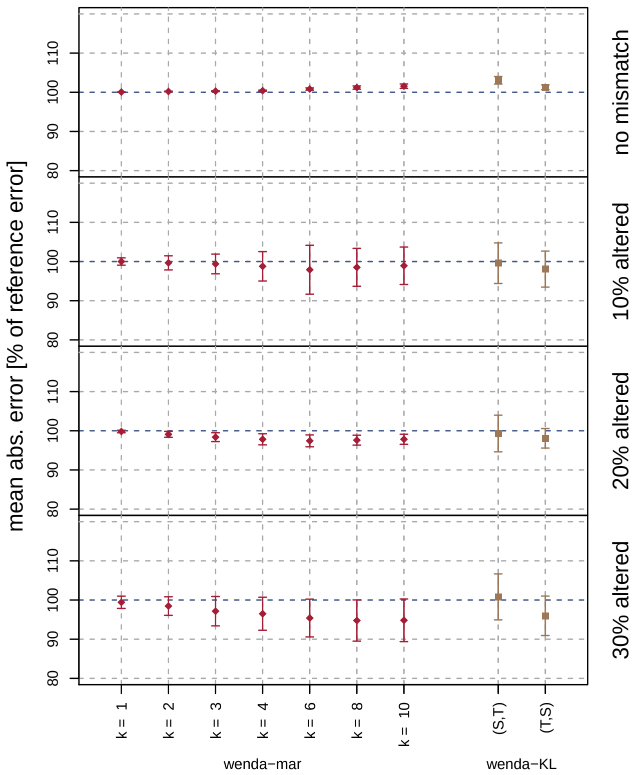
Weighted Elastic Net for Unsupervised Domain Adaptation with Application to Age Prediction from DNA Methylation Data

Supplementary Data

Lisa Handl, Adrin Jalali, Michael Scherer, Ralf Eggeling, Nico Pfeifer

Tissue in our data	**Mached tissue(s) in GTEx data**
Blood	WholeBlood
Brain	BrainAnteriorcingulatecortexBA24, BrainCaudatebasalganglia, BrainCerebellarHemisphere, BrainCerebellum, BrainCortex, BrainFrontalCortexBA9, BrainHippocampus, BrainHypothalamus, BrainNucleusaccumbensbasalganglia, BrainPutamenbasalganglia
Brain CRBM	BrainCerebellum, BrainCerebellarHemisphere
Brain Frontal	BrainCortex, BrainFrontalCortexBA9
Brain Frontal Cortex	BrainCortex, BrainFrontalCortexBA9
Brain Hippocampus	BrainHippocampus
Brain MedialFrontalCortex	BrainCortex, BrainFrontalCortexBA9, BrainAnteriorcingulatecortexBA24
Brain MidBrain	no match
Brain Occipital	BrainCortex
Brain Temporal	BrainCortex
Breast	BreastMammaryTissue
Buccal	no match
CD4+ cells	no match
Cord Blood	WholeBlood
Esophagus	EsophagusGastroesophagealJunction, EsophagusMucosa, EsophagusMuscularis
Fat	AdiposeSubcutaneous
Hair	no match
Kidney	no match
Liver	Liver
Lung	Lung
Menstrual Blood	WholeBlood
Muscle	MuscleSkeletal
Unknown (head and neck)	no match
Omentum	AdiposeVisceralOmentum
Pancreas	Pancreas
Saliva	no match
Spleen	Spleen
Vaginal Swab	Vagina
Whole Blood	WholeBlood

Supplementary Table 1. Table of how we matched tissues in our dataset with tissues in the data published by Aguet *et al.* (2017) to use the tissue similarities they reported as prior knowledge in *wenda-pn.*

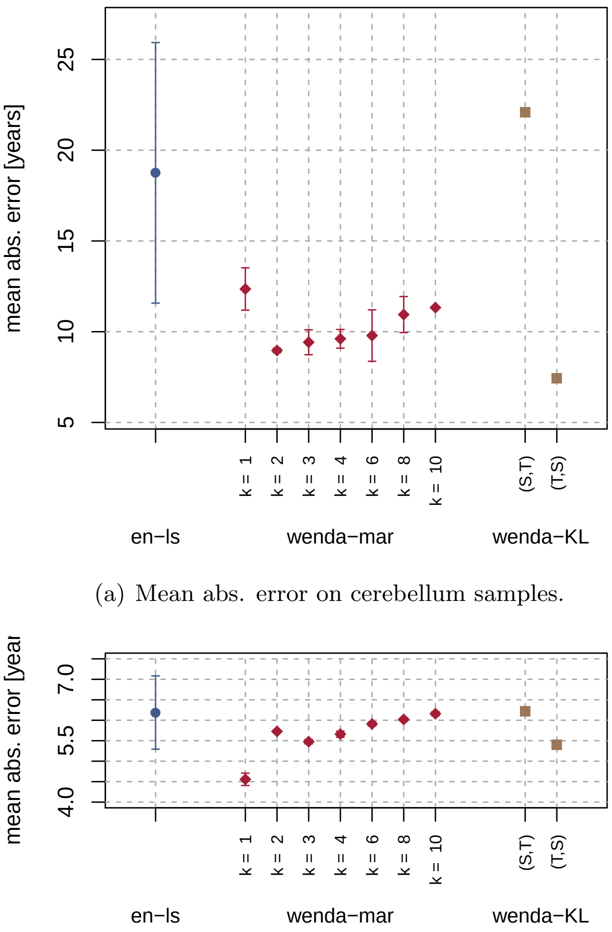


Supplementary Figure 1. As an alternative to the *wenda-mar* baseline, we also tested the KL-divergence $D_K(P||Q)$ between the discretized source and target distribution as possible feature weights (*wenda-KL*). Due to the asymmetry of the KL divergence, there are two variants, *wenda-KL*(S, T) and *wenda-KL*(T, S), depending on the order in which the discretized source (S) and target (T) distributions are compared.

This Figure shows the mean absolute errors of *wenda-mar* and *wenda-KL* on cerebellum samples (a) and on the full test set (b). We report all errors relative to the mean absolute error of *en* and show mean ± standard deviation over 10 runs of 10-fold cross-validation. In all simulation scenarios, both variants of *wenda-KL* perform similar or worse than *wenda-mar*.

Computing an empirical KL divergence for two samples of continuous variables requires the choice of a suitable discretization method. For the results presented here we chose a bin size based on the smaller sample (i.e., the target domain data) using the rule of thumb proposed by Sturges (1926) and applied it to the range of values of both samples.

The KL divergence is attractive from a theoretical perspective, but is not directly applicable as an alternative score in *wenda-pn* and *wenda-cv*, where we compare the value of each feature in a test sample to the conditional distribution (in the source domain) given the remaining features. This conditional distribution is predicted by g_f and is different for each test sample and feature, which means that we are repeatedly comparing a single value to a single predicted distribution.

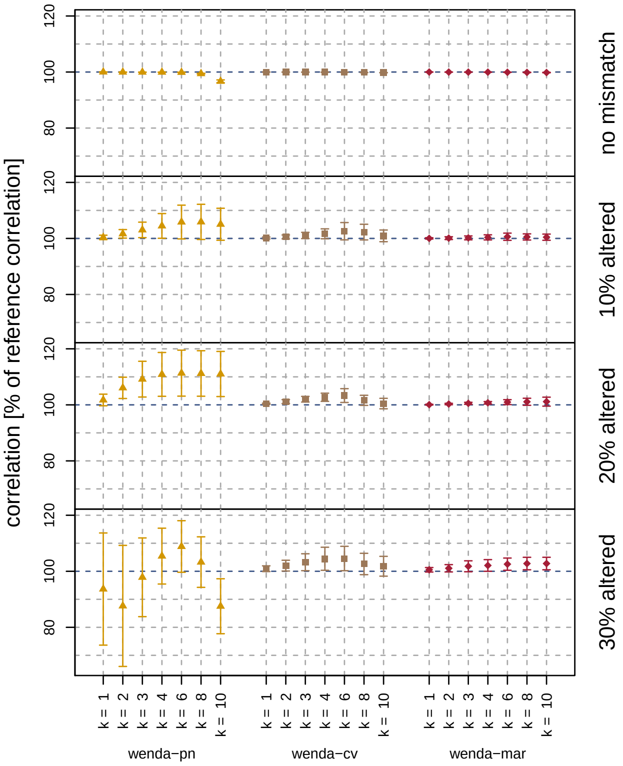


(a) Mean abs. error on cerebellum samples.

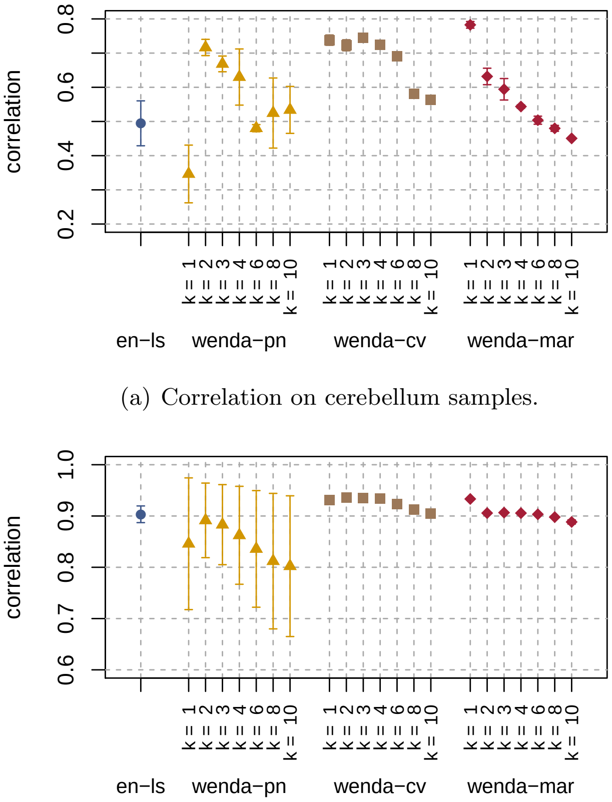
(b) Mean abs. error on the full test set.

Supplementary Figure 2. Comparison of the mean absolute errors of *en-ls*, *wenda-mar* and *wenda-KL* (see Supplementary Figure 1) on cerebellum samples (a) and on the full test set (b). Error bars indicate mean ± standard deviation over 10 runs of 10-fold cross-validation.

The results of *wenda-KL* differ dramatically between the two variants. On Cerebellum samples *wenda-KL*(T, S) performs surprisingly well, even slightly outperforming *wenda-mar*, but *wenda-KL*(S, T) performs substantially worse (in the range of *en-ls*). On the full set, both variants have a performance in a similar range as *wenda-mar*.



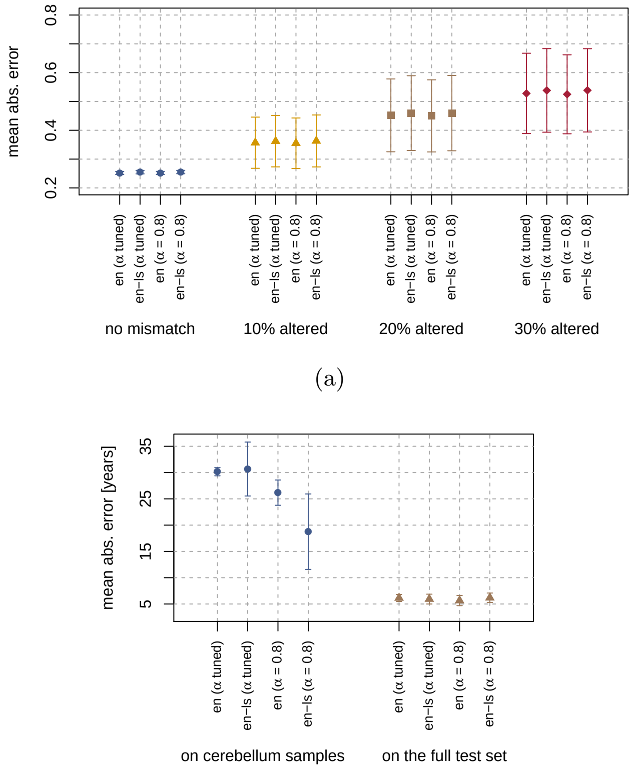
Supplementary Figure 3. As a second performance measure (in addition to mean absolute error), we report the correlation between true and predicted output. This figure shows correlations for *wenda-pn*, *wenda-cv* and *wenda-mar* on simulated test data. Each row shows results on one target domain (no mismatch, 10–30% altered variables). We report all correlations relative to the correlation of true output and predictions of *en*, showing the mean ± standard deviation over 10 simulations.



(a) Correlation on cerebellum samples.

(b) Correlation on the full test set.

Supplementary Figure 4. Correlation of predicted and true age for all models on cerebellum samples (a) and on the full test set (b). For *wenda-pn* we computed correlations only based on samples which were in the evaluation set. We show the mean and standard deviation over 10 runs of 10-fold cross-validation or, in case of *wenda-pn*, over all considered splits of the test tissues.



(a)

(b)

Supplementary Figure 5. Comparison of the mean absolute errors of multiple baseline models on simulated data (a) and on the DNA methylation data (b). The models included are *en* and *en-ls*, each with fixed $\alpha = 0.8$ and with α determined during cross-validation (in steps of 0.05). We show the mean and standard deviation over 10 simulations and 10 runs of 10-fold cross-validation, respectively.

On simulated data, all baseline models perform very similarly and tuning α during cross-validation does not lead to an improvement. This is consistent across all considered target domains. On the DNA methylation data, all baseline models show very similar performance on the full test set, but *en-ls* with $\alpha = 0.8$, which we use as reference in the main article, outperforms all other baseline models on cerebellum samples. It seems that even though tuning α in addition to λ may give the model slightly more flexibility to fit the training data well, it does not necessarily improve how the model generalizes to other domains.

On the DNA methylation data we found that only very small values of α were selected (either 0 or 0.05). This means that the resulting models were much closer (or even equal) to ridge regression than to LASSO and produced less sparse solutions. The sparsity of our baseline *en-ls* with $\alpha = 0.8$ might be the reason why it generalizes better. It could also explain why the subsequent least-squares fit of *en-ls* is beneficial for $\alpha = 0.8$, but not for tuned α. The final least-squares fit can help if many of noisy features were already excluded by the elastic net. If most features remain in the fit, removing the regularization penalty only increases variance.

References

Aguet, F., Brown, A. A., Castel, S. E., Davis, J. R., He, Y., *et al.* (2017). Genetic effects on gene expression across human tissues. *Nature*, **550**(7675), 204–213.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*. **21**(153), 65–66.