# Raw Data and Data Integration in TD-NMR Studies of Food Structure: the Machines are Learning; Are We?

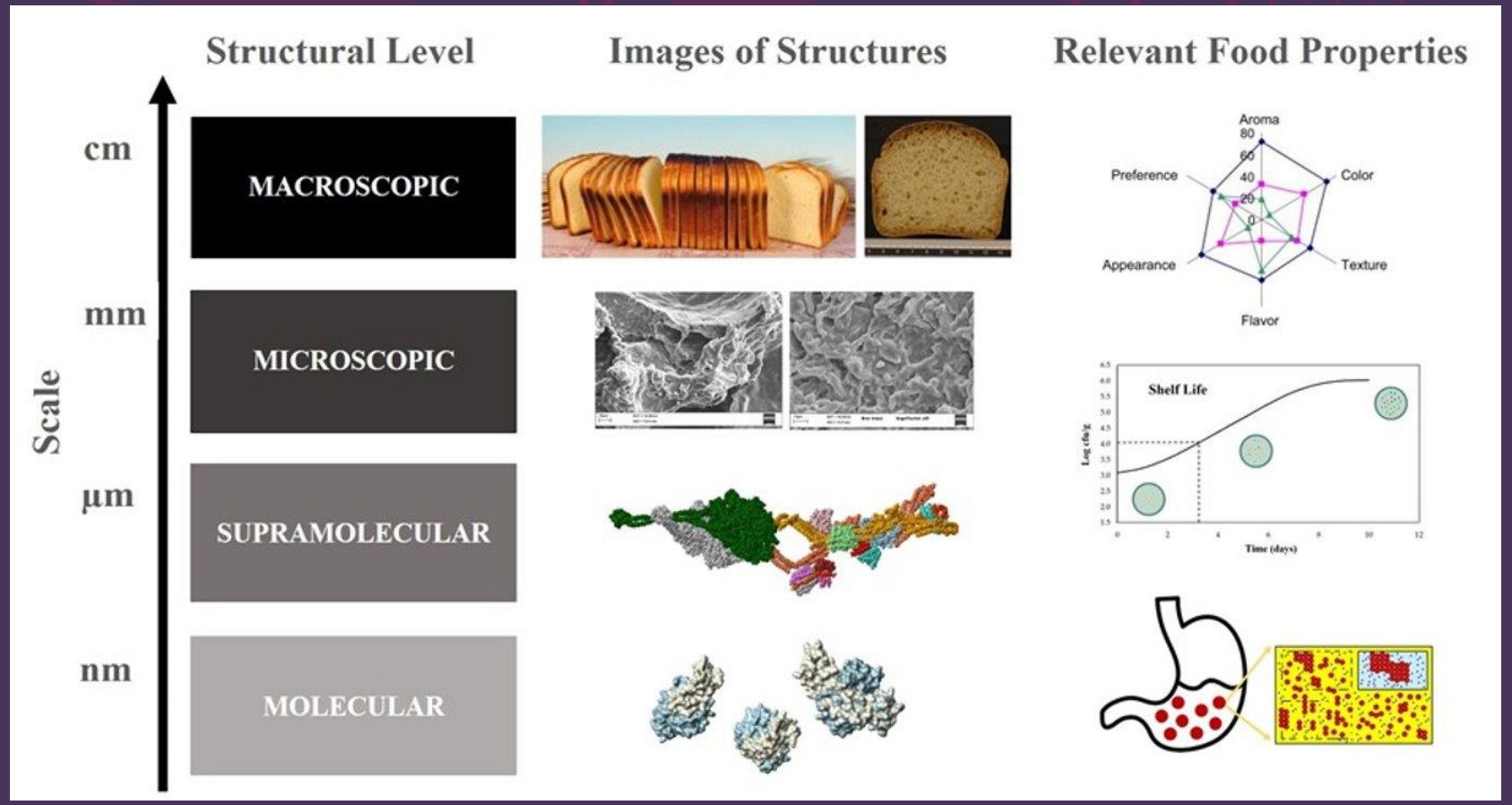
**Carlo Mengucci, PhD**

**Bio-NMR Group**

**DISTAL, UniBo**

*Workshop on Food Quality Analysis with Time Domain NMR and AI-Driven Mathematical Models*. PON Seminars, Bologna, 21 April 2022

# Studying food structure




## Food Matrix

**Defined by structures at different lenght scales**

**Properties are determined by structural elements at micro-scale**

**Can we link structures with food properties?**

**Figure:** structures in the food matrix. *Adapted from Mengucci et al., 2022, TIFS, Elsevier*

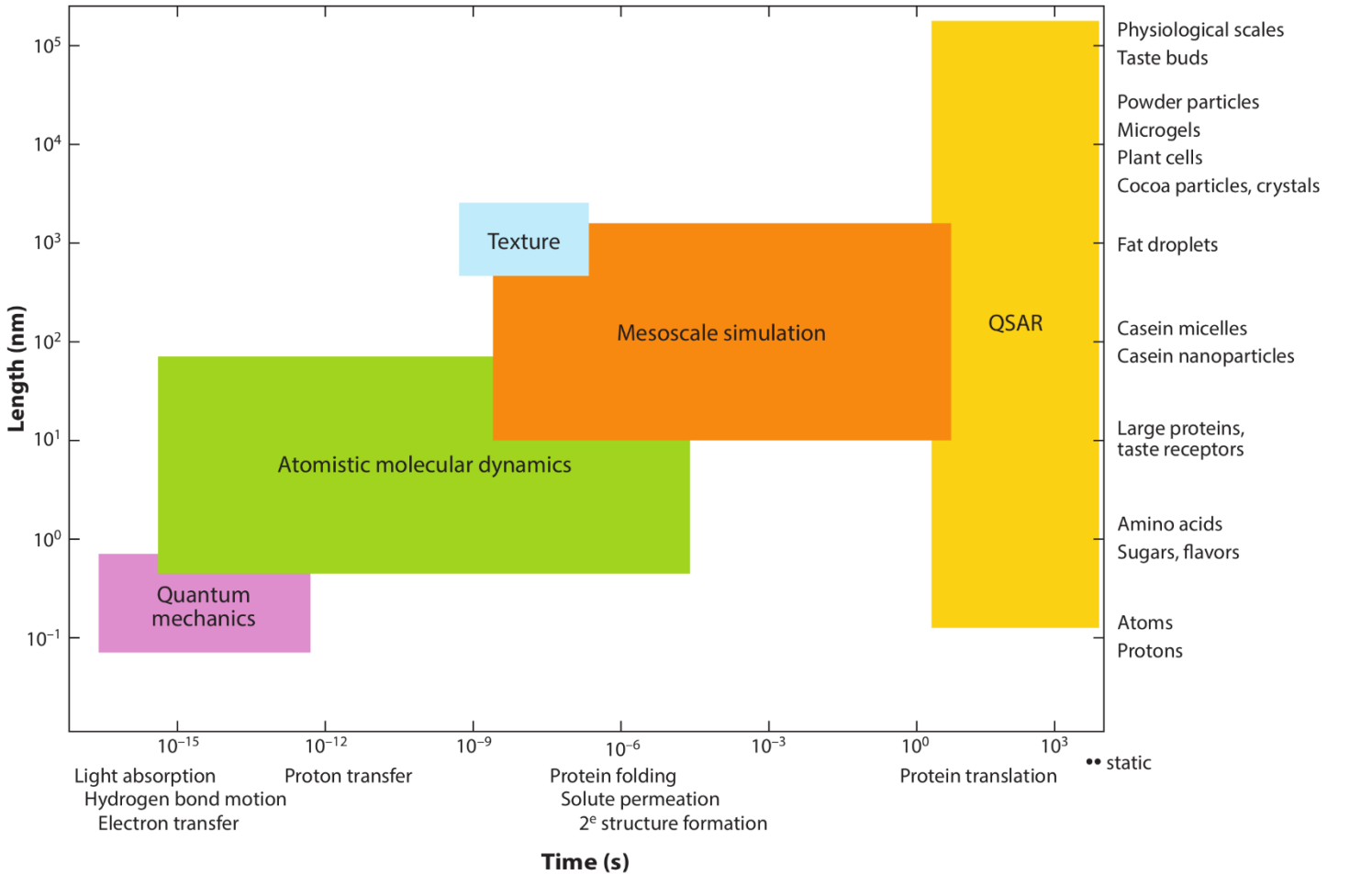
# Studying food structure

- Many length scales to investigate
- Many different techniques: some high-throughput, some not so high-throughput
- Data integration is required to determine properties affected by structures

| SCALE LENGTH | METHODS | PHYSICAL STATE/ STRUCTURAL ELEMENTS | INFORMATION ON: |
|---|---|---|---|
| >1 cm | • Texture analysis<br>• Image analysis<br>• Sensory panel | liquid, gel, solid, porous solid | - properties of network at large deformation<br>- size and shape macrostructural elements<br>- sensorial attributes (e.g., appearance, colour, firmness, overall acceptability) |
| 1 mm–1 cm | • Texture analysis<br>• Microscopy | liquid -aqueous matrix (*aqueous phase in fruit juices*), liquid -emulsion matrix (*mayonnaise*), gels (*desserts, processed meats*), porous matrix (*bread, extruded snacks*), viscoelastic matrix (*dough*), etc. | - properties of network at large deformation related to eating properties<br>- microstructure |
| 1–500 µm | • Confocal microscopy<br>• Light microscopy<br>• Rheology | micelles (*casein micelles*), droplets, air cells (*bread bubbles*), crystals (*salt*), fibres, granules (*starch granules*), etc. | - size and shape of structures<br>- properties of network at small deformation<br>- ingredient interaction |
| 10–500 nm | • Light scattering<br>• Electron microscopy | micelles (*casein micelles*), droplets, air cells (*bread bubbles*), crystals (*salt*), fibres, granules (*starch granules*), etc. | - aggregation, density, arrangement<br>- size of structures |
| <10 nm | • Raman<br>• Chromatography<br>• Thermal analysis<br>• SDS Page<br>• NIR | carbohydrates (*starch*), proteins (*gluten, caseins*), lipids, water, etc. | - molecular structure<br>- proportion of elementary parts<br>- unfolding vs. native<br>- denaturation/ transition temperature |

**Table.** Principal methods for structural analyses at characteristic length scales in foods, appearance of food matrix and structural elements. *Adapted from Mengucci et al., 2022, TIFS, Elsevier*

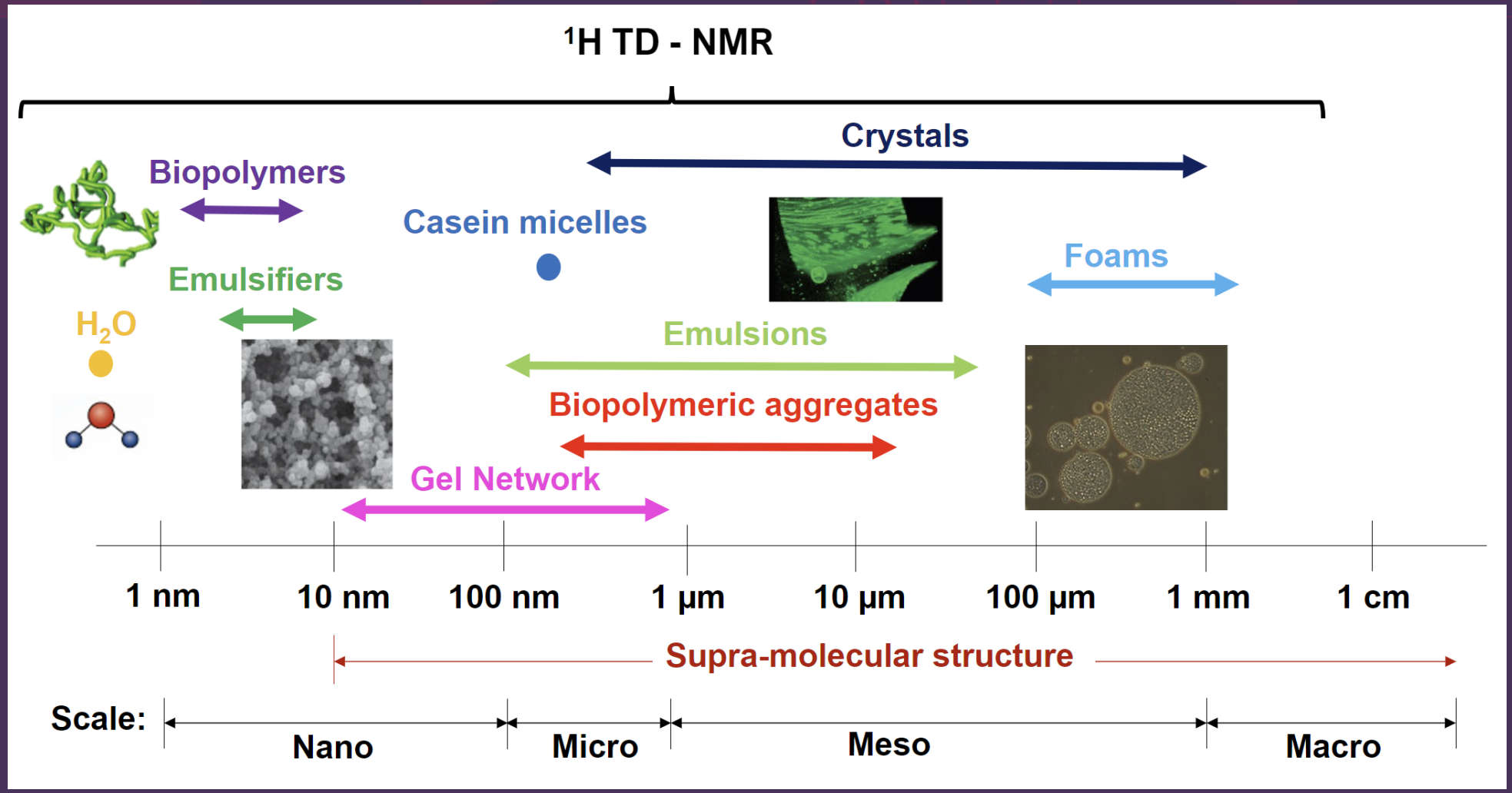
# Full scale simulation of food structure: a dream or a possibility?



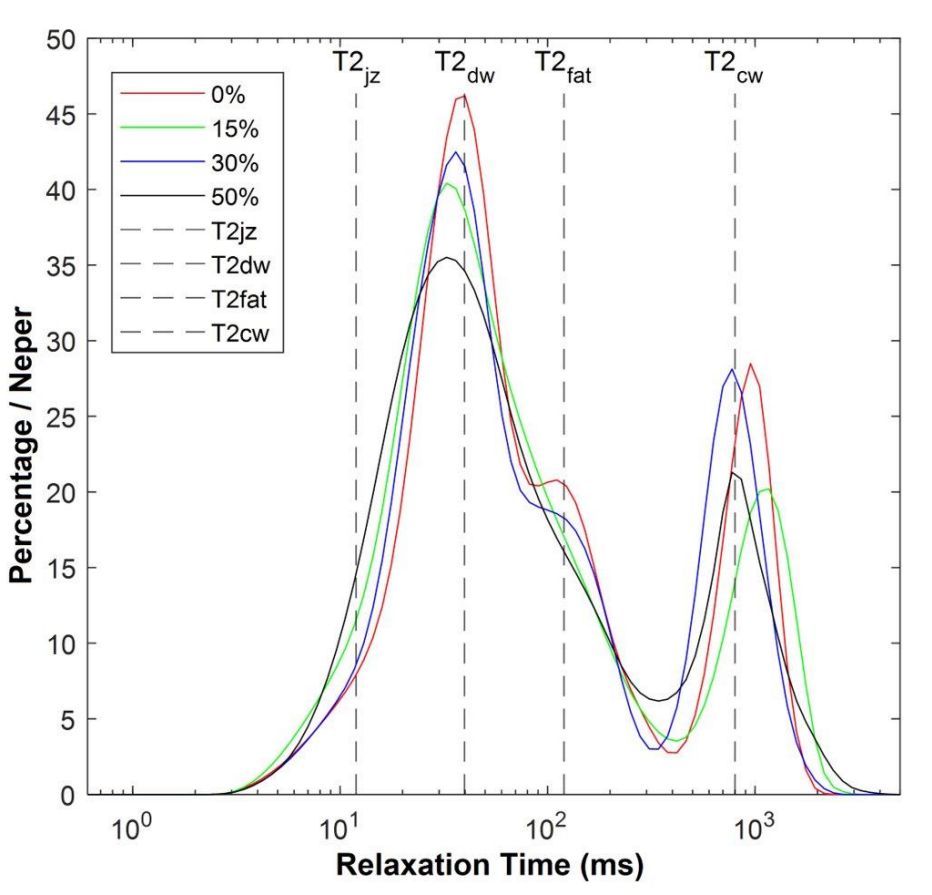

- **Different structures at different lenght scales**
- **Different phenomena at different time scales**
- **Different descriptive paradigms and methods**

**Figure**: Schematics of molecular interactions in food science phenomena across different time and length scales, with appropriate particle-based simulation methods. QSAR: quality structure–activity relationships . *Adapted from da Silva et al. 2020, "Understanding and controlling food protein structure and function in foods: Perspectives from experiments and computer simulations"*

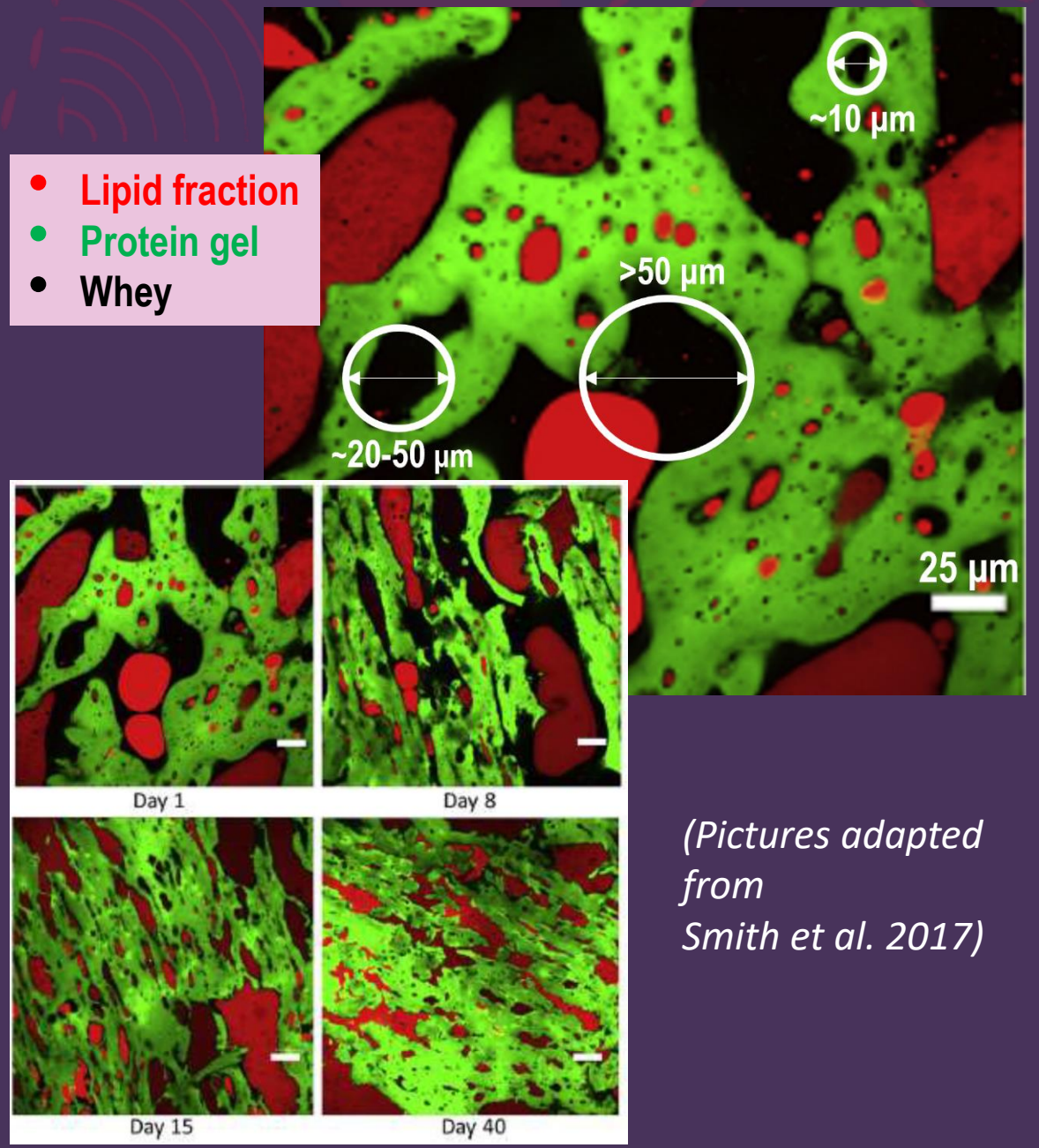
# The TD-NMR: a fast, cost-efficient base for description



# The TD-NMR: a fast, cost-efficient base for description

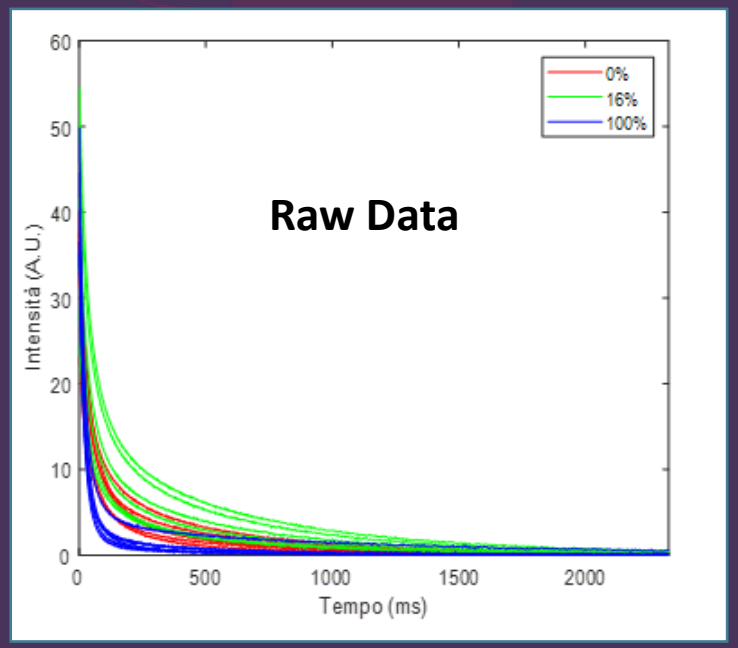



**(jz)** 1H of water molecules in fast chemical exchange, **(dw)** mainly water in compartments of size around 10 µm, minor contribution of the liquid lipid fraction, **(fat)** mainly liquid lipid fraction, minor contribution of water in compartments of dimension around 30 and 50 µm, **(cw)** water retained in compartments larger than 50 µm.

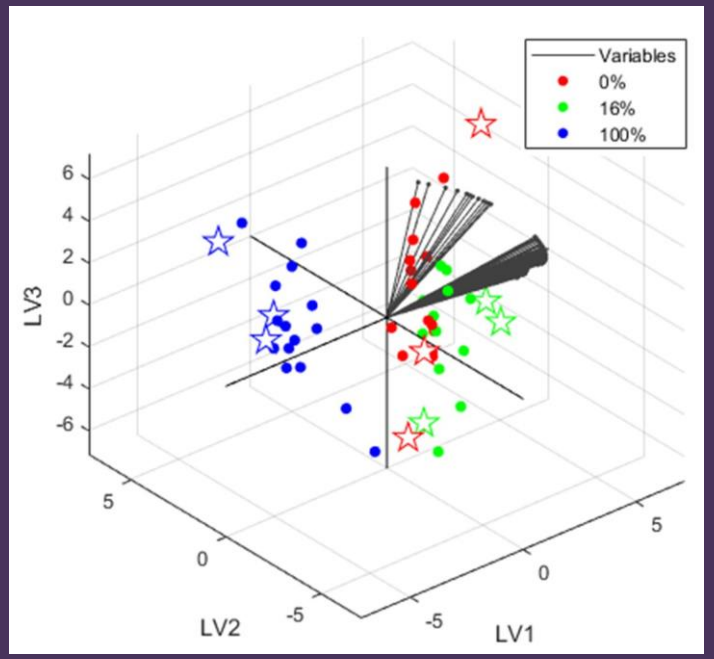



*(Pictures adapted from Smith et al. 2017)*

# The TD-NMR: high throughput data, machine learning and classification

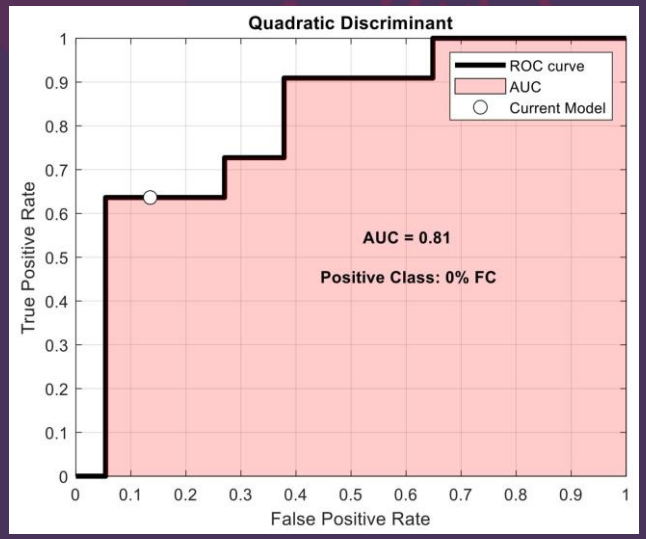



SVD Methods, Denoising...




Latent Space, Dimensionally Reduced Problem

Feature Selection, Feature Agglomeration, Optimization, Crossvalidation




Classification, Prediction: Food Fraud Detection

*Figures adapted from Mengucci et al., 2021, ApplSci, MDPI*

# The TD-NMR in the high throughput data era: the description and classification dilemma

**Description (e.g. Through Inversion of Multiexponential Decay Data)**

- Immediately gives info on structure through the interpretation of proton populations
- May be affected by parameters of smoothing and user-based choices
- May require a-priori knowledge on the observed system to deal with processing generated artifacts

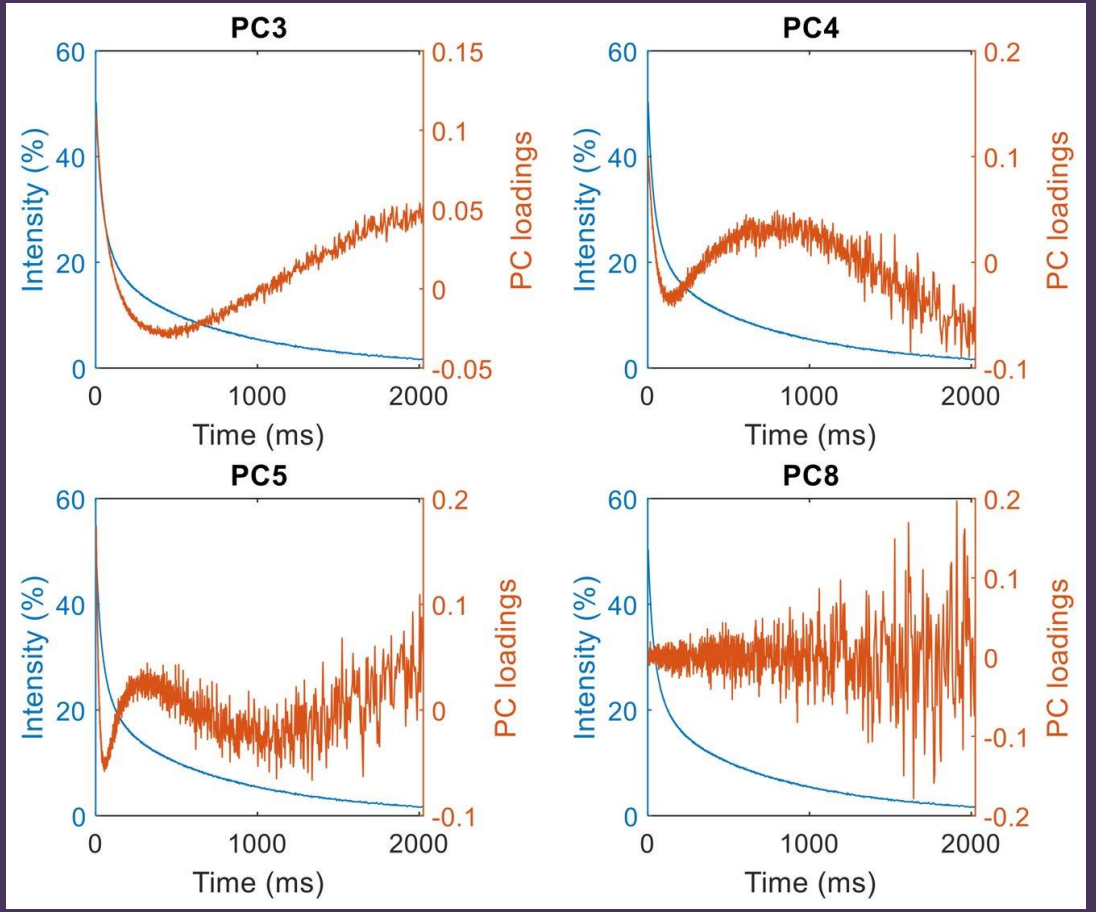
**Classification (DR + Classification Problem solved in lower-dimensional space)**

- Very general representation of data, self-optimizing methods
- Fast data acquisition. With the right amount of training data and a bit of know-how, we can solve real life problems (e.g. food fraud detection)
- Not so easy to interpret when considering each point of the decay as a feature

# The TD-NMR in the high throughput data era: the description and classification dilemma

**Classification (DR + Classification**
**Problem solved in lower-dimensional space)**

- Very general representation of data, self-optimizing methods
- Fast data acquisition. With the right amount of training data and a bit of know-how, we can solve real life problems (e.g. food fraud detection)
- Not so easy to interpret when considering each point of the decay as a feature



# Data Integration, RAW data, General Descriptors

- How can we get the best from the two worlds (description, classification)?

- Can we move forward and use ML to extract parameters from general representations of the data, for description?

# Beyond Classification, Toward Description: Dimensionality Reduction and Texture Analysis

- **Imaging:** straightforward way of parametrizing food structure. (**Problem:** not all imaging is high-throughput and suitable for general parameter extraction, see table)
- **TD-NMR:** efficient way (from many perspectives) to study changes in location and mobility of water in food matrices

**Table.** Main descriptors and (dis)advantages for electronic microscopy and magnetic resonance imaging. *Adapted from Mengucci et al., 2022, TIFS, Elsevier*

| | SEM | MRI |
|---|---|---|
| **Descriptors** | • Particle size and morphology<br>• Pore size and morphology<br>• Size distribution and morphology<br>• Shape orientation (e.g., fibres) and diameter distribution (e.g., beads) | • First order grey level statistics (e.g., Histogram of grey levels statistics, symmetry of grey levels centred about the mean, entropy of the image)<br>• Roughness of textures<br>• Degree of linearity<br>• Co-occurrence matrix statistics (e.g., Haralick moments)<br>• Structural or morphological features of ROIs (e.g., Bounding ellipsoid volume ratios)<br>• Transform features (features extracted in frequency domains) |
| **Pros & Cons** | • Not immediately suitable for high-throughput production (parameter dependent acquisitions: lighting, magnification etc.)<br>• No data harmonization standard due to heterogeneous necessities of application fields and experiments<br>• Widely applied in many fields<br>• Canonical descriptors immediately linkable with physical quantities<br>• Very high resolution<br>• Requires specific assumptions for image analysis (i.e., presence/absence of certain geometrical structures, pores, shapes etc.) | • Inherently suitable for high-throughput data production<br>• Data harmonization standards are widely supported in many biomedical fields (neuro imaging, imaging for oncology)<br>• Descriptors comes from low-level, general texture analysis and morphological studies alike<br>• Low resolution<br>• Does not require specific assumptions for image analysis, due to canonical analysis based upon general first order statistics of grey levels and moments of cooccurrence matrix. |

# Beyond Classification, Toward Description: Dimensionality Reduction and Texture Analysis

- **Aim:** integrate descriptors from high-resolution imaging techniques (like SEM) and TD-NMR high-throughput experiments, to enhance automation AND interpretation

- **Strategy:** define minimal SEM images processing and segmentation (texture parameters extraction), train a DR on raw TD-NMR decays capable of describing food matrix transformation

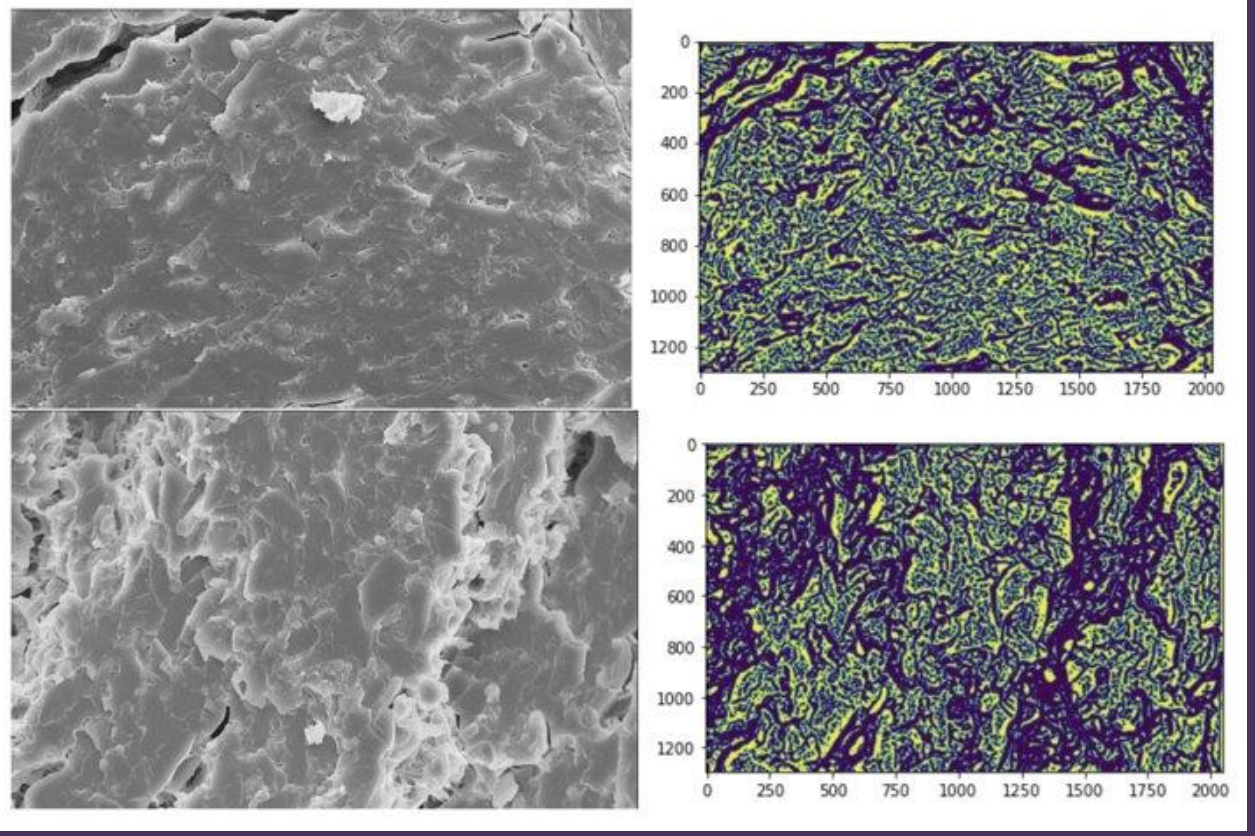
# A study on pasta: cooking and water-matrix interaction characterization in a ML framework

**Semola Spaghetti**:

- SEM images of different zones, acquired with a set of minimal acquisition specifics (i.e., zoom, lighting, well defined morphological regions of the pasta to acquire) at different cooking times

- T2 decays acquired at different cooking times

# A study on pasta: from images to descriptors




Images at different cooking times. Top: 1 min, Bottom: 10 mins

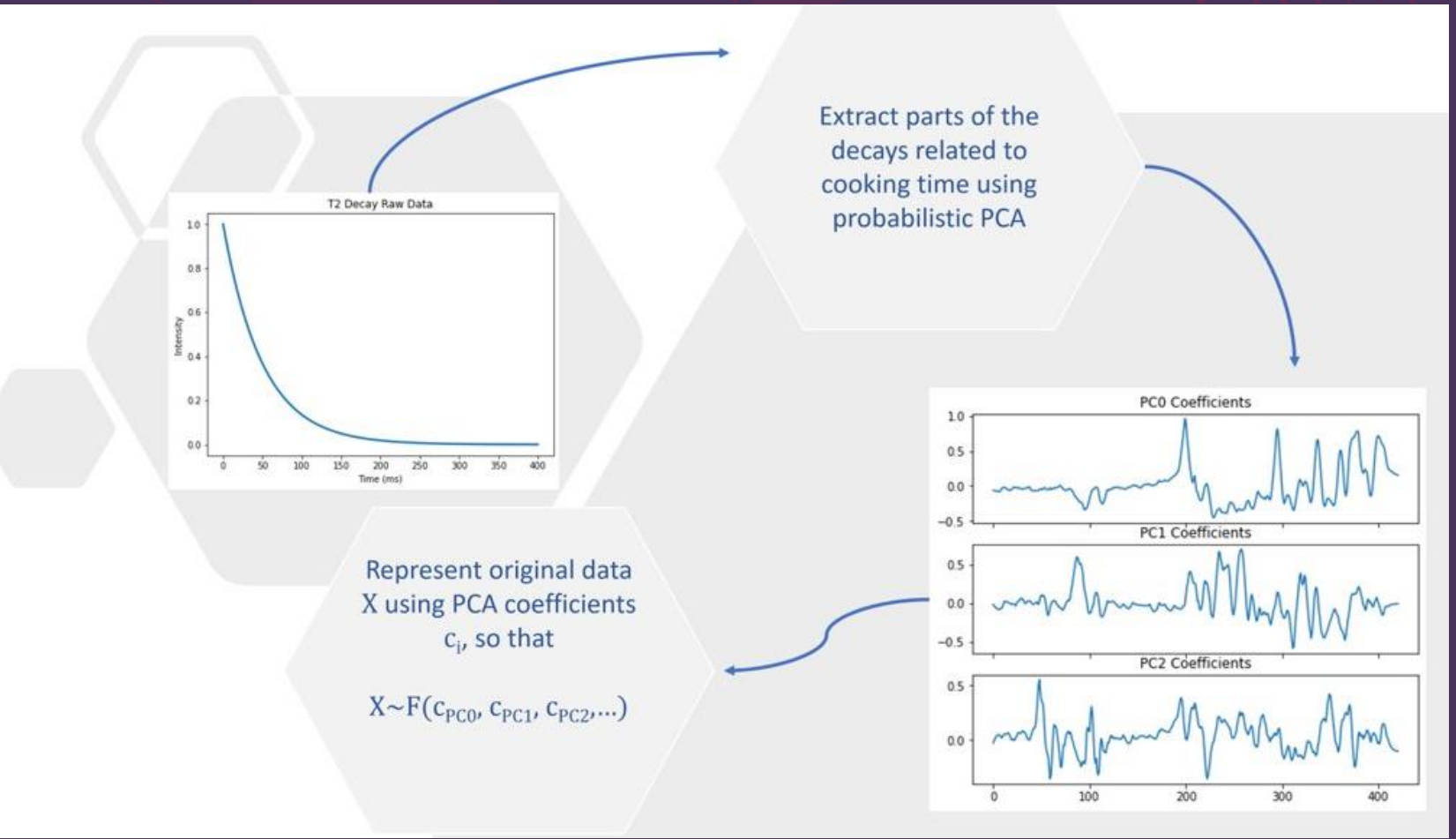
Thresholded and labelled images

## Co-occurrence matrix computation+ Haralick Moments

$$\mathbf{G} = \begin{bmatrix} p(1,1) & p(1,2) & \cdots & p(1,N_g) \\ p(2,1) & p(2,2) & \cdots & p(2,N_g) \\ \vdots & \vdots & \ddots & \vdots \\ p(N_g,1) & p(N_g,2) & \cdots & p(N_g,N_g) \end{bmatrix}$$

| | |
|---|---|
| Angular Second Moment | $\sum_i \sum_j p(i,j)^2$ |
| Contrast | $\sum_{n=0}^{N_g-1} n^2 \{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)\}, \lvert i-j \rvert = n$ |
| Correlation | $\frac{\sum_i \sum_j (ij)p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$ where $\mu_x$ , $\mu_y$ , $\sigma_x$ , and $\sigma_y$ are the means and std. deviations of $p_x$ and $p_y$ , the partial probability density functions |
| Sum of Squares: Variance | $\sum_i \sum_j (i-\mu)^2 p(i,j)$ |
| Inverse Difference Moment | $\sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j)$ |
| Sum Average | $\sum_{i=2}^{2N_g} i p_{x+y}(i)$ where $x$ and $y$ are the coordinates (row and column) of an entry in the co-occurrence matrix, and $p_{x+y}(i)$ is the probability of co-occurrence matrix coordinates summing to $x+y$ |
| Sum Variance | $\sum_{i=2}^{2N_g} (i - f_8)^2 p_{x+y}(i)$ |
| Sum Entropy | $-\sum_{i=2}^{2N_g} p_{x+y}(i) \log\{p_{x+y}(i)\} = f_8$ |
| Entropy | $-\sum_i \sum_j p(i,j) log(p(i,j))$ |
| Difference Variance | $\sum_{i=0}^{N_g-1} i^2 p_{x-y}(i)$ |
| Difference Entropy | $-\sum_{i=0}^{N_g-1} p_{x-y}(i) \log\{p_{x-y}(i)\}$ |
| Info. Measure of Correlation 1 | $\frac{HXY - HXY1}{\max\{HX,HY\}}$ |
| Info. Measure of Correlation 2 | $(1-\exp[-2(HXY2 - HXY)])^{\frac{1}{2}}$ where $HXY = -\sum_i \sum_j p(i,j) \log(p(i,j))$ , $HX$ , $HY$ are the entropies of $p_x$ and $p_y$ , $HXY1 = -\sum_i \sum_j p(i,j) \log\{p_x(i)p_y(j)\}$ $HXY2 = -\sum_i \sum_j p_x(i)p_y(j) \log\{p_x(i)p_y(j)\}$ |

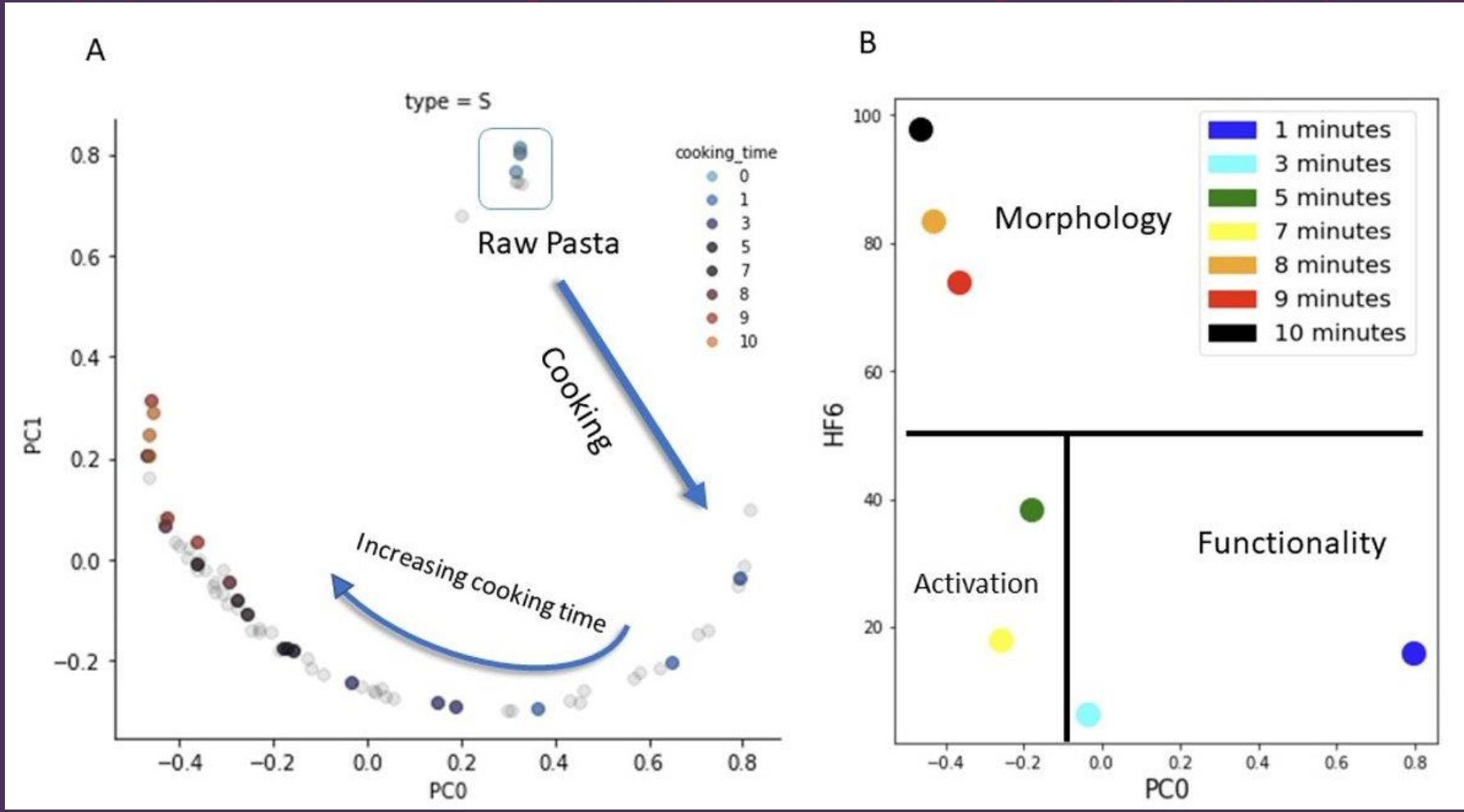
# A study on pasta: TD-NMR + dimensionality reduction of raw data of decays




**Train KPCA to find best separability of cooking times:** tune kernel and kernel hyperparameters

**Figure:** The process behind the decomposition of T2 decays raw data into a lower dimensional space. Each time point of each decay is interpreted as variable and fed to a kernel PCA with an RBF kernel. Data are transformed according to coefficients which are dependent on the kernel parameters, optimized through machine learning. *Adapted from Mengucci et al., 2022, TIFS, Elsevier*

# A study on pasta: Results




**Investigation of water mobility/dynamics and structures.** Qualitative interpretation of the relationship between PC0 (TD-NMR) and HF6 (Imaging) : in the functionality phase, water mobility is related to starch gelatinization and little morphological changes. After an activation phase, with the rupture of structures in the food matrix, morphological changes detected in images follow a strictly monotonous trend related to cooking time (morphology).

**Figure**. Resulting lower dimensional latent space (left). Scatter of PC0 VS the sixth Haralick moment (HF6, right). *Adapted from Mengucci et al., 2022, TIFS, Elsevier*

# Conclusions: Intertwining Description and Classification in the Era of High-Throughput Data

- Framework is useful to study properties related to changes in water mobility and their links with structures of food matrices, during their transformation
- Framework is useful to extract general descriptors for imaging techniques that normally require assumptions (e.g. morphology of elements in SEM images)
- Framework is useful to integrate data through general parameters and representations and interpret them in a lower dimensional space (and create suitable inputs for ML routines)

# Future Developments

- Preliminary stage results are promising, to be experimented with other imaging techniques
- Work in progress: find a smart way to get information on meaningful zones of the decay curves, through kernel matrix eigenvalues and inverse transform kernel coefficients (different from classic PCA!)
- Study relationships of these descriptors with inversion of multiexponential decay modelling parameters

# Take Home Messages

- Classification doesn't always mean interpretation (even with shallow methods)
- Data integration requires automation in extraction AND generality of parameters
- Machines always learn: it is our responsibility to be good teachers and learn along