

Fostering transparent medical image AI via an image-text foundation model grounded in medical literature

Chanwoo Kim¹, Soham U. Gadgil¹, Alex J. DeGrave^{1,2}, Zhuo Ran Cai³, Roxana Daneshjou^{4,5,*}, and Su-In Lee^{1,*}

¹Paul G. Allen School of Computer Science and Engineering, University of Washington

²Medical Scientist Training Program, University of Washington

³Program for Clinical Research and Technology, Stanford University

⁴Department of Dermatology, Stanford School of Medicine

⁵Department of Biomedical Data Science, Stanford School of Medicine

* indicates co-senior authorship

Abstract

Building trustworthy and transparent image-based medical 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. Here, we present a foundation model approach, named MONET (Medical cONcept rETriever), which learns how to connect medical images with text and generates dense concept annotations to enable tasks in AI transparency from model auditing to 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 on the basis of 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, outperforming supervised models built on previously concept-annotated dermatology datasets. We demonstrate how MONET enables AI transparency across the entire AI development pipeline from dataset auditing to model auditing to building inherently interpretable models.

Introduction

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 and methods needed to promote AI transparency and to de-mystify “black-box” models often require medical datasets with dense annotations of human-understandable concepts. For example, for building a melanoma classifier, it would be medically meaningful to understand the data and model using concepts such as “darker pigmentation”, “atypical pigment networks”, and “multiple colors”. Unfortunately, obtaining such labels requires a significant amount of time from domain experts, and consequently, most medical datasets limit annotations to little more than diagnoses. In contrast, *rich annotation* with the extensive and highly descriptive clinical concepts developed by the medical community could enable numerous benefits. Such rich annotations could promote understanding of key biases in datasets, empower detection of undesirable behavior in medical AI devices, and foster the development of AI devices that better align with physicians’ expectations. However, few medical image datasets include such extensive annotations, and the time expended in existing efforts [1] argues that obtaining this data via large-scale efforts by human experts is infeasible.

Here, we instead leverage the collective knowledge of the medical community, as encapsulated in publicly available medical literature and medical textbooks, to teach an AI model, MONET (Medical cONcept rETriever), to richly annotate medical images with semantically meaningful and medically relevant concepts (Fig. 1A-B). We focus on the application of dermatology to showcase its versatility since dermatology has heterogeneity in disease appearance across diverse skin tones and has no standardized imaging practices, leading to significant heterogeneity in imaging conditions (*e.g.*, lighting, blurriness). In this setting, examples of clinical concepts include lesion color (*e.g.*, brown) and

43 morphology (*e.g.*, nodule). MONET’s automatic concept generation capability empowers us to perform meaningful
 44 trustworthiness analysis across all stages of the medical AI pipeline, as demonstrated by three use cases (Fig.1C-E).

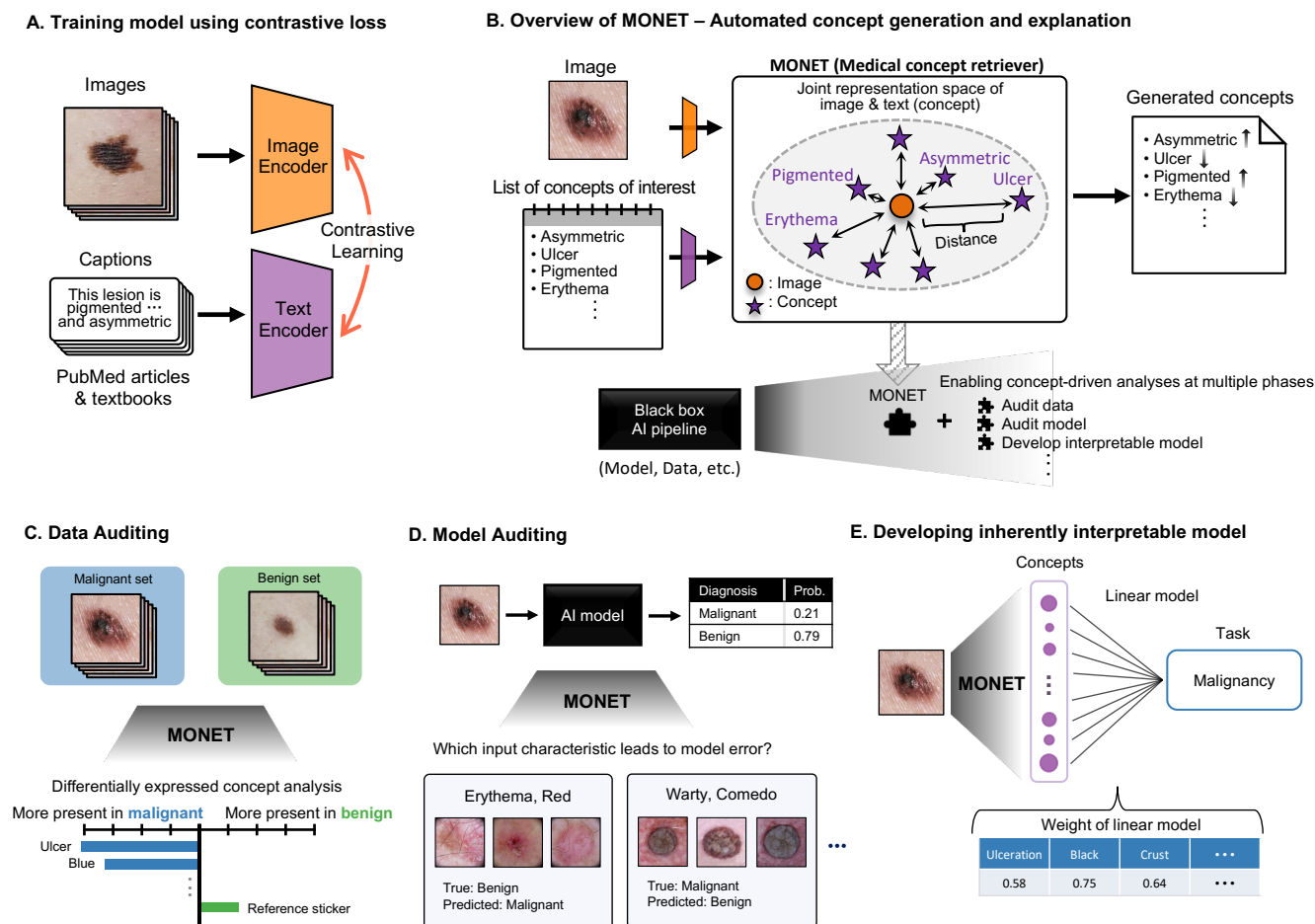


Fig. 1 | Overview of MONET framework and its usage examples. (A) *Training procedure.* MONET is trained using contrastive learning on an extensive set of dermatology image and text pairs collected from PubMed articles and medical textbooks. During the training process, the paired image and text are forced to be close in the joint representation space, while those from different pairs are forced to be far apart. (B) *Automatic concept generation.* MONET can map medical concepts and images onto a joint representation space, allowing it to determine the degree to which a concept is present in an image for any given concept by measuring the distance between the image and concept text prompts in the representation space. Its concept generation capability enables various concept-driven analyses at multiple stages of the medical AI pipeline. (C) *Concept-level data auditing.* MONET’s automatic concept generation 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. (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.

45 Dataset auditing can identify biases in the data before using it for any clinically relevant task, thereby improving
 46 the quality of the data, providing an opportunity for preliminary bias mitigation, and improving overall trustworthiness
 47 in the data. Prior work in dataset auditing has identified how particular concepts are associated, either appropriately
 48 or inappropriately, with data labels [2–4]. In medicine, data auditing has identified spurious correlations in AI training
 49 data [5, 6]. For example, the overrepresentation of chest drain in the x-rays of patients with pneumothorax led to AI
 50 algorithms that relied on their presence. However, chest tubes are a treatment used after a physician had diagnosed
 51 pneumothorax and not a causal feature [5]. Because data is not static, dataset auditing also allows the detection of
 52 dataset shifts or drifts by identifying the changes in the representation of a concept in the data [7–9]. MONET enables

53 us to examine datasets on the basis of a rich set of automatically retrieved concepts (Fig.1C).

54 Model auditing involves demystifying the “black-box” of AI models – understanding the factors involved in AI
55 decision-making [10–13]. AI models that fail during real-world deployment can lead to worse outcomes for patients.
56 AI models have been shown to make systematic errors on a subset of data with shared features, resulting in uneven
57 performance across the data [2–6, 13, 14]. To prevent this and make appropriate adjustments, models should be
58 audited to understand their failure modes prior to deployment. Model auditing is not only important immediately
59 after model development, but is a continual process, especially since models may undergo updating over time [7–9].
60 MONET powers model auditing: the dense concept annotations generated by MONET can be used to understand
61 which input characteristic leads to model errors (Fig.1D).

62 While most existing AI models are black boxes, newer methods in the field of explainable AI have attempted to
63 create inherently interpretable models that use concept-level features as input [15, 16]. The automatic generation of
64 a rich set of semantically meaningful concepts allows us to leverage and enhance these recent models (Fig.1E).

65 Because of the need to develop and test the aforementioned methods and tools, there have been prior attempts
66 to create densely annotated datasets in medicine. Such datasets include SkinCon [1], PH2 [17], derm7pt [18], and
67 Osteoarthritis Institute Knee X-ray dataset (OAI) [19]. However, because these medical datasets are annotated by
68 humans and usually require domain expertise, the number of densely annotated datasets, the number of images in
69 each dataset, and the number of “concepts” that can be labeled in each image are limited. MONET overcomes these
70 challenges by generating medical concepts automatically. Our framework is built upon *contrastive learning*, a recent AI
71 breakthrough that enables the direct utilization of natural language descriptions on images [20]. Since this approach
72 does not require manual labeling, it can unlock the potential of vast numbers of image and text pairs, allowing for the
73 harnessing of data of much larger scale than was possible with supervised learning.

74 To train MONET, we collect an extensive set of dermatology image and text pairs ($n = 105,550$) from PubMed
75 articles and medical textbooks. We map an image (or text) into a lower-dimensional vector or a *representation* through
76 a neural network, namely the *encoder* (Fig.1A), creating a *representation space*. During the training process, an image
77 and text from the same pair are forced into close proximity in the representation space, while those from different pairs
78 are forced to be farther apart (Methods). Once trained, MONET’s *zero-shot capability* (*i.e.*, the ability to generate
79 a medical concept without a separate learning procedure) generates concepts (Fig. 1B and Methods). When a user
80 provides an image and a list of concepts to generate, MONET determines the presence of each concept in the image
81 by calculating the distances between the image and concept text prompts in the joint representation space, where
82 images and texts are jointly mapped.

83 MONET’s automatic concept generation capability enables a whole range of capabilities in medical AI that were
84 previously infeasible in practice. We showcase MONET’s versatility by demonstrating its use in auditing data, auditing
85 models, and creating inherently interpretable models. MONET enables a sophisticated multi-point analysis and can be
86 used to probe any part of the medical processing workflow. For data auditing, we apply MONET to the International
87 Skin Imaging Collaboration (ISIC) dataset [21–26], the most widely used data in dermatology AI [27], to confirm
88 known trends and discover new ones. We also use MONET to identify which input characteristics lead to errors
89 in medical AI models. Finally, we integrate MONET with the concept bottleneck model (CBM) [15], a well-known
90 approach for building inherently interpretable models, and show MONET+CBM’s advantages over supervised models
91 in terms of *both* performance and interpretability. All of these tasks are central to the development and deployment
92 of trustworthy and transparent AI models in medicine.

93 Results

94 Automatic concept generation

95 We first assess MONET’s concept generation capability before demonstrating how this capability can improve the
96 transparency and interpretability of medical AI pipelines. The fundamental mechanism in MONET’s concept gener-
97 ation is the mapping of medical concepts and images onto a joint representation space. This allows the generation
98 of a *concept presence score*, *i.e.*, the degree to which a concept is present in an image, by measuring the distance
99 between the image and concept text prompts in the joint representation space (Methods). We evaluate MONET’s
100 concept generation ability by identifying those images with the highest concept presence scores using both *clinical*
101 and *dermoscopic* images, the two widely used dermatological image types. Dermoscopic images are captured using
102 digital photography with a specialized dermatological instrument called a dermoscope that magnifies skin lesions to
103 capture fine details, while clinical images are often taken at least 6 cm away with a digital camera. For our evaluation,
104 we employ clinical images ($n = 4,960$) from the Fitzpatrick17k and Diverse Dermatology Images (DDI) datasets and
105 dermoscopic images ($n = 71,242$) from the ISIC dataset (Methods).

106 Fig. 2 and Supplementary Fig. 1-2 display clinical and dermoscopic images with high concept presence scores

107 for each concept. These represent examples of widely used medical concepts in dermatology. Dermatologists use a
108 standardized terminology to describe the morphology, color, configuration and distribution of skin lesions. MONET
109 excels at recognizing these medical concepts in clinical and dermoscopic images. For example, “erythema” is a term
110 used by dermatologists to describe a red or violaceous color, which usually occurs in the presence of inflammation.
111 It can be found in various skin diseases, such as atopic dermatitis, psoriasis, and rosacea. Two board-certified
112 dermatologists confirmed that the images with large presence scores for erythema exhibit skin redness in both clinical
113 and dermoscopic images (Fig. 2). Similarly, images with the concept “blue” show dark blue lesions with pigmentation
114 in the dermis, resulting from the Tyndall effect. Moreover, MONET was able to retrieve images with primary
115 morphological features such as bullae (large, tense fluid-filled blisters) and pustules (small, pus-filled blisters), as well
116 as secondary morphological features including ulcers (open sores) and xerosis (dry, scaly skin).

117 We assess the performance of MONET’s concept generation using ground truth concept labels in SkinCon (Table
118 1). Of the 48 concepts in the dataset, we exclude any with less than 30 positive examples, leaving 21 concepts for our
119 analysis. We use 1,645 images from Fitzpatrick17k and DDI datasets with ground truth SkinCon concept labels. We
120 compare MONET’s performance to a supervised learning approach, training a ResNet-50 model using ground-truth
121 concept labels from SkinCon [28], and to a pre-existing contrastive image-text model that was not specifically trained
122 on dermatology images but on 400 million available image-text pairs on the web - the CLIP (Contrastive Language-
123 Image Pretraining) model by OpenAI [20] (Methods). We find that MONET outperforms the ResNet-50 model and
124 the CLIP model in terms of concept generation. Specifically, we compare the mean of the area under the receiver
125 operating characteristic curve (AUROC) across concepts with ground truth labels and count how many concepts
126 achieved an AUROC higher than 0.7. MONET achieves a mean AUROC of 0.766; in contrast, CLIP achieves a mean
127 AUROC of 0.692. The ResNet-50 model, trained to predict concept labels, achieves a mean AUROC of 0.692. Of the
128 21 concepts analyzed, MONET remarkably displays 19 concepts with an AUROC over 0.7, compared to 9 for CLIP
129 and 11 for the fully supervised model. Additionally, we conduct the same comparative analysis using disease labels,
130 which can be viewed as the most fine-grained concept labels; we map the disease labels instead of SkinCon concepts
131 to the image-text joint representation space. Our findings indicate that MONET’s performance is still comparable to
132 that of supervised models in this case (Supplementary Table 2). This observation is consistent with a previous study,
133 which found that a contrastive learning model trained on radiology images demonstrates comparable performance in
134 predicting pathologies in chest X-rays to that of supervised learning models [29].

135 We also evaluate the performance of MONET’s concept generation across diverse skin tones (Methods). A recent
136 study revealed that state-of-the-art dermatology AI models exhibit uneven performance across skin tones, particularly
137 underperforming on dark skin tones, potentially due to the insufficient representation of diverse skin tones in training
138 data [30]. One advantage of contrastive learning, the technique used to train MONET, is its ability to easily harness
139 heterogeneous data from diverse sources for training. This can help reduce performance disparities across demographics
140 compared to training on a single data source. To determine whether MONET is free from this issue, we compared
141 its performance per skin tone using the Fitzpatrick skin type labels included in the Fitzpatrick17k and DDI datasets.
142 MONET demonstrated even performance across skin tones (Supplementary Table 3).

143 Finally, we also explore MONET’s capability to recognize non-clinical concepts, such as artifacts that are irrelevant
144 to the diagnosis. Many studies have shown that medical AI systems use such non-clinical concepts to make predictions,
145 particularly when a spurious correlation exists between the artifacts and prediction labels [6, 31]. In dermatology AI,
146 it has been shown that artifacts, such as clinical marking or size reference stickers, can have a detrimental effect on
147 the model’s generalizability [32–34]. The ability of MONET to identify such artifacts, in addition to clinical concepts,
148 will facilitate more fine-grained auditing and debugging of medical AI pipelines. Supplementary Fig. 3 shows images
149 from the ISIC dataset that MONET identified as containing non-clinical concepts. Supplementary Fig. 3A shows
150 images with purple pen ink marking that MONET automatically identified; in dermatology, lesions that are biopsied
151 are often routinely marked with purple ink markers. Supplementary Fig. 3B shows orange stickers that MONET
152 identified; they serve as a lesion marker. In the ISIC dataset, these orange stickers predominantly show up in the
153 pediatric cases, which are largely benign. As these artifacts predominantly appear in certain types of images, they
154 may inadvertently cause AI algorithms to associate purple ink markings with malignancy and orange stickers with
155 benign lesions [34]. Also, MONET automatically identifies images with body location features (such as nails and
156 hair) (Supplementary Fig. 3C, D). A recent study has shown that anatomic locations may play a critical role in
157 the performance of dermatology AI algorithms; however, most datasets lack these annotations [35]. Further, MONET
158 identifies images with dermoscopic borders, which appear on a subset of dermoscopic images depending on the image
159 processing process (Supplementary Fig. 3E).

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



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 10 images for each concept. **(A)** Clinical images from the Fitzpatrick17k and DDI datasets. We exclude images inappropriate for public display due to the inclusion of sensitive body parts; for completeness, we denote the filenames of these files in Supplementary Table 1 **(B)** Dermoscopy images from the ISIC dataset.

Method	Mean AUROC
MONET	0.766 (19/21)
CLIP	0.692 (9/21)
ResNet-50 (Fully supervised)	0.692 (11/21)

Table 1 | Performance of MONET’s concept generation as compared to the baselines. We use concept labels in the SkinCon dataset as ground truth. Of the 48 concepts in the dataset, we exclude any with less than 30 positive examples, leaving 21 concepts for our analysis. The baselines 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. The numbers in parentheses represent the count of concepts for which the method achieves an AUROC over 0.7 over the total number of concepts examined.

162 Data auditing

163 Ensuring that training data aligns with users’ expectations is a crucial first step toward developing AI models since
164 many unreasonable model behaviors stem from unidentified pitfalls in the training data [6, 32, 34]. For example, in
165 dermatology, when preparing a dataset for training an AI model to diagnose malignancy, the differentiating features
166 in the data between classes (*i.e.*, malignant and benign images) should not contain any biases or spurious correlations,
167 such as the pen markings used to identify biopsied lesions [34]. Upon identifying any irregularities, adjustments can
168 be made, such as improving the data collection and processing [36, 37] or applying optimization techniques to improve
169 generalizability [31, 38].

170 However, examining large-scale datasets for irregularities is challenging and labor-intensive. One approach is to
171 manually label features of interest and create a contingency table between each feature and the target label to check for
172 spurious correlations; however, this is subjective and not easily scalable [5]. Another approach is to train a generative
173 model, such as CycleGAN [39], to learn the distribution of data for each class [6]; the trained generative model can
174 modify an image from one class to resemble an image from another class. By observing these changes, a data examiner
175 can identify the distinguishing features of each diagnostic group. However, generative models are difficult to train and
176 necessitate manual inspection of the transformed images.

177 To address the issue, we can employ MONET to automate the data examination process. MONET’s automatic
178 concept generation capability can explain the distinguishing features between any two arbitrary sets of images in the
179 language of human-interpretable concepts, which we refer to as *concept differential analysis* (see Methods). Supple-
180 mentary Fig. 4 shows benchmark analysis results.

181 As a practical use case, we employ MONET to analyze the ISIC dataset, the largest dermoscopic image dataset,
182 which consists of over 70,000 publicly available images that are commonly used to train dermatology AI models [21–
183 27]. We divide the images into a malignant ($n = 10,091$) and a benign set ($n = 61,151$), assuming malignancy as
184 the prediction target, and examine which concepts were more present in which set (Fig. 3A). We test for 48 concepts
185 listed in SkinCon along with eight artifacts, including red coloration, pinkish coloration, and purple ink markings,
186 nails, hair, orange sticker, gel, and dermoscopic border.

187 The top 5 concepts in the malignant images are ulcer, erosion, warty, pinkish coloration and blue coloration; in
188 contrast, the top 5 concepts in the benign images are orange sticker, hypopigmentation, the color salmon, xerosis, and
189 hyperpigmentation. These concepts represent key features that a prediction model, trained on the ISIC dataset, may
190 potentially use to differentiate benign from malignant lesions. They encompass both clinically pertinent features such
191 as ulcer, crust, erosion, warty, and black coloration, as well as irrelevant confounders such as orange sticker and nail.
192 Skin ulcerations and erosions are commonly linked to malignant skin tumors such as melanoma, basal cell carcinoma
193 and squamous cell carcinoma, making their association with malignancy logical. On the other hand, prediction models
194 learning from confounding concepts may lead to biases and detrimental consequences. For example, dermatologists use
195 orange stickers as a lesion marker, and with the ISIC dataset, this was predominantly used in the pediatric population
196 which had mostly benign lesions. The bias in the data could lead a model to erroneously associate orange stickers
197 with a low likelihood of malignancy.

198 Furthermore, we can use this approach to assess distinctive trends specific to different data sources. In medicine,
199 data sharing across institutions is limited due to the sensitive nature of medical data and regulatory constraints.
200 In many cases, a medical AI is developed within a few institutions and then distributed to other institutions for
201 deployment. For this reason, it is important to understand and monitor the shifts in the concept representations
202 between data and identify factors that can potentially compromise the transferability of medical AI. By doing so,
203 necessary adjustments can be made preemptively. The ISIC dataset is a collection of images from multiple hospitals
204 and research institutions, which serves as an ideal resource for simulating situations where the development and
205 deployment sites differ. In this analysis, we focus on two cohorts released in the ISIC Challenge 2019—the Medical
206 University of Vienna (Med U. Vienna, malignant: $n = 1,824$ / benign: $n = 8,049$) and the Hospital Clínic de
207 Barcelona (Hospital Barcelona, malignant: $n = 6,097$ / benign: $n = 6,205$)—since they represent the two largest
208 cohorts in the entire ISIC dataset when stratified by release year and data source. We perform concept differential
209 analysis between malignant and benign images, as noted above, for each cohort separately. We then compare the
210 obtained concept differential expression scores between the two cohorts (Fig. 3B).

211 When we sort the test concepts in the order of absolute differences, the top-listed concept was a “red” hue.
212 Redness is positively correlated with malignancy for the images from Hospital Barcelona but negatively correlated
213 with malignancy for the images from Med U. Vienna. This means that redness has the potential to compromise the
214 transferability of medical AI between two institutions. This trend is clearly visible in the sampled images from each
215 cohort, as well. Fig. 3C displays images sampled from the top 100 images with high concept expression scores for the
216 red coloration for each cohort, along with their diagnostic labels. The images that have more redness collected from
217 Med U. Vienna are often benign, while those collected from Hospital Barcelona are often malignant. The top 500 and
218 top 1,000 red images in the Med U. Vienna contain more benign than malignant ones, while the top 500 and top 1,000

219 red images from Hospital Barcelona still contain more malignant than benign samples. (Fig. 3D).

220 In sum, we demonstrate how MONET can assist with auditing large-scale datasets. Since concept differential
221 analysis is conducted simply by describing a concept in a natural language, the approach fosters the scalable discovery
222 of trends within the data. Using the insights gained through this process, AI model developers can enhance data
223 collection, processing, and optimization techniques, ultimately yielding more reliable and trustworthy medical AI
224 models.

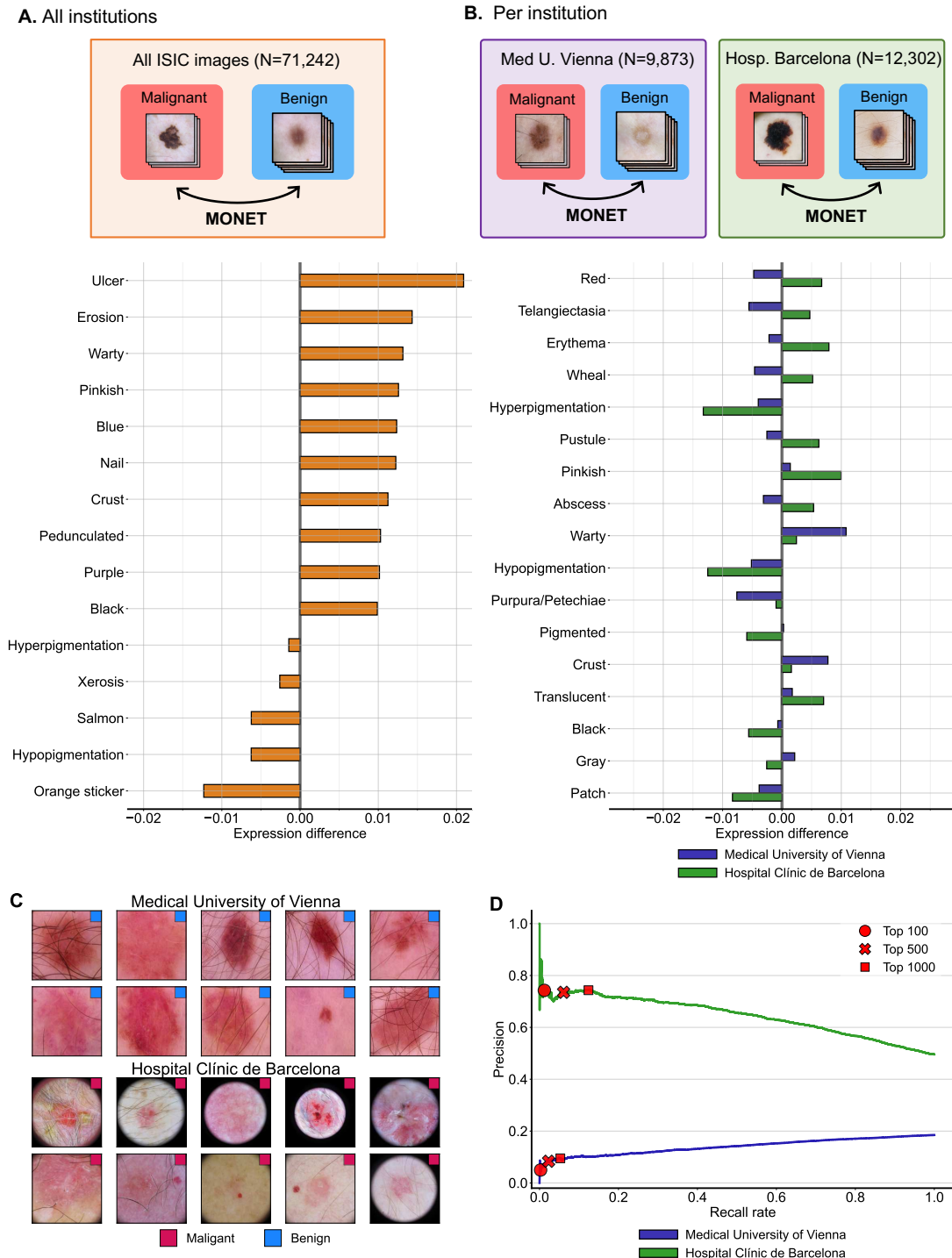


Fig. 3 | Concept-level data auditing. (A) We perform concept differential analysis between malignant images and benign images in the ISIC dataset. We show the top 10 concepts with positive values and the top 5 concepts with negative values. A positive value means the concept was more present in the malignant images than in the benign images, and vice versa. (B) We perform 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 the Medical University of Vienna, and the green bar represents the output from the Hospital Clínic de Barcelona. We show the top 15 concepts based on their absolute differences between the two cohorts. (C) Examples of red images in each cohort. We display 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 are sorted based on their concept presence scores for redness and then compared to their malignancy labels. Precision is defined as the proportion of malignant images above a certain threshold out of all images above that threshold, while recall rate is defined as the proportion of malignant images above the threshold out of all malignant images. The top 500 and top 1000 red images from Barcelona Hospital still contain more malignant than benign samples.

225 Model auditing

226 Various techniques for auditing AI models have been developed to understand the factors involved in AI decision-
227 making. One classical approach is the use of *saliency maps*, which highlight regions in an input image that significantly
228 contribute to the model’s prediction [40–42]. The saliency maps of each image help to identify which pixel-level
229 features lead to a correct or incorrect prediction. However, the highlighted pixels are often not easily translated into
230 semantically meaningful concepts understandable to a human [43].

231 To address this issue, we can use MONET to audit AI models through the lens of medical concepts. We developed
232 a method “model auditing with MONET” (MA-MONET) that uses MONET to automatically detect semantically
233 meaningful medical concepts that lead to model errors (Methods). MA-MONET starts by sorting images from a test
234 set into groups based on their visual similarity. It then labels the clusters that perform below the overall accuracy as
235 low-performing. For each low-performing cluster, MONET identifies medical concepts. Each low-performing cluster
236 is then compared to a high-performing counterpart containing similar images, with concepts separately identified in
237 the high-performing cluster. If two visually similar clusters (one high-performing, the other low-performing) differ in
238 terms of a few concepts, these differing concept terms can be hypothesized as leading to high error rates. Finally, we
239 produce a ranked list of medical concepts identified by MONET that differentiate the two clusters.

240 To validate our model auditing, we first perform a benchmarking analysis, using a situation where the ground
241 truth (*i.e.*, the concepts leading to model error) is already known (Fig. 4A and Methods). We create a training
242 dataset with spurious correlations from the Fitzpatrick17k and DDI datasets: 500 malignant images that feature a
243 particular SkinCon concept, while the 500 benign images do not. After training a CNN model to predict malignancy
244 on this confounded dataset, we test it on a dataset where the correlation is reversed (Methods); in the test set, 500
245 sampled benign images have the SkinCon concept, while 500 sampled malignant images do not. We cluster images
246 in the test set into 40 clusters, and about 20 of these clusters underperform, meaning their accuracy falls below the
247 average accuracy. For these low-performing image clusters, we apply the MONET-based error explanation method
248 to obtain a ranked list of medical concepts that would explain the model error. Finally, we observe if the concept of
249 spurious correlation we know is recovered.

250 We conduct the analysis using 5 concepts that remain after filtering out concepts with fewer than 30 samples in each
251 category required for creating the confounded training and test sets (*i.e.*, malignant–with concept, malignant–without
252 concept, benign–with concept, and benign–without concept) : “crust”, “hyperpigmentation”, “plaque”, “erythema”,
253 and “papule”. For each of the 5 concepts, we repeat this analysis 20 times with different random seeds changing the
254 training and test sets. Consequently, we test 100 settings in total. Across these settings, the mean AUROC of the
255 trained model is 0.779 for validation sets, but decreases to a mean of 0.458 for test sets.

256 We measure the frequency of the known spurious correlation being recovered by MA-MONET (Fig. 4B), checking
257 if the top-N concept lists of any low-performing clusters include the known spurious correlation concept. We compare
258 this outcome with that of the out-of-the-box CLIP model, which was not specifically trained on dermatology data
259 [20]. The low-performing clusters being analyzed in each setting are the same for both methods, but the enumeration
260 of concepts associated with errors is done using CLIP instead of MONET. The results for the top 1, 2, and 3 rankings
261 are markedly higher with MONET, at 0.590, 0.800, and 0.890, respectively, compared to those obtained with CLIP,
262 which are 0.270, 0.520, and 0.660, respectively.

263 To showcase its use in real-world scenarios, we consider a widely occurring situation where a model is trained
264 at one institution and deployed at another [44, 45] (Fig. 4C). For training and testing, we use the same datasets
265 we used in the data auditing section, specifically the two largest cohorts in the ISIC dataset: the Hospital Clínic de
266 Barcelona ($n = 12,302$) and the Medical University of Vienna ($n = 9,873$). We train CNN models on the data from
267 one institution using a standard training regimen and test them on the data from the other institution, and vice versa.

268 For the AI model trained on Med U. Vienna, it showed an AUROC of 0.885 in the internal validation, but the
269 value dropped to 0.707 during the external validation. This decline in performance would prompt AI model developers
270 to question which input characteristics led to model errors. To elucidate this, we use MA-MONET to pinpoint which
271 concepts are associated with model errors. For each cluster with high error rates, the misclassified images and the
272 terms associated with errors are shown in Fig. 4D (displaying the top 5 clusters sorted in the order of high error rates)
273 and Supplementary Fig. 5 (displaying the top 15 clusters sorted in the order of high error rates). For example, the
274 cluster with the highest error rate, displayed in the first row of Fig. 4D and Supplementary Fig. 5A, is characterized by
275 the concepts “blue”, “black”, “gray”, “pigmented”, and “flat-topped”. Remarkably, we notice several clusters where
276 high error rates are explained by concepts related to “red”. For instance, the cluster shown in Supplementary Fig. 5F
277 are characterized by “erythema”, “salmon”, “sclerosis”, “scar”, and “translucent”. Interestingly, we also find that the
278 malignant images are predominantly misclassified as benign. Out of the 74 malignant images in the cluster, 55 images
279 are misclassified to be benign. This observation aligns with the trends we noted in the data auditing experiment,
280 where red images from Med U. Vienna (*i.e.*, training dataset) were benign, while the red images in Hosp. Barcelona
281 (*i.e.*, test dataset) were malignant.

282 Conversely, we also train an AI model on Hosp. Barcelona data and tested it on Med U. of Vienna data. In
283 this case, the AUROC of 0.844 in internal validation drops to 0.741 during external validation. For each cluster with
284 high error rates, the misclassified images and the identified terms associated with errors are shown in Fig. 4E and
285 Supplementary Fig 6. The cluster with the highest error rate, shown in the first row of Fig. 4E, is characterized by the
286 concepts “pinkish”, “erythema”, “gray”, “red”, “atrophy”. Out of the 197 benign images in the cluster, 103 images
287 are misclassified as malignant. This observation also aligns with the trends we noted in the data auditing experiment.

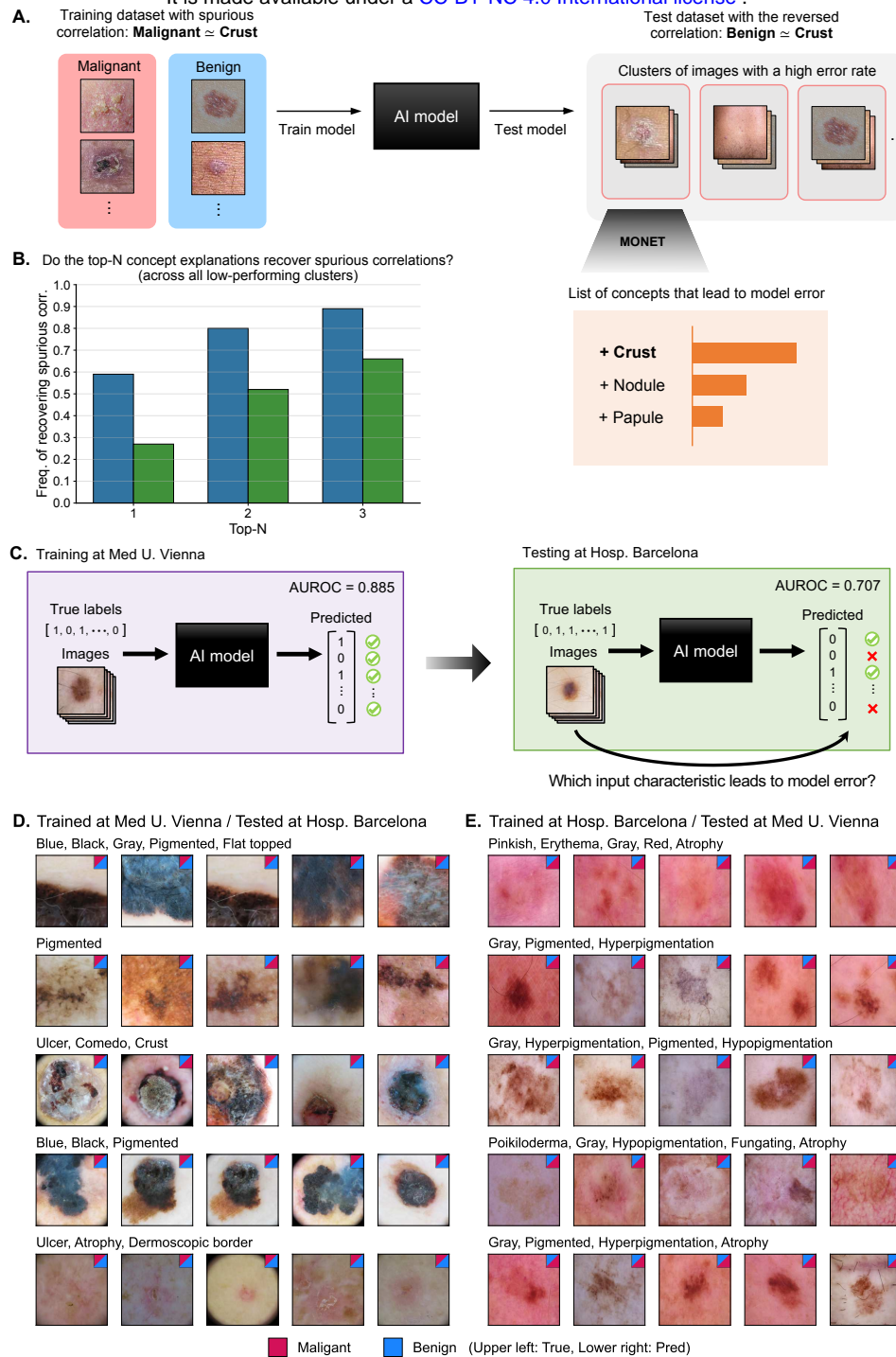


Fig. 4 | Concept-level model auditing. (A) We perform a benchmark analysis to see how well “model auditing with MONET” (MA-MONET) can identify the semantically meaningful concepts that lead to model error. To this end, we generate settings where we know the ground truth (*i.e.*, concepts that lead to model errors); we create a training and test dataset with spurious correlation. We use MA-MONET to identify which concepts lead to model error for an AI model trained on this confounded dataset. MA-MONET returns a ranked list of concepts that explain model errors. (B) The frequency of the known spurious correlation being recovered by MA-MONET is shown. (D)-(E) Each row displays one of the top 5 clusters, sorted by high error rates. For each cluster, we show the misclassified images and the corresponding concepts associated with errors. We represent 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 5 misclassified images shown for each are selected based on the average concept presence of the identified concepts.

288 Inherently interpretable model building

289 In medicine, inherently interpretable models are of particular interest since they allow physicians to easily decipher
290 the factors influencing a model’s decision. Rather than training a complex black-box model that requires post-hoc
291 explanation, an inherently interpretable model offers greater transparency and control in model behavior. *Concept*
292 *bottleneck models* (CBMs) are a well-known type of inherently interpretable model [15]. CBMs make predictions in
293 a two-step manner: first, they predict concepts from the input using a complex model such as a CNN (*i.e.*, input
294 \rightarrow concept); then, they use these predicted concepts to predict the target output via a linear model (*i.e.*, concept
295 \rightarrow output). As each node in the bottleneck layer represents a human-interpretable concept, CBMs offer greater
296 explainability. Further, CBMs facilitate the incorporation of domain knowledge into models by imposing constraints
297 on the concepts used, thereby improving the ability to control model behavior.

298 However, CBMs have a significant limitation: they require concept annotations in the training data. To achieve
299 high performance with CBMs, it is essential to train them on a sufficient number of samples and ensure they operate
300 with an adequate number of concept labels that are relevant to the prediction task. This constraint has hindered the
301 practical application of CBMs. We address this issue by using MONET’s automatic concept generation to eliminate
302 the need for manually annotated concept labels in the original training procedure of the CBMs.

303 We explore the application of MONET and CBMs for melanoma and malignancy prediction tasks, the most preva-
304 lent prediction tasks in dermatology AI. MONET+CBM approach predicts the target (*i.e.*, melanoma or malignancy)
305 by combining automatically generated concepts by MONET (*i.e.*, the concept presence score) via a linear model
306 (Fig. 5A and Methods). This gives CBM access to many concepts and many samples compared to a manual labeling
307 approach. The following comparison makes use of 4,960 clinical images sourced from the Fitzpatrick17k and DDI
308 datasets. For melanoma prediction, we further filter images to ensure that data mirrors a well-defined clinical task,
309 resulting in 775 images (Methods). For each setting, we repeated evaluations with 20 different train-test splits.

310 We observe that access to a large number of concepts and samples offers performance advantages. Fig. 5B-E
311 compares the performance of MONET+CBM to that of using manual concept annotations. For a fair comparison, we
312 use both methods on the same set of concepts, specifically the 48 concepts in SkinCon. We chose SkinCon concepts
313 because they already have manual annotations provided by experts. For MONET+CBM, we use all training samples
314 and concepts because it can automatically generate concepts without expert annotation. As we increase the number
315 of manually labeled samples used or the number of concepts used, the performance of the CBM created from manual
316 labeling improves. However, even when all manually labeled concepts and training samples are used, the manual
317 approach is not able to match the performance of MONET+CBM, which has access to more samples due to the
318 ability to produce automatic concept labels. As concepts in SkinCon are not annotated for all images, the manual
319 label approach is limited to 1,316 malignancy and 294 melanoma images that have manual concept labels; in contrast,
320 MONET makes use of all 3,968 malignancy and 620 melanoma images in our training set.

321 We compare the performance MONET+CBM to the other baselines, such as supervised models and CLIP-based
322 CBM, for the same prediction targets (as described in Methods) (Fig. 5F and G). Dermatologists selected 11 target-
323 relevant curated concepts for the bottleneck layer to compare MONET+CBM and CLIP+CBM, which can both
324 flexibly label concepts. Compared to using all 48 SkinCon concepts, the mean AUROC across runs using the 11 curated
325 concepts decreased from 0.854 to 0.805 for malignancy prediction and decreased from 0.896 to 0.892 for melanoma
326 prediction. Still, for both predicting malignancy and melanoma, MONET+CBM outperforms all other baseline
327 methods in terms of the mean AUROC (for malignancy, 0.805 with a standard deviation of 0.014; for melanoma,
328 0.892 with a standard deviation of 0.019). Out of 20 runs with different random splits of the train and test data,
329 MONET+CBM outperformed all other methods in 15 runs for malignancy prediction and 18 for melanoma prediction,
330 with the linear probing method outperforming in the remaining runs. We also conduct one-sided paired t-tests,
331 comparing the AUROC values of MONET+CBM to those of other methods, where the alternative hypothesis is that
332 MONET+CBM’s AUROC is higher than the other method. In all cases, the resulting p-values are less than 0.001.
333 Thus, MONET’s ability to automatically generate concepts enables the creation of models that are both interpretable
334 and high-performing.

335 While the concepts used for the concept bottleneck model were selected by dermatologists based on factors that can
336 help predict melanoma, we wanted to check that the way these concepts were used by this model align with established
337 clinical rules for the same task. Fig. 5H and I show the weights of the trained linear classifier corresponding to the
338 concepts used in the bottleneck layer. For the Melanoma target, the results are consistent with the ABCDEs of
339 melanoma [46], which define easily recognizable features—namely, **a**symmetry, **b**order irregularity, **c**olor variation,
340 **d**iameter, and **e**volution—that differentiate malignant melanomas from benign melanocytic nevi. From the concept
341 weights obtained, all concepts that coincide with the ABCDEs get a positive weight as expected, indicating a positive
342 correlation with the melanoma prediction target. The concept “blue” also has a positive weight referring to the
343 dermoscopy concept of blue-white veils observed in melanomas. The concepts “white” and “tiny” get an almost zero
344 weight, while the concept “regular” gets a negative weight, consistent with prior knowledge shared by dermatologists,

345 that regular borders and color indicate a benign lesion. For the malignancy target, no well-defined guidelines exist for
346 deriving concepts; thus, the same concept list as the Melanoma target is used. The results are similar, with a majority
347 of the concepts retaining their directionality, except for increased sparsity in concept weights.

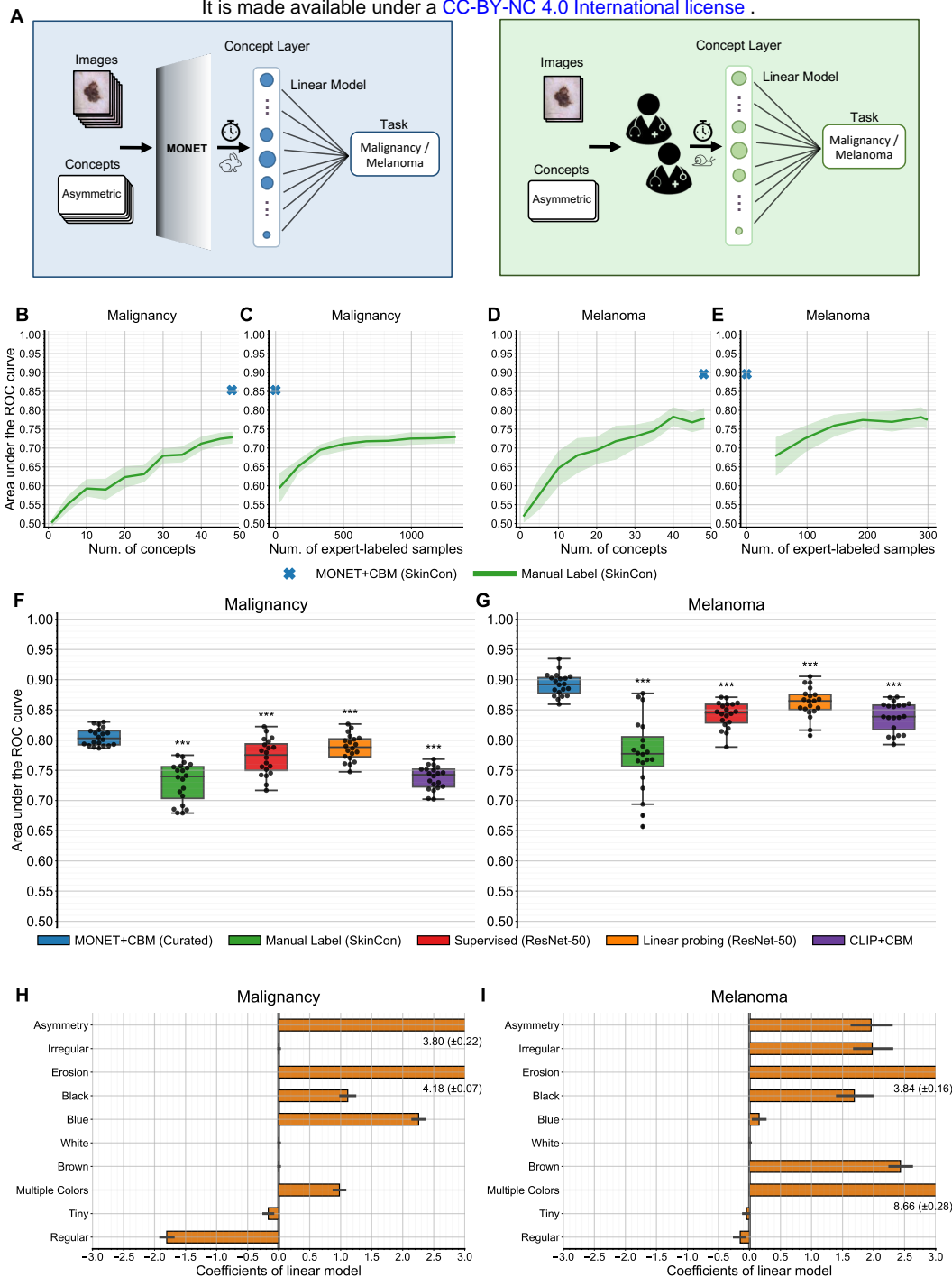


Fig. 5 | Concept bottleneck model. **(A)** Concept bottleneck model built using concepts generated by MONET (blue). The model first generates concepts using MONET and then predicts disease labels by combining them via a linear model. Concept bottleneck model built using concepts manually labeled by experts (green). The model uses manually annotated concept labels to predict disease labels using a linear model. Manual annotations take a lot longer than concept generation using MONET. **(B)-(C)** Performance of a malignancy prediction model trained using manual labels with respect to the number of concepts and the number of expert-labeled samples. **(D)-(E)** Performance of a melanoma prediction model trained using manual labels with respect to the number of concepts and the number of expert-labeled samples. **(B)-(E)** MONET+CBM is shown as a cross mark because it can utilize all concepts without expert annotation. The shaded area represents the 95% confidence interval. **(F)** Performance comparison of malignancy prediction models. **(G)** Performance comparison of melanoma prediction models. **(F)-(G)** Unlike **(B)-(E)**, MONET+CBM uses task-relevant concepts curated by dermatologists. Each dot represents the AUC measure for individual runs with a different train-test split. The box represents the interquartile range with its lower and upper bounds corresponding to the first quartile and third quartile, respectively. p values derived from one-sided paired t -tests comparing MONET+CBM and other methods are indicated: $* < 0.05$, $** < 0.01$, $*** < 0.001$; $n = 20$ runs of each method. **(H)** Coefficient of the linear model in MONET+CBM for malignancy prediction. **(I)** Coefficient of the linear model in MONET+CBM for melanoma prediction. **(H)-(I)** The error bars present the 95% confidence interval.

348 Discussion

349 Even with the regulatory approval of AI-supported medical devices, much of the AI pipeline is not transparent - from
350 large-scale datasets that may contain biases to so-called “black-box” models that are not easily audited or interpretable
351 [27]. One approach to improving transparency and trustworthiness is identifying semantically meaningful, human-
352 understandable concepts that are present in datasets or used by models. However, to date, all datasets and methods
353 developed using concepts have relied on human labeling and domain experts, which is not tractable for large-scale,
354 real-world deployment.

355 Here, we demonstrate the ability to develop automated concept labeling in a medical domain that would usually
356 require domain expertise, and we showcase how these automated concepts can be used to perform tasks for trustworthy
357 AI development at all stages, from developing new models to auditing existing datasets and models. We focus on
358 the field of dermatology due to the heterogeneity of the image data, the large number of potential concepts, and the
359 ability to validate our methods on existing datasets.

360 Prior work using image-text models in medicine focused on self-supervised training of diagnostic models that can
361 identify a handful of disease labels, such as in radiology or pathology [29, 47]. However, our challenge is to develop a
362 model that can label a vast number of human-understandable concepts across two image modalities within dermatology:
363 clinical images and dermoscopic images. To solve this challenge, we collect a large number of dermatology image-text
364 pairs from PubMed articles and medical textbooks and train an image-text model, MONET. This dermatology image-
365 text model facilitates automatic generation of concepts, and we show how it can be used to improve the transparency
366 of dermatology AI models. To our knowledge, we are the first to use a large biomedical image-text model to improve
367 the transparency and explainability of medical AI systems.

368 For a concept generation task where we had domain expert labels as the ground truth, we find that MONET,
369 which requires no domain expert labeling, outperforms the baseline CLIP model and a supervised model trained
370 from domain-expert labeled images. These findings are significant since the bottlenecks of data labeling and domain
371 expertise time can be overcome with image-text models developed from existing medical corpora.

372 After demonstrating the ability to generate concepts on par with supervised models, we showcase MONET’s ability
373 to facilitate AI auditing and transparency in the dermatology domain. For example, artifacts such as pen markings,
374 stickers, and hair are known to affect dermatology model performance [34, 48]. However, most studies do not assess
375 the influence of artifacts on their models because their datasets are not labeled for these anomalies. We demonstrate
376 MONET’s ability to automatically identify these artifacts, which can be useful for data and model auditing. As
377 an example of how this kind of auditing is useful, we analyzed data from the ISIC 2018 challenge and find that a
378 “red” hue appears more often in benign images for the Medical University of Vienna while images from the Hospital
379 Clínic de Barcelona more often have a “red” hue associated with malignant images. This leads to confounding if a
380 model is trained on one site’s dataset and tested on the other, as we see when we implement MA-MONET for model
381 auditing. These insights derived from using MA-MONET might not be readily achievable via conventional saliency
382 map techniques [43]. For instance, the “red” hue noticed using MA-MONET is not a localized attribute, so a saliency
383 map approach would not necessarily highlight this aspect [43]. Utilizing the insights gained through MONET and
384 MA-MONET, AI model developers can refine data collection and processing and also improve optimization techniques,
385 consequently fostering the development of more reliable and trustworthy medical AI models.

386 In medicine, inherently interpretable models are of particular interest since they allow physicians to easily decipher
387 the factors influencing a model’s decision. Concept bottleneck models (CBMs) are one such inherently interpretable
388 model but have been limited because they require a priori concept labels, which only a handful of medical datasets
389 contain. MONET overcomes this issue with automatic concept labeling, allowing the creation of CBMs that were not
390 previously possible.

391 MONET demonstrates the ability to automatically label numerous concepts across heterogeneous disease states and
392 across two modalities (clinical and dermoscopic) in dermatology. A limitation of our experiments is the availability
393 of diverse skin tones in dermoscopic images since no public datasets exist with diverse dermoscopic images [49].
394 Thus, when assessing MONET on clinical images, we utilize two datasets known to include a diversity of skin tones,
395 Fitzpatrick 17k and DDI, and find that MONET performs well with these datasets.

396 While MONET covers heterogeneous dermatology data across two modalities, clinical and dermoscopic, future
397 iterations can extend to other forms of medical imaging to improve AI transparency for those use cases. MONET
398 demonstrates that AI transparency and trustworthiness at scale is feasible in a way that was previously impossible:
399 through image-text models tailored to the medical domain of interest.

400 Methods

401 Dataset

402 Overview

403 We trained MONET on 105,550 pairs of image and text collected from PubMed articles and medical textbooks [50, 51].
404 We evaluated MONET using images from the International Skin Imaging Collaboration (ISIC) [21–26], Fitzpatrick17k
405 [52], and Diverse Dermatology Images (DDI) datasets [30].

406 PubMed

407 The PMC Open Access Subset is a dataset with millions of scientific articles released by PubMed Central (PMC) [50].
408 First, to find dermatology articles in the dataset, we queried papers in PMC using dermatology-related terms (*i.e.*,
409 dermatology, melanoma, skin cancer), 114 disease labels in the Fitzpatrick17k dataset [52], and 48 concept labels in
410 SkinCon [1]. We downloaded 496,510 articles found via this query. In total, the articles contained 3,172,490 figures.
411 Next, we filtered out non-dermatology-related figures (e.g., graphs, illustrations, diagrams, slide images, and X-ray
412 images). To this end, we repeated the process of running a clustering algorithm on the images and manually excluding
413 groups of non-dermatology ones. Specifically, we carried out the following procedure. The clustering features were
414 50 principal components of the embedding of the penultimate layer of the EfficientNetV2-S model pre-trained on
415 ImageNet [53]. Using the features, we ran a K-means clustering algorithm with the K (*i.e.*, the number of clusters) of
416 20. For fine-grained filtering, we further applied the K-means algorithm with the K of 20 on each cluster, resulting in
417 400 clusters in each step. For the K-means algorithm, we used the implementation in scikit-learn Python package (ver.
418 1.2.2) [54]. We manually inspected 50 samples of each cluster and filtered out clusters with non-dermatology images.
419 After going through this step three times, we determined that the remaining clusters contained mostly dermatology
420 images. Post-filtering, 50,265 images remained. Finally, we paired the figure captions to their corresponding images
421 based on the provided XML-formatted file. This file stores the article’s structure with components such as abstract,
422 sections, figures, and figure legends tagged.

423 Textbook

424 We first extracted images from 55 medical textbooks, yielding a total of 104,223 images. After undergoing the same
425 filtering procedure as we did for PubMed images, 55,285 images remained. The PDF format of the textbooks made
426 matching images with associated text difficult, since PDFs lack the structure information provided by XML formats.
427 To address this issue, we implemented the following procedure. We used PyMuPDF (ver. 1.21.1), an open-source PDF
428 rendering software, to parse the PDF files, extracting text and image objects along with their respective coordinates.
429 Then, we assigned text to images appropriately based on the following criteria. First, we included text that starts
430 with words indicative of figure legends, such as “Fig” or “Figure”. Next, we excluded text based on font and font size.
431 Also, since each textbook maintained a consistent layout for placing figure legends (*i.e.*, legends positioned above or
432 below the figure), we incorporated this into our filtering process. Lastly, from the remaining text, we selected the one
433 closest to the image. We customized the specific parameters (*i.e.*, figure identifier, font, font size, and caption position
434 relative to the image) for each textbook to ensure accurate text-image associations.

435 ISIC

436 The International Skin Imaging Collaboration (ISIC) archive is a repository of digital skin images, primarily consisting
437 of dermoscopic images, sourced from various institutions. ISIC represents the largest and the most commonly used
438 dataset for the development of dermatology AI [27]. We downloaded 71,671 images in total from all of the ISIC
439 collections, including ISIC Challenge datasets 2016, 2017, 2018, 2019, and 2020 [21–26]. The images have diagnostic
440 attributes, including binary malignancy versus benign labels and 27 granular disease labels. For per-institution
441 analysis, we grouped images by data sources based on the attribution column in the metadata. We selected the
442 two largest cohorts: the Department of Dermatology at the Medical University of Vienna (9,873 samples) and the
443 Department of Dermatology at the Hospital Clínic de Barcelona (12,302 samples).

444 Fitzpatrick17k

445 Since the PubMed and textbook datasets contain clinical (*i.e.*, non-dermoscopic) images, for evaluation purposes,
446 we required additional clinical images with ground-truth annotations. As the first of these datasets, we chose Fitz-
447 patrick17k[52], which contains dermatological images collected from online dermatology atlases accompanied by disease
448 annotations and Fitzpatrick skin type labels. To reduce the impact of artifacts in the images, we filtered the dataset

449 to exclude images with visible patient clothing, visible anatomy (*e.g.*, fingers, ears, eyes, etc.), or other elements
450 except for lesions and background skin. To filter these images, we first manually annotated 10% of the full dataset
451 (1,657 of 16,577 images), marking each image as include or exclude, then trained a machine-learning model to classify
452 the remaining 90% of images. In particular, we fine-tuned a DenseNet-121[55] (pre-trained on ImageNet) to predict
453 exclusions using our 80% of our hand-labeled data, then chose an operating point to maximize the F1 score (maximum
454 = 0.81) on the remaining 20% of the hand-labeled data. This filtering resulted in a total of 4,951 images, incorporating
455 both those that passed the classifier’s screening and 462 images from our hand-labeled set. Fitzpatrick17k contains
456 near duplicate images with slight differences in angle or cropping; we filter for duplicate images to prevent overlap
457 between the train and test sets for the concept bottleneck model experiments. To measure the distance between
458 images, we obtained the 50 principal components of the embedding of the penultimate layer of the EfficientNetV2-S
459 model (pre-trained on ImageNet) [53]. Then, we calculated cosine similarity between the 50 principal components of
460 EfficientNet embedding. To rigorously filter out duplicates, we used a loose threshold (cosine similarity = 0.9) and
461 manually identified any false positives. In total, among the 4,951 images in Fitzpatrick17k “clean” set, we identified
462 523 sets of duplicate images, with some sets containing up to 6 duplicates. When selecting which images to keep
463 among the duplicates, we prioritized keeping those images with SkinCon annotations. After this filtering, we had a
464 total of 4,386 images. Lastly, we excluded 62 images that were marked as ‘Do not consider this image’ (*i.e.*, images of
465 low quality or considered not appropriate) in the SkinCon dataset. This led to a final dataset containing 4,324 images.

466 Additionally, for melanoma prediction tasks, amongst the 113 fine-grained diagnosis labels, we further refined the
467 data to include only melanomas and melanoma look-alikes, such that the data mirrors a well-defined clinical task. In
468 line with the disease filtering criteria outlined by Degraeve et al. [43], we included melanomas, benign melanocytic
469 lesions, seborrheic keratoses, and dermatofibromas, resulting in a total of 500 images.

470 DDI

471 As a second set of clinical images with ground truth labels, we chose the Diverse Dermatology Images (DDI) dataset.
472 DDI contains 656 clinical images of diverse skin tones, obtained from Stanford Clinics [30], accompanied by anno-
473 tations of Fitzpatrick skin type and histopathologically proven diagnoses. Again, we excluded 20 images that were
474 marked as ‘Do not consider this image’ in the SkinCon dataset, resulting in the final dataset of 636 images. For
475 melanoma prediction tasks, we narrowed the dataset to include only melanomas and melanoma look-alikes, resulting
476 in a total of 275 images, in accordance with the approach taken by Degraeve et al. [43]. Among the 78 fine-grained
477 diagnosis labels in DDI, the melanoma category comprises the general label “melanoma” as well as the more spe-
478 cific labels acral-lentiginous melanoma, melanoma *in situ*, and nodular melanoma. Melanoma look-alikes consist of
479 acral melanotic macule, atypical spindle cell nevus of reed, benign keratosis, blue nevus, dermatofibroma, dysplastic
480 nevus, epidermal nevus, hyperpigmentation, keloid, inverted follicular keratosis, melanocytic nevi, nevus lipomatosus
481 superficialis, pigmented spindle cell nevus of reed, seborrheic keratosis, irritated seborrheic keratosis, and solar lentigo.

482 SkinCon

483 SkinCon is at present the most comprehensive dataset on dermatology concepts [1]. The dataset features 48 concepts,
484 curated by board-certified dermatologists, that are frequently used to describe skin lesion attributes such as shape,
485 size, color, and texture. The dermatologists manually annotated the ground-truth labels for these concepts on 3,230
486 images from the Fitzpatrick17k dataset, which originally consisted of 16,577 images, and all 656 images from the DDI
487 dataset. Of the 4,324 images in the Fitzpatrick17k dataset we obtained after filtering, 1,009 had SkinCon annotations.
488 Among the 500 images in the dataset used for the melanoma prediction task, 95 had annotations.

489 MONET

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

494 Architecture

495 We use the vision transformer architecture, ViT-L/14, as our image encoder [56]. This encoder takes an input image
496 of size 224 x 224 and outputs a 768-dimensional embedding. For the text encoder, we use a transformer architecture
497 with 12 self-attention layers. It takes tokenized text with a maximum limit of 77 tokens as input and outputs a 768-
498 dimensional embedding. We use the same architecture as CLIP [20] to take advantage of the weights from pre-trained
499 models.

500 Preprocessing

501 To meet the input requirements for the encoder architectures, we process image and text inputs as follows. Each input
502 image is re-sized and center-cropped to be 224x224 dimensions. It is then normalized using the mean and standard
503 deviation used in CLIP [20]. Throughout the training phase, we applied standard data augmentation steps instead,
504 such as random resized crops, vertical flips, horizontal flips, and color jittering for brightness, contrast, and saturation.
505 For each input text, we apply tokenization using lower-cased byte pair encoding [57]. In cases the text encountered
506 during training was longer than the text encoder’s maximum token limit of 77, we split the text into sentences and
507 chose half of them from the beginning. We repeated this process until the token count was reduced to fewer than 77.

508 Training

509 We use cosine similarity as the distance metric. Both encoders are jointly trained to maximize the cosine similarity
510 between the image and text embeddings of the correct pairs while minimizing the cosine similarity between the
511 embeddings of incorrect pairings. To this end, we use a symmetric cross-entropy loss on cosine similarity scores; after
512 calculating the cosine similarities between embeddings, we scale them by a temperature parameter λ and normalized
513 them into a probability distribution with the softmax function. The temperature parameter λ was also updated during
514 training. We optimize the loss using the Adam optimizer [58] with a cosine learning rate schedule for 10 epochs. This
515 implementation detail follows that of CLIP [20].

516 For hyper-parameter tuning, we split the dataset into training and validation sets and find hyper-parameters that
517 result in the best validation loss; we use validation loss for hyper-parameter tuning because there is no large-scale
518 ground truth label for evaluating concept generation performance. We tune the hyper-parameter of batch size (128,
519 256, 512, 1024) and learning rate (1e-3, 1e-4, 1e-5, 1e-6). We find that the larger batch size results in lower validation
520 loss until a batch size of 512 is reached. We also find that the learning rate of 1e-5 leads to the lowest validation loss.
521 Using the tuned hyper-parameters, we train the model on the whole dataset for 10 epochs. We use 6 Nvidia A40
522 GPUs with data parallelism. Model training takes 1 hour and 40 minutes.

523 Automatic concept generation

524 During the training procedure of image and text encoders, an image and a text from the same pair are forced to be
525 close to each other in the embedding space, while ones from different pairs are forced to be far apart. After training,
526 MONET can measure the proximity between an image and any arbitrary text. We use this capability to automatically
527 generate concepts for images.

528 To generate a concept c for a given batch of N images I_1, I_2, \dots, I_N , we first compute the image embeddings
529 $f_{\text{image}}(I_1), f_{\text{image}}(I_2), \dots, f_{\text{image}}(I_N)$ using the image encoder f_{image} . We also compute the concept prompt embedding
530 $f_{\text{text}}(T_c)$ and reference prompt embedding $f_{\text{text}}(T_r)$ using the text encoder, where T_c is a concept prompt (*e.g.*, “This
531 is a skin image of { }”) and T_r is a reference prompt (*e.g.*, “This is a skin image of”). Supplementary Table 4 shows the
532 terms used for each concept for filling templates. Next, we calculate the cosine similarity between image embeddings
533 and prompt embeddings. When multiple terms are used for each concept, we calculate the cosine similarity for each
534 term and average them. Finally, we obtain concept presence score $p_{i,c}$ that represents the degree to which a concept
535 is present in the image as follows:

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

536 where $\text{sim}_{\cos}(\cdot)$ is the cosine similarity score between image embeddings and text embeddings, $\text{sim}_{\cos}(I_i, T_c) =$
537 $\frac{f_{\text{image}}(I_i)^T f_{\text{text}}(T_c)}{\|f_{\text{image}}(I_i)\| \|f_{\text{text}}(T_c)\|}$, and λ is the temperature parameter learned during the training. We normalize by reference prompt
538 to remove the effect of templates being used. Further, we use multiple templates to minimize the effects of templates.
539 For clinical images, we used templates: “This is skin image of { }”, “This is dermatology image of { }”, and “This is
540 image of { }”. For dermoscopic images, we used the templates “This is dermatoscopy of { }” and “This is dermoscopy
541 of { }”. We use concept presence scores averaged across different templates in the end.

542 Quantitative evaluation

543 We use 1,645 images with SkinCon labels from Fitzpatrick17k and DDI datasets for the task of predicting SkinCon
544 concepts. We use 4,324 images from Fitzpatrick17k and 636 images from DDI datasets for the task of predicting disease
545 labels, respectively. We compare the performance of MONET’s concept generation to that of a supervised learning
546 approach, training a ResNet-50 model [28], and to that of a pre-existing image-text model CLIP [20]. MONET and
547 CLIP do not require additional training to perform these tasks; the output from MONET and CLIP models is obtained

548 via the automatic concept generation procedure described above. In contrast, we need to train a supervised learning
549 model. We train the model using a standard training recipe as follows. We initialize the model using ImageNet pre-
550 trained weights. We then replace the last layer of the model with a new MLP layer that matches the dimension of the
551 prediction target; for SkinCon concepts, we train each concept one by one (the dimension is 1), and for disease labels,
552 we train disease labels considered all at once (the dimension is 113 for Fitzpatrick17k and 78 for DDI). We finally
553 train the model using cross-entropy loss for 20 epochs. We use the Adam optimizer [58] with a ReduceLROnPlateau
554 learning rate scheduler implemented in Pytorch (ver. 1.13.0); the initial learning rate is 1e-3, and it is reduced based
555 on validation loss with the patience parameter of 2. Also, we use EarlyStopper implemented in PyTorch, which stops
556 the training when the validation loss does not improve 5 times. The available data for each task is split into train/test
557 sets with a ratio of 4:1, and 20% of the train set is left for calculating validation loss. While for MONET and CLIP,
558 we calculate AUROC across all available samples in one go, for the ResNet-50 model, we repeat the evaluation with
559 20 different train-test splits and calculate the average AUROC for each target to leverage all samples fully.

560 Data auditing

561 Concept differential analysis

562 MONET’s ability to map images and texts onto the co-embedding space enables us to describe the different char-
563 acteristics between two sets of images in natural language. Assume we have two sets of images, denoted as $\mathbf{I}_+ =$
564 $\{I_1, I_2, \dots, I_{N_+}\}$ and $\mathbf{I}_- = \{I_1, I_2, \dots, I_{N_-}\}$, and a list of concepts we want to investigate $[c_1, c_2, \dots, c_{N_c}]$. We
565 first obtain the prototype embedding of each image set by computing an average of normalized image embeddings,
566 $m_+ = \sum_{I_i \in \{\mathbf{I}_+\}} \frac{f_{\text{image}}(I_i)}{\|f_{\text{image}}(I_i)\|}$ and $m_- = \sum_{I_i \in \{\mathbf{I}_-\}} \frac{f_{\text{image}}(I_i)}{\|f_{\text{image}}(I_i)\|}$. Then, we calculate the displacement vector from m_-
567 to m_+ by subtracting out the two prototype embedding $m_\Delta = m_+ - m_-$. Finally, we get a differential concept
568 expression score by computing the dot product between the prototype and normalized embeddings of concept prompt
569 $C_{\Delta, i} = m_\Delta^T \cdot \frac{f_{\text{text}}(T_i)}{\|f_{\text{text}}(T_i)\|}$. This score measures how much more each concept is differentially expressed in S_+ than in S_- .
570 A similar technique for converting a set of images to text has been previously used by Eyuboglu et al. [13].

571 Benchmark analysis

572 To perform a benchmark study on concept differential analysis, we construct synthetic data using ground-truth concept
573 labels in the SkinCon dataset. For each concept in SkinCon, we create a dataset split into two groups: one with 100
574 images, many of which are associated with the concept, and another with 100 images, many of which are not associated
575 with the concept. We use the noise level parameter to control the degree to which the concept is correlated with the
576 grouping; it indicates the probability that images are randomly sampled from the opposite group. We run simulations
577 20 times for each combination of parameters with different random seeds.

578 Model auditing

579 Model auditing with MONET

580 We can use MONET to automatically detect semantically meaningful medical concepts that lead to model errors.
581 Model auditing with MONET (MA-MONET) starts by sorting images from a test set into groups based on their
582 visual similarity. To this end, we run the K-means clustering algorithm implemented in the scikit-learn Python
583 package (ver. 1.2.2) [54, 59]. We use 50 principal components of the embedding of the penultimate layer of the
584 EfficientNetV2-S model (pre-trained on ImageNet) [53] as clustering features. Next, we calculate the accuracy across
585 all samples and also per cluster; for thresholding the probability output from the trained classifier, we choose an
586 operating point that maximizes the F1 score. Following this, we identify medical concepts for the “low-performing”
587 cluster; we define low-performing clusters as ones with accuracy lower than overall accuracy. Each low-performing
588 cluster is compared to a high-performing counterpart containing similar images to understand what differentiates
589 them; among the clusters that perform better than the overall accuracy, we choose one whose centroid is closest in
590 Euclidean distance to the low-performing cluster. We conduct a concept differential analysis between the high and
591 low-performing clusters to pinpoint concepts that are more presented in the low-performing cluster. If two visually
592 similar clusters (one high-performing, the other low-performing) differ in terms of a few concepts, these differing
593 concept terms can be hypothesized as leading to high error rates. We then filter out concepts with a concept presence
594 score below 0.5 in the low-performing. Finally, we obtain a ranked list of medical concepts identified by MONET that
595 differentiate the two clusters.

596 Benchmark analysis

597 For benchmarking analysis, we use a situation where the ground truth (*i.e.*, the concepts leading to model error) is
598 already known. We create a training dataset with spurious correlations from the Fitzpatrick17k and DDI datasets:
599 500 malignant images that feature a particular SkinCon concept, while the 500 benign images do not. For the test set,
600 we reverse the correlation; 500 sampled benign images have the SkinCon concept, while 500 sampled malignant images
601 do not. For concepts of spurious correlation, we use 5 concepts that remain after filtering out concepts with fewer
602 than 30 samples in each category required for creating the confounded training and test sets (*i.e.*, malignant-with
603 concept, malignant-without concept, benign-with concept, and benign-without concept): “crust”, “hyperpigmenta-
604 tion”, “plaque”, “erythema”, and “papule”. For each of the 5 selected concepts, we repeat this analysis 20 times
605 with different random seeds varying the training and test sets. Consequently, we conduct analysis for a total of 100
606 settings. In addition to the concept we intentionally introduce as a confounder, there are other concepts that also
607 inadvertently become confounders. For example, when we had “papule” to be associated with malignancy in the train
608 set, “plaque” was associated with benign images in the training set. In such cases, we define all of them as the ground
609 truth. On average, there are two concepts across all 100 test settings. We consider the spurious correlations are
610 recovered if the top-N concepts identified by MA-MONET include at least one of these ground truth concepts across
611 all low-performing clusters.

612 Building inherently interpretable neural network

613 Concept Bottleneck Model

614 Concept Bottleneck Models (CBMs) [15] are inherently interpretable models that identify the importance of each con-
615 cept for the classifier’s prediction. They use a bottleneck layer to extract compact and discriminative representations
616 of the input data. The bottleneck layer, typically composed of a small number of units, imposes a constraint on
617 the amount of information transmittable through the network, forcing it to make predictions by using interpretable
618 features that align well with the users’ expectations. This technique can be used to reduce the dimensionality of the
619 data and improve the efficiency of the model while preserving its predictive power. CBMs have been successfully
620 applied to a wide range of tasks, including image and video classification [60, 61], natural language processing [62],
621 and being applied in different medical settings [1, 63, 64]. However, a caveat is that these models need a large set of
622 concept annotations to perform well, and collecting these labels is laborious and time intensive.

623 MONET lets us automatically generate concepts for images that can be used to scale to a large concept dataset
624 with an arbitrary number of concepts. Specifically, MONET helps to create the bottleneck layer, denoted by $b_c : \mathcal{X}_{\text{image}} \rightarrow \mathbb{R}^{N_c}$, that maps an input image I_i to a vector of dimension N_c , the number of concepts, where each
625 dimension corresponds to one of the N_c concepts. An interpretable linear model is then trained on the prediction
626 target to get importance scores for each concept corresponding to the trained model weights.

627 To create the bottleneck layer, we start with a concept list $[c_1, c_2, \dots, c_{N_c}]$, chosen with guidance from our derma-
628 tologist collaborators, containing concepts that are predictive of the target. Ideally, the bottleneck layer is binarized
629 using the concept labels. However, we lack access to the concept annotations, and thresholding the similarity score of
630 each concept with the input image is non-trivial. Instead, for each concept c_j , we curate a set of reference concepts
631 $[r_{j1}, r_{j2}, \dots, r_{jN_j}]$ where N_j is the number of reference concepts for concept c_j . Each reference concept is selected
632 such that it is sufficiently far from the concept of interest in the representation space while being closer to the other
633 reference concepts. We do this by choosing antonyms of the concept of interest as the reference concepts.

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

$$638 \quad p'_{i,c_j} = \frac{\exp(\text{sim}_{\cos}(I_i, c_j)/\lambda)}{\exp(\text{sim}_{\cos}(I_i, c_j)/\lambda) + \sum_k \exp(\text{sim}_{\cos}(I_i, r_{jk})/\lambda)} \quad (2)$$

639 where $\text{sim}_{\cos}(\cdot)$ is the cosine similarity score obtained using MONET, and λ is a temperature parameter used to
640 magnify the differences in similarity scores. λ is manually tuned to the value that performs the best on the train
641 set. Once the bottleneck layer is created, we train a simple linear classifier on the prediction target using stochastic
642 gradient descent. Specifically, for a classifier f and a given sample $x \in \mathbb{R}^{N_c}$, the prediction obtained is $w^T f(x) + b$. We
643 apply L1 regularization to favor sparsity in the trained model weights and make the model more interpretable. Once
644 the linear classifier is trained, the learned weights w can be analyzed to understand the importance of each concept
645 for the prediction target.

646 To demonstrate the efficacy of MONET, we use two different prediction targets: (1) Melanoma vs. Melanoma look-
647 alike, and (2) Malignant vs. Benign. We differentiate between these two targets since all melanomas are malignant,
648 but not all malignant lesions are melanoma. For this experiment, we use the clean Fitzpatrick 17k[52] and DDI[30]
649 datasets. We use 80% of the data for training and reserving the rest for testing. To create the bottleneck layer, we use
650 11 concepts that are known to be correlated to the prediction targets; specifically, we use the ABCDEs of melanoma
651 [46] as a guideline to compile the list of concepts for the bottleneck layer. Supplementary Table 5 lists these concepts
652 along with the reference concepts used for normalization. MONET’s ability to automatically generate concepts for
653 images lets us easily add more data or concepts as needed without any manual annotations.

654 We compare MONET+CBM to several other baseline methods of obtaining target predictions from input images:

- 655 • **Vanilla CLIP+CBM:** We use an out-of-the-box CLIP model to create the bottleneck layer and trains a linear
656 classifier, similar to MONET+CBM. The only difference is that the vanilla CLIP model is not fine-tuned on
657 dermatology images and thus lacks the context of the setting in which we run the experiment; as a result, it
658 cannot adequately capture the semantic differences between technical dermatology terms.
- 659 • **Supervised:** We train a deep learning model using the standard fully supervised approach without incorporating
660 concepts. We use ResNet-50 pre-trained on the ImageNet where the last classification head was replaced to match
661 the dimension of the prediction target. We train the model end-to-end to classify the input images into the target
662 classes. The implementation details are the same as described in the Qualitative evaluation subsection under
663 Automatic concept generation. We only change the maximum training epoch from 20 to 50
- 664 • **Linear Probing** We use the representation of the penultimate layer of ResNet-50 pre-trained on ImageNet
665 as the input for a linear model. The difference with supervised is that during the training, the backbone of
666 ResNet-50 is frozen.
- 667 • **Manual Labeling** We use the SkinCon dataset [1], which applies concept annotations covering 48 concepts for
668 3230 images from the Fitzpatrick 17k dataset to create the bottleneck layer. These concepts were chosen by two
669 board-certified dermatologists considering the clinical descriptor terms used to describe skin lesions.

670 Data availability

671 PMC Open Access Subset is publicly available from <https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/>.
672 Evaluation datasets are all publicly available and can be accessed from: ISIC ([https://challenge.isic-archive.c](https://challenge.isic-archive.com/data/)
673 [om/data/](https://github.com/mattgroh/fitzpatrick17k)), Fitzpatrick17k (<https://github.com/mattgroh/fitzpatrick17k>), and DDI(<https://stanfordaimi.a>
674 zurewebsites.net/datasets/35866158-8196-48d8-87bf-50dca81df965).

675 Code availability

676 The code used in our analysis is available at <https://github.com/suinleelab/MONET>. It includes various scripts for
677 data collection and preprocessing, training the MONET model, and conducting benchmark studies. Also, it provides
678 the MONET model weights.

679 Acknowledgements

680 We thank Chris Lin and other people in the Lee Lab for helpful discussions.

681 Funding

682 C.K., S.U.G., A.J.D., and S.-I.L. were supported by the National Science Foundation (CAREER DBI-1552309 and
683 DBI-1759487) and the National Institutes of Health (R35 GM 128638 and R01 AG061132). C.K. was supported by
684 the Asan Foundation Biomedical Science Scholarship. R.D. was supported by the National Institutes of Health (5T32
685 AR007422-38) and the Stanford Catalyst Program.

686 Ethics declarations

687 Competing interests

688 R.D. reports fees from L’Oreal, Frazier Healthcare Partners, Pfizer, DWA, and VisualDx for consulting; stock options
689 from MDAcne and Revea for advisory board; and research funding from UCB.

690 References

- 691 1. Daneshjou, R., Yuksekgonul, M., Cai, Z. R., Novoa, R. & Zou, J. Y. *SkinCon: A skin disease dataset densely an-*
692 *notated by domain experts for fine-grained debugging and analysis in Advances in Neural Information Processing*
693 *Systems* (eds Koyejo, S. et al.) **35** (Curran Associates, Inc., 2022), 18157–18167.
- 694 2. Goel, K., Gu, A., Li, Y. & Ré, C. *Model Patching: Closing the Subgroup Performance Gap with Data Augmentation*
695 *in 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*
696 (OpenReview.net, 2021).
- 697 3. Sagawa*, S., Koh*, P. W., Hashimoto, T. B. & Liang, P. *Distributionally Robust Neural Networks in International*
698 *Conference on Learning Representations* (2020).
- 699 4. Rajpurkar, P. et al. *MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs* May 22,
700 2018. arXiv: 1712.06957[physics].
- 701 5. Oakden-Rayner, L., Dunnmon, J., Carneiro, G. & Re, C. *Hidden stratification causes clinically meaningful failures*
702 *in machine learning for medical imaging in Proceedings of the ACM Conference on Health, Inference, and Learning*
703 *ACM CHIL ’20: ACM Conference on Health, Inference, and Learning* (ACM, Toronto Ontario Canada, Apr. 2,
704 2020), 151–159. ISBN: 978-1-4503-7046-2.
- 705 6. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal.
706 *Nature Machine Intelligence* **3**. Number: 7 Publisher: Nature Publishing Group, 610–619. ISSN: 2522-5839 (July
707 2021).
- 708 7. Pianykh, O. S. et al. Continuous Learning AI in Radiology: Implementation Principles and Early Applications.
709 *Radiology* **297**. PMID: 32840473, 6–14. eprint: <https://doi.org/10.1148/radiol.2020200038> (2020).
- 710 8. Feng, J. et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of
711 AI algorithms in healthcare. *npj Digital Medicine* **5**, 66. ISSN: 2398-6352 (May 2022).
- 712 9. Vokinger, K. N., Feuerriegel, S. & Kesselheim, A. S. Continual learning in medical devices: FDA’s action plan
713 and beyond. *The Lancet Digital Health* **3**. Publisher: Elsevier, e337–e338. ISSN: 2589-7500 (June 1, 2021).
- 714 10. Kim, B. et al. *Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors*
715 *(TCAV) in Proceedings of the 35th International Conference on Machine Learning* International Conference on
716 Machine Learning. ISSN: 2640-3498 (PMLR, July 3, 2018), 2668–2677.
- 717 11. Crabbé, J. & van der Schaar, M. *Concept Activation Regions: A Generalized Framework For Concept-Based*
718 *Explanations in NeurIPS* (2022).
- 719 12. Abid, A., Yuksekgonul, M. & Zou, J. *Meaningfully debugging model mistakes using conceptual counterfactual*
720 *explanations in Proceedings of the 39th International Conference on Machine Learning* International Conference
721 on Machine Learning. ISSN: 2640-3498 (PMLR, June 28, 2022), 66–88.
- 722 13. Eyuboglu, S. et al. *Domino: Discovering Systematic Errors with Cross-Modal Embeddings in The Tenth Interna-*
723 *tional Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022* (OpenReview.net,
724 2022).
- 725 14. Chung, Y., Kraska, T., Polyzotis, N., Tae, K. & Whang, S. Automated Data Slicing for Model Validation: A Big
726 Data - AI Integration Approach. *IEEE Transactions on Knowledge and Data Engineering* **32**, 2284–2296. ISSN:
727 1558-2191 (Dec. 2020).
- 728 15. Koh, P. W. et al. *Concept Bottleneck Models in Proceedings of the 37th International Conference on Machine*
729 *Learning* International Conference on Machine Learning. ISSN: 2640-3498 (PMLR, Nov. 21, 2020), 5338–5348.
- 730 16. *Post-hoc Concept Bottleneck Models in The Eleventh International Conference on Learning Representations,*
731 *ICLR 2023, Rwanda, May 1-5, 2023* (2023).
- 732 17. Mendonça, T., Ferreira, P. M., Marques, J. S., Marcal, A. R. & Rozeira, J. *PH 2-A dermoscopic image database*
733 *for research and benchmarking in 2013 35th annual international conference of the IEEE engineering in medicine*
734 *and biology society (EMBC)* (2013), 5437–5440.

- 735 18. Kawahara, J., Daneshvar, S., Argenziano, G. & Hamarneh, G. Seven-point checklist and skin lesion classification
736 using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics* (2018).
- 737 19. Nevitt, M., Felson, D. & Lester, G. The osteoarthritis initiative. *Protocol for the cohort study* **1** (2006).
- 738 20. Radford, A. *et al.* *Learning Transferable Visual Models From Natural Language Supervision* in *Proceedings of*
739 *the 38th International Conference on Machine Learning* International Conference on Machine Learning. ISSN:
740 2640-3498 (PMLR, July 1, 2021), 8748–8763.
- 741 21. Gutman, D. *et al.* *Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium*
742 *on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC)* May 4, 2016.
743 arXiv: 1605.01397[cs].
- 744 22. Codella, N. C. F. *et al.* *Skin lesion analysis toward melanoma detection: A challenge at the 2017 International*
745 *symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC) in 2018*
746 *IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* 2018 IEEE 15th International Sympo-
747 sium on Biomedical Imaging (ISBI 2018). ISSN: 1945-8452 (Apr. 2018), 168–172.
- 748 23. Codella, N. *et al.* *Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the Interna-*
749 *tional Skin Imaging Collaboration (ISIC)* Mar. 29, 2019. arXiv: 1902.03368[cs].
- 750 24. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic
751 images of common pigmented skin lesions. *Scientific Data* **5**. Number: 1 Publisher: Nature Publishing Group,
752 180161. ISSN: 2052-4463 (Aug. 14, 2018).
- 753 25. Combalia, M. *et al.* *BCN20000: Dermoscopic Lesions in the Wild* Aug. 30, 2019. arXiv: 1908.02288[cs, eess].
- 754 26. Rotemberg, V. *et al.* A patient-centric dataset of images and metadata for identifying melanomas using clinical
755 context. *Scientific Data* **8**. Number: 1 Publisher: Nature Publishing Group, 34. ISSN: 2052-4463 (Jan. 28, 2021).
- 756 27. Jones, O. T. *et al.* Artificial intelligence and machine learning algorithms for early detection of skin cancer in
757 community and primary care settings: a systematic review. *The Lancet Digital Health* **4**. Publisher: Elsevier,
758 e466–e476. ISSN: 2589-7500 (June 1, 2022).
- 759 28. He, K., Zhang, X., Ren, S. & Sun, J. *Deep Residual Learning for Image Recognition* in *2016 IEEE Conference*
760 *on Computer Vision and Pattern Recognition (CVPR)* 2016 IEEE Conference on Computer Vision and Pattern
761 Recognition (CVPR) (IEEE, Las Vegas, NV, USA, June 2016), 770–778. ISBN: 978-1-4673-8851-1.
- 762 29. Tiu, E. *et al.* Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised
763 learning. *Nature Biomedical Engineering*. Publisher: Nature Publishing Group, 1–8. ISSN: 2157-846X (Sept. 15,
764 2022).
- 765 30. Daneshjou, R. *et al.* Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science*
766 *Advances* **8**, eabq6147 (2022).
- 767 31. Janizek, J. D., Erion, G., DeGrave, A. J. & Lee, S.-I. *An Adversarial Approach for the Robust Classification of*
768 *Pneumonia from Chest Radiographs* in *Proceedings of the ACM Conference on Health, Inference, and Learning*
769 (Association for Computing Machinery, Toronto, Ontario, Canada, 2020), 69–79. ISBN: 9781450370462.
- 770 32. Bissoto, A., Fornaciali, M., Valle, E. & Avila, S. *(De) Constructing Bias on Skin Lesion Datasets* in *2019*
771 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 2019 IEEE/CVF
772 Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (IEEE, Long Beach, CA, USA,
773 June 2019), 2766–2774. ISBN: 978-1-72812-506-0.
- 774 33. Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J. & Yap, M. H. Analysis of the ISIC image
775 datasets: Usage, benchmarks and recommendations. *Medical Image Analysis* **75**, 102305. ISSN: 1361-8415 (Jan. 1,
776 2022).
- 777 34. Winkler, J. K. *et al.* Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Perform-
778 ance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatology* **155**,
779 1135–1141. ISSN: 2168-6068 (Oct. 1, 2019).
- 780 35. Navarrete-Dechent, C., Liopyris, K. & Marchetti, M. A. Multiclass Artificial Intelligence in Dermatology: Progress
781 but Still Room for Improvement. *Journal of Investigative Dermatology* **141**, 1325–1328. ISSN: 0022-202X (2021).
- 782 36. Singh, C., Balakrishnan, G. & Perona, P. *Matched sample selection with GANs for mitigating attribute confound-*
783 *ing* Mar. 24, 2021. arXiv: 2103.13455[cs, stat].
- 784 37. Leming, M., Das, S. & Im, H. Construction of a confounder-free clinical MRI dataset in the Mass General Brigham
785 system for classification of Alzheimer’s disease. *Artificial Intelligence in Medicine* **129**, 102309. ISSN: 0933-3657
786 (July 1, 2022).

- 787 38. Zhao, Q., Adeli, E. & Pohl, K. M. Training confounder-free deep learning models for medical applications. *Nature*
788 *Communications* **11**. Number: 1 Publisher: Nature Publishing Group, 6010. ISSN: 2041-1723 (Nov. 26, 2020).
- 789 39. Zhu, J., Park, T., Isola, P. & Efros, A. A. *Unpaired Image-to-Image Translation Using Cycle-Consistent Ad-*
790 *versarial Networks* in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October*
791 *22-29, 2017* (IEEE Computer Society, 2017), 2242–2251.
- 792 40. Lundberg, S. M. & Lee, S.-I. *A Unified Approach to Interpreting Model Predictions* in *Proceedings of the 31st*
793 *International Conference on Neural Information Processing Systems* (Curran Associates Inc., Long Beach, Cali-
794 fornia, USA, 2017), 4768–4777. ISBN: 9781510860964.
- 795 41. Sundararajan, M., Taly, A. & Yan, Q. *Axiomatic Attribution for Deep Networks* in *Proceedings of the 34th*
796 *International Conference on Machine Learning - Volume 70* (JMLR.org, Sydney, NSW, Australia, 2017), 3319–
797 3328.
- 798 42. Selvaraju, R. R. *et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization* in
799 *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 618–626.
- 800 43. DeGrave, A. J., Cai, Z. R., Janizek, J. D., Daneshjou, R. & Lee, S.-I. Dissection of medical AI reasoning processes
801 via physician and generative-AI collaboration. *medRxiv*. eprint: [https://www.medrxiv.org/content/early/2](https://www.medrxiv.org/content/early/2023/05/16/2023.05.12.23289878.full.pdf)
802 [023/05/16/2023.05.12.23289878.full.pdf](https://www.medrxiv.org/content/early/2023/05/16/2023.05.12.23289878.full.pdf) (2023).
- 803 44. Han, S. S. *et al.* The degradation of performance of a state-of-the-art skin image classifier when applied to
804 patient-driven internet search. *Scientific Reports* **12**. Number: 1 Publisher: Nature Publishing Group, 16260.
805 ISSN: 2045-2322 (Sept. 28, 2022).
- 806 45. Navarrete-Dechent, C. *et al.* Automated Dermatological Diagnosis: Hype or Reality? *Journal of Investigative*
807 *Dermatology* **138**. Publisher: Elsevier, 2277–2279. ISSN: 0022-202X, 1523-1747 (Oct. 1, 2018).
- 808 46. Rigel, D. S., Friedman, R. J., Kopf, A. W. & Polsky, D. ABCDE—an evolving concept in the early detection of
809 melanoma. *Archives of dermatology* **141**, 1032–1034 (2005).
- 810 47. Huang, Z., Bianchi, F., Yuksekogul, M., Montine, T. & Zou, J. *Leveraging medical Twitter to build a vi-*
811 *sual-language foundation model for pathology AI* Pages: 2023.03.29.534834 Section: New Results. Apr. 1, 2023.
- 812 48. Combalia, M. *et al.* Validation of artificial intelligence prediction models for skin cancer diagnosis using der-
813 moscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge. *The Lancet Digital Health*
814 **4**. Publisher: Elsevier, e330–e339. ISSN: 2589-7500 (May 1, 2022).
- 815 49. Daneshjou, R., Smith, M. P., Sun, M. D., Rotemberg, V. & Zou, J. Lack of Transparency and Potential Bias in
816 Artificial Intelligence Data Sets and Algorithms: A Scoping Review. *JAMA Dermatology* **157**, 1362–1369. ISSN:
817 2168-6068 (Nov. 1, 2021).
- 818 50. National Library of Medicine. *PMC Open Access Subset* [https://www.ncbi.nlm.nih.gov/pmc/tools/openft](https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/)
819 [list/](https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/) (2022).
- 820 51. Gamper, J. & Rajpoot, N. M. *Multiple Instance Captioning: Learning Representations From Histopathology*
821 *Textbooks and Articles* in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual,*
822 *June 19-25, 2021* (Computer Vision Foundation / IEEE, 2021), 16549–16559.
- 823 52. Groh, M. *et al. Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitz-*
824 *patrick 17k Dataset* in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*
825 *(CVPRW) 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*
826 *(IEEE, Nashville, TN, USA, June 2021)*, 1820–1828. ISBN: 978-1-66544-899-4.
- 827 53. Tan, M. & Le, Q. V. *EfficientNetV2: Smaller Models and Faster Training* in *Proceedings of the 38th International*
828 *Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event* (eds Meila, M. & Zhang, T.) **139**
829 (PMLR, 2021), 10096–10106.
- 830 54. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–
831 2830 (2011).
- 832 55. Huang, G., Liu, Z., Maaten, L. V. D. & Weinberger, K. Q. *Densely Connected Convolutional Networks* in *2017*
833 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE Computer Society, Los Alamitos,
834 CA, USA, July 2017), 2261–2269.
- 835 56. Dosovitskiy, A. *et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* in *9th In-*
836 *ternational Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021* (Open-
837 Review.net, 2021).

- 838 57. Sennrich, R., Haddow, B. & Birch, A. *Neural Machine Translation of Rare Words with Subword Units* in *Pro-*
839 *ceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*
840 (Association for Computational Linguistics, Berlin, Germany, Aug. 2016), 1715–1725.
- 841 58. Kingma, D. P. & Ba, J. *Adam: A Method for Stochastic Optimization* in *3rd International Conference on Learning*
842 *Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (eds Bengio,
843 Y. & LeCun, Y.) (2015).
- 844 59. Lloyd, S. Least squares quantization in PCM. *IEEE Transactions on Information Theory* **28**, 129–137 (1982).
- 845 60. Lanchantin, J., Wang, T., Ordonez, V. & Qi, Y. *General multi-label image classification with transformers* in
846 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), 16478–16488.
- 847 61. Jeyakumar, J. V. *et al.* Automatic Concept Extraction for Concept Bottleneck-based Video Classification. *arXiv*
848 *preprint arXiv:2206.10129* (2022).
- 849 62. Sun, X. *et al.* Interpreting deep learning models in natural language processing: A review. *arXiv preprint*
850 *arXiv:2110.10470* (2021).
- 851 63. Klimiene, U. *et al.* *Multiview Concept Bottleneck Models Applied to Diagnosing Pediatric Appendicitis* in *2nd*
852 *Workshop on Interpretable Machine Learning in Healthcare (IMLH)* (2022).
- 853 64. Wu, C., Parbhoo, S., Havasi, M. & Doshi-Velez, F. *Learning Optimal Summaries of Clinical Time-series with*
854 *Concept Bottleneck Models* in *Machine Learning for Healthcare Conference* (2022), 648–672.

A. Erythema



B. Bulla



C. Xerosis



D. Pustule

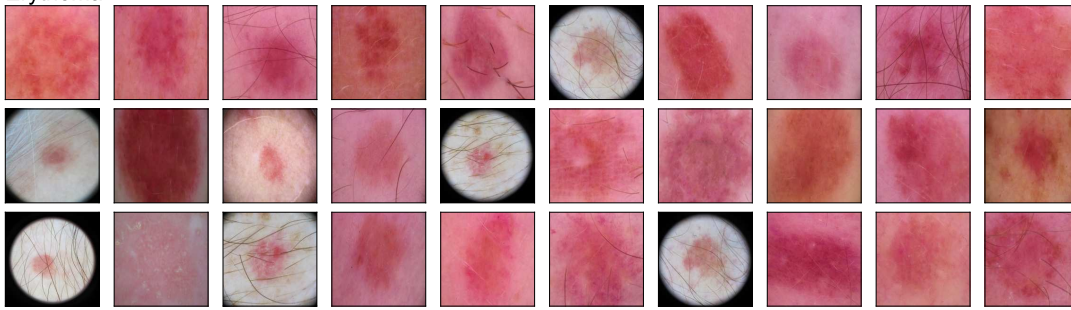


E. Ulcer

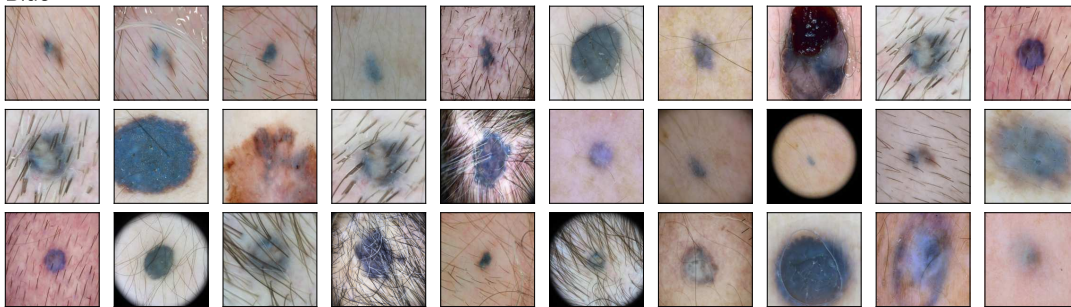


Supplementary Fig. 1 | Clinical images with high concept presence scores calculated using MONET. We show the top 30 images for each concept. (A) Erythema. (B) Bulla. (C) Xerosis. (D) Pustule. (E) Ulcer.

A. Erythema



B. Blue



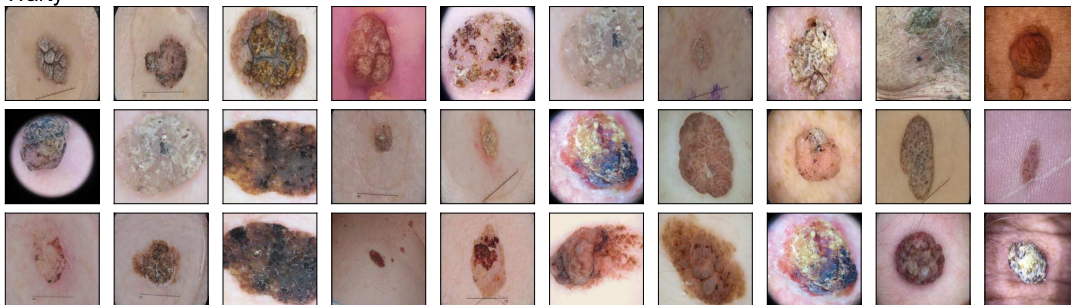
C. Nodule



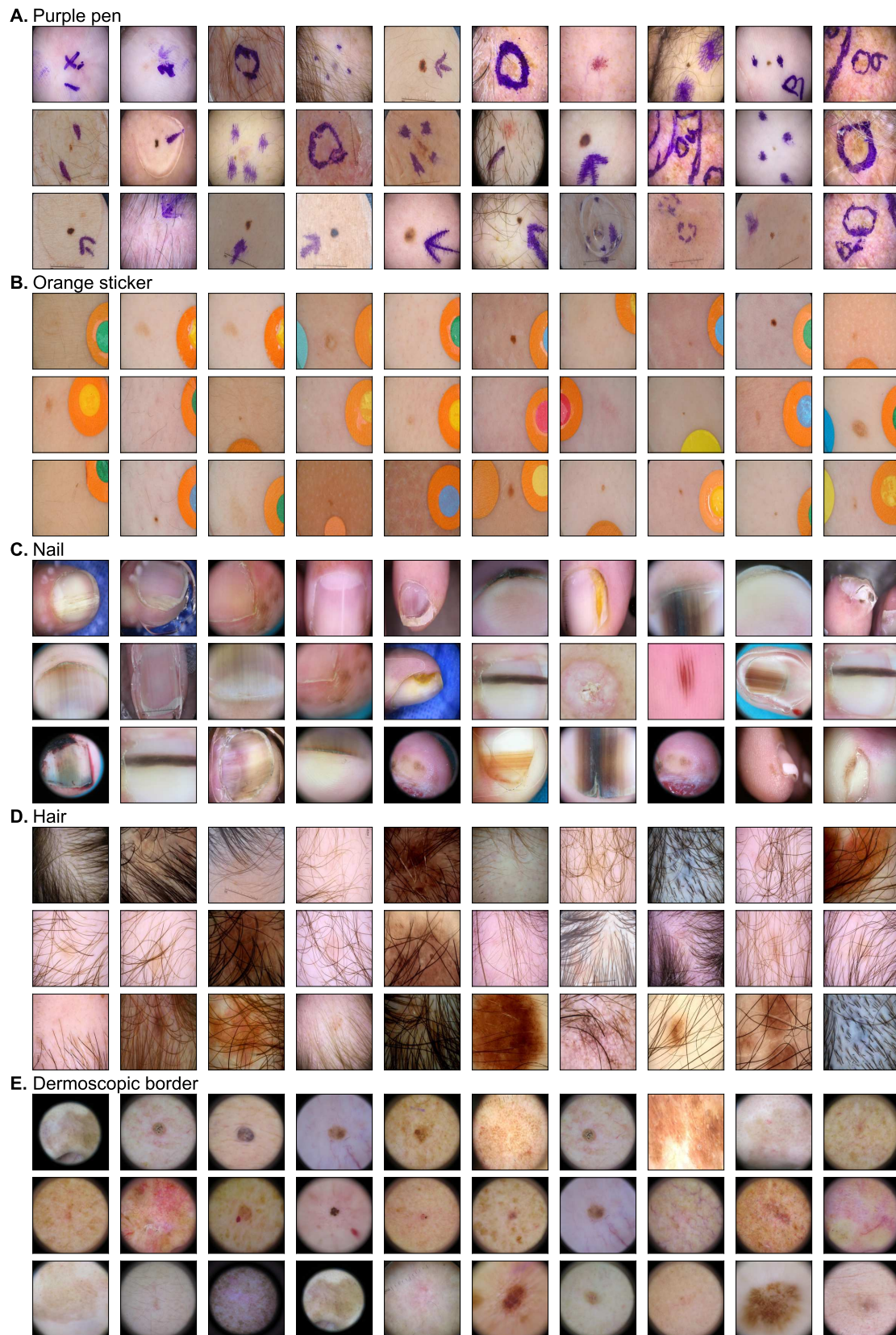
D. Ulcer



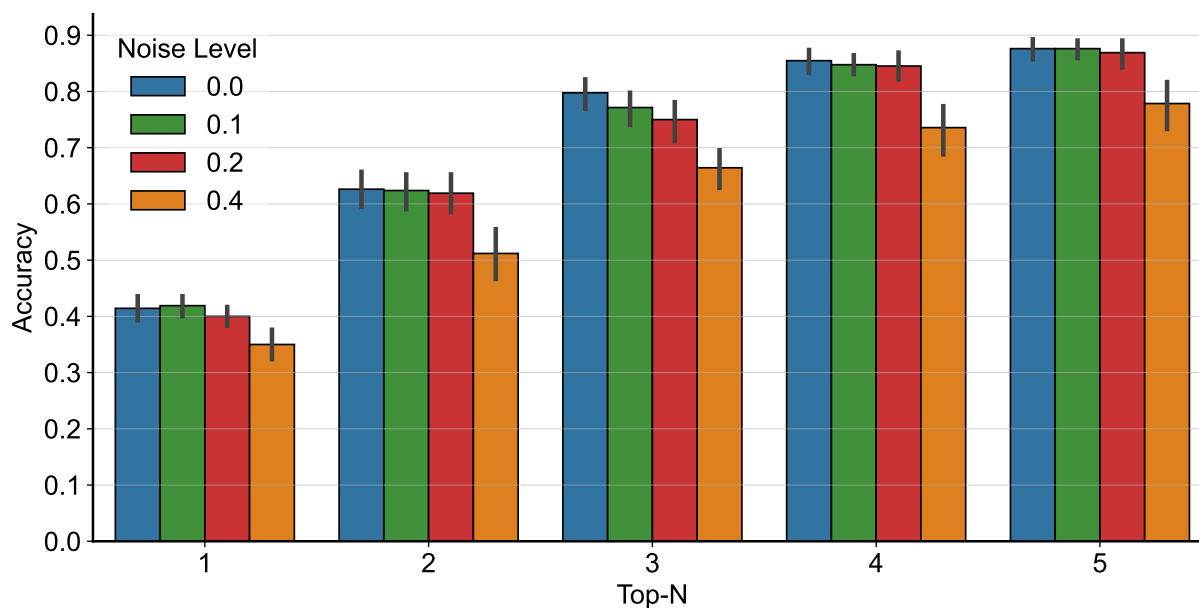
E. Warty



Supplementary Fig. 2 | Dermoscopic images with high concept presence scores calculated using MONET. We show the top 30 images for each concept. (A) Erythema. (B) Blue. (C) Nodule. (D) Ulcer. (E) Warty.



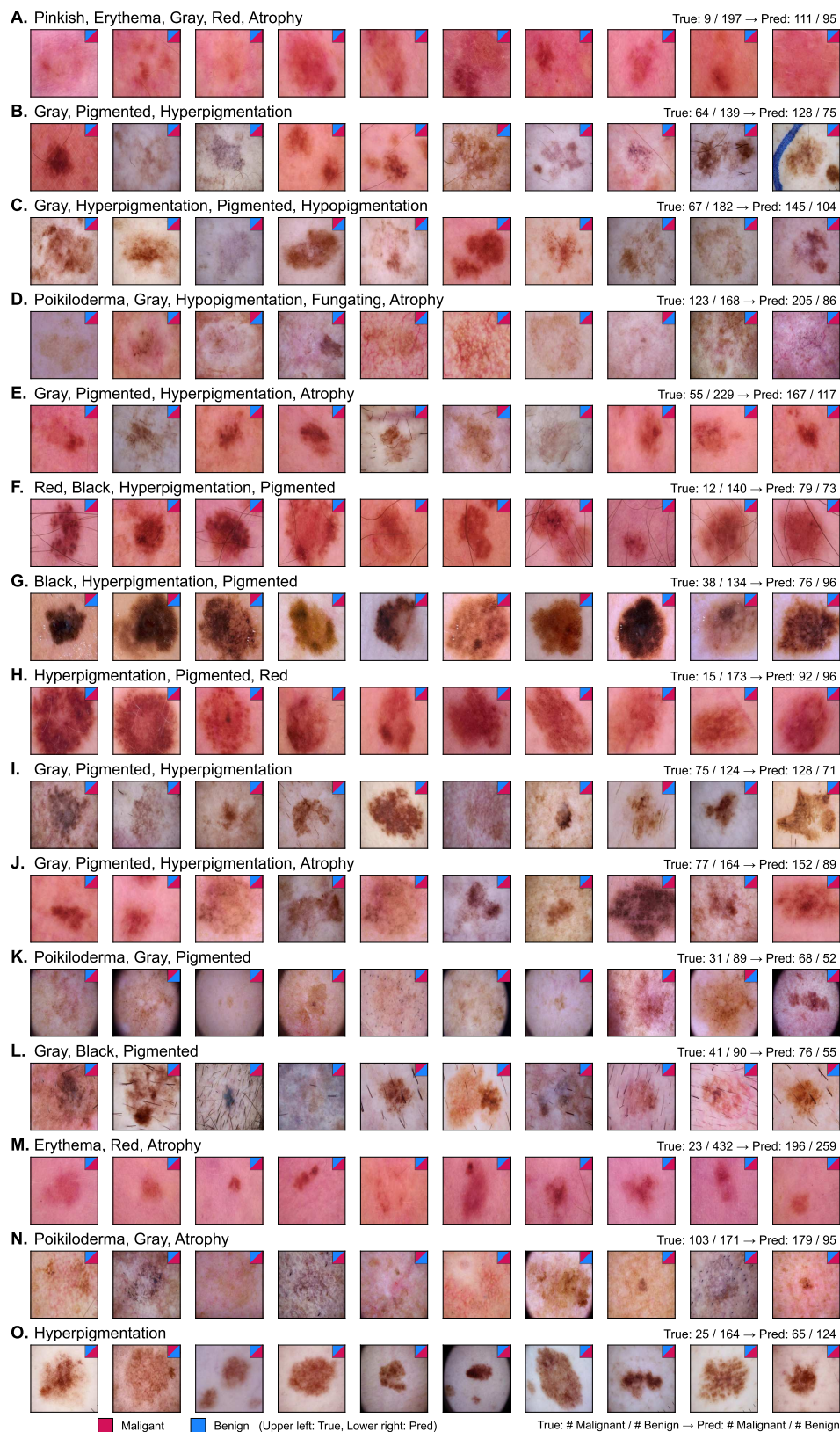
Supplementary Fig. 3 | Dermoscopic images with artifacts as determined by high concept presence scores calculated using MONET. We show the top 30 images for each artifact. (A) Purple pen. (B) Orange sticker. (C) Nail. (D) Hair. (E) Dermoscopic border.



Supplementary Fig. 4 | Accuracy of concept differential analysis. We perform a benchmark analysis to assess MONET’s ability to identify presented concepts correctly. To do this, we generate two paired datasets with known ground truth (i.e., a specific concept is differentially presented) and conduct concept differential analysis on these datasets, letting us determine how accurately the analysis recognizes the intended concept. This experiment is conducted on 21 out of 48 concepts from SkinCon that remained after excluding those with fewer than 30 positive examples. For each concept, we sample a set of 100 images where a concept is highly presented and another set of 100 images where a concept is highly absent from Fitzpatrick17k and DDI datasets, with replacement. Additionally, we varied the noise parameters (0, 0.1, 0.2, and 0.4), which control how correlated the concept is to each grouping. For example, noise = 0.1 means that in the “concept present” set, 90% of the images have the concept, while in the “concept absent” set, only 10% of the images have the concept. For each combination of settings (i.e., 21 intended concepts and 4 noise levels), we repeat this evaluation 20 times with different random seeds. The error bars represent the 95% confidence interval.



Supplementary Fig. 5 | Concept-level model auditing. We train a model on the Med U. of Vienna dataset and test it on the Hosp. Barcelona dataset. Each row displays one of the top 15 clusters, sorted by high error rates. For each cluster, we show the misclassified images and the corresponding concepts associated with errors. We represent 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 10 misclassified images shown for each are selected based on the average concept presence of the identified concepts.



Supplementary Fig. 6 | Concept-level model auditing. We train a model on the Hosp. Barcelona dataset and test it on the Med U. of Vienna dataset. Each row displays one of the top 15 clusters, sorted by high error rates. For each cluster, we show the misclassified images and the corresponding concepts associated with errors. The 10 misclassified images shown for each are selected based on the average concept presence of the identified concepts.

Concept	File name (Dataset)
Erythema	58b4bc079ca94e6e9377a42ca7564b40.jpg (Fitzpatrick17k)
	720cf31558966c82c118ab75b50632eb.jpg (Fitzpatrick17k)
	5f046cda32a3cc547205662e7be774f9.jpg (Fitzpatrick17k)
Ulcer	d8bf377acc45a3beb0c6e81bf7ac1ff5.jpg (Fitzpatrick17k)

Supplementary Table 1 | Images excluded from figures. We exclude 4 images inappropriate for public display due to the inclusion of sensitive body parts, such as genitals, breasts, and buttocks, from Fig. 2 and Supplementary Fig. 1. Their file names, as well as the dataset they belong to, are noted.

Method	Mean AUROC	
	Labels in Fitzpatrick17k	Labels in DDI
MONET	0.830 (52/59)	0.701 (4/6)
CLIP	0.680 (28/59)	0.595 (0/6)
Fully supervised (ResNet-50)	0.856 (58/59)	0.700 (2/6)

Supplementary Table 2 | Performance of MONET in annotating disease labels as compared to baselines

We use disease labels in the clinical image datasets, Fitzpatrick17k and DDI datasets, as ground truth. We exclude any with less than 30 positive examples, leaving 59 labels in Fitzpatrick17k and 6 labels in DDI for our analysis. We use 4,324 samples from Fitzpatrick17k and 636 samples from DDI. The baselines 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. The numbers in parentheses represent the count of concepts for which the method achieves an AUROC over 0.7 over the total number of diseases examined.

Fitzpatrick skin type	Mean AUROC
FST I-II	0.767 (17/21)
FST III-IV	0.759 (18/21)
FST V-VI	0.768 (16/21)

Supplementary Table 3 | Evaluation of MONET’s concept generation performance per skin tone.

We calculate AUROC metrics per each Fitzpatrick skin type (FST) separately: FST I-II (light skin tone, $n = 717$), FST III-IV (intermediate skin tone, $n = 607$), and FST V-VI (dark skin tone, $n = 283$). The numbers in parentheses represent the count of concepts for which the method achieves an AUROC over 0.7 over the total number of concepts examined.

Concept	Terms
abscess	abscess, swollen, pus-filled lump
acuminate	acuminate
atrophy	atrophic
black	black color, black
blue	blue, blue color
brown(hyperpigmentation)	hyperpigmented, hyperpigmentation, brown(hyperpigmentation)
bullae	bullae, blister
burrow	scabies, burrow
comedo	whitehead, blackhead
crust	dried crust, crust
cyst	cyst
dome-shaped	like dome
erosion	erosive, erosion, breakdown of the outer layers, impetigo
erythema	redness, erythematous
excoriation	excoriation
exophytic/fungating	fungating
exudate	exudate
fissure	dry and cracked skin
flat topped	flat topped
friable	friable
gray	gray
induration	edema, oedema
lichenification	lichenification, thickened and leathery
macule	freckle, macular, lentigo, macule
nodule	nodular, cyst, nodule
papule	papular
patch	hyperpigmented, melasma, vitiligo
pedunculated	pedunculated
pigmented	pigmented
plaque	plaque, dermatitis, psoriasis
poikiloderma	sun aging
purple	purple
purpura/petechiae	purpura
pustule	pustule
salmon	salmon patch
scale	flaky and scaly, scaly, hyperkeratosis
scar	scar, keloid scars, hypertrophic scars, contractures scars, acnescars scars
sclerosis	scleroderma, crest syndrome
telangiectasia	dilated or broken blood vessels
translucent	translucent, this bump is translucent
ulcer	ulcer, ulcerated
umbilicated	umbilicated
vesicle	vesicle, fluid-containing
warty/papillomatous	warty and papillomatous
wheal	urticaria
white(hypopigmentation)	white(hypopigmentation), hypopigmentation
xerosis	dry skin, abnormally dry skin, xerosis
yellow	yellow
purple pen	purple pen
nail	nail
pinkish	pinkish
red	red
hair	hair
orange sticker	orange sticker
dermoscope border	dermoscopy

Supplementary Table 4 | Terms used to generate concept prompts

Concept of Interest	Reference Concepts
Asymmetry	Symmetry, Regular, Uniform
Irregular	Regular, Smooth
Blue	Green, Red
White	Black, Colored, Pigmented
Brown	Pale, White
Black	White, Creamy, Colorless, Unpigmented
Erosion	Deposition, Buildup
Multiple Colors	Single Color, Unicolor
Tiny	Large, Big
Regular	Irregular

Supplementary Table 5 | Concepts used in the bottleneck layer for the Concept Bottleneck Model