A Global Atlas of Digital Dermatology to Map Innovation and Disparities

Fabian Gröger[1,2*], Simone Lionetti[2], Philippe Gottfrois[1,3], Alvaro Gonzalez-Jimenez[2,3], Lea Habermacher[3], Labelling Consortium[3‡], Ludovic Amruthalingam[2], Matthew Groh[4], Marc Pouly[2†], Alexander A. Navarini[1,3*†]

[1]Department of Biomedical Engineering, University of Basel, Hegenheimermattweg 167b, Allschwil, CH–4123, Basel, Switzerland.
[2]Department of Computer Science, Lucerne University of Applied Sciences and Arts, Suurstoffi 1, Rotkreuz, CH–6343, Zug, Switzerland.
[3]Department of Dermatology, University Hospital of Basel, Burgfelderstrasse 101, Basel, CH–4055, Basel, Switzerland.
[4]Kellogg School of Management, Northwestern University, 2211 Campus Drive, Evanston, US–60208, Illinois, United States of America.

*Corresponding author(s). E-mail(s): fabian.groeger@unibas.ch; alexander.navarini@usb.ch;
Contributing authors: simone.lionetti@hslu.ch; philippe.gottfrois@unibas.ch; alvaro.gonzalezjimenez@usb.ch; lea.habermacher@usb.ch; ludovic.amruthalingam@hslu.ch; matthew.groh@kellogg.northwestern.edu; marc.pouly@hslu.ch;
‡Labelling Consortium: Maximilian Boehm, Sara Cerminara, Georgios Nikolakis, and Alexander Romswinkel.
†These authors contributed equally to this work.

Abstract

The adoption of artificial intelligence in dermatology promises democratized access to healthcare, but model reliability depends on the quality and comprehensiveness of the data fueling these models. Despite rapid growth in publicly available dermatology images, the field lacks quantitative key performance indicators to measure whether new datasets expand clinical coverage or merely replicate what is already known. Here we present SkinMap, a multi-modal framework for the first comprehensive audit of the field's entire data basis. We unify the publicly available dermatology datasets into a single, queryable semantic atlas comprising more than 1.1 million images of skin conditions and quantify (i) informational novelty over time, (ii) dataset redundancy, and (iii) representation gaps across demographics and diagnoses. Despite exponential growth in dataset sizes, informational novelty across time has somewhat plateaued: Some clusters, such as common neoplasms on fair skin, are densely populated, while underrepresented skin types and many rare diseases remain unaddressed. We further identify structural gaps in coverage: Darker skin tones (Fitzpatrick V–VI) constitute only 5.8% of images and pediatric patients only 3.0%, while many rare diseases and phenotype combinations remain sparsely represented. SkinMap provides infrastructure to measure blind spots and steer strategic data acquisition toward undercovered regions of clinical space.

Keywords: Artificial intelligence, Dermatology, Health equity, Data diversity

Introduction

The integration of artificial intelligence (AI) into medicine revolutionizes many clinical specialties, including dermatology. However, the performance and clinical reliability of these models are fundamentally dependent on the quality, diversity, and scale of the data on which they are trained[1–4]. For years, the dermatology community has pointed to data scarcity, particularly diversity-related issues, as a primary obstacle to progress[5–8], but the claim has remained largely unquantified. This gap masks deeper issues: Without a systematic audit of the global data landscape, the research community risks investing resources in collecting redundant data, while biases and blind spots go unaddressed. Furthermore, models trained on unexamined corpora of data can inadvertently learn and amplify biases related to skin tones, rare conditions, or specific demographics, leading to tools that fail in real-world clinical settings and

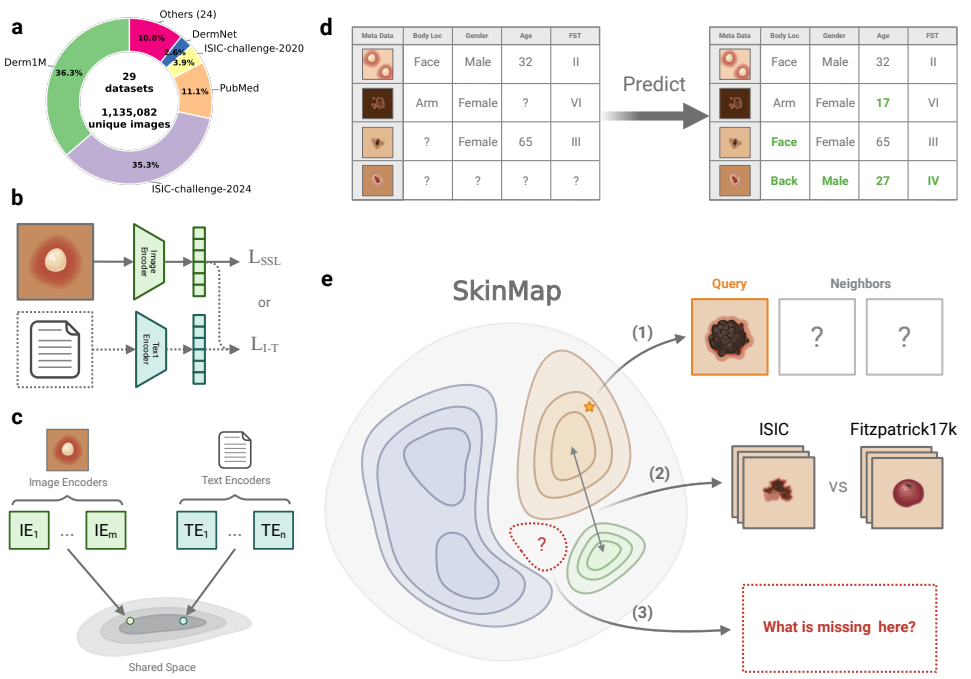


Fig. 1 The SkinMap multi-modal framework for auditing the global dermatology landscape. a, Aggregation of 29 public datasets yielding 1.1 million unique images with sparse original metadata. **b**, Multi-modal training pipeline utilizing self-supervised learning (SSL) on imaging data and contrastive learning on image-text pairs generated via templated captions. **c**, Construction of a unified latent space by projecting diverse encoders into a shared low-dimensional manifold. **d**, Imputation engine training, where linear probes learn to predict missing attributes (*e.g.*, Fitzpatrick skin type, age, and sex) using partial meta-information. **e**, Applications of the resulting digital SkinMap atlas: (1) A clinician-facing search tool for finding semantically similar cases; (2) a structural audit tool for identifying dataset overlaps and homologies; and (3) a strategic acquisition guide for detecting underrepresented latent regions to direct future data collection.

potentially worsen existing health disparities. Previous initiatives to improve the data landscape have been largely unguided, focusing on presumed gaps and biases rather than those identified through systematic quantitative analysis[6;9;10]. Although these contributions are valuable, they lack a comprehensive perspective because a reproducible and scalable methodology to measure dataset coverage and marginal utility is missing. Consequently, the field operates without key performance indicators to quantify whether new data expand the clinical map or reinforce existing blind spots.

Here, we introduce SkinMap, a framework that unifies the fragmented global archive into a queryable, interactive digital atlas. Visualizing the data landscape in real time enables clinicians and researchers to search for clinically similar cases and to transparently assess data availability across collections. To achieve this, we construct a shared semantic embedding space that captures clinical similarities across datasets. We aggregate 1,135,082 unique images from 29 public datasets, ranging from leading archives such as `ISIC`[11] to specialized cohorts[6;9;12–24], and harmonize them to enable joint analysis. By employing a multi-modal approach that aligns images with standardized natural language captions, and combining them in an ensemble with purely image-trained models, we project these disparate samples into a common latent space where proximity corresponds to visual and clinical similarity. Crucially, this shared space enables us to quantitatively assess the central issue of unharmonized and missing metadata. Using the latent representations of annotated samples, we infer missing clinical attributes, such as Fitzpatrick skin type (FST), age, gender, and geographic origin, across the entire collection with clinically validated, high-accuracy. This imputation transforms the sparsely annotated archive into an actionable resource, enabling us to analyze the demographics of the entire 1.1 million-image collection rather than only the labeled and harmonized subset. Following unification, our audit indicates that the dermatology data landscape remains unrepresentative of the global population. Current datasets are disproportionately composed of lightly pigmented skin images and are geographically concentrated in the Global North[5;9], whilst darker skin tones (FST V–VI) comprise only 5.8% of available image data. Furthermore, a temporal analysis of data acquisition reveals that despite exponential growth in dataset sizes, novelty has plateaued across both visual information and associated labels, revealing a systemic inefficiency in current data collection strategies. Rather than exploring high-gain gaps in rare conditions and underrepresented demographics, the community is investing scarce resources in the redundant accumulation of common neoplasms among

fair-skinned individuals. Finally, by quantifying how similar datasets truly are, we show that several nominally distinct datasets share near-identical representations, implying that many so-called external validations in the literature may effectively be internal validations. This unrecognized data leakage artificially inflates performance metrics and risks deploying models that fail to generalize to truly new patient populations.

Ultimately, SkinMap shifts the paradigm from unguided data accumulation to strategic acquisition. We delineate high-priority data needs, identify critical gaps such as healthy skin with FST V–VI and specific nail pathologies, and provide a concrete, evidence-based road map for guiding future data collection and collaboration. The SkinMap framework constitutes a generalizable blueprint for other medical imaging domains, such as radiology and pathology, setting a new standard for auditing and expanding the data foundations upon which the next generation of medical AI will be built. To support these research efforts, we release the complete SkinMap framework, including digital atlas and the model ensemble, under an open-source license.

Results

Quantifying demographic biases

The vast majority of the 1.1 million images lack a complete set of demographic metadata, such as skin tone, age, and gender. To address this, we utilized the SkinMap embedding space to impute missing attributes, expanding the metadata coverage of the dataset by an average of 50.7 percentage points (Fig. 2c). We validated this imputation engine on an evaluation set and two datasets (`DDI`[6] and `PAD-UFES-20`[17]) that were strictly held out, *i.e.*, excluded from pretraining. The SkinMap ensemble demonstrates superior performance, robustly outperforming state-of-the-art foundation models, including MONET[25] and PanDerm[26], in attribute prediction accuracy (Fig. 2a–b, Supplementary Table 6 for statistical confidence intervals). The strong concordance between the ground-truth subset and the fully imputed archive confirms

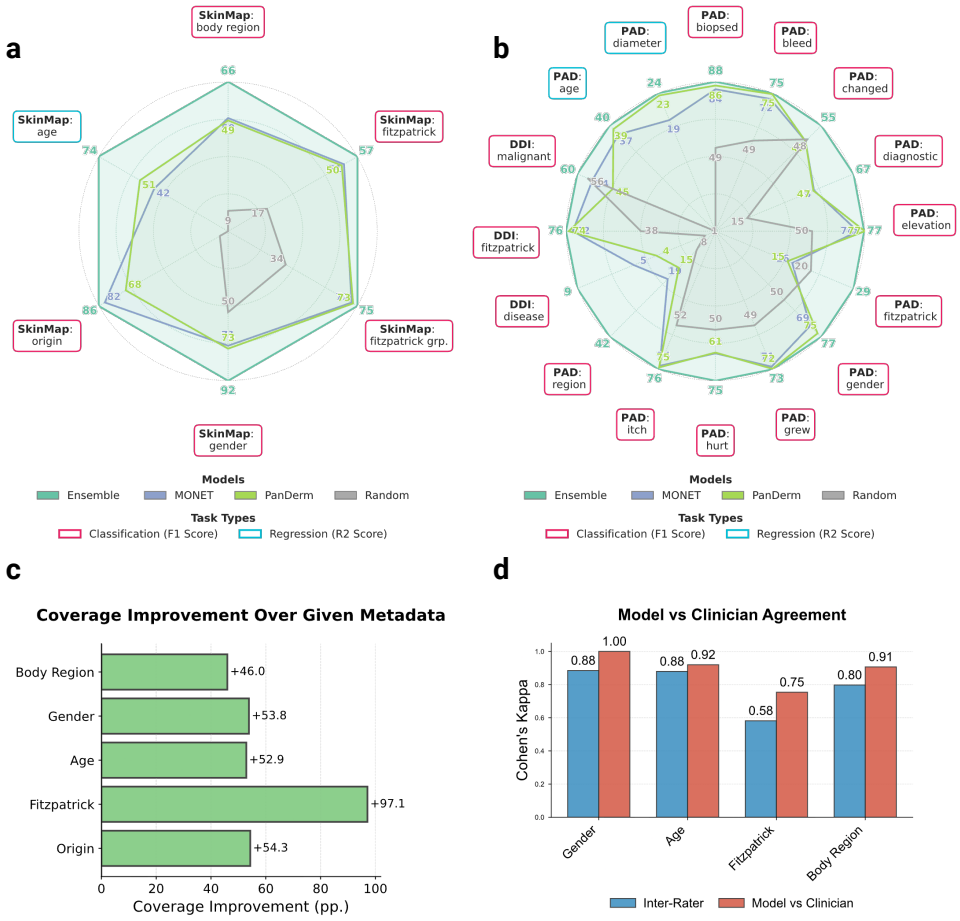


Fig. 2 Benchmarking imputation precision and metadata expansion. a, **b**, Radar plots comparing the predictive performance of the SkinMap ensemble against state-of-the-art foundation models MONET and PanDerm across multiple attributes on the internal validation set (**a**) and external hold-out datasets (**b**) and against random performance. The Ensemble model demonstrates superior performance in recovering missing demographic labels. **c**, Quantification of metadata coverage expansion achieved via imputation, showing significant gains for key attributes such as FST (+97.1 pp.) and geographic origin (+54.3 pp.). **d**, Comparative analysis of the model's imputation accuracy versus experienced dermatologists on a subset of 150 diverse cases.

that the ensemble faithfully estimates the demographics without hallucinating diversity or introducing algorithmic bias (Supplementary Fig. 2). This is further confirmed by validating the predictions for 150 cases against experienced, practicing dermatologists (n=5), who agree across metadata more with the model than among themselves

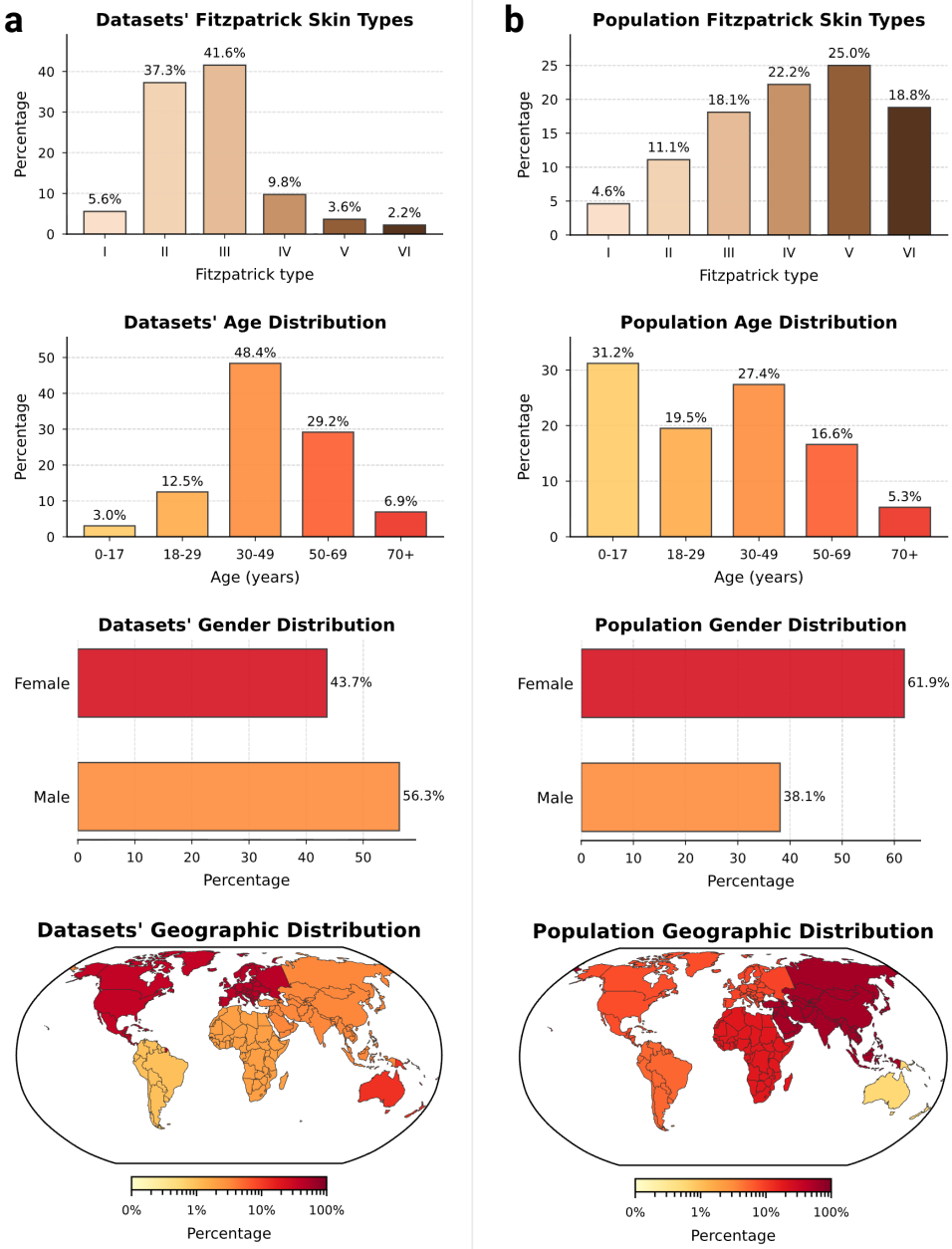


Fig. 3 Systemic demographic and geographic disparities in the global dermatological datasets. Distributions of demographic attributes across the aggregated digital datasets (left, **a**) compared with global population statistics for dermatological visits (right, **b**). The comparison reveals sharp divergences: FST V–VI are marginalized (5.8% in datasets vs. 43.8% globally), pediatric populations are underrepresented (3.0% vs. 31.2%), and geographic origins are heavily skewed toward the Global North.

(Fig. 2d). This densely annotated dataset overcomes the limitations of the original, sparse labels to assess the demographic composition of the entire digital archive.

Comparing the demographic distribution in the available data with expected population demographics reveals a significant divergence between dataset collection and global demographics (Fig. 3a–b). To refute the argument that data scarcity merely reflects workforce shortages, we benchmarked against clinical prevalence statistics, *i.e.*, specifically outpatient encounters[27] and diverse study cohorts[28], alongside WHO global standards[29;30]. The most profound disparity exists in skin tone: While FST III–VI comprise about 84.1% of the global dermatology-visit baseline, the available datasets are skewed predominantly toward FST II–III (78.9%), with darker skin tones (FST V–VI) constituting only 5.8%. There is an equally strong geographical bias, elsewhere often referred to as a form of "data colonialism"[31], where datasets originate almost exclusively from the Global North (North America, Western Europe) and Australia, leaving the majority of the human population in the Global South virtually unrepresented (Fig. 3c, bottom). The underrepresented regions of the map highlight an urgent need for targeted, infrastructure-level data collection partnerships with the Global South and across Asia. We similarly observe significant shifts in age and gender. The data collection centers on middle age (48.4% aged 30–49), with the pediatric population (0–17 years) marginalized at 3.0% compared to 31.2% globally. This gap likely reflects a combination of true prevalence differences in dominant skin-cancer-focused archives and stricter consent/privacy constraints around pediatric imaging. Nevertheless, even within general dermatology cohorts, common pediatric conditions (*e.g.*, eczema across diverse skin tones) remain sparsely represented, leaving important portions of pediatric clinical space effectively unmapped. While gender approaches parity (43.7% Female), intersectional imbalances compound in specific subgroups, *e.g.*, Female aged 50+ (Supplementary Information C.1.1). In summary, the demographic distribution represented by current dermatological data collections is a skewed

reflection of the dermatological population: It is predominantly white, western, and middle-aged. By quantifying these mismatches, SkinMap moves the field from assuming the existence of bias to precisely knowing where data support is missing, providing a road map for the targeted data acquisition required to build truly global medical AI.

Diminishing novelty in data acquisition

Model performance is often assumed to scale with data volume[32]. Our temporal audit suggests the field is nearing a saturation regime in which additional samples contribute less and less novel information. We quantify novelty as the embedding-space dissimilarity of collected samples relative to previously available data. To separate the exploration of new regions from the effect of increasing the sample size of the existing distribution, we compare the collected data against a dataset that is simply resampled from the historical pool. We define *yearly novelty* as the average distance of the samples released in a given calendar year from their neighbors, divided by the same average for the data resampled from previous years. Against this baseline, we observe a clear decline: Increases in raw sample counts no longer yield proportional gains in embedding-space novelty (Fig. 4a), indicating diminishing marginal informational return under our operational definition. Despite the large number of new samples collected, exceeding 1,000,000 in 2017–2025, novelty has plateaued near the random baseline. The main exception occurs in 2024, when the addition of total body photography (TBP) imagery from `ISIC-challenge-2024` produces a transient spike in novelty. Together, these results indicate that recent datasets predominantly land within already-populated regions of the data space, and that future gains will require targeted collection in underrepresented regions rather than indiscriminate scaling.

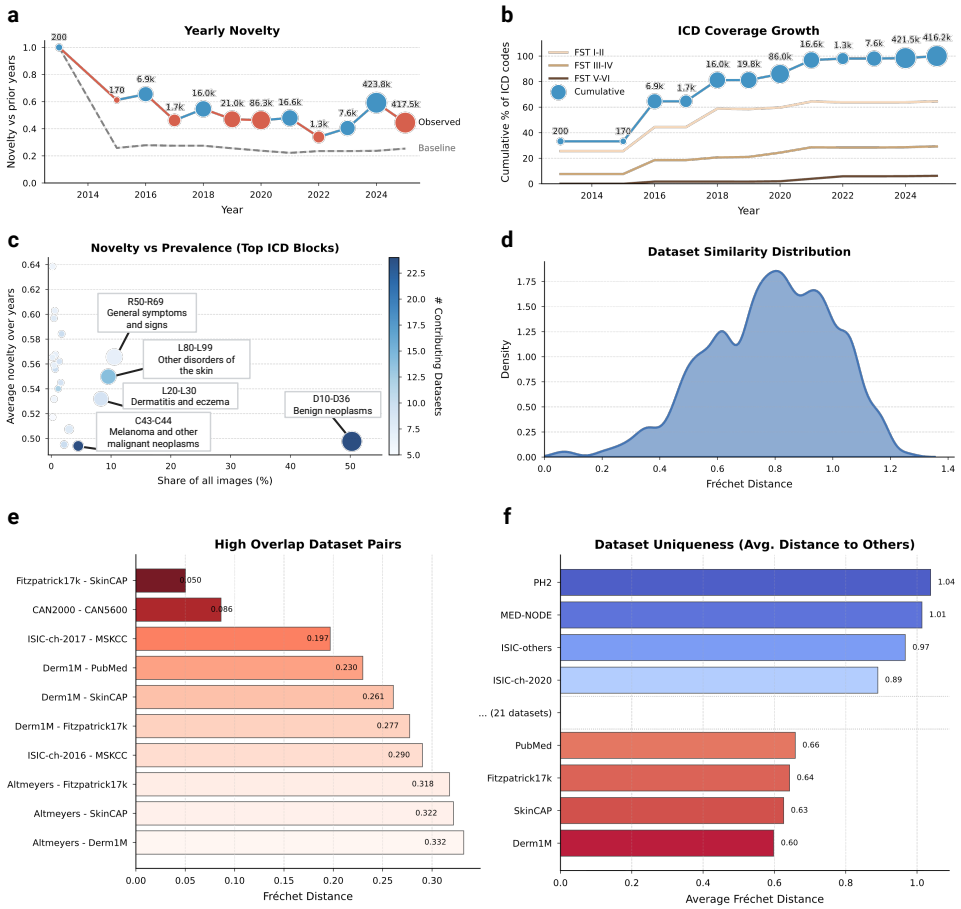


Fig. 4 Diminishing informational novelty and latent dataset redundancy. a, Temporal analysis of data acquisition showing the yearly novelty (blue or red depending on whether it increases or decreases relative to the previous year), defined as the average embedding distance of new samples relative to the historical pool. Gray shows the random baseline, which represents the expected novelty if new releases were merely random resamples of the historical data pool. **b**, Cumulative percentage of ICD-10 codes covered over time, showing a plateau in diagnostic diversity despite increasing data volume. **c**, Novelty versus prevalence for top ICD blocks; high-volume categories (*e.g.*, *benign neoplasms*) exhibit low novelty, while high-novelty categories (*e.g.*, *general symptoms*) remain rare. **d**, Density of pairwise distance between collections, skewed toward low values, indicating high redundancy. **e**, Identification of "high overlap" pairs, revealing hidden subset relationships. **f**, "Dataset uniqueness" scores, highlighting outliers like `PH2` as high-value sources of distributional variance.

To determine if this saturation stems from visual redundancy or a lack of clinical diversity, we analyzed the yearly international classification of diseases (ICD)

coverage growth stratified by skin tone (Fig. 4b). The analysis reveals a stark divergence: While disease diversity on fair skin (FST I–II) continues to expand, albeit at a reduced rate, coverage for darker skin tones (FST V–VI) has flatlined at near-zero levels. This confirms that the field is predominantly including missing diseases for white populations, while the clinical spectrum for the global majority remains unmapped. By investigating novelty against prevalence in datasets (Fig. 4c), we localized this saturation to specific diagnostic categories. The current landscape is heavily weighted towards collecting *benign neoplasms* (D10–D36) and *melanoma* (C43–C44), which provide high sample volumes but low novelty scores. Conversely, high-novelty categories such as *general symptoms and signs* (R50–R69) remain poorly addressed in terms of both image volume and contributing datasets.

These findings suggest that without a strategic shift towards underrepresented phenotypes, *i.e.*, specifically rare pathologies and diverse skin tones, scaling existing archives yields diminishing informational returns. Continued accumulation in high-density clusters (*i.e.*, FST II–III, *common neoplasms*) may lead to minor diagnostic performance gains in arguably important cases, but contributes little to the model's ability to recognize rare or atypical presentations.

Evaluating dataset similarity and redundancy

The current dermatology open data landscape comprises 29 nominally distinct datasets. To quantify how distinct these datasets are in practice, we measured pairwise distances in the SkinMap embedding space as a proxy for mutual diversity. The resulting similarity density (Fig. 4d) is skewed toward low distances, implying that most public datasets differ only modestly and do not constitute clearly separable clinical populations. Consistent with this, we identify several high-overlap pairs (Fig. 4e) with near-identical statistical structure, including established

relationships (*e.g.*, `Fitzpatrick17k`–`SkinCAP` and `CAN2000`–`CAN5600`) and substantial similarities between annual challenge datasets and their parent institutional archives (*e.g.*, `ISIC-challenge-2017`–`MSKCC`, `Derm1M`–`PubMed`, and `Derm1M`–`SkinCAP`). Although such overlap may be expected when data originate from shared institutions, it has direct implications for evaluation: Validation on a related dataset can masquerade as "external" testing while effectively reproducing internal validation, artificially inflating reported performance, and obscuring gaps in real-world generalization.

To guide future data selection, we propose a score for dataset uniqueness (Fig. 4f), computed as the average distance of a dataset to the global collection of dermatology. General-purpose datasets such as `Derm1M` are close to the statistical center of the map, indicating that they represent typical patient profiles currently available to AI. Undoubtedly useful for teaching models the basics, adding more such data yields limited gains in model robustness. In contrast, datasets such as `PH2` and `MED-NODE` populate the outer regions of the map. `PH2` and `MED-NODE` contribute distributional variance through a tightly controlled acquisition protocol and lesion characteristics that are comparatively uncommon in larger archives. These "outliers" are high-value assets: They contain the distinct, non-standard variations necessary to rigorously stress-test an AI's ability to handle diverse real-world populations.

Identifying structural gaps in the data landscape

To assess the coverage of the global dermatological data collection, we analyze the density of different regions of the SkinMap embedding space. We identify regions of hypersaturation and sparsity, *i.e.*, clusters with the highest and lowest 2.5% density quantiles (Supplementary Fig. 1), revealing that the atlas is polarized between massive redundancy and distinct structural voids. The high-density regions represent statistically overrepresented visual clinical presentations (Supplementary Fig. 1c). These regions are dominated by near-indistinguishable patches of skin or common

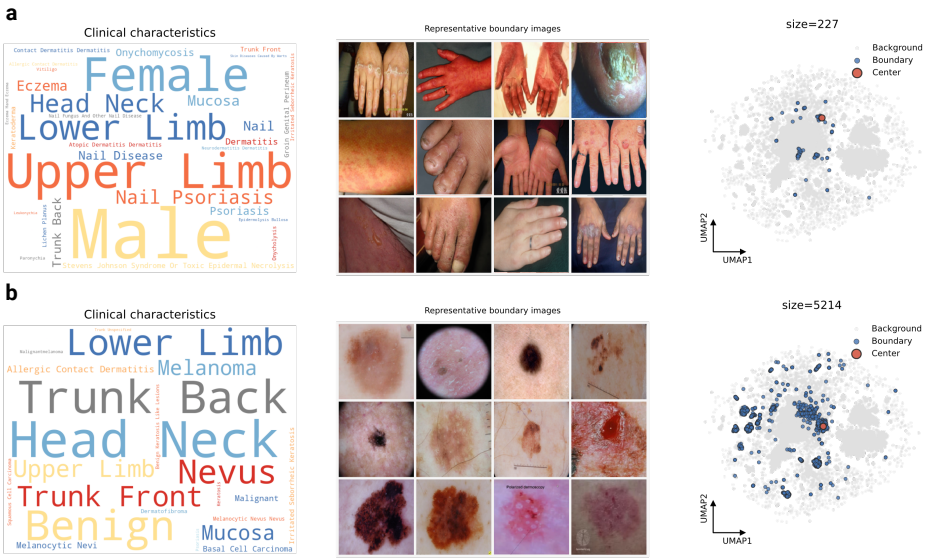


Fig. 5 Detection of topological holes in the embedding space. a, Detailed analysis of the highest-persistence void. The word cloud of metadata from boundary samples is dominated by *nail*, *nail disease*, and *nail psoriasis*, confirming that this topological void represents a structural absence of nail pathologies. **b**, Analysis of a second void bordered by diverse dermoscopy images of the *lower limb*, *head*, and *neck*. The existence of this gap indicates a structural lack of varying conditions and intermediate presentations within this modality in the current archive.

moles, confirming that continued accumulation in these regions yields diminishing returns. Conversely, low-density regions often highlight data-quality issues such as images unrelated to dermatology, reconstruction artifacts, and heavily occluded images (Supplementary Fig. 1d). Thus, embedding density can also serve as an automated quality-assurance mechanism, *i.e.*, by pruning redundant clusters and filtering outliers.

Moreover, we applied spectral persistent homology[33] to detect topological holes, *i.e.*, void regions with almost no cases (Fig. 5). Unlike simple low-density regions, these holes represent "unknown unknowns": Valid clinical phenotypes that are structurally absent from the training data. Crucially, this analysis independently confirms our metadata findings (Supplementary Information C.1.1). We identify a topological hole specifically circumscribed by nail pathologies (Fig. 5a). While the boundary contains scattered examples of *onychodystrophy* and *nail psoriasis*, its existence indicates

a systemic lack of variation in this anatomy. This confirms that the scarcity of nail conditions identified in our metadata audit is not merely a labelling issue, but a fundamental visual gap in the global atlas. By identifying these voids, SkinMap enables targeted data collection by specifying which phenotypes and demographic intersections are missing from current public archives. In the global dermatology data basis, high-priority gaps include nail disorders, pediatric dermatoses, especially in darker skin tones, and low-prevalence ICD blocks that appear only transiently in the public record.

Digital atlas

To transform these findings from a static audit into an actionable resource, we developed a web-based digital atlas that enables the community to interact directly with the SkinMap embedding space. This interactive platform serves two primary research purposes: Exploratory comparative analysis of dermatological presentations and strategic steering of data collection initiatives. To support diagnostic workflows, the application allows practitioners to upload cases, which are projected in real time onto the shared latent manifold (Fig. 6a–b). By leveraging the unified collection of 1.1 million images, the system retrieves and displays the most semantically similar cases along with their ground-truth diagnoses, also allowing for dynamic filtering of attributes such as body region or FST. This capability effectively provides clinicians with a global reference standard, moving beyond the limitations of local institutional archives.

To validate the clinical utility of this digital atlas, we benchmarked retrieval performance using two hold-out datasets, `DDI` and `PAD-UFES-20`, measuring precision and recall across both same-dataset and cross-dataset (atlas-wide) settings (Fig. 6c–e). Our results indicate that same-dataset retrieval remains a relatively effective baseline, even in the highly challenging, long-tailed disease distributions characteristic of the `DDI` dataset. However, the performance is fundamentally constrained by the limited

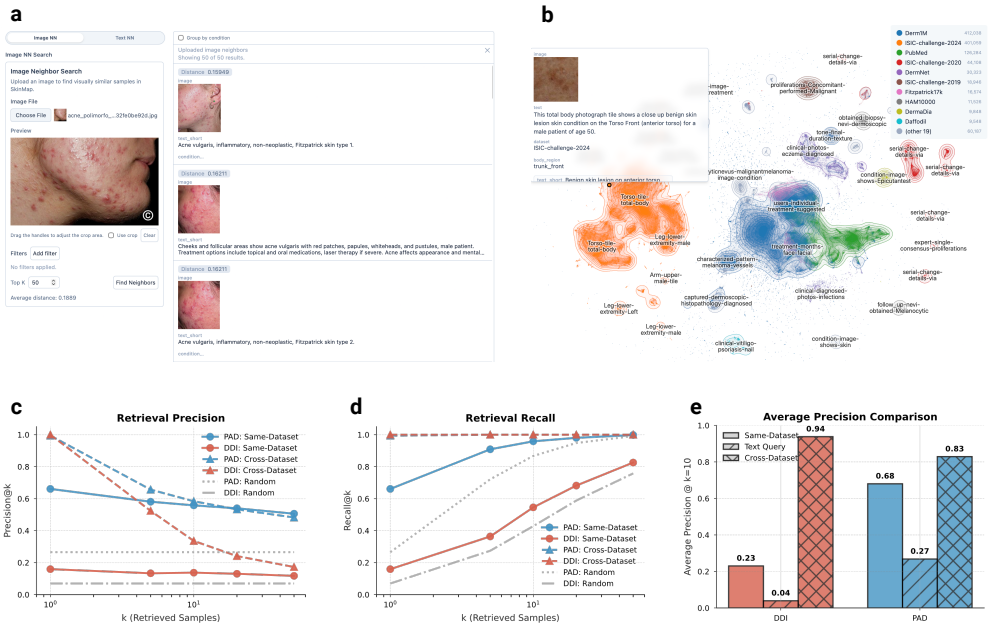


Fig. 6 The SkinMap digital atlas for clinical decision support and strategic data steering. a, Interactive web-based interface allowing for real-time case uploads and retrieval of semantically similar cases with associated ground-truth diagnoses. **b**, Visualization of the global latent space, projecting diverse datasets into a unified manifold where proximity corresponds to clinical similarity and allows practitioners to explore the atlas. **c**–**d**, Benchmarking of retrieval precision (**c**) and recall (**d**) on hold-out datasets, comparing baseline performance within the same dataset against cross-dataset retrieval using the complete atlas. **e**, Comparison of Average Precision (at $k = 10$) demonstrating that while same-dataset retrieval provides a baseline for specialized cohorts like `DDI`, cross-dataset retrieval through the exhaustive SkinMap atlas provides a significant performance boost, increasing `DDI` precision from 0.23 to 0.94.

sample volume of rare pathologies within single institutions. By expanding the search to the complete SkinMap atlas, we observe a substantial performance boost due to the system's ability to find semantically similar "neighbors" across the exhaustive 1.1 million-image collection (Fig. 6e). For the `DDI` dataset, average precision (at $k = 10$) increases from 0.23 in the same-dataset setting to 0.94 using the full atlas. Similarly, `PAD-UFES-20` retrieval precision improves from 0.68 to 0.83. This capability provides clinicians with a global reference standard, effectively bridging local data gaps by leveraging the collective informational density of the entire dermatological landscape.

By further enabling practitioners to query the SkinMap embedding space and compute average distances to known collections and samples, the platform provides the necessary infrastructure for strategic steering of data acquisition. By establishing quantitative KPIs for dataset diversity, the atlas enables funding bodies and research consortia to optimize resource investment, prioritizing high-gain samples that fulfill target goals over guesswork.

Discussion

For the past decade, the dermatology AI community has operated under the assumption of data bias, yet the scale and topography of this bias, as well as the marginal novelty of new data contributions, have remained unquantified. By unifying over 1.1 million fragmented images into a queryable semantic atlas, this study provides a quantitative audit of the open data foundations of modern dermatological AI systems. The findings expose a systemic divergence between the "digital population" used for model training and the global human population. Specifically, the representation of FST V and VI is restricted to 5.8% of the global dermatology collection, and pediatric data comprises only 3.0%. Rather than a simple lack of samples, these statistics indicate a structural exclusion where models are developed on a demographic distribution that is heavily skewed toward Western, middle-aged, and lightly pigmented populations. This geographic concentration, primarily in the Global North, suggests that without targeted intervention, algorithmic performance will likely not generalize to the populations bearing a significant portion of the global disease burden. Crucially, the temporal analysis challenges the prevailing assumption that dataset scale alone drives progress. The observed decoupling of data volume from informational novelty indicates that the strategy of indiscriminate data accumulation has reached a point of diminishing returns. Despite growth in dataset sizes, the novelty of new samples has plateaued. Similar patterns, in which output grows while disruptive novelty declines, have been

reported across science and technology[34;35], underscoring the need for explicit measures of marginal dataset contribution rather than relying solely on scale. Together with the density and demographic analyses, these results indicate that recent growth has disproportionately accrued in already high-density regions, specifically *benign neoplasms* on fair skin, whereas high-value, low-density regions, such as rare pathologies and diverse skin tones, remain stagnant. The persistence of these voids implies that continued aggregation of convenience samples will likely yield minimal improvements in model robustness and generalization.

Consequently, these findings advocate for a paradigm shift from volume-centric accumulation to strategic, precision data acquisition. SkinMap offers the infrastructure to operationalize this shift, serving as a mechanism to assess the marginal utility of potential data collections before resources are committed to annotation. By projecting new cohorts into a shared space, the community can distinguish between redundant contributions and high-priority samples that fill in topological voids. This metric-driven approach ensures that future infrastructure partnerships prioritize addressing structural gaps in the data landscape, thereby aligning the digital representation of dermatology with the diversity of real-world clinical practice. We hope that SkinMap will be a catalyst for a collaborative effort to maintain a comprehensive collection of digital skin datasets, thus accelerating dermatology in both clinics and research institutions. In practical terms, the atlas suggests a number of actionable acquisition directions: Expand geographically through infrastructure partnerships in underrepresented regions, build pediatric datasets under strengthened consent and privacy protections, increase coverage of underrepresented anatomical sites such as nails, and prioritize sustained acquisition of rare and non-neoplasm diagnostic categories to prevent "orphan" conditions from disappearing from the public record. These directions operationalize dataset innovation as measurable expansion of clinical coverage rather than continued densification of already-saturated clusters.

Methods

We built a global atlas of public dermatology imaging by (i) aggregating and harmonizing 29 datasets into a single corpus, (ii) converting standardized structured metadata into consistent image–text pairs, (iii) constructing the SkinMap embedding space via an ensemble of multi-modal and self-supervised encoders, (iv) imputing missing demographic and anatomical attributes using linear probes on frozen embeddings, and (v) auditing the resulting landscape with complementary metrics (*i.e.*, temporal novelty, dataset similarity, density, and topology analyses).

Data collection and harmonization

We conducted a comprehensive review and aggregation of the global public dermatology data landscape, identifying 29 publicly available datasets through a survey of biomedical repositories, challenge platforms, and published literature. We included datasets that were publicly accessible for research use and contained dermatology images (*i.e.*, clinical photography, dermoscopy, or TBP) with structured labels (*e.g.*, diagnosis) or textual description. A complete list of included datasets is provided in the Supplementary Information C.4. All images were retrieved from the official hosting source, and we preserved the original dataset identifiers to maintain traceability throughout the pipeline. To reduce duplicates, we performed identifier-based deduplication wherever reliable identifiers were available. Specifically, we used dataset-provided image/lesion/case identifiers and, where present, patient identifiers to detect repeated samples within the same source and across re-releases of the same organizational archive. For large multi-release archives (*e.g.*, `ISIC` and related derivatives), we cross-referenced identifiers across the different subsets/challenges and removed duplicates, retaining a single canonical instance per identifier. This effort yielded a collection of 1,135,082 unique images. The aggregated corpus is heterogeneous, incorporating dermoscopic images, clinical photography, and TBP from diverse sources

such as institutional archives or educational web resources. To use these disparate data sources for multi-modal training, we first manually harmonized the available structured metadata fields across datasets (*i.e.*, diagnosis, body site, age, gender, and origin), resolving dataset-specific conventions and synonymy into a unified schema. Body-site information and other free-text fields were manually mapped to a consistent controlled vocabulary (*e.g.*, anatomical regions), and age was standardized to common formats and units. After this harmonization step, we converted the structured metadata into natural-language descriptions to create consistent text-image pairs for multi-modal learning. We utilized a template description followed by a large language model (LLM) refinement step (Qwen2.5-7B-Instruct[36;37]) to generate concise, clinically coherent captions for all images, ensuring a consistent text-image pair format for subsequent contrastive learning (Supplementary Information C.3.1). The LLM step was constrained to stylistic normalization (*e.g.*, brevity and consistent phrasing) while preserving the underlying structured content. Additionally, dataset labels were manually assigned to ICD-10 codes whenever possible. Note, the ICD codes are used solely in subsequent analyses, whereas the original labels are used for model training.

SkinMap embedding space

To harmonize heterogeneous datasets into a single analyzable entity, we constructed the SkinMap embedding space using an ensemble learning approach that captures both high-level semantic concepts and low-level visual textures. All images were embedded using a standardized preprocessing pipeline, fixed input resolution, and normalization consistent with each encoder. The same pipeline was used for all analyses reported in this study. We first trained multiple multi-modal models from scratch on curated image-text pairs using different objectives, such as CLIP[38] and SigLIP[39], and on different partitions of the dataset. Partitions were constructed to reduce dataset-specific shortcut learning (*e.g.*, by mixing sources and holding out subsets during

individual model training). To leverage the full structural variance of the data without relying solely on possibly noisy metadata, we extended this with an ensemble of SSL models, specifically DINO[40], iBOT[41], and MAE[42], pretrained on the imaging data alone. For each encoder, we extracted a single global feature vector per image and ℓ_2-normalized features before aggregation. The final representation for each sample was generated by concatenating these diverse feature vectors and projecting them onto a shared, low-dimensional manifold via a learnable linear mapping, using a contrastive objective combined with the STRUCTURE regularization[43] to preserve consistency with the pretrained space. In practice, we used a fixed-dimensional representation for all analyses, *e.g.*, 1024 dimensions. Model and training hyperparameters, embedding dimensionalities for each backbone, and projector configuration are detailed in Supplementary Information C.3.2.

Metadata imputation

Given the sparsity of metadata in public collections, we used an imputation framework to estimate missing demographic labels. We trained a suite of linear probes on the frozen SkinMap embedding space using the subset of samples with ground-truth annotations. Separate probes were trained to predict each type of metadata to be recovered, namely origin, FST, age, gender, and anatomical region. This could be extended to other partially available information. Each probe was implemented as a linear model trained on top of the frozen embedding using `scikit-learn`[44]. We used `LogisticRegression` for categorical targets and `Ridge` regression for continuous targets. We validated the performance of the imputation engine on both an evaluation set and two external, distinct datasets (`DDI` and `PAD-UFES-20`) that were purposely excluded during pretraining. Model performance was quantified using standard metrics for each attribute (*e.g.*, macro-F1 for categorical labels and R^2 for continuous targets), and uncertainty was assessed via bootstrap confidence intervals. Detailed performance

comparison, including the evaluation of each model in the ensemble, is given in the Supplementary Table 4 and 5.

Audit metrics

Novelty over time

To assess diminishing returns in data acquisition, we defined a yearly novelty metric based on latent-space density. For each sample released in a given year, we computed the mean cosine distance to its k-nearest neighbors in the historical pool (all samples released in prior years). To distinguish genuine informational gain from increasing dataset density, we constructed a null hypothesis using bootstrap resampling. For each year, we generated a baseline distribution by randomly resampling points from the historical pool, matching the size of the new release, and computing their novelty scores. Observed novelty scores that converged toward this random baseline provided signal for saturation, indicating that new data releases were structurally redundant with existing archives.

Let $\{\mathbf{z}_i\}_{i=1}^{N} \subset \mathbb{R}^{d\times N}$ denote the samples' embeddings and $y_i \in \mathbb{Z}$ the associated release year. We first ℓ_2-normalized every embedding $\hat{\mathbf{z}}_i = \mathbf{z}_i/\|\mathbf{z}_i\|_2$, so cosine similarity between any two embeddings matched their inner product. For a fixed calendar year t, we defined the index sets

$$\mathcal{I}_t = \{i : y_i = t\}, \qquad \mathcal{P}_t = \{j : y_j < t\}, \tag{1}$$

corresponding to current-year samples $\mathcal{I}_t$ and the historical pool $\mathcal{P}_t$, respectively. For each sample $i \in \mathcal{I}_t$, we computed cosine similarities to $\mathcal{P}_t$, took the k nearest neighbors, and converted them to distances,

$$d_{i,\ell} = 1 - \hat{\mathbf{z}}_i^{\top}\hat{\mathbf{z}}_{\pi_{i,\ell}}, \qquad \ell = 1, \ldots, k, \tag{2}$$

where $\pi_{i,\ell}$ indexes the ℓ-th neighbor of i found via a cosine-similarity search. Averaging the k distances gave a per-sample novelty score

$$\bar{d}_i = \frac{1}{k} \sum_{\ell=1}^{k} d_{i,\ell}, \tag{3}$$

and the raw yearly novelty was defined as the mean over all current-year samples,

$$\nu_t = \frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} \bar{d}_i. \tag{4}$$

To disentangle true informational novelty from statistical artifacts of increasing pool sizes, we constructed a null hypothesis via bootstrap resampling. For each year t, we performed B bootstrap iterations (with $B = 200$). In each iteration b, we randomly sampled (with replacement) $|\mathcal{I}_t|$ points from the historical pool $\mathcal{P}_t$ to form a pseudo-cohort $\mathcal{B}_t^{(b)}$. We then computed their k-nearest neighbor distances within $\mathcal{P}_t$ (excluding self-matches), yielding a baseline novelty score

$$\nu_t^{(b)} = \frac{1}{|\mathcal{I}_t|} \sum_{j \in \mathcal{B}_t^{(b)}} \bar{d}_j, \tag{5}$$

where $\bar{d}_j$ is computed analogously to the observed case but with one additional neighbor to exclude the trivial self-match. The empirical distribution $\{\nu_t^{(b)}\}_{b=1}^{B}$ characterizes the expected novelty under the null hypothesis that new samples are merely redrawn from previous years. We summarize this via the mean

$$\nu_t^{\circ} = \frac{1}{B} \sum_{b=1}^{B} \nu_t^{(b)} \tag{6}$$

and construct confidence intervals using the $\alpha/2$ and $1 - \alpha/2$ quantiles (here $\alpha = 0.05$ for 95% CI). The ratio ν_t/ν_t° quantifies how much more novel the observed data is

compared to random resampling, with values near 1 indicating saturation and values > 1 indicating the population of previously uncovered regions.

Dataset similarity and uniqueness

We quantified the similarity between distinct datasets using the Fréchet distance (FD) to assess shifts in means and covariances of their distributions. For each dataset, we modeled its distribution in the SkinMap latent space as a multivariate Gaussian distribution with empirical mean and covariance. Specifically, pairwise Fréchet distance (FD) was computed across all 29 collections. Pairs exhibiting low FD (approaching zero) were identified as homologous, indicating that they occupy nearly identical latent manifolds and likely represent subsets or re-releases of the same data lineage. Conversely, we defined the uniqueness of a dataset as the average FD of a collection relative to all others. This metric was used to identify high-value outliers that contribute significant distributional variance to the global atlas.

For each dataset a, we considered the normalized embeddings belonging to that dataset, $\{\hat{\mathbf{z}}_i : i \in \mathcal{D}_a\}$. We approximated the distribution with its empirical first and second moments:

$$\boldsymbol{\mu}_a = \frac{1}{|\mathcal{D}_a|} \sum_{i \in \mathcal{D}_a} \hat{\mathbf{z}}_i, \qquad \Sigma_a = \frac{1}{|\mathcal{D}_a| - 1} \sum_{i \in \mathcal{D}_a} (\hat{\mathbf{z}}_i - \boldsymbol{\mu}_a)(\hat{\mathbf{z}}_i - \boldsymbol{\mu}_a)^\top. \tag{7}$$

Given any two datasets a and b, we summarized their distributional discrepancy with the Fréchet distance between the corresponding Gaussians[45],

$$\mathrm{FD}(a, b) = \|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|_2^2 + \mathrm{Tr}\Big(\Sigma_a + \Sigma_b - 2(\Sigma_a \Sigma_b)^{1/2}\Big). \tag{8}$$

This scalar captures both location shift (difference in means) and shape shift (covariance mismatch) between datasets under the assumed Gaussian approximation.

Density, redundancy, and hole detection

To identify regions of redundancy and data quality failures within the landscape, we modeled the embedding space density using a Gaussian mixture model. We defined operational thresholds at the 97.5th and 2.5th percentiles of log-density to partition the landscape into regions of hypersaturation and sparsity. High-density regions were analyzed to identify overrepresented visual concepts, while low-density regions were inspected to detect outliers, artifacts, and underrepresented clinical phenotypes.

To identify underrepresented regions (or *holes*) in the embedding space, we employ a topological data analysis approach based on the spectral persistent homology method recently developed by Damrich et al.[33]. Traditional persistent homology methods fail in high-dimensional spaces with noise because ambient Euclidean distances become uninformative. Damrich et al.[33] demonstrated that this limitation can be overcome by computing persistent homology on a filtered simplicial complex constructed from *effective resistance distances* on the data manifold, which remain robust to high-dimensional noise by capturing the intrinsic geometry of the data. We apply their method to the SkinMap embedding space to identify topological voids corresponding to underrepresented patient subgroups.

Effective resistance distance on k-nearest neighbor graphs. Let $Z = \{\mathbf{z}_i\}_{i=1}^N \subset \mathbb{R}^d$ denote the set of embedding vectors in the d-dimensional latent space. We construct a k-nearest neighbor graph $G = (V, E)$ where vertices V correspond to all data points and edges connect each point to its k nearest neighbors.

The graph is symmetrized by including edge (i, j) if j is among the k nearest neighbors of i *or* i is among the k nearest neighbors of j. Let $A \in \{0, 1\}^{N \times N}$ denote the resulting adjacency matrix. We define the degree matrix $D = \mathrm{diag}(d_1, \ldots, d_N)$ where $d_i = \sum_j A_{ij}$ and construct the graph Laplacian:

$$L = D - A. \tag{9}$$

The *effective resistance distance*[46] between vertices i and j is defined through the pseudoinverse of the Laplacian. Let $L^\dagger$ denote the Moore–Penrose pseudoinverse of L, computed via eigendecomposition with regularization to handle the zero eigenvalue corresponding to the constant eigenvector. The effective resistance distance is given by:

$$d_{ij}^{\text{eff}} = (L^\dagger)_{ii} + (L^\dagger)_{jj} - 2(L^\dagger)_{ij}. \tag{10}$$

As noted by Hein et al.[47], this naive formulation exhibits undesirable scaling properties in large graphs, where distances converge to approximately $1/d_i + 1/d_j$ independent of graph structure. Following Damrich et al.[33], we apply the von Luxburg degree correction[47], yielding the corrected effective resistance:

$$\tilde{d}_{ij}^{\text{eff}} = d_{ij}^{\text{eff}} - \frac{1}{d_i} - \frac{1}{d_j} + \frac{2A_{ij}}{d_i d_j} - \frac{A_{ii}}{d_i^2} - \frac{A_{jj}}{d_j^2}. \tag{11}$$

The corrected distance $\tilde{d}_{ij}^{\text{eff}}$ captures the intrinsic manifold geometry by measuring the "difficulty" of random walks between vertices, remaining stable in the presence of high-dimensional noise[33]. The resulting distance matrix $D^{\text{eff}} = [\tilde{d}_{ij}^{\text{eff}}]$ serves as input to the persistent homology computation.

Persistent homology and hole detection. Given the distance matrix D^{eff}, we construct a Vietoris–Rips filtration[48]. A k-simplex $\sigma = \{v_{i_0}, \ldots, v_{i_k}\}$ is included in the complex at filtration parameter τ if and only if the maximum pairwise effective resistance distance among its vertices satisfies:

$$\max_{\ell, m \in \{0,\ldots,k\}} \tilde{d}_{i_\ell i_m}^{\text{eff}} \leq \tau. \tag{12}$$

As τ increases from 0 to ∞, simplices are added to the complex, forming a nested sequence of simplicial complexes (the filtration). We compute the persistent homology

of this filtration using the Ripser algorithm[49], which efficiently extracts topological features across all scales.

A *hole* in the embedding space corresponds to a 1-dimensional homology class (denoted H_1)[48]. Each such class persists over a filtration interval $[\tau_{\text{birth}}, \tau_{\text{death}}]$, where τ_{birth} is the scale at which a loop first appears and τ_{death} is the scale at which it is filled. The *persistence* of the feature is defined as:

$$\text{persistence} = \tau_{\text{death}} - \tau_{\text{birth}}. \tag{13}$$

Features with high persistence represent topologically significant structures, while low-persistence features typically correspond to noise[48]. We rank all detected H_1 features by their persistence values and select the top k holes for further analysis, where k is determined based on the dataset characteristics and the desired granularity of the gap analysis.

Cocycle extraction and hole characterization. For each persistent H_1 feature, we extract its *cocycle* representation[49]: The set of edges (vertex pairs) that comprise the 1-cycle defining the hole. Let $\mathcal{C}_k = \{(i_1, j_1), (i_2, j_2), \ldots, (i_{|\mathcal{C}_k|}, j_{|\mathcal{C}_k|})\}$ denote the cocycle for the k-th hole. The unique vertices involved in this cocycle are:

$$\mathcal{V}_k = \{i : (i, j) \in \mathcal{C}_k \text{ or } (j, i) \in \mathcal{C}_k\}. \tag{14}$$

These vertices represent the data points forming the boundary of the hole. To enable clinical interpretation and prioritization of detected gaps, we characterize each hole by:

- *Hole center.* The mean position in latent space of all vertices in the cocycle:

$$\mathbf{c}_k = \frac{1}{|\mathcal{V}_k|} \sum_{i \in \mathcal{V}_k} \mathbf{z}_i. \tag{15}$$

- *Hole size.* The cardinality of the cocycle's vertex set: $s_k = |\mathcal{V}_k|$, representing the number of data points forming the hole.
- *Hole volume.* We estimate the volume occupied by the hole in latent space by computing the empirical covariance of the cocycle vertices and modeling it as a d-dimensional hypersphere. Let:

$$r_k = \text{median}_{i \in \mathcal{V}_k} \|\mathbf{z}_i - \mathbf{c}_k\| \tag{16}$$

  denote the characteristic radius. The hole volume is approximated as:

$$V_k = \frac{\pi^{d/2}}{\Gamma(d/2+1)} r_k^d, \tag{17}$$

  where Γ is the gamma function.
- *Boundary points.* To characterize the data surrounding each hole, we extract boundary samples by finding, for each vertex $i \in \mathcal{V}_k$, its k_b nearest neighbors in the full dataset Z. We filter these neighbors by distance, retaining only those within:

$$\|\mathbf{z}_j - \mathbf{z}_i\| \leq \alpha \cdot \text{median}_{i' \in \mathcal{V}_k} d_{i',k_b}, \tag{18}$$

  where d_{i',k_b} is the distance to the k_b-th nearest neighbor of vertex i' and α is a radius scale factor (here $\alpha = 1.5$). The union of these boundary points provides representative samples near the hole for qualitative inspection.

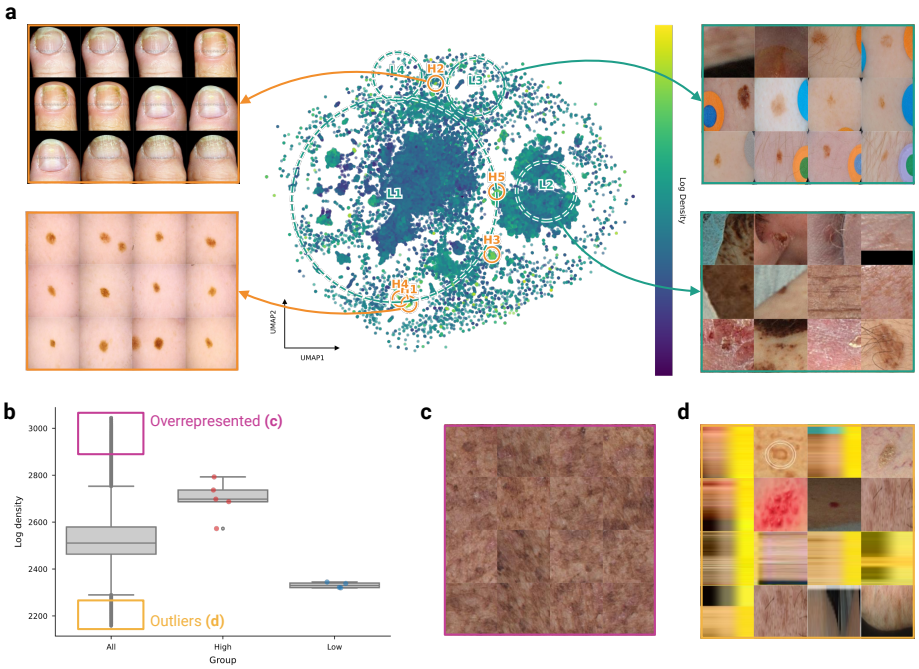
Acknowledgements. We thank Martin A. Walter for their valuable suggestions, which helped improve the clarity of the manuscript. This research was funded in part by the Swiss National Science Foundation (SNSF) under grant 20HW-1_228541. Figure 1 was created in BioRender. All other figures were generated programmatically (Python/matplotlib) from the analyses described.

Declarations

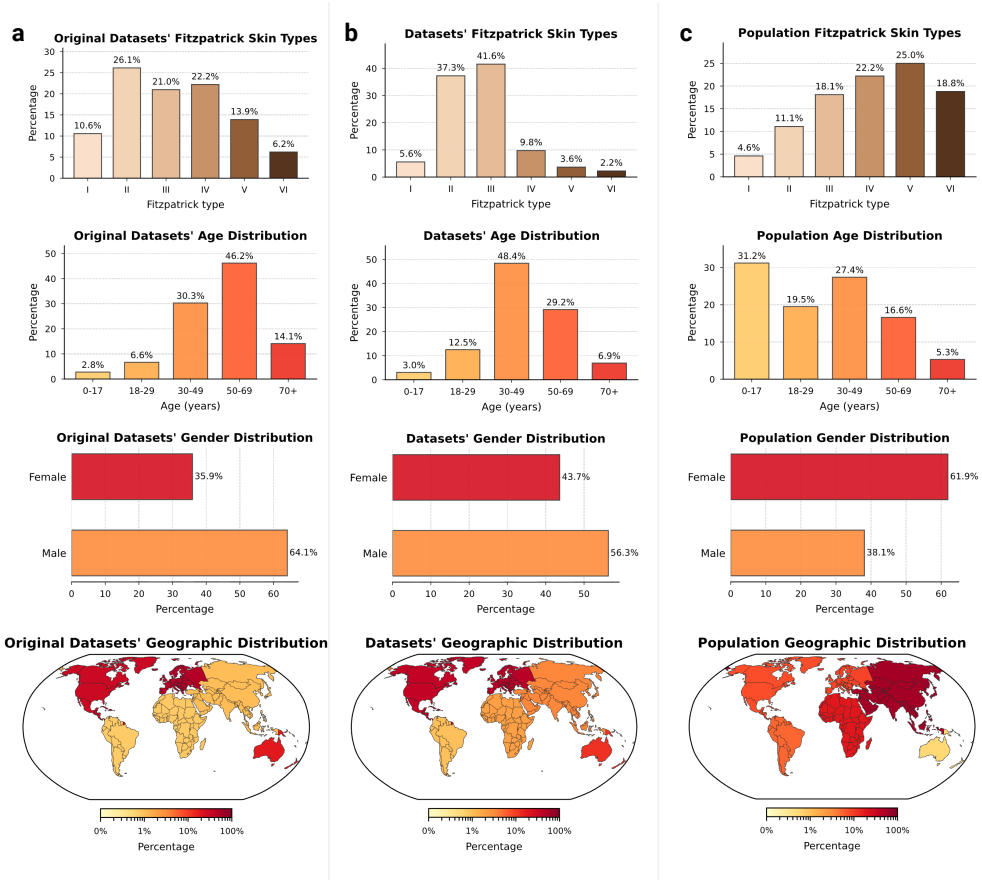
- **Funding:** This work was supported in part by the Swiss National Science Foundation (SNSF; Grant 20HW-1_228541).
- **Competing interests:** A.A.N. is a shareholder of Derma2Go AG and Derma.One GmbH. All other authors declare no competing interests.
- **Ethics approval and consent to participate:** Not applicable. This study used mainly publicly available datasets. For the verification of the model predictions with domain experts (Fig. 2), we solely collected answers with consent from the participants as categorical labels along with anonymized annotator identification. Thus, these annotations contain no personally identifiable information or offensive content. In discussions with experts from the co-authors' institutions, it was concluded that this verification process does not require ethics approval because the study examines publicly available datasets and does not involve human subjects beyond categorical annotations.
- **Consent for publication:** Not applicable.
- **Data availability:** All datasets used in this study are publicly available from their original sources under their respective terms of use. Complete list of datasets and access information is provided in the Supplementary Information C.4. The datasets used in this study are publicly available and were de-identified by the original data curators prior to their release. The authors of this study had no access to personally identifiable information (PII) or patient keys.

- **Materials availability:** Resources associated with SkinMap (including derived artifacts such as the embedding space and metadata harmonization resources, where redistribution is permitted) will be released under an open-source license and made publicly available upon publication.
- **Code availability:** The source code for the SkinMap framework, including inference scripts, pretrained model weights, and the projection head parameters used to generate the embedding space, will be made available upon publication. The interactive digital atlas will be accessible upon publication.
- **Author contributions:** Conceptualization: F. Gröger; Methodology: F. Gröger, S. Lionetti; Software: F. Gröger; Validation: F. Gröger, S. Lionetti, P. Gottfrois; Formal analysis: F. Gröger, S. Lionetti; Investigation: F. Gröger, P. Gottfrois; Data curation: F. Gröger, P. Gottfrois; Writing – original draft: F. Gröger; Writing – review & editing: All; Visualization: F. Gröger; Supervision: M. Groh, M. Pouly, A.A. Navarini; Funding acquisition: M. Pouly, A.A. Navarini.

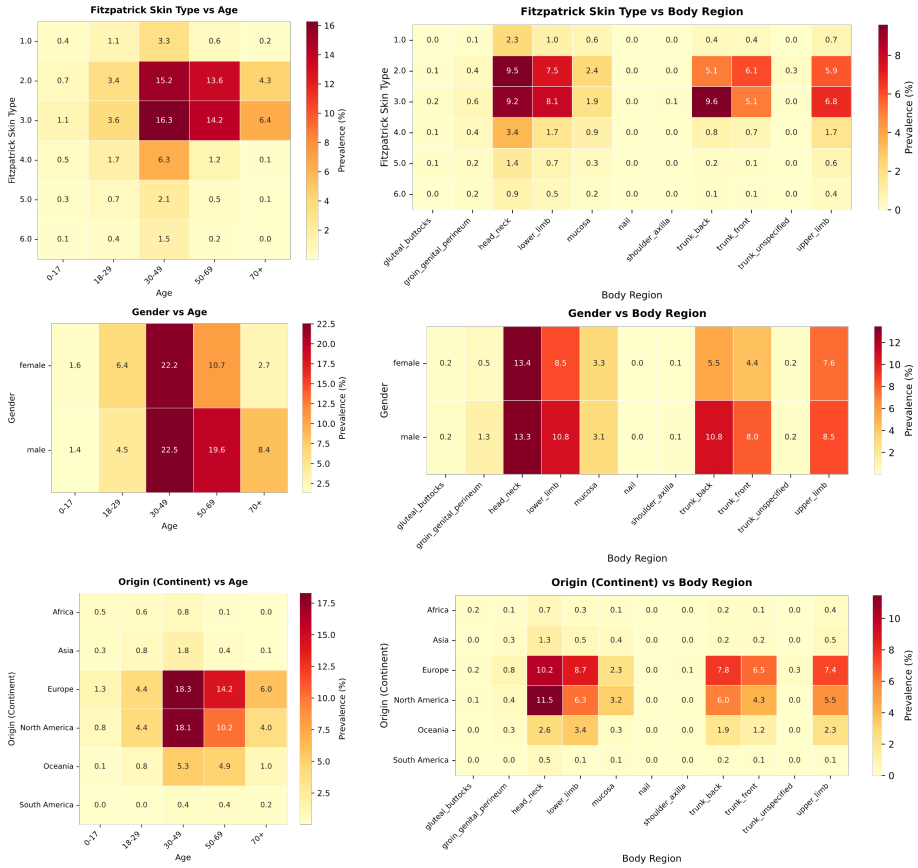
Appendix A Supplementary Figures



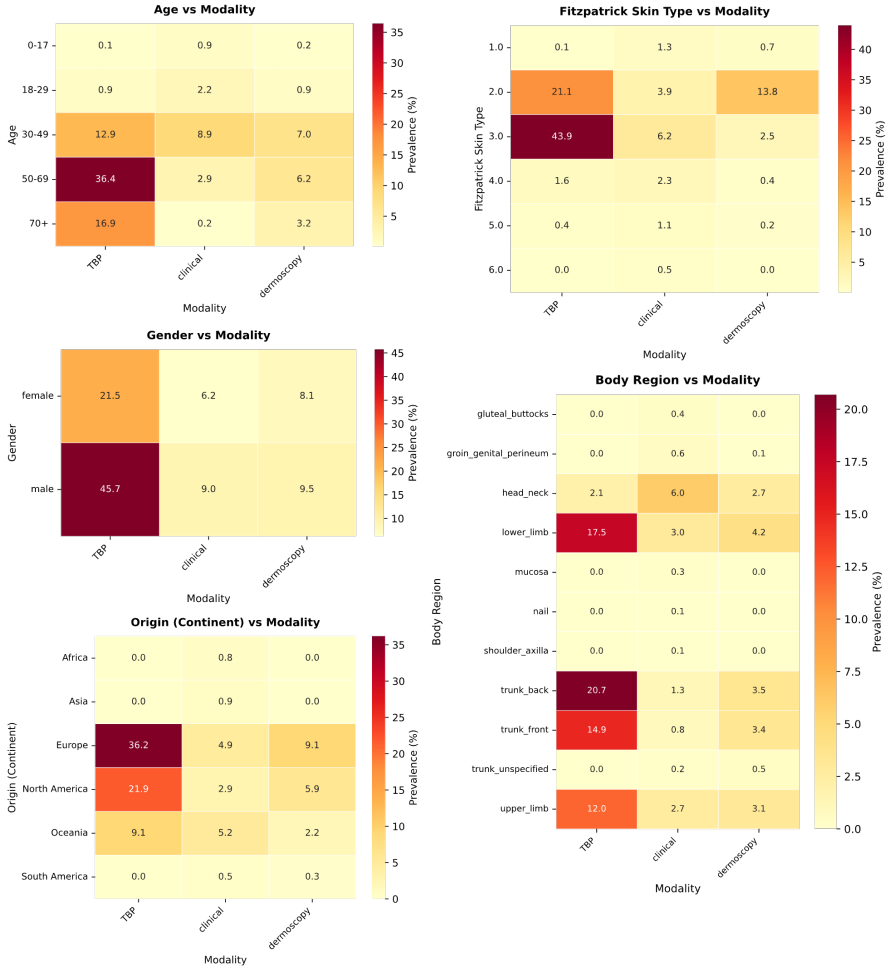
Supplementary Fig. 1 Embedding space density as an automated quality assurance mechanism. a, UMAP projection of the SkinMap embedding space overlaid with a Gaussian mixture model density map. Dashed circles indicate regions of hypersaturation (high density, $H1-H5$) and sparsity (low density, $L1-L4$). **b**, Boxplot distribution of log-density scores distinguishing core clusters from peripheral outliers. **c**, Representative samples from high-density peaks, revealing visual homogeneity and redundancy in common lesions. **d**, Representative samples from low-density regions, effectively isolating data quality failures, including non-clinical images, reconstruction artifacts, and heavy occlusions.



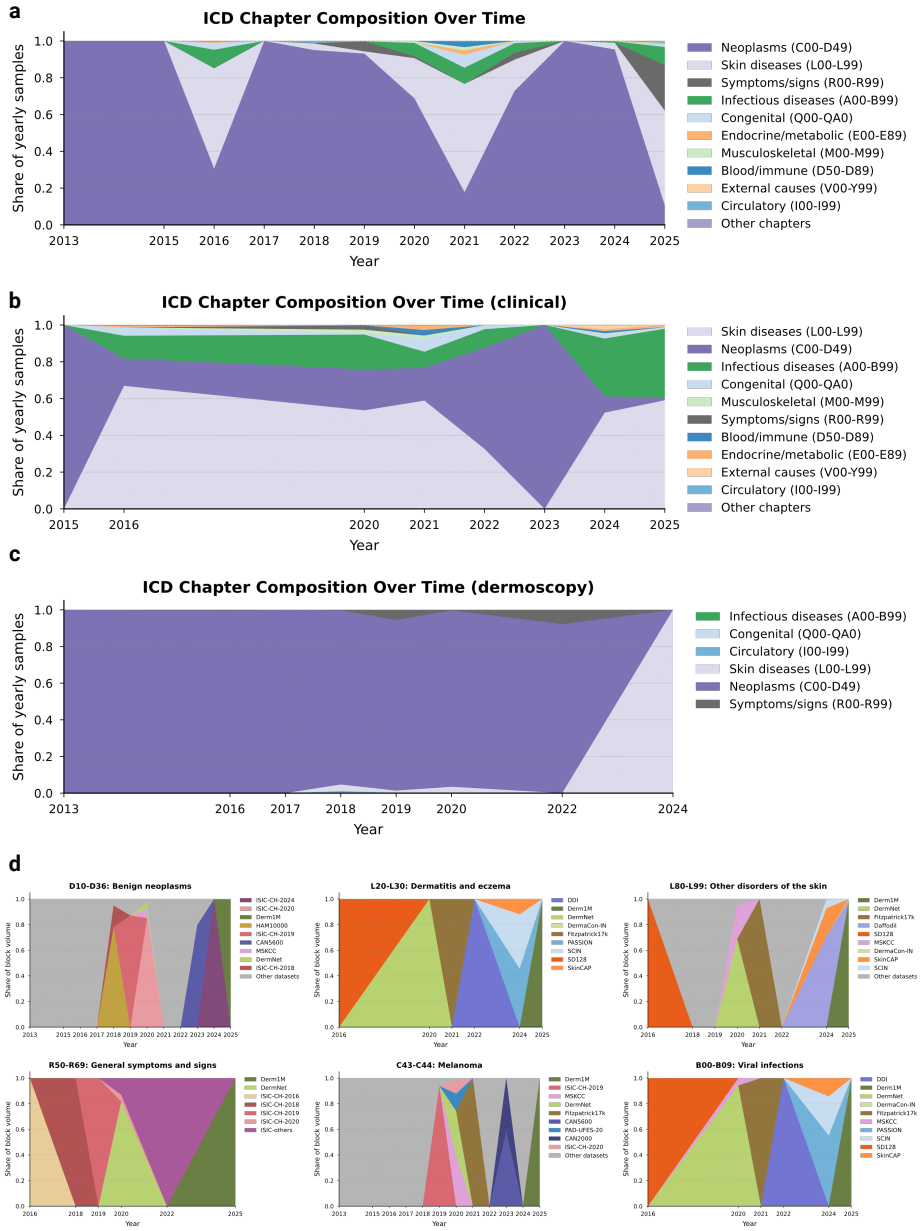
Supplementary Fig. 2 Validation of imputed demographic distributions. Comparison of demographic distributions derived from original sparse metadata (left column, **a**), the fully imputed SkinMap attributes (center column, **b**), and global population baselines (right column, **c**). The imputed distributions (center) effectively reconstruct the latent demographics of the dataset, revealing the true extent of exclusions (such as the scarcity of FST V–VI and pediatric data) that are otherwise obscured by missing metadata in the original archives.



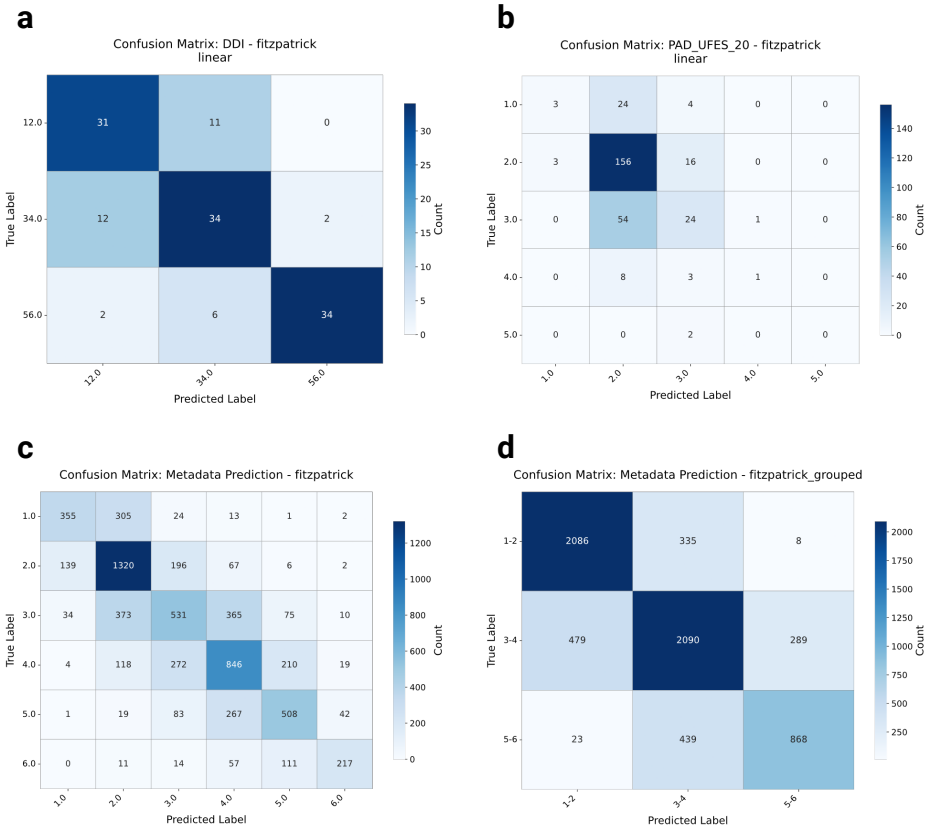
Supplementary Fig. 3 Intersectional metadata audit reveals compounded underrepresentation in demographic subgroups. Heatmaps displaying the prevalence of samples across intersecting demographic attributes (*i.e.*, FST, age, gender, geographic origin, and body region) in the global archive. The color gradient represents the prevalence of samples, ranging from high saturation (red) to scarcity (yellow). The analysis uncovers deep structural voids beyond single-axis statistics: The interaction between skin tone and age highlights the critical lack of pediatric data for darker skin tones (FST V-VI), where sample counts are negligible relative to adult FST II populations. Geographic origin versus skin tone confirms that, even when darker skin tones are present, they are often disconnected from their geographic centers of prevalence (*e.g.*, the Global South) and instead originate from Western datasets. Body region analysis reveals that non-standard anatomical locations on underrepresented demographics constitute "dark corners" of the latent space, potentially compromising diagnostic performance for complex clinical presentations. These intersectional gaps indicate that global metrics of diversity mask profound local voids, where specific patient subgroups (*e.g.*, elderly patients with FST VI) remain effectively invisible to current AI models.



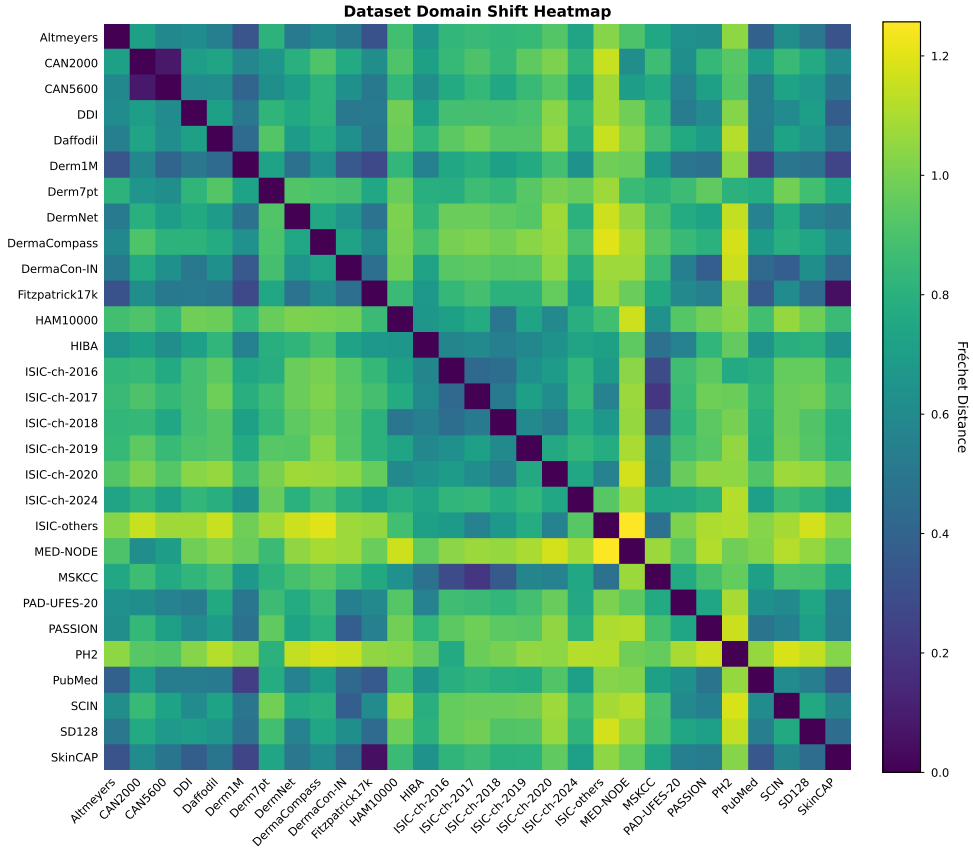
Supplementary Fig. 4 Modality-specific distribution and structural bottlenecks. Heatmaps displaying the prevalence of samples across intersections of imaging modality (*i.e.*, TBP, clinical, and dermoscopy) and demographic/clinical attributes. The color gradient represents sample prevalence, from scarcity (yellow) to hypersaturation (red).



Supplementary Fig. 5 Temporal evolution and provenance of ICD-10 disease categories. a, Stacked area chart displaying the proportional share of ICD-10 chapters in yearly data releases. The landscape has become increasingly dominated by *Neoplasms* (C00–D49) in recent years, correlating with the novelty saturation observed in the main analysis. **b**, Granular provenance of specific diagnostic blocks (colors represent source datasets). The visualization highlights that diversity is often driven by isolated, dataset-specific events (*e.g.*, `ISIC challenges` that drive melanoma counts) rather than by broad, continuous acquisition.



Supplementary Fig. 6 Confusion matrices for Fitzpatrick skin type imputation. Performance of the linear probe on **a**, the `DDI` external validation set; **b**, the `PAD-UFES-20` external validation set; and **c**, **d**, the internal hold-out test set for ungrouped (**c**) and grouped (**d**) skin tones. Errors are concentrated around the diagonal (adjacent skin types), indicating that the model learns a coherent ordinal representation of skin tone rather than making random categorical errors.



Supplementary Fig. 7 Pairwise Fréchet distance (FD) heatmap quantifying inter-dataset domain shifts. Matrix of FD between all 29 aggregated datasets. Dark blue indicates high similarity, statistically confirming subset relationships (*e.g.*, `Fitzpatrick17k`–`SkinCAP`) and institutional clusters (`ISIC archives`). Yellow/green bands identify high-uniqueness outliers that contribute significant distributional diversity to the global atlas.

Appendix B Supplementary Tables

Supplementary Table 1 Multi-modal backbone training hyperparameters. All CLIP and SigLIP models were trained with these settings unless otherwise specified.

Hyperparameter	Value
Optimizer	AdamW[50]
Base learning rate	5×10^{-6}
Weight decay	0.01
Batch size (per GPU)	32
Warmup ratio	0.1
LR schedule	Linear warmup + linear decay
Max gradient norm	1.0
Text token length	77
Mixed precision	FP16

Supplementary Table 2 Model ensemble composition. Each model contributes distinct feature representations to the final SkinMap space.

Architecture	Loss	Init	Epochs	Dataset
Multi-modal models				
ViT-Large/14[51]	CLIP[38]	Random	100	1/3 of dataset
ViT-Large/14[51]	CLIP[38]	Random	100	1/3 of dataset
ViT-Large/14[51]	CLIP[38]	Random	100	1/3 of dataset
ViT-Large/14[51]	SigLIP[39]	Random	100	1/3 of dataset
ViT-Large/14[51]	SigLIP[39]	Random	100	1/3 of dataset
ViT-Large/14[51]	SigLIP[39]	Random	100	1/3 of dataset
ViT-Base/32[51]	CLIP[38]	Random	100	full dataset
ViT-Base/32[51]	SigLIP[39]	Random	100	full dataset
Self-supervised models				
ViT-Base/16[51]	DINO[40]	Random	500	full image dataset
ViT-Base/16[51]	iBOT[41]	Random	500	full image dataset
ViT-Base/16[51]	MAE[42]	Random	500	full image dataset

Supplementary Table 3 Ensemble fusion and projector training hyperparameters. The projector maps concatenated multi-model embeddings to a unified 1024-dimensional representation.

Hyperparameter	**Value**
Projector architecture	Linear (single layer)
Output dimension	1024
Loss function	contrastive
STRUCTURE λ	10.0
STRUCTURE hierarchy levels	1
STRUCTURE warmup steps	500
Training epochs	20
Learning rate	5×10^{-4}
Batch size	512
PCA initialization	true

Supplementary Table 4 Comparative performance benchmarking of the SkinMap ensemble and foundation models. Detailed performance for attribute prediction across internal validation and external hold-out datasets (`DDI` and `PAD-UFES-20`). Metrics include macro-F1 scores for classification tasks and R^2 scores for regression-based attributes.

Model	DDI: disease	DDI: FST	DDI: malignant	PAD: age	PAD: biopsed	PAD: bleed
Random	0.01	0.38	0.56	-0.53	0.49	0.49
Ensemble	**0.09**	**0.76**	**0.60**	**0.40**	**0.88**	**0.75**
MONET	0.05	0.72	0.54	0.37	0.84	0.72
PanDerm	0.04	0.74	0.45	0.39	0.86	**0.75**
ViT-L CLIP 1/3	0.05	0.67	0.55	0.28	0.80	0.66
ViT-L CLIP 1/3	0.05	0.64	0.53	0.30	0.80	0.68
ViT-L CLIP 1/3	0.04	0.68	0.56	0.24	0.82	0.64
ViT-L SigLIP 1/3	0.05	0.66	0.55	0.30	0.78	0.63
ViT-L SigLIP 1/3	0.04	0.59	0.54	0.18	0.74	0.58
ViT-L SigLIP 1/3	0.04	0.57	0.58	0.21	0.77	0.63
ViT-B CLIP	0.05	0.73	0.54	0.39	0.87	0.68
ViT-B SigLIP	0.03	0.61	0.50	0.28	0.80	0.60
ViT-B DINO	0.04	0.65	0.50	0.38	0.83	0.73
ViT-B iBOT	0.05	0.73	0.58	0.39	0.80	0.63
ViT-B MAE	0.01	0.61	0.43	0.19	0.49	0.42

Model	PAD: changed	PAD: diagnostic n.	PAD: diameter	PAD: elevation	PAD: FST	PAD: gender
Random	0.48	0.15	-0.58	0.50	0.20	0.50
Ensemble	**0.55**	**0.67**	0.24	**0.77**	**0.29**	**0.77**
MONET	0.47	0.48	0.19	0.74	0.16	0.69
PanDerm	0.47	0.47	0.23	**0.77**	0.15	0.75
ViT-L CLIP 1/3	0.47	0.43	0.14	0.66	0.17	0.64
ViT-L CLIP 1/3	0.47	0.46	0.15	0.68	0.18	0.66
ViT-L CLIP 1/3	0.47	0.44	0.11	0.66	0.18	0.62
ViT-L SigLIP 1/3	0.47	0.45	0.12	0.69	0.18	0.67
ViT-L SigLIP 1/3	0.47	0.36	0.00	0.61	0.23	0.54
ViT-L SigLIP 1/3	0.47	0.41	0.04	0.66	0.20	0.58
ViT-B CLIP	0.47	0.59	0.20	0.75	0.16	**0.77**
ViT-B SigLIP	0.47	0.44	0.10	0.65	0.18	0.69
ViT-B DINO	0.47	0.55	**0.28**	0.74	0.16	0.71
ViT-B iBOT	0.47	0.45	0.25	0.71	0.17	0.71
ViT-B MAE	0.47	0.20	0.05	0.38	0.15	0.64

Supplementary Table 5 Comparative performance benchmarking of the SkinMap ensemble and foundation models (continued). Detailed performance for attribute prediction across internal validation and external hold-out datasets (`DDI` and `PAD-UFES-20`). Metrics include macro-F1 scores for classification tasks and R^2 scores for regression-based attributes.

Model	**PAD: grew**	**PAD: hurt**	**PAD: itch**	**PAD: region**	**SkinMap: age**	**SkinMap: body r.**
Random	0.49	0.50	0.52	0.08	-0.61	0.09
Ensemble	**0.73**	**0.75**	**0.76**	**0.42**	**0.74**	**0.66**
MONET	0.71	0.62	0.74	0.19	0.42	0.50
PanDerm	0.72	0.61	0.75	0.15	0.51	0.49
ViT-L CLIP 1/3	0.64	0.55	0.71	0.16	0.58	0.57
ViT-L CLIP 1/3	0.65	0.62	0.70	0.18	0.52	0.54
ViT-L CLIP 1/3	0.69	0.56	0.70	0.15	0.57	0.56
ViT-L SigLIP 1/3	0.67	0.58	0.70	0.16	0.52	0.53
ViT-L SigLIP 1/3	0.63	0.50	0.65	0.10	0.65	0.59
ViT-L SigLIP 1/3	0.67	0.59	0.67	0.12	0.52	0.54
ViT-B CLIP	0.72	0.60	0.74	0.22	0.59	0.58
ViT-B SigLIP	0.66	0.61	0.68	0.21	0.44	0.42
ViT-B DINO	**0.73**	0.66	0.72	0.40	0.45	0.49
ViT-B iBOT	0.71	0.57	0.71	0.33	0.45	0.49
ViT-B MAE	0.56	0.45	0.43	0.04	0.30	0.21

Model	**SkinMap: FST**	**SkinMap: FST grp.**	**SkinMap: gender**	**SkinMap: origin**
Random	0.17	0.34	0.50	0.06
Ensemble	**0.57**	**0.75**	**0.92**	**0.86**
MONET	0.51	0.72	0.71	0.82
PanDerm	0.50	0.73	0.73	0.68
ViT-L CLIP 1/3	0.49	0.70	0.83	0.79
ViT-L CLIP 1/3	0.50	0.70	0.78	0.81
ViT-L CLIP 1/3	0.50	0.70	0.82	0.80
ViT-L SigLIP 1/3	0.50	0.70	0.78	0.81
ViT-L SigLIP 1/3	0.49	0.67	0.90	0.79
ViT-L SigLIP 1/3	0.49	0.68	0.81	0.79
ViT-B CLIP	0.54	0.73	0.83	0.84
ViT-B SigLIP	0.45	0.65	0.74	0.70
ViT-B DINO	0.51	0.71	0.71	0.76
ViT-B iBOT	0.52	0.72	0.71	0.78
ViT-B MAE	0.34	0.62	0.63	0.38

Supplementary Table 6 Statistical reliability and confidence intervals (CI) for baseline model performance. Confidence intervals were generated using $B = 1,000$ bootstrap samples. Data are presented as point estimate (95% CI lower–upper). Performance is measured in terms of F1 for categorical labels and R^2 for continuous variables.

Model	DDI: disease	DDI: FST	DDI: malignant	PAD: age	PAD: biopsed	PAD: bleed
Ensemble	0.09 (0.06–0.12)	0.76 (0.69–0.83)	0.60 (0.50–0.70)	0.40 (0.28–0.45)	0.88 (0.85–0.91)	0.75 (0.71–0.80)
MONET	0.05 (0.03–0.08)	0.72 (0.65–0.80)	0.54 (0.45–0.63)	0.37 (0.27–0.44)	0.84 (0.80–0.87)	0.72 (0.67–0.77)
PanDerm	0.04 (0.02–0.06)	0.74 (0.67–0.81)	0.45 (0.40–0.52)	0.39 (0.29–0.46)	0.86 (0.83–0.89)	0.75 (0.70–0.80)

Model	PAD: changed	PAD: diagnostic n.	PAD: diameter	PAD: elevation	PAD: FST	PAD: gender
Ensemble	0.55 (0.49–0.63)	0.67 (0.60–0.72)	0.24 (0.08–0.35)	0.77 (0.73–0.80)	0.29 (0.23–0.39)	0.77 (0.72–0.82)
MONET	0.47 (0.46–0.48)	0.48 (0.44–0.51)	0.19 (0.09–0.28)	0.74 (0.70–0.78)	0.16 (0.14–0.21)	0.69 (0.64–0.75)
PanDerm	0.47 (0.46–0.48)	0.47 (0.44–0.51)	0.23 (0.09–0.35)	0.77 (0.73–0.82)	0.15 (0.14–0.19)	0.75 (0.70–0.80)

Model	PAD: grew	PAD: hurt	PAD: itch	PAD: region	SkinMap: age	SkinMap: body r.
Ensemble	0.73 (0.66–0.74)	0.75 (0.70–0.81)	0.76 (0.72–0.78)	0.42 (0.35–0.49)	0.74 (0.74–0.75)	0.66 (0.65–0.67)
MONET	0.71 (0.66–0.76)	0.62 (0.55–0.68)	0.74 (0.70–0.78)	0.19 (0.16–0.22)	0.42 (0.42–0.43)	0.50 (0.49–0.51)
PanDerm	0.72 (0.67–0.77)	0.61 (0.55–0.67)	0.75 (0.71–0.79)	0.15 (0.13–0.17)	0.51 (0.50–0.51)	0.49 (0.47–0.50)

Model	SkinMap: FST	SkinMap: FST grp.	SkinMap: gender	SkinMap: origin
Ensemble	0.57 (0.55–0.58)	0.75 (0.74–0.76)	0.92 (0.92–0.93)	0.86 (0.86–0.86)
MONET	0.51 (0.49–0.52)	0.72 (0.71–0.73)	0.71 (0.70–0.71)	0.82 (0.80–0.84)
PanDerm	0.50 (0.48–0.51)	0.73 (0.72–0.74)	0.73 (0.72–0.73)	0.68 (0.66–0.70)

Supplementary Table 7 Orphan ICD codes and transient acquisition patterns. List of diagnostic categories that exhibit an episodic trajectory, appearing in specific collection windows (*e.g.*, 2016 or 2020) but absent from subsequent acquisition efforts. The collection years column highlights the temporal discontinuity of these samples, indicating a lack of sustained surveillance for pathologies outside the dominant clusters of benign and malignant neoplasms.

ICD code	ICD description	ICD chapter	Samples (n)	Collection years
B09	Unspecified viral infection characterized by skin and mucous membrane lesions	Certain infectious and parasitic diseases (A00–B99)	1308	2020–2020
A64	Unspecified sexually transmitted disease	Certain infectious and parasitic diseases (A00–B99)	436	2020–2020
L23.7	Allergic contact dermatitis due to plants	Diseases of the skin (L00–L99)	325	2020–2020
A38	Scarlet fever	Certain infectious and parasitic diseases (A00–B99)	35	2020–2020
C7B.04	Secondary carcinoid tumors of peritoneum	Neoplasms (C00–D49)	34	2020–2020
B30.3	Acute epidemic hemorrhagic conjunctivitis (enteroviral)	Certain infectious and parasitic diseases (A00–B99)	32	2020–2020
L66.2	Folliculitis decalvans	Diseases of the skin (L00–L99)	15	2020–2020
L60.2	Onychogryphosis	Diseases of the skin (L00–L99)	11	2020–2020
Q89.7	Multiple congenital malformations, not elsewhere classified	Congenital malformations, deformations, chromosomal abnormalities, and genetic disorders (Q00–QA0)	8	2020–2020
L66.0	Pseudopelade	Diseases of the skin (L00–L99)	5	2020–2020
E88.3	Tumor lysis syndrome	Endocrine, nutritional and metabolic diseases (E00–E89)	4	2020–2020
L68.3	Polytrichia	Diseases of the skin (L00–L99)	2	2020–2020
L81.2	Freckles	Diseases of the skin (L00–L99)	1	2020–2020
I78.9	Disease of capillaries, unspecified	Diseases of the circulatory system (I00–I99)	177	2018–2018
L64	Androgenic alopecia	Diseases of the skin (L00–L99)	41	2016–2016
L85.9	Epidermal thickening, unspecified	Diseases of the skin (L00–L99)	40	2016–2016

Appendix C Supplementary Information

C.1 Extended analysis

C.1.1 Sparse metadata against imputed

To validate that our audit of the global landscape reflects the true underlying demographics of the data rather than artifacts of the imputation process, we compared the distributional characteristics of the originally annotated subset with those of the fully imputed archive. Public medical datasets frequently suffer from missing metadata. In our aggregation, a substantial fraction of the 1.1 million images lacked definitive labels for key attributes such as FST, age, and gender.

As illustrated in Supplementary Fig. 2, we observe a strong concordance between the distributions derived from the sparse, ground-truth metadata (left column) and those generated by our imputation ensemble across the entire dataset (center column). For instance, the dominance of FST II–III is consistent between the original subset and the full imputed atlas, as is the marginalization of FST V–VI. Similarly, the pediatric population (ages 0–17) remains low in both the partial (2.8%) and full (3.0%) views, confirming that the unlabelled majority of the data does not harbor a hidden reservoir of diversity.

This structural consistency implies that the "missingness" of metadata in public archives is not driving the observed disparities. Rather, the bias is intrinsic to the image acquisition process itself. The unlabelled data essentially reflect the same exclusionary patterns as the labelled data: A heavy skew toward Western, middle-aged, and lightly pigmented populations. Consequently, the stark divergence from global population statistics (right column) reflects a genuine, systemic failure of the current dermatological data landscape to capture the diversity of the human population.

C.1.2 Missing data from a metadata perspective

While the univariate analysis presented in the main text exposes severe imbalances, such as the restriction of FST V–VI to only 5.8% of the archive and pediatric patients to 3.0%, these aggregate statistics potentially mask even deeper intersectional voids. To investigate the availability of data for specific patient subgroups, we analyzed the joint distributions of imputed metadata attributes (Supplementary Fig. 3). This intersectional audit reveals that the global archive is not merely skewed, but highly concentrated around a specific "standard" patient profile. As illustrated in the FST vs. age and gender vs. age heatmaps, the vast majority of the 1.1 million samples are clustered in a narrow demographic window: Patients aged 30–70 with FST II–III. This region of hypersaturation suggests that model performance reported in current literature is heavily weighted toward this specific demographic intersection. Conversely, the analysis highlights "dark corners" of the latent space where data support collapses entirely. Most critically, the intersection of age and skin tone reveals a compounded exclusion of pediatric populations with darker skin tones. While pediatric data is scarce overall, the bin representing pediatric patients (0–17 years) with FST V–VI is effectively empty. This implies that current foundation models possess virtually no ground-truth visual signals for common childhood dermatoses on darker skin, rendering them potentially unsafe for this vulnerable subgroup. Furthermore, the origin (continent) vs. age analysis underscores the geographical homogeneity of the data. Even when diversity in age or body region is present, it is almost exclusively derived from datasets originating from the Global North (specifically Europe and North America). The distinct lack of samples from the Global South (Africa and South America) across all body regions suggests that environmental and genetic variations specific to these populations are absent from the training dataset. Ultimately, these intersectional voids define the strict operational limits of the current Digital Atlas. Outside the saturated center of Western, middle-aged, fair-skinned patients, the reliability of diagnostic

AI remains unsupported by the underlying data. The technical modality of data acquisition further compounds these structural exclusions (Supplementary Fig. 4). While the recent expansion of the atlas has been driven by the introduction of TBP, the intersectional analysis reveals that this modality is almost exclusively restricted to adult, fair-skinned populations (FST II–III) within Europe and North America. Specifically, TBP shows negligible coverage of pediatric patients and individuals with darker skin tones (FST V–VI), whereas it is heavily concentrated on the trunk and lower limbs.

C.1.3 ICD composition over dataset

To investigate the drivers behind the diminishing novelty observed in recent years, we analyzed the temporal evolution of disease categories and their specific data sources. Supplementary Fig. 5a–c illustrates a heavy and persistent skew toward *neoplasms* (C00–D49), particularly in the most recent acquisition phase (2022–2025), where this category dominates the vast majority of yearly samples. While earlier years display transient spikes in diversity, likely driven by specific releases containing *symptoms/signs* (R00–R69) or *skin diseases* (L00–L99), the overall trend demonstrates a consolidation around well-represented pathologies. Supplementary Fig. 5d further granularizes this by mapping the provenance of specific diagnostic blocks, revealing that the availability of these categories is often contingent on isolated dataset release events rather than continuous collection. We observe that high-volume categories such as *benign neoplasms* (D10–D36) and *melanoma* (C43–C44) are heavily dictated by the sequential release of `ISIC` challenges and the `Derm1M` archive. This dependency highlights a fragility in the global data landscape, where the representation of entire disease categories rests on single sources rather than a robust, diverse network of data collectors.

C.1.4 Orphan ICD codes

Beyond the broad saturation of common neoplasms, our analysis uncovers a phenomenon of "orphan" diagnostic categories, *i.e.*, pathologies that exhibit a transient presence in the global archive before vanishing from subsequent data collection efforts. As detailed in Supplementary Table 7, this pattern effectively represents lost knowledge, where specific diseases appear during a single release event (often linked to a specific dataset such as `PAD-UFES-20` or disparate `ISIC challenges`) but are not sustained in longitudinal acquisition. This discontinuity is most visible in infectious and inflammatory categories, which stand in stark contrast to the persistent accumulation of melanoma and nevi. For instance, *unspecified viral infections* (B09) and *sexually transmitted diseases* (A64) appeared as cohorts in 2020 ($n = 1,308$ and $n = 436$, respectively) but have seen no replenishment in the years since. Similarly, niche inflammatory conditions such as *allergic contact dermatitis due to plants* (L23.7) were captured in a single 2020 interval. The existence of these orphans underscores the structural fragility of the current data landscape: In the absence of strategic, sustained surveillance pipelines, the field remains dependent on episodic, uncoordinated releases that fail to provide the continuous representation necessary for robust model generalization.

C.2 Detailed results

C.2.1 Detailed results on Fitzpatrick skin type prediction

To ensure that the analysis based on skin tones is not confounded by overestimation of the types, we provide confusion matrices in Supplementary Fig. 6 for the FST prediction on hold-out datasets, *i.e.*, `DDI`, `PAD-UFES-20`, and the hold-out test set for the full collection. Results show that, across datasets, mispredictions mostly occur when confusing boundary types, *i.e.*, types that are similar to each other, such as FST V–VI, and less often when confusing higher types for lower ones. This provides evidence that, although there are mispredictions, they often occur near the ground truth, indicating that the estimates reflect the underlying distributions.

C.2.2 Dataset distances

To provide a comprehensive map of the relationships among the 29 aggregated datasets, we computed the pairwise FD between all collection pairs and visualized the results as a heatmap in Supplementary Fig. 7. This quantitative analysis reveals significant distributional homologies that challenge the assumption of dataset independence. We observe a distinct `ISIC` cluster in which annual challenges (2016–2020) and institutional archives such as `MSKCC` exhibit minimal pairwise distances, reflecting their shared lineage and patient populations. Furthermore, the analysis statistically confirms suspected subset relationships. The vanishingly small FD between `Fitzpatrick17k` and `SkinCAP` confirms that the latter functions as a subset of the former, while similar homologies link `CAN2000` and `CAN5600`. Conversely, datasets such as `PH2` and `MED-NODE` emerge as distributional outliers, with high average distances to the rest of the field, thereby identifying them as unique sources of variance and critical benchmarks for evaluating model robustness to genuine domain shifts.

C.3 Additional information

C.3.1 Template description and LLM prompt

To generate concise captions suitable for multi-modal model training (maximum 77 tokens), we employed an open-source LLM to compress the template image descriptions while maintaining medical accuracy. We used the Qwen2.5-7B-Instruct model[37] with the following configuration: Temperature = 0.3, maximum new tokens = 100, top-p sampling = 0.9, running inference in float16 precision on CUDA-enabled hardware. The compression task was structured using a two-component prompt template. The system prompt established the task requirements and quality criteria:

> You are a medical caption compression expert. Your task is to rewrite long medical image descriptions into concise captions suitable for multi-modal model training.
>
> Requirements:
>
> 1. Keep all medically relevant information (condition, body location, demographics)
> 2. Remove verbose phrasing and unnecessary words
> 3. Maintain medical accuracy and terminology
> 4. Target maximum length: 77 tokens (typically 60-70 words)
> 5. Use clear, direct language
>
> Example:
>
> Input: "This total body photograph tile shows a close-up benign skin lesion on the right leg (lower extremity) for a male patient in his 60s."
>
> Output: "Benign skin lesion on right leg, male, aged 60-69."

For each caption, the user prompt template provided the specific compression instruction:

> Compress this medical image caption while maintaining all medically relevant information: {original_caption}
>
> Provide only the compressed caption, no explanation.

Generated responses were post-processed to extract only the first line of output, removing any explanatory text. In case of generation failures, the system preserved the original description as a fallback.

C.3.2 Hyperparameters and configurations

We detail all model architectures, training configurations, and fusion parameters used to construct the SkinMap embedding space. The final ensemble comprises eight multi-modal models trained on different data partitions and three self-supervised learning models pretrained on the imaging data alone.

Multi-modal backbone training. We trained multiple CLIP[38] and SigLIP[39] models from scratch on the curated image-text pairs. All models used a fixed input resolution of 224×224 pixels with standard ImageNet normalization[52]. Supplementary Table 1 summarizes the core training hyperparameters shared across all multi-modal models.

Model ensemble composition. The ensemble includes models with different architectures (ViT-Large[51] and ViT-Base[51]), loss functions (CLIP[38] vs. SigLIP[39]), and training datasets. This diversity ensures robustness to model-specific biases and captures complementary aspects of the visual and semantic space. Supplementary Table 2 details the specific configuration of each model in the ensemble.

Embedding fusion and projection. Individual embeddings from all 11 encoders were concatenated and projected to a unified 1024-dimensional space using a learnable linear projector. The projector was trained with a combination of CLIP-style contrastive loss and STRUCTURE regularization[43] to preserve the semantic organization of the pretrained teacher spaces while enabling efficient fusion. Supplementary Table 3 summarizes the projection training configuration.

C.4 List of datasets

Our analysis integrates 29 publicly available dermatology image datasets, comprising 1,135,082 unique images after deduplication. Below, we provide a detailed description of each dataset, including imaging modality, geographic origin, and sample size. Dataset counts represent the final deduplicated samples retained in our corpus.

C.4.1 ISIC archive

The International Skin Imaging Collaboration (ISIC) archive[11] constitutes the largest single data source, aggregated across multiple challenge years (2016–2024) and contributing institutions. We treat each challenge year as a separate cohort to preserve temporal metadata.

- **ISIC-2024**[53] (401,059 images): Total body photography dataset from the 2024 challenge, primarily sourced from US institutions. Images employ TBP modality for longitudinal lesion monitoring.
- **ISIC-2020**[54] (44,108 images): Dermoscopic images from the 2020 challenge, predominantly contributed by institutions in Australia.
- **ISIC-2019**[11] (18,946 images): Dermoscopic lesion dataset from the 2019 challenge, sourced primarily from Hospital Clínic de Barcelona, Spain.
- **ISIC-2018**[55] (2,805 images): Dermoscopic images from the 2018 challenge, representing multi-institutional contributions.
- **ISIC-2017**[56] (1,696 images): Dermoscopic dataset from the 2017 challenge focusing on melanoma detection.
- **ISIC-2016**[57] (1,279 images): Inaugural ISIC challenge dataset with dermoscopic images.
- **ISIC-others**[11] (1,898 images): Aggregated samples from smaller ISIC collections not associated with annual challenges, predominantly from US institutions.

C.4.2 Clinical photography datasets

- **Derm1M**[58] (412,038 images): Large-scale vision-language dataset aligned with clinical ontologies, aggregated from diverse web-sourced educational materials.
- **Fitzpatrick17k**[9] (16,574 images): Benchmark dataset for evaluating algorithmic performance across skin tones, compiled from online dermatology atlases with expert-annotated Fitzpatrick skin type labels.
- **SCIN**[12] (5,033 images): Skin Condition Image Network dataset, crowdsourced from US internet users via Google Search advertisements. Includes self-reported demographics and dermatologist labels across diverse skin tones.
- **PASSION**[13] (4,145 images): Pediatric dermatology dataset from Sub-Saharan Africa, collected 2020–2023 across Madagascar (2,185 images), Guinea (480 images), Malawi (2,150 images), and Tanzania (86 images). Focus on common conditions in pigmented skin: Eczema, fungal infections, scabies, and impetigo.
- **DermaCon-IN**[14] (5,450 images): Clinical dermatology dataset from India, featuring diverse skin conditions with expert annotations across Fitzpatrick and Monk skin tone scales. Predominantly comprises infectious disorders (2,227 images), inflammatory disorders (1,592 images), and pigmentary disorders (831 images), representing common dermatological presentations in South Asian populations.
- **SD-128**[15] (5,619 images): Clinical skin disease classification benchmark derived from the DermQuest online atlas, comprising 128 disease categories.
- **SkinCAP**[16] (4,000 images): Mixed-modality dataset combining clinical photography and close-range imaging.
- **PAD-UFES-20**[17] (2,298 images): Brazilian smartphone-captured lesion dataset from Federal University of Espírito Santo, representing real-world telemedicine imaging conditions.

- **DDI**[6] (656 images): Diverse Dermatology Images dataset from Stanford, retrospectively curated 2010–2020 with pathologically confirmed diagnoses stratified across Fitzpatrick skin types I–VI.
- **MED-NODE**[18] (170 images): Computer-assisted melanoma diagnosis dataset from the Netherlands, employing non-dermoscopic clinical photography.

C.4.3 Dermoscopic datasets

- **HAM10000**[19] (11,526 images): Multi-source dermoscopic image collection of common pigmented lesions from Medical University of Vienna, Austria, and University of Queensland, Australia.
- **Derm7pt**[20] (2,022 images): Seven-point melanoma checklist dataset from Italy, featuring both clinical and dermoscopic modalities with structured diagnostic criteria annotations.
- **PH2**[21] (200 images): Dermoscopic benchmark from Portugal with expert-segmented melanocytic lesions.
- **Daffodil**[22] (9,548 images): Dermoscopic dataset aggregated from web-sourced educational materials.
- **MSKCC**[23] (8,984 images): Dermoscopic images from Memorial Sloan Kettering Cancer Center, New York.
- **HIBA**[24] (1,635 images): Dermoscopic dataset from Hospital Italiano de Buenos Aires, Argentina.

C.4.4 Educational datasets

- **PubMed** (126,284 images): Dermatology images extracted from scientific literature in the PubMed Central open-access repository, representing global clinical research.
- **DermNet** (30,324 images): Clinical educational resource from DermNet NZ, New Zealand.

- **DermaCompass** (2,450 images): Clinical educational resource from DermaCompass from Switzerland.
- **Altmeyers** (7,673 images): Clinical dermatology encyclopedia images from Germany.
- **CAN5600** (5,619 images): Web-sourced clinical images from approximately 80 countries.
- **CAN2000** (2,006 images): Subset of web-sourced clinical images from diverse international sources.

References

[1] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 2017.

[2] Tatsunori Hashimoto. Model Performance Scaling with Multiple Data Sources. In *International Conference on Machine Learning*, 2021.

[3] Sachin Goyal, Pratyush Maini, Zachary C Lipton, Aditi Raghunathan, and J Zico Kolter. Scaling Laws for Data Filtering–Data Curation cannot be Compute Agnostic. In *Conference on Computer Vision and Pattern Recognition*, 2024.

[4] Anirudh Subramanyam, Yuxin Chen, and Robert L Grossman. Scaling Laws Revisited: Modeling the Role of Data Quality in Language Model Pretraining. *arXiv preprint*, 2025.

[5] Roxana Daneshjou, Mary P Smith, Mary D Sun, Veronica Rotemberg, and James Zou. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. *JAMA Dermatology*, 2021.

[6] Roxana Daneshjou, Kailas Vodrahalli, Roberto A. Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M. Swetter, Elizabeth E. Bailey, Olivier Gevaert, Pritam Mukherjee, Michelle Phung, Kiana Yekrang, Bradley

Fong, Rachna Sahasrabudhe, Johan A. C. Allerup, Utako Okata-Karigane, James Zou, and Albert S. Chiou. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*, 2022.

[7] Maria Paz Salinas, Javiera Sepúlveda, Leonel Hidalgo, Dominga Peirano, Macarena Morel, Pablo Uribe, Veronica Rotemberg, Juan Briones, Domingo Mery, and Cristian Navarrete-Dechent. A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis. *NPJ Digital Medicine*, 2024.

[8] Fabian Gröger, Ludovic Amruthalingam, Simone Lionetti, Alexander A Navarini, Fabian Ille, and Marc Pouly. A review and systematic guide to counteracting medical data scarcity for AI applications. *Computer Methods and Programs in Biomedicine Update*, 2025.

[9] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Conference on Computer Vision and Pattern Recognition*, 2021.

[10] Roxana Daneshjou, Mert Yuksekgonul, Zhuo Ran Cai, Roberto Novoa, and James Y Zou. Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. *Advances in Neural Information Processing Systems*, 2022.

[11] ISIC Society. ISIC Archive, 2022.

[12] Abbi Ward, Jimmy Li, Julie Wang, Sriram Lakshminarasimhan, Ashley Carrick, Bilson Campana, Jay Hartford, Pradeep K. Sreenivasaiah, Tiya Tiyasirisokchai, Sunny Virmani, Renee Wong, Yossi Matias, Greg S. Corrado, Dale R. Webster, Margaret Ann Smith, Dawn Siegel, Steven Lin, Justin Ko, Alan Karthikesalingam, Christopher Semturs, and Pooja Rao. Creating an Empirical Dermatology Dataset Through Crowdsourcing With Web Search Advertisements. *JAMA*

Network Open, 2024.

[13] Philippe Gottfrois, Fabian Gröger, Faly Herizo Andriambololoniaina, Ludovic Amruthalingam, Alvaro Gonzalez-Jimenez, Christophe Hsu, Agnes Kessy, Simone Lionetti, Daudi Mavura, Wingston Ng'ambi, et al. Passion for dermatology: Bridging the diversity gap with pigmented skin images from sub-saharan africa. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024.

[14] Shanawaj S Madarkar, Mahajabeen Madarkar, Teli Prakash, Konda Reddy Mopuri, Vinaykumar MV, KVL Sathwika, Adarsh Kasturi, Gandla Dilip Raj, PVN Supranitha, Harsh Udai, et al. DermaCon-IN: A Multi-concept Annotated Dermatological Image Dataset of Indian Skin Disorders for Clinical AI Research. *Advances in Neural Information Processing Systems*, 12 2025.

[15] Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. A benchmark for automatic visual classification of clinical skin disease images. In *European Conference on Computer Vision*, 2016.

[16] Juexiao Zhou, Liyuan Sun, Yan Xu, Wenbin Liu, Shawn Afvari, Zhongyi Han, Jiaoyan Song, Yongzhi Ji, Xiaonan He, and Xin Gao. Skincap: A multi-modal dermatology dataset annotated with rich medical captions. *arXiv preprint*, 2024.

[17] Andre G. C. Pacheco, Gustavo R. Lima, Amanda S. Salomão, Breno Krohling, Igor P. Biral, Gabriel G. de Angelo, Fábio C. R. Alves Jr, José G. M. Esgario, Alana C. Simora, Pedro B. C. Castro, Felipe B. Rodrigues, Patricia H. L. Frasson, Renato A. Krohling, Helder Knidel, Maria C. S. Santos, Rachel B. do Espírito Santo, Telma L. S. G. Macedo, Tania R. P. Canuto, and Luíz F. S. de Barros. PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, 2020.

[18] I. Giotis, N. Molders, S. Land, M. Biehl, M.F. Jonkman, and N. Petkov. MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic

images". *Expert Systems with Applications*, 2015.

[19] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 2018.

[20] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 2019.

[21] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. PH 2-A dermoscopic image database for research and benchmarking. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, 2013.

[22] Khushbu and Sharun Akter. Skin Disease Classification Dataset, 2024.

[23] Memorial Sloan Kettering Cancer Center. MSKCC Skin Tone Labeling Dataset, 2025.

[24] María Agustina Ricci Lara, María Victoria Rodríguez Kowalczuk, Maite Lisa Eliceche, María Guillermina Ferraresso, Daniel Roberto Luna, Sonia Elizabeth Benitez, and Luis Daniel Mazzuoccolo. A dataset of skin lesion images collected in Argentina for the evaluation of AI tools in this population. *Scientific Data*, 2023.

[25] Chanwoo Kim, Soham U Gadgil, Alex J DeGrave, Jesutofunmi A Omiye, Zhuo Ran Cai, Roxana Daneshjou, and Su-In Lee. Transparent medical image AI via an image–text foundation model grounded in medical literature. *Nature Medicine*, 2024.

[26] Siyuan Yan, Zhen Yu, Clare Primiero, Cristina Vico-Alonso, Zhonghua Wang, Litao Yang, Philipp Tschandl, Ming Hu, Lie Ju, Gin Tan, et al. A multimodal vision foundation model for clinical dermatology. *Nature Medicine*, 2025.

[27] Lauren AV Orenstein, John S Barbieri, Meron Siira, Ethan Borre, Krittin J Supapannachart, Eric Viera, Courtney Ann Prestwood, Robert Swerlick, Rachel E Patzer, and Suephy C Chen. Outpatient Dermatology Productivity Measures by Patient Race, Sex, and Age. *JAMA Dermatology*, 2025.

[28] Steven Y He, Charles E McCulloch, W John Boscardin, Mary-Margaret Chren, Eleni Linos, and Sarah T Arron. Self-reported pigmentary phenotypes and race are significant but incomplete predictors of Fitzpatrick skin phototype in an ethnically diverse population. *Journal of the American Academy of Dermatology*, 2014.

[29] Department of Economic. *World Population Prospects 2024: Summary of Results*. Stylus Publishing, LLC, 2024.

[30] Omar B Ahmad, Cynthia Boschi-Pinto, Alan D Lopez, Christopher JL Murray, Rafael Lozano, Mie Inoue, et al. Age standardization of rates: a new WHO standard. *Geneva: World Health Organization*, 2001.

[31] Nick Couldry and Ulises A Mejias. The costs of connection: How data is colonizing human life and appropriating it for capitalism. In *The costs of connection*. Stanford University Press, 2019.

[32] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *arXiv preprint*, 2020.

[33] Sebastian Damrich, Philipp Berens, and Dmitry Kobak. Persistent homology for high-dimensional data based on spectral methods. *Advances in Neural Information Processing Systems*, 2024.

[34] Russell J Funk and Xiangting Wu. When disruption endures: Science of science. *Nature Computational Science*, 2025.

[35] Michael Park, Erin Leahey, and Russell J Funk. Papers and patents are becoming less disruptive over time. *Nature*, 2023.

[36] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 Technical Report. *arXiv preprint*, 2024.

[37] Qwen Team. Qwen2.5: A Party of Foundation Models, September 2024.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[39] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *International Conference on Computer Vision*, 2023.

[40] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision*, 2021.

[41] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT Pre-training with Online Tokenizer. In *International Conference on Learning Representations*, 2022.

[42] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Conference on*

Computer Vision and Pattern Recognition, 2022.

[43] Fabian Gröger, Shuo Wen, Huyen Le, and Maria Brbić. With Limited Data for Multimodal Alignment, Let the STRUCTURE Guide You. *Advances in Neural Information Processing Systems*, 12 2025.

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2011.

[45] DC Dowson and BV666017 Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 1982.

[46] Arpita Ghosh, Stephen Boyd, and Amin Saberi. Minimizing effective resistance of a graph. *SIAM Review*, 2008.

[47] Matthias Hein, Jean-Yves Audibert, and Ulrike Von Luxburg. From graphs to manifolds–weak and strong pointwise consistency of graph Laplacians. In *International Conference on Computational Learning Theory*, 2005.

[48] Herbert Edelsbrunner and John Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010.

[49] Ulrich Bauer. Ripser: efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology*, 2021.

[50] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2017.

[51] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

[52] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, 2009.

[53] Nicholas R Kurtansky, Brian M D'Alessandro, Maura C Gillis, Brigid Betz-Stablein, Sara E Cerminara, Rafael Garcia, Marcela Alves Girundi, Elisabeth Victoria Goessinger, Philippe Gottfrois, Pascale Guitera, et al. The SLICE-3D dataset: 400,000 skin lesion image crops extracted from 3D TBP for skin cancer detection. *Scientific Data*, 2024.

[54] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*, 2021.

[55] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint*, 2019.

[56] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *International Symposium on Biomedical Imaging*, 2018.

[57] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). *arXiv*

preprint, 2016.

[58] Siyuan Yan, Ming Hu, Yiwen Jiang, Xieji Li, Hao Fei, Philipp Tschandl, Harald Kittler, and Zongyuan Ge. Derm1m: A million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology. In *International Conference on Computer Vision*, 2025.