Feature Space Translation Framework for Cross-Device Alignment and Derivation of Clinically Relevant Retinal Biomarkers from Foundation Model Embeddings

Anonymous Authors[1]

Abstract

Oculomics enables non-invasive assessment of systemic diseases through retinal imaging, offering a portable, cost-effective alternative to traditional methods. However, its real-world deployment is limited by device-induced variability and poor generalizability across heterogeneous imaging systems. In recent years, large retinal foundation models such as RetFound and RetFound Green have shown strong performance on several downstream tasks. However, their generalization across imaging devices remains limited. Furthermore, the high-dimensional representations produced by these models do not directly correspond to clinically meaningful retinal biomarkers, further limiting their interpretability and clinical utility.

To address these challenges, we propose a feature space translation framework using tabular neural networks to align feature distributions across devices, thereby improving cross-device generalization. We further investigate whether the same framework can be used to translate final-stage foundation model embeddings into clinically meaningful retinal biomarkers, enabling more interpretable representations. Experiments on the AI-READI dataset show that our framework supports cross-device translation of retinal biomarkers, enables biomarker extraction from foundation model embeddings, and allows cross-device biomarker estimation directly from embeddings. Together, these results demonstrate seamless integration of our approach into existing oculomics pipelines while improving the consistency, interpretability, and clinical usability of retinal biomarkers.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Oculomics (Zhu et al., 2025) is a field that leverages retinal imaging to enable the non-invasive assessment of systemic disease burden by capturing microvascular changes through structural biomarkers. The retina is the only external organ in the human body where microvasculature can be directly visualized non-invasively, providing a unique view of blood vessels that feed blood to the heart, kidneys, brain and other vital organs. Changes in retinal biomarkers such as vessel tortuosity, fractal dimension, vessel density, and the arteriovenous ratio (AVR) (Estrada et al., 2015; Huang & Dashtbozorg, 2018; Srinidhi et al., 2019) have demonstrated strong diagnostic relevance for chronic conditions, including chronic kidney disease (Sasongko et al., 2012; Tan et al., 2024), type 2 diabetes, cardiovascular disease (Wong et al., 2002; Cheung et al., 2007), and neurological disorders (Liew et al., 2021). These biomarkers provide a unique window into systemic physiology, as retinal vasculature shares anatomical and functional similarities with other microvascular beds, making it particularly suitable for early disease detection and risk stratification.

Recent advances in deep learning, along with standardized extraction pipelines such as AutoMorph (Zhou et al., 2022), have enabled large-scale, automated, and reproducible quantification of these retinal features (Fraz et al., 2015; Perez-Rovira et al., 2011) from fundus images. In parallel, retinal foundation models (FMs) pretrained on large-scale datasets, such as RETFound (Zhou et al., 2023), have been proposed, leveraging extensive pretraining to shift the paradigm from handcrafted feature extraction to representation learning. These models learn rich, hierarchical representations that capture complex retinal patterns and achieve strong performance across diverse downstream tasks. Despite this progress, the deployment of these approaches in real-world clinical settings remains limited due to two major constraints. First, out-of-distribution (OOD) shifts arising from camera and device heterogeneity introduce systematic variability in image characteristics and extracted features, thereby reducing robustness and leading to model drift across datasets (Futoma et al., 2020). Second, despite their strong representational capacity, the high-dimensional

latent embeddings produced by these models lack direct correspondence with clinically interpretable retinal biomarkers, limiting their transparency, clinical validation, and broader adoption in practice.

To overcome these limitations, we propose a feature space translation framework that leverages tabular neural networks to map retinal biomarkers from the feature space of one imaging device to another, thereby mitigating device-induced variability and camera drift. Rather than introducing a new architectural design, our approach focuses on learning cross-device correspondences in feature space to improve robustness across heterogeneous imaging systems. Furthermore, we extend this framework to operate on the latent embeddings of retinal foundation models, such as RETFound (Zhou et al., 2023), to translate high-dimensional representations into clinically meaningful retinal biomarkers. In addition, the same formulation enables cross-device biomarker estimation directly from embeddings, allowing representations derived from one device to be calibrated and interpreted in the context of another. By decoupling representation learning from device-specific biases, our approach provides a simple yet effective way to improve the consistency, interpretability, and cross-device applicability of oculomics pipelines. The key contributions of this work could be summarized as follows:

- We perform cross-device feature translation of retinal biomarkers, extend the same framework to map latent embeddings from retinal foundation models such as RETFound to clinically meaningful biomarkers, and further enable cross-device biomarker estimation directly from embeddings.
- We investigate and benchmark a range of state-of-the-art tabular neural networks alongside classical machine learning approaches for the feature space translation task, providing a comprehensive evaluation of their effectiveness in this setting.
- Our framework is model-agnostic and can be directly integrated into existing oculomics pipelines without architectural modifications, enabling improved cross-device consistency and interpretability with minimal overhead.

2. Methodology

2.1. Problem Formulation

Let $\mathcal{D} = \{d_1, \ldots, d_K\}$ denote a set of imaging devices. For each device $d \in \mathcal{D}$, let $X^{(d)} \in \mathbb{R}^p$ represent the structured retinal biomarker space (e.g., AutoMorph features), and $E^{(d)} \in \mathbb{R}^q$ denote the corresponding latent representation obtained from a retinal foundation model such as RETFound.

We aim to learn a set of parametric mappings that enable (i) cross-device biomarker translation (Eq. 1), (ii) embedding-to-biomarker translation within a device (Eq. 2), and (iii) cross-device biomarker estimation directly from embeddings (Eq. 3).

$$f_{d_i \to d_j} : \mathbb{R}^p \to \mathbb{R}^p, \quad X^{(d_j)} \approx f_{d_i \to d_j}(X^{(d_i)}) \tag{1}$$

$$g_d : \mathbb{R}^q \to \mathbb{R}^p, \quad X^{(d)} \approx g_d(E^{(d)}) \tag{2}$$

$$h_{d_i \to d_j} : \mathbb{R}^q \to \mathbb{R}^p, \quad X^{(d_j)} \approx h_{d_i \to d_j}(E^{(d_i)}) \tag{3}$$

where $f_{d_i \to d_j}$, g_d, and $h_{d_i \to d_j}$ correspond to tasks (i), (ii), and (iii), respectively.

All mappings are learned using tabular predictive models by minimizing supervised reconstruction objectives over paired samples. For a given mapping $\phi \in \{f_{d_i \to d_j}, g_d, h_{d_i \to d_j}\}$, we define a general objective (Eq. 4) over input–target pairs $(u, v) \sim \mathcal{P}$, where $\ell(\cdot, \cdot)$ denotes a task-specific loss function.

$$\mathcal{L}(\phi) = \mathbb{E}_{(u,v)\sim\mathcal{P}} \left[\ell\big(v, \phi(u)\big) \right] \tag{4}$$

$$\mathcal{L}_{\text{NLL}}(\phi) = -\mathbb{E}_{(u,v)\sim\mathcal{P}} \left[\log p_\phi\big(v \mid u\big) \right] \tag{5}$$

Under a probabilistic formulation, we model the conditional distribution $p_\phi(v \mid u)$ and minimize the negative log-likelihood (Eq. 5). Assuming a Gaussian likelihood with mean $\phi(u)$ and fixed variance, this reduces to a mean squared error objective. To account for distributional discrepancies induced by cross-device variability, we additionally consider a Wasserstein-distance-based objective (Eq. 6) that promotes alignment between the predicted and target distributions. The overall optimization problem is defined in Eq. 7, where the mappings are learned over paired samples. This formulation provides a unified framework for addressing device-induced variability while improving the interpretability of foundation model representations.

$$\mathcal{L}_{\text{W}}(\phi) = \mathbb{E}_{(u,v)\sim\mathcal{P}} \left[W\big(\delta_{\phi(u)}, \delta_v\big) \right] \tag{6}$$

$$\min_{\phi} \ \mathcal{L}(\phi), \quad \phi \in \{f_{d_i \to d_j}, g_d, h_{d_i \to d_j}\} \tag{7}$$

2.2. Feature Translation Framework

The overall framework is illustrated in Fig. 1. We use retinal fundus images from the AI-READI dataset (AI-READI Consortium, 2024b;a), comprising 1,067 participants with imaging from multiple non-mydriatic devices, primarily Topcon

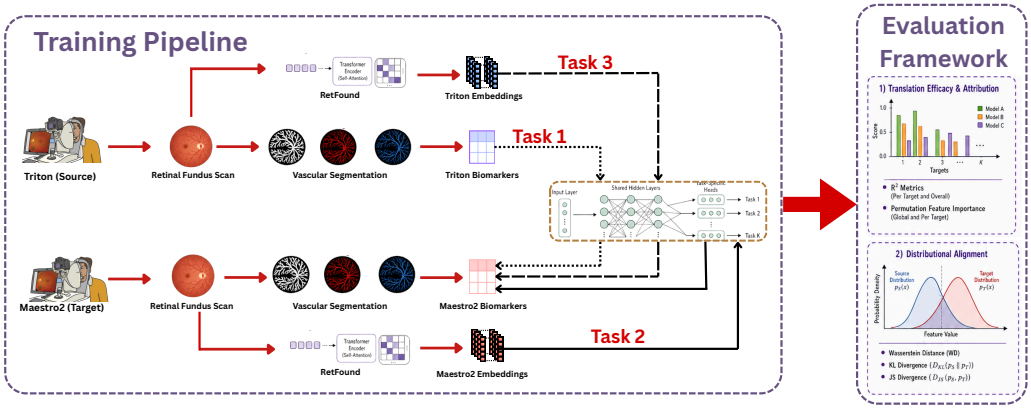

Figure 1. Overview of the proposed feature space translation framework. The training pipeline consists of three tasks: (i) cross-device biomarker translation, (ii) embedding-to-biomarker translation, and (iii) cross-device biomarker estimation from embeddings. The framework operates on retinal fundus images from multiple devices and is evaluated using both prediction accuracy and distributional alignment metrics.

Triton (source) and Topcon Maestro2 (target). These devices differ in acquisition characteristics, leading to systematic variations in extracted features. For each image, we extract clinically relevant vascular biomarkers (12 features) using AutoMorph, selected based on their established clinical relevance in characterizing retinal microvasculature. These include measures of vessel caliber (e.g., CRAE, CRVE, and AVR), structural complexity (e.g., fractal dimension and vessel density), and geometric properties such as vessel tortuosity. In parallel, we obtain latent embeddings using the pretrained RETFound model, resulting in two complementary representations per sample: a structured biomarker vector and a high-dimensional embedding.

We evaluate the framework through three complementary translation settings. First, we perform cross-device biomarker translation by learning mappings between AutoMorph-derived biomarkers extracted from Triton and Maestro2 images for paired samples. Second, we learn embedding-to-biomarker mappings independently for each device by regressing RETFound embeddings to the corresponding AutoMorph features, enabling assessment of how well latent representations capture clinically meaningful information. Third, we enable cross-device biomarker estimation directly from embeddings by combining these mappings, allowing embeddings from one device to be used to estimate biomarkers in the feature space of another device. All mappings are learned using tabular predictive models. Training is performed on paired samples available for participants imaged on multiple devices. The dataset is split at the patient level into 70%/15%/15% for training, validation, and testing, respectively, ensuring no subject overlap across splits. To ensure robustness, we further employ 5-fold cross-validation on the training set for model selection. Performance is evaluated using standard regression metrics, including mean absolute error (MAE) and coefficient of determination (R^2). In addition, we assess distributional alignment across devices using Wasserstein distance, Kullback–Leibler (KL) divergence, and Jensen–Shannon (JS) divergence, providing a comprehensive evaluation of both predictive accuracy and cross-device consistency.

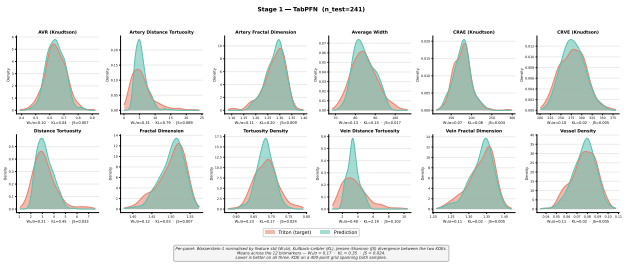

Figure 2. Distributional alignment for direct cross-device translation of AutoMorph-derived retinal biomarkers from Maestro2 to Triton.

3. Results and Discussion

We begin by evaluating cross-device biomarker translation using eight tabular learning approaches, including classical machine learning methods such as CatBoost, Gradient Boosting, Ridge Regression, and Linear Regression, along with state-of-the-art tabular neural network models including TabPFN (Hollmann et al., 2023), TabM (Gorishniy et al., 2025), RealMLP, and a lightweight MLP. These models were selected to span a diverse range of inductive biases, from linear baselines to recent architectures that have demonstrated strong performance on tabular benchmarks such as TabArena (Erickson et al., 2025).

Table 1 summarizes the performance for translation from Maestro2 to Triton. Strong predictive performance is observed across most vascular biomarkers, with features such

Table 1. Cross-device biomarker translation performance from Maestro2 to Triton across eight tabular learning models. Performance is reported in terms of coefficient of determination (R^2) and root mean squared error (RMSE). The best-performing method for each biomarker is shown in **bold**, and the second-best is underlined.

	TabPFN		Lightweight MLP		CatBoost		TabM		RealMLP		Ridge		Gradient Boosting		Linear Regression	
Biomarker	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
AVR (Knudtson)	**0.597**	**0.0462**	0.592	0.0465	0.530	0.0498	0.592	0.0465	0.568	0.0478	0.578	0.0472	0.503	0.0513	0.578	0.0472
Artery Distance Tortuosity	**0.296**	**3.0711**	0.256	3.1557	0.266	3.1342	0.278	3.1089	0.246	3.1777	0.256	3.1565	0.221	3.2300	0.253	3.1620
Artery Fractal Dimension	**0.688**	**0.0256**	0.669	0.0264	0.633	0.0278	0.683	0.0258	0.682	0.0258	0.675	0.0261	0.645	0.0273	0.675	0.0261
Average Width	**0.674**	**3.5798**	0.664	3.6346	0.610	3.9129	0.673	3.5824	0.648	3.7169	0.668	3.6114	0.649	3.7121	0.667	3.6162
CRAE (Knudtson)	0.729	11.3928	0.735	11.2756	0.725	11.4754	0.723	11.5155	**0.737**	**11.2203**	0.730	11.3733	0.698	12.0211	0.730	11.3659
CRVE (Knudtson)	0.782	14.3771	0.780	14.4633	0.768	14.8600	0.777	14.5561	0.781	14.4346	**0.785**	**14.3021**	0.748	15.4846	0.784	14.3234
Distance Tortuosity	**0.429**	**0.0135**	0.412	0.0138	0.395	0.0142	0.420	0.0136	0.405	0.0139	0.415	0.0137	0.372	0.0146	0.413	0.0138
Fractal Dimension	**0.701**	**0.0179**	0.694	0.0181	0.675	0.0188	0.699	0.0180	0.692	0.0182	0.697	0.0180	0.661	0.0191	0.696	0.0181
Tortuosity Density	**0.512**	**0.0217**	0.498	0.0220	0.470	0.0227	0.505	0.0218	0.488	0.0222	0.500	0.0219	0.452	0.0231	0.499	0.0219
Vein Distance Tortuosity	**0.318**	**0.0148**	0.302	0.0151	0.281	0.0156	0.310	0.0149	0.295	0.0152	0.305	0.0150	0.260	0.0160	0.303	0.0151
Vein Fractal Dimension	**0.645**	**0.0195**	0.632	0.0198	0.601	0.0206	0.640	0.0196	0.628	0.0199	0.636	0.0197	0.590	0.0209	0.635	0.0198
Vessel Density	**0.589**	**0.0321**	0.575	0.0326	0.542	0.0338	0.584	0.0323	0.567	0.0329	0.579	0.0324	0.530	0.0342	0.577	0.0325

as fractal dimension, vessel density, and vessel caliber (CRAE, CRVE) achieving R^2 values in the range of 0.6–0.78. TabPFN consistently achieves the best or near-best performance across most biomarkers, while TabM and the lightweight MLP also demonstrate competitive results. In contrast, linear models perform worse on more complex features such as tortuosity, indicating non-linear cross-device relationships. This is further supported by distributional analysis (Fig. 2), where predicted and target distributions show strong overlap with low divergence, suggesting effective mitigation of cross-device variability in the biomarker space.

Based on these observations, we restrict subsequent experiments to TabPFN, lightweight MLP, and CatBoost. We next evaluate embedding-to-biomarker translation to assess whether RETFound embeddings encode clinically meaningful retinal features. As shown in Table 2, several biomarkers, particularly fractal dimension and vessel caliber, can be recovered with high accuracy, with R^2 values exceeding 0.75. However, performance varies across features, with tortuosity-related measures showing weaker alignment. Distributional comparisons (see Fig. 3 in Appendix) indicate that non-linear models achieve better alignment, highlighting their importance for interpreting latent representations. These results suggest that RETFound embeddings encode clinically meaningful vascular information, although not uniformly across all biomarkers.

Finally, we evaluate the compositional setting, where embeddings are mapped to biomarkers and then translated across devices. As shown in Table 3, performance degrades relative to direct translation due to error accumulation, but several structural biomarkers such as fractal dimension and vessel density retain competitive performance. Distributional analysis (see Fig. 4 in Appendix) shows moderate divergence, while predicted distributions remain well-aligned for key features. Taken together, these results demonstrate that feature space translation mitigates cross-device variability and enables the extraction of clinically meaningful biomarkers from retinal foundation model embeddings. The feasibility of compositional translation further supports its use for cross-device clinical validation and deployment.

4. Conclusion

In this work, we presented a feature-space translation framework to address device-induced variability in oculomics while improving the interpretability of retinal foundation model representations. By learning mappings between structured biomarkers and latent embeddings, the framework enables consistent analysis across heterogeneous imaging devices. Experiments on the AI-READI dataset show that cross-device biomarker translation can be achieved reliably using tabular models, with strong predictive performance and distributional alignment. We further demonstrate that retinal foundation model embeddings encode clinically meaningful vascular information that can be recovered through appropriate mappings, enabling biomarker estimation directly from device-agnostic embeddings.

Overall, these results highlight feature-space translation as a practical solution for improving cross-device generalization in oculomics, supporting more reliable and clinically interpretable analysis. While this study uses an earlier release of the AI-READI dataset (1,067 participants) and focuses on one-directional translation from Maestro2 to Triton, future work will extend to bidirectional and multi-device settings and incorporate the newer 2,200+ participant release for more comprehensive validation. This study focused on Maestro2 and Triton due to quality limitations in other devices in AI-READI, future work will focus on improving robustness to lower-quality imaging and extend the framework to additional devices and datasets.

References

AI-READI Consortium. Flagship dataset of type 2 diabetes from the ai-readi project (2.0.0) [data set], 2024a. URL https://doi.org/10.60775/fairhub.2.

AI-READI Consortium. Ai-readi: rethinking data collection, preparation and sharing for propelling ai-based discoveries in diabetes research and beyond. *Nature Metabolism*, 2024b. doi: 10.1038/s42255-024-01165-x. URL https://doi.org/10.1038/s42255-024-01165-x.

Cheung, N., Bluemke, D. A., et al. Retinal arteriolar narrowing and left ventricular remodeling. In *Journal of the American College of Cardiology*, pp. 48–55, 2007.

Erickson, N., Purucker, L., Tschalzev, A., Holzmüller, D., Desai, P. M., Salinas, D., and Hutter, F. Tabarena: A living benchmark for machine learning on tabular data. *arXiv preprint arXiv:2506.16791*, 2025.

Estrada, R., Allingham, M. J., et al. Retinal artery-vein classification via topology estimation. In *IEEE Transactions on Medical Imaging*, pp. 2518–2534, 2015.

Fraz, M. M., Welikala, R. A., et al. Quartz: Quantitative analysis of retinal vessel morphology. In *Expert Systems with Applications*, pp. 7221–7234, 2015.

Futoma, J., Simons, M., et al. The myth of generalisability in clinical machine learning. In *Lancet Digital Health*, pp. e489–e492, 2020.

Gorishniy, Y., Kotelnikov, A., and Babenko, A. Tabm: Advancing tabular deep learning with parameter-efficient ensembling. In *International Conference on Learning Representations (ICLR)*, 2025.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. Tabpfn: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations (ICLR)*, 2023.

Huang, F. and Dashtbozorg, B. Artery/vein classification in fundus images. In *Machine Vision and Applications*, pp. 23–34, 2018.

Liew, G., Gopinath, B., et al. Retinal vasculature fractal and stroke mortality. In *Stroke*, pp. 1276–1282, 2021.

Perez-Rovira, A., MacGillivray, T., et al. Vampire: Vessel assessment platform for retina. In *IEEE EMBS Conference*, 2011.

Sasongko, M. B., Wong, T. Y., et al. Retinal arteriolar tortuosity and kidney dysfunction. In *American Journal of Ophthalmology*, pp. 176–183, 2012.

Srinidhi, C. L., P., A., and Rajan, J. Automated retinal artery/vein separation. In *IEEE Transactions on Image Processing*, pp. 2705–2718, 2019.

Tan, Y., Ma, Y., et al. Deep learning for detection of chronic kidney disease from retinal images. In *European Journal of Ophthalmology*, pp. 502–509, 2024.

Wong, T. Y., Klein, R., et al. Retinal arteriolar narrowing and risk of coronary heart disease. In *JAMA*, 2002.

Zhou, Y., Wagner, S. K., et al. Automorph: Automated retinal vascular morphology quantification. In *Translational Vision Science and Technology*, 2022.

Zhou, Y., Chia, M. A., Wagner, S. K., Ayhan, M. S., Williamson, D. J., Struyven, R. R., Liu, T., Xu, M., Lozano, M. G., Woodward-Court, P., et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.

Zhu, Z., Wang, Y., et al. Oculomics: Current concepts and evidence. In *Progress in Retinal and Eye Research*, 2025.

A. Additional Results

Table 2. Embedding-to-biomarker translation performance on Maestro2, where RETFound embeddings are mapped to AutoMorph-derived retinal biomarkers. Performance is reported using R^2 and RMSE. Best results are shown in **bold** and second-best in underline.

	TabPFN		Ridge		ElasticNet		PLS	
Biomarker	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
AVR (Knudtson)	0.058	0.0779	0.027	0.0792	0.033	0.0789	**0.083**	**0.0769**
Artery Distance Tortuosity	-0.036	3.9064	-0.342	4.4474	-0.000	3.8393	**0.019**	**3.8026**
Artery Fractal Dimension	0.749	0.0195	0.751	0.0195	**0.789**	**0.0179**	0.754	0.0193
Average Width	0.475	4.8116	**0.576**	**4.3267**	0.539	4.5109	0.426	5.0306
CRAE (Knudtson)	0.487	16.7132	**0.537**	**15.8784**	0.504	16.4320	0.511	16.3203
CRVE (Knudtson)	0.490	27.9760	**0.540**	**26.5600**	0.501	27.6794	0.539	26.6057
Distance Tortuosity	**0.233**	**0.8200**	0.126	0.8754	0.181	0.8474	0.129	0.8743
Fractal Dimension	0.857	0.0154	**0.875**	**0.0144**	0.835	0.0166	0.830	0.0168
Tortuosity Density	0.241	0.0299	-0.035	0.0349	0.149	0.0316	**0.243**	**0.0298**
Vein Distance Tortuosity	-0.046	1.4436	-0.159	1.5194	**0.039**	**1.3837**	-0.031	1.4327

Table 3. Cross-device biomarker estimation from embeddings (Maestro2 $\rightarrow$ Triton), where RETFound embeddings are first translated to biomarkers and then aligned across devices. Performance is reported using R^2 and RMSE. Best results are shown in **bold** and second-best in underline.

	TabPFN		CatBoost		Lightweight MLP		Ridge	
Biomarker	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
AVR (Knudtson)	0.042	0.0869	0.017	0.0880	**0.052**	**0.0864**	0.040	0.0869
Artery Distance Tortuosity	0.065	3.8367	0.023	3.9215	-0.006	3.9780	**0.085**	**3.7950**
Artery Fractal Dimension	0.573	0.0309	0.570	0.0311	0.567	0.0312	**0.577**	**0.0308**
Average Width	0.318	5.1273	**0.338**	**5.0522**	0.331	5.0780	0.322	5.1112
CRAE (Knudtson)	0.443	18.5751	0.442	18.5904	**0.461**	**18.2647**	0.452	18.4190
CRVE (Knudtson)	0.445	**27.0663**	0.442	27.1275	0.437	27.2463	**0.540**	26.5600
Distance Tortuosity	0.216	0.7865	0.191	0.7986	0.116	0.8348	**0.196**	**0.7962**
Fractal Dimension	0.715	0.0225	0.703	0.0230	**0.716**	0.0225	0.707	**0.0224**
Tortuosity Density	0.194	0.0322	0.176	0.0325	0.147	0.0331	**0.195**	**0.0322**
Vein Distance Tortuosity	-0.081	1.4492	-0.110	1.4691	-0.071	1.4429	**-0.060**	**1.4353**
Vein Fractal Dimension	0.719	0.0224	0.699	0.0232	**0.731**	**0.0219**	0.719	0.0224
Vessel Density	**0.760**	0.0068	0.747	0.0070	0.752	**0.0069**	0.750	0.0070

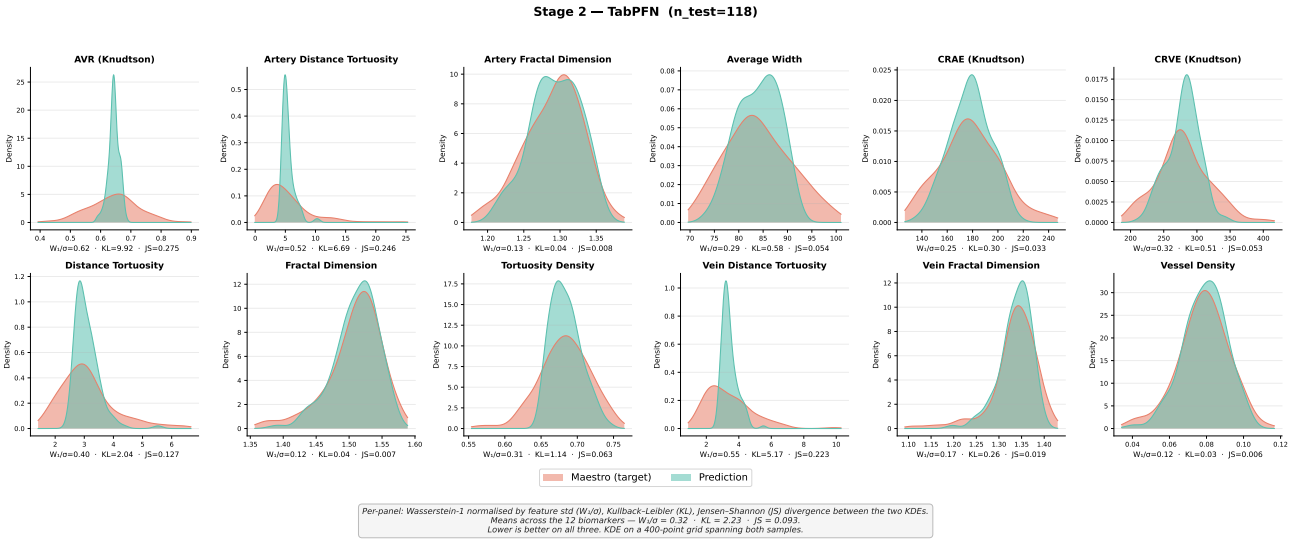

Figure 3. Distributional alignment between predicted biomarkers obtained from RETFound embeddings and ground-truth AutoMorph biomarkers. Strong overlap across key vascular features indicates that the embeddings encode clinically meaningful retinal information.

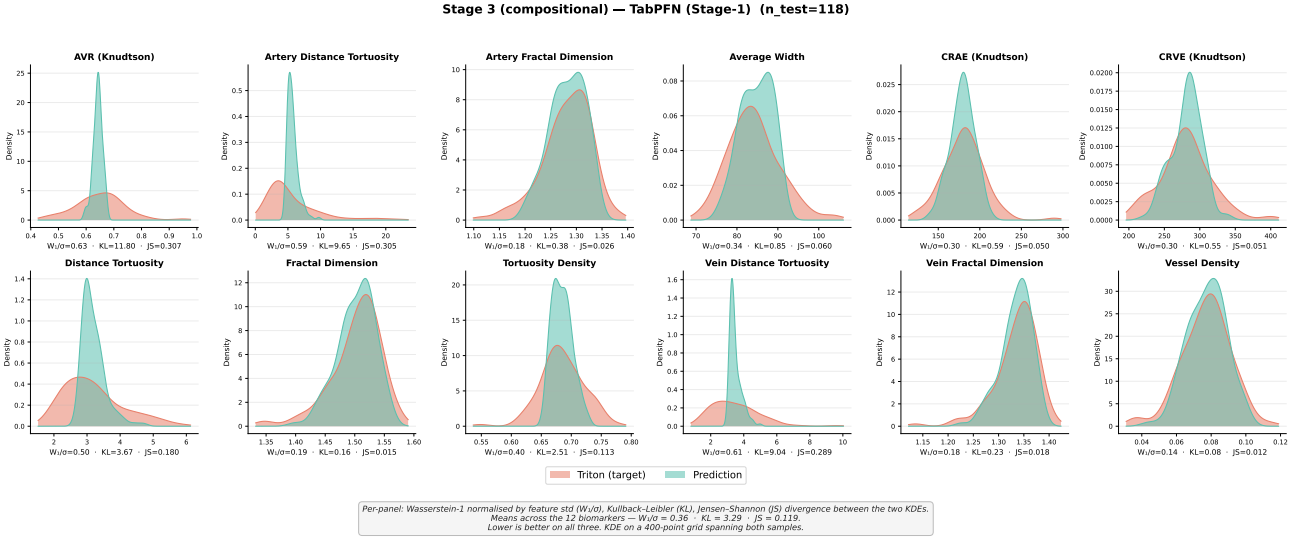

Figure 4. Distributional alignment for cross-device biomarker estimation from embeddings. RETFound embeddings are first translated to biomarkers and then aligned across devices. Predicted distributions remain well-aligned with target Triton biomarkers for key features.