Explainable AI-based analysis of human pancreas sections detects traits of type 2 diabetes

Lukas Klein[*1,2,3], **Sebastian Ziegler**[*2,3], **Felicia Gerst**[*4,5,6], **Yanni Morgenroth**[*6,7,8], **Karol Gotkowski**[2,3], **Eyke Schöniger**[6,7,8], **Nicole Kipke**[6,7,8], **Annika Seiler**[6,7,8], **Ellen Geibelt**[11], **Martin Heni**[9], **Silvia Wagner**[12], **Silvio Nadalin**[12], **Falko Fend**[13], **Daniela Aust**[14], **Andre Mihaljevic**[12], **Daniel Hartmann**[12], **Jurgen Weitz**[6,7,8], **Reiner Jumpertz-von Schwartzenberg**[4,5,6], **Marius Distler**[6,7,8], **Andreas Birkefeld**[4,5,6], **Susanne Ullrich**[4,5,6], **Paul F Jaeger**[**2,3], **Fabian Isensee**[**2,3], **Michele Solimena**[**6,7,8], **Robert Wagner**[**6,15,16]

[1] ETH Zürich, Department of Computer Science, Switzerland
[2] Helmholtz Imaging, DKFZ, Germany
[3] Division of Medical Image Computing, DKFZ, Heidelberg, Germany
[4] Institute for Diabetes Research and Metabolic Diseases of the Helmholtz Center Munich at the Eberhard Karls University of Tuebingen (IDM), Tuebingen, Germany
[5] Internal Medicine IV, Endocrinology, Diabetology and Nephrology, University Hospital Tuebingen, Tuebingen, Germany
[6] German Center for Diabetes Research (DZD e.V.) Neuherberg, Germany
[7] Department of Molecular Diabetology, University Hospital Carl Gustav Carus, Medical Faculty, Technische Universität Dresden, Dresden, Germany
[8] Paul Langerhans Institute Dresden (PLID) of the Helmholtz Center Munich at the University Hospital Carl Gustav Carus and Faculty of Medicine of the TU Dresden, Dresden, Germany
[9] Internal Medicine I, Endocrinology and Diabetology, University Hospital Ulm, Ulm, Germany
[10] Institute for Clinical Chemistry and Pathobiochemistry, Department for Diagnostic Laboratory Medicine, University Hospital Tübingen, Tübingen, Germany
[11] Center for Molecular and Cellular Bioengineering at the Technische Universität Dresden Light Microscopy Facility (TUD CMCB LMF)
[12] Department of General, Visceral and Transplant Surgery, University Hospital Tuebingen, Tuebingen, Germany
[13] Institute of Pathology and Neuropathology and Comprehensive Cancer Center, University Hospital Tübingen, Tübingen, Germany
[14] Department of Pathology, University Hospital Carl Gustav Carus, Medical Faculty, Technische Universität Dresden, Dresden, Germany
[15] Department of Endocrinology and Diabetology, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[16] Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Düsseldorf, Germany

lukas.klein@dkfz.de, sebastian.ziegler@dkfz.de

* contributed equally first
** contributed equally last

Interpretable AI: Past, Present and Future Workshop at NeurIPS 2024

Abstract

Type 2 diabetes (T2D) is a chronic disease currently affecting around 500 million people worldwide and potentially leading to severe health conditions. Yet, the causes for the underlying beta-cell failure leading to impaired insulin secretion are not fully understood, especially on a morphological level. While giga-pixel microscopy images may visualize such subtle morphological differences, the dimensionality and variability of the data quickly surpass the limits of human analysis. In response, we collected a dataset consisting of pancreas whole-slide images stained with multiple chromogenic and multiplex fluorescent stainings and trained various deep learning models to predict the T2D status. Using explainable AI (XAI) methods, we rendered the learned relationships humanly understandable, quantified them as comprehensive biomarkers, and utilized statistical modeling to assess their association with T2D. Our analysis reveals the contributions of adipocytes, pancreatic islets, and fibrotic patterns to T2D.

1 Introduction

Based on the current WHO classification of diabetes, over 90% of all persons with the disease fall into the category defined as type 2 diabetes (T2D). T2D is a major global health issue, affecting millions and placing a significant burden on healthcare systems [1]. Drivers of T2D are insulin resistance and impaired insulin secretion, with dysfunction of pancreatic islet beta-cells being a key feature. Despite extensive research, the exact nature of beta-cell failure in T2D remains unclear. Given the high degree of inter-individual variability, no specific traits are sufficient for pathologists to discriminate whether a pancreatic sample belongs to a subject with or without T2D. Pancreatic tissue, represented in giga-pixel whole-slide images (WSIs), lacks clear features that differ significantly between individuals with and without T2D.

Our goal was therefore to assess whether the exploitation of an imaging data-driven approach combining attention-based DL models with explainable artificial intelligence (XAI) would enable the attribution of a pancreatic patient to a patient with or without T2D with high reliability. While DL models could find patterns in data infeasible to detect by doctors, XAI methods subsequently rendered the learned relationships human-understandable, revealing regions of interest (ROIs) associated with the occurrence of T2D. Given the limited feasibility of qualitative analyses of large amounts of XAI results, we quantified the attention to the ROIs and subsequently computed specific biomarkers for the most important ones. At last, we analyzed these biomarkers in combination with clinical patient data using statistical models. Besides the integration of vast amounts of diverse WSIs, our approach offered the advantage that we do not bias ourselves to prior assumptions regarding T2D as the XAI application can also uncover unanticipated findings, leading to the formulation of new hypotheses about the disease.

2 Classification of T2D from WSIs of pancreatic tissue sections

Dataset We first applied both single chromogenic and multiplex fluorescence techniques to immunostain pancreas sections from 100 patients with (35) or without (65) T2D who underwent pancreatectomy [2–5] at two medical centers. Immunostained antigens included glucagon, insulin, and somatostatin as markers of islet α-, β- and δ-cells, respectively, as well as PECAM1 for endothelial cells (blood vessels), perilipin 1 for adipocytes (fat cells) and tubulin beta 3 for neuronal axons. In the case of the chromogenic stainings, each marker was detected individually in serial sections counterstained with hematoxylin. In the case of multiplex fluorescent stainings, serial sections were incubated with DAPI and primary antibodies either against glucagon, somatostatin, and tubulin beta 3 (Stainingset 1), or against insulin, PECAM1, and perilipin 1 (Stainingset 2) (see Figure 1 (a)). Brightfield and fluorescent WSIs were then used to train DL models distinguishing donors with or without T2D (see Figure 1 (b)). See Appendix A for a full description of the cohort and staining procedures. Patients were split into a training set (75) and a test set (25). For the chromogenic data, where each of the six stainings is available as an individual RGB WSI, separate models per staining were trained. In contrast, the multiplex fluorescent data contained multiple stainings in one WSI via different image channels. We compared different representations of these WSIs by either treating each of the three channels as color channel (RGB), encoding each of the channels individually and

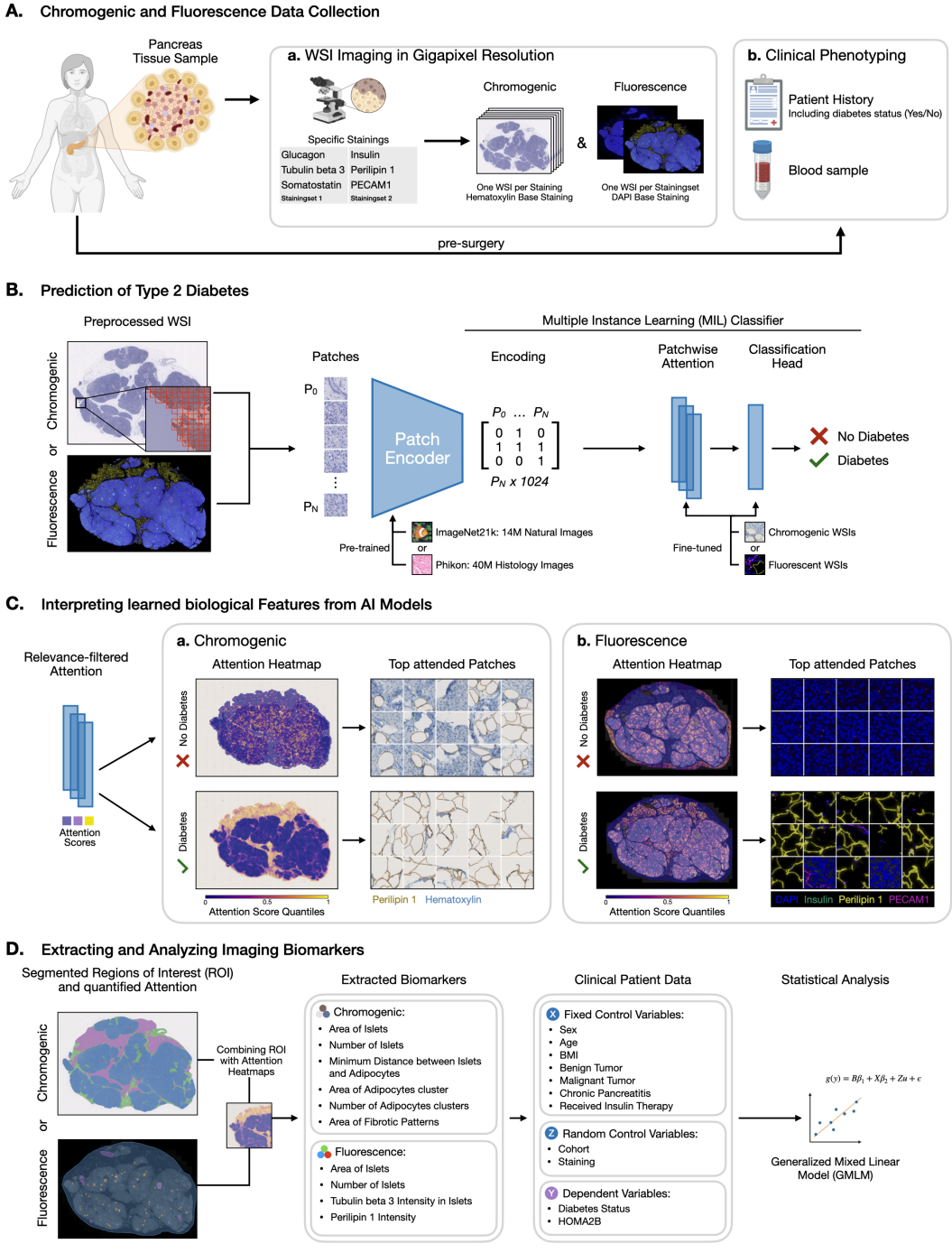

Figure 1: **(A)** We acquired serial pancreatic tissue sections, fasting blood samples, and clinical data from 100 patients. Further, we used six chromogenic stainings for brightfield and two staining sets for multiplex fluorescence microscopy to capture the WSIs. **(B)** Multiple Instance Learning (MIL) architecture with different pretrained encoders and fine-tune MIL classifier. **(C)** XAI methods were used to identify regions of interest (ROIs) utilized by the models for their prediction. **(D)** ROIs were segmented and quantified, with extracted biomarkers and clinical covariates analyzed using statistical models.

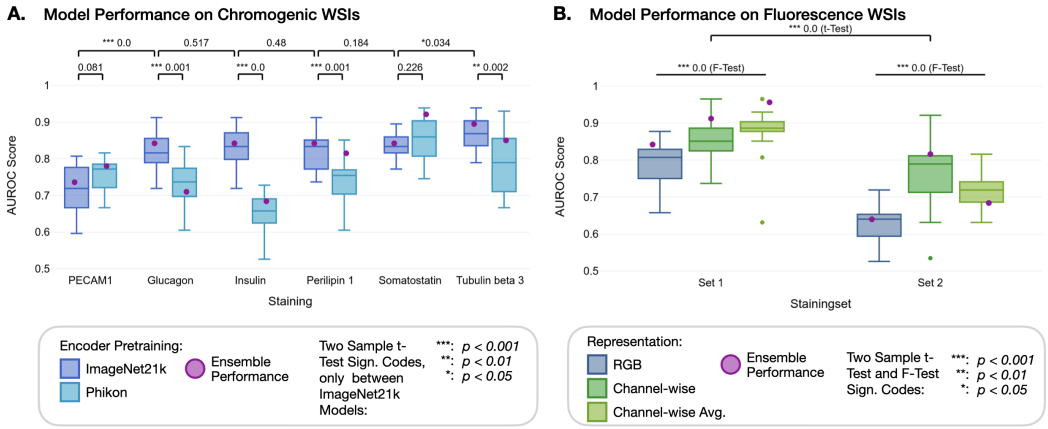

Figure 2: Performance efficiency of trained models to predict type 2 diabetes on chromogenic **(A)** and fluorescent **(B)** WSIs. AUROCs show ImageNet21k/CLAM outperforms Phikon/CLAM in most cases, with tubulin beta 3 achieving the highest performance, and Stainingset 1 with channel-wise average representation performing best on fluorescent WSIs.

append all resulting feature vectors (channel-wise) or encoding each of the channels individually and averaging the resulting feature vectors (channel-wise avg).

Model Performance WSI classification is a multiple instance learning (MIL) problem since an entire image is too large to be processed at once while a classification label is only available globally but not per patch. We used CLAM [6] which first encodes individual patches via a pre-trained feature extractor and then aggregates the features using attention mechanisms followed by a classification layer to obtain the WSI-level prediction. This allows for a direct interpretability of patch importance by visualizing the attention scores over the WSI. See Appendix B for further experiments with other MIL models. For the chromogenic data we tested two vision transformers either pre-trained on Imagenet21k [7] or Phikon [8], a foundation model trained on thousands of WSIs), while on the fluorescent data we only used Imagenet21k. When applied to chromogenic brightfield WSIs, the ImageNet21k pre-training delivered the best prediction performance on average (AUROC = 0.833; Figure 2 (a) and Appendix Table 3). Across all stainings except for PECAM1 and somatostatin, there was a significant increase in AUROC when using the ImageNet21k pre-trained encoder. When ImageNet21k/CLAM-based prediction was conducted using fluorescent WSIs, Stainingset 1, despite the absence of insulin staining, yielded significantly better classification results than Stainingset 2. The best performance was reached using the channel-wise avg representation on Stainingset 1 (Ensemble AUROC=0.956; Figure 2 (b) and Appendix Table 2).

3 AI models attend to specific biological traits

Upon completion of model training, we aimed to understand the biological features utilized by the models in predicting diabetes. For this purpose, we employed XAI techniques specifically within the domains of Attention and Attribution methods, to highlight regions critical to the models when predicting the T2D status Figure 1 (c). We applied these techniques to the best classification model among the evaluated training settings for both fluorescent and chromogenic modalities.

Initially, we focused on determining the significance of specific patches using the built-in attention mechanism of the multiple instance learning (MIL) classifier and creating attention heatmaps for the individual chromogenic and fluorescent WSIs of each patient Figure 3. The heatmaps of the chromogenic WSIs revealed both similarities and distinctions in the regions important to the model when classifying a WSI to diabetes or its absence. Specifically, for the "diabetes" outcome (Figure 3 (a), lower panels), several stainings showed heightened attention to fibrotic patterns (left side of the PECAM1-, glucagon-, insulin-, and somatostatin-stained WSIs) as well as a pronounced focus on adipocytes (top left of the perilipin 1-stained WSI). Both recognized features, i.e. fibrotic changes and adipocyte infiltration are mainly localized in the exocrine tissue. Notably, in the chromogenic WSIs

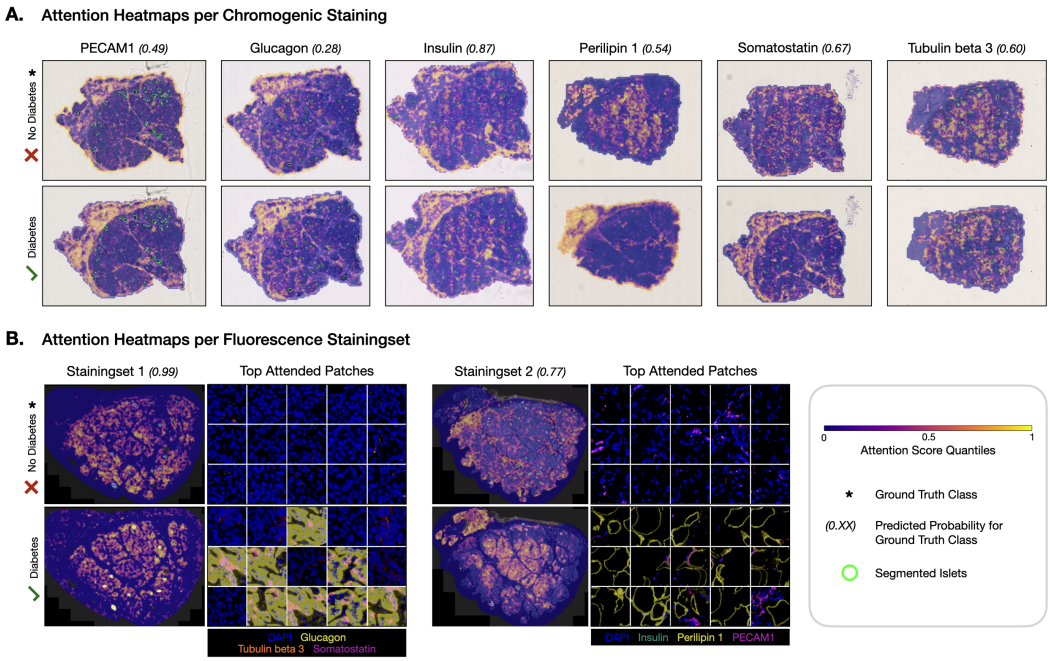

Figure 3: Representative attention heatmaps for the chromogenic **(A)** and fluorescent **(B)** WSIs of a single patient. Top 15 attended fluorescent patches associated with T2D or no T2D status.

the fibrotic patterns were revealed by low-intensity diffuse staining of the tissue by diaminobenzidine precipitates, regardless of antigen specificity, accounting for spatially restricted, high-intensity signals. Conversely, for the "no T2D" outcome, the models demonstrate significantly less attention to fibrotic patterns. Instead, they concentrate on tissue borders and display more dispersed attention patterns. The heatmaps for the fluorescent WSI of both Stainingsets revealed similar attention patterns such as on adipocytes (perilipin 1) in Stainingset 2 Figure 3 (b).

While these heatmaps facilitate the visualization of larger ROIs in the tissue, they do not provide detailed insights into the finer biological features that the model considers significant. To address this, we first sampled the top attended patches using attention scores (Figure 3 (b), see Appendix G for chromogenic WSIs), essentially zooming in from image-level to patch-level. In the case of fluorescent WSI, these heatmaps reveal that for Stainingset 2, the model indeed exhibits a high level of attention to adipocytes (perilipin 1). However, for Stainingset 1, the model focuses also on islet α-cells (glucagon), when classifying "diabetes". This focus of the model on α-cells when classifying a patient as having diabetes is not observable from the patch-level heatmap alone.

To further determine the specific pixel-level features utilized by the model in each of the top-attended patches, we employed multiple attribution methods to quantify the contribution of individual pixels to the predicted outcomes. The model was indeed able to recognize biological traits, such as cell nuclei, adipocyte cytosolic compartments, inter-cellular structures, or islets Appendix H. Integrating all XAI findings, we compiled a comprehensive report for each patient, imaging modality, and DL model.

4 Quantifying Attention to regions-of-interests

To quantitatively assess the global importance of specific regions across all heatmaps, we implemented a scalable, data-driven methodology. We examined highly attended patches to determine their corresponding tissue types. Those reoccurring were defined as ROIs and segmented in the WSI using a specifically trained DL segmentation model to enable quantitative analysis of total attention to different relevant tissue types (similar to localization metrics in XAI). The attention scores within ROIs were standardized for tissue size and total amount of attention. Our analysis of the attention scores revealed a heterogeneous distribution of attention toward tissue components such as islets, adipocytes, or fibrotic patterns in the individual chromogenic stainings. Attention to islets showed the highest z-scores for insulin and glucagon-stained WSIs, but it did not depend on diabetes status

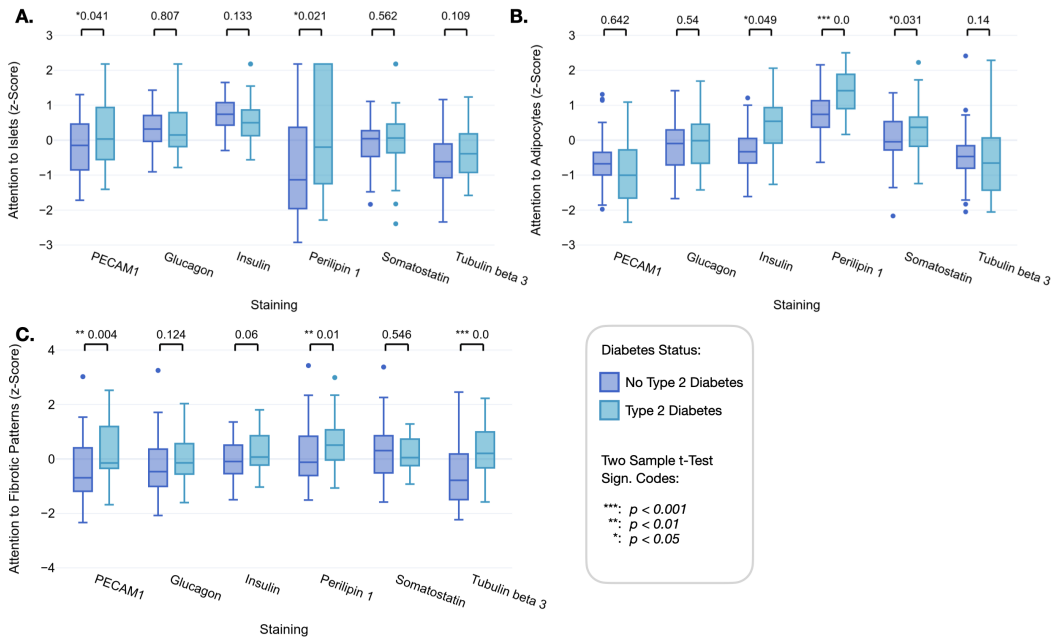

Figure 4: Z-scores for average attributed attention to **(A)** islets, **(B)** adipocytes, and **(C)** fibrosis for the individual chromogenic stainings of patients with and without T2D.

Figure 4 (a). The attention to adipocytes showed significant variability across different stainings, with the highest z-scores for the perilipin 1 staining Figure 4 (b). Notably, regardless of the stained antigens, the model's attention to fibrosis in the chromogenic WSIs was significantly higher in patients with T2D on the PECAM1, perilipin 1 and tubulin beta 3 stainings, indicating its potential relevance in the pathology of the disease Figure 4 (c).

5 Results

A. GMLM Regression for Diabetes Status (Chromogenic)

Fixed Effects	Estimate	SE	P-Value (*p*)	Sign.
Intercept	-1.607	0.290	0.000	***
Area of Islets	-0.237	0.192	0.216	
Number of Islets	0.088	0.105	0.404	
Area of Adipocytes clusters	0.266	0.126	0.035	*
Num. of Adipocytes clusters	-0.401	0.158	0.011	*
Min. Dist. Islets and Adipocytes	-0.633	0.237	0.007	**
Area of Fibrotic Patterns	0.524	0.124	0.000	***
Sex	0.610	0.226	0.007	**
Age	0.741	0.137	0.000	***
BMI	0.538	0.114	0.000	***
Malignant Tumor	0.330	0.240	0.170	
Chronic Pancreatitis	1.420	0.394	0.000	***

Random Effects	Type	Variance	STD
Staining	Intercept	0.000	0.000
Cohort	Intercept	0.176	0.483

Observations	600
Groups	{Staining: 6, Cohort: 2}
Log-likelihood	-298.010
AIC	624.010

Note * *p*<0.05, ** *p*<0.01, *** *p*<0.001

B. MLM Regression for HOMA2B (Chromogenic)

Fixed Effects	Estimate	SE	P-Value (*p*)	Sign.
Intercept	0.342	0.230	0.262	
Area of Islets	0.028	0.037	0.442	
Number of Islets	0.069	0.034	0.043	*
Area of Adipocytes clusters	0.216	0.042	0.000	***
Num. of Adipocytes clusters	-0.189	0.40	0.000	***
Min. Dist. Islets and Adipocytes	0.053	0.040	0.184	
Area of Fibrotic Patterns	-0.132	0.038	0.001	***
Size of Adipocytes clusters:BMI	-0.071	0.031	0.020	*
Sex	-0.179	0.067	0.008	**
Age	-0.125	0.036	0.001	***
BMI	0.218	0.039	0.000	***
Insulin Therapy	-1.460	0.098	0.000	***
Malignant Tumor	0.074	0.075	0.320	
Chronic Pancreatitis	0.254	0.124	0.041	*

Random Effects	Type	Variance	STD
Staining	Intercept	0.000	0.000
Cohort	Intercept	0.096	0.311

Observations	546
Groups	{Staining: 6, Cohort: 2}
Log-likelihood	-615.653
AIC	1265.307

Note * *p*<0.05, ** *p*<0.01, *** *p*<0.001

Table 1: Regression results of the generalized mixed linear model analysis for diabetes status **(A)** and HOMA2B **(B)** level based on chromogenic.

Based on XAI results and segmentation maps, we computed various imaging biomarkers related to T2D status, noting differences between chromogenic and fluorescence modalities, particularly where certain ROIs, like fibrotic patterns, couldn't be assessed in fluorescent WSIs due to the lack of dedicated staining. To evaluate the impact of these biomarkers on T2D status and insulin secretion (HOMA2B), we used a generalized mixed linear model (GMLM), controlling for patient traits and random effects associated with the cohort and staining method. See Figure 1 (d) for a list of all biomarkers and control variables and Table 1 for the chromogenic results (Appendix I for the fluorescence). Together with the XAI results, we derived the following three main findings:

Amount of Adipocytes affect Islets. The analysis of the chromogenic biomarkers revealed that the area of adipocyte clusters is positively associated with T2D, whereas the number of adipocyte clusters was inversely correlated with the disease (see Table 1 (a)). A higher number of adipocyte clusters may indicate a more advanced adipose infiltration of pancreatic tissue, i.e. intralobular adipocytes [9]. Such an intralobular infiltration would allow the adipocytes to get closer to the islets and thereby impact the islet microenvironment with negative consequences for insulin secretion [10]. Indeed we observe, that for patients with T2D, the pancreatic islets are on average closer located to adipocytes than for patients without T2D (see "Min. Dist. Islets and Adypocytes" Covariate in Table 1).

Fibrosis as a potential structural alteration by T2D. The peroxidase reaction used in chromogenic staining made it possible to see areas rich in extracellular matrix (ECM) within pancreatic tissue, clearly distinguishing between the main tissue structures and the surrounding connective tissue, also reported by de Haan et al. [11]. DL models showed greater attention to these ECM-rich areas in PECAM1 and perilipin 1 stainings when predicting T2D. Our results in Table 1 verify a positive association between the amount of fibrosis and T2D, independent of pancreatic tumors or insulin therapy, highlighting fibrosis as a potential structural alteration in T2D.

Significance of tubulin beta 3 for T2D detection. Chromogenic mono stainings had lower predictive performances than immunofluorescence images, with the highest average performance found for tubulin beta 3, which was also a key component of the best-performing fluorescence Stainingset 1. Notably, tubulin beta 3 and perilipin 1 received higher attention in T2D samples compared to non-T2D WSIs. These findings suggest that alterations in tubulin beta 3 structures may be a distinguishing feature of T2D, potentially impacting islet cell function [12, 13].

6 Conclusion

In this study, we analyzed a unique dataset of human pancreatic WSIs from T2D and non-T2D donors, enabling the training of deep learning models that accurately predict T2D status. This outcome is novel and represents a remarkable advancement in the field of DL-based biomarker discovery, especially because the pancreas, a central organ in the pathophysiology of T2D, exhibits only subtle anatomical changes in this condition, unlike in the case of cancer or pancreatitis. See Appendix F for a discussion of the limitations in the experiment setting. Our XAI-based analysis uncovered biomarkers that support hypotheses about the involvement of adipocytes, pancreatic islets, and enrichment of fibrotic patterns with T2D. This highlights the complexity of T2D pathogenesis and underscores the potential of DL and XAI methods in uncovering novel insights from large-scale microscopy data.

References

[1] International Diabetes Federation. *IDF Diabetes Atlas*, volume 10th edn. International Diabetes Federation, 2021. URL `https://www.diabetesatlas.org`.

[2] Michele Solimena, Anke M. Schulte, Lorella Marselli, Florian Ehehalt, Daniela Richter, Manuela Kleeberg, Hassan Mziaut, Klaus-Peter Knoch, Julia Parnis, Marco Bugliani, Afshan Siddiq, Anne Jörns, Frédéric Burdet, Robin Liechti, Mara Suleiman, Daniel Margerie, Farooq Syed, Marius Distler, Robert Grützmann, Enrico Petretto, Aida Moreno-Moral, Carolin Wegbrod, Anke Sönmez, Katja Pfriem, Anne Friedrich, Jörn Meinel, Claes B. Wollheim, Gustavo B. Baretton, Raphael Scharfmann, Everson Nogoceke, Ezio Bonifacio, Dorothée Sturm, Birgit Meyer-Puttlitz, Ugo Boggi, Hans-Detlev Saeger, Franco Filipponi, Mathias Lesche, Paolo Meda, Andreas Dahl, Leonore Wigger, Ioannis Xenarios, Mario Falchi, Bernard Thorens, Jürgen Weitz, Krister Bokvist, Sigurd Lenzen, Guy A. Rutter, Philippe Froguel, Manon von Bülow, Mark Ibberson, and Piero Marchetti. Systems biology of the IMIDIA biobank from organ

donors and pancreatectomised patients defines a novel transcriptomic signature of islets from individuals with type 2 diabetes. *Diabetologia*, 61(3):641–657, March 2018. ISSN 1432-0428. doi: 10.1007/s00125-017-4500-3.

[3] Felicia Gerst, Robert Wagner, Gabriele Kaiser, Madhura Panse, Martin Heni, Jürgen Machann, Malte N. Bongers, Tina Sartorius, Bence Sipos, Falko Fend, Christian Thiel, Silvio Nadalin, Alfred Königsrainer, Norbert Stefan, Andreas Fritsche, Hans-Ulrich Häring, Susanne Ullrich, and Dorothea Siegel-Axel. Metabolic crosstalk between fatty pancreas and fatty liver: effects on local inflammation and insulin secretion. *Diabetologia*, 60(11):2240–2251, November 2017. ISSN 1432-0428. doi: 10.1007/s00125-017-4385-1.

[4] Felicia Gerst, Benjamin A Jaghutriz, Harald Staiger, Anke M Schulte, Estela Lorza-Gil, Gabriele Kaiser, Madhura Panse, Sieglinde Haug, Martin Heni, Monika Schütz, Mandy Stadion, Annette Schürmann, Flavia Marzetta, Mark Ibberson, Bence Sipos, Falko Fend, Thomas Fleming, Peter P Nawroth, Alfred Königsrainer, Silvio Nadalin, Silvia Wagner, Andreas Peter, Andreas Fritsche, Daniela Richter, Michele Solimena, Hans-Ulrich Häring, Susanne Ullrich, and Robert Wagner. The Expression of Aldolase B in Islets Is Negatively Associated With Insulin Secretion in Humans. *The Journal of Clinical Endocrinology and Metabolism*, 103(12):4373–4383, September 2018. ISSN 0021-972X. doi: 10.1210/jc.2018-00791. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6915830/.

[5] Anna L. Gloyn, Mark Ibberson, Piero Marchetti, Alvin C. Powers, Patrik Rorsman, Maike Sander, and Michele Solimena. Every Islet Matters: Improving the Impact of Human Islet Research. *Nature metabolism*, 4(8):970–977, August 2022. ISSN 2522-5812. doi: 10.1038/s42255-022-00607-8. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11135339/.

[6] Jonathan H. Lu, Alison Callahan, Birju S. Patel, Keith E. Morse, Dev Dash, and Nigam H. Shah. Low adherence to existing model reporting guidelines by commonly used clinical prediction models. preprint, Health Informatics, July 2021. URL http://medrxiv.org/lookup/doi/10.1101/2021.07.21.21260282.

[7] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik. ImageNet-21K Pretraining for the Masses. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1, December 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/98f13708210194c475687be6106a3b84-Abstract-round1.html.

[8] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling, September 2023. URL https://www.medrxiv.org/content/10.1101/2023.07.21.23292757v2. Pages: 2023.07.21.23292757.

[9] Jordan J. Wright, Adel Eskaros, Annika Windon, Rita Bottino, Regina Jenkins, Amber M. Bradley, Radhika Aramandla, Sharon Philips, Hakmook Kang, Diane C. Saunders, Marcela Brissova, and Alvin C. Powers. Exocrine Pancreas in Type 1 and Type 2 Diabetes: Different Patterns of Fibrosis, Metaplasia, Angiopathy, and Adiposity. *Diabetes*, 73(7):1140–1152, October 2023. ISSN 0012-1797. doi: 10.2337/db23-0009. URL https://doi.org/10.2337/db23-0009.

[10] Róbert Wagner, Benjamin Assad Jaghutriz, Felicia Gerst, Morgana Barroso Oquendo, Jürgen Machann, Fritz Schick, Markus W Löffler, Silvio Nadalin, Falko Fend, Alfred Königsrainer, Andreas Peter, Dorothea Siegel-Axel, Susanne Ullrich, Hans-Ulrich Häring, Andreas Fritsche, and Martin Heni. Pancreatic Steatosis Associates With Impaired Insulin Secretion in Genetically Predisposed Individuals. *The Journal of Clinical Endocrinology & Metabolism*, 105(11):3518–3525, November 2020. ISSN 0021-972X. doi: 10.1210/clinem/dgaa435. URL https://doi.org/10.1210/clinem/dgaa435.

[11] Kevin de Haan, Yijie Zhang, Jonathan E. Zuckerman, Tairan Liu, Anthony E. Sisk, Miguel F. P. Diaz, Kuang-Yu Jen, Alexander Nobori, Sofia Liou, Sarah Zhang, Rana Riahi, Yair Rivenson, W. Dean Wallace, and Aydogan Ozcan. Deep learning-based transformation of H&E stained tissues into special stains. *Nature Communications*, 12(1):4884, August 2021.

ISSN 2041-1723. doi: 10.1038/s41467-021-25221-2. URL https://www.nature.com/articles/s41467-021-25221-2. Publisher: Nature Publishing Group.

[12] Rollie F. Hampton, Maria Jimenez-Gonzalez, and Sarah A. Stanley. Unravelling innervation of pancreatic islets. *Diabetologia*, 65(7):1069–1084, July 2022. ISSN 1432-0428. doi: 10.1007/s00125-022-05691-9. URL https://doi.org/10.1007/s00125-022-05691-9.

[13] Alexandra Alvarsson, Maria Jimenez-Gonzalez, Rosemary Li, Carolina Rosselot, Nikolaos Tzavaras, Zhuhao Wu, Andrew F. Stewart, Adolfo Garcia-Ocaña, and Sarah A. Stanley. A 3D atlas of the dynamic and regional variation of pancreatic innervation in diabetes. *Science Advances*, 6(41):eaaz9124, October 2020. doi: 10.1126/sciadv.aaz9124. URL https://www.science.org/doi/10.1126/sciadv.aaz9124. Publisher: American Association for the Advancement of Science.

[14] Rohit Babbar, Martin Heni, Andreas Peter, Martin Hrabě de Angelis, Hans-Ulrich Häring, Andreas Fritsche, Hubert Preissl, Bernhard Schölkopf, and Róbert Wagner. Prediction of Glucose Tolerance without an Oral Glucose Tolerance Test. *Frontiers in Endocrinology*, 9, March 2018. ISSN 1664-2392. doi: 10.3389/fendo.2018.00082. URL https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2018.00082/full. Publisher: Frontiers.

[15] Jonathan C Levy, David R Matthews, and Michel P Hermans. Correct Homeostasis Model Assessment (HOMA) Evaluation Uses the Computer Program. *Diabetes Care*, 21(12):2191–2192, December 1998. ISSN 0149-5992. doi: 10.2337/diacare.21.12.2191. URL https://doi.org/10.2337/diacare.21.12.2191.

[16] American Diabetes Association. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2021. *Diabetes Care*, 44(Supplement_1):S15–S33, December 2020. ISSN 0149-5992. doi: 10.2337/dc21-S002. URL https://doi.org/10.2337/dc21-S002.

[17] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31–71, January 1997. ISSN 0004-3702. doi: 10.1016/S0004-3702(96)00034-3. URL https://www.sciencedirect.com/science/article/pii/S0004370296000343.

[18] Pierre Courtiol, Eric W. Tramel, Marc Sanselme, and Gilles Wainrib. Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach, February 2020. URL http://arxiv.org/abs/1802.02212. arXiv:1802.02212 [cs, stat].

[19] Hila Chefer, Shir Gur, and Lior Wolf. Transformer Interpretability Beyond Attention Visualization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-66544-509-2. doi: 10.1109/CVPR46437.2021.00084. URL https://ieeexplore.ieee.org/document/9577970/.

[20] Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs), June 2023. URL http://arxiv.org/abs/1606.08415. arXiv:1606.08415 [cs].

[21] Karol Gotkowski, Carsten Lüth, Paul F. Jäger, Sebastian Ziegler, Lars Krämer, Stefan Denner, Shuhan Xiao, Nico Disch, Klaus H. Maier-Hein, and Fabian Isensee. Embarrassingly Simple Scribble Supervision for 3D Medical Segmentation, March 2024. URL http://arxiv.org/abs/2403.12834. arXiv:2403.12834 [cs].

[22] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, February 2021. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-020-01008-z. URL http://www.nature.com/articles/s41592-020-01008-z.

[23] Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23(4):550–560, December 1997. ISSN 0098-3500. doi: 10.1145/279232.279236. URL https://doi.org/10.1145/279232.279236.

Appendix

A Clinical Cohort

We analyzed clinical, laboratory, and histologic data obtained within the "Studying Islets from Living Donors" (SILDS) programs. Patients undergoing pancreatic surgery for different indications provided written informed consent to donate blood samples and pancreas tissue, and share health records and laboratory data for research purposes at both study sites. We obtained macroscopically healthy tissue resected during surgery but not required for further pathology workup. All patients were of European ethnicity. Additionally, fasting blood was drawn pre-surgery for detailed metabolic phenotyping. Fasting glucose and C-peptide levels were measured as previously described [14], and homeostatic model assessment (HOMA) of insulin secretion was calculated using the computer model-based HOMA-2B [15]. None of the participants had depleted endogenous insulin production as measured by C-peptide-based HOMA2B (lowest HOMA2B: 6% with a diabetes duration of 24 years), excluding type 1 diabetes among the participants. Information on medical history was collected by a physician. Documented by their health records, T2D patients were diagnosed as having T2D at least one year before admission to pancreatic surgery. This excludes diabetes in the context of exocrine pancreatic disease. In contrast, patients without diabetes neither had diabetes nor did they fulfill diagnostic criteria of T2D based on glycated hemoglobin (HbA1c) and fasting glucose, as defined by the American Diabetes Association [16]. Data Acquisition

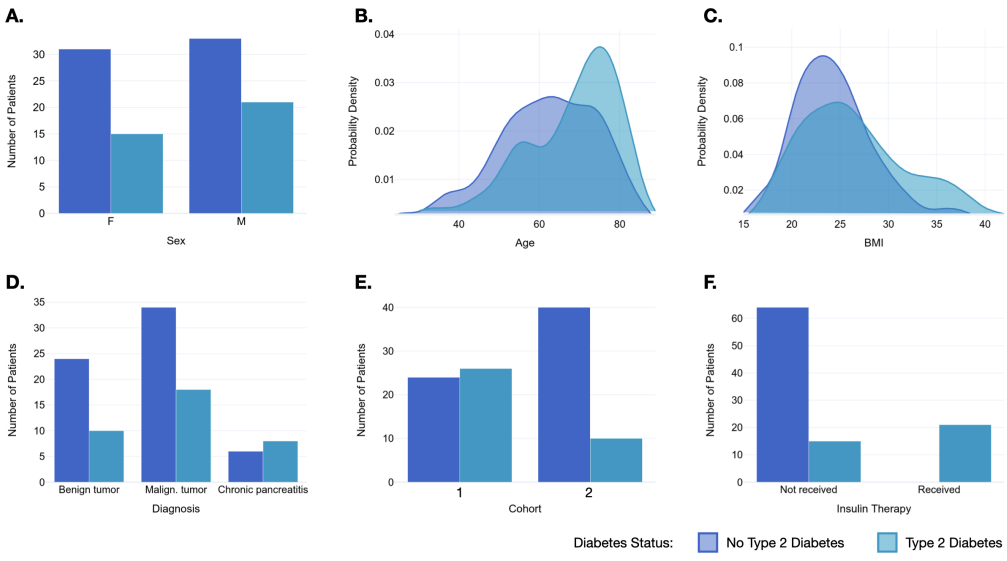

Figure 5: Distribution of Hospital Patient Data. Distribution of all control variables between both T2D statuses, excluding staining: sex **(A)**, age **(B)**, BMI **(C)**, diagnosis of related illness **(D)**, cohort **(E)**, and insulin therapy **(F)**.

Hospital Patient Data. The patient cohort-related metadata are summarized in Figure 5. The quantitative analysis of these parameters in Figure 5 revealed an unbalanced distribution of sex, age, BMI, tumor type, chronic pancreatitis, and cohort between the patients with and without diabetes. In 9 patients, fasting blood samples were not obtained prior to surgery and therefore HOMA2B was not calculated.

Chromogenic immunostaining and brightfield microscopy. Formalin-fixed, paraffin-embedded (FFPE) pancreatic sections (2-4 µm thick) were processed using an automatic slide stainer BenchmarkUltra (Ventana Technology, Roche Diagnostics). Deparaffinization was performed for four min at 72°C using EZPrep (Roche Ventana, #5279771001), followed by antigen-retrieval (AR) for 40 minutes at 100°C with TRIS-based CC1-buffer (Roche Ventana, #5424569001). After peroxidase-inhibition with I-View Inhibitor (Roche, #06396500001), the sections were incubated with primary

antibodies against insulin (1:1000; Dako, #A0564), glucagon (1:600; Santa Cruz, #sc13091), somatostatin (1:6000; Invitrogen, #14-9751-80), CD31 (1:100; Dako, #M0823), perilipin 1 (1:2000; Progen, #651156), and tubulin beta 3 (1:2500; R&D Systems, #MAB1195). The secondary horseradish peroxidase-linked antibody was detected via an Opti-View DAB IHC detection kit (Roche Ventana, #06396500001). The samples were counterstained with hematoxylin. WSI-acquisition was performed with a Hamamatsu NanoZoomer 2.0-HT using 20x magnification and NDP.scan 2.5 software.

Multiplex Fluorescence staining and fluorescence microscopy. Consecutive 2-4 µm FFPE sections derived from the identical pancreatic specimens as used for the chromogenic stainings underwent fluorescent labeling via an automated staining system DiscoveryUltra (Ventana Technology, Roche Diagnostics). Deparaffinization utilized EZPrep (Roche Ventana; #5279771001) for 32 min at 69°C, and subsequent AR was performed with TRIS-based CC1-buffer (Roche Ventana, #5424569001) at 91°C for 48 min. Primary antibody cocktails were applied after incubation with a human FC-receptor-blocking reagent (1:50; Miltenyi, #130-059-901).

The antibody cocktails for Stainingset 1 included for the first incubation antibodies against glucagon (1:200; Abcam plc. #Ab10988) and tubulin beta 3 (1:25; R&D Systems, MAB1195), for the second incubation mouse IgG1 con. AF555 (1:200; Invitrogen, #A-21127) and mouse IgG2a con. AF647 (1:50; Invitrogen, #A21241), and for the third incubation somatostatin con. AF750 (1:200; Novus Biologicals, #NBP2-99309 AF750) with DAPI (1 ng/ml; Invitrogen, #D1306). The cocktails for Stainingset 2 included for the first incubation antibodies against perilipin 1/PLIN1 (1:50; Progen, #690156) and PECAM1/CD31 (1:33; Abcam, #ab134168), for the second incubation mouse IgG1 con. AF555 (1:100; Invitrogen, #A21127) and rabbit IgG con. AF750 (1:50; Invitrogen, #A21039), and for the third incubation insulin con. AF488 (1:200; Invitrogen, #53-9769-82) with DAPI (1 ng/ml; Invitrogen, #D1306). WSI acquisition was performed with a slide scanner (Zeiss AxioScan.Z1 equipped with ZEN 3.1 software) at 20x magnification.

B Deep Learning Models

Model Architectures. A Multiple Instance Learning Approach (MIL) [17] is commonly used to tackle WSI classification. In MIL the individual patches are first encoded by a pre-trained feature extractor and subsequently pooled and fed together to a MIL classification algorithm that is trained to predict the image-level class label. We tested two different feature extractors on the chromogenic WSIs. The first is a Vision Transformer pre-trained on ImageNet21k [7], which consists of natural images, while the second is Phikon [8], a Vision Transformer pre-trained on over 6 thousand histologic WSIs and therefore specifically tailored to the domain. As MIL algorithm, we used Clustering-constrained Attention Multiple Instance Learning (CLAM) [6] as well as Chowder [18]. Both architectures deal with the variability in patch importance in different ways. CLAM uses an attention mechanism to focus on the most relevant patches within a bag while simultaneously imposing clustering constraints to ensure that similar patches receive comparable attention weights, enhancing interpretability and performance. We adapted the CLAM algorithm to further improve generalizability by only showing a random subset of patches of a WSI (usually 5%) to the model during an epoch. Moreover, a cosine annealing learning rate scheduler as well as gradient accumulation were added. The Chowder architecture employs a modular and hierarchical approach to MIL, where instance-level features are aggregated through a series of convolutional and pooling layers to capture complex patterns and dependencies. It also incorporates importance sampling to dynamically select the most informative instances from each bag, thereby reducing computational load and enhancing the model's ability to focus on critical data points.

Data Representations. Patches of chromogenic samples were encoded as RGB images (3 color channels) similar to natural images. In contrast, multiplex fluorescent immunostainings required further preprocessing, since in these cases each channel contains the intensity values of one of the respective stainings which does not coincide with color channels. Each of the fluorescent WSIs has 4 channels, containing 3 of the Stainingset-specific stainings as well as DAPI, as nuclear (DNA) marker. We tested 3 different data representations on the fluorescent WSIs:

- **RGB:** This representation treats each of the 3 stainings as RGB color channels and additionally overlays the DAPI channel as a grayscale image on each of the other 3 channels, resulting in an RGB image similar to what can be seen with a standard visualization software when showing all

stainings simultaneously. This representation contains the co-occurrence of the stainings, i.e. the model can learn which stainings occur together in the WSI.

- **Channel-wise:** The channel-wise representation treats each channel (staining) individually as a grayscale image. Consequently, each patch gets encoded 4 times during feature extraction. The resulting features (1-dimensional vectors) are then appended, leading to a representation where each channel is encoded in detail, but no co-occurrence information is preserved.
- **Channel-wise average:** The channel-wise average representation is very similar to the channel-wise representation but aims to include the co-occurrence of stainings by averaging the 4 different feature vectors respective to their occurrence in the image.

Training Procedure. We split the data consisting of 100 unique patients into a train (75 patients) and test (25 patients) set and performed the model development solely on the train data. For that, we used a 15-fold cross-validation, where in each fold we randomly used 60 patients from the train set for training and the remaining 15 patients for validation. Due to the different distributions of T2D within the two cohorts, we applied balanced sampling during training only for cohort 1 patients while the cohort 2 was already balanced. As the target metric, we chose the Area Under the Receiver Operating Characteristic Curve (AUROC), since for the subsequent XAI steps it is more important to have a model that has a high separability and can generally distinguish between the two classes (T2D or not) than having a model that works well with a specific cutoff as measured by the F1-Score or the Accuracy.

Representation	Stainingset	Encoder / MIL Algorithm	AUROC (higher is better)
Channel-wise Average	1	Imagenet / CLAM	**0.956**
	2	Imagenet / CLAM	0.684
Channel-wise	1	Imagenet / CLAM	**0.912**
	2	Imagenet / CLAM	0.816
RGB	1	Imagenet / CLAM	**0.842**
	2	Imagenet / CLAM	0.640

Table 2: Model performance on fluorescent WSIs.

C Attention Computation

Initially, we focused on determining the significance of specific patches using the built-in attention mechanism of the classification head. Nevertheless, raw attention data predominantly indicates general importance rather than class-specific relevance. To address the latter, we applied Attention Layer-wise Relevance Propagation (Attention LRP) [19], which filters the attention values to highlight patches distinctly associated with either diabetic or non-diabetic status within the same WSI. The class-specific relevance R is iteratively computed for each layer n. Here $x^{(n)}$ is the input at layer n with j elements and $x^{(n-1)}$ is the input of the downstream layer with i elements. w_{ji} is the respective weight matrix of the layer. Non-linear functions like GELU [20] produce both positive and negative outputs. To account for this, relevance propagation can be adjusted by forming a subset of indices q resulting in the following relevance propagation:

$$R_j^{(n)} = \sum_{\{i|(i,j)\in q\}} \frac{x_j w_{ji}}{\Sigma_{\{j'|(j',i)\in q\}} x_{j'} w_{j'i}} R_i^{(n-1)} \quad \text{with} \quad q = \{(i,j)|x_j w_{ji} \geq 0\} \tag{1}$$

We set $R^{(0)} = 1_t$ for initialization, where t is the one-hot-encoded outcome class. Based on the relevance, we compute the average relevance filtered attention $\bar{A}$ for each attention head b, with the number of heads h. To this end, we take the Hadamard product between the relevance and the attention's heads gradient $\Delta A^{(b)}$, considering only the positive relevance:

Staining	MIL Algorithm	Encoder	AUROC (higher is better)
Mean over Stainings	CLAM	Imagenet	**0.833**
	Chowder	Imagenet	0.830
	CLAM	Phikon	0.794
	Chowder	Phikon	0.781
Tubulin beta 3	CLAM	Imagenet	**0.895**
Insulin	CLAM	Imagenet	0.842
Glucagon	CLAM	Imagenet	0.842
Perilipin 1	CLAM	Imagenet	0.842
Somatostatin	CLAM	Imagenet	0.842
PECAM1	CLAM	Imagenet	0.737
Tubulin beta 3	CLAM	Phikon	0.851
Insulin	CLAM	Phikon	0.684
Glucagon	CLAM	Phikon	0.711
Perilipin 1	CLAM	Phikon	0.816
Somatostatin	CLAM	Phikon	**0.921**
PECAM1	CLAM	Phikon	0.781
Tubulin beta 3	Chowder	Imagenet	**0.877**
Insulin	Chowder	Imagenet	0.851
Glucagon	Chowder	Imagenet	0.640
Perilipin 1	Chowder	Imagenet	**0.877**
Somatostatin	Chowder	Imagenet	**0.877**
PECAM1	Chowder	Imagenet	0.851
Tubulin beta 3	Chowder	Phikon	0.833
Insulin	Chowder	Phikon	0.693
Glucagon	Chowder	Phikon	0.728
Perilipin 1	Chowder	Phikon	**0.860**
Somatostatin	Chowder	Phikon	0.842
PECAM1	Chowder	Phikon	0.728

Table 3: Model performance on chromogenic WSIs.

$$\bar{A}^{(b)} = I + \mathbb{E}_h(\nabla A^{(b)} \odot R^{(n_b)})^+ \tag{2}$$

D Imaging Biomarker

ROI Segmentation. To compute the identified biomarkers a segmentation map of the respective regions is needed. Due to the amount and size of the WSIs we used scribble annotation [21] to efficiently label 44 (out of 600 chromogenic WSIs) slides with the labels background, tissue, adipocytes, fibrotic patterns, and islets. Scribble annotation allows for a very fast annotation because regions are not annotated densely (i.e. each pixel gets annotated) but rather only smaller lines (scribbles) are used to approximately label the respective regions. During model training, we used the ignore label in nnU-Net [22] to only update the model weights based on annotated scribbles

while ignoring the remaining image. While annotation is based on scribbles, the model still returns a dense prediction which makes this approach well-suited for segmenting the ROIs. Since the nature of scribbles only allows a quantitative evaluation of the respective annotated areas, a qualitative analysis of the resulting segmentation maps was conducted. After adjusting some annotations and retraining the nnU-Net, a satisfactory segmentation quality was reached.

Quantifying Attention. The attention to specific ROIs for the chromogenic stainings was computed by summing up the raw attention within specific segmentation masks and standardizing it by the total attention in the WSI and the tissue size. We summed up the attention per staining channel for the fluorescence stainings and standardized it by the total attention over all stainings and the tissue size. For the exploratory analysis, we z-standardized all continuous variables on non-interpretable scales; however, we routinely z-standardized all continuous covariates for the statistical analysis.

Biomarker. For the chromogenic staining, we computed the area-related biomarkers based on the segmentation masks and counted the number of distinct islets and adipocyte clusters by applying connected components to the respective segmentation mask. All biomarkers are standardized by the total size of the tissue. For the islet-related biomarker, we only include islet segmentations larger than 5000 pixels, ignoring segmentation artifacts or errors. For the fluorescence-specific intensity biomarkers, we computed the channel-specific mean intensity as an approximation of the respective area of the trait and standardized it again by the tissue size. In the case of perilipin 1, we first threshold the intensity values, including only values larger than 1000 (intensity measured as 16-bit integers), filtering out unspecific stained traits. For tubulin beta 3 within islets, we only included the intensity within islets and additionally, standardized by the number of islets.

E Statistical Methods

Hypothesis Testing. To test differences between two groups, e.g. patients with and without diabetes, we applied a two-sample t-test. For testing, if at least one group is significantly different, we applied the F-test within a one-way ANOVA. To test if a coefficient in a linear model is significantly different from zero we used the Z-test. For all tests, we considered p-value > 0.05 as not significant.

Statistical Models. For statistical analysis of the biomarkers, we used a generalized mixed linear model (GMLM), as we have several dependence structures within the data. For the diabetes status as the outcome, we selected the canonical binomial link function. For the HOMA2B level, we selected the canonical Gaussian and Gamma link functions and reported results for the Gaussian link, as its Akaike Information Criterion (AIC) was lower. We accounted for random effects linked to the cohort and staining method. Due to no variation between the outcomes (i.e., diabetes status) within a patient, we could not model the patient as a random effect. In this case, the intercept would be enough to fit one model per patient perfectly. All models were fitted via maximum likelihood and the L-BFGS-B optimizer [23]. We tested for multicollinearity by computing variance inflation factor (VIF) and removed "benign tumor" due to perfect multicollinearity with "malignant tumor" combined with "chronic pancreatitis". Sensitivity analyses were performed using generalized estimating equation (GEE) models.

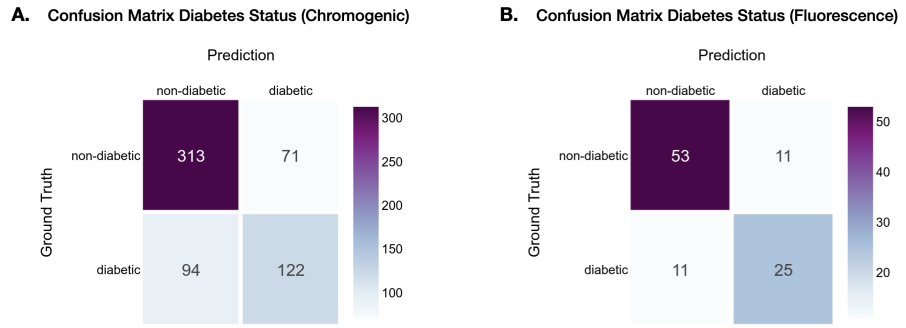

Figure 6: Top 15 attended chromogenic patches.

Statistical Model Evaluation. For all models with the same modality and response variable, we used the AIC for model selection. Additionally, we evaluated the logistic models through a confusion matrix (Figure 6) and the linear models through parity plots, residual plots, and Q–Q plots of the residuals (Figure 7).

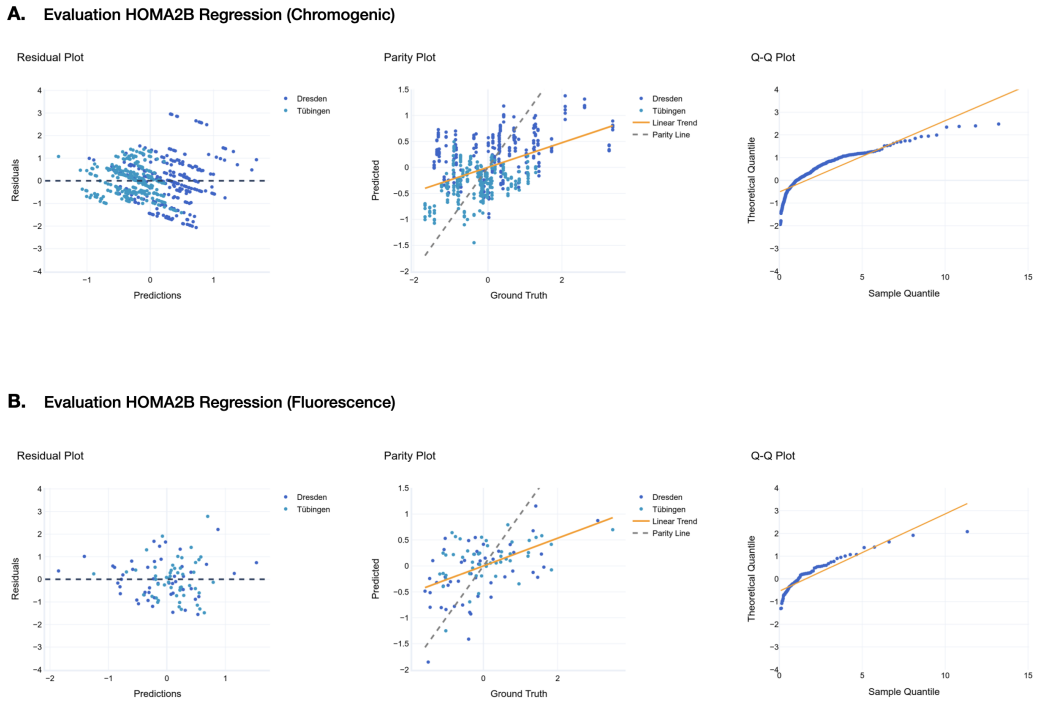

Figure 7: Top 15 attended chromogenic patches.

F Limitations

While this data-driven approach has the potential to accelerate and facilitate the understanding of T2D pathogenesis, several potential limitations must be considered. One is that immunostaining approaches can be associated with technical pitfalls such as antibody binding specificity, nonspecific background staining, or variations in the staining intensity across samples regardless of target antigen level. Furthermore, our histological preparations were not representative of the whole organ, since

they were obtained from limited regions of the pancreas, the anatomical location of which varied among living donors. Further, despite originating from the “healthy” margins of the surgical resection, as also verified by postoperative pathological assessment, the samples could be affected by changes in the tissue microenvironment such as inflammation, enrichment of connective tissuefocal fibrosis, and angiopathy.

G Top 15 attended chromogenic patches

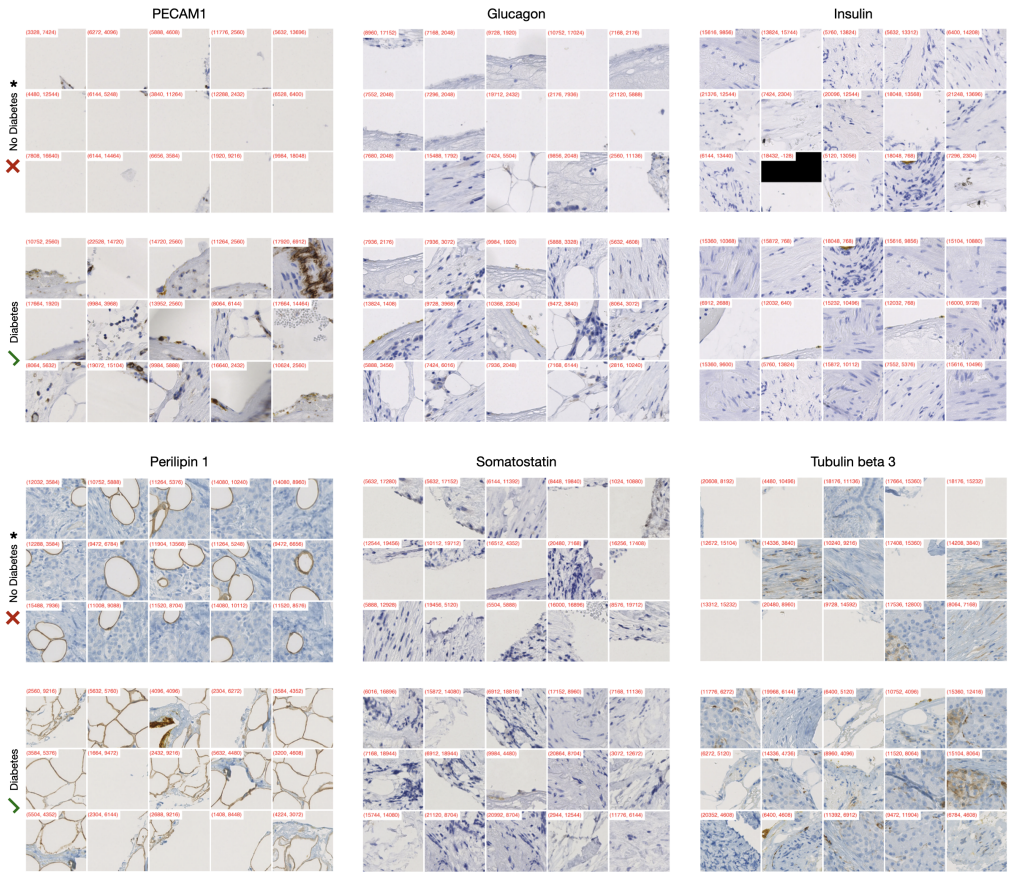

Figure 8: Top 15 attended chromogenic patches.

H Pixel-level saliency maps

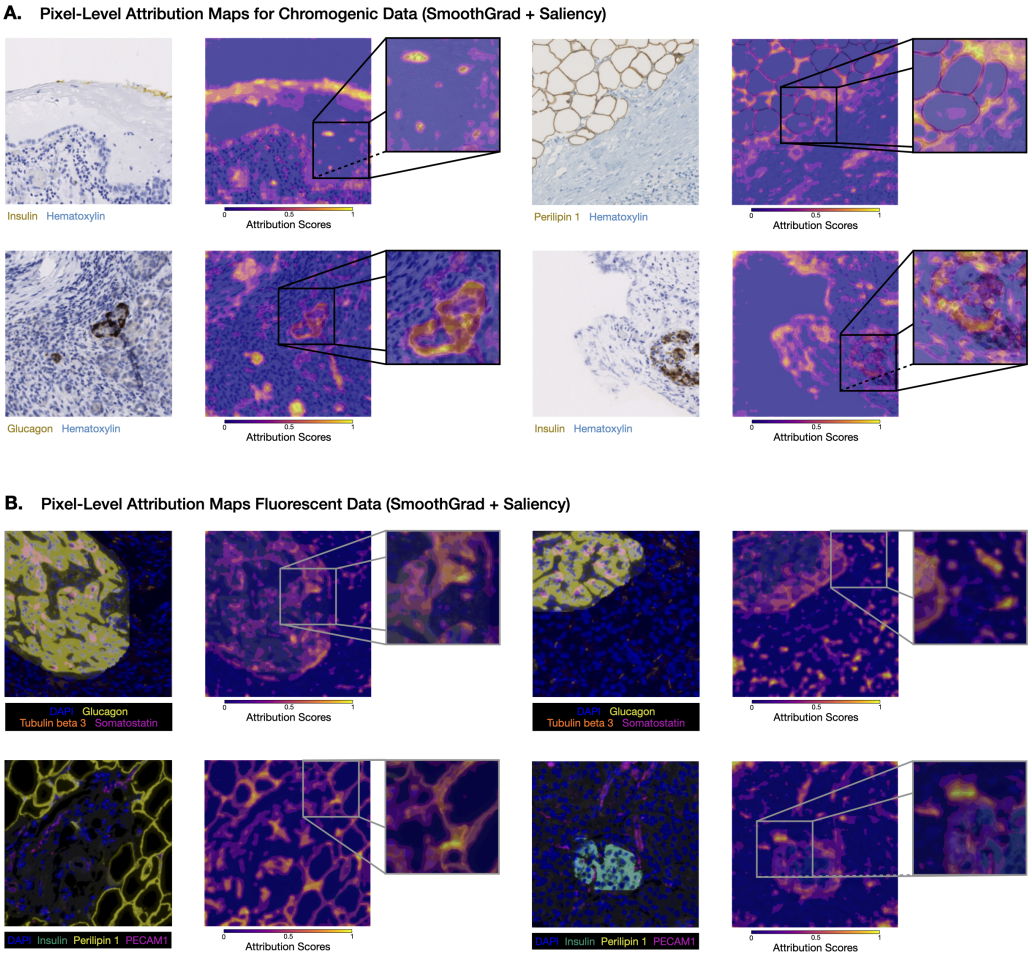

Figure 9: Pixel-level heatmaps of local regions and one single patch based on Saliency with SmoothGrad for chromogenic **(A)** and fluorescence **(B)** data.

I Regression Results Fluorescence

A. GMLM Regression for Diabetes Status (Fluorescence)

Fixed Effects	Estimate	SE	P-Value (*p*)	Sign.
Intercept	-1.889	0.877	0.031	*
Area of Islets	-0.446	0.426	0.295	
Number of Islets	0.085	0.288	0.768	
Perilipin 1 Intensity	0.408	0.265	0.124	
Tubulin beta 3 Intensity in Islets	-0.328	0.310	0.168	
Sex	0.709	0.554	0.201	
Age	0.832	0.330	0.012	*
BMI	0.593	0.263	0.024	*
Malignant Tumor	0.672	0.616	0.276	
Chronic Pancreatitis	1.915	0.967	0.048	*

Random Effects	Type	Variance	STD
Cohort	Intercept	0.754	0.868
Observations	100		
Groups	{Cohort: 2}		
Log-likelihood	-51.628		
AIC	125.256		

Note * *p*<0.05, ** *p*<0.01, *** *p*<0.001

B. MLM Regression for HOMA2B (Fluorescence)

Fixed Effects	Estimate	SE	P-Value (*p*)	Sign.
Intercept	0.306	0.173	0.080	
Area of Islets	-0.207	0.094	0.030	*
Number of Islets	0.280	0.103	0.008	**
Perilipin 1 Intensity	0.032	0.085	0.706	
Tubulin beta 3 Intensity in Islets	0.183	0.093	0.053	
Sex	-0.149	0.170	0.383	
Age	-0.154	0.088	0.085	
BMI	0.243	0.090	0.008	**
Malignant Tumor	0.076	0.187	0.685	
Chronic Pancreatitis	0.314	0.307	0.309	
Insulin Therapy	-1.435	0.244	0.000	***

Random Effects	Type	Variance	STD
Cohort	Intercept	0.000	0.000
Observations	91		
Groups	{Cohort: 2}		
Log-likelihood	-105.898		
AIC	237.797		

Note * *p*<0.05, ** *p*<0.01, *** *p*<0.001

Table 4: Regression results of the generalized mixed linear model analysis for diabetes status **(A)** and HOMA2B **(B)** level based on fluorescence.