Supporting Information Supplement Learning from heterogeneous data sources: An application in spatial proteomics

Lisa M. Breckels

July 7, 2015

1 Supporting Tables: Datasets

Labelled instances (markers)										Unlabelled	Total
40S R	60S R	CYT	ER	LYS	MT	CHR	NUC	PM	PROT		
23	36	25	44	16	150	18	18	43	14	722	1109

Supporting Table 1: Proteins identified in the E14TG2a mouse stem cell dataset including markers of protein sub-cellular localisation. 40S R = 40S Ribosome, 60S R = 60S Ribosome, CYT = Cytosol, ER = Endoplasmic reticulum, LYS = Lysosome, MT = Mitochondrion, CHR = Nucleus - Chromatin, NUC = Nucleus - Nucleolus, PM = Plasma membrane, PROT = Proteasome.

Labelled instances (markers)												Unlabelled	Total
CHR	CYT	CYT/NUC	END	ER	GA	LYS	MT	NUC	PM	RIB 40S	RIB 60S		
11	60	22	12	36	24	22	89	27	54	18	29	967	1371

Supporting Table 2: Proteins identified in the human HEK 293 dataset including markers of protein sub-cellular localisation. CHR = Chromatin associated, CYT = Cytosol, CYT/NUC = Cytosol or nucleus localised, END = Endosome, ER = Endoplasmic reticulum, GA = Golgi apparatus, LYS = Lysosome, MT = Mitochondria, NUC = Nucelus, PM = Plasma membrane, RIB 40S = Ribosome 40S, RIB 60S = Ribosome 60S.

Labelled instances (markers)									Unlabelled	Total
ER L	ER M	GA	MT	PL	PM	RIB	TGN	VA		
14	45	28	55	20	46	19	13	21	428	689

Supporting Table 3: Proteins identified in the *Arabidopsis thaliana* callus dataset including markers of protein sub-cellular localisation. ER L = Endoplasmic reticulum lumen, ER M = Endoplasmic reticulum membrane, GA = Golgi apparatus, MT = Mitochondria, PL = Plastid, PM = Plasma membrane, RIB = Ribosome, TGN = *Trans*-Golgi network, VA = Vacuole

Labelled instances (markers)					Unlabelled	Total
ER/VA	GA/CHL	MT	PM	TGN		
26	21	20	89	29	1155	1340

Supporting Table 4: Proteins identified in the *Arabidopsis thaliana* roots dataset including markers of protein sub-cellular localisation. ER/VA = Endoplasmic reticulum or vacuole, GA/CHL = Golgi apparatus or chloroplast, MT = Mitochondria, PM = Plasma membrane, TGN = *Trans*-Golgi network, VA = Vacuole

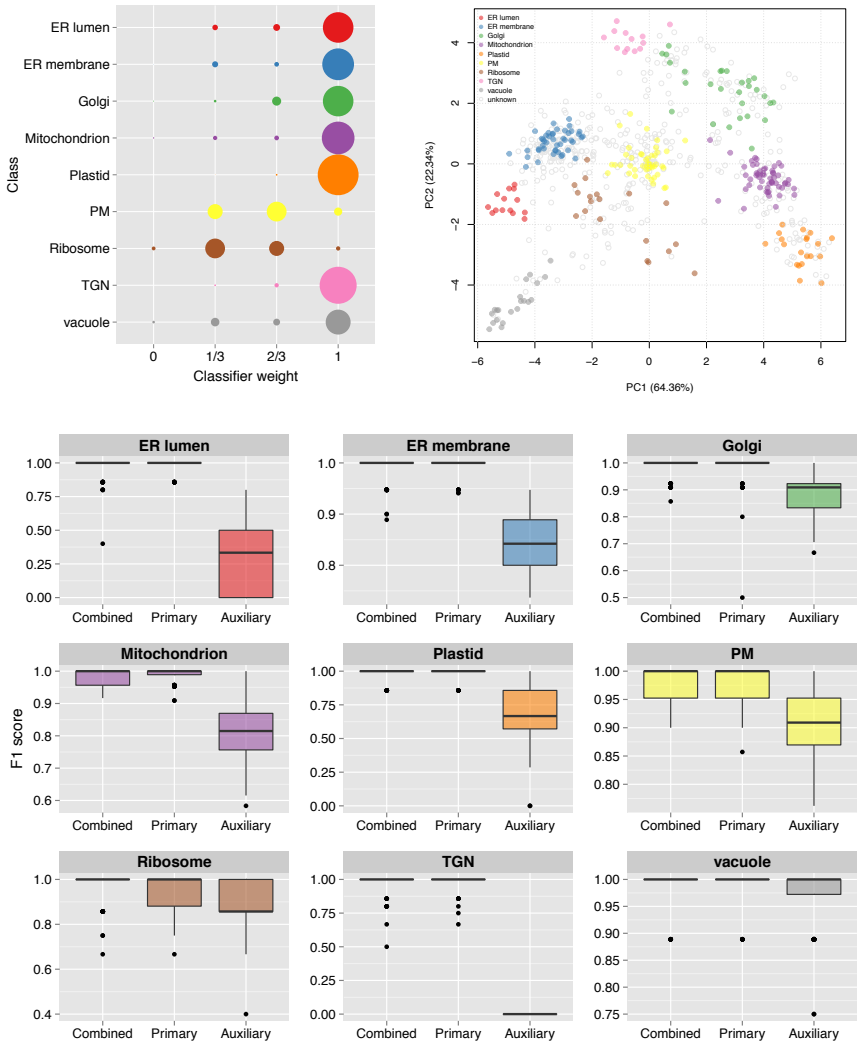
Labelled instances (markers)											Unlabelled	Total
CTK	ER	GA	LYS	MT	NUC	PER	PM	PROT	RIB 40S	RIB 60S		
7	28	13	8	29	21	4	34	15	20	32	677	888

Supporting Table 5: Proteins identified in the *Drosophila melanogaster* dataset including markers of protein sub-cellular localisation. CTK = Cytoskeleton, ER = Endoplasmic reticulum, GA = Golgi apparatus, LYS = Lysosome, MT = Mitochondria, NUC = Nucleus, PER = Peroxisome, PM = Plasma membrane, PROT = Proteasome, RIB 40S = Ribosome 40S, RIB 60S = Ribosome 60S

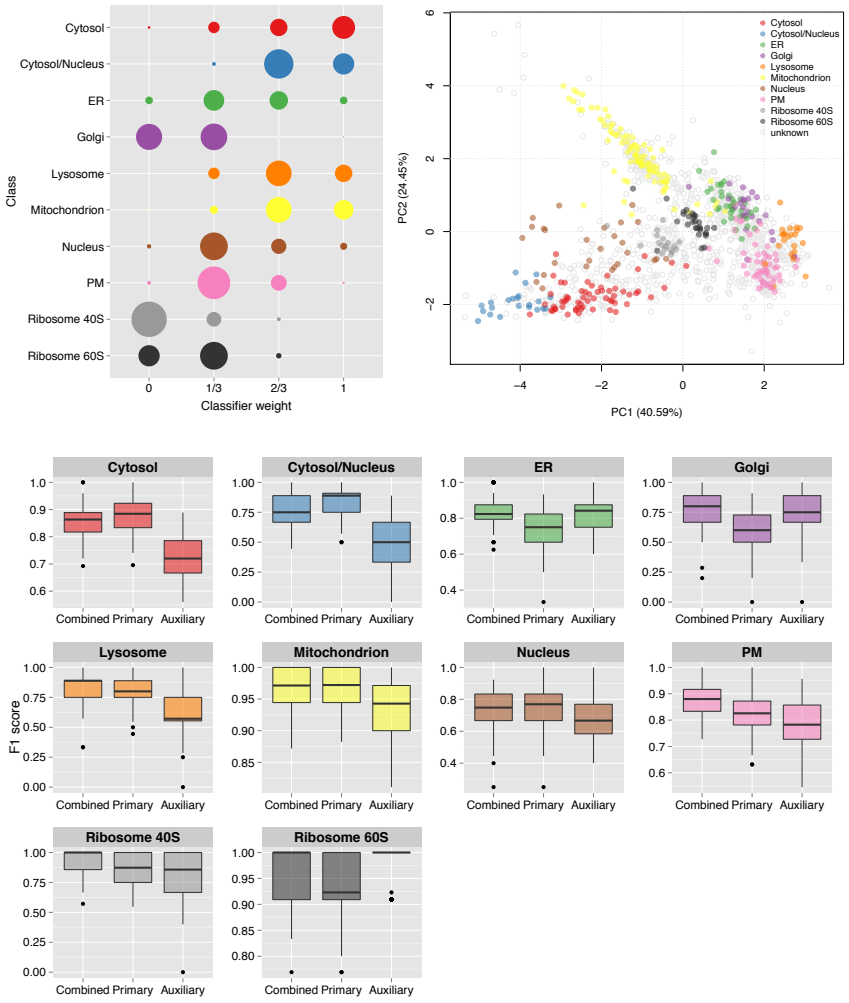
Dataset	# proteins		# features			
	Labelled	Unlabelled	Primary: LOPIT	Auxiliary: GO CC	HPA	YLoc
Mouse	387	722	8	314		387
Human*	404	967	8	355	18	
Fly	211	677	4	138		
Plant callus	261	428	16	70		
Plant roots	185	1155	6	153		

Supporting Table 6: Data dimensions for proteins identified in each primary dataset (LOPIT) and auxiliary datasets (GO CC: Gene ontology cellular compartment, HPA: Human Protein Atlas, YLoc: YLoc sequence and annotation features). *Only information from the HPA for 191 of the labelled markers was available, and for 479 of the unlabelled proteins.

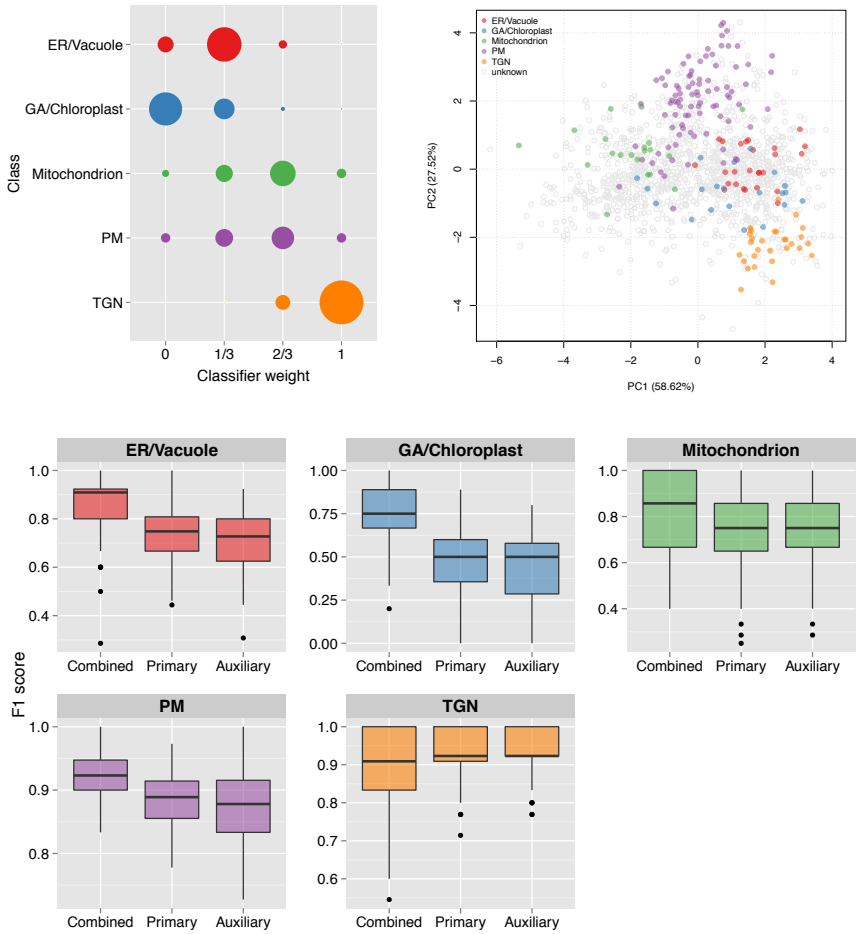
2 Supporting Figures: Results - k-NN TL



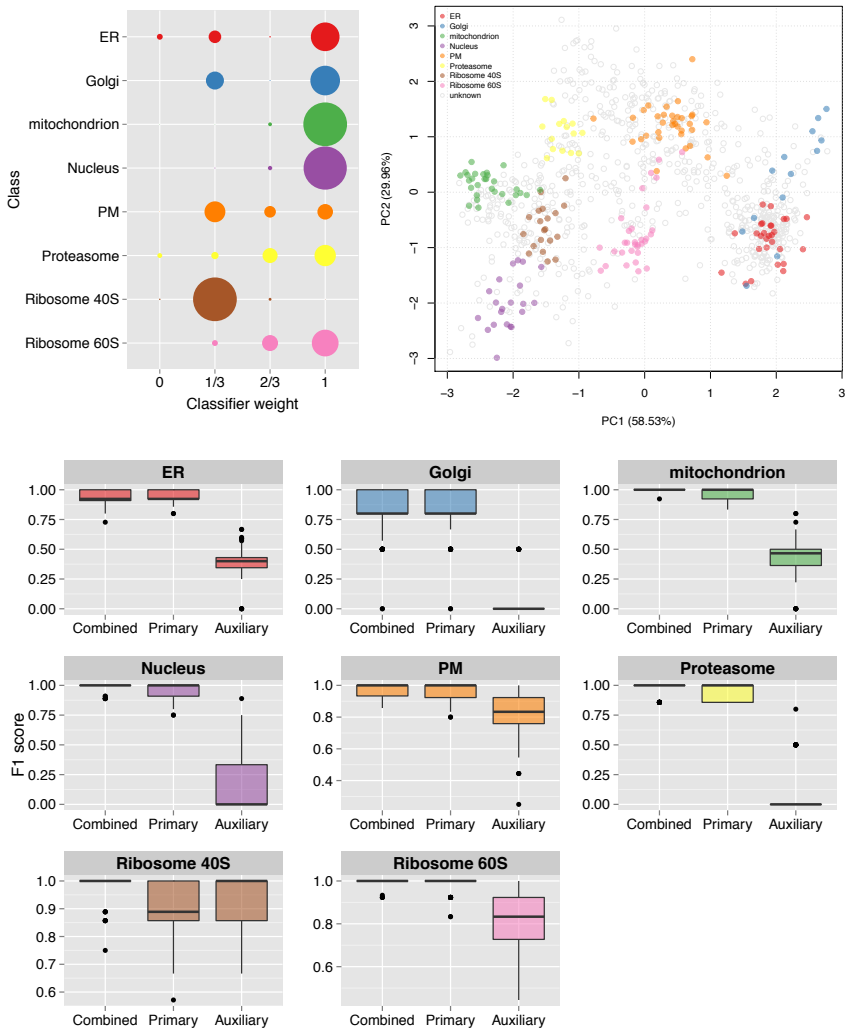
Supporting Fig. 1: Top left: Bubble plot, displaying the distribution of the optimised class weights over the 100 test partitions for the transfer learning algorithm applied to the *Arabidopsis thaliana* callus dataset. Top right: Principal components analysis plot (first and second components, of the possible eight), showing the clustering of proteins according to their density gradient distributions. Bottom: Sub-cellular class-specific box plots, displaying the estimated generalisation performance over 100 test partitions for the transfer learning algorithm applied with (i) optimised class-specific weights (combined), (ii) only primary data and (iii) only auxiliary data, for each sub-cellular class.



Supporting Fig. 2: Top left: Bubble plot, displaying the distribution of the optimised class weights over the 100 test partitions for the transfer learning algorithm applied to the human HEK293 dataset. Top right: Principal components analysis plot (first and second components, of the possible eight) of the human dataset, showing the clustering of proteins according to their density gradient distributions. Bottom: Sub-cellular class-specific box plots, displaying the estimated generalisation performance over 100 test partitions for the transfer learning algorithm applied with (i) optimised class-specific weights (combined), (ii) only primary data and (iii) only auxiliary data, for each sub-cellular class.

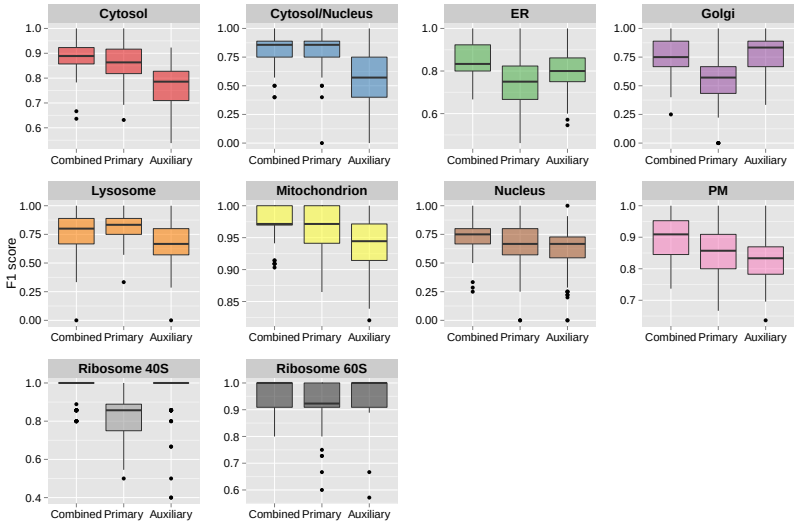


Supporting Fig. 3: Top left: Bubble plot, displaying the distribution of the optimised class weights over the 100 test partitions for the transfer learning algorithm applied to the *Arabidopsis thaliana* roots dataset. Top right: Principal components analysis plot (first and second components, of the possible eight) of the roots dataset, showing the clustering of proteins according to their density gradient distributions. Bottom: Sub-cellular class-specific box plots, displaying the estimated generalisation performance over 100 test partitions for the transfer learning algorithm applied with (i) optimised class-specific weights (combined), (ii) only primary data and (iii) only auxiliary data, for each sub-cellular class.

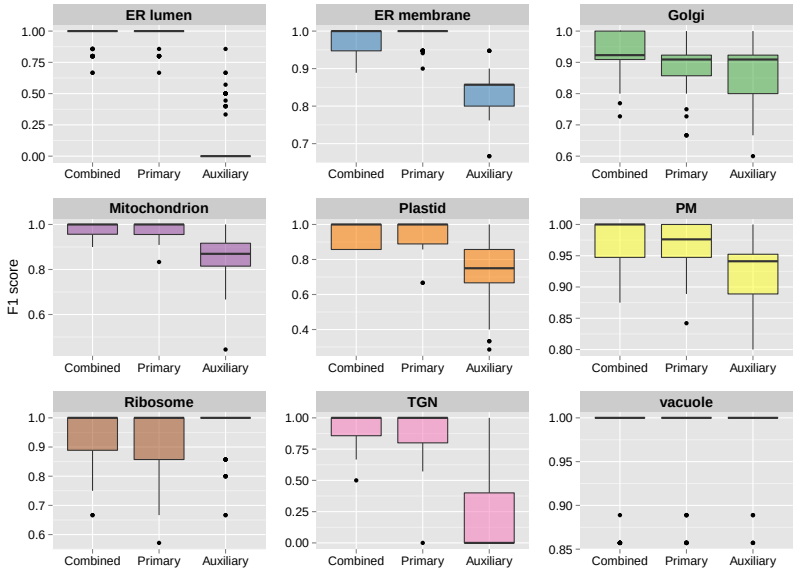


Supporting Fig. 4: Top left: Bubble plot, displaying the distribution of the optimised class weights over the 100 test partitions for the transfer learning algorithm applied to the *Drosophila melanogaster* dataset. Top right: Principal components analysis plot (first and second components, of the possible eight) of the *Drosophila* dataset, showing the clustering of proteins according to their density gradient distributions. Bottom: Sub-cellular class-specific box plots, displaying the estimated generalisation performance over 100 test partitions for the transfer learning algorithm applied with (i) optimised class-specific weights (combined), (ii) only primary data and (iii) only auxiliary data, for each sub-cellular class.

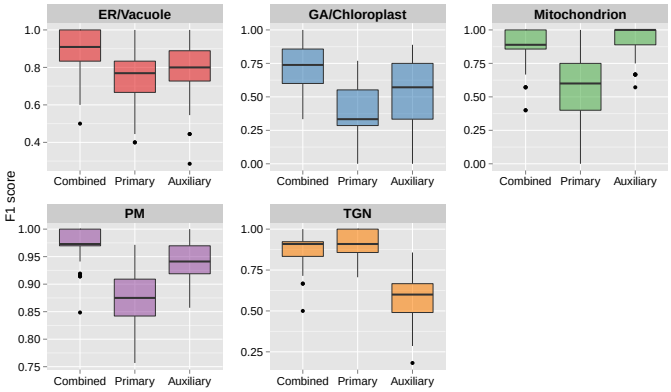
3 Supporting Figures: Results - SVM TL



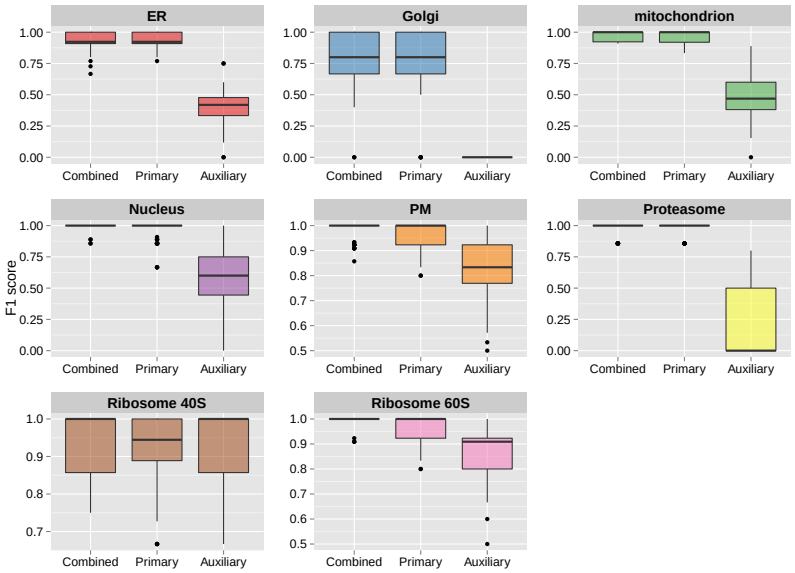
Supporting Fig. 5: Sub-cellular class-specific box plots, displaying the estimated generalisation performance over 100 test partitions for the SVM transfer learning algorithm applied to the human HEK293 dataset with (i) optimised class-specific weights (combined), (ii) only primary data and (iii) only auxiliary data, for each sub-cellular class.



Supporting Fig. 6: Sub-cellular class-specific box plots, displaying the estimated generalisation performance over 100 test partitions for the SVM transfer learning algorithm applied to the *Arabidopsis thaliana* callus dataset with (i) optimised class-specific weights (combined), (ii) only primary data and (iii) only auxiliary data, for each sub-cellular class.



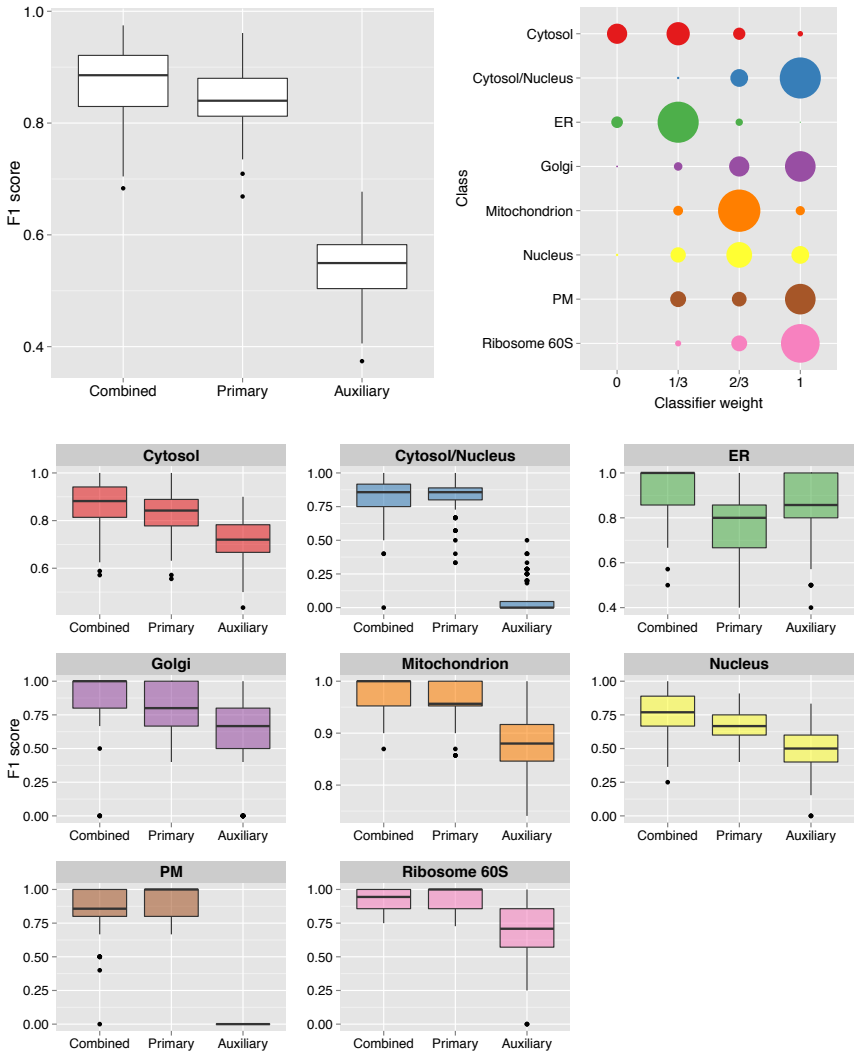
Supporting Fig. 7: Sub-cellular class-specific box plots, displaying the estimated generalisation performance over 100 test partitions for the SVM transfer learning algorithm applied to the *Arabidopsis thaliana* roots dataset with (i) optimised class-specific weights (combined), (ii) only primary data and (iii) only auxiliary data, for each sub-cellular class.



Supporting Fig. 8: Sub-cellular class-specific box plots, displaying the estimated generalisation performance over 100 test partitions for the SVM transfer learning algorithm applied to the *Drosophila melanogaster* dataset with (i) optimised class-specific weights (combined), (ii) only primary data and (iii) only auxiliary data, for each sub-cellular class.

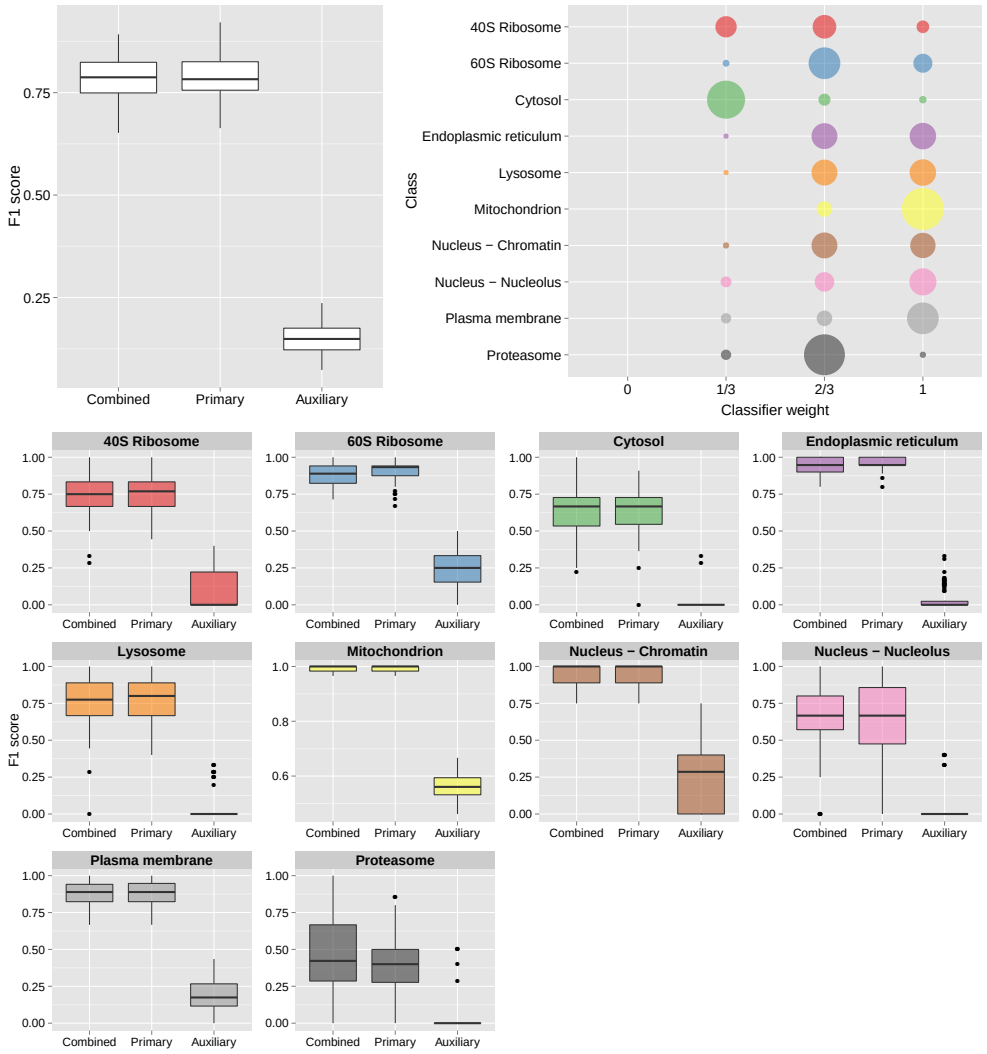
4 Supporting Figures: Other auxiliary data sources

4.1 The Human Proteome Atlas



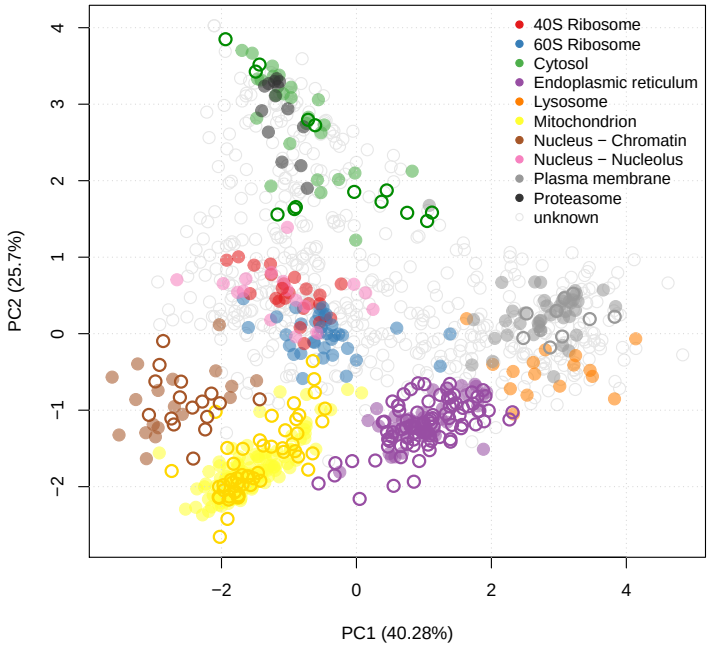
Supporting Fig. 9: Top left: Boxplot displaying the macro F1 scores over the 100 test partitions for the k-NN transfer learning algorithm applied with (i) optimised class-specific weights (combined), (ii) only primary data and (iii) only auxiliary data for the human HEK293 dataset. Top right: Bubble plot, displaying the distribution of the optimised class weights over the 100 test partitions for the transfer learning algorithm. Bottom: Boxplots, displaying the class specific generalisation performance over 100 test partitions for the k-NN TL experiments

4.2 YLoc



Supporting Fig. 10: Boxplots, displaying the overall (A) and class specific (C) estimated generalisation performance over 100 test partitions for the k-NN transfer learning (TL) algorithm applied with (i) optimised class-specific weights (combined), (ii) only primary data and (iii) only auxiliary YLoc data, for the E14TG2a mouse dataset. (B) Bubble plot, displaying the distribution of the optimised class weights over the 100 test partitions for the k-NN TL algorithm applied to the E14TG2a mouse dataset.

5 Biological applications



Supporting Fig. 11: Principal components analysis (PCA) plot of the E14TG2a mouse stem cell dataset highlighting the new localisations found using the k-NN TL method. Proteins are clustered according the their density gradient distributions. Each point on the PCA plot represents one protein. Proteins are coloured according to sub-cellular location, circled proteins represent new assignments and filled circles represent markers used in classifier training.

	40S R	60S R	CYT	ER	LYS	MT	CHR	NUC	PM	PROT	Total
SVM TL	0	0	56	37	1	45	8	0	57	0	204
SVM LOPIT only	0	4	0	46	0	47	9	5	38	0	149
Common assignments	0	0	0	28	0	42	7	0	31	0	108

Supporting Table 7: The number of sub-cellular assignments of the unlabelled proteins amongst the 10 known sub-cellular classes that are common from application of a SVM (on LOPIT only) and the SVM transfer learning (TL) (using LOPIT and GO CC) for the E14TG2a mouse dataset, and also assignments that are only found in the SVM TL and only using a SVM with LOPIT alone. The classification thresholds for the SVM and SVM TL were 0.850 and 0.785 respectively, based on a FDR of 5%.

	40S R	60S R	CYT	ER	LYS	MT	CHR	NUC	PM	PROT	Total
k-NN TL	0	0	14	85	0	52	16	0	9	0	176
k-NN LOPIT only*	0	0	0	0	0	0	0	0	0	0	0
Common assignments*	0	0	0	0	0	0	0	0	70	0	0

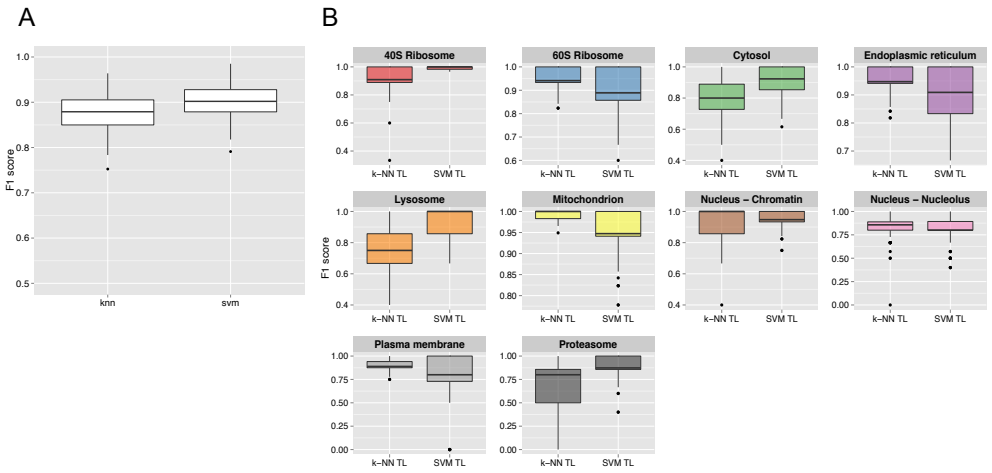
Supporting Table 8: The number of sub-cellular assignments of the unlabelled proteins amongst the 10 known sub-cellular classes that are common from application of a k-NN (on LOPIT only) and the k-NN transfer learning (TL) (using LOPIT and GO CC) for the E14TG2a mouse dataset, and also assignments that are only found in the k-NN TL and only using a k-NN with LOPIT alone. The classification threshold for k-NN TL was 0.805, based on a FDR of 5%. *A FDR of 5% was not achievable for the k-NN with LOPIT only thus no classifications were made. The lowest FDR achievable for k-NN with LOPIT only was 15%, which resulted in a classification threshold score of 1 (i.e. all neighbours must share the same class label for a protein to be assigned a class).

		SVM TL										
		40S R	60S R	CYT	ER	LYS	MT	CHR	NUC	PM	PROT	unknown
k-NN TL	40S R	0	0	0	0	0	0	0	0	0	0	0
	60S R	0	0	0	0	0	0	0	0	0	0	0
	CYT	0	0	5	0	0	0	0	0	0	0	9
	ER	0	0	0	35	0	0	0	0	0	0	50
	LYS	0	0	0	0	0	0	0	0	0	0	0
	MT	0	0	0	0	0	44	0	0	0	0	8
	CHR	0	0	0	0	0	0	8	0	0	0	8
	NUC	0	0	0	0	0	0	0	0	0	0	0
	PM	0	0	0	0	0	0	0	0	9	0	0
	PROT	0	0	0	0	0	0	0	0	0	0	0
	Unknown	0	0	51	2	1	1	0	0	48	0	443

Supporting Table 9: Contingency table showing the number of assignments of the unknowns in the E14TG2a mouse dataset using the k-NN transfer learning (TL) and SVM TL methods among the sub cellular classes that were included in the training data. The classification score threshold for the k-NN was 0.805 and for the SVM TL method was 0.785 (based on a FDR of 5%), proteins that did not achieve greater than equal to these scores were set to unknown. Reassuringly we found no counts off the diagonal except those in the unknown columns and rows, again highlighting that the results between the two classifiers are in high agreement. We see that many more assignments are made using the SVM TL method, however the extra assignments gained using SVM TL are all found to be labelled as unknown in the k-NN TL method, and visa versa. 40S R = 40S Ribosome, 60S R = 60S Ribosome, CYT = Cytosol, ER = Endoplasmic reticulum, LYS = Lysosome, MT = Mitochondrion, CHR = Nucleus - Chromatin, NUC = Nucleus - Nucleolus, PM = Plasma membrane, PROT = Proteasome.

6 k-NN TL vs SVM TL: A comparison

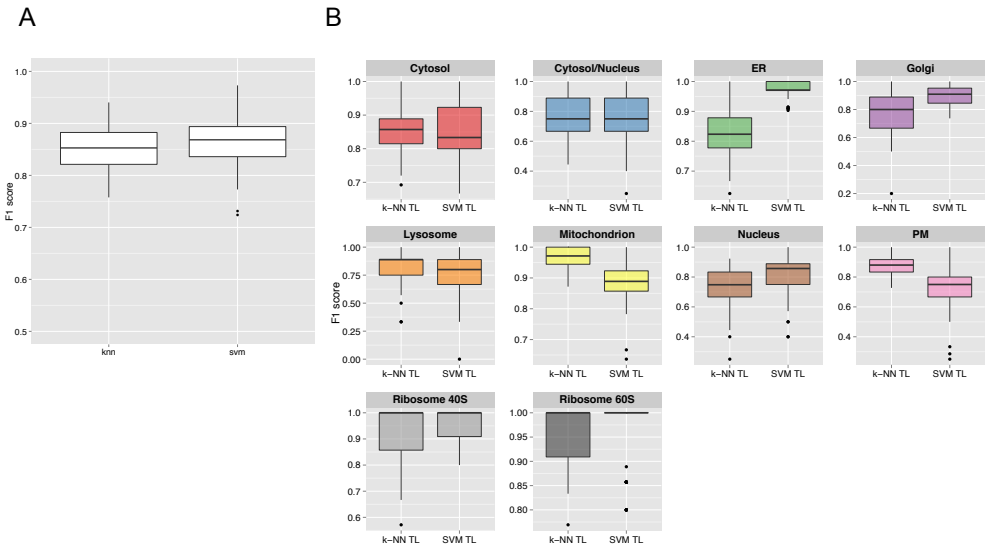
We compared the macro- and class-F1 scores for the k-NN TL and SVM TL methods on all datasets and found no single method outperformed the other, however, both methods outperformed using a classifier on one source alone. At the 0.01 significance level k-NN TL performed better overall than SVM TL for the callus dataset ($p = 4.365e^{-6}$) and SVM TL performs better than the k-NN TL method on the mouse dataset ($p = 6.298e^{-6}$). We also found that for the human, fly and roots datasets there was no significant difference in the performance between the two methods ($p = 0.0790$, $p = 0.396$, $p = 0.0108$, for each dataset respectively). Interestingly, the class-F1 scores showed that each TL method performed differently at the organelle level. For example, for the mouse dataset we found 4 of the 10 sub-cellular classes used in classifier creation performed significantly better with the k-NN TL, whereas another 4 of the sub-cellular classes performed better with SVM TL. For the 2 remaining classes both methods performed equally well. We see the same trend for all other datasets wherein no one method performs better on all sub-cellular classes than the other.



Supporting Fig. 12: Box plots displaying the macro-F1 (A) and class-F1 (B) scores for the k-NN transfer learning (TL) and SVM TL experiments over 100 test partitions on the E14TG2a mouse stem cell dataset.

	P value
40S Ribosome	4.196e-12
60S Ribosome	3.296e-07
Cytosol	3.016e-10
Endoplasmic reticulum	3.646e-05
Lysosome	2.726e-19
Mitochondrion	7.075e-10
Nucleus - Chromatin	1.186e-01
Nucleus - Nucleolus	2.734e-01
Plasma membrane	1.195e-04
Proteasome	6.258e-08

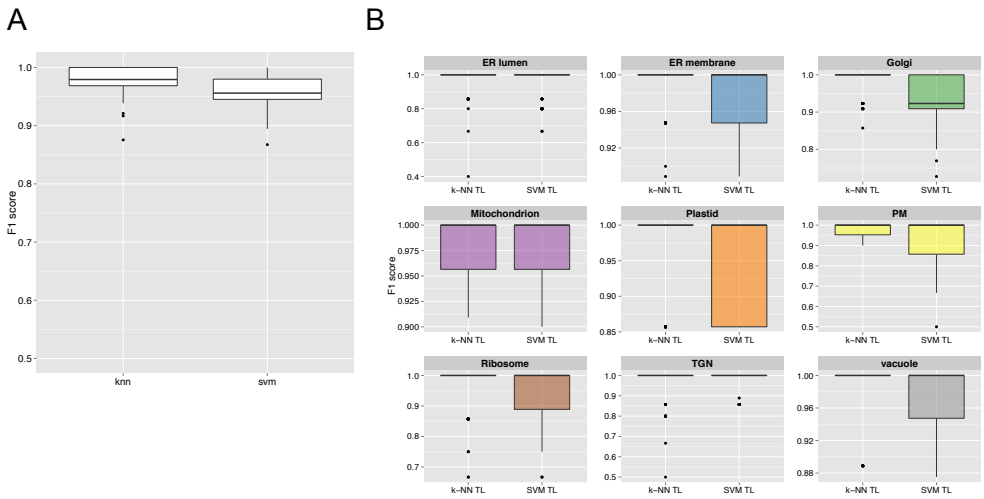
Supporting Table 10: P values from an unpaired two-sample t-test (with unequal variance) used to determine if the populations means between the k-NN TL and SVM TL methods are significantly different from one another for each sub-cellular class in the E14TG2a mouse stem cell dataset.



Supporting Fig. 13: Box plots displaying the macro-F1 (A) and class-F1 (B) scores for the k-NN transfer learning (TL) and SVM TL experiments over 100 test partitions on the human HEK293 LOPIT experiment

	p-value
Cytosol	6.779e-01
Cytosol/Nucleus	8.375e-01
ER	4.583e-32
Golgi	6.115e-13
Lysosome	3.461e-02
Mitochondrion	1.802e-21
Nucleus	1.507e-06
PM	6.383e-18
Ribosome 40S	4.961e-02
Ribosome 60S	9.155e-01

Supporting Table 11: P values from an unpaired two-sample t-test (with unequal variance) used to determine if the populations means between the k-NN TL and SVM TL methods are significantly different from one another for each sub-cellular class in the HEK283 human dataset.



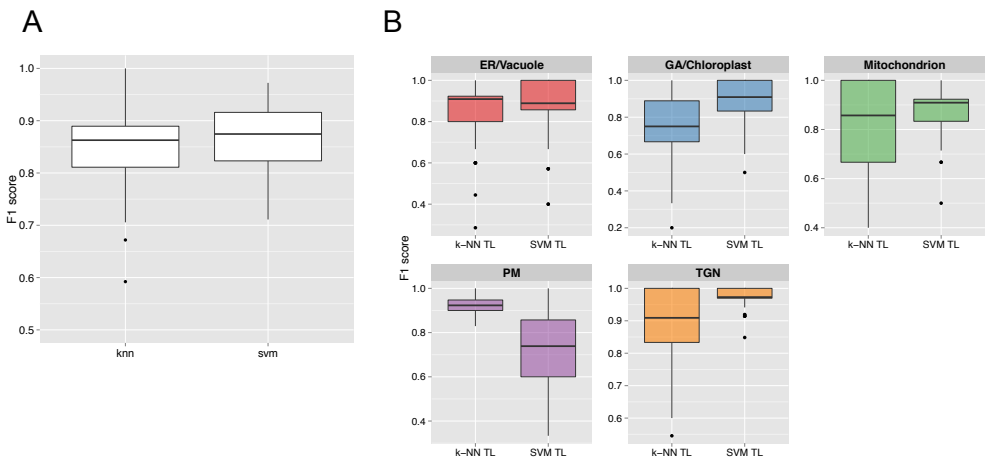
Supporting Fig. 14: Box plots displaying the macro-F1 (A) and class-F1 (B) scores for the k-NN transfer learning (TL) and SVM TL experiments over 100 test partitions on the *Arabidopsis thaliana* callus dataset.

	p-value
ER lumen	1.252e-01
ER membrane	2.830e-02
Golgi	3.693e-08
Mitochondrion	6.895e-02
Plastid	1.416e-02
PM	1.983e-05
Ribosome	4.760e-01
TGN	1.366e-01
vacuole	7.874e-02

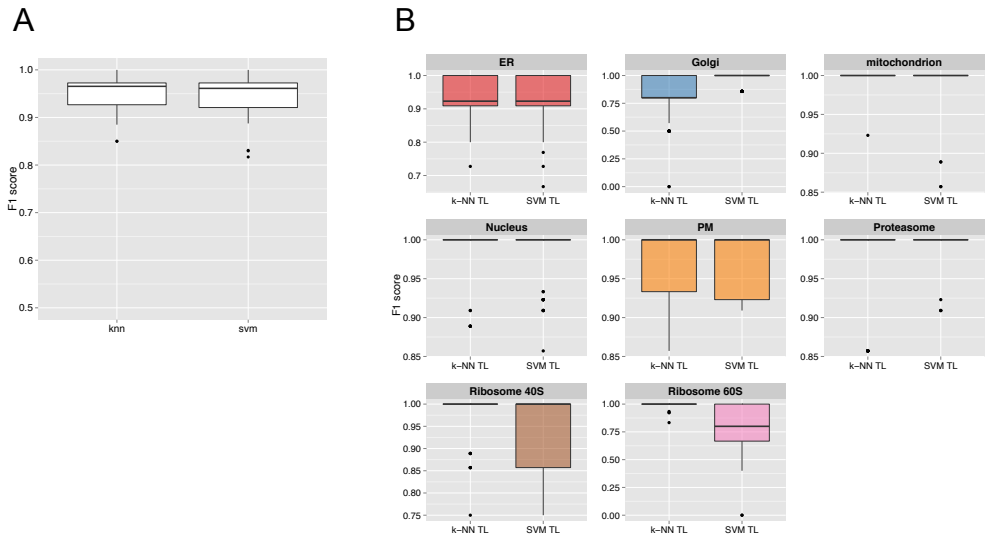
Supporting Table 12: P values from an unpaired two-sample t-test (with unequal variance) used to determine if the populations means between the k-NN TL and SVM TL methods are significantly different from one another for each sub-cellular class in the *Arabidopsis thaliana* proteome dataset.

	p-value
ER/Vacuole	1.200e-01
GA/Chloroplast	1.538e-13
Mitochondrion	6.697e-04
PM	9.863e-23
TGN	5.859e-13

Supporting Table 13: P values from an unpaired two-sample t-test (with unequal variance) used to determine if the populations means between the k-NN TL and SVM TL methods are significantly different from one another for each sub-cellular class in the *Arabidopsis thaliana* roots dataset.



Supporting Fig. 15: Box plots displaying the macro-F1 (A) and class-F1 (B) scores for the k-NN transfer learning (TL) and SVM TL experiments over 100 test partitions on the *Arabidopsis thaliana* roots dataset.



Supporting Fig. 16: Box plots displaying the macro-F1 (A) and class-F1 (B) scores for the k-NN transfer learning (TL) and SVM TL experiments over 100 test partitions on the *Drosophila melanogaster* dataset.

	p-value
ER	8.015e-01
Golgi	7.869e-15
mitochondrion	1.047e-01
Nucleus	1.466e-01
PM	6.400e-01
Proteasome	5.629e-05
Ribosome 40S	2.582e-05
Ribosome 60S	9.354e-16

Supporting Table 14: P values from an unpaired two-sample t-test (with unequal variance) used to determine if the populations means between the k-NN TL and SVM TL methods are significantly different from one another for each sub-cellular class in the *Drosophila melanogaster* datasets.