IMPERIAL

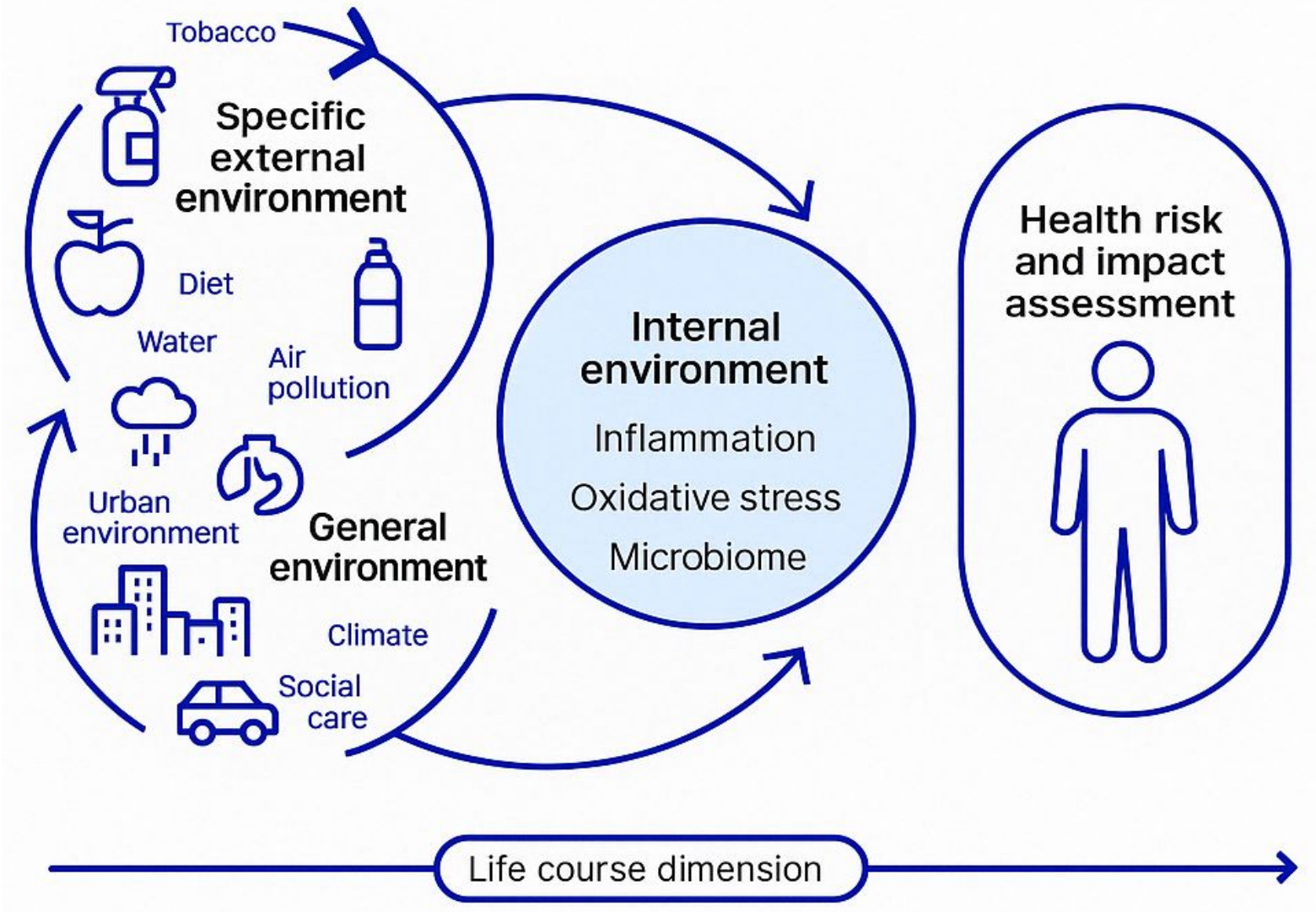
Expotype Clustering

Imogen Onno, Calix Tan, Tianshu Lu, Hannah Cooper
25/04/2025

Exposome

Why the exposome?

Traditional methods do not look at the joint effect of exposures and biological mechanisms, overlooking complex interactions



Exposome

What are we studying?

We aim to identify distinct exposure groups using unsupervised clustering, explore their differences through exposures and proteomic signatures, and how they affect disease risk

Cluster

Exposures

Proteomic Signatures

Diseases

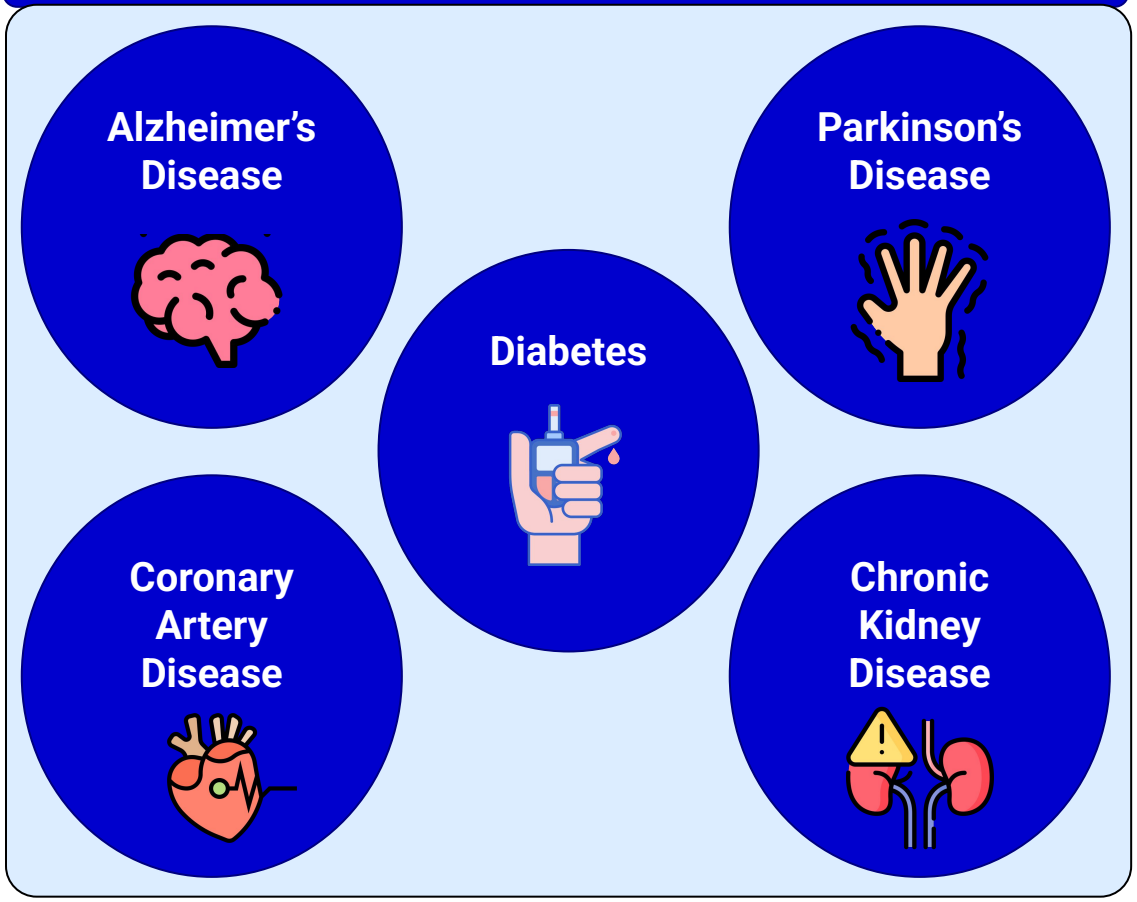
Introduction

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?

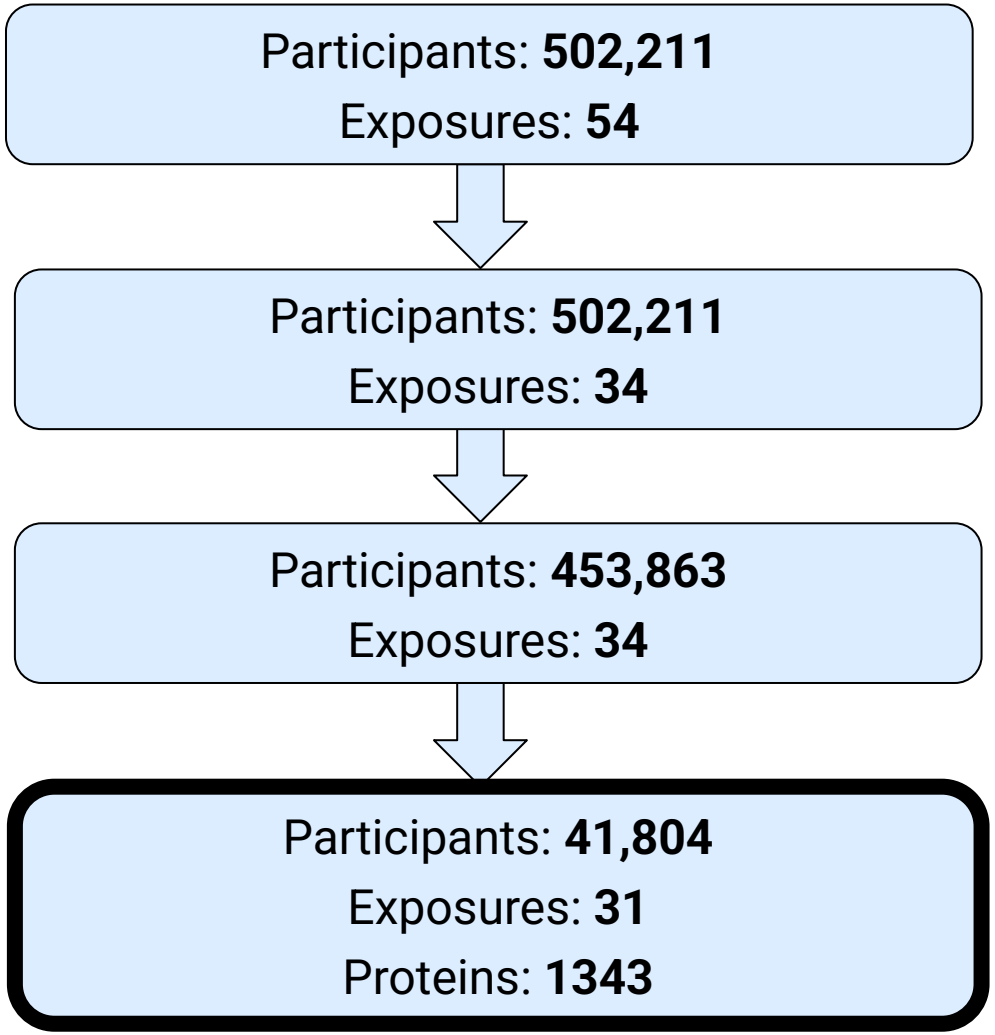
Chronic Health Outcomes



Methods

1	Data Preprocessing	Extraction, Missing data, Outlier Removal, Imputation, Scaling, Dummy Encoding
2	Clustering	Methods {Gaussian Mixture Model, HDDC, Biclustering, SOM + Hierarchical Clustering, Fuzzy Clustering}
3	Cluster Description	Univariate Analysis Stability Selection LASSO Logistic Regression with Selected Variables
4	Molecular Profiling	Univariate Analysis Stability Selection LASSO Logistic Regression with Selected Variables
5	Pathway Analysis	Reactome
6	Disease	Logistic Regression with Clusters Logistic Regression with Exposures

Data Preprocessing



1. Aggregation
2. Created sleep and distress score

1. Remove missing data
2. Imputation (MissRanger)

1. Merge with proteins
2. Remove outliers

Exposome Overview

Socio-economic

- Index of multiple deprivation
- Current employment status
- Qualifications
- Average household income before tax
- Own or rent accommodation
- Gas or solid-fuel cooking/heating
- Number in household
- Number of vehicles

Psychosocial

- Anxiety
- Able to confide
- Neuroticism score
- Illness, injury, bereavement, stress

Environmental

- Nitrogen dioxide air pollution 2010
- PM10 (2010)
- PM2.5
- Inverse distance to nearest major road
- Traffic intensity on nearest major road
- Average 24-hour sound level of noise pollution
- Greenspace percentage, buffer 1000m
- Water percentage, buffer 1000m
- Distance (Euclidean) to coast

Behavioural

- How many hours of sleep in a day
- Nap during the day
- Insomnia
- How often do you drink alcohol?
- Compared to 10 years ago, do you drink?
- Smoking status
- Pack years of smoking
- Exercise score
- Diet score

Childhood

- Maternal smoking
- Breastfed as a baby

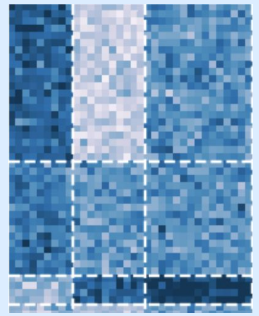
IMPERIAL

2. Clustering

Clustering: Method Explanations

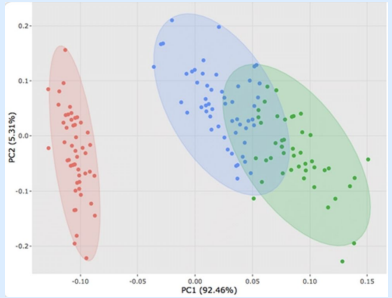
Biclustering

Clustering of both columns (exposures) and rows (individuals) to find local patterns



Fuzzy Clustering

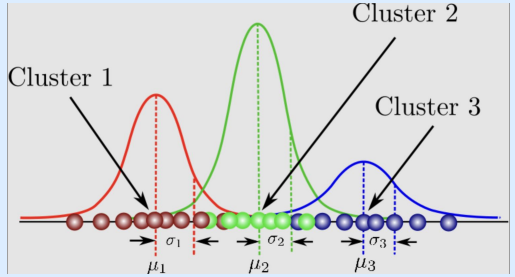
Each data point can belong to multiple clusters - suitable for handling uncertainty and overlapping distributions



Clustering: Method Explanations

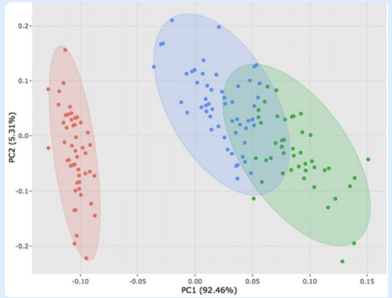
Gaussian Mixture Modeling

Probabilistic model that represents data as a combination of multiple Gaussian distributions



High-Dimensional Data Clustering (HDDC)

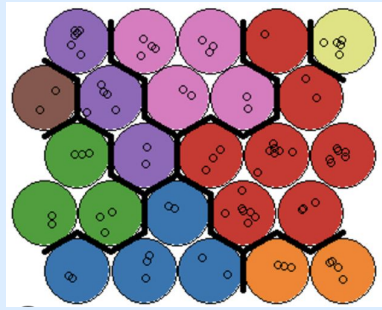
An extension of the Gaussian mixture model approach that aims to capture cluster-specific subspaces



Clustering: Method Explanations

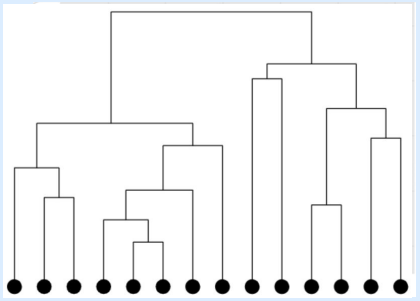
Self Organising Map (SOM)

First → Dimensionality Reduction
Artificial neural network that reduces dimensionality to represent distribution as a map



Hierarchical Clustering

Second → Clustering
Builds a hierarchy of clusters forming a tree-like structure in the form of a dendogram



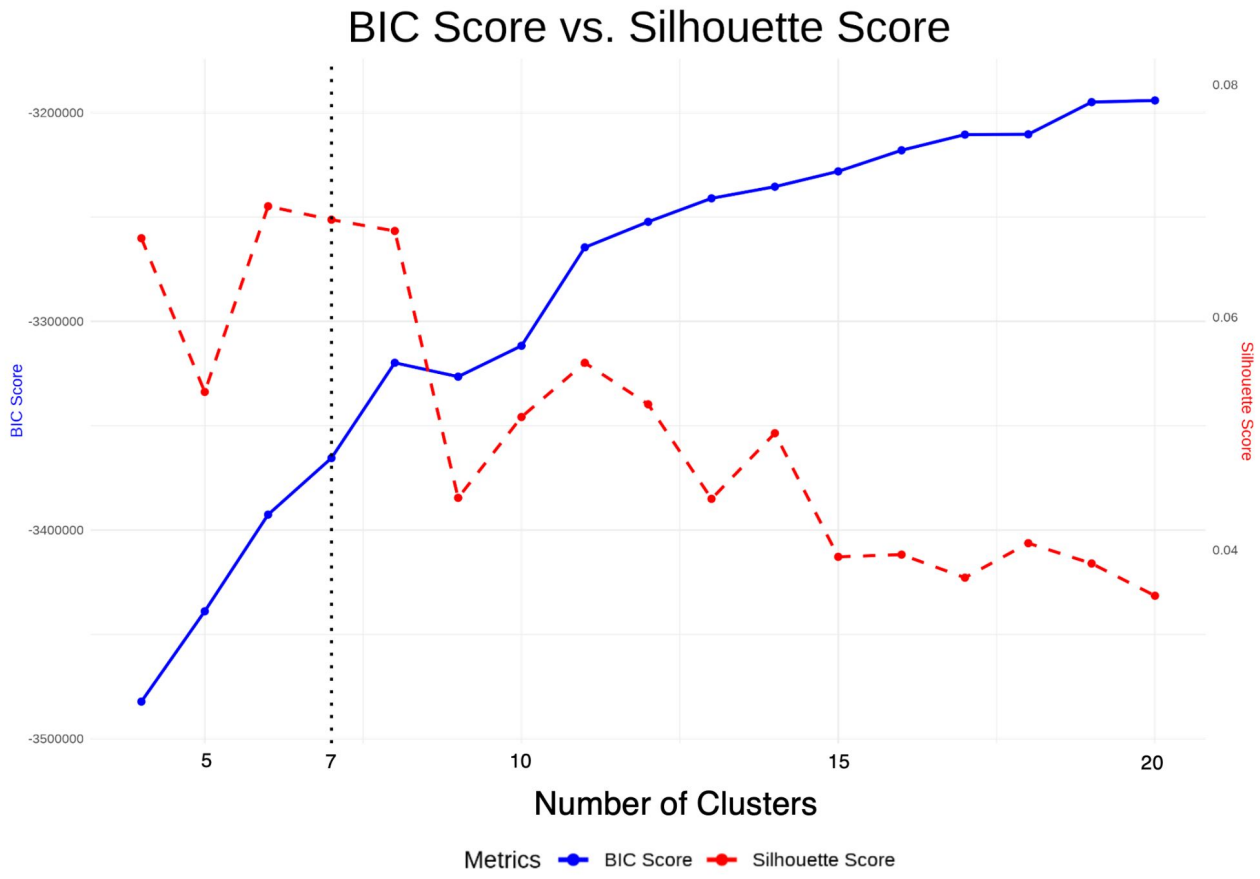
Clustering: Scoring Explanations

PAC Score	Silhouette Score	BIC Score
Measures cluster stability	Measures how well each point fits within it’s cluster	Evaluates model fit while penalising model complexity
Optimal clustering method	Optimal number of clusters	Optimal number of clusters
Lower = Better	Higher = Better	Lower = Better

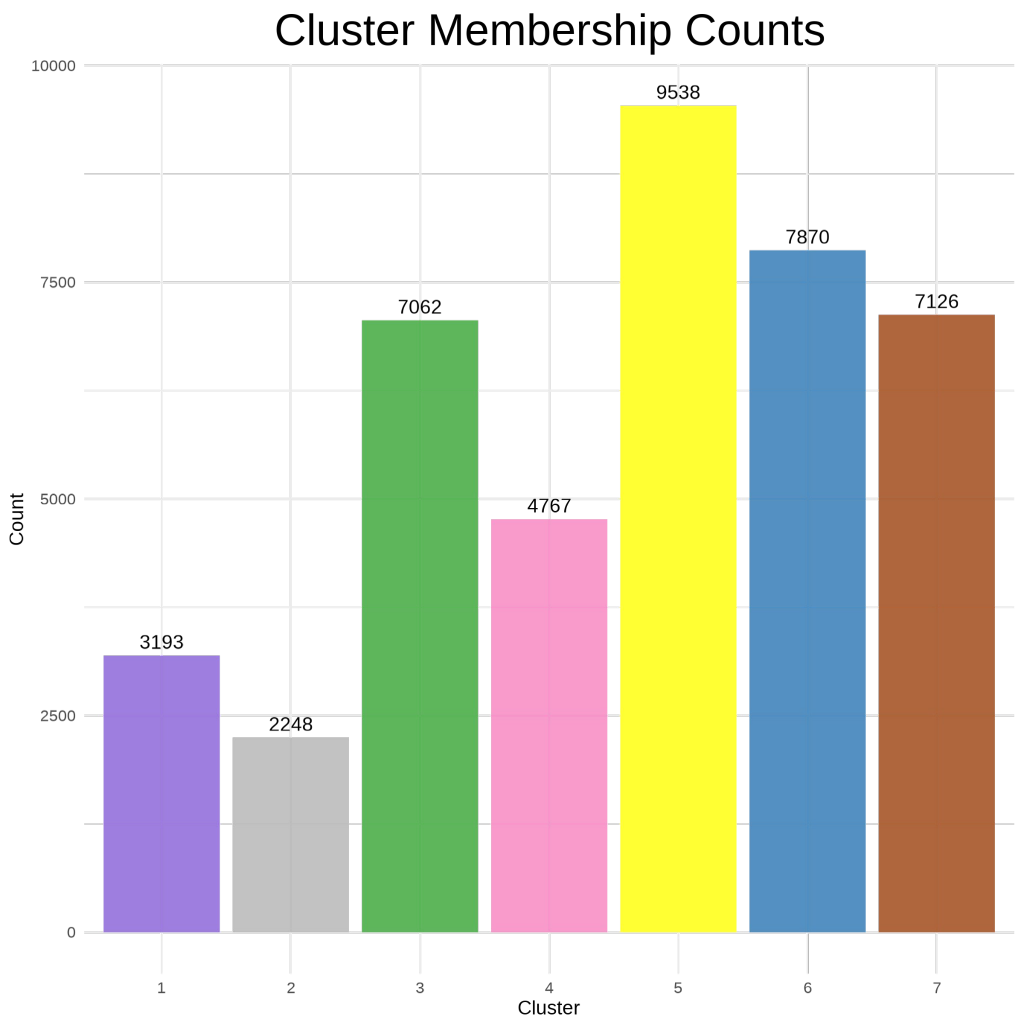
Clustering: Scoring Explanations

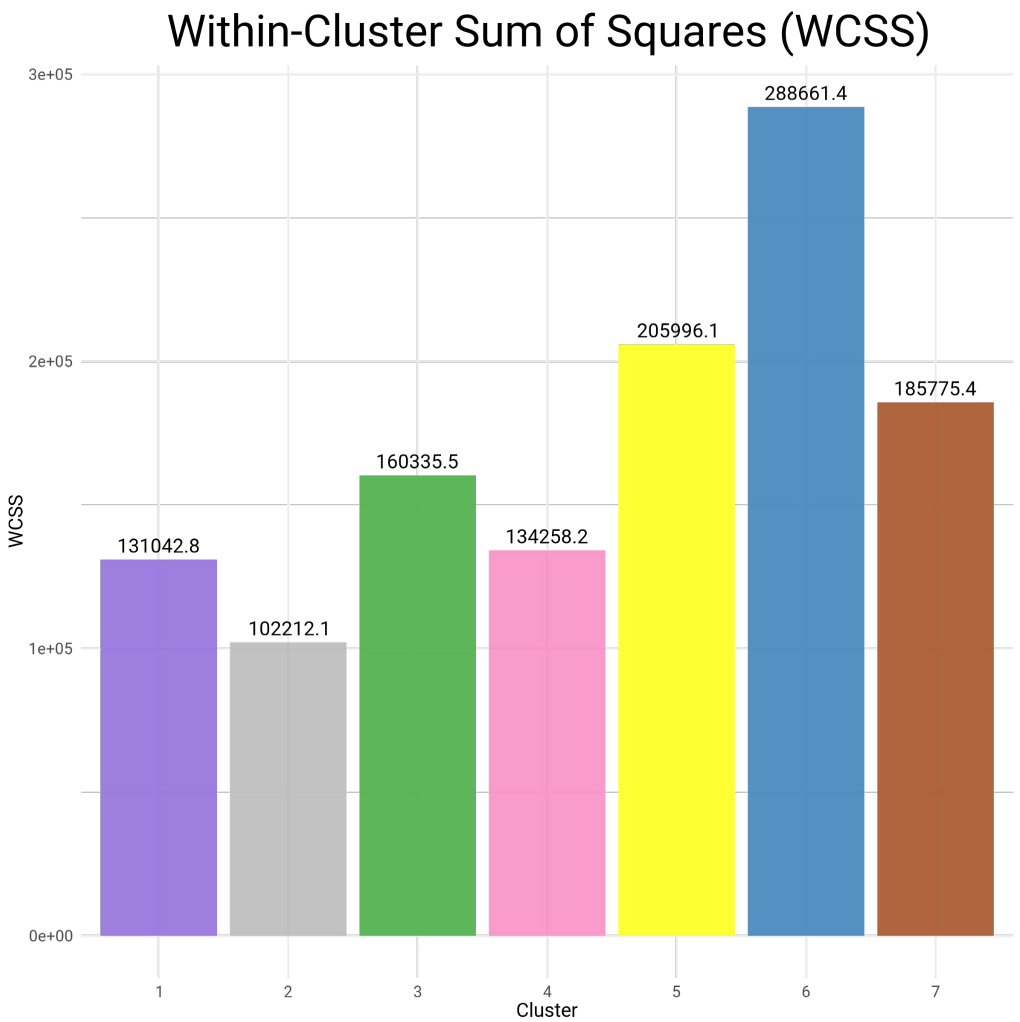
Chosen Method: Proportion of Ambiguous Clustering Scores

- **Gaussian Mixture Model**
  - PAC Score: 0.20
- **Bi Clustering**
  - PAC Score: 0.059
- **Fuzzy**
  - PAC Score: 0.531
- **SOM and Hierarchical**
  - PAC Score: 0.923
- **HDDC**



Cluster Overview
Gaussian Mixture Model

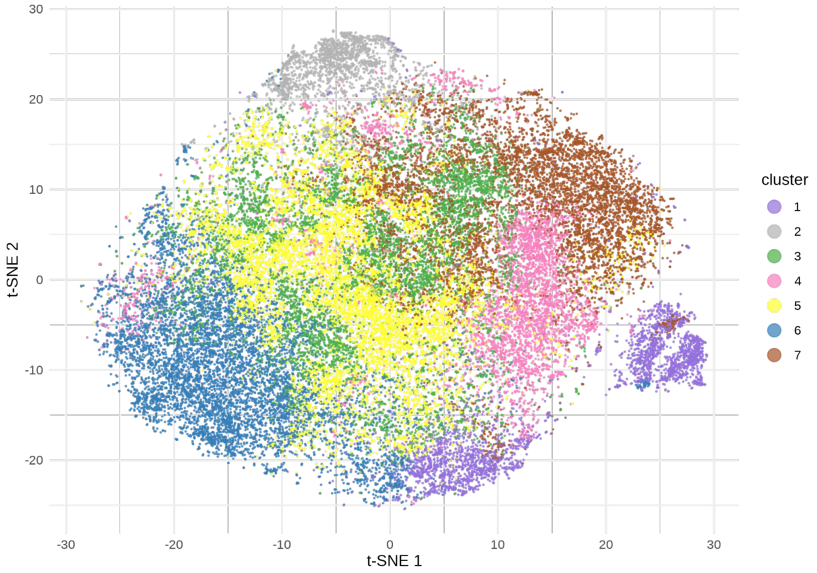




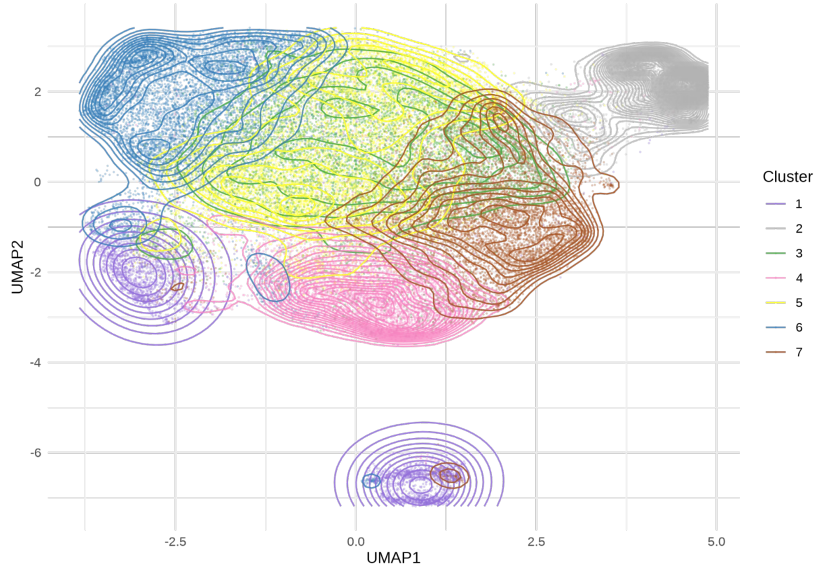
Cluster Overview
Gaussian Mixture Model

PCA Visualisation of GMM Clusters

t-SNE Visualisation of GMM Clusters



UMAP Visualisation of GMM Clusters



IMPERIAL

3. Cluster Description

Cluster Description: Exposures

Univariate Analysis
(32 x 7 models)

Logistic regression models, for exposure j to cluster i:

Y: Binary, allocation to cluster i (1) or any other cluster (0)

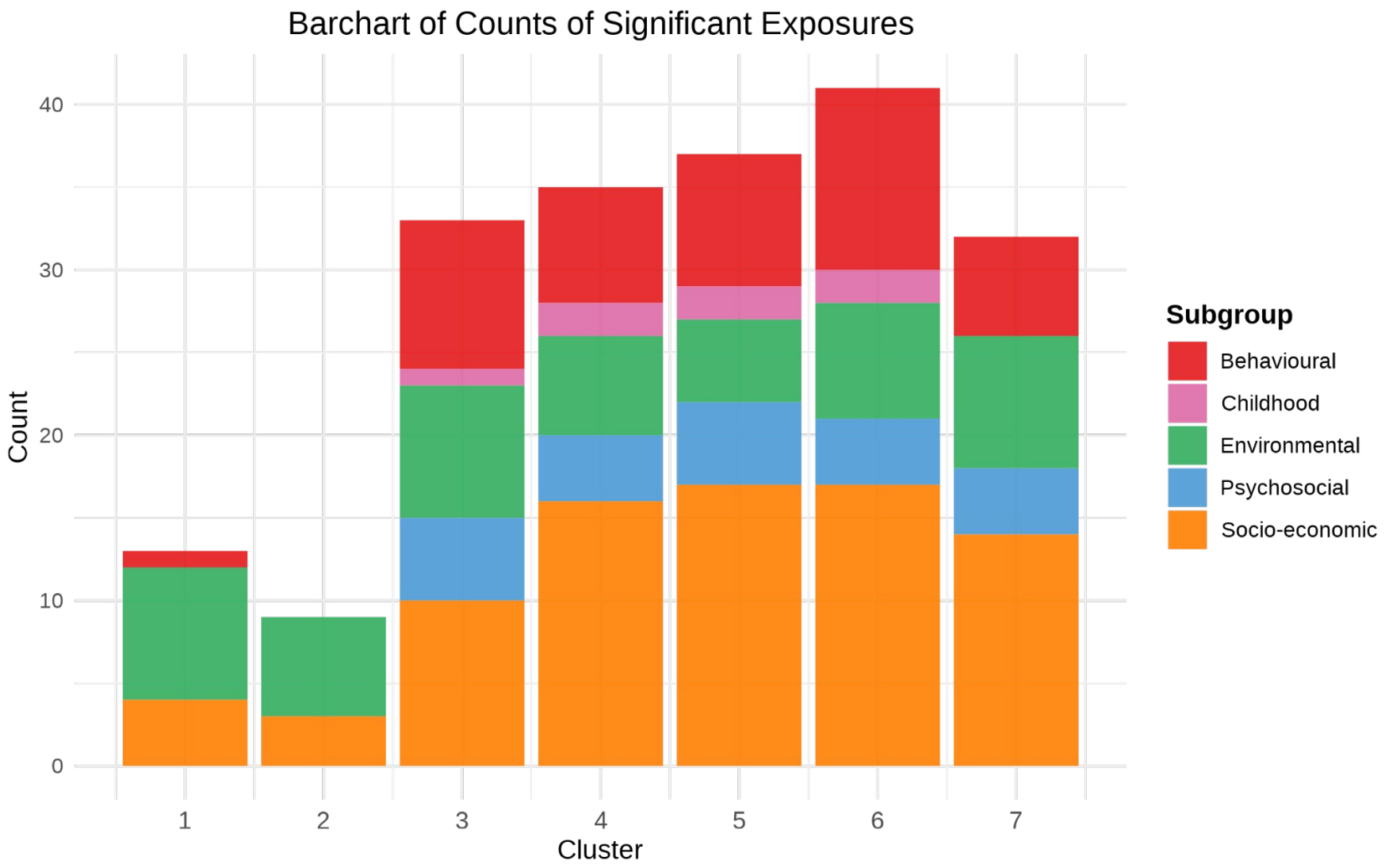
X: Exposure j

Confounders: Age, Sex, Ethnic background

Stability Selection LASSO

Logistic Regression

Univariate Analysis: Exposures



	Total number of exposures ($p < 0.0011$)
Cluster 1	13
Cluster 2	9
Cluster 3	33
Cluster 4	35
Cluster 5	37
Cluster 6	41
Cluster 7	32

Cluster Description: Exposures

Univariate Analysis

Stability Selection LASSO
(7 models)

Stability Selection LASSO for each cluster:

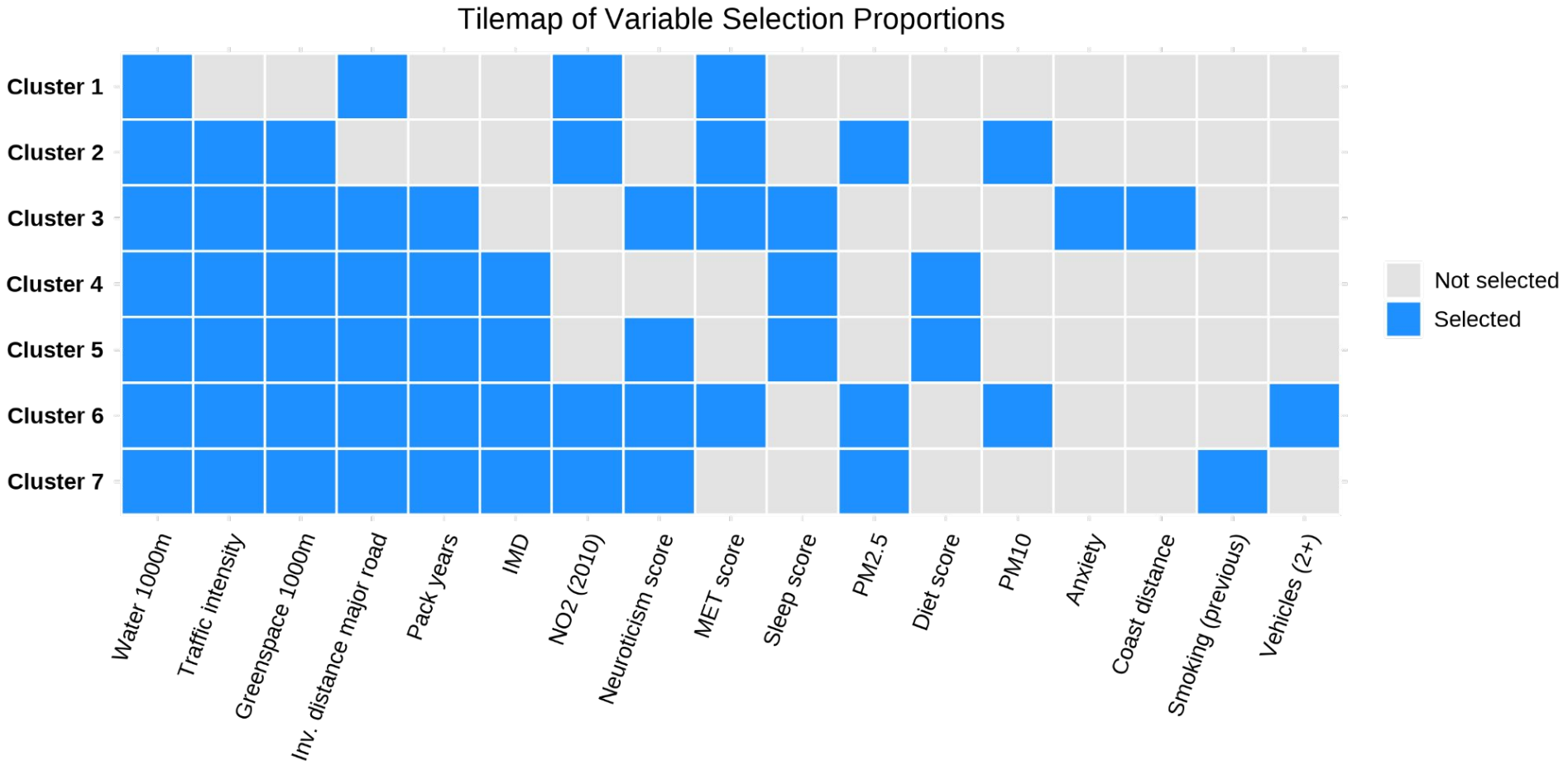
Y: Binary, allocation to cluster i (1) or any other cluster (0)

X: 32 exposures

Confounders: Age, Sex, Ethnic background

Logistic Regression

Stability Selection LASSO: Exposures



	Lambda	Pi
Cluster 1	0.0075	0.99
Cluster 2	0.0016	0.97
Cluster 3	0.013	0.99
Cluster 4	0.0080	0.99
Cluster 5	0.028	0.99
Cluster 6	0.0044	0.99
Cluster 7	0.00070	0.99

Cluster Description: Exposures

Univariate Analysis

Stability Selection LASSO

Logistic Regression
(7 models)

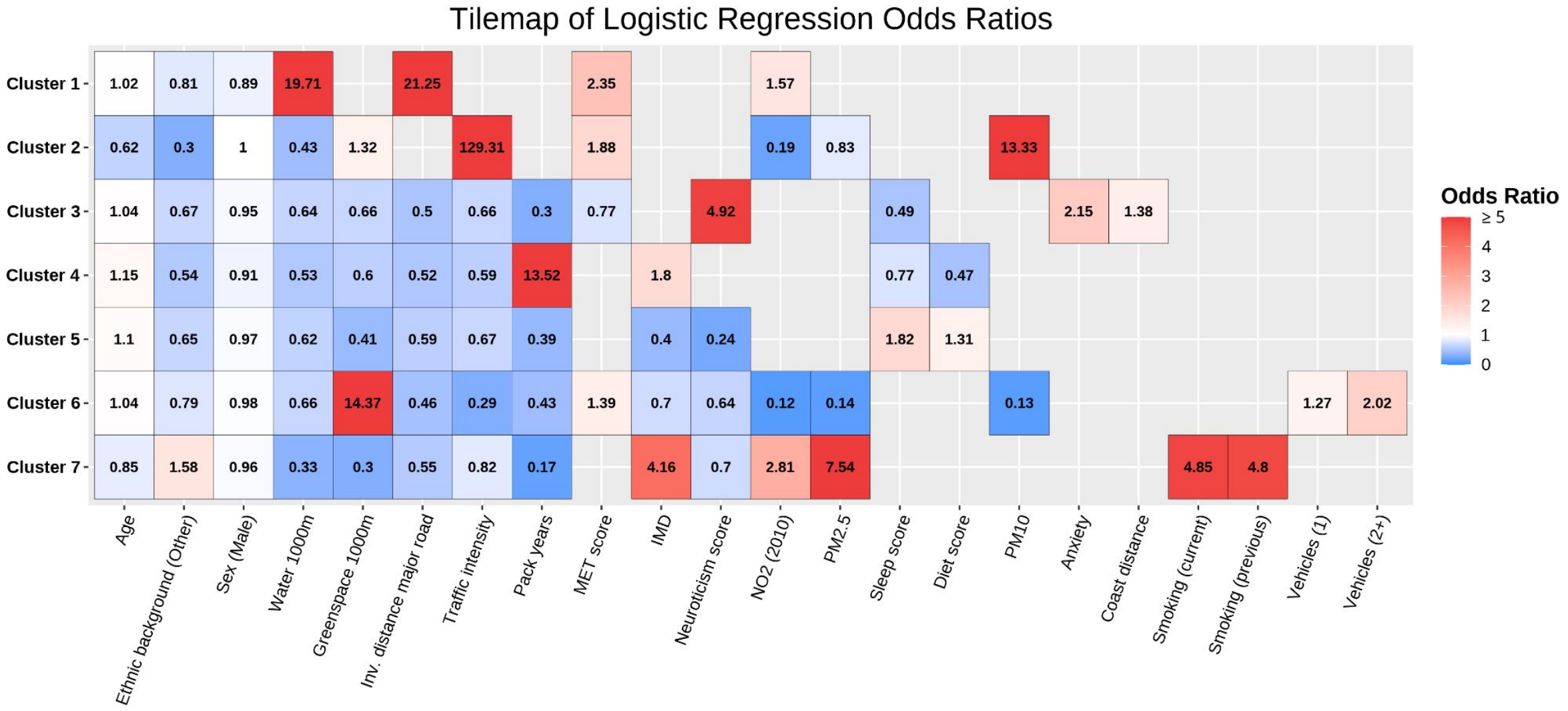
Logistic Regression for each cluster:

Y: Binary, allocation to cluster i (1) or any other cluster (0)

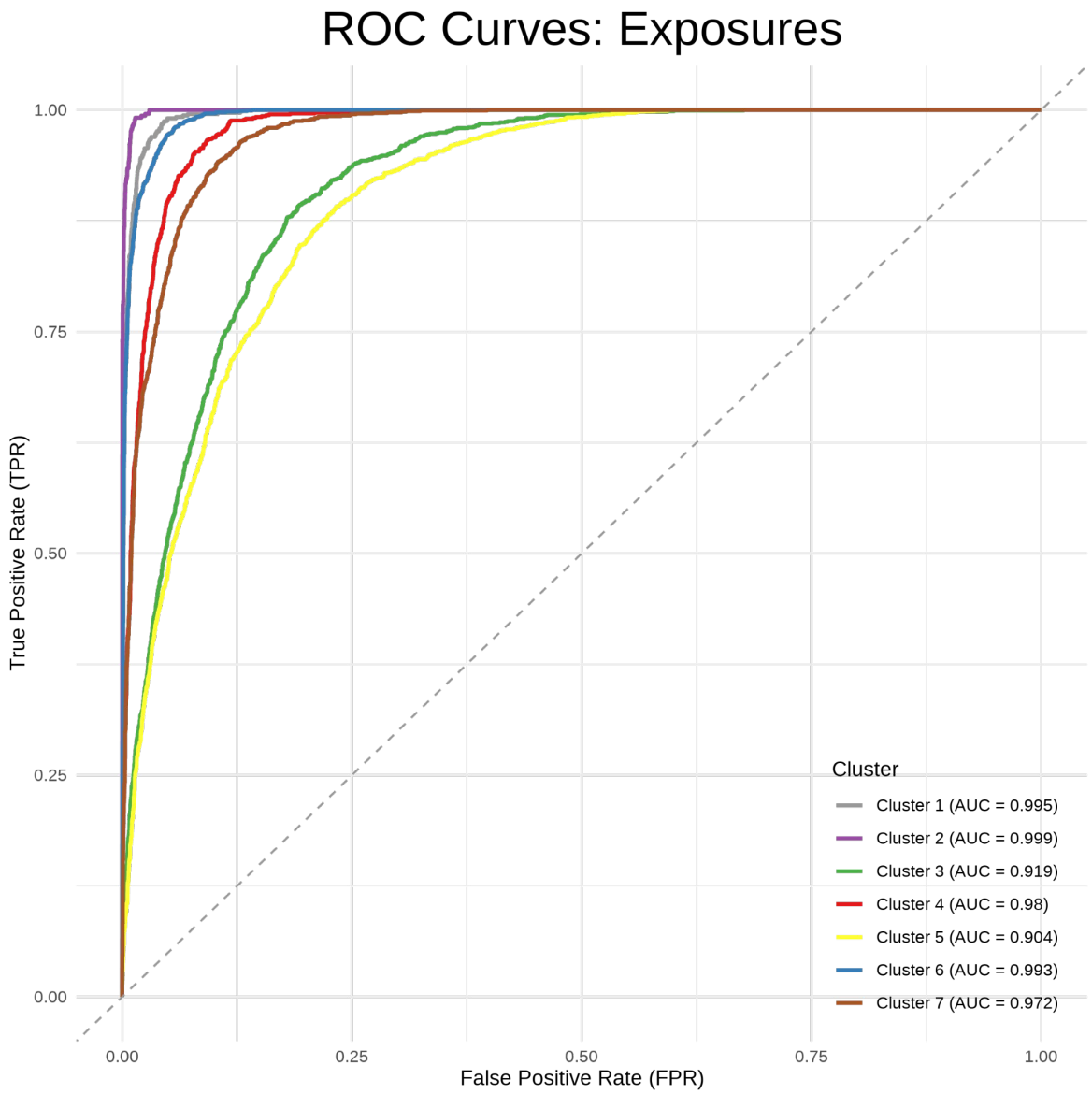
X: Exposures selected in Stability Selection LASSO

Confounders: Age, Sex, Ethnic background

Logistic Regression: Exposures



ROC Curves: Exposures



All AUC values > 0.90

Clusters 1, 2 and 6 have AUC values > 0.99

The logistic regression models with stably selected exposures are very successful at cluster allocation

IMPERIAL

4. Molecular Profiling

Molecular Profiling: Proteins

Univariate Analysis
(1343 x 7 models)

Logistic regression models, for protein j to cluster i:

Y: Binary, allocation to cluster i (1) or any other cluster (0)

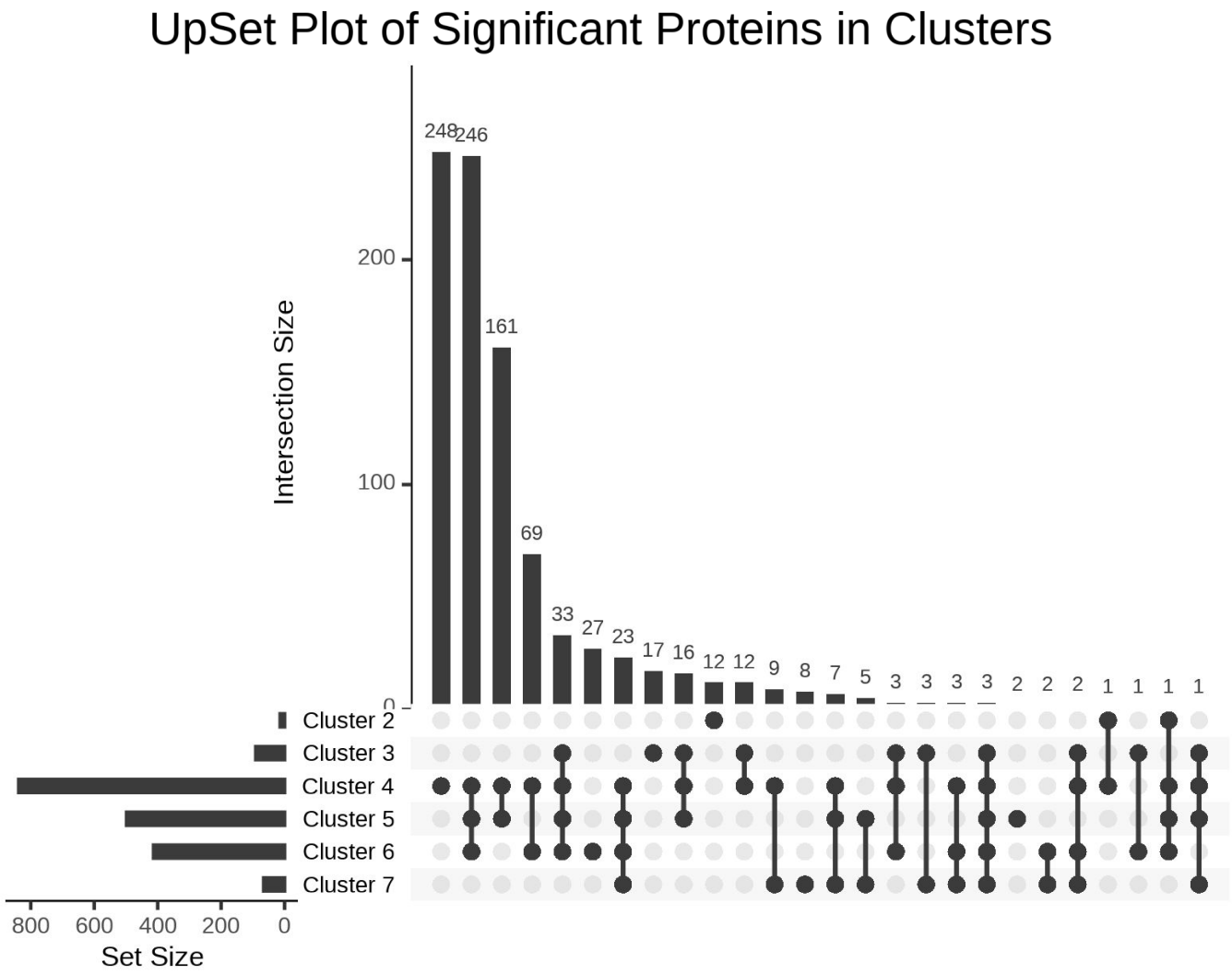
X: Protein j

Confounders: Age, Sex, Ethnic background

Stability Selection LASSO

Logistic Regression

Univariate Analysis: Proteins



	Total number of proteins ($p < 0.00004$)
Cluster 1	0
Cluster 2	14
Cluster 3	91
Cluster 4	838
Cluster 5	498
Cluster 6	413
Cluster 7	66

Molecular Profiling: Proteins

Univariate Analysis

Stability Selection LASSO
(7 models)

Stability Selection LASSO for each cluster:

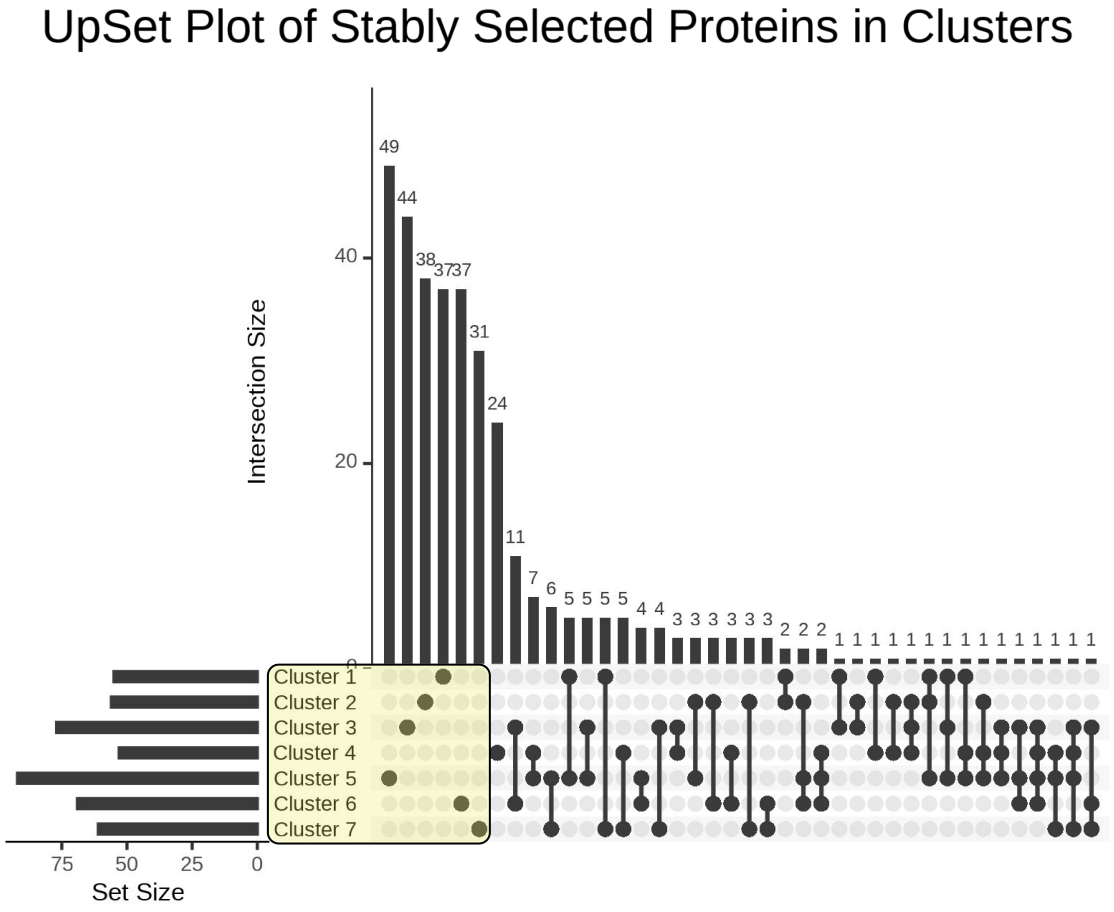
Y: Binary, allocation to cluster i (1) or any other cluster (0)

X: 1343 proteins

Confounders: Age, Sex, Ethnic background

Logistic Regression

Stability Selection LASSO: Proteins



	Total number of selected proteins
Cluster 1	55
Cluster 2	56
Cluster 3	77
Cluster 4	53
Cluster 5	92
Cluster 6	69
Cluster 7	61

	Lambda	Pi
Cluster 1	0.00120	0.85
Cluster 2	0.00089	0.85
Cluster 3	0.00153	0.86
Cluster 4	0.00174	0.88
Cluster 5	0.00159	0.85
Cluster 6	0.00173	0.85
Cluster 7	0.00167	0.88

Molecular Profiling: Proteins

Univariate Analysis

Stability Selection LASSO

Logistic Regression
(7 models)

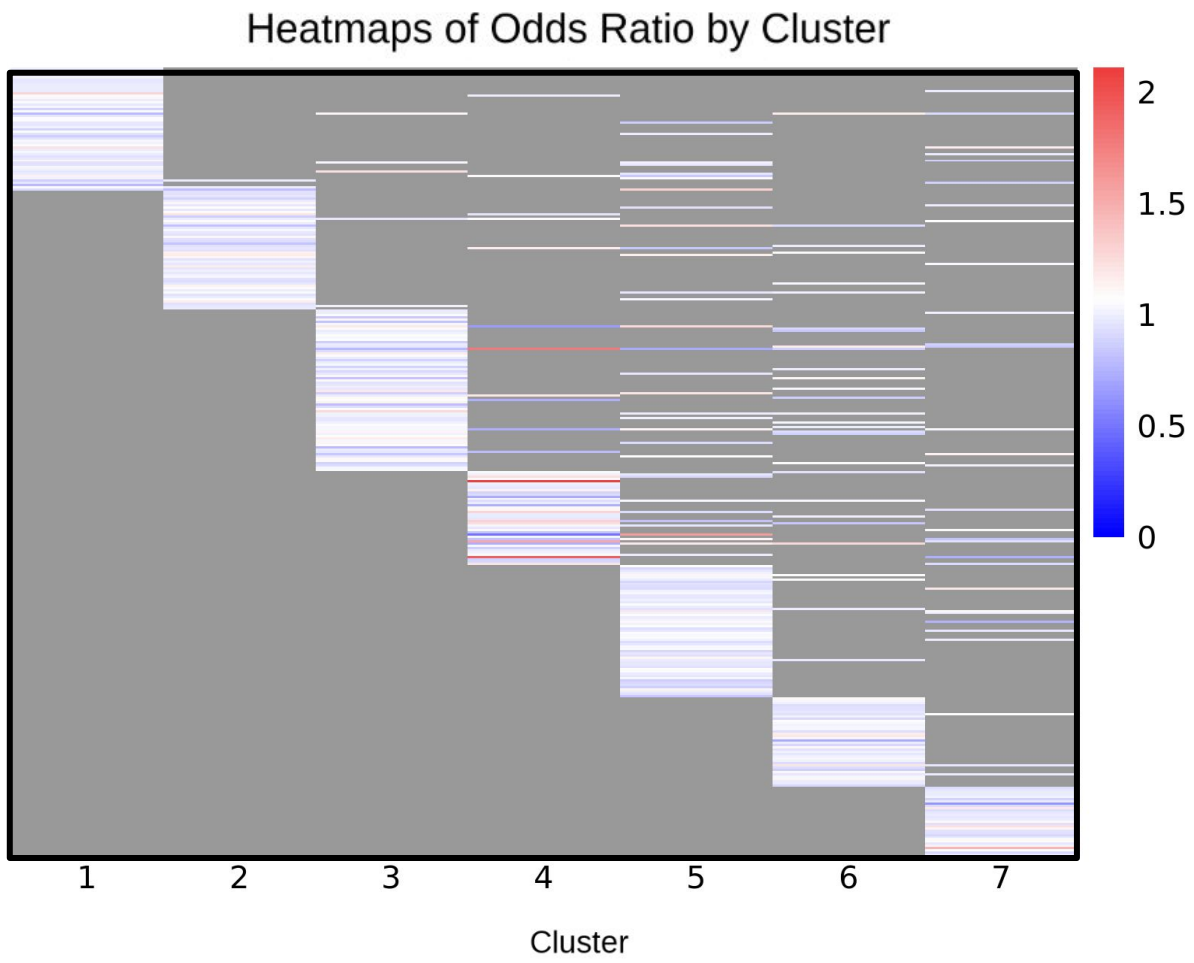
Logistic Regression for each cluster:

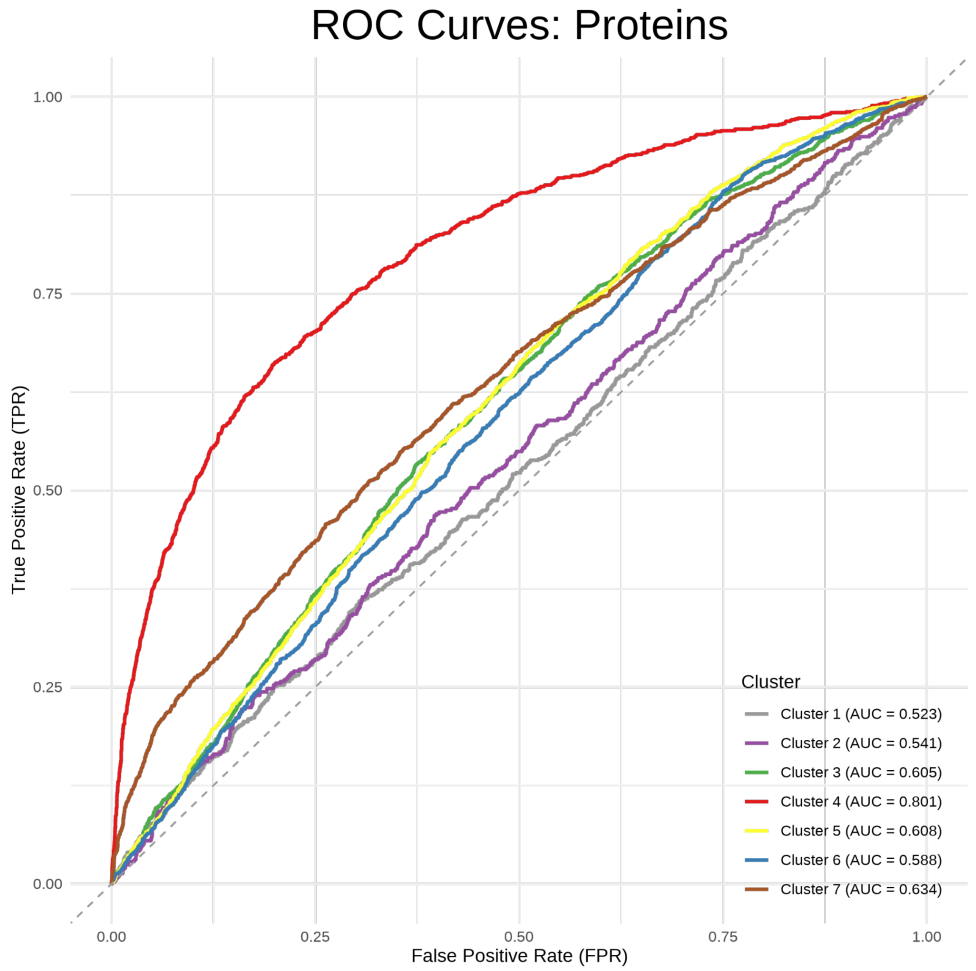
Y: Binary, allocation to cluster i (1) or any other cluster (0)

X: Proteins selected in Stability Selection LASSO

Confounders: Age, Sex, Ethnic background

Logistic Regression: Proteins



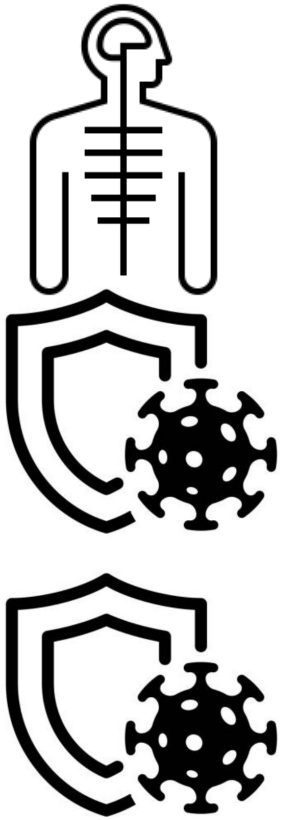


IMPERIAL

5. Pathway Analysis

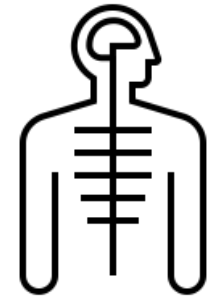
Enrichment Analysis of Clusters

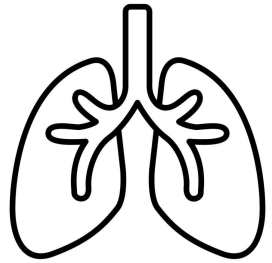
Overview	• Input each cluster's proteins into reactome.com • Pick top three significant pathways based on Entities FDR (<0.05) for each cluster	
Cluster 1	**Nervous system**	Downregulation of ERBB4 signaling; Interleukin-10 signaling; Nuclear signaling by ERBB4
Cluster 2	**Immune system**	Interleukin-10 signaling; Signaling by Interleukins; Immune System
Cluster 3	**Immune system**	Cytokine Signaling in Immune system; Signaling by Interleukins; Downregulation of ERBB4 signaling

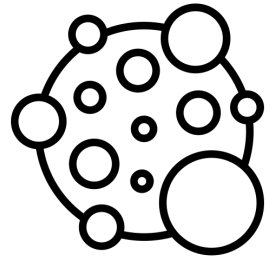


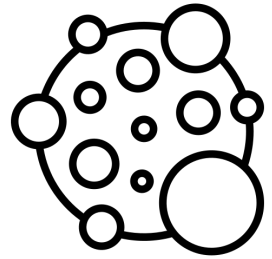
Enrichment Analysis of Clusters

Cluster 4	**Nervous system**	Synthesis, secretion, and deacylation of Ghrelin; NTF3 activates NTRK3 signaling; Post-translational modification: synthesis of GPI-anchored proteins
Cluster 5	**Respiratory system & Immune system**	TFAP2 (AP-2) family regulates transcription of growth factors and their receptors; Reversible hydration of carbon dioxide; Interleukin-33 signaling
Cluster 6	**Cancer system & Immune system**	PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling; Constitutive Signaling by Aberrant PI3K in Cancer; Negative regulation of the PI3K/AKT network
Cluster 7	**Cancer system**	Constitutive Signaling by Aberrant PI3K in Cancer; ERBB2 Activates PTK6 Signaling; ERBB2 Regulates Cell Motility









IMPERIAL

6. Diseases

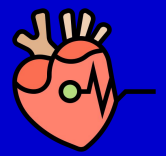
Disease Associations

Alzheimer's Disease

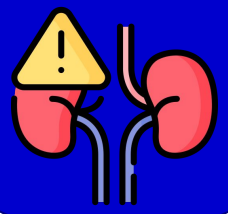
Diabetes

Parkinson's Disease

Coronary Artery Disease



Chronic Kidney Disease



Overview of Methods for Diseases

Incidence Rate

Calculated and standardised by the number of individuals per cluster

Prediction

Logistic regression used to assess if disease status can be predicted by:

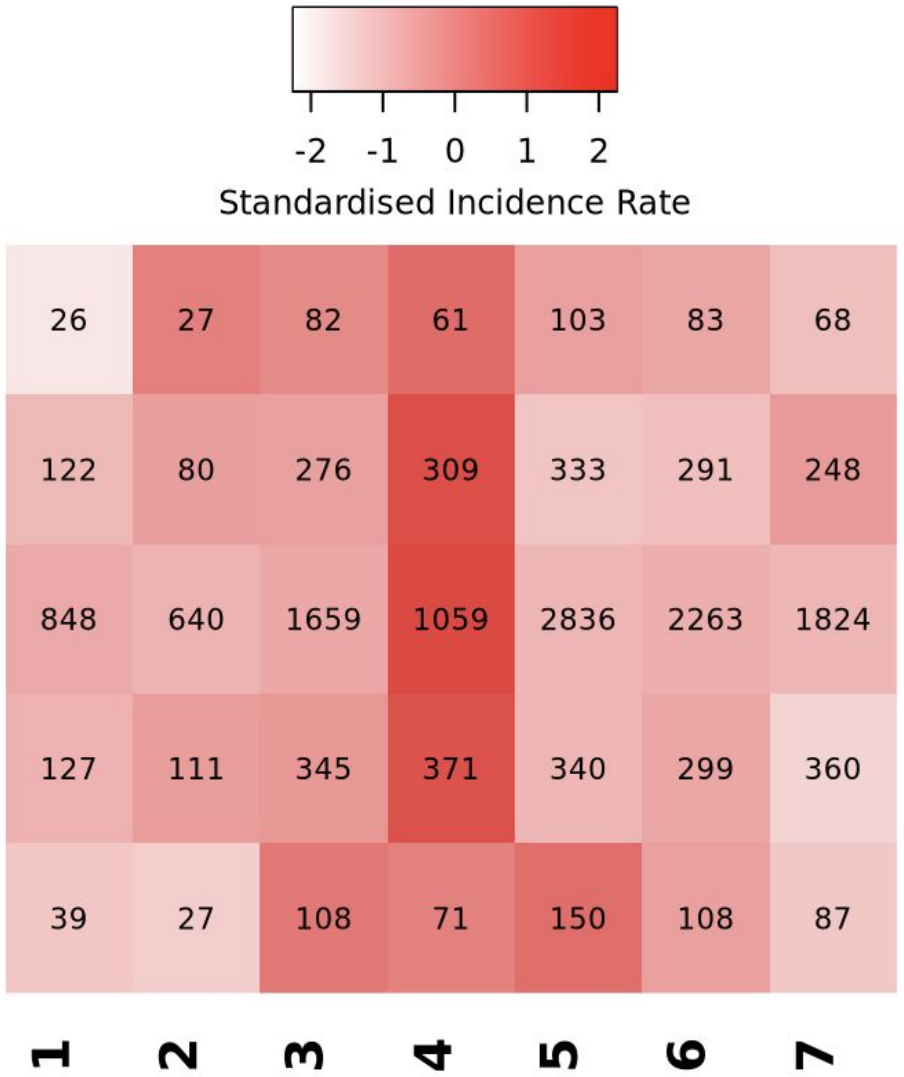
- Cluster membership
- Exposures
- Proteins

All models adjusted for age, sex, and ethnic background

Model Variants

Cluster
Cluster + Exposure
Cluster + Exposure + Proteins

Incidence Rate



AD (450; 41,698)

PD (1953; 39,518)

CKD (11,129; 15,380)

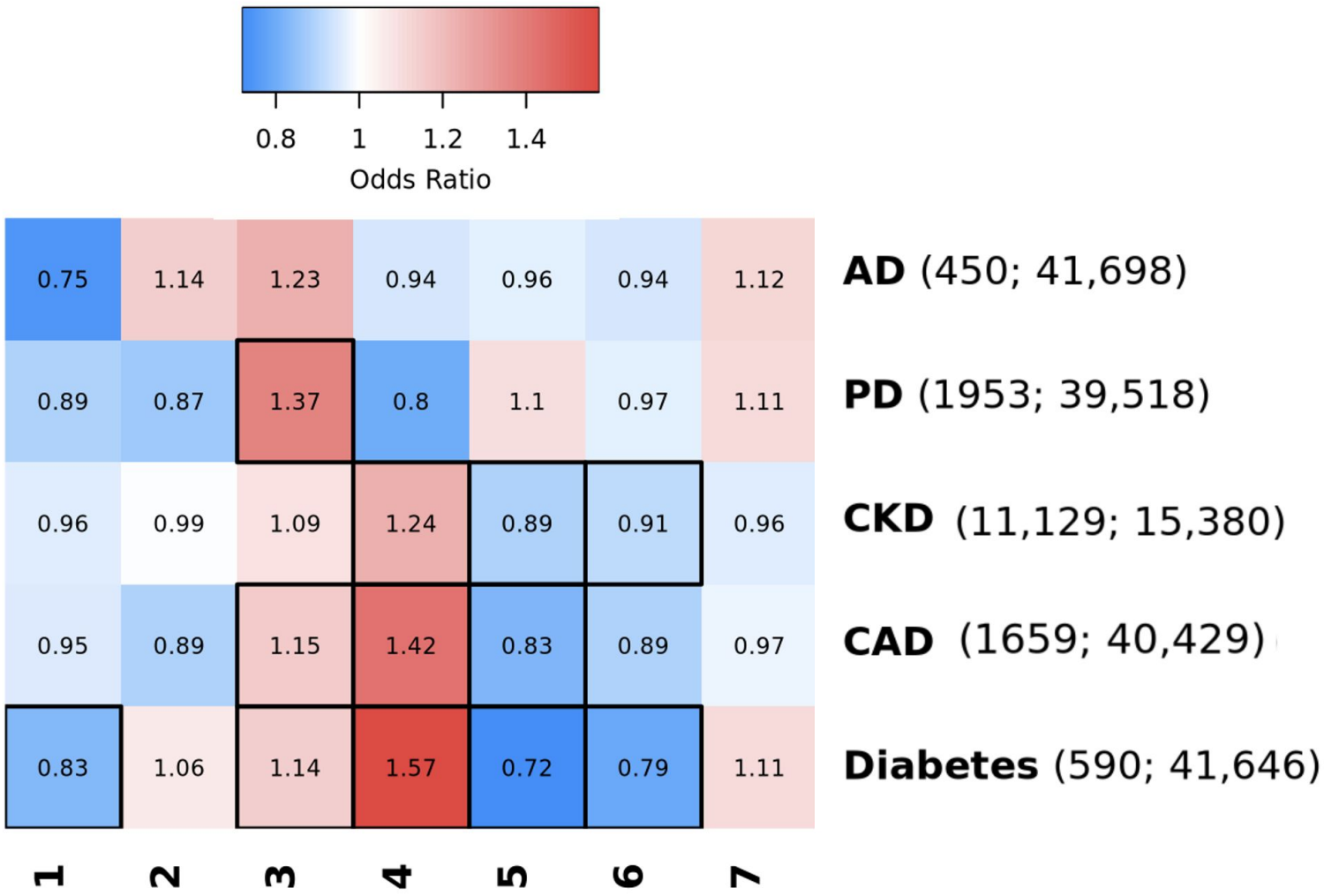
CAD (1659; 40,429)

Diabetes (590; 41,646)

Cluster 4 shows the highest incidence rates

Clusters 1, 2, and 7 show lower overall incidence rates, suggesting healthier profiles

Disease Prediction



AD (450; 41,698)

PD (1953; 39,518)

CKD (11,129; 15,380)

CAD (1659; 40,429)

Diabetes (590; 41,646)

Zero-sum encoding used for clusters

Clusters 3 and 4 are associated with an increased risk for certain diseases

Clusters 1, 5, and 6 appear to be protective against certain disease

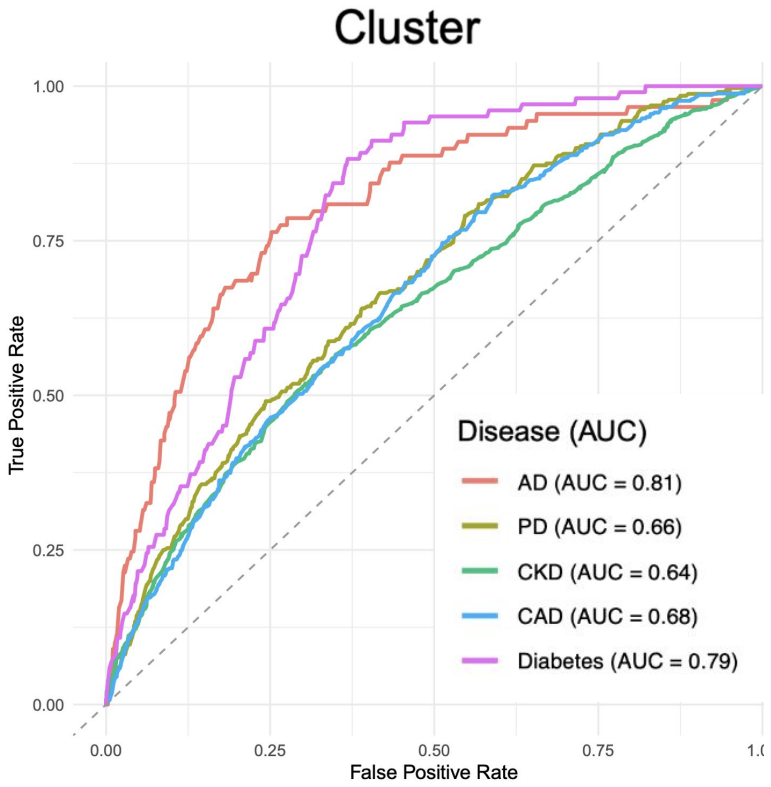
Highest predictive performance for AD

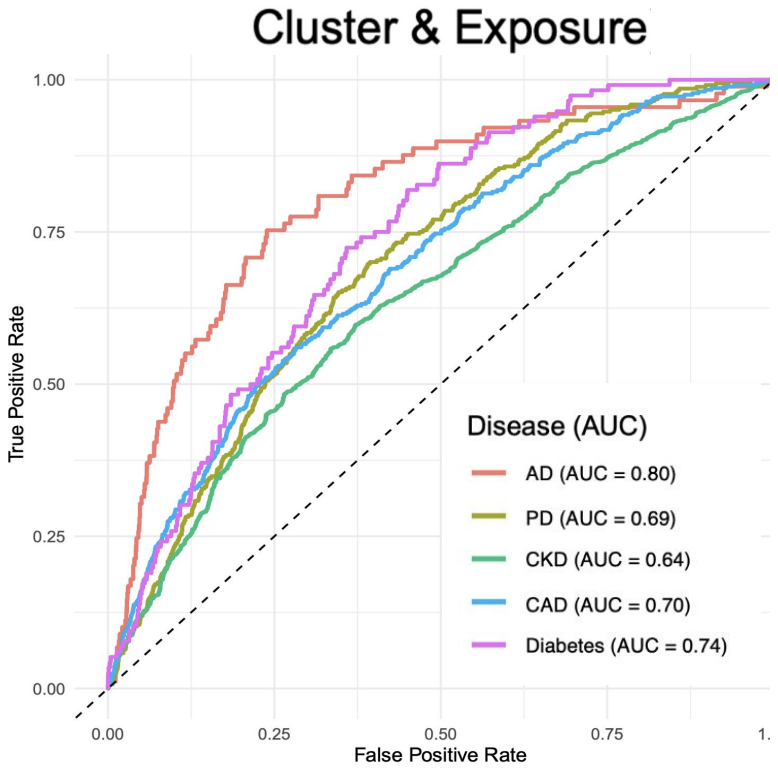
Lowest predictive performance for CKD

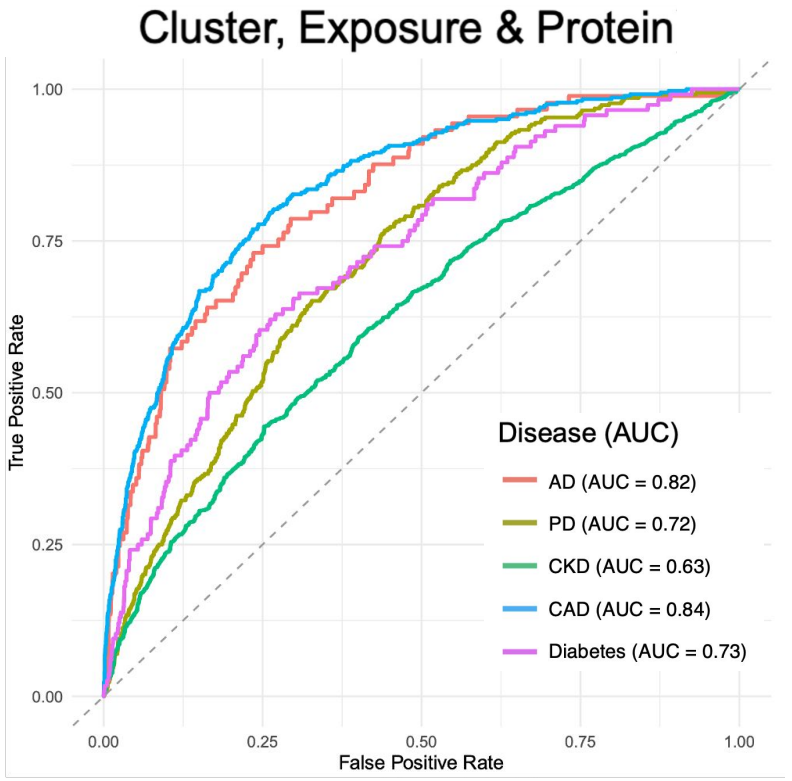
Cluster + Exposure + Protein yields best overall model performance

Disease Prediction Model Performance

ROC Curves by Disease



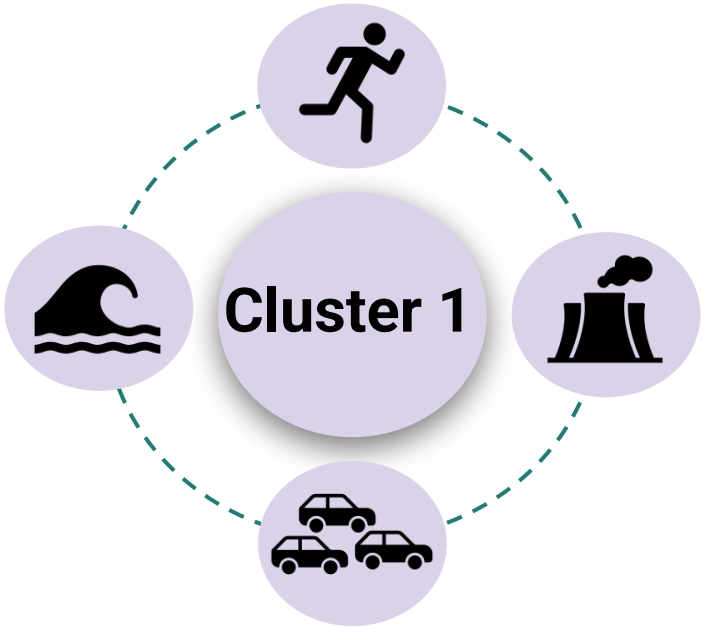




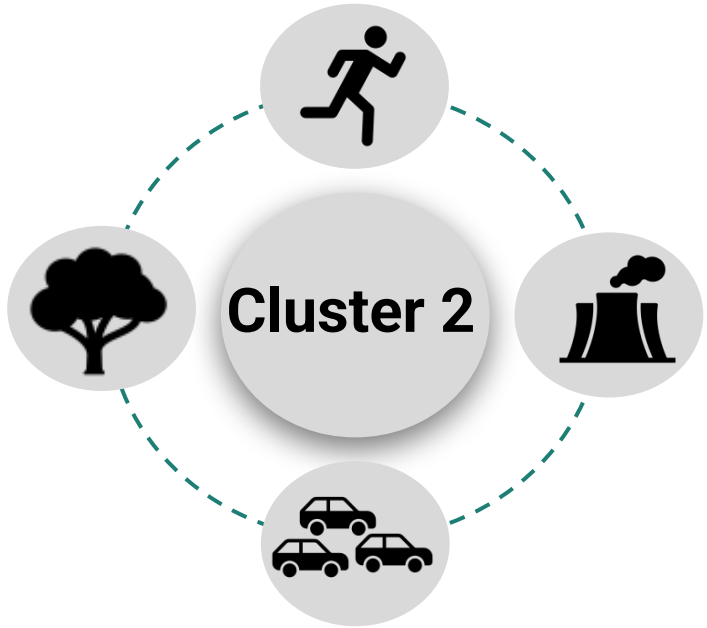
IMPERIAL

7. Discussion

Expotype Signatures

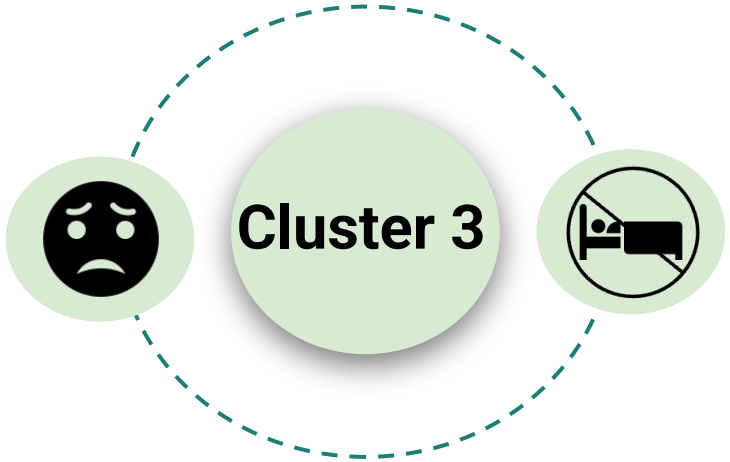


- **Exposures:** Higher pollution and exercise
- **Pathway analysis:** Nervous system
- **Disease:** Lower odds of Diabetes

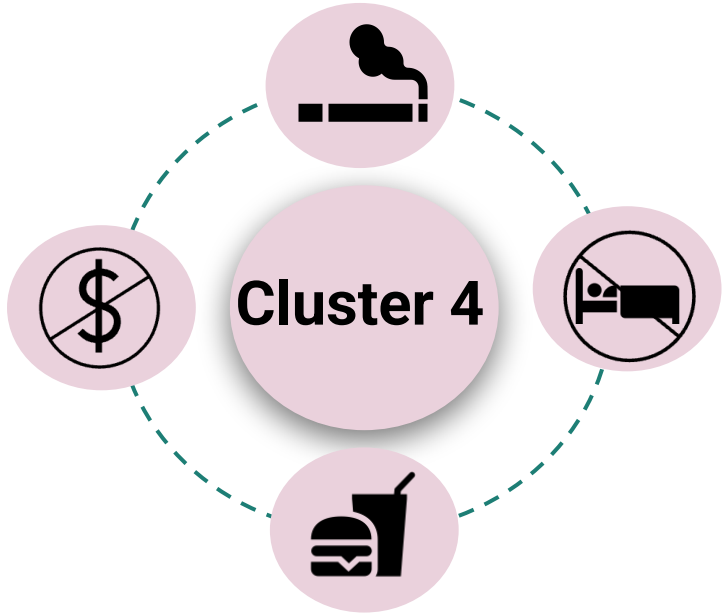


- **Exposures:** Higher pollution and exercise
- **Pathway analysis:** Immune system
- **Disease:** No significant associations with diseases

Expotype Signatures

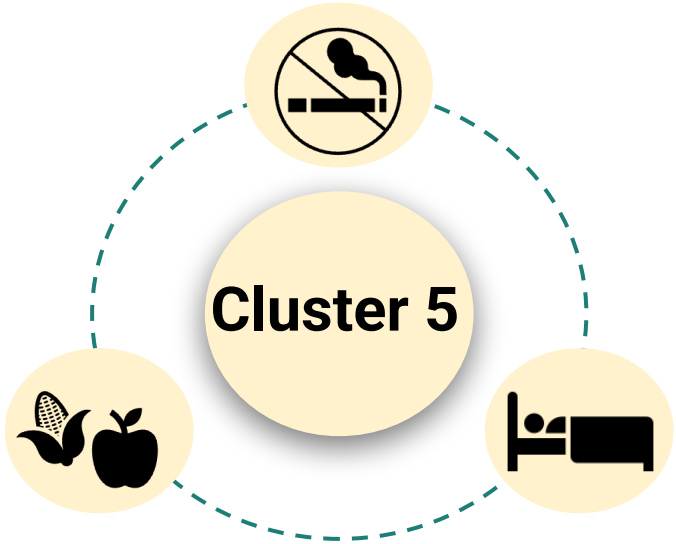


- **Exposures:** High neuroticism and anxiety
- **Pathway analysis:** Immune system
- **Disease:** Increased odds of PD, CAD, and Diabetes

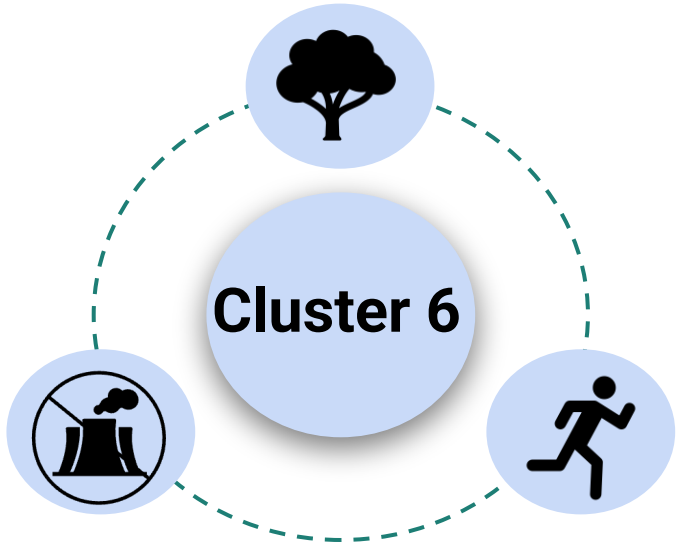


- **Exposures:** Current or previous smokers
- **Pathway Analysis:** Nervous system
- **Disease:** Increased odds of CKD, CAD, and Diabetes

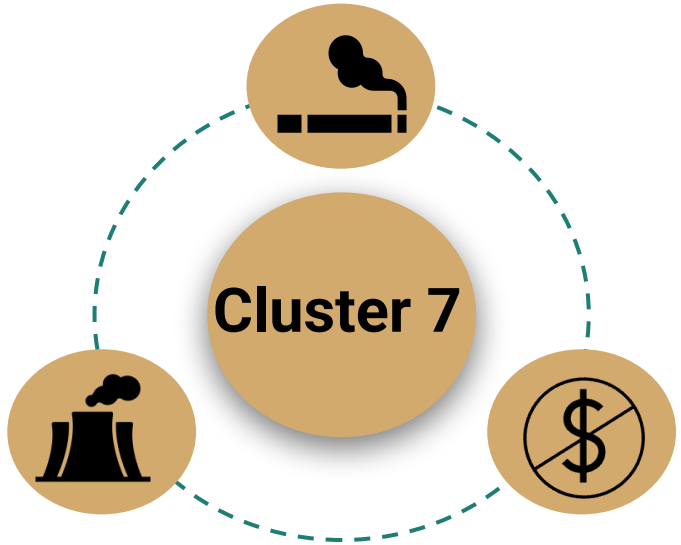
Expotype Signatures



- **Exposures:** Healthier exposure profile
- **Pathway Analysis:** Respiratory & Immune system
- **Disease:** Lower odds of CKD, CAD, & Diabetes



- **Exposures:** Healthier exposure profile
- **Pathway Analysis:** Cancer & Immune system
- **Disease:** Lower odds of CKD and Diabetes



- **Exposures:** High pollution and smoking
- **Pathway Analysis:** Cancer system
- **Disease:** No significant associations with diseases

Discussion and Implications

Disease-linked expotypes suggest policy campaigns to reduce smoking and improve mental health

The association with pollution and disease risk was lower than expected

Expotypes were mainly linked to the immune and nervous systems explaining biological effects

Limitations and Further Research

UK Biobank demographics are not representative of general population

Unbalanced cases and controls for the diseases

Supervised clustering with disease status as the outcome

IMPERIAL

Thank you

Expotype Clustering (Group 6)
07/03/2025

Contributions

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

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

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

Tianshu: Clustering (Fuzzy), Disease Extraction, Pathway Analysis

Appendix

References

1. Guillien A, et al. (2022) Exposome Profiles and Asthma among French Adults. Am J Respir Crit Care Med. Nov 15;206(10):1208-1219. doi: 10.1164/rccm.202205-0865OC.
   Erratum in: Am J Respir Crit Care Med. 2023 Dec 15;208(12):1347. Doi: 10.1164/rccm.v208erratum4. PMID: 35816632.
2. Song, J. et al. (2022) Using an Exposome-Wide Approach to Explore the Impact of Urban Environments on Blood Pressure among Adults in Beijing–Tianjin–Hebei and Surrounding Areas of China, Environmental Science & Technology 56 (12), 8395-8405 DOI: 10.1021/acs.est.1c08327
3. C. Bouveyron, S. Girard, C. Schmid, (2007) High-dimensional data clustering, Computational Statistics & Data Analysis, 52 (1), 502-519, https://doi.org/10.1016/j.csda.2007.02.009.
4. Herceg, Z. et al. (2018) Roadmap for investigating epigenome deregulation and environmental origins of cancer. Int. J. Cancer, 142: 874-882. https://doi.org/10.1002/ijc.31014

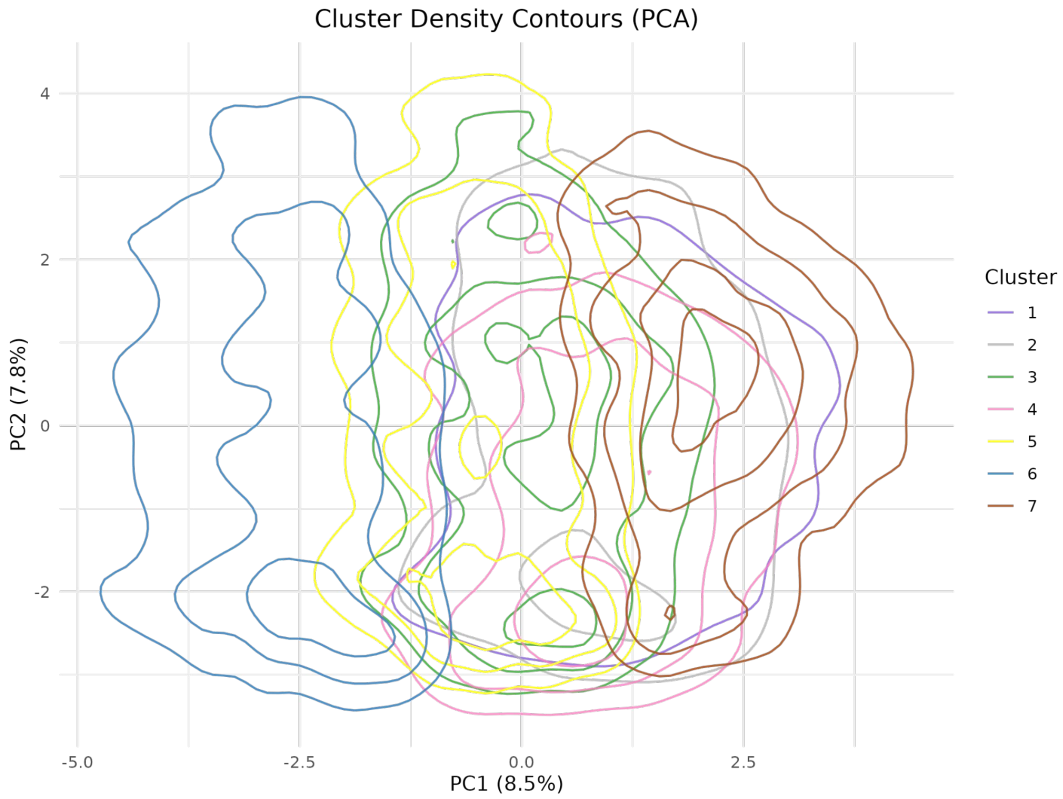
Table 1: Demographic & Socio Economic Variables

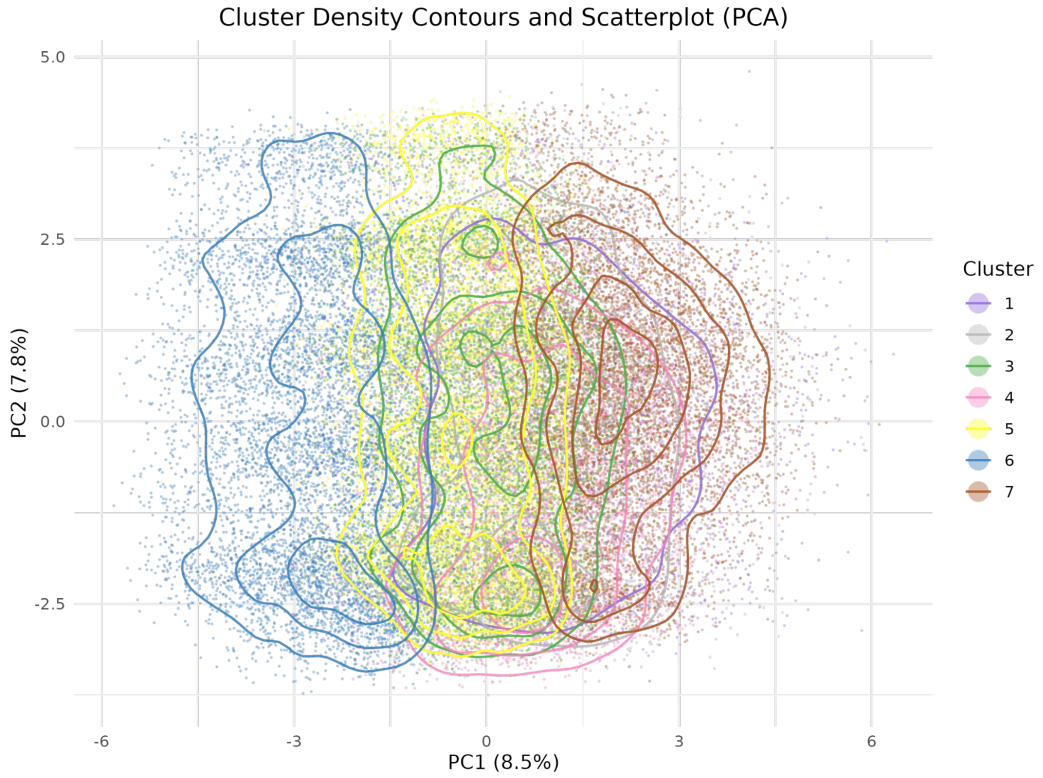
	1 (N=3193)	2 (N=2248)	3 (N=7062)	4 (N=4767)	5 (N=9538)	6 (N=7870)	7 (N=7126)	Overall (N=41804)	P-value
Sex									
Female	1752 (54.9%)	1221 (54.3%)	4652 (65.9%)	1998 (41.9%)	5080 (53.3%)	4318 (54.9%)	4044 (56.8%)	23065 (55.2%)	<0.001
Male	1441 (45.1%)	1027 (45.7%)	2410 (34.1%)	2769 (58.1%)	4458 (46.7%)	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%)	
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%)	
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%)	
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%)	
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%)	

Table 1: Environmental & Behavioural Variables

	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)	
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)	
Distance to Major Road									
Mean (SD)	221.04 (400.76)	731.54 (643.62)	564.87 (463.2)	533.99 (550.86)	552.18 (481.88)	1063.31 (1436.84)	314.97 (262.13)	592.39 (791.21)	<0.001
Median (Min, Max)	100 (33.3, 9090.9)	543.5 (100, 5882.4)	432.9 (100, 7692.3)	395.3 (50, 14285.7)	429.2 (100, 12500)	729.9 (50, 50000)	226.2 (50, 2222.2)	387.6 (33.3, 50000)	
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)	
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 (13.34)	8.35 (12.82)	3.74 (6.66)	34.73 (10.78)	3.5 (6.62)	6.39 (10.6)	5.79 (8.99)	8.72 (13.35)	<0.001
Median (Min, Max)	0 (0, 59)	0 (0, 58.5)	0 (0, 33.6)	34 (0, 59.6)	0 (0, 39.4)	0 (0, 59.5)	0 (0, 47)	0 (0, 59.6)	
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%)	
Neuroticism Score									
Mean (SD)	4.19 (3.32)	4.19 (3.2)	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 (0, 12)	4 (0, 12)	8 (0, 12)	4 (0, 12)	2 (0, 10)	3 (0, 12)	4 (0, 12)	4 (0, 12)	
Met Score									
Mean (SD)	2566.19 (2505.52)	2309.69 (2292.54)	1845.36 (1908.72)	2201.23 (2296.72)	2135.14 (1923.5)	2393.52 (2225.29)	2115.44 (2014.04)	2181.32 (2118.32)	<0.001
Median (Min, Max)	1687.5 (0, 10105)	1512 (0, 10026)	1207.5 (0, 10080)	1390 (0, 10080)	1569.8 (0, 10080)	1705 (0, 10080)	1493 (0, 10080)	1505 (0, 10105)	

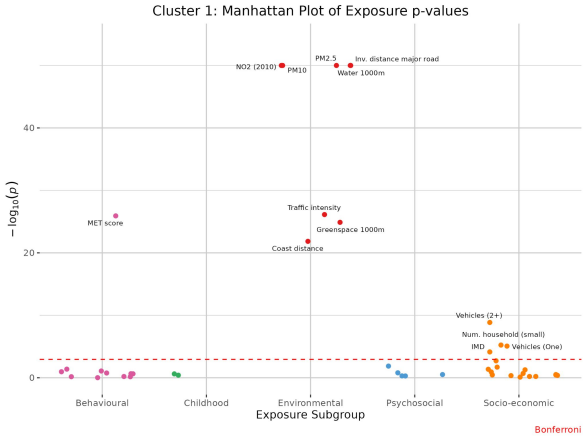
Cluster Overview
Gaussian Mixture Model

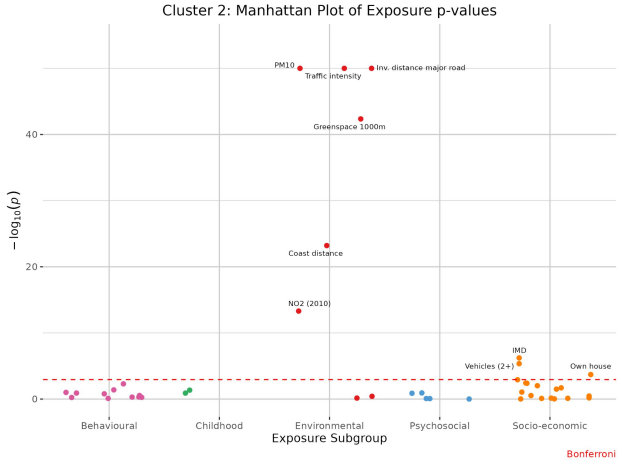


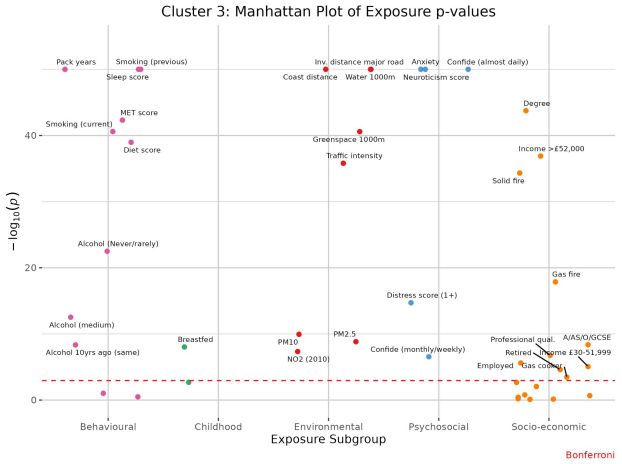


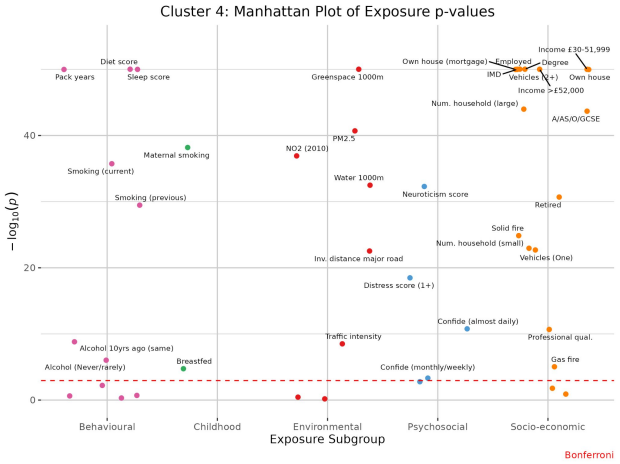
Univariate Analysis: Exposures

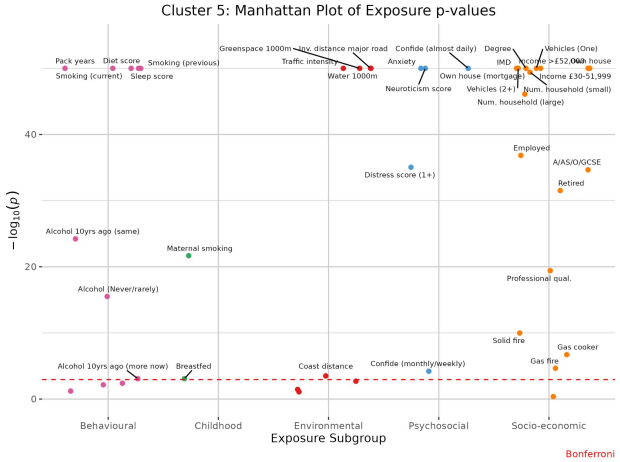
Manhattan Plots

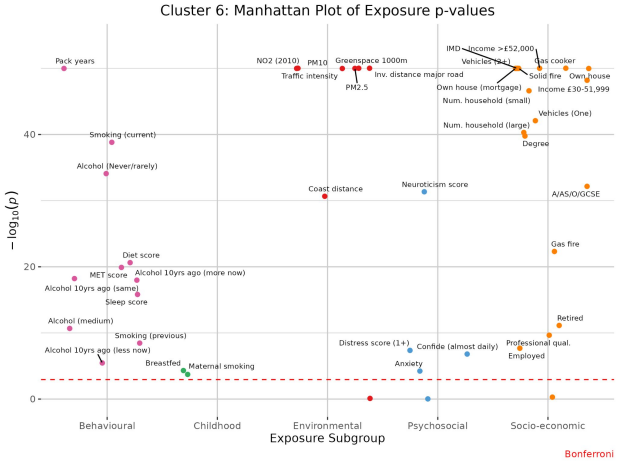


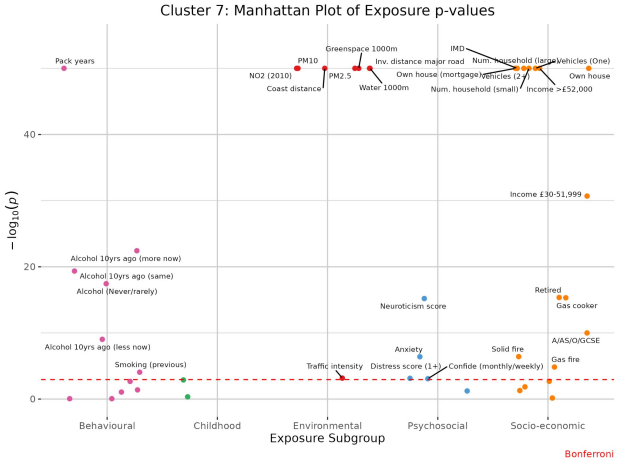






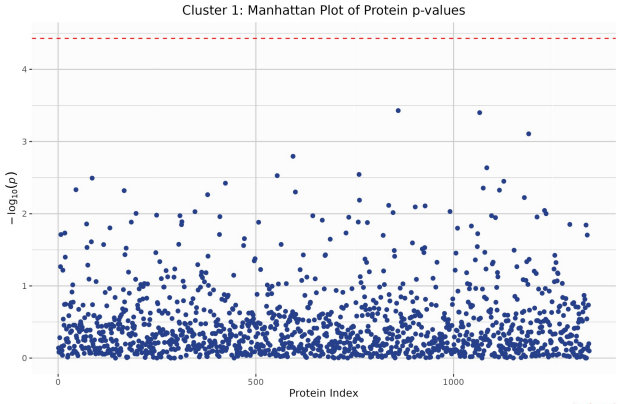


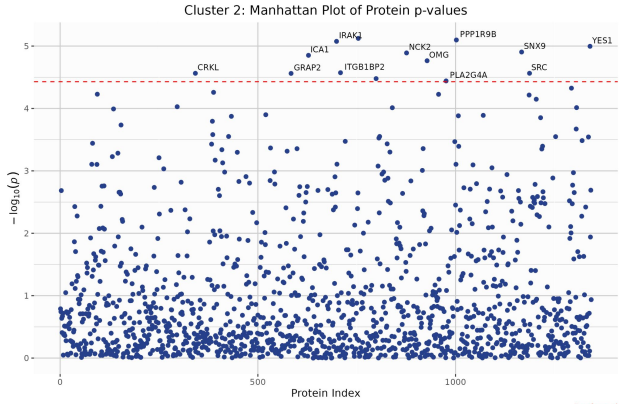


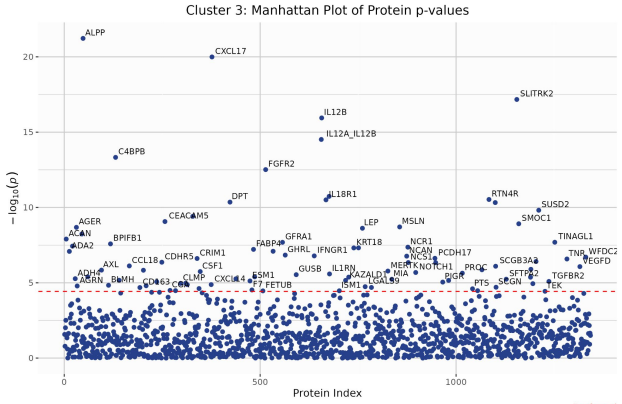


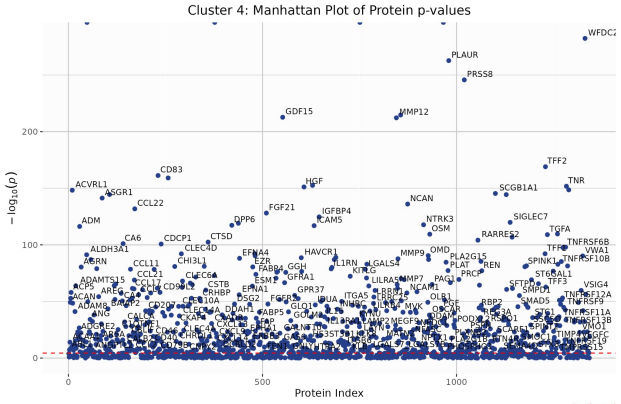
Univariate Analysis: Proteins

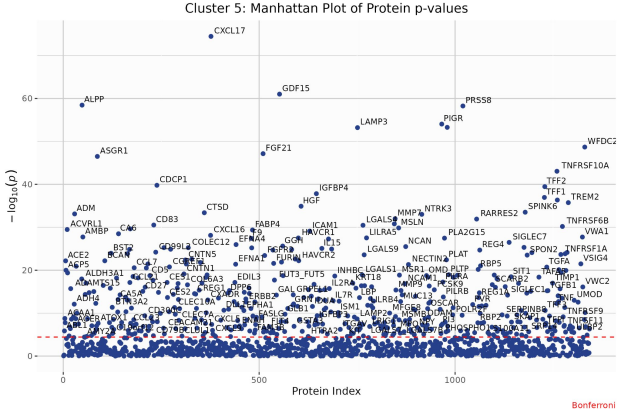
Manhattan Plots

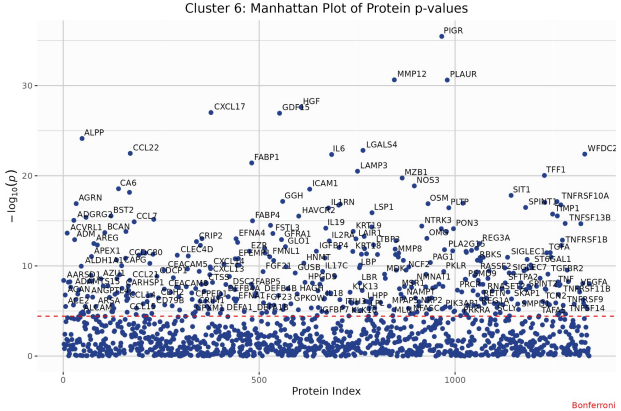


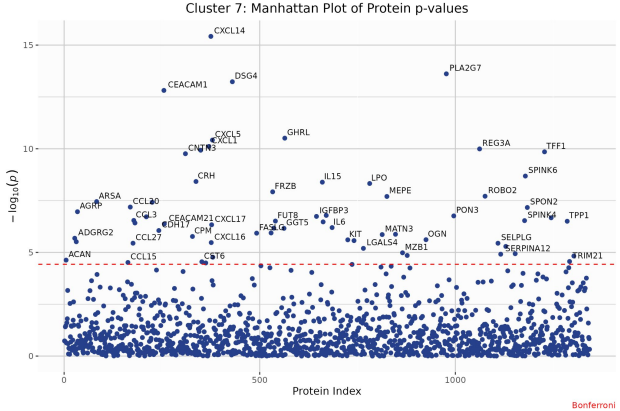




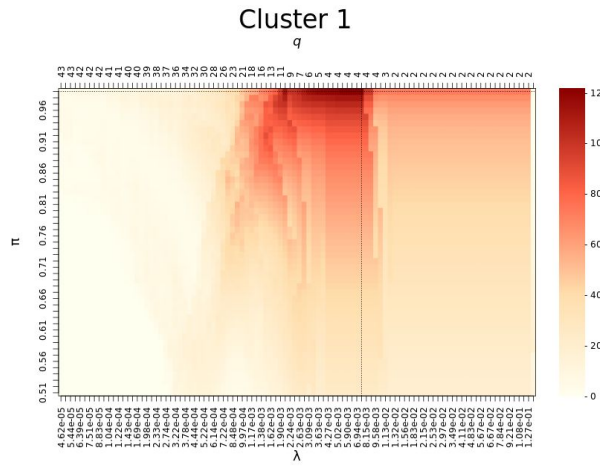


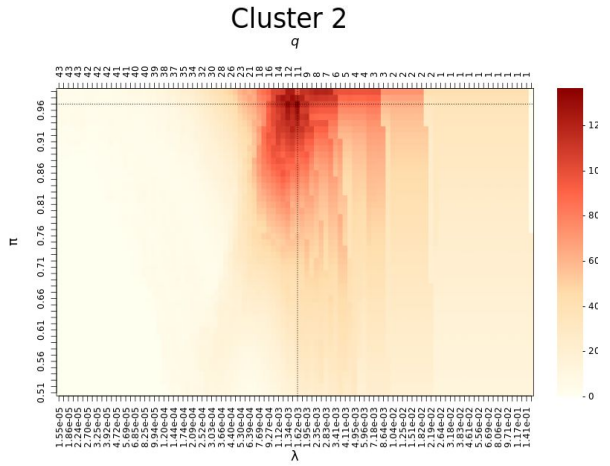


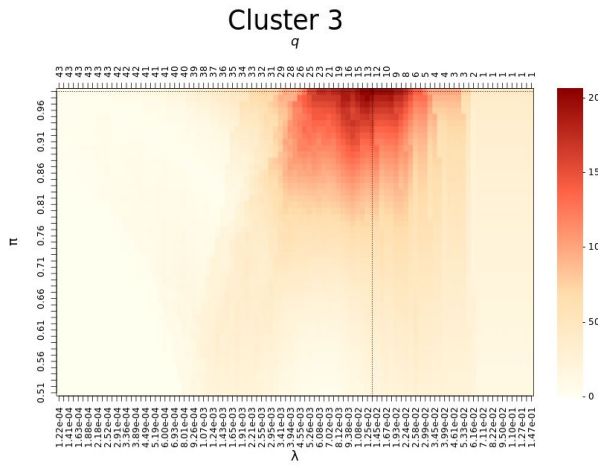


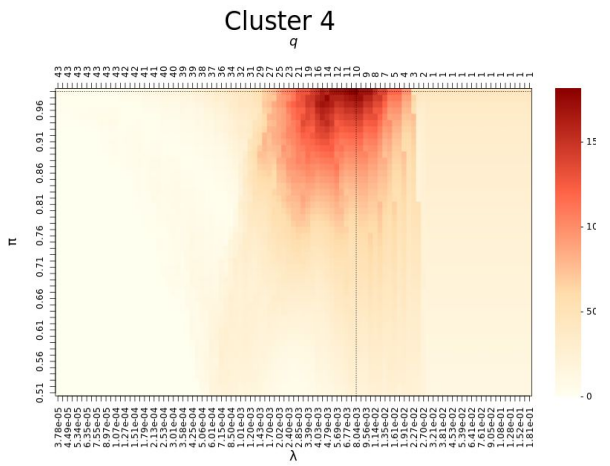


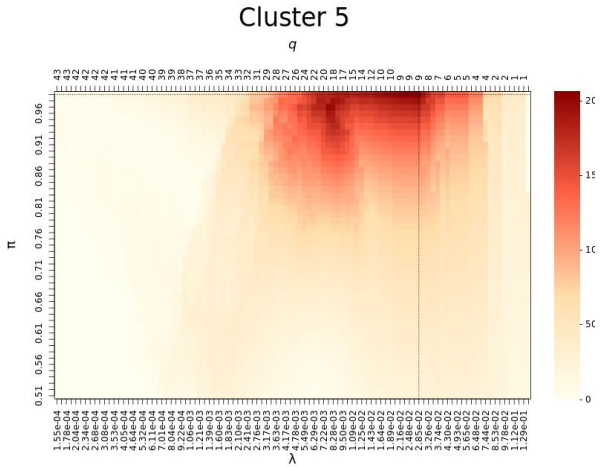
Stability LASSO Calibration Plots: Exposures

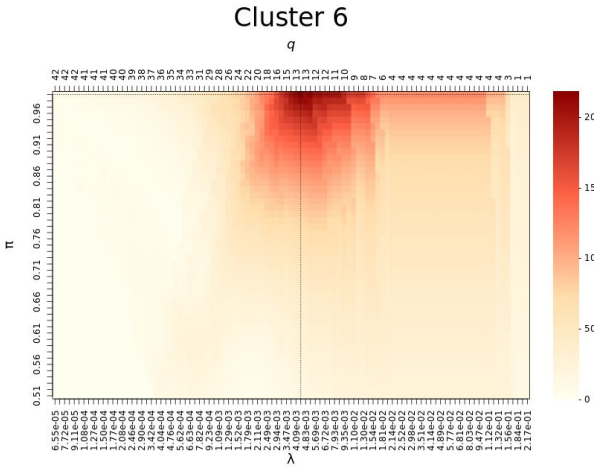


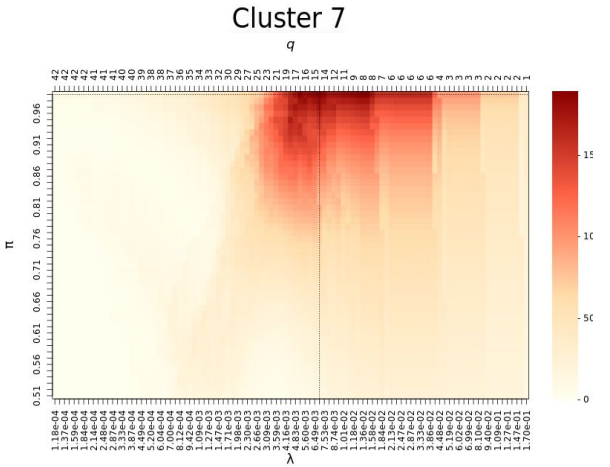




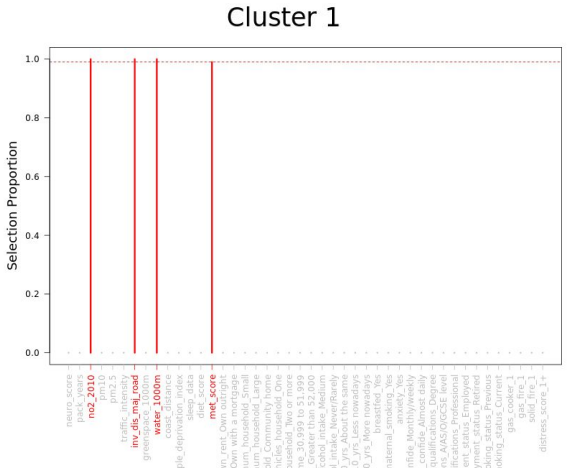


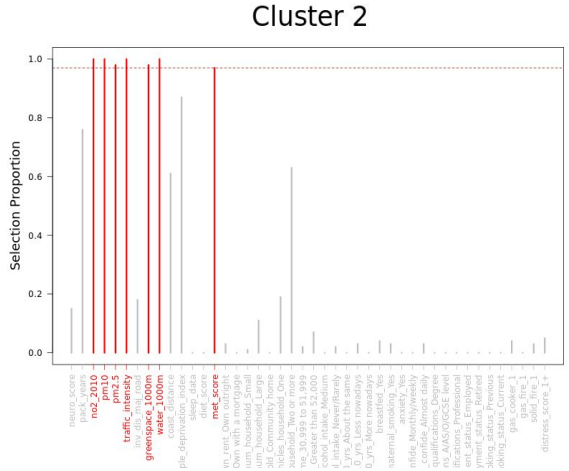


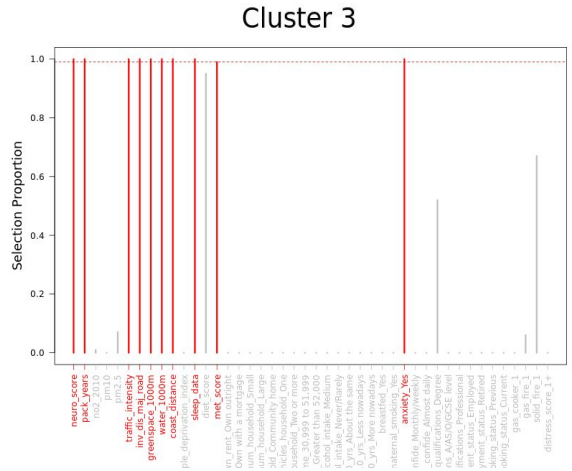


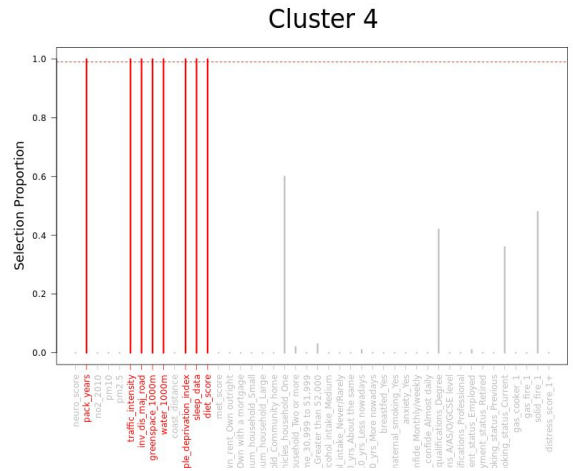


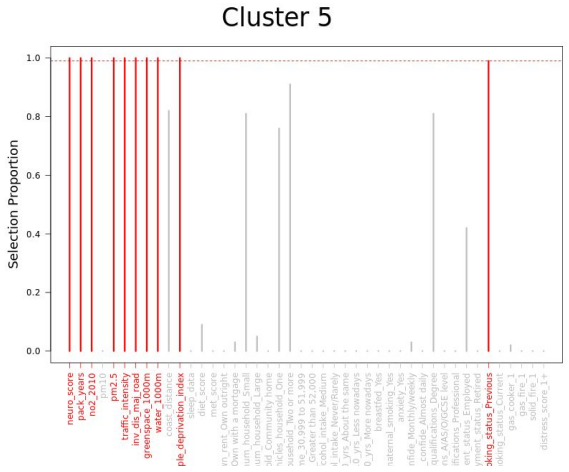
Stability LASSO Selection Proportion Plots: Exposures

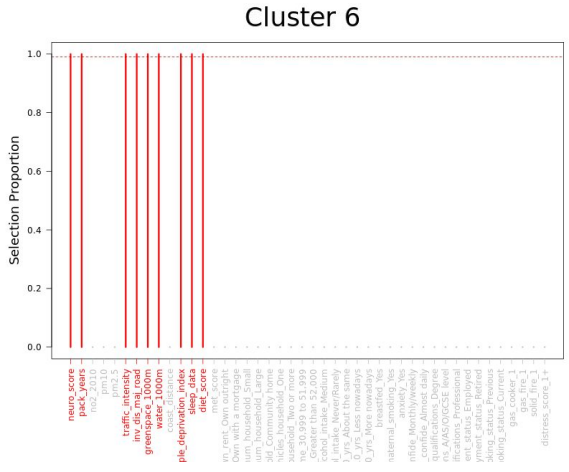


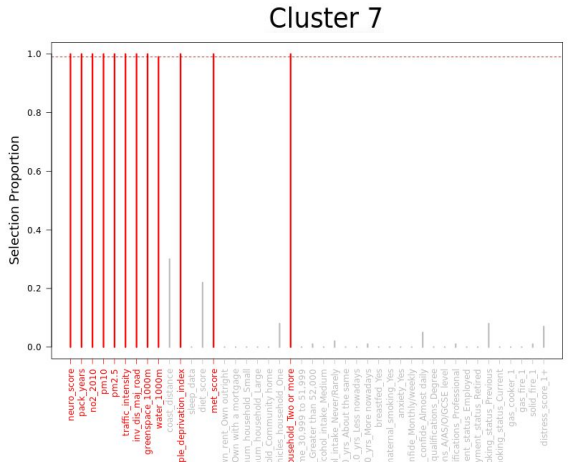




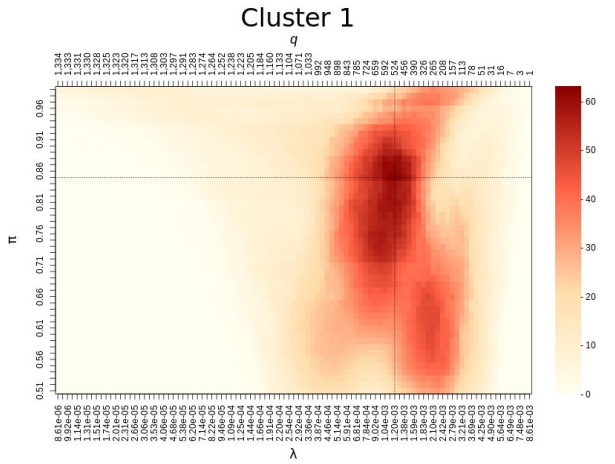


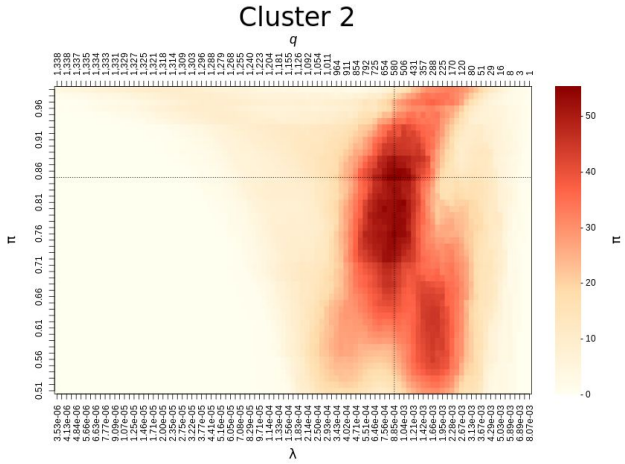


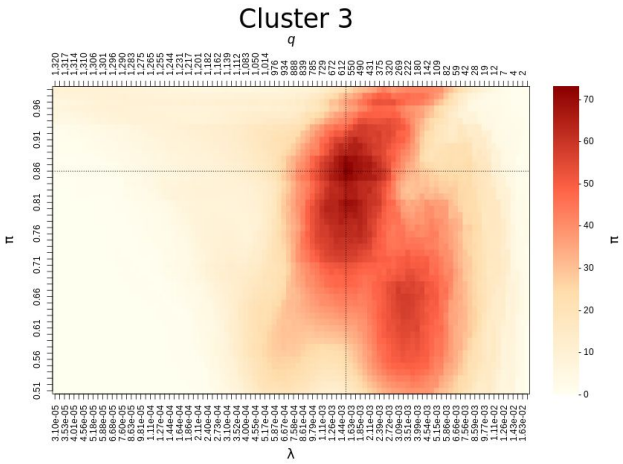


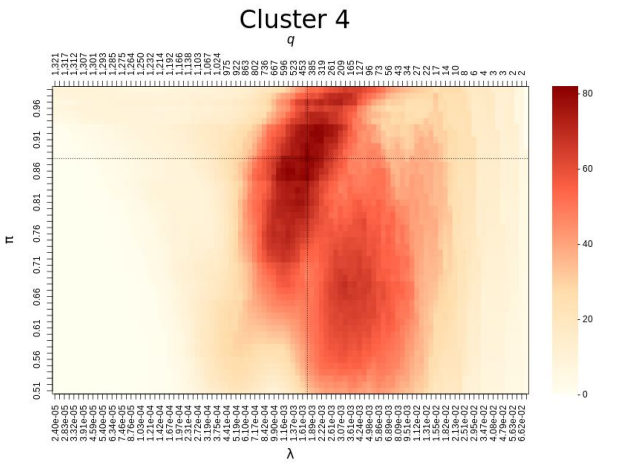


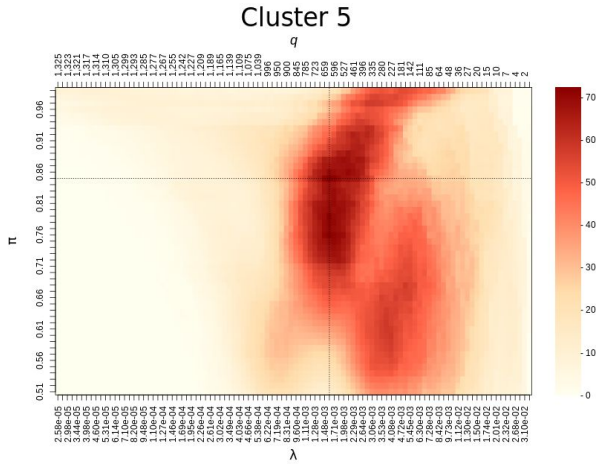
Stability LASSO Calibration Plots: Proteins

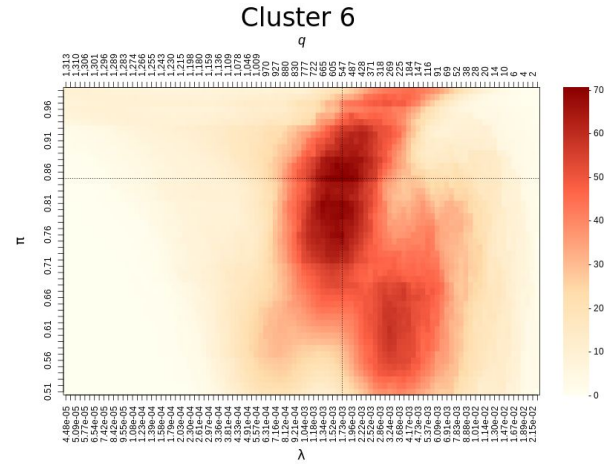


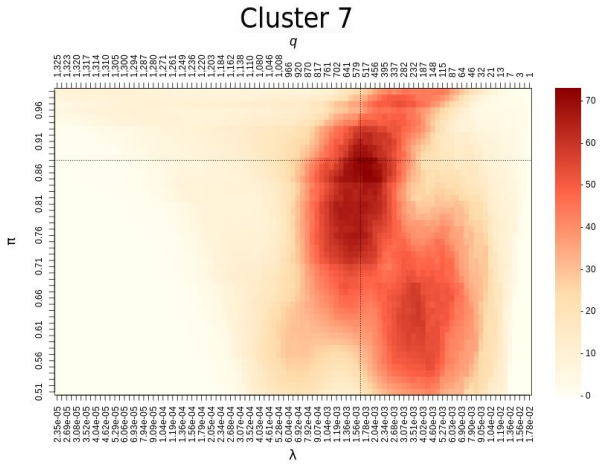




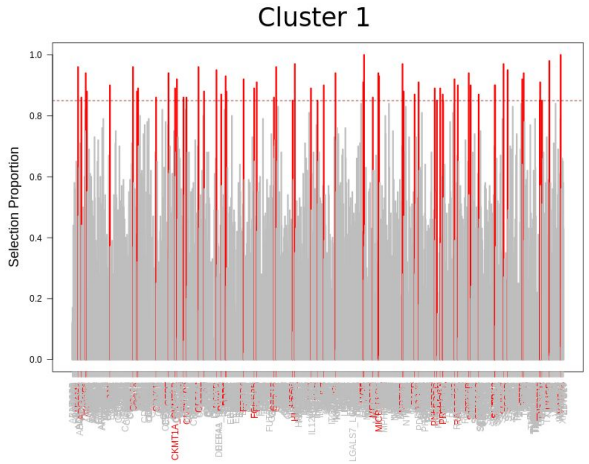


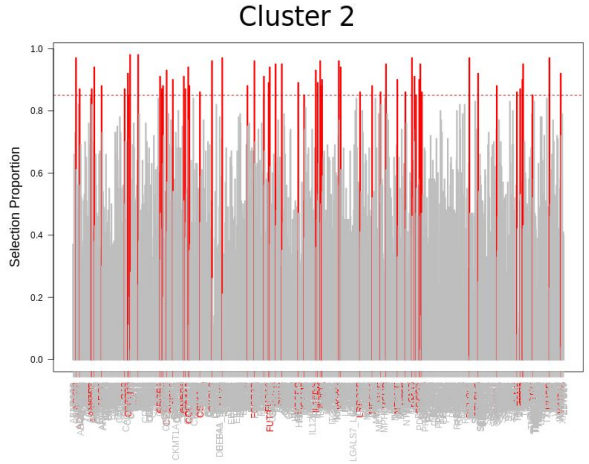


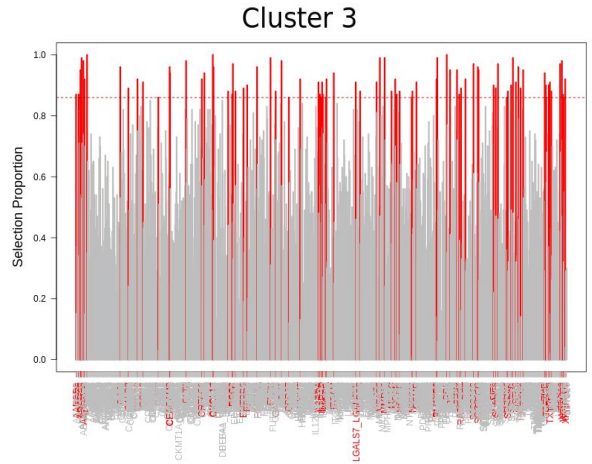


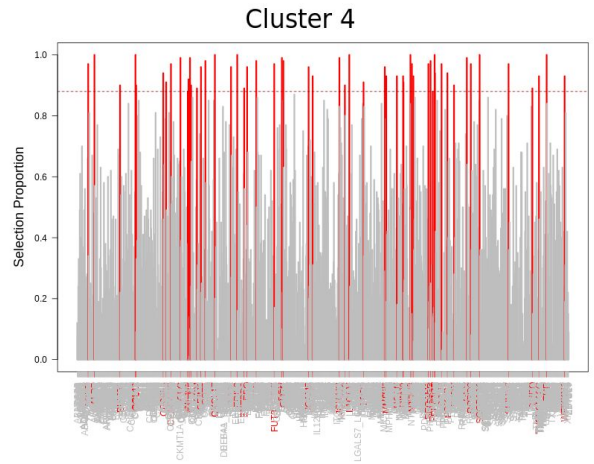


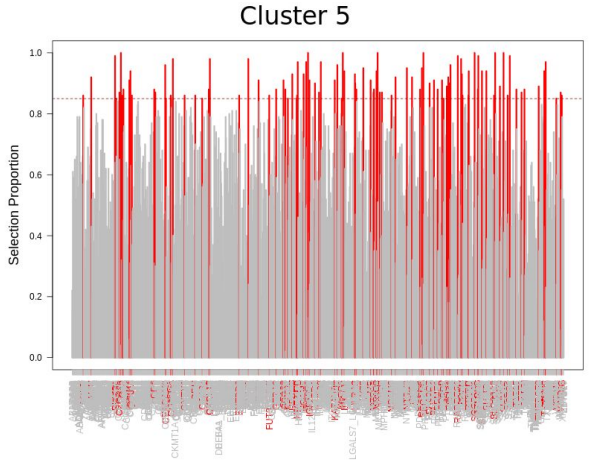
Stability LASSO Selection Proportion Plots: Proteins

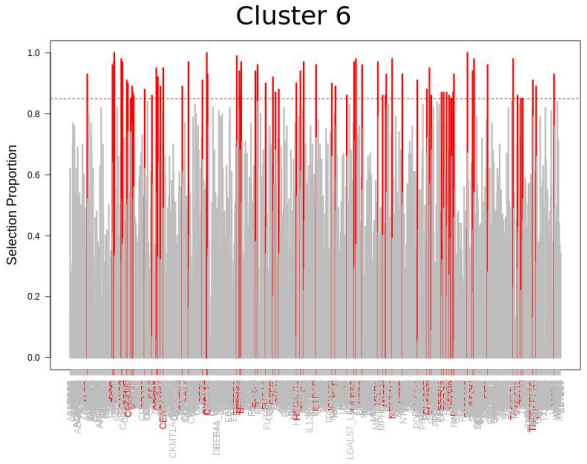


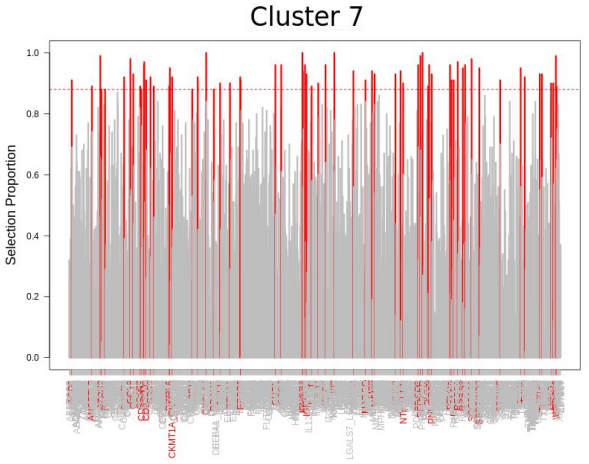






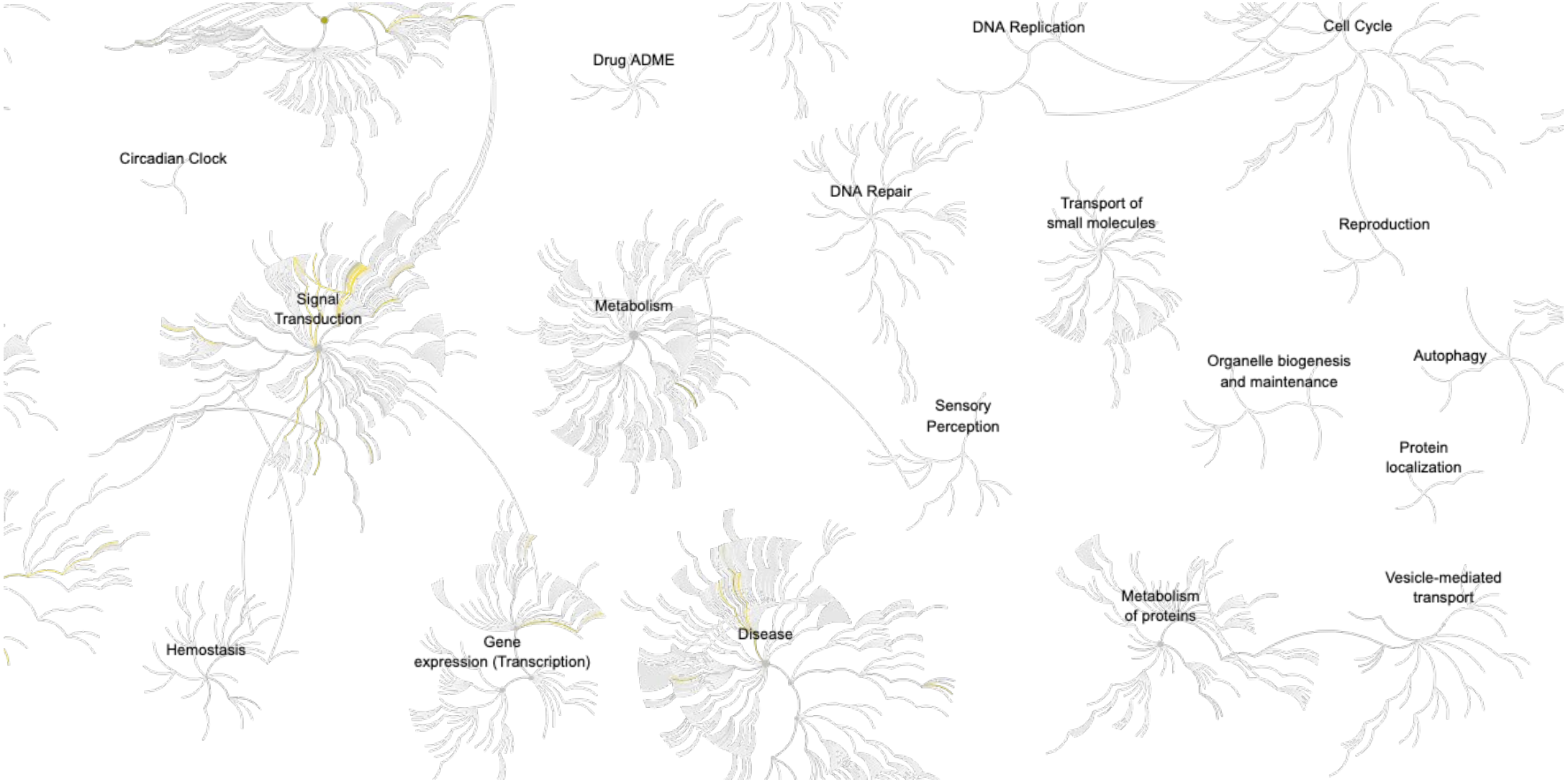






Pathway Analysis: Cluster 1 Example

reactome.com



Enrichment Analysis of Clusters

Entities FDR: adjusted p-value controlling false discoveries in pathway

	Pathway Name	No. of significant pathways	Entities FDR	Biology Role	Possible Implications on Health Outcome
Cluster 1	Downregulation of ERBB4 signaling Interleukin-10 signaling Nuclear signaling by ERBB4	26	0.003313 0.003313 0.003313	Nervous system, regulates cell differentiation and survival; immune system, suppresses inflammation; developmental system, modulates gene expression.	
Cluster 2	Interleukin-10 signaling Signaling by Interleukins Immune System	3	4.33E-05 9.93E-04 0.053569	Immune system, suppresses inflammation; immune system, mediates cytokine communication; immune system, coordinates defense and immune responses.	Pro-inflammatory pathways align with high pollution exposure, possibly increasing disease susceptibility.
Cluster 3	Cytokine Signaling in Immune system Signaling by Interleukins Downregulation of ERBB4 signaling	28	0.002407 0.005014 0.005014	Immune system, transmits cytokine signals; immune system, mediates interleukin responses; nervous system, reduces cell differentiation and survival.	Stress-related signalings may explain links to high anxiety, supporting the psychological vulnerability.
Cluster 4	Synthesis, secretion, and deacylation of Ghrelin NTF3 activates NTRK3 signaling Post-translational modification: synthesis of GPI-anchored proteins	7	0.001338 0.011738 0.011738	Endocrine system, regulates appetite hormone processing; nervous system, supports neuron survival and development; cellular system, modifies proteins for membrane anchoring.	Neuronal and metabolic signaling match smoking exposure, aligning with high risk of PD, CKD, CAD.
Cluster 5	TFAP2 (AP-2) family regulates transcription of growth factors and their receptors Reversible hydration of carbon dioxide Interleukin-33 signaling	10	0.006058 0.032018 0.032018	Developmental system, controls growth factor gene expression; respiratory system, maintains CO_2 balance; immune system, triggers inflammatory response.	Protective signaling pathways may enhance resilience, resulting a healthier lifestyle and lower disease odds.
Cluster 6	PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling Constitutive Signaling by Aberrant PI3K in Cancer Negative regulation of the PI3K/AKT network	86	3.37E-05 3.37E-05 3.37E-05	Cellular system, modulates survival signaling; cancer system, drives uncontrolled cell growth; immune system, inhibits PI3K/AKT pathway activation.	Anti-survival signaling and healthy environment jointly reduce CKD risk.
Cluster 7	Constitutive Signaling by Aberrant PI3K in Cancer ERBB2 Activates PTK6 Signaling ERBB2 Regulates Cell Motility	59	3.77E-04 3.77E-04 3.77E-04	Cancer system, promotes uncontrolled proliferation; epithelial system, activates intracellular kinase signaling; cellular system, enhances cell movement and invasion.	Pro-growth signaling may offset greenspace benefits, potentially elevating PD and CAD risk.