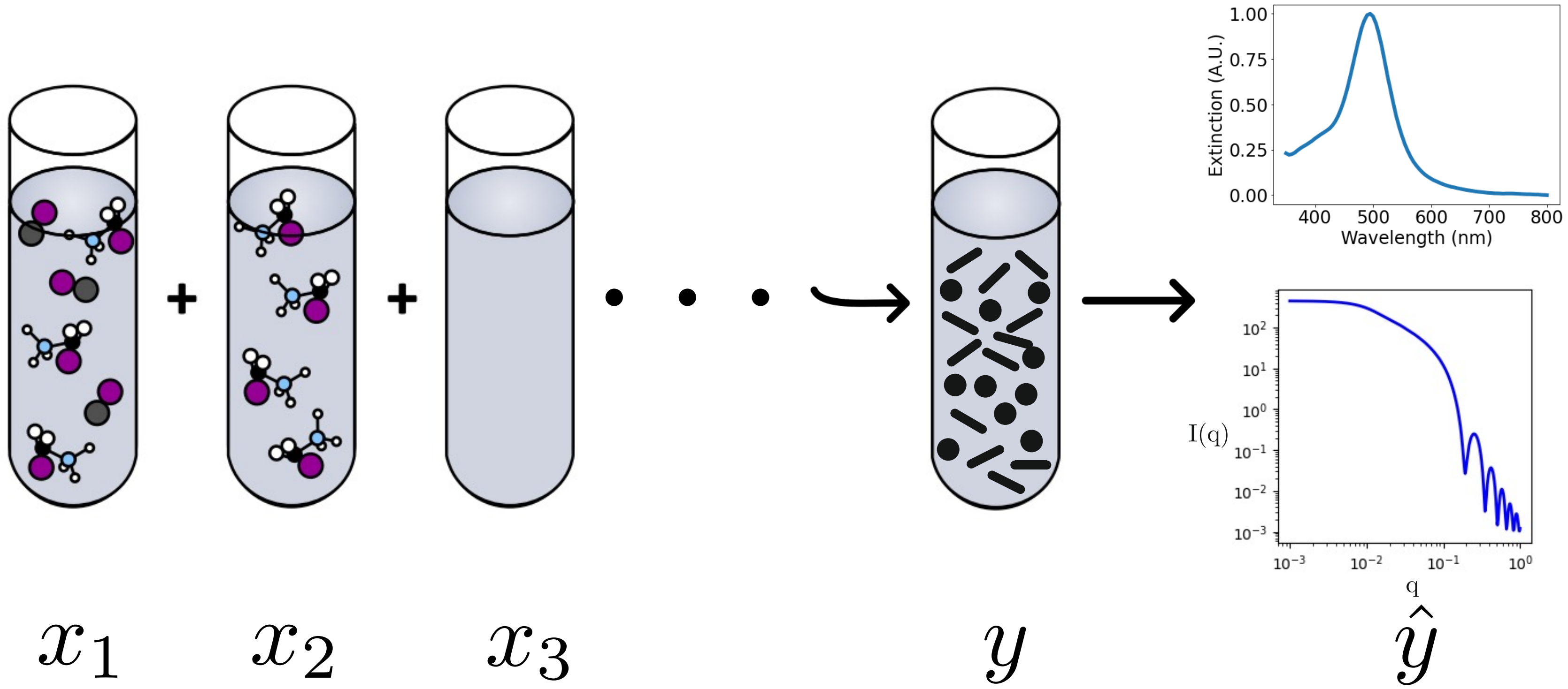
Jun-Fan Ding, et al., 27 June 2020, Nanoselect 94-110

Digitization of experimental material synthesis

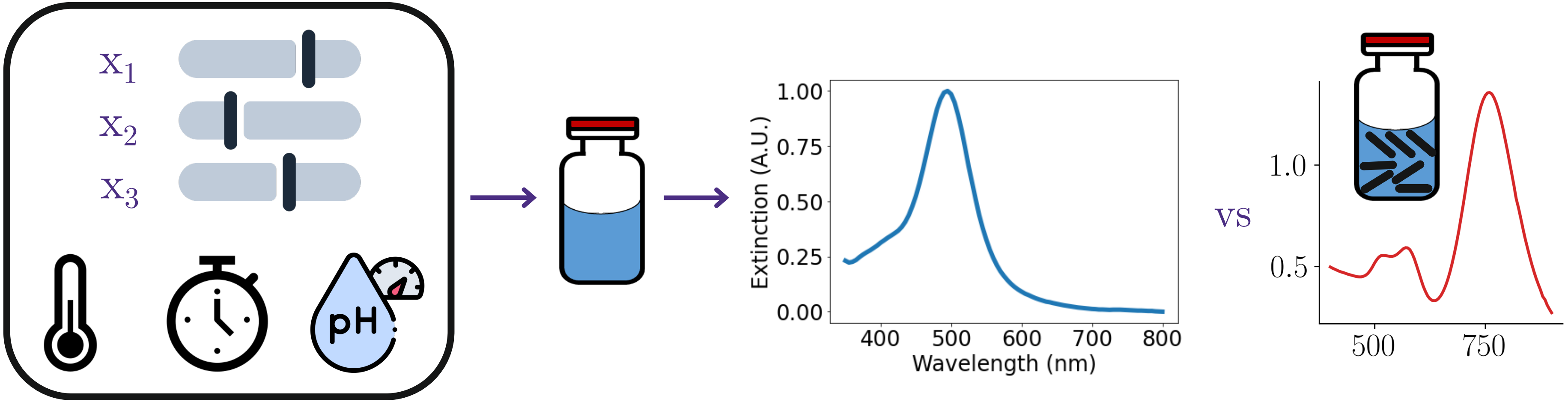




Digitization of experimental material synthesis



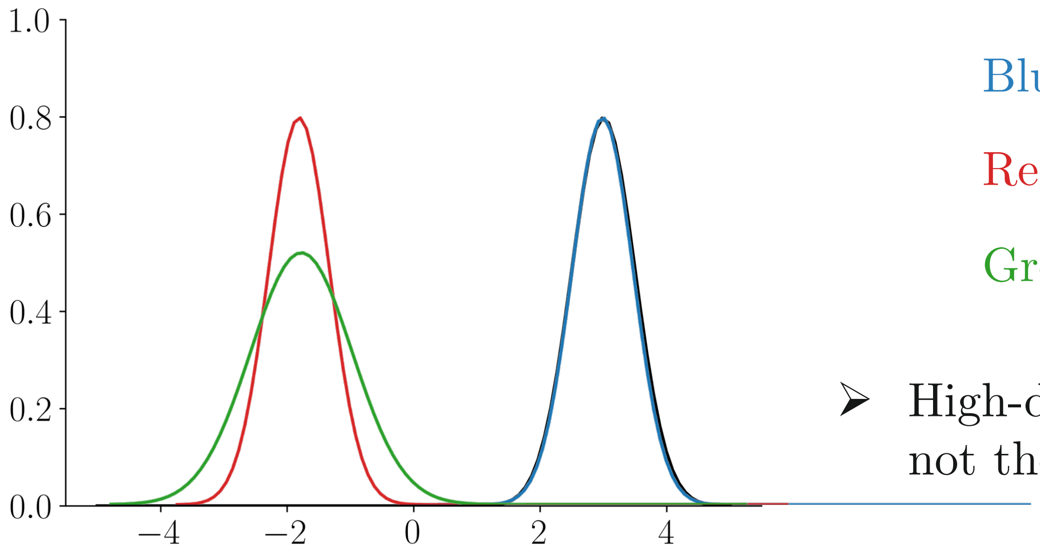
Optimizing experimental frameworks



➢ Challenge: Blackbox optimization and Comparing spectral data

Comparing spectra – Euclidean distance

$$d(y, y^*) = \sum_{i=1}^{n} (y_i - y_i^*)^2$$



Blue : 3.34

Red : 3.34

Green : 3.01

⁉

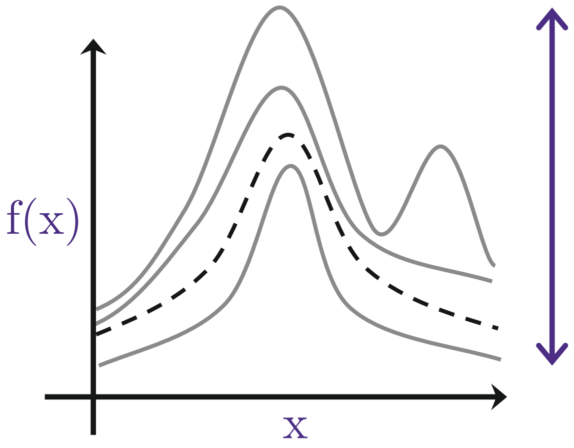
⁉

➢ High-dimensional Euclidean is not the "right" representation

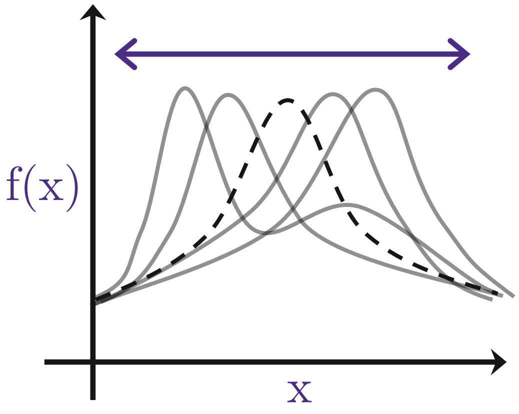
Comparing spectra – Functional Data Analysis

- Shape mismatch = distance along y-axis + x-axis

Amplitude distance



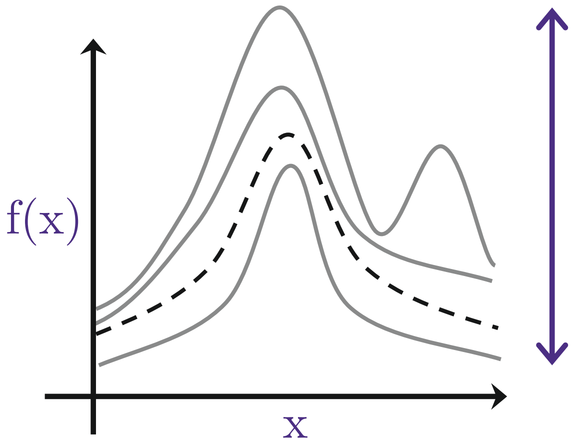
Phase distance



Amplitude distance

- Shape mismatch = distance along y-axis + x-axis

Amplitude distance



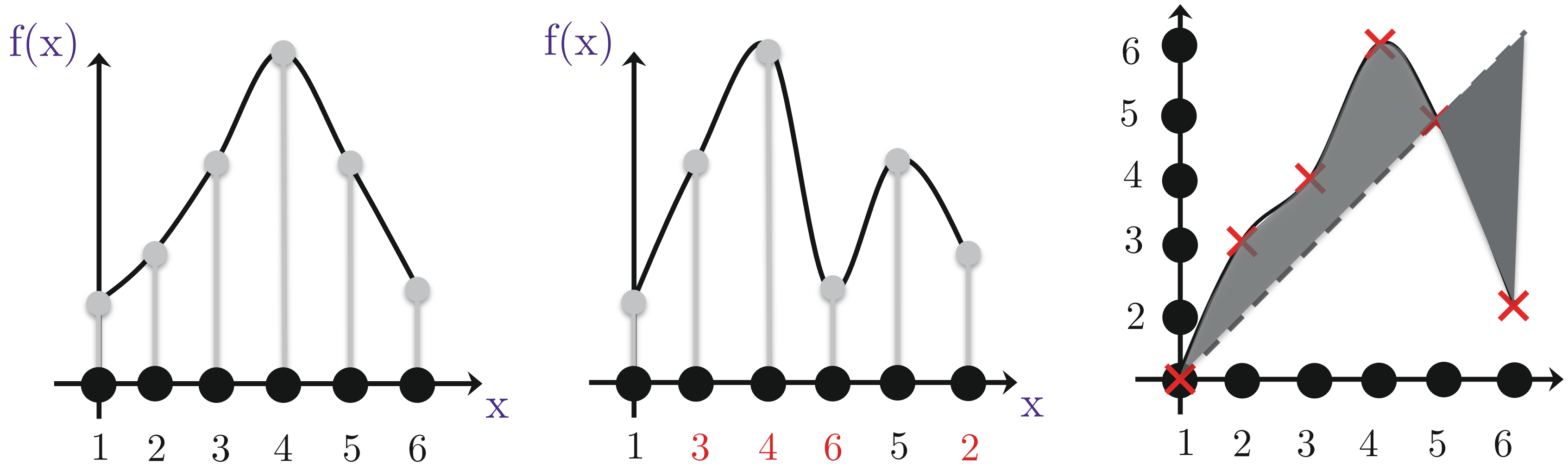
- Measures how fast the curve changes

$$q = \sqrt{\dot{f}(x)}$$

- Distance using function norm

$$d(q_1, q_2) = \int_0^1 \Big(q_1(x) - q_2(x)\Big)^2 \, \mathrm{d}x$$

Phase distance & Warping functions

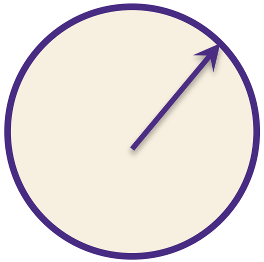


- ➢ Has identity, inverse, and a transformation – Group!

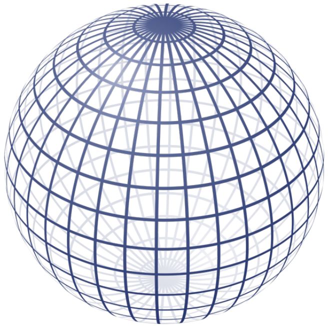
Differential geometry of Warping functions

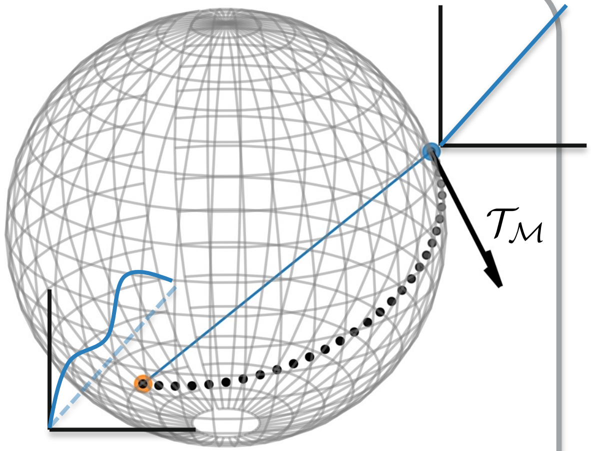
$\gamma : [0,1] \mapsto [0,1]$ with $\gamma(0) = 0$ and $\gamma(1) = 1$

$$x^2 + y^2 = 1$$



$$x^2 + y^2 + z^2 = 1$$



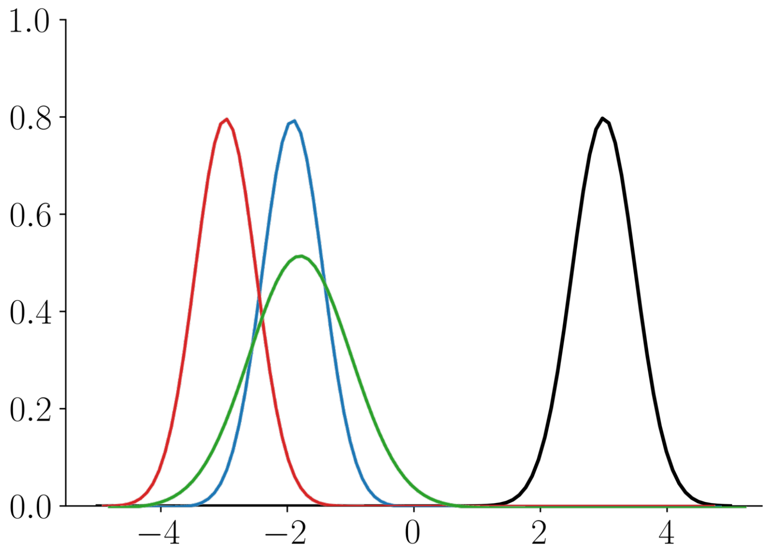


$$d_p = \cos^{-1}\left(\int_0^1 \sqrt{\dot{\gamma}(t)}\ \mathrm{d}t\right)$$

$$\|q_\gamma\|_{\mathbb{L}^2} = \int_0^1 q_\gamma^2(t)\ \mathrm{d}t = \int_0^1 \dot{\gamma}(t)\ \mathrm{d}t = \gamma(1) - \gamma(0) = 1$$

Comparing spectra – Amplitude-Phase distance

$$d(y, y^*) = d_{\text{amplitude}} + d_{\text{phase}}$$



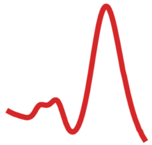
Blue : 0.21

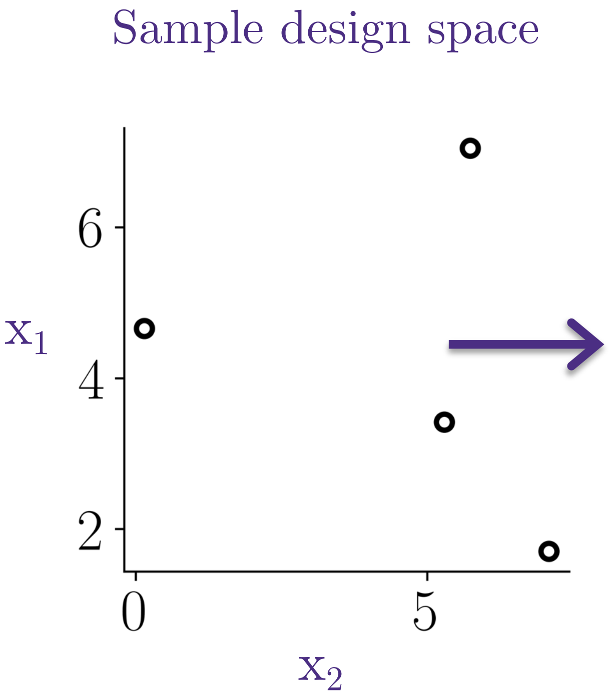
Red : 0.34

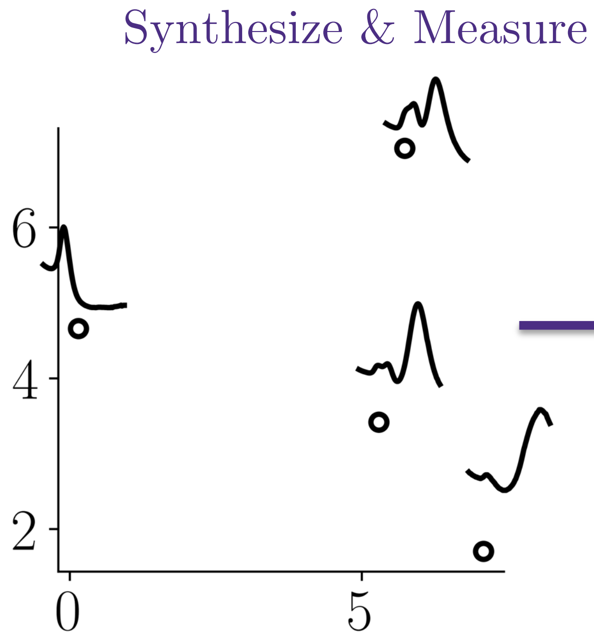
Green : 0.32

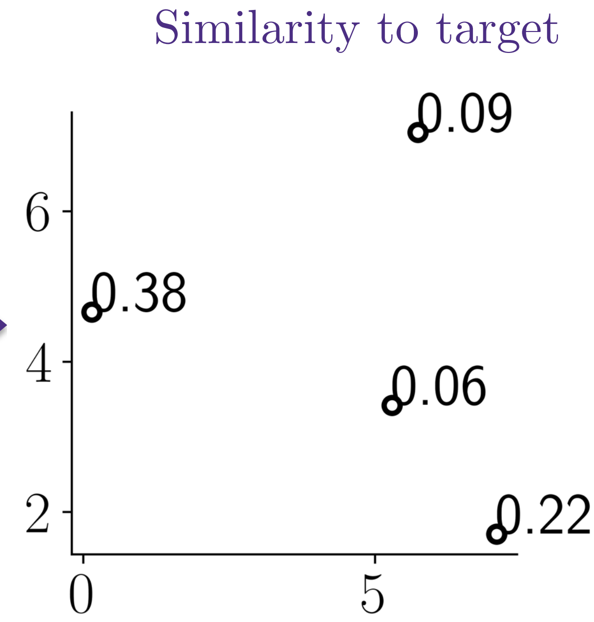
➢ AP distance clearly captures the shape based similarity

Solution : Batch synthesis + Bayesian opt.

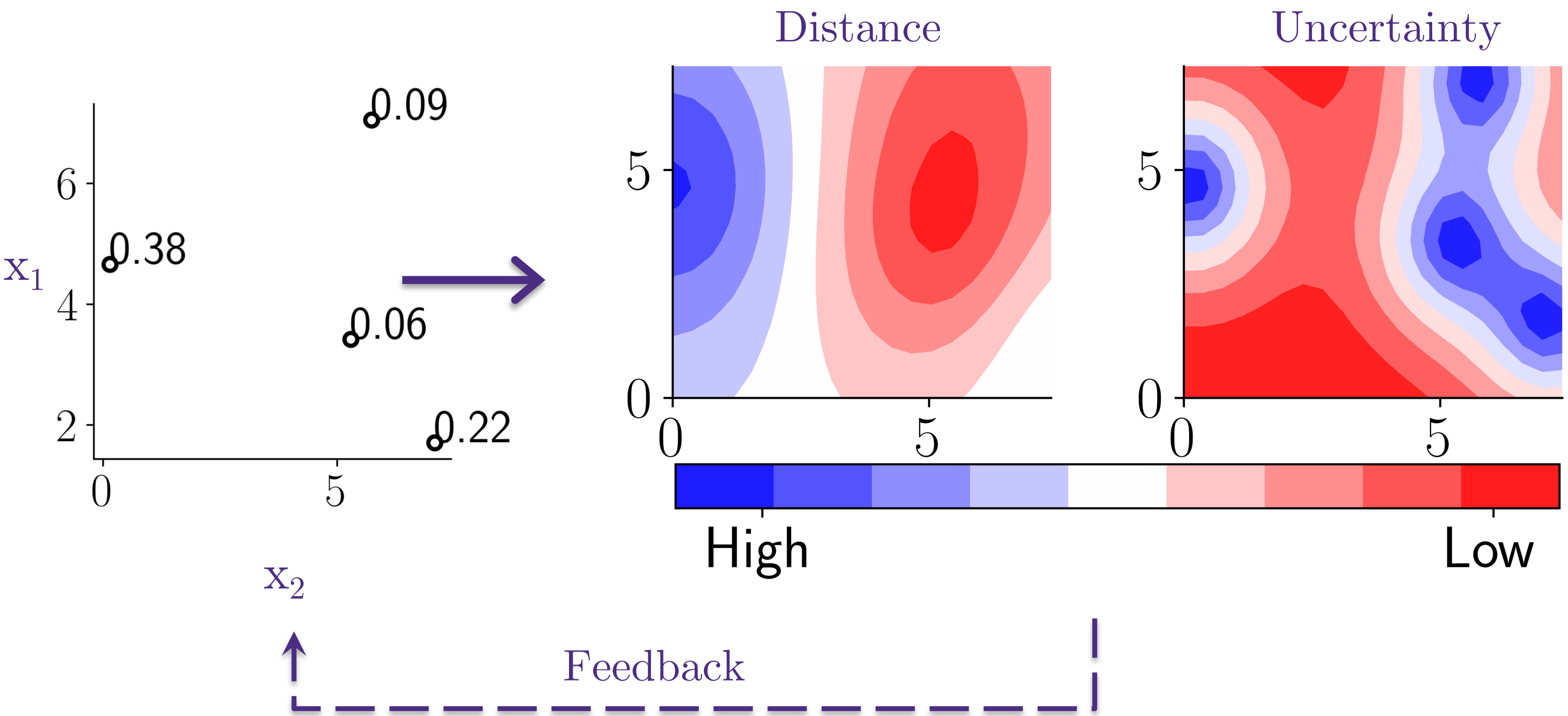








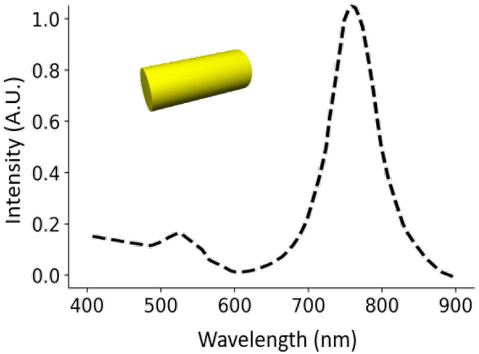
Solution : Gaussian process as surrogate

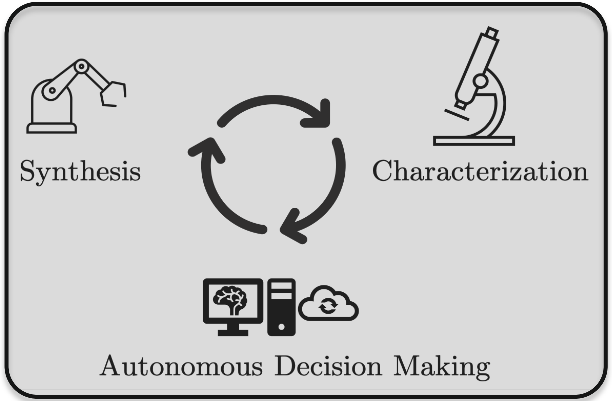


8D Optimization – Gold nanoparticles

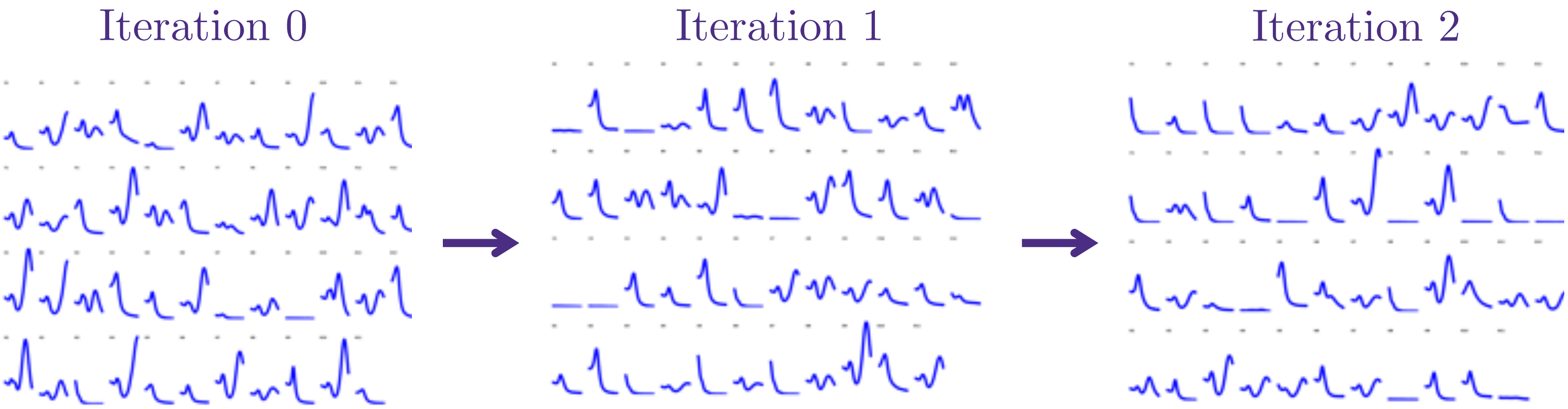

Huat Thart-Chiang

Experimental Design Space	
Reagent	Concentration Range (mM)
CTAB	0 – 75
Gold Chloride	0 – 0.15
Silver Nitrate	0 – 0.06
Ascorbic Acid	0 – 0.64
Gold Seeds	0 – 0.06
Hydrochloric Acid	0 – 14
Sodium Hydroxide	0 – 7.2
Sodium Chloride	0 – 14





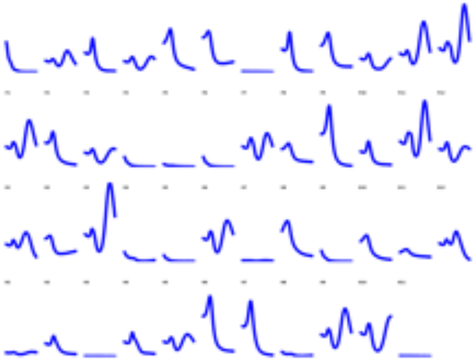
Using Euclidean distance as target



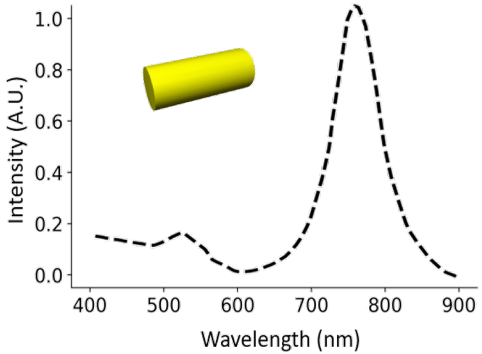
Using Euclidean distance as target

. . .

Iteration 6

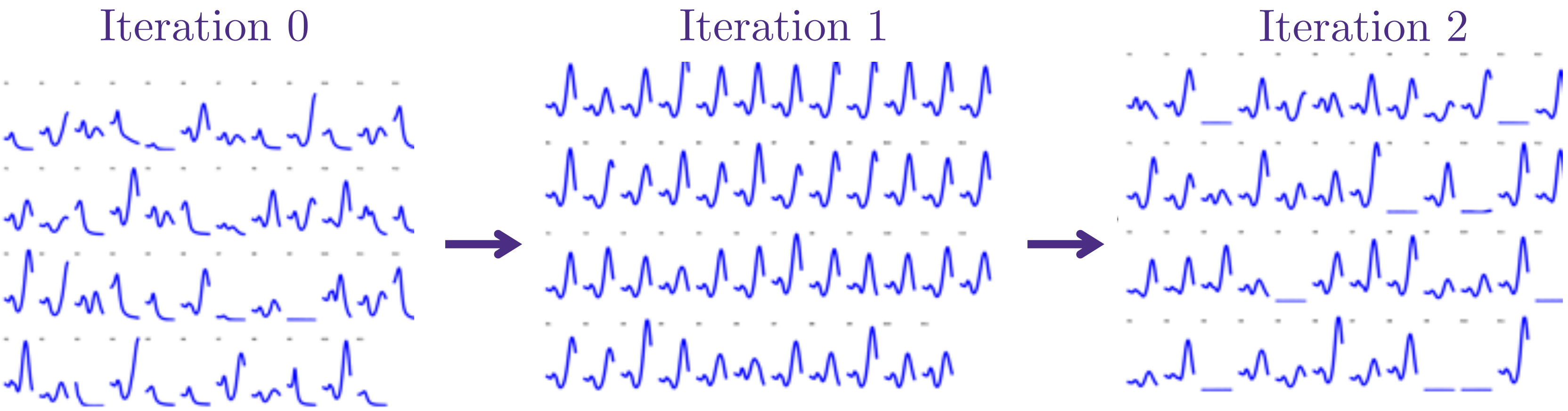


vs



➢ We make almost NONE that looks like our target

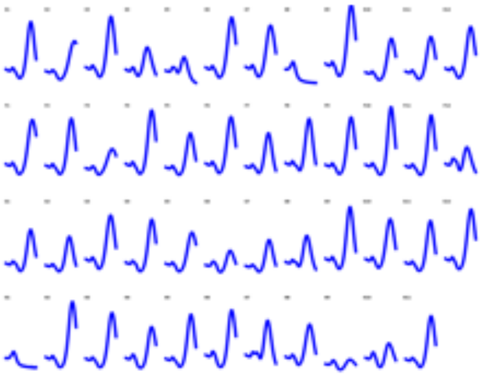
Using Amplitude-Phase distance as target



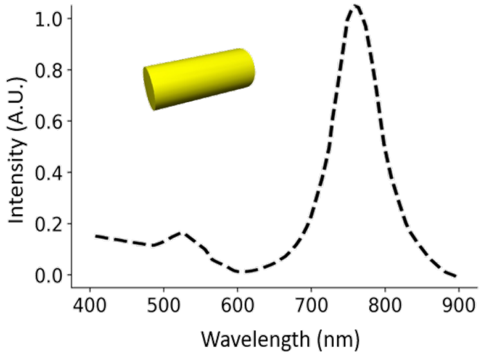
Using Amplitude-Phase distance as target

. . .

Iteration 6



vs



- We make almost EVERYTHING that looks like our target

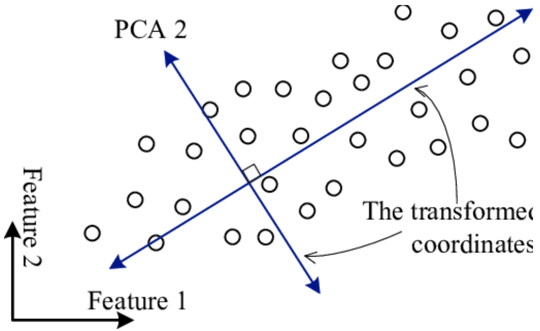
PCA for functional data

- Low-dimension projection of vectors

$$\widetilde{X}_{\text{nxd}} = VY_{\text{nxq}} \qquad \widetilde{X} = X - \mu$$

- Minimal covariance projection

$$V = \text{SVD}\left(\text{Var}(\widetilde{X})\right)$$

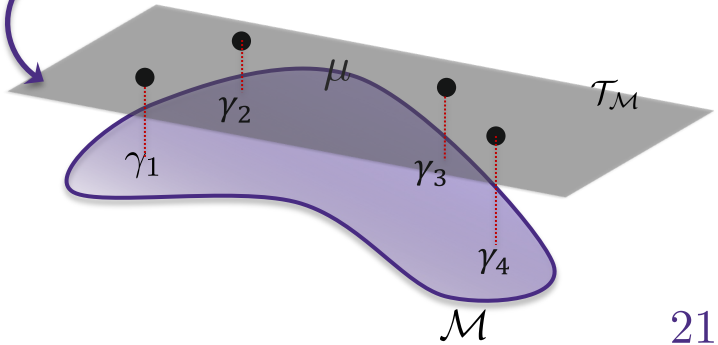


- PCA for Functional data:

$$\mu = \underset{f}{\text{argmin}} \sum_{i=1}^{n} \text{dist}(f, f_i)^2$$

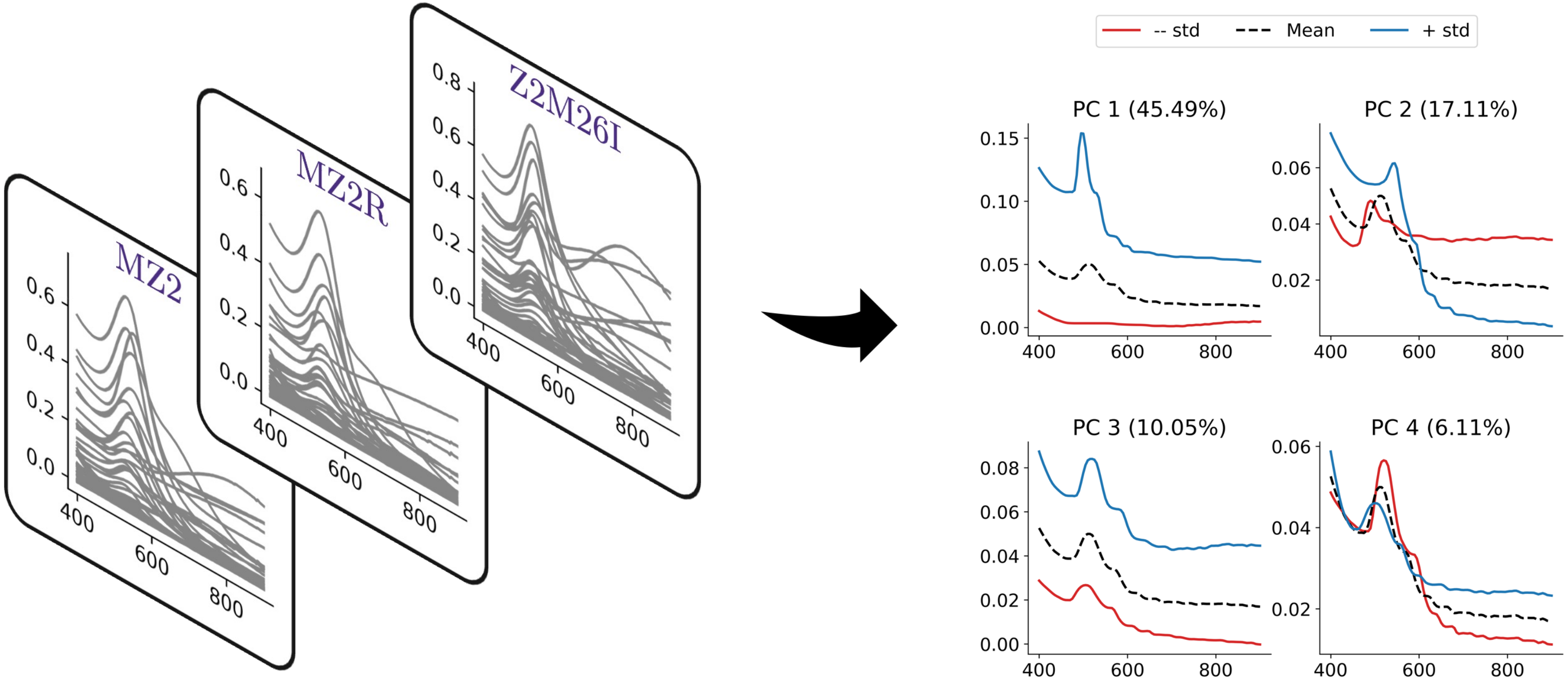
Amplitude variance: $\text{Var}(\{\tilde{f}_i\})$

Phase variance: $\text{Var}(\{\mu \circ \gamma_i\})$



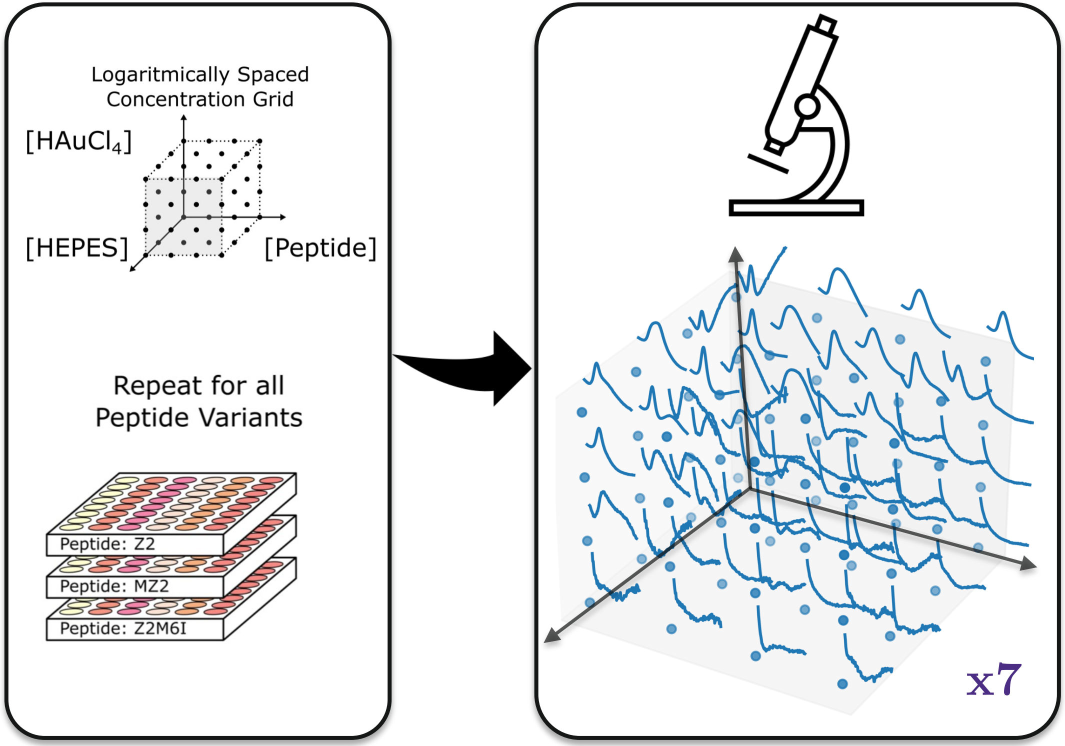
Data exploration using Functional PCA

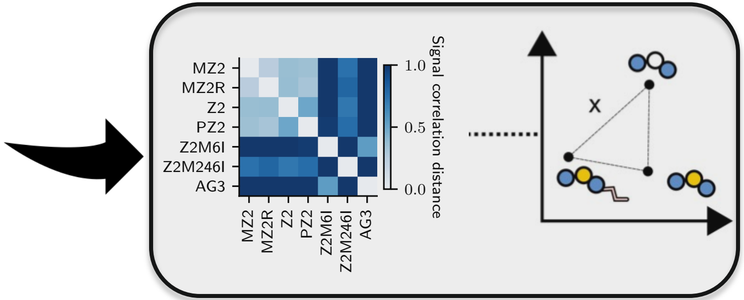
- Fitting a linear ‘generative’ model to functional data



Exploratory studies

Dr. Kacper Lachowski



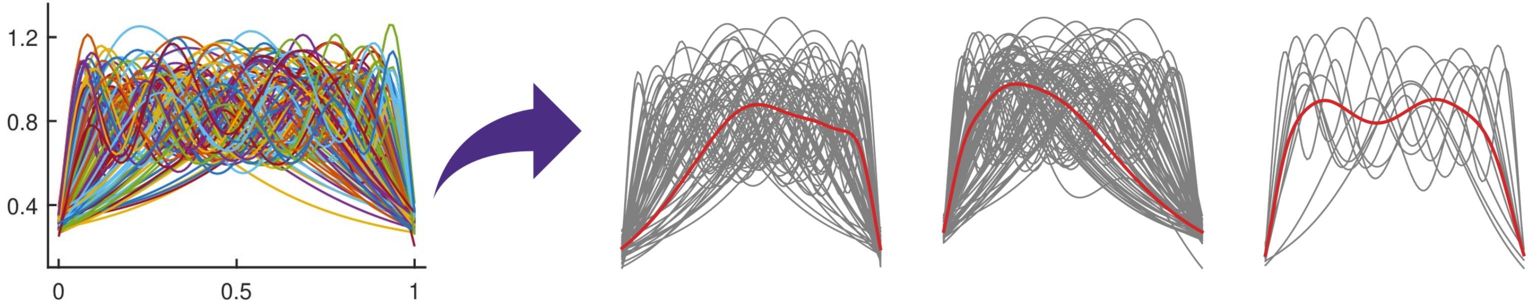


Principal Component Analysis
& Metric Dimensional Scaling

Digital Discovery 1.4 (2022): 427-439.

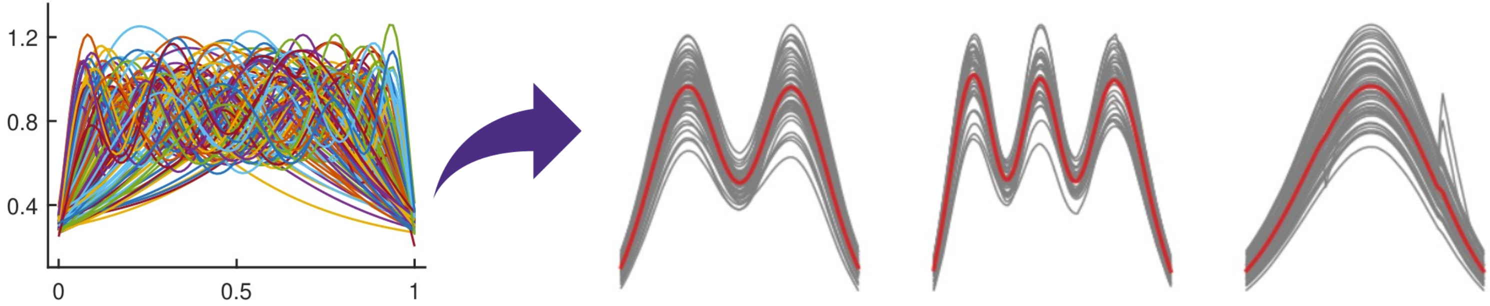
Functional k-means clustering

➢ Learn templates and assignment rules based on Euclidean distance



Functional k-means clustering

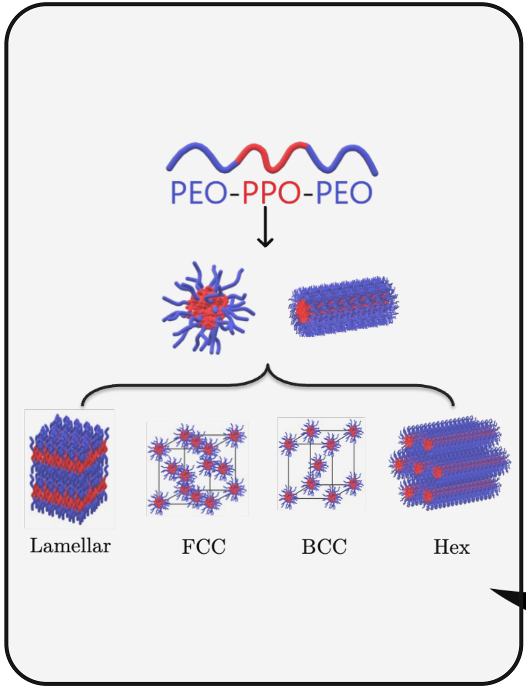
- Learn templates and assignment rules based on Shape distance

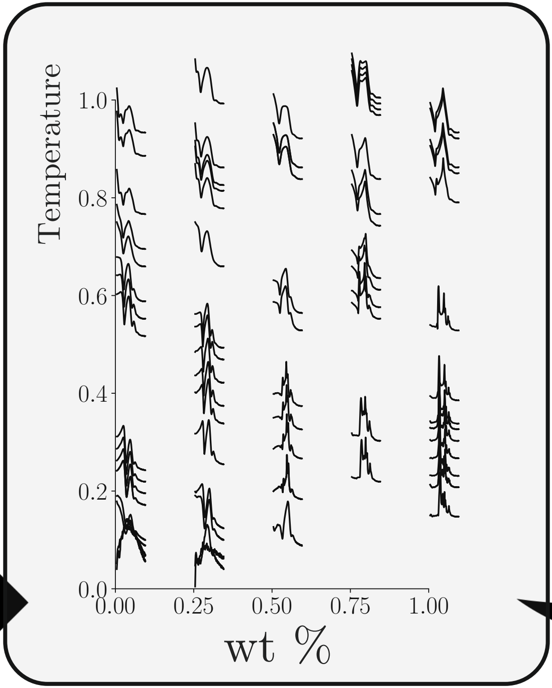


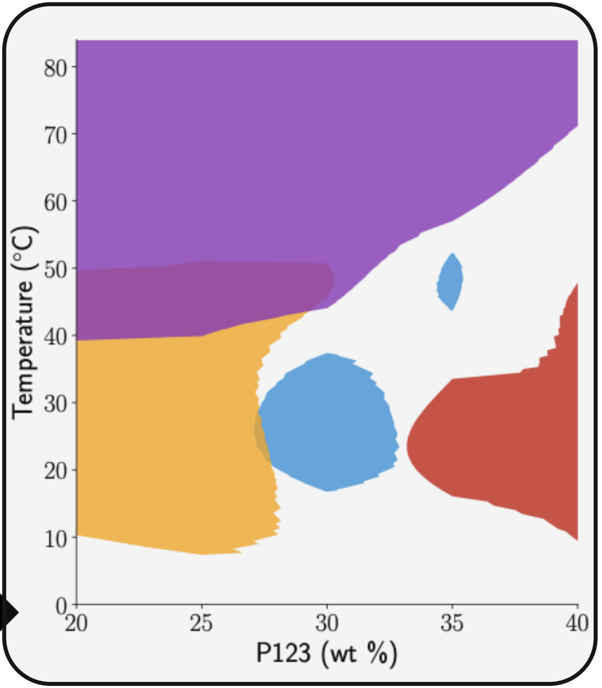
- Similarity to XRD/SAXS phase assignment: shifts via lattice expansion or broadened peaks are not a characteristic of the structure

Phase regions -- design of experiments & knowledge extraction

Karen Li





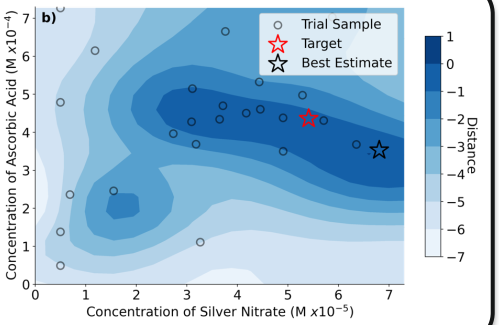


DOI: 10.1039/D3DD00105A (Paper) *Digital Discovery*, 2023, **2**, 1471-1483

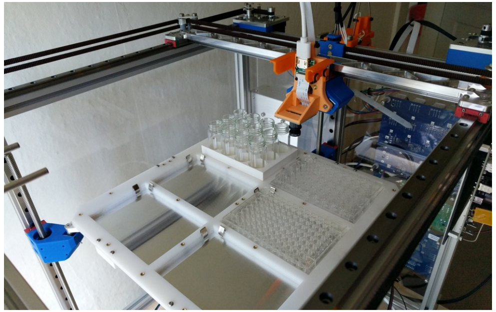
Conclusions

- High Throughput Experimentation needs **new computational tools**
- Combine **autonomous decision-making** with automation to unlock the full potential
- A careful rethink of **surrogate models** and **data representations**
- The **geometry of functions** -- encode the 'physics' into data-driven workflows

AI-Driven Materials Exploration



Robotic Automation



HTE Characterization

Thank you!

kiranvad@uw.edu kiranvad.github.io

CSSAS

U.S. DEPARTMENT OF ENERGY | Office of Science

Grant No : DE-SC0019911,
Neutron Scattering Program

UNIVERSITY of WASHINGTON
eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

pozzorg.com

Resources

DOI: 10.1039/D2DD00025C (Paper) *Digital Discovery*, 2022, **1**, 502-510

Autonomous retrosynthesis of gold nanoparticles *via* spectral shape matching†

Kiran Vaddi *[a], Huat Thart Chiang [a] and Lilo D. Pozzo *[ab]

DOI: 10.1039/D3DD00105A (Paper) *Digital Discovery*, 2023, **2**, 1471-1483

Metric geometry tools for automatic structure phase map generation†

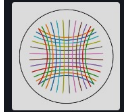
Kiran Vaddi *[a], Karen Li [b] and Lilo D. Pozzo [c]

DOI: 10.1039/D2DD00017B (Paper) *Digital Discovery*, 2022, **1**, 427-439

Multivariate analysis of peptide-driven nucleation and growth of Au nanoparticles†

Kacper J. Lachowski *[ac], Kiran Vaddi [a], Nada Y. Naser [a], François Baneyx [a] and Lilo D. Pozzo *[abc]

Geomstats



Geomstats is an open-source Python package for computations, statistics, and machine learning on nonlinear manifolds. Data from many application fields are elements of manifolds. For instance, the manifold of 3D rotations SO(3) naturally appears when performing statistical learning on articulated objects like the human spine or robotics arms. Likewise, shape spaces modeling biological shapes or other natural shapes are manifolds. Additional examples are introduced in Geomstats paper. Geomstats' source code is freely available on GitHub.

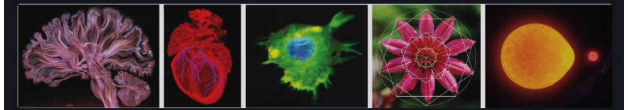
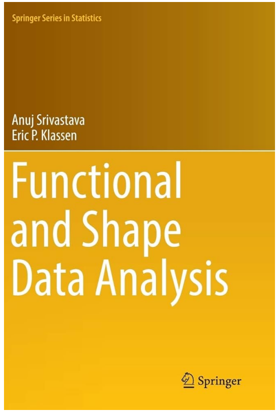
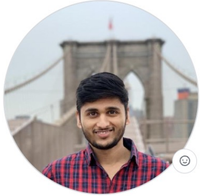


Figure: Shapes in natural sciences can be represented as data points on "manifolds". Images credits: Self Reflected, Greg Dunn Neuro Art, British Art Foundation, Ashok Prasad, Matematik Dunyasi, Gabriel Pérez.



GitHub

pozzo-research-group



Kiran Vaddi
kiranvad