

Host-transcriptome based classification of ocular pathogens: a proof-of-concept using machine learning analysis

Maria Taher Plumber¹, Ratnesh Kumar¹, Tejabhiram Yadavalli^{2,3*}

¹ *Department of Computer Science, University of Illinois Chicago, Chicago, IL, USA*

² *Department of Ophthalmology and Visual Sciences, University of Illinois Chicago, Chicago, IL, USA*

³ *Department of Biomedical Engineering, University of Illinois Chicago, Chicago, IL, USA*

** Corresponding author*

Supplementary Data

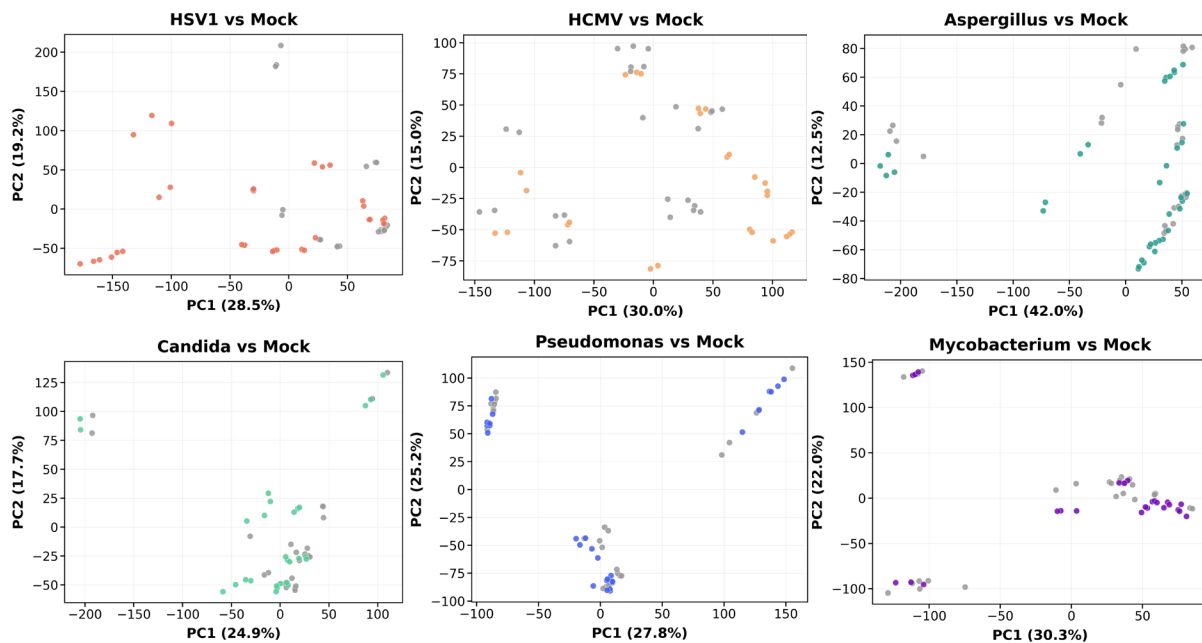
Important Notes for the manuscript

Data :

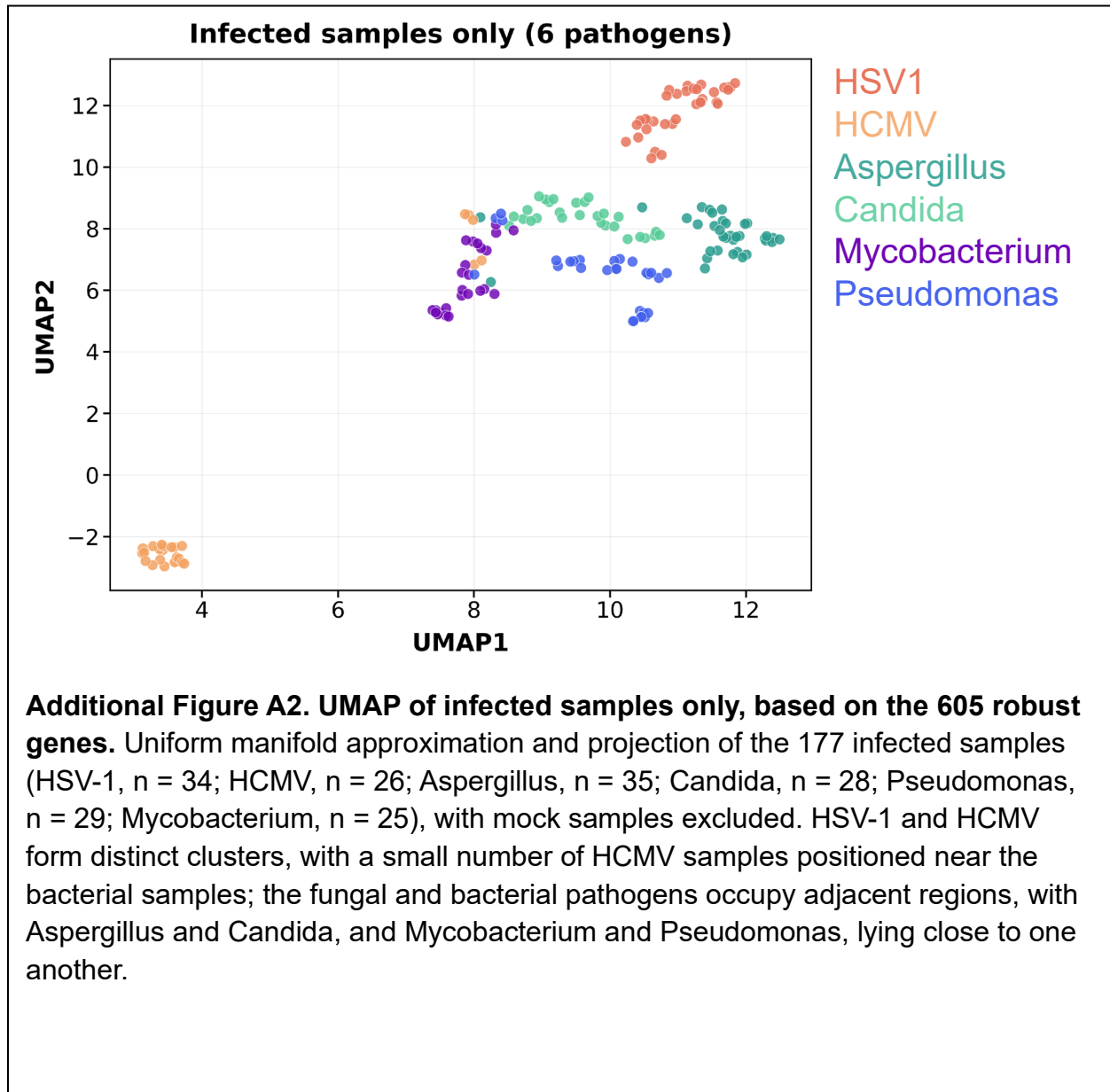
1. studydesign.xlsx contains all the metadata for the samples.
2. The bulk RNASeq data was fetched and downloaded from NCBI Sequence Read Archive with Organism: Homo Sapiens, Layout : Paired or Single, Source : Transcriptomics and Selection : cDNA
3. All the samples were matched with the Ensembl Human Transcriptome.
4. The data folders consist of excel files for each sample where all 330 samples are included in the “data” folder filename consisting of Pathogen_Condition_SampleID_Biopproject_CellType.xlsx in the respective pathogens folders: HHV1, HCMV5, Asper, Cand, Myc, Pseu.
5. All the transcripts were converted to genes using the summation of isoforms and mapped using tx2gene data.

Code :

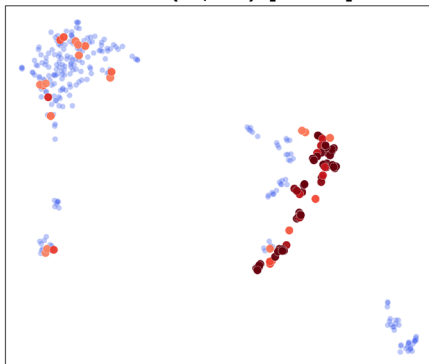
1. code1.py and code2.py are separate files with the entire work divided in 2 parts.
2. The codes are deliberately placed in the jupyter notebook cells with each cell showing the outputs in the notebook itself.
3. The main matrix of robust gene's log fold change values is combine_lfc_matrix.tsv attached in the folder.
4. The list of robust genes across each pathogen and 142 genes from grouped folds are attached as excel files.



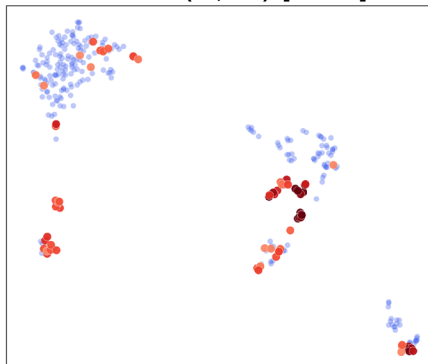
Additional Figure 1. Principal component analysis of infected versus mock samples for each pathogen prior to robust-gene selection. For each pathogen, PCA was performed on $\log_2(\text{TPM} + 1)$ expression of all genes passing the low-expression filter (CPM > 0.5 in at least 10% of that pathogen's samples), standardized per gene: HSV-1 (56 samples, 21447 genes; PC1 = 28.5%, PC2 = 19.2%), HCMV (52 samples, 18323 genes; PC1 = 30.0%, PC2 = 15.0%), Aspergillus (63 samples, 16574 genes; PC1 = 42.0%, PC2 = 12.5%), Candida (52 samples, 18506 genes; PC1 = 24.9%, PC2 = 17.7%), Pseudomonas (56 samples, 20702 genes; PC1 = 27.8%, PC2 = 25.2%), and Mycobacterium (51 samples, 17941 genes; PC1 = 30.3%, PC2 = 22.0%). Mock samples are shown in grey and infected samples in pathogen-specific colors. Samples form discrete groups in which infected and mock samples are intermixed, rather than separating by infection status, in contrast to the separation obtained after restriction to robust genes (Figure 2 a).



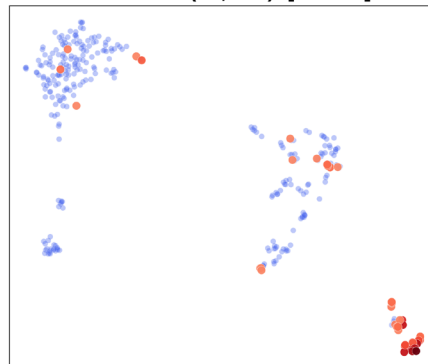
CCL3L3 (86/330) [5 folds]



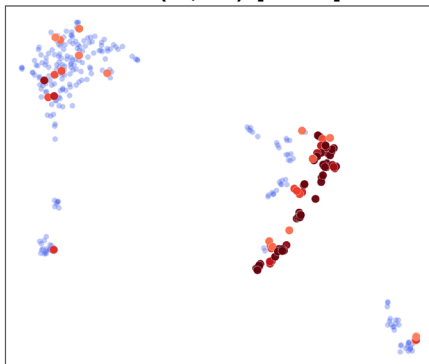
SERPINB7 (72/330) [5 folds]



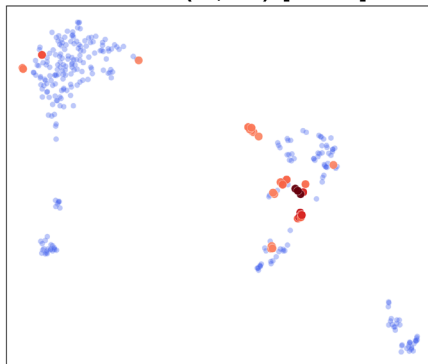
PAFAH1B1 (46/330) [5 folds]



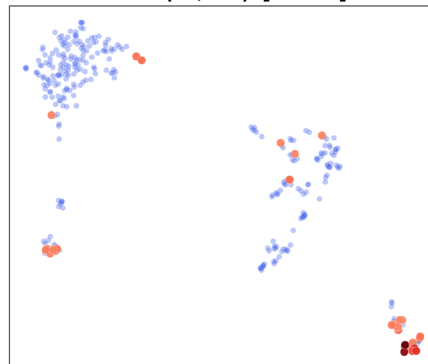
CCL4 (88/330) [5 folds]



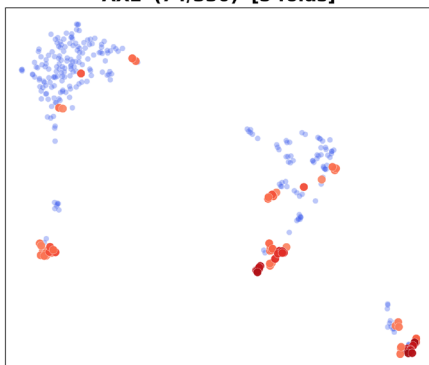
SERPINB4 (32/330) [5 folds]



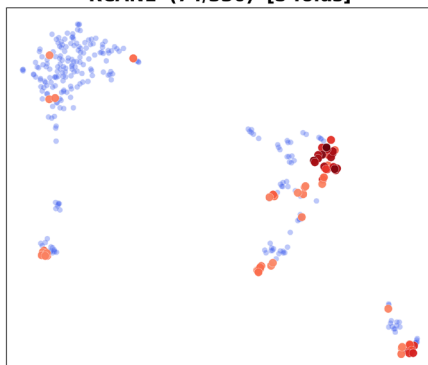
OTX1 (31/330) [5 folds]



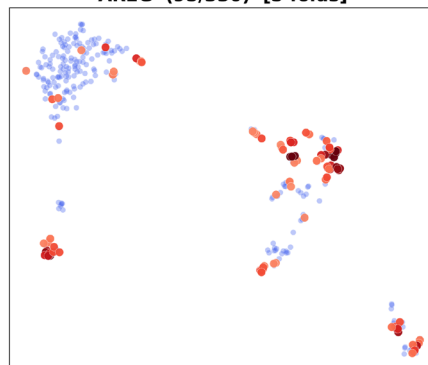
AXL (74/330) [5 folds]



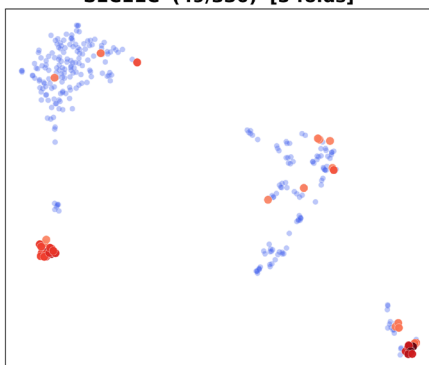
RCAN1 (74/330) [5 folds]



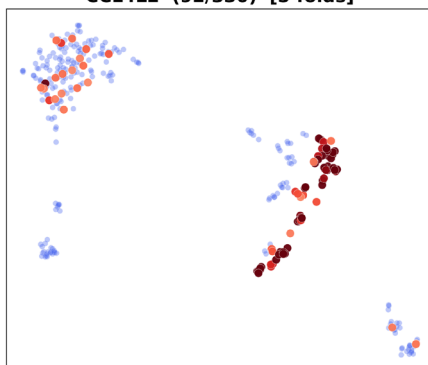
AREG (95/330) [5 folds]



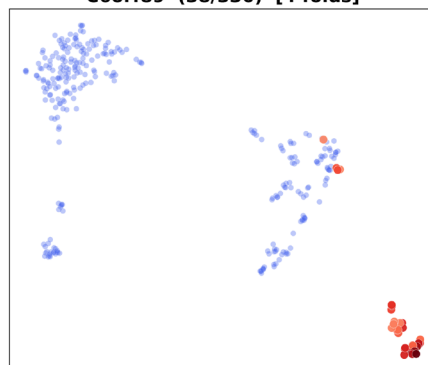
SEC11C (49/330) [5 folds]



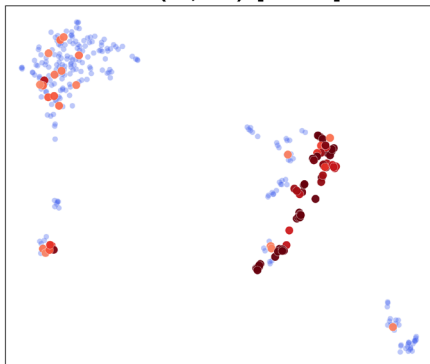
CCL4L2 (92/330) [5 folds]



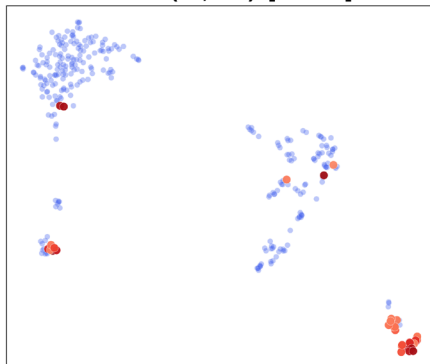
C6orf89 (38/330) [4 folds]



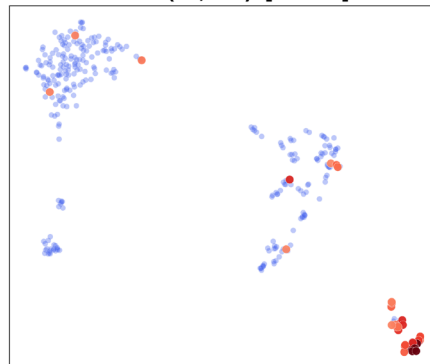
CCL3 (88/330) [4 folds]



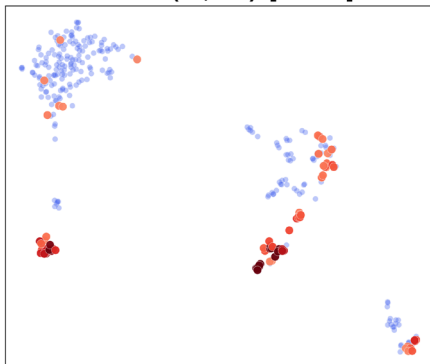
MAP1A (41/330) [4 folds]



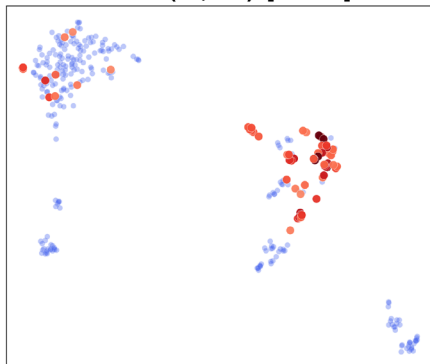
COMT (39/330) [4 folds]



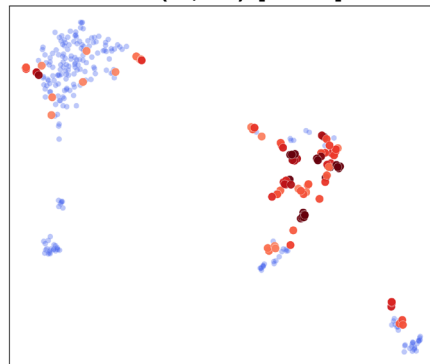
TRIM22 (79/330) [4 folds]



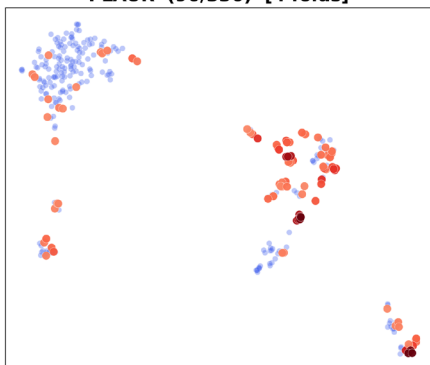
HCAR2 (67/330) [4 folds]



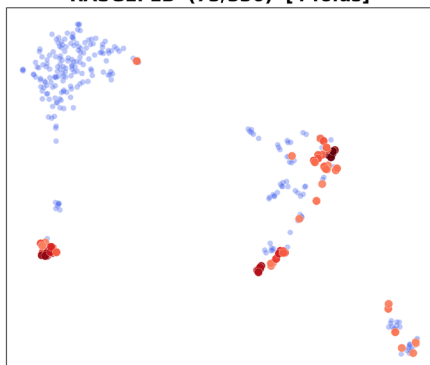
IL1A (96/330) [4 folds]



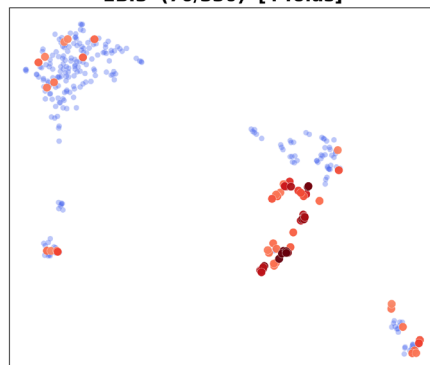
PLAUR (96/330) [4 folds]



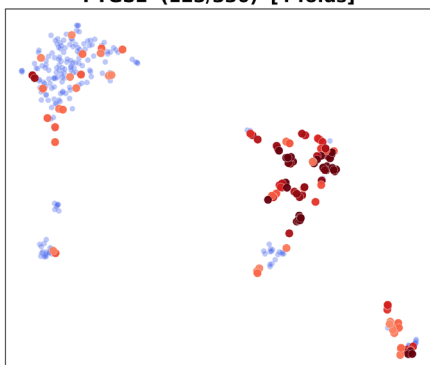
RASGEF1B (73/330) [4 folds]



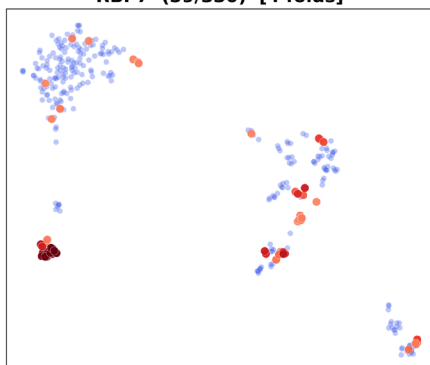
EBI3 (70/330) [4 folds]



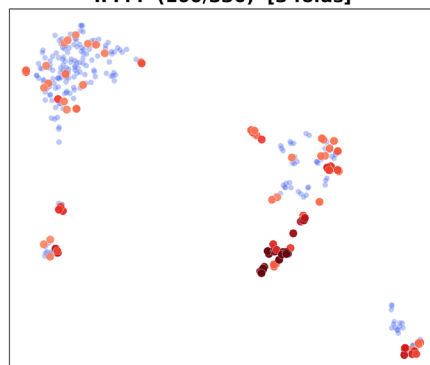
PTGS2 (123/330) [4 folds]



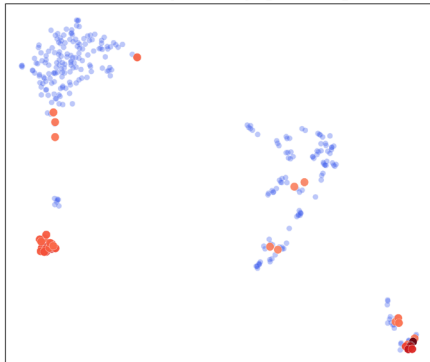
RBP7 (59/330) [4 folds]



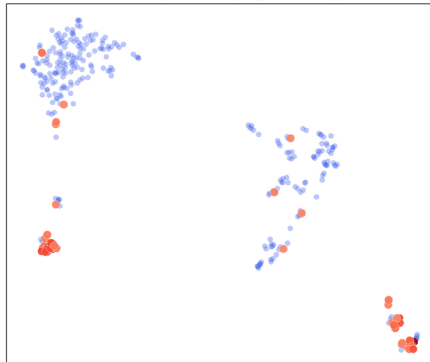
IFI44 (100/330) [3 folds]



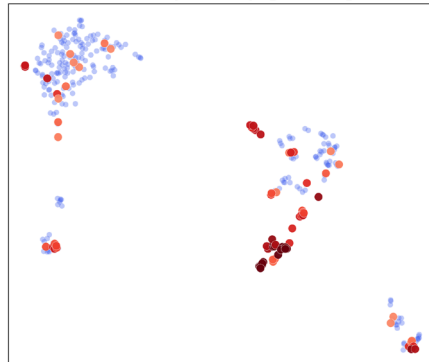
DSCC1 (41/330) [3 folds]



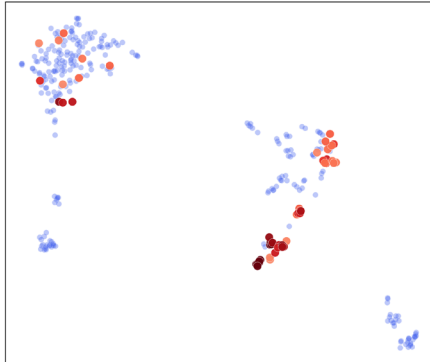
SVIP (50/330) [3 folds]



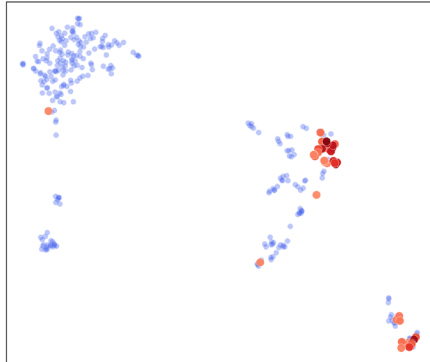
IFITM1 (81/330) [3 folds]



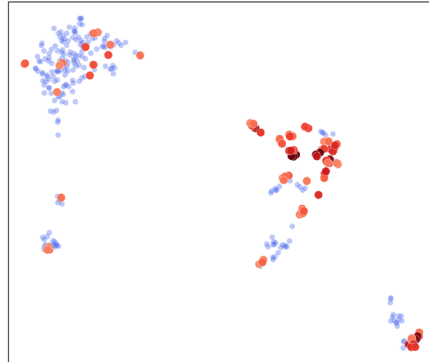
SIGLEC1 (54/330) [3 folds]



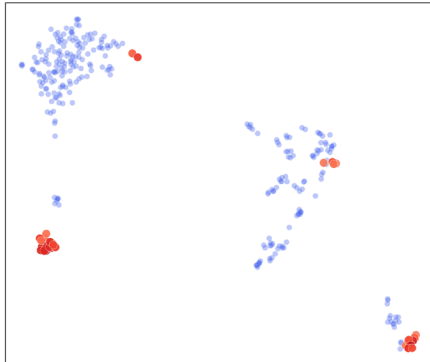
CYB5D1 (40/330) [3 folds]



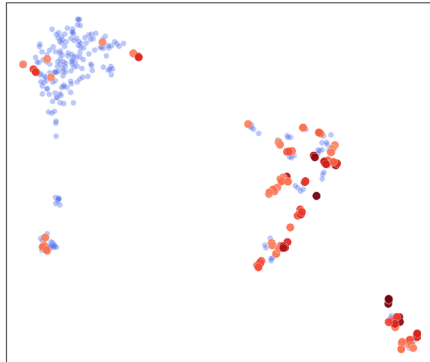
SLC2A3 (95/330) [3 folds]



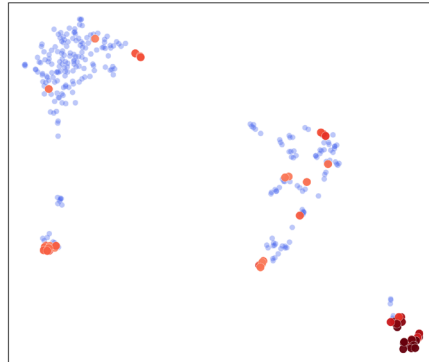
DOLPP1 (41/330) [3 folds]



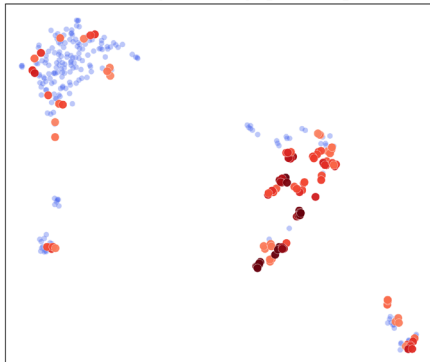
IRF1 (101/330) [3 folds]



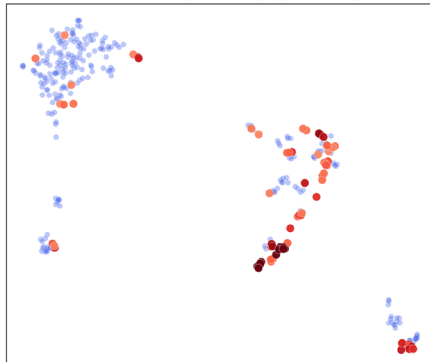
LSMEM1 (59/330) [3 folds]



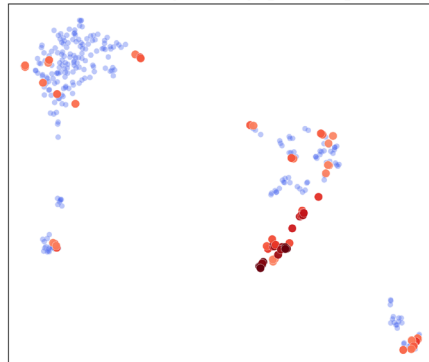
SOD2 (117/330) [3 folds]



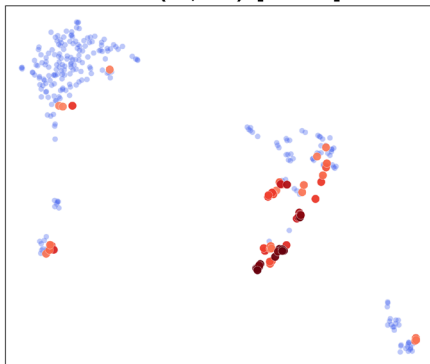
TNFSF10 (75/330) [3 folds]



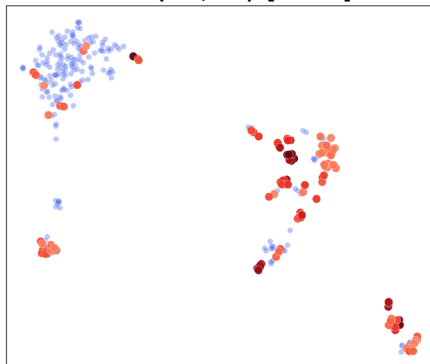
PARP9 (73/330) [3 folds]



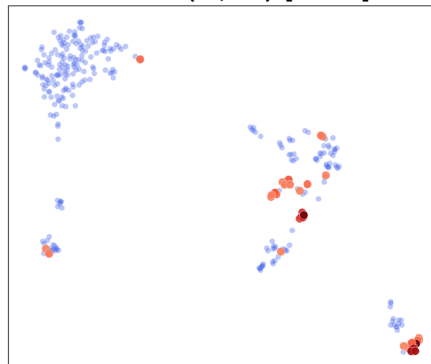
IL4I1 (62/330) [3 folds]



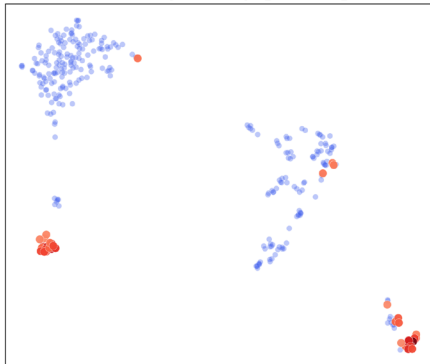
MAFF (125/330) [3 folds]



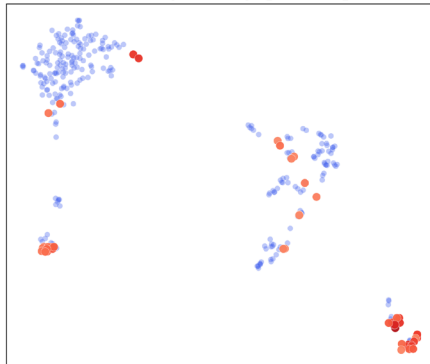
SLC25A37 (41/330) [3 folds]



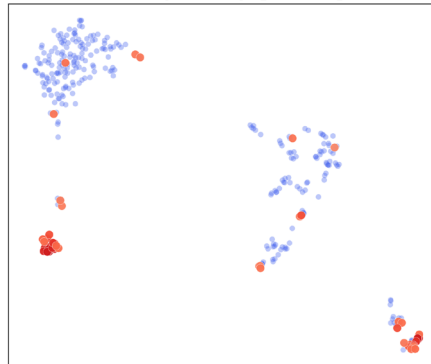
CERS6 (46/330) [3 folds]



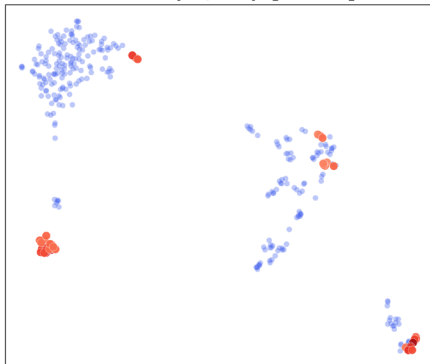
ZNF10 (47/330) [3 folds]



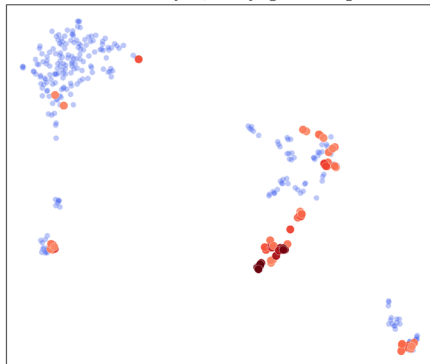
TAF5 (52/330) [3 folds]



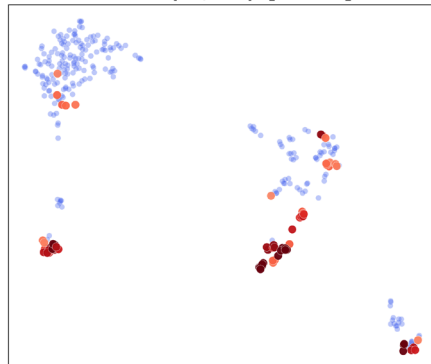
C9orf40 (44/330) [3 folds]



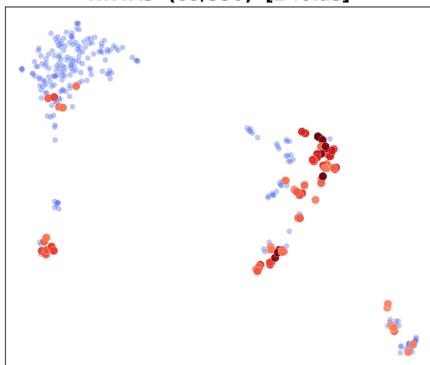
SAMD9 (67/330) [3 folds]



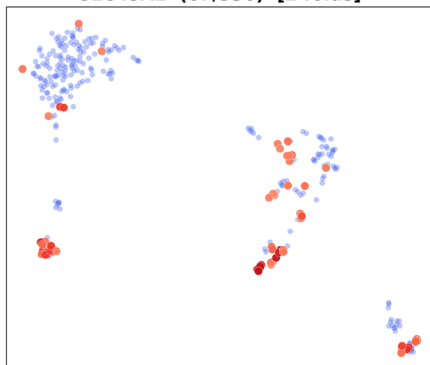
HERC5 (75/330) [2 folds]



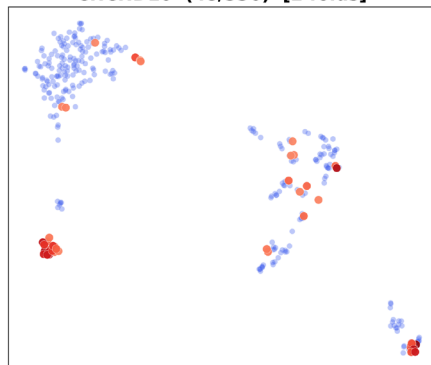
NR4A3 (85/330) [2 folds]



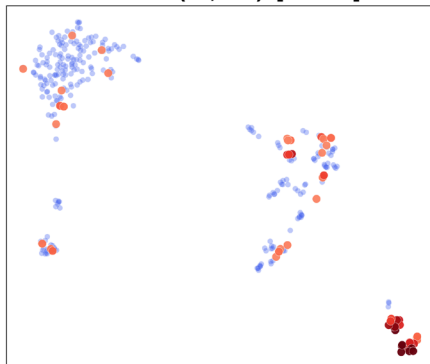
SLC43A2 (67/330) [2 folds]



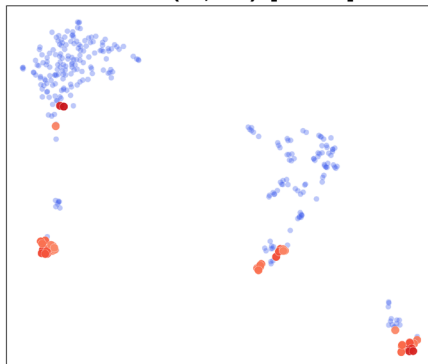
CHCHD10 (48/330) [2 folds]



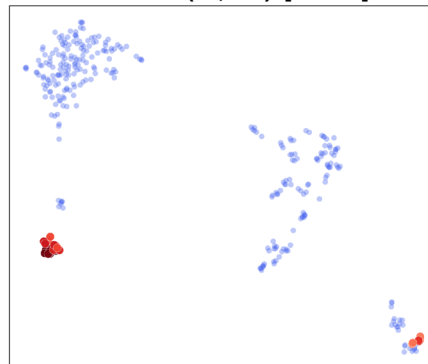
BCL2L11 (57/330) [2 folds]



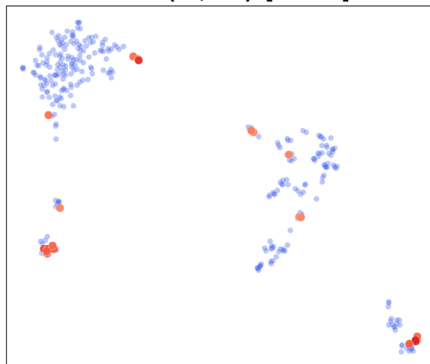
DIRAS1 (50/330) [2 folds]



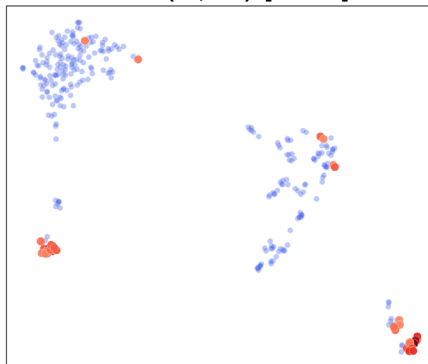
TMEM108 (28/330) [2 folds]



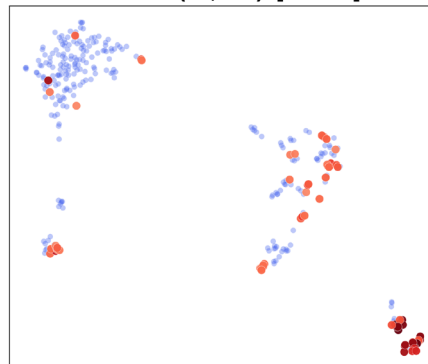
FGFR2 (25/330) [2 folds]



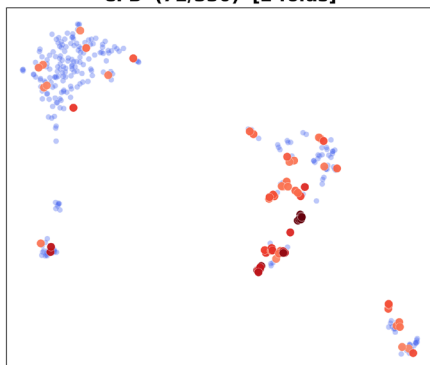
KLHL5 (43/330) [2 folds]



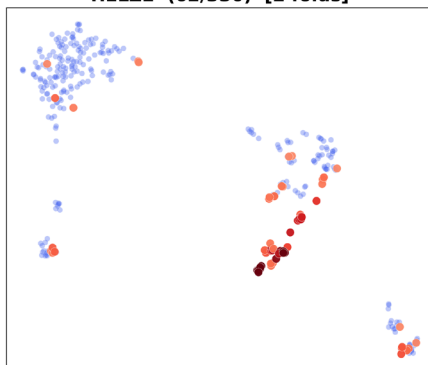
PYCARD (73/330) [2 folds]



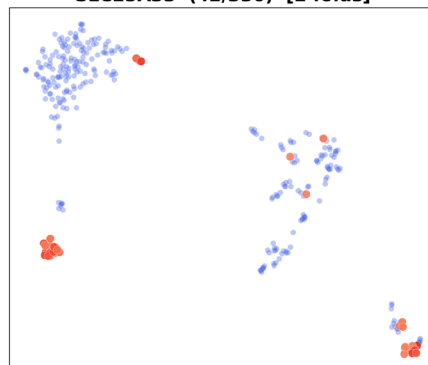
CFB (72/330) [2 folds]



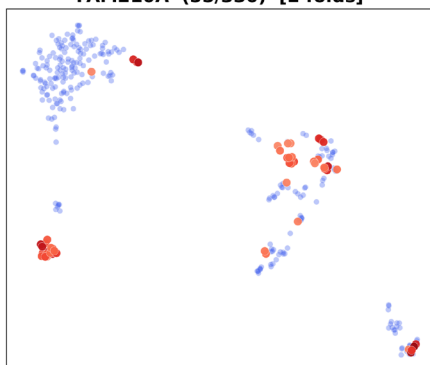
HELZ2 (62/330) [2 folds]



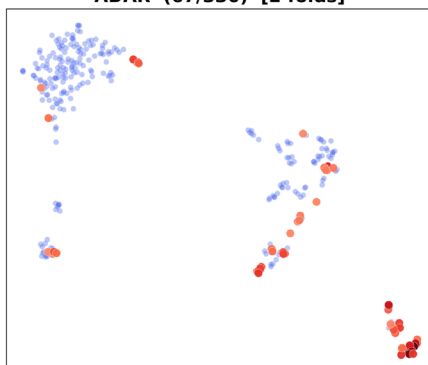
SLC25A33 (42/330) [2 folds]



FAM210A (55/330) [2 folds]



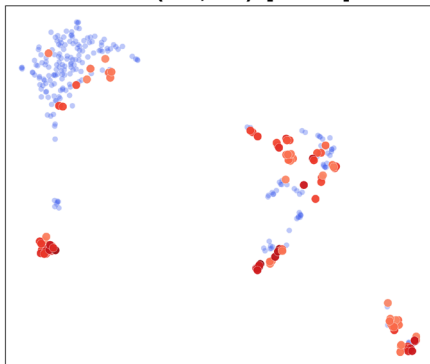
ADAR (67/330) [2 folds]



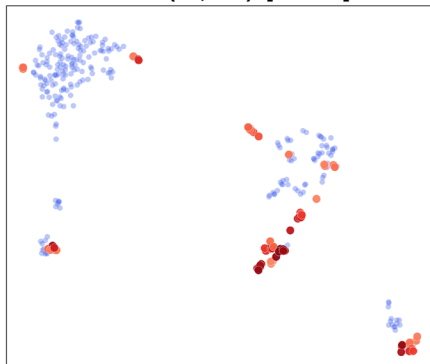
RP11-681L4.2 (16/330) [2 folds]



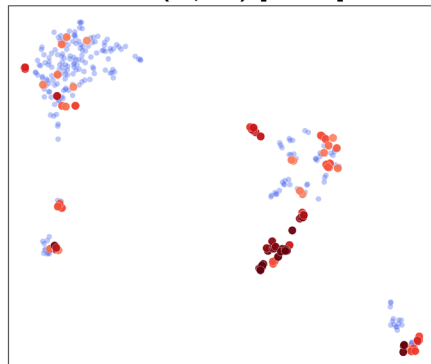
RND3 (102/330) [2 folds]



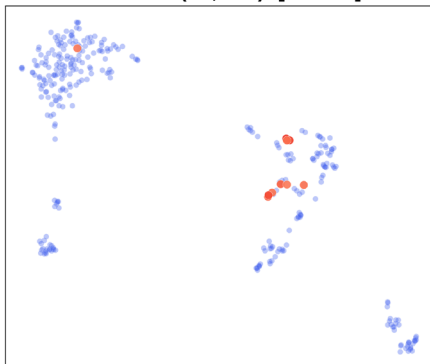
HERC6 (68/330) [2 folds]



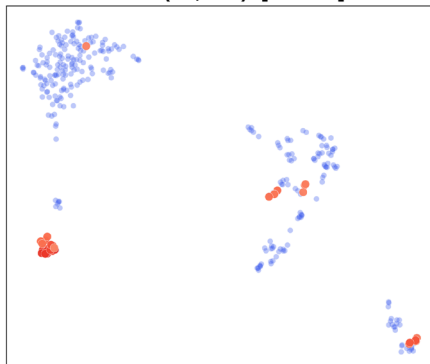
IFIT1 (93/330) [2 folds]



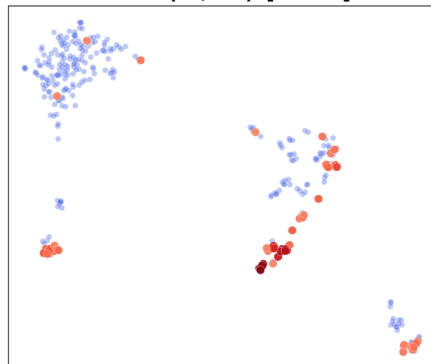
C6orf222 (10/330) [2 folds]



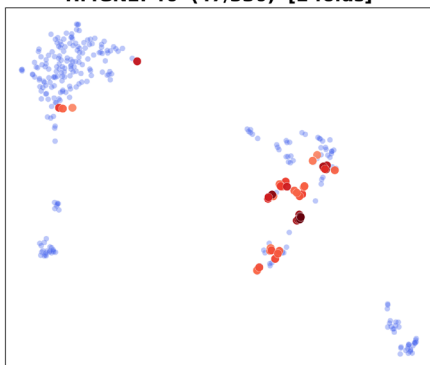
DOK5 (34/330) [2 folds]



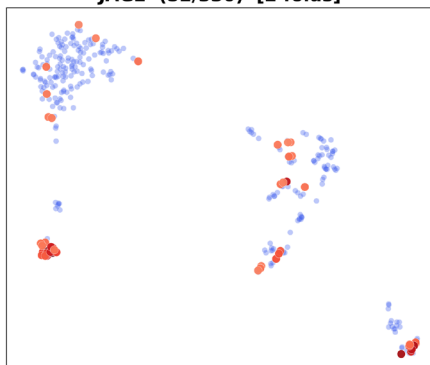
SP110 (64/330) [2 folds]



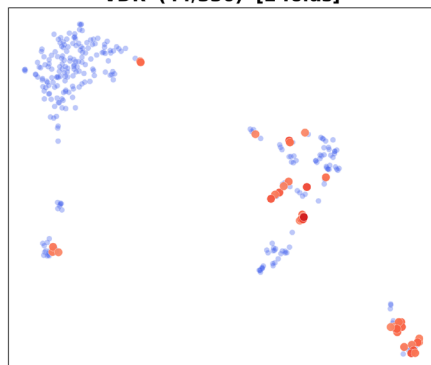
HMG2P46 (47/330) [2 folds]



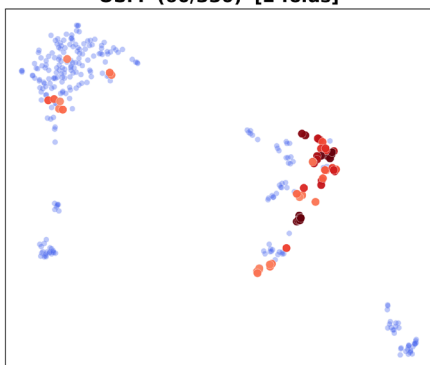
JAG2 (52/330) [2 folds]



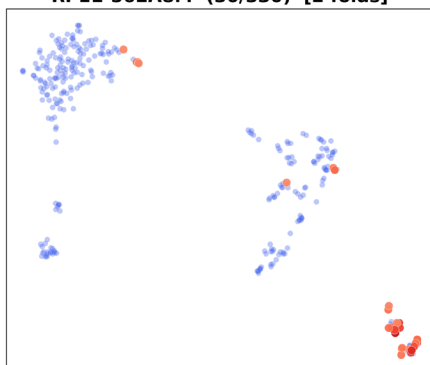
VDR (44/330) [2 folds]



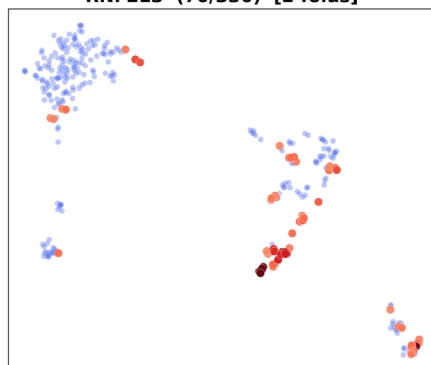
OSM (60/330) [2 folds]



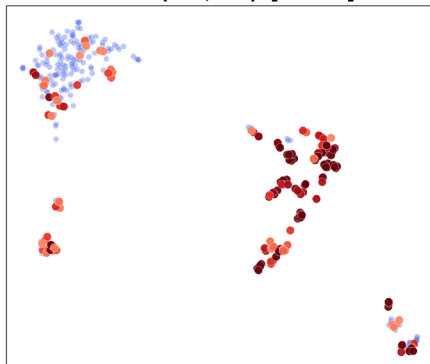
RP11-562A8.4 (30/330) [2 folds]



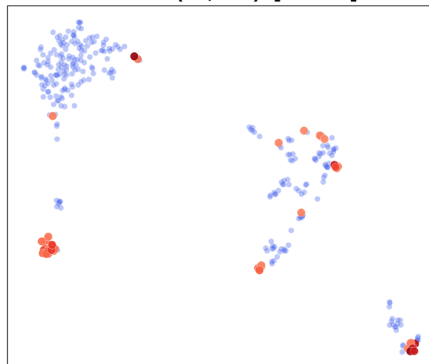
RNF213 (70/330) [2 folds]



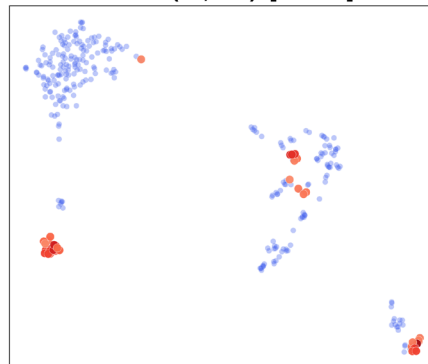
CXCL8 (161/330) [2 folds]



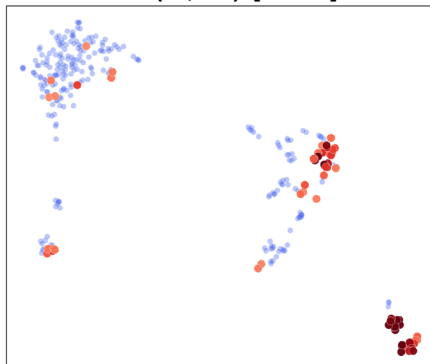
C7orf43 (45/330) [2 folds]



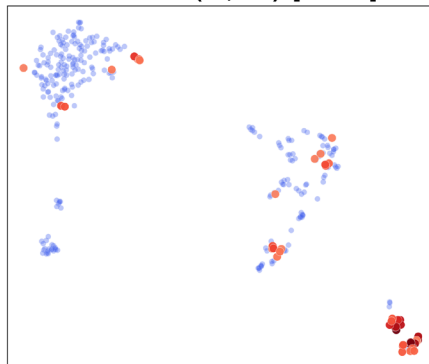
KCNK1 (42/330) [2 folds]



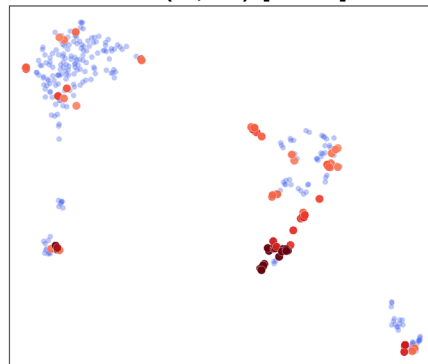
ARC (74/330) [2 folds]



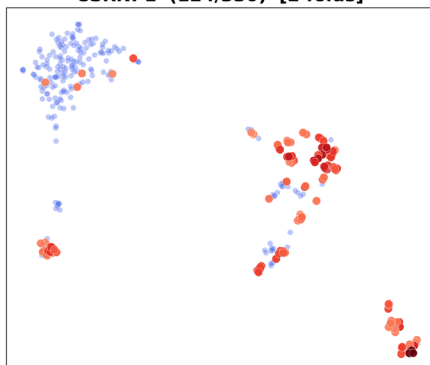
TNFRSF10D (50/330) [2 folds]



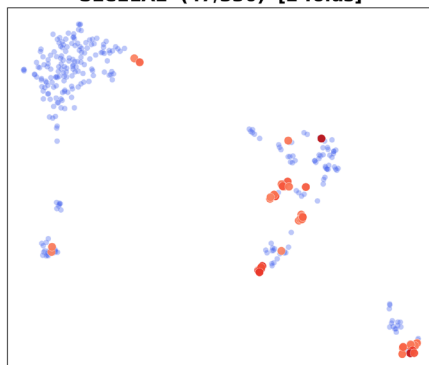
OAS2 (76/330) [2 folds]



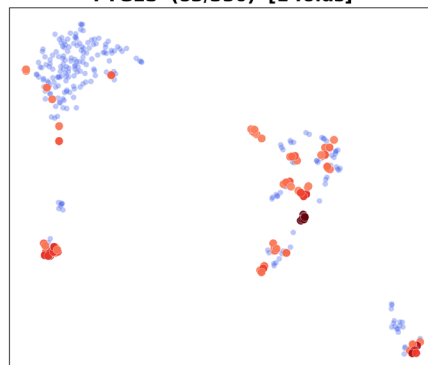
CSRNP1 (124/330) [2 folds]



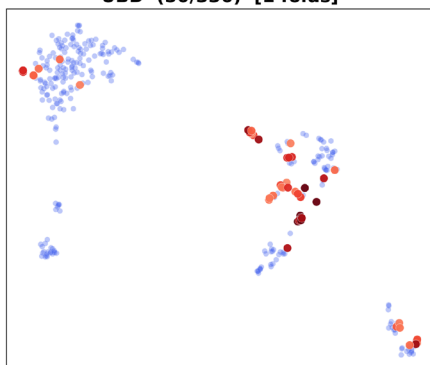
SLC11A2 (47/330) [2 folds]



PTGES (83/330) [1 folds]



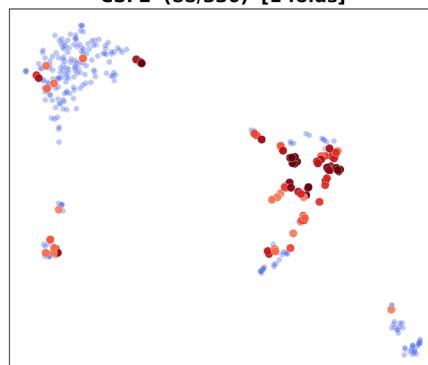
UBD (50/330) [1 folds]



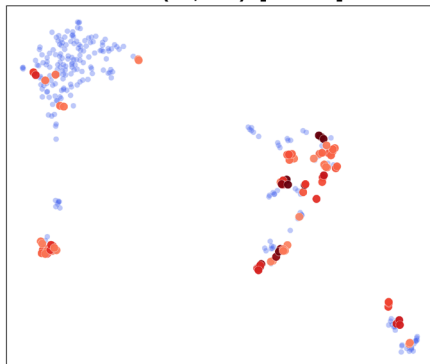
ATP6V0B (40/330) [1 folds]



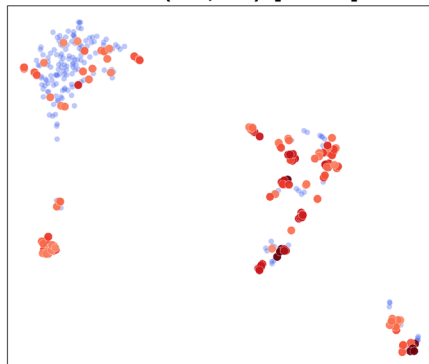
CSF2 (88/330) [1 folds]



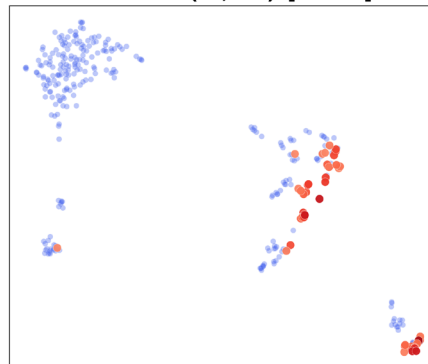
CD83 (93/330) [1 folds]



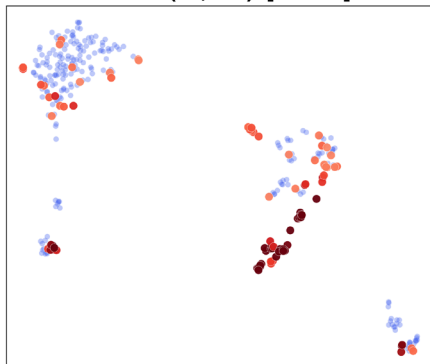
DUSP5 (147/330) [1 folds]



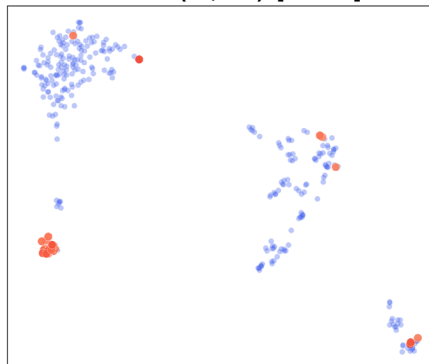
HS3ST3B1 (51/330) [1 folds]



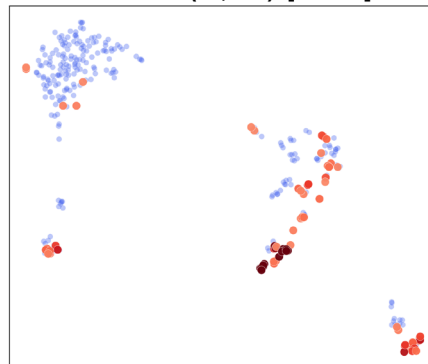
RSAD2 (93/330) [1 folds]



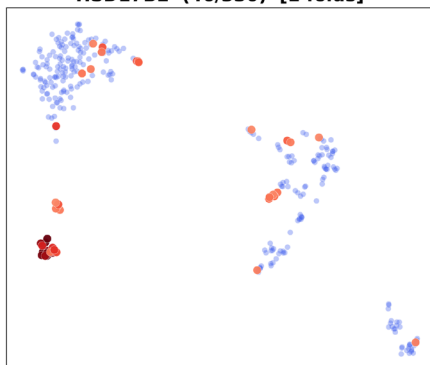
DYNC1I1 (27/330) [1 folds]



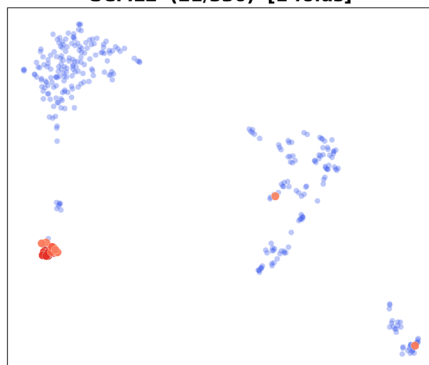
TNFSF13B (75/330) [1 folds]



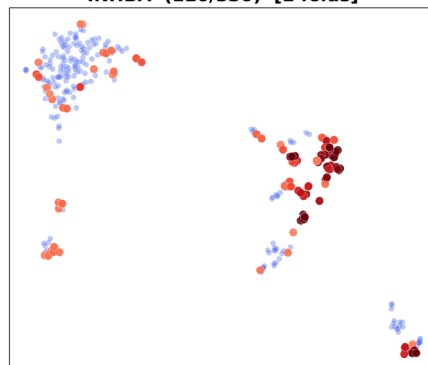
HSD17B2 (46/330) [1 folds]



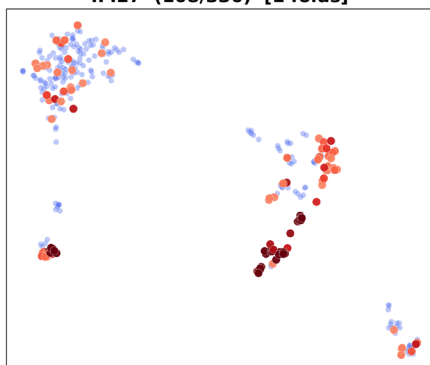
SCML2 (21/330) [1 folds]



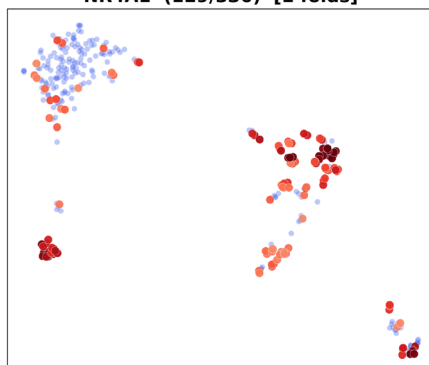
INHBA (110/330) [1 folds]



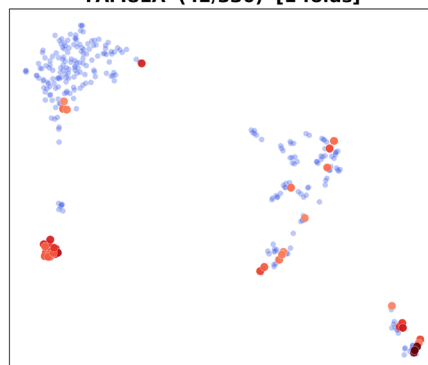
IFI27 (108/330) [1 folds]



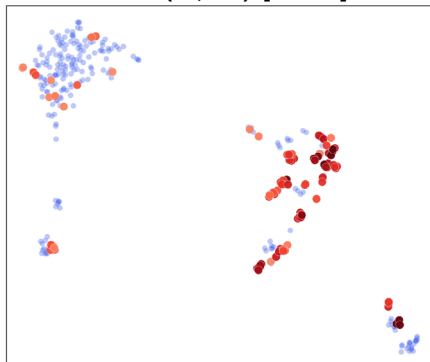
NR4A1 (129/330) [1 folds]



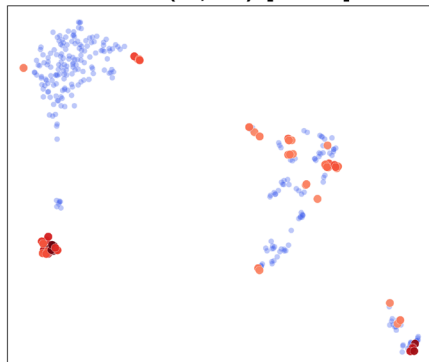
FAM81A (42/330) [1 folds]



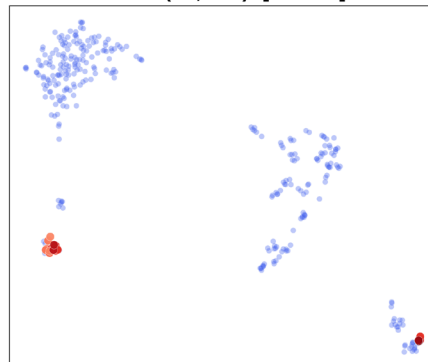
IRAK2 (99/330) [1 folds]



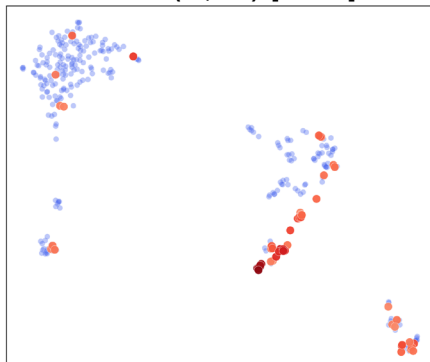
RPP25 (60/330) [1 folds]



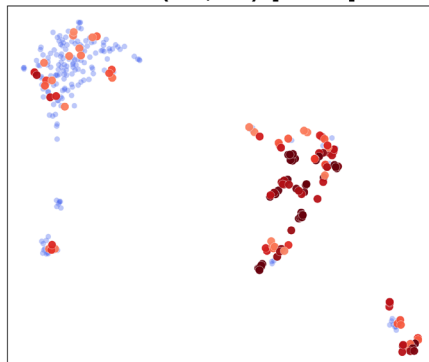
FZD9 (20/330) [1 folds]



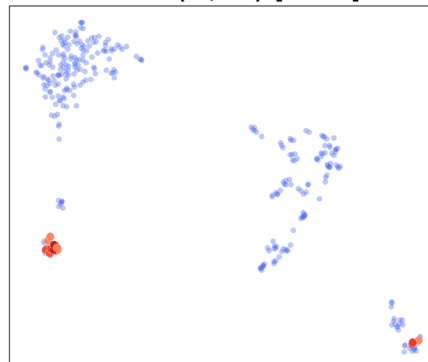
PARP12 (56/330) [1 folds]



CXCL1 (136/330) [1 folds]



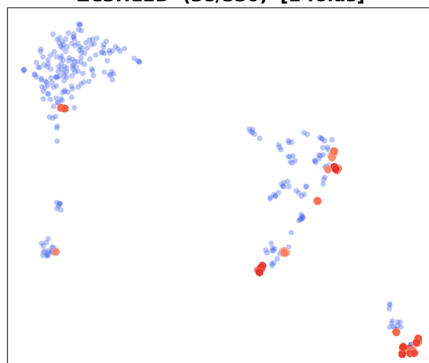
C2CD4C (19/330) [1 folds]



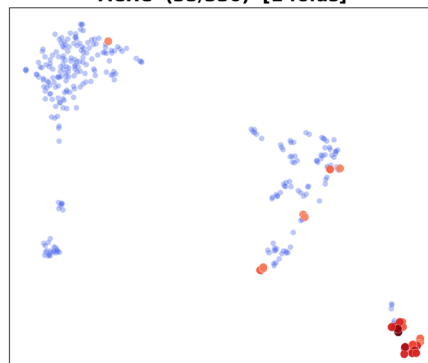
NDUFAF4 (45/330) [1 folds]



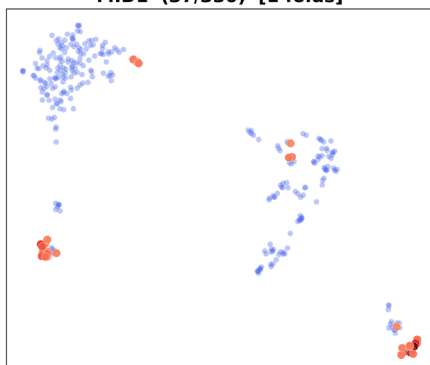
ZC3H12D (38/330) [1 folds]



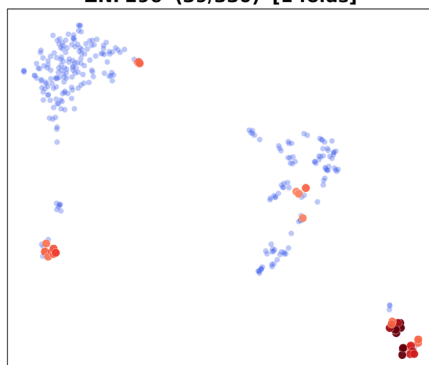
ACRC (35/330) [1 folds]



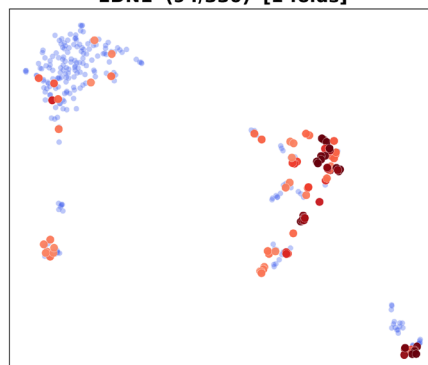
MID1 (37/330) [1 folds]



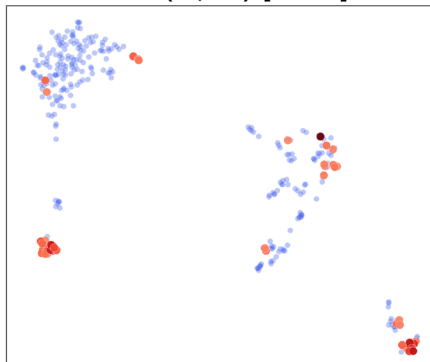
ZNF296 (39/330) [1 folds]



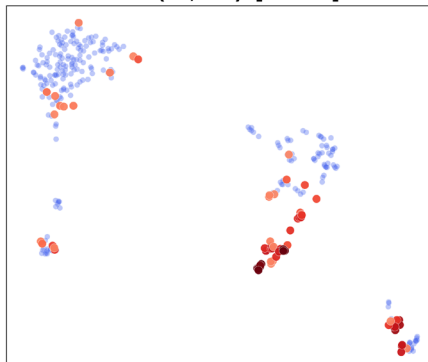
EDN1 (94/330) [1 folds]



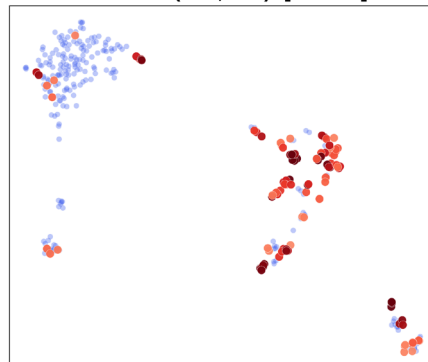
RELL2 (52/330) [1 folds]



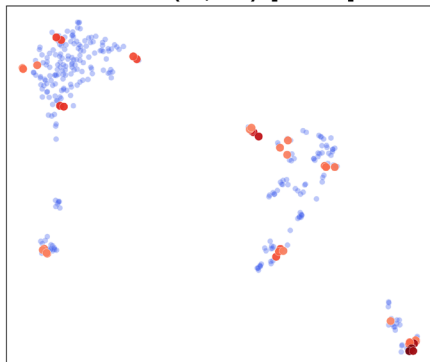
IRF7 (71/330) [1 folds]



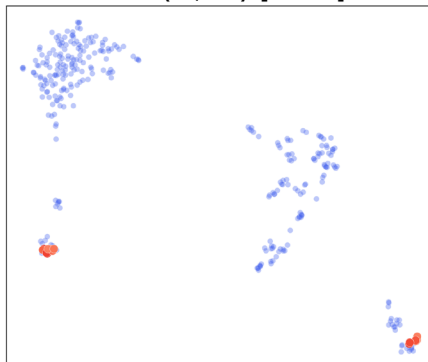
TNFAIP3 (109/330) [1 folds]



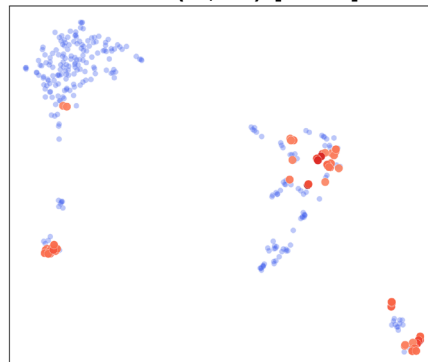
MYADM (41/330) [1 folds]



HCN4 (17/330) [1 folds]



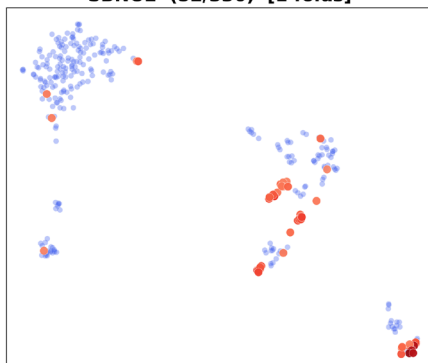
ZBTB10 (59/330) [1 folds]



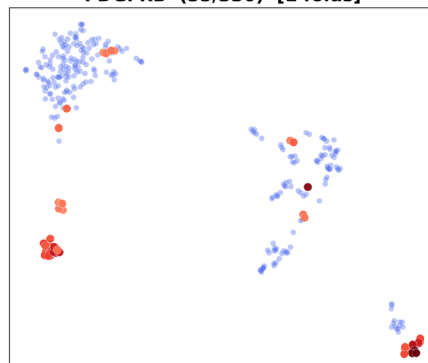
PKMP1 (25/330) [1 folds]



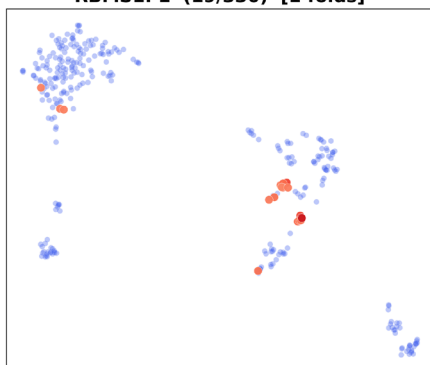
SBNO2 (52/330) [1 folds]



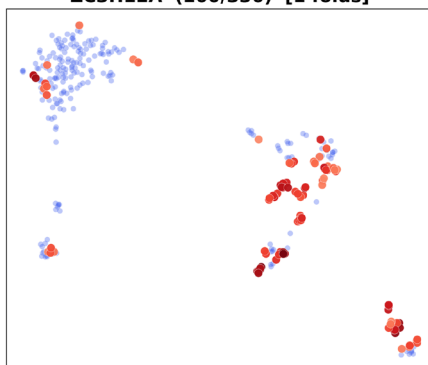
PDGFRB (55/330) [1 folds]



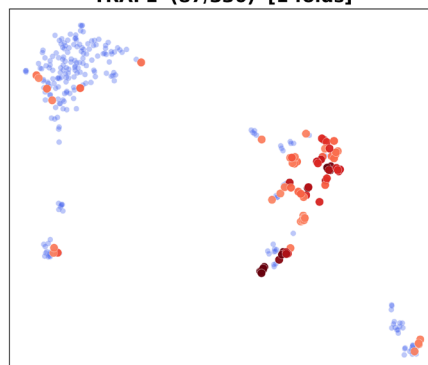
RBMS1P1 (19/330) [1 folds]



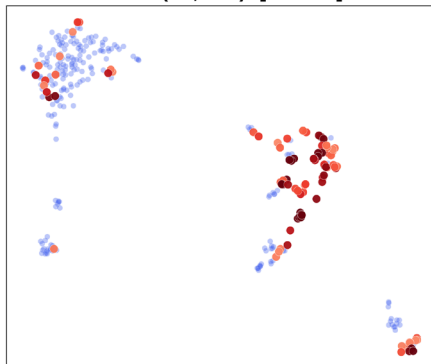
ZC3H12A (100/330) [1 folds]



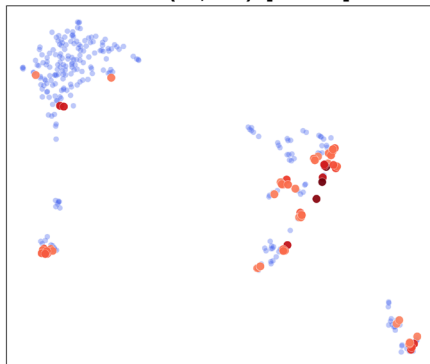
TRAF1 (87/330) [1 folds]



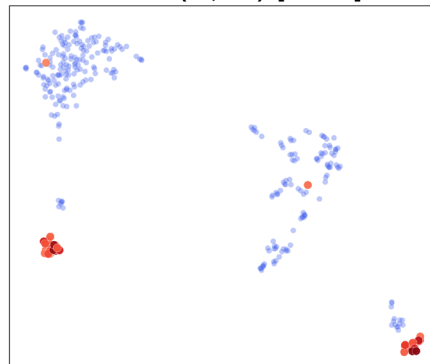
G0S2 (97/330) [1 folds]



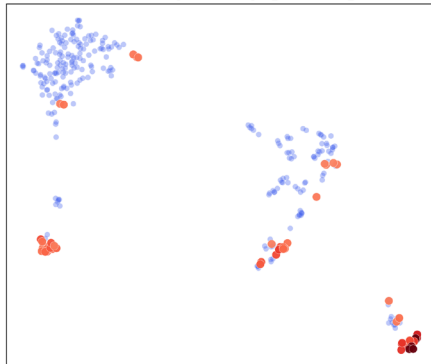
SGPP2 (64/330) [1 folds]



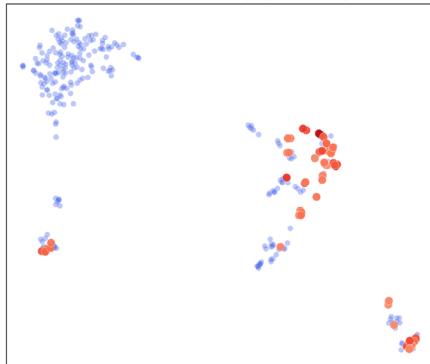
SEMA5A (39/330) [1 folds]



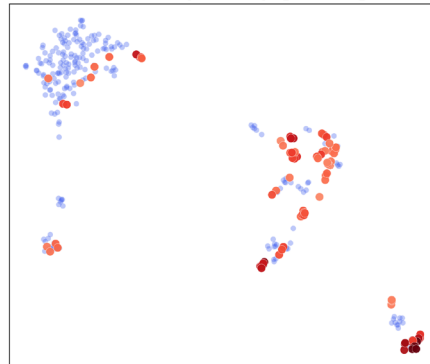
ATP2B4 (60/330) [1 folds]



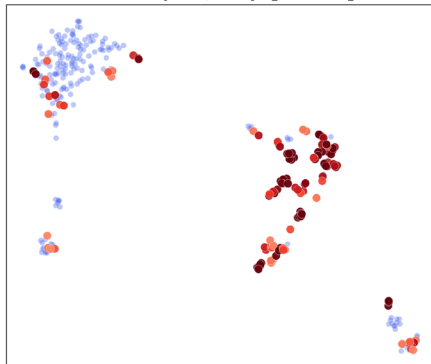
ARL5B (59/330) [1 folds]



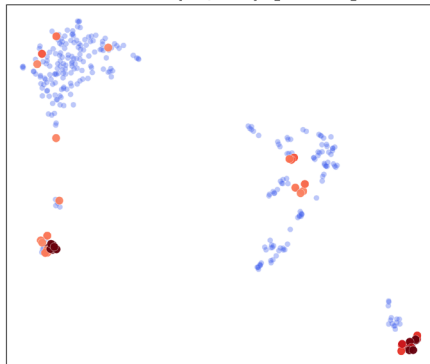
BHLHE40 (92/330) [1 folds]



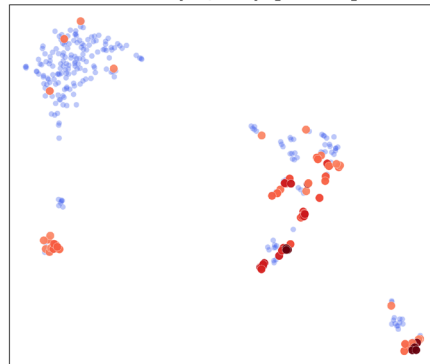
CXCL2 (129/330) [1 folds]



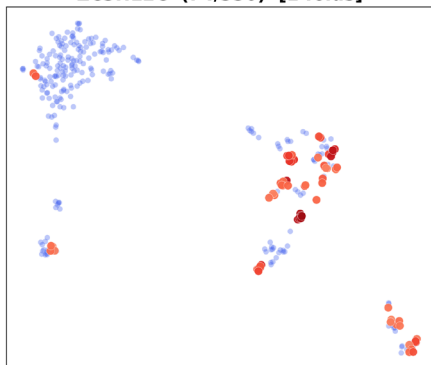
PCDH18 (45/330) [1 folds]



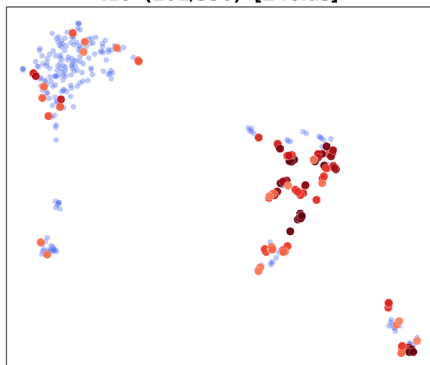
SLC2A6 (81/330) [1 folds]



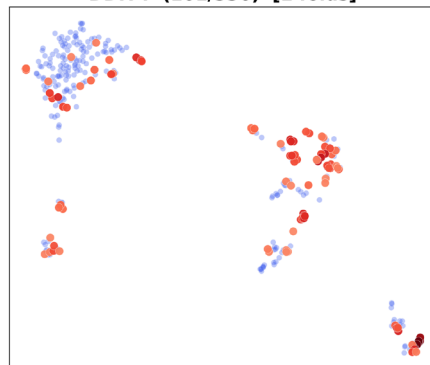
ZC3H12C (74/330) [1 folds]



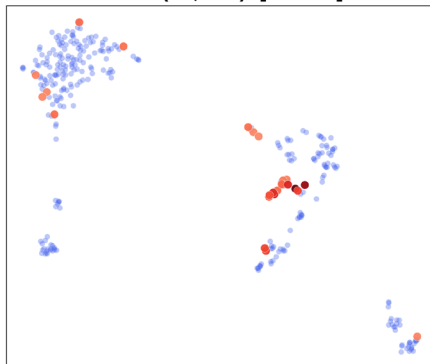
IL6 (101/330) [1 folds]



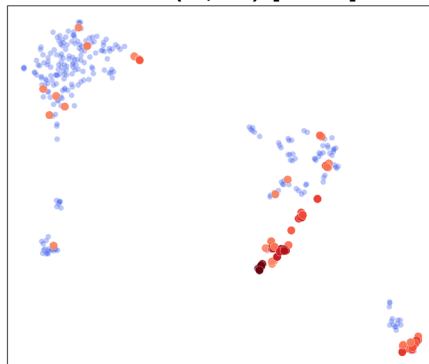
DDIT4 (101/330) [1 folds]



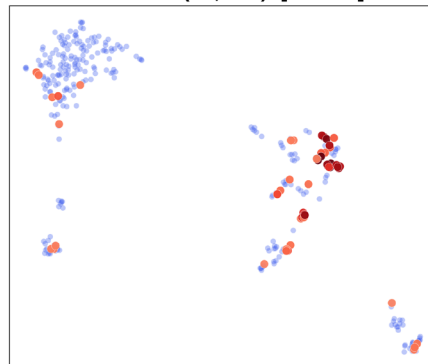
SAA4 (26/330) [1 folds]



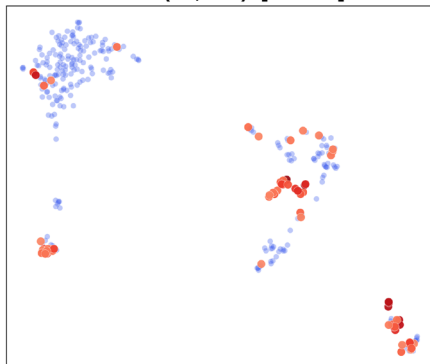
PARP14 (67/330) [1 folds]



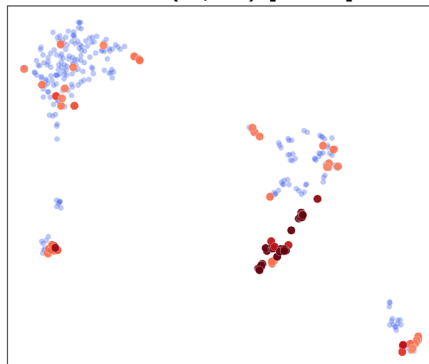
TNFSF15 (57/330) [1 folds]



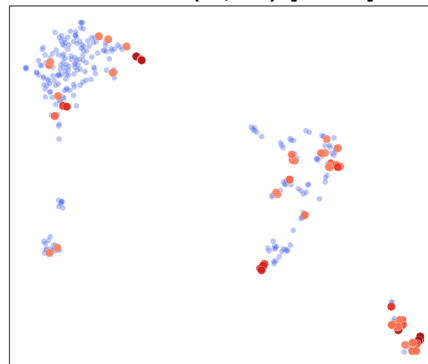
RND1 (67/330) [1 folds]



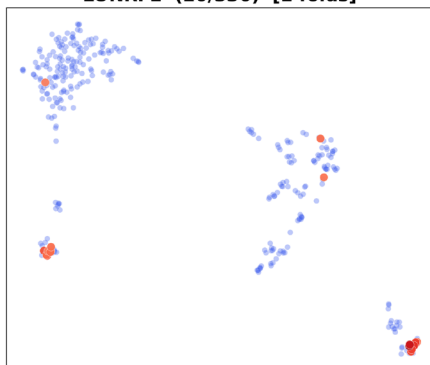
CMPK2 (83/330) [1 folds]



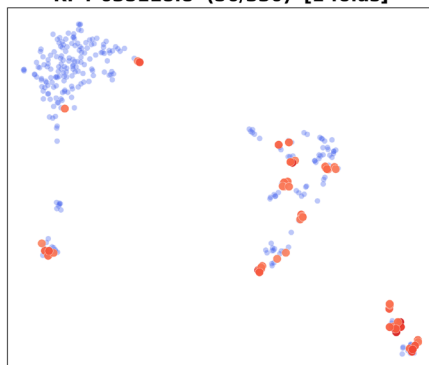
LINC00854 (66/330) [1 folds]



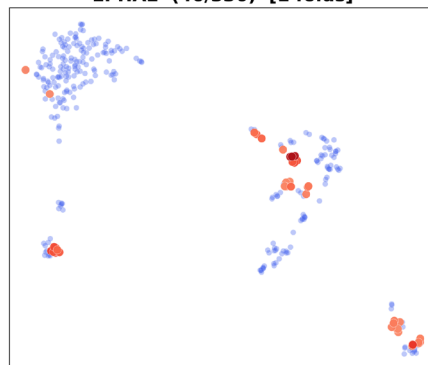
LONRF2 (20/330) [1 folds]



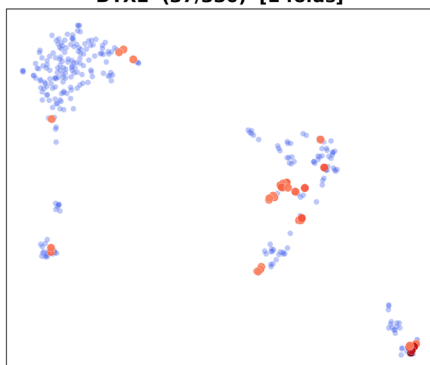
RP4-635E18.8 (56/330) [1 folds]



EPHA2 (46/330) [1 folds]



DTX2 (37/330) [1 folds]



Additional figure 3. Per-sample response for all 142 selected genes. Absolute $\log_2$ fold change for each of the 142 genes selected across cross-validation folds, overlaid on the sample-level UMAP embedding of all 330 samples. Each panel shows one gene, with samples reaching $|\log_2 \text{fold change}| \geq 1$ relative to study-matched mock shown in red and scaled by magnitude (capped at 4), and samples below this threshold shown in blue. Panel titles give the gene symbol, the number of samples reaching the threshold out of 330, and the number of cross-validation folds in which the gene was selected. Panels are ordered by selection frequency.