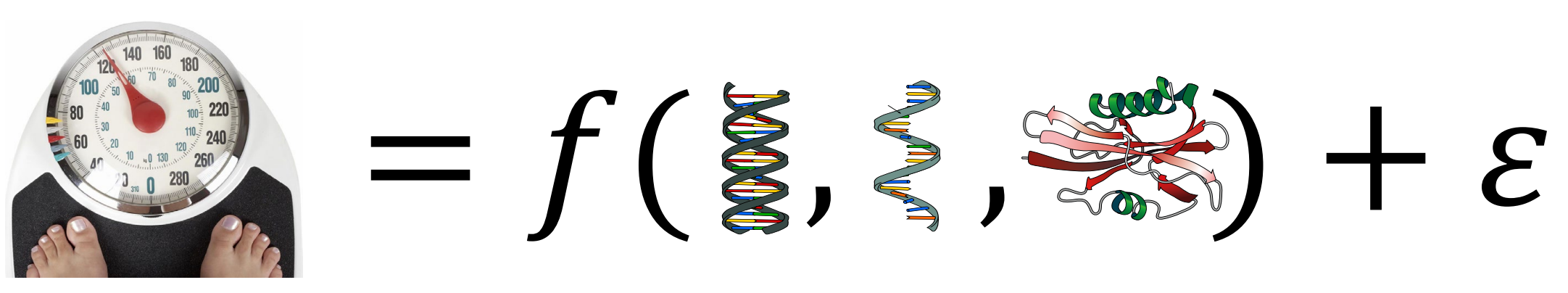
Integration of Multi-omics Data for Prediction of Metabolic Traits

Jelena Čuklina[1,2,3*], Yibo Wu[1], Evan G. Williams[1], María Rodríguez-Martínez[3], Ruedi Aebersold[1,4]

[1]ETH Zurich, Institute of Molecular Systems Biology, CH-8093 Zurich, Switzerland, [2]Ph.D. Program in Systems Biology, University of Zurich and ETH Zurich, CH-8057 Zurich, Switzerland,
[3]IBM Zurich Research Laboratory, Rüschlikon, CH-8803, Switzerland, [4]Faculty of Science, University of Zurich, Zurich, CH-8091, Switzerland

Motivation

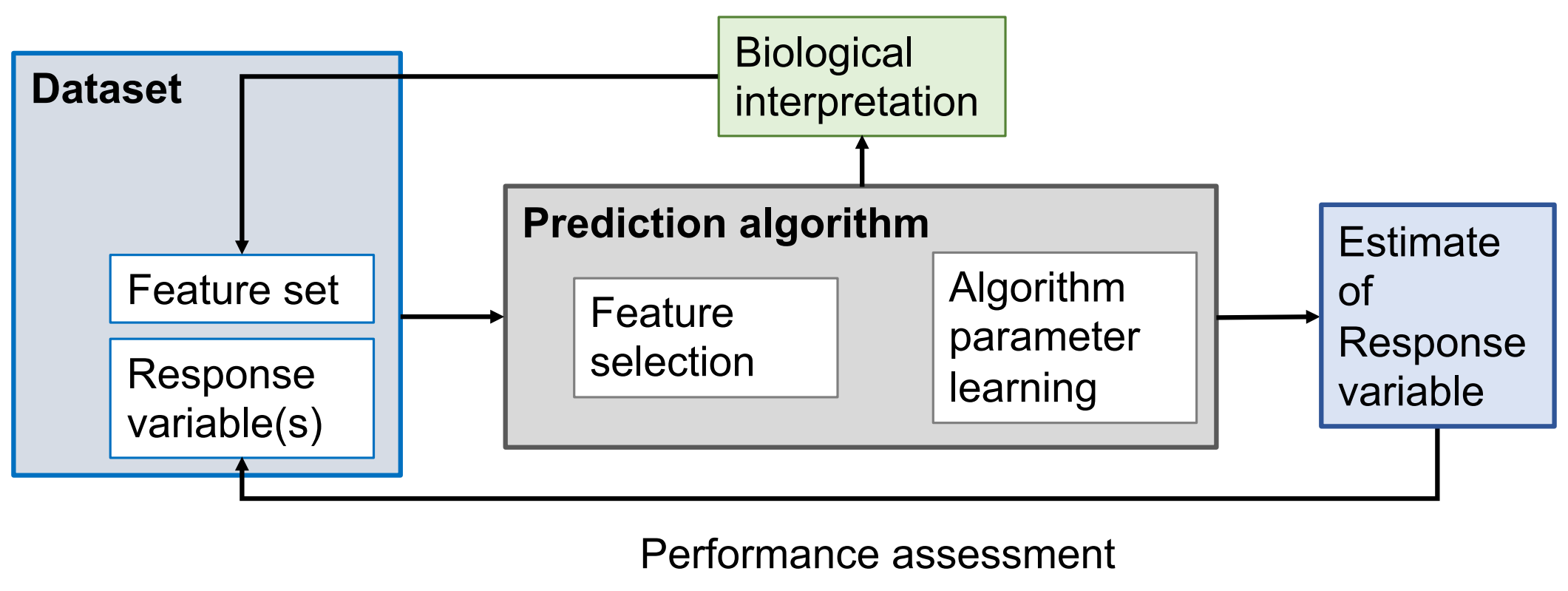
In biomarker research, the goal is to construct an prediction rule on the basis of a small number of predictors. Formally, this means representing a macro-level response as a function of molecular features (DNA variants, transcript or protein abundancies) with minimal error.



Aim

Develop a framework for selection of a composite biomarker: an ensemble of small number of predictors, that is able to predict the macro-level response.

Random Forest Algorithm



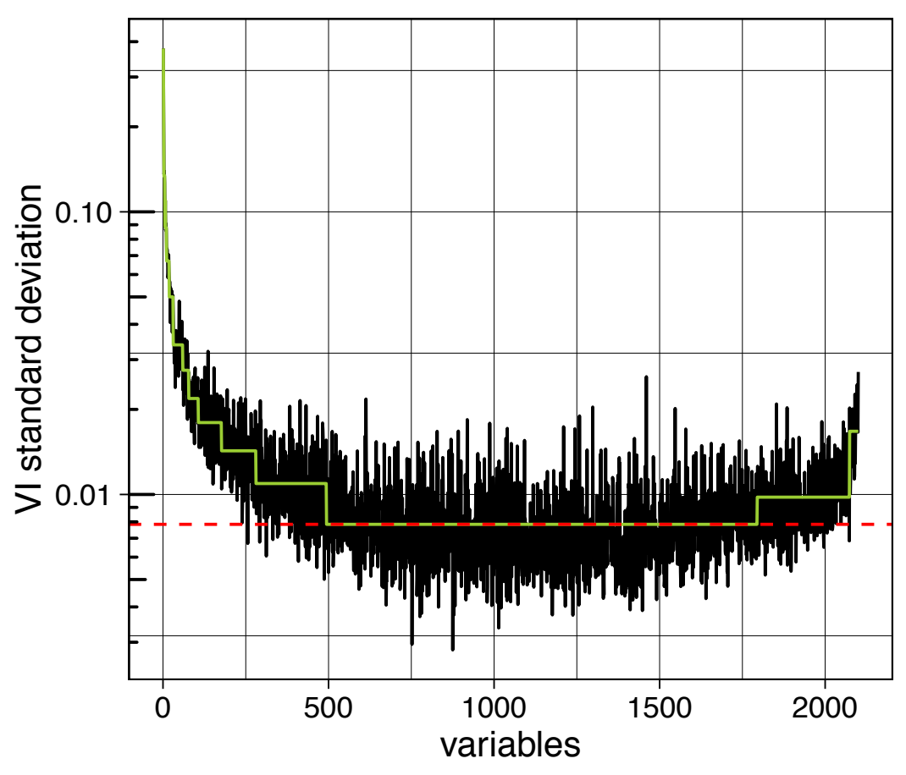
$$pseudoR^2 = 1 - \frac{MSE^*}{Var(y)} \approx Variance\ explained$$

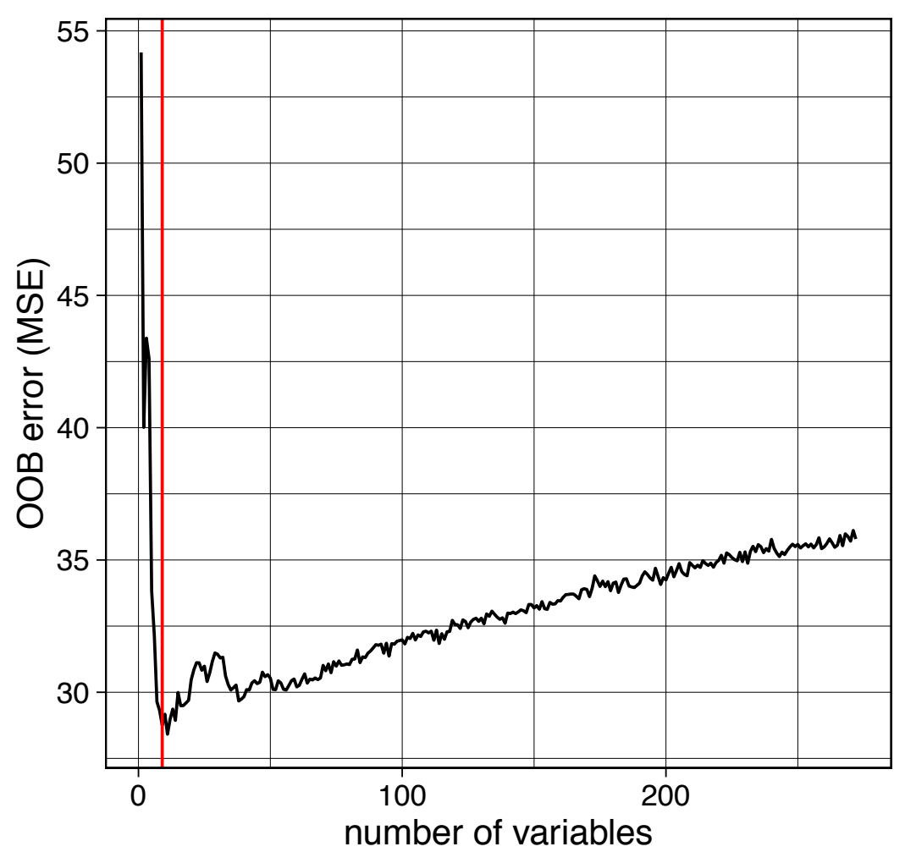
Random forest is an ensemble machine learning method, that constructs a multitude of decision trees [1]. Randomisation is achieved by using:
1) a random subset of features for split selection at each node;
2) a bootstrap of samples in each tree.
Variable importance (VI), used for feature selection is calculated as follows:

$$Importance = \sum_{trees} MSE_{variable\ permuted} - MSE$$

Variable selection

1. Important variables vary more in VI, so only variables with variation above **threshold** are maintained.
2. A set of nested models is constructed with a forward-selection procedure. The set of variables, that leads to minimum error is retained as the **interpretation** set.
3. A minimal-**predictive** variable set is determined by filtering out redundant variables from the interpretation set.

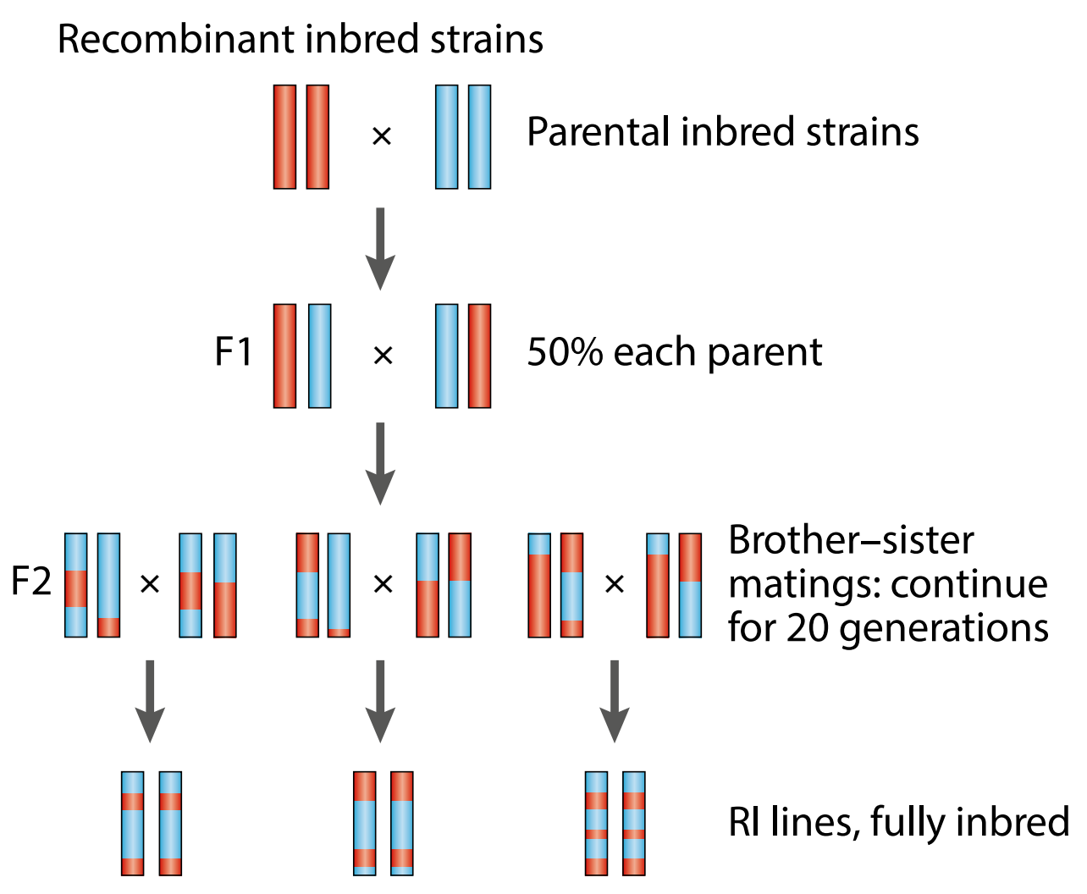




Data

To benchmark the process of construction of the composite biomarker, we use a mouse model. Mouse model has an advantage over human samples, as many confounding factors are controlled. Here we use measurements of 35 murine strains from the BXD recombinant inbred strain panel exposed to high-fat and chow diets.

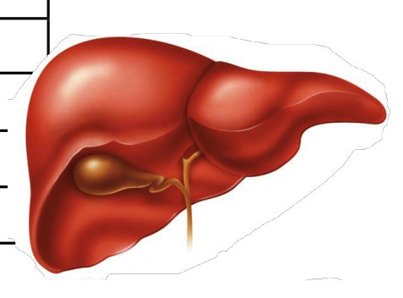
As explanatory variable set we use molecular profile of liver, and as response variables, we have selected 7 phenotypic traits related to metabolism.


Recombinant inbred strains
×
Parental inbred strains
F1
50% each parent
F2
Brother–sister matings: continue for 20 generations
RI lines, fully inbred

Molecular profiling

Profiling type	Number of features
Proteome	
(peptides)	22 227
(protein groups)	3 090
(proteins matching transcripts)	2100
Transcriptome	25 135
Genome loci	3 811

Table 1. Mice multi-omics profiling. Here we show preliminary results from proteomic level only



Trait selection

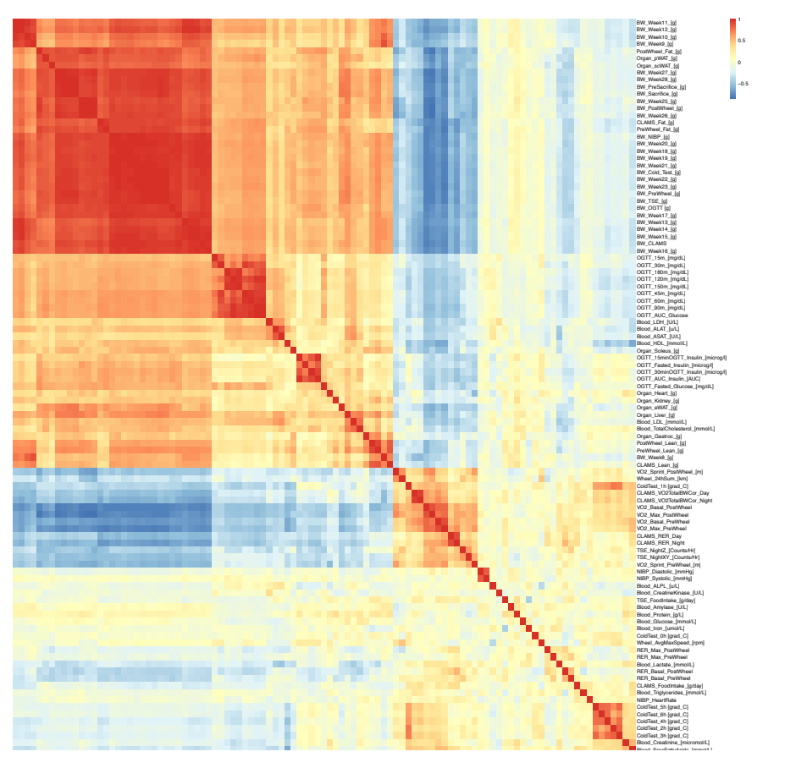
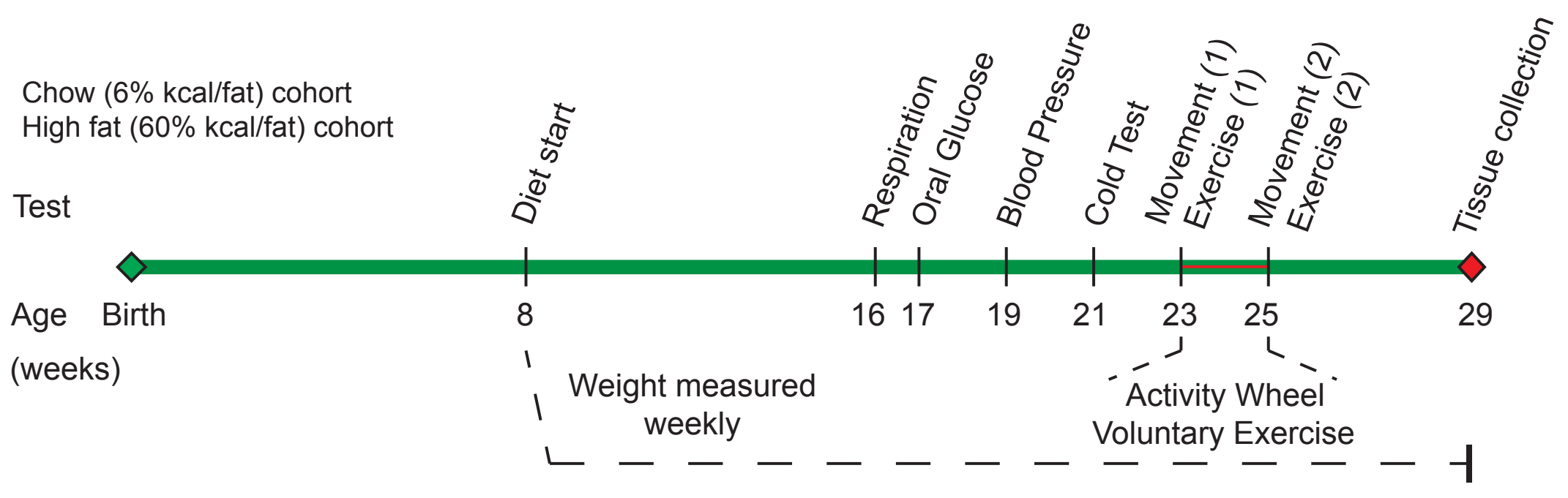


Table 2. Seven traits were selected for benchmarking process. Out of total 115 measured traits, whose correlation is shown on the left, we selected traits, representative of each hifly correlatied group.

Trait	Alias
BW_Week22_.g.	Body_Weight
Organ_pWAT_.g.	Fat
CLAMS_Lean_.g.	Lean_mass
ColdTest_5h..grad_C.	Cold_test
OGTT_AUC_Insulin_.AUC.	Insulin
OGTT_AUC_Glucose	Glucose
VO2_Sprint_PostWheel_.m.	Respiration



Results

Algorithm parameter choice

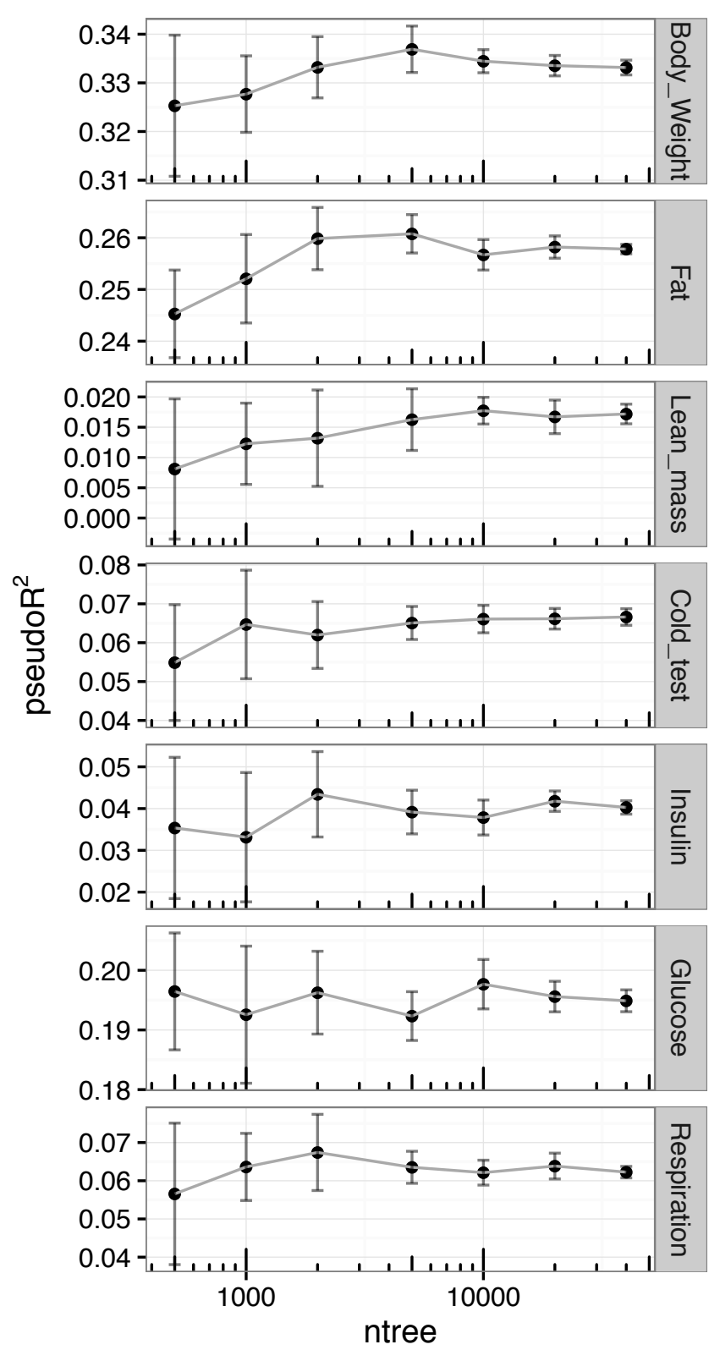


Fig 1. Random forest performance generally grows with increased number of trees and stability of prediction improves

Random Forest parameters:
- **ntree** (number of trees)
- **mtry** (number of features, tried at each node)

Higher ntree improves:
1. stability of prediction
2. performance (Fig. 1)
3. stability of variable importance list (Fig.2).

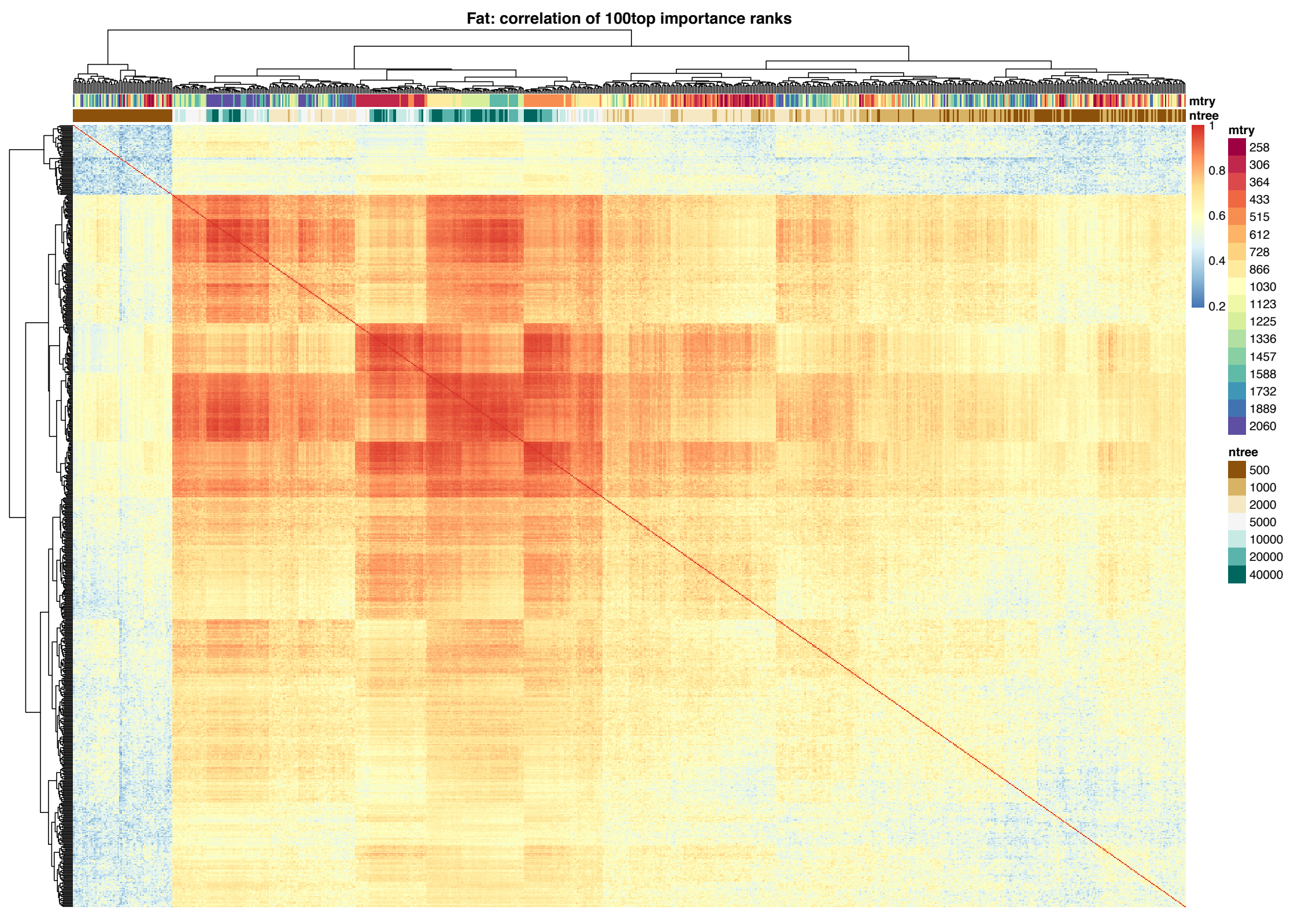
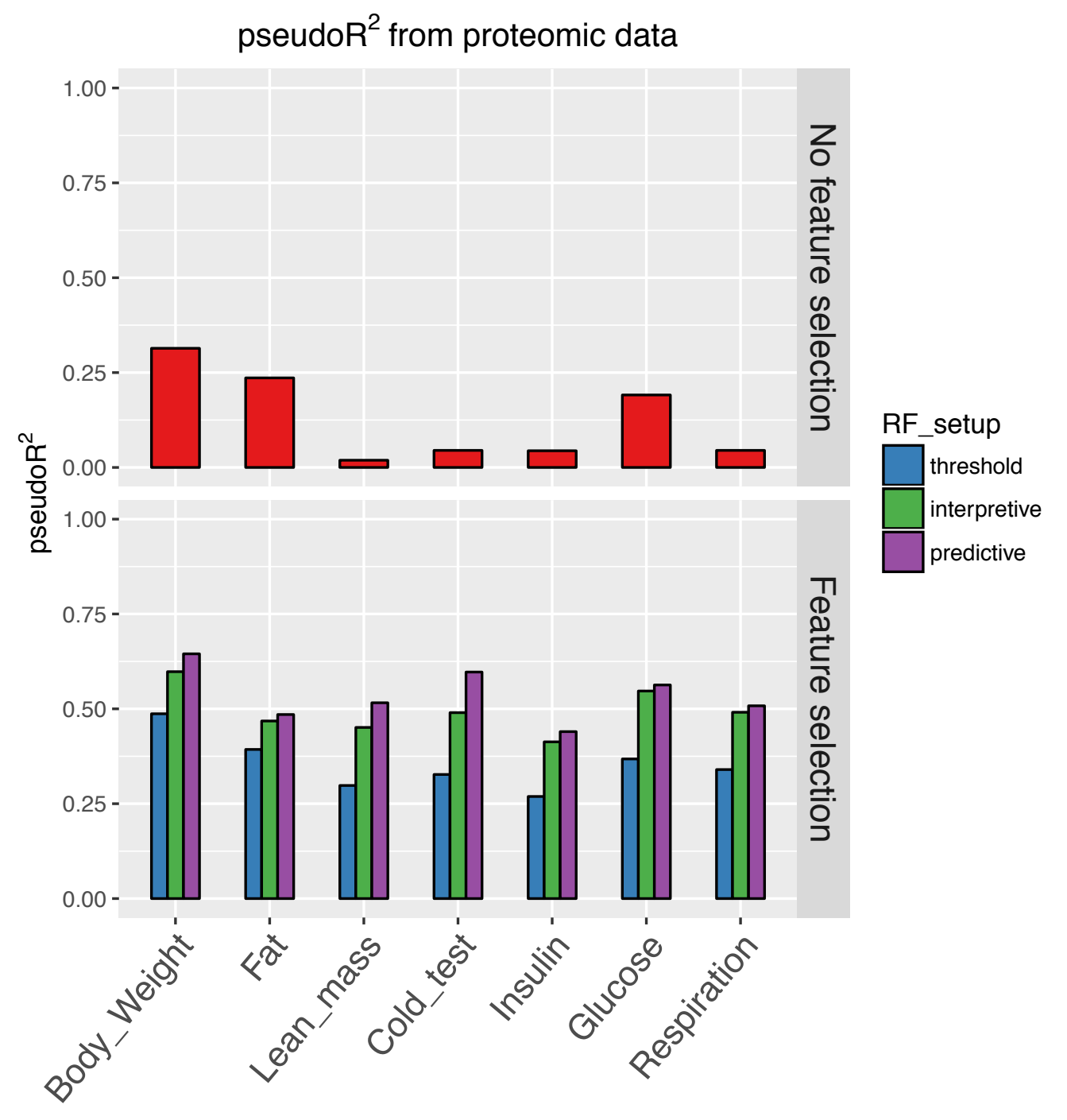


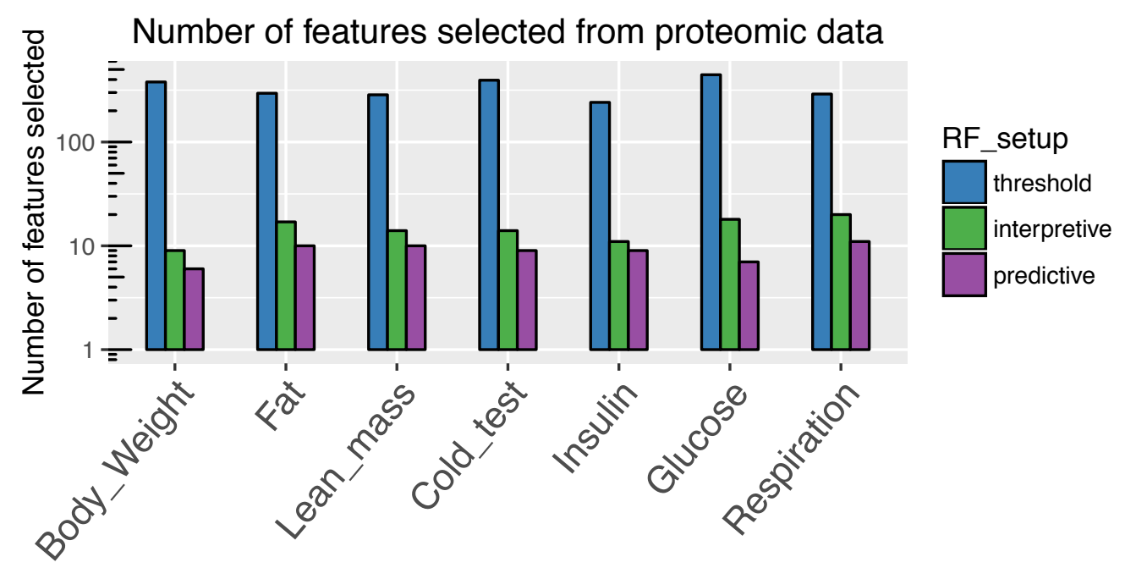
Fig 2. Correlation heatmap of variable importance rank for top 100 variables. Each ntree & mtry combination was repeated 10 times, the resulting variable ranks were compared for each setup by correlating ranks. Default forest size (500 trees) produces different importance ranking for each realisation: the resulting ranks don't correlate neither within 500-tree forests, nor with the ranks of bigger forests. However, forests of 5000-40000 of trees yield ranks that are more similar to each other. Thus, to get a reliable importance-ordered list of variables, bigger forest size is necessary. Mtry influence is lower. However, for mtry value, close to default (1/3 of variables, here 1030), a middle-size forest of 5000 trees yields a variable list, very similar to list of 20000-40000 trees.

Algorithm Performance



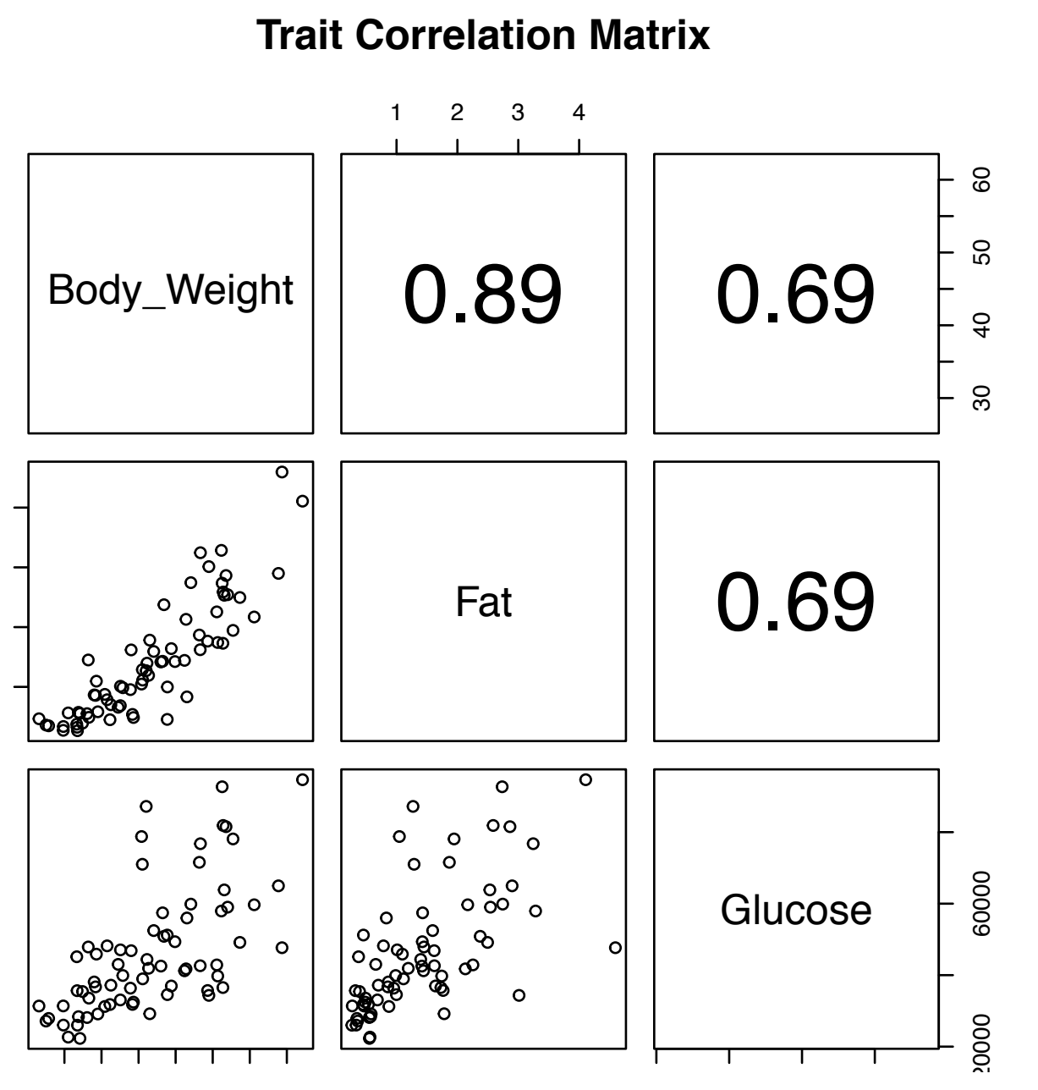
Random Forest explains up to 30% of variance with default parameters. Feature selection improves prediction substantially (by 21-53%).

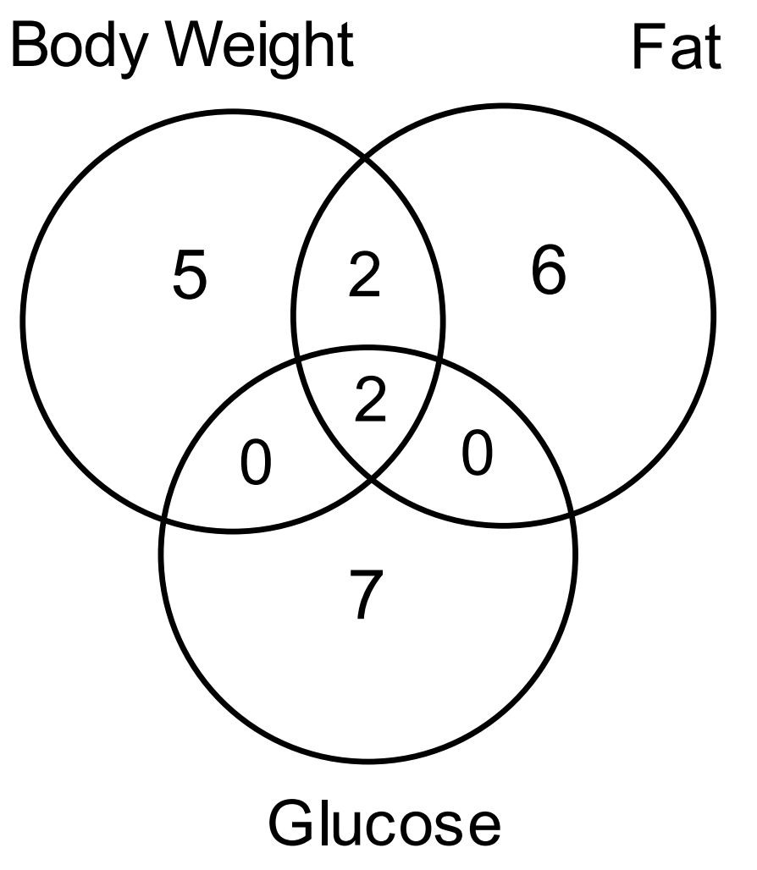
Feature selection



Surprisingly few features are required to achieve better prediction.

For highly correlated features, also predictors selected are shared (see Venn diagram)





References

1. Breiman L. Random Forests. Machine Learning.45(1):5-32.
2. Genuer R, Poggi JM, Tuleau-Malot C. Variable selection using random forests. Pattern Recogn Lett. 2010;31(14):2225-36.
3. Williams EG, Wu Y, Jha P, Dubuis S, Blattmann P, Argmann CA, et al. Systems proteomics of liver mitochondria function. Science. 2016;352(6291):aad0189.