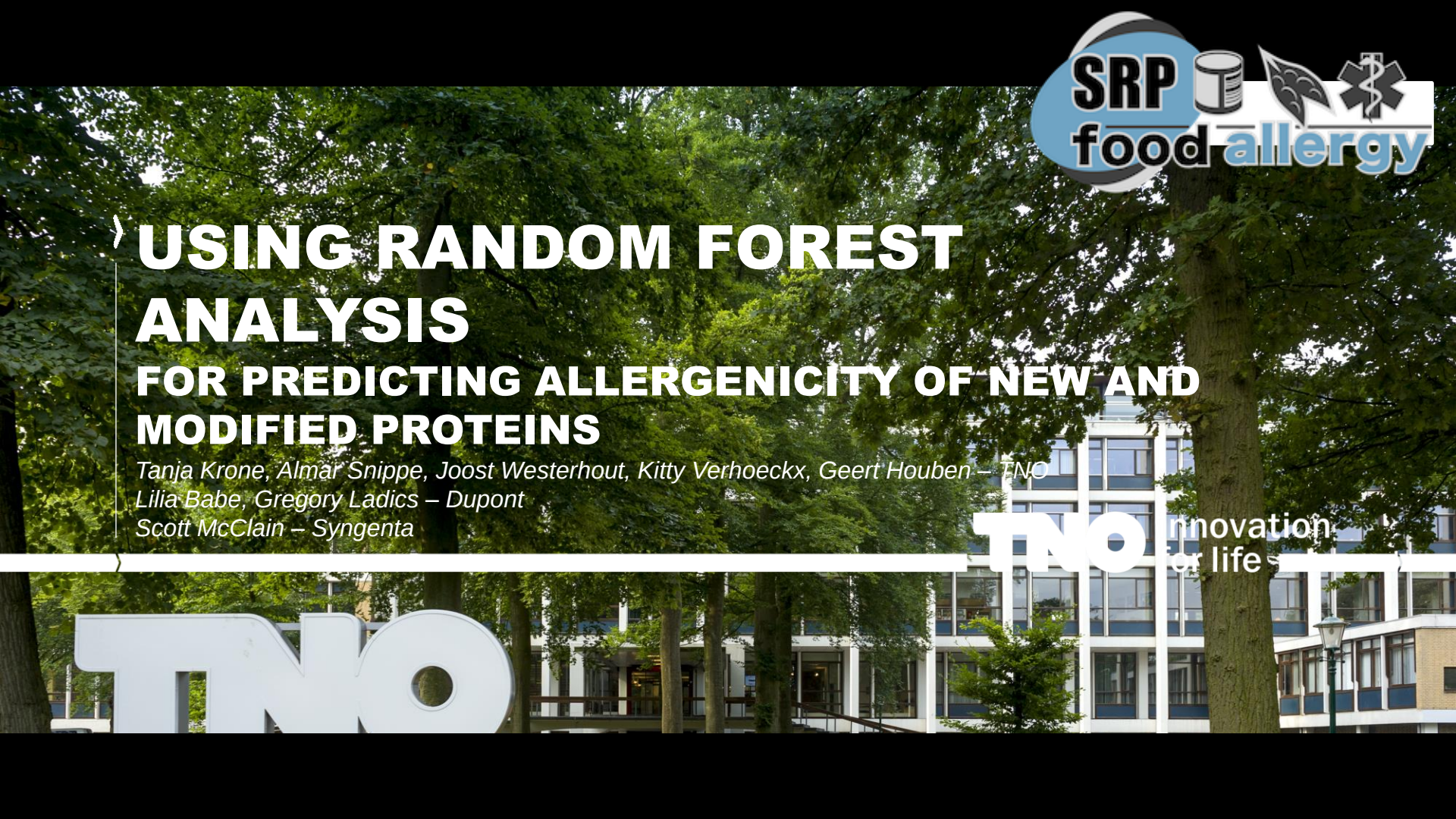


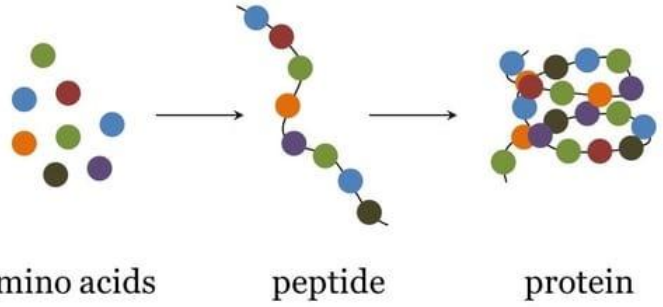
USING RANDOM FOREST ANALYSIS

FOR PREDICTING ALLERGENICITY OF NEW AND MODIFIED PROTEINS

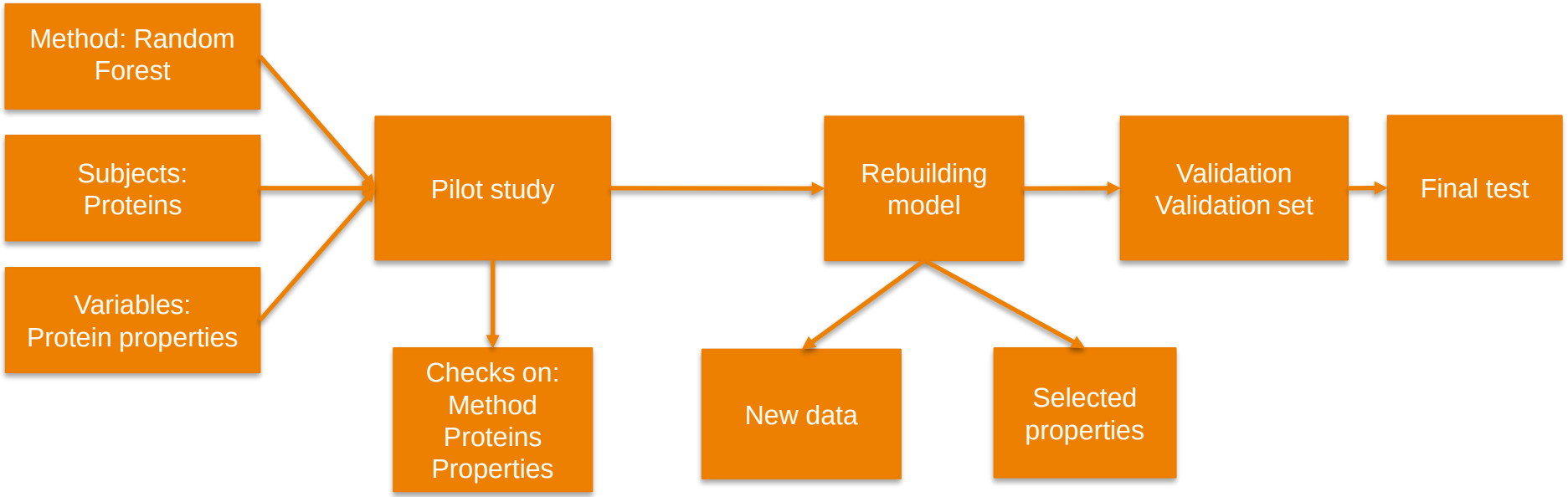
Tanja Krone, Almar Snippe, Joost Westerhout, Kitty Verhoeckx, Geert Houben – TNO
Lilia Babe, Gregory Ladics – Dupont
Scott McClain – Syngenta

BACKGROUND

- Protein allergenic reactions can be clinically important – e.g., peanut allergens can cause anaphylactic responses
- Typical Type I food allergy reactions are caused by proteins; some new food sources include new/broader food allergen exposure when distributed to new populations
- Predictive knowledge on distinguishing allergenic from non-allergenic proteins is lacking

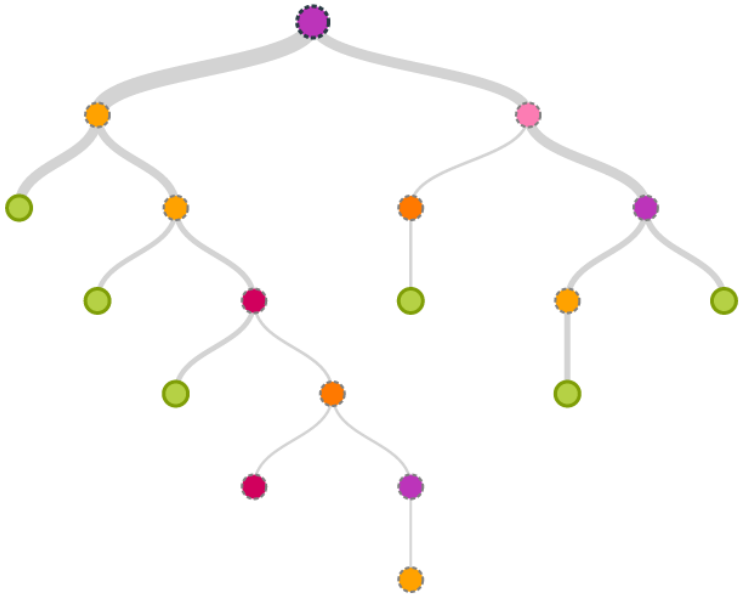


MODEL DEVELOPMENT STEPS



THE RANDOM FOREST MODEL

- Estimate classification based on combination of properties
- Create a large number of decision trees
  - Each tree consists of branches, splits and leaf
  - For each split a variable is selected,
    which is then split to diminish entropy
  - Each branch marks a sidepath after a split
  - Each leaf marks a final point (“decision”)



RANDOM FOREST: ACCURACY, SPECIFICITY, SENSITIVITY

- $Sensitivity = \frac{TP}{TP+FN} * 100\%$ → correctly predict positives
- $Specifity = \frac{TN}{TN+FP} * 100\%$ → correctly predict negatives
- $Accuracy = \frac{TP+TN}{TP+FP+TN+FN} * 100\%$ → correctly predict both

TP: True positives
TN: True negatives

FP: False positives
FN: False negatives

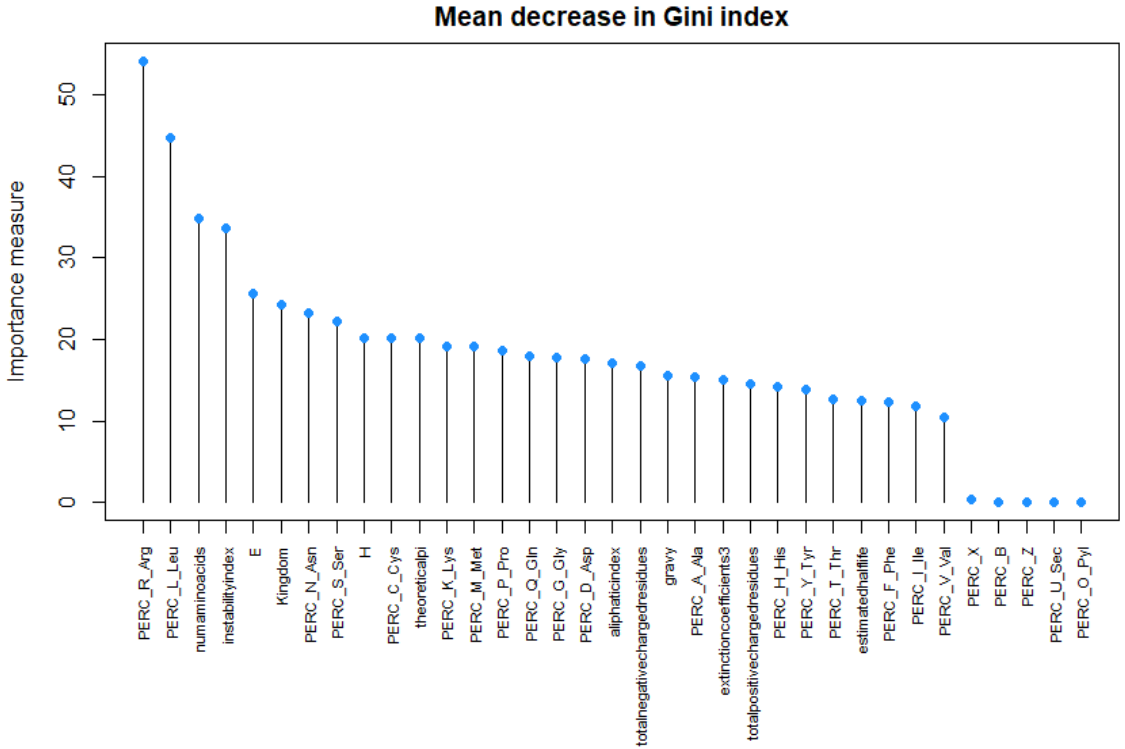
SUBJECTS: PROTEINS

- Resources include well-described, clinically relevant allergens and their sequences in available database(s).
- Open source data: Uni-prot
- 85.000.000 proteins
  - 550.000 reviewed proteins
  - 1680 allergenic proteins

BUILDING MODEL – DATA & VARIABLES

- Obtained subset of proteins:
  - Selected Training set: 40.000 non allergens, 839 allergens
- Only parameters obtained from amino acid sequence
  - Inclusion of parameters calculated by Protparam
  - Inclusion of secondary structure values obtained from PSI-PRED
- Three kingdoms
  - Animal, Plant, Fungi → Bacteria and virus hold too few allergens
- No need to reduce number of variables → all information is derived from the Amino Acid sequence

RESULTS



Performance measures	Final model	Six variable model
Accuracy	89%	87%
Specificity	89%	91%
Sensitivity	89%	84%

MODEL VALIDATION

- Predict allergenicity for new set
  - Animal: 10.000 non-allergenic proteins, 140 allergenic proteins
  - Fungi: 10.000 non-allergenic proteins, 50 allergenic proteins
  - Plant: 10.000 non-allergenic proteins, 229 allergenic proteins

Model	Accuracy	Specificity	Sensitivity
Training Set	89%	89%	89%
Animal	85% (-4%)	85% (-4%)	91% (+2%)
Fungi	86% (-2%)	86% (-3%)	88% (-1%)
Plant	89% (-0%)	89% (-0%)	91% (+2%)

  - These are good results for the validation: Accuracy is never below 85%

ALL INTACT PROTEINS WERE CORRECTLY PREDICTED

Name	species	Sequence comparable to known allergens	Predicted allergen	Allergenic
Larval cuticle protein A2B	Tenebrio molitor	N	Y	Y
Larval cuticle protein A1A	Tenebrio molitor	N	Y	Y
Larval cuticle protein A3A	Tenebrio molitor	N	Y	Y
Alpha-amylase	Tenebrio molitor	Y	Y	Y
Tropomyosin-1, isoforms 9A/A/B	Drosophila melanogaster	Y	Y	Y
Arginine kinase	Drosophila melanogaster	Y	Y	Y
Arginine kinase (Fragment)	Tenebrio molitor	?	N	Y
Cytochrome b	Tenebrio molitor	N	N	N
Elongation of very long chain fatty acids protein	Tenebrio molitor	N	N	N

LINKING STATISTICS WITH BIOLOGY

- The biological relevance of the biochemical properties with strongest effect on prediction model remain oftentimes a question. Some possible explanations:
  - The percentage of cysteine and the instability index are related to the stability of the protein. High stability of a protein is correlative with allergenic proteins.
  - The percentage of arginine and lysine are both involved in the fate of the protein in the gastrointestinal tract (stability and transport), but have opposite correlation with allergenic proteins.

TAKE HOME MESSAGES

- Important to predict allergenic potency of new proteins early in the development pipeline and to protect the allergic consumers.
- Using Data-driven methods, we created a model with over 85% accuracy, sensitivity and specificity
- The model might be applicable for (novel) food dossiers for safety assessment
- Statistical models and biological knowledge evolve over time, so new variables can be added in the future.
- Good collaboration between different areas of expertise is required for applied research
- Future steps: test on other, new proteins

ACKNOWLEDGEMENT

TNO innovation for life

Almar Snippe
Joost Westerhout
Jack Vogels
Eugene van Someren
Geert Houben
Kitty Verhoeckx

DUPONT

Greg Ladics
Lilia Babe

syngenta

Scott McClain