Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Measuring lesion–vascular context for diabetic retinopathy grading and retinal phenotyping

**Inam Ullah**

i1n23@soton.ac.uk

University Of Southampton https://orcid.org/0000-0001-9488-035X

**Shoaib Jameel**

University Of Southampton

**Article**

**Keywords:**

**Posted Date:** November 19th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11777044/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** There is no conflict of interest

RESEARCH ARTICLE

# Measuring lesion–vascular context for diabetic retinopathy grading and retinal phenotyping

Inam Ullah*, Shoaib Jameel
Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, United Kingdom
*Corresponding author: Inam Ullah, i1n23@soton.ac.uk

**BACKGROUND/OBJECTIVES:** Diabetic retinopathy (DR) grades summarise severity but do not quantify the vascular surroundings of individual lesions. We investigated whether measured lesion–vascular context improves lesion-based grading and characterises vascular variation beyond severity.

**SUBJECTS/METHODS:** This cross-sectional secondary analysis of public datasets matched 2,874 APTOS photographs to lesion predictions and AutoMorph vascular measurements (1,724 training, 575 validation, 575 test). Five models trained exclusively on DDR annotations predicted four lesion families. Registered anatomy yielded 24 spatial descriptors and 1,584 lesion–biomarker products. Structured ordinal models, an anatomical graph and six frozen image encoders evaluated grading. Separate grade-adjusted analyses assessed vascular associations and prediction.

**RESULTS:** Structured context increased test quadratic weighted kappa (QWK) from 0.764 to 0.827 (difference 0.063; 95% confidence interval [CI] 0.028–0.099). The graph achieved QWK 0.849 and macro area under the receiver operating characteristic curve 0.924 (QWK difference 0.085; 95% CI 0.050–0.119); both comparisons with lesion-only grading had false-discovery-rate-adjusted $q < 0.001$. Individual contextual additions were not separately supported. Graph sensitivity for proliferative DR was 0.316, compared with 0.605 for structured context. Grade-adjusted analysis supported 103 lesion–biomarker associations, and lesion counts improved prediction of 26/66 vascular measurements, predominantly calibre/width and fractal measures.

**CONCLUSIONS:** Combined lesion–vascular representations improved lesion-only grading and characterised vascular variation within severity categories. The measurements provide inspectable anatomical evidence, but segmentation limitations and classwise errors preclude claims of clinical readiness or prognostic validity.

## INTRODUCTION

Diabetes mellitus is characterised by chronic hyperglycaemia and complications affecting multiple organs, including the retina. A pooled population analysis estimated that 828 million adults had diabetes in 2022, underscoring the need for long-term surveillance of its complications [1]. Diabetic retinopathy is a sight-threatening neurovascular complication in which vascular leakage, occlusion and retinal injury produce clinically recognisable lesions [2]. The International Clinical DR scale distinguishes no DR, mild, moderate and severe non-proliferative DR, and proliferative DR (PDR), characterised by pathological neovascularisation [3]. These categories summarise severity but do not quantify the local vascular surroundings of individual lesions.

Microaneurysms reflect focal vascular abnormalities, haemorrhages indicate bleeding, hard exudates are deposits associated with leakage, and cotton-wool spots reflect disturbed nerve-fibre-layer transport [2]. Lesion burden, spatial distribution and vascular morphology therefore describe related but different aspects of retinal injury. A count cannot distinguish lesions overlapping a vascular skeleton from those separated from it. Conversely, a global measurement of calibre, density, tortuosity or network complexity cannot identify the vascular neighbourhood of a particular lesion. Eyes assigned the same severity grade may consequently differ in measurable lesion–vascular organisation.

Automated assessment has progressed from explicit retinal features [4] to image classifiers [5], lesion-aware systems [6] and pretrained retinal representations [7]. Saliency and lesion-concept graphs can expose regions or concepts associated with grading [8, 9], but neither is equivalent to a measured component-to-vessel distance. Joint anatomical and pathological annotations, including MAPLES-DR, facilitate analysis of visible retinal evidence [10]. The relevant gap is not the absence of interpretable or graph-based DR methods; it is the need to define and evaluate image-linked measurements connecting predicted lesions to their vascular setting.

The distinction also matters for interpreting automated outputs. A high grading score can arise from lesion burden without revealing the vascular structures adjacent to those lesions. A heatmap may identify influential tissue, but does not establish whether that tissue represents a lesion or quantify its anatomical surroundings. Explicit measurement provides a complementary level of description: a numerical value whose component, reference structure and calculation can be inspected together. It need not replace learned image features to be useful.

Clinical imaging studies provide complementary motivation. Verma and colleagues quantified haemorrhages and their optic-nerve-relative location when examining progression to PDR [11]. Retinal vascular geometry has been associated with subsequent DR [12], and AutoMorph-derived morphology with incident treatment for sight-threatening disease [13]. Lesion eccentricity and global vascular traits, however, differ from local correspondence to vessel, artery, vein and branch structures. Whether this correspondence contributes information beyond lesion burden remains an empirical question.

We aligned independently estimated lesions and vascular anatomy to measure proximity, skeleton overlap and surrounding occupancy. We evaluated these descriptors and lesion–biomarker products using matched ordinal grading comparisons, an anatomical graph and six frozen image encoders.

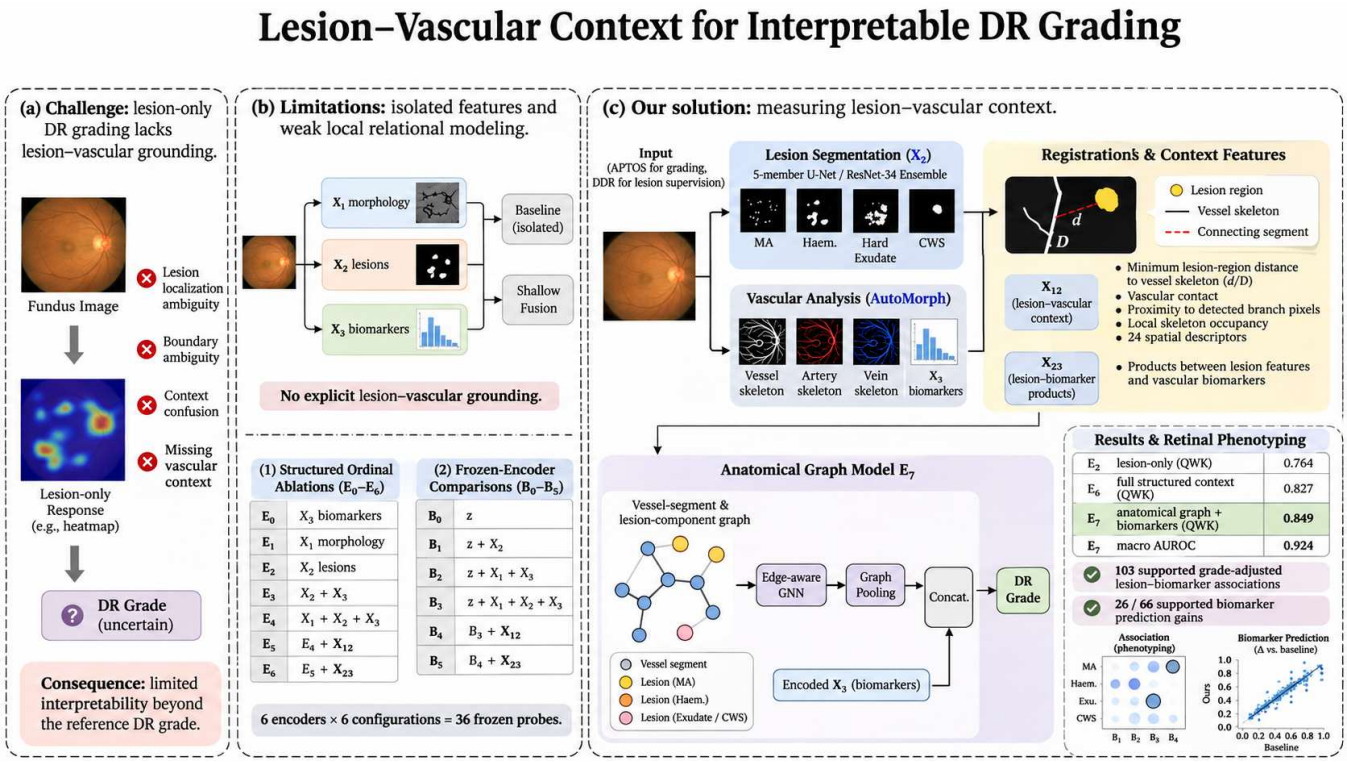



**Fig. 1 Study overview (a)** Lesion-only grading identifies disease-related image evidence but does not quantify the vascular surroundings of individual lesions. **(b)** Lesion burden, vascular morphology and vascular biomarkers provide complementary information, while models based on isolated features or their combination do not explicitly measure lesion–vascular correspondence. Structured ordinal models and six frozen image encoders evaluate progressively combined representations. **(c)** DDR-supervised predictions of microaneurysms, haemorrhages, hard exudates and cotton-wool spots are aligned with AutoMorph-derived vascular anatomy in APTOS photographs. Registered anatomy yields lesion–vascular spatial descriptors ($X_{12}$) and lesion–biomarker products ($X_{23}$), which are evaluated alongside lesion burden and vascular measurements using structured ordinal models, an anatomical graph and frozen image encoders. Separate grade-adjusted association and conditional prediction analyses assess whether lesion burden characterises vascular variation beyond reference DR grade. Arrows indicate analytical processing rather than causal relationships.

Separately, grade-adjusted associations and conditional prediction assessed whether lesion burden characterised vascular variation within severity categories. The study links anatomical measurement, grading and retinal phenotyping without treating cross-sectional associations as evidence of prognosis or systemic disease.The overall study design and analytical comparisons are summarised in Fig. 1.

## MATERIALS AND METHODS

### Study design, datasets and partitions

This cross-sectional secondary analysis used publicly available APTOS 2019 photographs for five-class grading and DDR images for lesion supervision. Measurements and labels were analysed at the image level; no new participants were recruited or longitudinal follow-up analysed. APTOS provides 3,662 labelled fundus photographs [14]. Matching images with vascular maps, biomarkers and lesion predictions yielded 2,874 observations, allocated by grade to training (1,724), validation (575) and testing (575). Test counts for grades 0–4 were 296/61/155/25/38. Every grading comparison used these non-augmented image partitions. Model fitting and preprocessing were performed using the training set, while the validation set was used for configuration selection. The test set was reserved for final evaluation and was not used for model fitting or selection. Exclusion reasons for the 788 unmatched images could not be fully recovered, and patient-level linkage was unavailable.

DDR supplied the sole expert supervision for microaneurysms (MA), haemorrhages(H), hard exudates (HE) and cotton-wool spots (CWS), retaining its 383/149/225 training/validation/test segmentation partitions [15]. DDR test masks and APTOS test grades did not select segmentation checkpoints. APTOS had no expert lesion masks. Supplementary Section S1 provides further details on cohort provenance, the original dataset partitions and complete grade distributions.

These measurements were combined at the image level.For photograph $I_i$ with reference grade $Y_i \in \{0, 1, 2, 3, 4\}$, the structured representation was

$$r_i = [X_{1,i}, X_{2,i}, X_{3,i}, \mathrm{vec}(X_{12,i}), X_{23,i}]. \quad (1)$$

The feature groups were varied under fixed partitions to compare their information. The graph used the corresponding anatomical objects rather than this tabular vector. Conditional vascular analysis used reference grade as a covariate, not as an outcome to be predicted. The two tasks therefore shared image evidence but had different targets.

### Vascular anatomy and lesion evidence

AutoMorph [16] prepared the retinal field, assessed image usability and supplied vessel, artery and vein skeletons: thin centreline representations. Morphology ($X_1$) comprised 16 summaries: presence, occupancy, connected components, endpoints and branchpoints for each vascular category, plus the artery-to-vein density ratio. Biomarkers ($X_3$) comprised 72 fields: six disc/cup measurements, 18 global vascular measurements and 24 each from optic-disc-referenced zones B

and C. These included density, width, tortuosity (vessel winding), fractal dimension (network complexity), central retinal arteriolar/venular equivalents and their ratio. Supplied width scales were retained; verified physical calibration was not assumed.

Five independently seeded U-Net/ResNet-34 models [17, 18] learned four independent sigmoid channels from DDR masks, with $c \in \{\mathrm{MA}, \mathrm{H}, \mathrm{HE}, \mathrm{CWS}\}$ indexing the lesion family. ImageNet-initialised encoders and randomly initialised decoders used lesion-balanced 512-pixel patches, horizontal/vertical flips and focal–Dice loss. Each trained for ten epochs with Adam, initial learning rate $10^{-4}$, cosine scheduling and gradient clipping. Validation selected checkpoints. Inference used 512-pixel tiles with 128-pixel overlap; probabilities were averaged before a common threshold:

$$\begin{aligned} \overline{P}_{ic}(p) &= \frac{1}{5}\sum_{s=1}^{5} P_{ic}^{(s)}(p), \\ M_{ic}(p) &= \mathbf{1}\left[\overline{P}_{ic}(p) \geq 0.5\right]. \end{aligned} \tag{2}$$

The 36 active lesion fields ($X_2$) included mean, 95th-percentile and maximum probability, component/image counts and quadrant counts. Confidence and count-based burden remained distinct. Eight-connected components are contiguous predicted regions, not necessarily individual clinical lesions. Supplementary Sections S2 and S5 specify feature accounting and segmentation. The three streams were extracted separately, without a shared foundation encoder or joint segmentation–grading objective. AutoMorph quality assessment concerned image usability, not disease severity. Because morphology and biomarkers were derived from the same vascular anatomy, statistical independence between $X_1$ and $X_3$ was not assumed.

### Registration and spatial measurements

Identifiers, image hashes and recorded crop, padding and resize transforms linked predictions to the $912 \times 912$ AutoMorph frame. Stored eight-bit probabilities used bilinear interpolation; field-of-view (FOV) masks used nearest-neighbour interpolation. Thresholding and FOV intersection followed alignment. Matching image dimensions alone did not establish registration (Supplementary Section S4). Saved probabilities were rounded to 0–255 values; threshold 128 represented the fixed 0.5 decision rule in the native archive. This threshold was not changed according to reference grade. Registration records connected the native prediction maps and anatomical outputs to the same photograph before calculating any spatial feature.

Figure 2 illustrates the component-level calculation for the main example, while Supplementary Fig. S2 shows the same procedure across all four lesion families. Let $F_i$ denote the retinal field and $V_i, A_i, W_i$ the vessel, artery and vein skeleton products. Skeletonising $V_i$ gives $S_i$; pixels with at least three occupied eight-neighbours define branch reference $B_i$, without merging adjacent branch pixels into anatomical bifurcations. For component $\ell$ and nonempty reference $T \in \{V_i, A_i, W_i, B_i\}$,

$$D_i = \sqrt{4|F_i|/\pi}, \qquad d_T(\ell) = \frac{1}{D_i}\min_{p\in\ell,\, t\in T} \|p - t\|_2. \tag{3}$$

Here $D_i$ is the equivalent-field diameter. The Euclidean distance transform implements the minimum over the entire component, not its centroid. Zero requires shared lesion/reference pixels; adjacency without overlap remains positive. Normalisation makes distances dimensionless.

Direct skeleton overlap and surrounding occupancy are

$$o(\ell) = \frac{|\ell \cap S_i|}{|\ell|}, \qquad \rho(\ell) = \frac{|N_r(\ell) \cap S_i|}{|N_r(\ell)|}, \tag{4}$$

where $N_r(\ell) = \{p \in F_i : \mathrm{dist}(p, \ell) \leq r\}$ and $r = \max(1, \mathrm{round}(0.03 D_i))$. The neighbourhood includes the lesion and is clipped to the retinal field; its radius is 3% of diameter, not area. These are thin-skeleton measurements, not perfusion or filled-vessel density.

The vector $\phi(\ell) = [d_V, d_A, d_W, d_B, o, \rho]$ contains six measurements. Let $\mathcal{L}_{ic}^{(k)}$ denote the set of components of lesion family $c$ in image $i$ for which descriptor $k$ is finite. Image-level aggregation was

$$X_{12,i}[c,k] = \frac{\sum_{\ell \in \mathcal{L}_{ic}^{(k)}} \phi_k(\ell)}{|\mathcal{L}_{ic}^{(k)}|}, \qquad X_{12,i} \in \mathbb{R}^{4\times 6}. \tag{5}$$

This gives 24 lesion–vascular spatial descriptors, with components weighted equally regardless of area. The minimum distance was calculated within each component before averaging across components, rather than as a single minimum over all lesion pixels in an image. When a lesion family was absent, all six descriptors for that family remained missing. When an anatomical reference was absent, only the corresponding distance remained missing. Missingness was retained until training-fitted preprocessing.Supplementary Algorithm Supplementary Algorithm S1 provides the full extraction procedure.

### Lesion–biomarker products

Image-level products capture joint variation rather than spatial geometry. Let $b_i \in \mathbb{R}^{24}$ contain lesion-count fields and $m_i \in \mathbb{R}^{66}$ vascular biomarkers, excluding disc/cup measurements. Training-fitted median imputation and standardisation yield

$$X_{23,i} = \mathrm{vec}(\widetilde{b}_i \widetilde{m}_i^{\top}) \in \mathbb{R}^{1584}. \tag{6}$$

Each product represents joint deviation from training distributions, not an identified biological mechanism. Correlated counts and regional measurements induce correlated predictors. For a given count and biomarker, a positive product can reflect either jointly high or jointly low standardised values; it is not a direct measure of lesion proximity. Forming products before feature selection allows the ordinal grader to consider these joint deviations alongside the original streams. Supplementary Algorithm S2 preserves field ordering and the training-only fitting boundary. Phenotyping instead uses four global lesion counts directly.

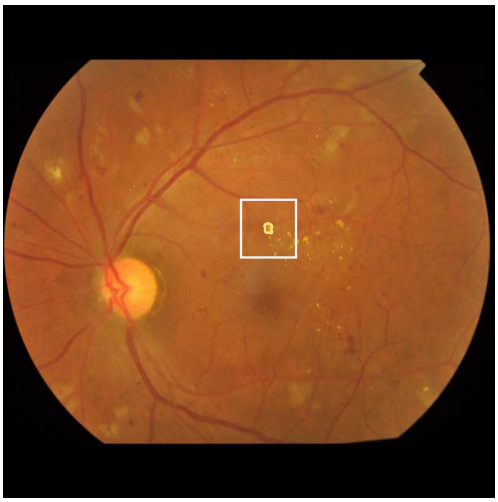
### Structured and graph-based grading

Table 1A defines E0–E7. E0–E6 shared an ordinal classifier. Invalid negative/non-finite biomarkers became missing; training-fitted median imputation, missingness indicators and standardisation preceded analysis-of-variance ranking against the five grades. The selected vector $z_i$ entered four L2-regularised logistic heads:

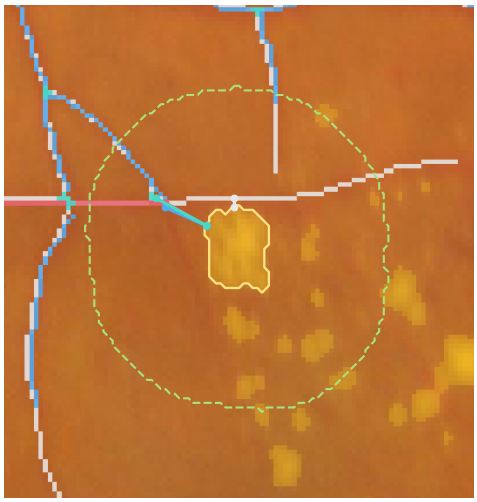
$$\begin{aligned} t_{ik} &= \mathbf{1}[Y_i > k], \\ u_{ik} &= \sigma(w_k^{\top} z_i + a_k), \quad k = 0, \ldots, 3. \end{aligned} \tag{7}$$

Each head used inverse-frequency binary class weights and `liblinear`, with at most 1,000 iterations. Validation QWK selected $C \in \{0.03, 0.1, 0.3, 1\}$ and $K \in \{64, 128, 256\}$, retaining

**A** Registered image and HE components



**B** Measured lesion neighbourhood



Vessel Artery Vein Branch pixels
Field-of-view-clipped neighbourhood

**C** Component-to-family aggregation

| Measurement | Component | HE family mean |
|---|---|---|
| Vessel distance / $D$ | 0.00234 | 0.02877 |
| Artery distance / $D$ | 0.01204 | 0.05076 |
| Vein distance / $D$ | 0.01151 | 0.05095 |
| Branch distance / $D$ | 0.01465 | 0.07242 |
| Skeleton overlap | 0.00000 | 0.00000 |
| Local skeleton occupancy | 0.02426 | 0.00809 |

HE components: 89; selected component: 25. Equivalent field diameter $D$ = 855.39 pixels; neighbourhood radius $r$ = 26 pixels. Family mean: equal component weights. Zero distance requires pixel overlap.
APTOS ee02294cc3d9, reference grade 2. Stored predictions; no lesions or vessels added.

**Fig. 2 Measurement of local lesion–vascular context.** (A) Registered APTOS photograph with predicted hard-exudate components. (B) One component, its field-of-view-clipped neighbourhood and the nearest vascular references. (C) Six component measurements and equal-weight family means over 89 components. Distances are normalised by the equivalent field diameter. Zero distance requires shared lesion/reference pixels; overlap and neighbourhood occupancy use vessel skeletons, not filled-vessel area or perfusion. The panels illustrate one fixed test image; predictions are not expert annotations.

all features when fewer were available. Ties favoured fewer features, then stronger regularisation. Independent heads need not be ordered, so inference used $\widetilde{u}_{ik} = \min_{j \le k} u_{ij}$ and

$$\begin{aligned} p_{i0} &= 1 - \widetilde{u}_{i0}, \qquad p_{i4} = \widetilde{u}_{i3}, \\ p_{ik} &= \widetilde{u}_{i,k-1} - \widetilde{u}_{ik}, \qquad k = 1, 2, 3. \end{aligned} \tag{8}$$

The non-negative probabilities sum to one; the largest determined grade. This is an inference correction, not jointly constrained training. Separate coefficients permit the same predictor to have different associations with successive grade transitions. E0–E2 assess individual streams; E3 combines lesions and biomarkers, E4 adds morphology, E5 adds geometry and E6 adds products. Adjacent comparisons retain the classifier but may select different feature subsets within their validation-selected budgets.

E7 represented predicted lesions and vascular segments as graph nodes. At 256-pixel resolution, branch removal partitioned the skeleton; size selection retained at most 48 segments and 48 lesions. Segment edges encoded adjacency; each lesion connected to its three nearest segment centroids. Eleven node attributes described anatomy, lesion identity and confidence; edge attributes encoded distance and relation type. Three edge-aware layers of width 64 used residual updates and attention pooling. A separate biomarker pathway (widths 64/32) joined the pooled graph representation before five-class softmax classification. Weighted cross-entropy and AdamW used learning rate $3 \times 10^{-4}$, decay $10^{-3}$, batch size 96 and at most 40 epochs with validation-QWK early stopping. Predictions averaged three seeds. The graph's centroid-based connections differ from the whole-component minimum distances in $X_{12}$; they are not distances along vascular paths. E7 changes representation, objective and averaging, rather than isolating topology (Supplementary Sections S6 and S9).

### Frozen-image comparisons

EfficientNet-B4, ResNet-152, ViT-B/16, DINO ViT-B/16, DINOv2 ViT-B/14 and RETFound MAE supplied fixed numerical image-feature vectors (embeddings) of 1,792/2,048/768/768/768/1,024 dimensions [7, 18–22]. For each embedding $h$, six configurations were evaluated:

$$\begin{aligned} B0 &: h, & B1 &: [h, X_2], \\ B2 &: [h, X_1, X_3], & B3 &: [h, X_1, X_2, X_3], \\ B4 &: [B3, X_{12}], & B5 &: [B4, X_{23}]. \end{aligned} \tag{9}$$

Brackets denote concatenation; B1 and B2 are separate additions to B0. All 36 probes used training-fitted preprocessing and the same ordinal classifier/search grid. Encoders remained in evaluation mode with fixed pretrained weights; only downstream graders learned APTOS labels. Convolutional encoders used global pooling, whereas transformer representations used class-token or recorded pooled outputs. RETFound preprocessing used bicubic resizing to 256 pixels,

centre cropping to 224 and ImageNet normalisation. Encoder-specific preprocessing remained fixed, and the validation-selected grading model retained its training fit at test evaluation. Model-specific transforms and checkpoint specifications are in Supplementary Section S8.

### Conditional vascular analysis and statistics

Four global lesion counts and 66 vascular biomarkers formed 264 pairs. Within training and test separately, variables were ranked and their grade-specific mean ranks removed. Residual-rank correlation estimated grade-adjusted association. Missing biomarkers were omitted pairwise. Significance used 9,999 within-grade permutations and Benjamini–Hochberg (BH) correction [23] across all pairs in each partition. Support required concordant signs and $q < 0.05$ in both; $q < 0.005$ was additionally reported. Training effects selected illustrative pairs, allowing up to two distinct vascular families per lesion. The full 264-pair correction was retained regardless of which pairs were illustrated; the selected display did not replace the family-wide analysis.

For each observed biomarker $m_{ik}$, ridge regression [24] compared five reference-grade indicators $g(Y_i)$ alone with their combination with four log-counts $c_i = \log(1 + n_i)$:

$$\begin{aligned} \widehat{m}_{ik}^{(0)} &= a_k + u_k^\top g(Y_i), \\ \widehat{m}_{ik}^{(1)} &= a'_k + v_k^\top g(Y_i) + w_k^\top c_i. \end{aligned} \tag{10}$$

Observed training rows fitted predictor/outcome scaling. Validation mean squared error selected each penalty from $\{0, 0.1, 1, 10, 100, 1000\}$ without refitting. Outcomes were never imputed. Added information was $\Delta R^2 = R_1^2 - R_0^2$; root mean squared error was also reported. Support required positive $\Delta R^2$, a paired 95% confidence interval (CI) above zero and BH $q < 0.05$ across 66 outcomes (Supplementary Section S10).

Quadratic weighted kappa (QWK), penalising larger grade discrepancies more strongly, was the primary grading metric. Macro one-versus-rest area under the receiver operating characteristic curve (AUROC), macro-F1, balanced accuracy, expected-grade error, classwise sensitivity/specificity and ten-bin expected calibration error (ECE) provided complementary assessment. F1 combines precision and sensitivity; ECE compares confidence with observed accuracy [25]. Macro averages give each grade equal weight. All comparisons used the same 575 test images, 2,000 paired image-bootstrap draws and prediction randomisation. BH correction used separate families: E3–E2/E5–E4/E6–E5/E7–E6, and post-hoc E5–E2/E6–E2/E7–E2. Conditional prediction used 2,000 bootstrap and 9,999 prediction-swap draws. Intervals describe test-image uncertainty, not training-run variation. Exploratory Spearman and Kruskal–Wallis analyses separately assessed 24 spatial descriptors, 90 lesion-count/vascular fields and 1,584 product features. These correction families were not pooled with the 264 grade-adjusted association tests. Supplementary Sections S6–S7 distinguish feature discovery, illustrative-case selection and the exploratory comparison families.

## RESULTS

### Anatomical evidence and segmentation quality

Figure 2 traces a hard-exudate measurement to the photograph: the selected component's normalised vessel distance was 0.00234, compared with a family mean of 0.02877 across 89 components. FOV diameter was 855.39 pixels and neighbourhood radius 26 pixels. Figure 3 retains errors and false-positive predictions across five reference grades; case selection used training-defined descriptors, not prediction correctness. A four-lesion worked example appears in Supplementary Fig. S2.

Ensemble macro pixel Dice was 0.501 on DDR validation and 0.332 on testing; the strongest individual member achieved validation Dice 0.506. Test Dice was 0.243/0.387/0.422/0.275 for MA/haemorrhage/HE/CWS. Precision across the four channels was 0.238/0.708/0.743/0.194, and sensitivity was 0.248/0.266/0.295/0.471. Thus the higher precision of haemorrhage and HE predictions did not imply high detection sensitivity. Segmentation estimates and failures remain in Supplementary Table S3 and Supplementary Fig. S3; APTOS overlays are predictions, not manual annotations.

### Grading from measured and image representations

Lesions were the strongest individual measured stream (Table 1A). E2 achieved QWK/AUROC 0.764/0.886, compared with 0.632/0.799 for biomarkers and 0.528/0.747 for morphology. Combining lesions and biomarkers gave 0.788/0.898. Adding morphology, geometry and products yielded QWK 0.789, 0.798 and 0.827 in E4–E6. AUROC was not monotonically improved: it fell to 0.883/0.879 in E4/E5 before reaching 0.905 in E6.

E6 exceeded E2 by 0.063 QWK (95% CI 0.028–0.099); E7 achieved QWK/AUROC 0.849/0.924, exceeding E2 by 0.085 (0.050–0.119). Both had BH $q < 0.001$. E5–E2 remained uncertain (0.034; $-0.008$–0.071; $q = 0.086$). E5–E4, E6–E5 and E7–E6 intervals also crossed zero (adjusted $q = 0.512$, 0.193 and 0.326). The E3–E2 biomarker addition likewise did not retain significance after its four-contrast correction ($q = 0.182$). Table 1C retains the primary contrasts; Supplementary Fig. S4 includes nonsignificant comparisons.

E7 improved grade-3 sensitivity relative to E6 (0.280 to 0.680), but reduced grade-4 sensitivity (0.605 to 0.316). E6 retained higher macro-F1 (0.589 versus 0.549) and accuracy (0.777 versus 0.730); E7 had lower ECE (0.036 versus 0.076). Confusion matrices and classwise results are in Supplementary Fig. S5 and Supplementary Table S4.

Table 1B reports all 36 frozen-encoder outcomes. Adding lesions increased numerical QWK for every encoder: EfficientNet-B4 rose from 0.831 to 0.841 and RETFound from 0.818 to 0.830. AUROC increased for five; RETFound decreased slightly from 0.8824 to 0.8816 before rounding. ViT-B/16 AUROC increased from 0.877 to 0.903. Broader context was encoder-dependent: EfficientNet-B4 peaked at B3 (QWK 0.849), while RETFound reached 0.848 in B4/B5. B4 and B5 reported identical QWK/AUROC for all encoders. These are numerical comparisons, not evidence of universal benefit or superiority over fine-tuning.

### Vascular variation beyond reference grade

Of 264 adjusted pairs, 139 met $q < 0.05$ in training and 108 in test. Concordant support retained 103: 28 MA, 27 haemorrhage, 21 HE and 27 CWS associations; 85 also met $q < 0.005$. The stricter threshold retained a subset of the same association analysis, rather than a separate discovery cohort. Training-selected test examples included MA–zone-B width ($\rho_{\mathrm{adj}} = 0.312$), haemorrhage–zone-C venular calibre (0.264), HE–zone-B venular calibre (0.265), and CWS–

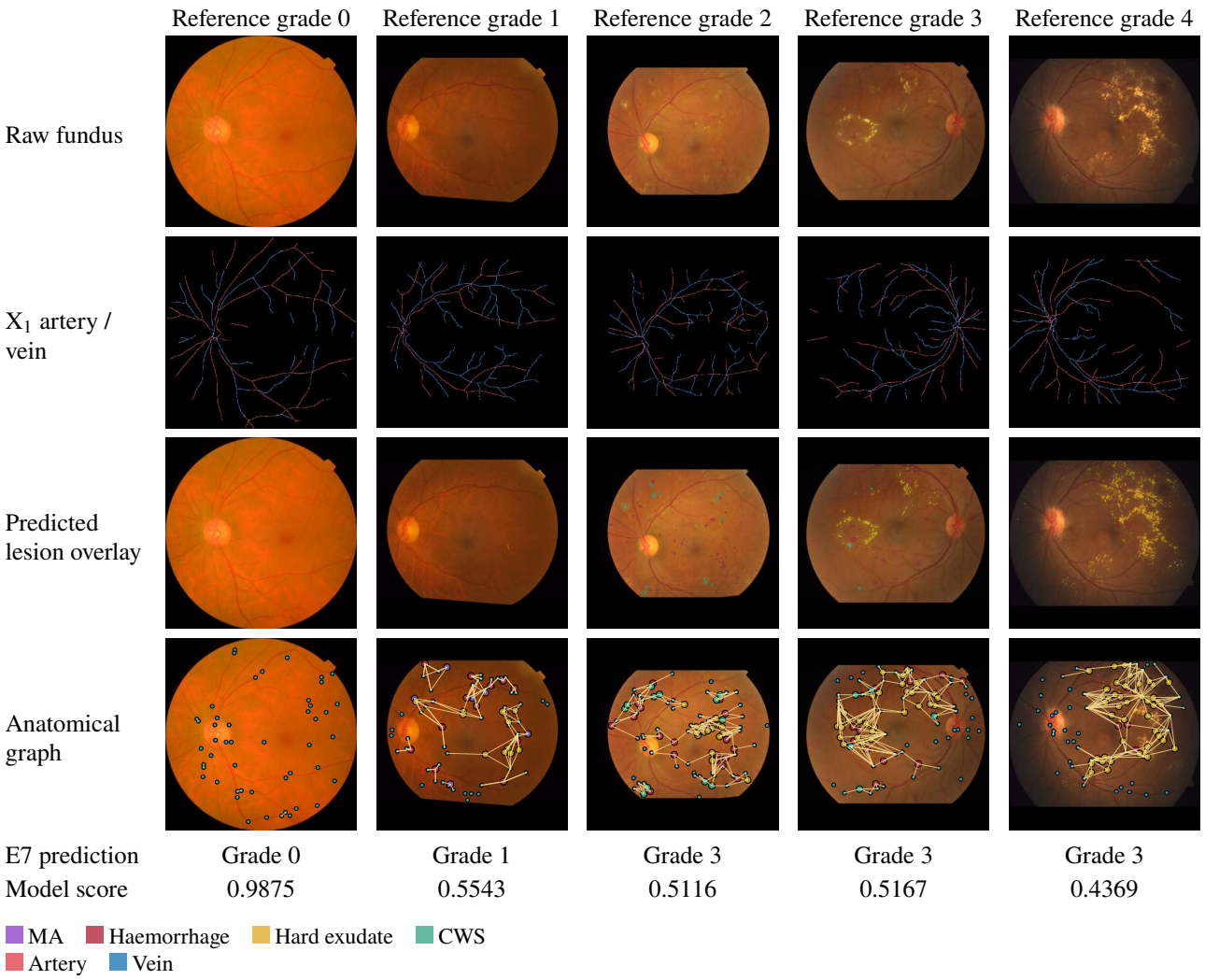



**Fig. 3 Image-derived evidence across the five reference grades.** Columns show one fixed APTOS test image per reference grade (0–4); rows show the photograph, registered artery/vein skeletons, predicted lesion overlay and anatomical graph. Cases were selected in a training-defined descriptor space, not by prediction correctness. Predicted E7 grades and scores are retained, including errors. Scores are model probabilities, not significance values. Arteries are red and veins blue; microaneurysms (MA) are purple, haemorrhages (Hem) red, hard exudates (HE) gold and cotton-wool spots (CWS) teal. Errors and false positives remain visible; cohort-level evaluation uses all 575 test images.

arterial density ($-0.203$).

Lesion counts gave positive prediction increments for 48/66 biomarkers. Twenty-six satisfied interval and BH criteria: 18/21 calibre/width, 6/9 fractal, 1/9 density and 1/27 tortuosity measurements (Figure 4). For zone-C central retinal venular equivalent using the Hubbard estimator, $R^2$ increased from 0.354 to 0.439 ($\Delta R^2 = 0.084$ from unrounded values; 95% CI 0.047–0.123; $q = 0.00165$; $n = 571$). Outcome-specific test sizes were 477–575. Gains concentrated in calibre/width and fractal measures; related regional estimates remain correlated (Supplementary Section S10).

## DISCUSSION

The study connects predicted retinal pathology to measurable vascular surroundings. Combined lesion–vascular representations improved lesion-only grading, with QWK increasing from 0.764 for lesion-only grading (E2) to 0.827 for the full structured representation (E6), while the anatomical graph (E7) reached 0.849 (Table 1A). Lesion burden carried vascular information within reference severity categories, with 103 supported grade-adjusted lesion–biomarker associations and prediction gains for 26 of 66 vascular measurements (Fig. 4). The contribution is a defined anatomical representation evaluated for predictive and phenotypic information, rather than a new clinical grading scale or evidence that every contextual addition improves performance.

The distinction between pathological burden and anatomical setting is important. Global counts and vascular summaries cannot reconstruct the local correspondence exposed by component-to-skeleton measurements. Haemorrhage eccentricity relative to the optic nerve has been examined in progression research [11]; our references are nearby vascular structures, and our endpoints are contemporaneous grades and vascular traits. A distance can be inspected on its source photograph, but does not identify the vessel of lesion origin, capillary non-perfusion or physiological connectivity. Equal component weighting also describes an average component's setting, not the proportion of total lesion area affected.

The grading comparisons support joint representations, not independent effects of geometry, products or graph topology. The E2-to-E6 QWK difference was 0.063 (95% CI 0.028–0.099; adjusted $q < 0.001$), supporting the combined repre-

**Table 1 Complete grading comparisons on the matched APTOS test set ($n = 575$).** (A) Structured ordinal and graph models, including discrimination, agreement and error measures. (B) All six frozen encoders under all six input configurations. (C) Paired lesion-baseline contrasts. Definitions and interpretation are given below the table.

**A. Structured evidence and anatomical graph**

| Model | Representation | QWK | AUROC | F1 | Bal. acc. | MAE | ECE | Acc. |
|---|---|---|---|---|---|---|---|---|
| E0 | $X_3$ | 0.632 | 0.799 | 0.351 | 0.418 | 0.686 | 0.137 | 0.603 |
| E1 | $X_1$ | 0.528 | 0.747 | 0.247 | 0.359 | 0.853 | 0.152 | 0.536 |
| E2 | $X_2$ | 0.764 | 0.886 | 0.448 | 0.514 | 0.499 | 0.127 | 0.677 |
| E3 | $X_2 + X_3$ | 0.788 | 0.898 | 0.502 | 0.544 | 0.448 | 0.089 | 0.722 |
| E4 | $X_1 + X_2 + X_3$ | 0.789 | 0.883 | 0.534 | 0.557 | 0.412 | 0.085 | 0.732 |
| E5 | E4 +$X_{12}$ | 0.798 | 0.879 | 0.539 | 0.557 | 0.388 | 0.105 | 0.741 |
| E6 | E5 +$X_{23}$ | 0.827 | 0.905 | **0.589** | **0.614** | **0.365** | 0.076 | **0.777** |
| E7 | Graph +$X_3$ | **0.849** | **0.924** | 0.549 | 0.603 | 0.368 | **0.036** | 0.730 |

**B. Frozen-image comparisons: QWK / macro AUROC**

| Encoder | B0 | B1 | B2 | B3 | B4 | B5 |
|---|---|---|---|---|---|---|
| EfficientNet-B4 | 0.831 / 0.913 | 0.841 / 0.914 | 0.838 / 0.908 | **0.849** / **0.920** | 0.847 / 0.909 | 0.847 / 0.909 |
| ResNet-152 | 0.829 / 0.871 | 0.833 / 0.888 | **0.838** / 0.885 | 0.826 / 0.907 | 0.830 / **0.916** | 0.830 / **0.916** |
| ViT-B/16 | 0.806 / 0.877 | 0.809 / **0.903** | 0.800 / 0.882 | 0.820 / 0.899 | **0.834** / **0.903** | **0.834** / **0.903** |
| DINO | 0.839 / 0.858 | 0.840 / 0.880 | **0.841** / 0.866 | 0.835 / **0.892** | 0.837 / 0.890 | 0.837 / 0.890 |
| DINOv2 | 0.826 / 0.866 | 0.830 / 0.874 | 0.829 / 0.892 | 0.831 / **0.895** | **0.835** / 0.877 | **0.835** / 0.877 |
| RETFound | 0.818 / 0.882 | 0.830 / 0.882 | 0.809 / 0.868 | 0.846 / 0.882 | **0.848** / **0.891** | **0.848** / **0.891** |

**C. Paired comparisons against lesion-only E2**

| Comparison | Representation | ΔQWK | 95% CI | BH-adjusted $q$ |
|---|---|---|---|---|
| E5–E2 | Three streams and geometry | 0.034 | $[-0.008, 0.071]$ | 0.086 |
| E6–E2 | Three streams, geometry and products | 0.063 | $[0.028, 0.099]$ | $< 0.001$ |
| E7–E2 | Anatomical graph and biomarkers | 0.085 | $[0.050, 0.119]$ | $< 0.001$ |

$X_1$: vascular morphology; $X_2$: lesion evidence; $X_3$: vascular biomarkers; $X_{12}$: spatial descriptors; $X_{23}$: lesion–biomarker products. Plus signs mean concatenation. E0–E6 use the same ordinal classification protocol; E7 averages three graph models. AUROC and F1 are macro averages. Bal. acc.: balanced accuracy; MAE: expected-grade mean absolute error; ECE: ten-bin expected calibration error; Acc.: accuracy. Bold marks the best displayed value (lower for MAE/ECE), including ties. B0: image only; B1: image + $X_2$; B2: image + $X_1/X_3$; B3: image + all three; B4: B3 + $X_{12}$; B5: B4 + $X_{23}$. Panel C uses paired test-image bootstrap intervals and a separate three-contrast BH family. These are fixed-test scores, not cross-validation means or independent repetitions.

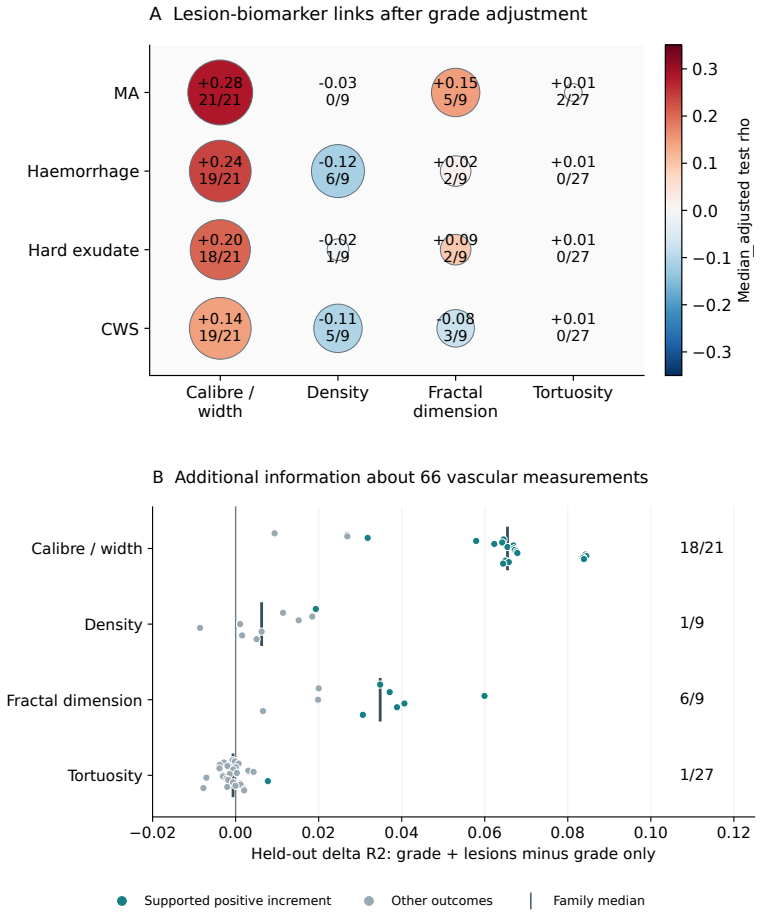
sentation rather than any individual contextual addition (Table 1A). E7 changes the model as well as the input structure; its size-limited graph and centroid edges approximate spatial relations rather than reconstruct a complete vascular tree. Frozen-encoder comparisons provide a further qualification: explicit lesions consistently increased numerical QWK, whereas broader context depended on the embedding and selected feature budget. Lesion-aware systems and retinal pretraining already provide strong representations [6, 7], and these probes do not establish superiority over optimised end-to-end models.

Clinical interpretation requires examining errors, not only aggregate agreement. E7's higher overall QWK (Table 1A) coexisted with reduced PDR sensitivity, 0.316 compared with 0.605 for E6 (Supplementary Table S4; Supplementary Fig. S5). The four supervised lesion families do not cover all proliferative findings, particularly neovascularisation. Modest segmentation Dice and low sensitivity in some channels propagate uncertainty into counts, geometry and graph nodes. Anatomical traceability explains what a measurement represents, but not whether the predicted lesion is correct or the classifier relies on appropriate evidence. Saliency or attention is likewise not sufficient validation of explanation fidelity [26, 27].

The two evaluation tasks also clarify what the measurements contribute. Grading asks whether vascular context helps distinguish ordered severity labels. Conditional phenotyping instead holds the reference label in the model and asks what lesion burden explains about vascular measurements beyond it. Here, 103 lesion–biomarker associations remained supported after grade adjustment, and lesion counts improved prediction for 26 of 66 vascular measurements (Fig. 3). Improvement in the latter cannot be interpreted as improved diagnosis, because the grade is already known. Keeping these questions separate avoids using a grading association as evidence of a biological pathway.

The phenotyping analysis is distinct from prognosis. Previous longitudinal work links vascular geometry and automated morphology to subsequent DR or treatment [12, 13]. Here, associations condition on a known contemporaneous grade. Prediction gains were concentrated in calibre/width and fractal measurements, supporting targeted retinal phenotyping without implying a generalised damage score. Grade adjustment removes between-grade mean-rank differences, but cannot remove all confounding, lesion co-occurrence, regional correlation or shared image-processing error. Outcome-linked oculomics would require independently measured clinical or molecular endpoints [28].

Limitations include a single matched grading cohort, a historical split inspection, incomplete exclusion provenance, and unavailable patient linkage. Image-level bootstrap intervals cannot account for unidentified dependence between




A: Bubble area represents the supported proportion; labels give the median adjusted correlation and supported/tested pairs. B: Each point is one vascular measurement. Related regional measurements are not independent mechanisms.

**Fig. 4 Vascular variation beyond reference DR grade.** (A) Grade-adjusted lesion–biomarker associations across 264 pairs. Bubble area is the proportion supported in both training and test with concordant signs and Benjamini–Hochberg $q < 0.05$; colour and labels give median adjusted test correlation and supported/tested counts. (B) All 66 changes in test $R^2$ when four lesion counts are added to reference-grade predictors. Teal denotes positive increments with paired 95% bootstrap intervals above zero and adjusted $q < 0.05$; ticks mark family medians. Outcome-specific test sizes are 477–575 images. Results come from one fixed cohort split; related regional measurements remain correlated and supported counts do not represent independent biological mechanisms.

eyes. DDR provides separate segmentation supervision, not external validation of grading. Acquisition calibration, age and image quality may influence vascular associations. Lesion-relocation controls, edge ablations, matched model capacity/objectives and expert APTOS annotation remain incomplete. Robustness to thresholds, registration error and aggregation choices requires evaluation.

In conclusion, measured lesion–vascular context provides an inspectable connection between local pathology, grading and retinal phenotype. Combined representations improved lesion-only grading, without isolating significant individual contextual effects. Independent anatomical validation and external, outcome-linked studies are needed before clinical, prognostic or systemic interpretation.

## PLAIN LANGUAGE SUMMARY

Diabetic retinopathy damages the retina, the light-sensitive tissue at the back of the eye. A disease grade summarises its severity, but does not describe precisely where each visible abnormality lies relative to nearby blood vessels. We used existing retinal photographs to measure these relationships and examined whether they helped computer-based grading. Combining predicted lesions with vascular measurements improved agreement with the reference grades. Lesion counts also provided information about blood-vessel measurements among images with the same grade. However, the lesion detector made errors, and the model with the highest overall agreement missed many images labelled as proliferative disease. These findings support further study of measurable retinal anatomy; they do not establish a replacement for clinical assessment or a method for predicting future sight loss.

## SUMMARY

### What was known before

- Lesion burden and retinal vascular morphology provide complementary information about diabetic retinopathy; lesion location and vascular geometry have also been investigated in relation to progression.
- A severity grade or global vascular measurement does not quantify the local relationship between an individual lesion and surrounding vessels.

### What this study adds

- Registered lesion and vascular predictions provide 24 inspectable spatial descriptors. Combined structured and graph-based representations improved lesion-only grading, without separately establishing the contribution of individ-

ual contextual additions.

- Lesion counts were associated with vascular variation beyond reference grade, principally calibre/width and fractal measurements. Segmentation errors and limited proliferative-disease sensitivity constrain clinical interpretation.

## ETHICS APPROVAL AND CONSENT

This study involved secondary analysis of previously released retinal photographs and annotations from the APTOS and DDR datasets. No new participants were recruited, and no new data were collected. The study did not involve direct participant contact or collection of identifiable information.

## DATA AND CODE AVAILABILITY

APTOS and DDR are publicly accessible third-party datasets available through their respective providers [14, 15]. The retinal images and associated annotations can be accessed through the dataset sources given in these references, subject to the providers' applicable access and reuse conditions. The Supplementary Information provides the measurement and feature definitions, analytical procedures, model specifications, segmentation and grading diagnostics, and additional results supporting the analyses. The study code is not publicly available at the time of submission but will be made available following acceptance in accordance with the journal's applicable data and code availability requirements.

## AUTHOR CONTRIBUTIONS

Inam Ullah conceived the study, developed the methodology, performed the analyses, interpreted the results, integrated the relevant literature, and drafted the manuscript. Shoaib Jameel supervised the study, provided methodological guidance and research oversight, and critically reviewed and revised the manuscript. Both authors read and approved the final manuscript.

## COMPETING INTERESTS

The authors declare that they have no competing interests.

## FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## USE OF AI-ASSISTED TOOLS

AI-assisted tools were used for language editing, manuscript preparation and limited code-related support. They were not used to generate the study data or reported experimental results. All content was reviewed and verified by the authors, who take responsibility for the final manuscript.

## SUPPLEMENTARY INFORMATION

Supplementary information is appended to this manuscript; all links open within this PDF. It contains Supplementary Algorithm S1 and Supplementary Algorithm S2, measurement and model specifications, segmentation and grading diagnostics, a four-lesion worked example, and complete grade-adjusted phenotyping analyses (Supplementary Section S10). All 36 frozen-encoder outcomes are in main Table 1B rather than repeated in the supplement.

## REFERENCES

[1] NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in diabetes prevalence and treatment from 1990 to 2022: a pooled analysis of 1108 population-representative studies with 141 million participants. Lancet. 2024;404(10467):2077-2093.

[2] Lim JI, Kim SJ, Bailey ST, Kovach JL, Vemulakonda GA, Ying GS, et al. Diabetic Retinopathy Preferred Practice Pattern. Ophthalmology. 2025;132(4):P75-P162.

[3] Wilkinson CP, Ferris FL, Klein RE, Lee PP, Agardh CD, Davis M, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology. 2003;110(9):1677-1682.

[4] Lee SC, Lee ET, Wang Y, Klein R, Kingsley RM, Warn A. Computer Classification of Nonproliferative Diabetic Retinopathy. Arch Ophthalmol. 2005;123(6):759-764.

[5] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402-2410.

[6] Dai L, Wu L, Li H, Cai C, Wu Q, Kong H, et al. A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nat Commun. 2021;12:3242.

[7] 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.

[8] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 618-626.

[9] Xu Q, Luo X, Lin Y, Liu C, Hu Y, Li J, et al. Vision-Language Models Guided Graph Concept Reasoning for Interpretable Diabetic Retinopathy Diagnosis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40; 2026. p. 27314-27322.

[10] Lepetit-Aimon G, Playout C, Boucher MC, Duval R, Brent MH, Cheriet F. MAPLES-DR: MESSIDOR anatomical and pathological labels for explainable screening of diabetic retinopathy. Sci Data. 2024;11:914.

[11] Verma A, Nittala MG, Mansoori Dara R, Facktor M, Ramachandra CA, Bhaskaranand M, et al. Predicting progression to proliferative diabetic retinopathy using automated versus manual quantification of retinal haemorrhages. Eye (Lond). 2026;40:682-688.

[12] Cheung CYL, Sabanayagam C, Law AKP, Kumari N, Ting DSW, Tan G, et al. Retinal vascular geometry and 6 year incidence and progression of diabetic retinopathy. Diabetologia. 2017;60(9):1770-1781.

[13] Nderitu P, Chia M, Karamchandani U, Kiire CA, Diafas A, Matta S, et al. Deep learning-derived retinal vascular morphology and incident treatment for sight-threatening diabetic retinopathy. Ophthalmol Sci. 2026;6(11):101385.

[14] Asia Pacific Tele-Ophthalmology Society. APTOS 2019 blindness detection; 2019. Kaggle competition dataset.

[15] Li T, Gao Y, Wang K, Guo S, Liu H, Kang H. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Inf Sci. 2019;501:511-522.

[16] 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. Transl Vis Sci Technol. 2022;11(7):12.

[17] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Medical Image

Computing and Computer-Assisted Intervention; 2015. p. 234-241.

[18] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 770-778.

[19] Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. vol. 97 of Proceedings of Machine Learning Research; 2019. p. 6105-6114.

[20] 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. In: International Conference on Learning Representations; 2021.

[21] Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, et al. Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 9650-9660.

[22] Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. DINOv2: learning robust visual features without supervision. Trans Mach Learn Res. 2024.

[23] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Methodol. 1995;57(1):289-300.

[24] Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55-67.

[25] Guo C, Pleiss G, Sun Y, Weinberger KQ. On Calibration of Modern Neural Networks. In: International Conference on Machine Learning. vol. 70 of Proceedings of Machine Learning Research; 2017. p. 1321-1330.

[26] Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. In: Advances in Neural Information Processing Systems. vol. 31; 2018.

[27] Jain S, Wallace BC. Attention Is Not Explanation. In: Proceedings of NAACL-HLT; 2019. p. 3543-3556.

[28] Wagner SK, Fu DJ, Faes L, Liu X, Huemer J, Khalid H, et al. Insights into Systemic Disease through Retinal Imaging-Based Oculomics. Transl Vis Sci Technol. 2020;9(2):6.

# Supplementary information

Measuring lesion–vascular context for diabetic retinopathy grading and retinal phenotyping

Back to the main article

This supplement provides reproducibility details and evidence not repeated in the main article. Sections S1–S4 describe the datasets, measured features and two extraction algorithms; S5–S9 cover segmentation, grading, feature screening and model specifications; S10 reports the complete grade-adjusted phenotyping analyses. Main Table 1 contains all E0–E7 results and all 36 frozen-encoder comparisons.

**Contents**

S1 Study Design and Dataset Provenance ..... S1
S2 Feature Accounting ..... S2
S3 Reproducible Algorithms ..... S2
S4 Operational Definition of Lesion–Vascular Context ..... S3
S5 Segmentation Protocol ..... S6
S6 Matched Grading Evaluation ..... S6
S7 Feature Screening and Illustrative-Case Selection ..... S8
S8 Frozen Encoder Comparisons ..... S10
S9 Graph Integrity and Remaining Controls ..... S10
S10 Full Grade-Adjusted Lesion–Biomarker Analysis ..... S10

**Key destinations:** Algorithm S1: spatial extraction; Algorithm S2: product construction; Table S3: segmentation; Table S4: classwise performance; Fig. S5: confusion matrices and ROC curves; Section S10: complete phenotyping results. *References cited below use the main article's bibliography. All navigation stays inside this PDF.*

## S1 STUDY DESIGN AND DATASET PROVENANCE

| Dataset | Train | Val. | Test | Supervision |
|---|---|---|---|---|
| APTOS | 1,724 | 575 | 575 | Grades 0–4 |
| DDR | 383 | 149 | 225 | Lesion masks |

**Table S1** Distinct dataset roles. APTOS is the grading cohort; DDR supplies lesion supervision. The grading cohort is matched by image identity across all evidence streams.

This is a cross-sectional secondary analysis of previously released fundus-image datasets. No new participants were recruited and no longitudinal follow-up was analysed. Lesion–vascular relationships were measured within images and evaluated against contemporaneous DR grades or vascular measurements; they are not estimates of disease progression.

DDR expert masks [15] train and evaluate lesion segmentation; APTOS image grades [14] train and evaluate DR grading. The canonical manifests retain original identifiers, partitions and image hashes, linking the source image, anatomical maps, lesion predictions and biomarker record throughout the analysis.

The released DDR segmentation partitions contain 383/149/225 images. Four mask folders are resolved into semantic names: MA to microaneurysm, HE to haemorrhage, EX to hard exudate, and SE to soft exudate/CWS. The segmentation model uses four sigmoid channels, not mutually exclusive softmax labels. Figure S3 compares real expert annotations with the corresponding ensemble predictions.

APTOS originates from the 2019 Kaggle competition released by the Asia Pacific Tele-Ophthalmology Society. All 2,874 matched study identifiers and grades agree with the original 3,662-row labelled release. The difference of 788 is reported without assigning unverified image-level exclusion causes. Grade counts for the study partitions are:

| Partition | G0 | G1 | G2 | G3 | G4 |
|---|---|---|---|---|---|
| Train | 889 | 182 | 465 | 76 | 112 |
| Validation | 296 | 61 | 155 | 26 | 37 |
| Test | 296 | 61 | 155 | 25 | 38 |

Grading observations are distinct, non-augmented images. DDR segmentation training uses sampled patches and flips. The APTOS image partitions are fixed; historical split inspection remains an acknowledged limitation of the exploratory evaluation.

Patient-level linkage is unavailable. These conditions limit confirmatory interpretation and do not establish patient-independent external replication.

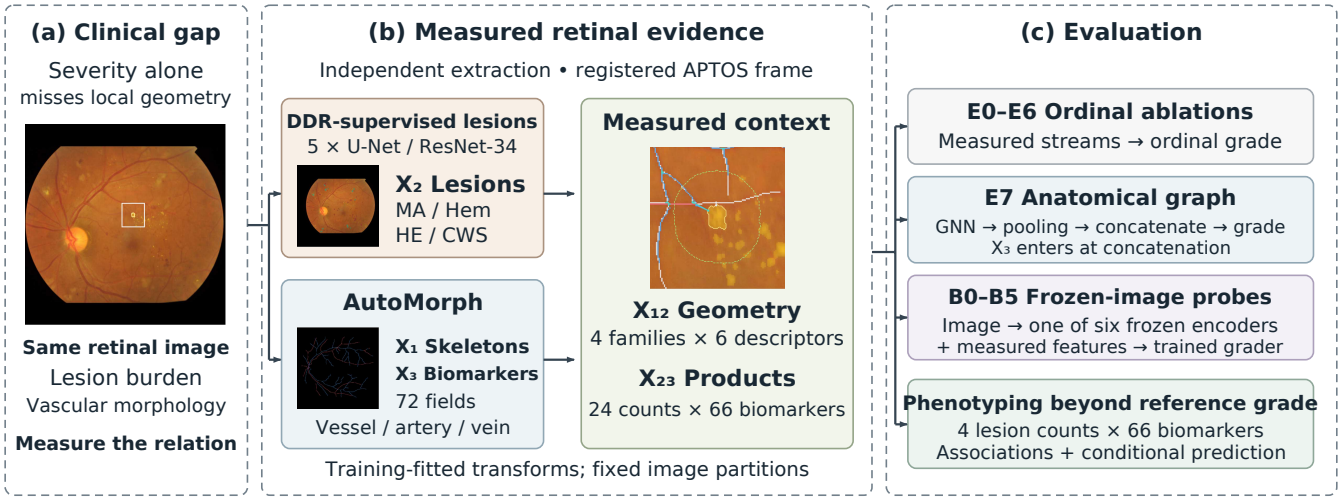



**Fig. S1** Study overview. DDR-supervised lesion predictions and AutoMorph anatomy are aligned for the same APTOS photograph. Spatial measurements and image-level products complement lesion burden and vascular measurements. Structured models, an anatomical graph and six frozen image encoders evaluate grading; separate conditional analyses examine vascular variation beyond reference grade. Insets show actual archived images and analysis outputs. Arrows indicate processing, not causality.

## S2 FEATURE ACCOUNTING

The AutoMorph processing sequence has four functions: preparing the retinal field, assessing whether image quality supports analysis, delineating anatomical structures, and measuring vascular morphology. Quality categories indicate image usability, not DR grades. Anatomical maps and skeleton summaries supply X1; the tabular global and optic-disc-referenced zone-B/C measurements supply X3. The DDR-trained ensemble independently supplies X2 for the same APTOS identifiers. The available manifest establishes a matched cohort of 2,874, but does not reconstruct a quality/failure reason for every omitted source image. No new exclusion causes are inferred from the stage names. Zones B and C are referenced to the optic disc, not the fovea. CRAE and CRVE denote central retinal arteriolar and venular equivalents, summary calibre estimates; AVR is the corresponding artery-to-vein ratio. Tortuosity describes vessel winding, while fractal dimension summarizes network complexity. These terms name vascular measurements, not additional lesion classes [16].

X1 contains 16 vascular summaries; X2 contains 36 active probability/count descriptors; X3 contains 72 biomarker fields. X12 has 24 measured descriptor types: six per lesion family. A wider archived schema contains eight entirely missing local-calibre/foveal placeholders. Those placeholders are not reported as measured anatomy and provide no observed spatial signal. Local tortuosity and verified foveal position are also unavailable.

A quadrant is one of four image regions used to summarize lesion distribution. The active X2 grading set contains probability summaries and predicted component/image/quadrant counts; area fractions are not active predictors. X23 multiplies 24 count fields by 66 vascular fields after training-fitted standardization, producing 1,584 correlated predictors. Redundant counts and related biomarkers are therefore interpreted jointly rather than as independent lesion mechanisms.

Missing distances when no lesion or reference structure is present are retained as missing until training-fitted imputation. Biomarker invalid-negative/non-finite handling also precedes train-fitted median imputation and missingness encoding. Neither held-out pixel labels nor APTOS test grades fit these transforms. Zones B/C are optic-disc-centred; the source filename does not establish foveal localisation. Width measurements are not uniformly converted to physical micrometres without verified acquisition calibration.

## S3 REPRODUCIBLE ALGORITHMS

Algorithms S1 and S2 expand spatial extraction and X23 construction. Both preserve missingness, feature order and the training-only fitting boundary. Finite means exclude unavailable component measurements; an empty set remains missing. Reference grades determine neither lesion locations nor product weights. X23 uses the cleaned input schema; the downstream grading transforms are fitted separately on training rows.

**Algorithm S1** Detailed lesion–vascular spatial descriptors (X12)

**Require:** Registered probabilities $P_c$ for four families; FOV $F$; skeleton products $V, A, W$
**Ensure:** Twenty-four ordered spatial descriptors, preserving unavailable measurements

Clip $V, A, W$ to $F$; $S \leftarrow \mathrm{Skeletonize}(V)$
$B \leftarrow \{p \in S : \deg_8(p) \geq 3\}$; $D \leftarrow \sqrt{4|F|/\pi}$
$r \leftarrow \max(1, \mathrm{round}(0.03D))$; initialize $X_{12}[4, 6]$ as missing
**for** each lesion family $c$ in the fixed channel order **do**
  $M \leftarrow (P_c \geq 0.5) \cap F$; $\mathcal{L} \leftarrow \mathrm{CC}_8(M)$
  **if** $\mathcal{L}$ is empty **then**
    **continue**
  **end if**
  **for** each component $\ell \in \mathcal{L}$ **do**
    **for** each reference $T \in \{V, A, W, B\}$ **do**
      **if** $T$ is empty **then**
        $d_T \leftarrow$ missing
      **else**
        $d_T \leftarrow \min_{p \in \ell, t \in T} \|p - t\|_2 / D$
      **end if**
    **end for**
    $N \leftarrow \{p \in F : \mathrm{dist}(p, \ell) \leq r\}$
    $o \leftarrow |\ell \cap S|/|\ell|$; $\rho \leftarrow |N \cap S|/|N|$
    $\phi(\ell) \leftarrow [d_V, d_A, d_W, d_B, o, \rho]$
  **end for**
  **for** each descriptor $k \in \{1, \ldots, 6\}$ **do**
    $X_{12}[c, k] \leftarrow$ unweighted mean of finite $\phi_k(\ell)$; missing if none are finite
  **end for**
**end for**
**return** $\mathrm{vec}(X_{12})$ in the recorded family/descriptor order

**Algorithm S2** Training-fitted lesion–biomarker products (X23)

**Require:** Count matrix $B \in \mathbb{R}^{n \times 24}$; vascular matrix $M \in \mathbb{R}^{n \times 66}$; fixed split indices
**Ensure:** $X_{23} \in \mathbb{R}^{n \times 1584}$ and frozen preprocessing transforms

Retain canonical count and vascular columns; exclude disc/cup fields
Mark invalid negative/non-finite biomarkers as missing using the upstream schema
Fit median imputers $I_b, I_m$ using training rows only
Fit standardizers $S_b, S_m$ to $I_b(B_{\mathrm{train}})$ and $I_m(M_{\mathrm{train}})$
Freeze $I_b, I_m, S_b, S_m$; do not refit on validation or test observations
**for** each partition $s \in \{\mathrm{train}, \mathrm{validation}, \mathrm{test}\}$ **do**
  $\tilde{B}_s \leftarrow S_b(I_b(B_s))$; $\tilde{M}_s \leftarrow S_m(I_m(M_s))$
  **for** each image $i$ in partition $s$ **do**
    **for** each count field $j = 1, \ldots, 24$ **do**
      **for** each vascular field $k = 1, \ldots, 66$ **do**
        $Q_i[j, k] \leftarrow \tilde{B}_s[i, j]\, \tilde{M}_s[i, k]$
      **end for**
    **end for**
    $X_{23}[i, :] \leftarrow \mathrm{vec}(Q_i)$, count-major with vascular fields varying fastest
  **end for**
**end for**
**return** $X_{23}$, ordered feature names and frozen transforms

## S4 OPERATIONAL DEFINITION OF LESION–VASCULAR CONTEXT

X12 measures relationships between two independently estimated structures in one image: DDR-supervised lesion predictions and AutoMorph vascular skeleton products. It is distinct from X2 probability/count summaries and X3 global/regional biomarkers. It does not infer a lesion's vessel of origin, diagnose local ischaemia, or identify clinical neovascularisation.

**Coordinate and identity contract.** The raw-image SHA256 identifies the lesion archive and registration record. The raw-to-crop affine transform incorporates the M0 border crop, square padding and resize to the $912 \times 912$ vessel frame. Rounded 8-bit probabilities are warped using bilinear interpolation, divided by 255 and thresholded at 0.5, matching archived grading-feature extraction. The M0 field of view is resized by nearest-neighbour interpolation; structures are clipped to it. Matching dimensions alone does not establish registration.

**Components and reference structures.** Lesions use eight-connectivity independently for MA, haemorrhage, hard exudate and CWS; touching predictions within one channel form one component, not necessarily one clinical lesion. Vessel, artery and vein inputs are skeleton products, not filled lumen masks. The binary product is skeletonized again for overlap, local occupancy and branch detection. Branch pixels have at least three occupied neighbours among eight surrounding pixels; adjacent branch pixels are not collapsed into one anatomical bifurcation. Vessel distances use the input product, whereas branch distances use the derived branch set.

| Descriptor per component | Exact operation | Meaning and scope |
|---|---|---|
| Vessel proximity | Minimum lesion-pixel to vessel-skeleton-pixel Euclidean distance, divided by $D_i$ | Shared lesion/reference pixels give zero; not vascular path length. |
| Artery proximity | Same minimum to artery skeleton | Proximity to predicted arterial structure, not proven lesion origin. |
| Vein proximity | Same minimum to vein skeleton | Independent of artery proximity; segmentation errors remain possible. |
| Branch proximity | Same minimum to derived branch pixels | Algorithmic branching structure, not expert bifurcation landmarks. |
| Skeleton overlap | Skeleton pixels inside the lesion divided by lesion pixels | Fraction of predicted lesion area intersecting the thin skeleton. |
| Local skeleton density | Skeleton pixels in the FOV-clipped lesion neighbourhood divided by neighbourhood pixels | Radius $\max(1, \mathrm{round}(0.03D_i))$; occupancy, not filled-vessel area or perfusion. |

**Table S2** Six implemented descriptors per lesion family. $D_i = \sqrt{4|F_i|/\pi}$ is the equivalent retinal-field diameter in pixels. Four families produce 24 measured descriptor types.

**Aggregation and missingness.** For descriptor $k$ and family $c$, let $\mathcal{L}_{ic}^{(k)}$ contain components with finite measurements. The image-level feature is

$$X_{12,ic,k} = \frac{\sum_{\ell \in \mathcal{L}_{ic}^{(k)}} \phi_k(\ell)}{|\mathcal{L}_{ic}^{(k)}|}. \tag{S1}$$

An empty denominator produces a missing value. Components have equal weight: a larger lesion does not dominate through its area. Each distance is the minimum within one component, followed by the mean across components, not the minimum over all lesions. An absent reference structure gives missing distances, while shared lesion/reference pixels legitimately give zero; adjacency without overlap is not zero. With lesions present but no skeleton, overlap and local occupancy can be zero. Entirely missing local-calibre/foveal fields are excluded from the observed descriptor count. Training-fitted preprocessing is unchanged.

**A checkable anatomical example.** The main paper's measurement plate traces a fixed, previously selected APTOS test case through aligned image maps, a lesion component, its neighbourhood and six numerical outputs. The largest hard-exudate component is highlighted only to explain the calculation; all components contribute to its family mean. The source analysis reports recomputation of family descriptors using the original implementation and comparison with the stored feature row before rendering. No mask is brightened, relocated or resized to improve its anatomical appearance. Component values are measurements, not image-specific significance tests. Main Figure 2 retains the complete five-grade anatomical panel.

**Four-lesion worked example.** Figure S2 extends main Figure 1 to all four predicted lesion families in the same fixed grade-2 image. The largest predicted component per family is highlighted using the original rule; family means use all finite component values with equal weights. The source analysis reports replay of the archived fields before rendering. Missing components remain missing, and a zero distance requires shared lesion/reference pixels. The five-grade anatomical comparison remains main Figure 2. These illustrations explain the calculation, not expert lesion confirmation, disease progression or a new grading experiment.

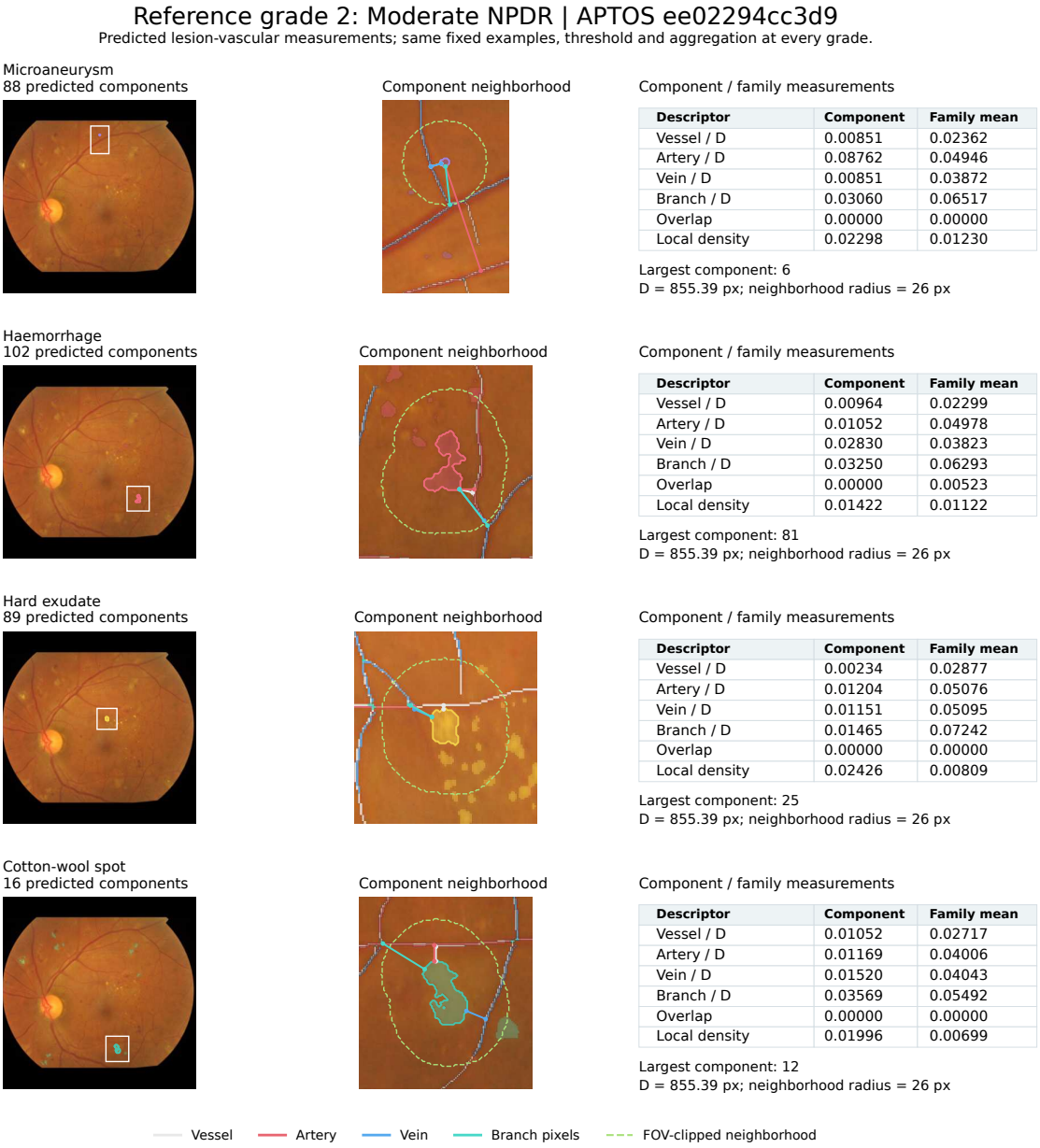



**Fig. S2** Four-lesion worked example from the fixed moderate-NPDR APTOS image. The hard-exudate row is the example in main Figure 1; the other rows show the same calculation for microaneurysms, haemorrhage and cotton-wool spots. Component selection and aggregation are unchanged. Predicted masks are not expert APTOS annotations.

## S5 SEGMENTATION PROTOCOL

| Lesion | Val. Dice | Test Dice | Precision | Sensitivity |
|---|---|---|---|---|
| MA | 0.372 | 0.243 | 0.238 | 0.248 |
| Haemorrhage | 0.509 | 0.387 | 0.708 | 0.266 |
| Hard exudate | 0.580 | 0.422 | 0.743 | 0.295 |
| CWS | 0.542 | 0.275 | 0.194 | 0.471 |

**Table S3** Segmentation quality of the five-member mean ensemble. Test values pool pixels over 225 DDR test images at fixed threshold 0.5. DDR is the only source of expert lesion masks used for model training.

Focal loss downweights well-classified pixels, while soft Dice loss measures overlap between predicted probabilities and lesion masks. The four sigmoid channels allow lesion families to overlap rather than forcing every pixel into a single class.

The five fixed segmentation seeds are 20261107–20261111. Members use U-Net with an ImageNet-initialized ResNet-34 encoder, lesion-balanced native-resolution 512-pixel crops, flips, focal loss with $\alpha = 0.75$, $\gamma = 2$, and soft Dice loss. Adam starts at $10^{-4}$ with cosine scheduling over ten epochs and gradient clipping at five. Recorded validation checks occur at epochs 1, 5 and 10.

The arithmetic mean of five sigmoid maps is thresholded at 0.5, fixed independently of the test masks. Validation macro pixel Dice for individual selected checkpoints is 0.506, 0.482, 0.407, 0.455 and 0.490; the ensemble is 0.501. DDR test macro pixel Dice is 0.332 and mean-image macro Dice is 0.291. The strongest individual member therefore slightly exceeds the ensemble on validation, and pixel-pooled and image-averaged metrics are reported separately.

**An expert-annotated failure example.** DDR case 007-3800-200 was the first image containing all four expert categories in a seed-20261111 random ordering of official test images. Selection used annotation coverage, not predictions or Dice. Native-resolution inference used 512-pixel tiles, 128-pixel overlap, equal sigmoid averaging and threshold 0.5. The selected-case Dice values were 0.029 for MA, 0.212 for haemorrhage, and zero for hard exudate and CWS. The source analysis reports matching the five checkpoint hashes to the ensemble. This deliberately retained failure illustrates error propagation; it does not replace the complete 225-image test metrics in Table S3.

APTOS examples in main Figure 2 use the stored ensemble predictions. Expert APTOS lesion masks are unavailable. Empty thresholded masks indicate no detected component, not an omitted output channel. Neither reference grade nor image-specific normalisation alters a mask. Fine lesions may require magnification.

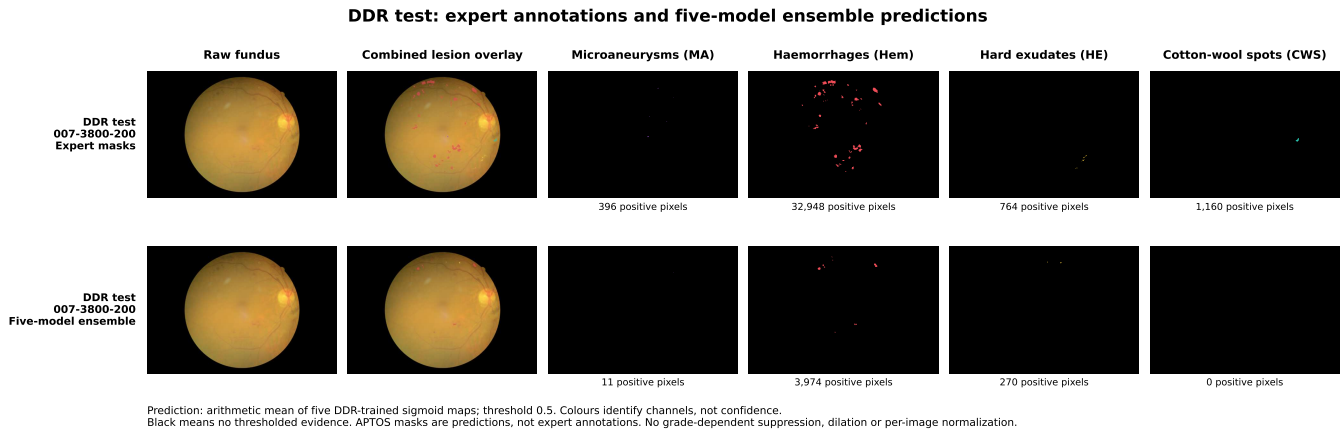



**Fig. S3** DDR test example: expert annotations (top) and five-member ensemble predictions (bottom). Columns show the photograph, combined overlay and separate MA, haemorrhage, hard-exudate and CWS masks. Counts denote native positive pixels. Missed CWS and inaccurate exudate localisation remain visible; colours identify channels rather than probability or clinical confirmation.

## S6 MATCHED GRADING EVALUATION

**Ordinal fitting and selection.** Each E0–E6 input is processed independently. Training medians impute missing continuous fields; missingness indicators record availability. Training-fitted standardization and ANOVA ranking against the five training grades precede four L2-regularized cumulative logistic heads. Heads use inverse-frequency binary class weights and the `liblinear` solver with at most 1,000 iterations. Validation QWK selects $C \in \{0.03, 0.1, 0.3, 1\}$ and $K \in \{64, 128, 256\}$; all available features are retained when fewer than $K$ exist. Exact ties favour smaller $K$, then smaller $C$. The selected training fit is retained for test evaluation. For selected feature vector $z_i$, head $k$ is trained with target $t_{ik} = \mathbb{1}[Y_i > k]$ and estimates

$$u_{ik} = \sigma(w_k^\top z_i + a_k), \qquad k = 0, 1, 2, 3. \tag{S2}$$

The independently fitted heads need not produce ordered cumulative probabilities. At inference, $\tilde{u}_{ik} = \min_{j \le k} u_{ij}$ gives non-increasing values, and

$$p_{ik} = \begin{cases} 1 - \tilde{u}_{i0}, & k = 0, \\ \tilde{u}_{i,k-1} - \tilde{u}_{ik}, & k = 1, 2, 3, \\ \tilde{u}_{i3}, & k = 4. \end{cases} \tag{S3}$$

The probabilities are nonnegative and sum to one; the largest defines the predicted grade. This is an inference-time correction, not a jointly constrained training objective.

**Statistical families.** The four model contrasts E3–E2, E5–E4, E6–E5 and E7–E6 form one BH family; the three post-hoc lesion-baseline contrasts E5–E2, E6–E2 and E7–E2 form another. These exploratory comparison families are not pooled. Paired intervals describe test-image sampling uncertainty, not training-seed or cross-validation variability. The separate feature-association families are 24 spatial descriptors, 90 lesion-burden/vascular fields and 1,584 products. They are distinct from the 264 grade-adjusted pair tests and 66 conditional prediction comparisons in Section S10.

All E0–E7 outcomes use the same 575 APTOS test images. Table S4 reports one-versus-rest sensitivity/specificity, including low-performing grades. Main Table 1A includes balanced accuracy, expected-grade absolute error, ECE and accuracy; those values are not repeated here. Expected-grade MAE uses the probability-weighted expected grade and is distinct from absolute argmax-grade error.

The main paper reports paired combined-context comparisons against lesion-only E2. Figure S4 supplies all four incremental component comparisons, including their uncertainty. These use the same 2,000 paired image-bootstrap draws and randomization protocol, with BH correction across four contrasts. Their intervals crossing zero qualify component attribution without changing the significant combined-context versus E2 results.

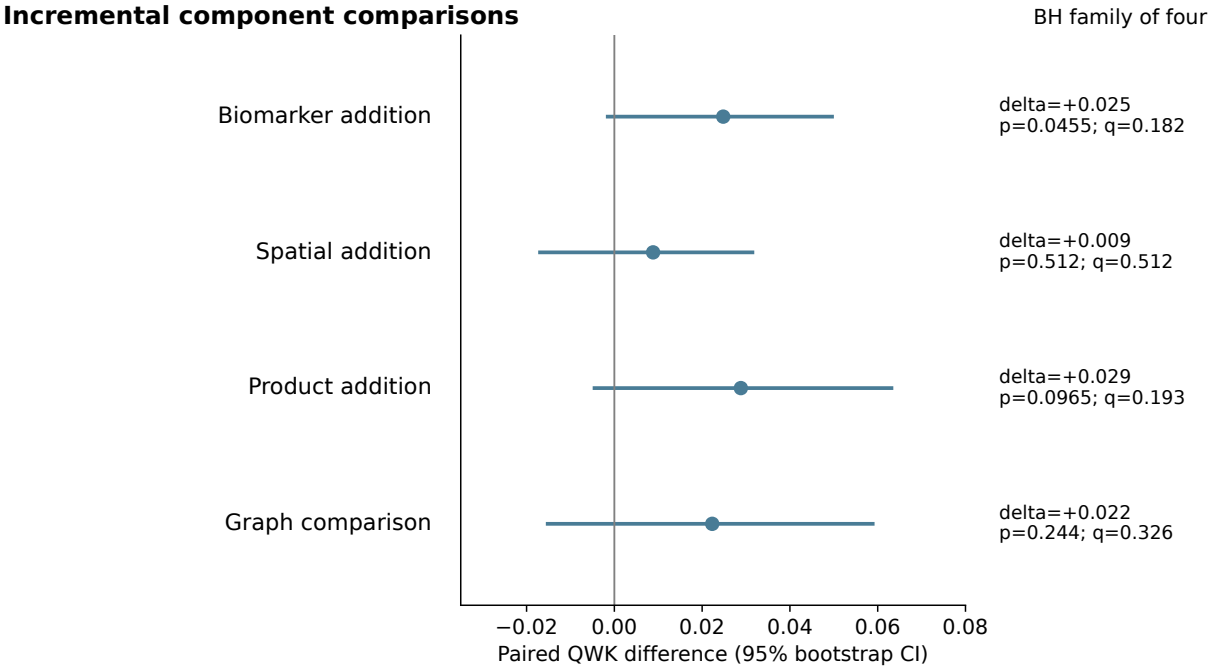



**Fig. S4** Complete incremental comparisons with paired bootstrap intervals, randomization p-values and BH q-values. The source predictions and correction families are unchanged from the main analysis.

| Model | Grade 0 | Grade 1 | Grade 2 | Grade 3 | Grade 4 |
|---|---|---|---|---|---|
| E0 | 0.939 / 0.849 | 0.082 / 0.975 | 0.213 / 0.955 | 0.120 / 0.969 | 0.737 / 0.745 |
| E1 | 0.889 / 0.738 | 0.000 / 1.000 | 0.090 / 0.986 | 0.000 / 1.000 | 0.816 / 0.650 |
| E2 | 0.976 / 0.957 | 0.492 / 0.957 | 0.245 / 0.974 | 0.040 / 0.991 | 0.816 / 0.747 |
| E3 | 0.976 / 0.968 | 0.574 / 0.947 | 0.406 / 0.962 | 0.080 / 0.976 | 0.684 / 0.823 |
| E4 | 0.980 / 0.950 | 0.508 / 0.955 | 0.465 / 0.938 | 0.280 / 0.965 | 0.553 / 0.866 |
| E5 | 0.976 / 0.961 | 0.492 / 0.949 | 0.523 / 0.931 | 0.320 / 0.962 | 0.474 / 0.885 |
| E6 | 0.983 / 0.971 | 0.639 / 0.955 | 0.561 / 0.938 | 0.280 / 0.969 | 0.605 / 0.899 |
| E7 | 0.966 / 0.978 | 0.623 / 0.928 | 0.432 / 0.917 | 0.680 / 0.891 | 0.316 / 0.968 |

**Table S4** Per-grade sensitivity / specificity. Grades are clinical reference classes; these rows are not lesion categories.

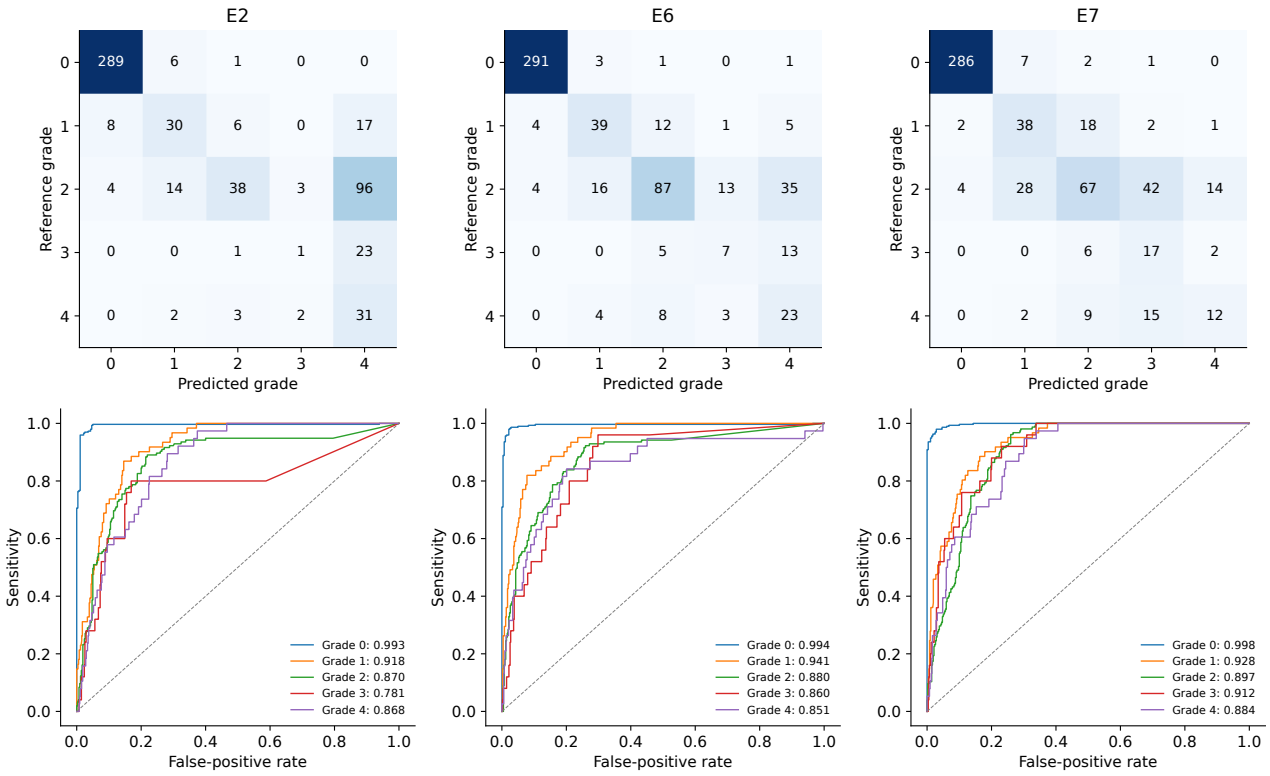



**Fig. S5** Actual test confusion matrices and class-specific ROC curves for lesion-only E2, structured-context E6 and graph-plus-biomarker E7. AUROC measures discrimination and does not replace sensitivity at the reported decision rule.

## S7 FEATURE SCREENING AND ILLUSTRATIVE-CASE SELECTION

Feature screening uses separate families of 24 spatial descriptors, 90 lesion-count/vascular fields and 1,584 product predictors. These exploratory tests are distinct from the full 264-pair grade-adjusted phenotyping analysis in Section S10. Training observations select descriptors and illustrative cases; test estimates are not used to select favourable images.

For main Figure 2, 17 X12 descriptor types satisfy both training Spearman $p < 0.005$ and training BH $q < 0.005$ across the 24-descriptor family. An imputer and scaler fit training rows only. Within each reference test grade, the displayed case is nearest the componentwise median in this fixed standardised space. Selection uses neither predicted grade, model confidence nor held-out feature $p$ values. It is not an image-specific significance test, nor evidence that every training-selected feature replicates at the same test threshold. All grading scores still use the full 575-image test partition, and incorrect predictions remain visible in the main montage.

Figure S6 includes every measured X12 descriptor, with its actual held-out sign, available sample size and BH-adjusted value. No unobserved foveal or local-calibre placeholder is plotted as anatomy. Figure S7 summarises product–grade correlations by lesion and vascular family. Products are correlated predictors, not independent biological interactions. Under the stricter training/test $p, q < 0.005$ criterion with concordant signs, the earlier screening analysis retained 734 product fields.

An earlier exploratory display selected eight direct lesion–biomarker pairs using unadjusted training associations and adjusted their within-grade tests across those eight pairs; six retained support. That result is not the primary 103-association finding, which uses correction over all 264 pairs in each partition. These selected-pair findings are not used to define support in the primary analysis.

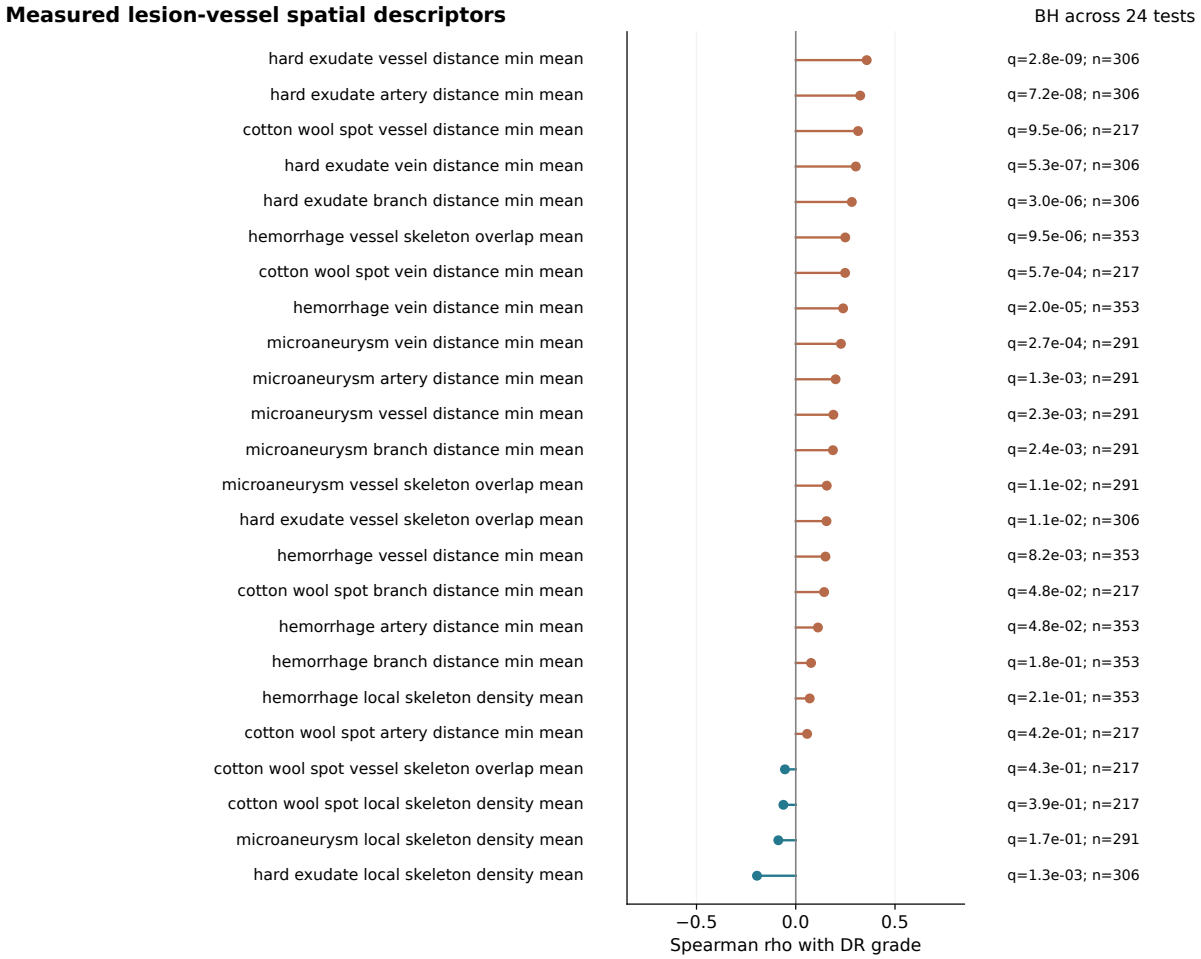



**Fig. S6** All measured X12 spatial descriptors and their held-out ordinal trends. The text column reports descriptor-specific sample sizes and BH-adjusted values across 24 tests. These are feature–grade associations, not the incremental grading effect of adding X12.

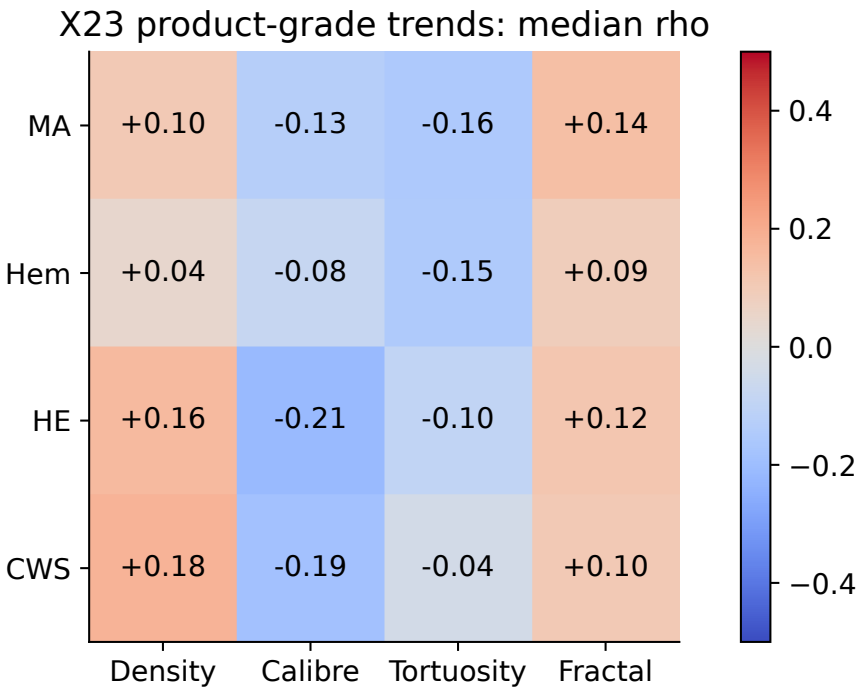



**Fig. S7** Median product–grade Spearman correlations, grouped by lesion and vascular family. The products are correlated predictors, not measured biological mechanisms. The grading contribution is evaluated separately in main Table 1.

## S8 FROZEN ENCODER COMPARISONS

Main Table 1B reports all 36 encoder–context outcomes under B0–B5; it is not duplicated here. Encoders remain frozen throughout extraction and probe fitting. EfficientNet-B4 uses ImageNet1K V1 weights, bicubic resize to 384 and centre crop 380; ResNet-152 uses ImageNet1K V2 weights, bilinear resize to 232 and crop 224; ViT-B/16 uses ImageNet1K V1 weights, bilinear resize to 256 and crop 224. All three use ImageNet channel normalisation. DINO and DINOv2 use their stored Hugging Face processors. RETFound uses the official CFP MAE checkpoint with bicubic resize to 256, crop 224 and ImageNet normalisation. Encoder dimensions are 1,792, 2,048, 768, 768, 768 and 1,024 in the order EfficientNet-B4, ResNet-152, ViT-B/16, DINO, DINOv2 and RETFound. No image-model parameters are updated.

Training-only median imputation with missingness indicators, standardization and ANOVA selection precede four balanced cumulative logistic heads. Every encoder/context evaluates the same grid: $C \in \{0.03, 0.1, 0.3, 1\}$ and $k \in \{64, 128, 256\}$. Validation QWK selects the configuration; exact ties favour smaller $k$, then smaller $C$. X23 builders are also fitted on training rows only. Validation QWK selects probe settings; no test observations refit the frozen encoders, preprocessing or classifier. Recorded B4 and B5 QWK/AUROC values are identical for all six encoders. This is an observed result, not an assumed benefit of interaction products.

| Frozen encoder | B0 QWK | B1 QWK | B0 AUC | B1 AUC |
|---|---|---|---|---|
| EfficientNet-B4 | 0.831 | **0.841** | 0.9126 | **0.9135** |
| ResNet-152 | 0.829 | **0.833** | 0.8708 | **0.8877** |
| ViT-B/16 | 0.806 | **0.809** | 0.8765 | **0.9029** |
| DINO | 0.839 | **0.840** | 0.8583 | **0.8799** |
| DINOv2 | 0.826 | **0.830** | 0.8663 | **0.8742** |
| RETFound | 0.818 | **0.830** | **0.8824** | 0.8816 |

**Table S5** Image-only B0 versus B1 with lesion evidence under the same frozen-probe framework. The AUROC entries retain four decimal places to show changes hidden by the three-decimal main table, including the small RETFound decrease. These are fixed-test outcomes, not fine-tuning results.

The configurations in main Table 1B are B0: image only; B1: image + X2; B2: image + X1/X3; B3: image + X1/X2/X3; B4: B3 + X12; B5: B4 + X23. Encoders are not fine-tuned in these comparisons. Each ordinal probe uses training-only transforms and validation selection. These are 36 outcomes on one APTOS test partition, not 36 independent replications or results from the earlier cross-cohort experiments.

## S9 GRAPH INTEGRITY AND REMAINING CONTROLS

**Construction and fitting.** The anatomical graph is formed at 256-pixel resolution. Removing branch pixels from the vascular skeleton defines vessel segments; size-based selection retains at most 48 segments and 48 predicted lesion components. Segment adjacency uses $5 \times 5$ dilation, and lesions connect bidirectionally to their three nearest segment centroids. The eleven node attributes describe object type, normalised location and size, vascular identity, branch contact or lesion confidence, and lesion category. Edge attributes contain distance and relation type.

Three edge-aware layers of width 64 apply residual updates and attention pooling. A parallel biomarker pathway has widths 64/32. Pooled graph and biomarker representations are concatenated for five-class softmax classification. Optimization uses weighted cross-entropy and AdamW with learning rate $3 \times 10^{-4}$, weight decay $10^{-3}$, batch size 96 and at most 40 epochs. Validation QWK controls early stopping. Predictions average three independently seeded models, and biomarker imputation/scaling are training-fitted.

**Integrity and interpretation.** Graph provenance checks require five ensemble member hashes, the raw-image hash and the corrected graph protocol. Legacy single-model archives are rejected. The source analysis reports regression tests checking that changing validation/test biomarker values cannot change imputer/scaler fitting, and exercising graph construction and forward/backward numerical behaviour.

The corrected graph uses centroid proximity edges, not vessel-path geodesics. Node type includes vessel and four lesion families. Biomarkers enter a separate residual pathway; they are not fabricated lesion nodes. Model attention is not validated explanation fidelity. Spatial-edge removal, within-image lesion relocation, matched loss/seed-averaging comparisons, and expert evaluation of APTOS lesion masks remain uncompleted controls and must not be reported as results.

The evaluation identifies lesion–vascular information within the stated study design. Independent cohorts, expert APTOS lesion assessment, and matched graph-topology controls would extend the current anatomical and predictive evidence.

## S10 FULL GRADE-ADJUSTED LESION–BIOMARKER ANALYSIS

This analysis asks whether lesion burden contains vascular information beyond the reference DR grade. It is separate from the 1,584 X23 products used in the grading experiments. The predictors here are four global component counts: MA, haemorrhage, hard exudate and CWS. The outcomes are 66 vascular measurements. Six disc/cup fields are excluded. All four counts are observed. Missing biomarker outcomes are omitted for each analysis and are never imputed.

**Conditional rank associations.** For each lesion–biomarker pair, both variables are ranked within the analysis split. Let $R_x$ and $R_m$ denote these ranks. The adjusted statistic is

$$r_{xm|Y} = \mathrm{corr}\left(R_x - \overline{R}_{x,Y},\ R_m - \overline{R}_{m,Y}\right), \tag{S4}$$

where the bars denote mean ranks within each reference-grade group. This removes between-grade mean rank differences. It does not remove all clinical or acquisition confounding. A two-sided permutation test uses 9,999 within-grade permutations. The finite-sample permutation p-value includes the observed ordering. BH correction [23] covers all 264 pairs separately in training and test data. Support requires $q < 0.05$ in both splits and concordant signs. A stricter summary uses $q < 0.005$ in both.

Training yields 139 discoveries and the test analysis yields 108. Of these, 103 have support in both splits with the same sign; 85 meet the stricter criterion. Supported pairs comprise 28 MA, 27 haemorrhage, 21 hard-exudate and 27 CWS associations. These are counts of measured relationships, not independent biological mechanisms. Regional versions of the same measurement can be strongly correlated.

The eight display pairs are chosen using training adjusted absolute correlations. Up to two pairs are retained per lesion, with distinct vascular families. Their test confidence intervals use 2,000 grade-stratified image bootstrap draws. Ranking and grade adjustment are repeated in every draw. This selection rule differs from the earlier eight-pair display, which used unadjusted training associations. That earlier analysis used an eight-comparison correction and is described separately in Section S7; its six supported pairs are not the full-family result.

**Predicting a biomarker conditional on grade.** Each observed biomarker is a separate regression outcome. The baseline includes reference grade only. The extended model adds $\log(1 + n_c)$ for each of the four lesion counts. Grade is encoded by five indicators. Both models include an intercept. Ridge regression [24] handles the redundant intercept/indicator coding; individual coefficients are not used for inference.

For each outcome, predictor scales and the outcome mean and standard deviation are fitted using its observed training rows. Separate baseline and extended penalties are selected from $\{0, 0.1, 1, 10, 100, 1000\}$ by validation standardized mean squared error. Exact ties favour the larger penalty. The models are not refitted on validation or test observations. Observed sample sizes range from 1,455–1,724 in training, 475–575 in validation and 477–575 in testing.

On the test observations, $R^2 = 1 - \mathrm{SSE}/\mathrm{SST}$ uses the observed test outcome mean in SST. The increment is $\Delta R^2 = R^2_{\mathrm{extended}} - R^2_{\mathrm{baseline}}$. RMSE is also reported in the recorded measurement units. Paired image bootstrap resampling gives 95% intervals for $\Delta R^2$ using 2,000 draws. A paired prediction-swap test uses 9,999 draws. BH correction covers all 66 outcomes. A supported gain requires positive $\Delta R^2$, an interval above zero and $q < 0.05$. Negative and uncertain changes remain in the tables and plots.

There are 48 positive increments and 26 supported gains. Calibre/width has 18 supported outcomes among 21 tested fields; fractal dimension has six among nine. Density has one of nine and tortuosity one of 27. Median increments across all outcomes in each family are 0.0655, 0.0348, 0.0063 and $-0.0006$, respectively. These results describe additional vascular information carried jointly by the four lesion burdens. They do not assign a unique effect to an individual lesion. Predictor overlap and nonuniform residual variation also limit coefficient-level interpretation. No features or observations were pruned for this presentation.

Reference grade is known in this conditional research comparison. It is not a deployable biomarker predictor for an ungraded eye. Neither this analysis nor the grading experiment includes linked systemic outcomes. Oculomics applications require new outcome-linked cohorts, relevant covariates and independent evaluation.

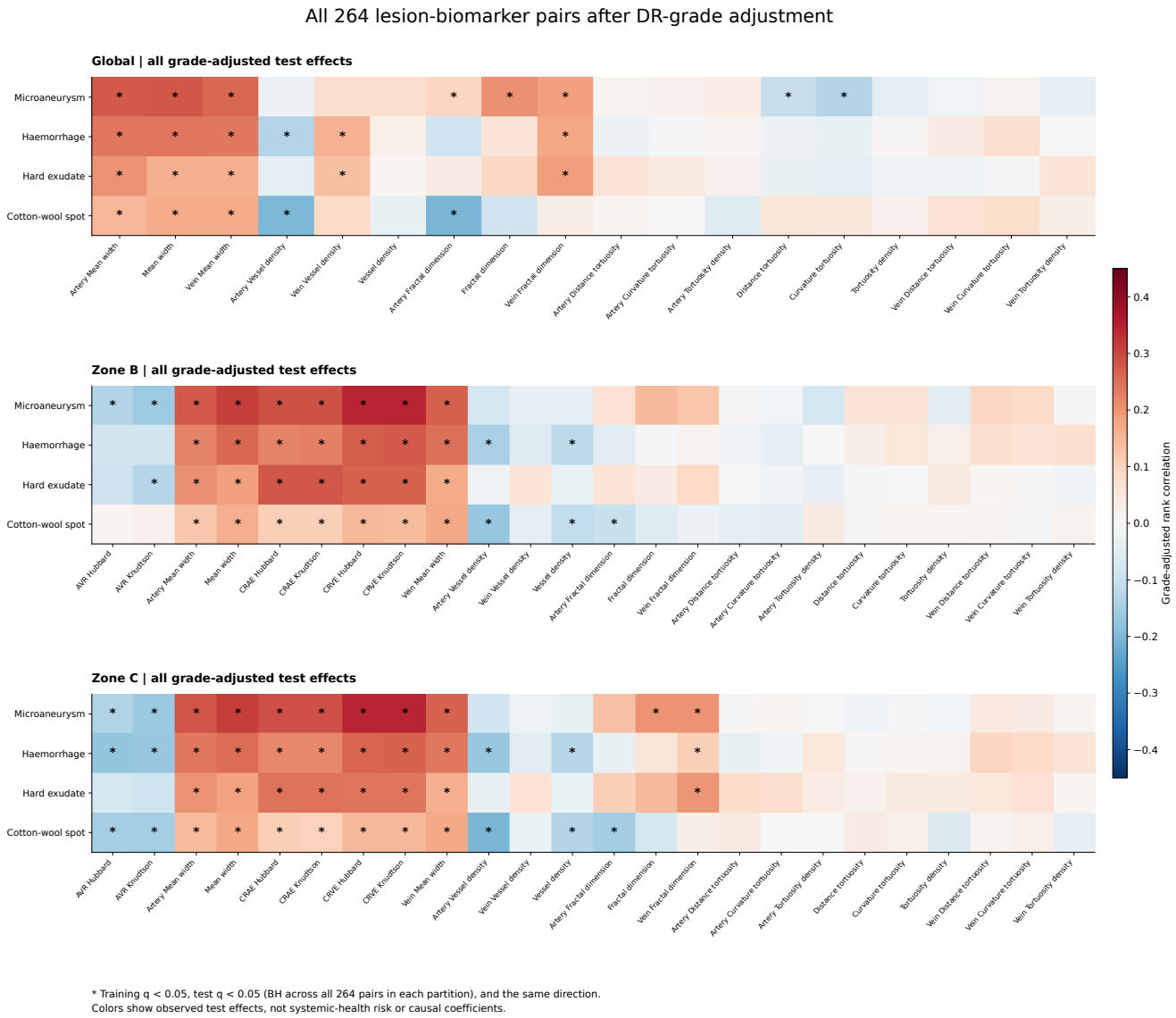



**Fig. S8** All 264 lesion–vascular pairs after grade adjustment. Estimates use the held-out APTOS observations with observed biomarker outcomes. Full-family correction and agreement with training determine support.

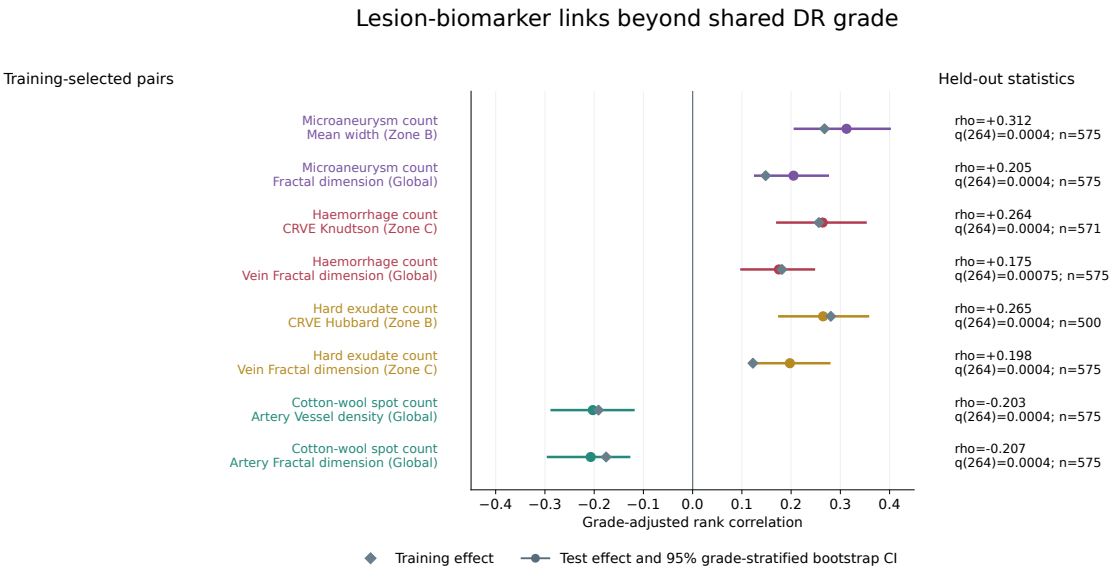



**Fig. S9** Eight training-selected adjusted associations. Test effects and grade-stratified bootstrap intervals are shown without selecting pairs by their test effects. Correction covers all 264 pairs, not only the displayed eight.

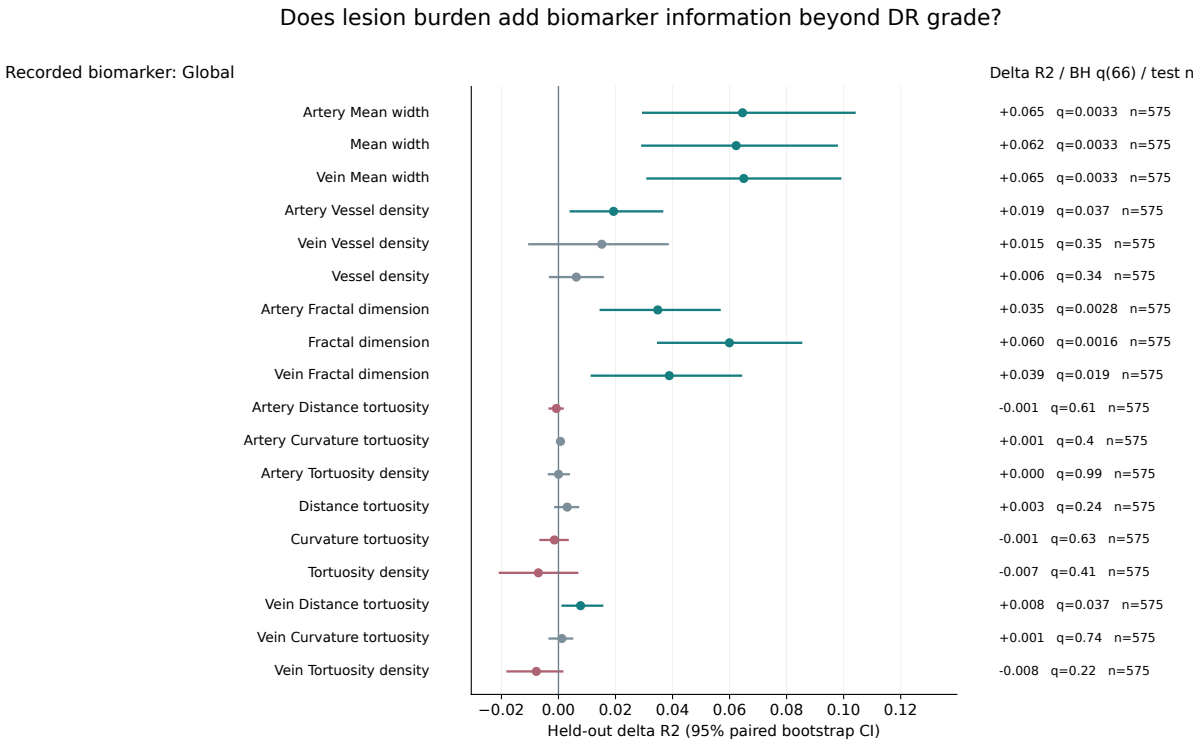



**Fig. S10** Global vascular outcomes: test $\Delta R^2$ from adding the four lesion burdens to a grade-only model. Intervals and supported/uncertain changes are retained. These are biomarker prediction increments, not DR grading improvements.

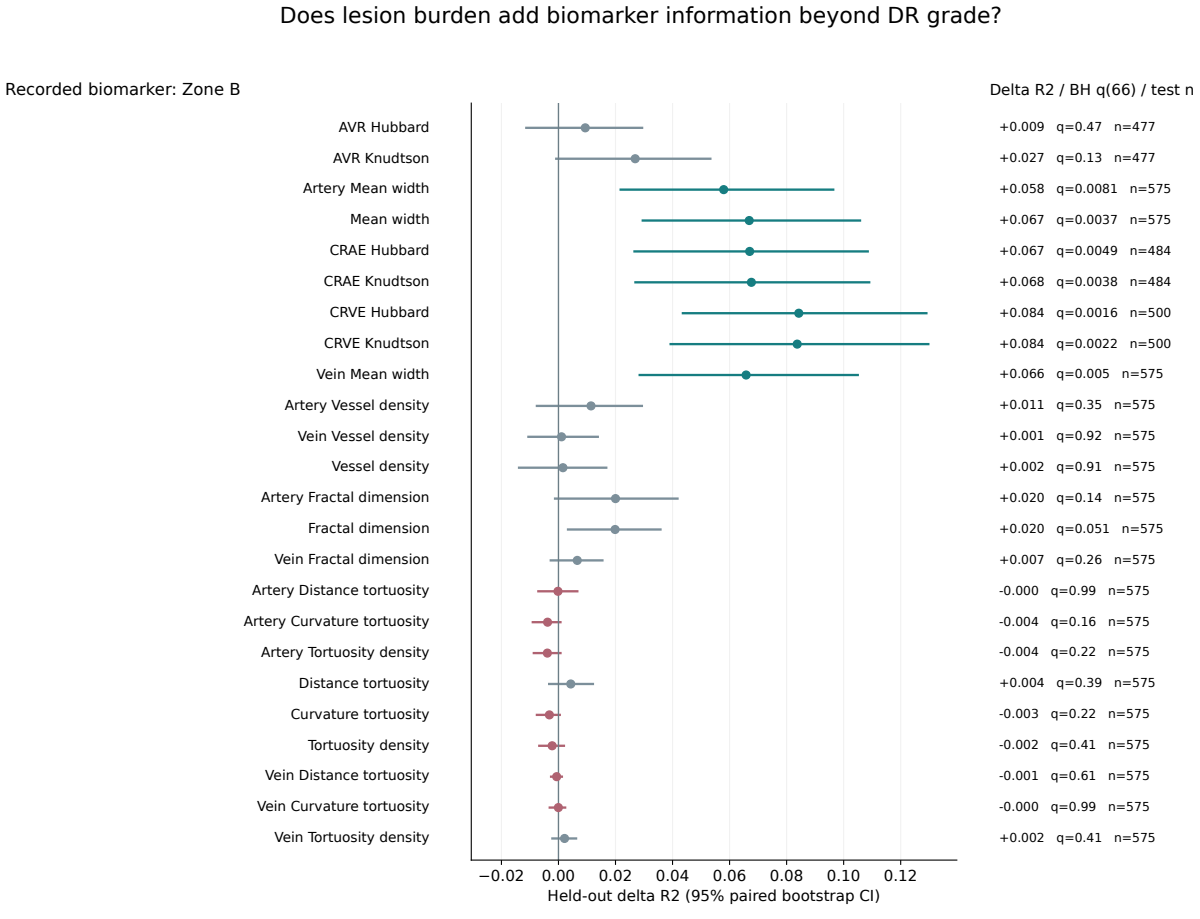



**Fig. S11** The same conditional prediction comparison for optic-disc-referenced zone B measurements. No missing outcome is imputed. Each outcome has its own observed test sample.

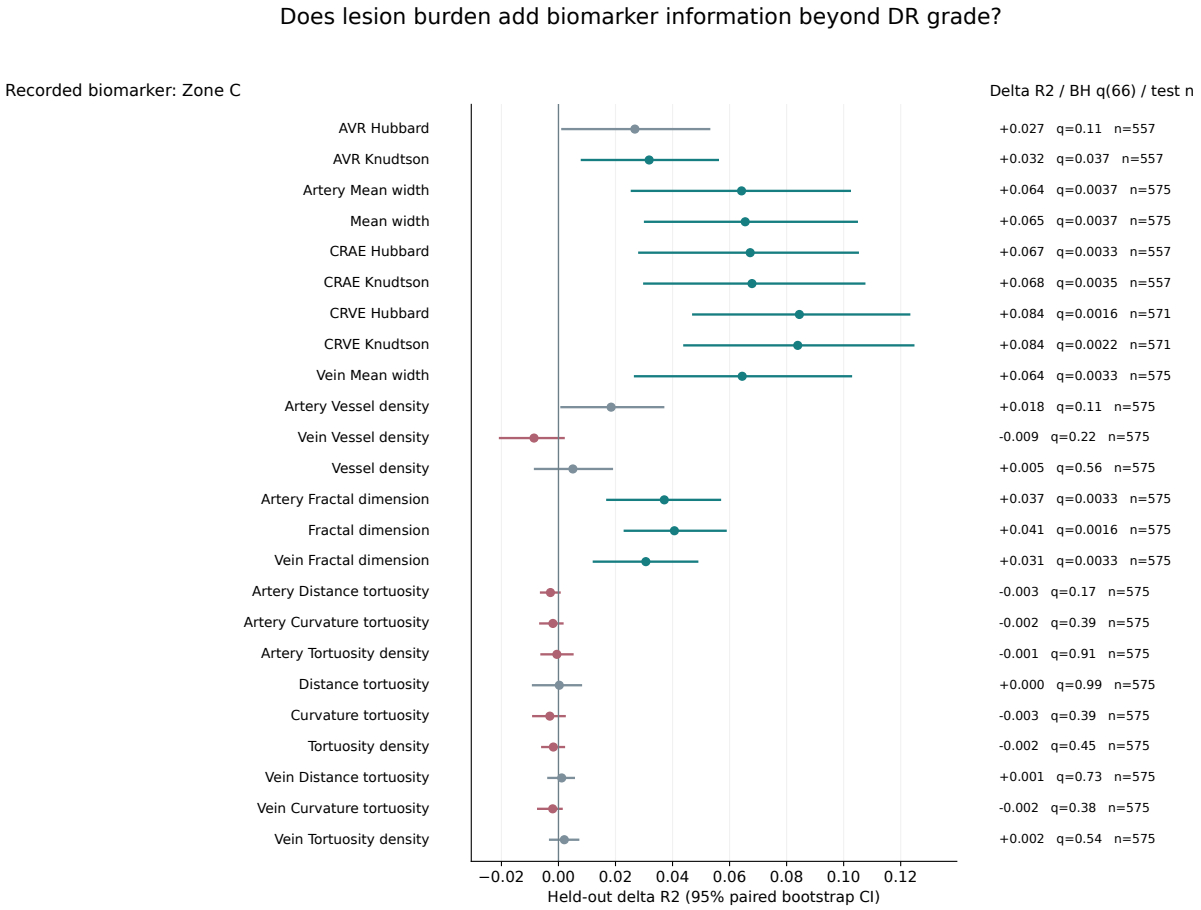



**Fig. S12** Conditional prediction increments for optic-disc-referenced zone C measurements. Shared regional or vascular-family information does not imply independent biological effects.