

Investigation of metabolic pathways from gut microbiome analyses regarding type 2 diabetes mellitus using artificial neural networks

Julienne Siptroth¹ · Olga Moskalenko² · Carsten Krumbiegel² · Jörg Ackermann³ · Ina Koch³ · Heike Pospisil¹

Received: 13 February 2023 / Accepted: 25 April 2023

Published online: 09 May 2023

© The Author(s) 2023 [OPEN](#)

Abstract

Background Type 2 diabetes mellitus is a prevalent disease that contributes to the development of various health issues, including kidney failure and strokes. As a result, it poses a significant challenge to the worldwide healthcare system. Research into the gut microbiome has enabled the identification and description of various diseases, with bacterial pathways playing a critical role in this context. These pathways link individual bacteria based on their biological functions. This study deals with the classification of microbiome pathway profiles of type 2 diabetes mellitus patients.

Methods Pathway profiles were determined by next-generation sequencing of 16S rDNA from stool samples, which were subsequently assigned to bacteria. Then, the involved pathways were assigned by the identified gene families. The classification of type 2 diabetes mellitus is enabled by a constructed neural network. Furthermore, a feature importance analysis was performed via a game theoretic approach (SHapley Additive exPlanations). The study not only focuses on the classification using neural networks, but also on identifying crucial bacterial pathways.

Results It could be shown that a neural network classification of type 2 diabetes mellitus and a healthy comparison group is possible with an excellent prediction accuracy. It was possible to create a ranking to identify the pathways that have a high impact on the model prediction accuracy. In this way, new associations between the alteration of, e.g. a biosynthetic pathway and the presence of diabetes mellitus type 2 disease can also be discovered. The basis is formed by 946 microbiome pathway profiles from diabetes mellitus type 2 patients (272) and healthy comparison persons (674).

Conclusion With this study of the gut microbiome, we present an approach using a neural network to obtain a classification of healthy and type 2 diabetes mellitus and to identify the critical features. Intestinal bacteria pathway profiles form the basis.

Keywords Type 2 diabetes mellitus · Gut · NGS · Artificial neural network · Machine learning · Explainable artificial intelligence (XAI)

Abbreviations

BMI	Body mass index
EC number	Enzym commission number
NGS	Next generation sequencing
rDNA	Ribosomal deoxyribonucleic acid

✉ Julienne Siptroth, siptroth@th-wildau.de; Olga Moskalenko, olga.moskalenko@biomes.world; Carsten Krumbiegel, carsten.krumbiegel@biomes.world; Jörg Ackermann, j.ackermann@bioinformatik.uni-frankfurt.de; Ina Koch, ina.koch@bioinformatik.uni-frankfurt.de; Heike Pospisil, heike.pospisil@th-wildau.de | ¹High Performance Computing in Life Sciences, Technical University of Applied Sciences Wildau, 15745 Wildau, Germany. ²BIOMES NGS GmbH, Schwartzkopfstraße 1, 15745 Wildau, Germany. ³Department of Molecular Bioinformatics, Institute of Computer Science, Goethe University Frankfurt, 60325 Frankfurt am Main, Germany.



T2D Type 2 diabetes
WHO World Health Organisation

1 Introduction

For a long time, type 2 diabetes mellitus was considered a disease of old age. However, risk factors such as diet, lack of exercise, and often associated obesity, are increasingly affecting younger people as well [1]. In recent years, this disease has become increasingly prevalent in children and young people. It is one of the most prevalent diseases in the world today. Thus, new challenges arise for the health care system. Often it comes in the consequence of a type 2 diabetes mellitus disease to the development of other diseases such as damage of kidneys and eyes, the diabetic foot, heart and vascular diseases, which can even lead to death [2]. Thus, diabetes mellitus was included by the World Health Organization (WHO) in 2019 for the first time in the top 10 leading causes of death [3]. However, studies have shown that the disease can be completely eliminated or at least significantly reduced. To achieve this, a change in lifestyle and reduction of body weight is necessary [4]. A division is made in diabetes mellitus, a metabolic disease, with partial hereditary predisposition, into two main forms. In addition to type 2 diabetes mellitus, which is the most common form with over 90%, there is also type 1 diabetes mellitus. In type 2 diabetes mellitus, the insulin processing of the cells is disturbed. The pancreas is still able to provide enough insulin, but it comes to an increasingly poor processing by the cell, this can even lead to a complete insulin resistance [5].

The human intestine is one of the most important organs and has an influence on many processes in the human body. It is not only responsible for digestion, but also controls inflammatory processes and supports the human immune system. The gut microbiome itself is influenced by a whole range of factors. Not only nutrition plays a role in its composition, but also many other factors such as the environment, age, gender and lifestyle. Depending on the influencing factors, the composition of the intestinal microbiota can vary greatly. The bacteria have the largest part in the human intestine, a total of about 100 trillion bacteria live there. Furthermore, this complex structure consists of fungi and animals. The diversity of bacteria in the human intestine is increased with a balanced diet. This is accompanied by a broad formation of a wide variety of metabolic products [6, 7].

These arise from a great diversity of bacterial pathways. The examination of the functional microbiome is becoming increasingly important. Especially, the alterations of biosynthesis pathways have characteristic impact on the profiling of type 2 diabetes mellitus. The altered occurrence of the biosynthesis pathways of amino acids (e.g. L-tyrosine, L-phenylalanine and L-isoleucine), the thiazole biosynthesis pathway or the pyrimidine deoxyribonucleotides de novo biosynthesis pathway are significant for type 2 diabetes mellitus.

2 Materials and methods

2.1 Data

There were more than 29,000 samples available, with information on the microbiome (relative counts per taxonomic level) and individual lifestyle (age, BMI, diet, etc.). A huge number of parameters on the individual lifestyle was included, e.g. diet, diseases and medication intake. The microbiome profiles were determined using NGS. For this purpose, the bacterial 16 S ribosomal rDNA is sequenced. Information on normalized counts per relative level (kingdom to species) is provided. The project partner BIOMES NGS GmbH provided the data, based on a self-test for the analysis of the intestinal microbiome. Only data with patients consent for scientific use was used. The customers of BIOMES NGS GmbH performed the test independently at home and enter the data regarding their individual lifestyle. There is no final verification by a medical doctor.

We classified the costumers into the group of healthy controls and type 2 diabetes mellitus (T2D) patients. For both groups an age between 18 and 80 years was considered. In the healthy group, classification was based on the following parameters: Age between 18 and 80 years, BMI between 18.5 and 27.5, no diseases, gastrointestinal complaints, gluten intolerances and medication intake, and no intake of antibiotics and/or probiotics in the last 3 months. Also, they had not to consume daily alcohol, and the well-being score had to be reported greater than 4 (out of 10) and the health score greater than or equal to 6 (out of 10). This resulted in 272 samples for the T2D group and 674 samples for the healthy group. A more even distribution

of the two groups would be desirable. However, further adjustment of the parameters or reduction of the healthy group would lead to an unacceptable sample size.

2.2 Methods

2.2.1 Sample preparation and sequencing

The submitted stool samples were stored and then prepared for lysis. After lysis has taken place, extraction was performed. This was followed by library preparation for sequencing using the Illumina MiSeq System followed by processing of the sequence reads.

2.2.2 Processing sequence reads

Subsequently, the determined paired-end reads were filtered. Using PANDAseq [8] the forward/reverse reads were merged. Then, an alignment was performed using BLASTn [9] against the SILVA rRNA database (version: 138.1) [10]. With CD-HIT [11, 12] the sequences were clustered. Followed by a calculation of the biologically normalized abundance applying the PICRUSt2 pipeline [13]. In parallel, the PICRUSt2 pipeline also determine the available pathways (MetaCyc [14]) for each sample. By Mapping the EC numbers of the gene families the abundances of the identified pathways were determined.

The steps of sample preparation and sequencing, as well as the processing of the sequence reads follow the description in a previous work [15].

Further analysis steps were performed with custom Python (3.7.7) scripts using the keras 2.3.1 [16], NumPy 1.18.1 [17], pandas 1.2.4 [18, 19], scikit-learn 0.22.1 [20], SciPy 1.4.1 [21], SHAP 0.40.0 [22] and tensorflow 2.1.0 [23] libraries.

2.2.3 Machine learning

A feedforward artificial neural network was used to assign a microbiome profile to the T2D group. The data were labeled for the classifier. A GridSearch was performed to optimize the hyperparameters and determine the most suitable architecture of the neural network. In Fig. 1 the used hyperparameters are listed with the tested settings. The hyperparameters optimized were activation function, optimizer, dense layer size, dropout, epochs and batch size. Between three (epochs) and up to seven (optimizer) different settings for the hyperparameters were tested. Furthermore, different numbers of layers were tested in the architecture of the neural network.

The chosen model (cf. Figure 2) consists of an input layer, six dense layers and a dropout layer. For all layers, the ReLU activation function was used except for the last layers, linear and sigmoid function was used there. Furthermore, the optimizer Adam [29] was applied. At the beginning 5% of the data set were taken out as test set. The neural network did not see this data in the training and validation phase. Training was done in 100 epochs. Evaluation of model performance was performed using repeated k -fold cross-validation. The size of k (number of splits) was set to 10 and the number of repetitions to 3. Accuracy was calculated as a measure to estimate the prediction accuracy. The validation data were then used to determine the accuracy of the model. Accuracy was calculated for each of the two groups (healthy and T2D) individually and for the entirety of the data. To determine which bacterial pathways have the greatest impact on prediction accuracy, feature importance was determined using SHAP [22]. The calculation of the classes of the top 50 pathways was done over several iterations. Only the first 50 pathways of each iteration were counted. In each iteration, the top 50 pathways were assigned a value between 1 and 50, depending on their calculated position. The values were summed up after each iteration. Afterwards, the first 50 pathways were used for the consideration of the classes.

3 Results

3.1 Classification with neural network

A neural network was trained to obtain a classification of healthy and T2D. For this purpose, the pathways for each sample were used as the basis.

The optimized network architecture (cf. Figure 2) contains one input layer, six dense layers, one dropout layer and one output layer. Adam was applied as optimizer and the ReLU function as the activation function, except in the last

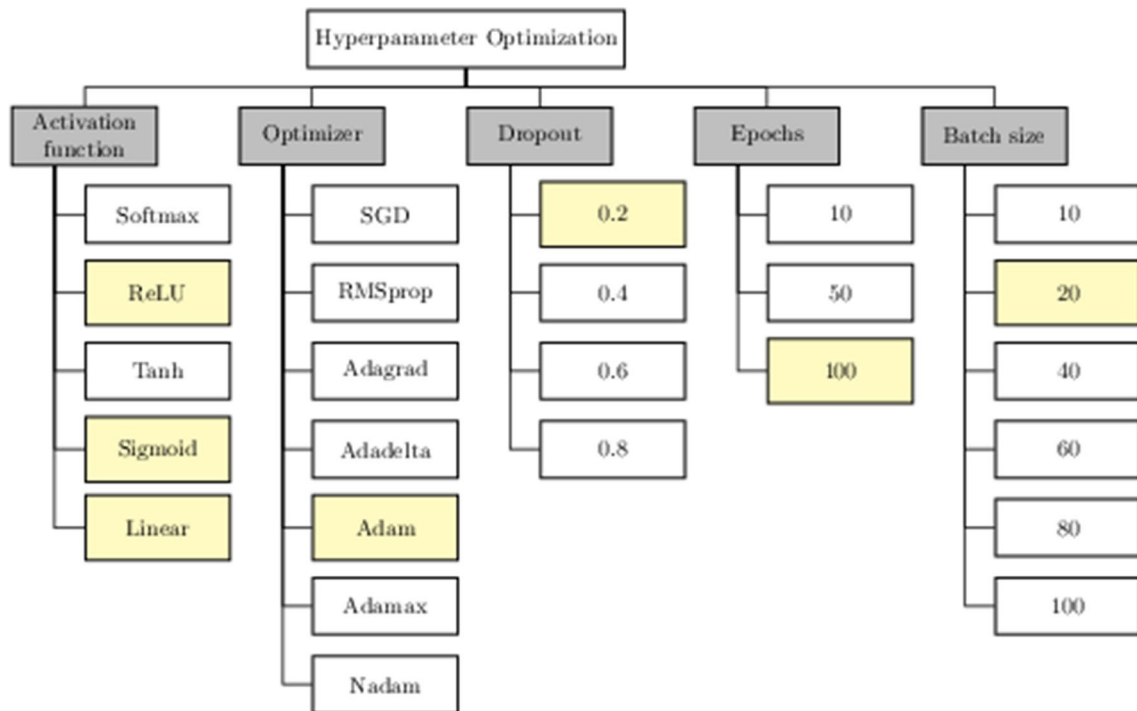


Fig. 1 Hyperparameter optimization with GridSearch with the tested settings for the activation function, optimizer (SGD [24, 25], RMSprop [26], Adagrad [27], Adadelta [28], Adam [29], Adamax [29], Nadam [30]), dropout, epochs and batch size. Used parameter (yellow)

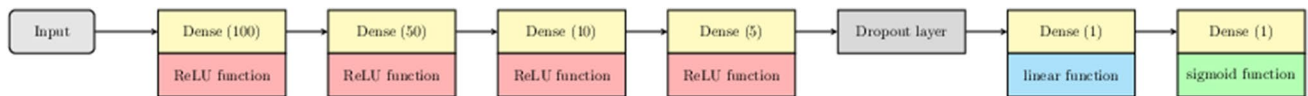


Fig. 2 Architecture of the chosen neural network with layers (dense layer size) and activation functions

two layers where the linear and the sigmoid function was used. The selected dropout was 0.2. 100 epochs were used for training. With this neural network we achieved an accuracy of 0.845. The precision for diabetes type 2 was 0.96, the recall (sensitivity) was 0.93 and the F1 score was 0.95. The specificity was 0.98. And for healthy the precision was 0.97, the recall was 0.98 and die F1 score was 0.95. The values for accuracy, precision and recall are shown in Fig. 3 for the T2D and the healthy group.

Other combinations of hyperparameters and layers achieved on average a prediction accuracy of 65% to 75%, and in some cases of about 80%.

3.2 Calculation of SHAP Feature Importance

To determine which bacterial pathways have the greatest influence on the model prediction accuracy, the feature importance was calculated using SHAP (Table 1).

The SHAP calculated top 10 pathways with the biggest impact on prediction accuracy were listed in Table 2. The table includes the BioCyc ID, the description and the occurrence in diabetes group.

The top 10 ranked pathways occurred in at least 97% of samples in both groups (healthy and T2D). For the 10 pathways with the lowest impact on model accuracy, on the other hand, these pathways occurred in less than 4% of the samples in each of the two groups. The SHAP-calculated 10 pathways with the lowest impact on prediction accuracy are listed in Table 3. The table includes the BioCyc ID and the description.

In Fig. 4 the classes of the top 50 pathways with the greatest influence on the accuracy are shown in a sunburst plot. The inner circle represents the classes into which the Top 50 pathways are categorized. The outer circle shows the subclasses that occur most frequently in the Top 50 pathways.

Fig. 3 Accuracy and scores for precision and recall of the neural network of the diabetes type 2 (green) and the healthy (blue) group

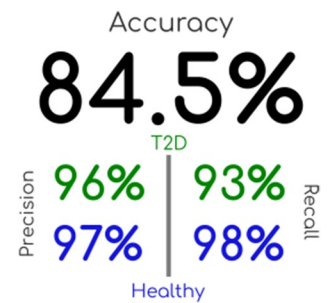


Table 1 Distribution of the selected parameters age, sex, BMI and nutrition for the two groups Healthy and T2D

Parameter	Healthy group (n=674)	T2D group (n=272)
Age, years	42.55 ± 12,12	59.71 ± 12.27
Women/Men/Divers, n	340/318/1	143/127/0
BMI, kg/m^2	23.13 ± 2.24	31.05 ± 6.38
Nutrition		
Omnivore	517	245
Vegan	40	2
Vegetarian	74	10
Pescetarian	43	8

Table 2 Top 10 pathways (biggest impact) ranked by SHAP

BioCyc ID	Pathways	Occurrence in Diabetes group
<i>PWY-5101</i>	L-isoleucine biosynthesis II	↑
<i>PWY-6628</i>	Superpathway of L-phenylalanine biosynthesis	↑
<i>PWY-6630</i>	Superpathway of L-tyrosine biosynthesis	↑
<i>PWY0-1479</i>	tRNA processing	↑
<i>P341-PWY</i>	Glycolysis V (Pyrococcus)	↓
<i>PWY-1861</i>	Formaldehyde assimilation II (RuMP Cycle)	↓
<i>PWY-6545</i>	Pyrimidine deoxyribonucleotides de novo biosynthesis III	↓
<i>PWY-6749</i>	CMP-legionaminat biosynthesis I	↓
<i>PWY-6891</i>	Thiazole biosynthesis II (Bacillus)	↓
<i>PWY0-1415</i>	Superpathway of heme biosynthesis from uroporphyrinogen-III	↓

Pathway BioCyc ID, description and occurrence in diabetes group (↑- increases; ↓- decreases)

The top 50 pathways belong to 4 different classes (out of 12 different classes). The *Biosynthesis* class is the most represented, followed by *Degradation/Utilization/Assimilation*. In the *Biosynthesis* class, the *Amino Acid Biosynthesis* and *Cofactor, Carrier, and Vitamin Biosynthesis* subclasses occurred most frequently, and in the *Degradation/Utilization/Assimilation* class, it was the *Fermentation* subclass. All other subclasses occurred not more than twice.

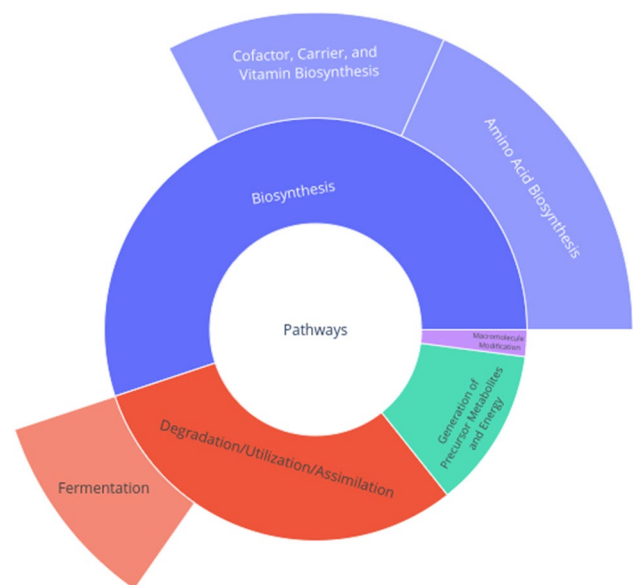
4 Discussion

Using a neural network, a classification of pathway microbiome profiles of individuals with diabetes mellitus type 2 disease could be realized. It was possible to distinguish these profiles from healthy comparison samples with excellent predictive accuracy. Furthermore, it is possible to rank the impact of the pathway on the model prediction accuracy. The 10 pathways with the greatest influence were *PWY-6891* (thiazole biosynthesis II (Bacillus)), *PWY0-1415* (superpathway of heme biosynthesis from uroporphyrinogen-III), *PWY-1861* (formaldehyde assimilation II (RuMP Cycle)), *PWY0-1479* (tRNA processing), *PWY-6630* (superpathway of L-tyrosine biosynthesis), *PWY-6545* (pyrimidine

Table 3 10 Pathways with the lowest impact calculated by SHAP

BioCyc ID	Pathways
<i>PWY-6942</i>	dTDP-D-desosamine biosynthesis
<i>PWY-5266</i>	p-cymene degradation
<i>PWY-7015</i>	Ribostamycin biosynthesis
<i>PWY-6660</i>	2-heptyl-3-hydroxy-4(1 H)-quinolone biosynthesis
<i>PWY-7020</i>	Superpathway of butirocin biosynthesis
<i>PWY-7014</i>	Paromamine biosynthesis I
<i>PWY-5499</i>	Vitamin B6 degradation
<i>PWY-5519</i>	D-arabinose degradation III
<i>PWY-622</i>	Starch biosynthesis
<i>PWY-7401</i>	Crotonate fermentation (to acetate and cyclohexane carboxylate)

Pathway BioCyc ID and description

Fig. 4 Classes of the Top 50 pathways with the greatest impact on model prediction depending on the mean absolute SHAP value. *Biosynthesis* (blue), *Degradation/Utilization/Assimilation* (red), *Generation of Precursor Metabolites and Energy* (green) and *Macromolecule Modification* (purple)

deoxyribonucleotides de novo biosynthesis III), *PWY-6749* (CMP-legionaminat biosynthesis I), *PWY-6628* (superpathway of L-phenylalanine biosynthesis), *P341-PWY* (glycolysis V (Pyrococcus)), and *PWY-5101* (L-isoleucine biosynthesis II).

The pathways *PWY-6630*, *PWY-6628* and *PWY-5101* have on average an increased occurrence in type 2 diabetes mellitus. These are biosynthesis pathways of amino acids. In the pathway *PWY-6630* L-tyrosine, in *PWY-6628* L-phenylalanine and in *PWY-5101*, L-isoleucine is synthesized. The amino acids phenylalanine and isoleucine are essential amino acids. These are amino acids that are necessary for life, but which the human body cannot produce itself. Thus, the intake through food is essential for the human body. Tyrosine is produced from phenylalanine. If not enough phenylalanine is available, tyrosine also becomes essential for the human body. In this case, it is called semi or partially essential. Furthermore, isoleucine, along with valine and leucine, is one of the branched-chain amino acids. Among other things, muscle formation and wound healing are involved in metabolism. Due to their ring structure (benzole ring), phenylalanine and tyrosine belong to the aromatic amino acids. These play an important role for example in energy metabolism and the formation of adrenaline. A sufficient amount of branched-chain and aromatic amino acids is therefore important for the human body, but studies have shown that a significantly increased concentration of these amino acids is characteristic of type 2 diabetes [31–33]. This negatively affects insulin processing. Thus, the increased occurrence of the pathways *PWY-6630*, *PWY-6628*, and *PWY-5101* is associated with type 2 diabetes and have an increased impact on model prediction accuracy. The *PWY-1479* pathway also occurs with increased occurrence in the diabetes mellitus type 2 group.

All other pathways, of the 10 with the biggest impact have on average a reduced occurrence in type 2 diabetes patients.

PWY-6891 is a thiazole biosynthesis pathway. In this process, thiazole is synthesized. This is a moiety of thiamine consisting of sulfur and nitrogen. This moiety is one of the most important structures in the human body. For example, it is found in vitamin B1 (thiamine) and has an influence on a wide variety of functions. But thiazoles are also used in the production of pigments and pharmaceuticals. Thus, it is in various forms also a component in drugs for the treatment of type 2 diabetes [34, 35]. A lower occurrence of the thiazole biosynthesis pathway is therefore strongly associated with type 2 diabetes and has a strong impact on the model prediction accuracy of the classification of type 2 diabetes.

PWY-6545 (pyrimidine deoxyribonucleotides de novo biosynthesis III) is a nucleoside and nucleotide biosynthesis pathway. Through it, pyrimidine nucleoside triphosphates are synthesized. Pyrimidine nucleoside triphosphates, along with purine nucleoside triphosphates, are the activated precursors of DNA and RNA. The change in the occurrence of these pathways has been shown in other studies. However, the exact relationship between type 2 diabetes mellitus disease and the decreased nucleotide biosynthesis pathway is not clear yet [36, 37].

The 10 pathways with the lowest impact on model prediction accuracy were *PWY-6942* (dTDP-D-desosamine biosynthesis), *PWY-5266* (p-cymene degradation), *PWY-7015* (ribostamycin biosynthesis), *PWY-6660* (2-heptyl-3-hydroxy-4(1 H)-quinolone biosynthesis), *PWY-7020* (superpathway of butirocin biosynthesis), *PWY-7014* (paromamine biosynthesis I), *PWY-5499* (vitamin B6 degradation), *PWY-5519* (D-arabinose degradation III), *PWY-622* (starch biosynthesis), *PWY-7401* (crotonate fermentation (to acetate and cyclohexane)). These pathways showed that the classifications with the chosen neural network, work for the type 2 diabetes. Thus, the pathway *PWY-7401* had the lowest impact (determined by SHAP) on model prediction accuracy. The value determined by SHAP was 0. Therefore, no impact on classification was associated with this pathway. This was confirmed by the data basis, in both groups (T2D and healthy) this pathway did not occur. So, there could be no impact on the model prediction accuracy. The remaining pathways show a similar result. These pathways are only present in very few samples (under 4%) and partly only in one group. Thus, no strong impact on the classification can be observed in this case as well. This shows that the classification by the selected neural network identifies the important factors of influence.

Acknowledgements Thanks to the laboratory team from BIOMES NGS GmbH, especially Philipp Franke. Further thanks go to Christian Rockmann and Prof. Dr. Marcus Frohme for the project acquisition.

Author contributions JS and HP Conceptualization. JS Machine Learning, Writing—Original Draft. OM and CK Providing the Data (Sample preparation and sequencing; processing sequence reads). IK and JA Writing—Review and Editing. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by the Federal Ministry of Education and Research of Germany under grant number 13FH209PX8.

Data availability The data used are not publicly available because the research project was carried out in collaboration with the company BIOMES NGS GmbH. They can be requested with a reasoned request to the corresponding author.

Declarations

Ethics approval and consent to participate The data are provided by the company BIOMES NGS GmbH. Customers pay for the microbiome test and give their approval for further scientific use in the purchase process. Only data with granted consent was used.

Consent for publication Not applicable

Competing of interests The authors declare that they have no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Fletcher Barbara, Gulanick Meg, Lamendola Cindy. Risk Factors for Type 2 diabetes mellitus. *J Cardiovasc Nurs.* 2002;16(2):17–23.
2. Anthony Cannon, Yehuda Handelsman, Michael Heile, Michael Shannon. Burden of illness in type 2 diabetes mellitus. *J Managed Care Special Pharm.* 2018;24:55.
3. World Health Organization (WHO). WHO reveals leading causes of death and disability worldwide: 2000-2019. fact sheet November 2020.
4. Lean MEJ, Leslie WS, Barnes AC, Brosnahan N, Thom G, McCombie L, Peters C, Zhyzhneuskaya S, Al-Mrabeh A, Hollingsworth KG, Rodrigues AM, Rehackova L, Adamson AJ, Sniehotta FF, Mathers JC, Ross HM, McIlvenna Y, Stefanetti R, Trenell M, Welsh P, Kean S, Ford I, McConnachie A, Sattar N, Taylor R. Primary care-led weight management for remission of type 2 diabetes (DiRECT): an open-label, cluster-randomised trial. *Lancet.* 2018;391(10120):541–51.
5. Zaccardi Francesco, Webb David R, Yates Thomas, Davies Melanie J. Pathophysiology of type 1 and type 2 diabetes mellitus: a 90-year perspective. *Postgrad Med J.* 2016;92(1084):63–9.
6. Lozupone Catherine A, Stombaugh Jesse I, Gordon Jeffrey I, Jansson Janet K, Knight Rob. Diversity, stability and resilience of the human gut microbiota. *Nature.* 2012;489(7415):220–30.
7. Shreiner Andrew B, Kao John Y, Young Vincent B. The gut microbiome in health and in disease. *Current Opin Gastroenterol.* 2015;31(1):69–75.
8. Masella Andre P, Bartram Andrea K, Truszkowski Jakub M, Brown Daniel G, Neufeld Josh D. PANDAsseq: paired-end assembler for illumina sequences. *BMC Bioinform.* 2012;13(1):31.
9. Altschul Stephen F, Gish Warren, Miller Webb, Myers Eugene W, Lipman David J. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–10.
10. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Jörg Peplies, Oliver Glöckner Frank. LVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 2013;41(D1):D590–6.
11. Li Weizhong, Godzik Adam. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics.* 2006;22(13):1658–9.
12. Limin Fu, Niu Beifang, Zhu Zhengwei, Sitao Wu, Li Weizhong. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics.* 2012;28(23):3150–2.
13. Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Curtis Taylor C.M., Huttenhower, and Morgan G. RUST2 for prediction of metagenome functions. *Nature Biotechnol.* 2020;38(6):685–8.
14. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Subhraveti P, Weaver DS, Karp PD. MetaCyc database of metabolic pathways and enzymes and the bioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 2016;44:D471–80.
15. Siptroth J, Moskalenko O, Krumbiegel C, Jörg Ackermann, Koch I, Pospisil H. Variation of butyrate production in the gut microbiome in type 2 diabetes patients. *Int Microbiol.* 2023. <https://doi.org/10.1007/s10123-023-00324-6>.
16. François Chollet and others. *Keras.* 2015
17. Harris CR, Millman KJ, van der Walt Stéfan J, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Allan Haldane, del Rio Jaime Fernández, Wiebe Mark, Peterson Pearu, Gérard-Marchant Pierre, Sheppard Kevin, Reddy Tyler, Weckesser Warren, Abbasi Hameer, Gohlke Christoph, Oliphant Travis E. *Array programming with NumPy.* London: Nature Publishing Group; 2020.
18. McKinney Wes. *Data structures for statistical computing in python.* 56–61, Austin. xas. 2010.
19. The pandas development team (2020) *pandas-dev/pandas: Pandas 1.0.3.* 2020.
20. Pedregosa Fabian, Varoquaux Gaël, Gramfort Alexandre, Michel Vincent, Thirion Bertrand, Grisel Olivier, Blondel Mathieu, Prettenhofer Peter, Weiss Ron, Dubourg Vincent, Vanderplas Jake, Passos Alexandre, Cournapeau David, Brucher Matthieu, Perot Matthieu, Duchesnay Édouard. *Scikit-learn: machine learning in Python.* *J Machine Learn Res.* 2011;12(85):2825–30.
21. Virtanen P, Gommers R, Oliphant TE., Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt StéfanJ., Brett M, Wilson J, Millman KJ, Mayorov N, Nelson Andrew RJ., Jones E, Kern R, Larson E, Carey CJ., Polat İlhan,Y, Moore EW., VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA., Harris CR., Archibald AM., Ribeiro Antônio H., Pedregosa Fabian, van Mulbregt Paul. *SciPy 1.0: fundamental algorithms for scientific computing in Python.* *Nature Method* 17(3):261–272. 2020.
22. Lundberg SM, Lee Su-In. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems.* 30 Curran Associates Inc 2017.
23. ...Abadi Martín, Agarwal Ashish, Barham Paul, Brevdo Eugene, Chen Zhifeng, Citro Craig, Corrado Greg S, Davis Andy, Dean Jeffrey, Devin Matthieu, Ghemawat Sanjay, Goodfellow Ian, Harp Andrew, Irving Geoffrey, Isard Michael, Jia Yangqing, Jozefowicz Rafal, Kaiser Lukasz, Kudlur Manjunath, Levenberg Josh, Mané Dandelion, Monga Rajat, Moore Sherry, Murray Derek, Olah Chris, Schuster Mike, Shlens Jonathon, Steiner Benoit, Sutskever Ilya, Talwar Kunal, Tucker Paul, Vanhoucke Vincent, Vasudevan Vijay, Viégas Fernanda, Vinyals Oriol, Warden Pete, Wattenberg Martin, Wicke Martin, Yuan Yu. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, TensorFlow: Large-scale machine learning on heterogeneous systems., TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems;* 2015.
24. Herbert Robbins, Sutton Monro. Stochastic approximation method. *Ann Math Stat.* 1951;22(3):400–7.
25. Jack Kiefer, Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *Ann Math Stat.* 1952;23(3):462–6.
26. Tieleman T, Hinton G. 2012. Lecture 65-rmsprop divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning.* 4(2):26
27. Duchi John, Hazan Elad, Singer Yoram. Adaptive subgradient methods for online learning and stochastic optimization. *J Machine Learn Res.* 2011;12(61):2121–59.
28. Zeiler Matthew D. ADADELTA: An adaptive learning rate method, 2012. [arXiv:1212.5701](https://arxiv.org/abs/1212.5701) [cs].
29. Kingma Diederik P, Ba Jimmy. Adam: A method for Stochastic optimization. 2017 [perimagedhttp://arxiv.org/abs/1412.6980](https://arxiv.org/abs/1412.6980)arXiv:1412.6980 [cs].
30. Dozat T. Incorporating Nesterov Momentum into Adam. 2016.

31. Ruiz-Canela Miguel, Guasch-Ferré Marta, Toledo Estefanía, Clish Clary B, Razquin Cristina, Liang Liming, Wang Dong D, Corella Dolores, Estruch Ramón, Hernández Álvaro, Edward Yu, Gómez-Gracia Enrique, Zheng Yan, Arós Fernando, Romaguera Dora, Dennis Courtney, Ros Emilio, Lapetra José, Serra-Majem Lluís, Papandreou Christopher, Portoles Olga, Fitó Montserrat, Salas-Salvadó Jordi, Hu Frank B, Martínez-González Miguel A. Plasma branched chain/aromatic amino acids, enriched Mediterranean diet and risk of type 2 diabetes: case-cohort study within the PREDIMED Trial. *Diabetologia*. 2018;61(7):1560–71.
32. Lin Rui, Liu Wentian, Piao Meiyu, Zhu Hong. A review of the relationship between the gut microbiota and amino acid metabolism. *Amino Acids*. 2017;49(12):2083–90.
33. Ashniev German A, Petrov Sergey N, Iablokov Stanislav N, Rodionov Dmitry A. genomics-based reconstruction and predictive profiling of amino acid biosynthesis in the human gut microbiome. *Microorganisms*. 2022;10(4):740.
34. Khatik G, Datusalia A, Ahsan W, Kaur P, Vyas M, Mittal A, Nayak S. A retrospect study on thiazole derivative as the potential antidiabetic agents in drug discovery and developments. *Current Drug Discovery Technol*. 14. 2017.
35. Pácal Lukáš, Kuricová Katarína, Kaňková Kateřina. Evidence for altered thiamine metabolism in diabetes: Is there a potential to oppose gluco- and lipotoxicity by rational supplementation? *World J Diabetes*. 2014;5(3):288–95.
36. Wang Lu, Pi Zifeng, Liu Shu, Liu Zhiqiang, Song Fengrui. Targeted metabolome profiling by dual-probe microdialysis sampling and treatment using *Gardenia jasminoides* for rats with type 2 diabetes. *Sci Rep*. 2017;7:10105.
37. Pillwein K, Reardon MA, Jayaram HN, Natsumeda Y, Elliott WL, Faderan MA, Prajda N, Sperl W, Weber G. Insulin regulatory effects on purine- and pyrimidine metabolism in alloxan diabetic rat liver. *Padiatrie Padologie*. 1988;23(2):135–44.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.