1. Materials and Methods

# Datasets

Both datasets used for the creation of this thesis was provided by ALK-Abell´o. The two datasets describe donor’s positive immune stimulation to specific Ragweed- and Tree allergen peptides, respec- tively. The immune stimulation is quantified by the T cell stimulation index (SI), which describes T cell proliferation calculated by:

*SI* =

*cpm stim*

*cpm medium control*

(1)

The Ragweed dataset is describing 53 anonymous donor’s immune responses to 226 different aller- gen peptides from 21 different allergens derived from four related ragweed species: Amb a (ambrosia artemisiifolia), Amb P (Ambrosia psilostachya), Amb t (Ambrosia trifida) and Art v (Artemisia vul- garis). In total, this dataset contains 6014 datapoints. Subsequently we will refer to this data set as the Ragweed data set.

Further, the second dataset contains 32 anonymous donors’ response to 72 different allergen pep- tides derived from allergens from the following tree species: Alnus glutinosa (alder), Malus domestica (Apple) and Quercus alba (white oak). This dataset contains 2844 datapoints in total.

Both datasets contain 7 attributes: *Donor ID*, *Ori protein*, *Ori pepseq*, *Ori SI*, *V ar protein*, *V ar pepseq* and *V ar SI*. Here, the ’ori protein’ and ’var protein’ attributes describes names of the allergens which the peptides are derived from. The sequence length of the allergen peptides varies from 19 to 23 amino acids. Subsequently we will refer to this data set as the Tree data set

# Data processing

The datapoints were aggregated into sets for each peptide pair in order to associate each peptide pair with the SIs of the set of donors reacting to each of the peptides in the pair. These sets were constructed for each dataset and used to analyse cross-reactivity between the respective peptides. Due to the exponential nature of the SI measurements, a logarithmic transformation was performed. Before data filtering, 375 peptide sets was constructed for the Ragweed dataset and 99 peptide sets was constructed for the Tree dataset.

# Quantitative assessment of cross-reactivity

The level of cross-reactivity between a pair of homologous peptides was quantified by calculating correlation coefficient between SI values towards the the peptide pairs across a series of donors. This approach is in accordance with the one used by Wu¨tzen et al.[3]. Specifically, both Pearson’s Correlation Coefficient (PCC) and Spearman’s Rank Correlation Coefficient (SCC) was used for the quantitative assessment of cross- reactivity between pairs of homologous allergen peptides, and cross-reactive peptides were defined from having an SCC > 0.5

# Data filtering

For data filtering, a permutation with 1000 re-samplings was carried out for the data of each peptide to comparison to combat false positive cross-reactions. Cross-reacting peptide pairs that showed statistical insignificance (P-value *>* 0.05) in the permutation test, was eliminated from the dataset and thus not used for further analysis. Further, all peptides which were predicted to not be binders of the MHC class II alleles from the 7 allele method (as discussed in 2.4) were also discarded from further analysis.

# Global alignment

Pairwise global alignment of each peptide pair was performed using the Needleman-Wunsch alignment algorithm. The pairwise global alignment was performed using both a naive approach with the following scoring scheme: gap penalty =-1, match score = 1 and mismatch score = -1, and a less naive approach using the BLOSUM50 substitution matrix for match and mismatch scoring. Peptide similarity was calculated by:

*Similarity*(%) =

*Aligned matches length of shortest sequence*

· 100 (2)

# Local alignment

Local sequence alignments of each peptide pair was performed using the Smith-Waterman(O2) algo- rithm. For the alignment, a gap penalty of -3 and a gap extention penalty of -1 was used. A Match and mismatch scoring was performed using the BLOSUM50 scoring matrix. Alignment performance was evaluated both by BLOSUM50 score and % identity.

# K-mer similarity

All possible 9-mers (substrings of 9 amino acids) of the ’ori pepseq’ was scored against all possible 9 mers of the ’var pepseq’ in the peptide sets using both percent similarity and BLOSUM50 score. This specific substring length was chosen to mimics the length of the peptide binding core for MHC class II.

As an extension to the 9-mer similarity feature, a kernel function was to calculate a score for the combination of all K-mer similarities ranging from *K* = 3 to *K* = 12 for all peptide sets.

* + 1. Peptide binding core prediction

NetMHCIIpan 4.1 was used for prediction of the MHC class II binding cores in each peptide in the datasets. This prediction was performed with basis in the seven HLA class II alleles from the seven allele method, listed in table ??.

Table 1: HLA class II variants used for peptide binding prediction

HLA-DRB5\*01:01

HLA-DRB4\*01:01

HLA-DRB3\*02:02

HLA-DRB3\*01:01

HLA-DRB1\*15:01

HLA-DRB1\*07:01

HLA-DRB1\*03:01

HLA Class II alleles

For each of the peptides, a predicted binding core and a respective rank was outputted for the seven HLA class II alleles. The binding core rank is an indicator for peptide:HLA binding affinity. The lower the rank the stronger the binding. The outputted rank of the peptide binding cores was considered in the following way:

* + - * Rank *<* 1 is a predicted strong binder
      * 1 ≥ Rank ≥ 5 is a predicted weak binder
      * Rank *>* 5 is a predicted non binder

Similarity features based on the predicted binding cores, was evaluated by both BLOSUM50 score and % identity. These similarity features include:

1. Similarity between the best ranking binding cores for each peptide in a set across all seven HLA class II alleles.
2. Similarity of the best ranking binding core across both peptides in the set and the corresponding binding core of the other peptide of same allele.
3. All the binding cores of one peptide were also scored against the against the binding cores of the other peptide and the best scores were used as a feature.

# Baseline assessment

In preparation for machine learning, a sequence based method was used for detection of data redun- dancy, and thus establishment of a performance baseline for the two datasets. Using 5 fold cross- validation, every peptide pair in the validation section was scored against every peptide pair in the training section using the peptide kernel score. Every peptide pair from the validation section then inherits the PCC value from the closest (highest scoring) peptide pair in the training section. High correlation in the baseline indicates high redundancy in the dataset.

# Random Forest

Random forest regression models were created with SciKit-Learn’s RandomForestRegressor function using the default hyperparameters. The specifics regarding these, can be found in appendix ??. MAN- GLER REFERENCE

5 fold cross-validation was used in order to ensure a correct estimate of the performance of the model. The 5 models yielded by the cross-validation were then used as an ensemble model for any predictions on other datasets.