Expotype Clustering

Hannah Cooper, Tianshu Lu, Imogen Onno, Calix Tan

Abstract

The exposome represents the totality of environmental and lifestyle exposures impacting human health, yet traditional methods often overlook their joint effects with biological mechanisms. We aimed to identify exposure-driven clusters (expotypes) in 41,804 UK Biobank participants, characterise their proteomic signatures, and examine associations with chronic disease risk. Analysing 28 exposures, 1,343 proteins, and demographics - we used unsupervised clustering, stability selection LASSO, logistic regression, and pathway enrichment analysis, to identify and explore seven distinct expotypes. Clusters 5 and 6 were characterised by healthier exposure profiles and lower disease risk, while clusters 3 and 4 exhibited unhealthy profiles and higher incidence of chronic conditions. Our findings highlight that both lifestyle exposures and proteomic profiles jointly contribute to disease risk, supporting the potential for exposome-informed risk stratification and prevention.

Introduction

Understanding the factors that influence human health requires an assessment of exposures, which are the environmental, behavioral and social factors encountered by an individual over their lifetime. These exposures are known to play a central role in the development and progression of many chronic diseases [1]. Unlike genetic factors, exposures are often modifiable, making them crucial targets for public health interventions.

Because of this, there has been a wide range of research on exposures and their effect on the development of chronic diseases. Traditional epidemiological studies often assess the link between exposure and disease by considering each exposure separately. Although this commonly used technique helps to understand the effect of a single exposure, it fails to account for the complex interactions between exposures and the effect of multiple exposures on disease, consequently underestimating true disease risk and obscuring potentially relevant biological pathways [2]. One solution to this is to explore the exposome.

The exposome, first defined by Wild in 2005, encompasses the totality of environmental, lifestyle, and occupational exposures an individual experiences from conception throughout life, along with their corresponding biological responses [3]. This concept acknowledges that chronic diseases often arise not from single exposures in isolation, but from complex, dynamic interactions among multiple external and internal factors.

To address this, we decided to integrate data on external exposures, such as diet, air pollution, and social environment, with internal proteomic signatures, using the UK Biobank cohort study.

In our study, we focused on five chronic health outcomes that represent a significant burden on public health globally: Alzheimer's Disease (AD), Parkinson's Disease (PD), Chronic Kidney Disease (CKD), Coronary Artery Disease (CAD), and Diabetes. These diseases are highly prevalent and contribute substantially to morbidity, mortality, and healthcare costs [4]. At the same time, they are known to be influenced by both environmental exposures and underlying biological mechanisms.

To capture the complexity of exposure profiles and their biological consequences, we employ unsupervised clustering techniques. Without any prior assumptions, this approach allows us to identify distinct groups in order to capture hidden patterns and co-exposures that drives the difference among participants. By linking these clusters with proteomic signatures and disease incidence, we aim to unravel how different exposure environments and distinct biological pathways leads to disease risks.

Therefore, we proposed two central research questions: How do exposures and proteomic signatures vary across expotypes in UK Biobank participants, and what biological pathways distinguish these clusters? Are the discovered expotypes from UK Biobank participants associated with different incidence and risk for chronic health outcomes?

Methods

The UK Biobank study collected data on over 500,000 adults at assessment centres across the UK from 2006 to 2010. The study measured lifestyle and physical characteristics through questionnaires, interviews and biological samples. These records are linked to hospital inpatient data including diagnoses and underlying conditions for each participant [5].

We analysed 28 exposure variables, 1343 protein levels and demographics (age, sex, ethnic background) of 41,804 UK Biobank participants. The exposures range across five domains: Behavioural, Childhood, Environmental, Psychosocial and Socio-Economic. Descriptive statistics of all variables are provided in Table 1 (Figures A1 to 3). The Metabolic Equivalent Minute (MET) score summarises physical activity using the IPAQ scoring protocol [6, 7]. The diet score summarises food intake and evaluates a score based on recommended intake levels [8]. We used a sleep score (0-3) with higher scores reflecting better sleeping habits [9, 10]. We also devised a neuroticism score to quantify individuals' levels of neurotic behaviour, with higher score indicating greater neuroticism. Protein data were obtained from participants' blood plasma samples using the antibody-based Olink Explore Proximity Extension Assay (PEA) [11].

Participants with five or more missing values were excluded. For the remaining data, missing exposures values were imputed using the *missRanger* package [12]. We then removed any participant with outlier values, defined as values more than three standard deviations above or below the mean for a given exposure. The dataset was split 50:30:20 for variable selection, training and testing - with the test data scaled on the training and selection to prevent data leakage.

Clustering

We implemented unsupervised clustering methods to identify groups of UK Biobank participants with similar lived experiences, excluding demographic variables. Due to the unknown underlying cluster structure, a range of methods were applied. First, partitional methods: Biclustering, with the *biclust* package [13], simultaneously groups exposures and individuals to reveal local patterns, whilst Fuzzy Clustering, with the *cluster* package [14], allows points to belong to multiple clusters, accommodating uncertain and overlapping distributions. Next, model-based approaches: Gaussian Mixture Models (GMM), using the *mclust* package [15], represent data as a combination of multiple Gaussian distributions, allowing membership probabilities to be estimated, while high-dimensional data clustering (HDDC), using the *HDclassif* package [16], extends this by fitting each cluster in its own low-dimensional subspace, improving stability and scalability when handling many exposures. Finally, a combination of the dimensionality reduction technique Self Organising Maps (SOM), with the *kohonen* package [17], which is an artificial neural network that reduces dimensionality and represents the distributions on a map, and hierarchical clustering, using the *factoextra* package [18], which forms dendrograms, allows for a clear summary of how clusters are formed and related.

The performance of each technique was assessed using the Proportion of Ambiguous Clustering (PAC) score, which evaluates cluster stability across resampling iterations, where lower scores represent greater stability. Once the preferred clustering method had been selected, the optimal number of clusters was determined by balancing minimisation of the Bayesian Information Criterion (BIC), which assesses model fit whilst penalising complexity, with maximisation of the Silhouette score, which measures cluster cohesion and separation.

Cluster Description

Having clustered the participants by their lifestyle exposures, we identified which exposures were associated with each cluster. For this analysis, a separate outcome variable Y_i was used for each cluster. Each binary outcome variable indicated assignment to a specified cluster i against assignment to any of the other clusters.

$$Y_i = \begin{cases} 1 & \text{Allocation to cluster } i \\ 0 & \text{Allocation to any other cluster} \end{cases}$$

First, as a univariate analysis, the association of every exposure and cluster indicator was tested via logistic regression models. Each model was adjusted for age, sex and ethnic background. Then, stability selection LASSO was used

to understand the expotype signatures of participants in each cluster. The devised stability score, the negative log-likelihood of the model against the null hypothesis of equiprobability of selection, is maximised in this approach [19]. Two parameters are calibrated for this optimisation: selection proportion cutoff (π) and penalty factor (λ). The resulting profiles include lifestyle exposures that are jointly and stably associated with each cluster allocation. Finally, a logistic regression model was fit for each cluster indicator, using the stably selected exposures for each model. Each model was adjusted for age, sex and ethnic background.

Molecular Profiling

After characterising the clusters by exposures, we identified the proteins associated with each cluster. The protein analysis followed very similar methods to the cluster description analysis, focusing on proteins rather than exposures. The same binary outcome variable Y_i was used, indicating assignment to a specified cluster i against assignment to any of the other clusters.

For the univariate analysis, we tested the association between each protein and cluster indicator using logistic regression models. This resulted in 9401 models, one for each protein and cluster combination. Each model was adjusted for age, sex, and ethnic background. Manhattan plots (Figures A25 to 31) highlighted the significantly associated proteins for each cluster.

We then utilised stability selection LASSO to stably select a group of proteins that are associated with the participants in each cluster, creating a molecular signature for each. The calibration plots and selection proportion threshold plots are provided (Figures A32 to 38). Lastly, we fit a logistic regression model for each cluster, using the stably selected proteins specific to each cluster as the predictors. Both the stability selection LASSO and logistic regression models were adjusted for age, sex, and ethnic background.

Pathway Enrichment Analysis

After selecting the proteins, pathway analysis was performed for each cluster to discover distinct biological mechanisms driving each group. Proteins were input into *Reactome*, a pathway database which provides intuitive bioinformatics tools for the visualisation, interpretation, and analysis of pathway knowledge [20]. Here, many biological pathways were tested to see if they were statistically enriched (overrepresented) in the selected proteins. Then, the enriched pathways were generated and visualised.

Both the selected proteins and the total known proteins in each pathway were considered. Through pathway analysis, it is determined whether the overlap between the selected proteins and the pathway was greater than would be expected by chance. Pathways with low FDR (<0.05), the adjusted p-value, were considered more reliably associated and the three lowest were chosen. Further literature review was conducted to investigate which biological mechanisms the pathways were associated with.

Disease Incidence and Prediction

To identify clusters with higher relative disease burden, standardised disease incidence rates were calculated to enable comparison across clusters of varying sizes. For each cluster, the disease incidence rate was calculated as the number of incident cases divided by the total number of individuals in that cluster. To compare relative incidence patterns across clusters within each disease, z-score standardisation was applied to the cluster-level rates by subtracting the disease-specific mean and dividing by the corresponding standard deviation.

Logistic regression models were developed to assess predictive capability for chronic disease outcomes. The dataset was split as described before, where 30% was used for model training and the remaining 20% reserved for evaluating model performance.

Cluster membership was used as the initial predictor, with zero-sum coding applied to ensure that all clusters were included in the model output without requiring a single reference category. This coding approach defined the overall mean odds across all clusters as the reference, allowing each cluster's effect to be interpreted relative to the average.

Stably selected exposures and proteins were sequentially added as additional predictors. All models were adjusted for potential confounders, including age, sex, and ethnic background. Predictive performance for each model and disease was evaluated using the Area Under the Receiver Operative Characteristic Curve (AUC).

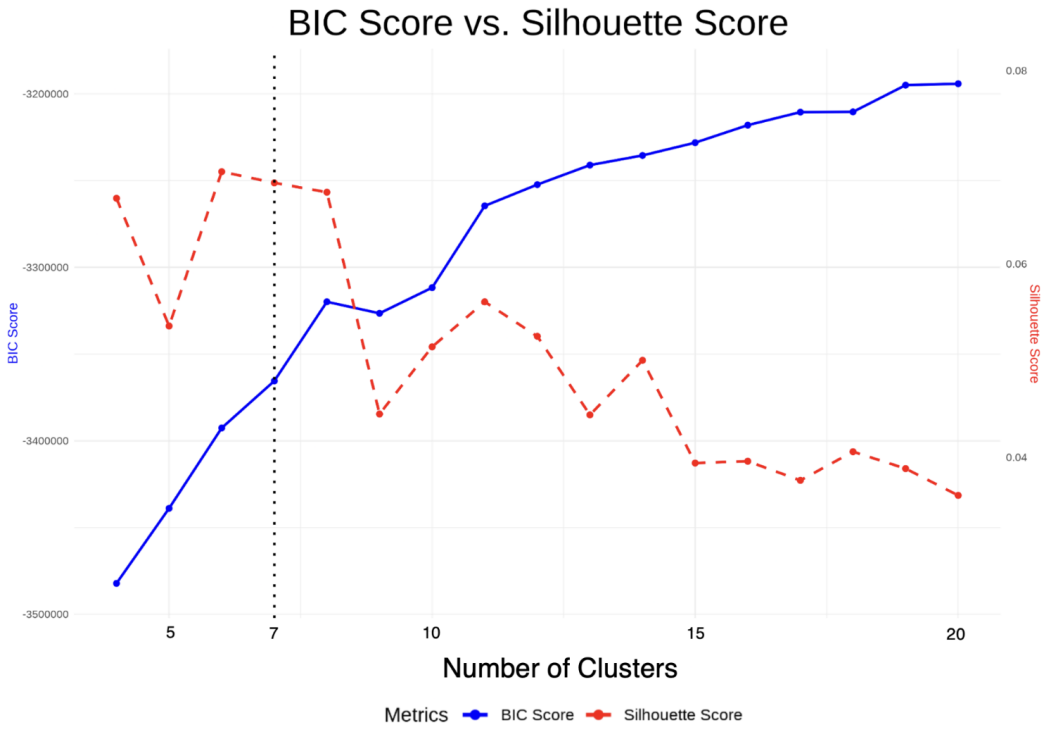

Figure 1: BIC and Silhouette scores across different cluster counts. The optimal number of clusters was identified at 7, balancing model fit and clustering quality.

Results

Clustering

GMM was selected as the preferred clustering method, achieving a low PAC score of 0.20 and clustering all individuals. Biclustering produced the most stable clusters with a PAC score of 0.059, however, it only clustered 60% of the data and was therefore excluded. In contrast, Fuzzy Clustering and SOM combined with Hierarchical Clustering demonstrated poor performance, with PAC scores of 0.531 and 0.923, respectively, indicating unstable clusters across resampling iterations. HDDC could not be run successfully due to convergence issues and was therefore excluded from further analysis.

Despite the greater distance between the BIC and Silhouette scores at six clusters, we determined that seven clusters was optimal for the GMM, providing a good trade-off between model fit and cluster separation (BIC: -3,250,000; Silhouette: 0.047), while allowing for an additional cluster to improve subgroup representation (Figure 1).

To visualise our clusters in two dimensions, we used three dimensionality reduction techniques. These included Uniform Manifold Approximation and Projection (UMAP) using the *umap* package [21] as shown in Figure 2, t-distributed Stochastic Neighbor Embedding (t-SNE) using the *Rtsne* package [22] as shown in Figure A5, and Principal Component Analysis (PCA) as shown in Figure A4. After plotting all participants, we colored each point according to its assigned cluster to highlight the groupings visually. Contour lines were added to the UMAP plot to illustrate density and boundaries of the cluster more clearly.

Cluster Description

The univariate analysis indicated higher levels of exposure association with clusters 3 to 7. Using the Bonferroni-adjusted threshold to account for multiple testing, these clusters had between 32 and 41 significant associations across all domains. Clusters 1 and 2 had fewer significant associations with 13 and 9 respectively. The Manhattan

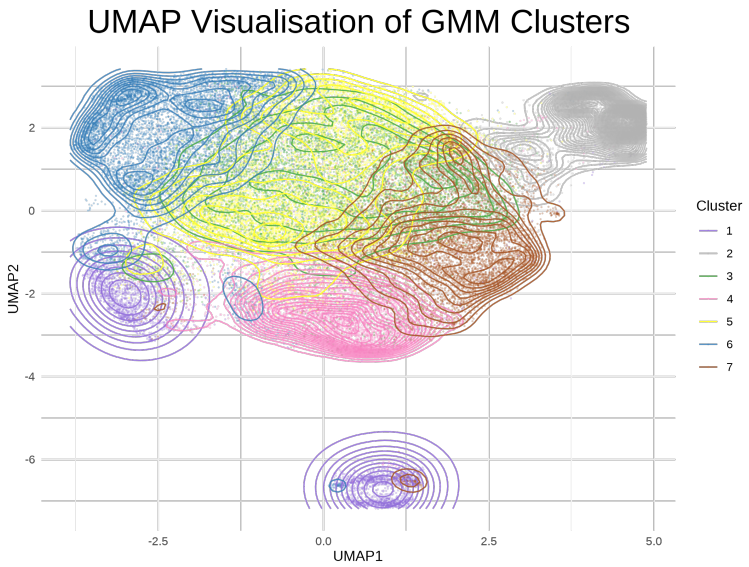

Figure 2: UMAP visualisation of clusters

plots from this analysis are provided (Figures A11 to 17). The total number of associations is shown in Figure A6 using a stacked bar chart to display domain-specific associations. Clusters 3 to 6 had associations across all five exposure domains and the environmental and socio-economic domains were associated with every cluster.

The stability selection LASSO reflected the patterns seen. The selection proportion plots and calibration plots for each model are provided (Figures A18 to 24). Environmental exposures including traffic intensity and land use were the most commonly selected across all clusters. Behavioural variables, including pack years, exercise and sleep score, were also stably selected across multiple clusters. Figure A7 summarises the selected variables across all selection models using a tile map.

The odds ratios (OR) for the stably selected exposures for each cluster allocation are displayed in Figure 3. Exposures that were selected across almost all clusters are associated with increased odds to one cluster (OR $>$ 1) and decreased odds with the remaining clusters. This is observed for the following variables: green space (1000m), water (1000m), inverse distance to major road, traffic intensity and pack years. This indicates that these exposures are useful to clearly indicate cluster allocation.

Clusters 5 and 6 have healthier exposure profiles. Increased sleep score [OR 1.82 (95% CI 1.71-1.93)] and diet scores [OR 1.31 (1.23-1.39)] and lower pack years [OR 0.39 (0.35-0.42)] are associated with higher odds of allocation to cluster 5. The odds of cluster 6 allocation are increased for participants living in areas with a high proportion of greenspace in the surrounding 1000m [OR 14.4 (11.3-18.2)] and higher exercise [OR 1.39 (1.25-1.54)]. Cluster 4's exposure profile is less healthy – increased pack years [OR 13.5 (11.9-15.4)], worse sleep patterns [OR 0.77 (0.70-0.85)] and lower diet score [OR 0.47 (0.42-0.52)] are associated with higher odds of allocation. Cluster 7 also has a less healthy exposure profile with higher pollution levels, NO_2 [OR 2.81 (2.30-3.43)] and $PM_{2.5}$ [OR 7.54 (6.35-8.95)], increasing odds of allocation. Allocation to clusters 4 [OR 1.80 (1.63-2.00)] and 7 [OR 4.16 (3.72-4.65)] is associated with IMD score, with higher deprivation increasing the odds of allocation.

Cluster 3 has a more unique expotype profile. Allocation is associated with anxiety [OR 2.15 (1.78-2.60)], neuroticism score [OR 4.92 (4.51-5.38)] and distance to the coast [OR 1.38 (1.29-1.47)]. Allocation to cluster 1 is associated with the inverse distance to the nearest major road [OR 21.2 (16.7-27.0)], the proportion of water in the surrounding 1000m [OR 19.7 (15.6-24.9)] and MET score [OR 2.35 (2.02-2.73)]. Cluster 2 allocation is associated with MET score [OR 1.88 (1.50-2.36)], traffic intensity [OR 129.3 (70.2-238.2)] and the proportion of green space in the surrounding 1000m [OR 1.32 (0.89, 1.96)].

The ROC curves (Figure A8) indicate that each logistic regression model is very successful at indicating cluster assignment. All AUC values are greater than 0.9 and clusters 1, 2 and 6 have AUC$>$ 0.99.

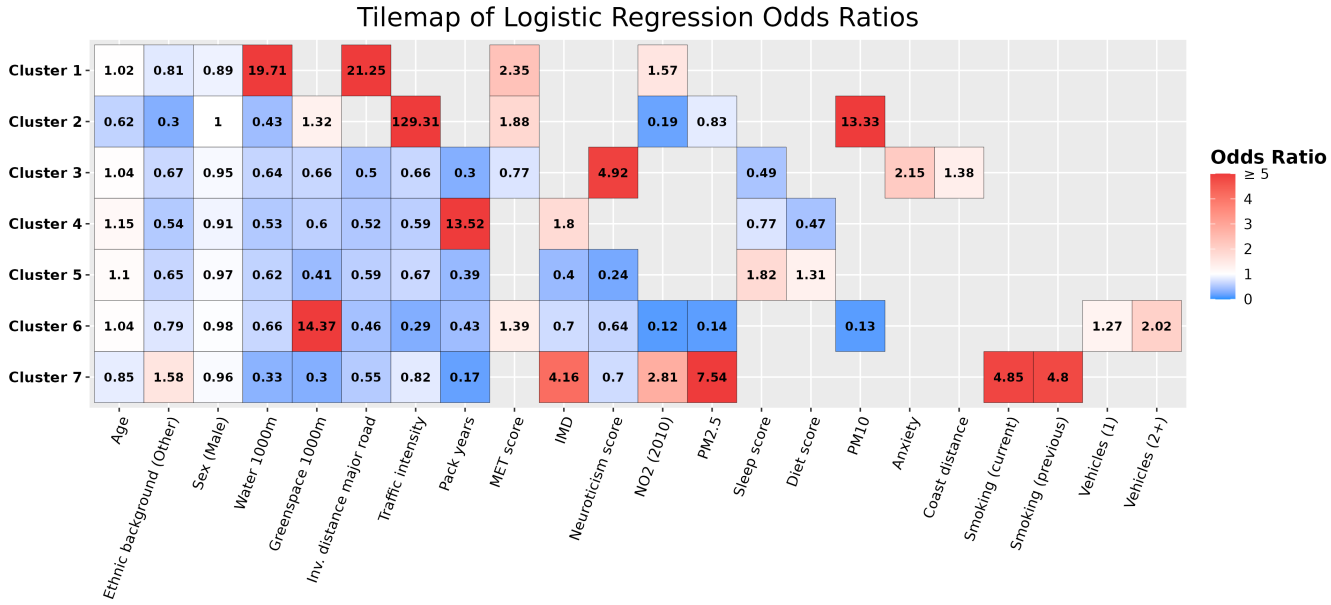

Figure 3: Tilemap displaying odds ratios of lifestyle exposures to each cluster allocation. (Reference categories: Smoking (Never), Vehicles owned (0))

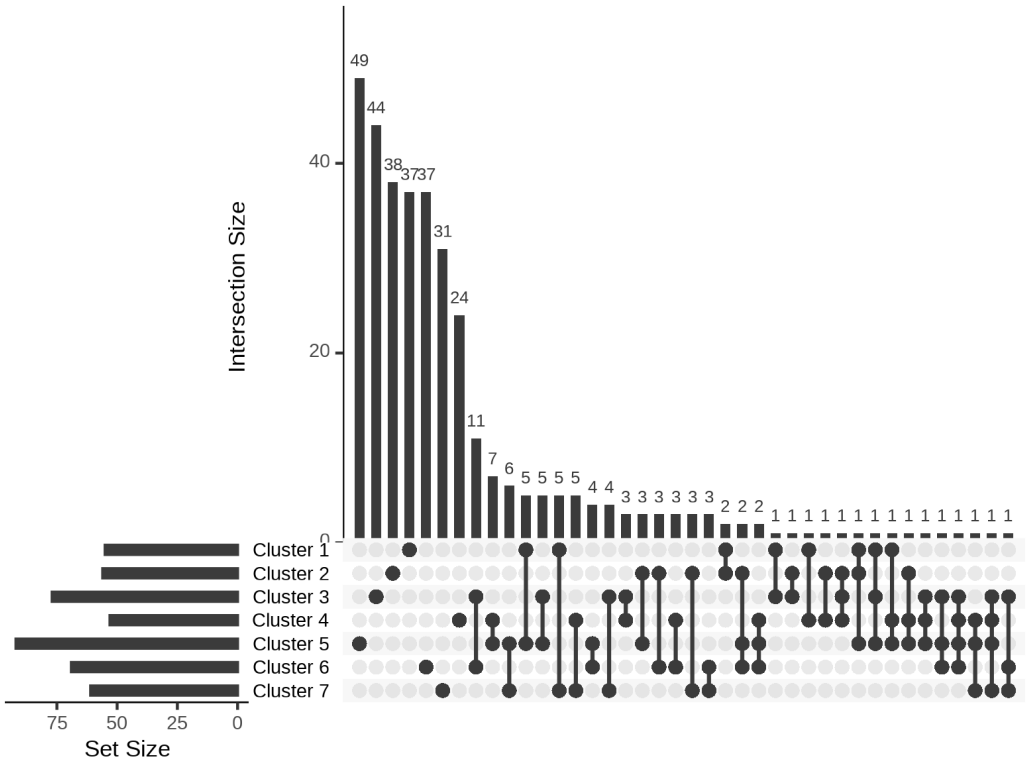

Figure 4: Upset plot displaying the stably selected proteins for each cluster.

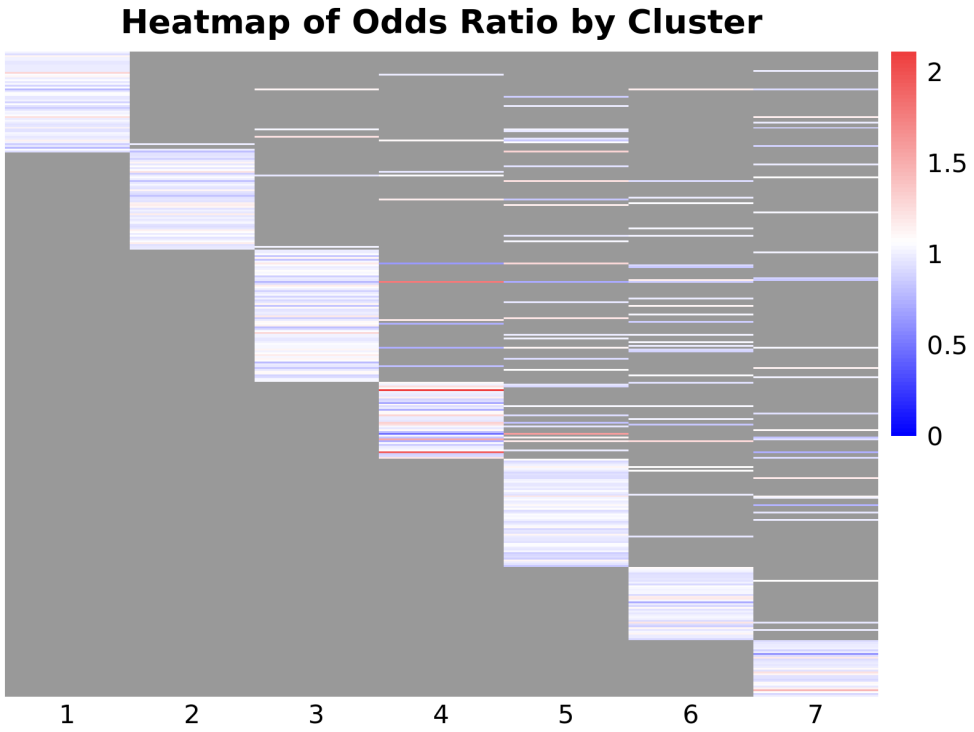

Figure 5: Heatmap displaying the odds ratios of proteins to each cluster allocation.

Molecular Profiling

The univariate analysis for proteins produced unstable results. The number of significant associations ranged from 0 proteins (cluster 1) to 838 proteins (cluster 3), using a Bonferroni-adjusted threshold. Figure A9 shows the overlapping and cluster-specific proteins among the significantly associated sets. The results reflected the exposure univariate analysis, as clusters 4-6 had the most significant associations for exposures and clusters. The Manhattan plots from this analysis are provided (Figures A25 to 31).

Using stability selection LASSO, between 50 to 100 proteins were selected for each cluster. Each cluster selected a relatively distinct set of proteins - 80% of stably selected proteins were unique to a single cluster. Figure A4 shows the overlapping and cluster-specific proteins among the stably selected sets. The selection proportion plots and calibration plots for each model are provided (Figures A32 to 38). Figure 5 shows the resulting odds ratios from our refit logistic regression models. Most of the odds ratios are around 1, with more downregulation than upregulation overall. Figure 5 also clearly shows the unique molecular signatures for each cluster. We then tested the logistic regression models, creating the ROC curves displayed in Figure A10. All of the AUC values were below 0.65, except cluster 4 had a value of 0.801.

Pathway Enrichment Analysis

The enriched pathways corresponded to seven functional clusters, revealing key biological mechanisms. In clusters 1 and 4, the pathways were primarily related to the nervous system. Clusters 2, 3, 5, and 6 were predominantly enriched for immune-related pathways. Additionally, cluster 5 showed enrichment in respiratory pathways, while clusters 6 and 7 also featured cancer-related pathways (Table 1).

Disease Incidence and Prediction

The standardised incidence rates showed that cluster 4 had the highest rates for AD, PD, CKD, and CAD, with moderately elevated rates of diabetes compared to other clusters. In contrast, clusters 1, 2, and 7 demonstrated lower overall incidence for all diseases (Figure 6).

Table 1: Pathway enrichment by cluster

Cluster	Pathway Name	# of Entities (found/total)	Entities FDR
1	Downregulation of ERBB4 signaling	3/11	0.0033137
	Interleukin-10 signaling	5/86	0.0033137
	Nuclear signaling by ERBB4	4/47	0.0033137
2	Interleukin-10 signaling	5/86	0.0033137
	Signaling by Interleukins	14/646	0.0050142
	Immune System	24/2661	0.0535697
3	Cytokine Signaling in Immune system	20/1095	0.0024067
	Signaling by Interleukins	14/646	0.0050142
	Downregulation of ERBB4 signaling	3/11	0.0033137
4	Synthesis, secretion, and deacylation of Ghrelin	4/26	0.0013384
	NTF3 activates NTRK3 signaling	2/4	0.0117385
	Post-translational modification: synthesis of GPI-anchored proteins	5/115	0.0117385
5	TFAP2 (AP-2) family regulates transcription of growth factors and their receptors	4/21	0.0060583
	Reversible hydration of carbon dioxide	3/17	0.0320181
	Interleukin-33 signaling	2/4	0.0320181
6	PI5P, PP2A, IER3 Regulate PI3K/AKT Signaling	7/137	0.0059400
	Constitutive Signaling by Aberrant PI3K in Cancer	7/104	0.0037700
	Negative regulation of the PI3K/AKT network	9/145	0.0337000
7	Constitutive Signaling by Aberrant PI3K in Cancer	7/104	0.0037700
	ERBB2 Activates PTK6 Signaling	4/18	0.0037700
	ERBB2 Regulates Cell Motility	4/19	0.0037700

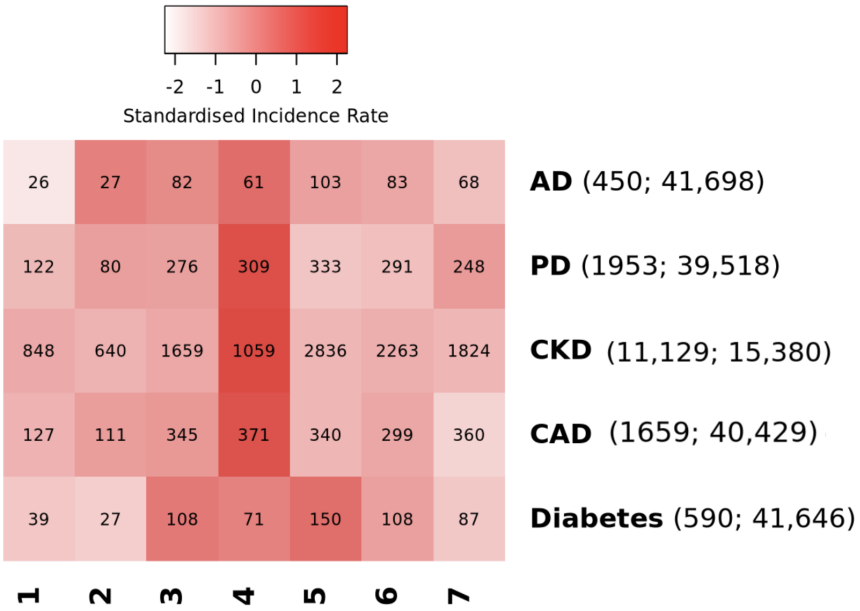

Figure 6: Standardised incidence rates across each cluster for each disease. Colour intensity indicates the standardised incidence rate, while numeric labels represent the raw case counts. Case counts and population sizes are shown in parentheses next to each disease.

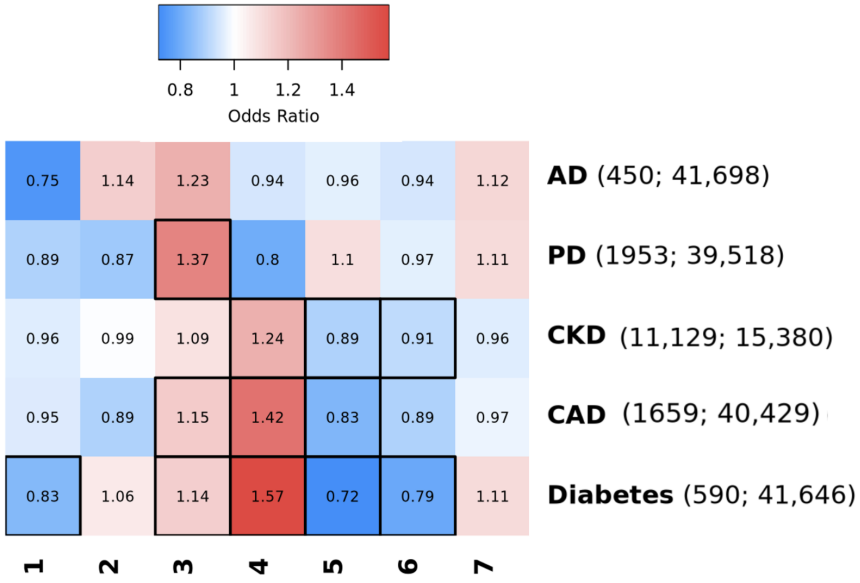

Figure 7: Odds ratios for each disease across clusters in the cluster-only model. Border outlines indicate statistically significant associations ($p < 0.05$).

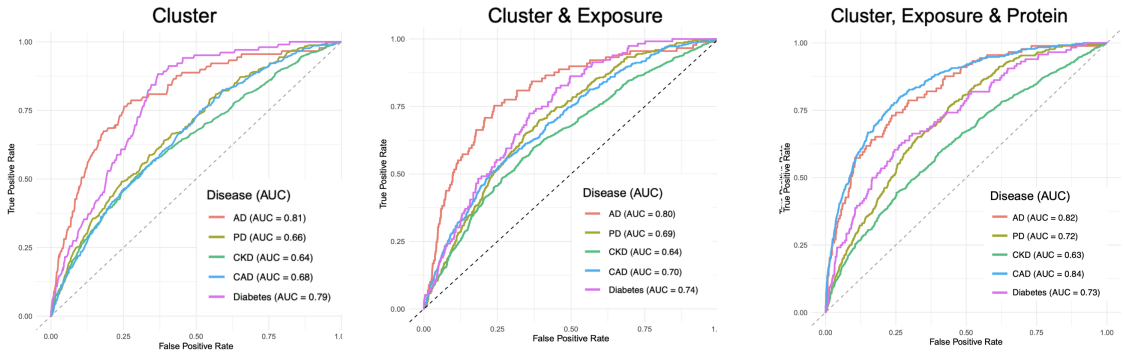

Figure 8: AUC for each disease across three predictive models: (1) cluster only, (2) cluster with exposure, and (3) cluster with exposure and proteins.

In the first logistic regression model, where cluster membership was used as the sole predictor, several clusters demonstrated either protective or risk associations with disease outcomes. Cluster 1 was associated with lower odds of diabetes (OR: 0.83), cluster 5 showed protective associations with CKD, CAD, and diabetes (OR: 0.89, 0.83, 0.72), and cluster 6 with CKD and diabetes (OR: 0.91, 0.79). On the other hand, cluster 3 was associated with higher odds of PD, CAD, and diabetes (OR: 1.37, 1.15, 1.14), while cluster 4 was associated with increased risk of CAD and diabetes (OR: 1.42, 1.57). Cluster 7 had no significant associations with any disease outcomes. (Figure 7).

Model performance, evaluated by the AUC, demonstrated that using clusters alone had good predictive performance for AD and diabetes (AUC: 0.81 and 0.79) but performed poorly for PD, CKD, and CAD (AUC: 0.66, 0.64, and 0.68). The second model, which incorporated clusters and exposures, showed similar results, with the highest performance for AD and diabetes (AUC: 0.80 and 0.74) and lowest for CKD (AUC: 0.64). In the third model, which included clusters, exposures, and proteins, performance further improved for CAD and AD (AUC: 0.84 and 0.83), though, as before, CKD remained the lowest performing (AUC: 0.63) (Figure 8).

Discussion

Protein analysis revealed distinct proteomic signatures for each cluster, highlighting the presence of cluster-specific molecular characteristics. However, our models predicting cluster membership from proteomic signatures performed poorly, indicating that proteomic data alone may not effectively distinguish between clusters. This shows the importance of looking at both exposures and proteins when characterising a cluster.

Our incidence rate analysis revealed that cluster 4, primarily composed of heavy smokers, had the highest incidence rates across four of the five diseases studied, capturing most of the cases of AD, PD, CKD, and CAD, while still having relatively high rates of diabetes. This highlights the well-established association between smoking and increased disease burden [23].

Analysis of the odds ratios from the cluster only disease prediction model revealed that both cluster 3 (high anxiety levels) and cluster 4 (heavy smokers) were significantly associated with increased odds of PD, CKD, CAD, and diabetes. These findings underscore the significant health impact of both psychological stress and smoking, reinforcing findings from other studies [23, 24]. Furthermore, cluster 3's heavy association with immune-related pathways, via cytokines and interleukins, may be linked to the presence of chronic inflammation, which has been shown to increase the risk of PD and diabetes [25, 26]. Cluster 4 was related to brain signaling pathways, mainly with NTF3-NTRK3, and Ghrelin regulation, both of which are important for brain health and metabolism. Disruption of these may increase the risk of AD, PD, and metabolic diseases like diabetes [27, 28].

On the other hand, cluster 1 which included individuals living near water and major roads with high physical activity levels, demonstrated lower odds of developing diabetes. This protective association may be directly attributed to the benefits of physical activity. Cluster 5 was associated with a healthier profile, including lower smoking, lower anxiety, and reduced socioeconomic deprivation. These characteristics contributed to the reduced odds of CKD,

CAD, and diabetes, reflecting the importance of both behavioural and social determinants in disease prevention. Furthermore, the association with IL-33, an immune signalling molecule, is linked to reduced cardio-renal and metabolic disease risk [29]. Similarly, cluster 6, defined by greater greenspace and lower exposure to traffic and air pollution, showed reduced odds of CKD and diabetes. These findings align with existing literature that suggest access to natural environments and lower air pollution can have protective effects on the kidney and cardiometabolic health [30, 31, 32]. The cluster's association with pathways related to immune and cancer signalling, such as the PI3K/AKT pathway, controls the level of inflammation and cell survival. Healthy regulation of these pathways may reduce the risk of CKD and diabetes [33].

Interestingly, clusters 2 and 7 were not significantly associated with any of the diseases analysed, despite being composed of individuals exposed to high levels of traffic and air pollution. This was unexpected, given the evidence linking such exposures to various chronic conditions [34]. These results may be explained by residual confounding, differences in exposure intensity, or other protective factors not captured in our clustering.

In the context of disease prediction using clusters or exposures as predictors, the similar performance observed between the second model (clusters and exposures) to the first (clusters) is expected, as cluster membership was determined based on exposures and therefore already encapsulates most, if not all the exposure-related information. Adding exposures as additional predictors does not provide new information to the model and likely introduces redundancy.

The improvement in predictive performance for CKD in the final model (clusters, exposures, and proteins) suggests that the protein data contained information not captured by clusters or exposures alone. This highlights the importance of incorporating protein data in disease prediction models and indicates that not all relevant information can be fully encoded through exposure variables.

However, our study is subject to several limitations. Firstly, the UK Biobank cohort is not fully representative of the UK population, as it predominantly consists of older, more affluent individuals of White ethnicity. This makes our findings less generalisable to the entire country's population. Secondly, for AD, PD, CAD, and diabetes, there was a substantial class imbalance, with relatively fewer incident cases compared to controls. This could have affected the predictive performance and stability of our models, potentially leading to biased estimates in identifying associations. Additionally, when analysing disease incidence and predictions, participants with additional comorbidities were not excluded due to concerns about reducing sample size. As a result, the presence of other diseases may have introduced confounding effects, potentially influencing the observed associations. Finally, the directionality of pathway regulation was not assessed via the Reactome platform, therefore the functional impact of pathways, whether protective or detrimental, to disease risks still remain unclear and warrants further investigation.

Future studies should explore the use of supervised clustering, which groups data based on feature similarity while incorporating the outcome during the clustering process, creating groups that are internally homogeneous but also strongly associated with the outcome. To validate and improve the generalisability of our findings to the broader UK population, future analyses should be conducted using datasets that more accurately reflects the country's demographics. Furthermore, we found that certain expotypes were associated with the body's cancer system, suggesting that future studies investigating cancer as an outcome could yield valuable insights.

Contributions

Hannah Cooper

Data Preprocessing, Clustering (GMM, SOM), Cluster Overview (Visualisation Plots), Table 1, Protein Analysis (Univariate, Stability Selection LASSO, Logistic Regression refit)

Tianshu Lu

Clustering (Fuzzy), Disease Extraction, Pathway Analysis

Imogen Onno

Data Preprocessing, Protein Extraction, Clustering (HDDC), Exposure Analysis (Univariate, Stability Selection LASSO, Logistic Regression refit)

Calix Tan

Data Imputation, Clustering (Biclustering), Cluster Scoring (PAC, BIC, Silhouette), Diseases (Incidence Rate, Prediction)

References

[1] Penny Webb, Chris Bain, and Andrew Page. *Essential epidemiology: an introduction for students and health professionals*. Cambridge University Press, 2017.

[2] Martine Vrijheid. "The exposome: a new paradigm to study the impact of environment on health". In: *Thorax* 69.9 (Sept. 2014), pp. 876–878. DOI: 10.1136/thoraxjnl-2013-204949.

[3] Christopher Paul Wild. "Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology". In: *Cancer Epidemiology Biomarkers & Prevention* 14.8 (2005), pp. 1847–1850.

[4] Baris Afsar et al. "An update on coronary artery disease and chronic kidney disease". In: *International Journal of Nephrology* 2014 (Mar. 2014), p. 767424. DOI: 10.1155/2014/767424.

[5] *UK Biobank - About our data*. https://www.ukbiobank.ac.uk/enable-your-research/about-our-data [Accessed: 21st April 2025]. 2024.

[6] IPAQ Research Committee et al. "Guidelines for data processing and analysis of the International Physical Activity Questionnaire (IPAQ)-short and long forms". In: *http://www. ipaq. ki. se/scoring. pdf* (2005).

[7] R. Wada. Exercise Score, Imperial College London.

[8] Caimei Yuan et al. "Associations of an overall healthy lifestyle with the risk of metabolic dysfunction-associated fatty liver disease". In: *BMC Public Health* 24.1 (2024), p. 3264.

[9] Raha West et al. "Sleep duration, chronotype, health and lifestyle factors affect cognition: a UK Biobank cross-sectional study". In: *BMJ Public Health* 2.1 (July 2024), e001000.

[10] Mengyu Fan et al. "Sleep patterns, genetic susceptibility, and incident cardiovascular disease: a prospective study of 385292 UK biobank participants". In: *European Heart Journal* 41.11 (Dec. 2019), pp. 1182–1189.

[11] Benjamin B Sun et al. "Plasma proteomic associations with genetics and health in the UK Biobank". In: *Nature* 622.7982 (2023), pp. 329–338.

[12] Marvin N Wright and Andreas Ziegler. "ranger: A fast implementation of random forests for high dimensional data in C++ and R". In: *Journal of statistical software* 77 (2017), pp. 1–17.

[13] Sebastian Kaiser. *Biclust: Bicluster algorithms*. 2023. URL: https://cran.r-project.org/web/packages/biclust/index.html.

[14] Martin Maechler et al. *cluster: Cluster Analysis Basics and Extensions*. R package version 2.1.8.1 — For new features, see the 'NEWS' and the 'Changelog' file in the package source). 2025. URL: https://CRAN.R-project.org/package=cluster.

[15] Chris Fraley et al. "Package 'mclust'". In: *Title Gaussian Mixture Modelling for Model Based Clustering, Classification, and Density Estimation. Available online* (2012).

[16] Laurent Bergé, Charles Bouveyron, and Stéphane Girard. "HDclassif: An R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data". In: *Journal of Statistical Software* 46.6 (2012), pp. 1–29. DOI: 10.18637/jss.v046.i06. URL: https://www.jstatsoft.org/article/view/v046i06.

[17] Ron Wehrens and Johannes Kruisselbrink. "Flexible Self-Organizing Maps in kohonen 3.0". In: *Journal of Statistical Software* 87.7 (2018), pp. 1–18. DOI: 10.18637/jss.v087.i07.

[18] Alboukadel Kassambara. *Factoextra: Extract and visualize the results of multivariate data analyses*. Apr. 2020. URL: https://cran.r-project.org/web/packages/factoextra/index.html.

[19] Barbara Bodinier et al. "Automated calibration for stability selection in penalised regression and graphical models." In: *J R Stat Soc Ser C Appl Stat* 72.5 (Nov. 2023), pp. 1375–1393.

[20] *Reactome Pathway Browser*. https://reactome.org/PathwayBrowser/. [Accessed: 1st May 2025]. 2025.

[21] Tomasz Konopka and Maintainer Tomasz Konopka. "R-package: umap". In: *Uniform Manifold Approximation and Projection* 836 (2018), p. 837.

[22] Jesse Krijthe, Laurens van der Maaten, and Maintainer Jesse Krijthe. "Package 'Rtsne'". In: *R package version* 13 (2018), p. 577.

[23] Leslie Y Kwan, Kathleen Stratton, and Richard J Bonnie. "Public health implications of raising the minimum age of legal access to tobacco products". In: (2015).

[24] Habib Yaribeygi et al. "The impact of stress on body function: A review". In: *EXCLI journal* 16 (2017), p. 1057.

[25] Iva Stojkovska, Brandon M Wagner, and Brad E Morrison. "Parkinson's disease and enhanced inflammatory response". In: *Experimental biology and medicine* 240.11 (2015), pp. 1387–1395.

[26] Sotirios Tsalamandris et al. "The role of inflammation in diabetes: current concepts and future perspectives". In: *European cardiology review* 14.1 (2019), p. 50.

[27] H Randall Griffith et al. "Brain metabolism differs in Alzheimer's disease and Parkinson's disease dementia". In: *Alzheimer's & Dementia* 4.6 (2008), pp. 421–427. DOI: 10.1016/j.jalz.2008.04.008.

[28] Débora Serrenho, Sandra D Santos, and Ana Luísa Carvalho. "The Role of Ghrelin in Regulating Synaptic Function and Plasticity of Feeding-Associated Circuits". In: *Frontiers in Cellular Neuroscience* 13 (2019), p. 205. DOI: 10.3389/fncel.2019.00205.

[29] Marella Marassi and Gian Paolo Fadini. "The cardio–renal–metabolic connection: a review of the evidence". In: *Cardiovascular Diabetology* 22.1 (2023), p. 195. DOI: 10.1186/s12933-023-01937-x.

[30] Xiang Qian Lao et al. "Environmental pollution to kidney disease: an updated review of current knowledge and future directions". In: *Kidney international* (2024).

[31] Yongze Li et al. "Association between air pollution and type 2 diabetes: an updated review of the literature". In: *Therapeutic advances in endocrinology and metabolism* 10 (2019), p. 2042018819897046.

[32] Howard Frumkin, Richard J Jackson, and Christine M Coussens. "Health and the environment in the southeastern united states". In: (2002).

[33] Xuegang He et al. "The PI3K/AKT signalling pathway in inflammation, cell death and glial scar formation after traumatic spinal cord injury: Mechanisms and therapeutic opportunities". In: *Cell Proliferation* 55.9 (2022), e13275. DOI: 10.1111/cpr.13275.

[34] Ioannis Manisalidis et al. "Environmental and health impacts of air pollution: a review". In: *Frontiers in public health* 8 (2020), p. 14.

Appendix

	1 (N=3193)	2 (N=2248)	3 (N=7062)	4 (N=4767)	5 (N=9538)	6 (N=7870)	7 (N=7126)	Overall (N=41804)	P-value
Sex									
Female	1752 (54.9%)	1221 (54.3%)	4652 (65.9%)	1998 (41.9%)	5080 (53.3%)	4318 (54.9%)	4044 (56.8%)	23065 (55.2%)	<0.001
Male	1441 (45.1%)	1027 (45.7%)	2410 (34.1%)	2769 (58.1%)	4458 (46.7%)	3552 (45.1%)	3082 (43.3%)	18739 (44.8%)	
Age									
Mean (SD)	56.6 (8.37)	56.9 (8.01)	55.9 (8.14)	59.1 (7.39)	57.1 (8.16)	57.3 (7.97)	54.8 (8.46)	56.7 (8.19)	<0.001
Median [Min, Max]	58.0 [40.0, 70.0]	58.0 [40.0, 70.0]	57.0 [40.0, 70.0]	61.0 [40.0, 70.0]	59.0 [40.0, 70.0]	59.0 [40.0, 70.0]	55.0 [39.0, 70.0]	58.0 [39.0, 70.0]	
Ethnicity									
British	2791 (87.4%)	2047 (91.1%)	6313 (89.4%)	4346 (91.2%)	8657 (90.8%)	7493 (95.2%)	5305 (74.4%)	36952 (88.4%)	<0.001
Other_ethnicity	402 (12.6%)	201 (8.9%)	749 (10.6%)	421 (8.8%)	881 (9.2%)	377 (4.8%)	1821 (25.6%)	4852 (11.6%)	
Multiple Deprivation Index									
Mean (SD)	5.71 (2.88)	5.18 (2.77)	5.48 (2.76)	6.82 (2.63)	4.11 (2.43)	4.22 (2.35)	7.93 (1.96)	5.50 (2.85)	<0.001
Median [Min, Max]	6.00 [1.00, 10.0]	5.00 [1.00, 10.0]	6.00 [1.00, 10.0]	7.00 [1.00, 10.0]	4.00 [1.00, 10.0]	4.00 [1.00, 10.0]	8.00 [1.00, 10.0]	5.00 [1.00, 10.0]	
Household Income									
Less than 18,000	707 (22.1%)	502 (22.3%)	1739 (24.6%)	1759 (36.9%)	1358 (14.2%)	1138 (14.5%)	1948 (27.3%)	9151 (21.9%)	<0.001
30,999 to 51,999	1667 (52.2%)	1183 (52.6%)	3970 (56.2%)	2489 (52.2%)	5252 (55.1%)	4271 (54.3%)	3705 (52.0%)	22537 (53.9%)	
Greater than 52,000	819 (25.6%)	563 (25.0%)	1353 (19.2%)	519 (10.9%)	2928 (30.7%)	2461 (31.3%)	1473 (20.7%)	10116 (24.2%)	
Qualifications									
Other	870 (27.2%)	687 (30.6%)	2337 (33.1%)	2245 (47.1%)	2065 (21.7%)	1807 (23.0%)	1976 (27.7%)	11987 (28.7%)	<0.001
Degree	1124 (35.2%)	670 (29.8%)	1908 (27.0%)	752 (15.8%)	3754 (39.4%)	2799 (35.6%)	2842 (39.9%)	13849 (33.1%)	
A/AS/O/GCSE level	1041 (32.6%)	769 (34.2%)	2478 (35.1%)	1495 (31.4%)	3154 (33.1%)	2814 (35.8%)	1991 (27.9%)	13742 (32.9%)	
Professional	158 (4.9%)	122 (5.4%)	339 (4.8%)	275 (5.8%)	565 (5.9%)	450 (5.7%)	317 (4.4%)	2226 (5.3%)	
Employment Status									
Other	311 (9.7%)	175 (7.8%)	794 (11.2%)	644 (13.5%)	521 (5.5%)	529 (6.7%)	866 (12.2%)	3840 (9.2%)	<0.001
Employed	1780 (55.7%)	1244 (55.3%)	4011 (56.8%)	2032 (42.6%)	5429 (56.9%)	4262 (54.2%)	4554 (63.9%)	23312 (55.8%)	
Retired	1102 (34.5%)	829 (36.9%)	2257 (32.0%)	2091 (43.9%)	3588 (37.6%)	3079 (39.1%)	1706 (23.9%)	14652 (35.0%)	
Own Rent									
Other	349 (10.9%)	165 (7.3%)	650 (9.2%)	863 (18.1%)	354 (3.7%)	310 (3.9%)	1476 (20.7%)	4167 (10.0%)	<0.001
Own outright	1720 (53.9%)	1263 (56.2%)	3592 (50.9%)	2497 (52.4%)	5766 (60.5%)	4711 (59.9%)	2907 (40.8%)	22456 (53.7%)	
Own with a mortgage	1124 (35.2%)	820 (36.5%)	2820 (39.9%)	1407 (29.5%)	3418 (35.8%)	2849 (36.2%)	2743 (38.5%)	15181 (36.3%)	
Num Household									
Alone	665 (20.8%)	374 (16.6%)	1287 (18.2%)	1162 (24.4%)	1194 (12.5%)	952 (12.1%)	1865 (26.2%)	7499 (17.9%)	<0.001
Small	1915 (60.0%)	1411 (62.8%)	4397 (62.3%)	3123 (65.5%)	6431 (67.4%)	5406 (68.7%)	3669 (51.5%)	26352 (63.0%)	
Large	594 (18.6%)	456 (20.3%)	1350 (19.1%)	472 (9.9%)	1895 (19.9%)	1496 (19.0%)	1535 (21.5%)	7798 (18.7%)	
Community home	19 (0.6%)	7 (0.3%)	28 (0.4%)	10 (0.2%)	18 (0.2%)	16 (0.2%)	57 (0.8%)	155 (0.4%)	
Vehicles Household									
None	354 (11.1%)	134 (6.0%)	567 (8.0%)	659 (13.8%)	275 (2.9%)	127 (1.6%)	1383 (19.4%)	3499 (8.4%)	<0.001
One	1382 (43.3%)	891 (39.6%)	3206 (45.4%)	2438 (51.1%)	3758 (39.4%)	2320 (29.5%)	3815 (53.5%)	17810 (42.6%)	
Two or more	1457 (45.6%)	1223 (54.4%)	3289 (46.6%)	1670 (35.0%)	5505 (57.7%)	5423 (68.9%)	1928 (27.1%)	20495 (49.0%)	

Figure 1: Table 1 (Demographics and Socio-Economic)

	1 (N=3193)	2 (N=2248)	3 (N=7062)	4 (N=4767)	5 (N=9538)	6 (N=7870)	7 (N=7126)	Overall (N=41804)	P-value
No2 2010									
Mean (SD)	30.37 (7.72)	25.15 (5.73)	25.92 (3.83)	27.23 (4.8)	26.06 (3.72)	17.44 (2.9)	34.81 (4.09)	26.32 (6.9)	<0.001
Median (Min, Max)	29.8 (12.9, 49.8)	25.7 (12.9, 43.5)	25.8 (12.9, 39)	27.2 (12.9, 42.9)	25.9 (13.1, 38.4)	17.3 (12.9, 30)	34.5 (22.6, 49.6)	26.1 (12.9, 49.8)	
Pm10									
Mean (SD)	16.75 (1.51)	19.16 (1.51)	15.87 (0.86)	15.98 (1.01)	15.94 (0.8)	14.1 (1.28)	16.89 (1.04)	15.98 (1.61)	<0.001
Median (Min, Max)	16.7 (11.9, 21.7)	19.3 (14.2, 21.7)	15.9 (12.1, 19.8)	16 (11.9, 20.2)	15.9 (12.1, 20.5)	14.2 (11.8, 20.2)	16.7 (12.3, 21.6)	15.9 (11.8, 21.7)	
Pm2.5									
Mean (SD)	10.36 (1)	9.92 (0.91)	9.88 (0.52)	10.08 (0.7)	9.9 (0.5)	8.74 (0.45)	11.1 (0.71)	9.94 (0.96)	<0.001
Median (Min, Max)	10.3 (8.2, 13)	9.9 (8.2, 13)	9.9 (8.2, 11.8)	10 (8.2, 13)	9.9 (8.2, 11.9)	8.7 (8.2, 10.5)	11 (9.4, 13)	9.9 (8.2, 13)	
Traffic Intensity									
Mean (SD)	18065.31 (9247.6)	60436.84 (13015.44)	18632.9 (7838.88)	19343.72 (8755.14)	18516.61 (7590.72)	13910.35 (6672.44)	21053.23 (8968.35)	20415.58 (12886.86)	<0.001
Median (Min, Max)	16217 (5050, 84554)	59142.5 (15689, 84554)	16989 (5038, 56220)	17201 (5038, 56220)	16989 (5050, 50291)	12326 (5018, 56047)	18820 (5050, 59064)	16956 (5018, 84554)	
Inv Dis Maj Road									
Mean (SD)	0.01 (0.01)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.01 (0)	0 (0)	<0.001
Median (Min, Max)	0 (0, 0)	0 (0, 0)	0 (0, 0)	0 (0, 0)	0 (0, 0)	0 (0, 0)	0 (0, 0)	0 (0, 0)	
Greenspace 1000m									
Mean (SD)	41.38 (18.75)	51.63 (19.71)	42.22 (14.72)	40.92 (16.02)	40.32 (14.29)	74.85 (14)	24.92 (9.64)	45.27 (21.47)	<0.001
Median (Min, Max)	38.5 (6.4, 97.3)	51.4 (8.9, 97.3)	41.6 (8.3, 96.9)	39.5 (7.9, 95.6)	40 (7.4, 97.6)	75.3 (11.4, 99.2)	23.6 (6.7, 75.5)	42.1 (6.4, 99.2)	
Water 1000m									
Mean (SD)	3.66 (2.77)	0.99 (1.19)	0.76 (0.86)	0.78 (0.92)	0.71 (0.83)	1.02 (1.11)	0.64 (0.78)	1.01 (1.4)	<0.001
Median (Min, Max)	3.9 (0, 8.9)	0.6 (0, 8.7)	0.5 (0, 5.2)	0.4 (0, 5.8)	0.4 (0, 5.4)	0.6 (0, 7.4)	0.3 (0, 5.2)	0.5 (0, 8.9)	
Coast Distance									
Mean (SD)	39.73 (28.64)	38.45 (29.62)	48.72 (27.51)	44.53 (27.89)	45.06 (27.24)	47.22 (26.77)	39.04 (24.25)	44.24 (27.28)	<0.001
Median (Min, Max)	42.9 (0, 101.6)	40.8 (0.2, 106.1)	53.7 (0.2, 102.8)	49.6 (0.2, 102.3)	49.2 (0.2, 102.5)	53.9 (0.1, 101.7)	37.7 (0.1, 102.7)	47.8 (0, 106.1)	
Gas Cooker									
0	932 (29.2%)	662 (29.4%)	1899 (26.9%)	1396 (29.3%)	2592 (27.2%)	2955 (37.5%)	1667 (23.4%)	12103 (29.0%)	<0.001
1	2261 (70.8%)	1586 (70.6%)	5163 (73.1%)	3371 (70.7%)	6946 (72.8%)	4915 (62.5%)	5459 (76.6%)	29701 (71.0%)	
Gas Fire									
0	1953 (61.2%)	1378 (61.3%)	3885 (55.0%)	2644 (55.5%)	5457 (57.2%)	5002 (63.6%)	4524 (63.5%)	24843 (59.4%)	<0.001
1	1240 (38.8%)	870 (38.7%)	3177 (45.0%)	2123 (44.5%)	4081 (42.8%)	2868 (36.4%)	2602 (36.5%)	16961 (40.6%)	
Solid Fire									
0	2904 (90.9%)	2060 (91.6%)	6732 (95.3%)	4566 (95.8%)	8894 (93.2%)	6511 (82.7%)	6655 (93.4%)	38322 (91.7%)	<0.001
1	289 (9.1%)	188 (8.4%)	330 (4.7%)	201 (4.2%)	644 (6.8%)	1359 (17.3%)	471 (6.6%)	3482 (8.3%)	

Figure 2: Table 1 (Environmental)

	1 (N=3193)	2 (N=2248)	3 (N=7062)	4 (N=4767)	5 (N=9538)	6 (N=7870)	7 (N=7126)	Overall (N=41804)	P-value
Alcohol Intake									
High	1438 (45.0%)	972 (43.2%)	2582 (36.6%)	2159 (45.3%)	4284 (44.9%)	3901 (49.6%)	2749 (38.6%)	18085 (43.3%)	<0.001
Medium	1146 (35.9%)	849 (37.8%)	2816 (39.9%)	1607 (33.7%)	3758 (39.4%)	2859 (36.3%)	2487 (34.9%)	15522 (37.1%)	
Never/Rarely	609 (19.1%)	427 (19.0%)	1664 (23.6%)	1001 (21.0%)	1496 (15.7%)	1110 (14.1%)	1890 (26.5%)	8197 (19.6%)	
Alcohol 10 Yrs									
0	266 (8.3%)	192 (8.5%)	651 (9.2%)	381 (8.0%)	602 (6.3%)	415 (5.3%)	906 (12.7%)	3413 (8.2%)	<0.001
About the same	1089 (34.1%)	772 (34.3%)	2160 (30.6%)	1291 (27.1%)	4008 (42.0%)	3094 (39.3%)	2207 (31.0%)	14621 (35.0%)	
Less nowadays	1343 (42.1%)	935 (41.6%)	3059 (43.3%)	2449 (51.4%)	3603 (37.8%)	3016 (38.3%)	3105 (43.6%)	17510 (41.9%)	
More nowadays	495 (15.5%)	349 (15.5%)	1192 (16.9%)	646 (13.6%)	1325 (13.9%)	1345 (17.1%)	908 (12.7%)	6260 (15.0%)	
Smoking Status									
Never	1759 (55.1%)	1290 (57.4%)	4781 (67.7%)	2 (0.0%)	6674 (70.0%)	4791 (60.9%)	4232 (59.4%)	23529 (56.3%)	<0.001
Previous	1108 (34.7%)	771 (34.3%)	1855 (26.3%)	3293 (69.1%)	2475 (25.9%)	2637 (33.5%)	2135 (30.0%)	14274 (34.1%)	
Current	326 (10.2%)	187 (8.3%)	426 (6.0%)	1472 (30.9%)	389 (4.1%)	442 (5.6%)	759 (10.7%)	4001 (9.6%)	
Pack Years									
Mean (SD)	9.04 (13.3)	8.35 (12.8)	3.74 (6.66)	34.7 (10.8)	3.50 (6.62)	6.39 (10.6)	5.79 (8.99)	8.72 (13.3)	<0.001
Median [Min, Max]	0 [0, 59.0]	0 [0, 58.5]	0 [0, 33.6]	34.0 [0, 59.6]	0 [0, 39.4]	0 [0, 59.5]	0 [0, 47.0]	0 [0, 59.6]	
Breastfed									
No	849 (26.6%)	611 (27.2%)	2103 (29.8%)	1231 (25.8%)	2311 (24.2%)	1901 (24.2%)	1742 (24.4%)	10748 (25.7%)	<0.001
Yes	2344 (73.4%)	1637 (72.8%)	4959 (70.2%)	3536 (74.2%)	7227 (75.8%)	5969 (75.8%)	5384 (75.6%)	31056 (74.3%)	
Maternal Smoking									
No	2214 (69.3%)	1521 (67.7%)	4816 (68.2%)	2949 (61.9%)	7024 (73.6%)	5573 (70.8%)	5124 (71.9%)	29221 (69.9%)	<0.001
Yes	979 (30.7%)	727 (32.3%)	2246 (31.8%)	1818 (38.1%)	2514 (26.4%)	2297 (29.2%)	2002 (28.1%)	12583 (30.1%)	
Anxiety									
No	1437 (45.0%)	950 (42.3%)	610 (8.6%)	2091 (43.9%)	6301 (66.1%)	3588 (45.6%)	3315 (46.5%)	18292 (43.8%)	<0.001
Yes	1756 (55.0%)	1298 (57.7%)	6452 (91.4%)	2676 (56.1%)	3237 (33.9%)	4282 (54.4%)	3811 (53.5%)	23512 (56.2%)	
Neuro Score									
Mean (SD)	4.19 (3.32)	4.19 (3.20)	7.61 (2.46)	4.52 (3.29)	2.06 (1.85)	3.82 (2.95)	4.02 (3.03)	4.22 (3.27)	<0.001
Median [Min, Max]	4.00 [0, 12.0]	4.00 [0, 12.0]	8.00 [0, 12.0]	4.00 [0, 12.0]	2.00 [0, 10.0]	3.00 [0, 12.0]	4.00 [0, 12.0]	4.00 [0, 12.0]	
Distress Score									
0	2761 (86.5%)	2004 (89.1%)	5971 (84.6%)	4085 (85.7%)	8766 (91.9%)	7106 (90.3%)	6071 (85.2%)	36764 (87.9%)	<0.001
1+	432 (13.5%)	244 (10.9%)	1091 (15.4%)	682 (14.3%)	772 (8.1%)	764 (9.7%)	1055 (14.8%)	5040 (12.1%)	
Sleep Data									
Mean (SD)	1.91 (0.722)	1.94 (0.707)	1.58 (0.627)	1.74 (0.711)	2.22 (0.604)	1.98 (0.658)	1.91 (0.687)	1.92 (0.694)	<0.001
Median [Min, Max]	2.00 [0, 3.00]	2.00 [0, 3.00]	2.00 [0, 3.00]	2.00 [0, 3.00]	2.00 [0, 3.00]	2.00 [0, 3.00]	2.00 [0, 3.00]	2.00 [0, 3.00]	
Diet Score									
Mean (SD)	5.85 (1.71)	5.89 (1.65)	5.68 (1.62)	5.02 (1.72)	6.28 (1.49)	6.04 (1.60)	5.88 (1.62)	5.87 (1.65)	<0.001
Median [Min, Max]	6.00 [1.00, 10.0]	6.00 [1.00, 10.0]	6.00 [1.00, 10.0]	5.00 [1.00, 10.0]	6.50 [1.00, 10.0]	6.00 [1.00, 10.0]	6.00 [1.00, 10.0]	6.00 [1.00, 10.0]	
Met Score									
Mean (SD)	2570 (2510)	2310 (2290)	1850 (1910)	2200 (2300)	2140 (1920)	2390 (2230)	2120 (2010)	2180 (2120)	<0.001
Median [Min, Max]	1690 [0, 10100]	1510 [0, 10000]	1210 [0, 10100]	1390 [0, 10100]	1570 [0, 10100]	1710 [0, 10100]	1490 [0, 10100]	1510 [0, 10100]	

Figure 3: Table 1 (Behavioural, Childhood, and Psycho-Social)

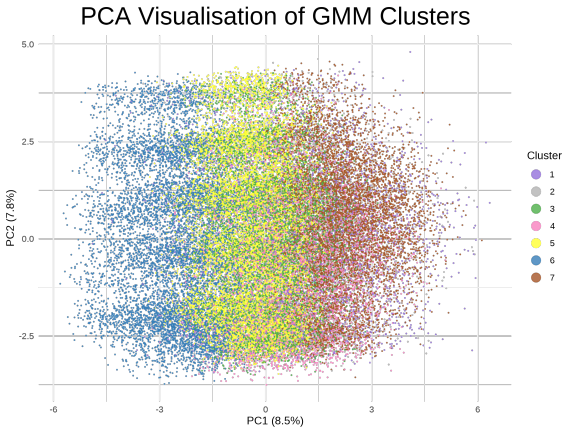

Figure 4: PCA visualisation of clusters

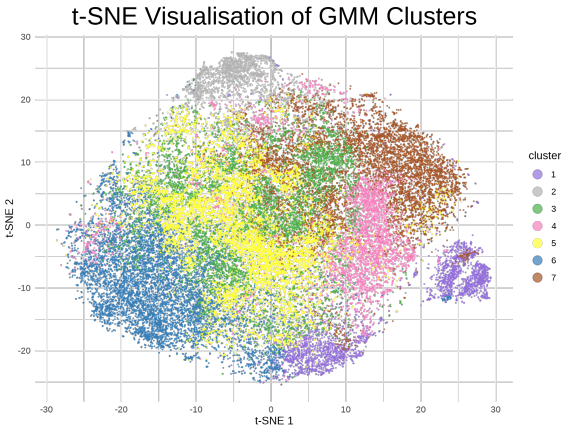

Figure 5: t-SNE visualisation of clusters

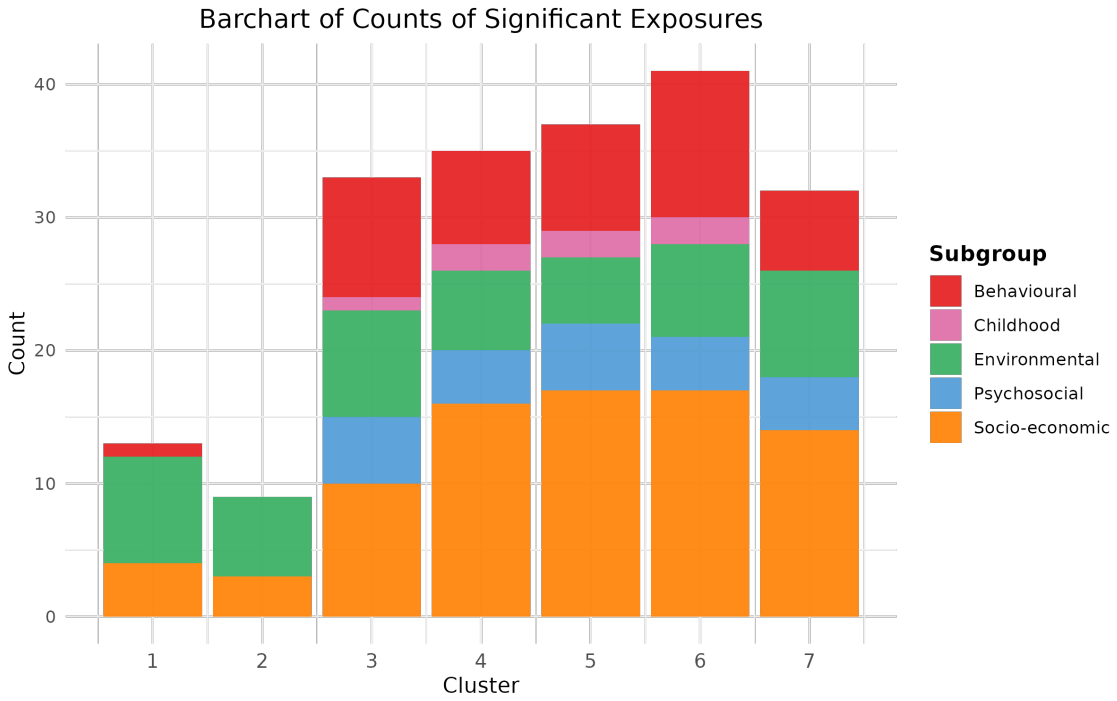

Figure 6: Stacked bar chart displaying total number of significant associations of exposures to each cluster allocation. Each bar is coloured by domain.

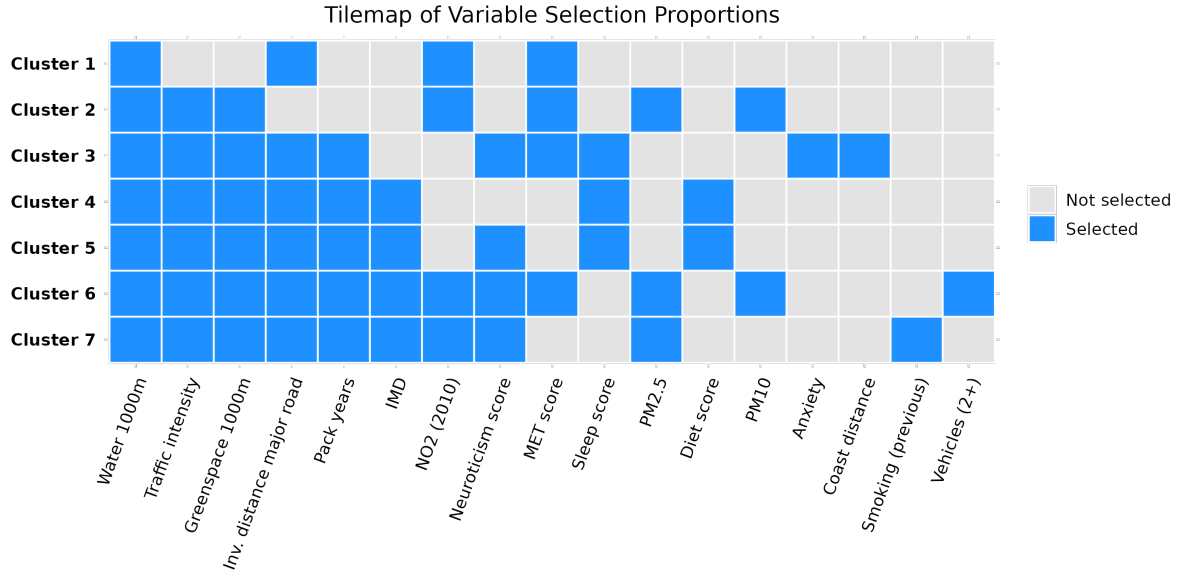

Figure 7: Tile map of stably selected exposures for each cluster allocation from stability selection LASSO

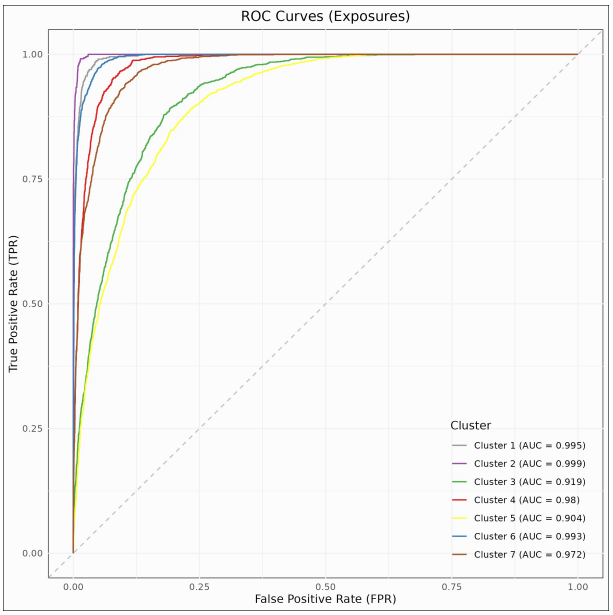

Figure 8: ROC curve for each refitted logistic regression model with stably selected exposures.

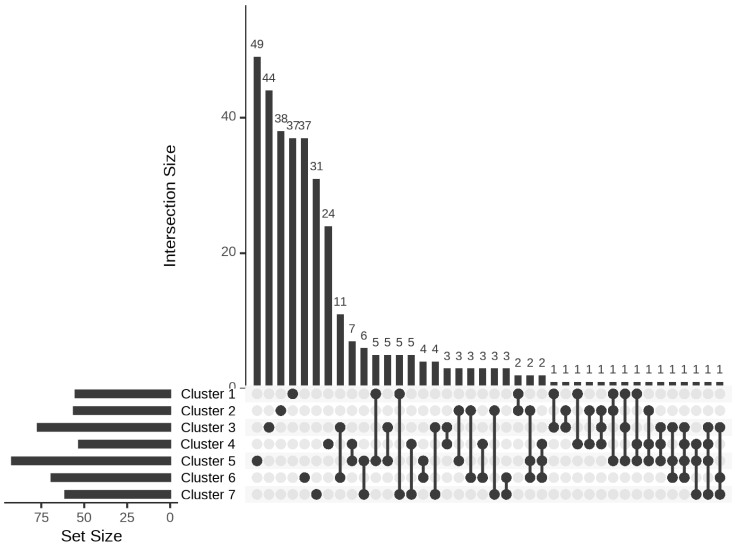

Figure 9: Upset plot displaying significant proteins for each cluster from univariate analysis.

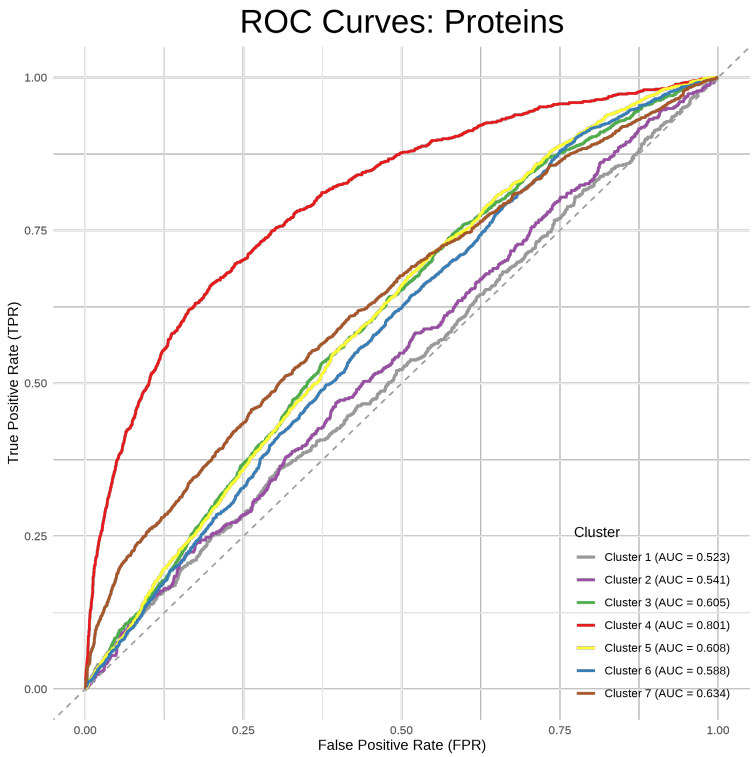

Figure 10: ROC curves for cluster prediction from proteins.

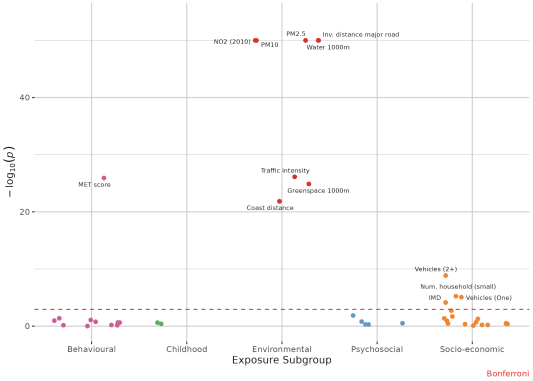

Figure 11: Manhattan plot exposures (cluster 1)

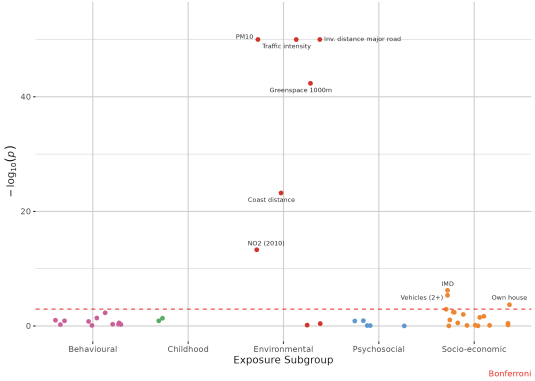

Figure 12: Manhattan plot exposures (cluster 2)

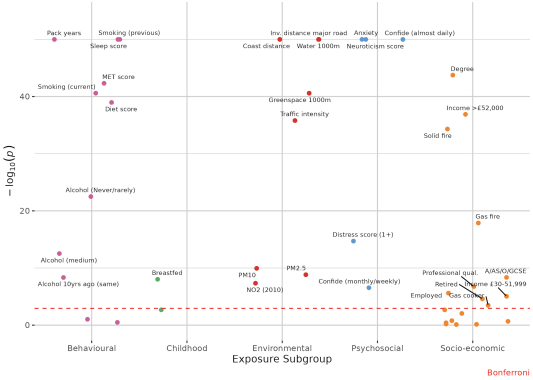

Figure 13: Manhattan plot exposures (cluster 3)

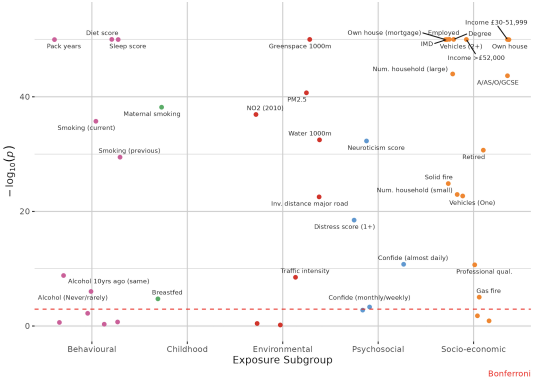

Figure 14: Manhattan plot exposures (cluster 4)

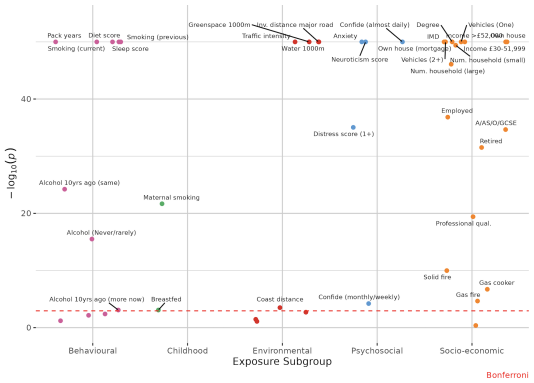

Figure 15: Manhattan plot exposures (cluster 5)

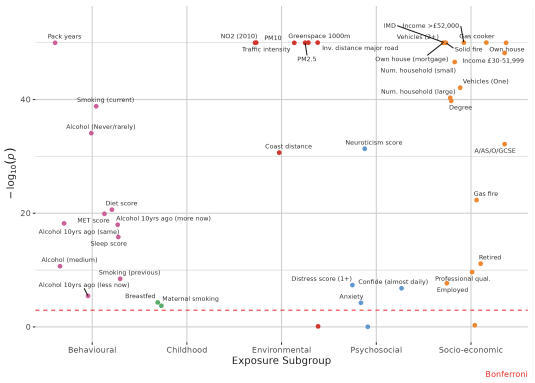

Figure 16: Manhattan plot exposures (cluster 6)

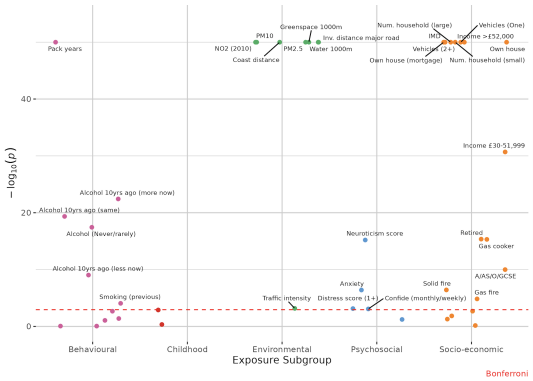

Figure 17: Manhattan plot exposures (cluster 7)

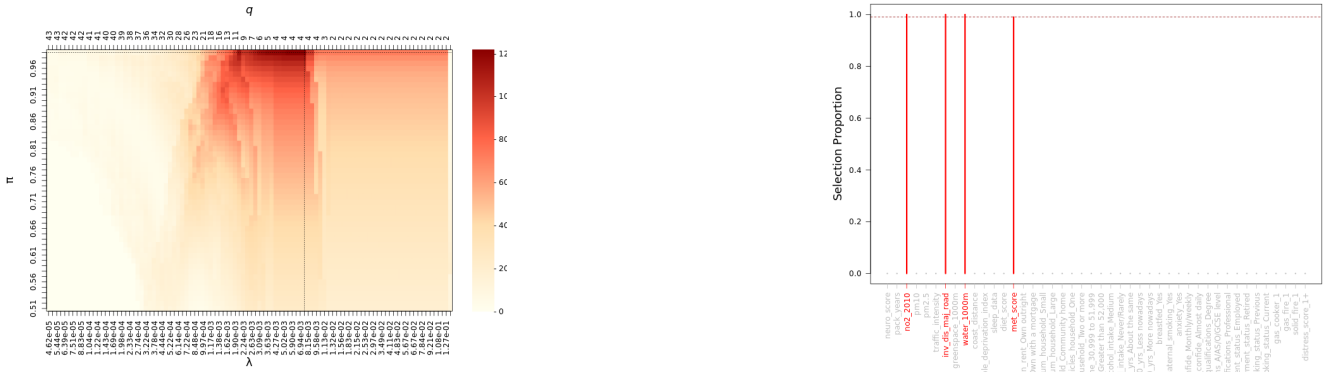

Figure 18: Calibration and selection plots exposures (cluster 1)

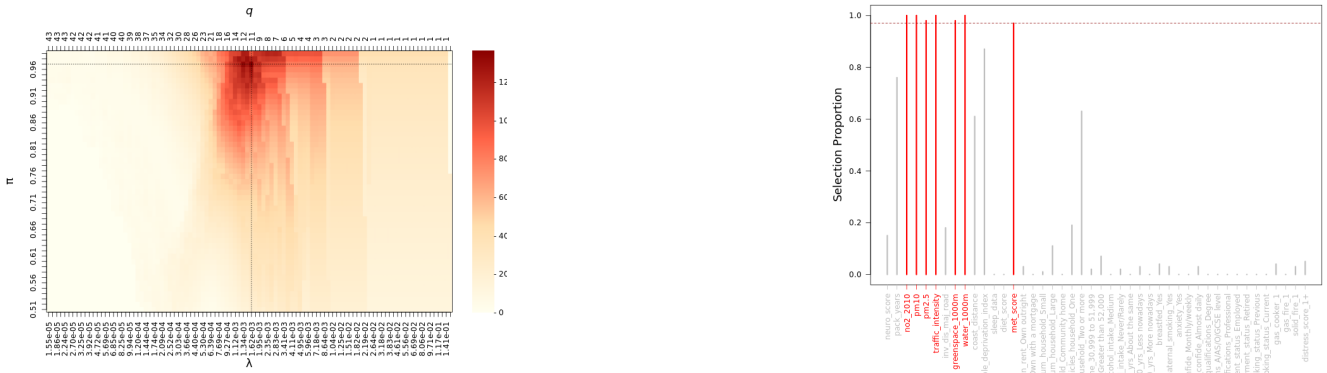

Figure 19: Calibration and selection plots exposures (cluster 2)

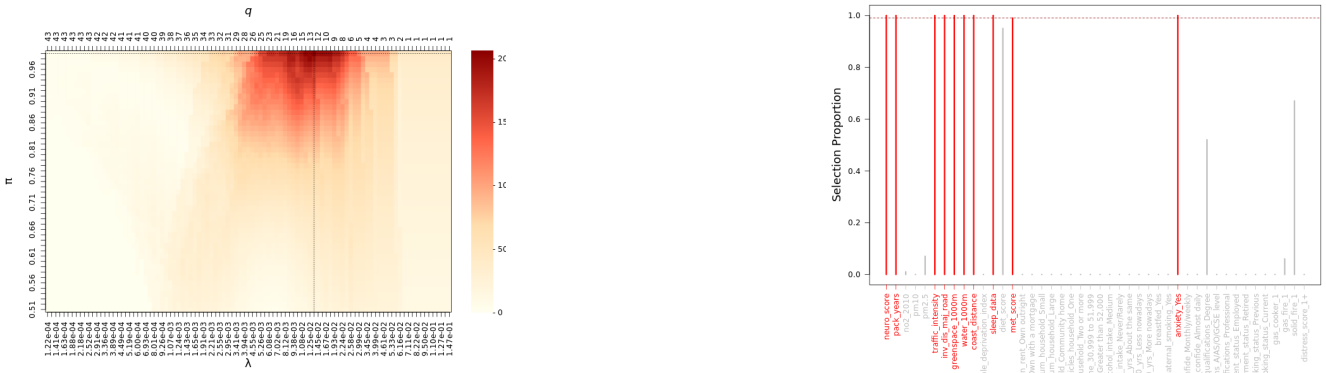

Figure 20: Calibration and selection plots exposures (cluster 3)

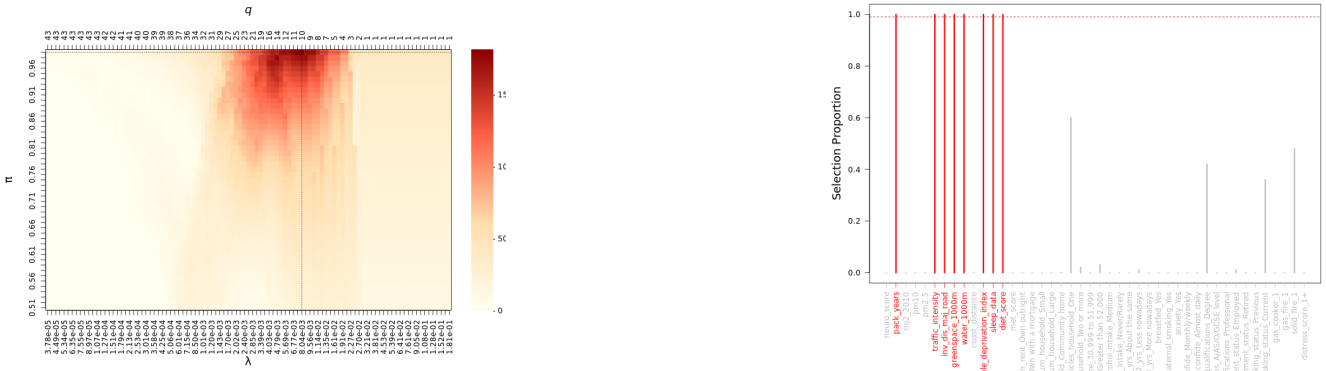

Figure 21: Calibration and selection plots exposures (cluster 4)

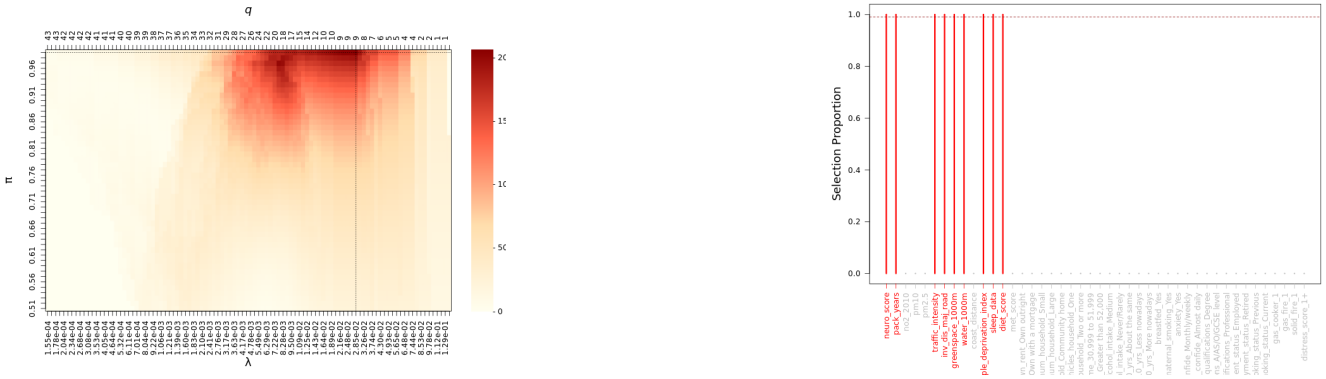

Figure 22: Calibration and selection plots exposures (cluster 5)

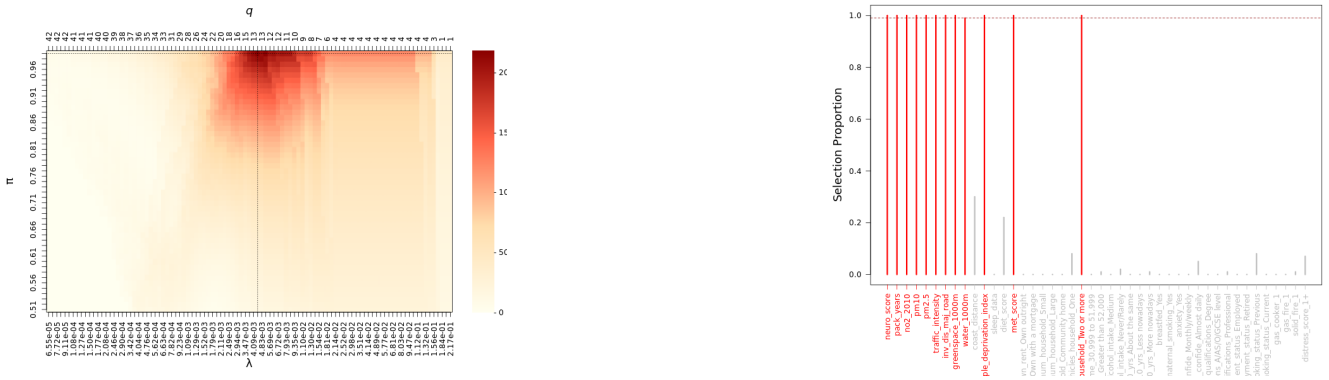

Figure 23: Calibration and selection plots exposures (cluster 6)

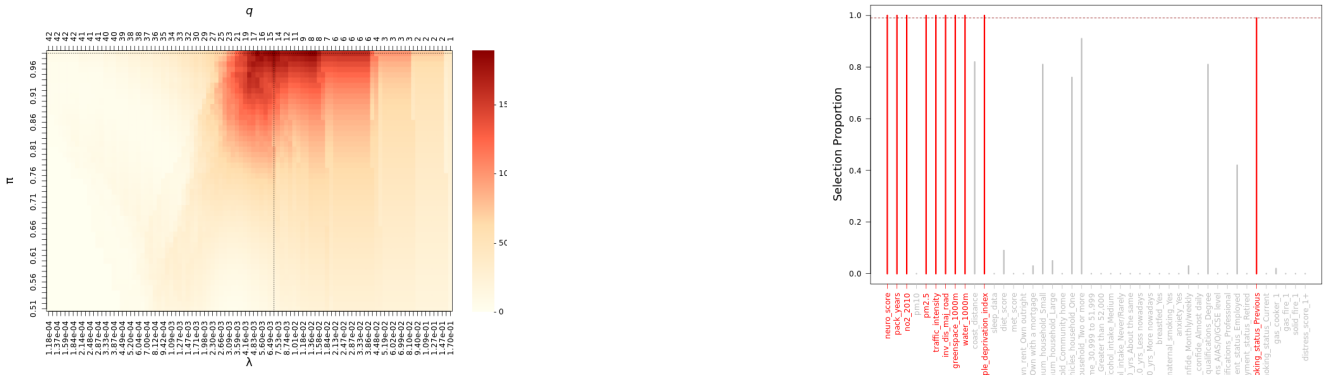

Figure 24: Calibration and selection plots exposures (cluster 7)

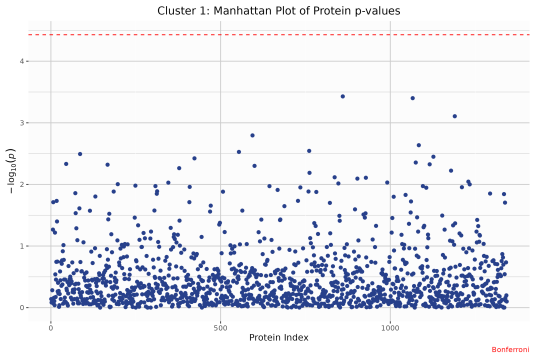

Figure 25: Manhattan plot proteins (cluster 1)

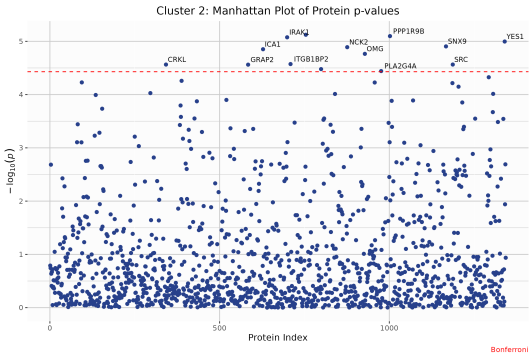

Figure 26: Manhattan plot proteins (cluster 2)

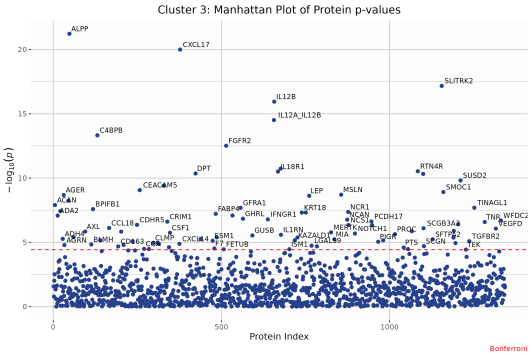

Figure 27: Manhattan plot proteins (cluster 3)

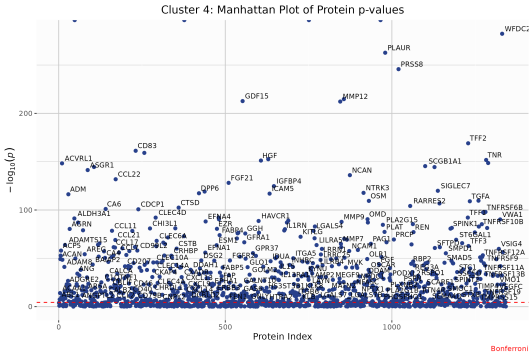

Figure 28: Manhattan plot proteins (cluster 4)

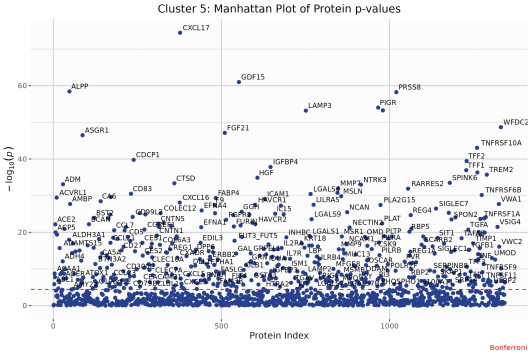

Figure 29: Manhattan plot proteins (cluster 5)

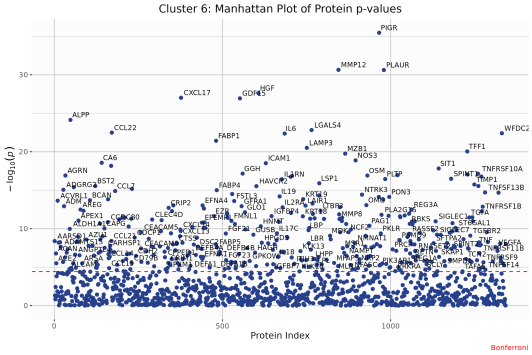

Figure 30: Manhattan plot proteins (cluster 6)

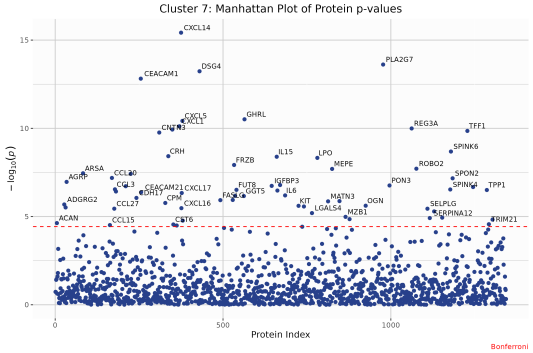

Figure 31: Manhattan plot proteins (cluster 7)

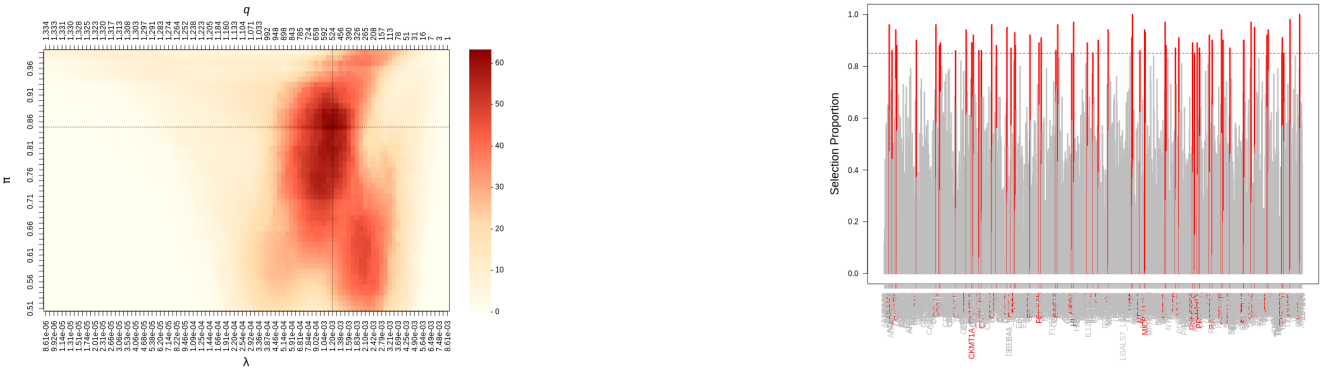

Figure 32: Calibration and selection plots proteins (cluster 1)

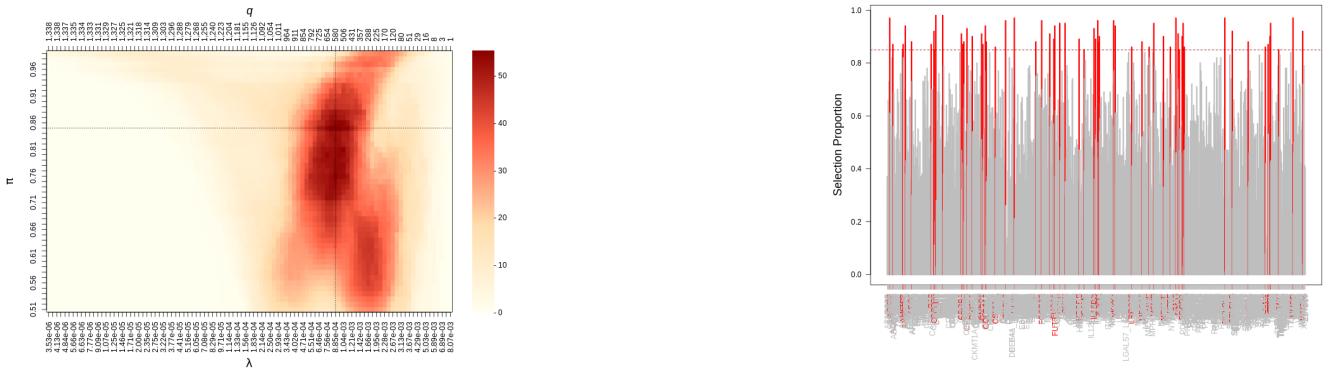

Figure 33: Calibration and selection plots proteins (cluster 2)

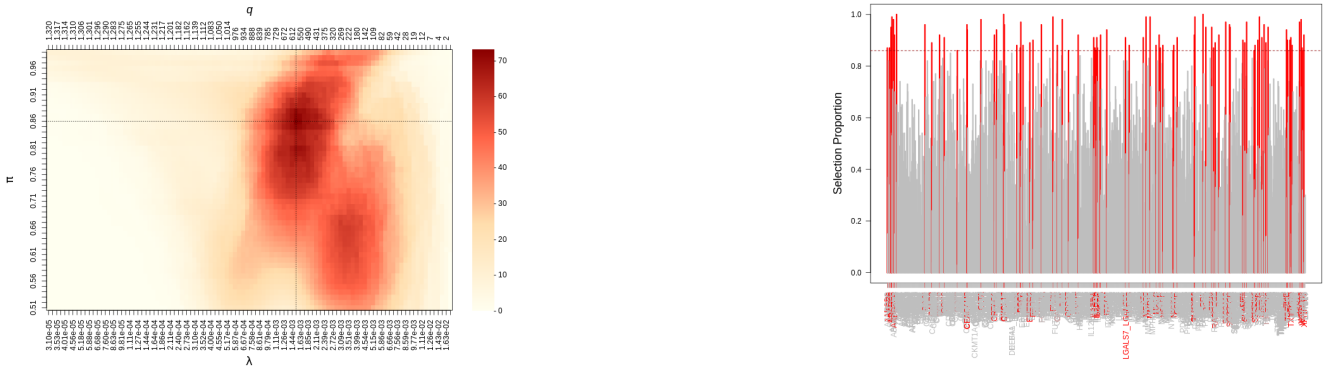

Figure 34: Calibration and selection plots proteins (cluster 3)

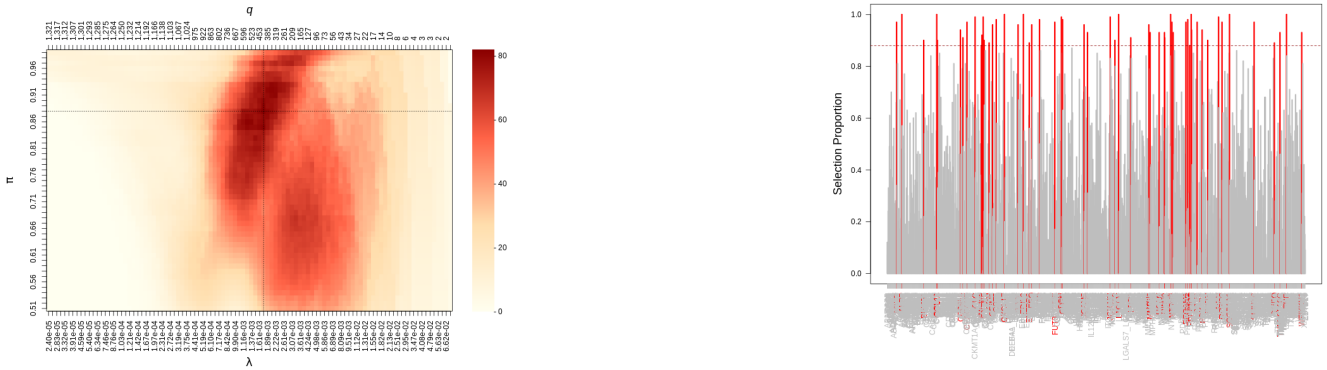

Figure 35: Calibration and selection plots proteins (cluster 4)

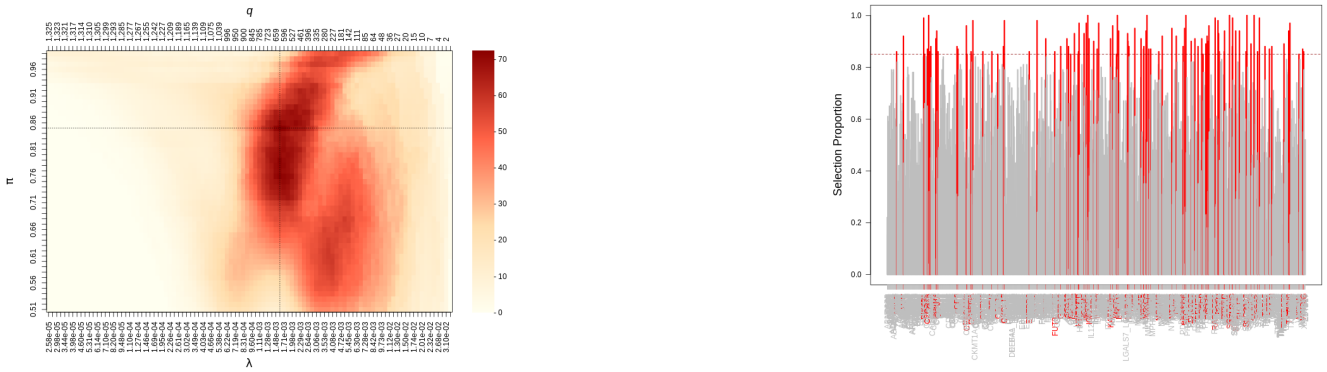

Figure 36: Calibration and selection plots proteins (cluster 5)

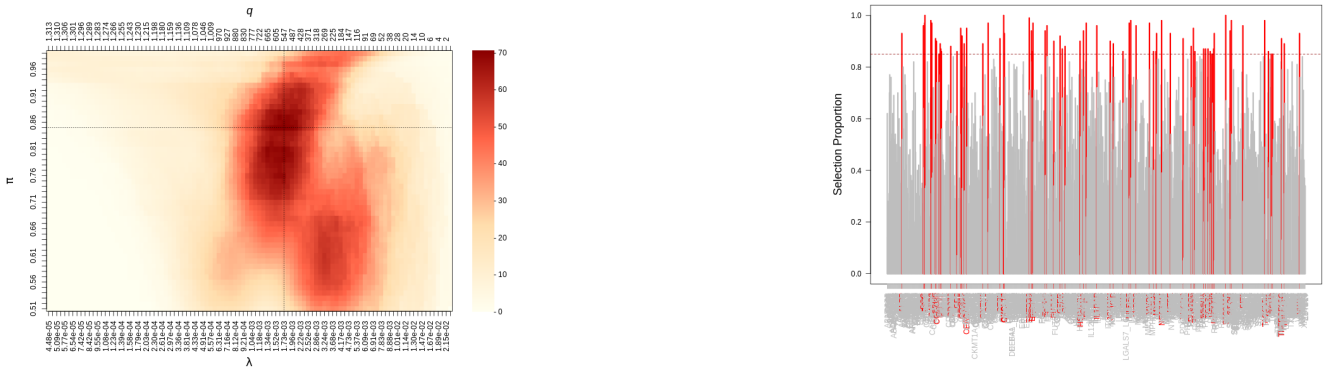

Figure 37: Calibration and selection plots proteins (cluster 6)

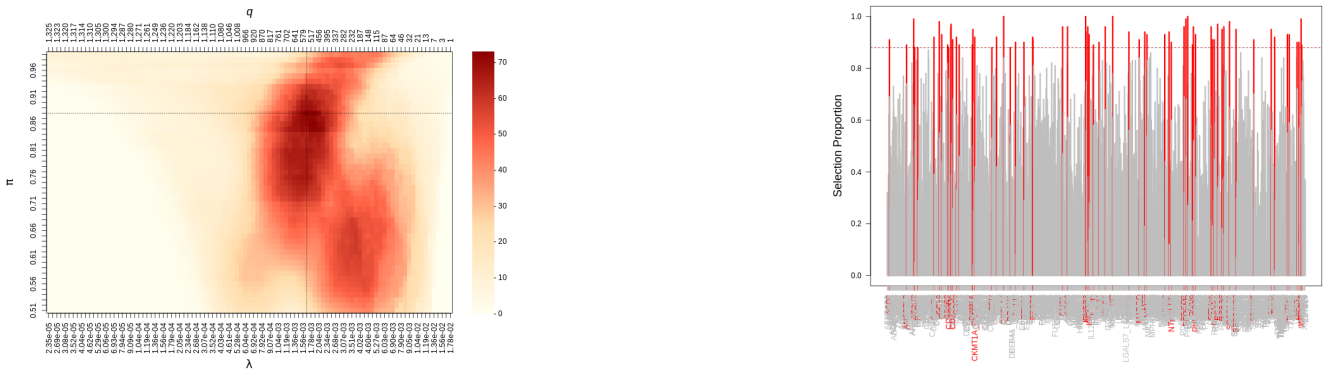

Figure 38: Calibration and selection plots proteins (cluster 7)