Machine learning algorithms in big data analyses identify determinants of insulin gene transcription

Wilson K.M. Wong[1,6], Vinod Thorat[2,6], Mugdha V. Joglekar[1,6], Charlotte X. Dong[1], Hugo Lee[3], Adwait Bhave[2], Feyza Engin[3,4], Aniruddha Pant[2], Louise T. Dalgaard[4], Sharda Bapat[2,6,7,*] and Anandwardhan A. Hardikar[1,4,6,7,*]

Running title: Machine-learning algorithms in biology

1- Diabetes and Islet Biology Group, School of Medicine, Western Sydney University, 30.1.27 Cnr Goldsmith and David Pilgrim Ave, Campbelltown, NSW 2560, Australia
2- AlgoAnalytics, Level 2, Ideas to Impacts Building, Pallod Farms, Baner, Pune, 411045, India
3- Department of Biomolecular Chemistry, University of Wisconsin-Madison, School of Medicine and Public Health, Madison, WI 53705, USA
4- Department of Medicine, Division of Endocrinology, Diabetes & Metabolism, University of Wisconsin-Madison, School of Medicine and Public Health, Madison, WI 53706, USA
5- Department of Science and Environment, Roskilde University, Universitetsvej 1, DK-4000 Roskilde, Denmark.

[6]: Contributed equally to this manuscript.

[7]: equal senior authors

* Address all correspondence to:

Dr. Sharda Bapat
Level 2, Alacrity India Innovation Center, Ideas to Impacts Building, Pallod Farms Lane 3, Baner, Pune, 411045, India.
E-mail: sbapat@algoanalytics.com

or
A/Prof. Anandwardhan A. Hardikar, PhD
Diabetes and Islet Biology Group, School of Medicine, Western Sydney University
30, Cnr Goldsmith and David Pilgrim Ave,
Campbelltown, NSW 2560, AUSTRALIA
Phone: +61 488119772
E-mail: A.Hardikar@westernsydney.edu.au
Web: https://www.isletbiology.info/

Total word count: 2582

Significance statements:

- A comparative analysis of 10 different machine learning (ML) approaches identifies most accurate and sensitive algorithms in big data (>490 million datapoints) analyses;
- Ensemble ML (Random Forest, ADA Boost and Gradient boost) models trained on 392,684,052 datapoints from human pancreatic scRNA-seq studies, identify features (genes) that predict insulin transcription in a separate (N=2,913) dataset;
- *In silico* validation of identified features confirmed predictors of insulin gene transcription;
- Known and novel key variables associated with insulin transcription are dysregulated in a Type 1 diabetes mouse model that shows transient β-cell dedifferentiation and in β-cells of individuals with type 2 diabetes.

Abstract

Machine learning (ML) workflows enable unprejudiced/robust evaluation of complex datasets and are increasingly sought in big data analyses. Here, we analyzed over 490,000,000 data points to compare 10 different ML algorithms in a large (N=11,652) training dataset of human pancreatic single-cell transcriptomes to identify features (genes) associated with the presence or absence of insulin transcript(s). Prediction accuracy/sensitivity of models were tested in a separate validation dataset (N=2,913) and the performance of each ML-workflow assessed. Overall, Ensemble ML workflows, and in particular, Random Forest ML algorithm delivered high predictive power in a receiver operator characteristic (ROC) curve analysis (AUC=0.83) at the highest sensitivity (0.98), compared to other algorithms. The top-10 features, (including *IAPP*, *ADCYAP1*, *LDHA* and *SST*) common to the three Ensemble ML workflows were significantly dysregulated in scRNA-seq datasets from Ire-1α$^{\beta-/-}$ mice that demonstrate de-differentiation of pancreatic β-cells as well as in pancreatic single cells from individuals with Type 2 Diabetes. Our findings provide a direct comparison of ML workflows in big data analyses, identify key determinants of insulin transcription and provide workflows for other regulatory analyses to identify/validate novel genes/features of endocrine pancreatic gene transcription.

Recent years have witnessed a surge in single-cell transcriptomic technologies; many already generating newer data and insights to address specific biological questions. Machine learning (ML) algorithms offer an unbiased mathematical workflow that facilitates the identification of complex relationships across variables. ML workflows involve an orderly set of instructions using automated, unbiased 'learning' processes usually targeted towards developing (training) a model that can be validated in a separate (test) dataset[1]. One goal of ML algorithms is to analyze big data to identify variables that cannot be recognized through conventional biostatistical techniques.

Currently, several ML algorithms are available to researchers handling big data in omics-based high content analyses. These can be broadly divided into two categories: supervised

and unsupervised algorithms[2]. Supervised methods (such as decision tree) derive relationships between one dependent and multiple independent variables using a training set and then apply that knowledge in the testing set for predictive/efficacy analysis. Unsupervised methods derive patterns/data clusters amongst all available variables. ML algorithms have been used to unravel patterns or to build associations or for predictions in several biological processes such as determining DNA methylation states in single cells[3], identifying signatures of lipid or metabolite species[4,5] or microRNAs[6] in predicting transition from gestational diabetes to type 2 diabetes as well as in genetic studies[7]. There are multiple ML algorithms available and it may present a challenge to select the most appropriate method for a particular dataset to answer a specific question. We, therefore, decided to compare different ML methodologies to (i) rank different ML methods for their performance on a large dataset (of 490,855,065 scRNA-sequencing data points) and (ii) understand the most important variables associated with insulin transcription.

Previous studies[8-11] from several laboratories have identified master regulatory transcription factors that regulate the embryonic development of insulin-producing islet β-cells. Although transcription factor-mediated insulin transcription regulation is a well-known mechanism during the development of insulin-producing cells, it is also recognized that active genes localized on different chromosomal regions can dynamically regulate gene transcription in post-natal life[12]. One approach to identify genes associated with insulin gene transcription is through single-cell (sc)RNA-seq-based big data analysis.

Here, we examined the performance of 10 different ML algorithms in a curated human pancreatic single-cell sequencing dataset of 490,855,065 data points (N=14,565 single cells and 33,701 expressed gene features). The aims of this study were (i) to provide a comparative account of the predictive potential of 10 different commonly used ML workflows

(**Supplementary Table 1**), and (ii) to use existing scRNA-seq datasets in identifying genes (variables) associated with or important for determining insulin transcript-containing cells.

Results

Machine learning (ML) algorithms yield varying performance outputs.

The scRNA-seq data were obtained from public databanks (GSE84133, GSE85241, E-MTAB-5061, GSE83139, GSE81608) of human pancreatic single-cell transcriptomes. We first randomized this available pancreatic scRNA-seq transcriptomic data and allocated 80% of samples to a discovery/training set (Training; N=11,652 samples) and remaining into a validation/testing set (Test; N=2,913 samples) as outlined in **Figure 1**. With the availability of several ML algorithms (**Supplementary Table 1**), we probed the discovery data using 10 different ML workflows (**Figure 2A**) to identify features highly associated with the presence of insulin transcripts in a single cell. Genes (features) identified as the most important/predictive variables for each of these ML workflows were used to identify insulin transcript-containing cells from the validation set (remainder 20% of the samples). Validation results of the identified gene features from each of the 10 ML workflows are presented in the form of a receiver operator characteristic (ROC) curve (**Figure 2B**). The top three ML algorithms; Gradient boosting, Random Forest and ADA boost (all Ensemble workflows), demonstrated similar performance returning an AUC of between 0.83 – 0.86. A confusion matrix is presented below each ROC curve dataset (**Figure 2B**) to present the false-positive and false-negative predictions within every workflow. These analyses show that although Ensemble machine learning workflows are the best in predicting insulin-transcribing cells, other workflows, such as logistic regression, also perform closely similar to the Ensemble methods.

Ensemble ML workflows to identify genes associated with insulin transcription.

The scRNA-seq datasets obtained from public databanks of human pancreatic single-cell transcriptomes were classified as insulin-transcribing (1) or those with no insulin (0) (**Figure 3A**). As described earlier, all the three Ensemble ML workflows presented with an AUC that was better than any of the other ML workflows tested in our ROC curve analysis. Ensemble workflows also presented with high accuracy ($\geq$87%), precision ($\geq$0.89), and sensitivity (>0.95), which was comparable to other popular workflows such as logistic regression (**Figure 3B**). As Ensemble ML workflows such as Random Forest use a collection of decision trees (forest), we decided to compare the performance of the top three (Ensemble) workflows to a single (Decision tree) algorithm. The relative contribution of the top 10 features (genes) from each of these ML workflows are presented as radar plots (**Figure 3C**), whilst the longer list of genes ranked by their importance is presented in **Supplementary Table 2**. *IAPP*, *ADCYAP1*, *LDHA* and *SST* were common to all three Ensemble workflows. To compare the expression of these features (genes) identified through each of the Ensemble and Decision Tree classifier, we examined the expression of genes identified to be associated with insulin transcription to those in a separate islet β-cell dataset (**Figure 3D**).

Insulin-associated genes are dysregulated during β-cell dedifferentiation.

Dedifferentiation of β-cells, characterized by the loss of expression of key β-cell maturation marker genes with an accompanying reduction in insulin secretion, has been observed in mouse models of type 1 (T1D) and type 2 (T2D) diabetes, as well as in individuals with diabetes[13-16]. We questioned if the expression of gene variables identified and validated (in silico) as being predictive of insulin gene transcription (**Figure 3B**) are dysregulated in a mouse model of T1D with evidence of islet dedifferentiation. Transient dedifferentiation of

islet β-cells was recently reported in an established T1D preclinical mouse model upon β-cell-specific deletion of a key stress response gene, *Ire1α*, (Ire1α$^{β-/-}$)[17]. These mice also demonstrated reduced β-cell number as well as diminished expression of insulin transcripts in β-cells compared to control (Ire-1α$^{fl/fl}$) mice. Therefore, we evaluated the expression of 25 gene transcripts that made up the top-10 features across the four different ML workflows (**Figure 3C**). Twelve of these features were not significantly regulated between Ire1α$^{β-/-}$ and Ire1α$^{fl/fl}$ islets. However, the remaining thirteen features (listed in **Figure 3C**) were significantly dysregulated in β-cells of Ire-1α$^{β-/-}$ mice that were undergoing dedifferentiation. De-differentiating β-cells showed significant downregulation of five key genes; *Iapp, MafA, Pcsk1n*, *Atp5e* and *Ldha*, whilst all other insulin-associated gene transcripts showed significantly higher levels. In Type 2 diabetes (T2D) it is known that *INS* transcript expression is reduced, therefore we validated the top gene features common (*IAPP*, *SST*, *MAFA*, *ADCYAP1* and *LDHA*) in three of the ML workflows analysed using a separate publicly available single-cell RNA-seq dataset from non-diabetic (ND) vs T2D adult human pancreas (GSE154126[18]). Four of the five genes (*IAPP*, *SST*, *MAFA*, *ADCYAP1*), were significantly lower in T2D insulin-transcribing cells compared to ND insulin-transcribing cells (**Supplementary Table 3**).

Discussion:

In this study, we compared the performance characteristics of 10 different ML algorithms, (**Supplementary Table 1**) that are currently used in big data analyses. We analyzed a scRNA-seq dataset that was randomly split to a larger (80%; 392,684,052 data points) training set involving model learning, and then a smaller (20%; 98,171,013 data points) validation set. All algorithms identified a set of genes (features) that identify insulin-

production (1) defined as the presence of one or more transcripts of insulin in a sample, or no insulin production (0) from the 11,652 single cells analysed in the training test. We validated the predictive features identified through each ML workflow in the validation/test set of 2,913 single cell transcriptomes. ML workflows that returned high performance (based on AUC, sensitivity/specificity) were selected and the top 10 genes (ranked by their importance) in each of those ML methods were re-validated in discrete mouse and human datasets that model beta cell dedifferentiation (**Figure 3D, Supplementary Table 3**).

Our analysis provides two major outcomes that are of interest to a broad range of data analysts and biologists. First, a comparison of the ML algorithms identified Ensemble-based ML methods as the best performing algorithms in our analyses. Logistic regression performed closest to Ensemble methods, in line with previous reports in clinical datasets[19]. We then compared Ensemble methodologies to Decision tree algorithm. Decision tree offers the often-desired simplistic model generation method as compared to Ensemble methods such as Random Forest. However, the latter builds multiple decision trees independently and offers an overall learning model that is closest to the best possible prediction. Indeed, Decision tree was determined to be a weaker predictor than the Random Forest as the latter reduces variance using different sample sets (bootstrap) in training, randomizing feature subsets, and combining the predictive learning by building multiple decision trees. Random Forest prediction outcomes were similar to gradient boosting, which also builds a set of decision trees, but one tree at a time. The bagging and boosting approach used in ADA/Gradient boosting methods seems to have offered better accuracy and performance in insulin prediction analysis than those observed using Random Forest, whereas the Random Forest algorithm offered the highest sensitivity (**Figure 3B**) amongst all methodologies tested.

The other outcome from this analysis is the identification of genes that are associated with and predictive of insulin gene transcription. Since bulk RNA-sequencing studies do not offer

the desired single-cell resolution to identify transcriptional regulation at a cellular level, our analyses provide a firsthand view of insulin gene transcriptional determinants identified through an unbiased, big data machine learning approach. The top three methodologies (based on high AUC values) belonged to Ensemble machine learning workflow. Weighted relative importance of the top-10 most important features are compared (**Figure 3C)**, and a longer list of all the gene features identified through the four ML workflows is provided in **Supplementary Table 2**. Interestingly, five genes were common to the top 10 features from all of the algorithms compared – *IAPP, ADCYAP1, MAFA, SST* and *LDHA*. The top-ranked gene associated with insulin gene transcription across all the Ensemble workflows was *IAPP*. Islet amyloid polypeptide (*IAPP*) and insulin are known to be expressed in pancreatic islet β-cells and co-secreted in response to changes in glucose concentration[20,21]. Their mRNA levels are also regulated by glucose. The promoters of both these genes share similar cis-acting sequence elements, and both bind the master regulatory transcription factor *PDX1*[20]. *FoxA2* (*HNF-3β*) negatively regulates *IAPP* promoter activity[22] and has also been shown to suppress insulin gene expression[23]. Although insulin gene expression is known to be regulated by several islet-enriched transcription factors, *MafA* is the most well recognized β-cell-specific activator of insulin gene expression[24]. The selection of *MAFA* as a key feature by three of the compared ML approaches tested through this analysis is therefore not surprising. The inclusion of *SST* in the top three gene features is intriguing. Somatostatin expression is known to be important in control of insulin release and ablation of somatostatin-expressing delta cells impairs pancreatic islet function and cause neonatal death in rodents[25]. SST analogs were shown to inhibit the release of insulin via the activation of both ATP sensitive K+ channels and G protein-coupled inward rectifier K+ channels[26]. Another candidate that was identified through these analyses is *MTRNR2L8,* a neuroprotective and antiapoptotic peptide derived from a portion of the mitochondrial *MT-RNR2* gene and reported in fetal as

well as adult beta cells[27]. *ADCYAP1* stimulates insulin secretion in a glucose-dependent manner[28] and genetic screening in type 2 diabetes Caucasians indicated the presence of two SNPs in exons 3 and 5 to be associated with type 2 diabetes[29]. Finally, *LDHA*, which was also selected through these unbiased analyses across the top-three ML workflows is a pancreatic β-cell disallowed gene[30-32] and human *LDHA* levels are predictive of insulin transcription[33]. Together, these algorithms help in identifying a set of genes expressed in or disallowed from insulin-producing pancreatic β-cells.

Mouse models often provide the validation to understand mechanisms that cannot be tested in through human studies. The Ire1α$^{β-/-}$ mouse model offers a unique model, wherein pancreatic β-cells transiently dedifferentiate during early post-natal life, allowing these knockout mice escape immune-mediated β-cell destruction and T1D in later life[17]. Analysis of islet single cell sequencing data from this model identified genes that were significantly dysregulated in β-cells of Ire1αβ-/- mice when compared to control (Ire1α$^{fl/fl}$) mice. Eight of thirteen features (from the Top 10 features in each of the four ML workflows, **Figure 3E**), which showed significant dysregulation between Ire1α$^{β-/-}$ and Ire1α$^{βfl/fl}$ mice are upregulated in β-cells of Ire1α$^{β-/-}$ mice. T2D islet single cell validation also revealed down-regulation of four common gene features (*IAPP*, *SST*, *MAFA* and *ADCYAP1* identified across our three top ML workflows) in T2D compared to ND insulin transcribing cells. Interestingly, Delta Like Non-Canonical Notch Ligand 1 (*DLK1*) was also significantly downregulated in T2D compared to ND insulin transcribing cells. The imprinted region of chromosome 14q32.2, contains microRNA cluster of *DLK1-MEG3* which are highly expressed and more specific in human β-cells compared to α-cells. Previous study had also shown that in Type 2 diabetes (T2D) human islets, the *MEG3*-microRNA locus expression levels are significantly lower[34]. The 14q32 locus of microRNAs (such as coexpression of miR-376a and miR-432) also have been shown to target and suppress the expression of *IAPP*[34].

Strength and Limitation: This is a first demonstration comparing multiple ML algorithms to identify key genes associated with insulin transcription using a large dataset of over 490 million data points. As anticipated, Ensemble methods perform better than most other workflows and identified a set of genes that corroborate with previous reports of transcriptional regulation of insulin in mouse and human β-cells. These findings indicate that unbiased ML workflows for big data analyses can generate biologically meaningful results, when applied to large training datasets. Our study provides the codes/scripts for other researchers to use in existing as well as emerging datasets for identification of gene candidates associated with other genetic pathways (eg. related to *GCG* or *GCK*) in future or to genes recognized to be associated with Type 2 diabetes GWAS datasets. We recognize that there are several limitations: we are unsure as to why some other well know candidates (such as *Pdx1*, and *NeuroD*) were not selected by our top predictive models. An explanation is that we used a whole pancreatic single cell dataset and that the predictive models generated through filtering out β-cells may be more enriched for known pro-endocrine gene regulators such as PDX1. The other explanation is that although PDX1 is a key regulator, the transcript levels analysed in these datasets using multiple scRNA-seq technologies may not be sufficient considering the sequencing depth offered by some of these scRNA-seq workflows.

We recognize that exhaustive (eg. LOOCV[35]) as well as non-exhaustive cross-validation approaches (such as K-fold cross-validation[36]) were not performed here. Such cross-validation approaches, although useful in assessing how results will generalize to an independent dataset, are mostly used in the validation of much smaller datasets. In big data analyses, the use of such cross-validation methodologies would limit the analyses to only those with an access to high-end cluster computing. The 10 different ML scripts used in these analyses are designed to work on a high-end personal computing device (i7 processor

with 4 cores and 32GB RAM or better). We believe that the application of such ML algorithms to the expanding scRNA-seq datasets would lead to the confirmation/validation of current as well as identification of novel determinants of gene transcription, thereby accelerating innovation in discovery of gene targets in biology and medicine.

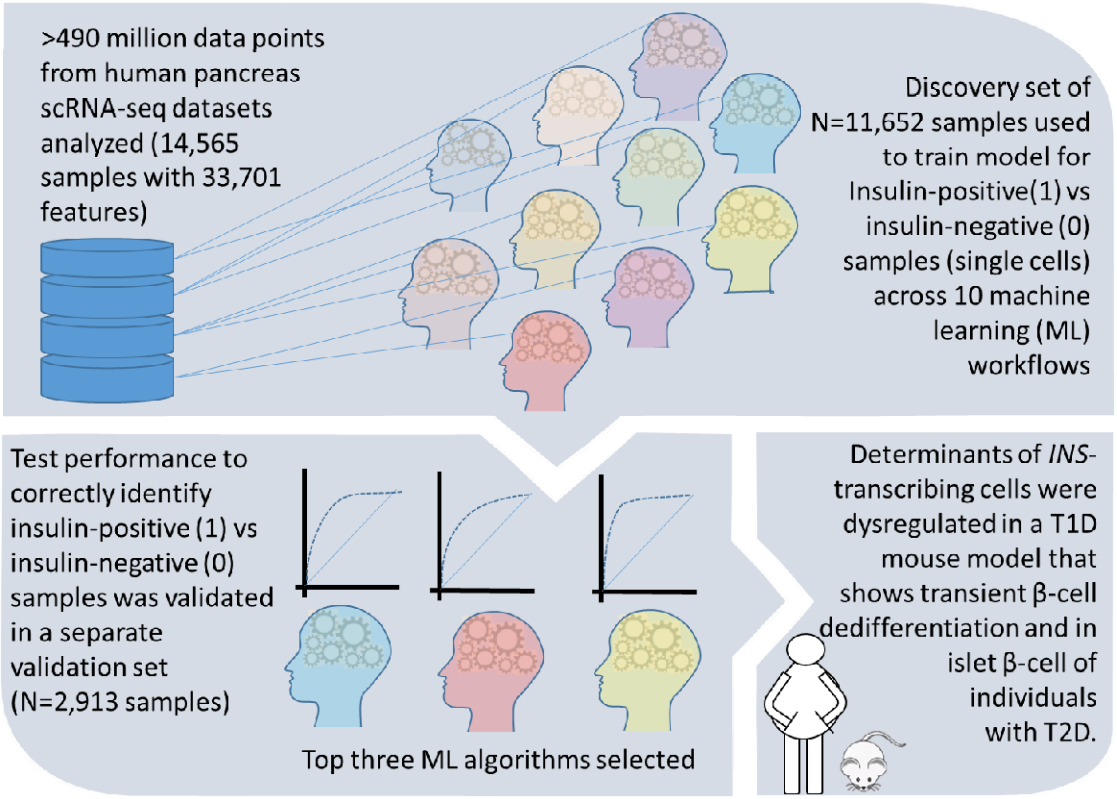
Acknowledgements:

The research presented herein has been funded through grants from the Australian Research Council Future Fellowship (FT110100254) the Juvenile Diabetes Research Foundation (JDRF) Australia T1D Clinical Research Network (JDRF/4-CDA2016-228-MB) and the Visiting Professorships (2016-18 and 2019-22) from the Danish Diabetes Academy, funded by the Novo Nordisk Foundation, grant number NNF17SA0031406 to AAH. WKMW is suppoted through JDRF Australia/Helmsley Charitable Trust innovation grant (to AAH). MVJ is supported through a JDRF International Advanced Post-doctoral award (3-APF-2016-178-A-N) and currently a transition award from JDRFI. AAH acknowledges the initial interactions with Tune Pers, University of Copenhagen, Denmark. AAH acknowledges the support through the Ainsworth Medical Research Fund, Western Sydney University School of Medicine, Australia. FE is supported by a grant from the JDRF-5-CDA-2014-184-A-N. HL is supported by NIH National Research Service Award T32 GM007215.

Author contributions

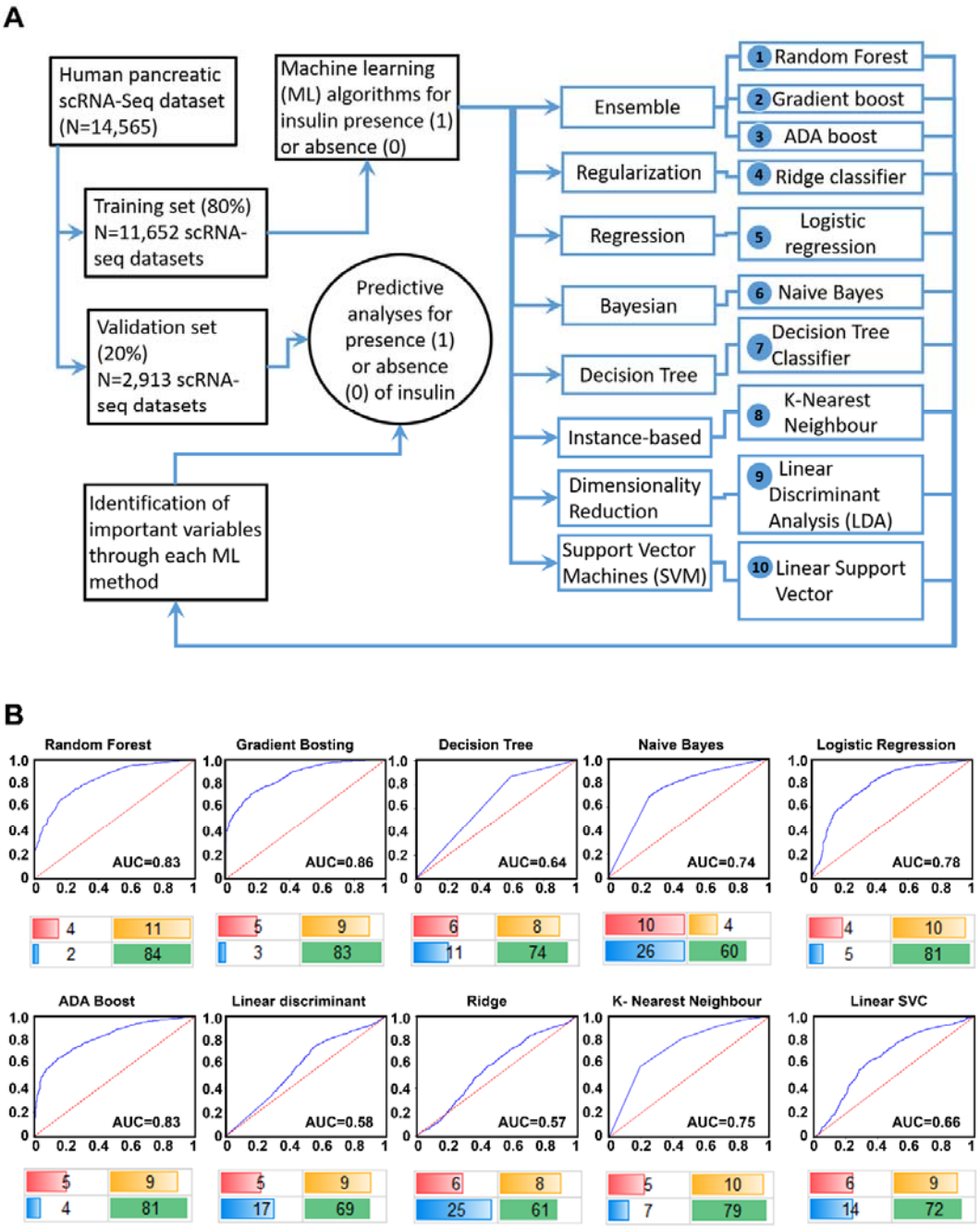
Conceptualization: AAH; Methodology: AAH, AP, SB, LTD; Software: VT, MVJ, WKMW, CXD, HL, LTD; Validation: WKMW, MVJ, VT, CXD, AB; Data Curation: WKMW, MVJ, VT, CXD, LTD, AAH; Writing – Original Draft: AAH, MVJ, WKMW, LTD; Review & Editing: All authors; Visualization: AAH; Supervision: AAH; Project Administration: AAH; Funding Acquisition: AAH.

Figure 1: Comparative analysis of scRNA-seq datasets



Five different human pancreatic single-cell(sc)RNA-sequencing datasets (GSE84133, GSE85241, E-MTAB-5061, GSE83139, GSE81608) were curated to test the performance of 10 different machine learning (ML) algorithms. A training set comprising randomly selected 80% datasets (N=11,652 samples) were used in the learning phase (Training set; top panel). Ten ML workflows used in this study are symbolically illustrated as coloured heads. Each of the workflows aimed to identify a list of weighted features that accurately identify the presence (1; "insulin-positive") or absence (0;insulin-negative) of insulin gene transcripts. Learning outcomes were validated (bottom-left panel) in 20% (N=2,913) of the total samples, and the top three ML algorithms providing the best area under the curve (AUC) in a receiver operator characteristic (ROC) curve analysis were selected to understand the importance of the selected features (genes) in each ML-workflow. Since our ML workflows identified genes associated with insulin gene transcription, we analyzed the expression of these genes in the Ireα$^{\beta-/-}$ mouse pancreatic scRNA-seq dataset (GSE144471) profiling dedifferentiating pancreatic β-cells.

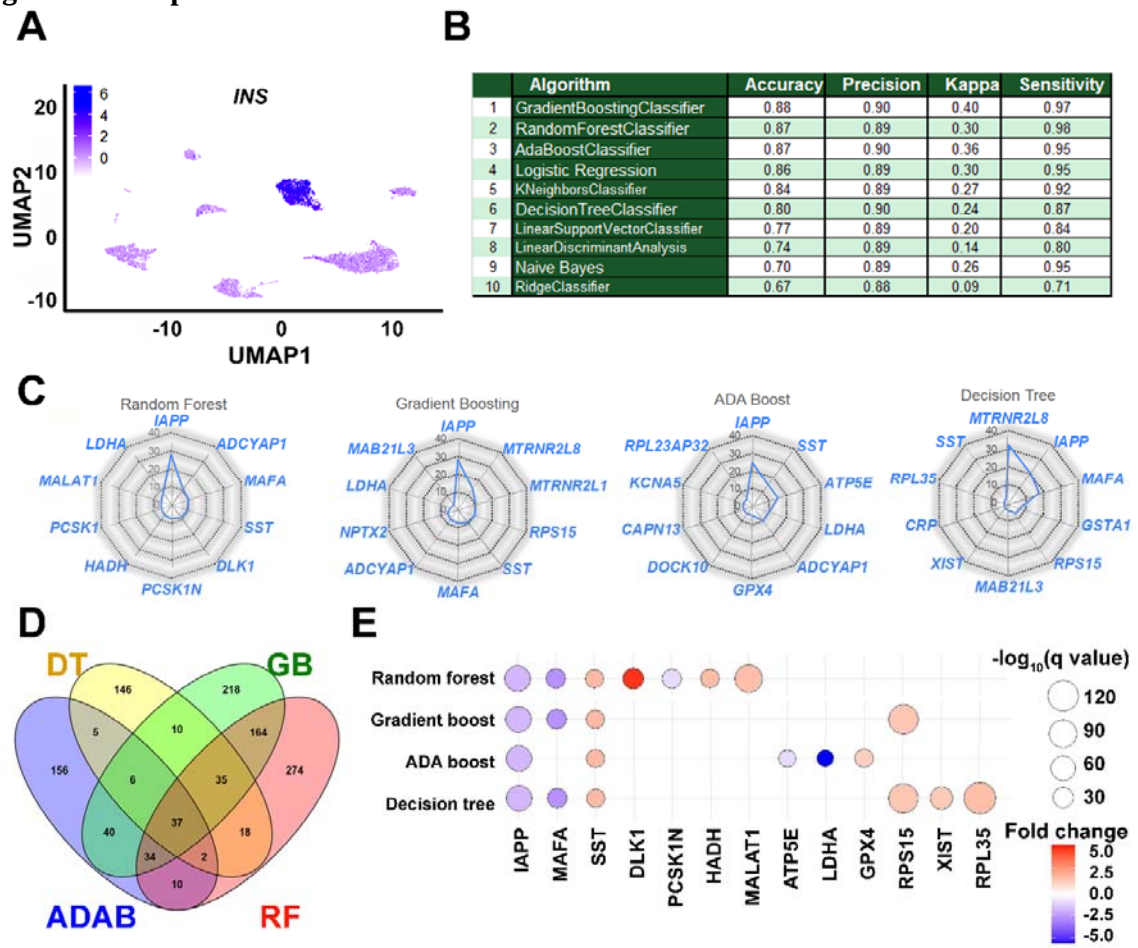
Figure 2: Study design and performance of different ML workflows.



A flowchart of our analytical plan is presented in (A). Previously published datasets of single-cell RNA-sequencing analyses from pancreatic islet cell preparations were randomly divided into a training (N=11,652) and a validation (N=2,913) set. The learning phase (Training) involved identifying features (genes) and their associated weights/coefficients in each of the ten machine learning (ML) methods (listed 1-10) used. Weighted features were used in the prediction of insulin transcription (across 10 ML algorithms) to test the performance of these models in an independent validation set of samples (N=2,913). ROC

curve plots for each ML algorithm using validation set data are presented in (B). The area under the curve (AUC) for the tested workflows are presented along with a confusion matrix below the plot. Percent values are rounded off to the nearest integer (and hence may not sum up to an absolute 100%) and represent true negative (red), true positive (green), false positive (yellow) and false negative (blue) samples identified in the validation set.

Figure 3: Performance and application of learned features in understanding insulin gene transcription.



	Algorithm	Accuracy	Precision	Kappa	Sensitivity
1	GradientBoostingClassifier	0.88	0.90	0.40	0.97
2	RandomForestClassifier	0.87	0.89	0.30	0.98
3	AdaBoostClassifier	0.87	0.90	0.36	0.95
4	Logistic Regression	0.86	0.89	0.30	0.95
5	KNeighborsClassifier	0.84	0.89	0.27	0.92
6	DecisionTreeClassifier	0.80	0.90	0.24	0.87
7	LinearSupportVectorClassifier	0.77	0.89	0.20	0.84
8	LinearDiscriminantAnalysis	0.74	0.89	0.14	0.80
9	Naive Bayes	0.70	0.89	0.26	0.95
10	RidgeClassifier	0.67	0.88	0.09	0.71

(A) A 2D clustering of pancreatic single cells assessed in this study using UMAP (Uniform Manifold Approximation and Projection plot). Cellular subtypes based on the UMAP clustering algorithm are labelled and graded (scale, inset) as per the level of insulin gene transcripts. (B) The performance of learnt models on accurately identifying insulin-positive (1) and insulin-negative (0) single cells from the validation dataset are presented. (C) Relative weighted rank contributions of the top 10 genes in each of the four listed ML algorithms are presented as spider plots plotted in the order of importance (starting clockwise at 12-O'clock position). Percent representation of each of the genes indicates their relative contribution in the set on the spider plot with a logarithmic scale (center=1% and outer circle=100%). A comparison of the gene features identified by the top three ensemble workflows is presented along with those identified by the Decision Tree classifier. (D) Pathways targeted by features from each of the four selected ML methods (RF: Random Forest; GB: Gradient Boosting; ADAB: ADA Boost; DT: Decision Tree) identified using gene ontology (GO) function analysis are presented in the Venn diagram. Number of GO terms enriched and common for top features (genes) in each ML method are plotted. (E) All significantly dysregulated genes identified from and common to the four ML algorithms presented herein were assessed in the scRNA-seq dataset from Ire1α$^{\beta-/-}$ mice. Bubble plot presenting fold-change and statistical significance (q-value) for each of the genes in Ire1α$^{fl/fl}$ and Ire1α$^{\beta-/-}$ mice are shown. Blue

colour represents downregulation while red colour indicates increased abundance of transcripts in Ire1α$^{\beta-/-}$ mice compared to control.

Methods

Pancreatic single-cell (sc)RNA sequencing datasets and analyses.

Human pancreatic single-cell sequencing datasets:

The pancreatic single-cell sequencing dataset (N=14,890) was extracted using the Panc8 data containing multiple publicly available scRNA-seq transcriptomes (GSE84133, GSE85241, E-MTAB-5061, GSE83139, GSE81608). Analysis was carried out by using R studio version 1.2.5033 using the SeuratData (version 0.2.1) and Seurat (3.2.3). Data normalization across multiple datasets in Panc8 are described previously[37]. Single-cell selection criteria involved all cells that contained transcript data (samples with all zeros eliminated). Single-cell datasets were randomly split to 80% samples selected in a training set with the remaining 20% retained for validation. ML-based identification of features predicting insulin transcript was carried by data scientists blinded to gene variables and sample information. Analytical plan (80/20 discovery and validation) were predetermined and gene identifiers/sample details were provided once models were ranked following in silico validation.

Ire1α$^{\beta-/-}$ mouse pancreatic single-cell dataset:

Single-cell RNA-seq dataset from pancreatic islets of Ire1α$^{fl/fl}$ (N=1,163 single-cell transcriptomes from 1 mouse) and Ire1α$^{\beta-/-}$ (N=1,683 single-cell transcriptomes from 2 mice) were obtained through GSE144471[17]. The β-cells (Ire1α$^{fl/fl}$: 830 cells; Ire1α$^{\beta-/-}$: 816 cells) were separated from the dataset and the expression values of selected genes were evaluated in the β-cell population.

T2D pancreatic single-cell (sc)RNA sequencing dataset:

Pancreatic single-cell normalized read dataset of adult ND (N=4) and T2D (N=10) donors were obtained from GSE154126 (PMID: 32739450). The adult ND (N=296) and T2D (N=505) insulin transcribing cells were compared and used for validation.

De-identified datasets were shared with data scientists. A random number generator function was used to allocate 80% of samples to a training set. Analyses were carried using Python (Ver:3.4), wherein the data file was imported into the data frame, transposed, edited to delete INS and INS-IGF2 columns from the data frame and labelled (label=0 where INS=0 and label=1 where INS>0). Classifiers were initialized and model trained using the discovery (80%) data set. Predictive analyses were then carried out on the validation (20%) set and the resulting accuracy metrics were saved to compare the feature importance. Classifiers selected (Random Forest, Gradient Boosting, Decision Tree Classifier, Logistic Regression, Multinomial Naive Bayes Classifier, ADA Boost Classifier, Linear Discriminant Analysis, Ridge Classifier, KNeighbors Classifier and Linear Support Vector Classifier) were analysed on the same set and codes for these analyses will be made available through GitHub on publication.

Pathway analysis. To analyze enrichment for β-cell pathways lists of pancreatic single-cell features generated by ML algorithms (Random forest, Gradient boosting, Decision tree classifier and ADA Boost classifier)were compared with β cell-expressed genes (from E-GEOD-20966) using Gene Ontology over-representation analysis on Pantherdb.org[38]. Pre-analytic workflows included cleaning up entries not mapping to protein-coding gene symbols. To analyze for overrepresentation among β-cell pathways lists of machine learning predicted genes were compared with β-cell expressed genes (N=13,165 from E-GEOD-20966) using Gene Ontology (GO)-analysis on Pantherdb.org[38]. Preanalytical workflows included cleaning up entries not mapping to protein-coding gene symbols in E-GEOD-20966. Gene lists for each ML algorithm consisted of the top 100 genes as predictors of insulin expression, which

were compared against the data set of β-cell expressed genes. Overrepresentation analysis using GO categories for biological processes (GO: BP) was performed using binomial testing using false-detection-rate to correct for multiple testing. Lists of significantly enriched pathways associated with each ML algorithm were compared using Venn diagrams[39].

Statistical analysis: The R software (ver. 3.6.1; R Foundation for Statistical Computing, Vienna, Austria) was used to create the categorical bubble plot using the packages ggplot2 (3.3.3), ggpubr (0.4.0) and proto (1.0.0). Statistical software, Microsoft Excel (ver. 2016; Microsoft, Redmond, WA, USA), the R software (ver. 3.6.1; R Foundation for Statistical Computing, Vienna, Austria) and/or GraphPad Prism (ver. 8.4.1; GraphPad Software, San Diego, CA, USA) were used for univariate test comparisons and Benjamini-Hochberg method for multiple testing.

References

1. Zou, J. *et al.* A primer on deep learning in genomics. *Nat Genet* **51**, 12-18 (2019).
2. Xu, C. & Jackson, S.A. Machine learning and complex biological data. *Genome Biol* **20**, 76 (2019).
3. Angermueller, C., Lee, H.J., Reik, W. & Stegle, O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol* **18**, 67 (2017).
4. Lai, M. *et al.* Amino acid and lipid metabolism in post-gestational diabetes and progression to type 2 diabetes: A metabolic profiling study. *PLoS Med* **17**, e1003112 (2020).
5. Khan, S.R. *et al.* The discovery of novel predictive biomarkers and early-stage pathophysiology for the transition from gestational diabetes to type 2 diabetes. *Diabetologia* **62**, 687-703 (2019).
6. Joglekar, M.V. *et al.* Postpartum circulating microRNA enhances prediction of future type 2 diabetes in women with previous gestational diabetes. *Diabetologia* (2021).
7. Schrider, D.R., Ayroles, J., Matute, D.R. & Kern, A.D. Supervised machine learning reveals introgressed loci in the genomes of Drosophila simulans and D. sechellia. *PLoS Genet* **14**, e1007341 (2018).
8. Stoffers, D.A., Zinkin, N.T., Stanojevic, V., Clarke, W.L. & Habener, J.F. Pancreatic agenesis attributable to a single nucleotide deletion in the human IPF1 gene coding sequence. *Nat Genet* **15**, 106-10 (1997).
9. Harrison, K.A., Thaler, J., Pfaff, S.L., Gu, H. & Kehrl, J.H. Pancreas dorsal lobe agenesis and abnormal islets of Langerhans in Hlxb9-deficient mice. *Nat Genet* **23**, 71-5 (1999).
10. Oliver-Krasinski, J.M. & Stoffers, D.A. On the origin of the beta cell. *Genes Dev* **22**, 1998-2021 (2008).
11. Doyle, M.J. & Sussel, L. Nkx2.2 regulates beta-cell function in the mature islet. *Diabetes* **56**, 1999-2007 (2007).
12. Osborne, C.S. *et al.* Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat Genet* **36**, 1065-71 (2004).
13. Wang, Y.J. *et al.* Single-Cell Transcriptomics of the Human Endocrine Pancreas. *Diabetes* **65**, 3028-38 (2016).
14. Cinti, F. *et al.* Evidence of beta-Cell Dedifferentiation in Human Type 2 Diabetes. *J Clin Endocrinol Metab* **101**, 1044-54 (2016).
15. Guo, S. *et al.* Inactivation of specific beta cell transcription factors in type 2 diabetes. *J Clin Invest* **123**, 3305-16 (2013).
16. Weir, G.C. & Bonner-Weir, S. Five stages of evolving beta-cell dysfunction during progression to diabetes. *Diabetes* **53 Suppl 3**, S16-21 (2004).
17. Lee, H. *et al.* Beta Cell Dedifferentiation Induced by IRE1alpha Deletion Prevents Type 1 Diabetes. *Cell Metab* **31**, 822-836 e5 (2020).
18. Avrahami, D. *et al.* Single-cell transcriptomics of human islet ontogeny defines the molecular basis of beta-cell dedifferentiation in T2D. *Mol Metab* **42**, 101057 (2020).
19. Lynam, A.L. *et al.* Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults. *Diagn Progn Res* **4**, 6 (2020).
20. Macfarlane, W.M. *et al.* Glucose regulates islet amyloid polypeptide gene transcription in a PDX1- and calcium-dependent manner. *J Biol Chem* **275**, 15330-5 (2000).

21. Mulder, H., Ahren, B. & Sundler, F. Islet amyloid polypeptide and insulin gene expression are regulated in parallel by glucose in vivo in rats. *Am J Physiol* **271**, E1008-14 (1996).
22. Shepherd, L.M., Campbell, S.C. & Macfarlane, W.M. Transcriptional regulation of the IAPP gene in pancreatic beta-cells. *Biochim Biophys Acta* **1681**, 28-37 (2004).
23. Wang, H., Gauthier, B.R., Hagenfeldt-Johansson, K.A., Iezzi, M. & Wollheim, C.B. Foxa2 (HNF3beta) controls multiple genes implicated in metabolism-secretion coupling of glucose-induced insulin release. *J Biol Chem* **277**, 17564-70 (2002).
24. Matsuoka, T.A. *et al.* The MafA transcription factor appears to be responsible for tissue-specific expression of insulin. *Proc Natl Acad Sci U S A* **101**, 2930-3 (2004).
25. Li, N. *et al.* Ablation of somatostatin cells leads to impaired pancreatic islet function and neonatal death in rodents. *Cell Death Dis* **9**, 682 (2018).
26. Smith, P.A., Sellers, L.A. & Humphrey, P.P. Somatostatin activates two types of inwardly rectifying K+ channels in MIN-6 cells. *J Physiol* **532**, 127-42 (2001).
27. Blodgett, D.M. *et al.* Novel Observations From Next-Generation RNA Sequencing of Highly Purified Human Adult and Fetal Islet Cell Subsets. *Diabetes* **64**, 3172-81 (2015).
28. Filipsson, K., Kvist-Reimer, M. & Ahren, B. The neuropeptide pituitary adenylate cyclase-activating polypeptide and islet function. *Diabetes* **50**, 1959-69 (2001).
29. Gu, H.F. Genetic variation screening and association studies of the adenylate cyclase activating polypeptide 1 (ADCYAP1) gene in patients with type 2 diabetes. *Hum Mutat* **19**, 572-3 (2002).
30. Rutter, G.A., Pullen, T.J., Hodson, D.J. & Martinez-Sanchez, A. Pancreatic beta-cell identity, glucose sensing and the control of insulin secretion. *Biochem J* **466**, 203-18 (2015).
31. Rutter, G.A. & Pullen, T.J. Comment on: Schuit et al. beta-Cell-specific gene repression: a mechanism to protect against inappropriate or maladjusted insulin secretion? Diabetes 2012;61:969-975. *Diabetes* **61**, e16; author reply e17 (2012).
32. Schuit, F. *et al.* beta-cell-specific gene repression: a mechanism to protect against inappropriate or maladjusted insulin secretion? *Diabetes* **61**, 969-75 (2012).
33. Cantley, J. *et al.* A preexistent hypoxic gene signature predicts impaired islet graft function and glucose homeostasis. *Cell Transplant* **22**, 2147-59 (2013).
34. Kameswaran, V. *et al.* Epigenetic regulation of the DLK1-MEG3 microRNA cluster in human type 2 diabetic islets. *Cell Metab* **19**, 135-45 (2014).
35. Zou, M. *et al.* A novel mixed integer programming for multi-biomarker panel identification by distinguishing malignant from benign colorectal tumors. *Methods* **83**, 3-17 (2015).
36. Dankers, F., Traverso, A., Wee, L. & van Kuijk, S.M.J. Prediction Modeling Methodology. in *Fundamentals of Clinical Data Science* (eds. Kubben, P., Dumontier, M. & Dekker, A.) 101-120 (Cham (CH), 2019).
37. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).
38. Mi, H. *et al.* Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat Protoc* **14**, 703-721 (2019).
39. Oliveros, J.C. Venny: An interactive tool for comparing lists with Venn's diagrams. *https://bioinfogp.cnb.csic.es/tools/venny/index.html* (2007-2015).

Supplementary Table 1.

#	Workflow	Build	Description
1	**Random Forest**	Ensemble	Random Forest classifier is a meta-estimator that fits several decision trees on various sub-samples of datasets and uses an average to improve the predictive accuracy of the model and controls over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement. Reduction in over-fitting and Random Forest classifier is more accurate than decision trees in most cases.
2	**Gradient Boosting**	Ensemble	The idea behind "gradient boosting" is to take a weak hypothesis or weak learning algorithm and make a series of tweaks to it that will improve the strength of the hypothesis/learner. This type of Hypothesis Boosting is based on the idea of Probability Approximately Correct Learning (PAC).
3	**Adaptive Boosting (Adaboost)**	Ensemble	For AdaBoost, many weak learners are created by initializing many decision tree algorithms that only have a single split. The instances/observations in the training set are weighted by the algorithm, and more weight is assigned to instances that are difficult to classify. More weak learners are added into the system sequentially, and they are assigned to the most difficult training instances. In AdaBoost, the predictions are made through majority vote, with the instances being classified according to which class receives the most votes from the weak learners.
4	**Ridge Classifier**	Regularization	Ridge classifier first converts binary targets to [-1, 1] and then treats the problem as a regression task, optimizing the same objective as above. The predicted class corresponds to the sign of the regressor's prediction. For multiclass classification, the problem is treated as multi-output regression, and the predicted class corresponds to the output with the highest value.
5	**Logistic Regression**	Regression	Logistic regression is a machine learning algorithm for classification. In this algorithm, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function. It is most useful for understanding the influence of several independent variables on a single outcome variable.
6	**Naive Bayes**	Bayesian	Naive Bayes algorithm based on Bayes' theorem with the assumption of independence between every pair of features. Naive Bayes classifiers work well in many real-world situations such as document classification and spam filtering. This algorithm requires a small amount of training data to estimate the necessary parameters. Naive Bayes classifiers are extremely fast compared to more sophisticated methods.

7	**Decision Tree Classifier**	Decision Tree	Given data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data. Decision Tree is simple to understand and visualise, requires little data preparation, and can handle both numerical and categorical data.
8	**K-Nearest Neighbours**	Instance-based	Neighbours based classification is a type of lazy learning as it does not attempt to construct a general internal model but simply stores instances of the training data. Classification is computed from a simple majority vote of the k nearest neighbours of each point. This algorithm is simple to implement, robust to noisy training data, and effective if training data is large.
9	**Linear Discriminant Analysis (LDA)**	Dimensionality Reduction	It is a classifier with a linear decision boundary, generated by fitting class conditional densities to the data and using Bayes' rule. The model fits a Gaussian density to each class, assuming that all classes share the same covariance matrix. The fitted model can also be used to reduce the dimensionality of the input by projecting it to the most discriminative directions, using the transform method.
10	**Linear Support Vector Classifier**	Support Vector Machines (SVM)	Support vector machine is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

Supplementary Table 2.

Rank	Gradient Boost	Random Forest	ADA Boost	Decision tree
1	IAPP	IAPP	IAPP	MTRNR2L8
2	MTRNR2L8	ADCYAP1	SST	IAPP
3	MTRNR2L1	MAFA	ATP5E	MAFA
4	RPS15	SST	LDHA	GSTA1
5	SST	DLK1	ADCYAP1	RPS15
6	MAFA	PCSK1N	GPX4	MAB21L3
7	ADCYAP1	HADH	DOCK10	XIST
8	NPTX2	PCSK1	CAPN13	CRP
9	LDHA	MALAT1	KCNA5	RPL35
10	MAB21L3	LDHA	RPL23AP32	SST
11	PCSK1N	NPTX2	NPM3	NPTX2
12	DDX3Y	CD99	GGPS1	CNPY3
13	MIF	C1QL1	RHOBTB1	PCSK1N
14	RBP4	EEF1A2	SIX3	TP73-AS1
15	ATP5E	CD151	RPL13AP17	CCDC28A
16	RPS28	GNAS	STMN3	IGSF3
17	CRP	RBP4	DDX6	RPL26
18	GSTA1	MIF	PLSCR4	HADH
19	MAST1	EEF2	FAM163A	MZT2B
20	CD24	UCHL1	ZNF578	RPL36A
21	RPS18	IGF2	MTRNR2L8	MIR663A
22	AES	LRRFIP1	RPL27A	MARK2
23	PRSS8	HNRPDL	CST3	ZNF682
24	TRIM47	UGDH-AS1	PRRC2C	C6orf1
25	CD151	MTRNR2L2	MTRNR2L6	BNIP2
26	GPX4	RPS17	KCTD21	EEF1A2
27	HADH	MAB21L3	ARPC3	KIAA1161
28	ATP5MD	TMEM66	KCTD1	TGFBI
29	MYO6	RPL3	LINC00342	RPS4Y1
30	L1TD1	ITGB1	ADSSL1	AHSA2
31	TTTY15	ABHD2	DLK1	RNFT2
32	XIST	ERO1LB	STRN	NFASC
33	CST3	ANXA4	SNHG14	PHC2
34	LOC101929224	TTR	MYL6	CCNG2
35	DLK1	ABCC8	LINC00294	BMS1
36	RPL8	YWHAZ	FXYD3	RSBN1
37	LINC01550	GPX4	WSCD2	SDF2L1
38	C4BPB	ATP5E	HIST1H1E	LINC01858
39	MTRNR2L10	CD59	CACNA2D3	FTH1
40	PRRC2C	HNRNPU	DCTN4	FBXL18
41	UBL7-AS1	KCNQ1OT1	MAFA	EIF1B
42	GNAS	MTRNR2L1	H1FX	BBX
43	ATAD2	PRRC2C	TTTY15	TAF7

44	AL353803.1	WSCD2	ANO6	FXYD6
45	HSPA6	TOMM6	SUPT20H	IRF2BP2
46	UCHL1	NLRP1	S100A14	NMB
47	LOC100287177	UGGT1	LYPD3	TMA7
48	THSD7A	SRRM2	SCARF1	KLF7
49	LINC00342	RPS15	SMC3	RPL17
50	RPS4Y1	CPE	MAST1	KIAA1919
51	GGPS1	CCAR1	CCDC38	SRPR
52	GCG	SCGN	ZNF790-AS1	GNG5
53	ZFP36L1	H3F3A	MARS	RASGRF1
54	BRWD3	PFKFB2	ZNF787	CXCL3
55	SACS	ANKRD12	AKAP1	ZPBP
56	IVNS1ABP	WAC-AS1	UGT2B17	ADCYAP1
57	ITGB1	LPP	CA5BP1	PNMA3
58	BBS12	GCG	TRAPPC12	PPP1CB
59	PCDHA8	C17orf76-AS1	PRPF19	ALDOA
60	EDA	RPS18	RAB30-AS1	CHGA
61	GAD2	ZC3HAV1	RDH13	C2CD2L
62	BRI3	TMSB4X	HADH	DOT1L
63	PRSS3P2	RPL36A-HNRNPH2	DENND6A	IGIP
64	RPS29	RPL8	RPS29	HERPUD1
65	TP73-AS1	RHOC	ATAD2	TMEM254
66	MTHFD1	REG3A	SLC16A3	EDARADD
67	GOLIM4	STX16	RPS15	GADD45B
68	AC025254.1	YWHAB	DROSHA	FOXA3
69	PTBP2	SRSF11	WBP11	SLC35B4
70	SERINC1	RNASEK	SEZ6	ABHD5
71	KCNJ2	PKD1P1	SNORA79	MALL
72	PSAP	PEMT	FAM83H	LOC101928069
73	EHF	MYO10	MTCH2	RPS20
74	AKR7A2	FTL	FAM114A1	SRP9
75	PKD1P1	C1orf127	PITPNA	A1BG
76	AC025259.3	LINC00657	LHPP	DDAH1
77	MYOF	RPL13AP20	BRPF3	C7orf65
78	PRPF19	ERO1B	IFITM1	TRIM33
79	MIR663A	PTPRN	BSG	ARHGAP35
80	YPEL3	SCG5	CSDE1	YOD1
81	EIF3J-DT	MYO6	MTHFD1	LRRC10
82	AC098934.2	PPIA	WBP1	PIGK
83	38596	TM4SF1	EHBP1	SERTM1
84	CPEB3	FXYD2	C4BPB	SERPINB6
85	LOC100506476	U2AF1	NEURL3	ING2
86	CENPC1	PTP4A2	EFNB3	GLTSCR2
87	BAI2	PPP1R1A	RBP4	SURF1
88	UEVLD	LOC643406	LINC00657	LOC100507463
89	MX1	GAD2	SPNS1	CYP2R1

Supplementary Table 3

Gene transcript	P-value	ND	T2D
INS	<0.0001	9.31 ± 0.29	6.52 ± 0.19
IAPP	<0.0001	6.88 ± 0.28	4.56 ± 0.20
ADCYAP1	<0.0001	3.24 ± 0.24	1.81 ± 0.16
MAFA	0.0021	0.89 ± 0.10	0.71 ± 0.08
SST	<0.0001	5.23 ± 0.25	3.02 ± 0.16
LDHA	0.2679	6.67 ± 0.26	6.85 ± 0.19

Supplementary Table 3: The key genes common in all top three ML workflows (**Figure 2C**) with *INS* gene accessed in Avrahami et al single-cell RNA-seq data (GSE154126[18]) between adult non-diabetic (ND; N=296) vs T2D (N=505) insulin transcribing single cells. Data are presented as $\log_2$ transformed normalized reads mean $\pm$ SEM. The P-values were calculated with the Mann-Whitney test. Genes with significant P-value (P<0.05) show here, remain significant after adjusting for multiple testing using the Benjamini-Hochberg method (FDR=0.05, on 21,153 tests).