Machine learning on large-scale proteomics data identifies tissue- and cell type-specific proteins
Tine Claeys¹,², Maxime Menu¹,², Robbin Bouwmeester¹,², Kris Gevaert¹,², Lennart Martens¹,²,*
¹ VIB-UGent Center for Medical Biotechnology, VIB, 9000 Ghent, Belgium
² Department of Biomolecular Medicine, Ghent University, 9000 Ghent, Belgium
* To whom correspondence should be addressed; lennart.martens@ugent.be
Supplementary data

Supplementary table 1. Overview of all used public proteomics datasets and their PRIDE accession number.

Supplementary table 2. Tsv file containing result of manual annotation effort on a raw file level.

Supplementary table 3. Overview of used algorithms and their hyperparameters.

Supplementary figure 1. Structure of the protein expression atlas built in MySQL.

Supplementary table 4. Structure of the MySQL database with the primary keys and size of each table.

Supplementary table 5. Overview of all tissues and cell types in the database.

Supplementary figure 2. Distribution of the raw files and projects per tissue. The most represented class is blood with 4,105 raw files and the least represented classes are epididymis, parathyroid gland, trachea, anus, uterus and seminal vesicle with only two raw files.

Supplementary table 6. Overview of the proteins, raw files, tissues and cell types for the six different atlases that were built using the MySQL database.

Supplementary figure 3. Heatmap visualising classifier performance on tissue fPexAt data

Supplementary figure 4. Heatmap visualising classifier performance on cell-type fPexAt data

Supplementary figure 5. Heatmap visualising classifier performance on the tissue PexAt data.

Supplementary figure 6. Heatmap visualising classifier performance on the cell-type PexAt data

Supplementary table 7. The ten most important proteins for brain classification in the tissue classifier with their mean importance, the standard deviation over all decision trees. The column ‘Brain enrichment’ contains a surface-level cross-check with three other sources: UniProtKB, the Human Protein Atlas (HPA) and the Bgee gene expression database.

Supplementary table 8. The ten most important proteins for brain classification in the cell type classifier with their mean importance, the standard deviation over all decision trees. The column ‘Brain enrichment’ contains a surface-level cross-check with three other sources: UniProtKB, the Human Protein Atlas (HPA) and the Bgee gene expression database.

Supplementary table 9. The ten most important proteins for substantia nigra classification in the cell type classifier with their mean importance, the standard deviation over all decision trees. The column ‘Brain enrichment’ contains a surface-level cross-check with three other sources: UniProtKB, the Human Protein Atlas (HPA) and the Bgee gene expression database.

Supplementary table 10. The ten most important proteins for frontal cortex classification in the cell type classifier with their mean importance, the standard deviation over all decision trees. The column ‘Brain enrichment’ contains a surface-level cross-check with three other sources: UniProtKB, the Human Protein Atlas (HPA) and the Bgee gene expression database.

Supplementary figure 7. t-SNE visualisation of NK-cells, heart and testis, coloured according to the assays.

Supplementary figure 8. Boxplots visualising the distribution of Kendall correlation coefficient comparison between ProteomicsDB and the Human Protein Atlas data (antibody atlas, unnormalized RNA atlas (tpm), normalized RNA atlas (nTPM) and protein-coding RNA atlas (ptpm) on the level of the organ (top) and the protein (bottom).
| PXD_accession     | PXD000072 | PXD006517 | PXD006776 | PXD006863 | PXD009258 | PXD009393 | PXD009261 |
|-------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
|                   | PXD003534 | PXD009118 | PXD008205 | PXD009396 | PXD010286 | PXD008568 | PXD010154 |
|                   | PXD001674 | PXD006805 | PXD001985 | PXD002437 | PXD009606 | PXD009348 | PXD020192 |
|                   | PXD006833 | PXD001325 | PXD002211 | PXD009438 | PXD010821 | PXD009917 | PXD004682 |
|                   | PXD004626 | PXD000533 | PXD007745 | PXD009112 | PXD009916 | PXD007975 | PXD009737 |
|                   | PXD009310 | PXD006576 | PXD006570 | PXD007123 | PXD009893 | PXD011214 | PXD009870 |
|                   | PXD006578 | PXD008311 | PXD002990 | PXD00529 | PXD005436 | PXD009813 | PXD007959 |
|                   | PXD002212 | PXD001326 | PXD006589 | PXD009196 | PXD011492 | PXD009061 | PXD008992 |
|                   | PXD007744 | PXD000004 | PXD010690 | PXD011309 | PXD009784 | PXD00561 | PXD015239 |
|                   | PXD006847 | PXD003215 | PXD003075 | PXD008252 | PXD003533 | PXD015523 | |
|                   | PXD010093 | PXD005733 | PXD009021 | PXD002014 | PXD008130 | PXD017213 | |
|                   | PXD000228 | PXD008440 | PXD010150 | PXD000440 | PXD009880 | PXD004928 | |
|                   | PXD002620 | PXD010271 | PXD006571 | PXD009709 | PXD006607 | PXD012971 | |
|                   | PXD007592 | PXD005477 | PXD010708 | PXD008443 | PXD011000 | PXD012970 | |
|                   | PXD005755 | PXD001524 | PXD004355 | PXD010284 | PXD011175 | PXD011178 | |
|                   | PXD011204 | PXD009343 | PXD005055 | PXD010260 | PXD007740 | PXD012960 | |
|                   | PXD004683 | PXD007846 | PXD004540 | PXD007080 | PXD001425 | PXD017261 | |
|                   | PXD003045 | PXD006176 | PXD007258 | PXD005323 | PXD007081 | PXD016865 | |
|                   | PXD005220 | PXD009645 | PXD008465 | PXD009144 | PXD005709 | PXD010508 | |
|                   | PXD006522 | PXD006465 | PXD002883 | PXD009781 | PXD004443 | PXD016261 | |
|                   | PXD009029 | PXD008832 | PXD009754 | PXD009440 | PXD010469 | PXD005970 | |
|                   | PXD006268 | PXD001324 | PXD008795 | PXD006911 | PXD006743 | PXD017268 | |
|                   | PXD009996 | PXD008723 | PXD006572 | PXD009254 | PXD011609 | PXD018678 | |
|                   | PXD006065 | PXD011183 | PXD008934 | PXD009056 | PXD008354 | PXD017976 | |
|                   | PXD006897 | PXD008996 | PXD009755 | PXD007122 | PXD009300 | PXD016437 | |
|                   | PXD009624 | PXD009350 | PXD009646 | PXD011241 | PXD009716 | PXD016555 | |
|                   | PXD004625 | PXD007624 | PXD010698 | PXD007705 | PXD007873 | PXD017804 | |
|                   | PXD009752 | PXD006265 | PXD010294 | PXD007894 | PXD005693 | PXD018682 | |
|                   | PXD004624 | PXD008029 | PXD008682 | PXD002133 | PXD007231 | PXD013649 | |

Supplementary Table S1. Overview of all used public proteomics datasets and their PRIDE accession number.

**Supplementary table 2.tsv**

Supplementary Table S2. Tsv file containg result of manual annotation effort on a raw file level.
| Classifier name                     | Code                                                                 |
|------------------------------------|----------------------------------------------------------------------|
| RandomForest unbalanced            | RandomForestClassifier(random_state=42, n_jobs=-1)                   |
| RandomForest dict balanced         | RandomForestClassifier(random_state=42, class_weight=dict_train_label_weights, n_jobs=-1) |
| RandomForest balanced              | RandomForestClassifier(random_state=42, class_weight='balanced', n_jobs=-1) |
| RandomForest balanced subsample    | RandomForestClassifier(random_state=42, class_weight='balanced_subsample', n_jobs=-1) |
| Balanced RandomForest              | BalancedRandomForestClassifier(random_state=42, n_jobs=-1)          |
| Balanced RandomForest balanced     | BalancedRandomForestClassifier(random_state=42, class_weight='balanced', n_jobs=-1) |
| Balanced Random Forest balanced subsample | BalancedRandomForestClassifier(random_state=42, class_weight='balanced_subsample', n_jobs=-1) |
| SVM unbalanced                     | SVC(random_state=42)                                                 |
| SVM                                | SVC(random_state=42, class_weight=dict_train_label_weights)          |
| XGBClassifier unbalanced           | XGBClassifier(random_state=42, objective='multi:softprob', eval_metric='mlogloss', num_class=num_classes, n_jobs=-1) |
| XGBClassifier dict balanced        | XGBClassifier(random_state=42, objective='multi:softprob', eval_metric='mlogloss', num_class=num_classes, weight=train_weights, n_jobs=-1) |
| LogisticRegression unbalanced      | LogisticRegression(random_state=42, multi_class='multinomial', n_jobs=-1) |
| LogisticRegression dict balanced   | LogisticRegression(random_state=42, multi_class='multinomial', class_weight=dict_train_label_weights, n_jobs=-1) |
| LogisticRegression balanced        | LogisticRegression(random_state=42, multi_class='multinomial', class_weight='balanced', n_jobs=-1) |

*Supplementary Table S3. Overview of used algorithms and their hyperparameters.*
Supplementary Figure S1. Structure of the protein expression atlas built in MySQL.
### Table

| Table                | Primary keys                        | Rows      |
|----------------------|-------------------------------------|-----------|
| Assay                | Assay_id, project_id                | 15,146    |
| Modifications        | Mod_id                              | 1,505     |
| Peptide              | Peptide_id                          | 262,329   |
| Peptide_modifications| Peptide_id                          | 8,020,452 |
| Peptide_to_assay     | Assay_id, peptide_id                | 14,259,946|
| Peptide_to_protein   | Peptide_id, uniprot_id              | 263,466   |
| Project              | Project_id, PXD_accession            | 183       |
| Protein              | Uniprot_id                          | 15,108    |
| Tissue               | Tissue_id                           | 156       |
| Tissue_to_assay      | Assay_id, tissue_id                 | 15,145    |

*Supplementary Table S4. Structure of the MySQL database with the primary keys and size of each table.*

| tissue_name          | cell_type            | tissue_name          | cell_type          | tissue_name     | cell_type          |
|----------------------|----------------------|----------------------|-------------------|----------------|-------------------|
| Adipose tissue       | Adipose tissue       | Eye                  | Anterior chamber  | Small intestine | Small intestine   |
| Adrenal gland        | Adrenal gland        | Eye                  | Sclera            | Smooth muscle   | Smooth muscle     |
| Anus                 | Anus                 | Eye                  | Eye               | Spleen          | Spleen            |
| Appendix             | Appendix             | Eye                  | Vitreous humor    | Sputum          | Sputum            |
| Ascites              | Ascites              | Follicular fluid     | Follicular fluid  | Stomach         | Stomach           |
| B-cells              | B-cells              | Gall bladder         | Gall bladder      | Synovial tissue | Synovial tissue   |
| Blood                | Blood plasma         | Heart                | Heart             | T-cells         | CD4 T-cells       |
| Blood                | Blood                | hMSC                 | Umbilical cord    | T-cells         | CD8 T-cell        |
| Blood                | Blood serum          | hMSC                 | Adipose tissue    | T-cells         | T-cells           |
| Blood                | Blood platelet       | Kidney               | Kidney            | Testis          | Testis            |
| Blood                | Blood serum          | Kidney               | Kidney (glomeruli)| Thyroid         | Thyroid           |
| Blood                | Blood platelets      | Kidney               | Kidney (single tubules) | Tonsil | Tonsil           |
| Blood                | Erythrocyte          | Kidney               | Kidney (podocytes)| Tooth           | Tooth enamel      |
| Blood                | Neutrophil           | Kidney               | Kidney (primary urinary cells)| Trachea | Trachea           |
| Blood                | Reticulocyte         | Liver                | Liver             | Trophoblast     | Trophoblast       |
| Blood                | Blood plasma (EDTA)  | Lung                 | Lung              | Umbilical cord  | Umbilical vein    |
| Blood                | Blood plasma (with heparin added) | Lymph node | Lymph node | Ureter | Ureter           |
| Bone                 | Bone                 | Microvessel          | Microvessel       | Urinary bladder | Urinary bladder   |
| Bone marrow          | Bone marrow          | Monocytes            | Monocytes         | Urine           | Urine             |
| Brain                | Brain                | Nasal polyps         | Nasal polyps      | Uterine cervix  | Hela              |
| Brain                | Substantia Nigra     | NK-cells             | NK-cells          | Uterine cervix  | Uterine cervix    |
| Brain                | Temporal Lobe        | Ovary                | Ovary             | Uterus          | Uterus            |
| Brain                | Prefrontal cortex    | Oviduct              | Oviduct           |                 |                   |
| Brain                | Frontal cortex       | Pancreas             | Pancreas          |                 |                   |
| Brain                | Spinal cord          | Pancreas             | Pancreatic islet  |                 |                   |
| Brain                | Temporal lobe        | Parathyroid gland    | Parathyroid gland |                 |                   |
| Brain                | Parietal             | Parotid gland        | Parotid gland     |                 |                   |
| Brain                | Occipital cortex     | Pituitary gland      | Pituitary gland   |                 |                   |
| Brain                | Cerebellum           | Placenta             | Placenta          |                 |                   |
| Brain                | Medulla oblongata    | Prostate             | Prostate          |                 |                   |
Supplementary Table S5. Overview of all tissues and cell types in the database.

| Tissue Type | Complete Atlas | Complete Atlas Healthy Tissue | Complete Atlas Diseased Tissue | Filtered Atlas | Filtered Atlas Healthy Tissue | Filtered Atlas Diseased Tissue |
|-------------|----------------|-------------------------------|-------------------------------|----------------|-------------------------------|-------------------------------|
| Proteins    | 11,891         | 11,809                        | 11,538                        | 5,110          | 5,021                         | 3,371                         |
| Raw Files   | 8,279          | 3,191                         | 5,088                         | 8,246          | 3,173                         | 5,073                         |
| Tissues     | 65             | 58                            | 32                            | 65             | 58                            | 32                            |
| Cell Types  | 97             | 81                            | 45                            | 97             | 81                            | 45                            |

Supplementary Figure S2. Distribution of the raw files and projects per tissue. The most represented class is blood with 4,105 raw files and the least represented classes are epididymis, parathyroid gland, trachea, anus, uterus and seminal vesicle with only two raw files.
Supplementary Table S6. Overview of the proteins, raw files, tissues and cell types for the six different atlases that were built using the MySQL database.

| Model                                 | Accuracy | F1_macro | precision_macro | recall_macro | F1_weighted | precision_weighted | recall_weighted |
|---------------------------------------|----------|----------|----------------|--------------|-------------|--------------------|-----------------|
| SVM unbalanced                        | 0.63     | 0.707    | 0.757          | 0.707        | 0.099       | 0.027              | 0.83            |
| SVM                                   | 0.942    | 0.918    | 0.933          | 0.947        | 0.942       | 0.947              | 0.942           |
| Randomforest unbalanced               | 0.903    | 0.976    | 0.902          | 0.975        | 0.903       | 0.905              | 0.903           |
| RandomForest dict balanced            | 0.983    | 0.976    | 0.902          | 0.975        | 0.982       | 0.984              | 0.983           |
| RandomForest balanced                 | 0.992    | 0.975    | 0.981          | 0.975        | 0.992       | 0.984              | 0.982           |
| Random Forest balanced subsample      | 0.939    | 0.935    | 0.948          | 0.943        | 0.945       | 0.967              | 0.939           |
| Balanced Randomforest                 | 0.917    | 0.919    | 0.935          | 0.928        | 0.915       | 0.953              | 0.917           |
| Balanced RandomForest balanced subsample | 0.916   | 0.917    | 0.933          | 0.926        | 0.924       | 0.951              | 0.916           |
| Balanced RandomForest balanced subsample | 0.955   | 0.946    | 0.954          | 0.955        | 0.955       | 0.972              | 0.955           |
| XGBClassifier unbalanced              | 0.968    | 0.942    | 0.953          | 0.941        | 0.966       | 0.969              | 0.968           |
| XGBClassifier dict balanced           | 0.968    | 0.942    | 0.953          | 0.941        | 0.966       | 0.969              | 0.968           |
| XGBClassifier balanced                | 0.996    | 0.985    | 0.993          | 0.985        | 0.993       | 0.985              | 0.994           |
| LogisticRegression unbalanced         | 0.248    | 0.117    | 0.354          | 0.131        | 0.178       | 0.232              | 0.248           |
| LogisticRegression dict balanced      | 0.879    | 0.85     | 0.845          | 0.859        | 0.901       | 0.960              | 0.879           |
| LogisticRegression balanced           | 0.604    | 0.522    | 0.604          | 0.536        | 0.692       | 0.707              | 0.604           |

Supplementary Figure S3. Heatmap visualising classifier performance on tissue fPexAt data

| Metrics                                 | Accuracy | F1_macro | precision_macro | recall_macro | F1_weighted | precision_weighted | recall_weighted |
|-----------------------------------------|----------|----------|----------------|--------------|-------------|--------------------|-----------------|
| SVM unbalanced                          | 0.906    | 0.651    | 0.697          | 0.653        | 0.778       | 0.798              | 0.806           |
| SVM                                     | 0.939    | 0.9      | 0.933          | 0.901        | 0.916       | 0.942              | 0.939           |
| Randomforest unbalanced                 | 0.903    | 0.97    | 0.973          | 0.971        | 0.982       | 0.985              | 0.983           |
| RandomForest dict balanced              | 0.904    | 0.972    | 0.975          | 0.971        | 0.983       | 0.984              | 0.984           |
| RandomForest balanced                   | 0.994    | 0.973    | 0.977          | 0.973        | 0.993       | 0.985              | 0.994           |
| Random Forest balanced subsample        | 0.826    | 0.818    | 0.852          | 0.835        | 0.84       | 0.901              | 0.826           |
| Balanced Randomforest                   | 0.734    | 0.702    | 0.786          | 0.738        | 0.730       | 0.822              | 0.734           |
| Balanced RandomForest balanced subsample | 0.728   | 0.716    | 0.771          | 0.731        | 0.712       | 0.82              | 0.728           |
| Balanced RandomForest balanced subsample | 0.791   | 0.787    | 0.823          | 0.799        | 0.837       | 0.879              | 0.791           |
| XGBClassifier unbalanced                | 0.962    | 0.919    | 0.928          | 0.921        | 0.966       | 0.963              | 0.962           |
| XGBClassifier dict balanced             | 0.962    | 0.919    | 0.928          | 0.921        | 0.966       | 0.963              | 0.962           |
| XGBClassifier balanced                  | 0.952    | 0.919    | 0.928          | 0.921        | 0.956       | 0.963              | 0.956           |
| LogisticRegression unbalanced           | 0.225    | 0.096    | 0.125          | 0.109        | 0.152       | 0.185              | 0.225           |
| LogisticRegression dict balanced        | 0.873    | 0.751    | 0.806          | 0.815        | 0.878       | 0.907              | 0.873           |
| LogisticRegression balanced             | 0.559    | 0.497    | 0.585          | 0.553        | 0.56       | 0.706              | 0.559           |

Supplementary Figure S4. Heatmap visualising classifier performance on cell-type fPexAt data
**Supplementary Figure S5.** Heatmap visualising classifier performance on the tissue PexAt data.

| Method                              | Accuracy | F1_macro | precision_macro | recall_macro | F1_weighted | precision_weighted | recall_weighted |
|-------------------------------------|----------|----------|-----------------|-------------|-------------|--------------------|-----------------|
| SVM unbalanced                      | 0.892    | 0.538    | 0.606           | 0.517       | 0.665       | 0.674              | 0.692           |
| SVM                                | 0.873    | 0.34     | 0.864           | 0.833       | 0.871       | 0.887              | 0.873           |
| RandomForest unbalanced             | 0.931    | 0.897    | 0.921           | 0.895       | 0.925       | 0.934              | 0.913           |
| RandomForest dist balanced          | 0.935    | 0.916    | 0.933           | 0.902       | 0.93        | 0.939              | 0.935           |
| RandomForest balanced               | 0.933    | 0.895    | 0.921           | 0.891       | 0.929       | 0.937              | 0.933           |
| Random Forest balanced subsample    | 0.771    | 0.729    | 0.742           | 0.773       | 0.777       | 0.821              | 0.771           |
| Balanced RandomForest               | 0.65     | 0.63     | 0.646           | 0.701       | 0.643       | 0.722              | 0.65            |
| Balanced RandomForest balanced      | 0.65     | 0.624    | 0.643           | 0.702       | 0.638       | 0.719              | 0.65            |
| Balanced RandomForest balanced subsample | 0.608     | 0.761    | 0.777           | 0.802       | 0.807       | 0.810              | 0.608           |
| XGBClassifier unbalanced            | 0.939    | 0.901    | 0.916           | 0.902       | 0.916       | 0.942              | 0.939           |
| XGBClassifier dist balanced         | 0.939    | 0.901    | 0.916           | 0.902       | 0.916       | 0.942              | 0.939           |
| Logistic Regression unbalanced      | 0.217    | 0.059    | 0.075           | 0.079       | 0.111       | 0.142              | 0.127           |
| Logistic Regression dist balanced   | 0.7      | 0.64     | 0.661           | 0.694       | 0.596       | 0.746              | 0.7             |
| Logistic/Regression balanced        | 0.491    | 0.6      | 0.444           | 0.669       | 0.468       | 0.537              | 0.491           |

**Supplementary Figure S6.** Heatmap visualising classifier performance on the cell-type PexAt data.

| Method                              | Accuracy | F1_macro | precision_macro | recall_macro | F1_weighted | precision_weighted | recall_weighted |
|-------------------------------------|----------|----------|-----------------|-------------|-------------|--------------------|-----------------|
| SVM unbalanced                      | 0.67     | 0.332    | 0.553           | 0.537       | 0.932       | 0.469              | 0.67            |
| SVM                                | 0.68     | 0.424    | 0.865           | 0.845       | 0.878       | 0.896              | 0.88            |
| RandomForest unbalanced             | 0.927    | 0.865    | 0.883           | 0.864       | 0.919       | 0.926              | 0.927           |
| RandomForest dist balanced          | 0.93     | 0.873    | 0.885           | 0.87        | 0.933       | 0.929              | 0.93            |
| RandomForest balanced               | 0.93     | 0.873    | 0.885           | 0.872       | 0.922       | 0.925              | 0.93            |
| Random Forest balanced subsample    | 0.434    | 0.398    | 0.427           | 0.502       | 0.427       | 0.538              | 0.434           |
| Balanced RandomForest               | 0.36     | 0.333    | 0.36            | 0.454       | 0.338       | 0.398              | 0.36            |
| Balanced RandomForest balanced      | 0.365    | 0.337    | 0.367           | 0.457       | 0.336       | 0.42               | 0.365           |
| Balanced RandomForest balanced subsample | 0.375     | 0.353    | 0.382           | 0.485       | 0.362       | 0.470              | 0.375           |
| XGBClassifier unbalanced            | 0.926    | 0.869    | 0.883           | 0.872       | 0.921       | 0.928              | 0.926           |
| XGBClassifier dist balanced         | 0.926    | 0.865    | 0.883           | 0.872       | 0.921       | 0.928              | 0.926           |
| Logistic Regression unbalanced      | 0.176    | 0.0572   | 0.0708          | 0.0694      | 0.112       | 0.152              | 0.176           |
| Logistic Regression dist balanced   | 0.69     | 0.62     | 0.64            | 0.676       | 0.697       | 0.748              | 0.698           |
| Logistic/Regression balanced        | 0.469    | 0.394    | 0.436           | 0.417       | 0.457       | 0.54               | 0.469           |
### Brain - tissue classifier

| uniprot ID | importance | standard deviation | mean abundance | protein name | Brain enrichment |
|------------|------------|---------------------|----------------|--------------|-----------------|
| Q16352     | 0.0139     | 0.0651              | 0.0161         | Alpha-internexin | yes, all sources |
| P80723     | 0.0116     | 0.0409              | 0.0218         | Brain acid soluble protein 1 | yes, all sources |
| P62328     | 0.0093     | 0.0239              | 0.0520         | Thymosin beta-4 | yes, Bgee (Uniprot and HPA: in blood) |
| Q12860     | 0.0085     | 0.0717              | 0.0094         | Contactin-1 | yes, all sources |
| Q96GW7     | 0.0082     | 0.0610              | 0.0020         | Brevican core protein | yes, all sources |
| P63313     | 0.0080     | 0.0212              | 0.0000         | Thymosin beta-10 | yes, Bgee (Uniprot and HPA: in blood) |
| P07196     | 0.0079     | 0.0577              | 0.0073         | Neurofilament light polypeptide (NF-L) | yes, all sources |
| P35612     | 0.0072     | 0.0608              | 0.0006         | Beta-adducin | yes, all sources |
| O94811     | 0.0072     | 0.0601              | 0.0041         | Tubulin polymerization-promoting protein | yes, all sources |
| Q7L0J3     | 0.0072     | 0.0595              | 0.0026         | Synaptic vesicle glycoprotein 2A | yes, all sources |

**Supplementary Table S7.** The ten most important proteins for brain classification in the tissue classifier with their mean importance, the standard deviation over all decision trees. The column ‘Brain enrichment’ contains a surface-level cross-check with three other sources: UniProtKB, the Human Protein Atlas (HPA) and the Bgee gene expression database.

### Supplementary Table S8. The ten most important proteins for brain classification in the cell type classifier with their mean importance, the standard deviation over all decision trees. The column ‘Brain enrichment’ contains a surface-level cross-check with three other sources: UniProtKB, the Human Protein Atlas (HPA) and the Bgee gene expression database.

### Brain - cell classifier

| uniprot ID | importance | standard deviation | mean abundance | protein name | Brain enrichment |
|------------|------------|---------------------|----------------|--------------|-----------------|
| Q16352     | 0.037583   | 0.12124             | 0.016075       | Alpha-internexin | yes, all sources |
| P07196     | 0.026597   | 0.123581            | 0.007261       | Neurofilament light polypeptide | yes, all sources |
| P78357     | 0.022235   | 0.111044            | 0.00183        | Contactin-associated protein 1 | yes, all sources |
| Q16720     | 0.018581   | 0.0992              | 0.000913       | Plasma membrane calcium-transporting ATPase 3 | yes, all sources |
| P80723     | 0.018152   | 0.052642            | 0.021842       | Brain acid soluble protein 1 | yes, all sources |
| P17600     | 0.017851   | 0.092861            | 0.004649       | Synapsin-1 | yes, all sources |
| P60201     | 0.016725   | 0.096025            | 0.050548       | Myelin proteolipid protein | yes, all sources |
| Q12860     | 0.016487   | 0.100304            | 0.009490       | Contactin-1 | yes, all sources |
| P05129     | 0.016079   | 0.084747            | 0.000902       | Protein kinase C gamma type | yes, all sources |
| Q92752     | 0.015677   | 0.092039            | 0.008958       | Tenascin-R | yes, all sources |

### Substantia nigra - cell classifier

| uniprot ID | importance | standard deviation | mean abundance | protein name | Brain enrichment |
|------------|------------|---------------------|----------------|--------------|-----------------|
| Q8N6N7     | 0.022      | 0.146683            | 0.002032       | Acyl-CoA-binding domain-containing protein 7 | Yes, HPA and Bgee |
Supplementary Table S9. The ten most important proteins for substantia nigra classification in the cell type classifier with their mean importance, the standard deviation over all decision trees. The column ‘Brain enrichment’ contains a surface-level cross-check with three other sources: UniProtKB, the Human Protein Atlas (HPA) and the Bgee gene expression database.

| Uniprot ID | Importance | Standard Deviation | Mean Abundance | Protein Name                                                                 | Brain Enrichment |
|------------|------------|--------------------|----------------|-------------------------------------------------------------------------------|------------------|
| P42684     | 0.030019   | 0.170584419        | 0.0000245      | Tyrosine-protein kinase ABL2                                                  | No               |
| Q9UNA1     | 0.02       | 0.14               | 0.0000522      | Rho GTPase-activating protein 26                                              | No               |
| Q6ICH7     | 0.018      | 0.132951119        | 0.0000757      | Aspartate beta-hydroxylase domain-containing protein 2                        | Yes, all sources|
| Q5TCQ9     | 0.01604    | 0.125472028        | 0.0000176      | Membrane-associated guanylate kinase, WW and PDZ domain-containing protein 3  | No               |
| Q9NVE7     | 0.016      | 0.125475097        | 0.0000359      | 4'-phosphopantetheine phosphatase                                             | No               |
| P10915     | 0.015987   | 0.125355759        | 0.0004557      | Hyaluronan and proteoglycan link protein 1                                   | Yes              |
| O96007     | 0.014043   | 0.117489241        | 0.0001154      | Molybdopterin synthase catalytic subunit                                      | Yes, all sources|
| Q9NZ56     | 0.014025   | 0.117488168        | 0.0001073      | Formin-2                                                                     | Yes, all sources|
| Q8IXS8     | 0.014021   | 0.117488866        | 0.0000485      | Protein FAM126B                                                              | Yes, Bgee        |
| Q96AB3     | 0.014      | 0.117490425        | 0.0000407      | Isochorismatase domain-containing protein 2                                  | No               |
Supplementary Table S10. The ten most important proteins for frontal cortex classification in the cell type classifier with their mean importance, the standard deviation over all decision trees. The column ‘Brain enrichment’ contains a surface-level cross-check with three other sources: UniProtKB, the Human Protein Atlas (HPA) and the Bgee gene expression database.

Supplementary figure 7. t-SNE visualisation of NK-cells, heart and testis, coloured according to the assays.
Supplementary figure 8. Boxplots visualising the distribution of Kendall correlation coefficient comparison between ProteomicsDB and the Human Protein Atlas data (antibody atlas, unnormalised RNA atlas (tpm), normalized RNA atlas (nTPM) and protein-coding RNA atlas (ptpm) on the level of the organ (top) and the protein (bottom).