Multi-view learning to unravel the different levels underlying hepatitis B vaccine response

Fabio Affaticati[1,2], Esther Bartholomeus[2,3], Kerry Mullan[1,2], Pierre Van Damme[2,4], Philippe Beutels[2,5], Benson Ogunjimi[2,5,6,7], Kris Laukens[1,2], Pieter Meysman[1,2]

[1]Adrem Data Lab, Department of Computer Science, University of Antwerp, Antwerp, Belgium; [2]Antwerp Unit for Data Analysis and Computation in Immunology and Sequencing, University of Antwerp, Antwerp, Belgium; [3]Department of Medical Genetics, University of Antwerp, Antwerp, Belgium; [4]Centre for the Evaluation of Vaccination (CEV), Vaccine and Infectious Disease Institute, University of Antwerp, Antwerp, Belgium; [5]Centre for Health Economics Research & Modeling Infectious Diseases, Vaccine & Infectious Disease Institute (VAXINFECTIO), University of Antwerp, Antwerp, Belgium; [6]Antwerp Center for Translational Immunology and Virology (ACTIV), Vaccine and Infectious Disease Institute, University of Antwerp, Antwerp, Belgium; [7]Department of Paediatrics, Antwerp University Hospital, Antwerp, Belgium

Abstract

Background

The high complexity of biological systems arises from the large number of spatially and functionally overlapping interconnected components constituting them. The immune system, which is built of reticular components working to ensure host survival from microbial threats, presents itself as particularly intricate. A vaccine response is likely governed by levels that, when considered separately, may only partially explain the mechanisms at play. Multi-view modelling can aid in gaining actionable insights on response markers shared across populations, capture the immune system diversity, and disentangle confounders. Hepatitis B virus (HBV) vaccination responsiveness acts as a feasibility study for such an approach.

Material and methods

Seroconversion to vaccine induced antibodies against HBV surface antigen (anti-HBs) in a vaccination cohort containing early-converters (n=21 ; <2 month) and late-converters (n=9 ; <6 months), was based on the anti-HBs titres (>10IU/L). Two approaches (principal component analysis and canonical correlation analysis) were used to interpret the multi-view data which encompassed bulk RNAseq, CD4+ T cell parameters (including T-cell receptor data), flow cytometry data, metadata including gender and age of the baseline parameters.

Results
Multi-view joint dimensionality reduction out-performed single-view methods in terms of AUC and balanced accuracy, confirming an increase in predictive power to be gained. The interpretation of the findings showed that age, gender, inflammation-related genesets and pre-existing vaccine specific T-cells were associated with vaccination responsiveness.

Conclusion
This multi-view dimensionality reduction approach complements the clinical seroconversion and all single-modalities and could identify what features underpin HBV vaccine response. This methodology could be extended to other vaccination trials to identify key features regulating responsiveness.

Introduction

The immune system is a complex network of interconnected components including effector cells, receptors, and mediators that underlie both humoral and cell-mediated immunity, innate and adaptive responses [1]. A single-level analysis of any of these components fails to fully capture the heterogeneity of the immune system and may not uncover meaningful markers. To address common research questions, a "systems immunology" approach is often necessary to leverage the interrelationships between different component levels. This framework will aid in detecting shared patterns and characterise the drivers of immune responses in a comprehensive manner [2]–[4]. Unfortunately, few methods are available within systems immunology that can directly interrogate immunological data across different levels.

Multi-view learning is a branch of machine learning techniques consisting in the fusion and integration of multiple datasets. By exploiting the additional intra-dataset information available, models thus built aim for a better generalisation performance [5]. Multi-view learning has received a lot of attention and several algorithms with encouraging performance have since been developed, many relating back to Canonical Correlation Analysis (CCA) [6]. Multi-view learning has been successfully applied for biological data integration, in particular, CCA based applications have been implemented for oncological, Alzheimer's studies and brain imaging tasks [7]–[11]. Principal Component Analysis (PCA) [12], [13] is a more well known dimensionality reduction technique that computes a linear transformation of the original features to produce a new set of orthogonal features of lower dimensionality that best explain the variability of the data. The derived variables are uncorrelated that capture maximal variance and are placed in descending order. CCA belongs to the same category of statistical

techniques of PCA and it can be qualified as Joint Dimensionality Reduction (JDR). The main difference lies in the input and the optimised statistic between the two methods. While PCA operates on a single multivariate dataset at a time, CCA works on two datasets. CCA has since been modified to function on more than two datasets at once and exploit their interdependencies by maximising the sum of pairwise correlations (SUMCOR) [14]. The dimensions of the different views are simultaneously reduced while maximising correlation across modalities instead of variance, irrespectively of the data types.

Unfortunately, the modelling for immunological datasets is restricted to the common multi-omics data integration approach, which does not make use of CCA modelling. This multi-omics method consists of a fine-tuned design aimed towards a specific set of data types, with Multi-Omics Factor Analysis (MOFA) as a prime example [15]. MOFA allows direct use of relational information to integrate the levels, such as the concept that a specific gene will be translated into a specific protein that belongs to a specific pathway. This specialisation locks their applicability to datasets of their field and hinders their use for immunology where other levels are more commonly measured, such as immune cellular composition or T cell and B cell receptor (TCR/BCR) repertoires.

Multi-view analysis is potentially beneficial to the vaccinology field. As highlighted in the literature, many factors influence how efficacious a vaccination is. These factors comprise identifying which immune cells are critical for a robust immune response including which memory cell population needs to develop (*e.g.* B cell vs T cells) and understanding how patient factors (*e.g.* gender, age, ethnicity) influence these findings. For instance, age significantly modifies the landscape of the immune system. These ageing changes are well documented in age-related pathologies such as Alzheimer's, arthritis, hypertension and atherosclerosis, which is linked to chronic inflammation and interferon activation [16]–[18]. Additionally, pre-existing cross-reactive T-cells, such as in the CD4+ memory compartment being required for a robust vaccine response to HBV [19]. Therefore, multi-level factors need to be considered when determining who would benefit the most from a vaccination. However, the current applied approaches only analysed one layer at a time, and hence might have led to potential confounders in interpreting the data.

Among the unsolved questions in vaccinology is what drives a good immune response to a vaccine. Prior studies have shown that a vaccine response can be, in part, predicted using mRNA expression from blood (cells) at baseline [20]–[22]. Thus, the response to a vaccine seems to be partially determined by the state of the immune system before the vaccine is administered. Predictive models based on a single data type, such as peripheral blood mononuclear cells (PBMCs), bulk transcription RNA expression, or flow cytometry (protein expression), have been previously built. However, as indicated before, these different measurement levels are not independent, with ~40% of mRNA levels being correlated to protein expression [23]. Therefore, it is not known what is causative, and what is a related

confounder. A multi-level approach may aid in deconvoluting these potentially confounding factors.

In this paper, we evaluate different multi-view approaches on multi-view data including bulk blood RNA sequencing (RNA-seq) gene expression, PBMC flow cytometry cell counts, memory CD4+ T-cell receptor repertoire and demographic metadata with reference to HBV vaccine response. We aimed to identify if this methodology could aid in determining if the baseline state is predictive of the vaccine response. As a proof of concept of our methodology, here we highlighted multi-view learning identified biological markers that correlate with response profiles developed after *de novo* HBV vaccination.

Methods

Study cohort

The study participants had been previously recruited to assess the HBV vaccination response [19], [24]. The cohort consisted of HBV-naïve 34 individuals, who were neither vaccinated nor infected (according to serology) against/by HBV. The original publications aimed to identify the response to a three-dose regimen of Engerix-B vaccine, with second and booster dose respectively administered 30 days and a year after the first dose. The vaccine response was measured based on anti-HBs titres at days 0, 60, 180 and 365 (Summarised in **Table 1**).

***Table 1**. Cut-off of converters and non-converters to hepatitis-B vaccination.*

Outcome	Ab titre	Time	# in group
non-converters	<10IU/L	-	4*
early-converters	>10IU/L	60 days	21
late-converters	>10IU/L	180 days	9

** Excluded from further analysis.*

For the non-converter class, it is impossible to determine if seroconversion was eventually reached at a later time point. For one non-converter lack of space on the kit during TCR sequencing also precluded access to the related data modality. Non-converters were thus discarded from the modelling.

Metadata of each participant were also considered (**Supplemental material S1**). All clinical features had been collected before administering the first dose. The features included age, gender, temperature, maximum and minimum blood pressure (Min and Max_BP). Other

clinical factors available included access to the absolute numbers of white blood cells (with differentiation in monocytes, lymphocytes and granulocytes), red blood cells (with separate attributes for hemoglobins and hematocrit) and platelets were determined with a haematology analyzer [19], [24] (**Supplemental material S1**). Lastly, flow cytometry identified the percentages of CD4+ T cell relevant information (**Supplemental material S1**), as this population is critical to developing a robust HBV antibody response. These CD4+ T cell parameters included the normalised ratio of vaccine-specific T-cells (HepBTCR), the total number of TCRs sequenced in the CD4+ memory (B0), the frequency of bystander TCRs (PPnrB0) and the frequency of vaccine-specific TCRs (PSB0).

RNAseq data quality control steps for

The first study [19] detailed the blood extraction and bulk RNA-sequencing of the PBMCs at day 0 and after vaccination at days three and seven. As stated, the pre-vaccination time point was considered. Of the 23812 transcripts, gender-specific and haemoglobin genes as well as removed transcripts with fewer than 100 counts across the samples were excluded, which left 13012 expressed transcripts for analysis. The counts were normalised using DEseq2's median of ratios to correct for sequencing depth and RNA composition [25], to allow reliable between-samples comparisons. Differential gene expression analysis was performed with the DESeq2 negative binomial models. To improve the interpretability, the gene counts were aggregated based on the published BloodGenModule3 features to reduce the high dimensionality effects that this data view would carry in the analysis [26]. BloodGenModule3 was chosen, as it contained a hierarchical, fixed repertoire of 382 functionally annotated gene sets (blood transcriptional modules [BTMs]) [27]. The aggregation BTMs by means was used.

Data sampling

The class distribution is relatively unbalanced. Synthetic Minority Oversampling TEchnique (SMOTE) [28] was used to upsample the minority class (late-conversion) in the training data set at each iteration of the Leave One Out Cross Validation (LOOCV) used to evaluate performance. It has been shown that SMOTE can be beneficial to high dimensional datasets when applied after feature reduction, such as in this case. New data points were generated by computing average measures of the five nearest neighbours of the same class in the training data set.

Integration methods

The 30 samples (**Supplemental material S1**), were used to evaluate several integration approaches for the applicability to the four processed views. The workflow is detailed in **Figure 1**. Each data view was also initially inputted in the Logistic Regression (LR), multivariable model commonly used for biological sciences studies, to establish a baseline

and examine the separate contributions. LR coefficients are readily interpreted, as the regression model requires few assumptions to determine the relationships between dependent and independent variables [29].

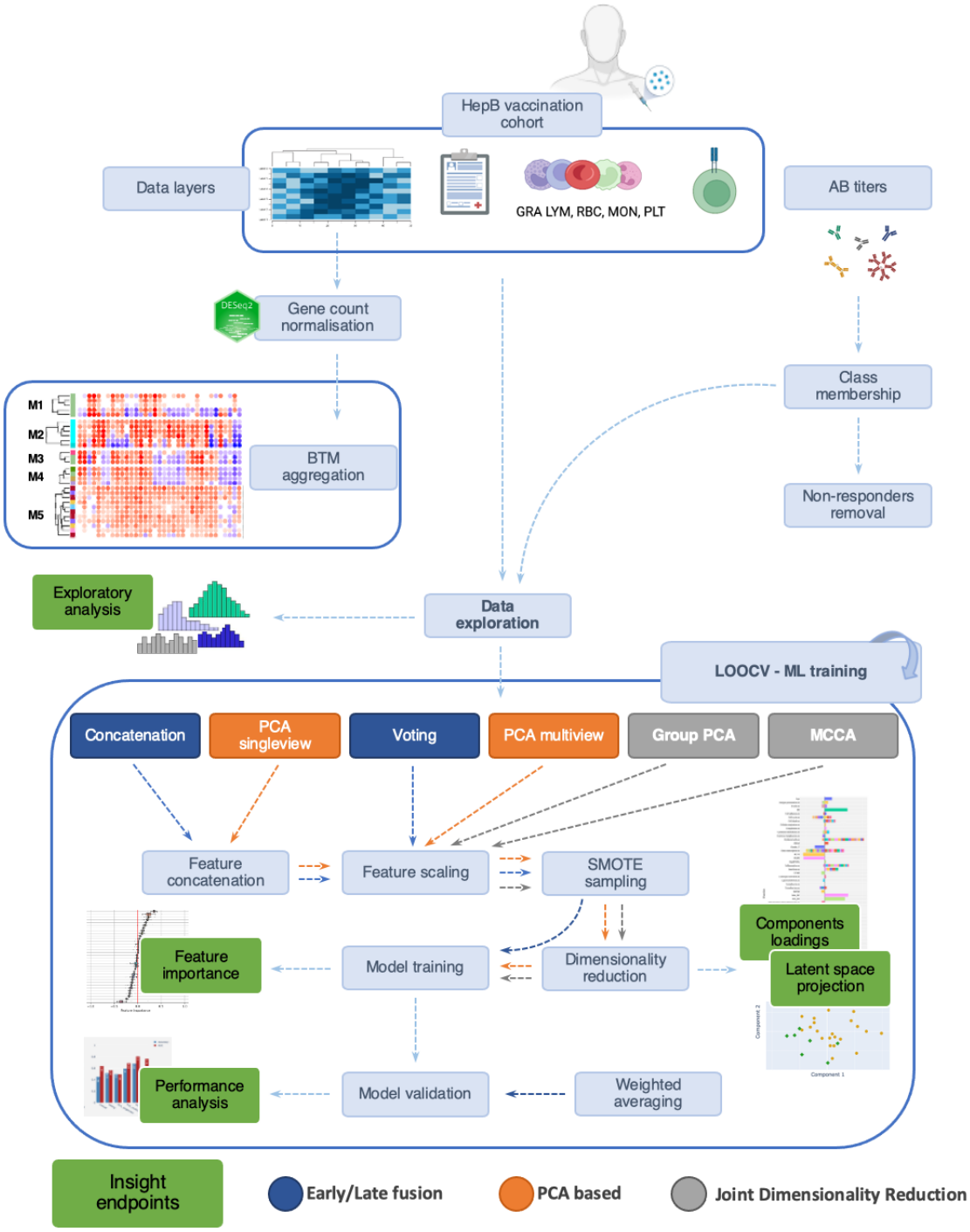

Figure 1. Research methodology workflow. *The colour coding represents the integration philosophies (blue, orange and grey) and, in green, highlights the interpretability endpoints. Figure created using BioRender (BioRender.com).*

The simplest and most common method to implement was an early fusion in which the views were concatenated at a feature level and a model is trained on all the variables. However, the advantageous ease of application is counterbalanced by the likely skewed fusion caused by high dimensional sets, such as our gene expression matrix [8], [30]. Mirroring this first technique a late fusion through voting was also applied [31]. The concatenation, in this case, happens at the decision level by merging independent predictions through weighted averaging. A classification model is built to exploit specific characteristics of the target view. The predictions are then weighted and combined in the final results.
Feature engineering was explored next. Variants of PCA were deemed most appropriate for our case. A sparse PCA variant [32], [33] that introduces a LASSO penalisation [34] was applied on the concatenated features and its output fed as training to the models (named "PCA_singleview"). A PCA was also applied on each separate view to discover equivalent dimensions of the view-specific components. In case of the gene expression set the sparse variant was used. The embeddings obtained were then concatenated along the component dimension before training (named "PCA_multiview"). As an even more compact alternative each view was whitened with PCA, then, latent features were concatenated and a PCA was performed on this matrix once more (Group PCA) [35]).

An adapted multi-view version of CCA, MultiviewCCA (MCCA)[14], [35], [36] in its regularised form was implemented. The regularisation parameter was introduced to reduce overfitting caused by differences in dimensionality between datasets. Regularised MCCA aims to discover a common signal across sources while also reducing view-specific noise [37]. Sparsity was introduced to denoise gene expression in particular [38]. The number of features in this case greatly exceeds the number of observations and CCA could not be applied otherwise. Shared components can then potentially show meaningful discriminant features. Similarly to PCA, each component can be individually studied to reveal the factors that most contributed to it. This aspect is fundamental to draw valuable insights in the task at hand. The decision to reduce the summed dimensionality (399 features) of the multi-view data to a few dimensions allows to obtain substantially more interpretable results.

Some of CCA's assumptions were considered during our analysis. Even though it was not strictly necessary for the machine learning algorithms used here, we applied standardisation on every dataset to respect CCA's assumption of multivariate normal distribution and homogeneity of variance. The relationship between the canonical variates calculated and the original features is assumed to be linear unless a kernel variant is used. Similarly, to multivariate regression, and many other statistical algorithms, CCA gains inference power and reliability from a high sample size. The application of SMOTE contributed in this sense by generating more data points. These assumptions can also be applicable for every other PCA-based transformation we implemented.

Results

Anti-HBs titers separation

We performed a PCA transformation of the antibody titer data. Principle component (PC) 1 (93% of the variation) segregates the data based on when the individuals responded to the vaccination (**Figure 2**). Therefore, they were used to segregate out the groups.

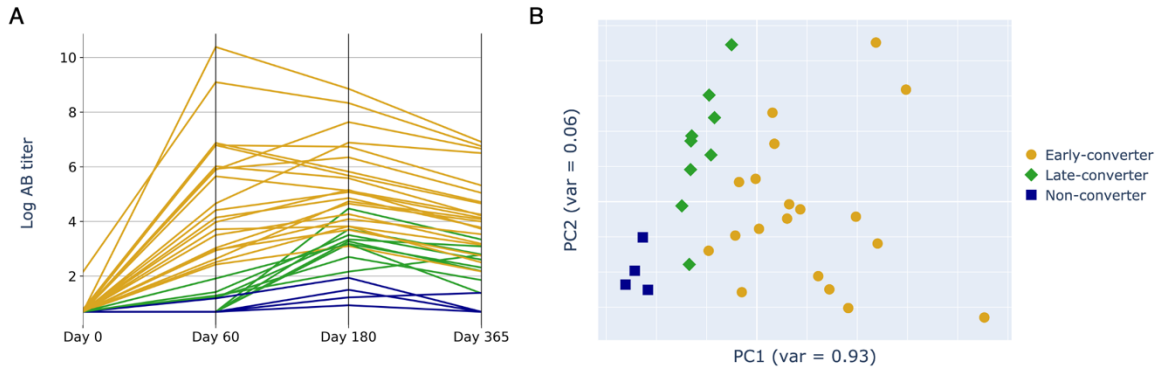

Figure 2. Principal components analysis of anti-HBs titres is correlated with responsiveness. *(**A**) Antibody titers at different sampling time points on a logarithmic scale, divided across the three response classes. (**B**) Principal component (PC) 1 and PC2 of the antibody titer data. Coloured by converter type (early-converter = orange circle; later-converter = green diamond; non-converter = blue square).*

Independent data levels show correlated features

Collinearity across views is an expected characteristic within the multi-view HBV vaccine data set, as each molecular or clinical level will influence the others. This step was to uncover dependencies within the metadata irrespective of the vaccine response. As expected, several variables are strongly correlated (**Figure 3**). For example, the lymphocytes (LYM0) and granulocytes (GRA0) counts are strongly anti-correlated (R = -0.98). As per the literature [38], the expected negative correlation of female gender (Gender_F) and red blood cell count (RBC0) (R=-0.63) was identified. TCR-derived parameters were also assessed. The day 0 total unique TCR sequences per individual (B0) is correlated with both the fraction of HBs-Ag-specific (PSB0, R=0.88) and aspecific (PPnrB0, R = 0.61). The normalised ratio of vaccine-specific TCRs (HepBTCRs) was poorly correlated to the initial TCR counts (B0; R = 0.16).

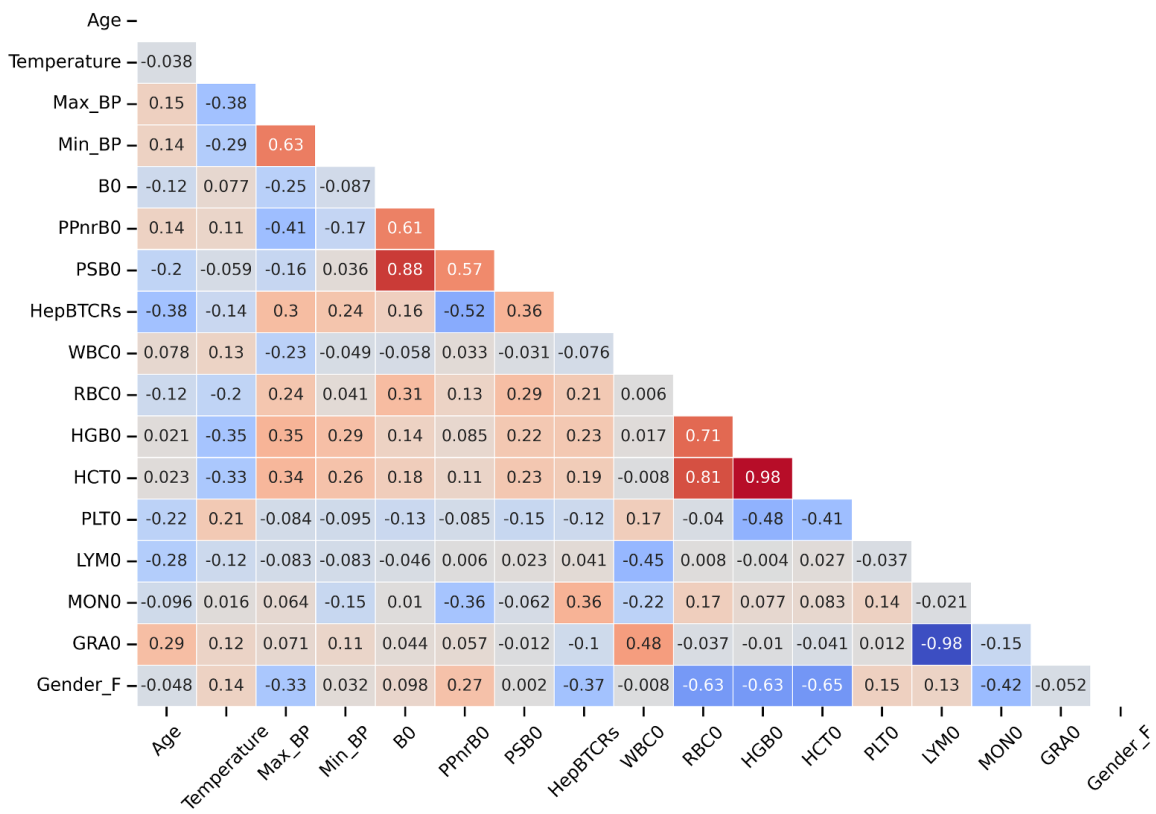

Figure 3. Correlation matrix shows overlap of cohort features. *Data represents the correlation coefficient across each non-genetic feature (see* ***Supplemental material S1*** *for cohort details).*

Individual levels contain elements that correlate with response

After implementing the multi-view integration, we interrogated the data to identify pernative features. As expected, *a*ge is higher for late-converters. Age is also anti-correlated to the anti-HBs titers presented at each timepoint, especially after the first two doses (**Supplemental material S2**). The age distribution differs across all 3 classes (**Figure 4**). Late-converters tend to be older than the other groups **Supplemental material S1**). Within the flow cytometry modality, lymphocyte counts present a positive association with the early response (**Supplemental material S1**). On the other hand, white blood cells and in particular granulocytes are associated with a delayed response (**Supplemental material S3**). The ratio HepBTCR and PSB0 can be associated with an early HBV vaccine response due to the higher HepB specificity of TCRs (**Supplemental material S4**). Just one gene was identified as differentially expressed prior to vaccination when comparing the early and late-converters: HLA-DRB5 (Log_2FoldChange = -4.96, adj. P-value = 2.50e-4 [FDR = 0.05]). This gene presents a very large negative log fold change of ~5, and is down-regulated in the late-converters. HLA-DRB5 encodes for a part of the MHC class II complex, related to antigen presentation.

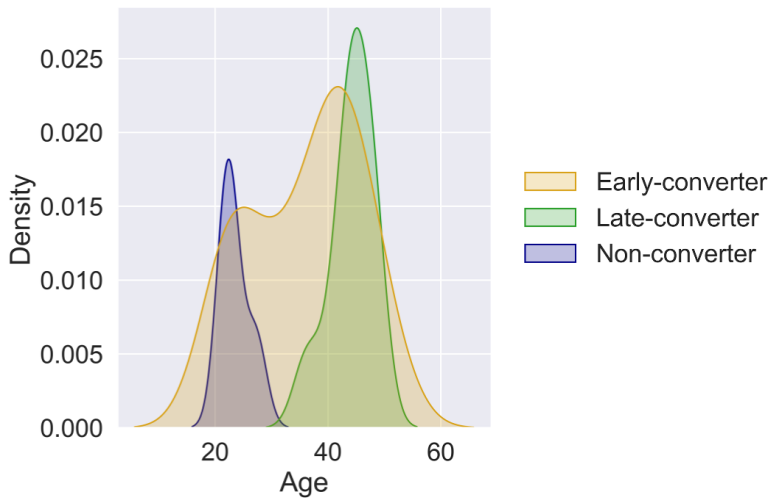

Figure 4. Age distributions reveal differences between classes. *Age distributions per seroconversion class (non-converters were then excluded from the study).*

Projection of data views provides insights into response classes

The different data levels were integrated and projected using two different approaches (**Figure 5**). The first is a PCA, where the principal components are defined by those that capture the most variance across the entire, concatenated, dataset. The first two PCs can be evaluated through visualisation of the samples in the lower dimensionality factor space (**Figure 5A**). The second is MCCA, which aims to detect a common latent structure across multiple datasets by following the SUMCOR definition. The two most informative canonical variates after same-time MCCA integration can then be inspected (**Figure 5B**). The MCCA approach had more prominent separation in the data.

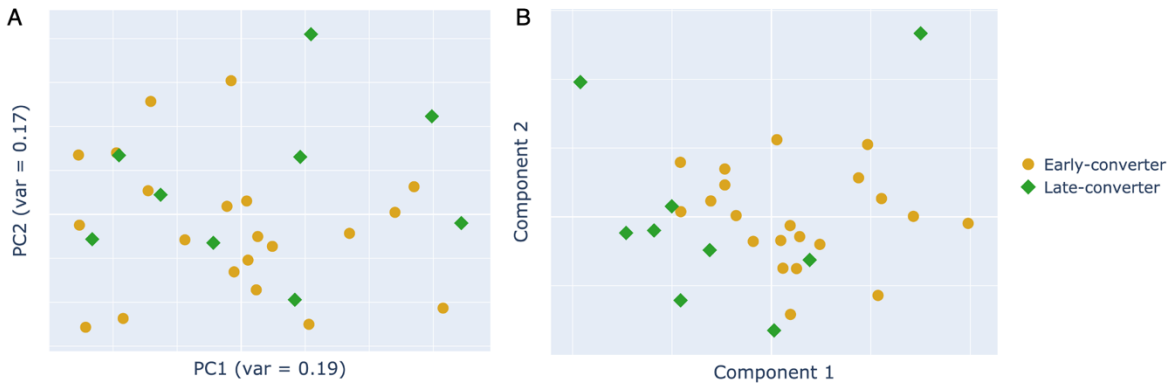

Figure 5. MCCA latent space projection provides superior class separation. *Two-dimensional data projection with (**A**) PCA and (**B**) MCCA of the integrated multi-view vaccine dataset. Each scatter plot showcases the first two dimensions. Data represents the early-converters (orange circle) and late-converters (green diamond) after applying each model.*

As MCCA has better separation of the converter status, it can further be explored by considering the loadings underlying the new dimensions which represent the contribution of each base feature. **Supplemental material S5** shows the stacked coefficients for each of the two first canonical variates of MCCA, grouped by function. These effects are attenuated if another of the implemented techniques is considered, GroupPCA, whose component analysis is shown in **Supplemental material S6**.

There are certain pre-existing features that are associated with vaccine efficacy [19], which include specific gene expression patterns (*e.g.* cytokine expression), immune cells overrepresentation and patient factors (*e.g.* age, gender, comorbidities). Upon inspection of the first component of the MCCA projection, we see that age, oxidative stress, inflammation and interferon modules are all negatively contributing (**Supplemental material S5A**). Lymphocytic BTMs present a negative sign to their coefficients as well. Furthermore, erythroid cells modules and blood cells related counts are relevant for both dimensions and in particular for the second (**Supplemental material S5B**). As can be expected based on the prior pairwise correlations, the higher blood cells counts are also associated with the male gender parameter [39]. The signal given by monocytes is weaker, but it is more consistent for the second component, with the monocyte count and monocyte-related modules having a positive contribution. Finally, CD4 lymphocytes are present in the contribution for the two components both in terms of modules and counts coming from the flow cytometry assays. This compound observation reinforces the importance of lymphocytes for the projection.

Moving to GroupPCA, we can identify, for the first PC (**Supplemental material S6A**), several age related markers. In the present study, inflammation, interferon, oxidative stress and phosphorylation signatures appeared positively tied to ageing. Other positive interactions included higher min and max blood pressure and a higher cell death-driven gene expression as well as lower RBC, HCT and HGB. Finally, the erythroid cells BTMs expression is noticeably relevant for this component. Platelets counts and BTMs appear negatively related to age.

For the second PC (**Supplemental material S6B**), with the decrease in age, we also see lower blood pressure, reduced cell death, oxidative stress and phosphorylation. Diminished T cells and lymphocyte related features seem to signal a less active adaptive immune system. On the other hand, heightened WBC counts, neutrophils and neutrophils activation modules, GRA counts together with the presence of inflammation and modules still show innate immunity characteristics for this dimension.

In comparison to MCCA, GroupPCA revealed distinct modules (**Supplemental material S6**). The top homogenous contributions included interferon and inflammation modules. This was distinct from MCCA (**Supplemental material S5**) where these features were not represented as having coherent contributions. GroupPCA presented a stronger T cell signal. Both methods

assigned rather consistent importance to gene transcription, protein synthesis and protein modification modules for both of their dimensions.

Multi-view integration allows superior classification performance

Thus far, the integration of this data set has been explored in an unsupervised fashion, *i.e.* without prior knowledge of the converter classes. However, as these are known, the various multi-view integration methods can be evaluated in their ability to capture and predict this ground truth. This can be done by feeding the integration output into a supervised machine learning model with a Leave-One-Out-Cross-Validation (LOOCV) approach. Integration approaches that successfully capture latent patterns within the data set that are related to vaccine response, should be able to obtain a high performance.

The classification task was originally skewed with a majority class of 21 early-converters versus a minority class of 9 late-converters. Evaluation of each data layer independently with a LR (**Table 2**), revealed that not all layers held the same predictive capacity. Overall, the TCR layer had the strongest predictive performance on validation data, with both balanced accuracy and AUC measures above 0.73. Conversely, the outputs produced using cell counts or metadata were slightly better than a random classifier, corresponding to a prediction accuracy or AUC marginally above the threshold of 0.5. The model using gene expression data consistently assigns a prediction with confidence (>0.6 decision value; **Supplemental material S7**). This results in a higher AUC but a lower balanced accuracy.

View	Cell counts	TCRseq	metadata	RNA-seq
Accuracy	0.524	0.738	0.508	0.413
AUC	0.407	0.751	0.529	0.619

Table 2. Performance evaluation in terms of accuracy and AUC of a LR trained on single modality data.

To improve the performance of the model, joint dimensionality reduction algorithms were applied. As expected, it improved the LR fitting for both GroupPCA (0.78 AUC; 0.71 accuracy) and MCCA (0.71 AUC; 0.68 accuracy) (**Figure 6A**). GroupPCA surpassed both the single level baselines and the early fusion, while also adding the interpretability highlighted in the previous section. However, using regular PCA methods was not as effective or even counterproductive, causing, in some instances, a drop in performance compared to the best single-view predictors (**Figure 6A**).

The data integration techniques produced consistent results for the more complex model, Random Forest (RF) (**Figure 6B**). Once again and alongside GroupPCA, MCCA presents itself as the best integration method. JDR implementations fell in the range of 0.62-0.8 for both AUC and accuracy, performing overall slightly worse than LR. After such a strong dimensionality reduction this result can be expected. Models are effectively trained on very few features and a simpler method should be sufficient in this circumstance.

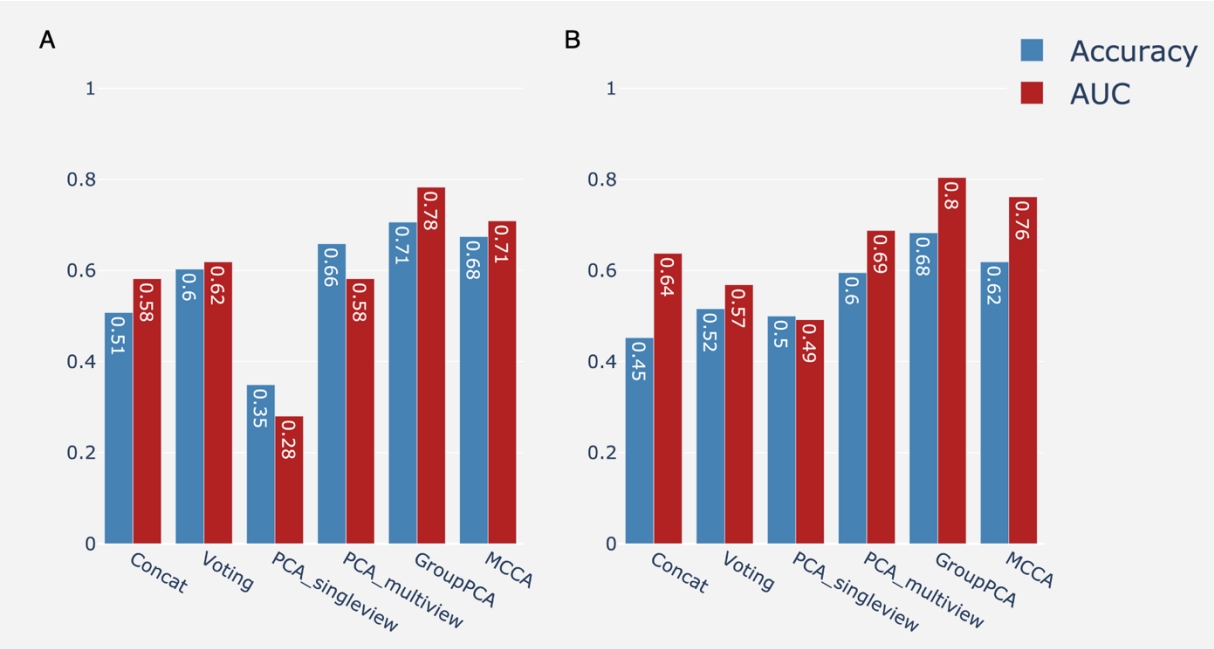

Figure 6. Classifiers performance comparison shows positive impact of joint data integration on predictions. *Methods performance comparison for the (**A**) Logistic Regression and (**B**) Random Forest classifiers in terms of AUC and Accuracy.*

In addition, RF and LR can be easily unravelled to further allow for interpretability of the underlying predictions. The coefficients of the LOOCV LR trained on single data views (**Figure 7**). Old age was the leading factor for a late response (negative class) in the unimodal analysis (**Figure 7C**), which was also apparent in our previous modelling.

Here feature importance within the dataset are showcased, which included granulocyte counts, neutrophils modules and other immune related BTMs linked with late response predictions. These features were distinct from the earlier converters which included the female gender, temperature, pre-existing vaccine specific T-cells (**Figure 7B**), platelets counts and BTMs as well as erythroid cells associated counts and modules. The coefficients for platelet counts, erythroid cells and platelet BTMs in particular, present the highest median for the single-view flow cytometry and RNA-seq models respectively (**Figure 7A and Figure 7D**).

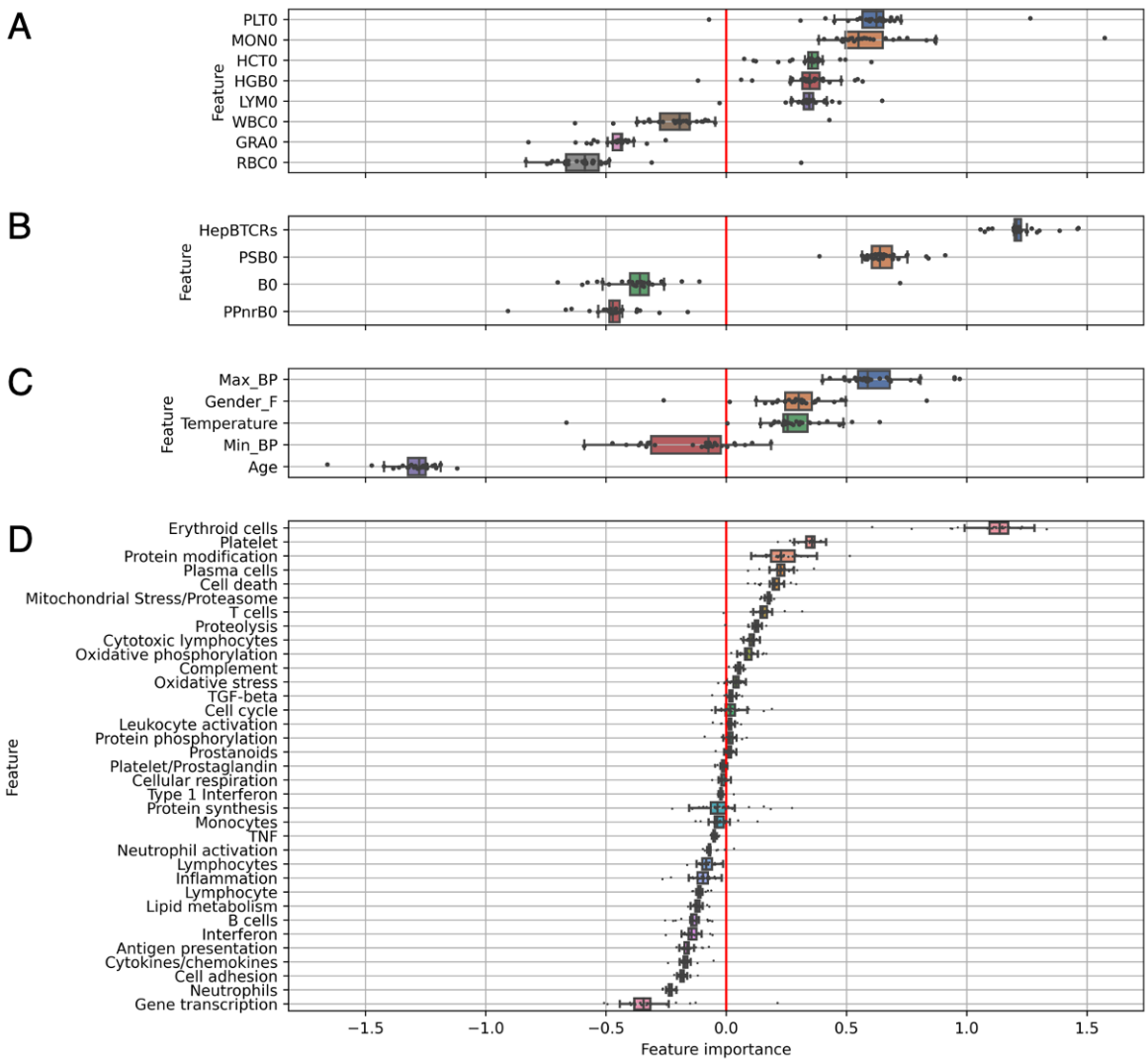

Figure 7. Unimodal feature importance analysis bolsters insights into biological markers tied to vaccine responsiveness. *Importance of leave-one-out cross validated LR coefficients for the unimodal models: (**A**) Cell counts, (**B**) CD4+ T-cells, (**C**) Metadata, (**D**) BTMs. Positive coefficients favour early-response predictions while negative coefficients favour late-response predictions. The vertical red line indicates the zero-importance threshold. BTMs with unknown functionality have been excluded from the visualisation.*

Discussion

Here we showcased the utility of using multi-view analysis to aid in the interoperability of a longitudinal dataset following response to HBV vaccination. This analysis involved incorporating HBV antibody titre levels, RNA-sequencing, CD4+ T cell-receptor, flow cytometry and patient defined features (*e.g. age, gender*). We implemented this multi-view analysis in this setting as to aid in finding interpretable patterns in a vaccination context,

which has not been previously interrogated together. This study bases itself on the notion that prior studies have also highlighted correlations between the different levels and the eventual HBV response, as measured by the anti-HBs titers [19], [24], [40], [41]. To point out the most meaningful approach and improvement of predictive power of multi-view datasets, we used several integration methods. Our approach identified key baseline features associated with earlier and later converters after HBV vaccination. This modelling was able to make informed interpretations of the data and account for confounding factors. For instance, our modelling was able to highlight that age and B-cells were more associated with late-converters.

Age is undoubtedly a major factor in the immune response, especially with “inflammaging” and “immunosenescence” being well documented phenomena [16], [17], [42]. The impact of age is so broad that, for instance, even though monocyte numbers have not been shown to differ in older populations when compared to younger subjects, their phenotype and functionality is considerably different with higher degree of senescent qualities [43]. Inflammation, interferon, oxidative stress and phosphorylation dysfunctions in elderly are also widely confirmed by the literature [44]. In our study, these signatures appear along with higher min and max blood pressure and a higher cell death-driven gene expression. Lower RBC, HCT and HGB, in conjunction with the positive contribution of age, let us infer a biological pattern back-traceable to ageing anaemia and reduced hemoglobins [45].

Granulocytes and lymphocytes appeared mutually exclusive, showing a very strong anti-correlation, a behaviour also detected in the literature [46]. According to [46]–[48], a high neutrophil(granulocytes)/lymphocyte ratio is tied to the inflammatory status of the patient and to an impaired cell-mediated immunity. Conversely, a lower ratio is associated with better outcomes for several diseases, including HBV. Likewise, higher temperatures can have a limiting effect on viral replication and an enhancing role on many immune related functions [49]. Our model contextually showed both patterns as important for conversion classification.

Gender presented itself as another important factor for the interpretation of the results. The male gender, relevant in this study, has previously been shown to have an effect on the presence of pro-inflammatory markers [50]. Additionally, females have been documented to have a better response to HBV booster vaccination [51] , developing greater anti-HBs titers. The models presented here showed the male gender appearing both as relevant in conjunction with inflammation-related features and as one of the most predictive features for early response when controlling for other variables during regression.

For the early response predictions, and in accordance with results for other pathogens [52], [53], pre-existing vaccine specific TCRs contributed positively. Our unimodal analysis, through T-cell gene expression and CD4 counts contributions, seems to suggest that a pre-vaccination priming of the T-cell compartment is conducive to an early vaccination response. Using this

modelling approach, we could also mine that erythroid cells and platelets counts and their respective BTMs were also impactful for an early seroconversion. In fact, platelets counts and BTMs appeared negatively related to age in a behaviour confirmed by the literature [54]. This finding can be interpreted as a reduced number of immune cells simultaneously circulating in the blood, leading to an overall less activated system.

Among the JDR techniques, GroupPCA performed slightly better than MCCA, although surpassing every other method and all four of the single level models in terms of AUC. However, fine tuning of the regularisation or the addition of learned weighting parameters to MCCA's algorithms can boost its performance. Furthermore, MCCA seems to have found a parsimonious representation of heterogeneity in the sample (significant latent space). The separation it produces is more prominent and even if it wasn't perfect (not perfect predictions) we can assume this separation could be even more trustworthy than our labelling.

Our modelling achieved a reasonable level of accuracy and AUC, but did not reach AUC>0.8, which highlights that improvements could be made with further refinements. For instance, a parameter of "importance" can be learned through self-attention mechanisms for each view to weight the contributions differently when applying JDR, instead of applying an imposed regularisation parameter. Starting from the assumption that similarity between samples is different in different views, [55] proposed a smooth representation by using a self-weighting method that enhances similarity grouping effects. An explanation for the reduced AUC could be attributed to incorrect converter classification, as this was based solely on the AB titres. However, our data identified screening for memory CD4+ T cells would identify if someone responded to the vaccination. Additionally, due to the limited understanding of all possible gene-gene networks (gene modules), our modelling could not map the contribution of genes with unknown biological function. Therefore, alternative ways to identify the most pernative gene sets is needed. Future versions of the method will have the capacity to include more gene modules, other than [27] and allow to use first dimension PCA instead of the BTM mean, capturing more in-set variability. To solve the linearity constraint of the combinations, Kernel MCCA (KMCCA) [35] could be used to first project the data in a higher dimensional space via the kernel trick and then apply MCCA to the new feature space. Alternatively, similarity network fusion (SNF) [56] has emerged as a specific data integration technique to learn underlying global and local structures when dealing with patient data [57], [58]. Producing embeddings, with encoders, could help us handle missing values, missing views and not only that, help us even reconstruct these missing aspects. In this field where data is scarce and missing data can be devastating it is extremely relevant. At training time all available views can be used to learn an embedding. During testing or generalisation, the embedding can help reconstruct the missing views' information since it was learned and it was incorporated in the embedding. Successful application of Deep CCA and Deep Canonically Correlated AutoEncoders (DCCAE) [59] prove the feasibility. Although these implementations are

promising, deep learning models had to be excluded at the time of this study, due to the limited sample size.

For the purpose of the analysis, a few drawbacks have to be considered. Such as the non-cohesive nature of the contributions coming from the BTMs. Firstly, the noise of gene expression data hinders full readability of their contributions. The high dimensionality aspect of RNA-seq data made the pattern harder to interpret. However, noisy results were expected when dealing with such noisy data. This problem is not unique to this study but is widely understood to be an unsolved challenge. Secondly, an effect akin to Simpson's paradox [60], [61] caused by multicollinearity, presents itself when inspecting some of MCCA's coefficients. For example, HCT, RBC and HGB are strongly positively correlated and we would expect them to appear with same-sign coefficients in our analysis. Yet, the sign of the relationship is reversed. CD4 lymphocytes are present in the contribution for the two components both in terms of modules and counts coming from the flow cytometry assays, but the aforementioned effect limits their unanimous interpretation. GroupPCA seemingly introduced less of this inversion paradox.

Ultimately, the classification of the response here presented should not be taken at face value. It may not always be representative of other cases for numerous reasons. Some vaccines have been shown to drive primarily a T-cell based response [62], [63]. Furthermore, the sampling timepoints that are standardised and come from older vaccines are not really applicable to mRNA vaccines which have different kinetics [64]. Therefore, they might completely miss some highly relevant changes in the immune state. Most attempts to predict vaccine immune responses are based on the sole antibody titers. This measurement tends to disregard the heterogeneity of the adaptive immune response. Integrating different data sources is a chance to characterise this variability. With the unsupervised approach, JDR has been shown to help uncover a space in which separation between classes of interest is more pronounced. However, the findings are currently restricted to HBV vaccination, and testing in other vaccination settings is needed. We effectively dealt with an unbalanced phase 1 trial-sized population for this analysis. The feature engineering and data integration steps had a decisive effect both in terms of performance and are paramount to gain actionable insights. Ideally a bigger sample should be sought after for the validation setting.

Conclusions

In conclusion, our modelling reflected many aspects of what is known about native HBV infections. Our findings on vaccination data, seem to mirror signatures coming from the literature. The results presented throughout this paper suggest the methodology is valid to identify biologically important features and why multi-level interrogations can aid in finding the most critical features. Our modelling does capture that early converters have more in common with those that respond well to the native infection (based on literature). We thus

recommend testing the modelling on other HBV vaccination cohorts or in other immunological settings to next corroborate the utility of this approach. After validation, we have envisioned this type of pipeline to be useful for guiding rational vaccine discovery, the use of targeted adjuvants and patient counselling, among many other applications.

Declarations

Competing interests

KL and PM hold shares in ImmuneWatch BV, an immunoinformatics company.

Funding

This work was supported by the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" Program.

Authors' contributions

FA performed the study and wrote the manuscript. EB performed the original study on hepatitis B vaccine response. KM supervised and revised the manuscript. KL and PM conceived and supervised the study and revised the manuscript. PVD, PB and BO revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

George Elias, Geert Mortier, Arvid Suls, Viggo Van Tenderloo & Hilde Jansens.

References

[1] J.-M. Anaya, Y. Shoenfeld, A. Rojas-Villarraga, R. A. Levy, and R. Cervera, Eds., *Autoimmunity: From Bench to Bedside*. Bogota (Colombia): El Rosario University Press, 2013. Accessed: Jan. 09, 2023. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK459447/

[2] B. Pulendran, S. Li, and H. I. Nakaya, 'Systems Vaccinology', *Immunity*, vol. 33, no. 4, pp. 516–529, Oct. 2010, doi: 10.1016/j.immuni.2010.10.006.

[3] B. Pulendran, 'Systems vaccinology: Probing humanity's diverse immune systems with vaccines', *Proc. Natl. Acad. Sci.*, vol. 111, no. 34, pp. 12300–12306, Aug. 2014, doi: 10.1073/pnas.1400476111.

[4] D. Furman and M. M. Davis, 'New approaches to understanding the immune response to vaccination and infection', *Vaccine*, vol. 33, no. 40, pp. 5271–5281, Sep. 2015, doi: 10.1016/j.vaccine.2015.06.117.

[5] J. Zhao, X. Xie, X. Xu, and S. Sun, 'Multi-view learning overview: Recent progress and new challenges', *Inf. Fusion*, vol. 38, pp. 43–54, Nov. 2017, doi: 10.1016/j.inffus.2017.02.007.

[6] H. Hotelling, 'Relations Between Two Sets of Variates', *Biometrika*, vol. 28, no. 3/4, p. 321, Dec. 1936, doi: 10.2307/2333955.

[7] B. Lei, S. Chen, D. Ni, and T. Wang, 'Discriminative Learning for Alzheimer's Disease Diagnosis via Canonical Correlation Analysis and Multimodal Fusion', *Front. Aging Neurosci.*, vol. 8, May 2016, doi: 10.3389/fnagi.2016.00077.

[8] A. Singanamalli *et al.*, 'Cascaded Multi-view Canonical Correlation (CaMCCo) for Early Diagnosis of Alzheimer's Disease via Fusion of Clinical, Imaging and Omic Features', *Sci. Rep.*, vol. 7, no. 1, p. 8137, Aug. 2017, doi: 10.1038/s41598-017-03925-0.

[9] C. Zhang, E. Adeli, T. Zhou, X. Chen, and D. Shen, 'Multi-Layer Multi-View Classification for Alzheimer's Disease Diagnosis', *Proc. AAAI Conf. Artif. Intell. AAAI Conf. Artif. Intell.*, vol. 2018, pp. 4406–4413, Feb. 2018.

[10] X. Zhuang, Z. Yang, and D. Cordes, 'A technical review of canonical correlation analysis for neuroscience applications', *Hum. Brain Mapp.*, vol. 41, no. 13, pp. 3807–3833, Sep. 2020, doi: 10.1002/hbm.25090.

[11] J. Yang *et al.*, 'Multi-view learning for lymph node metastasis prediction using tumor and nodal radiomics in gastric cancer', *Phys. Med. Biol.*, vol. 67, no. 5, p. 055007, Mar. 2022, doi: 10.1088/1361-6560/ac515b.

[12] K. Pearson, 'LIII. *On lines and planes of closest fit to systems of points in space*', *Lond. Edinb. Dublin Philos. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, Nov. 1901, doi: 10.1080/14786440109462720.

[13] I. T. Jolliffe, *Principal Component Analysis*. New York, NY: Springer New York, 1986. doi: 10.1007/978-1-4757-1904-8.

[14] J. R. Kettenring, 'Canonical analysis of several sets of variables', *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971, doi: 10.1093/biomet/58.3.433.

[15] R. Argelaguet *et al.*, 'Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets', *Mol. Syst. Biol.*, vol. 14, no. 6, Jun. 2018, doi: 10.15252/msb.20178124.

[16] L. Ferrucci and E. Fabbri, 'Inflammageing: chronic inflammation in ageing, cardiovascular disease, and frailty', *Nat. Rev. Cardiol.*, vol. 15, no. 9, pp. 505–522, Sep. 2018, doi: 10.1038/s41569-018-0064-2.

[17] C. Franceschi, P. Garagnani, P. Parini, C. Giuliani, and A. Santoro, 'Inflammaging: a new immune–metabolic viewpoint for age-related diseases', *Nat. Rev. Endocrinol.*, vol. 14, no. 10, pp. 576–590, Oct. 2018, doi: 10.1038/s41574-018-0059-4.

[18] S. M. M. Rasa *et al.*, 'Inflammaging is driven by upregulation of innate immune receptors and systemic interferon signaling and is ameliorated by dietary restriction', *Cell Rep.*, vol.

39, no. 13, p. 111017, Jun. 2022, doi: 10.1016/j.celrep.2022.111017.
[19] E. Bartholomeus *et al.*, 'Transcriptome profiling in blood before and after hepatitis B vaccination shows significant differences in gene expression between responders and non-responders', *Vaccine*, vol. 36, no. 42, pp. 6282–6289, Oct. 2018, doi: 10.1016/j.vaccine.2018.09.001.
[20] T. D. Querec *et al.*, 'Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans', *Nat. Immunol.*, vol. 10, no. 1, pp. 116–125, Jan. 2009, doi: 10.1038/ni.1688.
[21] J. S. Tsang *et al.*, 'Global Analyses of Human Immune Variation Reveal Baseline Predictors of Postvaccination Responses', *Cell*, vol. 157, no. 2, pp. 499–513, Apr. 2014, doi: 10.1016/j.cell.2014.03.031.
[22] J. S. Tsang *et al.*, 'Improving Vaccine-Induced Immunity: Can Baseline Predict Outcome?', *Trends Immunol.*, vol. 41, no. 6, pp. 457–465, Jun. 2020, doi: 10.1016/j.it.2020.04.001.
[23] A. Koussounadis, S. P. Langdon, I. H. Um, D. J. Harrison, and V. A. Smith, 'Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system', *Sci. Rep.*, vol. 5, no. 1, p. 10775, Jun. 2015, doi: 10.1038/srep10775.
[24] G. Elias *et al.*, 'Preexisting memory CD4 T cells in naïve individuals confer robust immunity upon hepatitis B vaccination', *eLife*, vol. 11, p. e68388, Jan. 2022, doi: 10.7554/eLife.68388.
[25] M. I. Love, W. Huber, and S. Anders, 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biol.*, vol. 15, no. 12, p. 550, Dec. 2014, doi: 10.1186/s13059-014-0550-8.
[26] S. Bailey, 'Principal Component Analysis with Noisy and/or Missing Data', *Publ. Astron. Soc. Pac.*, vol. 124, no. 919, pp. 1015–1023, Sep. 2012, doi: 10.1086/668105.
[27] M. C. Altman *et al.*, 'Development of a fixed module repertoire for the analysis and interpretation of blood transcriptome data', *Nat. Commun.*, vol. 12, no. 1, p. 4385, Jul. 2021, doi: 10.1038/s41467-021-24584-w.
[28] R. Blagus and L. Lusa, 'SMOTE for high-dimensional class-imbalanced data', *BMC Bioinformatics*, vol. 14, no. 1, p. 106, Dec. 2013, doi: 10.1186/1471-2105-14-106.
[29] H.-A. Park, 'An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain', *J. Korean Acad. Nurs.*, vol. 43, no. 2, p. 154, 2013, doi: 10.4040/jkan.2013.43.2.154.
[30] R. D. Markello, G. Shafiei, C. Tremblay, R. B. Postuma, A. Dagher, and B. Misic, 'Multimodal phenotypic axes of Parkinson's disease', *Npj Park. Dis.*, vol. 7, no. 1, p. 6, Jan. 2021, doi: 10.1038/s41531-020-00144-9.
[31] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, 'On combining classifiers', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998, doi: 10.1109/34.667881.
[32] I. T. Jolliffe, N. T. Trendafilov, and M. Uddin, 'A Modified Principal Component Technique Based on the LASSO', *J. Comput. Graph. Stat.*, vol. 12, no. 3, pp. 531–547, Sep. 2003, doi: 10.1198/1061860032148.
[33] H. Zou, T. Hastie, and R. Tibshirani, 'Sparse Principal Component Analysis', *J. Comput. Graph. Stat.*, vol. 15, no. 2, pp. 265–286, Jun. 2006, doi: 10.1198/106186006X113430.
[34] R. Tibshirani, 'Regression Shrinkage and Selection Via the Lasso', *J. R. Stat. Soc. Ser. B Methodol.*, vol. 58, no. 1, pp. 267–288, Jan. 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.
[35] R. Perry *et al.*, 'mvlearn: Multiview Machine Learning in Python', *J. Mach. Learn. Res.*, vol. 22, no. 109, pp. 1–7, 2021.
[36] A. Tenenhaus and M. Tenenhaus, 'Regularized Generalized Canonical Correlation Analysis', *Psychometrika*, vol. 76, no. 2, pp. 257–284, Apr. 2011, doi: 10.1007/s11336-011-9206-8.
[37] G. Lee *et al.*, 'Supervised Multi-View Canonical Correlation Analysis (sMVCCA): Integrating Histologic and Proteomic Features for Predicting Recurrent Prostate Cancer', *IEEE Trans. Med. Imaging*, vol. 34, no. 1, pp. 284–297, Jan. 2015, doi:

10.1109/TMI.2014.2355175.
[38] D. M. Witten and R. J. Tibshirani, 'Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data', *Stat. Appl. Genet. Mol. Biol.*, vol. 8, no. 1, pp. 1–27, Jan. 2009, doi: 10.2202/1544-6115.1470.
[39] W. G. Murphy, 'The sex difference in haemoglobin levels in adults — Mechanisms, causes, and consequences', *Blood Rev.*, vol. 28, no. 2, pp. 41–47, Mar. 2014, doi: 10.1016/j.blre.2013.12.003.
[40] Y. Tian *et al.*, 'Immune Response to Hepatitis B Virus Vaccine Among People Living With HIV: A Meta-Analysis', *Front. Immunol.*, vol. 12, p. 745541, Dec. 2021, doi: 10.3389/fimmu.2021.745541.
[41] P. Basireddy, S. Avileli, N. Beldono, and S. L. Gundela, 'Evaluation of Immune Response to Hepatitis B Vaccine in Healthcare Workers at a Tertiary Care Hospital', *Indian J. Med. Microbiol.*, vol. 36, no. 3, pp. 397–400, Jul. 2018, doi: 10.4103/ijmm.IJMM_17_431.
[42] C. Franceschi *et al.*, 'Inflamm-aging: An Evolutionary Perspective on Immunosenescence', *Ann. N. Y. Acad. Sci.*, vol. 908, no. 1, pp. 244–254, Jan. 2006, doi: 10.1111/j.1749-6632.2000.tb06651.x.
[43] R. P. H. De Maeyer and E. S. Chambers, 'The impact of ageing on monocytes and macrophages', *Immunol. Lett.*, vol. 230, pp. 1–10, Feb. 2021, doi: 10.1016/j.imlet.2020.12.003.
[44] W. Cao, 'IFN-Aging: Coupling Aging With Interferon Response', *Front. Aging*, vol. 3, p. 870489, May 2022, doi: 10.3389/fragi.2022.870489.
[45] D. L. Smith, 'Anemia in the elderly', *Am. Fam. Physician*, vol. 62, no. 7, pp. 1565–1572, Oct. 2000.
[46] M. Song, B. I. Graubard, C. S. Rabkin, and E. A. Engels, 'Neutrophil-to-lymphocyte ratio and mortality in the United States general population', *Sci. Rep.*, vol. 11, no. 1, p. 464, Jan. 2021, doi: 10.1038/s41598-020-79431-7.
[47] 'Clinical Predictors for Neutrophil–to–Lymphocyte Ratio Changes in Patients with Chronic Hepatitis B Receiving Peginterferon Treatment', *In Vivo*, vol. 31, no. 4, pp. 723–729, Jun. 2017, doi: 10.21873/invivo.11121.
[48] K. Liu, F.-S. Wang, and R. Xu, 'Neutrophils in liver diseases: pathogenesis and therapeutic targets', *Cell. Mol. Immunol.*, vol. 18, no. 1, pp. 38–44, Jan. 2021, doi: 10.1038/s41423-020-00560-0.
[49] S. S. Evans, E. A. Repasky, and D. T. Fisher, 'Fever and the thermal regulation of immunity: the immune system feels the heat', *Nat. Rev. Immunol.*, vol. 15, no. 6, pp. 335–349, Jun. 2015, doi: 10.1038/nri3843.
[50] J. C. Milan-Mattos *et al.*, 'Effects of natural aging and gender on pro-inflammatory markers', *Braz. J. Med. Biol. Res.*, vol. 52, no. 9, p. e8392, 2019, doi: 10.1590/1414-431x20198392.
[51] A. Trevisan, C. Frasson, D. De Nuzzo, A. Nicolli, and M. L. Scapellato, 'Significance of anti-HB levels below 10 IU/L after vaccination against hepatitis B in infancy or adolescence: an update in relation to sex', *Hum. Vaccines Immunother.*, vol. 16, no. 2, pp. 460–464, Feb. 2020, doi: 10.1080/21645515.2019.1656483.
[52] E. J. Yager, M. Ahmed, K. Lanzer, T. D. Randall, D. L. Woodland, and M. A. Blackman, 'Age-associated decline in T cell repertoire diversity leads to holes in the repertoire and impaired immunity to influenza virus', *J. Exp. Med.*, vol. 205, no. 3, pp. 711–723, Mar. 2008, doi: 10.1084/jem.20071140.
[53] S. M. Murray *et al.*, 'The impact of pre-existing cross-reactive immunity on SARS-CoV-2 infection and vaccine responses', *Nat. Rev. Immunol.*, Dec. 2022, doi: 10.1038/s41577-022-00809-x.
[54] C. I. Jones, 'Platelet function and ageing', *Mamm. Genome*, vol. 27, no. 7–8, pp. 358–366, Aug. 2016, doi: 10.1007/s00335-016-9629-8.
[55] J. Liu, S. Ge, Y. Cheng, and X. Wang, 'Multi-View Spectral Clustering Based on Multi-Smooth Representation Fusion for Cancer Subtype Prediction', *Front. Genet.*, vol. 12, p. 718915, Sep. 2021, doi: 10.3389/fgene.2021.718915.
[56] B. Wang *et al.*, 'Similarity network fusion for aggregating data types on a genomic scale',

Nat. Methods, vol. 11, no. 3, pp. 333–337, Mar. 2014, doi: 10.1038/nmeth.2810.
[57] M. Chierici *et al.*, 'Integrative Network Fusion: A Multi-Omics Approach in Molecular Profiling', *Front. Oncol.*, vol. 10, p. 1065, Jun. 2020, doi: 10.3389/fonc.2020.01065.
[58] S. Pai *et al.*, 'netDx: Interpretable patient classification using integrated patient similarity networks', *bioRxiv*, p. 084418, Nov. 2017, doi: 10.1101/084418.
[59] W. Wang, R. Arora, K. Livescu, and J. Bilmes, 'On Deep Multi-View Representation Learning: Objectives and Optimization', 2016, doi: 10.48550/ARXIV.1602.01024.
[60] E. H. Simpson, 'The Interpretation of Interaction in Contingency Tables', *J. R. Stat. Soc. Ser. B Methodol.*, vol. 13, no. 2, pp. 238–241, Jul. 1951, doi: 10.1111/j.2517-6161.1951.tb00088.x.
[61] Y.-K. Tu, D. Gunnell, and M. S. Gilthorpe, 'Simpson's Paradox, Lord's Paradox, and Suppression Effects are the same phenomenon – the reversal paradox', *Emerg. Themes Epidemiol.*, vol. 5, no. 1, p. 2, Dec. 2008, doi: 10.1186/1742-7622-5-2.
[62] S. C. Gilbert, 'T-cell-inducing vaccines - what's the future: T-cell vaccines', *Immunology*, vol. 135, no. 1, pp. 19–26, Jan. 2012, doi: 10.1111/j.1365-2567.2011.03517.x.
[63] P. Moss, 'The T cell immune response against SARS-CoV-2', *Nat. Immunol.*, vol. 23, no. 2, pp. 186–193, Feb. 2022, doi: 10.1038/s41590-021-01122-w.
[64] A. Y. Collier *et al.*, 'Differential Kinetics of Immune Responses Elicited by Covid-19 Vaccines', *N. Engl. J. Med.*, vol. 385, no. 21, pp. 2010–2012, Nov. 2021, doi: 10.1056/NEJMc2115596.