nature communications

Article | Article in Press

https://doi.org/10.1038/s41467-026-78602-w

A global atlas of digital dermatology to map innovation and disparities

Received: 9 January 2026

Accepted: 30 September 2026

Published online: 21 November 2026

Check for updates

Fabian Gröger[1,2] ✉, Simone Lionetti[2], Philippe Gottfrois[1,3], Alvaro Gonzalez-Jimenez[2,3], Lea Habermacher[3], Ludovic Amruthalingam[2], Matthew Groh[4], Marc Pouly[2,5] ✉ & Alexander A. Navarini[2,5] ✉

Artificial intelligence promises democratized access to dermatological care, but model reliability depends on the quality and comprehensiveness of the underlying data. Despite rapid growth in publicly available dermatology images, the field lacks quantitative indicators to measure whether new datasets expand clinical coverage or merely replicate what is already known. Existing collections are widely assumed to be biased, yet the scale and topography of these gaps have remained unquantified, leaving the community without a road map for strategic data acquisition. Here, we show, by unifying 1.1 million images from 29 public datasets into a queryable semantic atlas, that the global archive is markedly skewed and increasingly redundant: darker skin tones (Fitzpatrick V–VI) constitute only 11.0% of images, pediatric patients only 2.3%, and informational novelty has plateaued despite exponential growth, revealing where targeted data acquisition is most needed.

The integration of artificial intelligence (AI) into medicine holds the promise to transform 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 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 1,135,080 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

[1]Department of Biomedical Engineering, University of Basel, Hegenheimermattweg 167b, Allschwil, Basel, Switzerland. [2]Department of Computer Science, Lucerne University of Applied Sciences and Arts, Suurstoffi 1, Rotkreuz, Zug, Switzerland. [3]Department of Dermatology, University Hospital of Basel, Burgfelderstrasse 101, Basel, Basel, Switzerland. [4]Kellogg School of Management, Northwestern University, 2211 Campus Drive, Evanston, IL, USA. [5]These authors contributed equally: Marc Pouly, Alexander A. Navarini. ✉e-mail: fabian.groeger@unibas.ch; marc.pouly@hslu.ch; alexander.navarini@usb.ch

metadata. Using the latent representations of annotated samples, we infer missing clinical attributes, such as Fitzpatrick skin type (FST), age, sex, 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 11.0% 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 external validations in the literature may not capture true out-of-distribution performance. This distributional overlap risks inflating reported performance metrics, as models validated on nominally independent but representationally similar datasets may appear to generalize while merely performing well on familiar subpopulations rather than the broader clinical distribution.

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 the digital atlas and the model ensemble, under an open-source license (Fig. 1).

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 sex. To address this, we utilized the SKINMAP embedding space to impute missing attributes, expanding the metadata coverage of the dataset by an average of 60.8 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

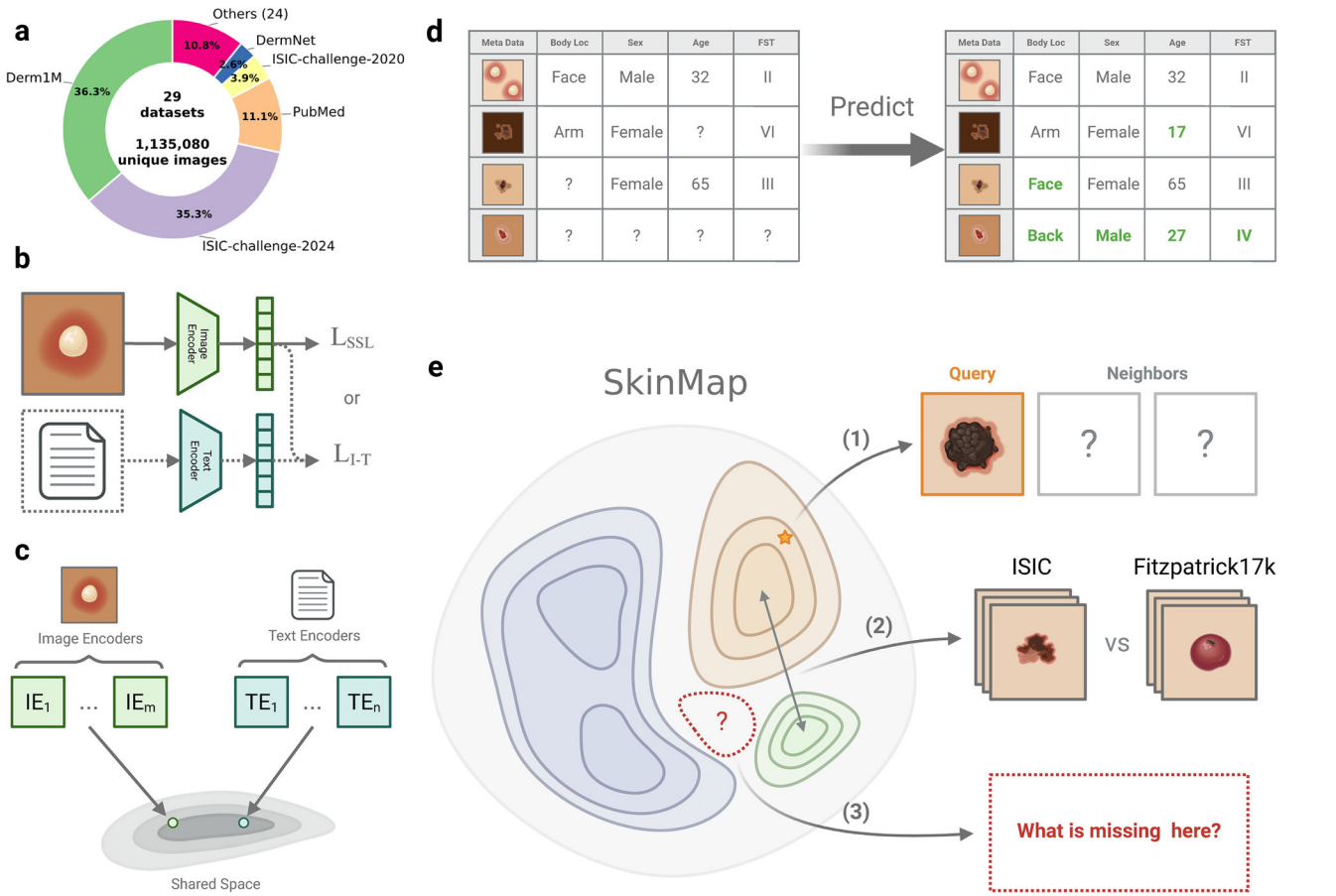

Fig. 1 | The SKINMAP multi-modal framework for auditing the global dermatology landscape. a Aggregation of 29 public datasets yielding 1.1 million unique images with sparse original metadata. **b** Multi-modal training pipeline utilizing self-supervised learning (SSL) on imaging data and contrastive 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 under-represented latent regions to direct future data collection. Created in BioRender. Gröger, F. (2026) https://BioRender.com/2udquxh.

https://doi.org/10.1038/s41467-026-78602-w

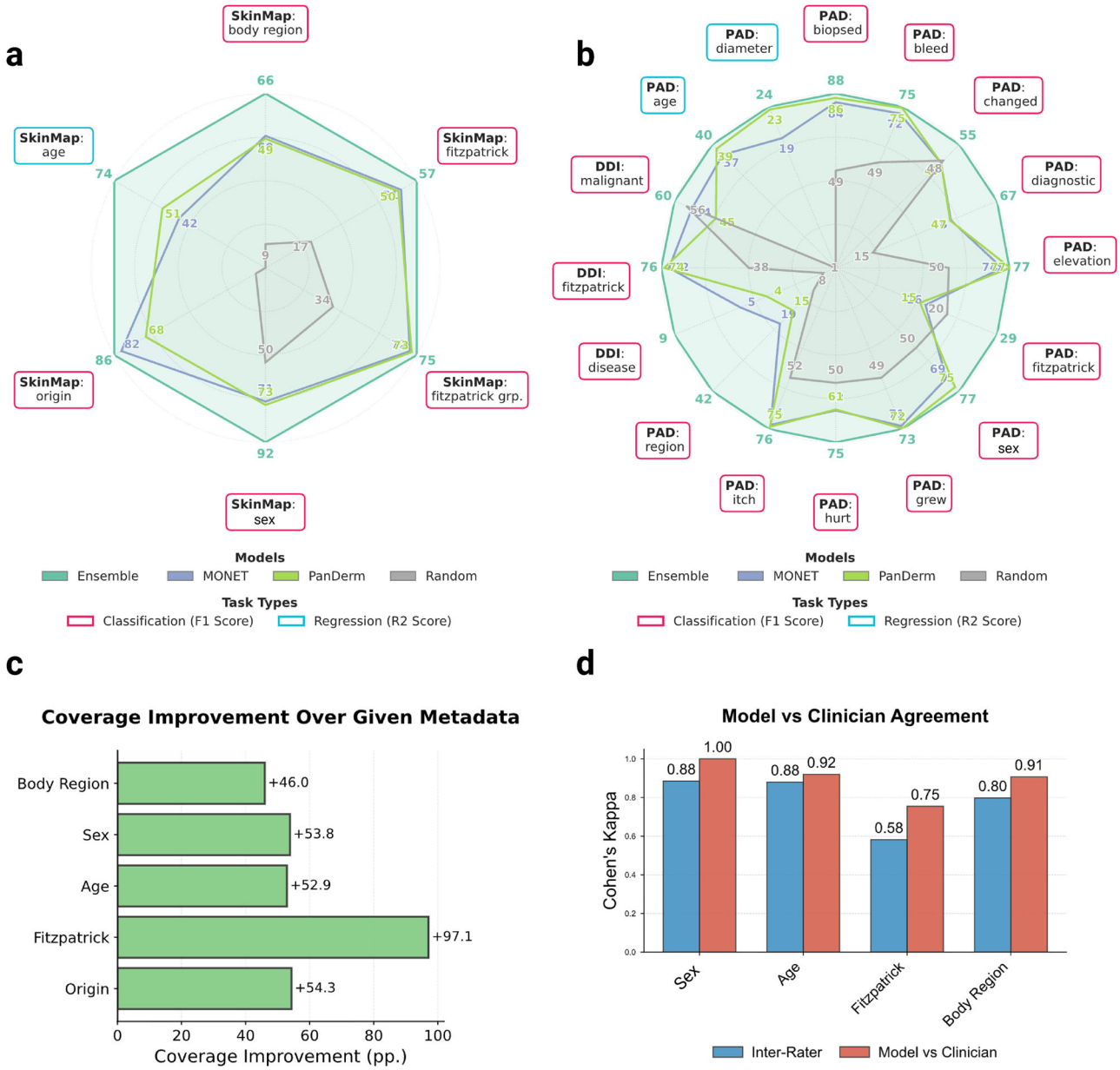

Fig. 2 | Benchmarking imputation precision and metadata expansion. a, **b** Radar plots comparing the predictive performance of the SKINMAP ensemble against state-of-the-art foundation models MONET and PanDerm across multiple attributes on the internal validation set (**a**) and external hold-out datasets (**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. Created in BioRender. Gröger, F. (2026) https://BioRender.com/2udquxh.

outperforming state-of-the-art foundation models, including MONET[25] and PanDerm[26], in attribute prediction accuracy (Fig. 2a, b and Supplementary Table 10 for statistical confidence intervals). The strong concordance between the ground-truth subset and the fully imputed archive confirms 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. 2d). This densely annotated dataset overcomes the limitations of the original, sparse labels to assess the demographic composition of the entire digital archive. To ensure that the demographic analysis is not distorted by low-quality images or near-duplicates, we report all population-level statistics on the quality-filtered subset of the atlas (see Methods, "Quality scoring and near-duplicate detection"; Supplementary Fig. 2 for a comparison of original, unfiltered, and filtered views). These processing steps alter the estimated demographic composition only marginally (the (FST) V–VI share by about one percentage point, the pediatric share negligibly) and do not drive the reported disparities, which are also robust to the choice of quality threshold (Methods, "Quality scoring and near-duplicate detection"). We note that imputation accuracy varies across subgroups (Supplementary Fig. 8): predictions for well-represented populations (e.g., fair-skinned, middle-aged) benefit from dense training signal and yield tighter confidence intervals, whereas estimates for underrepresented subgroups are subject to higher variance.

https://doi.org/10.1038/s41467-026-78602-w

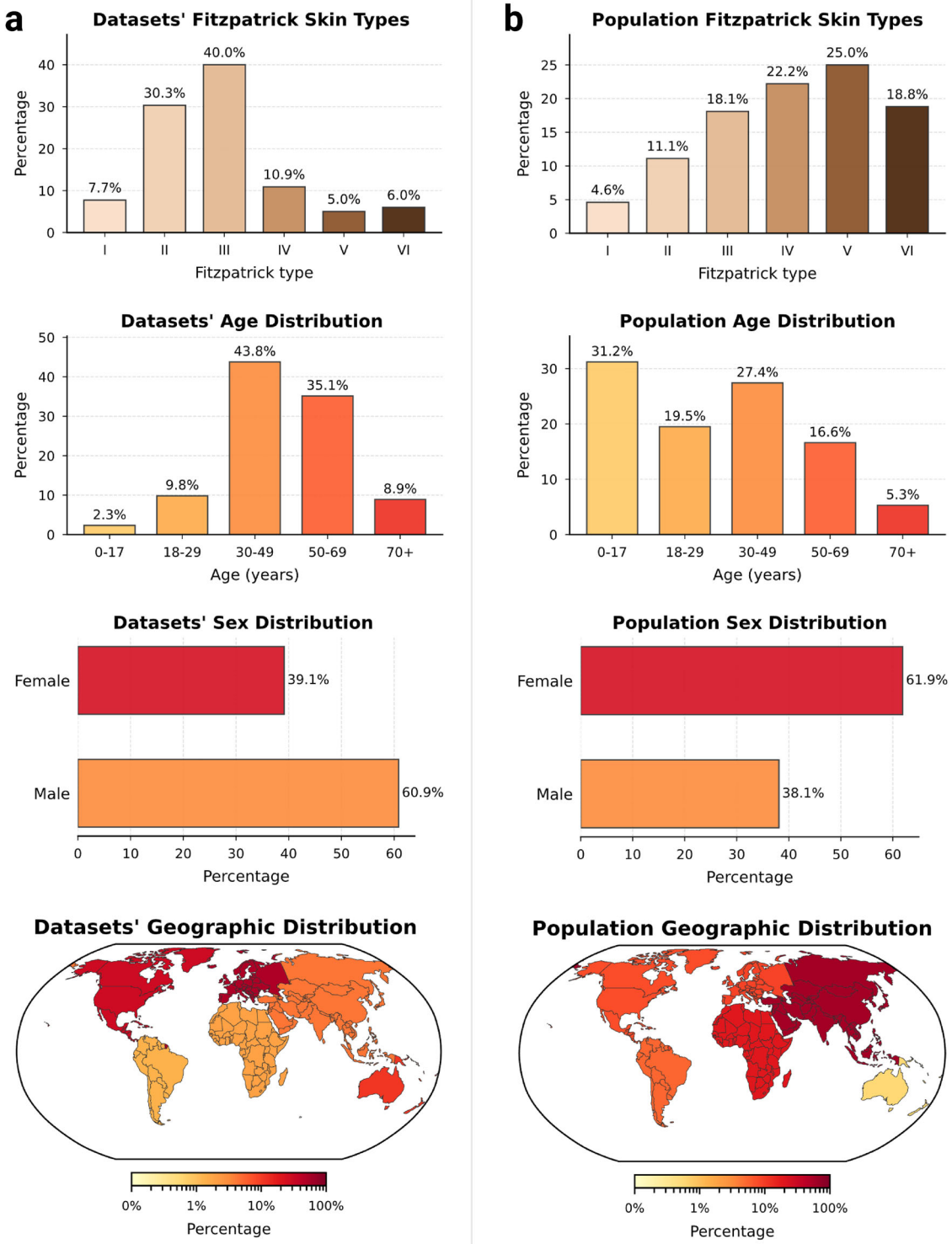

Fig. 3 | Systemic demographic and geographic disparities in the global dermatological datasets. Distributions of demographic attributes across the quality-filtered aggregated digital datasets (**a**) compared with global population statistics for dermatological visits (**b**). The comparison reveals sharp divergences: (FST) V--VI are underrepresented (11.0% in datasets vs. 43.8% globally), as are pediatric populations (2.3 vs. 31.2%), and geographic origins are heavily skewed toward the Global North. Created in BioRender. Gröger, F. (2026) https://BioRender.com/2udquxh.

https://doi.org/10.1038/s41467-026-78602-w

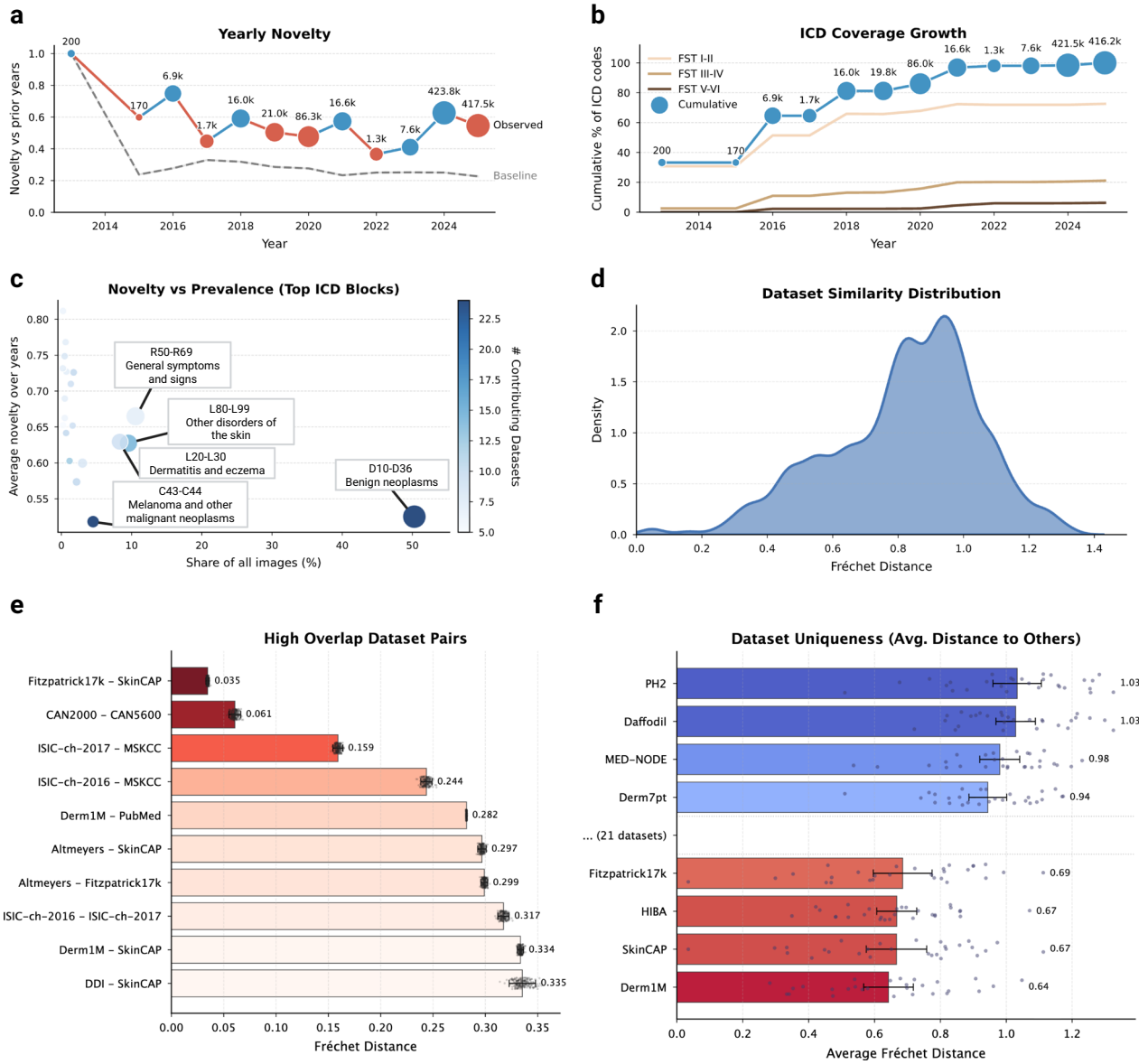

Fig. 4 | Diminishing informational novelty and latent dataset redundancy. **a** Temporal analysis of data acquisition showing the yearly novelty (blue or red depending on whether it increases or decreases relative to the previous year), defined as the average embedding distance of new samples relative 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. Bootstrap 95% confidence intervals on the observed novelty are omitted from the plot as they are narrower than the marker size (relative widths 0.05–3.8%); numeric values are reported in the text. **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 the highest-prevalence (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. Each bar is the pairwise Fréchet distance between two datasets (the measure of center); error bars denote 95% confidence intervals (point estimate ± 1.96 × bootstrap standard deviation; $n = 200$ bootstrap resamples). The intervals are two-sided and descriptive; no null-hypothesis statistical test was performed. The individual bootstrap resamples are shown as overlaid points. **f** Dataset uniqueness scores, highlighting outliers like PH2 as high-value sources of distributional variance. Each point is the mean pairwise Fréchet distance of a dataset to the other $n = 28$ datasets (the measure of center is the mean; the unit of analysis is the dataset, 29 in total); error bars denote the 95% confidence interval of the mean (±1.96 × standard deviation $/\sqrt{28}$). The intervals are two-sided and descriptive; no null-hypothesis statistical test was performed. The individual pairwise distances are shown as overlaid points. Created in BioRender. Gröger, F. (2026) https://BioRender.com/2udquxh.

Importantly, however, wider intervals are confined to small subgroups (e.g., (FST) prediction F1 for patients aged 70+: 0.34, 95% CI: 0.19–0.49, n = 204). Thus, while estimates for underrepresented subgroups should be interpreted with appropriate caution, the aggregate demographic findings are robust.

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 (70.3%), with darker skin tones ((FST) V–VI) constituting only 11.0%. There is an equally strong

geographical imbalance: 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 unrepresented (Fig. 3a, b). 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 sex. The data collection centers on middle age (43.8% aged 30–49), with the pediatric population (0–17 years) underrepresented at 2.3% 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 sex approaches parity (39.1% Female), intersectional imbalances compound in specific subgroups, e.g., Female aged 50+ (Supplementary Note 3). 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[31]. Our temporal audit suggests the field is nearing a regime of diminishing informational returns in which additional samples contribute progressively less distinct 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. Bootstrap 95% confidence intervals on both the observed novelty and the null baseline confirm that these comparisons are statistically robust: for every year, the observed novelty CI is well-separated from the baseline CI (Fig. 4a), with relative CI widths ranging from 0.05% (large-n years) to 3.8% (n = 1259). The main exception to the trend of diminishing returns 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.

To determine if this diminishing novelty 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 plateau 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 Fréchet distances in the SKINMAP embedding space as a proxy for mutual diversity. Bootstrap resampling (200 iterations per pair) confirms that these distances are stable, with a mean coefficient of variation of 0.58% across all dataset pairs. 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, or may still be valuable as existing datasets are enriched with valuable additional information, it has direct implications for evaluation. Indeed, validation on a nominally independent but substantially overlapping dataset does not constitute a genuine held-out evaluation, potentially inflating reported metrics 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`, `Daffodil`, and `MED-NODE` populate the outer regions of the map. These datasets 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 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

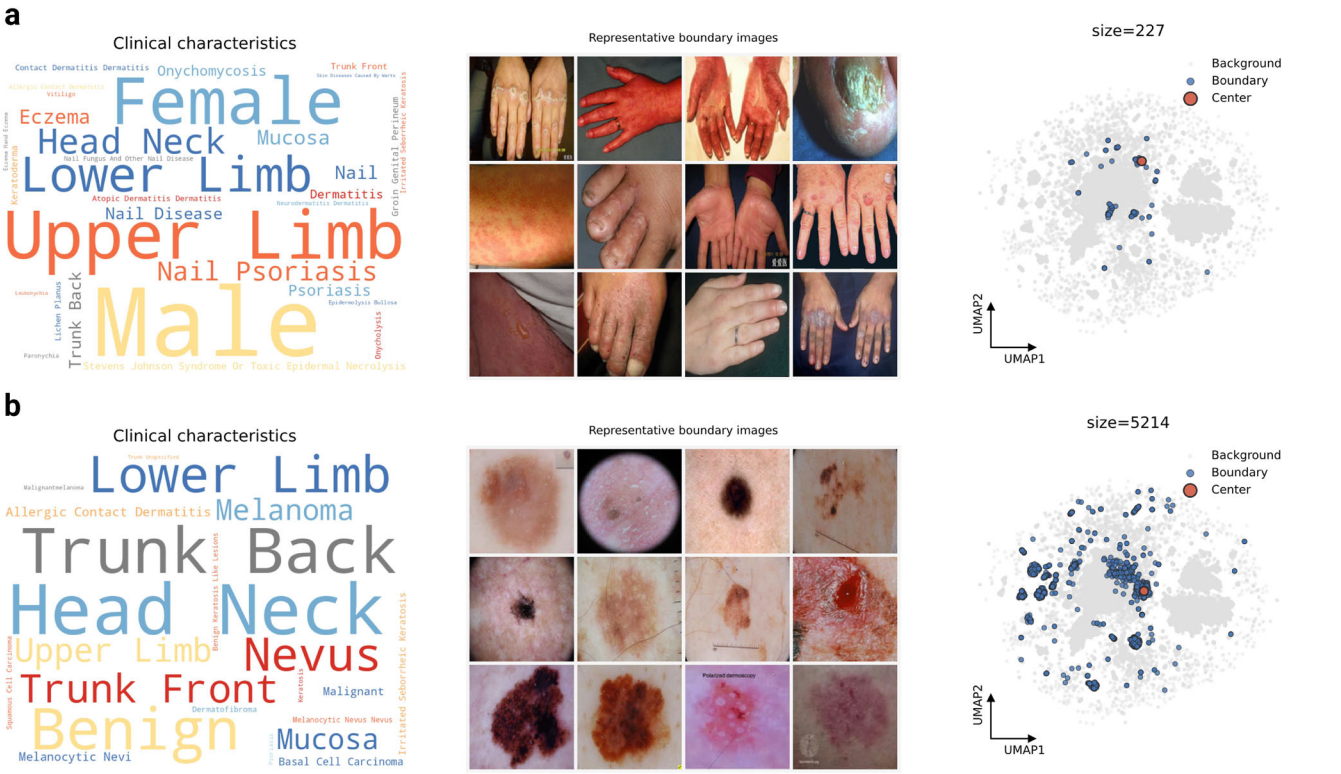

Fig. 5 | Detection of topological holes in the embedding space. a Detailed analysis of the highest-persistence void. The word cloud of metadata from boundary samples is dominated by nail, nail disease, and nail psoriasis, confirming that this topological void represents a structural absence of nail pathologies. **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. Created in BioRender. Gröger, F. (2026) https://BioRender.com/2udquxh.

pruning redundant clusters and filtering outliers. To assess whether these low-quality images and redundant samples affect the conclusions of this study, we conducted a sensitivity analysis comparing all key metrics across four atlas variants: the full atlas (n = 1,135,080), quality-filtered only, deduplicated only, and both filters combined (n = 851,735; 75.0%). The results demonstrate negligible impact across all variants: yearly novelty scores change by less than 0.001, mean Fréchet distances remain stable, and imputation metrics shift by less than 0.4 percentage points across all attributes (Supplementary Table 1). This robustness indicates that while low-quality and redundant images are present in the atlas, they do not distort the neighborhood structure or bias the downstream analyses reported here.

Moreover, we applied spectral persistent homology[32] 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. This topological analysis provides a complementary, label-free perspective that goes beyond what metadata counting alone can reveal. While the intersectional prevalence analysis (Supplementary Note 3) identifies gaps where labeled data were missing, persistent homology detects structural voids directly in the embedding-space geometry, without relying on metadata annotations. This is valuable because it can surface visual phenotypes that are absent from the atlas but not captured by any existing label vocabulary. When both analyses agree, this independent confirmation strengthens the conclusion that the gap is genuine rather than a labeling artifact. 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, i.e., the intersectional prevalence analysis of normalized sample counts across demographic and clinical attributes (Supplementary Fig. 3), is not merely a labeling issue, but a fundamental visual gap in the global atlas. To substantiate this, we compared the nail pathologies present in the atlas against a comprehensive clinical taxonomy of nail disorders. The atlas contains approximately 7800 nail-related images, dominated by unspecified nail disease (L60.9; n = 2776; reflecting imprecise labeling rather than clinical specificity), nail psoriasis (L40.8; n = 2767), and onychomycosis (B35.1; n = 2290). However, many clinically common nail conditions are absent or severely underrepresented (<200 images each), including subungual melanoma (absent), yellow nail syndrome (absent), onychogryphosis, trachyonychia, glomus tumors, koilonychia, habit-tic deformity, myxoid cysts, and Beau's lines. This confirms that the topological void reflects gaps in clinical coverage. 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

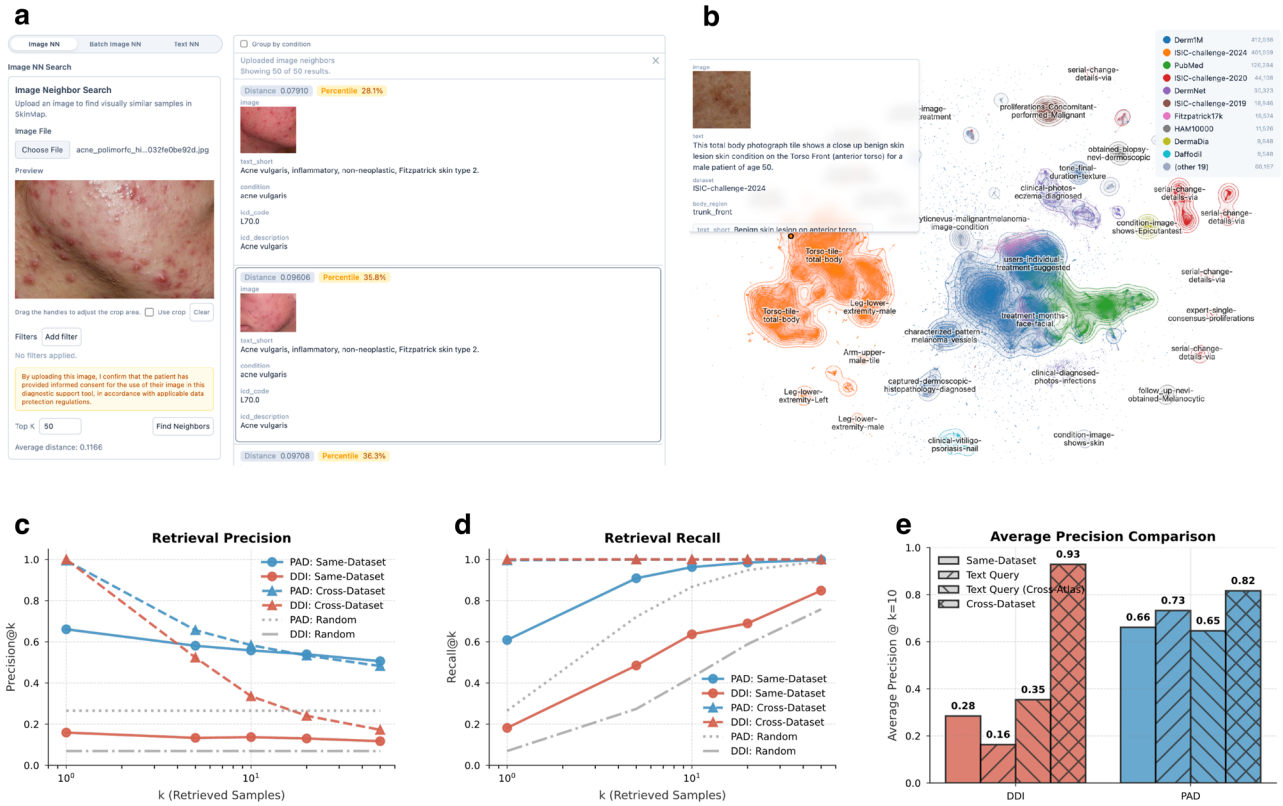

Fig. 6 | The SKINMAP digital atlas for clinical decision support and strategic data steering. a Interactive web-based interface allowing for real-time case uploads and retrieval of semantically similar cases with associated ground-truth diagnoses. **b** Visualization of the global latent space, projecting 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 baseline retrieval within a single dataset provides a starting point for specialized cohorts like `DDI`, cross-dataset retrieval through the exhaustive SKINMAP atlas provides a significant performance boost, increasing `DDI` precision from 0.16 to 0.93. Created in BioRender. Gröger, F. (2026) https://BioRender.com/2udquxh.

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 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 0.16 in the within-dataset baseline (text-based query on `DDI` alone) to 0.93 when retrieving by image similarity across the full atlas. Similarly, `PAD-UFES-20` retrieval precision improves from 0.65 to 0.82. 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. When a query image maps to a low-density or void-adjacent region of the embedding space, the returned neighbors naturally exhibit larger cosine distances, providing an implicit confidence signal to the user. The web interface displays these distances alongside the retrieved cases, enabling practitioners to assess whether the atlas has sufficient coverage for their query. Images landing in well-populated regions yield tightly clustered, low-distance neighbors, whereas queries near structural voids produce distant, heterogeneous matches, signaling that the atlas may lack relevant reference cases and that the retrieval results should be interpreted with caution.

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 11.0% of the global dermatology collection, and pediatric data comprises only 2.3%. 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[33,34], 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.

Standard machine learning techniques for handling class imbalance, such as oversampling, undersampling, or cost-sensitive learning, can improve classification performance, but they cannot generate visual information that was never collected. The structural gaps identified by SKINMAP are fundamentally data-availability problems: images of rare nail conditions on darker skin tones require extrapolation if they do not exist in any public archive. Moreover, the atlas reveals that the imbalance is not merely between classes but between intersections of demographic and clinical attributes (e.g., pediatric patients with (FST) V–VI), compounding the problem beyond what single-axis rebalancing can address.

Beyond characterizing structural gaps, the embedding-space density can serve as a proxy for retrieval reliability. We analyzed diagnostic retrieval accuracy as a function of local density (mean cosine similarity to the 50 nearest neighbors) and measured k-nearest neighbor diagnostic precision across density deciles. Overall, diagnostic precision increases monotonically from 0.34 in the sparsest decile to 0.89 in the densest (Spearman $\rho = 0.54$, $p < 10^{-10}$; Supplementary Tables 3 and 4). The effect is modality-dependent: clinical images show a wide range (precision 0.37–0.64, $\rho = 0.23$), dermoscopy exhibits a narrower range (0.71–0.98, $\rho = 0.42$), whereas (TBP) have near-perfect precision across all density levels (≥0.99), reflecting the homogeneity of the dominant (TBP) collection. The sparsest regions are dominated by heterogeneous and ill-defined categories (e.g., R69, unmapped codes), while the densest regions are dominated by common neoplasms (D23), confirming that the reliability of atlas-derived conclusions is tied to representation density across both demographic and diagnostic axes.

Taken together, 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 heavily represented clusters.

Methods

Ethics

This research complies with all relevant ethical regulations. This study is a secondary analysis of 29 publicly available, fully de-identified dermatological image datasets, each collected and released by its original curators under that source's own ethical approvals and terms of use (a complete list is provided in Supplementary Methods 3). The authors had no access to personally identifiable information or patient identifiers, and no new data involving human participants were collected. Because the study used only publicly available, de-identified data and involved no interaction with human participants, it does not constitute human-subjects research under our institutional policy, and separate ethics approval was therefore not required. Informed consent for the original image acquisition was obtained by the source-dataset curators where applicable. No additional consent was required for this secondary analysis of anonymized, publicly available data.

In accordance with the SAGER guidelines, sex was considered in the study design as one of the audited demographic attributes. Sex was recorded as a binary variable (female or male) as provided in the source-dataset metadata, where it was self-reported or assigned by the original data collectors, and was imputed by the SKINMAP ensemble for images lacking this annotation. Sex-disaggregated distributions are reported in Fig. 3 and in the intersectional prevalence analyses (Supplementary Note 3). Gender identity was not available in the source datasets and was not analyzed. The audited population comprises 1,135,080 images; age is likewise reported and disaggregated (Fig. 3).

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 Supplementary Methods 3. 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 1,135,080 unique images. Beyond identifier-based deduplication, we performed visual near-duplicate detection by searching each image's $k = 5$ nearest neighbors in the embedding space and flagging pairs with cosine similarity exceeding 0.98 (see Methods, "Quality scoring and near-duplicate detection"). This identified 157,629 images falling into 54,825 duplicate groups. After retaining one representative per group, 102,804 (9.1%) are flagged for removal. These visual duplicates are concentrated in large-scale collected datasets: Derm1M accounts for 74.8% (76,858), followed by DermNet (10,498) and Fitzpatrick17k (3412), while curated clinical datasets contain negligible duplicates. A sensitivity analysis confirms that neither these duplicates nor low-quality images materially affect the downstream metrics reported in this study (Supplementary Table 1). The aggregated corpus is heterogeneous, incorporating dermoscopic images, clinical photography, and (TBP) from diverse sources such as institutional archives or educational web resources. Of the images with known imaging modality (52.5%), 401,059 are (TBP) (35.3%), 104,638 are dermoscopic (9.2%), and 91,062 are clinical photographs (8.0%). The remaining 47.5% lack modality annotations in the original datasets. 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, sex, 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[35,36]) to generate concise, clinically coherent captions for all images, ensuring a consistent text-image pair format for subsequent contrastive learning (Supplementary Methods 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.

Quality scoring and near-duplicate detection. To assess image quality, we designed a composite scoring pipeline with four equally weighted components: (1) CLIP zero-shot classification score indicating whether the image depicts a dermatological photograph versus non-medical content, (2) dermatology modality prediction confidence from a linear probe, (3) nearest-neighbor label purity, i.e., the fraction of the k-nearest neighbors in the embedding space that share the same diagnostic label, and (4) linear probe diagnostic confidence. Each component is scaled to [0, 1] and the composite score is their unweighted mean. Applying a threshold of 0.46, we retain 937,044 of 1,135,080 images (82.6%). This threshold is the 17.4th percentile of the composite-score distribution. Quality filtering and near-duplicate removal change the atlas's demographic composition only marginally: the imputed (FST) V–VI share is 12.1% for the full corpus, 10.7% after quality filtering alone, 12.4% after near-duplicate removal alone, and 11.0% after both, and the pediatric (0–17) share stays between 2.1 and 2.3% throughout. Although Derm1M contributes 74.7% of all detected duplicates, near-duplicate removal changes the (FST) V–VI share by less than 0.5 percentage points, and is therefore not the primary determinant of the demographic composition. The demographic figures are likewise insensitive to the quality threshold, remaining within 11.0–11.2% ((FST) V–VI) and near 2.3% (pediatric) across thresholds from 0.30 to 0.50, and changing appreciably only above 0.50, where more than 25% of images are removed (Supplementary Table 2).

For near-duplicate detection, we performed a k-nearest neighbor search ($k = 5$) in the embedding space for every image and flagged pairs whose cosine similarity exceeded a threshold of 0.98. Connected components in the resulting similarity graph defined duplicate groups. Of the 1,135,080 images, 157,629 fall into 54,825 duplicate groups. After retaining one representative per group, 102,804 (9.1%) are removed. The sensitivity of our downstream analyses to both quality filtering and deduplication is reported in Supplementary Table1.

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[37] and SigLIP[38], and on different partitions of the dataset. Image-text pairs were constructed by generating natural-language descriptions from the structured metadata available for each image. For every dataset, a modality-aware template was used to convert the available fields, such as diagnosis, body location, (FST), age, sex, and imaging modality, into a descriptive sentence (e.g., "This dermoscopic image shows a melanocytic nevus on the trunk for a female patient of age 42 with Fitzpatrick type III."). Fields missing for a given sample were simply omitted from the template. The resulting descriptions were subsequently refined using a (LLM) to improve linguistic fluency, as detailed in Supplementary Methods 2. Partitions were constructed by randomly splitting the full dataset into three non-overlapping subsets, stratified by dataset source to ensure that each partition preserved the overall dataset composition. Each multi-modal model was then trained on a different partition, reducing dataset-specific shortcut learning by preventing any single model from memorizing source-specific artifacts. 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[39], iBOT[40], and MAE[41], 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[42] 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 Methods 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, sex, and anatomical region. Throughout this study, sex refers to the biological attribute as recorded in the source datasets. It was self-reported in the original dataset metadata where available, and otherwise estimated by the linear probe described above. Subgroup-stratified accuracy of the sex predictor across (FST) groups is reported in Supplementary Fig. 8. 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`[43]. 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 pre-training. 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 Supplementary Tables 8 and 9.

Yearly novelty. 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 diminishing returns, 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 $\hat{\mathbf{z}}_i = \mathbf{z}_i / \parallel \mathbf{z}_i \parallel_2$, 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 distinct the observed data is compared to random resampling, with values near 1 indicating overlap 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[44],

$$\mathrm{FD}(a,b) = \parallel \boldsymbol{\mu}_a - \boldsymbol{\mu}_b \parallel_2^2 + \mathrm{Tr}\left(\Sigma_a + \Sigma_b - 2(\Sigma_a\Sigma_b)^{1/2}\right). \tag{8}$$

This scalar captures both location shift (difference in means) and shape shift (covariance mismatch) between datasets under the assumed Gaussian approximation. To quantify the stability of the Fréchet distance estimates, we performed bootstrap resampling: for each dataset pair, we drew 200 bootstrap samples (with replacement) from each dataset, recomputed the mean and variance statistics, and derived the Fréchet distance under a diagonal covariance approximation. The bootstrap standard deviation was then used to construct 95% confidence intervals centered on the full-covariance point estimate.

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 over-represented 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 ref. 32. Traditional persistent homology methods fail in high-dimensional spaces with noise because ambient Euclidean distances become uninformative. Damrich, et al.[32] 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[45] 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:

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

As noted by ref. 46, 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 ref. 32, we apply the von Luxburg degree correction[46], yielding the corrected effective resistance:

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

The corrected distance $\tilde{d}_{ij}^{\mathrm{eff}}$ captures the intrinsic manifold geometry by measuring the difficulty of random walks between vertices, remaining stable in the presence of high-dimensional noise[32]. The resulting distance matrix $D^{\mathrm{eff}} = [\tilde{d}_{ij}^{\mathrm{eff}}]$ 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[47]. 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:

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

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[48], which efficiently extracts topological features across all scales.

A hole in the embedding space corresponds to a 1-dimensional homology class (denoted H_1)[47]. Each such class persists over a filtration interval $[\tau_{\mathrm{birth}}, \tau_{\mathrm{death}}]$, 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:

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

Features with high persistence represent topologically significant structures, while low-persistence features typically correspond to noise[47]. 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[48]: The set of edges (vertex pairs) that comprise the one-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:

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

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:

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

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.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All datasets used in this study are publicly available from their original sources under their respective terms of use. A complete list of datasets and access information is provided in Supplementary Methods 3. The datasets were de-identified by the original data curators prior to release. The authors had no access to personally identifiable information or patient keys. We do not redistribute the source images. The metadata harmonization and ICD mapping resources required to reconstruct the unified atlas from these public sources are included in the released code (see Code availability), and together with the pre-trained model, they reproduce the atlas embeddings and imputed metadata. Source data are provided with this paper.

Code availability

The SKINMAP source code, comprising the atlas-construction pipeline (multi-teacher embedding fusion, metadata harmonization, and ICD mapping) and the full analysis suite, is publicly available under the Apache-2.0 license at https://github.com/Digital-Dermatology/skinmap-codeand archived at Zenodo (https://doi.org/10.5281/zenodo.21454008). The pretrained SKINMAP model is available on the Hugging Face Hub at https://huggingface.co/Digital-Dermatology/SkinMap.

References

1. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
2. Hashimoto, T. Model performance scaling with multiple data sources. *Int. Conf. Mach. Learn.* **139**, 4107–4116 (2021).
3. Goyal, S., Maini, P., Lipton, Z. C., Raghunathan, A. & Kolter, J. Z. Scaling laws for data filtering–data curation cannot be compute agnostic. In *Conference on Computer Vision and Pattern Recognition* 22702–22711 (IEEE, 2024).
4. Subramanyam, A., Chen, Y. & Grossman, R. L. Scaling Laws revisited: modeling the role of data quality in language model pretraining. Preprint at arXiv:2510.03313 (2025).
5. Daneshjou, R., Smith, M. P., Sun, M. D., Rotemberg, V. & Zou, J. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. *JAMA Dermatol.* **157**, 1362 (2021).
6. Daneshjou, R. et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Sci. Adv.* **8**, eabq6147 (2022).
7. Salinas, M. P. et al. A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis. *NPJ Digit. Med.* **7**, 125 (2024).
8. Gröger, F. et al. A review and systematic guide to counteracting medical data scarcity for AI applications. *Comput. Methods Prog. Biomed. Update* **8**, 100220 (2025).
9. Groh, M. et al. Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. In *Conference on Computer Vision and Pattern Recognition* 1820–1828 (IEEE, 2021).
10. Daneshjou, R., Yuksekgonul, M., Cai, Z. R., Novoa, R. & Zou, J. Y. Skincon: a skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. *Adv. Neural Inf. Process. Syst.* **35**, 18157–18167 (2022).
11. ISIC Society. ISIC archive (2022).
12. Ward, A. et al. Creating an empirical dermatology dataset through crowdsourcing with web search advertisements. *JAMA Netw. Open* **7**, e2446615 (2024).
13. Gottfrois, P. et al. Passion for dermatology: bridging the diversity gap with pigmented skin images from sub-Saharan Africa. *Int. Conf. Med. Image Comput. Comput. Assist. Interv.* **15003**, 703–712 (2024).
14. Madarkar, S. S. et al. DermaCon-IN: a multi-concept annotated dermatological image dataset of Indian skin disorders for clinical AI research. *Adv. Neural Inf. Process. Syst.* **38**, 124163–124181 (2025).
15. Sun, X., Yang, J., Sun, M. & Wang, K. A benchmark for automatic visual classification of clinical skin disease images. *Eur. Conf. Comput. Vis.* **9910**, 206–222 (2016).
16. Zhou, J. et al. Skincap: a multi-modal dermatology dataset annotated with rich medical captions. Preprint at arXiv:2405.18004v2 (2024).
17. Pacheco, A. G. C. et al. PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data Brief.* **32**, 106221 (2020).
18. Giotis, I. et al. MED-NODE: a computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert Syst. Appl.* **42**, 6578–6585 (2015).
19. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **5**, 180161 (2018).
20. Kawahara, J., Daneshvar, S., Argenziano, G. & Hamarneh, G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE J. Biomed. Health Inform.* **23**, 538–546 (2019).
21. Mendonça, T., Ferreira, P. M., Marques, J. S., Marcal, A. & Rozeira, J. PH 2-A dermoscopic image database for research and benchmarking. In *International Conference of the IEEE Engineering in Medicine and Biology Society* 5437–5440 (IEEE, 2013).
22. Khushbu and Sharun Akter. Skin disease classification dataset. *Mendeley Data* (2024).
23. Memorial Sloan Kettering Cancer Center. MSKCC skin tone labeling dataset (2025).
24. Lara, M. A. R. et al. A dataset of skin lesion images collected in Argentina for the evaluation of AI tools in this population. *Sci. Data* **10**, 712 (2023).
25. Kim, C. et al. Transparent medical image AI via an image–text foundation model grounded in medical literature. *Nat. Med.* **30**, 1154–1165 (2024).
26. Yan, S. et al. A multimodal vision foundation model for clinical dermatology. *Nat. Med.* **31**, 2691–2702 (2025).
27. Orenstein et al, L. A. Outpatient dermatology productivity measures by patient race, sex, and age. *JAMA Dermatol.* **161**, 274 (2025).
28. He, S. Y. et al. Self-reported pigmentary phenotypes and race are significant but incomplete predictors of Fitzpatrick skin phototype in an ethnically diverse population. *J. Am. Acad. Dermatol.* **71**, 731–737 (2014).
29. Department of Economic. *World Population Prospects 2024: Summary of Results* (Stylus Publishing, LLC, 2024).
30. Ahmad, O. B. et al. *Age Standardization of Rates: A New WHO Standard* (World Health Organization, 2001).
31. Kaplan, J. et al. Scaling laws for neural language models. Preprint at arXiv:2001.08361 (2020).
32. Damrich, S., Berens, P. & Kobak, D. Persistent homology for high-dimensional data based on spectral methods. *Adv. Neural Inf. Process. Syst.* **37**, 41954–42014 (2024).
33. Funk, R. J. & Wu, X. When disruption endures: science of science. *Nat. Comput. Sci.* **5**, 440–441 (2025).
34. Park, M., Leahey, E. & Funk, R. J. Papers and patents are becoming less disruptive over time. *Nature* **613**, 138–144 (2023).
35. Yang, A. et al. Qwen2 technical report. Preprint at arXiv:2407.10671 (2024).
36. Qwen Team. Qwen2.5: a party of foundation models. (2024).
37. Radford, A. *Int. Conf. Mach. Learn.* **139**, 8748–8763 (2021).
38. Zhai, X., Mustafa, B., Kolesnikov, A. & Beyer, L. Sigmoid loss for language image pre-training. In *International Conference on Computer Vision* 11941–11952 (IEEE, 2023).
39. Caron, M. et al. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision* 9630–9640 (IEEE, 2021).
40. Zhou, J. et al. Image BERT pre-training with online tokenizer. Preprint at arXiv:2111.07832 (2022).
41. He, K. et al. Masked autoencoders are scalable vision learners. In *Conference on Computer Vision and Pattern Recognition* 15979–15988 (IEEE, 2022).
42. Gröger, F., Wen, S., Le, H. & Brbić, M. With limited data for multimodal alignment, let the STRUCTURE guide you. *Adv. Neural Inf. Process. Syst.* **38**, 168473–168502 (2025).
43. Pedregosa, F. et al. Scikit-learn: machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
44. Dowson, D. & Landau, B. The Fréchet distance between multivariate normal distributions. *J. Multivar. Anal.* **12**, 450–455 (1982).

45. Ghosh, A., Boyd, S. & Saberi, A. Minimizing effective resistance of a graph. *SIAM Rev.* **50**, 37–66 (2008).
46. Hein, M., Audibert, J.-Y. & Luxburg, U. V. From graphs to manifolds–weak and strong pointwise consistency of graph Laplacians. *Int. Conf. Comput. Learn. Theory* **3559**, 470–485 (2005).
47. Edelsbrunner, H. & Harer, J. *Computational Topology: An introduction* (American Mathematical Society, 2010).
48. Bauer, U. Ripser: efficient computation of Vietoris–Rips persistence barcodes. *J. Appl. Comput. Topol.* **5**, 391–423 (2021).

Acknowledgements

We thank Martin A. Walter for valuable suggestions that helped improve the clarity of the manuscript. S.L., L.A. and M.P. are supported by the Swiss National Science Foundation (SNSF) under grant 20HW-1_228541.

Author contributions

Conceptualization, software, visualization, labeling consortium, and writing—original draft: F.G.; Methodology and formal analysis: F.G. and S.L.; Validation: F.G., S.L., P.G. and A.G.J.; Investigation: F.G. and P.G.; Data curation: F.G., P.G. and L.H.; Writing—review & editing: F.G., S.L., P.G., A.G.J., L.H., L.A., M.G. and A.A.N.; Supervision: M.G., M.P. and A.A.N.; Funding acquisition: M.P. and A.A.N.

Funding

The authors received no specific funding for this work.

Competing interests

A.A.N. is a shareholder of Derma2Go AG and Derma.One GmbH. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-78602-w.

Correspondence and requests for materials should be addressed to Fabian Gröger, Marc Pouly or Alexander A. Navarini.

Peer review information *Nature Communications* thanks Lynn Petukhova, who co-reviewed with Rahul Biswas and Jason Weed, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.