Microsoft Research
Summit 2022

Developing ML models to approximate and predict protein digestibility

Sara Malvar

Food insecurity a slow-burning emergency now

How can we produce enough protein to feed 10 billion people?

Why we need to focus on nutrition, not hunger

Hungry nation: 70% Indians cannot afford healthy diet

Rising food prices fuelled this upward trend, with the cost of a healthy diet increasing by 3.3% in 2020

Importance of nutritious diet for a healthy life

There is more than enough food produced in the world to feed everyone on the planet. Yet as many as 828 million people still go hungry.

We need to increase production and decrease environmental impact

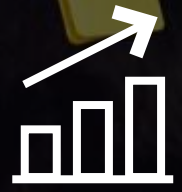
Over **820 million** people worldwide suffer from hunger

More than **2 billion** people lack vital nutrients

50% more food is needed by 2050. **40%** food wasted.

31% GHG emissions from agri-food systems, **70%** fresh water use

Our current meat-based diet is not sustainable



By 2050 the population will increase **30%,** while poultry consumption will increase **120%**

More than **70%** of medically important antibiotics are consumed by animals in the US

Plant-based meat market is expected to reach **$85B by 2030** and **$370B by 2035**

Change is being led by the **65%** omnivorous and **29%** flexitarians

New food development

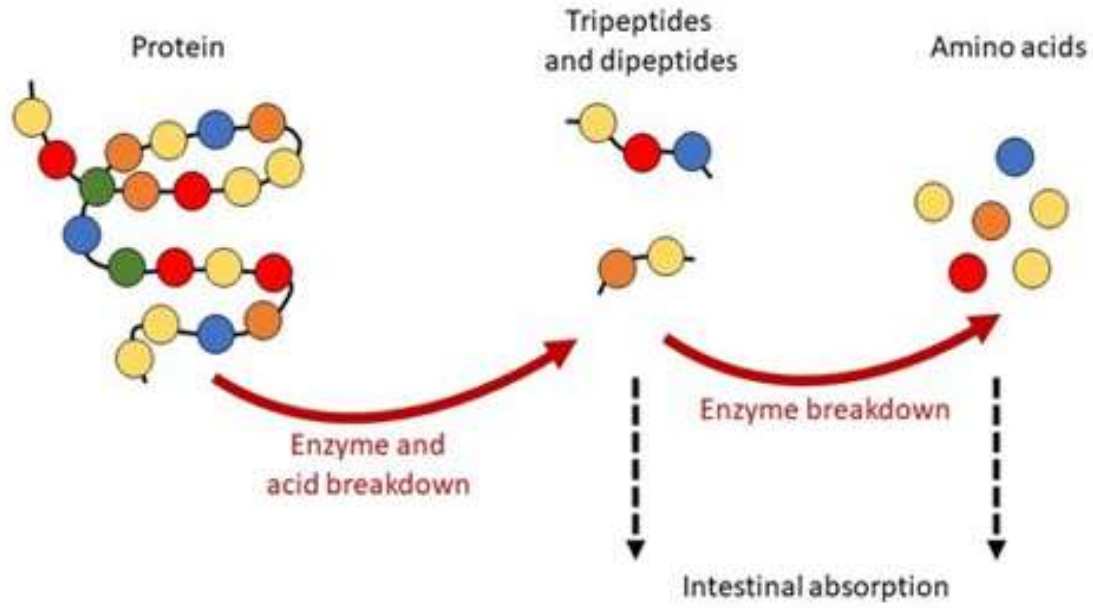
Food scientists need to check for

- *texture*
- *flavor*
- *bioavailability*

Proteins

The building blocks of proteins are amino acids linked together through peptide bonds

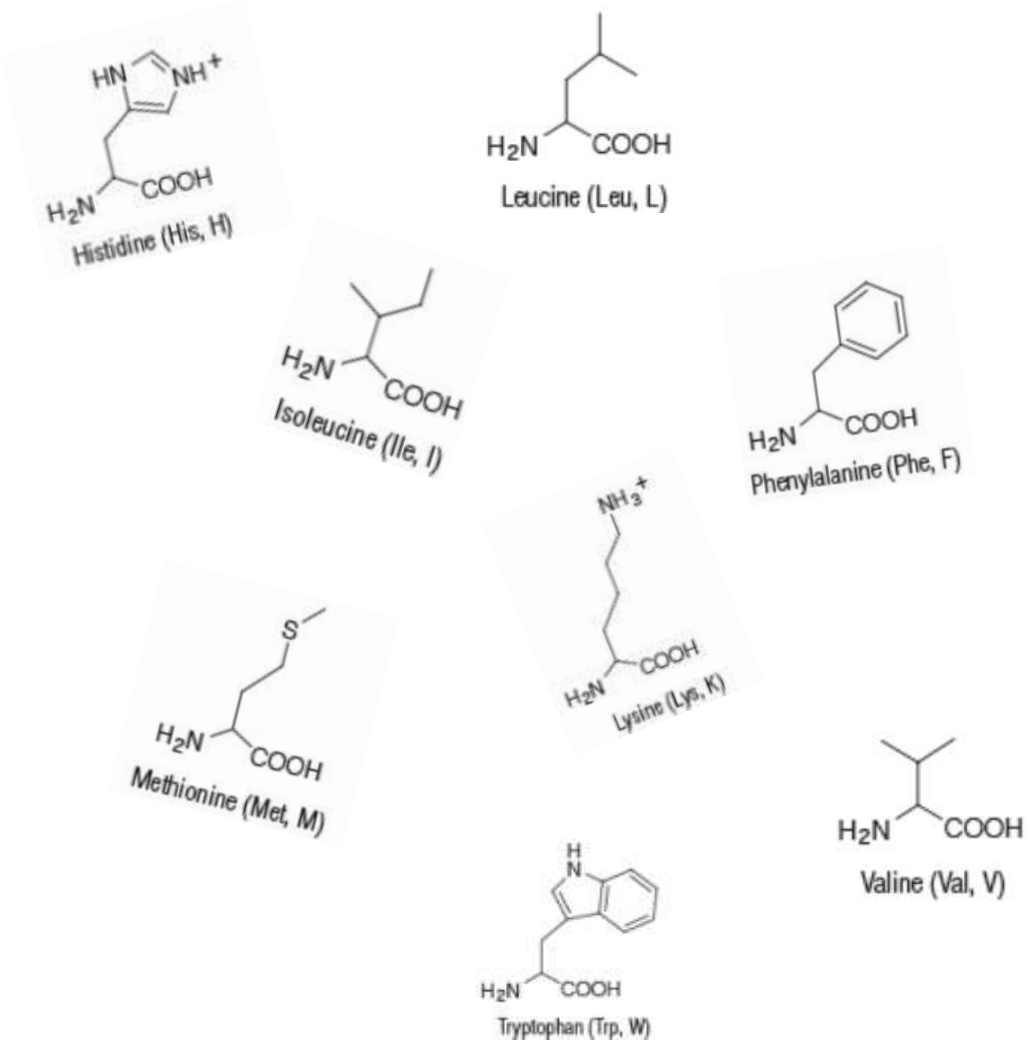
During digestion, these proteins are broken into amino acids



Essential amino acids

9 amino acids are considered essential: *phenylalanine, valine, threonine, tryptophan, methionine, leucine, isoleucine, lysine, and histidine*.

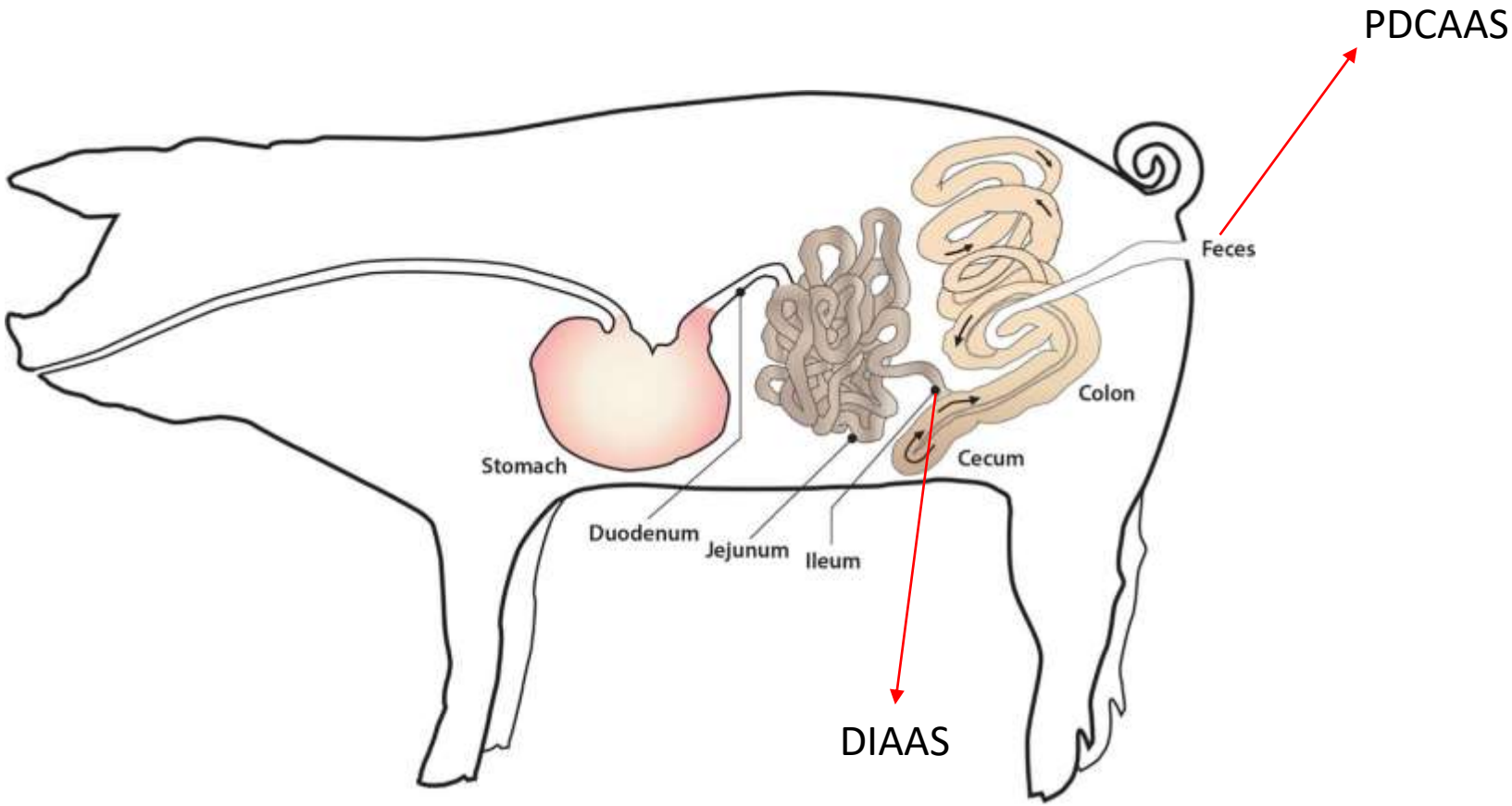
6 are conditional, being essential only under special pathophysiological conditions, such as prematurity in the infant or individuals in severe catabolic distress: *arginine, cysteine, glycine, glutamine, proline, and tyrosine*.



Why they are important?

- Break down food.
- Grow and repair body tissue.
- Make hormones and brain chemicals (neurotransmitters).
- Provide an energy source.
- Maintain healthy skin, hair and nails.
- Build muscle.
- Boost your immune system.
- Sustain a normal digestive system.

Protein digestibility



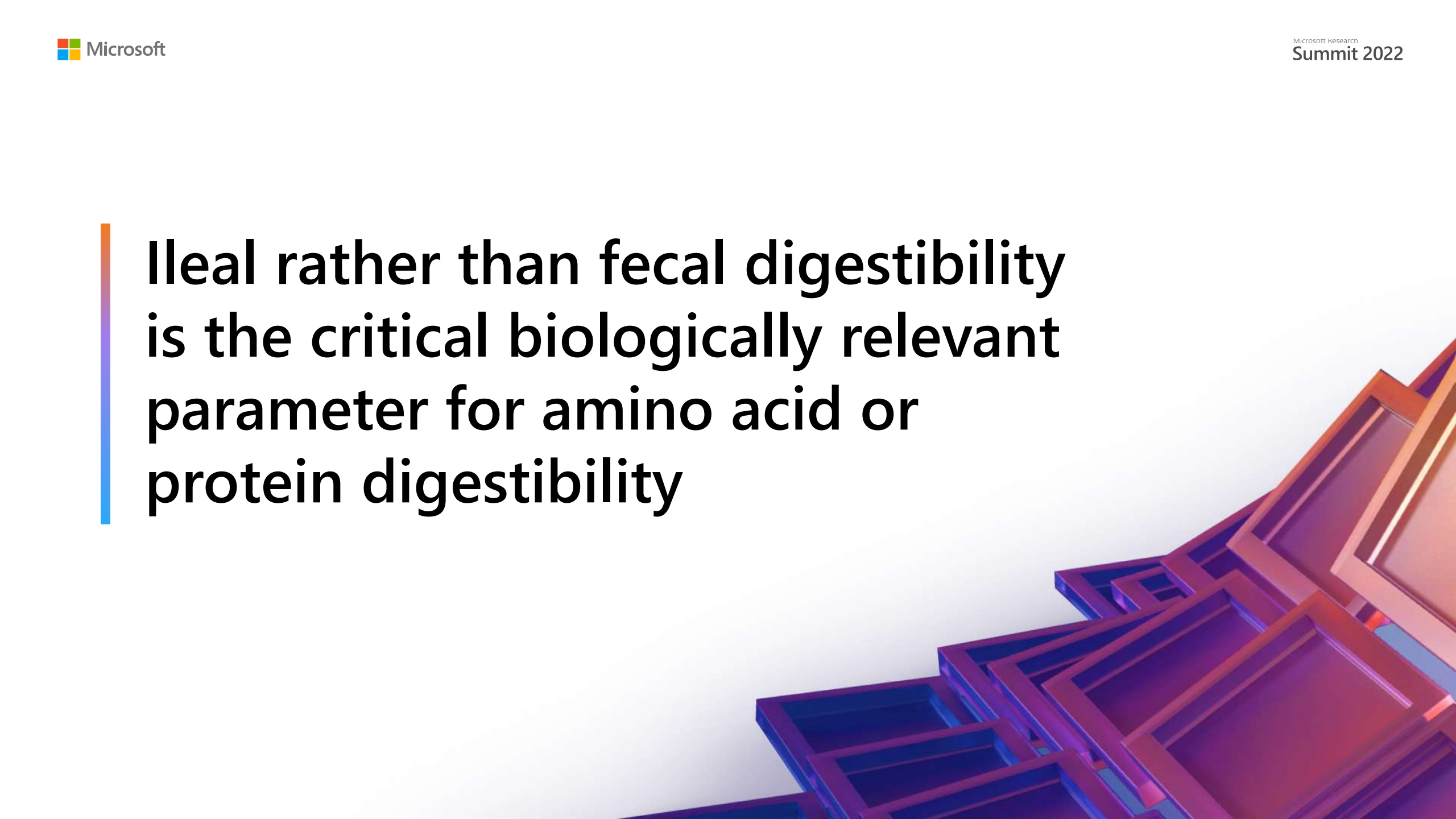
Protein digestibility

Protein digestibility-corrected amino acid score (PDCAAS)

- First limiting essential amino acid of the test protein as a percentage of the content of the same amino acid in a reference pattern.
- Percentage is corrected for the true fecal digestibility of the test protein, as measured in a rat assay.
- Truncated to 100%

$$\text{PDCAAS}[\%] = \frac{\text{mg of limiting amino acid in 1 g of test protein}}{\text{mg of same amino acid in 1 g of reference protein}} \times \text{fecal true digestibility}[\%] \times 100$$

Ileal rather than fecal digestibility is the critical biologically relevant parameter for amino acid or protein digestibility



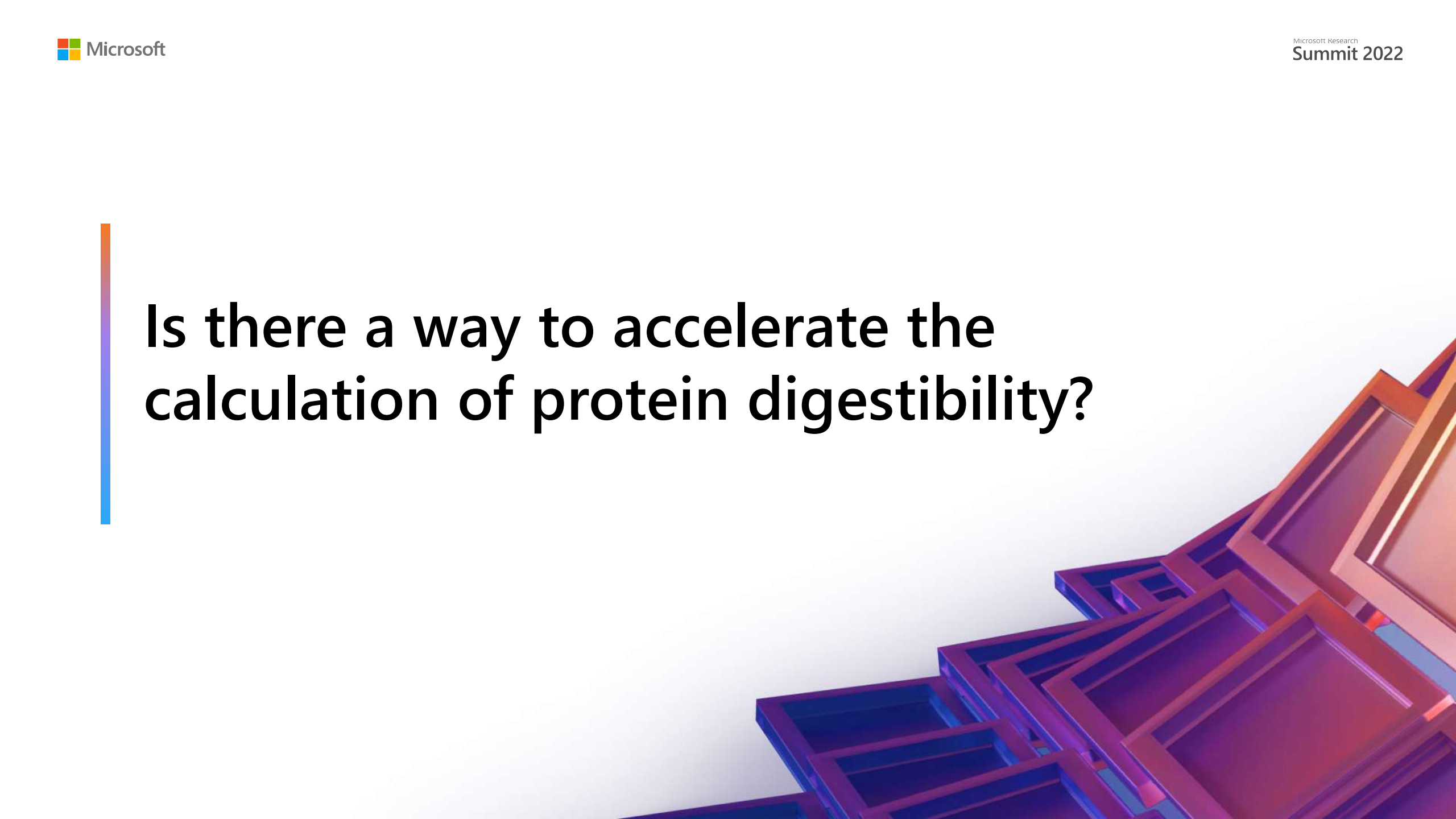
Protein digestibility

Digestible Indispensable Amino Acid Score (DIAAS)

- The DIAAS method involves determining the ratio of digestible amino acid.
- This test is typically done in pigs, which have a stronger biological similarity to humans than rats and other test subjects.
- We look for values above 75%

Fundamental challenge: calculating the ileal digestibility values for each amino acid in each food requires extensive animal testing using pigs.

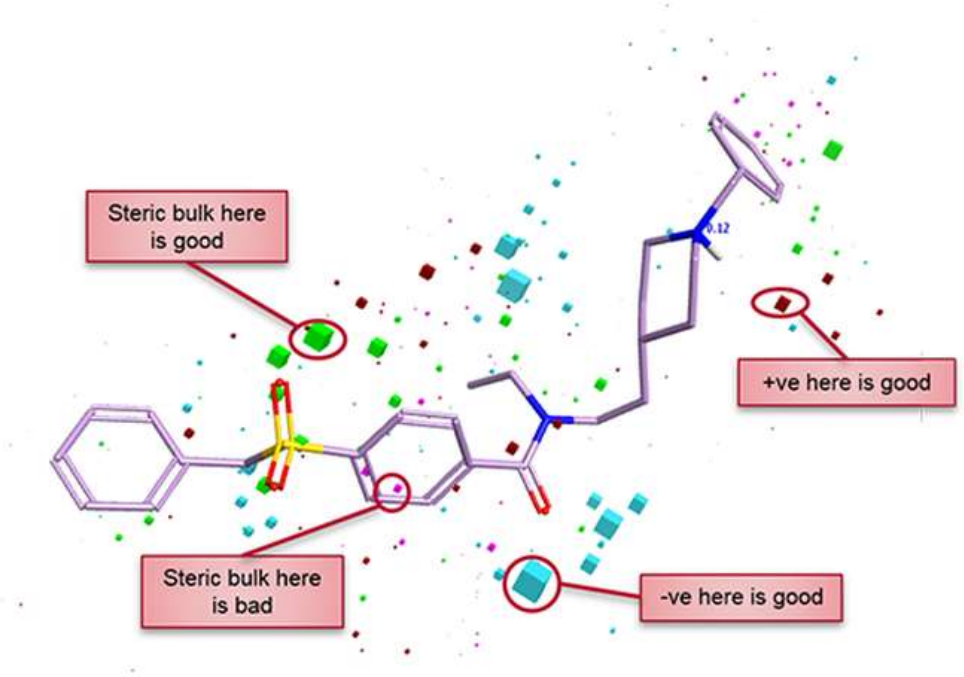
Is there a way to accelerate the calculation of protein digestibility?



Quantitative Structure-Activity Relationship (QSAR)

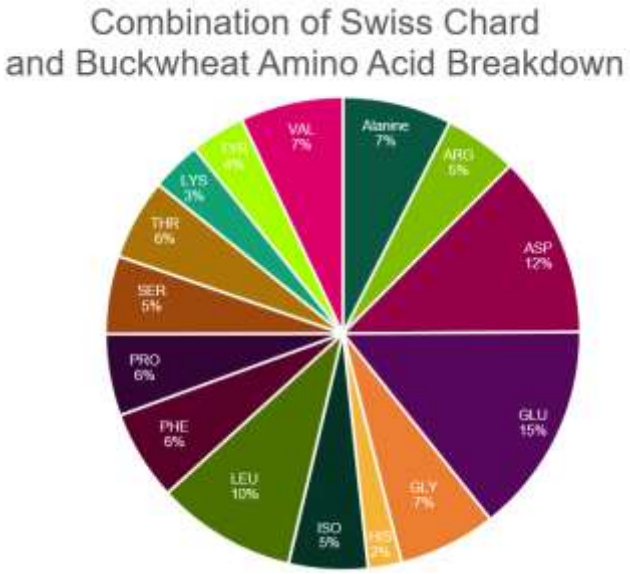
Approach commonly used in drug discovery and medicinal chemistry.

Mathematical models that relate biological activity / function to physiochemical characteristics of chemical compound.



Creating the dataset

1. Amino acid composition



Amino acid composition and ileal digestibility coefficients of ~200 raw foods. Dataset was curated based on experimental results from published papers.

2. Protein sequence

VAEGEASEQLQCERELQELQERELKACQQVMDQQLRDISP
PPKGGSFYPGETTPPQQLQQRIFWGIPALLKRYYPSVTCP
PGQGQQGYYPTSPQQPGQWQQPEQGQPRYYPTSPQQSGQL
GQPGYYPTSSQLQPGQLQQPAQGQQGQQPGQAQQGQQPGQ
QGQQGQQLGQGQQGYYPTSLQQSGQGQPGYYPTSLQQLGQ
QLQQPAQGQQPGQGQQGQQPGQGQQGQQPGQGQQPGQGQP
SSQQPTQSQQPGQGQQGQQVGQGQQAQQPGQGQQPGQGQP
SPQQSGQGQQPGQLQQSAQGQKGQQPGQGQQPGQGQQGQQ
SPQQSGQGQQPGQWQQPGQGQPGYYPTSPLQPGQGQPGYD

Nutrient characteristics scraped from food databases such as the USDA one for these foods. These characteristics help further characterize our foods

3. Nutrient characteristics

Nuts, pine nuts, dried (1)

New Search

Refuse: 23% (Shells)
Scientific Name: *Pinus spp.*
NDB No: 12147 (Nutrient values and weights are for edible portion)

Nutrient	Units	1.00 X 1 cup ------ 135g
Proximates		
Water	g	3.08
Energy	kcal	909
Energy	kJ	3802
Protein	g	18.48
Total lipid (fat)	g	92.30
Ash	g	3.50
Carbohydrate, by difference	g	17.66
Fiber, total dietary	g	5.0
Sugars, total	g	4.85

FASTA sequences of these foods collected from UniProt database. Physiochemical features extracted from the protlearn package.

Baseline models

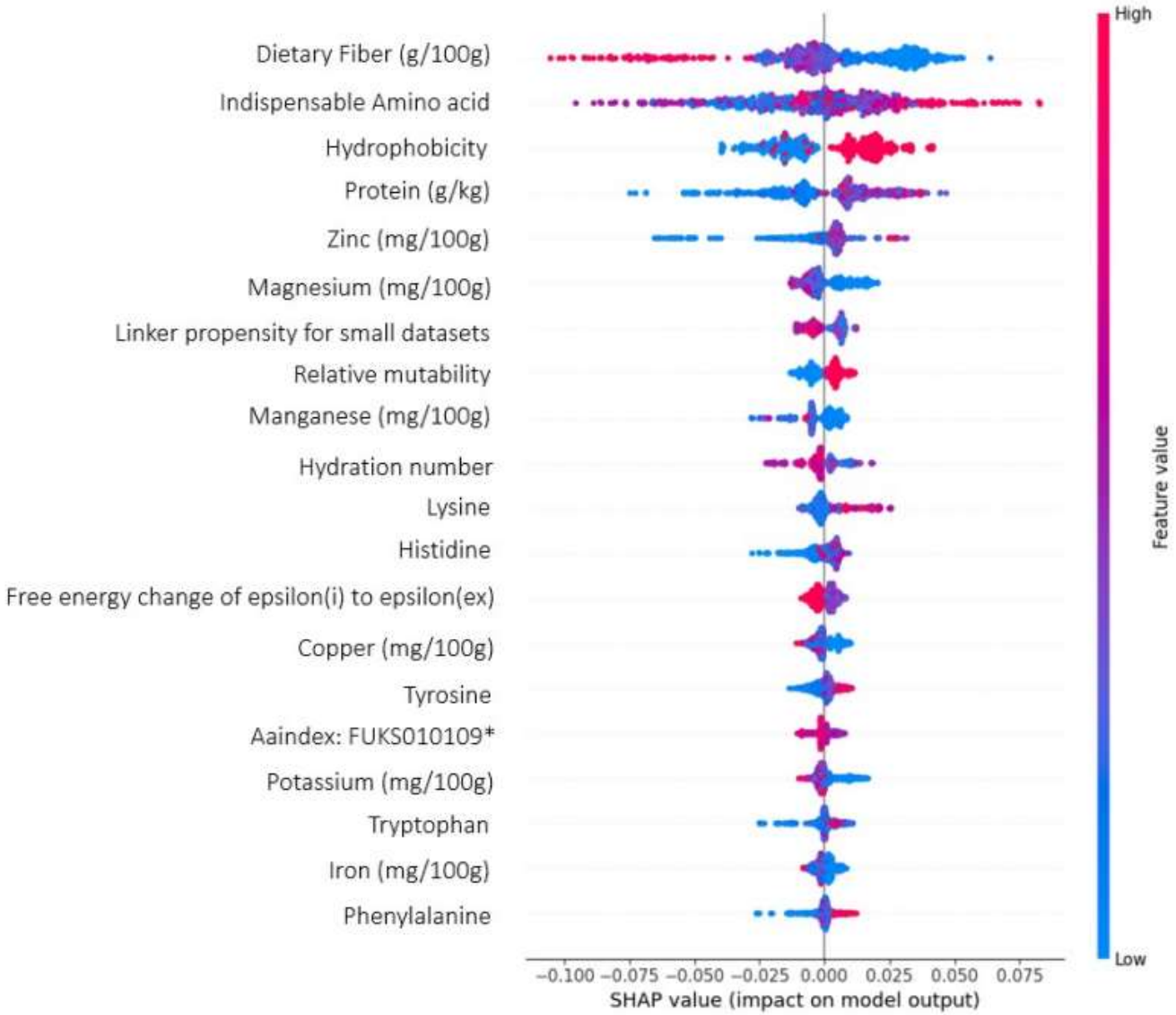
Preliminary results showed that tree-based classical models looked the most promising at predicting the ileal digestibility coefficients of foods.

Model	R^2	RSME
LGBM	0.88	0.04
XGBoost	0.87	0.04
Random Forest	0.82	0.05
Gradient Boosting	0.82	0.05
Bagging	0.81	0.05
KNeighbors	0.73	0.06
Decision Tree	0.72	0.06
Nu Support Vector Regression	0.70	0.07
Linear Regression	0.66	0.07
Lasso model (with Lars)	0.66	0.07
SVR	0.62	0.07
Poisson	0.49	0.09

SHAP analysis

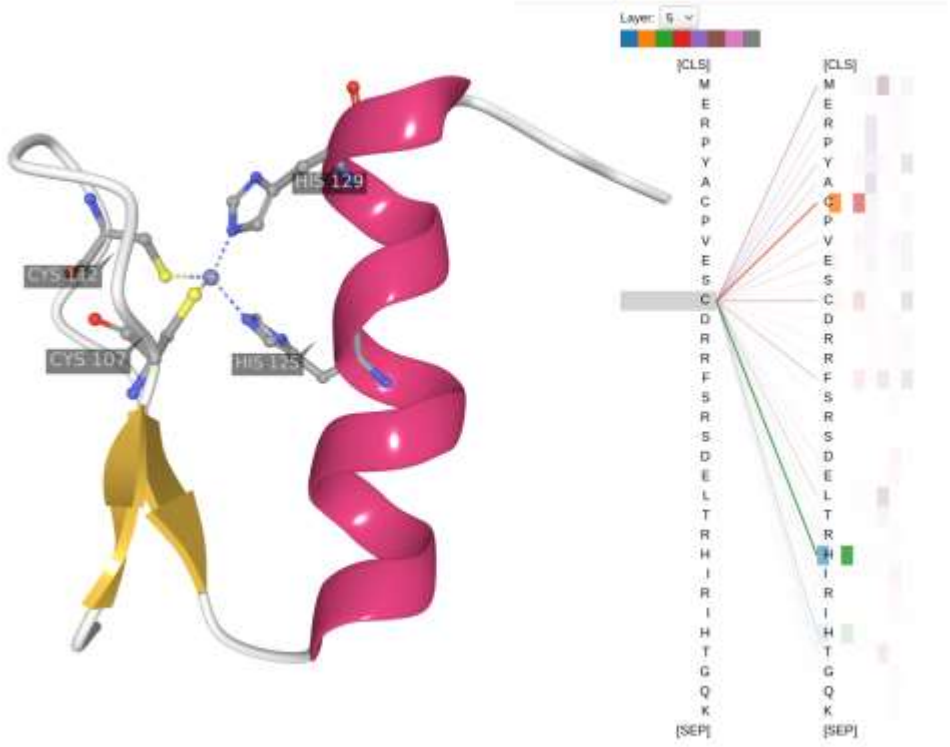
The second step of many ML methods is the feature extraction, the aim of which is to get the most effective features from the obtained raw segment.

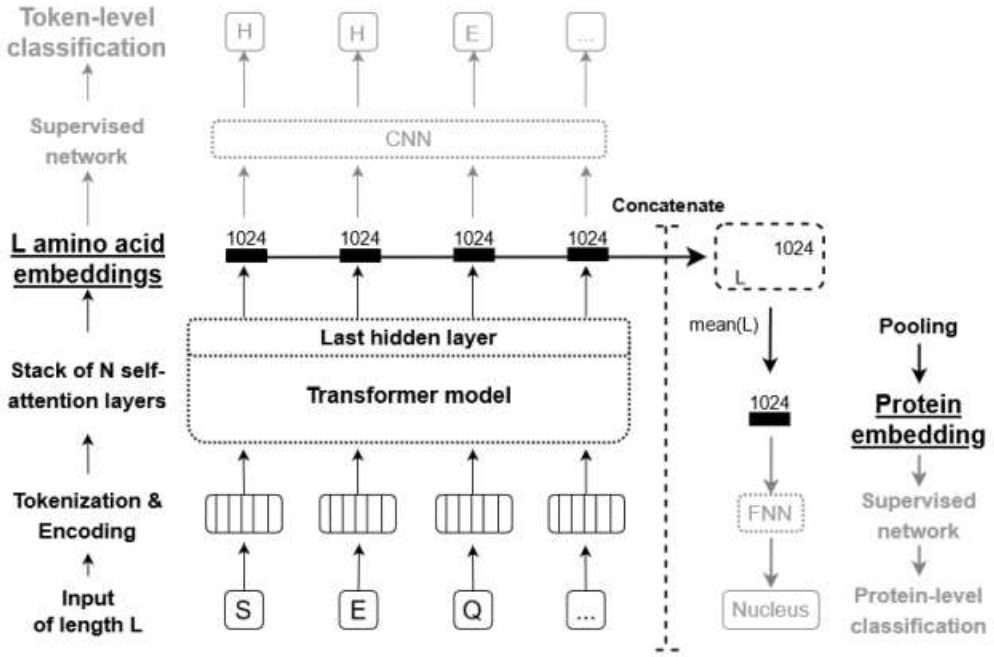
This problem was approached as a feature selection or dimensionality reduction method



Language model

Extracted 1024 embeddings for each food/protein family from a pretrained transformer model called ProtTrans.





Final architecture

Nutritional Information

AA composition

Protein sequence

ProtLearn

Biochemical info.

ProtTrans

Regressor

Features

SHAP

Final model

Equation

DIAAS for infants (0 - 6 months)

DIAAS for children (6 months - 3 years)

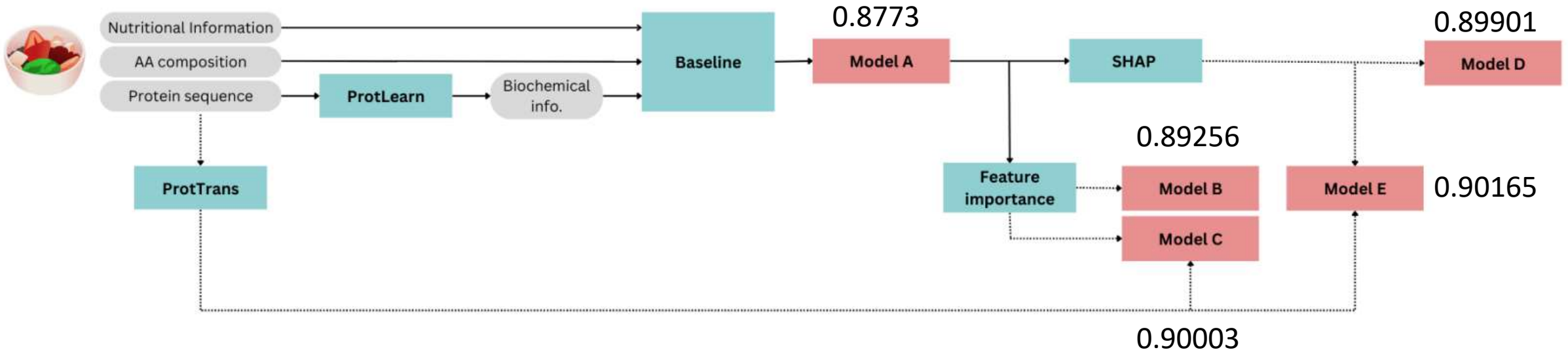
DIAAS for adults

$$\frac{\text{IAA}_{lim} \text{ in TP}[mg/g \text{ TP}]}{\text{IAA}_{lim} \text{ in RP}[mg/g \text{ RP}]} \times \text{ileal digestibility IAA}_{lim}[\%]$$

TP: Test Protein RP: Reference Protein

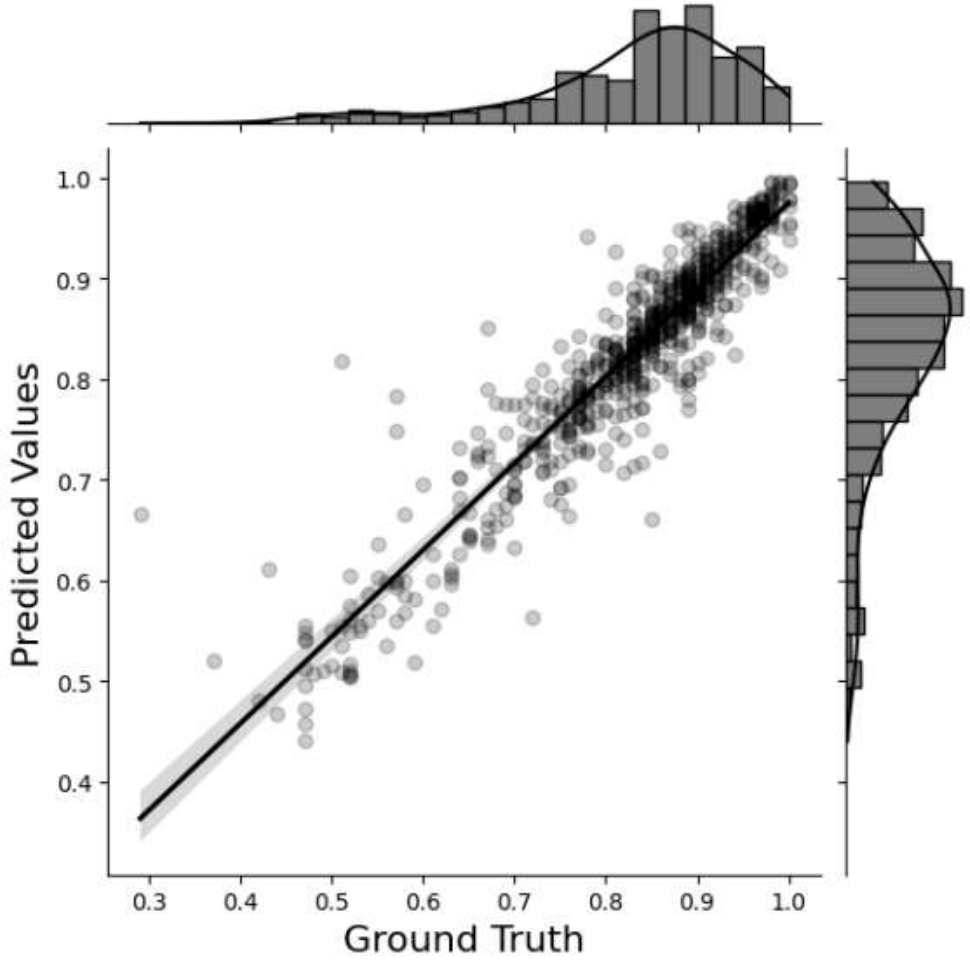
IAA_{lim}: First limiting indispensable amino acid

Ablation study



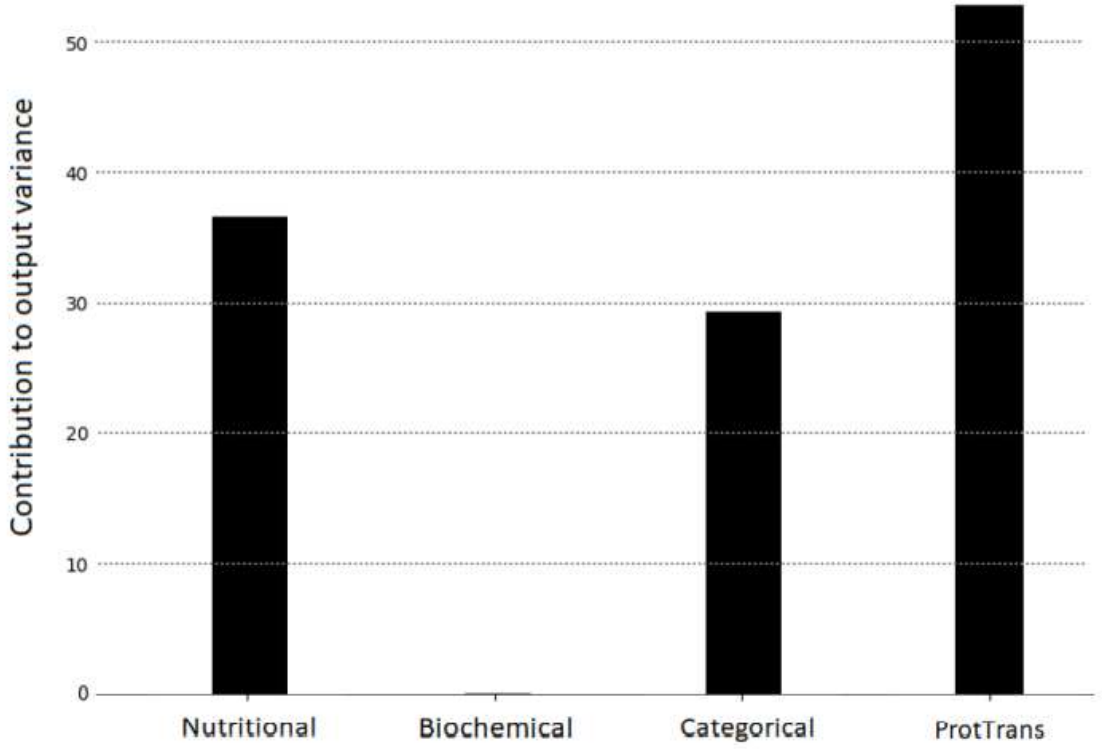
Best model

The R^2 of the model can be visually represented by the trend analysis of the ground truth data plotted against the predicted values. For each data point the values should be similar. Besides a few outliers, both ground truth and predicted distributions are similar.



Sensitivity analysis

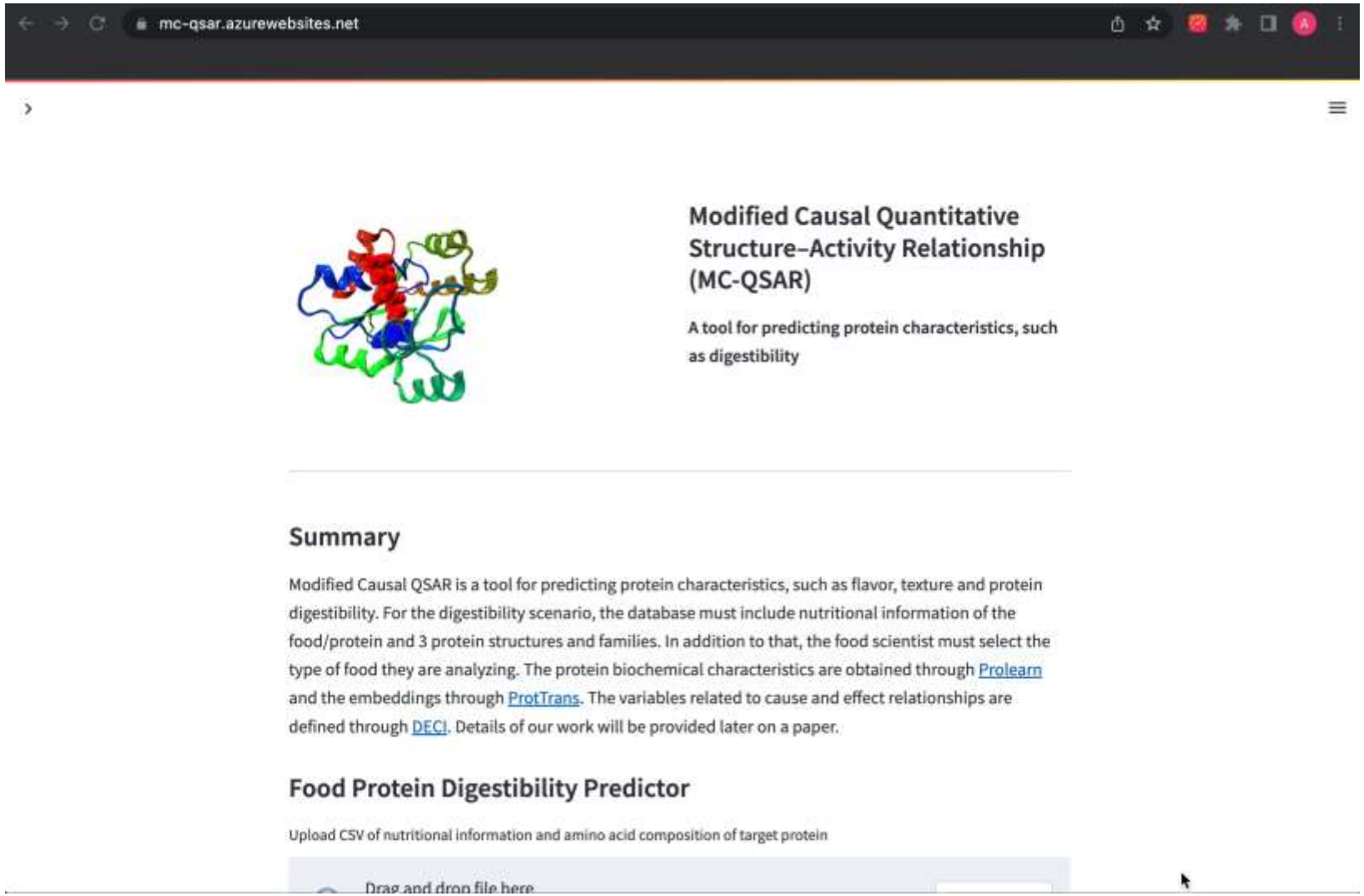
SALib package was used to perform variance sensitivity analysis. It is clear the ProtTrans features are responsible for over 50% of the variance, followed by the nutritional information and the categorical features



Microsoft

Microsoft Research
Summit 2022

Demo


mc-qsar.azurewebsites.net
Modified Causal Quantitative Structure–Activity Relationship (MC-QSAR)
A tool for predicting protein characteristics, such as digestibility
Summary
Modified Causal QSAR is a tool for predicting protein characteristics, such as flavor, texture and protein digestibility. For the digestibility scenario, the database must include nutritional information of the food/protein and 3 protein structures and families. In addition to that, the food scientist must select the type of food they are analyzing. The protein biochemical characteristics are obtained through Prolearn and the embeddings through ProtTrans. The variables related to cause and effect relationships are defined through DECI. Details of our work will be provided later on a paper.
Food Protein Digestibility Predictor
Upload CSV of nutritional information and amino acid composition of target protein
Drag and drop file here

Maize (yellow dent)

48

47.09

Tofu

88

94.58

Pork (raw belly)

119

118.16

Microsoft Research
Summit 2022

Thank you

Read our paper: Machine learning can guide experimental approaches for protein digestibility estimations