

## Prediction of type 2 diabetes mellitus onset using simple logistic regression models

Yochai Edlitz<sup>1,2</sup>, Eran Segal<sup>1,2,†</sup>

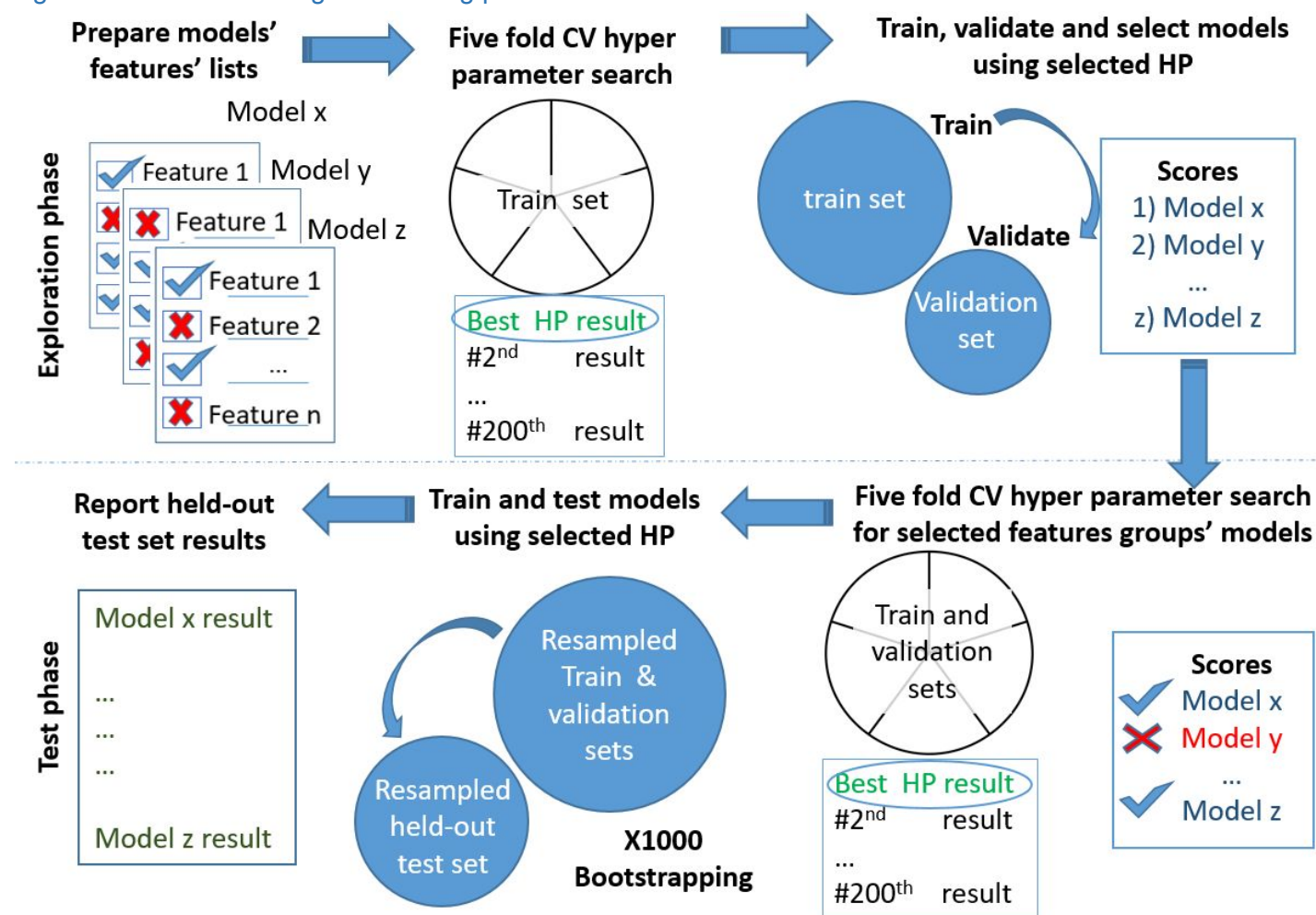
<sup>1</sup>Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel.

<sup>2</sup>Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel.

<sup>†</sup> Corresponding author

Supplementary material

Figure S1 - Models testing and training process



**Figure S1 |Models' extraction flow.** Scheme of the models' exploration and evaluation process. We use a fivefold CV with two hundred iterations of random hyperparameters (HP) selection process for each of the features groups. We then select the top-scored HP for each feature's group and Train a new model based on the train set and measure the auROC using the validation set. Out of the models validated, we chose the models that comply with objectives of minimal features and high performance. The reported results are of the held-out test set.

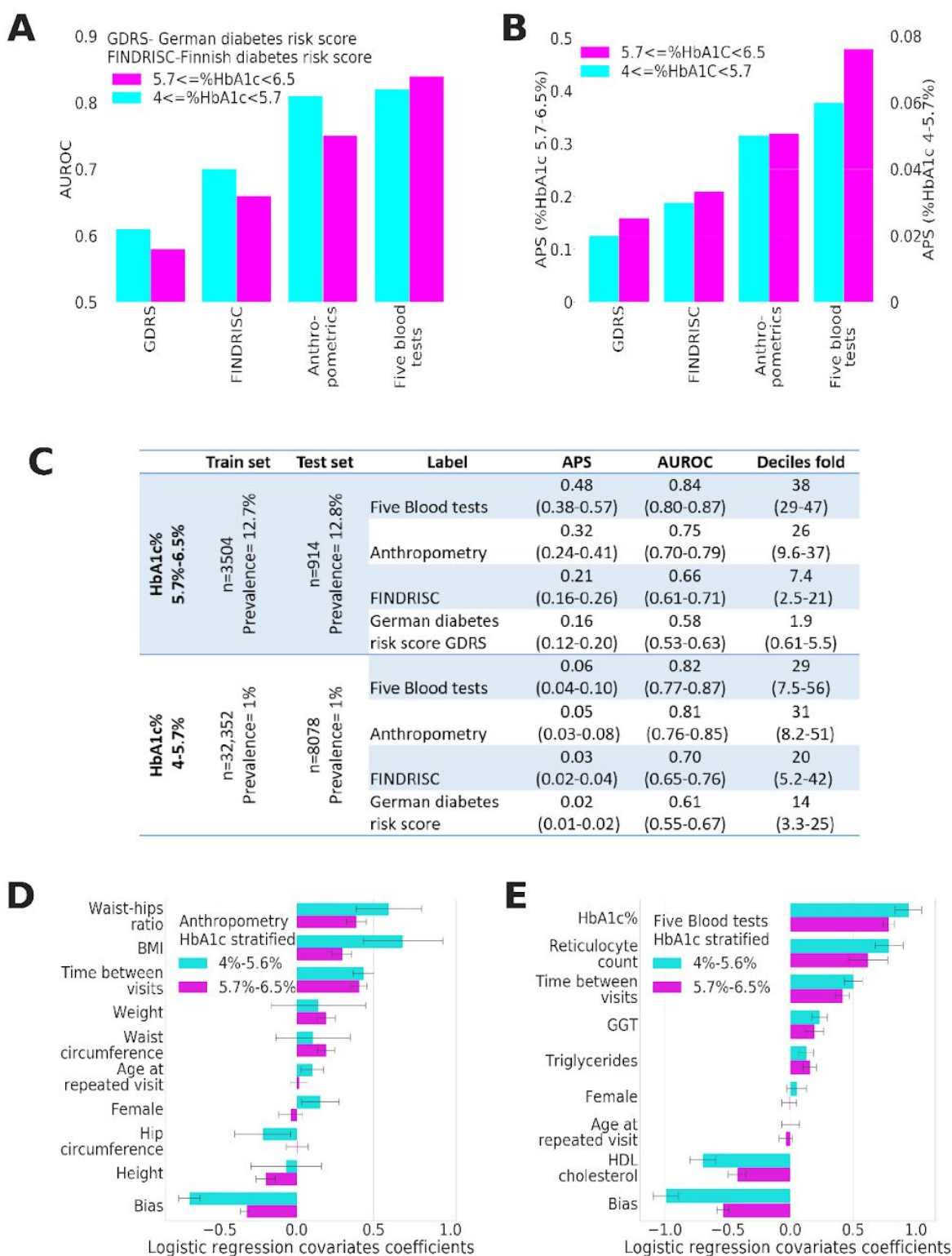
## Prediction within an HbA1c% stratified population

To verify that our model not only discriminates between participants who were already having elevated HbA1c% levels but is also capable of discriminating within a group of participants with normal or prediabetes HbA1c% levels we stratified our data to normoglycemic and prediabetic groups according to their HbA1c% levels. As patients are diagnosed by their HbA1c% levels, this measure is a strong predictor of T2D in itself with a 12% prevalence within the group of participants with HbA1c% levels 5.7-6.5%, versus only 1% of T2D onset for participants with HbA1c% levels 4-5.7% in our cohort.

We tested our models on these two T2D risk populations from our cohort. We examine what are the driving factors of T2D in each of these stratified groups (Figure S2). Within the normoglycemic group, the anthropometry model provides auROC of 0.81 (0.76-0.85) with an APS of 0.05 (0.03-0.08) and deciles' OR of x31 (8.2-51). These results outperform the FINDRISC results with auROC, APS and deciles' OR of 0.7 (0.65-0.76), 0.03 (0.02-0.04), and x20 (5.2-42) respectively. These results also surpass the GDRS results with auROC, APS, and deciles' OR of 0.61 (0.55-0.67), 0.02 (0.01-0.02), and x14(3.3-25) respectively. The Anthropometry model results in this normal HbA1c% range are similar to those of the five blood tests model results with auROC of 0.82 (0.77-0.87), APS of 0.06 (0.04-0.1) and deciles' OR of x29 (7.5-56). (Figure 2S A-C).

When we test the models within the pre-diabetic group, the anthropometry model achieves auROC of 0.75 (0.7-0.79), APS of 0.32 (0.24-0.41) deciles' OR of x26 (9.6-37). These results are again superior to the FINDRISC results with auROC, APS, and deciles' OR of 0.66 (0.61-0.71), 0.21 (0.16-0.26), and x7.4(2.5-21) respectively. It also surpasses the GDRS results with auROC, APS and deciles' OR of 0.58 (0.53-0.63), 0.16 (0.12-0.2) ,and x1.9 (0.61-5.5) respectively. In this HbA1c% range, the five blood tests model outperform the Anthropometry model with auROC, APS, and deciles' OR of 0.84 (0.80-0.87), 0.48 (0.38-0.57), and x38 (29-47) respectively. (Figure 2S a-c).

Comparing the feature importance of the Anthropometry model in the normoglycemic and pre-diabetic subpopulations indicates that the main differences are in the bias coefficient, which reflects the higher prevalence of T2D in the pre-diabetic group and the time between visits feature is also becoming more dominant than waist to hips ratio and BMI. Suggesting that in the pre-diabetic group, it is mostly a matter of time until T2D emerges. Within the normoglycemic group, we see that the most predictable feature is the WHR, followed by BMI - both taking into account not only the body measures but also implicit body measures ratios which indicate about the effect of the body shape (Figure S2 D). When we compare the five-blood-tests model's features importance, it appears that the features importance values did not vary much apart from the negative impact of HDL cholesterol in the normoglycemic group and the bias term which indicates the higher prevalence in the pre-diabetic group (Figure S2-E).



**Figure S2 | Stratified populations according to HbA1c% - models comparison. A)** auROC in both normoglycemic and pre-diabetic groups. Although the five-blood-tests model achieves more accurate scores than the anthropometric model in both populations, the anthropometry model's results approach the five blood tests model in the normoglycemic group. Both models outperform the FINDRISC and GDRS models.

**B) Comparison of model's APS within the HbA1c% stratified groups.** Left Y-axis indicates the APS of the models' in the prediabetics group, the right Y-axis indicates APS in the normoglycemic group, we see again that the order of model's predictability remains Five blood tests; Anthropometry; FINDRISC and lastly the GDRS model. **C)** A summary table for the various models AuROC, Average-Precision and deciles risk fold scores for the HbA1c% stratified subpopulations. **D)** A feature importance comparison of the Anthropometric model within the HbA1c% stratified groups. showing a stronger impact of WHR and BMI in the normoglycemic group and a stronger negative impact of the bias - indicating the higher prevalence in the pre-diabetic group. **E)** A feature importance comparison of the five-blood-tests model within the HbA1c% stratified groups. showing a stronger negative impact of HDL cholesterol in the normoglycemic group and a stronger negative impact of the bias - indicating the lower prevalence in the normoglycemic group.

Comparing the five blood tests model's features importance, the features importance values do not vary much. The main differences are the stronger negative impact of HDL cholesterol on the prediction results and the trivial bias term, which indicates the lower prevalence in the normoglycemic group (Figure S2 E).

#### Exploring the full features' space using gradient boosted decision trees.

To select features for our simple models, we analysed the features' importance of features that we sought of having some relation to T2D. We analyse what is the power of a predictive model with a vast amount of information, to compare it to our minimal features models.

We started by sorting out a list of 279 covariates from the first visit to the UKB assessment centre to be used as our preliminary features. On top of these features, we used the UKB single-nucleotide polymorphisms (SNPs) genotyping data and their calculated PRS.

We inspected the impact of various features' groups, using the lightGBM<sup>30</sup> gradient decision trees model using SHAP<sup>32,31</sup> (See methods). We aggregated the features into thirteen separate groups: age and sex; genetics; early life factors; sociodemographics; mental health; blood pressure and heart rate; family and ethnicity; medication; diet; lifestyle and physical activity; physical health; anthropometry; blood tests. All of these groups included age and sex.

We also tested the impact of HbA1c% with age and sex; genetics without age and Sex (Table S1). The top five predictive GBDT models, according to their auROC, APS, and decile folds, in descending order are: the "All features" model with auROC 0.9 (0.88-0.92 95%CI), APS of 0.28 (0.22-0.35 95% CI) and a deciles' OR of x65 (49-73); The "full-blood-tests" model with auROC 0.89 (0.86-0.9), APS 0.25 (0.2-0.31), and deciles' OR of x62 (48-69); The "five blood tests" model with auROC 0.87(0.85-0.89), APS 0.21 (0.17-0.27) and deciles' OR of x59(49-66); the HbA1c% based model with auROC 0.81 (0.78-0.84), APS 0.18 (0.13-0.23) and deciles' OR of x32(11-59); and the anthropometry model with auROC 0.76 (0.73-0.79), APS 0.07 (0.06-0.1) and deciles' OR x30 (12-42) (Table 1S).

To explore the heritability impact on the predictability of the models, we used two main features groups. One is a broad sense heritability based on SNPs and PRSs, and the other one is heritability which includes family and ethnicity. Adding the SNPs and PRSs data to the age and sex group, elevated the APS and auROC from 0.03 (0.02-0.04), 0.58 (0.54-0.62) to 0.04 (0.03-0.05) and 0.61 (0.57-0.65) respectively.

Adding this DNA array data to all other features together did not provide a considerable contribution to prediction. Using the family and ethnicity features with age and sex provides an APS of 0.04 (0.03-0.05),

auROC of 0.63(0.58-0.67), and deciles' OR of x5.4 (2-14) (Table S1), considerably lower than other features' groups such as the anthropometry group with auROC 0.76(0.73-0.79) and deciles' OR of x30 (12-42) and from the lifestyle and physical activity group, scoring auROC of 0.69 (0.65-0.72) and deciles' OR of x16 (6-29). A possible explanation for the low predictability of heritable features is that T2D genetics plays a role mainly in the younger population and the greater role played by environmental factors in those who develop diabetes late in life <sup>33</sup>

Another interesting result is a lifestyle and physical activity model, which includes ninety-eight features related to: physical activity; addictions; alcohol, smoking and cannabis use; electronic device use; employment; sexual factors; sleeping; social support and sun exposure. This model achieves an auROC of 0.69 (0.65-0.72) and deciles' OR of x16 (6-29), and it provides better prediction scores than the diet features' group. This model includes thirty-two diet features from the UKB touchscreen questionnaire on the reported frequency of type and intake of a range of common food and drink items. The diet based model achieved auROC of 0.64 (0.6-0.67) and deciles' OR of x6 (2-17) (Table S1).

<b>Label</b>	<b>APS</b>	<b>AUROC</b>	<b>Deciles fold</b>
<b>All</b>	0.28 (0.22-0.35)	0.90 (0.88-0.92)	65.48 (49-73)
<b>Blood Tests</b>	0.25 (0.20-0.31)	0.89 (0.86-0.90)	62.18 (48-69)
<b>Five blood tests</b>	0.21 (0.17-0.27)	0.87 (0.85-0.89)	59.30 (49-66)
<b>HbA1c%</b>	0.18 (0.13-0.23)	0.81 (0.78-0.84)	31.60 (11-59)
<b>Anthropometry</b>	0.07 (0.06-0.10)	0.76 (0.73-0.79)	29.84 (12-42)
<b>Physical health</b>	0.05 (0.04-0.07)	0.69 (0.65-0.73)	15.44 (5.-29)
<b>Lifestyle and physical activity</b>	0.05 (0.04-0.07)	0.69 (0.65-0.72)	16.40 (6.-29)
<b>Diet</b>	0.04 (0.03-0.05)	0.64 (0.60-0.67)	6.03 (2.-17)
<b>Medication</b>	0.04 (0.03-0.05)	0.63 (0.59-0.67)	5.60 (2.-14)
<b>Family and Ethnicity</b>	0.04 (0.03-0.05)	0.63 (0.58-0.67)	5.35 (2.-14)
<b>BP and HR</b>	0.04 (0.03-0.05)	0.62 (0.58-0.66)	8.54 (3.-20)
<b>Mental health</b>	0.04 (0.03-0.05)	0.62 (0.57-0.66)	6.12 (2.-17)
<b>Socio demographics</b>	0.04 (0.03-0.05)	0.61 (0.57-0.65)	5.19 (2.-15)
<b>Genetics Age and Sex</b>	0.04 (0.03-0.05)	0.61 (0.57-0.65)	5.31(2.-13)
<b>Early Life Factors</b>	0.03 (0.02-0.04)	0.58 (0.54-0.62)	2.57 (1.1-5.9)
<b>Age and Sex</b>	0.03 (0.02-0.04)	0.58 (0.54-0.62)	2.16 (0.9-4.4)
<b>Only genetics</b>	0.03 (0.02-0.04)	0.56 (0.52-0.60)	2.50 (1.1-5.5)

*Table S1: Predicting using features domain groups, Results table of GBDT models for various features domains. The logistic regression models provided better results than the GBDT models for the blood tests and anthropometrics based models.*

We then examined the additive contribution of each predictive group to the total predictive power of the "all features" model (Figure S4-A). We started with the baseline model of "age and sex" and added models one

after the other, sorted by their predictive power as a single features' group using GBDT. We conclude that using the five-blood-test model substantially increase the prediction results when compared to a model based only on HbA1c% with age, and sex.

The auROC, APS and deciles' OR increases from 0.81(0.78-0.84), 0.18(0.13-0.23) and x32(11-59) to 0.87(0.85-0.89), 0.21(0.17-0.27) and x59(49-66) respectively.

When we use the full blood tests model, the performance slightly increases to auROC, APS, and deciles' OR of 0.89(0.86-0.9), 0.25(0.2-0.31), and x62(48-69) respectively. We did not identify any major increase in results adding any other specific group to this list, suggesting that most of the predictive power of our models are captured by the blood test features or has collinearity with it. Using all features together provided an increase of performance to auROC, APS, and the deciles' OR of 0.9(0.88-0.92), 0.28(0.22-0.35), and x65(49-73) respectively (Figure S4 A-B).

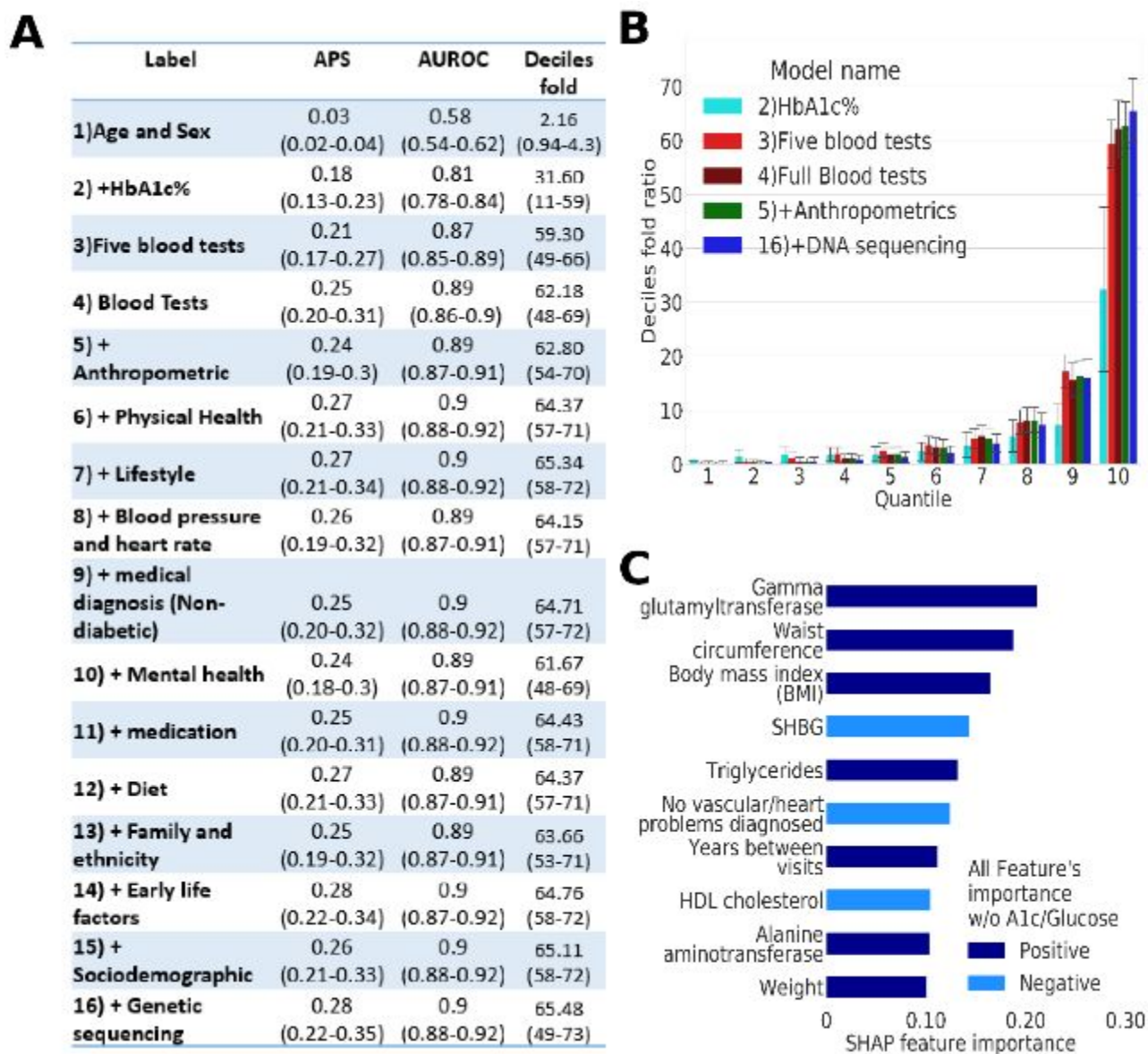
To get some insights for the features that drive the T2D prediction, we built an additional model using all features excluding the HbA1c% and non-fasting glucose, as these features are highly predictive and might screen out other clinically interesting covariates. The auROC, APS, and the deciles' prevalence OR results that we achieved using this model are 0.13 (0.1-0.16), 0.84 (0.82-0.86), and x49(41-56), respectively.

To gain understanding into the features that contribute the most to the GBDT model predictions, we used the SHAP package feature importance framework. SHAP uses Shapley values, which are essentially the average marginal contribution of a feature over all possible combinations (see methods)<sup>31</sup>.

The top ten important features include Gamma-glutamyl-transferase (GGT) (1st in the list), and Alanine-aminotransferase (ALT) (9th in the list), positively related to T2D onset prediction; high levels of these enzymes usually signal for liver damage, which might occur from fatty liver. Increased values of waist circumference (2nd), body mass index (BMI) (3rd), and weight (10th) are all in the ten leading predictors for T2D in this model. Sex hormone-binding globulin (SHBG) (4th), which is a protein produced by the liver and transports sex hormones blood, is also related to a decreased probability of developing T2D in our model. This result aligns with research showing that raised SHBG levels reduce the risk of T2D<sup>34</sup>. High levels of triglycerides (5th), which is long known to precede T2D<sup>35</sup>, also drive the prediction of developing T2D in this model. High levels of HDL (8th) cholesterol reduce the probability of developing T2D in this model, which aligns with the literature findings regarding correlations; although, a genetic mendelian randomization study found it as correlated but non-causal for T2D<sup>36,37</sup>.

Having no vascular or heart problems diagnosed reduces the probability of developing T2D. According to the National Health Interview Survey (NHIS) published in 1989, 14% of the diabetic population at the ages 45-64, also reported having ischemic heart disease, and both conditions are known to be related to the metabolic syndrome<sup>38,39,40</sup>. The time between visits, i.e., time from taking the features measurements to time of prediction is also in the top ten list, which implies that the cumulative probability of developing T2D increases over time.

Two features that are absent from this list are age and sex. While SHBG might confound sex - age seems to be irrelevant once we provide a model with the blood test results as features, and probably reflect a T2D "biological age" of the participant<sup>41,42</sup>(Figure S4-C)



**Figure S4 | Summary of Incremental feature's model:** **A.** Comparison table of APS, auROC, and deciles' OR metrics for GBDT models, where each model is embedding the preceding model's features plus additional features domain. The largest increase in prediction is upon adding the HbA1C% feature, which is also a biomarker of T2D diagnosis. Adding the DNA sequencing data did not contribute much to the prediction power of the model. **B.** Deciles-prevalence-fold-ratio (DFPR) graph. The five blood tests model considerably outperforms the HbA1C% model. **C.** Shapley values feature's importance of the "All features without HbA1c% nor glucose" GBDT model. Removal of the HbA1c% and non-fasting glucose enables other major contributors to T2D prediction to be seen having an impact on T2D prediction.

[References for PRS summary statistics articles.](#)

HbA1c<sup>43,44,45</sup>; Cigarettes per day, ever smoked, age start smoking<sup>46</sup>; HOMA-IR, HOMA-B, diabetes BMI unadjusted, diabetes BMI adjusted, fasting glucose<sup>47</sup>; Fasting glucose, 2 hours glucose level, fasting insulin,

fasting insulin adjusted BMI'-(MAGIC\_Scott)<sup>48</sup>; Fasting glucose, fasting glucose adjusted for BMI, fasting insulin adjusted for BMI<sup>49</sup>; Two hours glucose level<sup>50</sup>; Fasting insulin<sup>51</sup>; Fasting Proinsulin<sup>52</sup>; Leptin adjusted for BMI, Leptin unadjusted for BMI<sup>53</sup>; Triglycerides, Cholesterol, ldl, hdl<sup>54</sup>; BMI<sup>55</sup>; Obesity class1, obesity\_class2, overweight<sup>56</sup>; Anorexia<sup>57</sup>; Height<sup>58</sup>; Waist circumference, hips circumference<sup>59</sup>; Cardio<sup>60</sup>; Heart\_Rate<sup>61</sup>; Alzheimer<sup>62</sup>; Asthma<sup>63</sup>

## Bibliography

1. Zimmet, P., Alberti, K. G., Magliano, D. J. & Bennett, P. H. Diabetes mellitus statistics on prevalence and mortality: facts and fallacies. *Nat. Rev. Endocrinol.* **12**, 616–622 (2016).
2. International Diabetes Federation - Type 2 diabetes. at  
<<https://www.idf.org/aboutdiabetes/type-2-diabetes.html>>
3. WHO | Diabetes programme. at  
<<https://web.archive.org/web/20140329084830/http://www.who.int/diabetes/en/>>
4. Home | ADA. at <<https://www.diabetes.org/>>
5. Knowler, W. C. *et al.* Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N. Engl. J. Med.* **346**, 393–403 (2002).
6. Lindström, J. *et al.* Sustained reduction in the incidence of type 2 diabetes by lifestyle intervention: follow-up of the Finnish Diabetes Prevention Study. *Lancet* **368**, 1673–1679 (2006).
7. Diabetes Prevention Program Research Group. Long-term effects of lifestyle intervention or metformin on diabetes development and microvascular complications over 15-year follow-up: the Diabetes Prevention Program Outcomes Study. *Lancet Diabetes Endocrinol.* **3**, 866–875 (2015).
8. Noble, D., Mathur, R., Dent, T., Meads, C. & Greenhalgh, T. Risk models and scores for type 2 diabetes: systematic review. *BMJ* **343**, d7163 (2011).
9. Collins, G. S., Mallett, S., Omar, O. & Yu, L.-M. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med.* **9**, 103 (2011).
10. Kengne, A. P. *et al.* Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a



- validation of existing models. *The Lancet Diabetes & Endocrinology* **2**, 19–29 (2014).
11. Bernabe-Ortiz, A., Perel, P., Miranda, J. J. & Smeeth, L. Diagnostic accuracy of the Finnish Diabetes Risk Score (FINDRISC) for undiagnosed T2DM in Peruvian population. *Prim. Care Diabetes* **12**, 517–525 (2018).
  12. Lindström, J. & Tuomilehto, J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes Care* **26**, 725–731 (2003).
  13. Meijnikman, A. S., De Block, C. E. M., Verrijken, A., Mertens, I. & Van Gaal, L. F. Predicting type 2 diabetes mellitus: a comparison between the FINDRISC score and the metabolic syndrome. *Diabetol. Metab. Syndr.* **10**, 12 (2018).
  14. Schulze, M. B. *et al.* An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. *Diabetes Care* **30**, 510–515 (2007).
  15. Mühlenbruch, K. *et al.* Update of the German Diabetes Risk Score and external validation in the German MONICA/KORA study. *Diabetes Res. Clin. Pract.* **104**, 459–466 (2014).
  16. Eckel, R. H., Grundy, S. M. & Zimmet, P. Z. The metabolic syndrome. *Lancet* **365**, 1415–1428 (2005).
  17. Cheng, C.-H. *et al.* Waist-to-hip ratio is a better anthropometric index than body mass index for predicting the risk of type 2 diabetes in Taiwanese population. *Nutr. Res.* **30**, 585–593 (2010).
  18. Jafari-Koshki, T., Mansourian, M., Hosseini, S. M. & Amini, M. Association of waist and hip circumference and waist-hip ratio with type 2 diabetes risk in first-degree relatives. *J. Diabetes Complicat.* **30**, 1050–1055 (2016).
  19. Qiao, Q. & Nyamdorj, R. Is the association of type II diabetes with waist circumference or waist-to-hip ratio stronger than that with body mass index? *Eur. J. Clin. Nutr.* **64**, 30–34 (2010).
  20. Fekete, T. & Sopon, E. Glycaemic control and reticulocyte count in diabetic patients. *Horm. Metab. Res.* **18**, 141 (1986).
  21. Kontush, A. & Chapman, M. J. Why is HDL functionally deficient in type 2 diabetes? *Curr. Diab. Rep.* **8**, 51–59 (2008).

22. Bitzur, R., Cohen, H., Kamari, Y., Shaish, A. & Harats, D. Triglycerides and HDL cholesterol: stars or second leads in diabetes? *Diabetes Care* **32 Suppl 2**, S373-7 (2009).
23. Understanding A1C | ADA. at <<https://www.diabetes.org/a1c>>
24. Diabetes Prevalence 2019 | Diabetes UK. at <<https://www.diabetes.org.uk/professionals/position-statements-reports/statistics/diabetes-prevalence-2019>>
25. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
26. Hernán, M. A., Hernández-Díaz, S. & Robins, J. M. A structural approach to selection bias. *Epidemiology* **15**, 615–625 (2004).
27. Alex, F., ALEX, G., Bertr, R. G. F., BERTR, T. & THIRION. Scikit-learn: Machine Learning in Python.
28. Rufibach, K. Use of Brier score to assess binary predictions. *J. Clin. Epidemiol.* **63**, 938–9; author reply 939 (2010).
29. Davidson-Pilon, C. *et al.* CamDavidsonPilon/lifelines: v0.24.16. *Zenodo* (2020).  
doi:10.5281/zenodo.3937749
30. Ke, G. *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree. (2017).
31. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. (2017).
32. Lundberg, S. M. *et al.* From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
33. Ali, O. Genetics of type 2 diabetes. *World J. Diabetes* **4**, 114–123 (2013).
34. Perry, J. R. B. *et al.* Genetic evidence that raised sex hormone binding globulin (SHBG) levels reduce the risk of type 2 diabetes. *Hum. Mol. Genet.* **19**, 535–544 (2010).
35. Kannel, W. B. Lipids, diabetes, and coronary heart disease: insights from the Framingham Study. *Am. Heart J.* **110**, 1100–1107 (1985).
36. Femlak, M., Gluba-Brzózka, A., Ciałkowska-Rysz, A. & Rysz, J. The role and function of HDL in patients

- with diabetes mellitus and the related cardiovascular risk. *Lipids Health Dis.* **16**, 207 (2017).
37. Haase, C. L., Tybjærg-Hansen, A., Nordestgaard, B. G. & Frikke-Schmidt, R. HDL cholesterol and risk of type 2 diabetes: A mendelian randomization study. *Diabetes* **64**, 3328–3333 (2015).
  38. Wannamethee, S. G., Shaper, A. G., Lennon, L. & Morris, R. W. Metabolic syndrome vs Framingham Risk Score for prediction of coronary heart disease, stroke, and type 2 diabetes mellitus. *Arch. Intern. Med.* **165**, 2644–2650 (2005).
  39. Micha, R. & Mozaffarian, D. Trans fatty acids: effects on metabolic syndrome, heart disease and diabetes. *Nat. Rev. Endocrinol.* **5**, 335–344 (2009).
  40. Aubert, R. *Diabetes in America*. (DIANE Publishing, 1995).
  41. Sebastiani, P. *et al.* Biomarker signatures of aging. *Aging Cell* **16**, 329–338 (2017).
  42. Jylhävä, J., Pedersen, N. L. & Hägg, S. Biological Age Predictors. *EBioMedicine* **21**, 29–36 (2017).
  43. Soranzo, N. *et al.* Common variants at 10 genomic loci influence hemoglobin A<sub>1c</sub> levels via glycaemic and nonglycaemic pathways. *Diabetes* **59**, 3229–3239 (2010).
  44. Walford, G. A. *et al.* Genome-Wide Association Study of the Modified Stumvoll Insulin Sensitivity Index Identifies BCL2 and FAM19A2 as Novel Insulin Sensitivity Loci. *Diabetes* **65**, 3200–3211 (2016).
  45. Wheeler, E. *et al.* Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. *PLoS Med.* **14**, e1002383 (2017).
  46. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet.* **42**, 441–447 (2010).
  47. Morris, G. P. *et al.* Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc Natl Acad Sci USA* **110**, 453–458 (2013).
  48. Scott, R. A. *et al.* Large-scale association analyses identify new loci influencing glycaemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44**, 991–1005 (2012).
  49. 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).
50. 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).
51. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
52. Strawbridge, R. J. *et al.* Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes* **60**, 2624–2634 (2011).
53. Kilpeläinen, T. O. *et al.* Genome-wide meta-analysis uncovers novel loci influencing circulating leptin levels. *Nat. Commun.* **7**, 10494 (2016).
54. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
55. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
56. Berndt, S. I. *et al.* Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat. Genet.* **45**, 501–512 (2013).
57. Boraska, V. *et al.* A genome-wide association study of anorexia nervosa. *Mol. Psychiatry* **19**, 1085–1094 (2014).
58. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
59. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).
60. CARDIoGRAMplusC4D Consortium *et al.* Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat. Genet.* **45**, 25–33 (2013).
61. den Hoed, M. *et al.* Identification of heart rate-associated loci and their effects on cardiac conduction and

- rhythm disorders. *Nat. Genet.* **45**, 621–631 (2013).
62. Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
63. Moffatt, M. F. *et al.* A large-scale, consortium-based genomewide association study of asthma. *N. Engl. J. Med.* **363**, 1211–1221 (2010).

## Acknowledgements:

This research has been conducted using the UK Biobank Resource under Application Number 28784