Machine Learning in Multi-Omics Data to Assess Longitudinal Predictors of Glycaemic Trait Levels

Laurie Prélot[1], Harmen Draisma[1], Mila D. Anasanti[1], Zhanna Balkhiyarova[1], Matthias Wielscher[2], Loic Yengo[3], Sylvain Sebert[4], Mika Ala-Korpela[5,6,7,8,9,10], Philippe Froguel[1,11], Marjo-Riitta Jarvelin[2,4,12,13], Marika Kaakinen[1,14], Inga Prokopenko[1]

[1]Section of Genomics of Common Disease, Department of Medicine, Imperial College London, London, United Kingdom;
[2]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom;
[3]Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia;
[4]Center for Life Course Health Research, University of Oulu, Oulu, Finland
[5]Systems Epidemiology, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia;
[6]Population Health Science, Bristol Medical School, University of Bristol, Bristol, UK;
[7]Medical Research Council Integrative Epidemiology Unit at the University of Bristol, Bristol, UK;
[8]Computational Medicine, Faculty of Medicine, University of Oulu and Biocenter Oulu, Oulu, Finland;
[9]NMR Metabolomics Laboratory, School of Pharmacy, University of Eastern Finland, Kuopio, Finland;
[10]Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Faculty of Medicine, Nursing and Health Sciences, The Alfred Hospital, Monash University, Melbourne, Victoria, Australia;
[11]CNRS, Pasteur Institute of Lille, University of Lille, Lille, France;
[12]Oulu University Hospital, Oulu, Finland;
[13]Department of Life Sciences, College of Health and Life Sciences, Brunel University London, Uxbridge, United Kingdom;
[14]Centre for Pharmacology and Therapeutics, Department of Medicine, Imperial College London, London, United Kingdom.

Corresponding author:
Inga Prokopenko, i.prokopenko@imperial.ac.uk

Abstract

Type 2 diabetes (T2D) is a global health burden that will benefit from personalised risk prediction. We aimed to identify longitudinal predictors of glycaemic traits relevant for T2D by applying machine learning (ML) to multi-omics data from the Northern Finland Birth Cohort 1966 at 31 (T1) and 46 (T2) years old. We predicted fasting glucose/insulin (FG/FI), glycated haemoglobin (HbA1c) and 2-hour glucose/insulin from oral glucose tolerance test (2hGlu/2hIns) at T2 in 595 individuals from 1,010 variables at T1 and T2: body-mass-index (BMI), waist-hip-ratio, sex; nine blood plasma measurements; 454 NMR-based metabolites (228 at T1 and 226 at T2); 542 methylation probes established for BMI/FG/FI/HbA1c/T2D/2hGlu/2hIns (277 at T1 and 264 at T2). Metabolic and methylation data were used in their raw form (Mb-R, Mh-R) or in scores (Mb-S, Mh-S). We used six ML approaches: random forest (RF), boosted trees (BT) and support vector regression (SVR) with the kernels of linear/linear with L2 regularization/polynomial/radial-basis function. RF and BT showed consistent performance while most SVRs struggled with high-dimensional data. The predictions worked best for FG and FI (average R^2 values of six ML models: 0.47 and 0.30 for Mb-S). With Mb-S/Mb-R data, sex, branched-chain and aromatic amino acids, HDL-cholesterol, VLDL, glycoprotein acetyls, glycerol, ketone bodies at T2 and measurements of obesity already at T1 were amongst the top predictors. Addition of methylation data, did not improve the predictions (P>0.3, model comparison); however, 15/17 markers were amongst the top 25 predictors of FI/FG when using Mb-S+Mh-R data. With ML we could narrow down hundreds of variables into a clinically relevant set of predictors and demonstrate the importance of longitudinal changes in prediction.

Key Words

Glycaemic traits, longitudinal, machine learning, metabolomics, methylation, prediction, type 2 diabetes

Introduction

Diabetes accounts for the yearly deaths of about four million people between 20 and 79 years old (2017) world-wide. Prevalence of diabetes is expected to increase from 8.8% to 9.9% by 2045. The main challenge of diabetes health care is its growing burden in low and middle income countries. Besides, glucose tolerance impairment is progressing in young individuals, leading to high risk of developing Type 2 diabetes (T2D) later in life[1]. T2D is defined by insulin resistance and deficiency of insulin secretion in the pancreas. The T2D diagnosis criteria encompass fasting plasma glucose (FG) levels≥7.0mmol/L or two-hours post-prandial plasma glucose (2hGluc)≥11.1mmol/L or glycated Hemoglobin (HbA1c) ≥48mmol/L. HbA1c is an indicator for long-term control of glycemic state in diabetes patients.

Classical risk factors for T2D encompass sex, age, obesity, family history, hypertension, lifestyle factors, and are sometimes extended to cholesterol and blood pressure levels. Recent advances in omics technologies have allowed to explore diabetes risk factors in more detail. Large genome-wide association studies (GWAS) have estimated that genetics accounts for <15 % of diabetes heritability[2]. Currently, 128 distinct signals at 113 loci associated with T2D have been reported by GWAS meta-analyses[3,4]. GWAS also unveiled DNA loci associated with quantitative glycaemic traits in individuals without diabetes, including FG[5], fasting insulin (FI)[5], FG adjusted for body-mass-index (BMI)[5], FI adjusted for BMI[5], 2hGluc[6] and HbA1c[7]. These studies are based on the hypothesis that quantitative glycaemic traits may reflect mechanisms involved in diabetes pathogenesis. A weak correlation between genetic effects on glycaemic traits and T2D has been found by previous studies; however, the overlap may shine light on the mechanisms influencing glucose homeostasis and its dysregulation in diabetes[8].

Environment and lifestyle are likely to contribute to a large part of the T2D onset[9]. Environmental cues can affect gene expression by addition of a methyl group on a CpG-dinucleotides sites of DNA. This is called DNA methylation and is the most widely studied type of epigenetic modification. Studies in peripheral blood have found a mean absolute difference of 0.5%-1.1% in methylation levels between individuals with and without T2D[10]. Epigenome-wide association studies reported associations at 65 methylation markers for T2D[10,11] and provided support for overlap in epigenetic effects between T2D and glycaemic traits[11,12]. The epigenetic effect on glycaemic traits was smaller upon BMI adjustment[12]. The investigation of the link between BMI and methylation levels demonstrates that methylation at the majority of CpG sites in blood is consequential to higher BMI[13]. Weighted methylation risk scores have stronger contribution to incident T2D than traditional risk factors including overweight, obesity, central obesity, impaired fasting glucose and hyperinsulinaemia[10,13].

The metabolomic alterations underlying T2D have also emerged as a promising area of investigation. A large number of metabolites, including amino acids, especially branched-chain amino acids (BRACA) and aromatic amino acids, fatty acids, glycerophospholipids, ketone bodies and mannose have been associated with T2D incidence[14–16] and are risk factors. However, whether metabolites can be effective and reliable T2D predictors, remains unclear. From 2011, T2D prediction has been investigated and a small number of metabolic

compounds, which overlapped with the ones identified in association analyses, were highlighted[17–20].

Recently, the studies of T2D risk leveraged association analyses as well as machine learning algorithms (ML) for prediction of binary T2D phenotypes. ML used so far for T2D classification include: logistic regression with and without Lasso regularization[10,13,18,19,21–23], Regularized least-squares (RLS)[20], Cox regression[22], naive Bayes[18] and J48-decision tree[18]. Machine learning, unlike traditional statistical modelling, aims at optimizing the parameters of a model rather than estimating parameters from a given distribution. "Supervised" machine learning is designed to make a prediction about classes or continuous values of unseen data points ("target values"/"labels"/"outcomes"), based on training on an example dataset ("variables"/"predictors"). The strength of ML relies in the optimization process as well as the ability to use a relatively high-dimensional set of predictor variables without knowledge of the joint distribution of these variables[24].

In predictive studies using ML models classical risk factors, genetic risk scores (GRS)[21], methylation risk scores (MRS)[10,13] and metabolomic data[18–20,23], have been used as predictors of T2D incidence after a follow-up window of two to fourteen years. A few studies have suggested that metabolites improve prediction performance[18–20,22], while others have reported negligible or no improvement in prediction[23]. GRS have been shown to bring no incremental value over classical non-invasive factors and metabolic markers[21]. MRS combining CpG loci have been found to be associated with future type 2 diabetes incidence[10,13].

In this study we aimed to identify longitudinal predictors of glycaemic traits relevant for T2D by applying machine learning approaches to multi-omics data. We focused on epigenetic and metabolomic markers, from the Northern Finland Birth Cohort 1966 (NFBC1966), at 31 (T1) and 46 (T2) years for prediction of HbA1c, FG, 2hGluc, FI and 2hIns (two-hours insulin) at T2. We implement and compare three machine learning approaches: Boosted trees (BT), Random forest (RF) and support vector regression (SVR) with different combination of input variables.

Methods

Study Population

NFBC1966 is a birth cohort (N=12,058 births in 1966) with participants from northern Finland[25]. From the medical examination at 31 (T1, N=6,007) and 46 years (T2, N=5,861) we included participants with demographic, medication, epidemiological, blood biochemistry, metabolomics and epigenetic information available at both time points (N=626). Consent was obtained and the study was approved by the ethical committees of the University of Oulu and Imperial College London (Approval:18IC4421).

Data Collection

The biochemical assays[26,27], oral glucose[28], and HbA1c measurements[29] are detailed elsewhere. Metabolites were quantified by a high-throughput serum nuclear magnetic resonance (NMR) platform[30–33]. DNA methylation was measured in whole blood from samples from 807 randomly selected individuals after overnight fasting. IlluminaInfiniumHumanMethylation450 Beadchip and EPIC arrays were used at T1 and T2,

respectively. Methylation data was pre-processed on genome build CGCh37/hg19. Prior to any analysis, 25 individuals pregnant at T1 were excluded.

Quality Control and Imputation of Epidemiological, Blood Biochemistry and Metabolomics Data

Imputation of epidemiological, biochemical and metabolomics variables was performed jointly with random forest (MisForest in R[34]). Type 1 diabetes (T1D), T2D, gender, blood pressure, lipid and diabetes medication were included as factors. Each gender was imputed separately. Post-imputation, FI was log transformed to reduce skewness. All measures of fasting/post-prandial glucose and insulin at T2 were removed from the set of predictor variables. Pyruvate, which exhibited a high correlation with glucose, was removed as well.

Quality Control and Imputation of Epigenomic Data

For methylation array data, duplicate samples (N=9 at T1, N=8 at T2), gender mismatches (N=7/1 at T1/T2), and samples with <95% call rate (a detection *P*-value threshold of $P<10^{-16}$) (N=67/40 at T1/T2) were removed. Probes with <95% call rate were removed (N=14,486/14,586 at T1/T2). Intensity values were normalized with subset quantile normalization within array (SWAN in Minfi R[34]), and beta values were computed from methylated and unmethylated normalized probes intensities. Probes further than 4SD from the mean were removed. Batch and sex effect were corrected by including the principal components of the control probes intensities and the gender as linear predictors in the regression analysis of the samples[35]. Blood cell composition was corrected by using the Houseman estimates[36] of blood cell type in the regression. Imputation of methylation data was performed using the methylation residuals corrected for sex, and blood cell type.

Study Variables

HbA1c, 2hGluc, 2hIns, FG, FI levels were used as continuous outcomes to predict. A total of 1,010 variables from T1 and T2 were used as predictors in the current study. "Metabolic" predictors included: epidemiological data - sex, measures of obesity (BMI and waist-to-hip ratio), biochemical data - nine blood measurements of triglycerides, total cholesterol, high and low-density lipoprotein cholesterol (HDL-C and LDL-C), metabolomics data - 454 metabolites (228 at T1 and 226 at T2). The methylation dataset included 541 probes, including 277 at T1 and 264 at T2. The selection of methylation probes was based on previous association of the probes with seven phenotypes: 187 probes associated with BMI[13], 21 with FG[11], 11 with HbA1c[11] and 67 with T2D[11], one with 2hGluc[12], eight with FI[12], 21 with 2hIns[12]. Thirty-one of the 626 individuals, who had missingness rate≥50% in the selected methylation probes, were excluded. In total, 595 individuals were included in ML analysis.

Association Analysis of Selected Methylation Probes with continuous phenotypes

The association between selected methylation probes and continuous phenotypes was assessed in the NFBC using regression analysis, including relationships between T1-methylation and T1-phenotype (FG, FI or BMI available), T1-methylation and T2-phenotype (BMI, FG, HbA1c, T2D, 2hGluc, FI or 2hIns); and T2-methylation and T2-phenotype (**Table 1**).

For the ML analysis, we used all probes to compute scores in order to get the most powerful scores possible.

Predictors Combination and Prediction Frameworks

Metabolic (Mb) and Methylation (Mh) data were combined under their raw (R) form or transformed into scores (S) (see below). The following combinations were tested: Mb-R/ Mb-S/ Mh-R / Mh-S/ Mb-R + Mh-R/ Mb-R + Mh-S/ Mb-S + Mh-R/ Mb-S + Mh-S (**Figure 1**). Methylation and Metabolic data were either adjusted for BMI and waist-hip-ratio at T1 and T2, or kept unadjusted.

Calculation of The Scores

Unweighted methylation risk scores were used. Scores at T1 and T2 were based on the established associations with seven phenotypes, including BMI, FG, HbA1c, T2D, 2hGluc, FI, 2hIns. Metabolic risk scores at T1 and T2 grouped variables based on the following biochemical classes: lipoparticules, lipids, blood proteins, carbohydrates and insulin, keton-bodies, BRACA, other amino acids.

ML Approaches

Three ML methods were used for regression analysis: Boosted trees (BT), Random Forest (RF) and Support Vector Regression (SVR). SVR was implemented with Linear Kernel with and without L2 regularization (SVR-Linear, SVR-L2Linear, respectively), with Polynomial Kernel (SVR-Polynomial) and with Radial Basis function Kernel (SVR-RBF) (**Figure 1**). The algorithms were chosen for their ability to handle a large number of predictors, to account for non-linear relationships, the absence of the assumption regarding data distribution, and for their computational times.

The decision tree unit (in BT and RF) is a hierarchical framework. At each step the sample is split based on some threshold in one of the variables (feature). At each level, the feature examined must lead to the best possible prediction at the “leaves” level. Overfitting is avoided by a stopping criterion and recursive pruning of the tree. “Boosting” methods seek to construct iteratively predictors (e.g. trees, in BT) by focusing mis-predicted examples at the previous step[24]. Random forests are an ensemble of decision trees. Each tree is grown on a bootstrapped sample of the data. The subset of features examined is generated randomly. The final prediction is based on the voting majority or averaging the predictions[24]. Support vector regression, aims at fitting the data in a “regression” hyperplane by minimizing a margin. Linear separation is made possible by mapping the data to a high dimensional space via a Kernel function[24].

Optimization of the ML Algorithms

Nested cross validation was implemented. The data set was split into a training (80%) and testing set (20%) with a 5-fold cross validation. The performance of the ML models was estimated on the testing set, while parameter tuning was implemented on the training set by splitting it further into a 5-fold cross validation (nested). Random search method was used to find the model parameter combination which minimized the error of the model. The Root Mean Square Error (RMSE) was used to assess model performance during training. Both Rsquared (R^2) and RMSE were computed in the testing set to estimate performance. The

packages KernelLab, LibLinear, RandomForest and Xboost in R[34] were used with Caret as a wrapper.

Variable Importance in the ML Models

In boosted trees, the information gain was used as a measure of importance. Gain is based on the decrease in entropy after a dataset is split on a feature j at a branch of the tree. Random forest variables were ranked with the Increase in Mean Square Error (MSE). It estimates the increase of prediction error when the values of the feature j are randomly permuted. For SVR, each feature is evaluated based on its independent association with the outcome. The slope of the regression is used to rank the features.

Statistical Analysis

The performance of each model was computed as the average R^2 over the 5 testing folds of the cross validation. In result section, we report the R^2 pooled for the six ML algorithms. Comparison of the models was performed with a one-way ANOVA and post-hoc Tukey HSD test.

Results

We analyzed metabolic and methylation data from 595 individuals to predict the levels of five glycaemic traits: HbA1c, FG, 2hGluc, FI and 2hIns at T2 from metabolic and methylation variables at T1 and T2 (**Figure 1**).

ML Model Performance and Input Dataset

We first evaluated the performance of all ML models for each of the datatypes. We found that models with metabolic data ("Mb-R and Mb-S") had a performance that reached a maximum R^2 of 0.47 R^2 (**Figure 2a**). In contrast, models with methylation data only ("Mh-R and Mh-S"), reached up to 0.12 R^2. Thus, metabolic models performed significantly better than epigenomic models ($P\text{-value}_{TukeyHSD}<5\times10^{-8}$) (**Table 2**, Comparison 1-4). Then we assessed the performance of models combining both data types. We did not observe any significant difference ($P\text{-value}_{TukeyHSD}>0.30$) between "Mb-R" and "Mh-R + Mb-R, "Mb-R" and "Mh-S + Mb-R" (**Table 2**, Comparison 5-6). This result suggested that addition of methylation information does not increase the predictive ability of tested models. Next, we explored the effect of variables transformation to scores (**Table 2**, comparison 7-8). The Metabolic model "Mb-S" performed significantly better than "Mb-R" in the prediction of HbA1c, FG, 2hGluc ($P\text{-value}_{TukeyHSD}<0.050$) but showed no difference in the case of FI and 2hIns prediction, whereas the methylation model "Mh-R" performed better than "Mh-S" ($P\text{-value}_{TukeyHSD}<0.050$) for FI and 2hGluc and exhibited no difference for prediction of the other phenotypes. Therefore, our findings do not allow us to generalize over the power "scored" data compared to "raw" data. Finally, we found that "Mh-S + Mb-S" model performed significantly better than model with "Mh-R + Mb-

S" (P-value$_{TukeyHSD}$<0.050). This observation reflects the decrease in performance of the models upon inclusion of a large number of predictors.

ML Model Performance and Predicted Outcome

We compared the performance of the models for each of the outcomes (HbA1c, FG, 2hGluc, FI and 2hIns) (**Figure 2a**). In the context of the model with metabolic data as input we observed that performance was the best for FI (**Figure 2a**). Models with different outcomes were ranked as following: FI > FG > 2hIns > 2hGlc > HBA1c (P-value$_{TukeyHSD}$<5.0x10^{-3}). The average coefficient of determination R^2 over all machine learning algorithm was 0.47, 0.30, 0.21, 0.16, 0.11 ("Mb-S") // 0.43, 0.24, 0.19, 0.13, 0.06 ("Mb-R") for FI, FG, 2hIns, 2hGlc, HBA1c respectively.

ML Model Performance and Algorithm

Finally, we compared the prediction performance of each of the ML models (**Figure 2a**). In the context of the models including at least metabolic data as input ("Mb-R or -S + X"), we found that RF and BT and performed similarly for all phenotypes (P-value$_{TukeyHSD}$>0.18). Besides, no significant difference was found between SVR-L2Linear and the two previous models (P-value$_{TukeyHSD}$>0.050). Among SVR models, we found that SVR-L2Linear either performed equally or outperformed the other SVRs, depending on the input dataset. In particular, for datasets with a large number of predictors SVR-L2Linear was the best performing SVR (P-value$_{TukeyHSD}$<0.050).

ML Models and Variable Importance

We investigated the contribution of metabolic and epigenomic variables to the prediction of glycaemic traits. We discuss predictors importance only in the context of FG and FI outcomes, for which prediction algorithms reached the best R^2 (**Figure 2a**). On the one hand, we found that FG prediction was mostly explained by metabolic variables: leucine, isoleucine, valine, tyrosine, BMI and WHR, HDLs and VLDL, Glycoprotein acetyls, glycerol at T2, WHR at T1 and sex (**Figure 3**).

The metabolic models with "scored" variables were driven by variables which mirrored the top "raw" predictors. FI prediction, on the other hand, was explained by BMI, WHR, HDL, VLDL, BRACA, phenylalanine, tyrosine, glycoprotein acetyls, glycerol, lactate, several lipidic ratios at T2, BMI and FI at T1. Overall, the model with "scored" variables for FI supported the importance of the former variables, as well as ketones bodies (acetoacetate and 3-hydroxybutyrate) at T2.

Interestingly, the model combining metabolic scores and methylation raw variables showed the relative importance of a few methylation probes for prediction of FG and FI (**Figure 3**). In the context of FI prediction, one methylation probe was ranked above non BRACA-amino acids, six above ketone bodies and seven above lipids at T2. In FG prediction -one, two, four and six- methylation probes were respectively ranked above -blood protein, BRACA, carbohydrates and insulin, and non-BRACA amino acids- (**Figure 3**).

ML and Prediction from Variables at T1

To test whether variables at T1 only can provide information about glycaemic traits at T2, we restricted the dataset to the best ranked variables among those at T1 in the context of the "Mb-R" model. From the full model which predicted FI, FG, 2hIns, 2hGlc, HBA1c with a R^2 of

0.43, 0.24, 0.19, 0.13, 0.06 ("Mb-R"), the restriction to T1 variables caused a drop in R^2 to 0.23, 0.19, 0.13, 0.04, 0.04 (Data not shown). Together it suggests that prediction from T1 variables only is not achievable in our dataset.

ML and Body Measurement Adjustment

We aimed at understanding the influence of the measures of obesity in the models. All Mb-R/ Mb-S/ Mh-R / Mh-S/ Mb-R + Mh-R/ Mb-R + Mh-S/ Mb-S + Mh-R/ Mb-S +Mh-S, models were adjusted for T1-BMI, T2-BMI, T1-WHR and T2-WHR. All adjusted models exhibited $R^2<0.060$ (**Figure 2b**), including models predicting FI and FG. Thus, the measures of obesity at T1 and T2 appear to be the main drivers of prediction.

Replication Analysis in the NFBC: Epigenomics Data and Methylation Probes Selected by ML

Analysis of variable importance (**Figure 3**) showed that in the context of the "Mb-S + Mh-R" model, 15 and 17 methylation probes were amongst the top 25 predictors for FG and FI respectively. To better understand the link between these probes and the glycaemic phenotypes we looked at the traits which have been previously reported to be associated with them[11–13]. Among the probes selected by the model predicting FG, 11 methylation probes were reported to be associated with BMI, one probe with FG, one probe with FI, one probe with HbA1c, four probe with T2D, two probes with 2hIns. In the case of the model predicting FI, 11 of the methylation probes were reported to be associated with BMI, three with FG, four with T2D, two with 2hIns. Next, we tested whether these probes also passed the replication threshold in the NFBC cohort with a simple regression analysis. Interestingly, 3/3 and 5/8 of the T1 probes selected by ML respectively for FI and FG, did replicate the association in at least one of the three phenotype available at T1 (FI, FG, BMI). For probes at T2, association with at least one of the seven phenotypes at T2 (FI, FG, BMI, T2D, 2hGluc, 2hIns, HbA1c) was replicated in 12/14, 6/7 probes selected by ML for FI and FG prediction respectively.

Discussion

Our study is the first to implement machine learning for prediction of continuous glycaemic traits. In this work we have analysed metabolic and methylation data at 31 and 46 years old to predict five glycaemic traits indicative for diabetes diagnosis at 46 years old with six Machine learning approaches. We found that the models with the best predictive ability included raw or score metabolic data as input, BT, RF, or SVR-L2Linear as algorithms, FI and FG traits as outcomes. We identified metabolites and epidemiological variables which drove the prediction of the models. We showed that prediction exclusively from variables at T1 did not lead to good performance and that adjustment for measures of obesity reduced significantly the predictive capacity of the models. Finally, we found that methylation probes were selected among the top 25 predictors for FG and FI and demonstrated that 72% and 89% of these probes at T1 and T2 were associated with at least one of the traits tested for replication in NFBC.

Power of the Methodology

Very few prediction studies have stood out in the field of type 2 diabetes risk prediction largely dominated by association studies[18,19,21–23]. This fact is also known as so-called "missing

heritability issue" in genome-wide association studies[37]. However, all published studies targeted T2D onset prediction as a discrete value and rely on the categorization of patients based on diagnosis threshold for FG, 2hGluc, random glucose and HbA1c. Continuous phenotypes, on the contrary, have the potential to reflect the progressive onset of the disease and have better power. Moreover, prediction of continuous phenotypes allow removal of HbA1c, FG, FI, and OGTT-derived measures from the set of predictors which may hide more modest effects of other variables[23]. Indeed, FG[18–20,22,23] and 2hGluc[18,23] are good predictors of T2D.

Overlap of Metabolic Variables with Other Studies

Our study leverages machine learning ability to perform variable selection independently of a pre-filtering, with RF, BT and SVR-L2Linear algorithms. To date, RLS (a variant of SVR-L2Linear algorithms)[20] , J48-decision tree[18], and logistic regression with regularization[19,22] have highlighted importance of specific metabolites consistent with our findings. Indeed, branched-chain amino acids (Leucine, Valine, Isoleucine)[18,19], HDL, VLDL, glycerol, ApoA and Apo B, 3-hydroxybutyrate[19], aromatic amino acids (phenylalanine, tyrosine)[19,22] are established as important predictors by machine learning algorithms. In this study, we report glycoprotein acetyls, acetoacetate as good predictors of glycaemic trait levels. Although these markers are associated with T2D[38–40], for the first time here, we show that they are not only associated, but are also predictors of glycaemic health.

Epigenetic Markers as Predictors of Glycaemic Health

The machine learning algorithms assigned a high rank (first 25) to the established "metabolic health-associated" methylation probes in "Mh-R + Mb-S" model. Additionally, replication of associations for these probes in our study suggests that epigenetic variability may hold a predictive value in risk models for glycaemic health. Although the methylation markers did not improve the prediction of our model, weighting scores by the combined contribution of epigenetic regulation derived from established methylation marks might help accounting for their effect, and in building prediction models.

Epigenomic and Metabolic Scores

The metabolomics scores overperformed raw data for prediction of 3/5 outcomes, while the reverse pattern was found for methylation scores in 2/5 outcomes. This trend might be explained by the score computation, which has been performed from biochemical classes in metabolomics data and per traits previously reported to be associated with the probes[11–13] in methylation data.

Effect of BMI and WHR

Measures of body adiposity and those of obesity are established risk factors for T2D[41] and have a well-known impact on glycaemic trait variability[42]. In all six machine learning approaches and within all data combinations, we confirmed high predictive value of BMI and WHR at T1 in glycaemic trait levels change. On the one hand, outcome modelling with BMI and WHR at T1 and T2 adjustment has highlighted the primary role of obesity in the drive of

metabolic changes leading to T2D. On the other hand, this finding has helped us highlighting the modest predictive value of the metabolites and epigenetic variables available to us.

Limitations and Challenges of the Study

Our analysis has some limitations, such as relatively small sample size (595 subjects), which is a drawback for taking full advantage of machine learning prediction with high-dimensional data layers. Next, drawing of a threshold regarding variable importance in our machine learning models is not trivial. Indeed, depending on the algorithm, final variable importance will depend on the number of variables resampled by the algorithms or the regularization parameters chosen. Limitations inherent to the samples and the study design are the use of whole blood which is a heterogeneous tissue, and relatively young age of the participants (31 years old) at T1. No study so far has estimated the best time point for prediction of T2D.

In conclusion, we conducted the first study which aims at prediction of glycaemic trait levels with machine learning modelling from combined metabolic and methylation datasets from at least two time points. In the future, we expect that improvements in methylation scores computation, longer model tuning and replication in an external dataset will lead to stronger predictive ability of our models for glycaemic traits, and will unveil novel prognostic omics biomarkers for T2D endophenotypes.

Acknowledgments

IP is funded by the World Cancer Research Fund (WCRF UK) and World Cancer Research Fund International (2017/1641), the Wellcome Trust (WT205915), and the European Union's Horizon 2020 research and innovation programme (DYNAhealth, project number 633595). MAK works in a Unit that is supported by the University of Bristol and UK Medical Research Council (MC_UU_12013/1). The computational work was performed using the Imperial College Research Computing Service, DOI: 10.14469/hpc/2232.

We thank all cohort members and researchers who participated in the 31 and 46 years study. We also wish to acknowledge the work of the NFBC project center. NFBC1966 31 years old study received financial support from University of Oulu Grant no. 65354, Oulu University Hospital Grant no. 2/97, 8/97, Ministry of Health and Social Affairs Grant no. 23/251/97, 160/97, 190/97, National Institute for Health and Welfare, Helsinki Grant no. 54121, Regional Institute of Occupational Health, Oulu, Finland Grant no. 50621, 54231. NFBC1966 46 years old study received financial support from University of Oulu Grant no. 24000692, Oulu University Hospital Grant no. 24301140, ERDF European Regional Development Fund Grant no. 539/2010 A31592.

Data Availability: Data is available from the Northern Finland Birth Cohort (NFBC) for researchers who need the criteria for accessing confidential data. Please, contact NFBC project center (NFBCprojectcenter@oulu.fi) and visit the cohort website (www.oulu.fi/nfbc) for more information.

References

1. International Diabetes Federation - Home. Available at: https://www.idf.org/. (Accessed: 31st May 2018)
2. Voight, B. F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42,** 579–589 (2010).
3. Magic Investigators - Home Page. Available at: https://www.magicinvestigators.org/. (Accessed: 31st May 2018)
4. Scott, R. A. *et al.* An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes* (2017).
5. Manning, A. K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44,** 659–669 (2012).
6. Saxena, R. *et al.* Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat. Genet.* **42,** 142–148 (2010).
7. Soranzo, N. *et al.* Common variants at 10 genomic loci influence hemoglobin A1C levels via glycemic and nonglycemic pathways (Diabetes (2010) 59, (3229-3239)). *Diabetes* **60,** (2011).
8. Marullo, L., El-Sayed Moustafa, J. S. & Prokopenko, I. Insights into the Genetic Susceptibility to Type 2 Diabetes from Genome-Wide Association Studies of Glycaemic Traits. *Curr. Diab. Rep.* **14,** (2014).
9. Lowry, E. *et al.* Understanding the complexity of glycaemic health - Systematic bio-psychosocial modelling of fasting glucose in middle-age adults; a DynaHEALTH study. *Int. J. Obes.* **In press,** (2018).
10. Chambers, J. C. *et al.* Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. *lancet. Diabetes Endocrinol.* **3,** 526–534 (2015).
11. Walaszczyk, E. *et al.* DNA methylation markers associated with type 2 diabetes, fasting glucose and HbA1c levels: a systematic review and replication in a case–control sample of the Lifelines study. *Diabetologia* **61,** 354–368 (2018).
12. Kriebel, J. *et al.* Association between DNA Methylation in Whole Blood and Measures of Glucose Metabolism: KORA F4 Study. *PLoS One* **11,** (2016).
13. Wahl, S. *et al.* Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541,** 81–86 (2017).
14. Suhre, K. *et al.* Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. *PLoS One* **5,** e13953 (2010).
15. Fiehn, O. *et al.* Plasma metabolomic profiles reflective of glucose homeostasis in non-diabetic and type 2 diabetic obese African-American women. *PLoS One* **5,** e15234 (2010).
16. Drogan, D. *et al.* Untargeted metabolic profiling identifies altered serum metabolites of type 2 diabetes mellitus in a prospective, nested case control study. *Clin. Chem.* **61,** 487–497 (2015).
17. Wang, T. J. *et al.* Metabolite Profiles and the Risk of Developing Diabetes. *Nat. Med.* **17,** 448–453 (2011).
18. Allalou, A. *et al.* A Predictive Metabolic Signature for the Transition From Gestational Diabetes Mellitus to Type 2 Diabetes. *Diabetes* **65,** 2529–2539 (2016).
19. Liu, J. *et al.* Metabolomics based markers predict type 2 diabetes in a 14-year follow-up study. *Metabolomics* **13,** 104 (2017).
20. Peddinti, G. *et al.* Early metabolic markers identify potential targets for the prevention of type 2 diabetes. *Diabetologia* **60,** 1740–1750 (2017).
21. Mühlenbruch, K., Jeppesen, C., Joost, H.-G., Boeing, H. & Schulze, M. B. The Value of Genetic Information for Diabetes Risk Prediction – Differences According to Sex, Age, Family History and Obesity. *PLoS One* **8,** e64307 (2013).
22. Yengo, L. *et al.* Impact of statistical models on the prediction of type 2 diabetes using non-targeted metabolomics profiling. *Mol. Metab.* **5,** 918–925 (2016).
23. Savolainen, O. *et al.* Biomarkers for predicting type 2 diabetes development-Can

metabolomics improve on existing biomarkers? *PLoS One* **12,** e0177738 (2017).
24. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. (2009).
25. Northern Finland Cohorts. Available at: http://www.oulu.fi/nfbc/. (Accessed: 11th June 2018)
26. Taponen, S. *et al.* Hormonal Profile of Women with Self-Reported Symptoms of Oligomenorrhea and/or Hirsutism: Northern Finland Birth Cohort 1966 Study. *J. Clin. Endocrinol. Metab.* **88,** 141–147 (2003).
27. Taponen, S. *et al.* Metabolic Cardiovascular Disease Risk Factors in Women with Self-Reported Symptoms of Oligomenorrhea and/or Hirsutism: Northern Finland Birth Cohort 1966 Study. *J. Clin. Endocrinol. Metab.* **89,** 2114–2118 (2004).
28. Rautio, N. *et al.* Accumulated exposure to unemployment is related to impaired glucose metabolism in middle-aged men: A follow-up of the Northern Finland Birth Cohort 1966. *Prim. Care Diabetes* **11,** 365–372 (2017).
29. Perkiömäki, N. *et al.* Association between Birth Characteristics and Cardiovascular Autonomic Function at Mid-Life. *PLoS One* **11,** (2016).
30. Soininen, P. *et al.* High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. *Analyst* **134,** 1781–1785 (2009).
31. Soininen, P., Kangas, A. J., Würtz, P., Suna, T. & Ala-Korpela, M. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ. Cardiovasc. Genet.* **8,** 192–206 (2015).
32. Wurtz, P. *et al.* Quantitative Serum NMR Metabolomics in Large-Scale Epidemiology: A Primer on -Omic Technology. *Am. J. Epidemiol.* (2017).
33. Wang, Q., Holmes, M. V., Smith, G. D. & Ala-Korpela, M. Genetic support for a causal role of insulin resistance on circulating branched-chain amino acids and inflammation. *Diabetes Care* **40,** 1779–1786 (2017).
34. R: The R Project for Statistical Computing. Available at: https://www.r-project.org/. (Accessed: 10th June 2018)
35. Aslibekyan, S. *et al.* Association of Methylation Signals With Incident Coronary Heart Disease in an Epigenome-Wide Assessment of Circulating Tumor Necrosis Factor α. *JAMA Cardiol.* (2018). doi:10.1001/jamacardio.2018.0510
36. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13,** 86 (2012).
37. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456,** 18–21 (2008).
38. Bentley-Lewis, R. *et al.* Metabolomic profiling in the prediction of gestational diabetes mellitus. *Diabetologia* **58,** 1329–1332 (2015).
39. Akbay, E. *et al.* The relationship between levels of alpha1-acid glycoprotein and metabolic parameters of diabetes mellitus. *Diabetes. Nutr. Metab.* **17,** 331–335 (2004).
40. Mahendran, Y. *et al.* Association of ketone body levels with hyperglycemia and type 2 diabetes in 9,398 Finnish men. *Diabetes* **62,** 3618–3626 (2013).
41. Eckel, R. H. *et al.* Obesity and Type 2 Diabetes: What Can Be Unified and What Needs to Be Individualized? *J. Clin. Endocrinol. Metab.* **96,** 1654–1663 (2011).
42. Scott, R. A. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44,** 991–1005 (2012).

Table 1. Association analysis between 541 probes (277 at T1 and 264 at T2) and seven phenotypes. The number of methylation probes at T1 associated with T1 phenotypes (FG, FI or BMI available) as well as the number of methylation probes at T1 associated with T2 phenotypes (BMI, FG, HbA1c, T2D, 2hGluc, FI or 2hIns) and the number of methylation probes at T2 associated with T2 phenotypes (BMI, FG, HbA1c, T2D, 2hGluc, FI or 2hIns) is displayed.

Outcome T2 or T1 ~ Methylation T1 or T2	HbA1c		BMI		FG		FI		2hGluc		2hIns		T2D	
	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
Methylation data at T1		0/11	81/187	70/187	0/21	4/21	0/8	0/8		0/1		5/21		2/67
Methylation data at T2		0/11		133/187		2/21		1/8		1/1		9/21		6/67

Probes for which association with the outcome has been replicated with P-value < 0.05 in NFBC cohort

Phenotype Non Available

Association Not Tested

FG/FI: Fasting glucose/insulin; HbA1c: glycated haemoglobin; 2hGlu/2hIns: 2-hour glucose/insulin; T1: data of participants at age 31 years; T2: data of participants at age 46 years.

Table 2. Effect of the input dataset on the prediction performance of the five glycaemic traits. Mb-R: Metabolic Raw Variables, Mb-S: Metabolic Scored Variables, Mh-R: Methylation Raw Variables, Mh-S: Methylation Scored Variables. Metabolic predictors include epidemiological data, biochemical data and metabolomics data.

Comparison index	Models	Datasets included in the models				Number of outcomes for which the model A/B performs the **best**	Number of outcomes in which A and B perform **equally**
		Mb-R	**Mb-S**	**Mh-R**	**Mh-S**		
1) Comparison of:	Model A	X				5/5 ($P<5.0x10^{-8}$)	0/5
	Model B			X		0/5	
2) Comparison of:	Model A	X				5/5 ($P<5.0x10^{-13}$)	0/5
	Model B				X	0/5	
3) Comparison of:	Model A		X			5/5 ($P<1.0x10^{-16}$)	0/5
	Model B			X		0/5	
4) Comparison of:	Model A		X			5/5 ($P<5.0x10^{-13}$)	0/5
	Model B				X	0/5	
5) Comparison of:	Model A	X				0/5	5/5 (P>0.80)
	Model B	X		X		0/5	
6) Comparison of:	Model A	X				0/5	5/5 (P>0.30)
	Model B	X			X	0/5	
7) Comparison of:	Model A		X			3/5 ($P<5.0x10^{-3}$)	2/5 (P>0.050)
	Model B	X				0/5	
8) Comparison of:	Model A			X		2/5 (P<0.050)	3/5 (P>0.50)
	Model B				X	0/5	
9) Comparison of:	Model A		X		X	5/5 ($P<5.0x10^{-2}$)	0/5
	Model B		X	X		0/5	

FG/FI: Fasting glucose/insulin; HbA1c: glycated haemoglobin; 2hGlu/2hIns: 2-hour glucose/insulin.
The performance of all Machine learning (ML) algorithms upon inclusion of different datatypes was evaluated. Selected comparison of models in pairs are displayed on this figure to illustrate:
(Comparison 1-4)-The comparison between models with metabolic or methylation data only.
(Comparison 5-6)-The effect of combination of two data types.
(Comparison 7-8)-The effect of variables transformation to scores.
(Comparison 9) The decrease in performance of the models upon inclusion of a large number of predictors.
For a given phenotype (FG, FI, HbA1c, 2hGlu or 2hIns), the effect of an input dataset was assessed. Column 4 shows which model performed the best, and the number of outcomes for which this pattern is observed. Column 5 lists the number of phenotypes in which no difference between "A" and "B" was observed.
Models were compared with Tukey HSD test following a one-way ANOVA. To test the effect of a given dataset, we run all six ML algorithms in a nested cross validation framework (5 outer, 5 inner folds), thereby each group compared included six (ML algorithms) x five (testing errors) = 30 R^2 measures.

Figure legends

Figure 1. Experimental set-up for Machine learning (ML) analysis. ML was applied to multi-omics data from the Northern Finland Birth Cohort 1966 at 31 and 46 years. Fasting glucose/insulin (FG/FI), glycated haemoglobin (HbA1c) and 2-hour glucose/insulin (2hGlu/2hIns) phenotypes at T2 were predicted in 595 individuals using up to 1,010 variables from T1 and T2: Body-mass-index (BMI), waist-hip-ratio, sex; 10 plasma measurements; 453 NMR-based metabolites; 542 methylation probes established for BMI/FG/FI/HbA1c /2hGlu/2hIns/Type 2 diabetes. Six ML approaches were used, random forest, boosted trees and support vector regression (SVR) with the kernels of linear/linear with L2 regularization/polynomial/radial basis function.

Figure 2. Performance in R^2 of the different machine learning models. (a) Unadjusted for measurements of obesity (Waist-to-hip-ratio and Body mass index) at T1 and T2; (b) Adjusted for measurements of obesity (Waist-to-hip-ratio and Body mass index) at T1 and T2. Training of the algorithm was performed with a nested cross validation (5-folds outer, and 5-folds inner cross validation) and the R^2 of 5 outer testing folds is displayed for each machine learning model. "Metabolic predictors" include epidemiological data, biochemical data and metabolomics data. SVR: Support Vector Regression with linear/linear with L2 regularization/polynomial/radial basis function kernels.

Figure 3. Variable Importance for fasting glucose and fasting insulin prediction from two different datasets. For each Machine learning (ML) method, the normalized variable importance over five outer fold of cross validation was averaged into the "Variable-Model-Importance" (var.mod.Imp). Then for each of the six machine learning models, the variables were ranked based on the var.mod.Imp. The rank was averaged over the six models to obtain the "mean variable rank". The latter was used to select top 25 variables for display. For these variables, we display the variable importance after (1) weighting the var.mod.Imp by the R^2 obtained for each of the individual ML algorithms (2) averaging variable importance across the six ML models.
FI: Fasting Insulin; FG: Fasting Glucose; RF: random forest; BT: boosted trees, SVM: support vector regression models with linear/linear with L2 regularization/polynomial/radial basis function kernels. Metabolic predictors include epidemiological data, biochemical data and metabolomics data. T2: 46 years old, T1: 31 years old. BMI: Body Mass Index according to clinical examination, postal questionnaire if missing; WHR: Waist-to-hip ratio; Glycerol: Glycerol, mmol/l; GlycoproteinAcetyls: Glycoprotein acetyls, mainly a1-acid glycoprotein, mmol/l; Isoleucine: Isoleucine, mmol/l; Large_HDL_TotChol: Total cholesterol in large HDL , mmol/l; Large_HDL_CholEsters: Cholesterol esters in large HDL , mmol/l; Large_HDL_CholEsters_%: Cholesterol esters to total lipids ratio in large HDL , %; Large_HDL_FreeChol: Free cholesterol in large HDL , mmol/l; Large_HDL_FreeChol_%: Free cholesterol to total lipids ratio in large HDL , %; Large_HDL_Lipids: Total lipids in large HDL , mmol/l; Large_HDL_Particules: Concentration of large HDL particles, mol/l; Large_HDL_PhospholLipids_%: Phospholipids to total lipids ratio in large HDL , %; Large_HDL_Trigycerides_%: Triglycerides to total lipids ratio in large HDL , %; Large_VLDL_PhospholLipids: Phospholipids in large VLDL , mmol/l; Lactate: Lactate, mmol/l; Leucine: Leucine, mmol/l; Medium_VLDL_TotChol_%: Total cholesterol to total lipids ratio in medium VLDL , %; Medium_VLDL_CholEsters_%: Cholesterol esters to total lipids ratio in medium VLDL , %; Medium_VLDL_Trigycerides: Triglycerides in medium VLDL , mmol/l; Medium_VLDL_Trigycerides_%: Triglycerides to total lipids ratio in medium VLDL, %; Phenylalanine: Phenylalanine, mmol/l; Small_VLDL_Lipids: Total lipids in small VLDL , mmol/l; Small_VLDL_Particules: Concentration of small VLDL particles, mol/l; TrigyceridestoPhosphoglycerides: Ratio of triglycerides to phosphoglycerides, NA; Tyrosine: Tyrosine, mmol/l; Valine: Valine, mmol/l; VLDL_D: Mean diameter for VLDL particles, nm; XL_HDL_FreeChol: Free cholesterol in very large HDL, mmol/l; XL_HDL_PhospholLipids_%: Phospholipids to total lipds ratio in very large HDL, %; XL_HDL_Trigycerides_%: Triglycerides to total lipids ratio in very large HDL , %; XL_VLDL_Trigycerides:

Triglycerides in very large VLDL , mmol/l; XS_VLDL_FreeChol_%: Free cholesterol to total lipids ratio in very small VLDL , %; XXL_VLDL_FreeChol: Free cholesterol in chylomicrons and extremely large VLDL , mmol/l; XXL_VLDL_PhospholLipids: Phospholipids in chylomicrons and extremely large VLDL , mmol/l; XXL_VLDL_Trigycerides: Triglycerides in chylomicrons and extremely large VLDL , mmol/l; BRACA: Branched-chain amino acids (Leucine, Valine, Isoleucine), MesuresOfAdiposity: Body Mass Index + BMI, OtherAminoAcids: Alanine, Glutamine, Glycine, Histidine, Phenylalanine, Tyrosine, + Creatinine, T1_ CarbohydratesAndInsulin: Glucose from metabolomics platform, Lactate, Pyruvate, Citrate, Glycerol, Acetate, Fasting Plasma Glucose and Insulin from biochemical measurements. T2_Carbohydrates: Lactate, Citrate, Glycerol, Acetate. BloodProteins: Albumin, Glycoprotein Acetyls (mainly a1-acid glycoprotein) , KetonBodies: 3-hydroxybutyrate, Acetoacetate, Lipids: [Fasting Serum Triglycerides, Esterified cholesterol, Serum total triglycerides, Total phosphoglycerides, Ratio of triglycerides to Phosphoglycerides, Phosphatidylcholine and other Cholines, Sphingomyelins, Total fatty acids, Estimated degree of unsaturation in Lipids, 22:6, docosahexaenoic acid, 18:2, linoleic acid, Omega-3 fatty acids, Omega-6 fatty acids, Polyunsaturated fatty acids, Monounsaturated fatty acids; 16:1,18:1, Saturated fatty acids, Ratio of 22:6 docosahexaenoic acid to total fatty acids, Ratio of 18:2 linoleic acid to total fatty acids, Ratio of omega-3 fatty acids to total fatty acids, Ratio of omega-6 fatty acids, Ratio of polyunsaturated fatty acids to total fatty acids, Ratio of monounsaturated fatty acids to total fatty acids, Ratio of saturated fatty acids to total fatty acids.

Figures

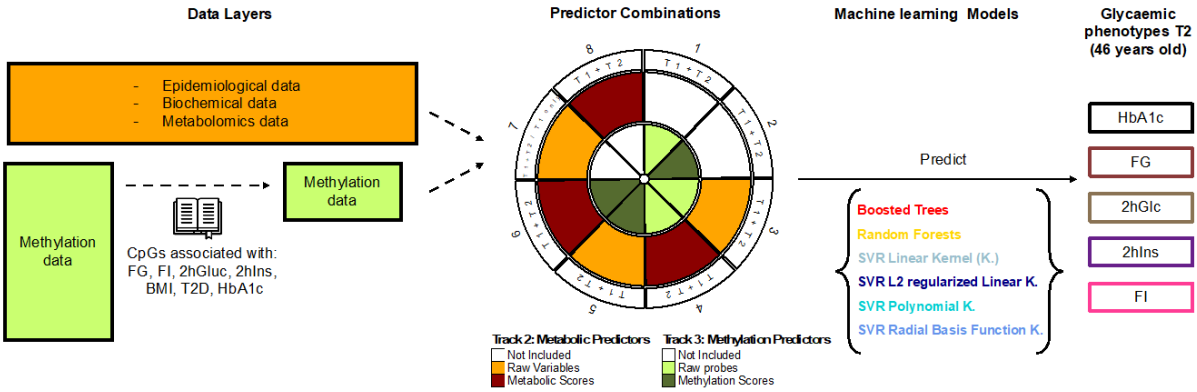

Figure 1. Experimental set-up for Machine learning (ML) analysis. ML was applied to multi-omics data from the Northern Finland Birth Cohort 1966 at 31 and 46 years. Fasting glucose/insulin (FG/FI), glycated haemoglobin (HbA1c) and 2-hour glucose/insulin (2hGlu/2hIns) phenotypes at T2 were predicted in 595 individuals using up to 1,010 variables from T1 and T2: Body-mass-index (BMI), waist-hip-ratio, sex; 10 plasma measurements; 453 NMR-based metabolites; 542 methylation probes established for BMI/FG/FI/HbA1c /2hGlu/2hIns/Type 2 diabetes. Six ML approaches were used, random forest, boosted trees and support vector regression (SVR) with the kernels of linear/linear with L2 regularization/polynomial/radial basis function.

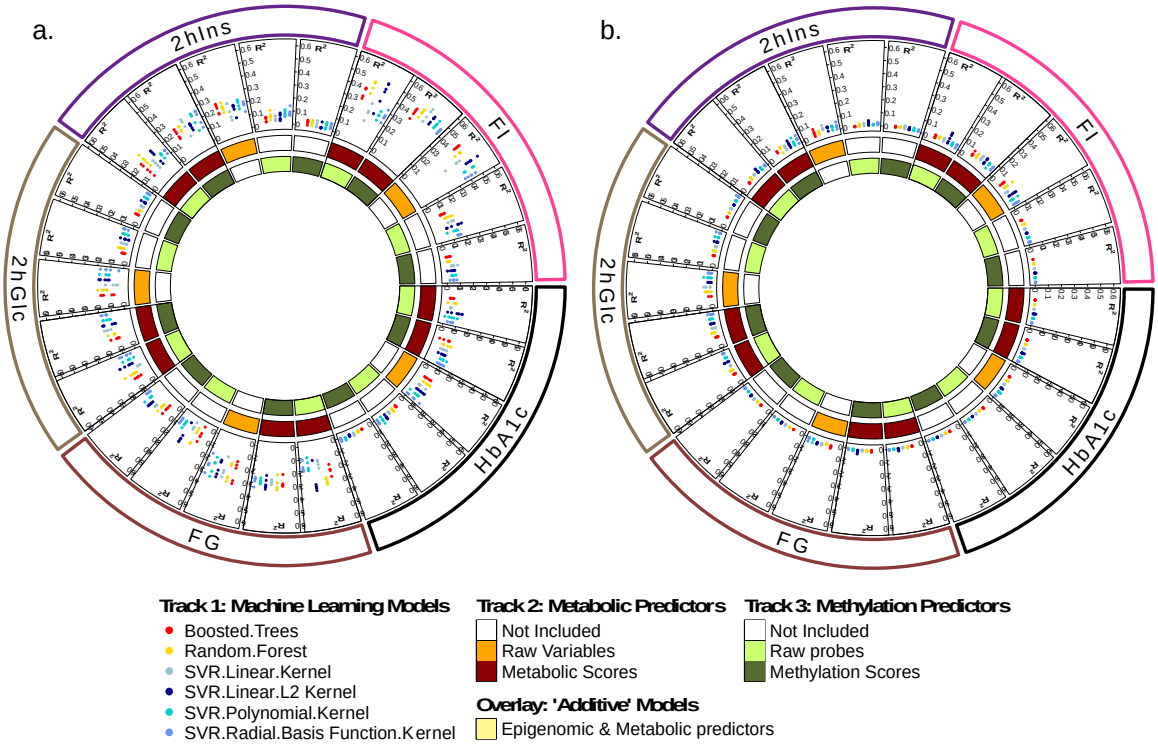

Figure 2. Performance in R^2 of the different machine learning models. (a) Unadjusted for measurements of obesity (Waist-to-hip-ratio and Body mass index) at T1 and T2; (b) Adjusted for measurements of obesity (Waist-to-hip-ratio and Body mass index) at T1 and T2. Training of the algorithm was performed with a nested cross validation (5-folds outer, and 5-folds inner cross validation) and the R^2 of 5 outer testing folds is displayed for each machine learning model. "Metabolic predictors" include epidemiological data, biochemical data and metabolomics data. SVR: Support Vector Regression with linear/linear with L2 regularization/polynomial/radial basis function kernels.

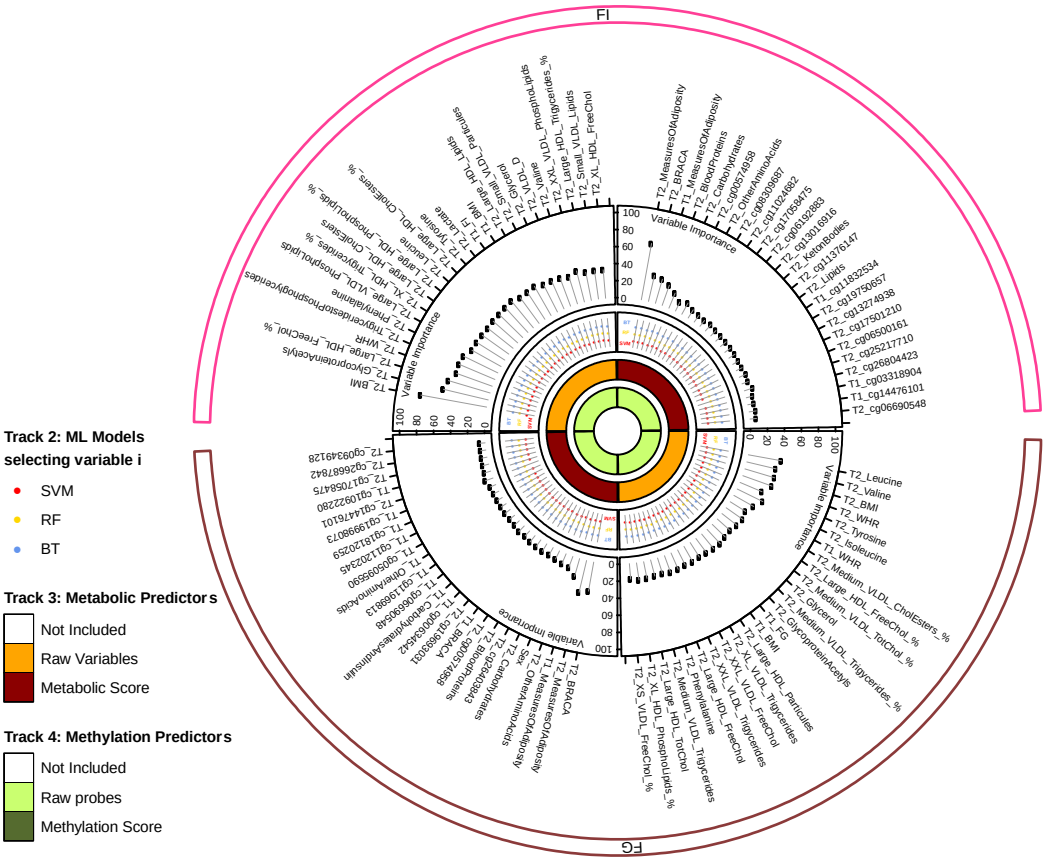

Figure 3. Variable Importance for fasting glucose and fasting insulin prediction from two different datasets. For each Machine learning (ML) method, the normalized variable importance over five outer fold of cross validation was averaged into the "Variable-Model-Importance" (var.mod.Imp). Then for each of the six machine learning models, the variables were ranked based on the var.mod.Imp. The rank was averaged over the six models to obtain the "mean variable rank". The latter was used to select top 25 variables for display. For these variables, we display the variable importance after (1) weighting the var.mod.Imp by the R^2 obtained for each of the individual ML algorithms (2) averaging variable importance across the six ML models.

FI: Fasting Insulin; FG: Fasting Glucose; RF: random forest; BT: boosted trees, SVM: support vector regression models with linear/linear with L2 regularization/polynomial/radial basis function kernels. Metabolic predictors include epidemiological data, biochemical data and metabolomics data. T2: 46 years old, T1: 31 years old. BMI: Body Mass Index according to clinical examination, postal questionnaire if missing; WHR: Waist-to-hip ratio; Glycerol: Glycerol, mmol/l; GlycoproteinAcetyls: Glycoprotein acetyls, mainly a1-acid glycoprotein, mmol/l; Isoleucine: Isoleucine, mmol/l; Large_HDL_TotChol: Total cholesterol in large HDL , mmol/l; Large_HDL_CholEsters: Cholesterol esters in large HDL , mmol/l; Large_HDL_CholEsters_%: Cholesterol esters to total lipids ratio in large HDL , %; Large_HDL_FreeChol: Free cholesterol in large HDL , mmol/l; Large_HDL_FreeChol_%: Free cholesterol to total lipids ratio in large HDL , %; Large_HDL_Lipids: Total lipids in large HDL , mmol/l; Large_HDL_Particules: Concentration of large HDL particles, mol/l; Large_HDL_PhospholLipids_%: Phospholipids to total lipids ratio in large HDL , %; Large_HDL_Trigycerides_%: Triglycerides to total lipids ratio in large HDL , %; Large_VLDL_PhospholLipids: Phospholipids in large VLDL , mmol/l; Lactate: Lactate, mmol/l; Leucine: Leucine, mmol/l; Medium_VLDL_TotChol_%: Total cholesterol to total lipids ratio in medium VLDL , %; Medium_VLDL_CholEsters_%: Cholesterol esters to total lipids ratio in medium VLDL , %; Medium_VLDL_Trigycerides: Triglycerides in medium VLDL, mmol/l;

Medium_VLDL_Trigycerides_%: Triglycerides to total lipids ratio in medium VLDL , %; Phenylalanine: Phenylalanine, mmol/l; Small_VLDL_Lipids: Total lipids in small VLDL , mmol/l; Small_VLDL_Particules: Concentration of small VLDL particles, mol/l; TrigyceridestoPhosphoglycerides: Ratio of triglycerides to phosphoglycerides, NA; Tyrosine: Tyrosine, mmol/l; Valine: Valine, mmol/l; VLDL_D: Mean diameter for VLDL particles, nm; XL_HDL_FreeChol: Free cholesterol in very large HDL, mmol/l; XL_HDL_PhosphoLipids_%: Phospholipids to total lipds ratio in very large HDL, %; XL_HDL_Trigycerides_%: Triglycerides to total lipids ratio in very large HDL , %; XL_VLDL_Trigycerides: Triglycerides in very large VLDL , mmol/l; XS_VLDL_FreeChol_%: Free cholesterol to total lipids ratio in very small VLDL , %; XXL_VLDL_FreeChol: Free cholesterol in chylomicrons and extremely large VLDL , mmol/l; XXL_VLDL_PhosphoLipids: Phospholipids in chylomicrons and extremely large VLDL , mmol/l; XXL_VLDL_Trigycerides: Triglycerides in chylomicrons and extremely large VLDL , mmol/l; BRACA: Branched-chain amino acids (Leucine, Valine, Isoleucine), MesuresOfAdiposity: Body Mass Index + BMI, OtherAminoAcids: Alanine, Glutamine, Glycine, Histidine, Phenylalanine, Tyrosine, + Creatinine, T1_ CarbohydratesAndInsulin: Glucose from metabolomics platform, Lactate, Pyruvate, Citrate, Glycerol, Acetate, Fasting Plasma Glucose and Insulin from biochemical measurements. T2_Carbohydrates: Lactate, Citrate, Glycerol, Acetate. BloodProteins: Albumin, Glycoprotein Acetyls (mainly a1-acid glycoprotein) , KetonBodies: 3-hydroxybutyrate, Acetoacetate, Lipids: [Fasting Serum Triglycerides, Esterified cholesterol, Serum total triglycerides, Total phosphoglycerides, Ratio of triglycerides to Phosphoglycerides, Phosphatidylcholine and other Cholines, Sphingomyelins, Total fatty acids, Estimated degree of unsaturation in Lipids, 22:6, docosahexaenoic acid, 18:2, linoleic acid, Omega-3 fatty acids, Omega-6 fatty acids, Polyunsaturated fatty acids, Monounsaturated fatty acids; 16:1,18:1, Saturated fatty acids, Ratio of 22:6 docosahexaenoic acid to total fatty acids, Ratio of 18:2 linoleic acid to total fatty acids, Ratio of omega-3 fatty acids to total fatty acids, Ratio of omega-6 fatty acids, Ratio of polyunsaturated fatty acids to total fatty acids, Ratio of monounsaturated fatty acids to total fatty acids, Ratio of saturated fatty acids to total fatty acids.