Comparing tangible retinal image characteristics with deep learning features reveals their complementarity for gene association and disease prediction

Michael J Beyeler[1,2], Olga Trofimova[1,2], Dennis Bontempi[1,2], Leah Böttger[1,2], Sacha Bors[1,2], Ilaria Iuliani[1,2], Sofia Ortin Vela[1,2], David M Presby[1,2,&], Sven Bergmann[1,2,3,&]

[1]Dept. of Computational Biology, University of Lausanne, Lausanne, Switzerland;
[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland;
[3]Dept. of Integrative Biomedical Sciences, University of Cape Town, Cape Town, South Africa;
[&]joint corresponding (david.presby@unil.ch, sven.bergmann@unil.ch) and equal senior authors

Abstract

Advances in computer-aided analyses, including deep learning (DL), are transforming medical imaging by enabling automated disease risk predictions and aiding clinical interpretation. However, DL's outputs and latent variables (LVs) often lack interpretability, impeding clinical trust and biological insight. In this study, we evaluated *RETFound*, a foundation model for retinal images, using a dataset annotated with clinically interpretable tangible image features (TIFs). Our findings revealed that individual LVs poorly represent complex TIFs but achieve higher accuracy when combined linearly. Fine-tuning *RETFound* to predict TIFs, providing “deep TIFs” provided better, but far from perfect surrogates, highlighting the limitations of DL approaches to fully characterise retinal images. Yet, our genetic analyses showed that deep TIFs exhibit heritability comparable to or exceeding measured TIFs but highlighted non-genetic variability in LVs. While measured and deep TIFs, as well as LVs, showed overlapping genetic and disease associations, their complementarity enhances prediction models. Notably, deep TIFs excelled in ocular disease prediction, emphasising their potential to refine retinal diagnostics and bridge gaps in conventional assessments of vascular morphology.

Introduction

Medical imaging has become essential for diagnosing diseases and providing presymptomatic risk prognoses. The use of computer-aided analysis, particularly through deep learning (DL), has made these processes not only more accurate but also scalable to large datasets. Recently, foundation models [1] have emerged as a powerful DL approach generating initial models pre-trained on extensive image datasets using self-supervised learning. Such models can then be applied across a wide range of use cases as they provide a robust and efficient latent space that captures image variability in a low-dimensional form. The basic idea is to use the “encoder” of the model that transforms an image into latent variables (LVs), which are then used as features for disease identification or prediction using a relatively simple model architecture. In some cases, the training procedure also involves “fine-tuning” the weights of the encoder, effectively co-adapting the latent space to provide the most pertinent features for the end-point prediction.

While there are ample examples showing the power of this approach [1–3], it also has several disadvantages: First, the LVs are typically difficult to interpret. Although there exists techniques to elucidate which regions of an image most strongly contribute to variation in a given LV, it is generally not obvious what exact properties of the image in these regions are captured. This is in contrast to more classical approaches where image processing is used to measure well defined properties in medical images, such as the size or shape of previously identified anatomic objects, for which there is typically well-established medical evidence for being relevant for certain diseases. Second, the number of LVs usually is still relatively large, typically between 2^7 and 2^{10}, and some of the LVs may be strongly correlated to each other, further complicating interpretability. Finally, there is no guarantee that LVs that efficiently capture the information needed to reconstruct medical images also include potentially very subtle signals that are most relevant for certain diseases.

In this work we compare LVs with “classical” tangible image features (TIFs) within a specific domain, namely retinal Color Fundus Images (CFIs). These images allow for inexpensive and noninvasive in-vivo assessment of the superficial inner layer of the retina, enabling the monitoring of ocular diseases such as diabetic retinopathy, macular degeneration, and glaucoma, as well as early risk prediction for systemic diseases, including cardio- and cerebrovascular disease, chronic kidney disease and diabetes [4]. As a result, CFIs have been acquired in several large cohorts and there is a rapidly expanding body of research on automatic extraction of LVs and TIFs from these images [5–11] and their associations with medically relevant information. Specifically, Zhou et al. recently published *RETFound* [2], a foundation model for retinal images that was trained on 1.6 million unlabelled retinal images using self-supervised learning with a transformer [12] architecture. Fine-tuning this model with labelled images the authors

demonstrated state-of-the-art performance in the diagnosis and prognosis of several eye diseases, as well as incident prediction of complex systemic disorders such as heart failure and myocardial infarction. Furthermore, we recently published a complementary study of 17 different morphological vascular phenotypes extracted from over 130k fundus images of close to 72k UK Biobank subjects. We found that a substantial portion of the variability in these TIFs—features of longstanding interest to ophthalmologists, including median vessel diameter, diameter variability, main temporal angles, vascular density, central retinal equivalents, the number of bifurcations, and tortuosity—can be attributed to genetic factors, with SNP-based heritability estimates ranging from 5% to 25%. Moreover, we observed a large number of genetic association signals with genes and pathways that appear plausible to be involved in modulating the TIFs, and many significant correlations with both ocular and systemic diseases or their risk factors [11].

Building on these two studies allows us to shed light on the relationship between LVs and TIFs in the concrete, data-rich setting of CFIs by addressing the following questions: To what extent do LVs capture similar or complementary information to that of TIFs? Do LVs also have sizable heritabilities? Do their genetic associations point to similar or different sets of genes and pathways? Finally, for a given disease, whether ocular or systemic, do LVs or TIFs have stronger predictive power, and is there added value in combining them? To address these questions, we first leverage the *RETFound* model using its pretrained weights to generate LVs for more than 40k CFIs which have been characterised in terms of TIFs. We complement these features by fine-tuning *RETFound* to predict TIFs, generating surrogate “deep TIFs”, and explore these three types of features in terms of their correlations with each other, genetics and disease (see Fig. 1).

Results

Phenotypic data generation in terms of TIFs and LVs

Using the methodology described in our previous work [10] we generated TIF profiles for 41 527 genotyped participants from the UK Biobank. Here we only included CFIs from the right eye for which we were able to measure all of our 17 TIFs. 96% of these images passed the quality control threshold we had used in our previous work (see Suppl. Fig. 1). This procedure effectively removed CFIs for which our algorithm failed to identify the optic disk or enough major blood vessels to measure the temporal angles, which happens more often for CFIs from older and diseased subjects. As a result the baseline characteristics of our cohort were different to those of our original work (see Suppl. Table 1 for details). Each CFI was then characterised in terms of 1024 LVs using *RETFound* prior to any fine-tuning using the weights of the network provided by the authors [2] (see Methods for details).

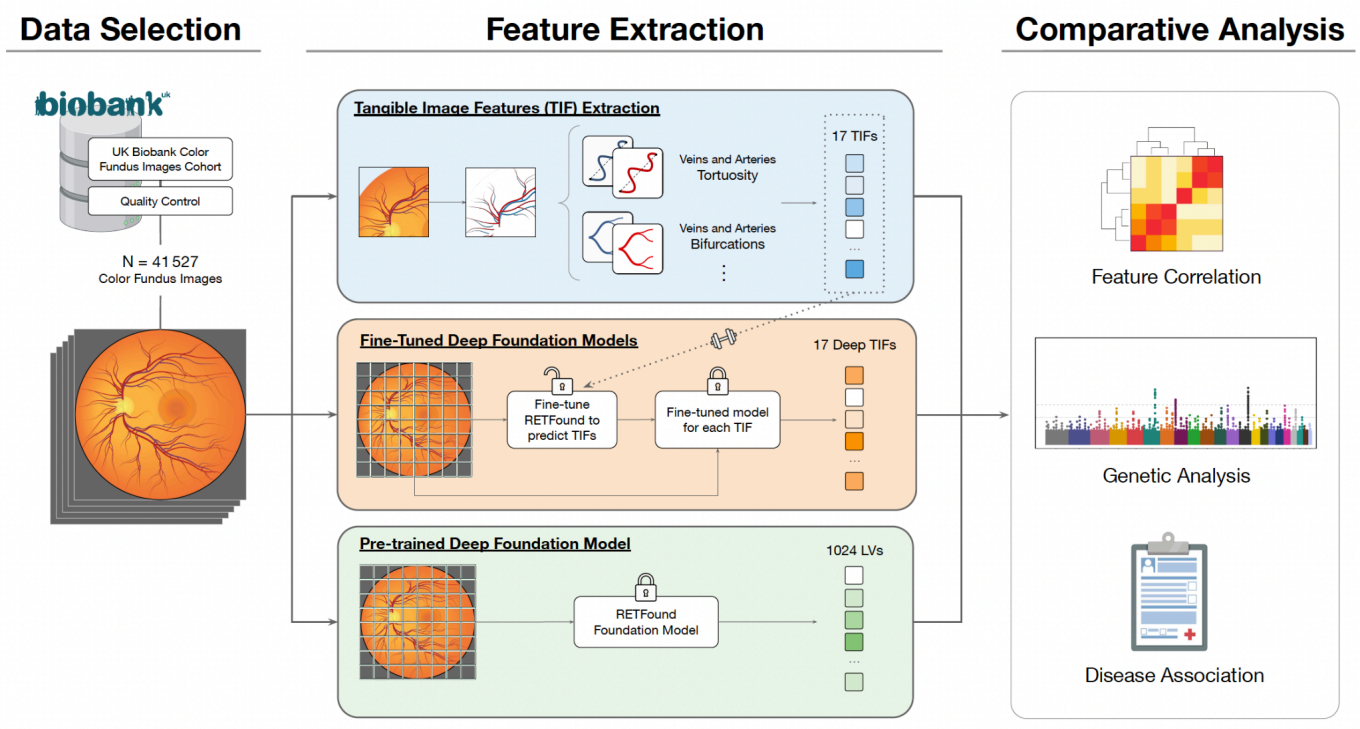

Figure 1 | Graphical overview of the study. (a) We selected a subset of genotyped subjects from the UK Biobank who had color fundus images (CFIs) taken, as well as available health outcome information and risk factors. (b) Top panel: Using a previously developed analysis pipeline to characterise the retinal vasculature [11], we extracted tangible image features (TIFs), such as vessel tortuosity or the number of bifurcations, from each CFI. We also used the retinal foundation model *RETFound* [2] as a DL approach to characterise CFIs both (middle) in terms of "deep TIFs" by fine-tuning *RETFound* to predict our 17 measured TIFs and (bottom) by extracting a set of 1024 latent variables (LVs) of the pre-trained (but not yet fine-tuned) *RETFound*. (c) For downstream comparative analyses, we employed direct correlations, genetic, and disease associations.

Modelling TIFs by LVs

By design, *RETFound* LVs capture much of the information contained in CFIs. This information is not limited to the retinal vasculature, but includes it. Thus, in principle, these LVs should also be informative of our vascular TIFs. In order to evaluate to what extent any individual TIF can be represented by a single LV we computed all pairwise correlations (shown as a bi-clustered cross-correlation matrix in Fig. 2a). We then picked for each TIF the LV with the largest squared correlation, which corresponds to the explained variance R^2_{SLR} of a Simple Linear Regression (SLR) model (top bars in Fig. 2b). Our analysis indicates that *individual* LVs can represent a sizable portion of the variability of relatively simple vascular measures, like the vascular densities $(max\, R^2_{SLR} = 0.36)$. Yet, this is not the case for more complex vascular TIFs, such as arterial central retinal equivalents $(max\, R^2_{SLR} = 0.09)$, tortuosity $(max\, R^2_{SLR} = 0.01)$, and even less for their artery/vein (A/V) ratios.

Next we asked to what extent combining multiple LVs to predict TIFs can increase the explained variance. As a first step we trained Multiple Linear Regression (MLR) models with all LVs as features

using five-fold cross-validation. The corresponding out-of-sample estimates for the explained variance R^2_{MLR} for each TIF (middle bars in Fig. 2b) are consistently higher than those from the best SLR model. They follow a similar trend with more than 90% explained variance for the vascular densities, but less than 25% for some of the complex TIFs, notably the A/V ratios for tortuosity $(R^2_{MLR} = 0.09)$ and central retinal equivalents $(R^2_{MLR} = 0.06)$. In a second step we directly trained *RETFound* to predict each of our 17 TIFs. This resulted in 17 "fine-tuned" *RETFound* models providing TIF estimates as their output that we coin "deep TIFs". For most of these deep TIFs the out-of-sample estimates R^2_{deep} for the explained variances (bottom bars in Fig. 2b) are larger than those of the corresponding MLR models. Specifically, the fine-tuned *RETFound* models outperformed MLR models for 11 out of 17 TIFs, with a mean improvement in R^2 of 0.05 (95% CI = [0.01, 0.08], t(16) = 2.45, p = 0.03)). We observed the most substantial increases for the temporal angles (arterial: $R^2_{deep} = 0.34$, venous: $R^2_{deep} = 0.61$) and arterial tortuosity $(R^2_{deep} = 0.43)$, while the improvements for the other TIFs are marginal or even absent.

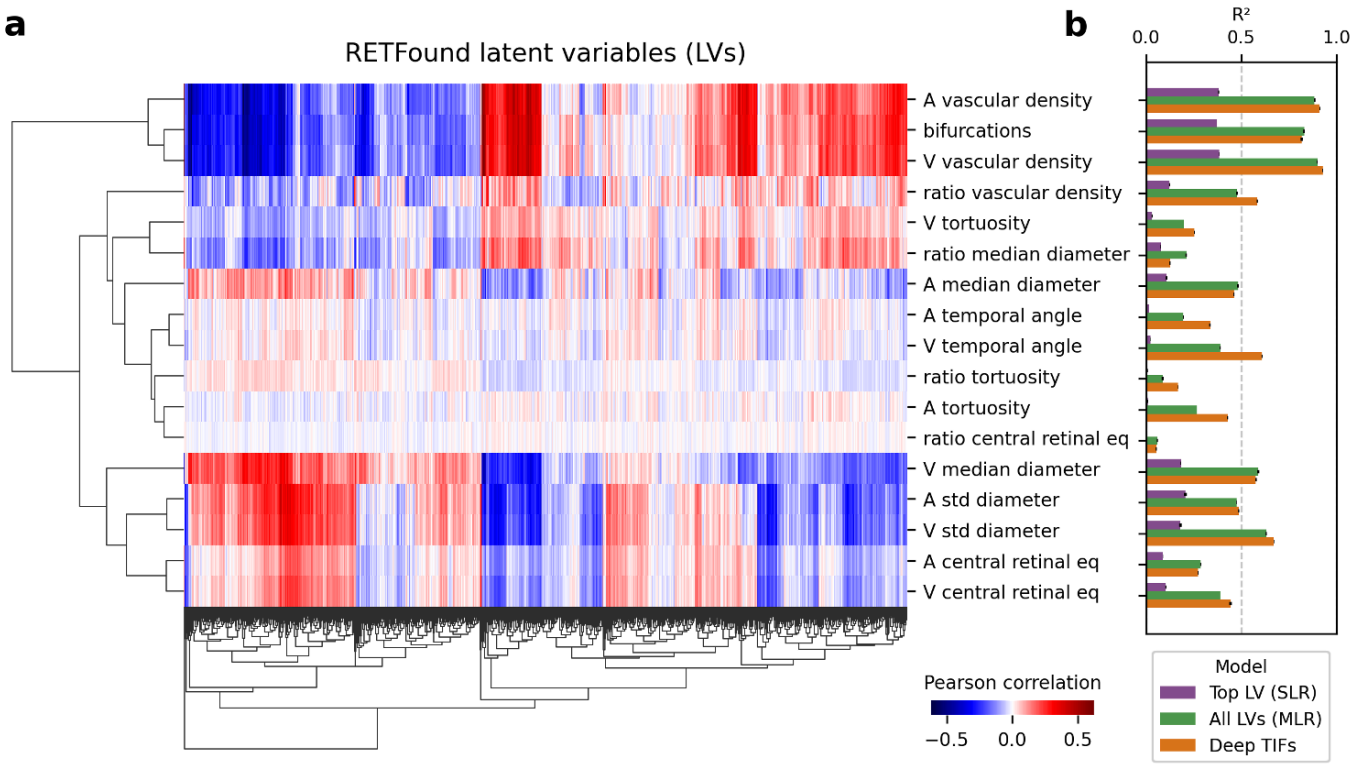

Figure 2 | Correlations analysis. (a) Bi-clustered cross-correlation matrix between LVs (x-axis) and tangible image features (TIFs, y-axis). (b) Variance of TIFs explained by the best simple linear regression (SLR) models using the single best baseline LV (purple), multiple linear regression (MLR) models using all baseline LVs (green) and fine-tuning *RETFound* models (orange). Out-of-sample R^2 and their standard errors across the five folds are displayed. A: arterial, V: venous, eq: equivalent, std: standard deviation.

Comparing the genetic associations signals of TIFs and LVs

We previously showed that many of our TIFs have a sizable genetic component, with heritability estimates ranging from 5% for median vessel diameter to >25% for tortuosity [10]. Also, our genome-wide association study (GWAS) revealed many relevant genes and pathways for the vast majority of TIFs. These genetic association signals indicate that our TIFs capture physiologically relevant traits that are partially modulated by genetic variants. Given that LVs capture some but not all of the information represented by TIFs we sought to elucidate their genetic association signals. To this end we performed GWAS, heritability estimates, and gene analysis for all 1024 LVs of the *RETFound* model (before any fine-tuning) and our 17 deep TIFs.

At the level of the SNP-wise association signals (Fig. 3a-c), we observed that overall there are more SNPs significantly associated with TIFs ($n = 5949$ with $p < 5/17 \cdot 10^{-8}$) than LVs ($n = 4731$ with $p < 5/1024 \cdot 10^{-8}$). Since LVs are not independent we also performed a GWAS on their 17 leading principal components (PCs), explaining 65% of their variance, which resulted in $n = 2130$ SNPs with $p < 5/17 \cdot 10^{-8}$. Overall we observed that of all the significantly associated loci only 7.4% were linked to both TIFs and LVs, while 52.4% were specific to TIFs and 40.2% to LVs. Yet, by far the most significant associations were observed for a subset of LVs contributing strongly to PCs 3 and 4 with an effect size of $\beta=0.45$ and nominal significance of $p<10^{-300}$. This association pertains to the *HERC2/OCA2* locus, which also modulates some of our TIFs, albeit not as strongly. In contrast, the GWAS signals for deep TIFs were relatively similar, but not identical, to those of the measured TIFs, with a few additional, as well as some missing hits. Aggregating SNP-wise association signals at the level of genes using our *PascalX* tool [13,14] we observed a similar scenario (Fig. 3d-f): The overlap between the sets of significantly associated genes is much larger between measured and deep TIFs than either of them to the LVs.

The distributions of heritability estimates (using linkage disequilibrium (LD) score regression) indicate that the vast majority of LVs have a small genetic contribution ($h^2 < 0.1$ for 77.6%, see Fig. 3g). Yet, it is sizable for a few LVs reaching values up to 21.4%, which is similar to that of the TIFs with the highest heritability estimates. In contrast, the corresponding heritability estimates for deep TIFs are either close to that of the corresponding measured TIF, or, in some cases, higher (Fig. 4a). The latter is most pronounced for the temporal angles and central retinal equivalents. Strikingly, a similar pattern is observed at the level of significantly associated genes (Fig. 3h): with the exception of the ratio of arterial and venous vascular density and venous tortuosity, deep TIFs have more exclusively associated genes than their corresponding measured TIFs (Fig. 4b). Lastly, genes associated with at least one trait show different patterns of pleiotropy between LVs and TIFs: for LVs, these genes are often involved in multiple traits, whereas for TIFs, they tend to be limited to one or two of the 17 traits (Fig. 3i).

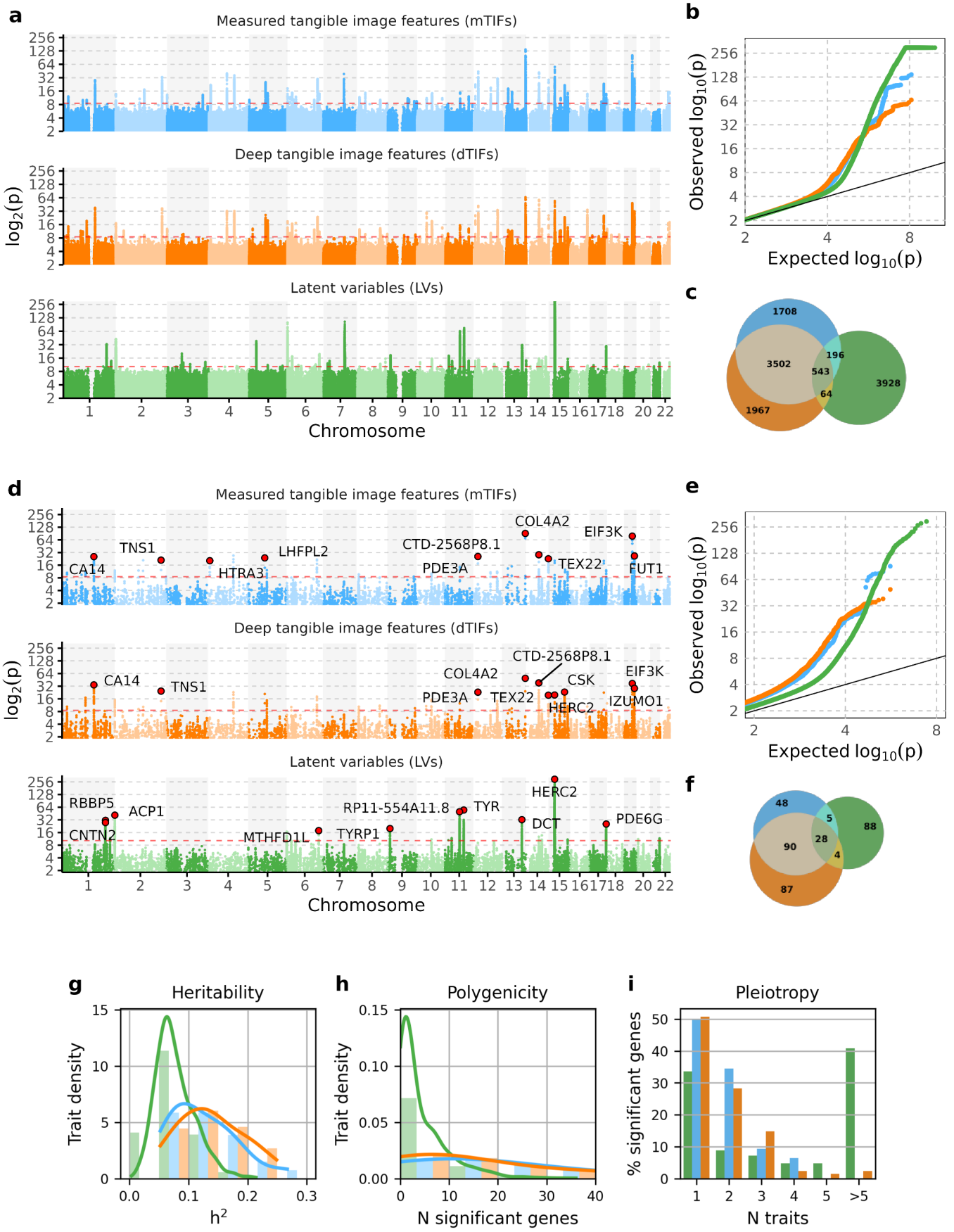

Figure 3 | Genetics associations of measured TIFs, deep TIFs, and LVs. (a) Manhattan plots, (b) QQ-plots and (c) Venn-diagrams for SNP-wise associations. (d) Manhattan plots, (e) QQ-plots and (f) Venn diagrams for gene-wise associations. (g) Density (i.e. normalised count) of TIFs, LVs and deep TIFs within heritability (h^2) range. (h) Density for number of significantly associated genes with TIFs, LVs and deep TIFs. (i) Percent of significant genes (see Fig. 3f) associated with a certain number of traits.

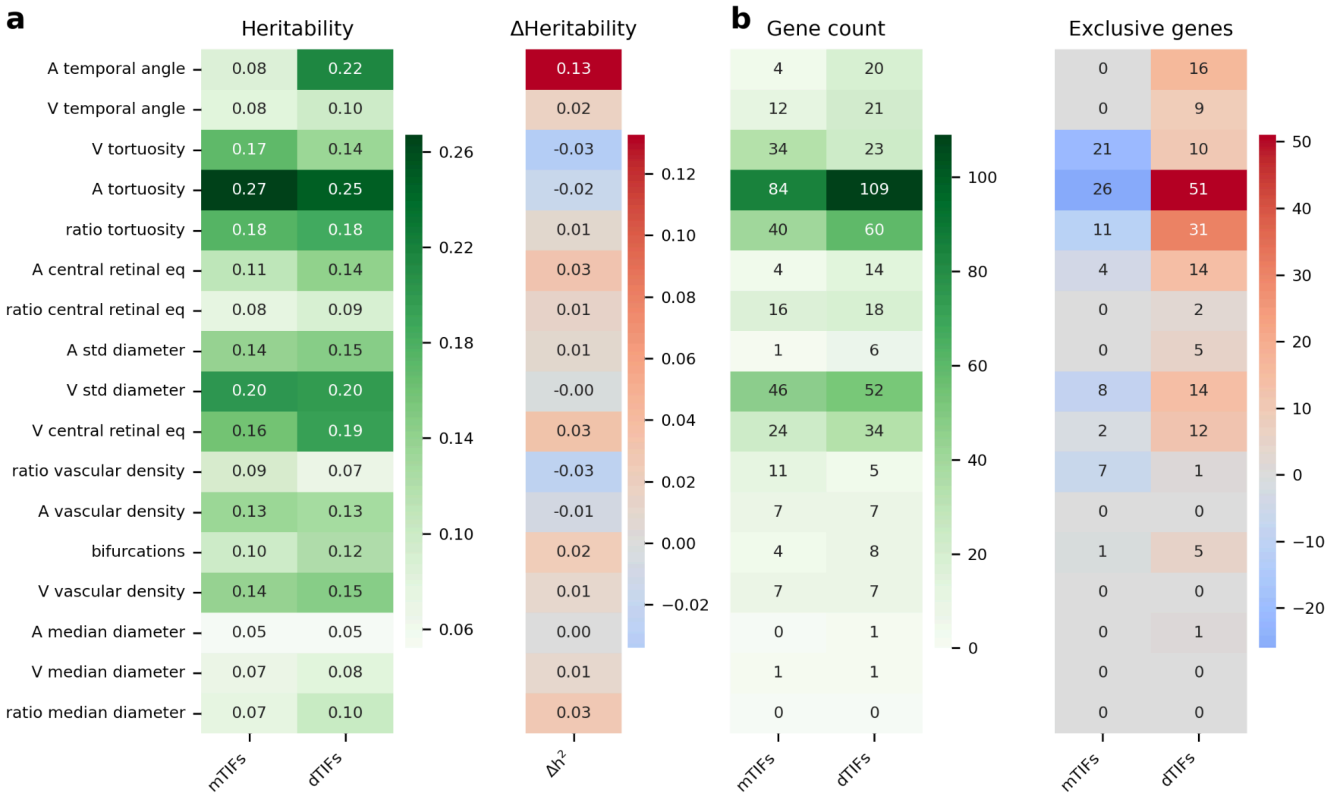

Figure 4 | Heritabilities and gene associations per TIF. (a) SNP heritability estimates obtained via LD score regression and the difference between deep and measured equivalents. (b) Number of associated genes of measured TIFs versus deep TIFs, including the count of genes exclusive to each. Abbreviations: A = Arteries; V = Venous; std = Standard Deviation; mTIFs = Measured Tangible Image Features; dTIFs = Deep Tangible Image Features.

Associating disease risk with measured and deep TIFs

We next asked to what extent ocular and general diseases and their risk factors can be explained in terms of measured and deep TIFs. To this end we used linear models that incorporated covariates alongside a single TIF as explanatory variables. We assessed the strength of associations using standardised effect sizes and statistical significance. Finally, we quantified the total number of unique significant associations across diseases and TIFs to summarise the findings. Although not statistically significant, the largest difference in number of significant diseases was found with Venous (V) temporal angle, with five diseases being significantly associated only with deep, but not measured TIFs (OR = 2.74, p = 0.21). While also not significant, diabetes-related eye disease ("diabetes-eye") displayed the largest number of differences between TIF types, with five more deep TIFs than measured TIFs (OR = 3.42, p = 0.16). When breaking down the diseases into general categories, we observed that ocular diseases tended to have significantly more cases with associated deep TIFs but no measured TIFs (24) than the reverse (3 cases, OR = 1.78, p = 0.02). In contrast no such bias was observed across risk factors (measured TIF = 13, deep TIF = 16; OR = 1.06, p = 0.84) and general diseases (measured TIF = 11, deep TIF = 8; OR = 0.91, p = 0.80). Overall, these findings suggest that deep TIFs may offer enhanced associative strength for certain disease categories, particularly ocular diseases, while demonstrating comparable performance to measured TIFs for risk factors and general diseases.

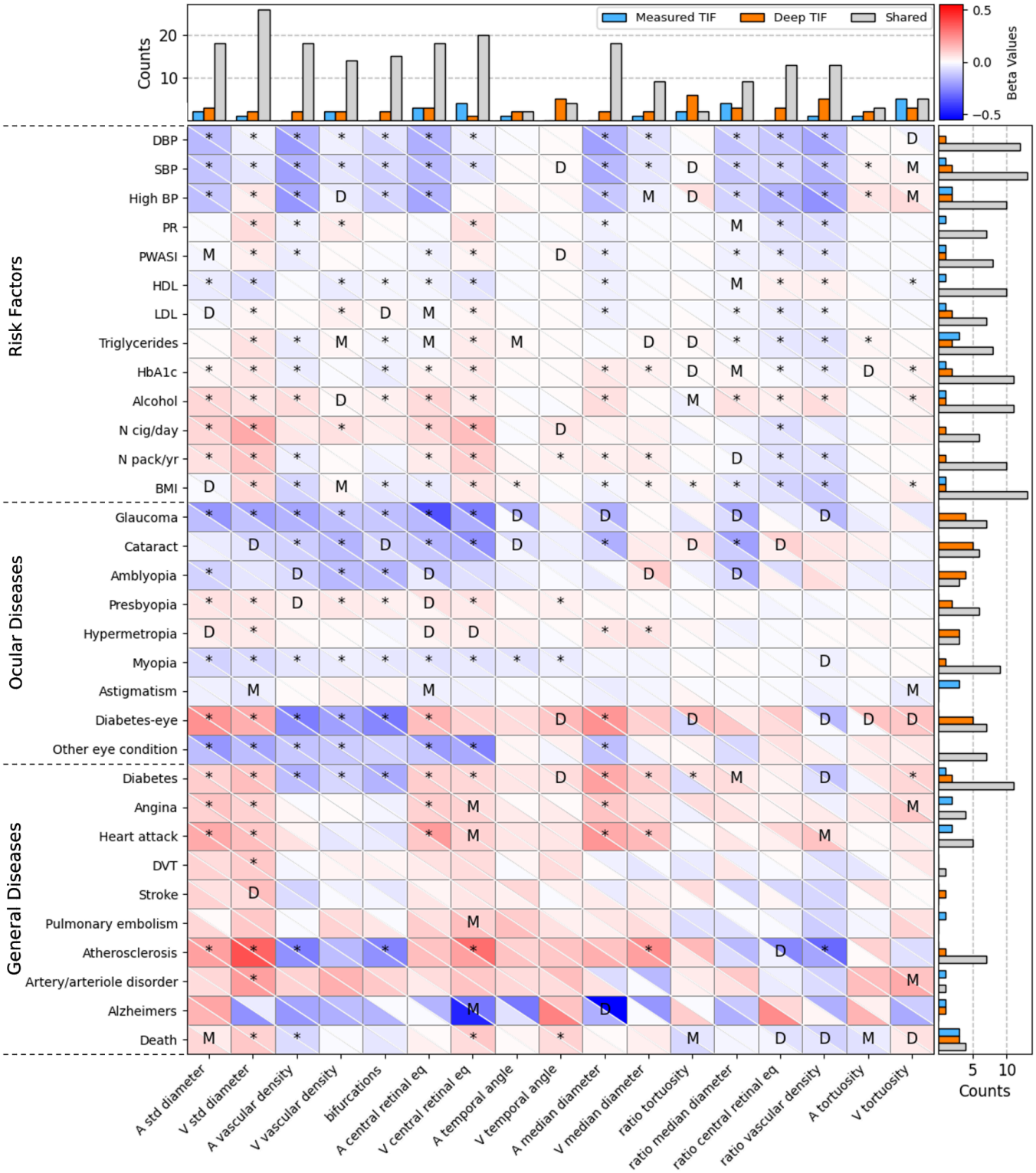

Figure 5 | Disease associations with measured and deep TIFs. Each cell presents the standardised effects ("beta values") of a linear model that explains a risk factor, ocular or general disease incident with a single measured (lower left) or deep (upper right) TIF alongside with covariates, including age and sex (see color map in the top right). Associations significant only for the deep or measured TIFs are indicated as with 'D' and 'M', respectively and a star (*), if significant for both. The bar graphs on the top and right indicate the total number of unique counts for each type of these associations across all diseases and TIFs, respectively. Abbreviations: A = Arteries; V = Veins; DBP = Diastolic Blood Pressure; SBP = Systolic Blood Pressure; Age BP = Age of High Blood Pressure Diagnosis; PR = Pulse Rate; PWASI = Pulse Wave Arterial Stiffness Index; HDL = High Density Lipoprotein; LDL = Low Density Lipoprotein; HbA1c = Glycated hemoglobin; N cig/day = self-reported number of cigarettes per day; N pack/yr = self-reported number of packs of cigarettes per year; BMI = Body Mass Index; DVT = Deep Vein Thrombosis.

Finally, we sought to elucidate whether models that combine multiple image features, including combinations of different types, have the potential to generate increased explanatory power over the simple models explored above. To the end we used multilinear models with the following seven sets of explanatory features: (i) a baseline model including only standard covariates (Covar), such as age and sex; (ii) the baseline model plus all 17 measured TIFs (mTIFs); (iii) the baseline model plus all 17 deep TIFs (dTIFs); (iv) the baseline model plus all 1024 LVs; (v) the baseline model plus all mTIFs and LVs; (vi) the baseline model plus all dTIFs and LVs; (vii) the baseline model plus all mTIFs and dTIFs; and (viii) all features. Fig. 6 show the explained variance added by each of the eight models, indicating that adding CFI features can improve substantially the explanatory power for some risk factors (in particular blood pressure), and that this improvement is most pronounced when combining TIFs (whether measured or deep) with LVs.

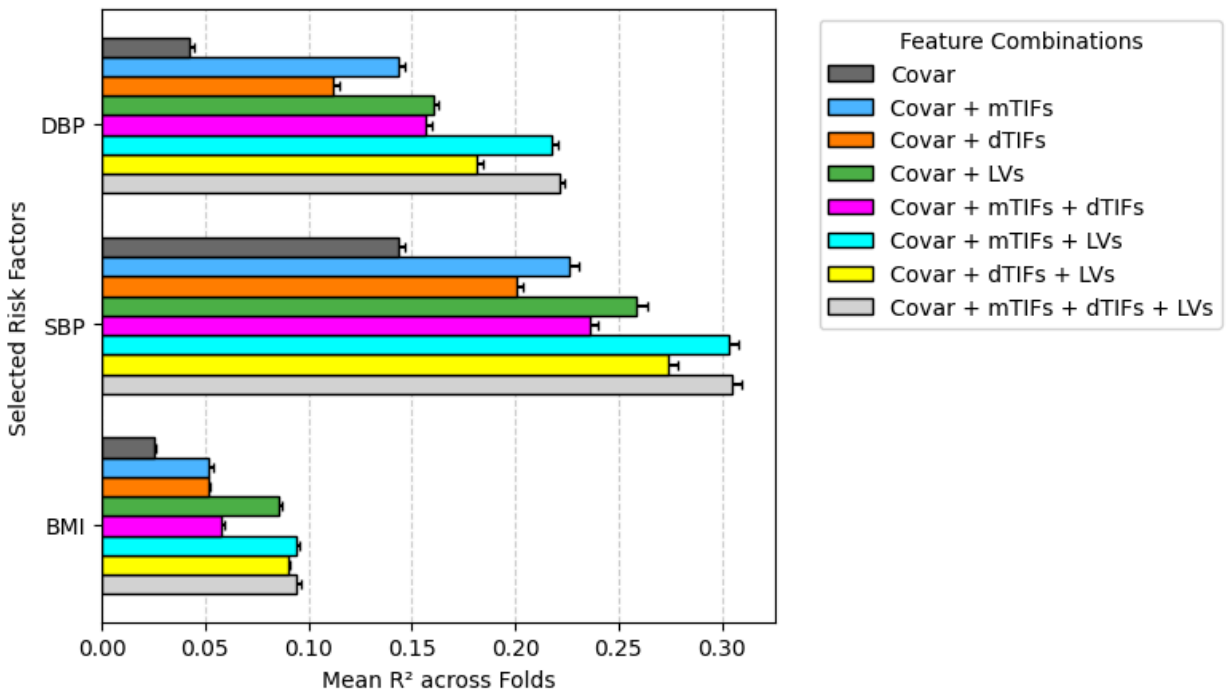

Figure 6 | Risk factors associations with CFI features. Explained variance added by linear models explaining disease traits (y-axis) in terms of CFI feature sets (x-axis), including: (i) a baseline model including only standard covariates, (ii) all 17 measured TIFs, (iii) all 17 deep TIFs, (iv) all 1024 LVs, (v) all measured TIF and LVs, (vi) all deep TIFs and LVs, (vii) all measured and deep TIFs, (viii) all features. Abbreviations: Covar = Covariates; dTIF = Deep Tangible Image Features; mTIF = Measured Tangible Image Feature; LV = Latent Variable; DBP = Diastolic Blood Pressure; SBP = Systolic Blood Pressure; BMI = Body Mass Index.

Discussion

Harnessing the rich information provided by medical images for disease diagnoses or risk assessment is being revolutionised by advances in computer-aided analyses. The latter range from tasks that support human physicians, such as segmentation of anatomical structures and taking well-defined

measurements on them, all the way to making fully automated disease risk predictions using end-to-end machine learning pipelines. Deep learning (DL) algorithms typically hold internal “latent space” representations of the images for this task, but their nonlinear entanglement makes them appear as “black boxes” lacking straightforward explanations. This is problematic in the biomedical domain where trust by the clinician is essential [15–18]. Furthermore, interpretability is key for deciphering the underlying biological or physiological mechanisms, which is pivotal for advancing treatment [19]. Visual interpretability methods, such as attention maps [20–22] and counterfactual explanations [23], have been developed to address this but often fall short in clarifying exactly how latent representations drive predictions (see [18] for recent advances in tackling this issue).

To shed some light on what latent variables (LVs) represent, in this study we investigated a recently proposed foundation model for retinal images, and applied it to a large collection of colour fundus images (CFIs) which we had previously characterised in terms of “tangible image features” (TIFs) of high clinical interpretability. Specifically, we asked to what extent LVs can predict TIFs, individually or jointly. Moreover, the fact that these images stem from genotyped subjects with extensive clinical phenotyping allowed us to directly compare the associations of LVs with genotypes and disease states or risk factors with those of TIFs.

Our analysis revealed that, prior to fine-tuning, individual LVs of *RETFound* do not closely match any of our TIFs, with the vascular densities and number of bifurcations having the closest LV representation explaining close to 40% of their variance, while it is less than 20% for all the other, more complex TIFs. Linear combinations of multiple LVs have substantially higher predictive power for TIFs, in particular for the densities, explaining close to 90% of their variance. Strikingly, fine-tuning *RETFound* to predict TIFs yielded marginal or no improvements compared to the LVs (c.f. [24–26] for similar negative results on fine tuning in LLMs), with the notable exception of the temporal angles between the main vascular branches and, to a lesser degree, vessel tortuosity. Thus, even the advanced transformer DL architecture of *RETFound* and its foundation model parameters trained on more than a million of retinal CFIs, do not allow for fine-tuning it to provide close approximation to some of the well-established classical ophthalmological parameters of the retinal vasculature. This is remarkable, because, in contrast to clinical endpoints, TIFs are fully determined by the images, i.e. there exist deterministic algorithmic procedures based on conventional image processing.

Our analysis of genetic associations elucidates to what extent DL features are modulated genetically and the degree of overlap with the association signals of the TIFs. We argue that genetic associations support physiological relevance, because they point to a causal chain from a DNA difference (i.e. a genetic variant at the level of the DNA) over a molecular variation (such as differential gene expression)

to a (typically minute) phenotypic variability. Heritability can be seen as a global summary statistic of such genetic effects and our results indicate that while deep TIFs are at least as heritable as their measured counterparts, supporting similar if not higher physiological pertinence, this is not true for the full spectrum of *RETFound* LVs. Indeed, it makes sense that many LVs may pertain to image features that have nothing to do with genetics, but rather summarise environmental, or even technical variability (such as illumination or orientation), to which TIFs should not be sensitive by design.

A complementary means of evaluating the pertinence of vascular features is in terms of their capacity to facilitate disease prediction. In our study we contrasted measured and deep TIFs by comparing their standardised effects in linear and logistic models with diseases or risk factors as response variables. Our findings demonstrate that overall they performed comparably for cardiovascular diseases and their risk factors, while deep TIFs appeared to outperform their measured counterparts in predicting ocular diseases. This outcome may be attributed to *RETFound'*s initial training on a diverse set of retinal images and tasks [2,27]. We hypothesise that its fine-tuned weights, while focused on predicting vascular morphology, may retain some of *RETFound'*s broader representation of retinal features. This may allow the model to capture subtle patterns or relationships between vascular morphology and other retinal structures that are relevant to ocular pathologies but not explicitly measured in traditional vascular assessments. The enhanced performance in ocular disease prediction using our fine-tuned model suggests that AI-derived vascular morphology metrics may provide additional informative features for ocular condition detection compared to conventional morphological measurements alone.

Our findings reveal that deep TIFs can predict some complex measured TIFs, such as vessel tortuosity, with reasonable accuracy; however, deep TIFs appear to perform poorly with certain measured features, such as assessing the measures of ratios between arteries and veins. We also find that deep TIFs often exhibit heritability comparable to their measured TIF counterparts, and occasionally have higher heritability than the measured TIFs, such as in the case of temporal angles of the arteries. Furthermore, while genetic and disease associations between deep and measured TIFs display some overlap, we observed complementarity between these two TIF measures, suggesting that deep TIFs may provide additional insights beyond traditional measurements. Even more striking are our initial results demonstrating that the integration of LVs with TIFs as explanatory variables can outperform models that only use either of these CFI features. Overall, this work underscores the potential of DL approaches to enhance our understanding of retinal biology and disease mechanisms, potentially leading to improved diagnostic and prognostic tools in ophthalmology.

Methods

Data and quality control

The UK Biobank is a population-based cohort of ~488k subjects with rich, longitudinal phenotypic data including medical history, and a median 10-year follow-up [28]. Standard retinal 45° CFIs were captured using a Topcon 3D-OCT 1000 Mark II, resulting in a total of 173 814 CFIs from 84 813 individuals. For this study we used 41 527 of these images from the right eye for which we were able to measure all of our 17 TIFs using our previously described analysis pipeline [11]. This effectively removed CFIs for which our algorithm failed to identify the optic disk or enough major blood vessels to measure the temporal angles. Since this happens more often for CFIs from older and diseased subjects, the cohort of the present study appears to be more "healthy" with respect to its baseline risk factors (see Suppl. Table 1 for comparison to our previous study). Since some of the LVs produced from *RETFound* encoded information specific to eye laterality (left vs. right), we opted to analyze a single eye rather than averaging across both eyes. Genotyping was performed on Axiom arrays for a total of 805 426 markers, from which ~96 million genotypes were imputed. We used the subset of 15 599 830 SNPs that had been assigned an rsID.

Measured phenotypes

The 17 measured TIFs comprise tortuosity, temporal angles, vessel diameters, its variability, vascular density, and the number of bifurcations. With the exception of the number of bifurcations, all of these measurements were performed independently for arteries and veins. To explore asymmetries between arteries and veins, the arterio-venous ratio between diameters, tortuosities and vascular densities were also included. Complete algorithmic descriptions of how these measurements were extracted are given in our recent publication [11].

Deep learning model

We utilise *RETFound*, a novel foundation model that leverages Google's vision transformer large patch 16 architecture [29]. Briefly, *RETFound* uses a masked autoencoder (MAE) approach [30] to optimise the model by reconstructing retinal images. In this process, the model masks a random subset of image patches and then reconstructs the missing patches, thereby learning effective representations of the images provided as training data. We then use the encoder portion of this model for extracting LVs and making predictions.

Extracting Latent Variables

To generate LVs, we first resize each image to 224×224 pixels and normalise the image to the means and standard deviations of ImageNet's RGB channels. After preprocessing the image, we then propagate each individual image through the encoder portion of the pre-trained *RETFound* model to generate 1024 LVs for each image.

Predicting tangible image features

For predicting TIFs, we also use just the encoder portion of the *RETFound* model and add a multi-layer perception to the head of the model to learn each TIF from an image. We utilised the base hyperparameters established in the original *RETFound* manuscript [2]. Specifically, we trained the model over 50 epochs, with a "warm up" in the 10 first epochs (with the learning rate r monotonically increasing from 0 to 5×10^{-4}), followed by a cosine annealing schedule (where r decreases from 5×10^{-4} to 1×10^{-6}). The batch size was set to 16, the dropout rate was set to 0.2, and image size was set to 224×224 pixels. After each epoch, we evaluate the model on a validation dataset and then keep the model that has the lowest mean absolute error across all epochs. Individual models were built for each of the 17 TIFs using 5-fold cross-validation.

Covariate corrections

All traits were corrected for a set of covariates known to confound the biological system in general, or the eye specifically. They comprise sex, age, age-squared, sex-by-age, sex-by-age-squared, spherical power, spherical power-squared, cylindrical power, cylindrical power-squared, imaging instance, assessment centre, genotype measurement batch, and the first 20 genomic PCs (see Suppl. Fig. 3).

Genomic analyses

For genome-wide association studies (GWAS), raw phenotypes were transformed using the rank-based inverse normal transformation. GWAS were conducted using BGENIE [31]. SNP-specific heritabilities and genetic correlations between IDPs were calculated using LD score regression [32]. LD score regression estimates SNP heritabilities by fitting an ordinary least squares (OLS) regression of SNP LD scores against their average chi-squared statistics. Gene and pathway scores were computed using *PascalX* [13,14,33]. Both protein-coding genes and lincRNAs were scored using the approximate "saddle" method, incorporating all SNPs within a 50 kb window surrounding each gene. All pathways from MSigDB v7.2 [34–36] were scored using *PascalX*'s ranking mode, merging and rescoring co-occurring genes separated by less than 100 kb. PascalX utilises LD structure for accurate computation of gene scores, which in these analyses was derived from the UK10K (hg19) reference panel [37].

Disease associations

Three distinct groups of disease traits were included in this work. First, a set of general systemic risk factors, second, a set of ocular diseases, and third, a set of systemic diseases, including mortality (see Suppl. Fig. 2). All disease data were collected from the UK Biobank (see Ref. [11] for how we mapped the data field identifiers to each disease included in this study).

In the case of the continuous risk factors, linear regression was used to estimate standardised effects using the 'statsmodels.formula.api' library in Python 3.8.13. For both ocular and general diseases, time-to-event data were binarised, and logistic regression was applied using the logit function from the 'statsmodels.formula.api' library.

For all regression analyses, traits were corrected for the listed covariates. To determine the significance of regression coefficients, p-values were computed, and the Benjamini-Hochberg procedure was applied using a false discovery rate of 0.05.

References

1. Azad B, Azad R, Eskandari S, Bozorgpour A, Kazerouni A, Rekik I, et al. Foundational Models in Medical Imaging: A Comprehensive Survey and Future Vision. 2023. Available: http://arxiv.org/abs/2310.18689

2. Zhou Y, Chia MA, Wagner SK, Ayhan MS, Williamson DJ, Struyven RR, et al. A foundation model for generalizable disease detection from retinal images. Nature. 2023;622: 156–163.

3. Parres D, Paredes R. Fine-Tuning Vision Encoder–Decoder Transformers for Handwriting Text Recognition on Historical Documents. Document Analysis and Recognition - ICDAR 2023. 2023; 253–268.

4. Ho H, Cheung CY, Sabanayagam C, Yip W, Ikram MK, Ong PG, et al. Retinopathy Signs Improved Prediction and Reclassification of Cardiovascular Disease Risk in Diabetes: A prospective cohort study. Sci Rep. 2017;7: 41492.

5. Fhima J, Van Eijgen J, Billen Moulin-Romsée M-I, Brackenier H, Kulenovic H, Debeuf V, et al. LUNet: deep learning for the segmentation of arterioles and venules in high resolution fundus images. Physiol Meas. 2024;45. doi:10.1088/1361-6579/ad3d28

6. Quiros JV, Liefers B, van Garderen K, Vermeulen J, Center ER, Consortium S, et al. VascX Models: Model Ensembles for Retinal Vascular Analysis from Color Fundus Images. 2024. Available: http://arxiv.org/abs/2409.16016

7. Tomasoni M, Beyeler MJ, Vela SO, Mounier N, Porcu E, Corre T, et al. Genome-wide Association Studies of Retinal Vessel Tortuosity Identify Numerous Novel Loci Revealing Genes and Pathways Associated With Ocular and Cardiometabolic Diseases. Ophthalmology Science. 2023;3: 100288.

8. Zhou Y, Wagner SK, Chia MA, Zhao A, Woodward-Court P, Xu M, et al. AutoMorph: Automated Retinal Vascular Morphology Quantification Via a Deep Learning Pipeline. Trans Vis Sci Tech. 2022;11: 12–12.

9. Van Eijgen J, Fhima J, Billen Moulin-Romsée M-I, Behar JA, Christinaki E, Stalmans I. Leuven-Haifa High-Resolution Fundus Image Dataset for Retinal Blood Vessel Segmentation and Glaucoma Diagnosis. Sci Data. 2024;11: 257.

10. Ortin Vela S, Beyeler MJ, Trofimova O, Tomasoni M. Phenotypic and Genetic Characteristics of Retinal Vascular Parameters and their Association with Diseases. medRxiv. 2023. Available: https://www.medrxiv.org/content/10.1101/2023.07.07.23292368.abstract

11. Ortín Vela S, Beyeler MJ, Trofimova O, Iuliani I, Vargas Quiros JD, de Vries VA, et al. Phenotypic and genetic characteristics of retinal vascular parameters and their association with diseases. Nat Commun. 2024;15: 9593.

12. Advances in medical image analysis with vision Transformers: A comprehensive review. Medical Image Analysis. 2024;91: 103000.

13. Krefl D, Bergmann S. Cross-GWAS coherence test at the gene and pathway level. PLoS Comput Biol. 2022;18: e1010517.

14. Krefl D, Brandulas Cammarata A, Bergmann S. PascalX: a Python library for GWAS gene and pathway enrichment tests. Bioinformatics. 2023;39. doi:10.1093/bioinformatics/btad296

15. Andrews B, Chang J-B, Collinson L, Li D, Lundberg E, Mahamid J, et al. Imaging cell biology. Nature Cell Biology. 2022;24: 1180–1185.

16. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28: 31–38.

17. DeGrave AJ, Cai ZR, Janizek J, Daneshjou R, Lee S-I. Auditing the inference processes of medical-image classifiers by leveraging generative AI and the expertise of physicians. Nature Biomedical Engineering. 2023; 1–13.

18. Rotem O, Schwartz T, Maor R, Tauber Y, Shapiro MT, Meseguer M, et al. Visual interpretability of image-based classification models by generative latent space disentanglement applied to in vitro fertilization. Nature Communications. 2024;15: 1–19.

19. Rotem O, Zaritsky A. Visual interpretability of bioimaging deep learning models. Nat Methods. 2024;21: 1394–1397.

20. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016. doi:10.1109/cvpr.2016.319

21. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. 2016 [cited 17 Dec 2024]. doi:10.1007/s11263-019-01228-7

22. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. Precup D, Teh YW, editors. arXiv [cs.CV]. 06--11 Aug 2017. pp. 3145–3153. Available: https://proceedings.mlr.press/v70/shrikumar17a.html

23. Lang O, Gandelsman Y, Yarom M, Wald Y, Elidan G, Hassidim A, et al. Explaining in Style: Training a GAN to explain a classifier in StyleSpace. arXiv [cs.CV]. 2021. Available: http://arxiv.org/abs/2104.13369

24. Fine-Tuning or Fine-Failing? Debunking Performance Myths in Large Language Models. [cited 18 Dec 2024]. Available: https://arxiv.org/html/2406.11201v1

25. Li X, Chan S, Zhu X, Pei Y, Ma Z, Liu X, et al. Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks. 2023. Available: http://arxiv.org/abs/2305.05862

26. Uppaal R, Hu J, Li Y. Is Fine-tuning Needed? Pre-trained Language Models Are Near Perfect for Out-of-Domain Detection. 2023. Available: http://arxiv.org/abs/2305.13282

27. Fu Y, Ma L, Wan S, Ge S, Yang Z. A novel clinical artificial intelligence model for disease detection via retinal imaging. Innovation (Camb). 2024;5: 100575.

28. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562: 203–209.

29. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2020. Available: http://arxiv.org/abs/2010.11929

30. He K, Chen X, Xie S, Li Y, Dollar P, Girshick R. Masked Autoencoders Are Scalable Vision Learners. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2022. doi:10.1109/cvpr52688.2022.01553

31. Band G, Marchini J. BGEN: a binary file format for imputed genotype and haplotype data. bioRxiv. 2018. p. 308296. doi:10.1101/308296

32. Bulik-Sullivan B, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47: 291–295.

33. Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. 2016. doi:10.1371/journal.pcbi.1004714

34. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102: 15545–15550.

35. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1: 417–425.

36. Castanza AS, Recla JM, Eby D, Thorvaldsdóttir H, Bult CJ, Mesirov JP. Extending support for mouse data in the Molecular Signatures Database (MSigDB). Nature Methods. 2023;20: 1619–1620.

37. UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015;526: 82–90.