**Introduction**

The alternative dairy market faces significant challenges in replicating the complex functionality of traditional dairy products, particularly cheese. Precision fermentation of caseins offers a promising solution, but the unique properties of these proteins and their interactions within food matrices require further investigation. This project aims to bridge the knowledge gap by developing machine learning models that predict the functionality of protein mixtures, combining precision fermentation-derived caseins with plant proteins. Through the use of data science, product development cam be streamlined, costs reduced, and the goal of creating of high-quality, sustainable dairy alternatives achieved in an accelerated timeframe. This research will explore key variables such as protein structure, matrix interactions, and processing conditions, utilizing advanced machine learning algorithms to model complex food matrices and predict sensory and textural properties.

Platforms such as AlphaFold and Shiru can provide information on the structure and functionality of proteins which can be used to streamline the experimental process by finding candidate proteins much faster than would be possible experimentally. The complex nature of food matrices is well-suited to analysis using machine learning (ML), given that the large number of variables and the complexity of their interactions is very challenging to parse through experimental and traditional data analysis techniques alone. Another issue that is hindering the progress of ML for alternative proteins is the lack of academic research in this area. A great deal of the cutting-edge research is being performed by companies who do not wish to share their methods or findings. By investigating this topic, a significant contribution to the area could be made.

There are many unanswered questions and topics which are still underexplored in the field of precision fermentation for alternative dairy. While caseins are a relatively well-studied group of proteins, there are still unknowns and the particular characteristics of caseins produced via precision fermentation (notably as being nonmicellar), require further study.

**Research questions:**

**I. Protein Functionality and Structure-Function Relationships:**

1. **Fundamental Functionality:**

* What specific functionalities (e.g., gelation, emulsification, meltability) are essential for replicating the desired characteristics of targeted dairy products using precision fermentation-derived caseins and plant proteins?

1. **Casein Behavior:**

* How do precision fermentation-derived caseins, particularly in their non-micellar form, contribute to these functionalities, and what are the key differences in their interactions compared to conventional caseins?

1. **Micellar Structure Significance:**

* How does the micellar structure of caseins influence the functionality of dairy products, and how do interactions (hydrogen bonding, mineral interactions, hydrophobic interactions etc) affect the micelle?

1. **Plant Protein Synergy:**

* How can specific aspects of plant proteins be used to enhance their functionality in combination with precision fermentation-derived caseins, and are there non-protein components (e.g. polyphenols) that can effectively interact with caseins or calcium phosphate nanoclusters?

1. **Processing Impact:**

* How do different processing methods (e.g., chymosin, acid, enzymatic cross-linking, fractionation of plant protein) affect the functionality of protein mixtures and the final product characteristics?

1. **Individual Casein Roles:**

* What are the individual contributions of different casein types to specific functionalities in alternative dairy products?

**II. Modelling Protein Functionality with Machine Learning:**

1. **Current Modelling Limitations:**

* What are the limitations and knowledge gaps in current methods used to model protein functionality in food matrices, particularly in the context of alternative dairy?

1. **Machine Learning Applications:**

* How can machine learning algorithms be effectively employed to predict protein functionality in complex food matrices, and what are the optimal feature extraction and encoding methods for this purpose?

1. **Model Validation and Optimization:**

* What validation strategies can be used to ensure the accuracy and reliability of machine learning models in predicting protein functionality, and how can these models be optimized for practical application in product development?

**III. Fermentation and Maturation Effects:**

1. **Fermentation Contribution:**

* To what extent can fermentation using microbial cultures enhance the functional and sensory properties of alternative dairy gels, compared to gelation via acid or chymosin?

1. **Maturation Modelling:**

* How can the complex processes of cheese maturation be modelled to predict flavour and texture profiles, and what are the key variables that influence these characteristics?

**Complexity in cheese and the structure-function relationship**

The cheese matrix is primarily composed of a casein network with inclusions of water (wherein some components may be dissolved), fat globules, minerals, bacteria, lactose, and lactic acid. The matrix can vary considerably, depending on the relative amounts and locations of components as well as their properties which can include the charge on the caseins, the presence and composition of fat globule membranes, and the states of water, fat, and minerals. A variety of other factors also play a role, such as the processing history of the cheese, pH, temperature, and ionic strength. Cheeses can be considered to contain 4 phases (protein, fat, aqueous, and gas phase) with different interactions predominating in each (Lamichhane et al., 2018).

Hydrophobic and Ca interactions are the most important for stabilising rennet-gel cheeses while electrostatic and hydrophobic interactions along with hydrogen bonds are more important for acid-gel cheeses. In rennet-gel cheeses, the strength of the interaction between Ca and caseins decreases with increasing ionic strength and decreases with temperature. The hydrophobic portions of the caseins participate in hydrophobic interactions. As hydrophobic interactions increase in strength with increasing temperature, these interactions may play an important role in the melting behaviour of cheese (Lamichhane et al., 2018). The hydration of the cheese plays a significant role in its behaviour. In the case of PF and hybrid caseins (micellar and nonmicellar), the hydrogen bonding capacity of the proteins and their superstructure may affect the hydration. Texture and flavour release during chewing play a major role in consumer perception and is determined not only by components of the food, but the arrangement in the food matrix. The components and matrix can be modified by many different factors during processing and maturation (Lamichhane et al., 2018). The structure of the cheese has important implications for the flavour release and sensory perception of the cheese. Given the complexity of the different structures in the wide variety of cheeses available, using modelling techniques to predict flavour profiles may be of great help in producing products with the desired characteristics.

The CCP nanoclusters dissolve at low pH and thus do not play as large a role in maintaining structural integrity for acid-gel cheeses.

In the protein phase, a gel network is formed from destabilised casein micelles. The properties of the gel are variable and are influenced by the pH, concentrations of chymosin, casein, and minerals, the temperature, and properties of the casein micelles such as size (Lamichhane et al., 2018). These aspects may be of interest to model in order to elucidate the relationship between them and allow for predictions of a gel’s characteristics based on these variables.

In the fat phase, the microstructure of the milk fat globule plays an important role in the behaviour of the fat, along with the fatty acid composition, and amount of fat (Lamichhane et al., 2018). In the case of precision fermentation, the source of fat is likely to be plant-based and have a different composition and structure from the native fat in milk. Replacing dairy fat for precision fermentation-derived cheeses is thus no simple matter with the type of fat used having ramifications for the sensorial characteristics of the cheese, especially for artisanal cheeses.

The aqueous phase in cheese can be divided into bound water or bulk water that is available as a solvent. Increasing the moisture content of cheese generally results in an increase in softness due to the plasticising effect of water and the corresponding decrease in casein concentration. However, increasing the moisture content does not necessarily produce a cheese with higher meltability. The melting properties of cheese are more closely linked to the nature of the casein-water interactions where factors such as ionic strength, pH, and concentration of CCP are more important and thus cannot be predicted from the moisture content alone (Lamichhane et al., 2018). The variety of molecules that are dissolved into the aqueous phase include various water-soluble nitrogenous compounds (amino acids, enzymes, proteins, peptides), minerals, mono and disaccharides (glucose, galactose, lactose), and organic acids. These components have a marked influence on the ionic strength, pH, and water activity of the food matrix and thus influence the textural and sensorial properties of the cheese. The NaCl content of cheese for taste and preservation purposes exerts an influence on the ionic strength of cheese which has a variety of potential effects on the proteins, textural properties, and flavour (Lamichhane et al., 2018).

The fermentation and maturation process provides the complexity and variety of cheeses. Modelling this process will likely be extremely complex.

While the variety of ingredients and processing steps that are used in cheesemaking adds a great deal of complexity to any attempt to recreate these products from different ingredients, it may also provide opportunities due to the number of tools in the toolbox, so to speak, that can be used to finetune a product.

**Machine learning in product development**

The object of the project is to use some or all of these parameters as inputs (e.g. pH, protein size) and targets (e.g. gel strength) to train a ML algorithm on. The relative importance of these variables must be determined to understand how certain features of the ingredients and processing conditions affect the final product. Thus, a trained model can predict which ingredients and processing conditions would be most effective for obtaining a specific product. A wide variety of protein functionalities can be modelled using ML and much of the research has so far been focused on medical research. For the purposes of food science, the variables of interest range from the nanoscale (e.g. amino acid sequence and structure, zeta potential) to the mesoscale (e.g. aggregate size, gel structure), to the macroscale (e.g. rheological behaviour, colour). Depending on the desired application, all of these properties can be modelled.

In food science, the protein characteristics most likely to be of interest include the structure, protein-protein interactions, protein-ligand (minerals, carbohydrates, fats) interactions, flexibility, stability, and conformational dynamics. In current protein-function prediction research, two distinct steps can be outlined: the first step involves encoding the data in an appropriate digital format and the second step trains algorithms on the data (Chen et al., 2023). Previous studies have used a support vector machine classifier to distinguish between DNA and RNA-binding proteins using characteristics such as amino acid sequence, polarizability, hydrophobicity, and solvent-accessible surface area as inputs. However, using current known features of proteins to predict functionality based on ML does not result in the desired accuracy and ML can be used in both the feature extraction of proteins which can then be used as inputs for ML algorithms to predict functionality. Thus, a range of protein encoding methods are available that can be used to extract features, such as substitution matrix representation (SMR), position-specific scoring matrix (PSSM), and orthogonal local preservation projection (OLPP), amongst others. A protein representation method that may be of particular use in determining features for use in food science is the Conjoint Triad Feature (CTF), where the characteristics of three consecutive amino acids at a time are considered, thus providing not just the sequence of the protein but also information on the relationships between consecutive amino acids which can be used when modelling protein-protein interactions (PPI). The multi-scale local descriptor (MLD) method also divides the protein into parts to provide information on the individual parts, but these sections are of variable length. The essential aspects of these different methods are certain empirical inputs and the amino acid sequence which is further examined for relevant features using ML (Chen et al., 2023).

There is a variety of ways in which ML can be used to improve ingredient selection, processing, and formulation. Models have been built to select ingredients and formulations to optimise for techno-functional properties while reducing the environmental impact (Lie-Piang et al., 2024). These types of applications using these models, such as spline regression, neural networks, and random forests, are of particular interest for alternative proteins using plant proteins (such as in hybrid micelles) as they can guide the product developer in choosing the most functional and environmentally-friendly ingredients and processes. Functional properties that are crucial to the success of alternative dairy products, such as gelation, foaming capacity, and emulsion stability can be modelled using these techniques even for ingredients that are complex mixtures of many components (Lie-Piang et al., 2023). These are by no means the only algorithms used for formulation optimisation and many others can be found in the literature (Cao et al., 2020; Chew et al., 2025).

**Potential variables of interest to model:**

* Protein characteristics:
  + Amino acid sequence
  + Protein structure
  + Protein hydrophobicity/hydrophilicity (from structure and AA composition)
  + Molecular size via SDS-page
  + Protein content and purity of isolates (Dumas, SDS-PAGE)
  + Flexibility/rigidity (from 3D structure)
  + pI
  + Thermal lability using differential scanning calorimetry
  + AA sequences with interaction potential (e.g. binding CCP through phosphate centres, hydrogen bonds, hydrophilic, hydrophobic interactions)
  + Post-translational modification present/absent
    - Addition of phosphate centres, glycosylation can be tested
  + Protein secondary structure via FTIR or circular dichroism spectroscopy to assess folding
  + Solubility under conditions of temperature, pH, ionic strength, concentration (BCA)
  + Protein-protein interactions via dynamic light scattering or spectroscopy
  + Analysis of micellar and serum phases (separated by centrifugation) and measured via HPLC (serum phase only – micellar phase by deduction)
  + Concentrations and ratios of proteins, possibly including different ratios of the different caseins
* Rheological characteristics:
  + Viscosity of solutions and emulsions before and after coagulation
  + Gel strength
  + Elastic modulus, G’
  + Viscous modulus, G”
  + Strain at break
  + Time taken for gelation
  + Effects of different [chymosin], [lactic acid], [Ca], temperatures
* Emulsion properties:
  + Droplet size distribution
  + Stability
  + Capacity
  + Oil retention in gel matrix
* Water and hydration properties:
  + Hydration of micelles especially if artificial and hybrid artificial casein micelles are used
  + Water-holding capacity via centrifugation
  + [water] in solutions, emulsions, and gels
* pH and ionic strength:
  + pH of solutions, emulsions, and gels (chymosin and/or acid-induced)
  + Ionic strength due to mineral content, notably NaCl and CCP
* Fat concentration and type:
  + Concentrations of fat
  + Type of fat – source and fatty acid composition
  + Fat crystallisation in emulsions and gels via microscopy or DSC
* Microstructural analysis:
  + Protein-fat interaction via CLSM
  + Surface topography via SEM
* Sensory evaluation:
  + Flavour profile via GC-MS to monitor flavour changes with addition of plant protein
  + Texture profile to measure chewiness, hardness etc.

**References**

Cao, L., Russo, D., Mauer, W., Gao, H. H., & Lapkin, A. A. (2020). Machine Learning-aided Process Design for Formulated Products. In S. Pierucci, F. Manenti, G. L. Bozzano, & D. Manca (Eds.), *Computer Aided Chemical Engineering* (Vol. 48, pp. 1789–1794). Elsevier. https://doi.org/10.1016/B978-0-12-823377-1.50299-8

Chen, J., Gu, Z., Lai, L., & Pei, J. (2023). In silico protein function prediction: The rise of machine learning-based approaches. *Medical Review*, *3*(6), 487–510. https://doi.org/10.1515/mr-2023-0038

Chew, A. K., Afzal, M. A. F., Kaplan, Z., Collins, E. M., Gattani, S., Misra, M., Chandrasekaran, A., Leswing, K., & Halls, M. D. (2025). Leveraging high-throughput molecular simulations and machine learning for the design of chemical mixtures. *Npj Computational Materials*, *11*(1), 72. https://doi.org/10.1038/s41524-025-01552-2

Lamichhane, P., Kelly, A. L., & Sheehan, J. J. (2018). *Symposium review:* Structure-function relationships in cheese. *Journal of Dairy Science*, *101*(3), 2692–2709. https://doi.org/10.3168/jds.2017-13386

Lie-Piang, A., Boom, R., & van der Padt, A. (2024). Towards low-impact food products through reverse engineering: A functionality-driven approach. *Journal of Food Engineering*, *367*, 111857. https://doi.org/10.1016/j.jfoodeng.2023.111857

Lie-Piang, A., Hageman, J., Vreenegoor, I., van der Kolk, K., de Leeuw, S., van der Padt, A., & Boom, R. (2023). Quantifying techno-functional properties of ingredients from multiple crops using machine learning. *Current Research in Food Science*, *7*, 100601. https://doi.org/10.1016/j.crfs.2023.100601