nature genetics

TECHNICAL REPORT

https://doi.org/10.1038/s41588-021-01010-x

Check for updates

GestaltMatcher facilitates rare disease matching using facial phenotype descriptors

Tzung-Chien Hsieh[1,25], Aviram Bar-Haim[2,25], Shahida Moosa[3], Nadja Ehmke[4], Karen W. Gripp[5], Jean Tori Pantel[1,4], Magdalena Danyel[4,6], Martin Atta Mensah[4,7], Denise Horn[4], Stanislav Rosnev[4], Nicole Fleischer[2], Guilherme Bonini[2], Alexander Hustinx[1], Alexander Schmid[1], Alexej Knaus[1], Behnam Javanmardi[1], Hannah Klinkhammer[1,8], Hellen Lesmann[1], Sugirthan Sivalingam[1,8,9], Tom Kamphans[10], Wolfgang Meiswinkel[10], Frédéric Ebstein[11], Elke Krüger[11], Sébastien Küry[12,13], Stéphane Bézieau[12,13], Axel Schmidt[14], Sophia Peters[14], Hartmut Engels[14], Elisabeth Mangold[14], Martina Kreiß[14], Kirsten Cremer[14], Claudia Perne[14], Regina C. Betz[14], Tim Bender[14,15], Kathrin Grundmann-Hauser[16], Tobias B. Haack[16], Matias Wagner[17,18], Theresa Brunet[17,18], Heidi Beate Bentzen[19], Luisa Averdunk[20], Kimberly Christine Coetzer[3], Gholson J. Lyon[21,22], Malte Spielmann[23], Christian P. Schaaf[24], Stefan Mundlos[4], Markus M. Nöthen[14] and Peter M. Krawitz[1]✉

Many monogenic disorders cause a characteristic facial morphology. Artificial intelligence can support physicians in recognizing these patterns by associating facial phenotypes with the underlying syndrome through training on thousands of patient photographs. However, this 'supervised' approach means that diagnoses are only possible if the disorder was part of the training set. To improve recognition of ultra-rare disorders, we developed GestaltMatcher, an encoder for portraits that is based on a deep convolutional neural network. Photographs of 17,560 patients with 1,115 rare disorders were used to define a Clinical Face Phenotype Space, in which distances between cases define syndromic similarity. Here we show that patients can be matched to others with the same molecular diagnosis even when the disorder was not included in the training set. Together with mutation data, GestaltMatcher could not only accelerate the clinical diagnosis of patients with ultra-rare disorders and facial dysmorphism but also enable the delineation of new phenotypes.

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], and all informative clinical features have to be taken into consideration. A large fraction of patients, particularly those with neurodevelopmental disorders, exhibit craniofacial abnormalities[3]. If the facial phenotype ('gestalt') is highly recognizable, such as in Down syndrome, it may also play an important role in establishing the diagnosis. Sometimes the gestalt is so characteristic or distinct that it reduces the search space of candidate genes or can be used to delineate new phenotype–gene associations[4]. 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

[1]Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany. [2]FDNA Inc., Boston, MA, USA. [3]Division of Molecular Biology and Human Genetics, Stellenbosch University and Medical Genetics, Tygerberg Hospital, Tygerberg, South Africa. [4]Institute of Medical Genetics and Human Genetics, Charité-Universitätsmedizin Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health, Berlin, Germany. [5]A.I. DuPont Hospital for Children/Nemours, Wilmington, DE, USA. [6]Berlin Center for Rare Diseases, Charité-Universitätsmedizin Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health, Berlin, Germany. [7]BIH Biomedical Innovation Academy, Digital Clinician Scientist Program, Berlin Institute of Health at Charité-Universitätsmedizin Berlin, Berlin, Germany. [8]Institute for Medical Biometry, Informatics and Epidemiology, Medical Faculty, University of Bonn, Bonn, Germany. [9]Core Unit for Bioinformatics Data Analysis, Medical Faculty, University of Bonn, Bonn, Germany. [10]GeneTalk, Bonn, Germany. [11]Institut für Medizinische Biochemie und Molekularbiologie (IMBM), Universitätsmedizin Greifswald, Greifswald, Germany. [12]CHU Nantes, Service de Génétique Médicale, Nantes, France. [13]l'Institut du Thorax, INSERM, CNRS, Université de Nantes, Nantes, France. [14]Institute of Human Genetics, University of Bonn, Medical Faculty & University Hospital Bonn, Bonn, Germany. [15]Center for Rare Diseases Bonn, University Hospital Bonn, Bonn, Germany. [16]Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany. [17]Institute of Human Genetics, School of Medicine, Technical University Munich, Munich, Germany. [18]Institute of Neurogenomics, Helmholtz Zentrum München GmbH, German Research Center for Environmental Health, Neuherberg, Germany. [19]Norwegian Research Center for Computers and Law, Faculty of Law, University of Oslo, Oslo, Norway. [20]Department of General Pediatrics, Neonatology and Pediatric Cardiology, Medical Faculty, University Hospital, Heinrich-Heine-University, Düsseldorf, Germany. [21]Department of Human Genetics and George A. Jervis Clinic, NYS Institute for Basic Research in Developmental Disabilities, Staten Island, NY, USA. [22]Biology PhD Program, The Graduate Center, The City University of New York, New York, NY, USA. [23]Institute of Human Genetics, University of Lübeck, Lübeck, Germany. [24]Institute of Human Genetics, Heidelberg University, Heidelberg, Germany. [25]These authors contributed equally: Tzung-Chien Hsieh, Aviram Bar-Haim. ✉e-mail: pkrawitz@uni-bonn.de

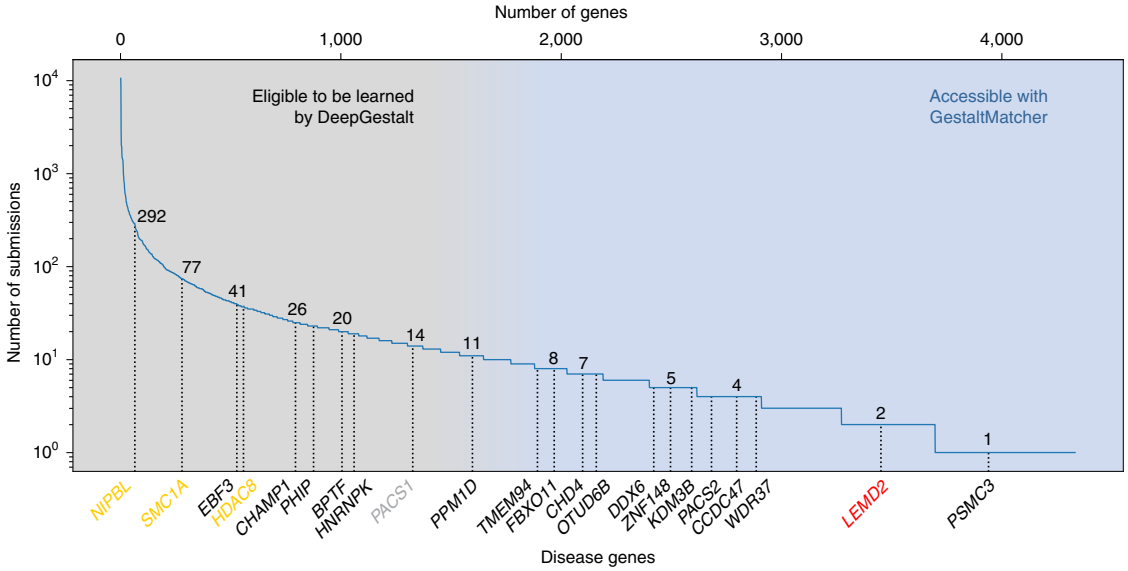


Fig. 1 | Subsets of disorders supported by DeepGestalt and GestaltMatcher. The lower *x* axis shows examples of disease genes, and the upper *x* axis is the cumulative number of genes. The *y* axis shows the number of pathogenic submissions in ClinVar for each gene. The numbers on the curve indicate the number of submissions for each of the indicated genes. Most of the rare disorders that DeepGestalt supports have relatively high prevalence based on their ClinVar submissions, for example, CdLS is caused by a mutation in *NIPBL*, *SMC1A* or *HDAC8* (yellow), among other genes. Disease genes such as *PACS1* (gray) cause highly distinctive phenotypes but are ultra-rare, representing the limit of what current technology can achieve. The first new disease that was characterized by GestaltMatcher is caused by mutations in *LEMD2* (red). A candidate disease gene associated with a characteristic phenotype that can be identified by GestaltMatcher is *PSMC3*.

previously seen a patient with an ultra-rare disorder or if the patient presents with a new disorder, both of which are increasingly common scenarios.

With the rapid development of machine learning and computer vision, a considerable number of next-generation phenotyping tools have emerged that can analyze facial dysmorphology using two-dimensional portraits of patients[5–13]. 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[5]. Since then, facial recognition technologies have improved substantially and constitute the core of the deep-learning revolution in computer vision[14,15]. The current state-of-the-art framework for syndrome classification, DeepGestalt (Face2Gene (F2G), FDNA Inc., USA), has been trained on more than 20,000 patients and currently achieves high accuracy in identifying the correct syndrome for roughly 300 syndromes[12,16]. DeepGestalt has also demonstrated a strong ability to separate specific syndromes and subtypes, surpassing human experts' performance[12]. Hence, pediatricians and geneticists increasingly use such next-generation phenotyping 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 images of at least seven different patients. The number of submissions to diagnostic databases of pathogenic variants, such as ClinVar[17], 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 (Fig. 1). For instance, Cornelia de Lange syndrome (CdLS), which has been modeled by multiple tools[5,12], is caused by mutations in *NIPBL*, *SMC1A* or *HDAC8*, as well as in 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 (Fig. 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. To model a new syndrome in a deep convolutional neural network (DCNN), the developer has to go through six separate steps (Supplementary Fig. 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 new 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[18], 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.

Our objective is to improve phenotypic decision support for rare disorders. 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. The encoder can also be thought of as the penultimate layer of a DCNN that was trained on known syndromes, such as DeepGestalt. The vectors resulting from the encoder 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.

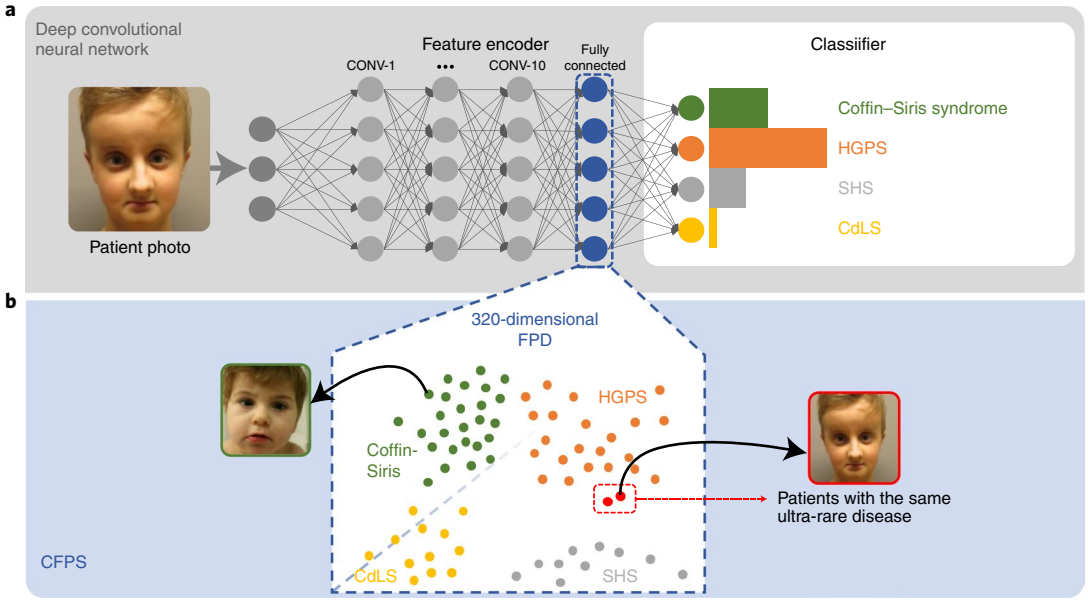


Fig. 2 | Concept of GestaltMatcher. a, Architecture of a DCNN consisting of an encoder and a classifier. Facial dysmorphic features of 299 frequent syndromes were used for supervised learning. The last fully connected layer in the feature encoder was taken as an FPD, which forms a point in the CFPS. **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 new phenotypes. Take the input image shown in the figure as an example: the patient's ultra-rare disease, which is caused by mutations in *LEMD2*, was not in the classifier, but was matched with another patient with the same ultra-rare disorder in the CFPS[4]. CONV-1, convolutional layer-1; CONV-10, convolutional layer-10; HGPS, Hutchinson–Gilford progeria syndrome; SHS, Schuurs-Hoeijmakers syndrome.

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

Overview. The feature encoder of GestaltMatcher computes a Facial Phenotype Descriptor (FPD) for each portrait image (Fig. 2a). Each FPD can be thought of as one coordinate in the CFPS (Fig. 2b). The distances between the FPDs in the CFPS form the basis for syndrome classification, delineation of new phenotypes and patient clustering. All benchmarking results described in this section, as well as those available through the web service, are based on data from F2G. The F2G dataset was used to construct a CFPS consisting of 26,152 images from 17,560 individuals who had been diagnosed with a total of 1,115 different syndromes, each supported by at least two cases. We divided the dataset into two categories: the rare dataset consisting of 816 ultra-rare and new syndromes, representing syndromes that we aim to identify, and the frequent set, consisting of 299 syndromes already identified by DeepGestalt. The latter set of known syndromes was also used to train the encoder. Each category was further split into a gallery (90% of each syndrome) and a test set (the remaining 10% of each syndrome) (Methods). The performance of the three use cases described below, that is, matching patients with diagnosed or undiagnosed individuals, and quantifying syndromic similarity, depends on the composition of the training set and the gallery.

Because F2G data cannot be shared, we also compiled the GestaltMatcher database (GMDB), consisting of 4,306 images from 3,693 individuals with 257 different syndromes. This second dataset is based on 902 publications and additional unpublished cases for which we obtained consent for sharing. All findings described in this section that are based on the F2G data can be reproduced qualitatively in the GMDB data; results obtained with the GMDB data are included in the Supplementary Information.

Training with dysmorphic images improves the performance. To investigate the importance of using a syndromic features encoder rather than a normal facial features encoder, we compared FPDs that are based on the same architecture but trained on different data. The first encoder, which we refer to as Enc-healthy, was only trained on data from healthy individuals in CASIA-WebFace[19]. The second encoder, which we refer to as Enc-F2G, was first trained on the faces of healthy individuals and then fine-tuned by training on dysmorphic faces from the gallery of patients with frequent syndromes. All images were encoded separately for each encoder. We then evaluated the performance of the encoders on test sets of syndromes from the frequent set and from the rare set. 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 Enc-F2G performed better in the matching process than those created with Enc-healthy (Table 1). The features created by Enc-F2G improved the accuracy of matching within the top-10 closest images from 31.46% to 49.12% for the frequent category and from 21.77% to 29.56% for the rare syndromes, which do not overlap with the frequent syndromes. This emphasizes the importance of training the encoder on data from faces with dysmorphic phenotypes and not only on healthy faces. The larger relative improvement of 56% on the frequent test set versus 36% for the rare set could possibly be explained as Enc-F2G being better suited to encode syndromes of the frequent set because it was previously trained on these disorders. Likewise, for some of the 816 new disorders, the characteristic features were not yet optimally represented by Enc-F2G because features of these disorders were not part of the training set.

Table 1 | Performance comparison between classification and clustering with different encoders on sets of known disorders

Test set	Model	Images		Supported syndromes	Null top-1 accuracy	Top-1	Top-5	Top-10	Top-30
		Gallery	Test						
F2G-frequent	Enc-F2G (softmax)	–	2,669	299	0.33%	**35.94%**	**52.45%**	**63.91%**	**78.13%**
F2G-frequent	Enc-F2G	19,950	2,669	299	0.33%	21.06%	39.62%	49.12%	67.98%
F2G-frequent	Enc-healthy	19,950	2,669	299	0.33%	10.69%	23.69%	31.46%	50.80%
F2G-rare	Enc-F2G	2,348.8	1,183.3	816	0.12%	**13.66%**	**23.62%**	**29.56%**	**40.94%**
F2G-rare	Enc-healthy	2,348.8	1,183.3	816	0.12%	9.46%	16.87%	21.77%	31.77%
F2G-frequent	Enc-F2G	22,298[a]	2,669	1,115[c]	0.09%	**20.15%**	**37.81%**	**46.85%**	**64.21%**
F2G-frequent	Enc-healthy	22,298[a]	2,669	1,115[c]	0. 09%	9.70%	22.51%	29.80%	48.24%
F2G-rare	Enc-F2G	22,298.8[b]	1,183.3	1,115[c]	0. 09%	**7.07%**	**14.19%**	**17.67%**	**24.41%**
F2G-rare	Enc-healthy	22,298.8[b]	1,183.3	1,115[c]	0. 09%	4.02%	8.84%	11.73%	16.61%

The DCNNs of Enc-F2G (softmax), Enc-F2G and Enc-healthy have the same architecture. Enc-healthy was trained on CASIA-WebFace. Training of Enc-F2G (softmax) and Enc-F2G was also initiated with CASIA-WebFace and further fine-tuned on photos of patients in the F2G-frequent set. The Enc-F2G (softmax) model is the same as Enc-F2G, but using the softmax values of the layer instead of cosine distances between the FPDs in the CFPS. For the top-1 to top-30 columns, the best performance in each set is boldfaced. The numbers of images and syndromes in the rare set are averaged over ten splits. Enc-F2G outperformed Enc-healthy on both types of syndromes, showing the importance of fine-tuning on patient photos for learning facial dysmorphic features. The top-10 accuracy of Enc-F2G only drops by 2.27 percentage points (from 49.12% to 46.85%) after increasing the number of cases in the gallery and almost quadrupling the number of supported syndromes from 299 to 1,115. [a] Number of images in the frequent gallery + rare gallery. [b] Average of ten splits in the frequent gallery + rare gallery. [c] Number of syndromes in the frequent gallery + rare gallery.

The same trend of improvement by fine-tuning on a diverse but smaller set of syndromic photos is also seen with the public GMDB dataset (Enc-GMDB versus Enc-F2G in Supplementary Table 1). These results suggest that an encoder that is fine-tuned on as many syndromic faces as possible, such as DeepGestalt, is a better fit for the task of syndrome classification than one trained only on healthy faces. Moreover, for rare syndromes not previously seen by the encoder, DeepGestalt's FPD provides a better generalization or clustering than the FPD encoded by CASIA-WebFace.

Syndromic diversity improves matching with new phenotypes. Earlier definitions of the FPD were mainly based on training a network with a small selection of common and highly characteristic syndromes[5,9]. In principle, we could train GestaltMatcher's encoder on all 1,115 different syndromes in our dataset. However, most of the facial 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 the encoder's fine-tuning by incrementally increasing their number starting with the most frequent ones. Due to the imbalance in prevalence among the disorders added each time, the improvement could be affected by the additional number of training subjects. Therefore, we used the same number of subjects for each syndrome. In this section, the test set consists only of disorders from the rare set that the encoder has not seen. The training procedure and averaging of the readout are described in detail in the Methods.

When we increased the number of training syndromes, the accuracy increased (Fig. 3). In general, the performance was also higher when more individuals per syndrome were used for training. Particularly when more than 50 syndromes were used, the curve for training with 20 subjects per syndrome was above the curve for 10 subjects per syndrome, and so on. The same trend is also shown in the public GMDB dataset (Supplementary Figs. 2 and 3).

Moreover, using double the number of syndromes is better than using double the number of subjects for most of the combinations (Supplementary Fig. 4), and the effect of doubling the number of syndromes used for training is greater when the base sample size is larger than 1,200 subjects (Extended Data Fig. 1 and Supplementary Fig. 5). Both of these findings suggest that increasing the syndromic diversity in the training set improves the performance for new disorders. However, in the real-world scenario, the numbers of subjects per syndrome are not imbalanced. Therefore, we also tested the effect of syndromes with fewer cases and found that they contributed only marginally to the performance (Supplementary Note and Extended Data Fig. 2). In the following section, the Enc-F2G encoder is based on the 299 previously described syndromes.

Comparing performance between GestaltMatcher and DeepGestalt. To validate the GestaltMatcher approach for the first-use case (matching to known syndromes), we first worked with the 323 images of patients with 91 syndromes from the London Medical Database (LMD)[20] that were already used for benchmarking the performance of DeepGestalt[12]. When using the frequent gallery, which contains syndromes that DeepGestalt currently supports, GestaltMatcher achieved 64.30% and 86.59% accuracy within the top-10 and top-30 ranks, respectively, which was lower than the 81.28% top-10 accuracy and 88.34% top-30 accuracy achieved by DeepGestalt with an Enc-F2G softmax approach (Supplementary Tables 2 and 3). However, when we used the gallery of all 1,115 syndromes for GestaltMatcher (frequent + rare), which is a search space that is roughly four times larger, the top-10 and top-30 dropped by only 2.40 percentage points and 5.17 percentage points, respectively (Supplementary Table 2). Moreover, we performed the same evaluation on the F2G-frequent test set and the GMDB-frequent test set and obtained similar results. When the number of syndromes in the gallery was increased from 299 to 1,115, the top-10 and top-30 also dropped slightly, by 2.27 and 3.77 percentage points, for the F2G-frequent test set (Table 1). The results with the GMDB-frequent test set also dropped only slightly while supporting more than twice the number of syndromes (Supplementary Table 1). These results indicate that the GestaltMatcher clustering approach is highly scalable and robust to adding new disorders, without the limitations of a classification approach.

Matching undiagnosed patients from unrelated families. In the second use case, we envision GestaltMatcher as a phenotypic complement to GeneMatcher[21]. 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 15 recent GeneMatcher publications with titles containing the phrase 'facial dysmorphism'[4,22–35]. In contrast to the benchmarking of the previous section, the gallery now consists of individuals with rare syndromes

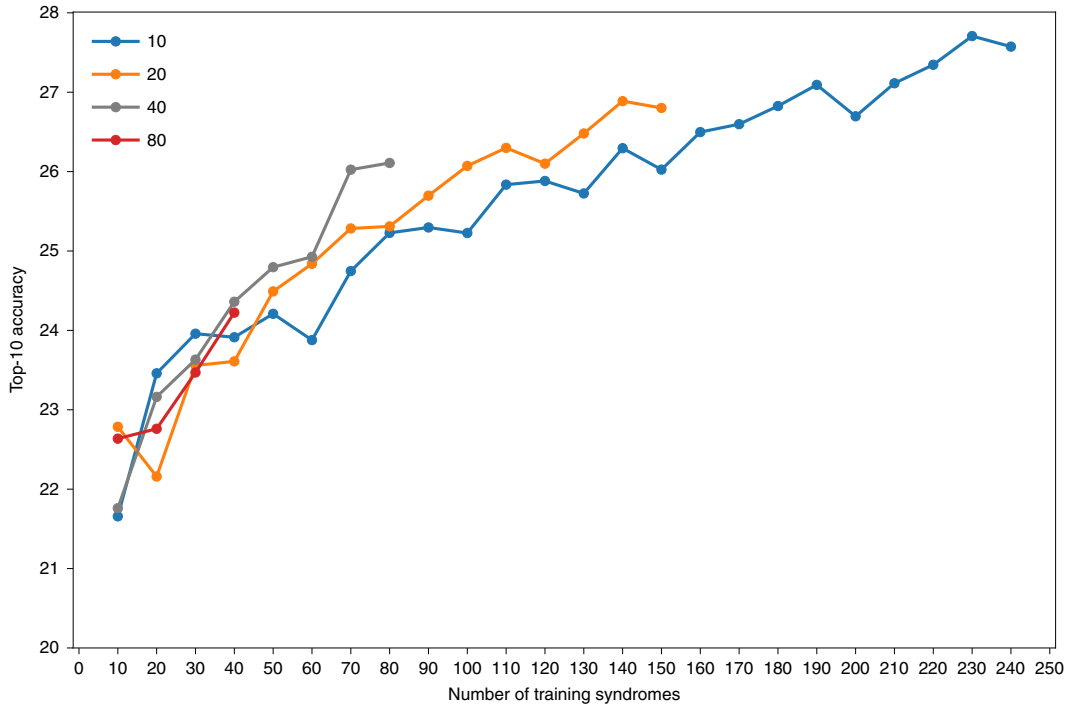


Fig. 3 | Influence of the number of syndromes included in model training. The *x* axis is the number of syndromes used in model training. The *y* axis shows the average top-10 accuracy of testing images in the rare set. Each line uses the same number of subjects per syndrome, which is shown in the key. For each point, we train the models five times with five different splits and average the results. The null accuracy (the expected value if the encoder returned random predictions) is 1.2% (10 of 816).

to simulate undiagnosed cases and, as a consequence, ranks refer to individuals and not disorders. For the evaluation, we still have to reveal in the end whether or not an individual from the gallery is a match for a test case, and nonmatching cases can harm the performance more when matching to individuals rather than disorders. For instance, if the first matching individual is at rank 30, but the 29 nonmatching individuals with higher similarity to the test case together have only four nonmatching disorders, then this match would contribute to the top-5 accuracy in matching on disorders, as in the previous section, but to the top-30 accuracy in matching to individuals, as in this section. Only the top-1 accuracy remains the same in both benchmarks.

In this scenario, we matched 30 of 91 subjects and connected 26 of 79 families when using the top-10 criterion (Table 2, Fig. 4 and Supplementary Fig. 6). When using the top-30 rank, 48 of 91 subjects were matched, and 40 of 79 families were connected. Enc-healthy, which is trained only with healthy individuals, matched only 40 out of 91 subjects and connected 34 out of 79 families using the top-30 rank (Supplementary Table 4). Hence, using the encoder trained with facial dysmorphic individuals improves the matching considerably.

As an example, in a study of *TMEM94* (ref. [33]), eight of the ten photos in six different families were matched, and five of 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 (Fig. 4). The youngest brother, F-2-5, matched families 1, 3, 5 and 6, and one sister, F-2-7, matched families 1, 4 and 6. Another sister, 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 (National Institutes of Health Undiagnosed Diseases Network) and Switzerland, indicating that GestaltMatcher can also connect patients of different ancestries. However, a more systematic analysis of pairwise distances still revealed considerably smaller distances between subjects with de novo mutations and their affected family members than between these subjects and unrelated individuals (Extended Data Fig. 3). This reflects similarities in the nonclinical features of the face, which is also higher within the same ancestry group and is a known confounding factor for the GestaltMatcher approach. However, it is a bias that can be attenuated[36], and will also diminish over time when more diverse training data become available[37].

GestaltMatcher and human experts agree on distinctiveness. We hypothesized that some of the ultra-rare disorders that were linked to their disease-causing genes early on, such as Schuurs-Hoeijmakers syndrome in 2012 (ref. [38]), have particularly distinctive facial phenotypes. To systematically analyze the dependence of disease-gene discovery on the distinctiveness of a facial gestalt, we asked three expert dysmorphologists (S. Moosa, N.E. and K.W.G.) to grade 299 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 dysmorphologists agreed on the same score for 195 of 299 syndromes, yielding a concordance of 65.2%. We then selected 50 syndromes as a test set and trained the model with the remaining 249 syndromes. We analyzed the correlation of the mean of the distinctiveness score from human experts with the top-10 accuracy that GestaltMatcher achieves for these syndromes without having been trained on them (Fig. 5a and Supplementary Table 5). The Spearman's rank correlation coefficient was 0.400 ($P=0.004$), 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 among the best-performing syndromes in GestaltMatcher. The analysis on 20 selected syndromes from the GMDB dataset also showed a positive correlation between distinctiveness score and top-5 accuracy (Supplementary Fig. 7 and Supplementary Table 6).

Table 2 | Matching of new phenotypes on a GeneMatcher validation set

Gene	Total families (subjects)	Connected families (subjects)[a]	
		Top-10	Top-30
BPTF (ref. [22])	6 (6)	0 (0)	2 (2)
CCDC47 (ref. [23])	4 (4)	0 (0)	0 (0)
CHAMP1 (ref. [24])	4 (4)	2 (2)	4 (4)
CHD4 (ref. [25])	3 (3)	0 (0)	0 (0)
DDX6 (ref. [26])	4 (4)	4 (4)	4 (4)
EBF3 (ref. [27])	6 (7)	0 (0)	0 (0)
FBXO11 (ref. [28])	17 (17)	5 (5)	9 (9)
HNRNPK (ref. [29])	3 (3)	3 (3)	3 (3)
KDM3B (ref. [30])	9 (9)	0 (0)	2 (3)
LEMD2 (ref. [4])	2 (2)	2 (2)	2 (2)
OTUD6B (ref. [31])	4 (9)	3 (4)	3 (6)
PACS2 (ref. [32])	6 (6)	0 (0)	2 (2)
TMEM94 (ref. [33])	**6 (10)**	**5 (8)**	**6 (10)**
WDR37 (ref. [34])	4 (4)	2 (2)	3 (3)
ZNF148 (ref. [35])	3 (3)	0 (0)	0 (0)
Total	79 (91)	26 (30)	40 (48)
Average	-	32.91% (32.97%)	50.63% (52.75%)

In the discovery mode for new phenotypes (second use case), all cases in the gallery are without diagnosis. For the performance readout, only the correct disease gene of a match is revealed. As an example, for individuals of the *TMEM94* study (shown in bold in the table), eight out of ten subjects had an image from another family within the top-10 rank, and five of the six families had at least one subject from another family in their top-10 rank. All subjects and families matched within the top-30. This table is based on the ranks from the similarity matrices in Fig. 4 and Supplementary Fig. 6. The accuracy of connected subjects corresponds to the accuracy of using Enc-F2G on the F2G-rare test set (shown in Table 1), but in discovery mode in a gallery of almost the same size as the F2G-rare gallery set. [a] Number of families (subjects) matched by a photo from another family in the top-10 or top-30 rank.

The correlation for GestaltMatcher accuracy and disease prevalence was not significant ($P=0.130$; Fig. 5b). This also means that ultra-rare disorders share a similar distribution of distinctiveness with more common ones, which is important for estimates about the performance of GestaltMatcher on new phenotypes in the real world.

Characterization of phenotypes in the CFPS. When syndromologists cannot find a molecular cause for a patient's phenotype in diagnostic-grade genes after extensive work-up in the laboratory, it becomes a research case, and they may compare the patient's condition to known disorders. For example, a potentially new phenotype could be described as 'syndrome *XY*-like' to build a case group for further molecular analysis through genome sequencing. In GestaltMatcher, this is the third use case, and such comparisons can be supported by cluster analysis in the CFPS with the cosine distance as a similarity metric (Supplementary Table 7).

If a new disease gene has been identified and the similarities of the patients to known phenotypes outweigh the differences, Online Mendelian Inheritance in Man (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 or cohesinopathies for CdLS. For our cluster analysis, we sampled individuals in our database with subtypes of four large phenotypic series and found high intersyndrome separability in addition to considerable intrasyndrome substructure in Noonan syndrome, CdLS, Kabuki syndrome and mucopolysaccharidosis. A *t*-distributed stochastic neighbor embedding (*t*-SNE)[39] projection of the FPDs into two dimensions yielded the best visualization results (Extended Data Fig. 4). Although any projection into a smaller dimensionality might cause a loss of information, the clusters are still clearly visible for the 743 individuals 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 syndromes from the frequent set with the largest number of subjects and 872 images of ten nondistinct syndromes (syndromes without facial dysmorphism) into two-dimensional space. In addition, we calculated the Silhouette index[40] for both of these datasets. The FPDs of the frequent syndromes showed ten clear clusters of subjects, but the *t*-SNE projection of subjects with nondistinct syndromes created no clear clusters (Extended Data Fig. 5). Moreover, the Silhouette index of the frequent syndromes (0.11) was higher than that of the nondistinct syndromes (−0.005); the negative Silhouette index indicates poor separation of the nondistinct 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 undiagnosed and unrelated patients in the CFPS who share a molecular abnormality until statistical significance is reached. We illustrate this process for the new disease gene *PSMC3* in a demonstration on the GestaltMatcher web service (Extended Data Fig. 6, www.gestaltmatcher.org). Ebstein et al.[41] report 22 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 *PSMC3* patients have the same facial phenotype, the proximity of two unrelated patients in the CFPS who share the same de novo *PSMC3* mutation is exceptional. Their distance is comparable to the pairwise distances of patients with the recurring 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 12 out of

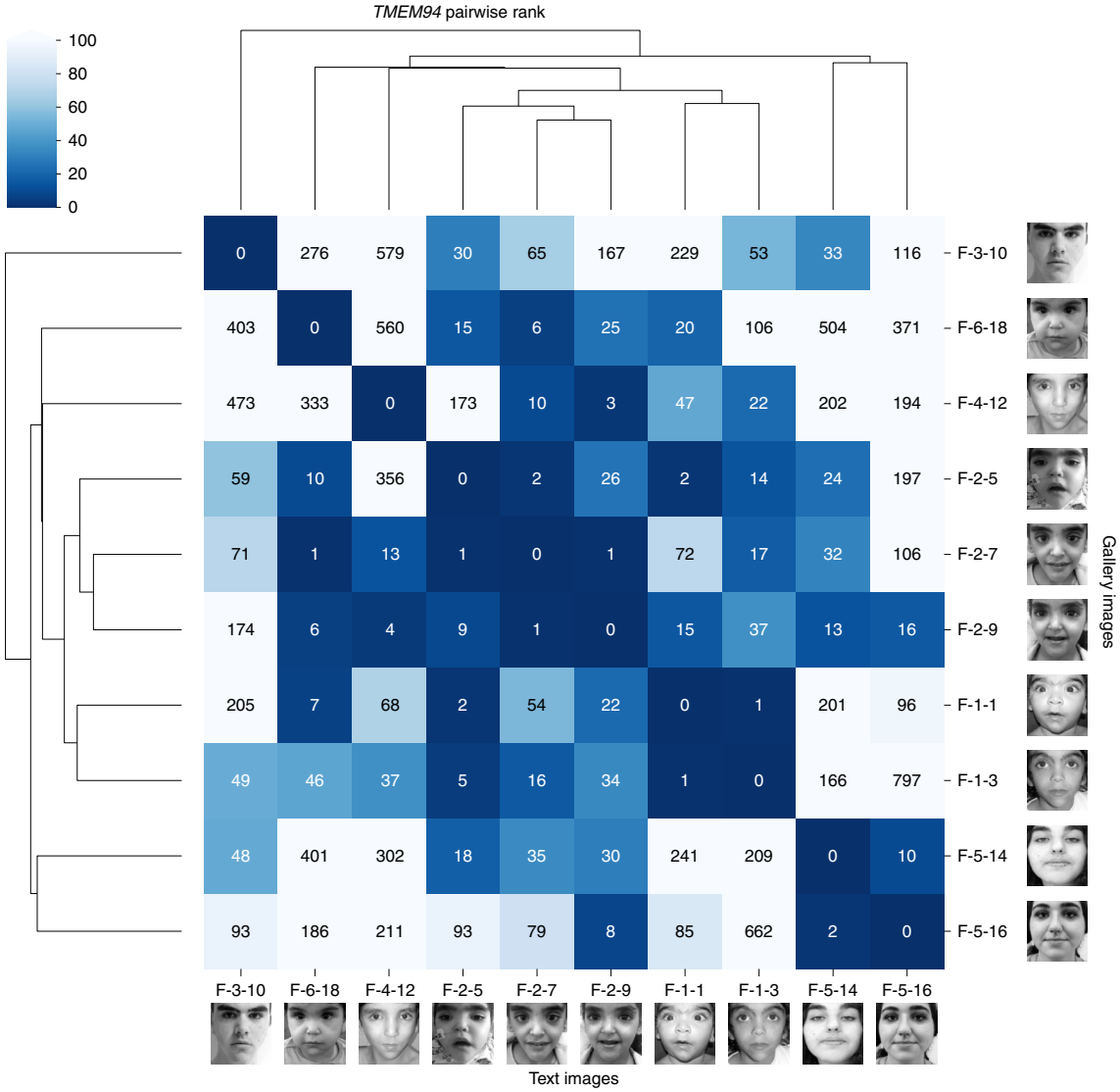


Fig. 4 | Pairwise ranks of individuals with mutations in *TMEM94*. Each label consists of family numbering and subject numbering, which are the same as in the original publication[33]. For example, F-2-7 means the seventh subject in the second family. Each column is the result of testing the image indicated at the bottom of the column. The number in the box is the rank to the corresponding image in the gallery. The fourth column starting from the left is the result of testing F-2-5, and the fourth row from the bottom shows that F-1-1 has a rank of 2 for F-2-5. In the fifth to seventh rows from the bottom are the ranks from family 2, which is the same family that F-2-5 is from.

22 patients in the CFPS for which portraits were available also hint 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, that is, those for which no patient is included in the training set, distinguishes it from other approaches. Matching of unseen syndromes is not only of importance for identifying ultra-rare disorders but can also be useful for the discovery of new diseases. Thus, GestaltMatcher could also speed up the process of delineating new disorders.

Importantly, GestaltMatcher provides the flexibility to easily scale up the number of supported syndromes or the number of unsolved cases without substantial loss in performance. The LMD validation analysis revealed that the use of the softmax approach, that is, classification based on the values of the last layer representing disorders, outperformed GestaltMatcher. However, the GestaltMatcher encoder, that is, clustering in the CFPS with values of the penultimate layer representing features, demonstrated high scalability by yielding similar performance when the number of supported syndromes was increased from 299 to 1,115. Furthermore, the distinctiveness of a syndrome correlated with the performance (Fig. 5a), whereas syndrome prevalence did not (Fig. 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 new syndromes (Supplementary Fig. 1). Instead, matching can be offered instantly for all unsolved cases with available frontal images, as long as consent has been provided for inclusion in the tool. If the gallery is populated by cases with a disease-causing mutation in a diagnostic-grade gene, we consider this a diagnostic work-up. In contrast, if the gallery is populated by further undiagnosed cases, it is a use case comparable to GeneMatcher.

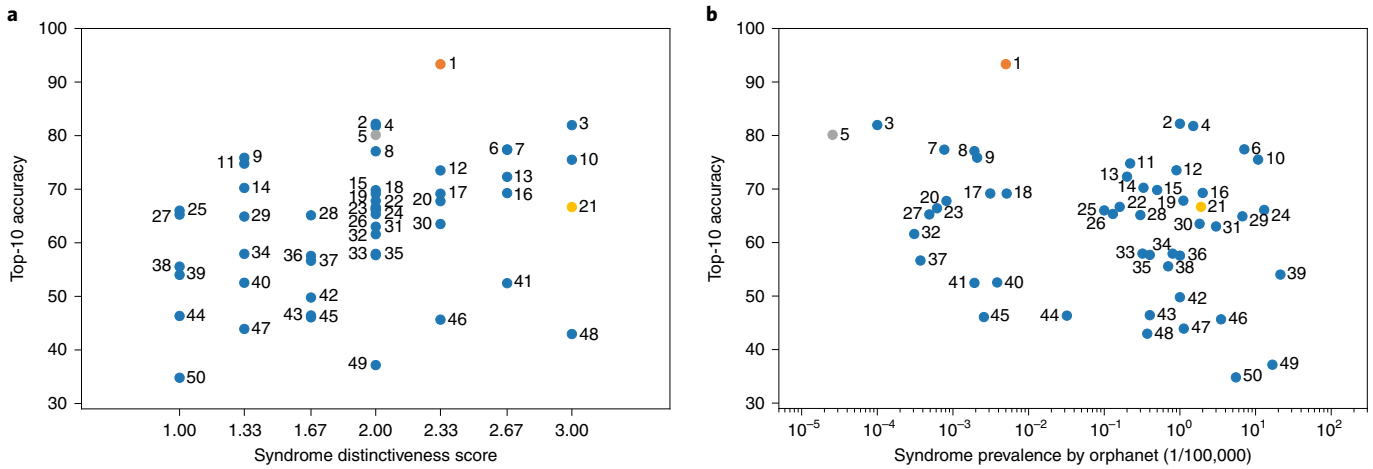


Fig. 5 | Correlation among syndrome prevalence, distinctiveness score and top-10 accuracy. a, Distribution of top-10 accuracy and distinctiveness score. The Spearman rank correlation coefficient was 0.400 ($P = 0.004$). **b**, Distribution of top-10 accuracy and prevalence. The Spearman rank correlation coefficient was −0.217 ($P = 0.130$) The details of each syndrome can be found in Supplementary Table 5 using the syndrome ID shown in the figure; syndrome 5 is Schuurs-Hoeijmakers syndrome. The *y* axis shows the average top-10 accuracy of the experiments over 100 iterations.

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 (that is, 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, for example, 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 ancestry, disease-causing genes or age) to further customize the evaluation.

One of the key features of GestaltMatcher is the ability to match patients and quantify their syndromic similarity. Clinician scientists often face two different tasks in their daily practice: (1) Assessing whether the patient's phenotype is specific for a known disorder. If, for example, a variant of unclear clinical significance is found in a diagnostic-grade gene, a match in GestaltMatcher would be considered as supporting evidence for the pathogenicity[42,43]. (2) Assessing whether the phenotypic similarity of an unsolved case to other individuals also lacking a diagnosis is high enough to form a case group that can be further analyzed. This could, for example, result in the identification of potentially deleterious variants in a new disease gene and would represent the phenotypic complement to existing matching approaches on the molecular level. Several online platforms, such as GeneMatcher, MyGene2 (https://mygene2.org/MyGene2) and Matchmaker Exchange[44], already allow physicians to look for similar patients based on sequencing information, and over the past few years these platforms have enabled the matching of thousands of patients. However, automated facial matching technology has not yet been included in any of these platforms, although phenotypic data, for example, encoded in Human Phenotype Ontology terms, are usually exchanged after contact has been established.

Since its first proof of concept, in which GestaltMatcher was used to identify two unrelated patients from different countries with the same new disease caused by the same de novo mutation in *LEMD2* (ref. [4]), our approach has successfully been applied to other ultra-rare disorders (Fig. 1). We matched 40 of 79 different families in 15 GeneMatcher publications by top-30 rank (Fig. 4 and Supplementary Fig. 6), and 11 candidate genes are currently under evaluation. This result shows the power and potential of GestaltMatcher to identify new syndromes. Although the number of individuals and the diversity of their phenotypes will affect the performance, cases with a high syndromic similarity will remain matchable due to the high dimensionality of the CFPS.

We therefore hope that GestaltMatcher will be readily integrated into other 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.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-01010-x.

Received: 31 December 2020; Accepted: 16 December 2021;
Published online: 10 February 2022

References

1. Ferreira, C. R. The burden of rare diseases. *Am. J. Med. Genet. 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. C. & Hart, P. S. Genetic studies of craniofacial anomalies: clinical implications and applications. *Orthod. Craniofac. Res.* **12**, 212–220 (2009).
4. 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).
5. Ferry, Q. et al. Diagnostically relevant facial gestalt information from ordinary photos. *eLife* **3**, e02020 (2014).
6. 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).
7. Cerrolaza, J. J. et al. Identification of dysmorphic syndromes using landmark-specific local texture descriptors. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)* 1080–1083 (IEEE, 2016).
8. Wang, K. & Luo, J. Detecting visually observable disease symptoms from faces. *EURASIP J. Bioinform. Syst. Biol.* **2016**, 13 (2016).
9. 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**, 90 (2017).

10. Shukla, P., Gupta, T., Saini, A., Singh, P. & Balasubramanian, R. A deep learning frame-work for recognizing developmental disorders. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* 705–714 (IEEE, 2017).
11. 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).
12. Gurovich, Y. et al. Identifying facial phenotypes of genetic disorders using deep learning. *Nat. Med.* **25**, 60–64 (2019).
13. van der Donk, R. et al. Next-generation phenotyping using computer vision algorithms in rare genomic neurodevelopmental disorders. *Genet. Med.* **21**, 1719–1725 (2019).
14. 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).
15. Huang, G. B., Ramesh, M., Berg, T. & Learned-Miller, E. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. *University of Massachusetts, Amherst, Technical Report* 07–49 (2007).
16. 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).
17. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
18. McKusick, V. A. On lumpers and splitters, or the nosology of genetic disease. *Perspect. Biol. Med.* **12**, 298–312 (1969).
19. Yi, D., Lei, Z., Liao, S. & Li, S. Z. Learning face representation from scratch. Preprint at arXiv [cs.CV], http://arxiv.org/abs/1411.7923 (2014).
20. Winter, R. M. & Baraitser, M. The London Dysmorphology Database. *J. Med. Genet.* **24**, 509–510 (1987).
21. 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).
22. 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).
23. 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).
24. Tanaka, A. J. et al. De novo pathogenic variants in *CHAMP1* are associated with global developmental delay, intellectual disability, and dysmorphic facial features. *Cold Spring Harb. Mol. Case Stud.* **2**, a000661 (2016).
25. 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).
26. 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).
27. 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).
28. 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).
29. 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).
30. 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).
31. 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).
32. 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).
33. 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).
34. 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).
35. 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).
36. Alvi, M., Zisserman, A. & Nellåker, C. Turning a blind eye: explicit removal of biases and variation from deep neural network embeddings. In *Computer Vision – ECCV 2018 Workshops* 556–572 (Springer International Publishing, 2019).
37. Lumaka, A. et al. Facial dysmorphism is influenced by ethnic background of the patient and of the evaluator. *Clin. Genet.* **92**, 166–171 (2017).
38. 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).
39. van der Maaten, L. & Hinton, G. Visualizing data using *t*-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
40. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
41. Ebstein, F. et al. De novo variants in the *PSMC3* proteasome AAA-ATPase subunit gene cause neurodevelopmental disorders associated with type I interferonopathies. Preprint at *medRxiv* https://doi.org/10.1101/2021.12.07.21266342 (2021).
42. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
43. Tavtigian, S. V. et al. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet. Med.* **20**, 1054–1060 (2018).
44. Philippakis, A. A. et al. The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum. Mutat.* **36**, 915–921 (2015).

Methods

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

F2G datasets. We collected images of individuals with clinically or molecularly confirmed diagnoses from the F2G 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 26,152 images from 17,560 individuals with a total of 1,115 syndromes (Supplementary Table 8).

GestaltMatcher was designed to distinguish syndromes with different properties. We separated syndromes by the number of affected individuals and whether they had already been learned by the DeepGestalt model. Extended Data Fig. 7 provides an overview of how the dataset was divided. The current DeepGestalt approach requires at least seven subjects to learn a new syndrome. We first used this threshold to separate the syndromes into 'frequent' and 'rare' syndromes. The objective of our study was to improve phenotypic decision support for 'rare disorders'. However, frequent syndromes that are not associated with facial dysmorphic features cannot be modeled by DeepGestalt. We therefore further selected 299 frequent syndromes that possess characteristic facial dysmorphism recognized by DeepGestalt to use as 'frequent syndromes'. The frequent 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. For rare 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 (Fig. 1). Of the 1,115 syndromes in the entire dataset, 299 were frequent and 816 were rare. DeepGestalt cannot yet be applied to 'rare' syndromes category.

We further divided each of these two datasets into a gallery and a 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 frequent syndrome were used to train the models, and the remaining 10% of subjects were used to validate the DeepGestalt training; the 90% then became the frequent gallery and the 10% were assigned to the frequent test set. For the rare dataset, we performed ten-fold cross-validation. In each syndrome, 90% and 10% of subjects were assigned to the gallery and test set, respectively. The test sets were designed to have the same distribution of distinctiveness as the training sets.

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

Please note that the threshold of seven subjects to divide the dataset into frequent and rare is to compare GestaltMatcher to DeepGestalt, which both use the same training data. We could adjust this threshold higher or even remove this threshold in the future.

GMDB dataset. We collected images of individuals with clinically or molecularly confirmed diagnoses from publications and individuals that gave appropriate informed consent for the purpose of this study. This dataset can be used as a public training and test set for benchmarking and is available at GMDB (www.gestaltmatcher.org).

At the time of the data freeze on 9 June 2021, the dataset consisted of 4,306 images of 3,693 individuals with a total of 257 syndromes from 902 publications (Supplementary Table 8). Six of the 3,693 individuals have not yet been published, but appropriate consent has been obtained. For a fair comparison with the F2G dataset, we performed the data separation in the same way. The dataset was first split by the same threshold (seven subjects) into frequent and rare datasets, giving 139 syndromes in the frequent dataset and 118 syndromes in the rare set. Both datasets were also later separated into gallery and test sets. The data split is shown in Supplementary Fig. 8. Of the 3,693 individuals in GMDB, 963 are also in the F2G dataset. To use the GMDB rare set as the test set for both the GMDB-frequent set and the F2G-frequent set, we made sure that no syndrome was in both the GMDB rare set and the F2G-frequent set (Extended Data Fig. 8).

DeepGestalt encoder. The preprocessing pipeline of DeepGestalt includes point detection, facial alignment (frontalization) and facial region cropping. During inference, a facial region crop is forward passed through a DCNN and ultimately gives 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 (FPD), resulting in a vector of size 320.

DeepGestalt was first trained on images of healthy individuals from CASIA-WebFace[19], and later fine-tuned on a dataset with patient images (F2G or GMDB). The encoder without fine-tuning on patient images was called Enc-healthy. The encoder later trained on 299 frequent syndromes in the F2G dataset was named Enc-F2G. The encoder trained on 139 frequent syndromes in GMDB was named Enc-GMDB. In the following sections, we have several encoders trained on different subsets of the F2G and GMDB datasets. The summary of all the encoders used in this study is shown in Supplementary Table 9. To compare GestaltMatcher and DeepGestalt, we employed a model that uses softmax for predicting syndromes, which we called 'Enc-F2G (softmax)'. This model is the same as Enc-F2G; the only difference is that Enc-F2G (softmax) used softmax in the last layer for prediction, as in DeepGestalt, and Enc-F2G used the cosine distance of FPDs for prediction.

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: CFPS. 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 CFPS, with each FPD a point located in the CFPS, as shown in Fig. 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 (for which the diagnoses were known in this proof-of-concept study) and all of the gallery images. Then, for each test image, if an image from another individual 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 averaging the top-*k* accuracy (percentage of test images with correct matches within the top *k*) of each syndrome to avoid biasing predictions toward the major class. We further compared the accuracy of each syndrome in the frequent and rare syndrome subsets to investigate whether GestaltMatcher can extend DeepGestalt to support more syndromes. To compare its performance on predicting syndromes with that of DeepGestalt, we first performed image aggregation on the syndrome level before calculating top-*k* accuracy, so that only the nearest image of each syndrome was taken into account.

LMD validation analysis. We compiled 323 images of patients diagnosed with 91 frequent syndromes from the LMD publication test set[12,20] and used this as the validation set for frequent 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 frequent syndromes consisting of 19,950 images of patients with 299 syndromes, and a unified gallery consisting of 22,298 images of patients with 1,115 syndromes. We then reported the top-*k* accuracy and compared the results of these three settings (DeepGestalt with softmax, GestaltMatcher with the frequent gallery and GestaltMatcher with the unified gallery).

Rare syndromes analysis. To understand the potential for matching rare syndromes, we trained an encoder, denoted Enc-F2G-rare, on 467 out of 816 rare syndromes with more than two and fewer than seven subjects. Ninety percent of the subjects were used to train Enc-F2G-rare 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-F2G-rare and Enc-F2G using both cosine distance and the softmax classifier.

Matching undiagnosed patients from unrelated families. We selected 15 articles published from 2015 to 2019 in which GeneMatcher was used to establish an association between a gene and a new phenotype with facial dysmorphism in patients from unrelated families. In total, these studies contained 108 photos of 91 subjects from 79 families. The details are shown in Table 2. The 15 genes were not among the F2G-frequent syndromes, so we can consider them each as a new phenotype to the model. We performed leave-one-out cross-validation on this dataset; that is, we kept one photo as the test set, and we assigned the rest of the photos to a gallery of 3,533 photos with 816 rare 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. When using the GeneMatcher data, we did not perform syndrome aggregation because aggregation cannot be performed if the syndrome is not known. Instead, we matched patients rather than predicting disorders.

Syndrome facial distinctiveness score. To evaluate the importance of the facial gestalt for clinical diagnosis of the patient, we asked three dysmorphologists (coauthors S. Moosa, N.E. and K.W.G.) 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 from the facial phenotype alone.

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[45].

Unseen syndromes correlation analysis. To investigate the influence of prevalence and distinctiveness score on the performance of new syndromes with facial dysmorphism, we selected 50 frequent 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 Fig. 9 and Supplementary Table 5. The encoder (Enc-F2G-exclude-50) was trained on 90% of the subjects from the other 249 frequent 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: between top-10 accuracy and the syndrome's distinctiveness score, and between top-10 accuracy and the prevalence of syndromes collected from Orphanet.

The same analysis was also performed on the GMDB dataset. We selected 20 syndromes from GMDB-frequent instead of 50 syndromes because the GMDB dataset is smaller than the F2G dataset, and we trained the Enc-GMDB-exclude-20 on the remaining 119 frequent syndromes. The details of the 20 selected syndromes and the results are reported in Supplementary Table 6. Please note that we report the top-5 accuracy in the GMDB dataset instead of top-10 accuracy because of the smaller number of syndromes in the gallery.

Analysis of number of training syndromes and subjects. In this analysis, we evaluated the influence of training with additional syndromes and subjects to the new disorders. To avoid an imbalance among the syndromes, we used the same number of subjects for each syndrome. We first used four different settings for the number of subjects: 10, 20, 40 and 80. However, some syndromes have fewer subjects than the four settings used for training: for 10, 20, 40 and 80 subjects, there are 242, 156, 84 and 40 syndromes. We then defined the ordering of syndromes we added each time. To add the same syndromes for the four numbers of subjects each time, we first sorted syndromes with the number of subjects in descending order. To avoid bias due to having specific disorders added at each position, we then performed random sorting five times within each of the intervals [1, 40], [41, 80], [81, 150] and [151, 240] to generate five different lists of syndromes. Thus, the ordering from common disorders to rare disorders was by interval rather than by syndrome. For example, Kabuki syndrome might be in the ninth position in the first list, but in the 20th position in the second list, but in each randomly sorted list Kabuki syndrome is in the first interval.

For each of five different lists of training syndromes, we performed the same training described as follows. We first trained X number of syndromes with 10 subjects, where $X = 10$ to 240, incremented at an interval of 10 syndromes. As mentioned above, there are only 156 syndromes with more than 20 subjects. Thus, we trained syndromes with 20 subjects with $X = 10$ to 150 syndromes with the same increment of 10 syndromes. We performed the same process for 40 and 80 subjects, with maximums of 80 and 40, respectively.

For each setting (number of subjects, number of syndromes), we had five models. We then encoded the photos separately with each model and tested them on the rare syndromes, which had not been seen by the models. In the end, we averaged the performance by the five models and report the average as the top-10 accuracy for each setting in Fig. 3. We also used the models described above to encode the GMDB dataset, tested them with the GMDB rare set and report the results in Supplementary Fig. 2.

Because the GMDB dataset is smaller than the F2G dataset, we were not able to use the same number of subjects and syndromes to perform the analysis. For the GMDB dataset, we used 10, 20, 40 for the number of subjects, and syndrome intervals of [1, 10], [11, 40] and [41, 80]. The results of training on GMDB and testing of the GMDB rare set are shown in Supplementary Fig. 3.

We next wanted to compare two scenarios: double the number of training syndromes and double the number of training subjects. For example, we first set training on ten subjects for each of ten syndromes as the base setting, then compared this performance to training ten subjects for each of 20 syndromes (double syndromes) and training 20 subjects for each of ten syndromes (double subjects). The base setting had 100 subjects in total. Double syndromes and double subjects each had 200 subjects. This comparison allows us to understand the different influences of adding more syndromes and adding more subjects. The results are shown in Extended Data Fig. 1 and Supplementary Figs. 4 and 5.

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

Data availability

The data that support the findings of this study are divided into two groups, nonsharable data (F2G) and sharable data (OMIM, CASIA-WebFace, GMDB). F2G data are from Face2Gene users and cannot be shared to protect patient privacy. OMIM data can be downloaded at https://omim.org/downloads. CASIA-WebFace and GMDB are available for noncommercial, research and educational purposes, and subject to controlled access. For CASIA-WebFace, user conditions are available at http://www.cbsr.ia.ac.cn/english/casia-webFace/casia-webfAce_AgreEmeNtS.pdf, and requests should be sent to cbsr-request@authenmetric.com. For GMDB, please contact info@gestaltmatcher.org and specify which analyses you intend to perform. The board of GestaltMatcher will check and respond within 10 business days whether your request is compatible with the user conditions.

Code availability

GestaltMatcher can be subdivided into its algorithmic part, data that are required to train the neural network and a service that can be used for matching patients. The project's landing page, www.gestaltmatcher.org, redirects to separate pages for each category. The web service for matching patients is based on Enc-F2G and is accessible for health care professionals. Parts of this service are proprietary and cannot be shared. However, the architecture of the CNN, as well as the code for evaluation, is available under a creative commons license.

References

45. 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).

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through individual grants to P.M.K. (grant nos. KR 3985/7-3, KR 3985/6-1). M.M.N. and R.C.B. are supported by the DFG through grants under the auspices of the Germany Excellence Strategy (grant nos. EXC2151–390873048, ImmunoSensation2). A. Schmidt received additional support by the BONFOR program of the Medical Faculty of the University of Bonn (grant no. 2020-1A-15). We also acknowledge support from the TRANSLATE-NAMSE project. We are also grateful for the language editing provided by N. Ruff.

Author contributions

N.E., J.T.P., M.D., M.A.M., D.H., S.R., A.K., B.J., H.L., F.E., E.K., S.K., S.B., A. Schmidt, S.P., H.E., E.M., M.K., K.C., C.P., R.C.B., T. Bender, K.G.-H., T.B.H., M.W., T. Brunet, L.A., K.C.C., K.W.G. and G.J.L. collected and managed samples and data. T.-C.H., A.B.-H., G.B., A.H., H.K., S.S. and A. Schmid conducted data analysis. A.B.-H., G.B., T.K. and W.M. developed the software. N.E., K.W.G., D.H., N.F., H.B.B., M.S., C.P.S., S. Mundlos, S. Moosa, M.M.N. and P.M.K. provided intellectual input on clinical dysmorphology and translational, ethical and legal aspects. T.-C.H., A.B.-H., N.F., S. Moosa and P.M.K. wrote the manuscript with input from all authors. P.M.K. conceived and directed the study with input from all authors.

Competing interests

A.B.-H., N.F. and G.B. are employees of FDNA. T.K. is an employee of GeneTalk GmbH. M.A.M. is a participant in the BIH Charité Digital Clinician Scientist Program founded by the late Prof. Duska Dragun and funded by the Charité-Universitätsmedizin Berlin and the Berlin Institute of Health. M.M.N. reported receiving personal fees from the Lundbeck Foundation, Robert Bosch Stiftung, Shire GmbH, Life & Brain GmbH and HMG Systems Engineering GmbH outside the submitted work. The other authors declare no conflicts of interest.

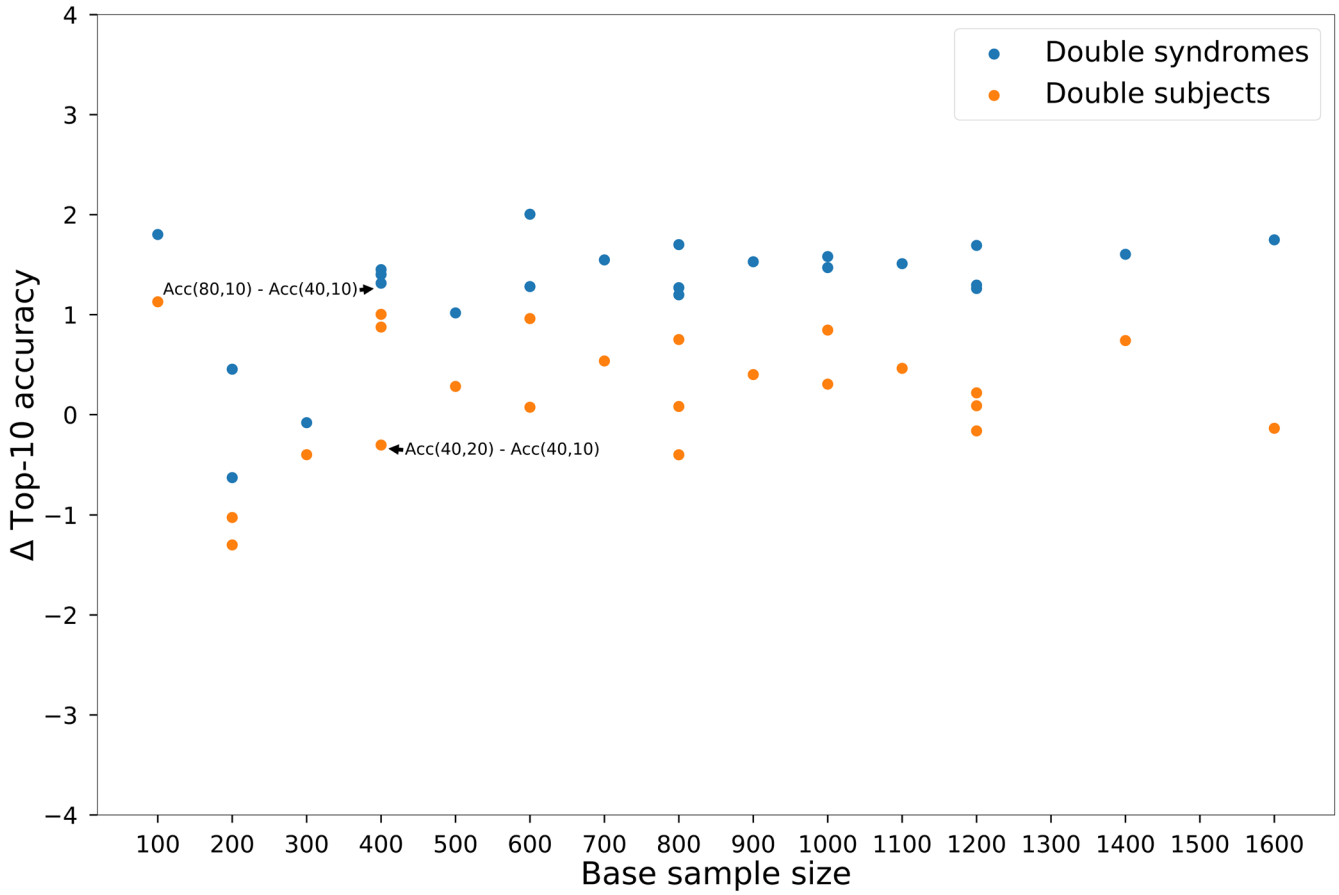
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-021-01010-x.

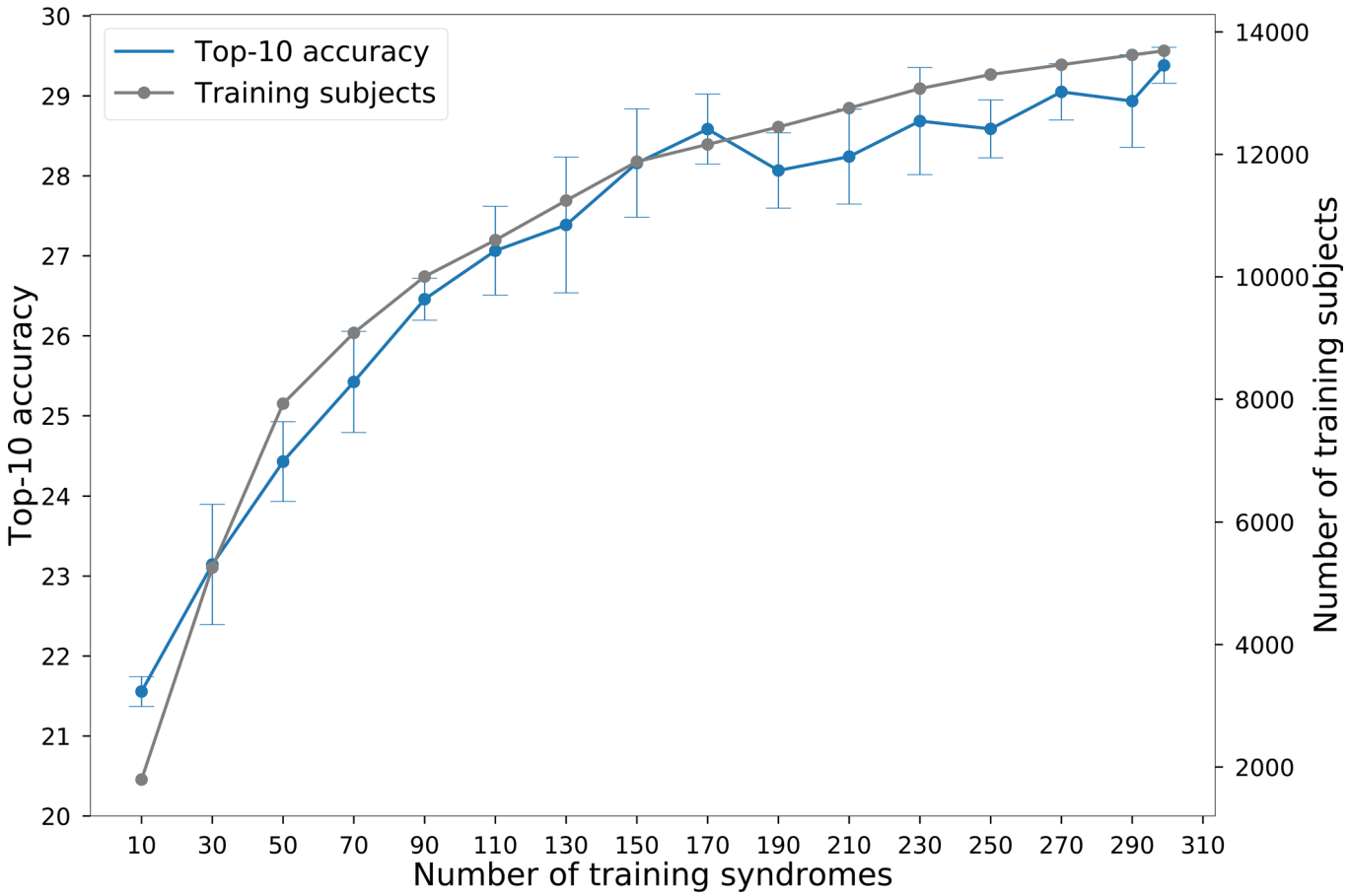
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-01010-x.

Correspondence and requests for materials should be addressed to Peter M. Krawitz.

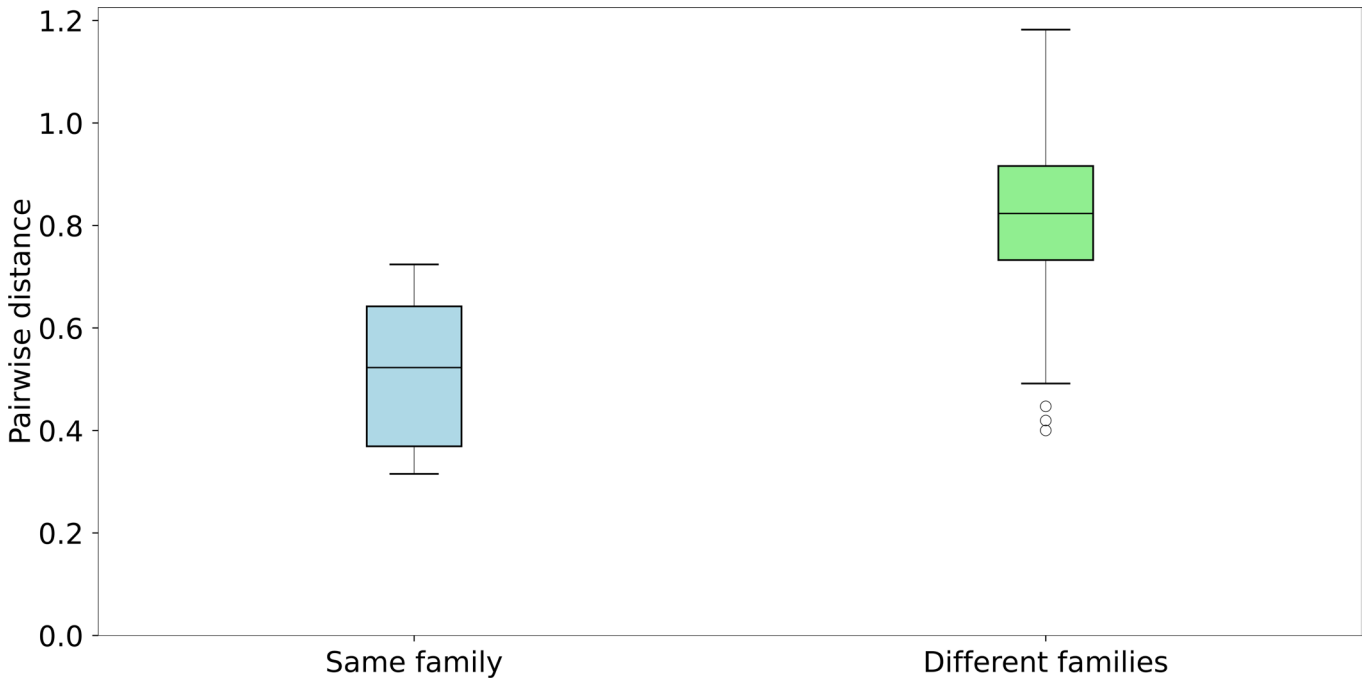
Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.



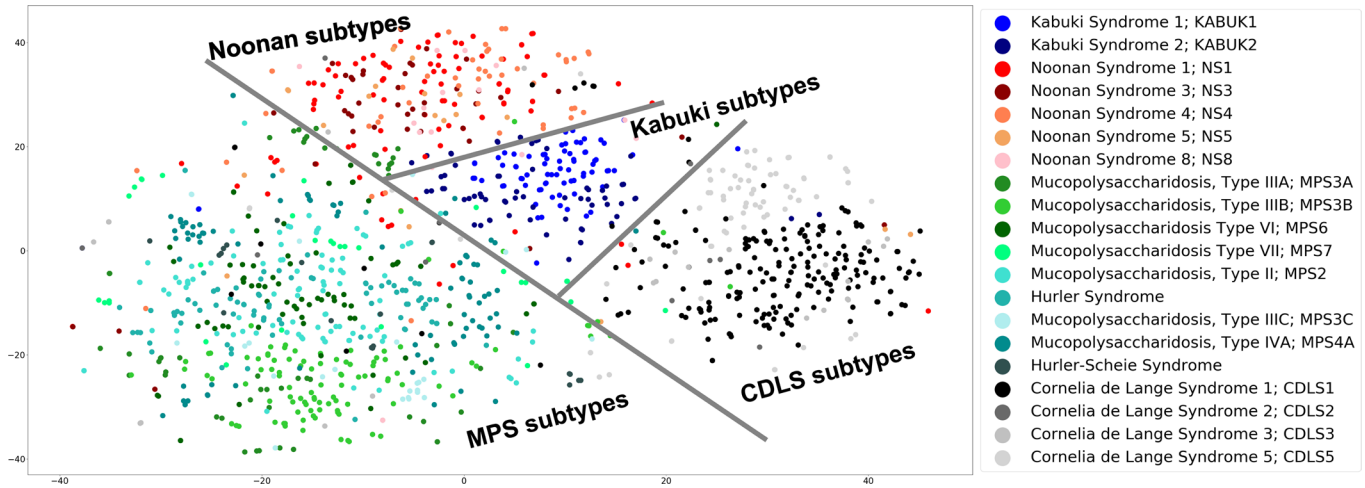
Extended Data Fig. 1 | Performance improvement of double syndromes and double subjects when using different base sample sizes with Face2Gene models and the Face2Gene rare set. Base sample size is calculated as the number of subjects multiplied by the number of syndromes. For example, the point of 40 subjects and 10 syndromes has sample size of 400, and it equals both the point of 10 subjects and 40 syndromes and the point of 20 subjects and 20 syndromes. ΔTop-10 accuracy is the difference of accuracy between the double syndromes or subjects and the base point, and is calculated based on Fig. 3. Take the two points annotated in the figure as two examples. The base point is 10 subjects and 40 syndromes with sample size 400. The upper indicated point is subtracting the point of 10 subjects and 40 syndromes from the point of 10 subjects and 80 syndromes in Fig. 3. The lower point is subtracting the point of 10 subjects and 40 syndromes from the point of 20 subjects and 40 syndromes in Fig. 3. In this graph, doubling the number of syndromes always improves top-10 accuracy more than doubling the number of subjects, particularly at larger base sample sizes. Thus, adding more syndromes is more effective than adding more subjects when enlarging the training set.



Extended Data Fig. 2 | 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 for five models, and the error bars show the standard deviation over five models. The right *y*-axis is the cumulative number of subjects in the training syndromes. Each point is the average of testing five different models with different data splits. The null accuracy is 1.23% (10/816).

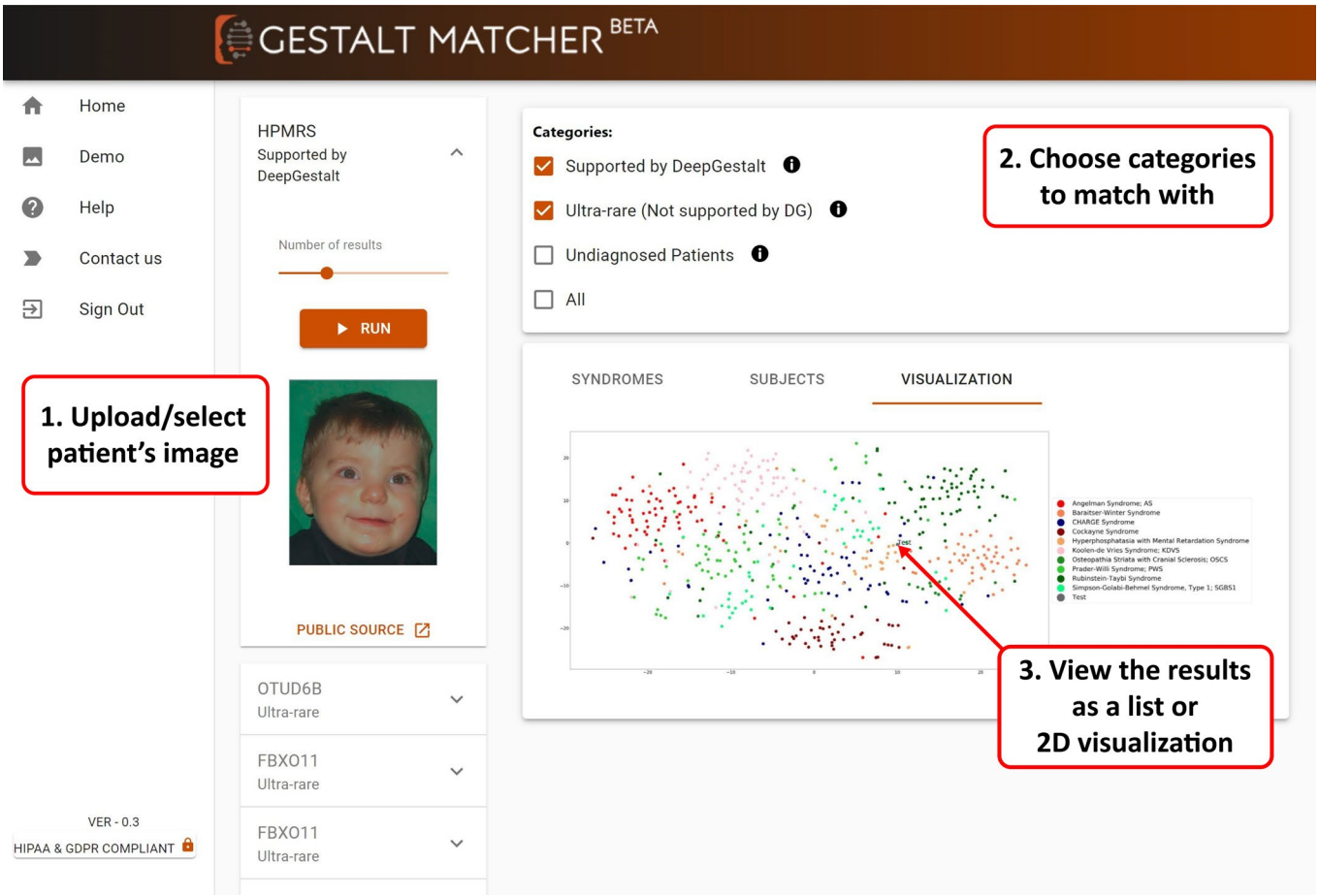


Extended Data Fig. 3 | Comparison of the pairwise distance distribution between subjects in the same family and subjects in different families with the same disease-causing gene. The median distance between affected individuals from the same family is 0.522, and the median distance between individuals from different families is 0.823. In the box plots, the center line indicates the median values, and the bottom and top edge of the box are the first (25%) and the third (75%) quartiles. The whiskers extend the data points outside the 1st to the 3rd quartiles. The total number of data points (*n*) for the same family is 28, and *n* is 928 for the different families.

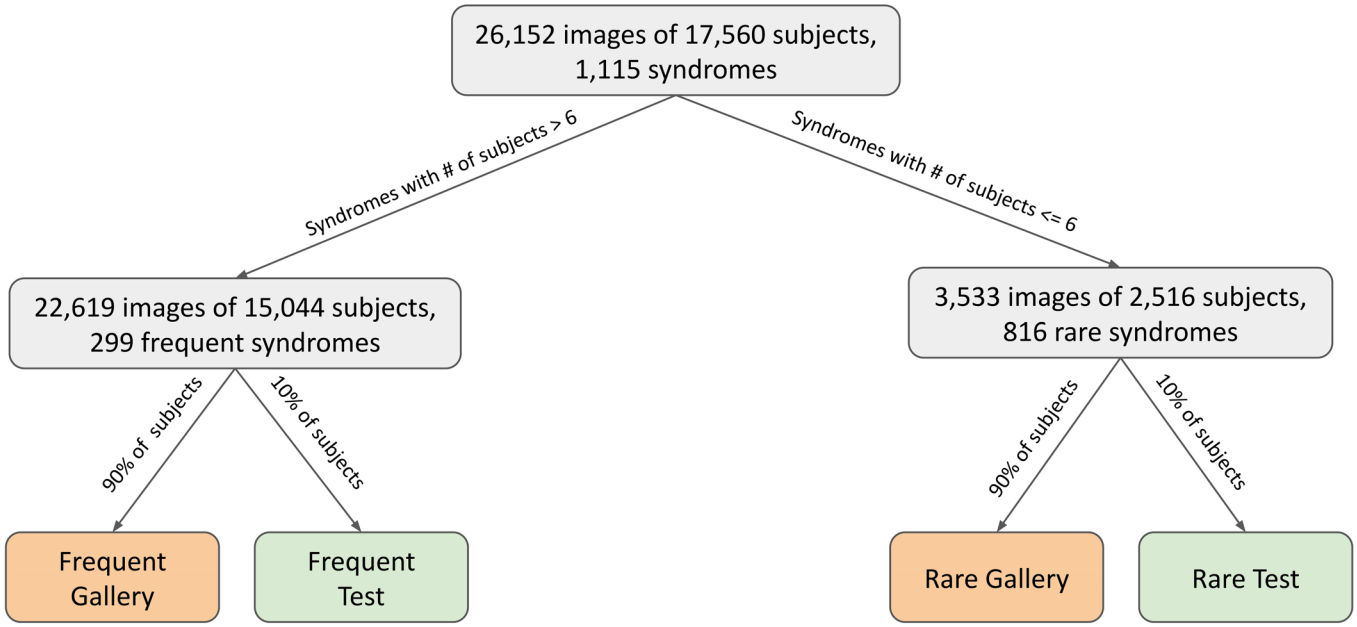


Extended Data Fig. 4 | Hierarchical clustering of four phenotypic series using a *t*-SNE projection of the Facial Phenotype Descriptors. The projection shows clustering of FPDs for Kabuki syndrome, Noonan syndrome, mucopolysaccharidosis, and Cornelia de Lange syndrome.

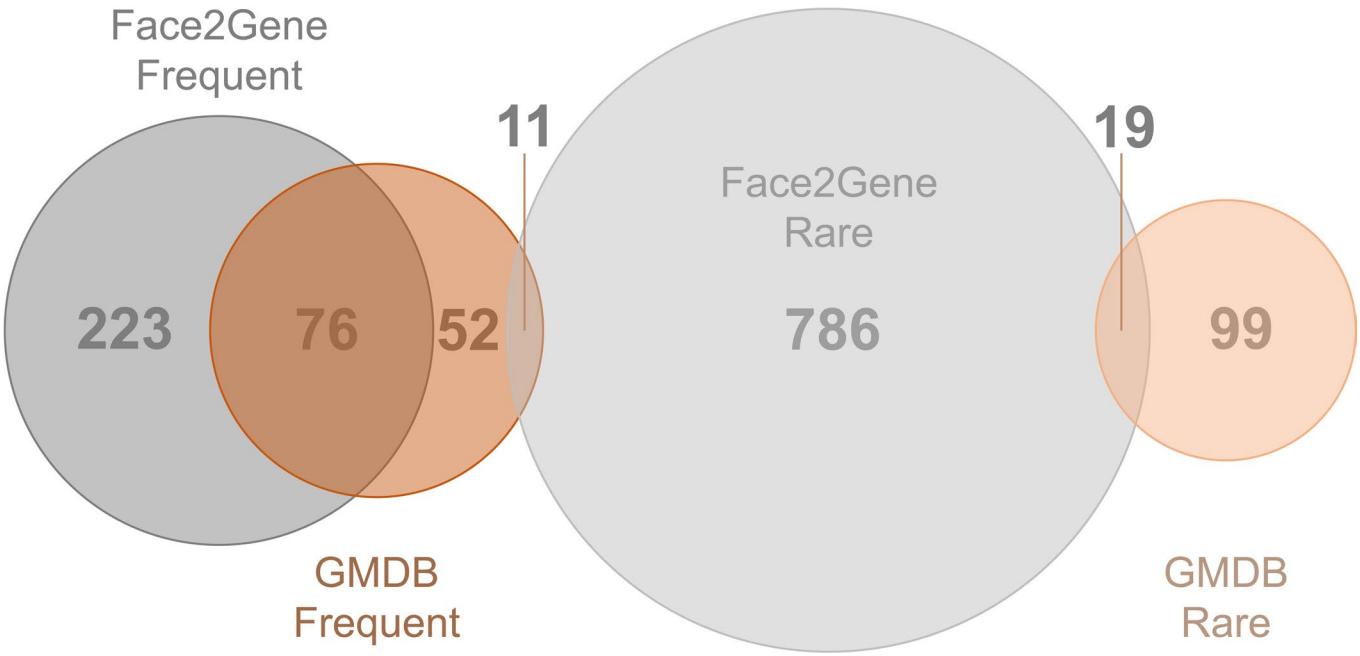
Extended Data Fig. 5 | *t*-SNE visualization of Facial Phenotype Descriptors of syndromes with or without facial dysmorphism. a, Ten syndromes with facial dysmorphism. **b**, Ten syndromes without facial dysmorphism.



Extended Data Fig. 6 | Screenshot of the GestaltMatcher web service. Users can upload a patient photo to match against patients in the selected categories and can also visualize the clustering of patients by *t*-SNE. Access can be requested from www.gestaltmatcher.org. If the category DeepGestalt is selected, only cases with one of the frequent 299 diagnoses that DeepGestalt supports populate the gallery. If category Ultra-rare is chosen, the gallery is populated by cases with one of the 816 diagnoses not supported by DeepGestalt. The category of Undiagnosed Patients is suitable for a research setting if no match with a known disorder could be made (see, for example, *PSMC3* in the online demo).



Extended Data Fig. 7 | Overview of Face2Gene data categorization in GestaltMatcher. The data were first divided by the number of subjects in each syndrome. Syndromes with more than six subjects were denoted frequent syndromes, and those with six or fewer as rare syndromes. Frequent syndromes were also recognized by DeepGestalt. Each category was further divided into a gallery and a test set. For each frequent syndrome, 90% of subjects were assigned to the gallery and used for model training; the remaining 10% of subjects were kept for validating the model training and were sampled in the test set. We performed 10-fold cross-validation on rare syndromes. In each syndrome, 90% of subjects were assigned to the gallery and 10% of subjects were assigned to the test set.



Extended Data Fig. 8 | Venn diagram of numbers of syndromes in the Face2Gene and GMDB datasets. Within each dataset, frequent syndromes are defined as those with seven or more subjects, and rare syndromes are defined as those with six or fewer subjects.

nature research

Corresponding author(s): Peter Krawitz

Last updated by author(s): Nov 1, 2021

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	GestaltMatcher can be subdivided into its algorithmic part, data that is required to train the neural network and a service that can be used for matching patients. The project's landing page www.gestaltmatcher.org redirects to separate pages for each category. The web service for matching patients is based on Enc-F2G and is accessible for health care professionals. Parts of this service are proprietary and cannot be shared. However, the architecture of the CNN, as well as the code for evaluation is available under a creative commons license.
Data analysis	All analysis with data from GeneMatcher has been cited within the main manuscript. See references 22 to 35. For performance comparisons we trained GestaltMatcher encoders and DeepGestalt with multiple splits of data as described in detail in the online methods.

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data that support the findings of this study are divided into two groups, non-sharable data (F2G) and sharable data (OMIM, CASIA-WebFace, GMDB). F2G data is from Face2Gene users and cannot be shared in order to protect patient privacy. OMIM data can be downloaded at https://omim.org/downloads. CASIA-WebFace and GMDB are available for non-commercial, research, and educational purposes, and subject to controlled access. For CASIA-WebFace, user conditions are available at http://www.cbsr.ia.ac.cn/english/casia-webFace/casia-webfAce_AgreEmeNtS.pdf, and requests should be sent to cbsr-request@authenmetric.com. For

GMDB, please contact info@gestaltmatcher.org and specify which analyses you intend to perform. The board of gestaltmatcher will check and respond within ten business days whether your request is compatible with the user conditions.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	In all applications of deep learning the performance usually increases with the size of the data set. In total, we had access to photographs of 21,836 patients with 1,362 different rare diseases. It was the objective of this work to analyze how the setup of these data collections influences the performance of the GestaltMatcher encoder.
Data exclusions	Data of patients with disorders without facial dysmorphism was excluded from further analysis.
Replication	All findings of the analysis of F2G data could be replicated on the GMDB data.
Randomization	Data was randomly split into training (90%) and test (10%) sets. This was done repeatedly to control for covariates.
Blinding	Splitting data into training and test sets is equivalent to blinding data for the machine.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	Male and female individuals of European, African, and Asian ethnic background, aged between 3 months and 41 years were analyzed in this study. Age, sex, and ethnic background can confound the performance. However, since these characteristics are independent of the disease, they do not affect the findings about matching accuracy.
Recruitment	Participants were recruited worldwide via the platform Face2Gene and from the literature. Participants with a European ethnic background are overrepresented, because differences in healthcare systems. This might results in a better matching performance for patients of European descent, than for patients of Asian or African descent.
Ethics oversight	This study is governed by the approval of the following Institutional Review Boards: Charité–Universitätsmedizin Berlin, Germany (EA2/190/16); UKB Universitätsklinikum Bonn, Germany (Lfd.Nr.386/17). The authors have obtained written informed consent from the patients or their guardians, including permission to publish photographs.

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

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com