nature medicine

Article

https://doi.org/10.1038/s41591-024-02887-x

Transparent medical image AI via an image–text foundation model grounded in medical literature

Received: 9 June 2023

Accepted: 27 February 2024

Published online: 16 April 2024

Check for updates

Chanwoo Kim[1], Soham U. Gadgil[1], Alex J. DeGrave[1,2], Jesutofunmi A. Omiye[3,4], Zhuo Ran Cai[5], Roxana Daneshjou[3,4,6] & Su-In Lee[1,6]

Building trustworthy and transparent image-based medical artificial intelligence (AI) systems requires the ability to interrogate data and models at all stages of the development pipeline, from training models to post-deployment monitoring. Ideally, the data and associated AI systems could be described using terms already familiar to physicians, but this requires medical datasets densely annotated with semantically meaningful concepts. In the present study, we present a foundation model approach, named MONET (medical concept retriever), which learns how to connect medical images with text and densely scores images on concept presence to enable important tasks in medical AI development and deployment such as data auditing, model auditing and model interpretation. Dermatology provides a demanding use case for the versatility of MONET, due to the heterogeneity in diseases, skin tones and imaging modalities. We trained MONET based on 105,550 dermatological images paired with natural language descriptions from a large collection of medical literature. MONET can accurately annotate concepts across dermatology images as verified by board-certified dermatologists, competitively with supervised models built on previously concept-annotated dermatology datasets of clinical images. We demonstrate how MONET enables AI transparency across the entire AI system development pipeline, from building inherently interpretable models to dataset and model auditing, including a case study dissecting the results of an AI clinical trial.

Ensuring the transparency and robustness of medical AI systems involves assessing data and models at every stage, from model training to post-deployment monitoring. However, the tools needed to de-mystify 'black-box' medical AI systems often require medical datasets with dense annotations of human-understandable concepts. For example, for a melanoma classifier, it would be medically meaningful to understand the data and model using clinically relevant concepts such as 'darker pigmentation', 'atypical pigment networks' and 'multiple colors'. Unfortunately, obtaining such labels requires a substantial amount of time from domain experts and, consequently, most medical datasets have limited annotations beyond diagnoses. In contrast, rich annotation with the extensive clinical concepts developed by

[1]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA. [2]Medical Scientist Training Program, University of Washington, Seattle, WA, USA. [3]Department of Dermatology, Stanford School of Medicine, Stanford, CA, USA. [4]Department of Biomedical Data Science, Stanford School of Medicine, Stanford, CA, USA. [5]Program for Clinical Research and Technology, Stanford University, Stanford, CA, USA. [6]These authors jointly supervised this work: Roxana Daneshjou, Su-In Lee. e-mail: roxanad@stanford.edu; suinlee@cs.washington.edu

https://doi.org/10.1038/s41591-024-02887-x

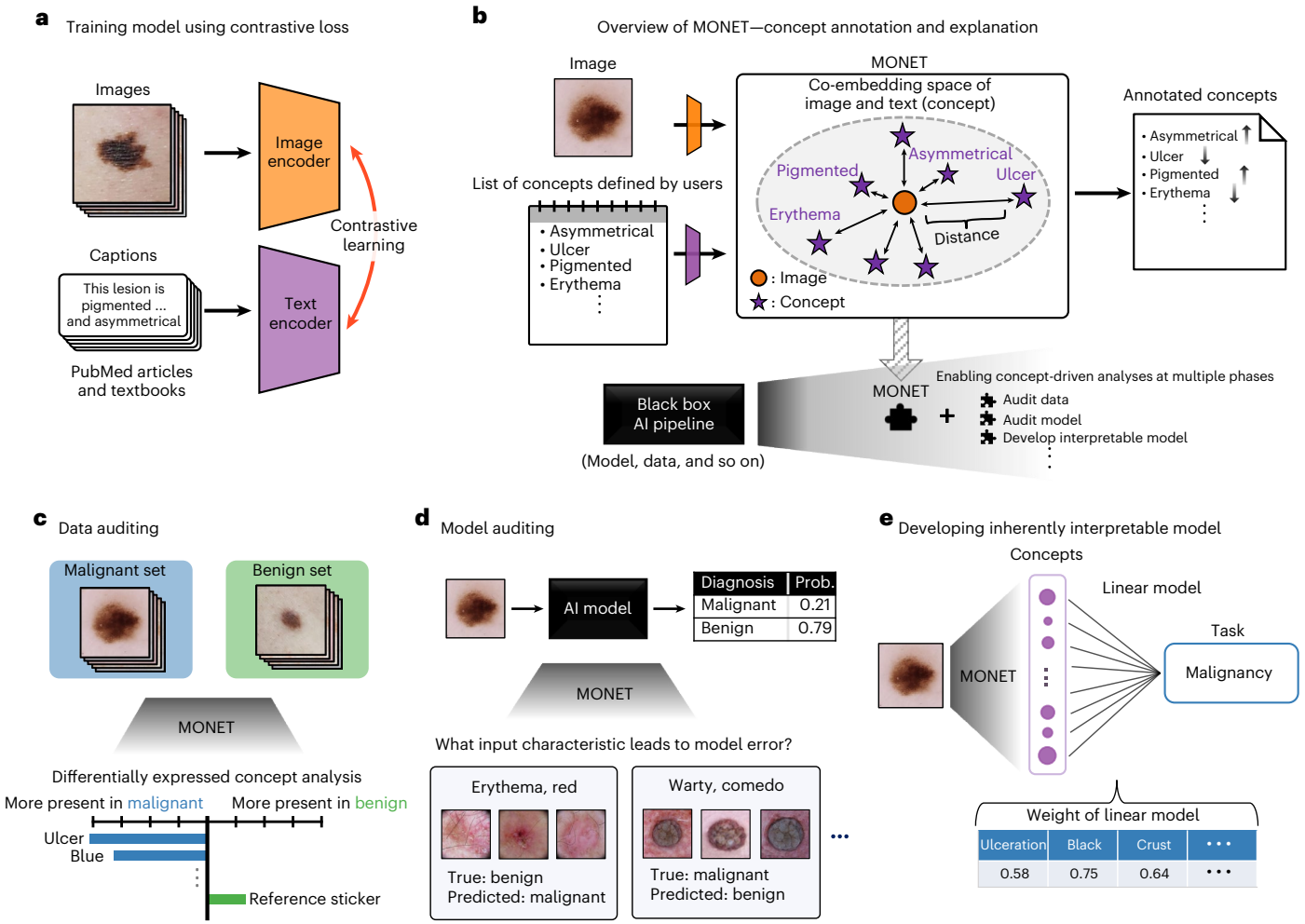


Fig. 1 | Overview of MONET framework and its usage examples. a, Training procedure. MONET is trained using contrastive learning on an extensive set of dermatology image and text pairs collected from PubMed articles and medical textbooks. During the training process, the paired image and text are forced to be close in the joint representation space, whereas those from different pairs are forced to be far apart. **b**, Automatic concept annotation. MONET can map medical concepts and images on to a joint representation space. This allows us to determine the degree to which a user-defined concept is present in an image, by measuring the distance between the image and concept text prompts in the representation space. Its concept annotation capability enables various concept-driven analyses at multiple stages of the medical AI pipeline. **c**, Concept-level data auditing. MONET's automatic concept annotation capability makes it possible to explain the distinguishing features between two sets of data in the language of human-interpretable concepts. This approach facilitates the auditing of large-scale datasets with ease. **d**, Concept-level model auditing. MONET can be used to identify which input characteristic leads to the errors of medical AI models. Prob., Probability. **e**, Developing inherently interpretable models. MONET can be used to develop inherently interpretable medical AI models that operate on human-interpretable concepts aligning with physicians' expectations. These models allow physicians to easily decipher the factors influencing the models' decisions, ensuring high transparency. Figures adapted from refs. 11,12 with permission and from ref. 16 under CC-BY license.

the medical community could enable numerous analyses. Such rich annotations could help researchers developing medical AI systems to understand biases in datasets better and detect undesirable behavior in these systems. Furthermore, these annotations could foster the development of AI systems that better align with physicians' expectations. However, prior efforts[1] have shown that obtaining these data via large-scale efforts by human experts is unfeasible.

In the present study, we instead leverage the collective knowledge of the medical community, as encapsulated in publicly available medical literature and textbooks, to teach an AI model, MONET, to richly annotate medical images with medically relevant concepts (Fig. 1a,b). Given a predefined list of concepts by users, MONET assigns scores to images for each concept, indicating the degree to which the image represents the concept. We focus on dermatology to showcase its versatility because dermatology has heterogeneity in disease appearance across diverse skin tones and imaging conditions (for example, lighting, blurriness) owing to no standardized imaging practices. In this setting, examples of clinical concepts include lesion color (for example, brown) and morphology (for example, nodule). MONET's concept annotation capability allows meaningful trustworthiness analyses across all stages of the medical AI pipeline, as demonstrated by three use cases (Fig. 1c–e): dataset auditing, model auditing and construction of interpretable models.

Prior efforts to densely annotate datasets in medicine[1–4] relied on human experts, limiting the scale of annotation. MONET overcomes these challenges by automatically annotating user-chosen medical concepts. Our framework is built on contrastive learning, an AI technique that enables the direct utilization of natural language descriptions on images[5]. This approach bypasses manual labeling, unlocking the potential of vast numbers of image and text pairs—data on a scale much larger than that accessible with supervised learning.

To train MONET, we collected dermatology image–text pairs (n = 105,550) from PubMed articles and medical textbooks. We mapped an image (or text) into a lower-dimensional vector (a 'representation')

through a neural network (the 'encoder'; Fig. 1a), creating a 'representation space'. During training, an image and text from the same pair are forced into proximity in the representation space, whereas those from different pairs are forced to be farther apart (Methods). Once trained, MONET's zero-shot capability enables annotation of a medical concept without a separate learning procedure (Fig. 1b and Methods). When a user provides an image and a list of concepts to annotate, MONET determines the presence of each concept in the image by calculating the distances between the image and concept text prompts in the joint representation space. After training MONET, we assessed MONET's performance in annotation, as well as its performance in a variety of use cases related to AI transparency.

Results

Automatic concept annotation

We first assessed MONET's concept annotation capability before demonstrating how it could improve the transparency and interpretability of medical AI pipelines. We evaluated MONET's concept annotation ability by identifying those images with the highest concept presence scores using both clinical and dermoscopic images, two widely used dermatological image types (Methods). For our evaluation, we employed clinical images (n = 4,910) from the Fitzpatrick 17k[6] and Diverse Dermatology Image (DDI)[7] datasets and dermoscopic images (n = 71,665) from the International Skin Imaging Collaboration (ISIC)[8–16] and Derm7pt[3] datasets (Methods).

We observed that MONET successfully retrieves relevant clinical and dermoscopic images for a variety of dermatological terms (Fig. 2 and Supplementary Figs. 1–3). For instance, when prompted with 'erythema', describing a type of redness or violaceousness, MONET retrieves images exhibiting such redness, as verified by two board-certified dermatologists. Prompting with the concept 'blue' retrieves images of dark-blue lesions, which appear to be the result of pigmentation in the dermis. MONET was also able to retrieve images with primary morphological features such as papules and nodules, as well as secondary morphological features such as ulceration.

We assessed the performance of MONET's concept annotation using ground-truth concept labels from two datasets: SkinCon[1] for clinical images and Derm7pt[3] for dermoscopic images (Table 1). After excluding concepts with <30 positive examples, the test comprised 21 concepts over 1,635 clinical images and 7 concepts over 1,011 dermoscopic images. We compared MONET's performance with a supervised learning approach, training a ResNet-50 model using ground-truth concept labels[17], and with a pre-existing contrastive image–text model that was not specifically trained on dermatology images but on 400 million available image–text pairs on the web—the CLIP (Contrastive Language-Image Pretraining) model by OpenAI[5] (Supplementary Methods). We found that MONET typically outperforms CLIP on both clinical and dermoscopic images (Table 1). This is expected, given that MONET is essentially a fine-tuned version of CLIP on dermatology data. MONET also performs competitively with the fully supervised baseline, with MONET often leading on clinical images and the fully supervised model often leading on dermoscopic images.

In addition, we conducted the same comparative analysis using disease labels, which can be viewed as the most fine-grained concept labels; we mapped the disease labels instead of SkinCon concepts to the image–text joint representation space. Our findings indicated that MONET's performance is superior to the baseline CLIP model for both clinical and dermoscopic images and still comparable to that of supervised models for clinical images (Supplementary Table 2). Our finding with clinical images is consistent with a previous study, which found that a contrastive learning model trained on radiology images performed competitively with supervised learning models[18].

We also evaluated the performance of MONET's concept annotation across diverse skin tones (Methods). A recent study revealed that state-of-the-art dermatology AI models exhibit uneven performance across skin tones, particularly underperforming on dark skin tones[7], which can lead to disparities in diagnosis and clinical outcomes for patients with skin of color[19,20]. We compared MONET's performance per skin tone using the Fitzpatrick skin type (FST) labels included in the Fitzpatrick 17k and DDI datasets and found that it demonstrates consistent performance (Supplementary Table 4).

Furthermore, we explored MONET's capability to recognize nonclinical concepts, such as artifacts that are irrelevant to the diagnosis. Many studies have shown that medical AI systems use such nonclinical concepts to make predictions, particularly when a spurious correlation exists between the artifacts and the prediction labels[21,22], compromising the model's generalizability[23–25]. The ability of MONET to identify such artifacts, in addition to clinical concepts, would facilitate more fine-grained auditing and debugging of medical AI pipelines. MONET successfully retrieved images with common dermatological artifacts, such as purple ink markings (Extended Data Fig. 1a), which dermatologists commonly use to mark lesions before biopsy, and orange stickers (Extended Data Fig. 1b), which appear predominately on the pediatric cases of the ISIC dataset. As these artifacts predominantly appear in certain types of images, they may inadvertently cause AI algorithms to associate purple ink markings with malignancy and orange stickers with benign lesions[23–25]. Also, MONET automatically identifies images with body location features (such as nails and hair) (Extended Data Fig. 1c,d). A recent study has shown that anatomical locations may play a critical role in the performance of dermatology AI algorithms; however, most datasets lack these annotations[26]. Furthermore, MONET identifies images with dermoscopic borders, which appear on a subset of dermoscopic images depending on the image processing pipeline (Extended Data Fig. 1e). We also compared quantitatively against CLIP's ability to identify these artifacts and found that MONET typically outperforms this baseline, as measured by the number of top 100 images that indeed exhibit the following artifacts: purple pen markings (92 accurate with MONET versus 83 with CLIP), orange sticker (70 versus 54, respectively), hair (100 versus 97) and dermoscopic borders (95 versus 12).

Last, we examined MONET's ability to annotate skin tones, motivated by the need for such annotations when analyzing skin-tone-related biases[6,7,27–32]. MONET achieved an area under the receiver operating characteristic curve (AUROC) of 0.805 in annotating FSTs, whereas CLIP achieved 0.736 (Supplementary Table 3). A supervised baseline (ResNet-50) outperformed MONET with an AUROC of 0.874, although this performance may be biased upward as a result of the 'internal' testing scenario (that is, the test data are held out from the training data, but originate from the same distribution), whereas the same test data result in a more rigorous 'external' test scenario for MONET and CLIP.

In the following sections, we showcase how MONET can be used to improve the transparency and trustworthiness of dermatology AI using real-world data.

Data auditing

Ensuring that training data align with users' expectations is a crucial first step toward developing AI models because many unreasonable model behaviors stem from unidentified pitfalls in the training data[21,23,25,33]. For example, in dermatology AI, differentiating features between classes (that is, malignant and benign images) in the training data should not contain any spurious correlations, such as the pen markings used to identify biopsied lesions[25]. On identifying any irregularities, adjustments can be made, such as improving the data collection and processing[34,35] or applying optimization techniques to improve generalizability[22,36–38]. However, existing approaches for examining datasets for irregularities, which require manual labeling[39] or inspection[21,40], are labor intensive and thus not scalable to large datasets (Supplementary Discussion).

To address the issue, we employed MONET to automate data examination. MONET's capability to automatically annotate concepts

https://doi.org/10.1038/s41591-024-02887-x

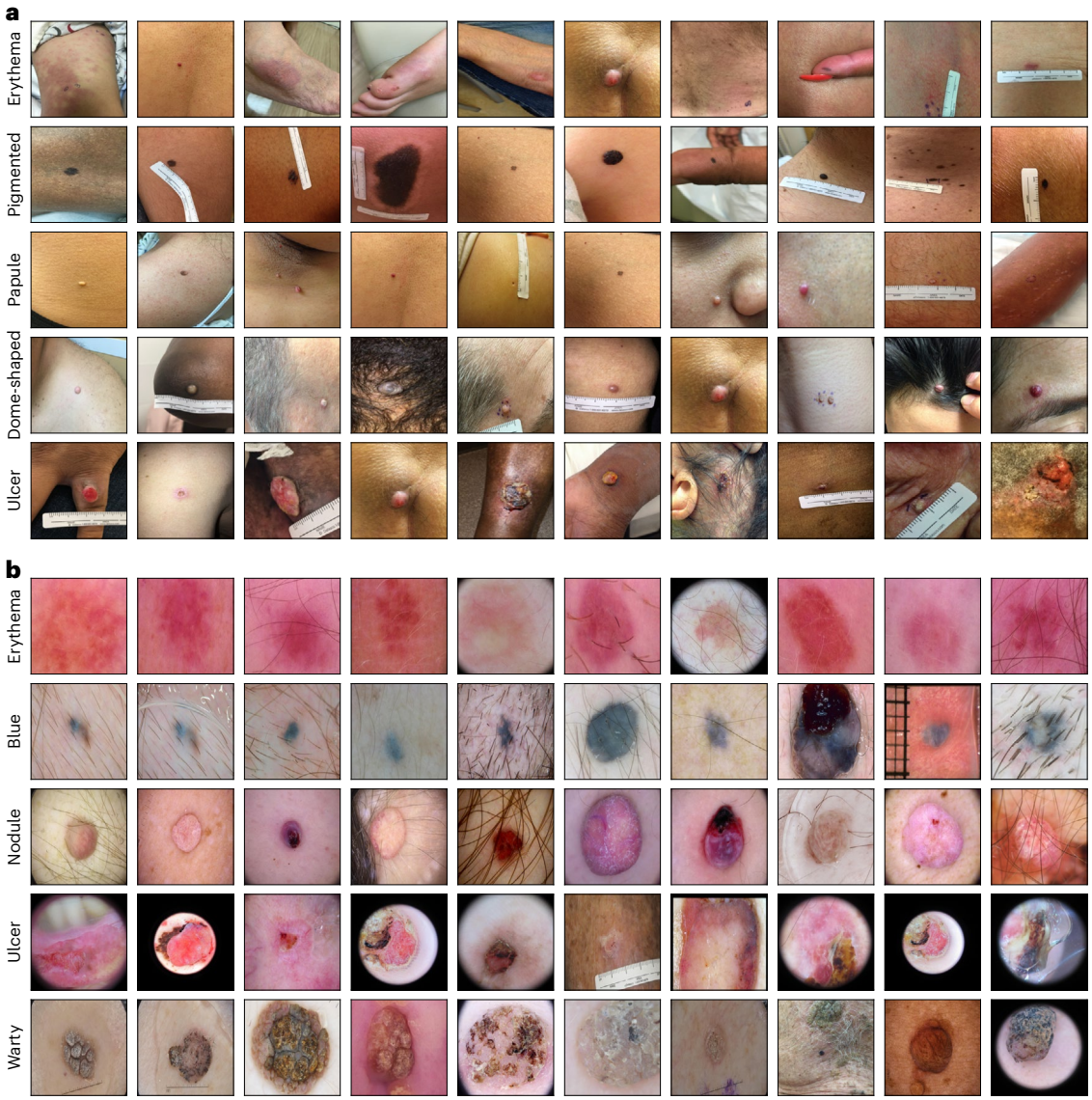


Fig. 2 | Images with high concept presence scores calculated using MONET. The concept presence score represents the degree to which a concept is present in an image. Each row displays the top ten images for each concept. **a**, Clinical images from the DDI datasets. **b**, Dermoscopy images from the ISIC and Derm7pt datasets. We excluded images for public display as a result of license constraints; for completeness, we list the filenames of these files in Supplementary Table 1. The images shown are adapted from refs. 3,7,11,12 with permission and from refs. 13,14 under CC-BY license.

helped identify distinguishing features between any two arbitrary sets of images in human-understandable terms, which we refer to as concept differential analysis (Methods).

We first assessed MONET's performance at concept differential analysis via a benchmark: we constructed two datasets such that a predefined concept was differentially represented, by using ground-truth concept labels in SkinCon or Derm7pt, and then we checked whether MONET's concept differential analysis correctly identified that known concept. We observed that MONET successfully detects differentially expressed concepts in both clinical and dermoscopic datasets (Supplementary Figs. 4 and 5).

As a practical use case, we employed MONET to analyze the ISIC dataset, the largest dermoscopic image dataset, which consists of >70,000 publicly available images commonly used to train dermatology AI models[8–16,41]. We divided the images into a malignant (n = 9,990) and a benign set (n = 60,525), assuming malignancy as the prediction target, and examined which concepts were more present in which set (Fig. 3a). We tested for 48 concepts listed in SkinCon and 7 concepts listed in Derm7pt along with 8 artifacts, including red coloration, pink coloration, purple ink markings, nails, hair, orange sticker, gel and dermoscopic borders.

The top five concepts in the malignant images are ulcer, blue-whitish veil, erosion, warty and pink coloration, whereas the top five concepts in the benign images are orange sticker, hypopigmentation, the color salmon, xerosis and pigment network. They encompass clinically pertinent features as well as spurious confounders. Skin ulcerations and erosions are commonly linked to malignant skin tumors such as melanoma, basal cell carcinoma and squamous cell carcinoma, making their association with malignancy logical. On the other hand, prediction models learning from confounding concepts may lead to biases and detrimental consequences. For example, dermatologists use orange stickers as a lesion marker and, with the ISIC dataset, this was predominantly used in the pediatric population which had mostly benign lesions. The bias in the data could lead a model to erroneously associate orange stickers with a low likelihood of malignancy.

Furthermore, we used this approach to assess trends specific to different data sources. In medicine, data sharing across institutions is limited owing to the sensitive nature of medical data and, in many cases,

Table 1 | Performance of MONET's concept annotation compared with the baselines

Method	AUROC±s.d. (no.)	
	Labels in SkinCon	Labels in Derm7pt
MONET	0.767±0.086 (19/21)	0.704±0.081 (4/7)
CLIP	0.691±0.113 (8/21)	0.563±0.069 (1/7)
ResNet-50 (fully supervised)	0.703±0.086 (11/21)	0.793±0.069 (6/7)

We used concept labels in the SkinCon and Derm7pt datasets as ground truth. Of the 48 concepts in the SkinCon dataset, we excluded any with <30 positive examples, leaving 21 concepts for our analysis. All 7 concepts in Derm7pt had ≥30 positive samples. The baseline methods are CLIP, an image–text model not specifically trained on dermatology images, and the ResNet-50 model trained on ground-truth labels in a fully supervised manner. We compared the performance using the AUROC across concepts with ground-truth labels. We examined AUROC to report the models' performance across all operating points. Performances were reported as mean AUROC±s.d. across concepts. The numbers in parentheses represent the count of concepts for which the method achieves an AUROC >0.7 and the total number of concepts examined.

medical AI models are developed within a few institutions and then distributed to other institutions for deployment. For this reason, it is important to understand and monitor the shifts in the concept representations between datasets and identify factors that can compromise the transferability of medical AI models. The ISIC dataset contains images from multiple institutions and serves as an ideal resource for simulating situations where the development and deployment sites differ. In this analysis, we focused on two cohorts released in the ISIC Challenge 2019—the Medical University of Vienna (Med. U. Vienna, malignant: $n = 1{,}824$; benign: $n = 8{,}187$) and the Hospital Clínic de Barcelona (Hospital Barcelona, malignant: $n = 6{,}021$; benign: $n = 6{,}250$)—because they represent the two largest cohorts in the entire ISIC dataset when stratified by release year and data source. We performed concept differential analysis between malignant and benign images for each cohort separately and then compared the obtained concept differential expression scores between the two cohorts (Fig. 3b).

Sorting the test concepts by absolute differences, the top listed concept with opposite signs between two institutions was a 'red' hue. At Hospital Barcelona, redness positively correlated with malignancy, whereas at Med. U. Vienna it correlated negatively. This posits that redness has the potential to compromise the transferability of medical AI models between two institutions. This trend is visible in the red images from each cohort, as well (Fig. 3c,d). These experiments thus demonstrate that MONET can assist with auditing large-scale datasets.

Model auditing

Although multiple techniques have been developed to 'audit' a model by understanding the factors involved in AI decision-making[42–49], numerous issues have been highlighted with regard to the de facto standard family of techniques: saliency maps. For example, these techniques highlight the regions of an input image that contribute most to a model's prediction[42–44], but the highlighted pixels are often not easily translated into semantically meaningful concepts[50–52].

To address this issue, we developed a method 'model auditing with MONET' (MA-MONET), which uses MONET to automatically detect semantically meaningful medical concepts that lead to model errors (Methods). MA-MONET compares concept presence scores between two visually similar clusters of images with large differences in model performances—if two visually similar clusters (one high performing, the other low performing) differ in terms of a few concepts, these differing concept terms can be hypothesized as leading to high error rates.

To validate MA-MONET, we first performed a benchmarking analysis, using a situation where the ground truth (that is, the concepts leading to model error) was already known (Fig. 4a and Methods).

In the benchmarking analysis, we trained a convolutional neural network (CNN) to predict malignancy on a dataset where a particular SkinCon concept was spuriously correlated with malignancy and tested it on a dataset where the correlation was reversed (Supplementary Methods). We observed a substantial drop in mean AUROC from 0.782 for validation sets to 0.463 for test sets, across 100 simulation settings. We applied MA-MONET to observe whether it identified the known concepts that were spuriously correlated. We measured the frequency of the known spurious correlation being identified by MA-MONET (Fig. 4b and Supplementary Fig. 6).

To showcase MONET-MA's use in real-world scenarios, we considered a common situation in which a model is trained at one institution and deployed at another[53,54] (Fig. 4c). Similar to our data-auditing experiments, we trained CNN models on one institution's data and tested on data from another.

The performance of an AI model trained on Med. U. Vienna dropped between internal and external validation (AUROC 0.883 and 0.697, respectively), which may prompt developers to question which input characteristics led to errors. Application of MA-MONET identified characteristics associated with high error rates (Fig. 4d and Supplementary Fig. 7), such as a cluster of images labeled 'blue-whitish veil', 'blue', 'black', 'gray' and 'flat topped'. Identified characteristics also included terms related to 'red' (for instance, Supplementary Fig. 7i) where malignant images are predominantly misclassified as benign, in alignment with our data-auditing experiments.

Conversely, we trained an AI model on Hospital Barcelona data and tested it on Med. U. Vienna data, observing the AUROC drop from 0.832 in internal validation to 0.717 in external validation. In MA-MONET results, the cluster with the highest error rate was characterized by the concepts 'erythema', 'regression structure', 'red', 'atrophy' and 'hyperpigmentation' (Fig. 4e and Supplementary Fig. 8). This observation also aligned with the trends noted in the data-auditing experiment.

Inherently interpretable model building

In medicine, inherently interpretable models, such as concept bottleneck models (CBMs)[55,56], are of particular interest because they allow physicians to easily decipher the factors influencing a model's decision. Unlike complex black-box models, CBMs make predictions in a two-step manner: first, they predict concepts from the input using a CNN (that is, input → concept); then, they use these predicted concepts to predict the target output via a linear model (that is, concept → output). As each node in the bottleneck layer represents a human-interpretable concept, CBMs offer a greater ability to explain. Furthermore, CBMs facilitate the incorporation of domain knowledge into models by imposing constraints on the concepts used, giving better control over model behavior.

However, CBMs are limited because they require concept annotations in the training data and such data are not always available. We thus evaluated whether MONET's automatic concept annotation could eliminate the need for manual concept labeling and facilitate CBM development.

We explored the application of MONET and CBMs for melanoma and malignancy prediction, the most common prediction tasks for dermatology AI models. We first compared 'MONET + CBM', our automatic concept scoring approach (Fig. 5a and Methods) with the manual labeling method (Fig. 5b). For both targets, the manual approach's training set was about a third of MONET + CBM's because it is limited to images with SkinCon labels; MONET + CBM uses all training samples and concepts because it can automatically annotate concepts without expert labeling. We observed that access to a larger number of concepts and samples enhanced performance (Fig. 5b–e). Although the manual approach improved with more data, it still fell short of MONET + CBM, which benefits from a larger sample size (Methods).

Next, we benchmarked MONET + CBM against other baselines, such as supervised black-box models and CLIP-based CBM, neither of which requires ground-truth concept labels (Methods). This allowed

https://doi.org/10.1038/s41591-024-02887-x

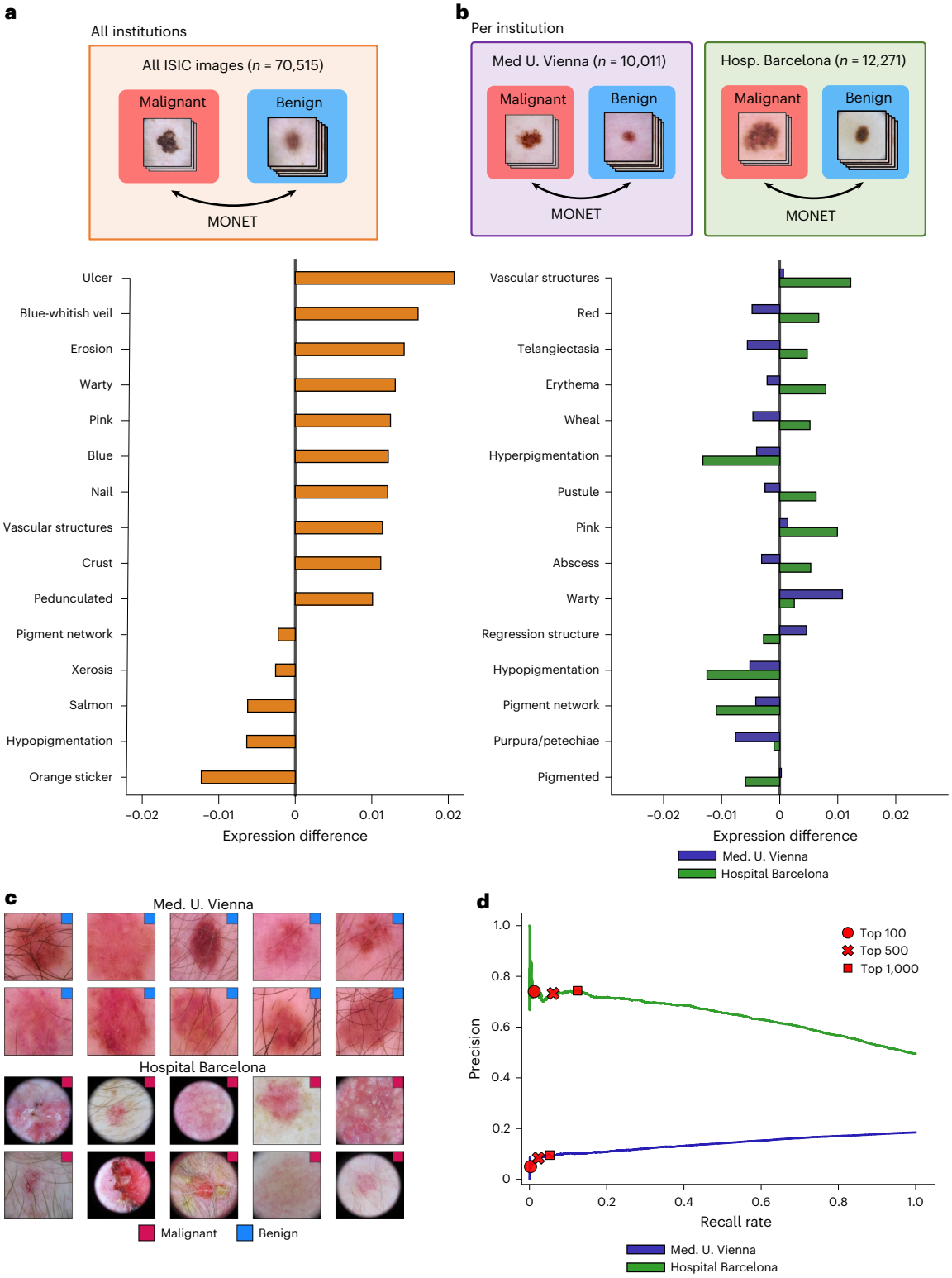
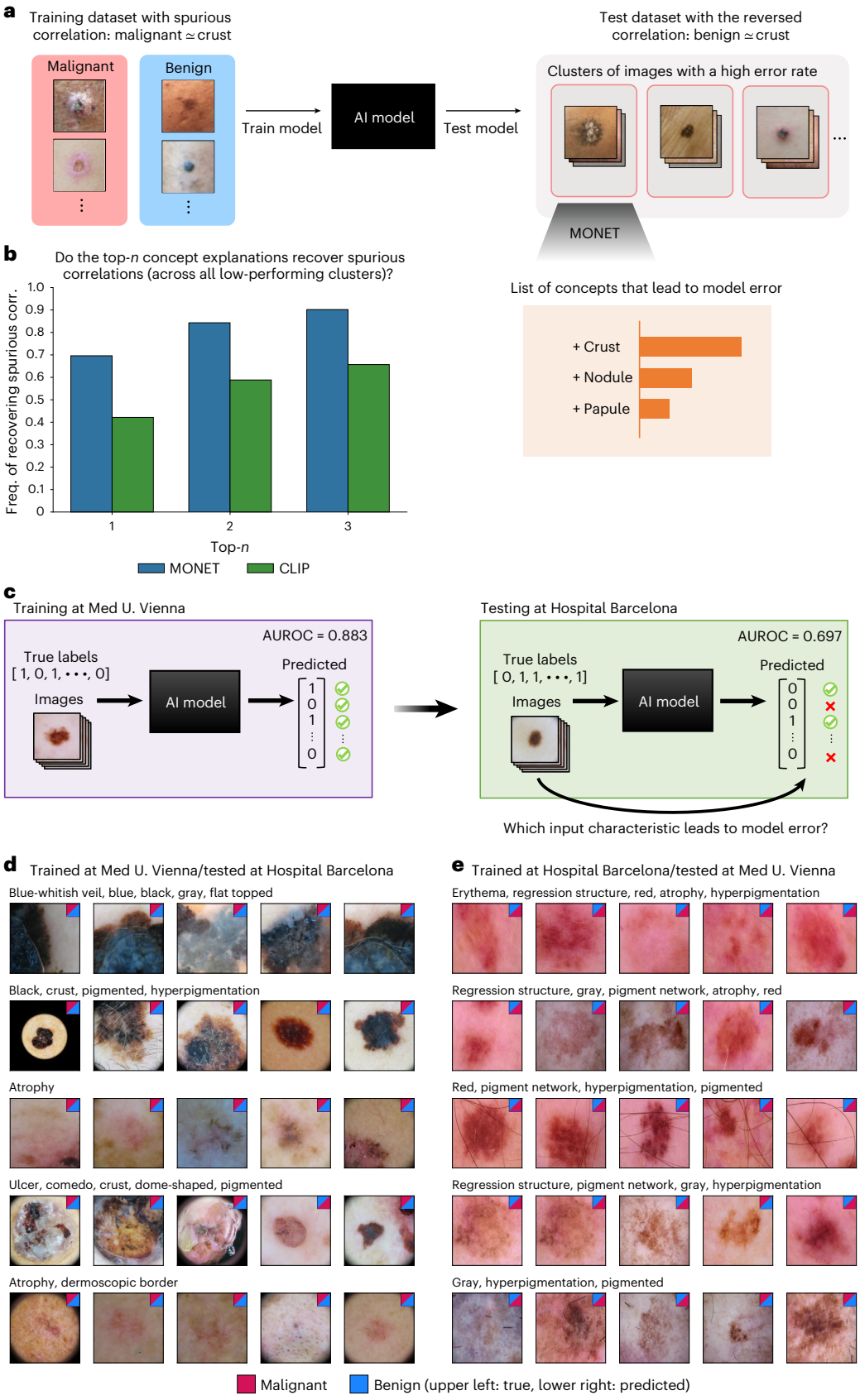


Fig. 3 | Concept-level data auditing. a, Concept differential analysis between malignant images and benign images in the ISIC dataset. We showed the top ten concepts with positive values and the top five concepts with negative values. A positive value means that the concept was more present in the malignant images than in the benign images, and vice versa. **b**, Concept differential analysis between malignant and benign images per data source in the ISIC dataset to identify data-source-specific trends. The purple bar represents the output from Med. U. Vienna and the green bar the output from Hospital Barcelona. We showed the top 15 concepts based on their absolute differences between the 2 cohorts. **c**, Examples of red images in each cohort. We displayed 10 randomly selected images from the top 100 images in each cohort that had high concept expression scores for redness. **d**, Precision-recall curve for images in each cohort. The images in each cohort were sorted based on their concept presence scores for redness and then compared with their malignancy labels. Precision was defined as the proportion of malignant images above a certain threshold out of all images above that threshold, whereas recall rate was defined as the proportion of malignant images above the threshold out of all malignant images. The top 500 and top 1,000 red images from Hospital Barcelona still contain more malignant than benign samples. The figures have been adapted from refs. 11,12 with permission and from ref. 16 under CC-BY license.

https://doi.org/10.1038/s41591-024-02887-x



us to use other larger datasets lacking concept labels and task-relevant curated concepts in the bottleneck layer. Dermatologists selected ten target-relevant curated concepts for the bottleneck layer of MONET + CBM and CLIP + CBM, which can both flexibly label concepts.

For predicting malignancy and melanoma from clinical images, MONET + CBM outperforms all baselines in terms of the mean AUROC (malignancy: 0.815, melanoma: 0.881) and one-sided, paired Student's *t*-tests ($P < 0.01$) (Fig. 5f,g). This demonstrates that MONET enables the

Fig. 4 | Concept-level model auditing. a, A benchmark analysis to see how well MA-MONET can identify the semantically meaningful concepts that lead to model error. To this end, we generated settings where we knew the ground truth (that is, concepts that lead to model errors); we created a training and a test dataset with spurious correlation (corr.). We used MA-MONET to identify which concepts led to model error for an AI model trained on this confounded dataset. MA-MONET returned a ranked list of concepts that explain model errors. **b**, The frequency (freq.) of the known spurious correlation recovered by MA-MONET. We compared MONET's frequency with that of the out-of-the-box CLIP model[5] and found that the results are markedly inferior to MONET: the low-performing clusters being analyzed in each setting are the same for both methods, but the comparison of concept presence between the high-performing and low-performing clusters is done using CLIP instead of MONET. **c**, A malignancy prediction model trained using images from one institution and tested on images from another. **d**, A model trained on the Med. U. Vienna dataset and tested on the Hospital Barcelona dataset. **e**, A model trained on the Hospital Barcelona dataset and tested on the Med. U. Vienna dataset. **d**,**e**, Each row displaying one of the top five clusters, sorted by high error rates. For each cluster, we showed the misclassified images and the corresponding concepts associated with errors. We represented the true and predicted labels for each image by the color of the upper left and lower right triangles in the small box, respectively. The numbers at the top right compare the number of malignant and benign samples for the true and predicted labels. The five misclassified images shown for each are selected based on the average concept presence of the identified concepts. The figures have been adapted from refs. 7,11,12 with permission.

creation of models that are both interpretable and high performing. We also verified whether the way MONET + CBM uses these concepts aligns with established clinical rules (Fig. 5h,i). For the melanoma target, the results are consistent with the ABCDEs of melanoma[57]—asymmetry, border irregularity, color variation, diameter and evolution—that differentiate malignant melanomas from benign melanocytic nevi. All concepts that coincide with the 'ABCDEs' of melanoma (for example, asymmetry) get a positive weight, whereas the concepts 'white' and 'regular' get a negative weight. For the malignancy target, no well-defined guidelines exist for deriving concepts, but we observed similar trends to melanoma prediction.

For dermoscopic images, we employed images from the ISIC dataset (Methods). In the same analysis as with the clinical images, we found that the supervised black-box models consistently outperformed in both malignancy and melanoma prediction tasks (Fig. 5j–k). Our investigation into the linear classifiers' weights shows general alignment with prior knowledge (Fig. 5l,m); for melanoma prediction, the concepts 'erosion' and 'multiple colors' have high positive weights, 'tiny' has a high negative weight and 'blue' has a positive weight, referring to the dermoscopy concept of blue-whitish veils observed in melanomas.

We further evaluated our models for a common scenario: training at one institution and deploying at another. Following the setup from the model auditing section, we trained models on images from Hospital Barcelona (n = 12,271; 6,021 malignant versus 6,250 benign) and tested on images from Med. U. Vienna (n = 10,011; 1,824 malignant versus 8,187 benign) and vice versa (Extended Data Fig. 2). For both prediction targets in each scenario, we found that black-box models exhibited substantial generalization gaps (that is, performance drops between internal and external test sets), whereas MONET + CBM exhibited better external performance and smaller generalization gaps (Extended Data Fig. 2).

Last, we explored how concept selection affects MONET + CBM's performance, particularly whether task-relevant concepts in bottleneck layers enhanced a model's performance. We compared CBMs operating on our curated, task-relevant concepts against those using SkinCon concepts, which, although clinically relevant, were not specifically tailored for melanoma or malignancy predictions. Our extensive testing with various concept subset sizes showed that CBMs with our curated concepts consistently outperformed those with SkinCon concepts in all cases (Extended Data Fig. 3). Notably, in interinstitution transfer scenarios, CBMs leveraging all 10 of our curated concepts even surpassed those using all 48 SkinCon concepts. In addition, we investigated the implications of including the concept 'red' in the bottleneck for MONET + CBM's performance in interinstitution transfer settings (Supplementary Table 5), because the concept 'red' was identified as one that exhibits differing associations with malignancy between the two institutions, which could potentially degrade the model's generalization capabilities. We observed that adding 'red' increases validation AUROC but decreases test AUROC in both transfer settings, as expected. These observations demonstrate that careful curation of concepts in MONET + CBM enables control of model behavior and leads to benefits in performance.

In summary, MONET enables building an inherently interpretable model by annotating concepts at scale.

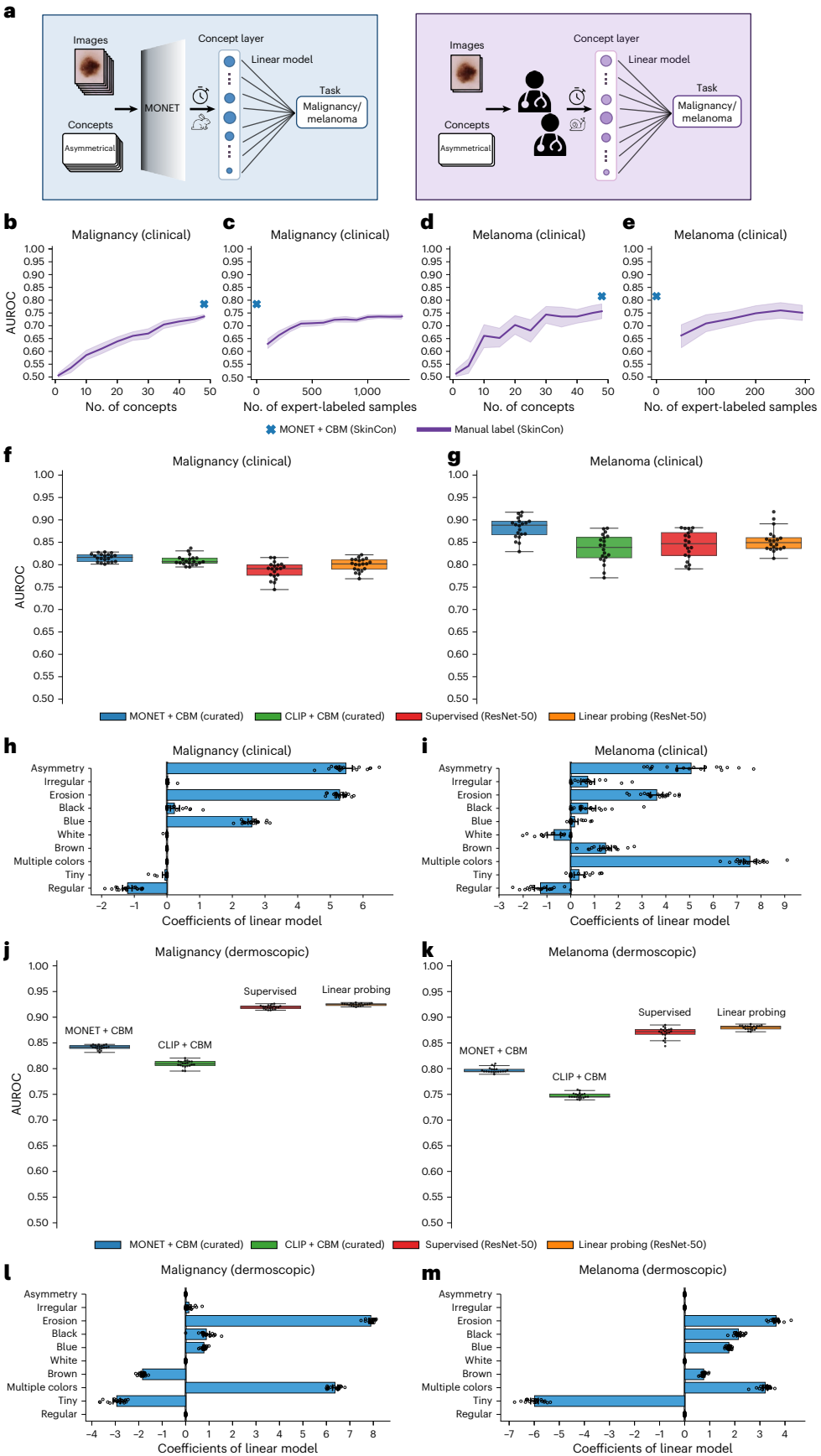
Assessing a clinical trial of a dermatology AI algorithm

To further demonstrate applications of MONET in the real world, we applied MONET to better understand failure cases in a recent prospective clinical trial (PROVE-AI) of a dermatology AI algorithm (ADAE)[15]. As this model had very high sensitivity, we expanded on the subgroup analyses performed in the trial to help explain the lower-than-desired specificity (0.341). Based on MONET's analysis, additional concepts (Supplementary Table 6 and Supplementary Figs. 9 and 10), including

Fig. 5 | Concept bottleneck model. a, CBM built using concepts annotated by MONET (blue). The model first annotates concepts using MONET and then predicts disease labels by combining them via a linear model. The CBM was built using concepts manually labeled by experts (purple). **b**–**e**, Performance of malignancy (**b**,**c**) and melanoma (**d**,**e**) prediction models trained using manual labels (SkinCon) with respect to the number of concepts and the number of expert-labeled samples. MONET + CBM is shown as a cross mark because it can utilize all concepts without expert annotation. We used clinical images from the Fitzpatrick 17k and DDI datasets, which accompany 48 SkinCon concept labels for a subset of the images, and used those 48 concepts in the bottleneck layer. For malignancy prediction, we used all images (n = 4,910; 1,129 malignant versus 3,781 benign), whereas, for melanoma prediction, we used filter images to mirror well-defined clinical tasks (n = 769; 321 melanomas versus 448 melanoma look alikes). **f**,**g**, Performance comparison of malignancy (**f**) and melanoma (**g**) prediction models for clinical images. Unlike **b**–**e**, MONET + CBM uses task-relevant concepts curated by dermatologists. We compared AUROC values between MONET + CBM and other methods using one-sided, paired Student's t-tests (Methods): for malignancy, P values were 5.182×10^{-3} for CLIP + CBM, 1.732×10^{-6} for supervised (ResNet-50) and 9.985×10^{-6} for linear probing (ResNet-50); for melanoma, 2.315×10^{-9}, 2.084×10^{-5} and 2.615×10^{-5}, respectively. **h**,**i**, Coefficient of the linear model in MONET + CBM for malignancy (**h**) and melanoma (**i**) predictions on clinical images. Each dot represents the coefficient from one of n = 20 different train–test splits. **j**,**k**, Performance comparison of malignancy (**j**) and melanoma (**k**) prediction models for dermoscopic images. We used images from the ISIC dataset: for malignancy prediction, 70,515 images (9,999 malignant versus 60,516 benign) and, for melanoma prediction, 43,678 images with disease-specific labels (5,900 melanoma versus 37,778 non-melanomas). In one-sided, paired Student's t-tests, as performed in **f** and **g**, the following P values were observed: for malignancy: 1.105×10^{-18} for CLIP + CBM, 1 for supervised (ResNet-50) and 1 for linear probing (ResNet-50); for melanoma, 4.419×10^{-22}, 1 and 1, respectively. **l**,**m**, Coefficient of the linear model in MONET + CBM for malignancy (**l**) and melanoma (**m**) predictions on dermoscopic images. For each setting of **b**–**m**, evaluations are repeated with n = 20 different train–test splits. The shaded areas, error bars and box plots are shown for the 20 evaluations. The shaded areas and error bars indicate the 95% confidence interval, with the center line representing the mean. In the box plots, the boxes represent the interquartile range (IQR) with their lower and upper bounds corresponding to the first and third quartiles, respectively, and the center lines indicate the medians. The whiskers extend to 1.5× IQR from the box. Each dot represents the value from each evaluation. The figures have been adapted from ref. 16 under CC-BY license.

https://doi.org/10.1038/s41591-024-02887-x

'poikiloderma', 'pink' and 'erythema', are associated with low specificity. These concepts were not present in the original metadata and thus could not have been identified by the original subgroup analyses without costly annotation of numerous concepts. Indeed, a review of the top 100 benign images for each of those above concepts shows a statistically significant level of misclassification (Supplementary Table 6).



Discussion

Even with the regulatory approval of AI-supported medical devices, much of the AI pipeline is not transparent from large-scale datasets, which may contain biases, to so-called 'black-box' models that are not easily audited or interpretable[41]. One approach to improving transparency and trustworthiness is identifying semantically meaningful, human-understandable concepts present in datasets or used by models. However, to date, all datasets and methods developed using concepts have relied on human labeling and domain experts, which is intractable for large-scale, real-world deployment.

In the present study, we used an image–text foundation model for automated concept labeling in a medical domain that would usually require domain expertise and we showcased how these automatically annotated concepts could be used to perform tasks for trustworthy AI development at all stages, from developing new models to auditing existing datasets and models. We focused on dermatology owing to the heterogeneity of the image data, the large number of potential concepts and the ability to validate our methods on existing datasets.

Prior work using image–text foundation models in medicine focused on self-supervised training for diagnostic tasks, such as in radiology or pathology[18,58]. MONET is not intended to be a diagnostic model; rather, we aimed to develop a model that can label a vast number of human-understandable concepts across two image modalities within dermatology: clinical and dermoscopic images. To solve this challenge, we collected a large number of dermatology image–text pairs from PubMed articles and medical textbooks and trained an image–text model, MONET. MONET facilitates large-scale annotation of concepts and we showed its utility in improving the transparency and explainability of dermatology AI systems. We anticipated that similar approaches could be applied to other medical domains and to updated models as image–text AI systems evolve. For instance, recently proposed updates to CLIP solve technical issues—such as failure to account for synonyms—by leveraging external medical databases[59] or knowledge embedded in large language models[60,61], and our core idea of concept annotation is readily adaptable to these updates.

For a concept annotation task, using domain-expert labels as the ground truth, we found that MONET, which requires no expert labeling, outperformed the baseline CLIP model for both clinical and dermoscopic images and a supervised model trained on expert labels for clinical images. Supervised models performed best on dermoscopic images, possibly because the training data for MONET contain a larger proportion of clinical images than dermoscopic images. However, in our tests, the performance of MONET for labeling concepts in dermoscopic images was sufficient for tasks such as data and model auditing. Our findings demonstrate that image–text foundation models, built from medical corpora, can overcome the bottlenecks of data labeling and domain expertise.

After demonstrating MONET's ability to annotate concepts, we showcased its ability to facilitate AI auditing and transparency in dermatology. For example, artifacts such as pen markings, stickers and hair are known to affect dermatology model performance[25,62]. However, most studies do not assess the influence of artifacts on their models because their datasets are not labeled with these. We demonstrated MONET's ability to automatically identify these artifacts, aiding in both data and model auditing. As an example of how this kind of auditing is useful, we analyzed data from the ISIC 2019 challenge and found differences in the association of a 'red' hue and malignant images across different institutions. This may lead to confounding if a model is trained on one site's dataset and tested on another, as we see when we implement MA-MONET for model auditing. These insights derived from using MA-MONET might not be readily achievable via conventional saliency map techniques[50–52] because the 'red' hue is not a localized attribute. Utilizing the insights gained through MONET and MA-MONET, AI model developers can refine data collection and processing and also improve optimization techniques, consequently fostering the development of more reliable and trustworthy medical AI models.

In medicine, inherently interpretable models could enable physicians or developers to decipher the factors influencing a model's decision. CBMs are one such inherently interpretable model but have been limited because they require a priori concept labels, available in only a handful of medical datasets. MONET overcomes this issue with automatic concept labeling, allowing the creation of CBMs that were previously impossible. In our study, we showcase CBMs for malignancy and melanoma detection to demonstrate what is possible; however, these models are not meant for diagnostic use without additional rigorous testing across larger datasets.

We envision that MONET will impact the clinical deployment and monitoring of medical image AI systems by enabling more thorough auditing[63–65]. For instance, MONET or derivative systems could be integrated to help monitor for shifts in input data that may affect the performance of clinically deployed models[66–68]. We anticipate that MONET will facilitate active involvement of physicians in the development and oversight of medical AI systems. Their expertise will be crucial in guiding MONET's applications and contextualizing its outputs.

In addition, MONET could be used to analyze the fairness of medical AI systems. We found that MONET could annotate skin-tone labels and this could be used for auditing the fairness of medical datasets, for instance by identifying concepts that are differentially expressed across skin tones. In our view, the performance necessary for real-world use of these labels is unclear and we caution users to evaluate MONET's suitability on a case-by-case basis.

As with all methods, MONET has limitations. Although MONET successfully annotated a variety of concepts in our testing, these tests do not guarantee the quality of all annotations and, in particular, we expect that MONET will struggle to score concepts absent in its training data. Although our use of a large, diverse body of training data mitigates this concern, users should assess performance in their specific use cases. Moreover, MONET was developed specifically for medical concept annotation and is not meant to be used for diagnostic tasks. As MONET is trained on data that may exhibit biases, MONET too may exhibit biases. Although we verified that MONET performs well across skin tones when testing on clinical images, we were unable to examine performance across skin tones in dermoscopic images as a result of the unavailability of diverse dermoscopic datasets[27]. In addition, MONET could also demonstrate biases unrelated to skin tone.

In summary, our study demonstrates that AI transparency and trustworthiness at scale are feasible in a way that was previously impossible: through image–text models tailored to the medical domain of interest.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-024-02887-x.

References

1. 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. In *Advances in Neural Information Processing Systems* (eds Koyejo, S. et al.) 18157–18167 (Curran Associates, Inc., 2022).
2. Mendonça, T., Ferreira, P. M., Marques, J. S., Marcal, A. R. & Rozeira, J. PH 2-A dermoscopic image database for research and benchmarking. In *35th Annual International Conference of the IEEE* 5437–5440 (Engineering in Medicine and Biology Society, 2013).

3. 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).
4. Nevitt, M., Felson, D. & Lester, G. *The Osteoarthritis Initiative. Protocol for the cohort study* V 1.1 6.21.06 (accessed 1 Nov 2023); https://nda.nih.gov/static/docs/StudyDesignProtocolAndAppendices.pdf
5. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. 38th International Conference on Machine Learning* (eds Meila, M. & Zhang, T.) 8748–8763 (PMLR, 2021).
6. Groh, M. et al. Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 1820–1828 (IEEE, 2021).
7. Daneshjou, R. et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Sci. Adv.* **8**, eabq6147 (2022).
8. Gutman, D. et al. 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). Preprint at https://arxiv.org/abs/1605.01397 (2016).
9. Codella, N. C. F. 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 *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* 168–172 (ISBI, 2018).
10. Codella, N. et al. Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the International Skin Imaging Collaboration (ISIC). Preprint at https://arxiv.org/abs/1902.03368 (2019).
11. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* https://doi.org/10.1038/sdata.2018.161 (2018).
12. Combalia, M. et al. BCN20000: dermoscopic lesions in the wild. Preprint at https://arxiv.org/abs/1908.02288 (2019).
13. Rotemberg, V. et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Sci. Data* https://doi.org/10.1038/s41597-021-00815-z (2021).
14. Memorial Sloan Kettering Cancer Center. Consecutive biopsies for melanoma across year 2020. *ISIC Archive* https://doi.org/10.34970/151324 (2022).
15. Marchetti, M. A. et al. Prospective validation of dermoscopy-based open-source artificial intelligence for melanoma diagnosis (PROVE-AI study). *npj Digit. Med.* **6**, 127 (2023).
16. Ricci Lara, M. A. 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).
17. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016).
18. Tiu, E. et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat. Biomed. Eng*. https://doi.org/10.1038/s41551-022-00936-9 (2022).
19. Agbai, O. N. et al. Skin cancer and photoprotection in people of color: a review and recommendations for physicians and the public. *J. Am. Acad. Dermatol.* **70**, 748–762 (2014).
20. Sierro, T. J. et al. Differences in health care resource utilization and costs for keratinocyte carcinoma among racioethnic groups: a population-based study. *J. Am. Acad. Dermatol.* **86**, 373–378 (2022).
21. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **3**, 610–619 (2021).
22. Janizek, J. D., Erion, G., DeGrave, A. J. & Lee, S.-I. An adversarial approach for the robust classification of pneumonia from chest radiographs. In *Proc. ACM Conference on Health, Inference, and Learning* (ed. Ghassemi, M.) 69–79 (Association for Computing Machinery, 2020).
23. Bissoto, A., Fornaciali, M., Valle, E. & Avila, S. (De) Constructing bias on skin lesion datasets. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 2766–2774 (IEEE, 2019).
24. Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J. & Yap, M. H. Analysis of the ISIC image datasets: usage, benchmarks and recommendations. *Med. Image Anal.* **75**, 102305 (2022).
25. Winkler, J. K. et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol.* **155**, 1135–1141 (2019).
26. Navarrete-Dechent, C., Liopyris, K. & Marchetti, M. A. Multiclass artificial intelligence in dermatology: progress but still room for improvement. *J. Invest. Dermatol.* **141**, 1325–1328 (2021).
27. 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**, 1362–1369 (2021).
28. Massie, J. P. et al. Patient representation in medical literature: are we appropriately depicting diversity? *Plast. Reconstr. Surg. Glob. Open* **7**, e2563 (2019).
29. Massie, J. P. et al. A picture of modern medicine: race and visual representation in medical literature. *J. Natl Med. Assoc.* **113**, 88–94 (2021).
30. Lester, J., Jia, J., Zhang, L., Okoye, G. & Linos, E. Absence of images of skin of colour in publications of COVID19 skin manifestations. *Br. J. Dermatol.* **183**, 593–595 (2020).
31. Louie, P. & Wilkes, R. Representations of race and skin tone in medical textbook imagery. *Soc. Sci. Med.* **202**, 38–42 (2018).
32. Groh, M., Harris, C., Daneshjou, R., Badri, O. & Koochek, A. Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. In *Proc. ACM Human Computer Interactions* 1–26 (Association for Computing Machinery, 2022).
33. Rajpurkar, P. et al. MURA: large dataset for abnormality detection in musculoskeletal radiographs. In *1st Conference on Medical Imaging with Deep Learning* (MIDL, 2018).
34. Singh, C., Balakrishnan, G. & Perona, P. Matched sample selection with GANs for mitigating attribute confounding. Preprint at https://arxiv.org/abs/2103.13455 (2021).
35. Leming, M., Das, S. & Im, H. Construction of a confounder-free clinical MRI dataset in the Mass General Brigham system for classification of Alzheimer's disease. *Artif. Intell. Med.* **129**, 102309 (2022).
36. Zhao, Q., Adeli, E. & Pohl, K. M. Training confounder-free deep learning models for medical applications. *Nat. Commun.* **11**, 6010 (2020).
37. Goel, K., Gu, A., Li, Y. & Ré, C. Model patching: closing the subgroup performance gap with data augmentation. In *9th International Conference on Learning Representations* (ICLR, 2021); https://openreview.net/forum?id=9YlaeLfuhJF
38. Sagawa, S., Koh, P. W., Hashimoto, T. B. & Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations* (ICLR, 2020); https://dblp.org/rec/conf/iclr/SagawaKHL20.html?view=bibtex

39. Oakden-Rayner, L., Dunnmon, J., Carneiro, G. & Re, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proc. ACM Conference on Health, Inference, and Learning ACM CHIL '20* (ed. Ghassemi, M.) 151–159 (ACM, 2020).
40. Zhu, J., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017* 2242–2251 (IEEE Computer Society, 2017).
41. Jones, O. T. et al. Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review. *Lancet Digit. Health* **4**, e466–e476 (2022).
42. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proc. 31st International Conference on Neural Information Processing Systems* (eds Guyon, I. et al.) 4768–4777 (Curran Associates Inc., 2017).
43. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *Proc. 34th International Conference on Machine Learning* (eds Precup, D. & Teh, Y. W.) 3319–3328 (PMLR, 2017).
44. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)* 618–626 (IEEE, 2017).
45. Kim, B. et al. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In *Proc. 35th International Conference on Machine Learning* (eds Dy, J. G. & Krause, A.) 2668–2677 (PMLR, 2018).
46. Crabbé, J. & van der Schaar, M. Concept activation regions: a generalized framework for concept-based explanations. In *Advances in Neural Information Processing Systems* (eds Koyejo, S. et al.) 2590–2607 (Curran Associates Inc., 2022).
47. Abid, A., Yuksekgonul, M. & Zou, J. Meaningfully debugging model mistakes using conceptual counterfactual explanations. In *Proc. 39th International Conference on Machine Learning* (eds Chaudhuri, K. et al.) 66–68 (PMLR, 2022).
48. Eyuboglu, S, et al. Domino: discovering systematic errors with cross-modal embeddings. In *10th International Conference on Learning Representations, ICLR 2022* (ICLR, 2022); https://openreview.net/forum?id=FPCMqjI0jXN
49. Chung, Y., Kraska, T., Polyzotis, N., Tae, K. & Whang, S. Automated data slicing for model validation: a big data–AI integration approach. In *IEEE Transactions on Knowledge and Data Engineering* 2284–2296 (IEEE, 2020).
50. DeGrave, A. J., Cai, Z. R., Janizek, J. D., Daneshjou, R. & Lee, S.-I. Auditing the inference processes of medical-image classifiers by leveraging generative AI and the expertise of physicians. *Nat. Biomed. Eng.* https://doi.org/10.1038/s41551-023-01160-9 (2023).
51. Reyes, M. et al. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiol. Artific. Intell.* **2**, e190043 (2020).
52. Arun, N. et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiol. Artific. Intell.* **3**, e200267 (2021).
53. Han, S. S. et al. The degradation of performance of a state-of-the-art skin image classifier when applied to patient-driven internet search. *Sci. Rep.* **12**, 16260 (2022).
54. Navarrete-Dechent, C. et al. Automated dermatological diagnosis: hype or reality? *J. Invest. Dermatol.* **138**, 2277–2279 (2018).
55. Koh, P. W. et al. Concept bottleneck models. In *Proc. 37th International Conference on Machine Learning International Conference on Machine Learning* 5338–5348 (PMLR, 2020).
56. Yuksekgonul, M., Wang, M. & Zou, J. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations* (ICLR, 2023); https://dblp.org/rec/conf/iclr/YuksekgonulW023.html?view=bibtex
57. Rigel, D. S., Friedman, R. J., Kopf, A. W. & Polsky, D. ABCDE—an evolving concept in the early detection of melanoma. *Arch. Dermatol.* **141**, 1032–1034 (2005).
58. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual–language foundation model for pathology image analysis using medical Twitter. *Nat. Med.* **29**, 2307–2316 (2023).
59. Wang, Z., Wu, Z., Agarwal, D. & Sun, J. MedCLIP: contrastive learning from unpaired medical images and text. In *Proc. 2022 Conference on Empirical Methods in Natural Language Processing* (eds Goldberg, Y. et al.) 3876–3887 (Association for Computational Linguistics, 2022).
60. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
61. Jiang, L. Y. et al. Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362 (2023).
62. Combalia, M. et al. Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge. *Lancet Digit. Health* **4**, e330–e339 (2022).
63. Corbin, C. K. et al. DEPLOYR: a technical framework for deploying custom real-time machine learning models into the electronic medical record. *J. Am. Med. Inform. Assoc.* **30**, 1532–1542 (2023).
64. Coalition for Health AI. *Blueprint for Tustworthy AI Implementation Guidance and Assurance for Healthcare* (MITRE Corporation, 2023); https://tinyurl.com/CHAI-paper
65. Bedoya, A. D. et al. A framework for the oversight and local deployment of safe and high-quality prediction models. *J. Am. Med. Inform. Assoc.* **29**, 1631–1636 (2022).
66. Pianykh, O. S. et al. Continuous learning AI in radiology: implementation principles and early applications. *Radiology* **297**, 6–14 (2020).
67. Feng, J. et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. *npj Digit. Med.* **5**, 66 (2022).
68. Vokinger, K. N., Feuerriegel, S. & Kesselheim, A. S. Continual learning in medical devices: FDA's action plan and beyond. *Lancet Digit. Health* **3**, e337–e338 (2021).

Methods

Dataset

Overview. We trained MONET on 105,550 pairs of images and texts collected from PubMed articles and medical textbooks[69,70]. We evaluated MONET using both clinical and dermoscopic images, two widely used dermatological image types. Dermoscopic images are captured using digital photography with a specialized dermatological instrument called a dermatoscope. This magnifies skin lesions to capture fine details, whereas clinical images are often taken at least 6 cm away with a digital camera. For evaluation, we used clinical images from the Fitzpatrick 17k[6] and DDI[7] datasets and dermoscopic images from the Derm7pt[3] and ISIC[8–16] datasets.

PubMed. The PMC Open Access Subset is a dataset with millions of scientific articles released by PubMed Central (PMC)[69]. First, to find dermatology articles in the dataset, we queried papers in the PMC using dermatology-related terms (that is, dermatology, melanoma, skin cancer), 114 disease labels in the Fizpatrick17k dataset and 48 concept labels in SkinCon[1]. We downloaded 496,510 articles found via this query. In total, the articles contained 3,172,490 figures. Next, we filtered out non-dermatology-related figures (for example, graphs, illustrations, diagrams, slide images and X-ray images). To this end, we repeated the process of running a clustering algorithm on the images and manually excluding groups of non-dermatology ones (see Supplementary Methods for full details). Postfiltering, 50,265 images remained. Finally, we paired the figure captions with their corresponding images based on the provided XML-formatted files. This file stores the article's structure with components such as abstract, sections, figures and figure legends tagged.

Textbook. We first extracted images from 55 medical textbooks, yielding a total of 104,223 images. After undergoing the same filtering procedure as we did for PubMed images, 55,285 images remained. The PDF format of the textbooks made matching images with associated text difficult, because PDFs lack the structure information provided by XML formats. To address this issue, we implemented the following procedure. We used PyMuPDF (v.1.21.1), an open-source PDF rendering software, to parse the PDF files, extracting text and image objects along with their respective coordinates. Then, we assigned text to images appropriately based on the handcrafted rules described in Supplementary Methods. Finally, by combining these two refined datasets from PubMed articles and medical textbooks, we compiled the training set comprising 105,550 pairs of images and texts, with the quality evaluation of this training data detailed in Supplementary Methods.

Fitzpatrick 17k. As the PubMed and textbook datasets contain clinical (that is, non-dermoscopic) images, for evaluation purposes we required additional clinical images with ground-truth annotations. As the first of these datasets, we chose Fitzpatrick 17k[6], which contains dermatological images collected from online dermatology atlases accompanied by disease annotations and FST labels.

We performed multiple steps of quality control on the dataset to ensure the validity of our analysis. First, to reduce the impact of artifacts in the images, we filtered the dataset to exclude images with visible patient clothing, visible anatomy (for example, fingers, ears, eyes), or other elements except for lesions and background skin. We fine-tuned a DenseNet-121 (ref. 71) (pretrained on ImageNet) to identify images with such artifacts (Supplementary Methods). This filtering resulted in a total of 4,951 images. Also, Fitzpatrick 17k contains near duplicate images with slight differences in angle or cropping; we filtered for duplicate images to prevent overlap between the train and test sets for the CBM experiments. The filtering was done by calculating pairwise cosine similarity between the images, based on the EfficientNetV2-S model embedding (pretrained on ImageNet)[72] (Supplementary Methods). After this filtering, we had a total of 4,386 images. In addition, we excluded images overlapping with the MONET training data because the overlaps have the potential to inflate the performance of MONET. To this end, we used an approach similar to the duplicate removal process described above (Supplementary Methods); for each image in Fitzpatrick 17k, we measured the distance to every image from the training data. Through this process, we removed the identified 50 images, leaving 4,336 images. Last, we excluded 62 images that were marked as 'Do not consider this image' (that is, images of low quality or considered inappropriate) in the SkinCon dataset. This led to a final dataset of 4,274 images.

In addition, for melanoma prediction tasks, among the 113 fine-grained diagnosis labels, we further refined the data to include only melanomas and melanoma look alikes, such that the data mirror a well-defined clinical task. In line with the disease-filtering criteria outlined by DeGrave et al.[50], we included melanomas, benign melanocytic lesions, seborrheic keratoses and dermatofibromas, resulting in a total of 494 images.

DDI. As a second set of clinical images with ground-truth labels, we chose the DDI dataset. DDI contains 656 clinical images of diverse skin tones, obtained from Stanford Clinics[7], accompanied by annotations of FST and histopathologically proven diagnoses. Again, we excluded 20 images that were marked as 'Do not consider this image' in the SkinCon dataset, resulting in the final dataset of 636 images. Similar to our approach with the Fitzpatrick 17k dataset, we investigated and found no duplicate images. In addition, we checked for overlapping images with the MONET training dataset and confirmed that none existed. For melanoma prediction tasks, we narrowed down the dataset to include only melanomas and melanoma look alikes, resulting in a total of 275 images, following the approach used by DeGrave et al.[50]. The diagnosis labels used for filtering are described in Supplementary Methods.

SkinCon. SkinCon is at present the most comprehensive dataset on dermatology concepts[1]. The dataset features 48 concepts, curated by board-certified dermatologists, that are frequently used to describe skin lesion attributes such as shape, size, color and texture. The dermatologists manually annotated the ground-truth labels for these concepts on 3,230 images from the Fitzpatrick 17k dataset, which originally consisted of 16,577 images, and all 656 images from the DDI dataset. Of the 4,274 images in the Fitzpatrick 17k dataset that we obtained after filtering, 999 had SkinCon annotations. Among the 494 images in the Fizpatrick17k dataset used for the melanoma prediction task, 93 had annotations. All the images in DDI had annotations.

Derm7pt. As the SkinCon dataset had concept annotations for clinical images only, we needed an additional concept dataset to evaluate MONET's performance on dermoscopic images. For this purpose, we used Derm7pt, namely the seven-point dermatology checklist dataset, consisting of 1,011 dermoscopic images and 1,002 corresponding clinical images (some were missing in the dataset), all derived from the same 1,011 lesion cases[3]. The dataset provides annotations for seven dermatological concepts: 'pigment network', 'blue-whitish veil', 'vascular structure', 'pigmentation', 'streaks', 'dots and globules' and 'regression structures'. For quantitative evaluations, we converted concepts with multi-class annotations into binary labels (that is, present and absent). For our analysis, we utilized the 1,011 dermoscopic images from Derm7pt. We restricted our analysis to dermoscopic images because those concepts are primarily dermoscopic features. Although Derm7pt covered fewer concepts than SkinCon, it served as a valuable resource for assessing MONET's concept annotation capability on dermoscopic images. As we did for the Fitzpatrick 17k dataset, we thoroughly checked the images and found that there are no duplicates or overlaps with the MONET training data. The dataset also contained labels for 14 diseases, which we used for evaluating MONET's disease label annotation performance.

ISIC. The ISIC archive is a repository of digital skin images, primarily consisting of dermoscopic images, sourced from various institutions. ISIC represents the largest and the most commonly used dataset for the development of dermatology AI models[41]. We used this dataset to demonstrate that MONET facilitates transparency analyses for large-scale datasets. We downloaded 71,670 images in total from all of the ISIC collections, including ISIC Challenge datasets 2016, 2017, 2018, 2019 and 2020 (refs. 8–16). The images have diagnostic attributes, including binary malignancy versus benign labels, and 27 granular disease labels. We first excluded 36 images labeled as 'overview', which depicted the whole torso. In addition, as we did for the Fitzpatrick 17k dataset, we examined duplicates or overlaps with the MONET training data. We found 959 sets of duplicates and 6 overlaps. After exclusion, 70,654 images remained. For per-institution analysis, we grouped images by data sources based on the attribution column in the metadata. We selected the two largest cohorts: the Department of Dermatology at Med. U. Vienna (10,011 samples) and the Department of Dermatology at Hospital Barcelona (12,271 samples). For the analysis of All Data Are Ext (ADAE) algorithms, we used the PROVE-AI study dataset available in the ISIC archive. It had 603 images, consisting of 95 melanomas and 508 non-melanomas[15]. There were no duplicates or overlaps with the MONET training data.

MONET

Formally, let $f_{\text{image}} : \mathcal{X}_{\text{image}} \to \mathbb{R}^d$ be the MONET image encoder and $f_{\text{text}} : \mathcal{X}_{\text{text}} \to \mathbb{R}^d$ be the MONET text encoder. Given a dataset of paired images $I \in X_{\text{image}}$ and text descriptions $T \in \mathcal{X}_{\text{text}}$, $D_{\text{paired}} = (I_i, T_i)_{i=1}^{n_{\text{paired}}}$, our goal is to train the two encoders such that the distances between pairs of embeddings, $\text{dist}\left(f_{\text{image}}(I_i), f_{\text{text}}(T_j)\right)$, reflect the semantic similarity between I_i and T_j for all $i,j \le n_{\text{paired}}$.

Architecture. We used the vision transformer architecture, ViT-L/14, as our image encoder[73]. This encoder took an input image of size 224 × 224 and output a 768-dimensional embedding. For the text encoder, we used a transformer architecture with 12 self-attention layers. It took tokenized text with a maximum limit of 77 tokens as input and output a 768-dimensional embedding. We used the same architecture as CLIP[5] to take advantage of the weights from pretrained models.

Preprocessing. To meet the input requirements for the encoder architectures, we processed image and text inputs as follows. Each input image was resized and center cropped to dimensions 224 × 224. It was then normalized using the mean and s.d. used in ImageNet[74]. Throughout the training phase, we applied standard data augmentation steps instead, such as random resized crops, vertical flips, horizontal flips and color jittering for brightness, contrast and saturation. For each input text, we applied tokenization using lower-case byte pair encoding[75]. In cases where the text encountered during training was longer than the text encoder's maximum token limit of 77, we split the text into sentences and chose half of them from the beginning. We repeated this process until the token count was reduced to ≤77.

Training. We used cosine similarity as the distance metric. Both encoders were jointly trained to maximize the cosine similarity between the image and text embeddings of the correct pairs while minimizing the cosine similarity between the embeddings of incorrect pairings. To this end, we used a symmetric cross-entropy loss on cosine similarity scores; after calculating the cosine similarities between embeddings, we scaled them by a temperature parameter and normalized them into a probability distribution with the softmax function. The temperature parameter was also updated during training. We optimized the loss using the Adam optimizer[76] with a cosine learning rate schedule for ten epochs. This implementation detail followed that of CLIP[5].

For hyperparameter tuning, we split the dataset into training and validation sets and found hyperparameters that result in the best validation loss; we used validation loss for hyperparameter tuning because there is no large-scale ground-truth label for evaluating concept annotation performance. We tuned the hyperparameter of batch size (128, 256, 512, 1024) and learning rate (1×10^{-3}, 1×10^{-4}, 1×10^{-5}, 1×10^{-6}). We found that the larger batch size resulted in lower validation loss until a batch size of 512 was reached. We also found that the learning rate of 1×10^{-5} leads to the lowest validation loss. Using the tuned hyperparameters, we trained the model on the whole dataset for ten epochs. We used six Nvidia A40 graphics processing units with data parallelism. Model training took 1 h and 40 min.

Automatic concept annotation

During the training procedure of image and text encoders, an image and its corresponding text from the same pair were forced to be close to each other in the embedding space, whereas those from different pairs were forced to be far apart. After training, MONET could measure the proximity between an image and any arbitrary text. We used this capability to automatically annotate concepts for images.

To annotate a concept c for a given batch of N images $I_1, I_2, \cdots, I_N$, we first computed the image embeddings $f_{\text{image}}(I_1), f_{\text{image}}(I_2), \cdots, f_{\text{image}}(I_N)$ using the image encoder f_{image}. We also computed the concept prompt embedding, $f_{\text{text}}(T_c)$, and reference prompt embedding, $f_{\text{text}}(T_r)$, using the text encoder, where T_c is a concept prompt (for example, 'This is skin image of …') and T_r is a reference prompt (for example, 'This is skin image'). Supplementary Table 7 shows the terms used for each concept for filling templates. Next, we calculated the cosine similarity between image embeddings and prompt embeddings. When multiple terms were used for each concept, we calculated the cosine similarity for each term and averaged them. Finally, we obtained concept presence score, $p_{i,c}$ which represents the degree to which a concept is present in the image as follows:

$$p_{i,c} = \frac{\exp(\text{sim}_{\cos}(I_i, T_c)/\lambda)}{\exp(\text{sim}_{\cos}(I_i, T_c)/\lambda) + \exp(\text{sim}_{\cos}(I_i, T_r))/\lambda)} \quad (1)$$

where $\text{sim}_{\cos}(\cdot)$ is the cosine similarity score between image embeddings and text embeddings, $\text{sim}_{\cos}(I_i, T_c) = \frac{f_{\text{image}}(I_i)^T f_{\text{text}}(T_c)}{|f_{\text{text}}(I_i)||f_{\text{text}}(T_c)|}$, and λ is the temperature parameter learned during the training. We normalized by reference prompt to remove the effect of templates being used. Furthermore, we used multiple templates to minimize the effects of templates. For clinical images, we used the templates: 'This is skin image of …', 'This is dermatology image of …' and 'This is image of …'. For dermoscopic images, we used the templates 'This is dermatoscopy of …' and 'This is dermoscopy of …'. We used concept presence scores averaged across different templates in the end. We describe the quantitative evaluation of concept differential analysis against baseline methods in Supplementary Methods.

Data auditing

MONET's ability to map images and texts on to the co-embedding space enabled us to describe the different characteristics between two sets of images in natural language (that is, concept differential analysis). Assuming that we have two sets of images, denoted as $\mathbf{I}_+ = \{I_1, I_2, \ldots, I_{N^+}\}$ and $\mathbf{I}_- = \{I_1, I_2, \ldots, I_{N^-}\}$, and a list of concepts we wanted to investigate $[c_1, c_2, \ldots, c_{N_c}]$, we first obtained the prototype embedding of each image set by computing an average of normalized image embeddings: $m_+ = \frac{1}{N^+}\sum_{I_i \in \{\mathbf{I}_+\}} \frac{f_{\text{image}}(I_i)}{\|f_{\text{image}}(I_i)\|}$ and $m_- = \frac{1}{N^-}\sum_{I_i \in \{\mathbf{I}_-\}} \frac{f_{\text{image}}(I_i)}{\|f_{\text{image}}(I_i)\|}$. Then, we calculated the displacement vector from m_- to m_+ by subtracting out the two prototype embedding $m_\Delta = m_+ - m_-$. Finally, we got a differential concept expression score by computing the dot product between the displacement vector and normalized embeddings of concept prompt: $C_{\Delta,i} = m_\Delta^T \cdot \frac{f_{\text{text}}(T_i)}{\|f_{\text{text}}(T_i)\|}$. This score measures how much more each concept is differentially expressed in $\mathbf{I}_+$ than in $\mathbf{I}_-$. A similar technique for converting a set of images to text has been previously used by

Eyuboglu et al.[48]. We describe the benchmarking analysis of concept differential analysis in Supplementary Methods.

Model auditing

We can use MONET to automatically detect semantically meaningful medical concepts that lead to model errors. MA-MONET starts by sorting images from a test set into groups based on their visual similarity. To this end, we ran the K-means clustering algorithm implemented in the scikit-learn Python package (v.1.2.2)[77,78]. We used 50 principal components of the embedding of the penultimate layer of the EfficientNetV2-S model (pretrained on ImageNet)[72] as clustering features. Next, we calculated the accuracy across all samples and also per cluster; for thresholding the probability output from the trained classifier, we chose an operating point that maximizes the $F1$ score. After this, we identified medical concepts for the 'low-performing' cluster; we defined low-performing clusters as those with accuracy lower than overall accuracy. Each low-performing cluster was compared with a high-performing counterpart containing similar images to understand what differentiates them; among the clusters that performed better than the overall accuracy, we chose the cluster with a centroid closest in Euclidean distance to the low-performing cluster. We compared the mean concept presence scores of each concept between the high- and low-performing clusters to pinpoint concepts that are more present in the low-performing cluster. If two visually similar clusters (one high performing, the other low performing) differ in terms of a few concepts, these differing concept terms can be hypothesized as leading to high error rates. We then filtered out concepts with a concept presence score <0.5 in the low-performing cluster. Finally, we obtained a ranked list of medical concepts identified by MONET that differentiate the two clusters. We describe the benchmarking analysis of MA-MONET in Supplementary Methods.

Building inherently interpretable neural network

Concept bottleneck models. CBMs[55] are designed to be inherently interpretable, which can be applied to a wide range of tasks, including general vision and natural language-processing tasks[79–81] and medical applications[1,82,83]. However, a caveat is that these models need a large set of concept annotations to perform well, which are expensive and time-consuming to obtain.

MONET allows us to automatically annotate concepts for images, which enables scaling to a large concept dataset with an arbitrary number of concepts. Specifically, MONET helps to create the bottleneck layer, denoted by $b_c : \mathcal{X}_{\text{image}} \to \mathbb{R}^{N_c}$, that maps an input image I_i to a vector of dimension N_c, the number of concepts. An interpretable linear model is then trained on the prediction target to get importance scores for each concept corresponding to the learned weights.

To create the bottleneck layer, we started with a concept list $[c_1, c_2, \ldots, c_{N_c}]$, chosen with guidance from our dermatologist collaborators, containing concepts that are predictive of the target. Ideally, the bottleneck layer is binarized using the concept labels. However, we lacked access to the concept annotations and thresholding the similarity score of each concept with the input image is non-trivial. Instead, for each concept c_j, we curated a set of reference concepts $[r_{j1}, r_{j2}, \ldots, r_{jN_j}]$ where N_j is the number of reference concepts for concept c_j. Each reference concept is selected such that it is sufficiently far from the concept of interest in the representation space while being closer to the other reference concepts. We do this by choosing antonyms of the concept of interest as the reference concepts.

Once the set of reference concepts has been created, MONET calculates the similarity scores of the input image to the concept of interest and the corresponding reference concepts. The former is then normalized by taking a softmax with the reference concept scores. The resulting normalized score, p'_{i,c_j}, is then used in the bottleneck node for that concept, as shown in equation 2:

$$p'_{i,c_j} = \frac{\exp\left(\text{sim}_{\cos}\left(I_i, c_j\right)/\lambda\right)}{\exp\left(\text{sim}_{\cos}\left(I_i, c_j\right)/\lambda\right) + \sum_k \exp\left(\text{sim}_{\cos}\left(I_i, r_{jk}\right)/\lambda\right)} \quad (2)$$

where $\text{sim}_{\cos}(\cdot)$ is the cosine similarity score obtained using MONET and λ is a temperature parameter used to magnify the differences in similarity scores. λ is manually tuned to the value that performs the best on the train set. Once the bottleneck layer has been created, we trained a simple linear classifier on the prediction target using stochastic gradient descent. Specifically, for a learned weight w and a normalized score $p'_i \in \mathbb{R}^{N_c}$, the prediction obtained is $w^T p'_i + b$. We applied L1 regularization to favor sparsity in the trained model weights and make the model more interpretable. Once the linear classifier had been trained, the learned weights w could be analyzed to understand the importance of each concept for the prediction target.

Evaluation setting. To demonstrate the efficacy of MONET, we used two different prediction targets: (1) malignant versus benign and (2) melanoma versus melanoma look alikes. We differentiated between these two targets because all melanomas are malignant, but not all malignant lesions are melanoma.

We conducted separate experiments on two types of images: clinical images and dermoscopic images. For the clinical image experiment, we used images ($n = 4{,}910$) from the clean Fitzpatrick 17k[6] and DDI[7] datasets, which include 1,129 malignant and 3,781 benign images. For melanoma prediction, we further filtered images to ensure that data mirrored a well-defined clinical task, resulting in 769 images, consisting of 321 melanomas and 448 melanoma look alikes. We used 80% of the data for training and reserved the rest for testing. Thus, for malignancy, we used 3,928 samples for training and 982 samples for testing; for melanoma, we used 615 samples for training and 153 samples for testing. For each setting, we repeated evaluations with 20 different train–test splits. We noted that, for the comparison of MONET + CBM with the manual labeling depicted in Fig. 5b–e, the number of training samples used for the manual labeling approach is smaller because it relies on the images labeled in SkinCon (Supplementary Methods).

For dermoscopic images, we employed dermoscopic images ($n = 70{,}515$) from the ISIC dataset, with 9,999 classified as malignant and 60,516 as benign. For the melanoma prediction task, we excluded images lacking disease-specific labels, narrowing our analysis to 43,678 images consisting of 5,900 melanomas and 37,778 identified as non-melanomas. Mirroring the manual labeling comparison, we performed evaluations using 20 separate train–test splits, maintaining an 80:20 ratio; in this case, test set images are used for evaluation because none of the methods required ground-truth concept labels. Thus, for malignancy, we used 56,412 samples for training and 14,103 samples for testing; for melanoma, we used 34,942 samples for training and 8,735 samples for testing.

To create the bottleneck layer, we used ten concepts that are known to be predictive of targets (Supplementary Table 8); specifically, we use the ABCDEs of melanoma[57] as a guideline to compile the list of concepts. We compared MONET + CBM with several other baseline methods of obtaining target predictions from input images. The baseline methods are described in Supplementary Methods.

Statistics. From 20 runs with different train–test sets, we obtained 20 pairs of AUROC values. To compare the performance of the MONET + CBM method against other methods, we conducted paired sample Student's t-tests on these AUROC values. Specifically, we used one-sided tests, where the alternative hypothesis is that MONET + CBM's AUROC is higher than from the other method. As the number of pairs is 20, the degree of freedom for these tests is 19.

Assessing a clinical trial of a dermatology AI algorithm. We replicated the PROVE-AI clinical trial, a recent prospective clinical study that evaluated the ADAE algorithm, an open-source, noncommercial model for melanoma diagnosis[15]. In the original study, the ADAE model

was thresholded to achieve a high 95% sensitivity on retrospective data, before being evaluated in a prospective observational clinical study. To replicate this study, we obtained the model and threshold used for the sensitivity from the authors. We then ran this model in test mode on the exact set of prospective images retrieved from the ISIC archive[15]. We evaluated the model on 603 dermoscopic images (95 melanomas and 508 non-melanomas), calculating metrics such as AUROC, sensitivity, specificity and accuracy. We compared these metrics with the values reported by the authors and found them to be almost identical. Subsequently, we performed subgroup analyses for the ADAE algorithm to identify concepts associated with lower specificity. For each concept, we calculated the specificity of the top 100 images among benign images, ranked by MONET, and compared it with that of the remaining benign images. We obtained *P* values via a one-sided Fisher's exact test, comparing the odds of false positives with true negatives between the two groups. The alternative hypothesis is that the odds ratio is >1, indicating a higher rate of false positives in the top 100 images than in the remaining images. We adjusted the *P* values for multiple comparisons using Bonferroni correction, accounting for the 62 concepts tested.

Reporting summary

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

Data availability

The PMC Open Access Subset is publicly available from https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist. Evaluation datasets are all publicly available and can be accessed from: ISIC (https://challenge.isic-archive.com/data), Derm7pt (https://derm.cs.sfu.ca), Fitzpatrick 17k (https://github.com/mattgroh/fitzpatrick17k) and DDI (https://stanfordaimi.azurewebsites.net/datasets/35866158-8196-48d8-87bf-50dca81df965).

Code availability

The code used in our analysis is available at https://github.com/suinleelab/MONET (ref. 84). It includes various scripts for data collection and preprocessing, training the MONET model and conducting benchmark studies. Also, it provides the MONET model weights. The ADAE algorithm can be publicly accessed from https://github.com/ISIC-Research/ADAE.

References

69. PMC Open Access Subset. *National Library of Medicine* www.ncbi.nlm.nih.gov/pmc/tools/openftlist (2022).
70. Gamper, J. & Rajpoot, N. M. Multiple instance captioning: learning representations from histopathology textbooks and articles. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021* 16549–16559 (Computer Vision Foundation/IEEE, 2021).
71. Huang, G., Liu, Z., Maaten, L. V. D. & Weinberger, K. Q. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2261–2269 (IEEE Computer Society, 2017).
72. Tan, M. & Le, Q. V. EfficientNetV2: smaller models and faster training. In *Proc. 38th International Conference on Machine Learning, ICML 2021* (eds Meila, M. & Zhang, T.) 10096–10106 (PMLR, 2021).
73. Dosovitskiy, A. et al. An image is worth 16×16 words: transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021* (ICLR, 2021); https://openreview.net/forum?id=YicbFdNTTy
74. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)* 248–255 (IEEE Computer Society, 2009).
75. Sennrich, R., Haddow, B. & Birch, A. Neural machine translation of rare words with subword units. In *Proc. 54th Annual Meeting of the Association for Computational Linguistics, Vol. 1: Long Papers* 1715–1725 (Association for Computational Linguistics, 2016).
76. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015* (eds Bengio, Y. & LeCun, Y.) (ICLR, 2015).
77. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Machine Learn. Res.* **12**, 2825–2830 (2011).
78. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Informa. Theory* **28**, 129–137 (1982).
79. Lanchantin, J., Wang, T., Ordonez, V. & Qi, Y. General multi-label image classification with transformers. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 16478–16488 (IEEE, 2021).
80. Jeyakumar, J. V. et al. Automatic concept extraction for concept bottleneck-based video classification. Preprint at https://arxiv.org/abs/2206.10129 (2022).
81. Sun, X. et al. Interpreting deep learning models in natural language processing: a review. Preprint at https://arxiv.org/abs/2110.10470 (2021).
82. Klimiene, U. et al. Multiview concept bottleneck models applied to diagnosing pediatric appendicitis. In *2nd Workshop on Interpretable Machine Learning in Healthcare* (IMLH, 2022).
83. Wu, C., Parbhoo, S., Havasi, M. & Doshi-Velez, F. Learning optimal summaries of clinical time-series with concept bottleneck models. In *Proc. 7th Machine Learning for Healthcare Conference* (eds Lipton, Z. C. et al.) 648–672 (PMLR, 2022).
84. suinleelab/MONET. *Github* https://github.com/suinleelab/MONET (2024).

Acknowledgements

We thank C. Lin and other people in S.-I.L.'s lab for helpful discussions. The members of S.-I.L.'s lab, including C.K., S.U.G., A.J.D. and S.-I.L., received support from the National Science Foundation (grant nos. CAREER DBI-1552309 and DBI-1759487) and the National Institutes of Health (NIH, grant nos. R35 GM 128638 and R01 AG061132). R.D. was supported by the NIH (5T32 AR007422-38) and the Stanford Catalyst Program.

Author contributions

C.K., R.D. and S.-I.L. conceived the initial study. C.K., S.U.G., A.J.D. and J.A.O. performed the experiments. Z.R.C. and R.D. evaluated the training data, provided dermatological insights and clinical context in all steps of the analyses, and analyzed images from concept retrieval experiments. C.K., S.U.G., A.J.D., J.A.O., Z.R.C., R.D. and S.-I.L. wrote the paper. S.-I.L. secured funding. R.D. and S.-I.L. co-supervised the study.

Competing interests

R.D. reports fees from L'Oreal, Frazier Healthcare Partners, Pfizer, DWA and VisualDx for consulting, stock options from MDAcne and Revea for advisory board and research funding from UCB. The remaining authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-024-02887-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-024-02887-x.

Correspondence and requests for materials should be addressed to Roxana Daneshjou or Su-In Lee.

Peer review information *Nature Medicine* thanks Titus Brinker and Ben Glocker for their contribution to the peer review of this work. Primary Handling Editor: Lorenzo Righetto, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-024-02887-x

Extended Data Fig. 1 | Dermoscopic images with artifacts from the ISIC dataset as determined by high concept presence scores calculated using MONET. We show the top 30 images for each artifact. **a**, Purple pen. **b**, Orange sticker. **c**, Nail. **d**, Hair. **e**, Dermoscopic border. Figures adapted from ref. 11,12 with permission and from ref. 13,14 under CC-BY license.

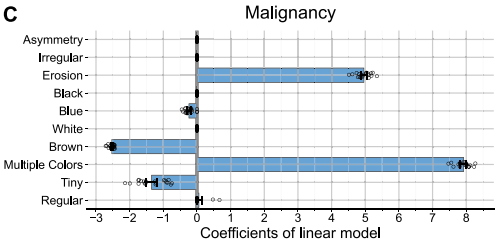
https://doi.org/10.1038/s41591-024-02887-x

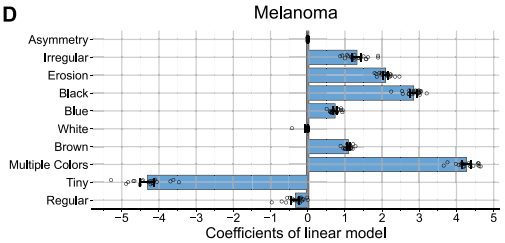
A Malignancy

Methods	Validation AUROC	Test AUROC
MONET+CBM	0.770	**0.760**
CLIP+CBM	0.749	0.700
Supervised (ResNet-50)	0.832	0.729
Linear probing (ResNet-50)	0.814	0.720

B Melanoma

Methods	Validation AUROC	Test AUROC
MONET+CBM	0.774	**0.761**
CLIP+CBM	0.752	0.662
Supervised (ResNet-50)	0.834	0.723
Linear probing (ResNet-50)	0.843	0.742



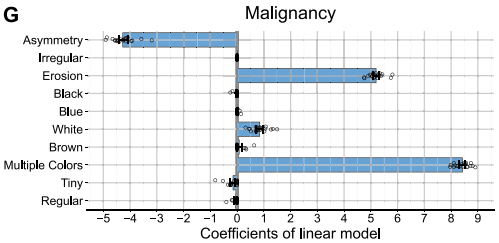


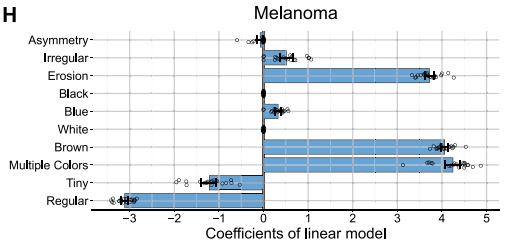
E Malignancy

Methods	Validation AUROC	Test AUROC
MONET+CBM	0.787	**0.761**
CLIP+CBM	0.726	0.717
Supervised (ResNet-50)	0.864	0.691
Linear probing (ResNet-50)	0.874	0.650

F Melanoma

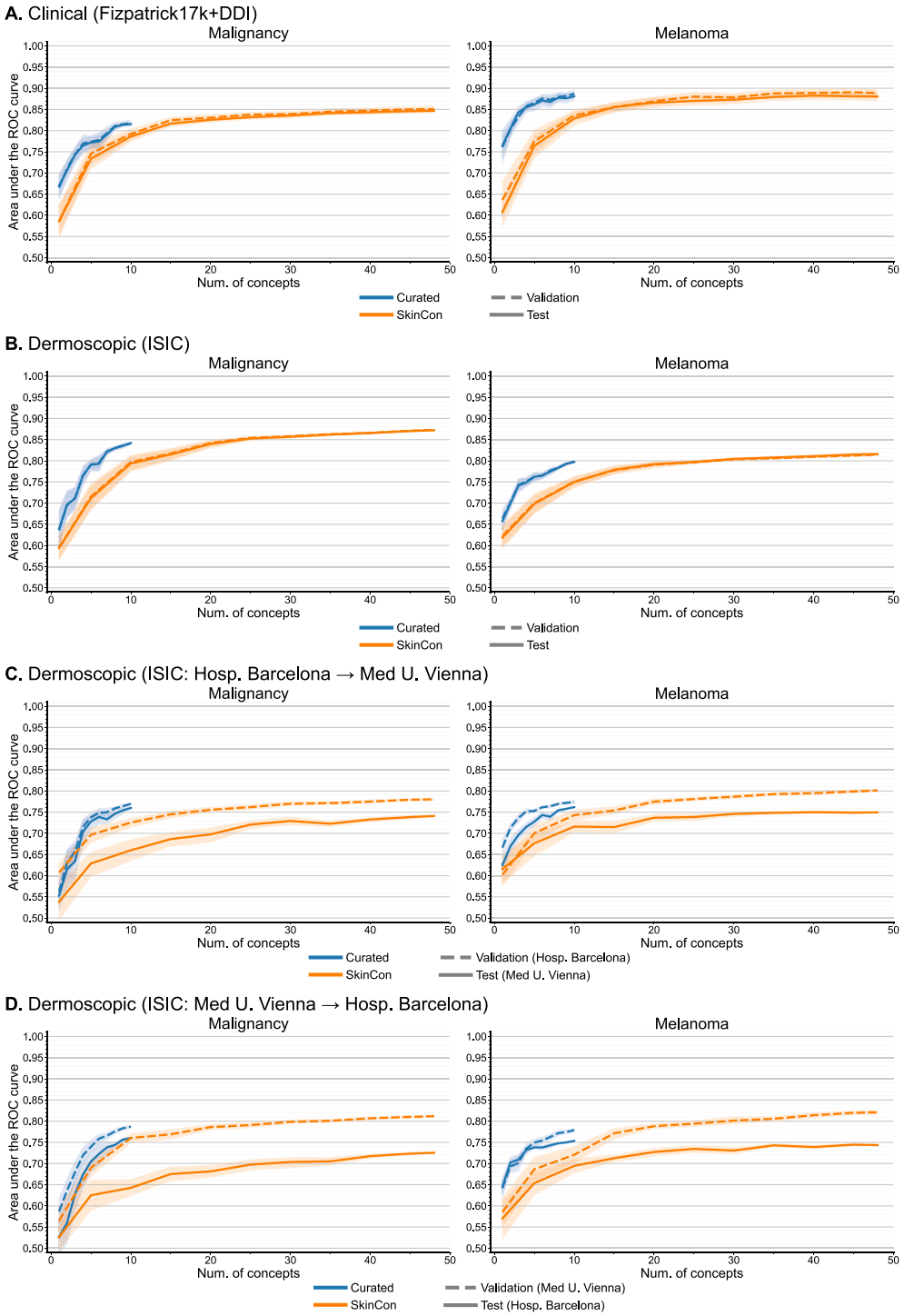
Methods	Validation AUROC	Test AUROC
MONET+CBM	0.780	**0.754**
CLIP+CBM	0.726	0.717
Supervised (ResNet-50)	0.881	0.668
Linear probing (ResNet-50)	0.862	0.680





Extended Data Fig. 2 | Concept bottleneck models in the interinstitution model transfer setting. a–**d**, We train models using images from Hospital Barcelona and test their performance using images from Med. U. Vienna. In each setting, we repeated the evaluations 20 times, using a different split of images from the training site into train and validation sets each time. **a**–**b**, Performance comparison of malignancy and melanoma prediction models. We measure the validation AUROC on a 25% validation set from Hospital Barcelona data. We measure the test AUROC on the entire Med. U. Vienna data. **c**–**d**, Coefficient of the linear model in MONET + CBM for malignancy and melanoma predictions. The error bars, obtained from $n = 20$ different runs, indicate the 95% confidence interval, extending from the mean. **e**–**h**, We train models using images from Med. U. Vienna and test their performance using images from Hospital Barcelona. **e**–**f**, Performance comparison of malignancy and melanoma prediction models. We measure the validation AUROC on a 25% validation set from Med. U. Vienna data. We measure the test AUROC on the entire Hospital Barcelona data. **g**–**h**, Coefficient of the linear model in MONET + CBM for malignancy and melanoma predictions. The error bars, obtained from $n = 20$ different runs, indicate the 95% confidence interval, extending from the mean.

https://doi.org/10.1038/s41591-024-02887-x



Extended Data Fig. 3 | Effect of concept sets on MONET + CBM's performance. We compare CBMs operating on our curated, task-relevant, concepts with those operating on SkinCon concepts. From each concept list, we sample subsets of concepts with varying sizes and train CBMs separately on these subsets, repeating this process 20 times for each subset size. For our 10 curated concepts, we use the subset size ranging from 1 to 10, and for SkinCon concepts, we use the subset size of interval 5, including 1 and the full set of 48 (that is, 1, 5, 10, …, 40, 45, 48). The center line indicates the mean across $n = 20$ runs, with the shaded area covering the 95% confidence interval. **a**, Performance of MONET + CBM for malignancy and melanoma predictions on clinical images with respect to the number of concepts. **b**, Performance of MONET + CBM for malignancy and melanoma predictions on dermoscopic images with respect to the number of concepts. **c**, Performance of MONET + CBM for malignancy and melanoma predictions on dermoscopic images in an interinstitution model transfer setting with respect to the number of concepts. We train models on Hospital Barcelona and test them on Med. U. Vienna. **d**, Performance of MONET + CBM for malignancy and melanoma predictions on dermoscopic images in an interinstitution model transfer setting with respect to the number of concepts. We train models on Med. U. Vienna and test them on Hospital Barcelona.

nature portfolio

Corresponding author(s): Roxana Daneshjou
Su-In Lee

Last updated by author(s): Nov 23, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The source data was collected from publicly available online sources using Python (ver. 3.9) with the following Python packages: ftplib (included with Python ver. 3.9), h5py (ver. 3.7.0), Selenium (ver. 4.7.2), and isic-cli (ver. 8.2.0).
Data analysis	We analyzed data via custom, open-source code, as described in our Code Availability statement. Our code can be accessed at https://github.com/suinleelab/MONET. The data analysis was implemented using Python (ver. 3.9). We used the following Python packages for data analysis and plotting: PyTorch (ver. 1.13.0), PyMuPDF (ver. 1.21.1), BeautifulSoup4 (ver. 4.11.1), scikit-learn (ver. 1.2.2), SciPy (ver. 1.9.3), NumPy (ver. 1.24.1), pandas (ver. 1.5.2), h5py (ver. 3.7.0), hydra (ver. 1.2.0), OmegaConf (ver. 2.3.0), Pillow (ver. 9.3.0), Matplotlib (ver. 3.6.2), and seaborn (ver. 0.12.1)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

PMC Open Access Subset is publicly available from https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/.
Evaluation datasets are all publicly available and can be accessed from: ISIC (https://challenge.isic-archive.com/data/), Derm7pt (http://derm.cs.sfu.ca/), Fitzpatrick17k (https://github.com/mattgroh/fitzpatrick17k), and DDI(https://stanfordaimi.azurewebsites.net/datasets/35866158-8196-48d8-87bf-50dca81df965). The ADAE algorithm can be publicly accessed from https://github.com/ISIC-Research/ADAE.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We determined sample sizes based on the total number of images available in each of the publicly available datasets.
Data exclusions	We excluded images that not dermatology-related from training dataset. After filtering, training dataset consisted of 105,550 pairs of image and text collected from PubMed articles and medical textbooks. For the clinical image datasets, Fitzpatrick17k and DDI dataset, we excluded images that are labeled as 'Do not consider this image' in the SkinCon dataset (i.e., images of low quality or considered not appropriate). Additionally, we filtered out duplicated images present in the Fitzpatrick17k dataset. We also excluded images overlapping with those in the training dataset. Detailed procedure is described in Methods section.
Replication	When training the MONET model, we independently retrained the model with different random seeds and confirmed each run shows consistent loss curves. Also, during benchmarking analysis conducted for the data auditing, model auditing, and concept bottleneck model, we conducted the analysis 20 times with different random seeds and reported the results. All results can be reproduced using our source code and model weights provided.
Randomization	In the benchmarking analysis of data auditing and model auditing, we randomly sampled data conditioned on a target and a concept to generate synthetic datasets with correlated prediction target labels and concept labels. We conducted the analysis using 20 different random seeds to obtain robust results. Additionally, for evaluations that involved training models, we randomly allocated samples to training and test sets in an 80%:20% ratio.
Blinding	The experiments that involved grouping were done using computer code so blinding was not required.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging