GestaltMatcher: Overcoming the limits of rare disease matching using facial phenotypic descriptors

Peter Krawitz (pkrawitz@uni-bonn.de)
University of Bonn https://orcid.org/0000-0002-3194-8625
Tzung-Chien Hsieh
University of Bonn https://orcid.org/0000-0003-3828-4419
Aviram Bar-Haim
FDNA Inc.
Shahida Moosa
Division of Molecular Biology and Human Genetics, Stellenbosch University and Medical Genetics
Nadja Ehmke
Charité
Karen Gripp
A.I. DuPont Hospital for Children/Nemours
Jean Tori Pantel
Charité - Universitätsmedizin Berlin https://orcid.org/0000-0002-2674-4660
Magdalena Danyel
Charité
Martin-Atta Mensah
Charité
Denise Horn
Institut fur Humangenetik
Nicole Fleischer
FDNA
Guilherme Bonini
FDNA Inc.
Alexander Schmid
Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn, Rheinische Friedrich-Wilhelms-Universität Bonn
Alexej Knaus
University of Bonn
Sugirthan Sivalingam

Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn, Rheinische Friedrich-Wilhelms-Universität Bonn

Tom Kamphans
GeneTalk

Frédéric Ebstein
Institut für Medizinische Biochemie und Molekularbiologie, Universität Greifswald

Elke Krüger
Institut für Medizinische Biochemie und Molekularbiologie, Universität Greifswald
https://orcid.org/0000-0002-2551-242X

Sébastien KURY
CHU de Nantes https://orcid.org/0000-0001-5497-0465

Stephane Bezieau
CHU Nantes https://orcid.org/0000-0003-0095-1319

Axel Schmidt
Institute of Human Genetics, University of Bonn, Medical Faculty & University Hospital Bonn

Sophia Peters
Institute of Human Genetics, University of Bonn, Medical Faculty & University Hospital Bonn

Hartmut Engels
Institute of Human Genetics, University of Bonn, Medical Faculty & University Hospital Bonn

Elisabeth Mangold
University of Bonn

Martina Kreiß
Institute of Human Genetics, University of Bonn, Medical Faculty & University Hospital Bonn

Kirsten Cremer
University Hospital Bonn

Claudia Perne
Institute of Human Genetics, University of Bonn, Medical Faculty & University Hospital Bonn

Regina Betz
University of Bonn https://orcid.org/0000-0001-5024-3623

Kathrin Grundmann-Hauser
Institute of Medical Genetics and Applied Genomics, University of Tübingen

Tobias Haack
Helmholtz Zentrum München

Matias Wagner
Institute of Human Genetics, School of Medicine, Technical University Munich

T. Brunet
Institute of Human Genetics, Technical University of Munich

Heidi Beate Bentzen
University of Oslo https://orcid.org/0000-0001-8285-818X

Malte Spielmann
Institute of Human Genetics, University of Lübeck
Christian Schaaf
Department of Human Genetics, University Hospital of Heidelberg
Stefan Mundlos
Charite
Markus Nöthen
University of Bonn

Technical Report

Keywords: monogenic disorders, craniofacial abnormalities, phenotyping tools, DeepGestalt, clinical diagnosis

Posted Date: February 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-138785/v1

Version of Record: A version of this preprint was published at Nature Genetics on February 10th, 2022.
See the published version at https://doi.org/10.1038/s41588-021-01010-x.

Abstract

The majority of monogenic disorders cause craniofacial abnormalities with characteristic facial morphology. These disorders can be diagnosed more efficiently with the support of computer-aided next-generation phenotyping tools, such as DeepGestalt. These tools have learned to associate facial phenotypes with the underlying syndrome through training on thousands of patient photographs. However, this “supervised” approach means that diagnoses are only possible if they were part of the training set. To improve recognition of ultra-rare diseases, we created GestaltMatcher, which uses a deep convolutional neural network based on the DeepGestalt framework. We used photographs of 21,836 patients with 1,362 rare disorders to define a “Clinical Face Phenotype Space”. Distance between cases in the phenotype space defines syndromic similarity, allowing test patients to be matched to a molecular diagnosis even when the disorder was not included in the training set. Similarities among patients with previously unknown disease genes can also be detected. Therefore, in concert with mutation data, GestaltMatcher could accelerate the clinical diagnosis of patients with ultra-rare disorders and facial dysmorphism.

Introduction

Rare genetic disorders affect more than 6.2% of the global population[1]. Because genetic disorders are rare and diverse, accurate clinical diagnosis is a time-consuming and challenging process, often referred to as the “diagnostic odyssey.[2]” Craniofacial abnormalities are present in 30–40% of genetic disorders[3]. Patients with these syndromic disorders usually have recognizable facies, such as the typical features associated with Down syndrome or Fragile X syndrome. Hence, the facial manifestation can provide a crucial visual hint to help a clinician identify possible underlying disorders, which reduces the search space of candidate genes and speeds up the genetic diagnostic workup. However, the ability to recognize these syndromic disorders relies heavily on the clinician’s experience. Reaching a diagnosis is very challenging if the clinician has not previously seen a patient with an ultra-rare disorder or if the patient presents with a novel disease, both of which are increasingly common scenarios.

With the rapid development of machine learning and computer vision, a considerable number of next-generation phenotyping (NGP) tools have emerged that can analyze facial dysmorphology using two-dimensional (2D) portraits of patients[4–12]. These tools can aid in the diagnosis of patients with facial dysmorphism by matching their facial phenotype with that of known disorders. In 2014, Ferry *et al.* proposed using a Clinical Face Phenotype Space (CFPS) formed by facial features extracted from images to perform syndrome classification; the system in that study was trained on photos of more than 1,500 controls and 1,300 patients with eight different syndromes[4]. Since then, facial recognition technologies have improved significantly and constitute the core of the deep-learning revolution in computer vision[13,14]. The current state-of-the-art framework for syndrome classification, DeepGestalt, has been trained on more than 20,000 patients and currently achieves high accuracy in identifying the correct syndrome for roughly 300 syndromes[11,15]. DeepGestalt has also demonstrated a strong ability to

separate specific syndromes and subtypes, surpassing human experts' performance. Hence, pediatricians and geneticists increasingly use such NGP tools for differential diagnostics in patients with facial dysmorphism. However, most existing tools, including DeepGestalt, need to be trained on large numbers of photographs, and are therefore limited to syndromes with at least seven submissions. The number of submissions to diagnostic databases of pathogenic variants, such as ClinVar[16], has become a good surrogate for the prevalence of rare disorders. When submissions to ClinVar of disease genes with pathogenic mutations are plotted in decreasing order, most of the supported syndromes are on the left, indicating relatively high prevalence (Figure 1). For instance, Cornelia de Lange syndrome (CdLS), which has been modeled by multiple tools[4,11], is caused by mutations in *NIPBL*, *SMC1A*, and *HDAC8,* as well as other genes, and has been linked to hundreds of reported mutations. However, more than half of the genes in ClinVar have fewer than ten submissions each (Figure 1). As a result, most phenotypes have not been modeled because sufficient data are lacking. Thus, the need to train on large numbers of photographs is a major limitation for the identification of ultra-rare syndromes.

A second limitation of classifiers such as DeepGestalt is that their end-to-end, offline-trained architecture does not support new syndromes without additional modifications. In order to model a new syndrome in a deep convolutional neural network (DCNN), the developer has to go through six separate steps (Supplementary Figure 1), including collecting images of the new syndrome; changing the classification head, which is the last layer of the DCNN; retraining the network; and more. In addition, the model cannot be used to quantify similarities among undiagnosed patients, which is crucial in the delineation of novel syndromes.

A third shortcoming of current approaches is that they are not able to contribute to the longstanding discussion within the nosology of genetic diseases about distinguishability. Syndromic differences have been hard to measure objectively[17], and decisions to "split" syndromes into separate entities on the basis of perceived differences or to "lump" syndromes together on the basis of similarities have been made subjectively. Current tools are unable to quantify the similarities between syndromes in a way that could shed light on the underlying molecular mechanisms and guide classification.

Here we describe GestaltMatcher, an innovative approach that uses an image encoder to convert all features of a facial image into a vector of numbers. These vectors are then used to build a CFPS for matching a patient's photo to a gallery of portraits of solved or unsolved cases. The distance between cases in the CFPS quantifies the similarities between the faces, thereby matching patients with known syndromes or identifying similarities between multiple patients with unknown disorders and thereby helping to define new syndromes. Because GestaltMatcher quantifies similarities between faces in this way, it addresses all three of the limitations described above: (1) it can identify "closest matches" among patients with known or unknown disorders, regardless of prevalence; (2) it does not need new architecture or training to incorporate new syndromes; and (3) it creates a search space to explore similarity of facial gestalts based on mutation data, which can point to shared molecular pathways of phenotypically similar disorders.

Results

The feature encoder of GestaltMatcher computes a Facial Phenotypic Descriptor (FPD) for each portrait image (Figure 2a). Each FPD can be thought of as one coordinate in the CFPS (Figure 2b). The distances between the FPDs in the CFPS form the basis for syndrome classification and patient clustering.

The complete dataset used to construct the CFPS consisted of 33,350 images from 21,836 subjects who had been diagnosed with a total of 1,362 syndromes, each supported by at least two cases. We divided the dataset into categories of distinct (rare syndromes with facial dysmorphism recognized by DeepGestalt), non-distinct (rare syndromes without described facial dysmorphism, not recognized by DeepGestalt), and target (ultra-rare syndromes with facial dysmorphism that we hope to be able to identify, not evaluable by DeepGestalt). Each category was further split into the gallery (90% of each syndrome) and a test set (the remaining 10% of each syndrome) (see the Online methods for details).

Training on images of dysmorphism improves the performance of the FPD

To investigate the importance of using a syndromic features encoder rather than a normal facial features encoder, we compared FPDs created by the DeepGestalt encoder (Enc-DeepGestalt) with those created by the CASIA-WebFace[18] encoder (Enc-CASIA), which has the same architecture. DeepGestalt was first trained on the faces of healthy subjects and then fine-tuned by training on dysmorphic faces from a gallery of patients with 296 distinct syndromes, whereas Enc-CASIA was trained on the faces of healthy subjects only. All images were encoded separately for each encoder. We then evaluated the performance of the encoders with the distinct, non-distinct, and target test sets. The performance metric was the percentage of test cases (with known diagnosis) for which an FPD with the matching disorder was within the *k* closest diagnoses in the CFPS (the top-*k* accuracy). The features created by DeepGestalt performed better in the matching process than those created with Enc-CASIA (Table 1). This emphasizes the importance of training the encoder on data from faces with dysmorphic phenotypes and not only on healthy faces. The features created by DeepGestalt improved the accuracy of matching within the top-10 closest images by 33% for the distinct category. Furthermore, the top-10 accuracy was improved by 33% for the target syndromes, which do not overlap with the distinct syndromes. These results suggest that the features encoded by DeepGestalt are a better fit for the task of syndrome classification than the features encoded by the modern CASIA face recognition model. Moreover, DeepGestalt's FPD provides a better generalization than the FPD encoded by CASIA for target syndromes that it had not previously seen.

Top-10 accuracy plateaus when GestaltMatcher is trained on more than 100 syndromes

Earlier definitions of the FPD were mainly based on training a network with a small selection of common and highly characteristic syndromes[4,8]. In principle, we could train GestaltMatcher's encoder on all 1,362 different syndromes in our dataset. However, most of the phenotypes that have recently been linked to a gene are either ultra-rare or less distinctive, and using a very unbalanced training set with many ultra-rare disorders linked to only few cases may add noise without substantial additional benefit. We therefore analyzed the influence of the number of syndromes on model training by incrementally increasing their number starting with the most frequent ones (Figure 3). The top-10 accuracy improved with an increase in the number of syndromes until 110 syndromes was reached, fluctuated as the number of syndromes further increased to 190, and became saturated after 190 syndromes. From these dynamics, we can conclude that including additional syndromes for defining the FPD will provide little benefit, and we decided to model the encoder of GestaltMatcher with the previous 296 syndromes of DeepGetalt, rather than all 1,362.

GestaltMatcher performs similar to DeepGestalt with better scalability

To validate the GestaltMatcher approach, we first worked with the 323 images of patients with 90 syndromes from the London Medical Database (LMD)[19] that were already used for benchmarking the performance of DeepGestalt[11]. When using the distinct gallery, which contains syndromes that DeepGestalt currently supports, GestaltMatcher achieved 74.30% and 89.78% accuracy within the top-10 and top-30 ranks, respectively, which was lower than the 84.52% top-10 accuracy and 91.64% top-30 accuracy achieved with DeepGestalt (Table 2 and Supplementary Table 1). However, when we used the gallery of all 1,362 syndromes for GestaltMatcher (distinct, non-distinct, and target), the top-10 and top-30 dropped by only 3.78% and 4.98%, respectively, indicating that the GestaltMatcher approach is highly scalable.

Matching undiagnosed patients from unrelated families

We envision the use of GestaltMatcher as a phenotypic complement to GeneMatcher[20]. To prove that we can match patients from unrelated families who have the same disease by using only their facial photos, we selected syndromes from 14 recent GeneMatcher publications with a title containing the phrase "facial dysmorphism". In this test set, we matched 27 of 104 photos and connected 27 of 77 families when using the top-10 criterion (Table 3, Figure 4, and Supplementary Figure 2). When using the top-30 rank, 47 of 104 photos were matched, and 41 of 77 families were connected. Enc-CASIA, which is trained only with healthy subjects, matched only 30 out of 104 photos and connected 32 out of 77 families using the top-30 rank (Supplementary Table 2). Hence, using the encoder trained with facial dysmorphic subjects improves the matching considerably.

As an example, in a study of *TMEM94*[21], nine of the ten photos in six different families were matched, and all six families were connected within the top-10 rank. When the three test images in family 2 (F-2-5, F-2-7, F-2-9) were tested, the other five families were among those in the top-30 rank (Figure 4). The youngest brother, F-2-5, matched families 1, 3, 5, and 6, and both sisters, F-2-7 and F-2-9, matched families 1, 4, 5, and 6. The six families were recruited at five different institutes in India, Qatar, the United States (NIH Undiagnosed Diseases Network), and Switzerland, indicating that GestaltMatcher can also connect patients of different ethnic origins. However, a more systematic analysis of pairwise distances still revealed considerably smaller distances between subjects with *de novo* mutations and their unaffected family members than between these subjects and unrelated individuals (Supplementary Figure 3). Hence, ethnicity could be a potential confounding factor for the GestaltMatcher approach. However, it is a bias that can be attenuated[22] and will also diminish over time when more diverse training data becomes available[23].

Syndrome distinctiveness assessed by GestaltMatcher correlates with expert opinion

We hypothesized that ultra-rare disorders that were linked to their disease-causing genes early on, such as Schuurs-Hoeijmakers syndrome in 2012[24], have particularly distinctive facial phenotypes. To systematically analyze the dependency of disease-gene discovery on the distinctiveness of a facial gestalt, we asked three expert dysmorphologists to grade 296 syndromes on a scale from 1 to 3. The more easily they could distinguish the diseases, and the more characteristic of the disease they deemed the facial features, the higher the score. All three syndromlogists agreed on the same score for 195/296 syndromes, yielding a concordance of 65.8%. We then analyzed the correlation of the mean of this distinctiveness score from human experts with the top-10 accuracy that GestaltMatcher achieves for these syndromes without having been trained on them (Figure 5a). The Spearman's rank correlation coefficient was 0.421 (P = 0.002), indicating a clear positive correlation between distinctiveness score and top-10 accuracy. Syndromes with a higher average score tended to perform better, with Schuurs-Hoeijmakers syndrome being amongst the best-performing syndromes in GestaltMatcher. In contrast, there was no significant correlation for GestaltMatcher accuracy and disease prevalence (P = 0.126; Figure 5b).

Characterization of phenotypes in the CFPS

When syndromologists cannot reach a final diagnosis for a patient after extensive genetic sequencing, they may compare the patient's condition to a known molecular disorder, for example describing a "syndrome *XY*–like phenotype". In GestaltMatcher, such comparisons can be supported by cluster analysis in the CFPS with the cosine distance as a similarity metric (Supplementary Table 3).

If a novel disease gene has been identified and the similarities of the patients to known phenotypes outweigh the differences, OMIM groups them into a phenotypic series. On the gene or protein level, such phenotypic series often correspond to molecular-pathway diseases, such as GPI-anchor deficiencies for Hyperphosphatasia with mental retardation syndrome (HPMRS) or cohesinopathies for CdLS. For our cluster analysis, we sampled subjects in our database with subtypes of four large phenotypic series and found high inter-syndrome separability in addition to considerable intra-syndrome substructure in e.g. Noonan syndrome, CdLS, or mucopolysaccharidosis. A *t*-SNE[25] projection of the FPDs into two dimensions yielded the best visualization results (Supplementary Figure 4). Although any projection into a smaller dimensionality might cause a loss of information, the clusters are still clearly visible for the 743 subjects sampled from these four phenotypic series. This observation provides further evidence that characteristic phenotypic features are encoded in the FPDs.

To demonstrate the separability of syndromes with facial dysmorphism, we also used *t*-SNE to project 4,353 images of the ten distinct syndromes with the largest number of subjects and 872 images of ten non-distinct syndromes into 2D space. In addition, we calculated the Silhouette index[26] for both of these datasets. The FPDs of the distinct syndromes showed ten clear clusters of subjects (Supplementary Figure 5), but the *t*-SNE projection of subjects with non-distinct syndromes created no clear clusters. Moreover, the Silhouette index of the distinct syndromes (0.11) was higher than that of the non-distinct syndromes (−0.005); the negative Silhouette index indicates poor separation of the non-distinct syndromes.

GestaltMatcher as a tool for clinician scientists

The transition of a research case to a diagnostic case is best described by the process of matching unrelated patients in the CFPS with a shared molecular cause until statistical significance is reached. We illustrate this process for the novel disease gene *PSMC3* in a demonstration on the GestaltMatcher website (Supplementary Figure 6, www.gestaltmatcher.org). Ebstein *et al.* (not yet published) report 18 patients with a neurodevelopmental disorder of heterogeneous dysmorphism that is caused by *de novo* missense mutations in *PSMC3*, which encodes a proteasome 26S subunit. Although not all patients have a single facial gestalt in common, the proximity of two unrelated patients in the CFPS who share the same *de novoPSMC3* mutation is exceptional. Their distance is comparable to the pairwise distances of patients with the reoccurring missense mutation R203W in *PACS1*, which is the only known cause of Schuurs-Hoeijmakers syndrome. On the one hand, the high distinctiveness of these two *PSMC3* cases with the same mutation allows direct matching by phenotype. On the other hand, the pairwise similarities of 10 out of 18 patients in the CFPS for which portraits were available, also hints that the protein domains have more than one function. The previously described scalability of GestaltMatcher makes an exploration of such similarities in the CFPS possible for any number of cases as soon as they have been added to the gallery of undiagnosed patients.

Discussion

GestaltMatcher's ability to match previously unseen syndromes, i.e., those for which no patient is included in the training set, distinguishes it from other tools. Because matching of unseen syndromes can be considered the discovery of novel diseases, GestaltMatcher could speed up the process of defining new diseases.

Importantly, GestaltMatcher provides the flexibility to easily scale up the number of supported syndromes. Although the LMD validation analysis revealed that the use of softmax to predict syndromes trained in the model outperformed GestaltMatcher, GestaltMatcher demonstrated high scalability by yielding similar performance when the number of supported syndromes in the CFPS was increased from 296 to 1,362. Furthermore, the distinctiveness of a syndrome correlated with the performance (Figure 5a), whereas syndrome prevalence did not (Figure 5b). Thus, GestaltMatcher can match a syndrome with a distinguishable facial gestalt even if it is of extremely low prevalence. This enables us to avoid the long development flow currently required to support and discover novel syndromes (Supplementary Figure 1). Instead, matching can be offered instantly for all undiagnosed cases with available frontal images for which consent has been provided for inclusion in the tool.

GestaltMatcher's framework also allows us to abstract the encoding of a dataset away from the classification task. For example, one can evaluate both phenotypic series and pleiotropic genes within a single CFPS, or obtain the most-similar patients for each of the matched syndromes, with minor computational cost (i.e., in real time). Furthermore, the GestaltMatcher framework computes the similarity between each of the test set images across the entire dataset of images. This similarity can be computed using different metrics, e.g., cosine or Euclidean distance. The results are then aggregated according to the chosen configuration. For example, image similarity can be aggregated at the patient level or the syndrome level. Furthermore, the dataset can be filtered according to different parameters (such as ethnicity, number of affected genes, or age) to further customize the evaluation.

One of the most important features of GestaltMatcher is the ability to match patients with highly similar facial features. Clinicians are often faced with the challenge of finding enough patients with a similar phenotype to statistically link the phenotype to a gene. This is especially true when dealing with presumed novel or extremely rare Mendelian disorders. Several online platforms, such as GeneMatcher, MyGene2 (https://mygene2.org/MyGene2), and Matchmaker Exchange[27], allow physicians to look for similar patients based on phenotypic data, such as HPO terms, or genomic sequencing information, and over the past few years, these platforms have facilitated the matching of thousands of patients. However, although facial phenotypes are crucial for allowing physicians to determine whether two patients have a similar disorder, automated facial matching technology has not yet been included in any of these "gene matching" platforms. We expect that GestaltMatcher will be readily integrated into these matching platforms to aid in determining which phenotypes should be grouped together into a syndrome or phenotypic series, as well as linking individual patients to a molecular diagnosis.

Since its first proof of concept, in which GestaltMatcher was used to identify two unrelated patients from different countries with the same novel disease, caused by the same *de novo* mutation in *LEMD2*[28], our

approach has successfully be applied to further ultra-rare disorders (Figure 1). We matched 41 of 77 different families in 14 GeneMatcher publications by top-30 rank, and 11 candidate genes are currently under evaluation. This result shows the power and potential of GestaltMatcher to identify novel syndromes.

Online Methods

Study approval

This study is governed by the following Institutional Review Board (IRB) approval: Charité–Universitätsmedizin Berlin, Germany (EA2/190/16); UKB Universitätsklinikum Bonn, Germany (Lfd.Nr.386/17). The authors have obtained written informed consent given by the patients or their guardians, including permission to publish photographs.

Datasets

We collected images of subjects with clinically or molecularly confirmed diagnoses from the Face2Gene database (https://www.face2gene.com). Extracted, deidentified data were used to remove poor-quality or duplicated images from the dataset without viewing the photos. After removing images of insufficient quality, the dataset consisted of 33,350 images from 21,836 subjects with a total of 1,362 syndromes (Supplementary Table 4).

GestaltMatcher was designed to distinguish syndromes with different properties. We separated syndromes by the number of affected subjects and whether they had already been learned by the DeepGestalt model. Supplementary Figure 7 provides an overview of how the dataset was divided. The current DeepGestalt approach requires at least seven subjects to learn a novel syndrome. We first used this threshold to separate the syndromes into rare and ultra-rare syndromes. We denoted ultra-rare syndromes as “target” syndromes because the objective of our study was to improve phenotypic decision support for these disorders. However, rare syndromes that are not associated with facial dysmorphic features cannot be modeled by DeepGestalt. We therefore further divided rare syndromes into “distinct” (possessing characteristic facial dysmorphism recognized by DeepGestalt) and “non-distinct” (without facial dysmorphic features or that cannot be recognized by DeepGestalt). The distinct syndromes were used to validate syndrome prediction and the separability of subtypes of a phenotypic series because these syndromes are known to have facial dysmorphic features that are well recognized by the DeepGestalt encoder. We excluded autism from the non-distinct group of syndromes in this study because it had many more subjects than other non-distinct syndromes, leading to an imbalanced dataset. For target syndromes, we sought to demonstrate that GestaltMatcher could predict a syndrome even if facial images were publicly available for only a few subjects. It is noteworthy that, for more than half of all known disease-causing genes, fewer than ten cases with pathogenic variants have been submitted to

ClinVar (Figure 1). Of the 1,362 syndromes in the entire dataset, 296 were distinct, 242 non-distinct, and 824 target. DeepGestalt cannot yet be applied to non-distinct and target syndromes.

We further divided each of these three datasets into a gallery and test set. The gallery is the set of subjects that we intend to match, given a subject from the test set. First, 90% of subjects with each distinct syndrome were used for training models, and the remaining 10% of subjects were used to validate DeepGestalt training; the 90% then became the distinct gallery and the 10% were assigned to the distinct test set. For the target and non-distinct datasets, we performed 10-fold cross-validation. In each syndrome, 90% and 10% of subjects were assigned to the gallery and test set, respectively.

Matching only within a dataset would not represent a real-world scenario. Therefore, the galleries of the three datasets were later combined into a unified gallery that was used to search for matched patients.

DeepGestalt encoder

The preprocessing pipeline of DeepGestalt includes point detection, facial alignment (frontalization), and facial region cropping. During inference, facial region crop is forward passed through a deep convolutional network (DCNN), and ultimately got the final prediction of the input face image. The DeepGestalt network consists of ten convolutional layers (Conv) with batch normalization (BN) and a rectified linear activation unit (ReLU) to embed the input features. After every Conv-BN-ReLU layer, a max pooling layer is applied to decrease spatial size while increasing the semantic representation. The classifier part of the network consists of a fully connected linear layer with dropout (0.5). In this study, we considered the DeepGestalt architecture as an encoder–classification composition, pipelined during inference. We chose the last fully connected layer before the softmax classification as the facial feature representation (facial phenotypic descriptor, FPD), resulting in a vector of size 320. The encoder trained on 296 distinct syndromes was named Enc-DeepGestalt.

Our first hypothesis was that images of patients with the same molecularly diagnosed syndromes or within the same phenotypic series, and who also share similar facial phenotypes, can be encoded into similar feature vectors under some set of metrics. Moreover, we hypothesized that DeepGestalt's specific design choice of using a predefined, offline-trained, linear classifier could be replaced by other classification "heads", for example, *k*-Nearest Neighbors using cosine distance, which we used for GestaltMatcher.

Descriptor projection: Clinical Face Phenotype Space

Each image was encoded by the DeepGestalt encoder, resulting in a 320-dimensional FPD. These FPDs were further used to form a 320-dimensional space called the Clinical Face Phenotype Space (CFPS), with each FPD a point located in the CFPS, as shown in Figure 2. The similarity between two images is quantified by the cosine distance between them in the CFPS. The smaller the distance, the greater the

similarity between the two images. Therefore, clusters of subjects in the CFPS can represent patients with the same syndrome, similarities among different disorders, or the substructure under a phenotypic series.

Evaluation

To evaluate GestaltMatcher, we took the images in the test set as input and positioned them in the CFPS defined by the images of the gallery. We calculated the cosine distance between each of the test set images and all of the gallery images. Then, for each test image, if an image from another subject with the same disorder in the gallery was among the top-*k* nearest neighbors, we called it a top-*k* match. We then benchmarked the performance by top-*k* accuracy (percent of test images with correct matches within the top *k*). We further compared the accuracy of each syndrome in the distinct, non-distinct, and target syndrome subsets to investigate whether GestaltMatcher can extend DeepGestalt to support more syndromes.

London Medical Dataset validation analysis

We compiled 323 images of patients diagnosed with 90 distinct syndromes from the LMD[19] and used this as the validation set for distinct syndromes. We first evaluated the validation set using softmax, which is a DeepGestalt method. To compare the performance with that of GestaltMatcher, we evaluated the performance of GestaltMatcher on two different galleries: a gallery of distinct syndromes consisting of 20,091 images of patients with 296 syndromes, and a unified gallery consisting of 27,826 images of patients with 1,362 syndromes. We then reported the top-*k* accuracy and compared the results of these three conditions (DeepGestalt with softmax, GestaltMatcher with distinct gallery, and GestaltMatcher with unified gallery).

Target syndromes analysis

To understand the potential for matching target syndromes, we trained an encoder, denoted Enc-Target, on 477 out of 824 target syndromes with more than three and fewer than seven subjects. Ninety percent of the subjects were used to train Enc-Target and were later assigned to the gallery. The remaining 10% of subjects were assigned to the test set. We then compared the performance of Enc-Target and Enc-DeepGestalt (see previous section) using cosine distance and the softmax classifier.

Syndrome facial distinctiveness score

To evaluate the importance of the facial gestalt for clinical diagnosis of the patient, we asked three dysmorphologists to score the usefulness of each syndrome's facial gestalt for establishing a diagnosis. Three levels were established:

1. Facial gestalt can be supportive in establishing the clinical diagnosis.
2. Facial gestalt is important in establishing the clinical diagnosis, but diagnosis cannot be made without additional clinical features.
3. Facial gestalt is a cardinal symptom, and a visual or clinical diagnosis is possible based only on the facial phenotype.

We then averaged the grades from the three dysmorphologists for each syndrome.

Syndrome prevalence

The prevalence of each syndrome was collected from Orphanet (www.orpha.net). Birth prevalence was used when the actual prevalence was missing. If only the number of cases or families was available, we calculated the prevalence by summing the numbers of all cases or families and dividing by the global population, using 7.8 billion for the global population and a family size of ten for each family[29].

Unseen syndromes correlation analysis

To investigate the influence of prevalence and distinctiveness score on the performance for novel syndromes with facial dysmorphism, we selected 50 distinct syndromes and kept them out of the training set. The 50 syndromes were selected to have evenly distributed distinctiveness scores and prevalence distribution; the distributions are shown in Supplementary Figure 7 and Table 4. The encoder (Enc-unseen) was trained on 90% of the subjects from the other 246 distinct syndromes. In addition, we performed random downsampling to remove the confounding effect of prevalence. For each iteration, we randomly downsampled each syndrome by assigning five subjects to the gallery and one subject to the test set. We then averaged the top-10 accuracy of 100 iterations. We calculated Spearman rank correlation coefficients for the following two pairs of data: the first between top-10 accuracy and the syndrome's distinctiveness score, and the second between top-10 accuracy and the prevalence of syndromes collected from Orphanet.

Analysis of number of training syndromes

In this analysis, we trained the encoders with different numbers of syndromes. We first sorted the syndromes by the number of subjects in each syndrome, in descending order. We then trained 13 encoders, each with a different number of training syndromes. We used the ten most common syndromes in the training set for the first encoder. For the second encoder, we trained on the top 30 syndromes, and continually increased the number of syndromes for each subsequent encoder by 20 until we reached 246 syndromes. Thus, we simulated how syndromes would be included in model training in the real world. We took the 50 selected distinct syndromes as the test set and performed random downsampling as

described in the previous section; the only difference was that we used encoders trained from ten to 246 syndromes.

GeneMatcher validation analysis

We selected 14 publications in which GeneMatcher was used to match patients with facial dysmorphism from unrelated families. In total, these studies contained 104 photos of 89 subjects from 77 families. The details are shown in Table 3. We performed leave-one-out cross-validation on this dataset, i.e., we kept one photo as the test set, and we assigned the rest of the photos to a gallery of 3,636 photos with 824 target syndromes to simulate the distribution of patients with unknown diagnosis. We then evaluated the performance by top-1 to top-30 rank. If a photo of another subject with the same disease-causing gene from an unrelated family was among the top-*k* rank, we called it a match.

Moreover, we used top-k rank to measure how many unrelated families were connected. If one unrelated family was among the test photo's top-*k* rank, the families were considered to be connected at that rank. How many families were matched to at least one unrelated family was also represented.

Code availability

GestaltMatcher is a partially proprietary framework. While the source code for cropping the face cannot be shared, the architecture of the CNN, as well as a web service of the trained version of the tool is accessible for use by healthcare professionals free of charge at www.gestaltmatcher.org.

Data availability

The data that support the findings of this study are divided into two groups, published data, and restricted data. Published data are available from the reported references and also from www.gestaltmatcher.org. Restricted data are curated from Face2Gene users under a license and cannot be published, to protect patient privacy.

References

1. Ferreira, C. R. The burden of rare diseases. *Am. J. Med. Genet. Part A***179**, 885–892 (2019).
2. Baird, P. A., Anderson, T. W., Newcombe, H. B. &Lowry, R. B. Genetic disorders in children and young adults: A population study. *Am. J. Hum. Genet.***42**, 677–693 (1988).
3. Hart, T. & Hart, P. Genetic studies of craniofacial anomalies: clinical implications and applications. *Orthod. Craniofac. Res.***12**, 212–220 (2009).
4. Ferry, Q. *et al.* Diagnostically relevant facial gestalt information from ordinary photos. 1–22 (2014). doi:10.7554/eLife.02020

5. Kuru, K., Niranjan, M., Tunca, Y., Osvank, E. & Azim, T. Biomedical visual data analysis to build an intelligent diagnostic decision support system in medical genetics. *Artif. Intell. Med.***62**, 105–118 (2014).
6. Cerrolaza, J. J. *et al.* Identification of dysmorphic syndromes using landmark-specific local texture descriptors. *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)* 1080–1083 (2016). doi:10.1109/ISBI.2016.7493453
7. Wang, K. & Luo, J. Detecting Visually Observable Disease Symptoms from Faces. *EURASIP J. Bioinform. Syst. Biol.***2016**, 13 (2016).
8. Dudding-Byth, T. *et al.* Computer face-matching technology using two-dimensional photographs accurately matches the facial gestalt of unrelated individuals with the same syndromic form of intellectual disability. *BMC Biotechnol.***17**, 1–9 (2017).
9. Shukla, P., Gupta, T., Saini, A., Singh, P. & Balasubramanian, R. A Deep Learning Frame-Work for Recognizing Developmental Disorders. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* 705–714 (2017). doi:10.1109/WACV.2017.84
10. Liehr, T. *et al.* Next generation phenotyping in Emanuel and Pallister-Killian syndrome using computer-aided facial dysmorphology analysis of 2D photos. *Clin. Genet.***93**, 378–381 (2018).
11. Gurovich, Y. *et al.* Identifying facial phenotypes of genetic disorders using deep learning. *Nature Medicine***25**, 60–64 (2019).
12. van der Donk, R. *et al.* Next-generation phenotyping using computer vision algorithms in rare genomic neurodevelopmental disorders. *Genet. Med.***21**, 1719–1725 (2019).
13. Taigman, Y., Yang, M., Ranzato, M. & Wolf, L. DeepFace: Closing the gap to human-level performance in face verification. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 1701–1708 (IEEE Computer Society, 2014). doi:10.1109/CVPR.2014.220
14. Learned-Miller, E., Huang, G. B., RoyChowdhury, A., Li, H. & Hua, G. Labeled faces in the wild: A survey. *Adv. Face Detect. Facial Image Anal.* 189–248 (2016). doi:10.1007/978-3-319-25958-1_8
15. Pantel, J. T. *et al.* Efficiency of Computer-Aided Facial Phenotyping (DeepGestalt) in Individuals with and without a Genetic Syndrome: Diagnostic Accuracy Study. *J. Med. Internet Res.***22**, e19263 (2020).
16. Landrum, M. J. *et al.* ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.***46**, D1062–D1067 (2018).
17. McKusick, V. A. On lumpers and splitters, or the nosology of genetic disease. *Perspect. Biol. Med.***12**, 298–312 (1969).
18. Yi, D., Lei, Z., Liao, S. & Li, S. Z. Learning Face Representation from Scratch. (2014).
19. Winter, R. M. & Baraitser, M. The London Dysmorphology Database. *Journal of medical genetics***24**, 509–510 (1987).
20. Sobreira, N., Schiettecatte, F., Valle, D. & Hamosh, A. GeneMatcher: A Matching Tool for Connecting Investigators with an Interest in the Same Gene. *Hum. Mutat.***36**, 928–930 (2015).

21. Stephen, J. *et al.* Bi-allelic TMEM94 Truncating Variants Are Associated with Neurodevelopmental Delay, Congenital Heart Defects, and Distinct Facial Dysmorphism. *Am. J. Hum. Genet.***103**, 948–967 (2018).
22. Alvi, M., Zisserman, A. & Nellaker, C. Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)***11129 LNCS**, 556–572 (2018).
23. Lumaka, A. *et al.* Facial dysmorphism is influenced by ethnic background of the patient and of the evaluator. *Clin. Genet.***92**, 166–171 (2017).
24. Schuurs-Hoeijmakers, J. H. M. *et al.* Recurrent de novo mutations in PACS1 cause defective cranial-neural-crest migration and define a recognizable intellectual-disability syndrome. *Am. J. Hum. Genet.***91**, 1122–1127 (2012).
25. Van Der Maaten, L. & Hinton, G. *Visualizing Data using t-SNE*. **9**, (2008).
26. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.***20**, 53–65 (1987).
27. Philippakis, A. A. *et al.* The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery. *Hum. Mutat.***36**, 915–921 (2015).
28. Marbach, F. *et al.* The Discovery of a LEMD2-Associated Nuclear Envelopathy with Early Progeroid Appearance Suggests Advanced Applications for AI-Driven Facial Phenotyping. *Am. J. Hum. Genet.***104**, 749–757 (2019).
29. Nguengang Wakap, S. *et al.* Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur. J. Hum. Genet.***28**, 165–173 (2020).
30. Stankiewicz, P. *et al.* Haploinsufficiency of the Chromatin Remodeler BPTF Causes Syndromic Developmental and Speech Delay, Postnatal Microcephaly, and Dysmorphic Features. *Am. J. Hum. Genet.***101**, 503–515 (2017).
31. Morimoto, M. *et al.* Bi-allelic CCDC47 Variants Cause a Disorder Characterized by Woolly Hair, Liver Dysfunction, Dysmorphic Features, and Global Developmental Delay. *Am. J. Hum. Genet.***103**, 794–807 (2018).
32. Tanaka, A. J. *et al.* De novo pathogenic variants in CHAMP1 are associated with global developmental delay, intellectual disability, and dysmorphic facial features . *Mol. Case Stud.***2**, a000661 (2016).
33. Weiss, K. *et al.* De Novo Mutations in CHD4, an ATP-Dependent Chromatin Remodeler Gene, Cause an Intellectual Disability Syndrome with Distinctive Dysmorphisms. *Am. J. Hum. Genet.***99**, 934–941 (2016).
34. Balak, C. *et al.* Rare De Novo Missense Variants in RNA Helicase DDX6 Cause Intellectual Disability and Dysmorphic Features and Lead to P-Body Defects and RNA Dysregulation. *Am. J. Hum. Genet.***105**, 509–525 (2019).
35. Harms, F. L. *et al.* Mutations in EBF3 Disturb Transcriptional Profiles and Cause Intellectual Disability, Ataxia, and Facial Dysmorphism. *Am. J. Hum. Genet.***100**, 117–127 (2017).

36. Jansen, S. *et al.* De novo variants in FBXO11 cause a syndromic form of intellectual disability with behavioral problems and dysmorphisms. *Eur. J. Hum. Genet.***27**, 738–746 (2019).
37. Au, P. Y. B. *et al.* GeneMatcher Aids in the Identification of a New Malformation Syndrome with Intellectual Disability, Unique Facial Dysmorphisms, and Skeletal and Connective Tissue Abnormalities Caused by De Novo Variants in HNRNPK. *Hum. Mutat.***36**, 1009–1014 (2015).
38. Diets, I. J. *et al.* De Novo and Inherited Pathogenic Variants in KDM3B Cause Intellectual Disability, Short Stature, and Facial Dysmorphism. *Am. J. Hum. Genet.***104**, 758–766 (2019).
39. Santiago-Sim, T. *et al.* Biallelic Variants in OTUD6B Cause an Intellectual Disability Syndrome Associated with Seizures and Dysmorphic Features. *Am. J. Hum. Genet.***100**, 676–688 (2017).
40. Olson, H. E. *et al.* A Recurrent De Novo PACS2 Heterozygous Missense Variant Causes Neonatal-Onset Developmental Epileptic Encephalopathy, Facial Dysmorphism, and Cerebellar Dysgenesis. *Am. J. Hum. Genet.***102**, 995–1007 (2018).
41. Kanca, O. *et al.* De Novo Variants in WDR37 Are Associated with Epilepsy, Colobomas, Dysmorphism, Developmental Delay, Intellectual Disability, and Cerebellar Hypoplasia. *Am. J. Hum. Genet.***105**, 413–424 (2019).
42. Stevens, S. J. C. *et al.* Truncating de novo mutations in the Krüppel-type zinc-finger gene ZNF148 in patients with corpus callosum defects, developmental delay, short stature, and dysmorphisms. *Genome Med.***8**, 131 (2016).

Tables

Table 1: Performance comparison of the DeepGestalt and CASIA encoders on distinct, non-distinct, and target test sets.

Test set	Model	Gallery		Test images	Top-1	Top-5	Top-10	Top-30
		Images	Syndromes					
Distinct	Enc-DeepGestalt	20,091	296	3,083	**33.56%**	**57.74%**	**68.03%**	**82.43%**
Distinct	Enc-CASIA	20,091	296	3,083	16.65%	38.76%	51.06%	71.15%
Non-distinct	Enc-DeepGestalt	5,488.2	238.3	879.8	**8.70%**	**22.05%**	**30.56%**	**49.84%**
Non-distinct	Enc-CASIA	5,488.2	238.3	879.8	5.72%	15.87%	23.94%	42.97%
Target	Enc-DeepGestalt	2,395.3	820.4	1,186.2	**11.36%**	**20.12%**	**25.25%**	**36.13%**
Target	Enc-CASIA	2,395.3	820.4	1,186.2	7.98%	6.72%	19.00%	29.44%

Enc-DeepGestalt and Enc-CASIA have the same architecture. Enc-DeepGestalt training was initiated with CASIA-WebFace and further fine-tuned on photos of patients. For the top-1 to top-30 columns, the better performance in each pair is boldfaced. The numbers of images and syndromes in non-distinct and target sets are averaged over ten splits. Enc-DeepGestalt outperformed Enc-CASIA on all three types of syndromes, showing the importance of fine-tuning on patient photos for learning facial dysmorphic features.

Table 2: Comparison of GestaltMatcher and DeepGestalt on the LMD validation set.

Method	Gallery images	Supported Syndromes	Top-1	Top-5	Top-10	Top-30
DeepGestalt	-	296	**54.49%**	**77.09%**	**84.52%**	**91.64%**
GestaltMatcher	20091	296	35.91%	64.71%	74.30%	89.78%
GestaltMatcher	**27826**	**1,362**	33.74%	60.18%	70.52%	84.80%

The results of 323 images from LMD, validated by GestaltMatcher and DeepGestalt. We evaluated the GestaltMatcher approach on two different galleries, distinct (n = 296) and unified (n = 1,362). The best performance and the largest number of images and supported syndromes among the three conditions is boldfaced.

Table 3: GeneMatcher validation set.

Gene	PMID	Subject	Connected families[a]		
			Top-10	Top-30	Total
BPTF[30]	28942966	6	0	0	6
CCDC47[31]	30401460	4	0	2	4
CHAMP1[32]	27148580	4	4	4	4
CHD4[33]	27616479	3	0	0	3
DDX6[34]	31422817	4	4	4	4
EBF3[35]	28017373	7	0	0	6
FBXO11[36]	30679813	17	6	9	17
HNRNPK[37]	26173930	3	3	3	3
KDM3B[38]	30929739	9	2	4	7
OTUD6B[39]	28343629	9	0	3	4
PACS2[40]	29656858	6	0	2	6
TMEM94[21]	30526868	10	6	6	6
WDR37[41]	31327508	4	2	2	4
ZNF148[42]	27964749	3	0	2	3
Total	-	89	27	41	77
Average	-	-	35.06%	53.25%	-

[a] Number of families matched by a photo from another family in the top-10 or top-30 rank.

For example, in the *TMEM94* study, ten out of ten images had an image from another family within the top-30 rank, and all six families had at least one subject from another family in their top-30 rank.

Table 4: The 50 selected syndromes used in the random downsampling experiment, sorted by top-10 accuracy.

ID	Syndrome	Top 10	Score[a]	Prevalence[b]
1	Hutchinson-Gilford Progeria Syndrome; HGPS	92.62	2.33	0.005
2	Mucopolysaccharidosis Type VI; MPS6	87.63	2	0.16
3	Nijmegen Breakage Syndrome; NBS	86.18	2	1
4	Barth Syndrome; BTHS	79.67	1.33	0.22
5	Williams-Beuren Region Duplication Syndrome	78.95	1.33	0.00209
6	Crouzon Syndrome	78.76	2.33	0.9
7	Williams-Beuren Syndrome; WBS	78.53	3	10.8
8	Baraitser-Winter Syndrome	78.14	2.67	0.00077
9	Schuurs-Hoeijmakers syndrome; SHMS	78.05	2	0.00002564
10	Oculodentodigital Dysplasia	76.92	2.33	0.00312
11	Campomelic Dysplasia	75.66	1.33	0.33
12	Laron Syndrome	74.22	1.67	0.3
13	Cornelia De Lange Syndrome	73.55	3	1.9
14	Coffin-Lowry Syndrome; CLS	73.46	2	1.5
15	Pycnodysostosis	72.64	2	0.13
16	Mucolipidosis III Alpha/beta	72.41	2	13
17	Wolf-Hirschhorn Syndrome; WHS	72.25	2.67	2
18	Renpenning Syndrome 1; RENS1	69.17	2.33	0.00082
19	Blepharophimosis, Ptosis, and Epicanthus Inversus; BPES	68.75	3	0.0001
20	Dubowitz Syndrome	68.63	2	0.2
21	Branchiooculofacial Syndrome; BOFS	68.35	2	0.00192
22	Lubs X-Linked Mental Retardation Syndrome; MRXSL	68.22	1.67	0.00256
23	Weaver Syndrome; WVS	67.86	2	0.00062
24	Hallermann-Streiff Syndrome; HSS	67.16	2.67	0.00192
25	Hyper-IgE Recurrent Infection Syndrome	67	1	0.1
26	Sotos Syndrome	66.86	2.67	7.1
27	Seckel Syndrome	66.67	2.67	0.2
28	Koolen-de Vries Syndrome; KDVS	66.42	2.33	1.82
29	Ectrodactyly, Ectodermal Dysplasia, and Cleft Lip/palate Syndrome 1; EEC1	64.81	2	1.11
30	Opitz GBBB Syndrome, Type II; GBBB2	64.29	2	3
31	Ehlers-Danlos syndrome, vascular type; EDSVASC	63.5	1.67	1
32	Simpson-Golabi-Behmel Syndrome, Type 1; SGBS1	62.98	2	0.00321
33	Johanson-Blizzard Syndrome; JBS	61.64	2	0.4
34	Waardenburg Syndrome	59.26	3	0.37
35	Rothmund-Thomson Syndrome; RTS	59.26	2	0.00513
36	Mental Retardation X-Linked 102; MRX102	58.27	1	0.00049
37	Myotonic Dystrophy	54.86	1.33	6.7
38	Alagille Syndrome	54.17	1.33	0.8
39	Larsen Syndrome; LRS	51.85	1.67	0.4
40	Joubert Syndrome	51.49	1.33	1.125
41	Neurofibromatosis, Type I; NF1	50	1	21.3
42	Fetal Alcohol Syndrome; FAS	49.33	2.33	1.6
43	Filippi Syndrome; FLPIS	48.25	1.67	0.00037
44	Mucopolysaccharidosis, Type IIIA; MPS3A	48.21	2	0.32
45	Focal Dermal Hypoplasia; FDH	47.15	1.33	0.00385
46	Miller-Dieker Lissencephaly Syndrome; MDLS	41	1.67	1
47	Holt-Oram Syndrome; HOS	40.56	1	0.7
48	Trisomy 18 Syndrome	39.91	2	16.7
49	Tetrasomy 18p	38.79	1	0.03205
50	Turner Syndrome	32.88	1	5.5

[a] Average of the scores from three clinicians.

[b] Obtained from Orphanet; prevalence is per 100,000 population.

Figures

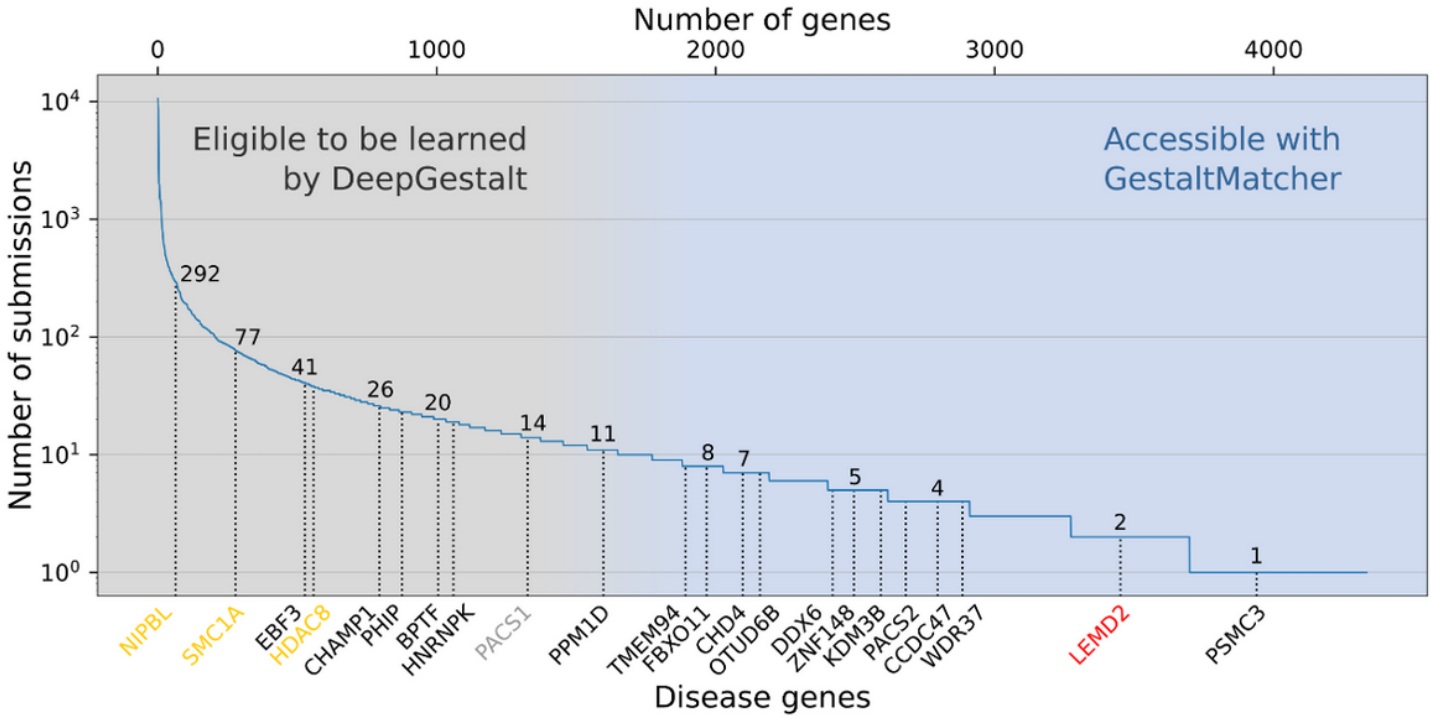

Figure 1

Subsets of disorders supported by DeepGestalt and GestaltMatcher. The lower x-axis shows examples of disease genes, and the upper x-axis is the cumulative number of genes. The y-axis shows the number of pathogenic submissions in ClinVar for each gene. The numbers on the curve indicate the number of submissions for each of the indicated genes. Most of the rare disorders that DeepGestalt supports, have a higher prevalence based on their ClinVar submissions, e.g. Cornelia de Lange syndrome CdLS which is caused by mutation in e.g. NIPBL, SMC1A, and HDAC8. Disease genes such as PACS1, cause highly distinctive phenotypes but are ultra-rare, representing the limit of what current technology can achieve. The first novel disease that was characterized by GestaltMatcher, is caused by mutations in LEMD2. A candidate disease gene with a characteristic phenotype feasible to identification by GestaltMatcher is PSMC3.

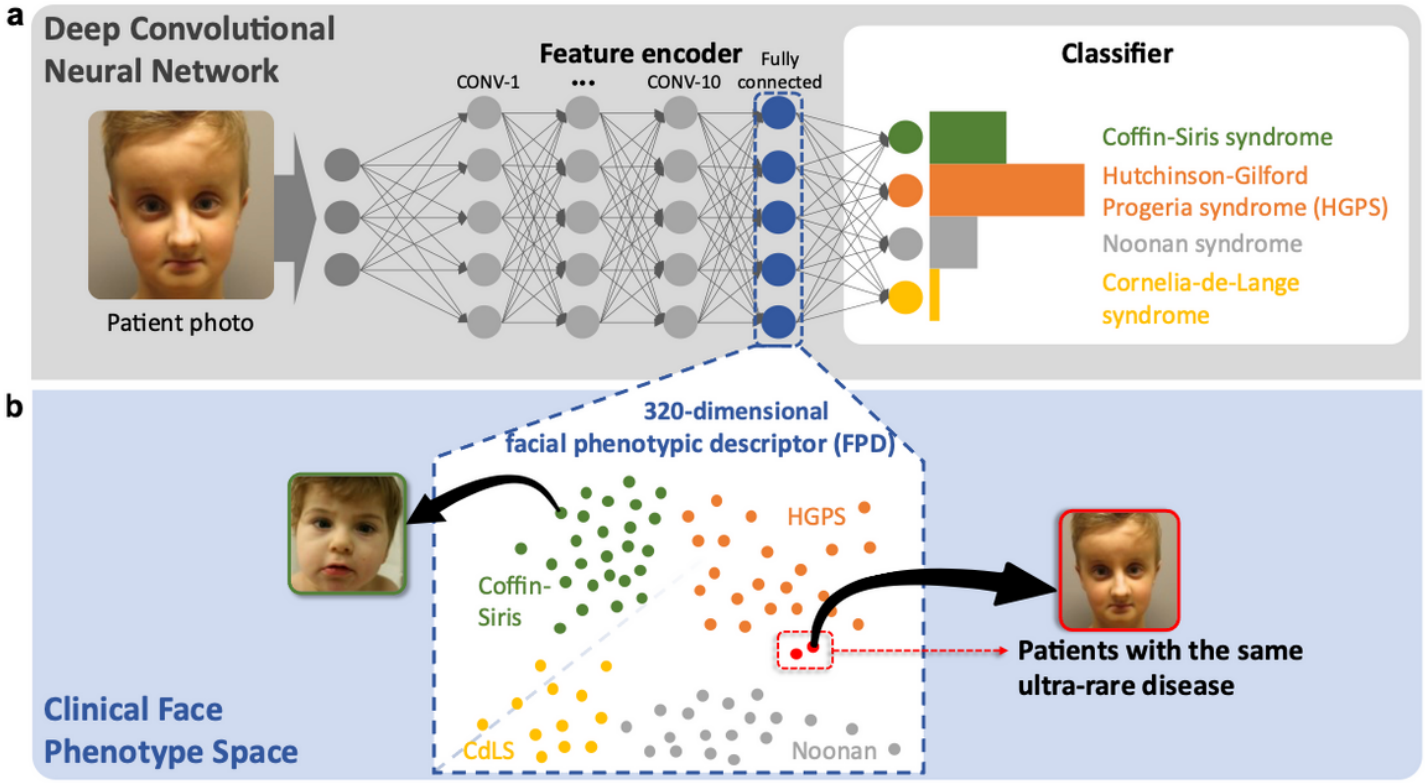

Figure 2

Concept of GestaltMatcher. a, Architecture of a deep convolutional neural network (DCNN) consisting of an encoder and a classifier. Facial dysmorphic features of 296 distinct rare syndromes were used for supervised learning. The last fully connected layer in the feature encoder was taken as a Facial Phenotypic Descriptor (FPD), which forms a point in the Clinical Face Phenotype Space (CFPS). b, In the CFPS, the distance between each patient's FPD can be considered as a measure of similarity of their facial phenotypic features. The distances can be further used for classifying ultra-rare disorders or matching patients with novel phenotypes. Take the input image as an example: the patient's ultra-rare disease, which is caused by mutations in LEMD2, was not in the classifier, but it could match another patient with the same ultra-rare disorder in CFPS.

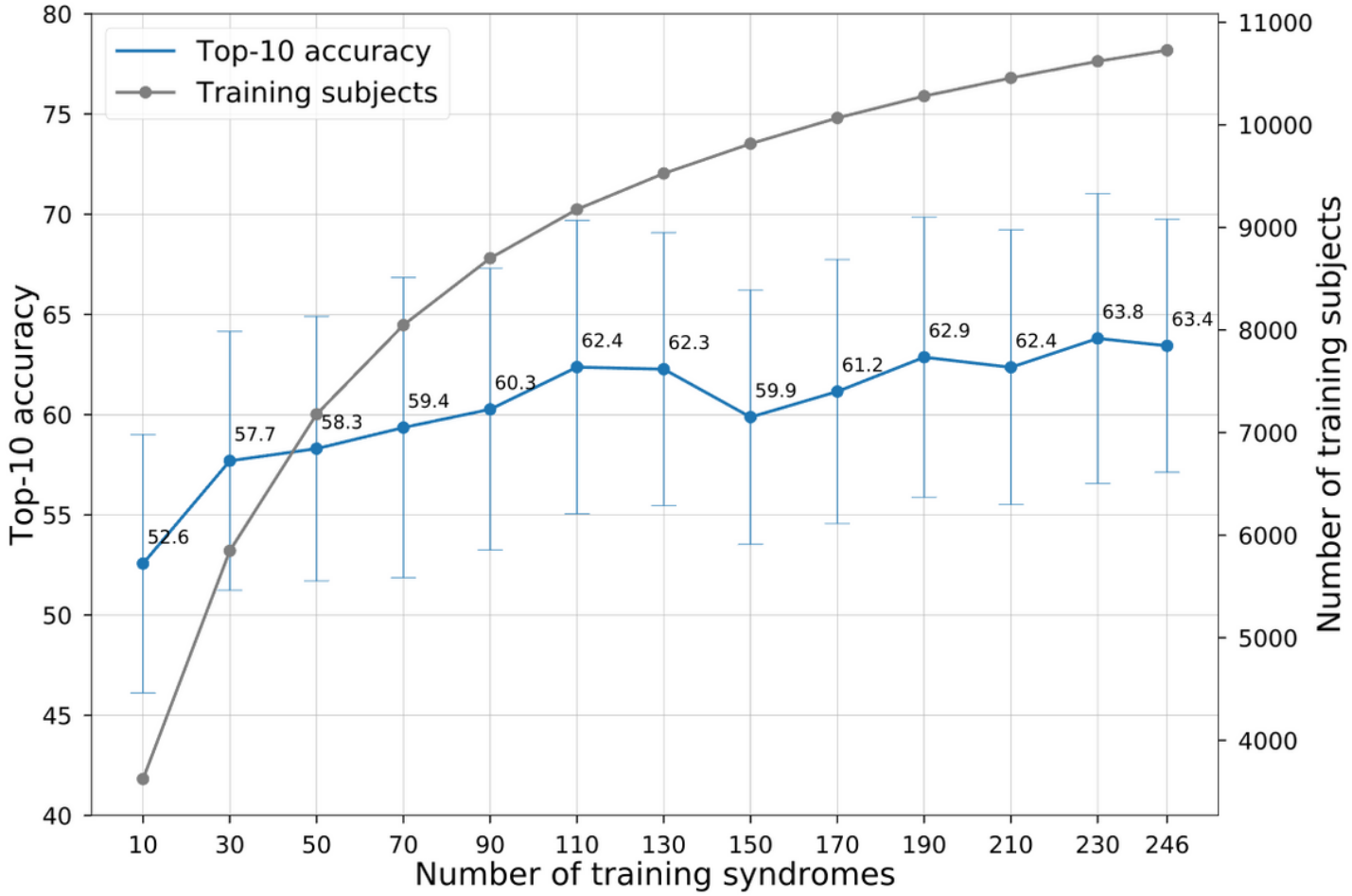

Figure 3

Influence of the number of syndromes included in model training. The x-axis is the number of syndromes used in model training. The left y-axis shows the average top-10 accuracy over 100 iterations, and the error bars show standard deviation. The right y-axis is the cumulative number of subjects in the training syndromes.

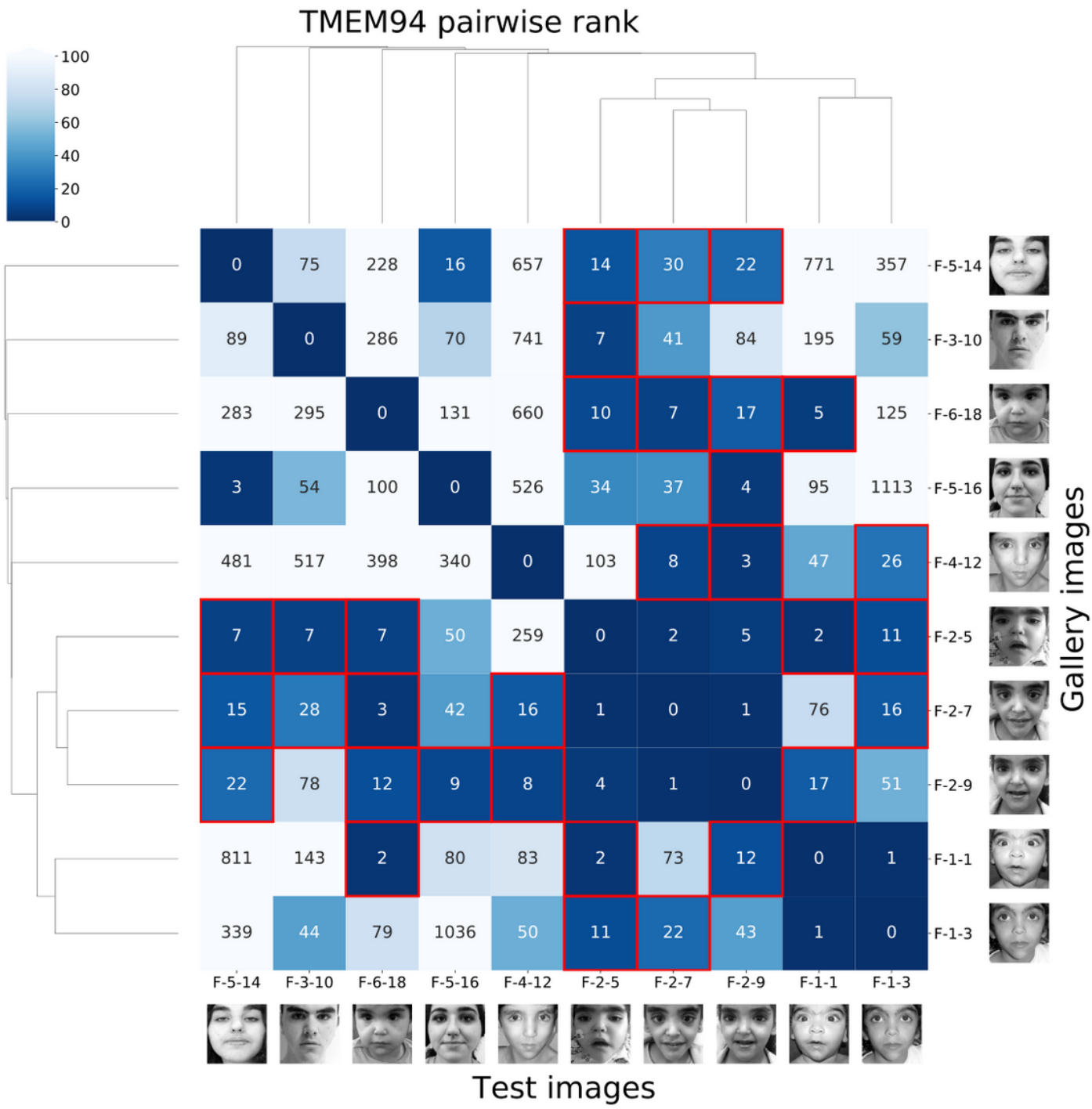

Figure 4

Pairwise ranks of subjects with TMEM94. Each label consists of family numbering and subject numbering, which are the same as in the original publication21. For example, F-2-7 means the seventh subject in the second family. Each column is the result of testing the image indicated at the bottom of the column. The number in the box is the rank or distance to the corresponding image in the gallery. When the rank was less than 30 and the two subjects were from different families, we added a red border to the cell. Let us take F-2-5 as an example. The sixth column starting from the left is the result of testing F-2-5, and the second row from the bottom shows that F-1-1 has a rank of 2 for F-2-5; because 2 is less than 30,

a red border was added. In the third to fifth rows from the bottom are the ranks from family 2, which is the same family that F-2-5 is from, so the cells do not have red borders.

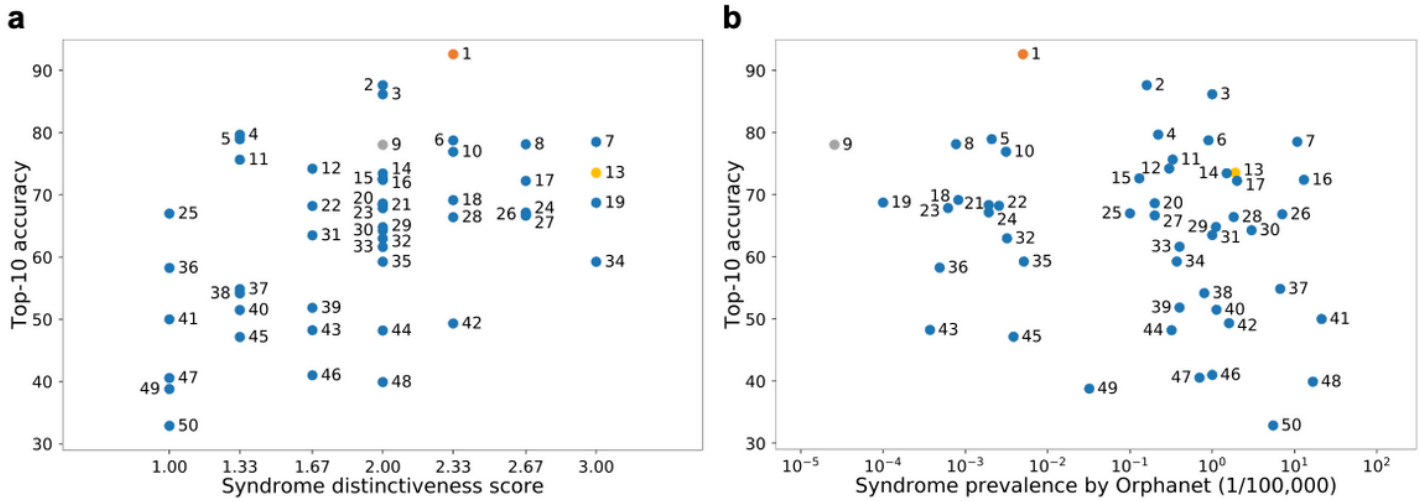

Figure 5

Correlation among syndrome prevalence, distinctiveness score, and top-10 accuracy. a, Distribution of top-10 accuracy and distinctiveness score. The Spearman rank correlation coefficient was 0.421 (P = 0.002). b, Distribution of top-10 accuracy and prevalence. The Spearman rank correlation coefficient was −0.219 (P = 0.126) The details of each syndrome can be found in Table 4 using the syndrome ID shown in the figure; syndrome 9 is Schuurs-Hoeijmakers syndrome. The y-axis shows the average top-10 accuracy of the experiments over those 100 iterations.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- gestaltmatchsupplementarymaterials.pdf
- SupplementaryTable4.xlsx