

# **Next-generation phenotyping for rare Mendelian disorders**

Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)  
der Mathematisch-Naturwissenschaftlichen Fakultät  
der Rheinischen Friedrich–Wilhelms–Universität, Bonn

vorgelegt von

**Tzung-Chien Hsieh**

aus

Miaoli, Taiwan

Bonn, Februar 2022



Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der  
Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Prof. Dr. Peter M. Krawitz
2. Gutachter: Prof. Dr. Christian Bauckhage

Tag der Promotion: 01.07.2022

Erscheinungsjahr: 2022





## Acknowledgements

I would like to thank Prof. Peter Krawitz and Prof. Christian Bauckhage for their supervision and support during my Ph.D. study. They taught me not only the knowledge in medical and computer science research fields but also helped me integrate into the German culture. Besides, I want to thank Prof. Thomas Schultz and Prof. Markus Nöthen for being part of my Ph.D. evaluation committees.

Moreover, I appreciate the support from Dr. Jean Tori Pantel, Prof. Shahida Moosa, Dr. Alexej Knaus, Prof. Karen W. Gripp, Dr. Hellen Lesmann, Dr. Tom Kamphans, Dr. Martin Atta Mensah, Dr. Nadja Ehmke, Stanislav Rosnev, Max Zhao, and all the colleagues in IGSB and the Human Genetics department of University Hospital of Bonn for the PEDIA and GestaltMatcher projects.

Our industry partner FDNA was also an indispensable collaborator in my Ph.D. works. During the collaboration, I received tremendous support from Aviram Bar-Haim, Nicole Fleischer, Guy Nadav, Yair Hanani, Yaron Gurovich, and all the colleagues in FDNA.

The accompany of my family and friends was massive support during my Ph.D. study. I want to thank my parents and my brother. They took care of me from birth and even after coming to Germany. My parents frequently sent me supplies from Taiwan to help me pass COVID time and took care of me when I was in the incredible strict quarantine in Taiwan.

I want to thank my wife, Dr. Jing-Mei Li. She suffered from my terrible cooking skill and was forced to cook if she wanted to eat something edible. I would also like to thank my friends, Dr. Pei-Chen Peng, Dr. Ju-Yi Peng, Yu-Chung Yang, Yung-Hsiang Yang, Chia-Mau Ni, Shyh-En Lin, Hsin Miao, and Hao Jing. They wasted lots of time chatting nonsense with me that are important for relieving the pressure.

In the end, I would like to thank again to all of my supervisors, my family, colleagues, and friends. I will not achieve a Ph.D. study without them.



## Abstract

Worldwide, rare genetic disorders affect more than 6.2% of the population. The long diagnostic process is often called the ‘diagnostic odyssey.’ With the recent advances in computer vision, many next-generation phenotyping (NGP) approaches such as DeepGestalt have shown a strong ability to differentiate rare disorders and are widely used by clinicians in clinics. However, the current NGP approaches for rare disorders still have limitations on three aspects: current approaches do not support ultra-rare and novel disorders; no publicly available dataset; lack of automatic diagnostic pipeline that integrates exome and facial analysis. Therefore, we proposed GestaltMatcher, GestaltMatcher Database (GMDB), and Prioritization of Exome Data by Image Analysis (PEDIA) to tackle the current difficulties.

We first developed GestaltMatcher as an extension to DeepGestalt to support ultra-rare and novel disorders. GestaltMatcher first encoded the frontal image into a 320-dimensional Facial Phenotype Descriptor (FDP). We further formed a Clinical Face Phenotype Space by the FDPs and quantified the facial syndromic similarities among the patients by calculating the cosine distance between two FDPs in the space. This approach can support ultra-rare disorders and novel diseases and analyze the patients’ similarities to explore the novel gene-phenotype relationship.

To solve the problem of lacking a public medical image dataset, we proposed GMDB to host the medical images curated from the publication and the consented patients. GMDB is an open-access medical image database to the research community for deep learning purposes and reference material for clinician-scientists to easily see the medical images.

In order to support the facial phenotyping approach in the automatic exome diagnosis, the PEDIA approach was proposed to integrate facial analysis into the exome prioritization pipeline. We further showed GeneTalk platform as an example of implementing the PEDIA approach into an existed variant analysis platform.

In the end, we envision that GestaltMatcher, GMDB, and PEDIA can be integrated into a diagnostic platform and further connected with the patient match platforms such as MatchMaker Exchange to enable global collaboration and further improve the diagnosis of rare Mendelian disorders.



# Contents

Contents .....	ix
List of Figures.....	xiii
List of Tables .....	xv
Chapter 1    Introduction .....	17
1.1    Motivations .....	17
1.2    Contributions.....	21
1.3    List of publications .....	22
Chapter 2    Background and related works .....	24
2.1    Next-generation phenotyping approaches.....	24
2.2    DeepGestalt: facial phenotyping framework .....	25
2.2.1    Face datasets .....	25
2.2.2    Face detection and alignment .....	26
2.2.3    Training procedure .....	27
2.2.4    Evaluation on Face2Gene and GMDB datasets .....	29
2.3    Discussion .....	31
Chapter 3    The discovery of a novel phenotype by AI-driven facial phenotyping .....	32
3.1    Summary .....	32
3.2    DeepGestalt analysis .....	33
3.3    FaceNet analysis .....	35
3.4    Discussion .....	36
Chapter 4    GestaltMatcher facilitates rare disease matching using facial phenotype descriptors    37	
4.1    Summary .....	37
4.2    Abstract .....	38
4.3    Introduction.....	39
4.4    Results.....	42
4.4.1    Training with dysmorphic images improves the performance. ....	43
4.4.2    Syndromic diversity improves matching with novel phenotypes.....	45
4.4.3    Comparing performance between GestaltMatcher and DeepGestalt .....	49
4.4.4    Matching undiagnosed patients from unrelated families.....	50
4.4.5    GestaltMatcher and human experts agree on distinctiveness .....	54
4.4.6    Characterization of phenotypes in the CFPS.....	55
4.4.7    GestaltMatcher as a tool for clinician scientists .....	57

4.5	Discussion .....	59
4.6	Methods.....	61
4.6.1	Study approval .....	61
4.6.2	Face2Gene datasets.....	61
4.6.3	GMDB datasets.....	63
4.6.4	DeepGestalt encoder .....	64
4.6.5	Descriptor projection: Clinical Face Phenotype Space .....	66
4.6.6	Evaluation .....	66
4.6.7	London Medical Dataset validation analysis.....	66
4.6.8	Rare syndromes analysis.....	67
4.6.9	Matching undiagnosed patients from unrelated families.....	67
4.6.10	Syndrome facial distinctiveness score .....	68
4.6.11	Syndrome prevalence.....	68
4.6.12	Unseen syndromes correlation analysis.....	68
4.6.13	Analysis of number of training syndromes and subjects .....	69
4.7	Acknowledgements .....	71
4.8	Code availability .....	71
4.9	Data availability .....	71
Chapter 5	GestaltMatcher Database: medical imaging data for deep learning on rare disorders	73
5.1	Introduction .....	73
5.2	Methods.....	74
5.3	Online platform and database .....	76
5.3.1	GUI for patient data annotation .....	77
5.3.2	Digital consent for easier patient recruitment.....	79
5.3.3	Visualize patients in gallery view .....	81
5.3.4	Training data for next-generation phenotyping .....	81
5.4	Discussion .....	82
Chapter 6	Prioritization of exome data by image analysis .....	84
6.1	Summary .....	84
6.2	Abstract .....	85
6.3	Introduction .....	85
6.4	Materials and methods .....	86
6.5	Results .....	90
6.6	Discussion .....	94
6.7	Code availability .....	97
6.8	PEDIA in variants analysis platform .....	97

6.8.1	PEDIA platform.....	97
6.8.2	PEDIA in GeneTalk.....	101
Chapter 7	Discussion and the future of next-generation phenotyping .....	104
7.1	Modernizing DeepGestalt approach .....	105
7.2	Bias removal .....	106
7.3	Synthesizing faces with facial dysmorphism.....	109
7.4	Enabling global collaboration.....	110
Chapter 8	Conclusion .....	112
	Bibliography .....	113
	Appendix .....	127
A.1	Supplementary information of GestaltMatcher .....	127
A.2	Supplementary information of PEDIA .....	127





## List of Figures

Figure 1: Next-generation phenotyping tool diagnoses patients by facial image.....	18
Figure 2: Screenshot of suggested syndromes in Face2Gene.....	30
Figure 3: Facial phenotypes of individuals 1 (left) and 2 (right).....	33
Figure 4: Similarity analysis of two <i>LEMD2</i> patients. ....	34
Figure 5: Histogram of the pairwise distances among all cohort cases.....	36
Figure 6: Subsets of disorders supported by DeepGestalt and GestaltMatcher. ....	41
Figure 7: Concept of GestaltMatcher.....	43
Figure 8: Influence of the number of syndromes included in model training. ....	47
Figure 9: Performance improvement of double syndromes and double subjects when using different base sample sizes with Face2Gene models and the Face2Gene rare set. ....	48
Figure 10: Influence of the number of syndromes included in model training. ....	49
Figure 11: Pairwise ranks of individuals with mutations in <i>TMEM94</i> .....	52
Figure 12: Comparison of the pairwise distance distribution between subjects in the same family and subjects in different families with the same disease-causing gene. ....	53
Figure 13: Correlation among syndrome prevalence, distinctiveness score, and top-10 accuracy. ....	55
Figure 14: Hierarchical clustering of four phenotypic series, Kabuki syndrome, Noonan syndrome, mucopolysaccharidosis, and Cornelia de Lange syndrome, using a <i>t</i> -SNE projection of the Facial Phenotype Descriptors.....	56
Figure 15: <i>t</i> -SNE visualization of Facial Phenotype Descriptors of (a) ten syndromes with and (b) ten syndromes without facial dysmorphism.....	57
Figure 16: Screenshot of the GestaltMatcher web service. ....	58
Figure 17: Overview of Face2Gene data categorization in GestaltMatcher. ....	62
Figure 18: Venn diagram of numbers of syndromes in the Face2Gene and GMDB datasets. ....	64
Figure 19: GMDB database schema.....	76
Figure 20: Patient information in GMDB.....	77
Figure 21: Patient photos in GMDB.....	78
Figure 22: Phenotypic and genomic information in GMDB. ....	78
Figure 23: Patient invitation link. ....	79
Figure 24: Digital consent. ....	80

Figure 25: Gallery view in GMDB.....	81
Figure 26: Download section of GMDB dataset. ....	82
Figure 27: Prioritization of exome data by image analysis (PEDIA): cohort and classification approach.....	88
Figure 28: Performance readout and visualization of test results for a representative prioritization of exome data by image analysis (PEDIA) case.....	93
Figure 29: The flowchart of PEDIA platform. ....	98
Figure 30: PEDIA Manhattan plot in PEDIA platform. ....	99
Figure 31: List of genes with PEDIA scores in the PEDIA platform.....	99
Figure 32: Variants sorted by clinical significance in VCF viewer.....	100
Figure 33: Variant annotation from external databases.....	100
Figure 34: PhenoBot in GeneTalk. ....	101
Figure 35: PEDIA Manhattan plot in GeneTalk.....	102
Figure 36: List of genes with PEDIA scores in GeneTalk. ....	103
Figure 37: Results of baseline and after applied JLU.....	108
Figure 38: <i>t</i> -SNE visualization of baseline and after applied JLU. ....	108

## List of Tables

Table 1. The architecture of DeepGestalt. ....	28
Table 2: Performance comparison between training on Face2Gene and GMDB datasets..	29
Table 3: Performance comparison between classification and clustering with different encoders on sets of known disorders. ....	45
Table 4: Matching of novel phenotypes on a GeneMatcher validation set. ....	51
Table 5: Benchmarking on GMDB dataset. ....	82
Table 6: Ethnicity distribution in training and test set.....	107

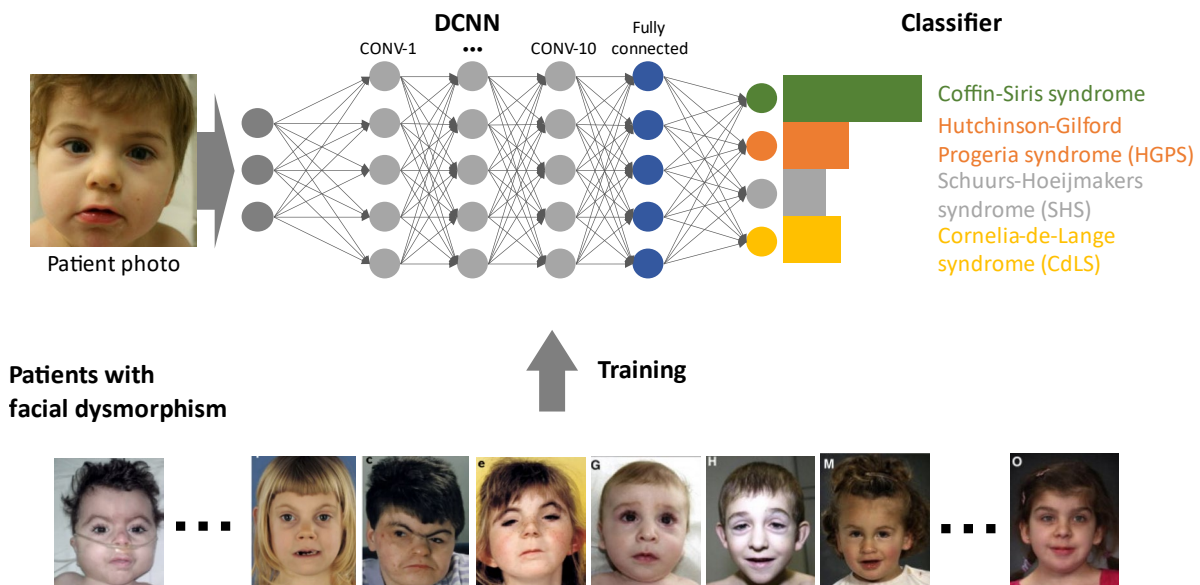


# Chapter 1 Introduction

This thesis aims to develop the next-generation phenotyping (NGP) approach based on facial image analysis to improve the diagnosis of patients with rare Mendelian disorders. It contains five topics: (1) the introduction to next-generation phenotyping approaches, (2) overcoming the limits of rare disease matching using facial phenotypic descriptor, (3) GestaltMatcher database, (4) prioritization of exome data by image analysis (PEDIA), and (5) the future of next-generation phenotyping. These topics drove the NGP technology from proof of concept to the application in clinical settings that help clinicians diagnose patients with rare disorders and further explore the unknown genotype-phenotype relationship. This cumulative thesis comprises three published studies and one not yet published work.

## 1.1 Motivations

Diagnosing a patient with a rare disorder is difficult due to the vast search space and the lack of advanced computer-assisted approaches. To date, there are more than 8000 rare Mendelian disorders. The large search space results in a challenging and long journey to identify the correct diagnosis that is frequently called diagnostic odyssey. Another reason is the lack of advanced computer-assisted approaches. Back to 15 years ago, there was no machine-readable language that could describe and analyze phenotypic features. Hence, making the diagnosis relied heavily on the clinician's experience. In 2008, Peter Robinson invented Human Phenotype Ontology (HPO) to describe the clinical phenotypic features, and the HPO terms became the global standard for the phenotypic description (Robinson et al. 2008). Numerous computational approaches have been developed based on HPO terms (Köhler et al. 2009; Bauer et al. 2012; Smedley et al. 2015; Cipriani et al. 2020). However, these tools suffered from the information lost during the conversion to HPO terms. For example, for the facial dysmorphic features with no HPO terms to describe, the clinicians always refer to them as “characteristic faces.” Hence, more advanced approaches to preserve the information are still required.



**Figure 1: Next-generation phenotyping tool diagnoses patients by facial image.** The deep convolutional networks (DCNN) were trained on patients with facial dysmorphism. By passing the patient photo into the networks, we can obtain the similarity scores as the possibilities to the syndromes trained in the networks. This patient has Coffin-Siris syndrome, and after passing this patient into the network, Coffin-Siris syndrome has the highest similarity score. The patients' photos in this figure are taken from the publications (Hoyer et al. 2012; Kline et al. 2018).

With the rapid development of computer vision and machine learning, a considerable number of next-generation phenotyping (NGP) approaches have emerged for analyzing rare genetic disorders by using two-dimensional frontal facial images (Ferry et al. 2014; Kuru et al. 2014; Cerrolaza et al. 2016; K. Wang and Luo 2016; Dudding-Byth et al. 2017; Shukla et al. 2017; Liehr et al. 2018; Gurovich et al. 2019; van der Donk et al. 2019; Porras et al. 2021; Hong et al. 2021). Ferry *et al.* proposed a clinical face phenotype space that converted frontal faces into a phenotype space by training on more than 1,000 photos with eight syndromes in 2014 (Ferry et al. 2014). Later in 2019, FDNA published the facial analysis framework, DeepGestalt (Figure 1), in Nature Medicine that trained the deep convolutional neural networks on 17,000 patients with more than 200 syndromes (Gurovich et al. 2019). This tool was launched in the Face2Gene platform (<https://www.face2gene.com>) and is already widely used by thousands of clinicians in their daily diagnostic routines. These approaches opened the door to enable frontal images analysis to aid the diagnosis.

However, the current approaches struggled in the three different aspects: algorithms, data resources, and application for variant prioritization.

**Limitations to the current NGP approaches:** Although NGP technology such as DeepGestalt is compelling, it still encounters three significant limitations: (1) no support for ultra-rare disorders, (2) difficulty to be scaled, and (3) no explanation to patients' similarity. The reason is that DeepGestalt trains the networks in a supervised manner, and it will be difficult to include an ultra-rare disorder in the networks when there are only very few photos for this disorder available. Moreover, we have to retrain and re-evaluate the networks that require lots of time and effort to include new disorders. In the end, it cannot quantify the similarities between patients or between syndromes that are crucial to answering the lumpner and splitter question (McKusick 1969; Oti, Huynen, and Brunner 2008).

To overcome these limitations, my colleagues and I published these two works, “The Discovery of a LEMD2-Associated Nuclear Envelopathy with Early Progeroid Appearance Suggests Advanced Applications for AI-Driven Facial Phenotyping,” (Marbach et al. 2019), and “GestaltMatcher facilitates rare disease matching using facial phenotype descriptors,” (T.-C. Hsieh et al. 2022). These two publications will be introduced in Chapter 3 and Chapter 4.

**Limited publicly available image resources:** Even though we have seen significant progress in the next-generation phenotyping technology in the last decade, we still cannot overlook the fundamental problem of lacking a high-quality public dataset in this research field. When we talk about the boost of image recognition, most people might mention ImageNet (Jia Deng et al. 2009). This publicly available database contains over one million images and one thousand classes. Based on this dataset, the “ImageNet Large Scale Visual Recognition Challenge” has become an essential annual benchmarking competition since 2010 (Russakovsky et al. 2014). Numerous classical architectures have emerged in this competition. Moreover, for the face recognition task, many public face datasets such as Labeled Faces in the Wild (LFW), CASIA-WebFace, and VGGFace pushed the face recognition technology forward (Huang et al. 2007; Yi et al. 2014; Parkhi, Vedaldi, and Zisserman 2015). Therefore, the requirement of many images, cleanliness of the dataset, and public availability are the key factors to push computer vision research.

However, we have none of the three factors mentioned above in this facial dysmorphology analysis of rare disorders. The two most well-known databases are London Medical Database (Winter and Baraitser 1987) and Face2Gene. The London Medical Database is also called LMD as an abbreviation. Although LMD was an essential resource for clinicians and NGP technology, people now often question its outdated resource, and many syndromes in the database are based on differential diagnoses. The Face2Gene database contains more than 30,000 patients with around 1,500 disorders. The patients are contributed by the clinicians who used the platform. It means that it requires data curation, and it is not able to be shared due to legal restrictions.

Moreover, none of these two is publicly available. The lacking of data not only raises the threshold to enter this research field, but we can hardly benchmark the methods proposed by different research teams. Hence, a facial dysmorphology version of ImageNet or LFW is an urgent need to push this field to the next level.

For this purpose, we developed the GestaltMatcher Database (<https://gestaltmatcher.org>), a collection of medical images curated by medical experts, and it is accessible to the scientific community. This work will be presented in Chapter 5.

**Lacking facial phenotyping application for variant prioritization:** Several published works have shown the power of NGP tools to reduce the search space of candidate genes. It provides tremendous help for clinicians to decide the genes selected in the gene panel. However, as the cost of whole-exome sequencing is continuously dropping, there is a need to efficiently and automatically integrate the facial analysis into the variant prioritization pipeline. That is, in practice, the clinicians need this kind of automatic pipeline for the diagnostic workup. Therefore, we develop the PEDIA approach in “PEDIA: Prioritization of Exome Data by Image Analysis” (T. C. Hsieh et al. 2019) to enable the automatic diagnostic pipeline that integrates facial image analysis, clinical feature analysis, and exome sequencing analysis.



## 1.2 Contributions

This thesis contains my seven published works. These publications are essential works in my doctoral period for the facial analysis of rare disorders. In this thesis, I only presented and discussed three publications (Marbach et al. 2019; T.-C. Hsieh et al. 2022; T. C. Hsieh et al. 2019) because these three already included the other four publications (Jean T. Pantel et al. 2018; Knaus et al. 2018; L. Guo et al. 2021; Ebstein et al. 2021).

Chapter 2 introduces the recent next-generation phenotyping approaches. I took DeepGestalt as an example to demonstrate the whole facial analysis framework: face cropping, network architecture, training procedure, and benchmarking.

The first current difficulty on the algorithmic level is addressed in Chapter 3 and Chapter 4. Chapter 3 presents the proof of concept of matching two patients with the novel disease by the facial features extracted from the DeepGestalt model (Marbach et al. 2019). Later in Chapter 4, I will introduce the solution with more detail and experiments that we conducted in the publication, “GestaltMatcher facilitates rare disease matching using facial phenotype descriptors” (T.-C. Hsieh et al. 2022). GestaltMatcher is an extension of DeepGestalt. In this work, we used the same network architecture as in DeepGestalt. We constructed a Clinical Face Phenotype Space (CFPS) using facial phenotypic descriptors extracted from the feature layer. The cosine distance in the CFPS quantified the similarity between two patients. By this approach, we are no longer limited to the syndromes with enough images and are flexible to the novel diseases. With the similarities among the patients, we can investigate the facial gestalt with the underlying molecular mechanism or disease pathway.

Moreover, as GestaltMatcher can match patients with a similar phenotype, it can be used as a facial image version of GeneMatcher (Sobreira et al. 2015), which helps clinicians find the second patient of ultra-rare disorders. We envision integrating GestaltMatcher into the Matchmaker Exchange platform (Philippakis et al. 2015) to enhance phenotypic matching.

To solve the lack of publicly available image resources, we developed GestaltMatcher Database (<https://gestaltmatcher.org>), and I will introduce this database in Chapter 5. Until January of 2021, it contained over 9,173 images with 620 different disorders. More than 30

clinician-scientists from different countries have curated patient data from the publications or the patients with proper consent. Most importantly, this database is available for the research community. We believe this database can enable global collaboration and accelerate the process of data collection and curation. It was a not yet published work.

Chapter 6 will introduce “PEDIA: Prioritization of Exome Data by Image Analysis,” published in *Genetics in Medicine* in 2019 (T. C. Hsieh et al. 2019). This work demonstrated how we integrated analysis from three different kinds of patient data: facial photo (facial phenotype), HPO terms (clinical description), and exome sequencing data (molecular data). This analysis pipeline was already implemented in the University Hospital of Bonn and can be considered an example of a standard diagnostic pipeline in the future.

In the end, I discussed the current status of NGP approaches, the problems not addressed in this thesis, such as the con-founder analysis. I also proposed possible future works to strengthen NGP approaches and synthesize frontal images with facial dysmorphism to analyze rare Mendelian disorders.

### 1.3 List of publications

- **Hsieh, Tzung-Chien**, Aviram Bar-Haim, Shahida Moosa, Nadja Ehmke, Karen W. Gripp, Jean Tori Pantel, Magdalena Danyel, et al. 2022. “GestaltMatcher Facilitates Rare Disease Matching Using Facial Phenotype Descriptors.” *Nature Genetics*, February. <https://doi.org/10.1038/s41588-021-01010-x>.
- Ebstein, Frédéric, Sébastien Küry, Victoria Most, Cory Rosenfelt, ..., **Tzung-Chien Hsieh**, et al. 2021. “De Novo Variants in the PSMC3 Proteasome AAA-ATPase Subunit Gene Cause Neurodevelopmental Disorders Associated with Type I Interferonopathies.” medRxiv.
- Guo, Lily, Jiyeon Park, Edward Yi, Elaine Marchi, Yana Kibalnyk, Anastassia Voronova, **Tzung-Chien Hsieh**, Peter M. Krawitz, and Gholson J. Lyon. 2021. “KBG Syndrome: Prospective Videoconferencing and Use of AI-Driven Facial Phenotyping in 25 New Patients.” medRxiv.

- 
- **Hsieh, Tzung-Chien**, Martin A. Mensah, Jean T. Pantel, Dione Aguilar, Omri Bar, Allan Bayat, Luis Becerra-Solano, et al. 2019. “PEDIA: Prioritization of Exome Data by Image Analysis.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 21 (12): 2807–14.
  - Marbach, Felix, Cecilie F. Rustad, Angelika Riess, Dejan Đukić, **Tzung-Chien Hsieh**, Itamar Jobani, Trine Prescott, et al. 2019. “The Discovery of a LEMD2-Associated Nuclear Envelopathy with Early Progeroid Appearance Suggests Advanced Applications for AI-Driven Facial Phenotyping.” *American Journal of Human Genetics* 104 (4): 749–57.
  - Knaus, Alexej, Jean Tori Pantel, Manuela Pendziwiat, Nurulhuda Hajjir, Max Zhao, **Tzung-Chien Hsieh**, Max Schubach, et al. 2018. “Characterization of Glycosylphosphatidylinositol Biosynthesis Defects by Clinical Features, Flow Cytometry, and Automated Image Analysis.” *Genome Medicine* 10 (1): 3.
  - Pantel, Jean T., Max Zhao, Martin A. Mensah, Nurulhuda Hajjir, **Tzung-Chien Hsieh**, Yair Hanani, Nicole Fleischer, et al. 2018. “Advances in Computer-Assisted Syndrome Recognition by the Example of Inborn Errors of Metabolism.” *Journal of Inherited Metabolic Disease*, April. <https://doi.org/10.1007/s10545-018-0174-3>.

## Chapter 2 Background and related works

### 2.1 Next-generation phenotyping approaches

With the advance of computer vision in the last decade, face recognition technology has enabled the disorder prediction by analyzing two-dimensional facial images (Ferry et al. 2014; Kuru et al. 2014; Cerrolaza et al. 2016; K. Wang and Luo 2016; Dudding-Byth et al. 2017; Shukla et al. 2017; Liehr et al. 2018; Gurovich et al. 2019; van der Donk et al. 2019; Porras et al. 2021; Hong et al. 2021). One of the essential studies at the beginning of the computer vision trend was the work published by Ferry and his colleagues in 2014. They proposed Clinical Face Phenotype Space (CFPS) formed by the feature vectors encoded by the model trained on 1,363 photos with eight different syndromes (Ferry et al. 2014).

Since then, face recognition technologies were improved significantly and were the core of the deep learning revolution in computer vision. DeepFace (Taigman et al. 2014) demonstrated, for the first time, human-level performance in identity verification on the Labeled Faces in the Wild dataset (Huang et al. 2007). As a result, the face recognition system trained on CCTV images was utilized to match patients with one of ten syndromic disorders with intellectual disability (Dudding-Byth et al. 2017). In addition, the facial recognition model from healthy individuals can also be integrated with the CFPS as a hybrid model, and it was proven able to discriminate the facial gestalt of three novel disease-causing genes (van der Donk et al. 2019).

Although many novel approaches were proposed after Ferry's work, the scale of the training dataset and the number of disorders did not increase so much until DeepGestalt was published in 2019 (Gurovich et al. 2019). DeepGestalt is the facial analysis framework proposed by FDNA Inc. that trained the deep convolutional neural networks on over 17,000 facial photos representing more than 200 disorders. DeepGestalt was considered the current state-of-the-art facial analysis approach, and the training dataset was the most extensive collection. This approach was already launched in the Face2Gene platform

(<https://www.face2gene.com>) and used by clinicians in their daily diagnostic routine. In this thesis, I took DeepGestalt as an example of the facial analysis framework and will introduce how DeepGestalt works in this chapter.

## 2.2 DeepGestalt: facial phenotyping framework

### 2.2.1 Face datasets

The training procedure requires two different kinds of datasets. The first is the dataset of healthy faces used for training, and the second is the patients with genetic disorders. Precisely, the genetic disorders should be the disorders with facial dysmorphism. Here I used “genetic disorders” as the disorders with facial dysmorphism to make it short.

The first dataset is used to train the networks to learn general facial features. There are many publicly available face datasets such as Labeled Faces in the Wild (LFW), CASIA-WebFace, CelebFaces+, VGGFace, MS-Celeb-1M, VGGFace2 (Huang et al. 2007; Yi et al. 2014; Sun, Wang, and Tang 2014; Parkhi, Vedaldi, and Zisserman 2015; Y. Guo et al. 2016; Cao et al. 2017). In DeepGestalt, they used CASIA-WebFace as the dataset for the face recognition task. So I also took it as an example for the following parts. As there are many more up-to-date and larger datasets compared to CASIA-WebFace, a benchmark of using different face datasets for this step is needed in the future.

The second dataset is for fine-tuning the networks to learn facial dysmorphic features. There were two datasets used in this thesis, the Face2Gene dataset and the GMDB dataset. The Face2Gene dataset is a private dataset owned by FDNA Inc., and it contains more than 38,000 facial photos with around 1,300 disorders. GMDB dataset is a publicly accessible dataset first introduced in GestaltMatcher publication (T.-C. Hsieh et al. 2022). More than 30 clinician-scientists from different countries have curated patient data from the publications or the patients with proper consent. As the result of international collaboration, it contained over 5,000 images with 500 different disorders until September of 2021. These two datasets serve different purposes. Face2Gene dataset is more extensive than the GMDB dataset, almost by a factor of eight. Therefore, the Face2Gene dataset is more suitable for

developing new methods or for use in production. However, it is a private dataset that only FDNA could access. The research community cannot reproduce the proposed methods and benefit from this dataset. Hence, we proposed the GMDB dataset mainly curated from medical publications. It could be used as a public benchmark dataset and accessible to the research community.

### **2.2.2 Face detection and alignment**

The images collected from medical publications or the photos taken directly from patients usually have different face sizes, poses, occlusions, and lightening, so face detection and alignment are crucial preprocessing steps. Face detection detects the face in the given image, and face alignment is to rotate or scale the image to the desired angle and size. We can obtain the faces less influenced by the pose and image size by these two steps. As face detection is a crucial step in face recognition, many approaches can detect and align the faces (Huang et al. 2012; Li et al. 2015; K. Zhang et al. 2016; Jiankang Deng, Guo, Ververas, et al. 2020). Face detection for our task is more straightforward than most face detection benchmarking, and we can obtain high-quality face crops with most of the existing tools. Because the source code of face detection and alignment methods described in DeepGestalt were not open to the public, instead of the unpublic resource, I introduced how we utilized RetinaFace (Jiankang Deng, Guo, Ververas, et al. 2020) to crop the face below. The source code of RetinaFace can be found in the official Git repository (<https://github.com/serengil/retinaface>).

RetinaFace can detect the bounding box of the face and the five facial landmarks: right eye, left eye, nose, right end of the mouth, and the left end of the mouth. The shape of the network's input layer is (100, 100), so we have to use the bounding box and facial landmarks to perform the scaling and alignment. We first rotate the image to let the two eyes horizontally aligned and then scale the image into (100, 100). Moreover, DeepGestalt takes a grayscale image as input, so we later convert the image into a grayscale image. The final step can be optional when using different architectures with different input shapes.

### 2.2.3 Training procedure

As briefly introduced in Section 2.2.1, we utilized transfer learning for training the DeepGestalt model. We first trained the networks on CASIA-WebFace that contains only healthy individuals to learn the general facial features, and later trained the same networks with the weights preserved from previous steps on the Face2Gene dataset to learn the facial dysmorphic features. The networks used in DeepGestalt are similar to the networks introduced in the CASIA-WebFace publication (Yi et al. 2014).

The architecture is shown in Table 1. DeepGestalt network consists of ten convolutional layers with batch normalization (BN) and ReLU for embedding the input features. After every Conv-BN-ReLU layer, a max-pooling layer is applied for reducing the spatial size while increasing the semantic representation. The classifier part of the network consists of a fully connected linear layer with dropout (0.5).

Because there are two steps in DeepGestalt training, it results in two different output sizes of fully-connected layer and softmax. The size depends on the training dataset. When we use the CASIA-WebFace to train the normal face recognition model, the output size is 10575 that is the number of classes in CASIA-WebFace. When using the Face2Gene dataset as a training dataset, the output size is 299. 299 is the number of syndromes we used for training in Chapter 4. This number will change when we include more syndromes in the training set or remove syndrome from the training set.

During inference, the facial region crop is forward passed through a deep convolutional network (CNN), and finally, we obtain the final prediction for the input face image. The DeepGestalt paper (Gurovich et al. 2019) also introduced using different face regions as input and aggregated the final prediction. They proved that this aggregated method gained higher accuracy than only using the whole face as input. I will not introduce this aggregated method in the following sections. Therefore, we can consider the DeepGestalt in this thesis only takes the whole face as input.

**Table 1. The architecture of DeepGestalt.** In the fully-connected layer and softmax, the output size equals the training dataset classes. The number of classes in CASIA-WebFace is 10575, and the number of classes in the Face2Gene dataset is 299. The classes in the Face2Gene dataset can vary from time to time.

Name	Size	Output size
Conv-1	3x3/1	100x100x32
Conv-2	3x3/1	100x100x64
Max-pooling	2x2/2	50x50x64
Conv-3	3x3/1	50x50x64
Conv-4	3x3/1	50x50x128
Max-pooling	2x2/2	25x25x128
Conv-5	3x3/1	25x25x96
Conv-6	3x3/1	25x25x192
Max-pooling	2x2/2	13x13x192
Conv-7	3x3/1	13x13x128
Conv-8	3x3/1	13x13x256
Max-pool	2x2/2	7x7x256
Conv-9	3x3/1	7x7x160
Conv-10	3x3/1	7x7x320
Average-pooling	7x7/1	1x1x320
Dropout (50%)		1x1x320
FC		10575 <sup>1</sup> (299) <sup>2</sup>
Softmax		10575 <sup>1</sup> (299) <sup>2</sup>

---

<sup>1</sup> The number of classes in CASIA-WebFace dataset.

<sup>2</sup> The number of classes in Face2Gene dataset.



### 2.2.4 Evaluation on Face2Gene and GMDB datasets

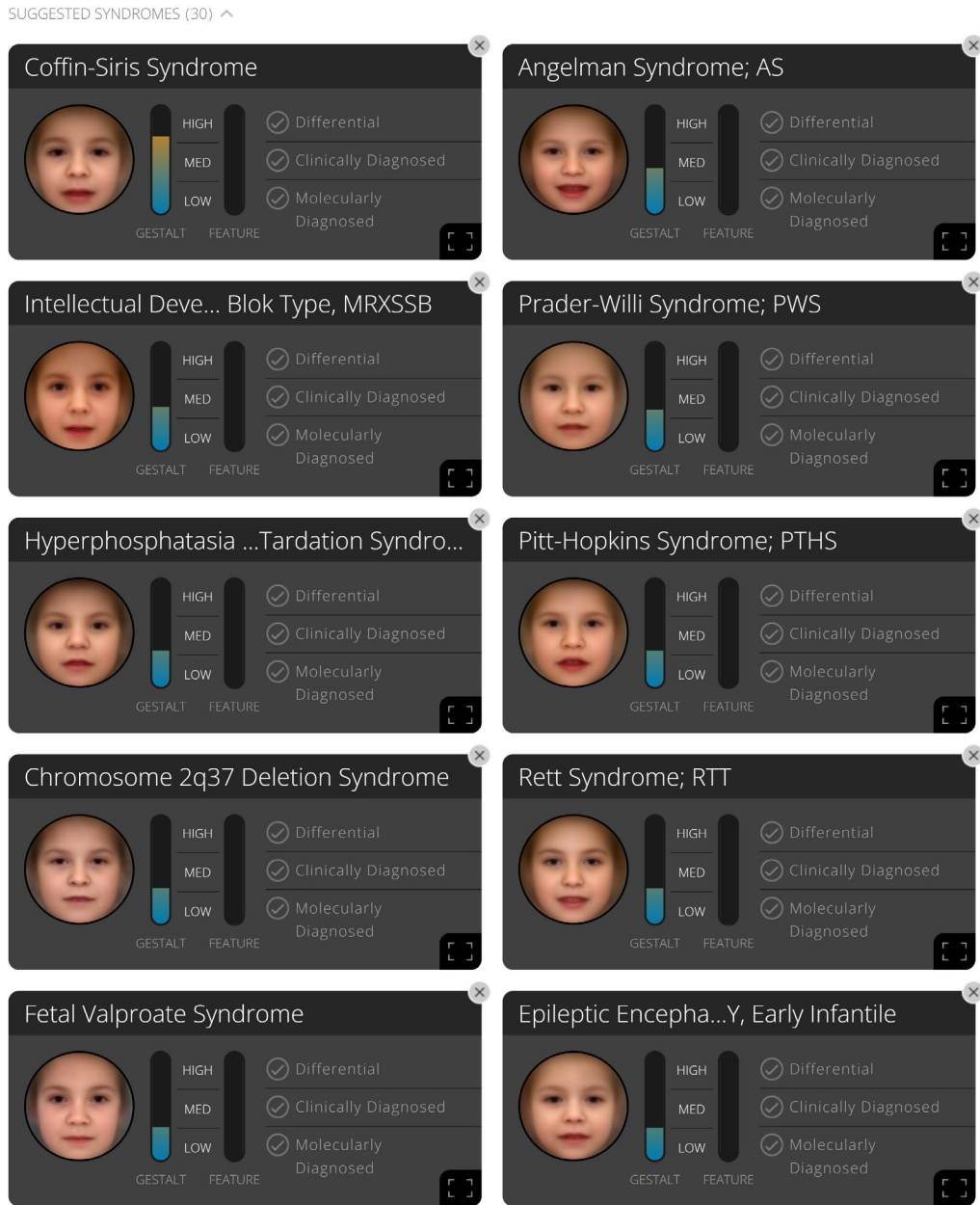
To evaluate the performance of DeepGestalt, I first trained and tested on the Face2Gene dataset. The training dataset consisted of 19,950 images from 299 different disorders, and the test dataset consisted of 2,669 images. Because the Face2Gene dataset is not open to the public, I later applied the same training and testing on the GMDB dataset. The version of the GMDB dataset is the same as the one published in the GestaltMatcher study (T.-C. Hsieh et al. 2022). The training dataset contained 3,438 images from 139 disorders, and the test dataset contained 360 images.

The top- $k$  accuracy is used to evaluate the performance. For each test image, the output is a list of gestalt scores that indicate the disorders' possibility. After sorting the list by gestalt scores in descending order, we can obtain a list of suggested syndromes. The disorder with a higher rank has a higher possibility of the correct diagnosis. If the correct diagnosis is among the top- $k$  ranks, we called it a top- $k$  match. The examples of the top- $k$  match are shown in Figure 2. The performance can be benchmarked by averaging each syndrome's top- $k$  accuracy (percent of test images with correct matches within the top- $k$ ) to avoid biasing predictions toward the major class.

The top- $k$  accuracy ( $k = 1, 5, 10$ , and  $30$ ) is reported in Table 2. When using the Face2Gene dataset as the training set, the top-1 accuracy was 35.94%, and the top-10 accuracy was 63.91%. On the GMDB dataset, the top-1 and top-10 accuracy was 25.13% and 59.79%, respectively.

**Table 2: Performance comparison between training on Face2Gene and GMDB datasets.**

Model	Classes	Training images	Test images	Top-1	Top-5	Top-10	Top-30
Face2Gene	299	19,950	2,669	35.94%	52.45%	63.91%	78.13%
GMDB	139	3,438	360	25.13%	47.23%	59.79%	77.26%



**Figure 2: Screenshot of suggested syndromes in Face2Gene.** By sorting gestalt scores in descending order, we can obtain a list of suggested syndromes. This list shows the top-10 syndromes. Suppose the correct diagnosis of this patient is Coffin-Siris syndrome which is at the first rank. Then we call it a top-1 match. It will also contribute to any  $k$  larger than one. Take Rett Syndrome for the correct diagnosis as another example. The rank of Rett Syndrome is at eighth place, so we can call it a top-10 match. We usually take 1, 5, 10, 30 for the value of  $k$ .

## 2.3 Discussion

This chapter introduced the next-generation phenotyping approach, such as DeepGestalt, by presenting the network architecture and the benchmarking on Face2Gene and GMDB datasets. It is noted that the accuracy of the Face2Gene dataset was not comparable to the results reported in the original DeepGestalt publication because the training and testing sets were different. Moreover, the model was trained on the whole face as an example that is different from the DeepGestalt method introduced in the original publication. They first cropped the face into different face regions and later aggregated the prediction results.

Besides, we should not directly compare the performance between training on Face2Gene and GMDB datasets because the numbers of the syndrome and test sets were different. However, even though the classes of Face2Gene were two times that of GMDB, the performance of Face2Gene was still higher than GMDB. It indicates that the model trained on a more extensive and diverse dataset performs better because the Face2Gene dataset is approximately seven to eight times larger than the GMDB dataset.

Although the results reported in the previous section were hard to compare to the original work, we could still gain hints for further improvement from both data and method perspectives. The current publicly available dataset to the scientific community is still too small compared to the private Face2Gene dataset, and it is a considerable concern for the further development of next-generation phenotyping approaches. Hence, the collection of images for the scientific community is an urgent need.

In addition, most of the syndromes in the Face2Gene dataset are the common ones among rare disorders. The speed of collecting data would become more and more difficult and slow due to the rareness of the data. So it is crucial that the method should support ultra-rare disorders and novel diseases. The following two chapters will tackle this problem.

In the end, the architecture used in DeepGestalt is from a study presented in 2014, and it might be needed to update the network architecture or the way of aggregating the different face regions. This topic will be further discussed in Section 7.1.

## Chapter 3 The discovery of a novel phenotype by AI-driven facial phenotyping

### 3.1 Summary

This study is the proof of concept of how to utilize NGP tools to explore the novel gene-phenotype association. We introduced a novel phenotype presented in two unrelated patients caused by a *de novo* mutation in *LEMD2*. These two patients shared the same *de novo* disease-causing mutation and were initially matched by GeneMatcher. This phenotype was later named Marbach-Rustad progeroid syndrome (MIM: 619322).

These two patients were first analyzed by DeepGestalt (Face2Gene), and both gestalt scores of suggested syndromes provided by DeepGestalt were very similar. That provided a hint that these two patients had a similar facial phenotype. However, DeepGestalt can only recognize the disorder trained in the model and quantify the similarity between patient and disorder. Hence, to prove these two patients had the same novel disorder, we utilized FaceNet (Schroff, Kalenichenko, and Philbin 2015) to obtain the facial embedding of each photo and compare it to the other 265 patients with 66 monogenic disorders in the PEDIA cohort. We found that the pairwise distance between these two patients was significantly smaller than the random pairwise comparison of the other 265 patients. We then concluded that these two patients shared similar facial gestalt and suggested *LEMD2* can link to this novel disease.

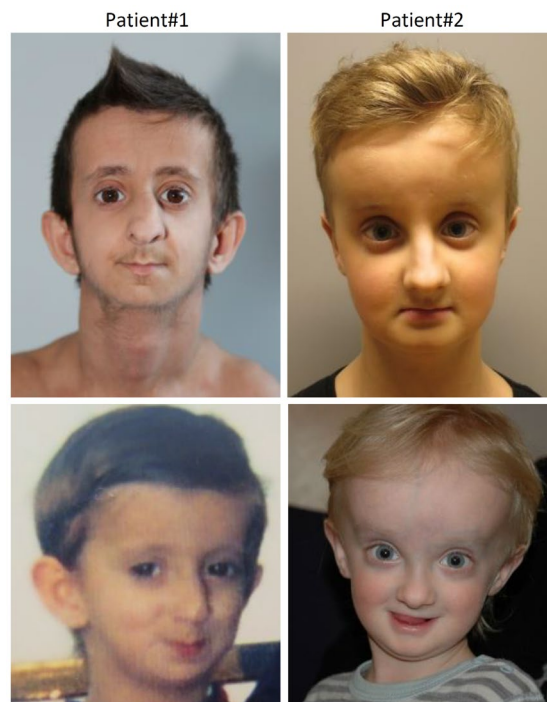
This work was already published in the American Journal of Human Genetics in 2019 with the title ‘The Discovery of a *LEMD2*-Associated Nuclear Envelopathy with Early Progeroid Appearance Suggests Advanced Applications for AI-Driven Facial Phenotyping’ (Marbach et al. 2019). I am one of the co-authors and performed the facial analysis in this study.

With this study’s inspiration, instead of using FaceNet trained on healthy persons only, we would like to take DeepGestalt trained on patients with facial dysmorphism as a better encoder to convert images to feature vectors. Therefore, in the next chapter, we will

introduce the GestaltMatcher approach (T.-C. Hsieh et al. 2022) that tackles the limitation to ultra-rare disorders and novel diseases.

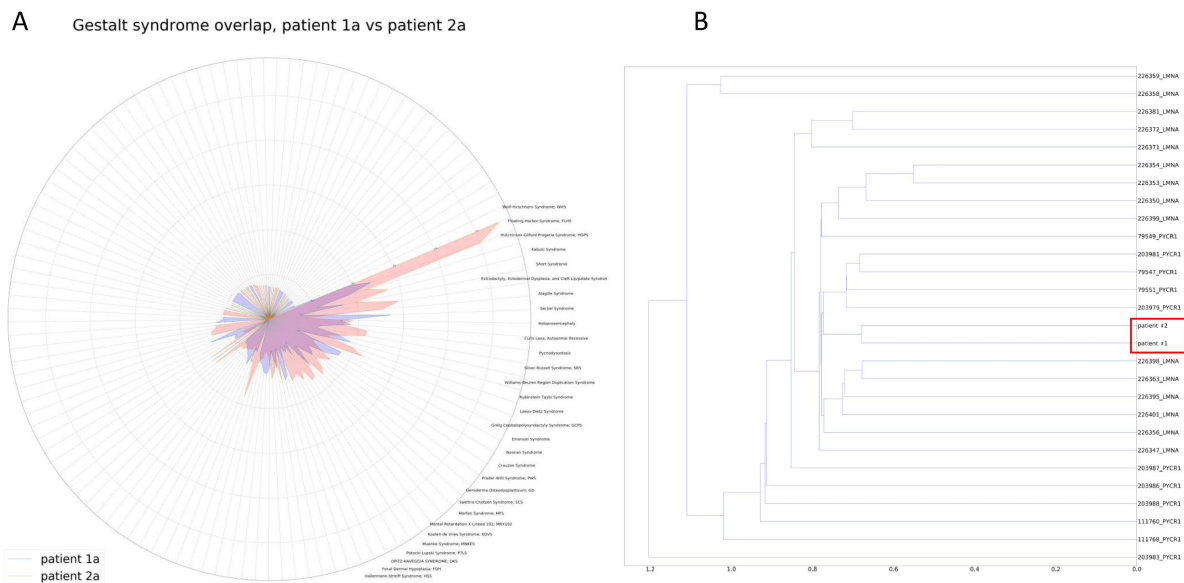
## 3.2 DeepGestalt analysis

The two patients analyzed in this study were initially presented in two different university hospitals (Bologna, Italy and Oslo, Norway). The German group later diagnosed the first individual in Cologne. The same *de novo* disease-causing mutation c.1436C>T, (p. Ser479Phe) in *LEMD2* was first identified independently on these two patients. Later, both groups found each other by the match with GeneMatcher (Sobreira et al. 2015). These two patients both presented similar progeria-like appearances. The frontal images of both patients are shown in Figure 3.



**Figure 3: Facial phenotypes of individuals 1 (left) and 2 (right).** Individual 1 is shown at the ages of 16 years (upper panel) and 3 years (lower panel); individual 2 is shown at the ages of 10 years and 2 years (upper and lower panel, respectively).

In 2017, the German group analyzed frontal images of individual #1 by Face2Gene (Gurovich et al. 2019) and found that the well-established nuclear envelopathy HGPS was listed among the most likely diagnoses. Later, the image analysis of individual #2 had similar results. We were intrigued by the DeepGestalt results. We then contacted Face2Gene and obtained the vector of similarity scores for each patient. Each vector had 216 syndromes similarity scores indicating the similarities to 216 syndromes trained in DeepGestalt. Both individuals' DeepGestalt similarity scores vectors are shown on a radar plot in Figure 4A. The figure shows the high overlap of the similarity scores and also suggests that the novel phenotype presented in both patients might be related to progeria-like syndromes such as HGPS.



**Figure 4: Similarity analysis of two *LEMD2* patients.** **A:** Overlap of DeepGestalt similarity scores of patients#1 (blue) and #2 (red) to other disorders. The computed facial similarity of each patient to the indicated disorders is visualized as colored areas extending outward on the radicular axes, and purple color indicates overlap. **B:** A dendrogram is used to visualize the computed phenotypic “distance” of patients #1 and #2 in a sample containing 265 individuals with 66 different syndromes. The close-up shows both patients are in close proximity to each other, as well as to patients with progeroid disorders (HGPS, *PYCRI*-related autosomal recessive cutis laxa).

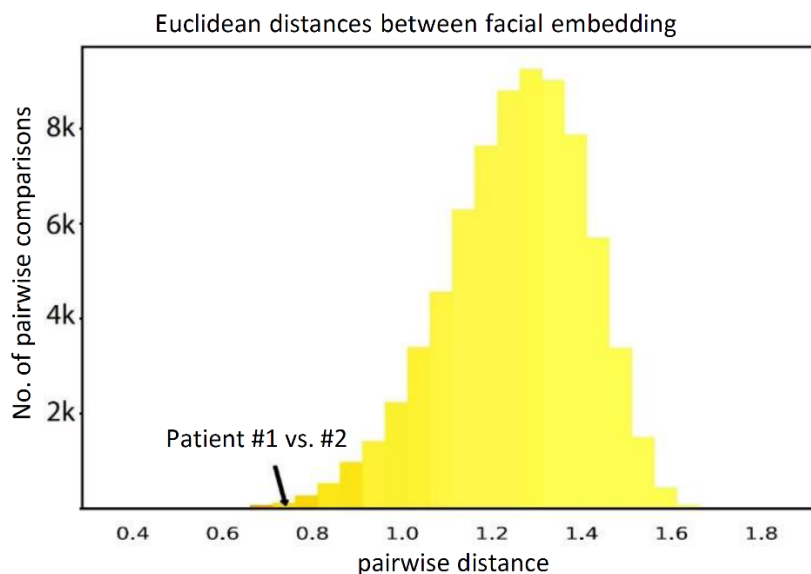
### 3.3 FaceNet analysis

Because the genetic disorder of individuals #1 and #2 was previously unknown, we further investigated the similarity between both cases in an unsupervised way. To do this, we wanted to measure the similarity between the two cases and compare to the patients with other disorders to see whether the two patients are more similar compared to the others.

To compare with the patients with other syndromes, we selected a cohort of 265 individuals from the PEDIA cohort (T. C. Hsieh et al. 2019). We are also interested in comparing the two patients to the patient with progeroid-related syndromes such as HGPS and *PYCR1*-related autosomal recessive cutis laxa. Therefore, these 265 patients have 66 different monogenic syndromes with facial dysmorphism, including 22 *LMNA* and 11 *PYCR1* patients.

We used FaceNet (Schroff, Kalenichenko, and Philbin 2015) to measure the similarities among patients. FaceNet was initially trained on healthy individuals to perform the intra-person recognition task, and each photo was encoded by the pre-trained FaceNet model (version 20170512-110547) into a 128-dimensional embedding vector. We measured the patient similarity by calculating the Euclidean distance between two vectors and hypothesized that individuals with the same disorder are smaller than those between individuals with different disorders. We found that individuals #1 and #2 were almost the most similar among 34,980 random pairwise comparisons and were much more similar than most individuals inside other disease entities, including some related individuals (Figure 5).

We further performed the clustering analysis to prove that these two patients were more similar to other progeroid-related syndromes. Figure 4B shows that the two *LEMD2* were the most similar patients to each other compared to *LMNA* and *PYCR1* patients. Therefore, we concluded that there two *LEMD2* patients presented a novel phenotype similar to progeroid-related syndromes and suggested that *LEMD2* could link to this new phenotype.



**Figure 5: Histogram of the pairwise distances among all cohort cases.** Comparisons of the same disease entity were used for adding a red blend to the respective bins according to their proportion. At the left side of the distribution, where the two individuals with the *LEMD2* mutation are also posed, the percentage of pairs with the same disease-causing gene increases.

### 3.4 Discussion

This study presented an example of linking *LEMD2* to a novel disorder related to progeroid syndrome. Although DeepGestalt can point to a similar syndrome to a patient’s phenotype, DeepGestalt cannot identify the novel phenotype. Moreover, DeepGestalt cannot measure the similarity between two patients. It is crucial that the NGP approaches can support ultra-rare diseases and novel disorders because these disorders are more and more common in the diagnostic workup.

We showed that the deep learning approach such as FaceNet initially trained on healthy individuals for intra-person verification could quantify the similarities among patients. It is a proof-of-concept study that moves from the classification method to the clustering approach. The next chapter will introduce GestaltMatcher developed based on this study and a more comprehensive analysis of this patient matching analysis.



## **Chapter 4   GestaltMatcher facilitates rare disease matching using facial phenotype descriptors**

### **4.1   Summary**

The previous chapter showed an example of identifying a novel phenotype by matching two unrelated patients with the facial feature vectors. Inspired by the previous chapter, we developed the GestaltMatcher approach to enable matching patients with the same facial phenotype. GestaltMatcher can be seen as an extension of DeepGestalt (Gurovich et al. 2019) that aims to overcome the following three limitations: (1) not able to support ultra-rare syndrome; (2) cannot support novel disorder; (3) cannot measure the similarities among patients.

This work first introduced the three limitations listed above to the current NGP approaches, such as DeepGestalt. We then presented GestaltMatcher that uses the DCNN trained on patients as an image encoder to convert facial photos into 320-dimensional feature vectors, and the vectors were referred to as Facial Phenotype Descriptors (FPDs). The Clinical Face Phenotype Space was formed by the FPDs, and the cosine distance in this space can further measure the patients' facial syndromic similarities.

We then proved that with GestaltMatcher, we could support the 299 syndromes trained in the model and additional 816 ultra-rare syndromes that the model has not seen before, indicating that GestaltMatcher can support the ultra-rare and novel disorders. Moreover, we validated GestaltMatcher on the 15 recent publications that utilized GeneMatcher (Sobreira et al. 2015) to match the patients who presented facial dysmorphism to show that GestaltMatcher can be seen as the facial image version of GeneMatcher. We further proved that GestaltMatcher could distinguish the disorders under the same phenotypic series. In the end, we also presented GestaltMatcher as a tool for clinician-scientists to diagnose patients and match patients with an unknown diagnosis. In the end, we envision GestaltMatcher can

be connected to the MatchMaker Exchange platform (Philippakis et al. 2015) to provide the matching by facial images.

The following sections will present the GestaltMatcher approach, and it is already published in Nature Genetics in 2022 with the title ‘GestaltMatcher facilitates rare disease matching using facial phenotype descriptors’ (T.-C. Hsieh et al. 2022). I am one of the co-first authors in this work, and I conducted the data analysis and wrote the manuscript. I am also responsible for organizing the entire project and communicating among the clinicians-scientists, from collecting the patient data to analyzing the phenotypes. In the following sections, the Method section (Section 4.6) is presented after the Discussion section (Section 4.5) to keep the same order of the paragraph as presented in the original publication. The supplementary materials can be found in Appendix A.1.

## 4.2 Abstract

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 novel phenotypes.

## 4.3 Introduction

Rare genetic disorders affect more than 6.2% of the global population (Ferreira 2019). Because genetic disorders are rare and diverse, accurate clinical diagnosis is a time-consuming and challenging process, often referred to as the “diagnostic odyssey” (Baird et al. 1988), and all informative clinical features have to be taken into consideration. A large fraction of patients, particularly those with neurodevelopmental disorders, exhibit craniofacial abnormalities (Hart and Hart 2009). 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 novel phenotype-gene associations (Marbach et al. 2019). 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 disorder, both of which are increasingly common scenarios.

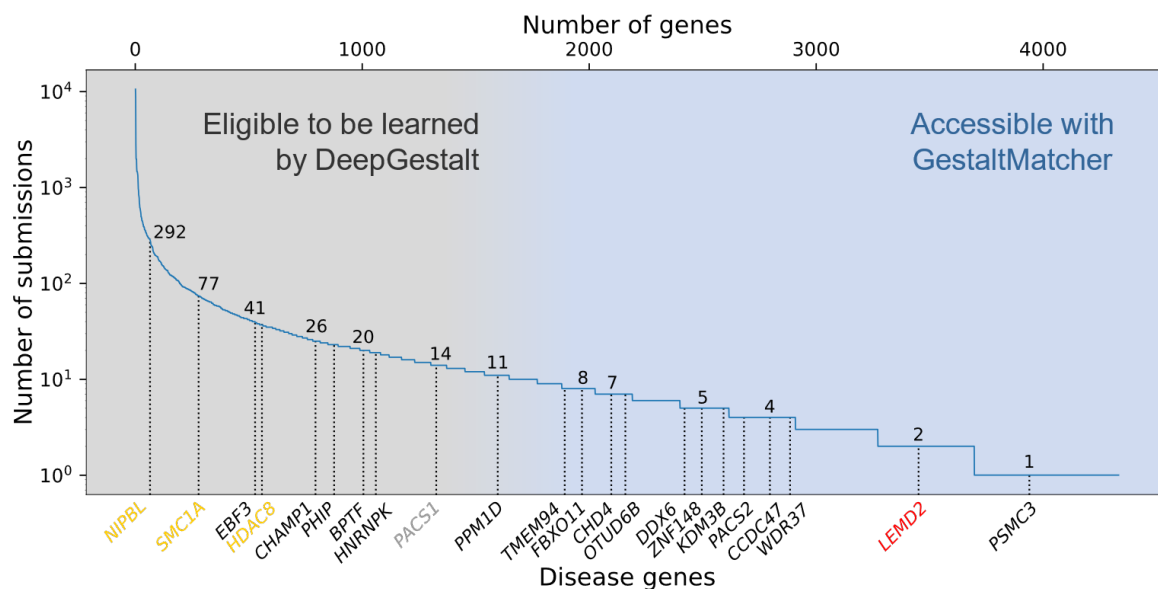
With the rapid development of machine learning and computer vision, a considerable number of next-generation phenotyping tools have emerged that can analyze facial dysmorphology using two-dimensional (2D) portraits of patients (Ferry et al. 2014; Kuru et al. 2014; Cerrolaza et al. 2016; K. Wang and Luo 2016; Dudding-Byth et al. 2017; Shukla et al. 2017; Liehr et al. 2018; Gurovich et al. 2019; van der Donk et al. 2019). 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 (Ferry et al. 2014). Since then, facial recognition technologies have improved significantly and constitute the core of the deep-learning revolution in computer vision (Taigman et al. 2014; Huang et al. 2007). The current state-of-the-art framework for syndrome classification, DeepGestalt (Face2Gene, 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 (Gurovich et al. 2019; Jean Tori Pantel et al. 2020). DeepGestalt has also demonstrated a strong ability to

separate specific syndromes and subtypes, surpassing human experts' performance (Gurovich et al. 2019). 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 (Landrum et al. 2018), 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 6). For instance, Cornelia de Lange syndrome (CdLS), which has been modeled by multiple tools (Ferry et al. 2014; Gurovich et al. 2019), 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 (Figure 6). 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 (Appendix A.1 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 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 (McKusick 1969), 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.



**Figure 6: 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, e.g., Cornelia de Lange syndrome (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 novel 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*.

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

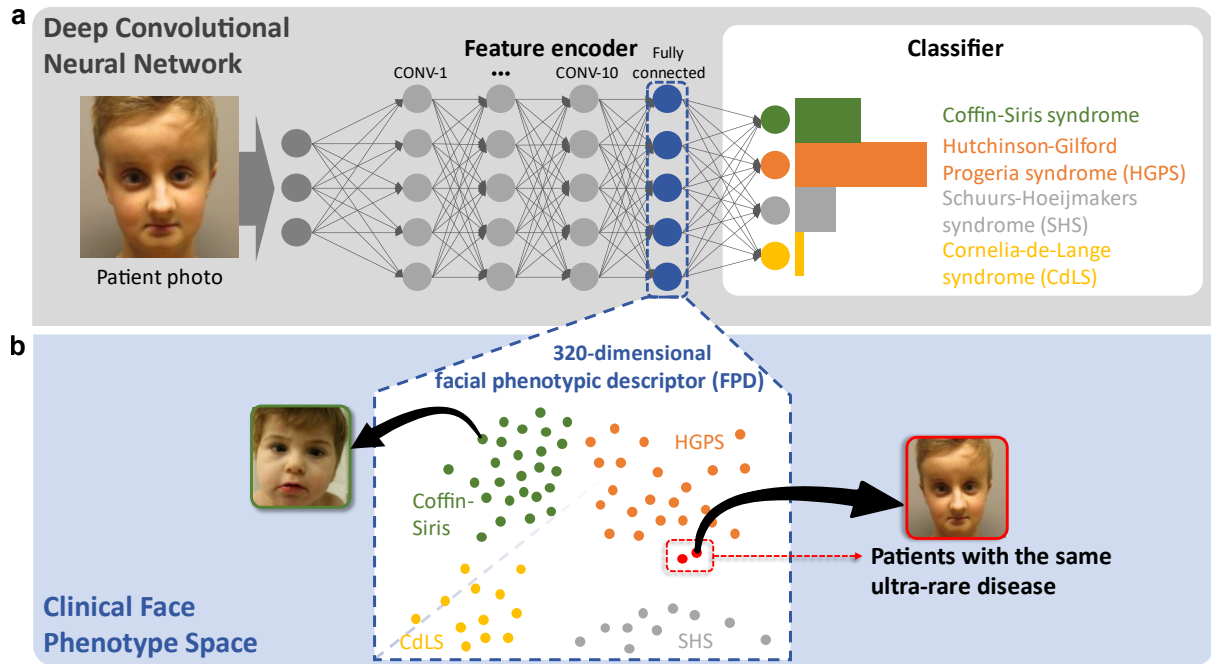
syndromes or identifying similarities between multiple patients with unknown disorders and thereby helping to define new syndromes. Because GestaltMatcher quantifies similarities between faces in this way, it addresses all three of the limitations described above: (1) it can identify “closest matches” 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.

## 4.4 Results

**Overview.** The feature encoder of GestaltMatcher computes a Facial Phenotype Descriptor (FPD) for each portrait image (Figure 7a). Each FPD can be thought of as one coordinate in the CFPS (Figure 7b). The distances between the FPDs in the CFPS form the basis for syndrome classification, delineation of novel 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 Face2Gene (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 novel 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) (see Methods for details). 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 data set 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 on the GMDB data; results obtained with the GMDB data are included in Appendix A.1 Supplementary Information.



**Figure 7: Concept of GestaltMatcher.** **a**, Architecture of a deep convolutional neural network 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 a Facial Phenotype 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 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 (Marbach et al. 2019).

#### 4.4.1 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 (Yi et al. 2014). 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 3). 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 novel disorders, the characteristic features were not yet optimally represented by *Enc-F2G* because features of these disorders were not part of the training set.

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* vs. *Enc-F2G* in Appendix A.1 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.



**Table 3: 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%	<b>35.94%</b>	<b>52.45%</b>	<b>63.91%</b>	<b>78.13%</b>
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%	<b>13.66%</b>	<b>23.62%</b>	<b>29.56%</b>	<b>40.94%</b>
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 <sup>a</sup>	2,669	1,115 <sup>c</sup>	0.09%	<b>20.15%</b>	<b>37.81%</b>	<b>46.85%</b>	<b>64.21%</b>
F2G-frequent	Enc-healthy	22,298 <sup>a</sup>	2,669	1,115 <sup>c</sup>	0.09%	9.70%	22.51%	29.80%	48.24%
F2G-rare	Enc-F2G	22,298.8 <sup>b</sup>	1,183.3	1,115 <sup>c</sup>	0.09%	<b>7.07%</b>	<b>14.19%</b>	<b>17.67%</b>	<b>24.41%</b>
F2G-rare	Enc-healthy	22,298.8 <sup>b</sup>	1,183.3	1,115 <sup>c</sup>	0.09%	4.02%	8.84%	11.73%	16.61%

The deep convolutional neural networks of Enc-F2G (softmax), Enc-F2G, and Enc-healthy have the same architecture. Training of Enc-F2G (softmax) and Enc-F2G was initiated with CASIA-WebFace and further fine-tuned on photos of patients in the Face2Gene 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.

<sup>a</sup> Number of images in the frequent gallery + rare gallery.

<sup>b</sup> Average of ten splits in the frequent gallery + rare gallery.

<sup>c</sup> Number of syndromes in the frequent gallery + rare gallery.

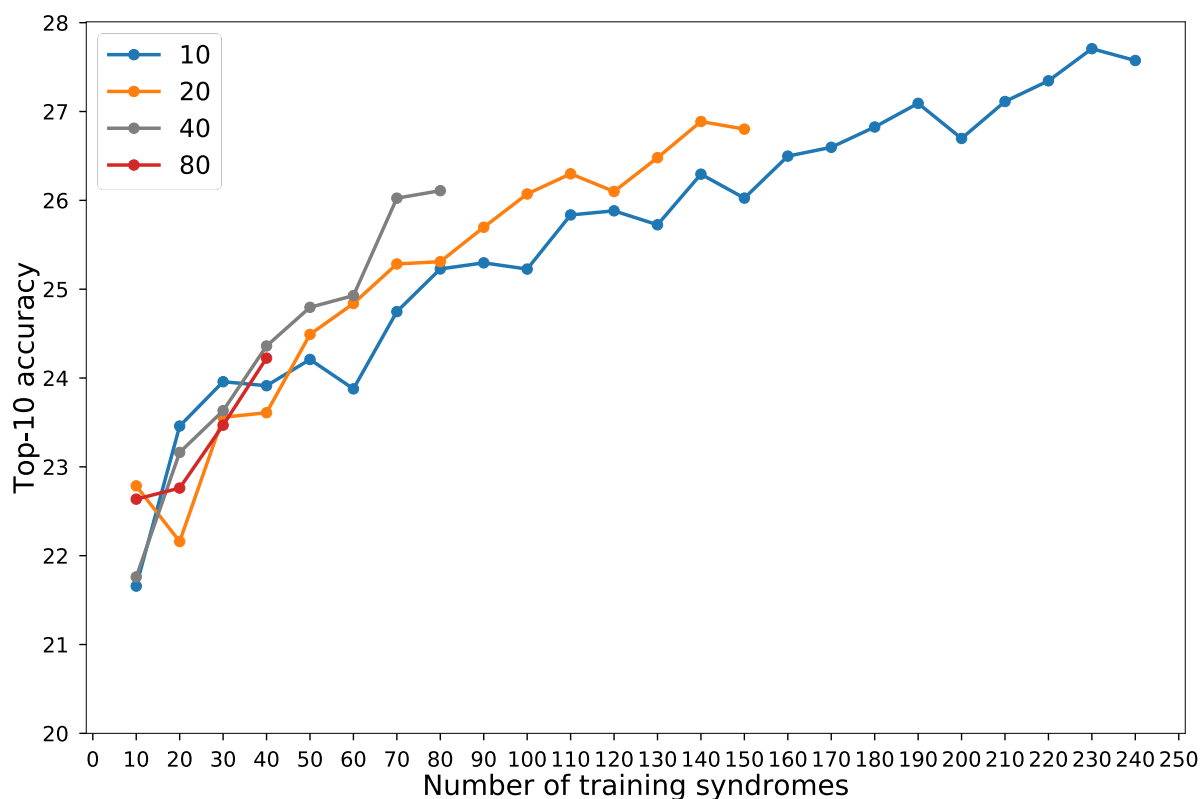
#### 4.4.2 Syndromic diversity improves matching with novel phenotypes

Earlier definitions of the FPD were mainly based on training a network with a small selection of common and highly characteristic syndromes (Ferry et al. 2014; Dudding-Byth et al. 2017). 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

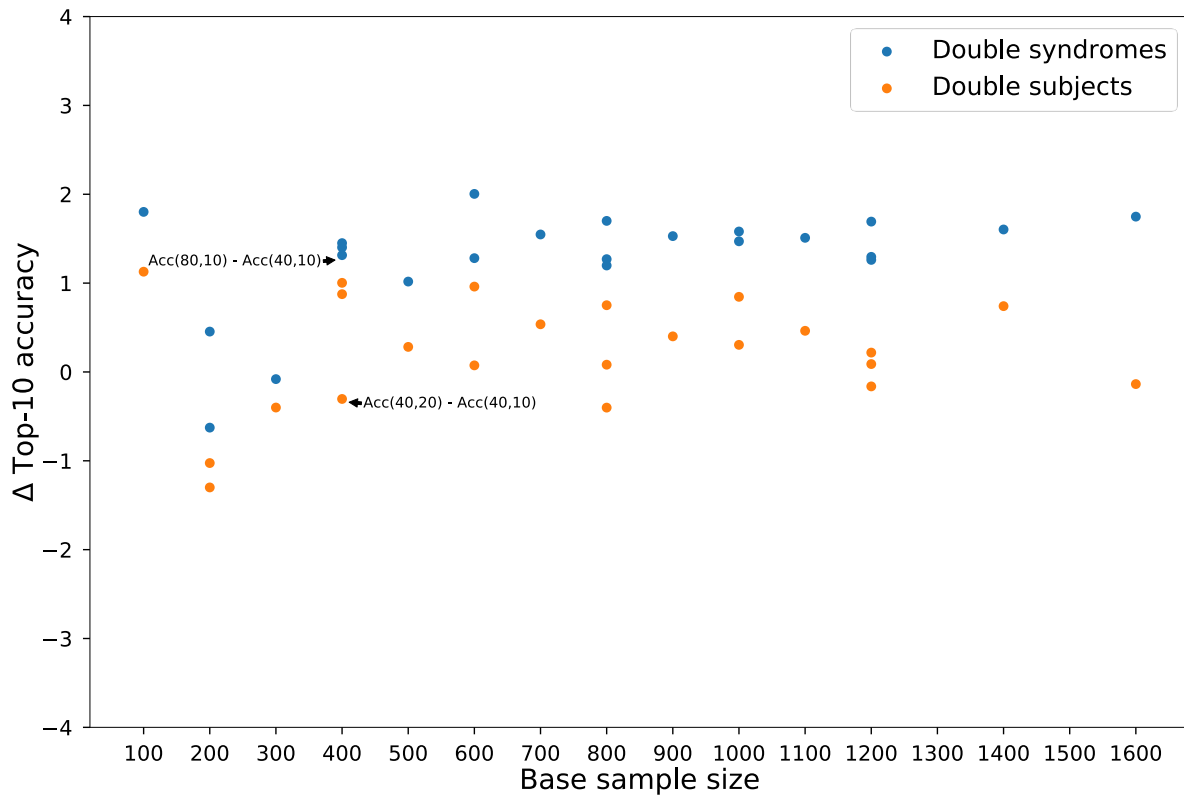
encoder's fine-tuning by incrementally increasing their number starting with the most frequent ones. Due to the imbalance in prevalence among the disorders added each time, the improvement could be affected by the additional number of training subjects. Therefore, we used the same number of subjects 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 is described in detail in the Methods.

When we increased the number of training syndromes, the accuracy increased (Figure 8). In general, the performance was also higher when more individuals per syndrome were used for training. Particularly when more than 50 syndromes are used, the curve for training with 20 subjects/syndrome was above the curve for 10 subjects/syndrome, and so on. The same trend is also shown in the public GMDB dataset (Appendix A.1 Supplementary Figs. 2 and 3).

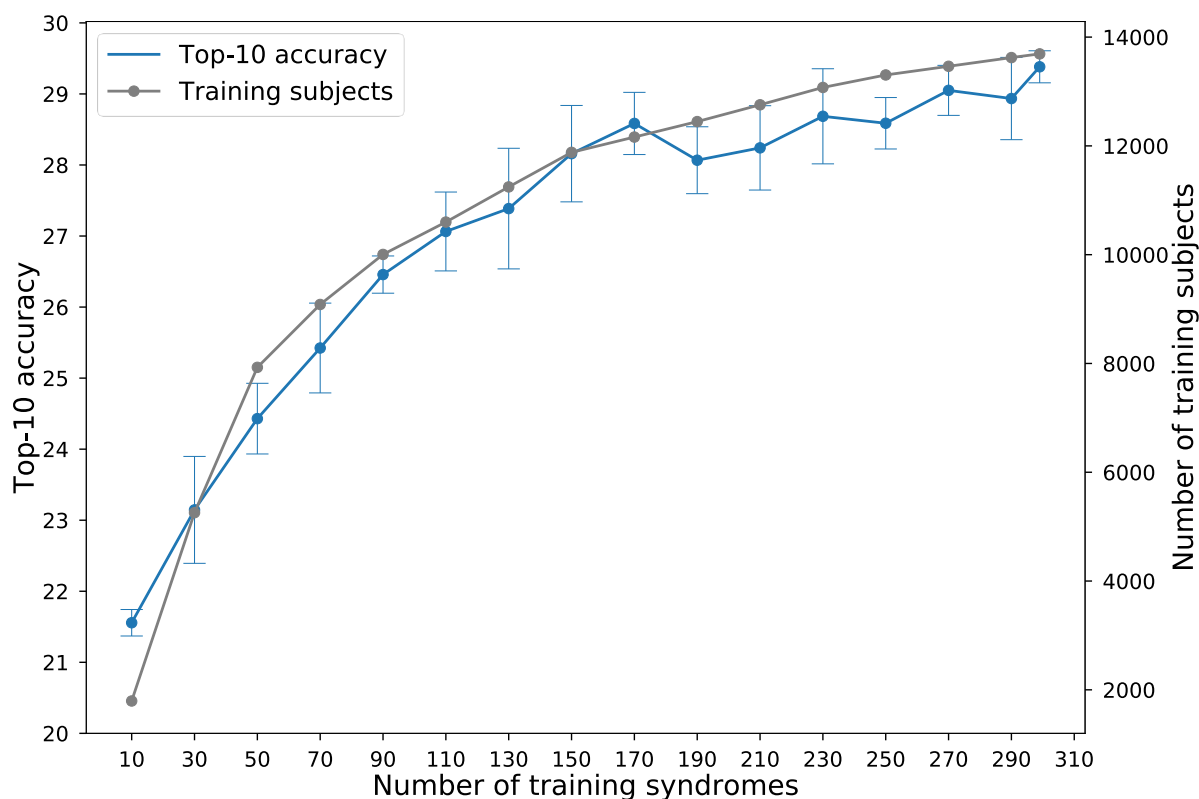
Moreover, using double the number of syndromes is better than using double the number of subjects for most of the combinations (Appendix A.1 Supplementary Fig. 4), and the effect of doubling the number of syndromes used for training is greater when the base sample size is larger than 1,200 subjects (Figure 9 and Appendix A.1 Supplementary Fig. 5). Both of these findings suggest that increasing the syndromic diversity in the training set improves the performance for novel 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 (Appendix A.1 Supplementary Note and Figure 10). In the following section, the Enc-F2G encoder is based on the 299 previously described syndromes.



**Figure 8: 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% (10/816).



**Figure 9: 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 by 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.  $\Delta$ Top-10 accuracy is the difference of accuracy between the double syndromes or subjects and the base point, and is calculated based on Figure 8. 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 Figure 8. The lower point is subtracting the point of 10 subjects and 40 syndromes from the point of 20 subjects and 40 syndromes in Figure 8. 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.



**Figure 10: 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).

### 4.4.3 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) (Winter and Baraitser 1987) that were already used for benchmarking the performance of DeepGestalt (Gurovich et al. 2019). 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 a Enc-F2G softmax approach (Appendix A.1 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 (Appendix A.1 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 3). The results with the GMDB-frequent test set also dropped only slightly while supporting more than twice the number of syndromes (Appendix A.1 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.

#### **4.4.4 Matching undiagnosed patients from unrelated families**

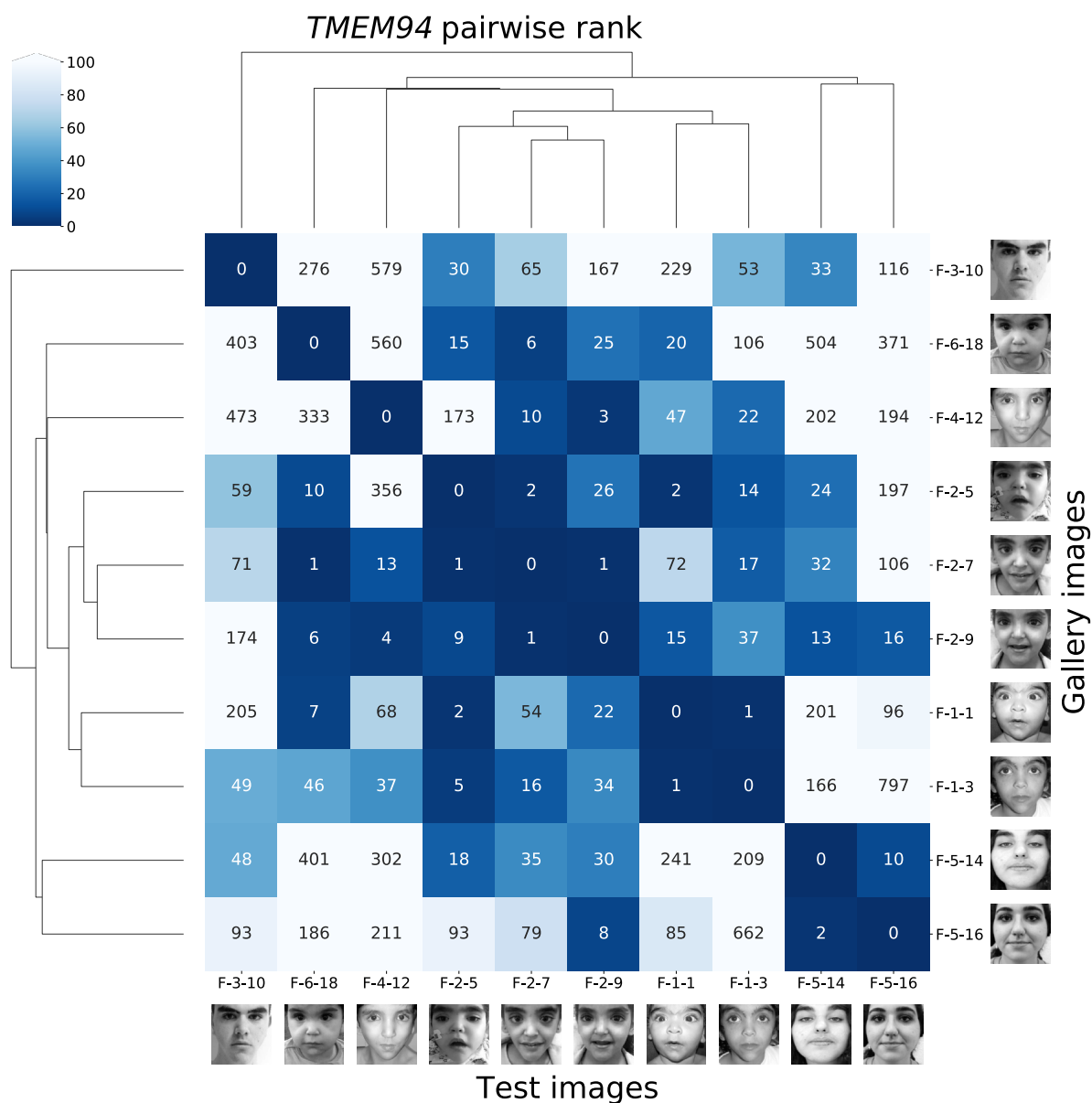
In the second use case, we envision GestaltMatcher as a phenotypic complement to GeneMatcher (Sobreira et al. 2015). To prove that we can match patients from unrelated families who have the same disease by using only their facial photos, we selected syndromes from 15 recent GeneMatcher publications with titles containing the phrase “facial dysmorphism” (Stankiewicz et al. 2017; Morimoto et al. 2018; Tanaka et al. 2016; Weiss et al. 2016; Balak et al. 2019; Harms et al. 2017; Jansen et al. 2019; Au et al. 2015; Diets et al. 2019; Marbach et al. 2019; Santiago-Sim et al. 2017; Olson et al. 2018; Stephen et al. 2018; Kanca et al. 2019; Stevens et al. 2016). In contrast to the benchmarking of the previous section, the gallery now consists of individuals with rare syndromes 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 non-matching 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 non-matching individuals with higher similarity to the test case together have only four non-matching 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.

**Table 4: Matching of novel phenotypes on a GeneMatcher validation set.**

Gene	Total families (subjects)	Connected families (subjects) <sup>a</sup>	
		Top-10	Top-30
<i>BPTF</i> (Stankiewicz et al. 2017)	6 (6)	0 (0)	2 (2)
<i>CCDC47</i> (Morimoto et al. 2018)	4 (4)	0 (0)	0 (0)
<i>CHAMP1</i> (Tanaka et al. 2016)	4 (4)	2 (2)	4 (4)
<i>CHD4</i> (Weiss et al. 2016)	3 (3)	0 (0)	0 (0)
<i>DDX6</i> (Balak et al. 2019)	4 (4)	4 (4)	4 (4)
<i>EBF3</i> (Harms et al. 2017)	6 (7)	0 (0)	0 (0)
<i>FBXO11</i> (Jansen et al. 2019)	17 (17)	5 (5)	9 (9)
<i>HNRNPK</i> (Au et al. 2015)	3 (3)	3 (3)	3 (3)
<i>KDM3B</i> (Diets et al. 2019)	9 (9)	0 (0)	2 (3)
<i>LEMD2</i> (Marbach et al. 2019)	2 (2)	2 (2)	2 (2)
<i>OTUD6B</i> (Santiago-Sim et al. 2017)	4 (9)	3 (4)	3 (6)
<i>PACS2</i> (Olson et al. 2018)	6 (6)	0 (0)	2 (2)
<b><i>TMEM94</i></b> (Stephen et al. 2018)	<b>6 (10)</b>	<b>5 (8)</b>	<b>6 (10)</b>
<i>WDR37</i> (Kanca et al. 2019)	4 (4)	2 (2)	3 (3)
<i>ZNF148</i> (Stevens et al. 2016)	3 (3)	0 (0)	0 (0)
Total	79 (91)	26 (30)	40 (48)
Average	-	32.91% (32.97%)	50.63% (52.75%)

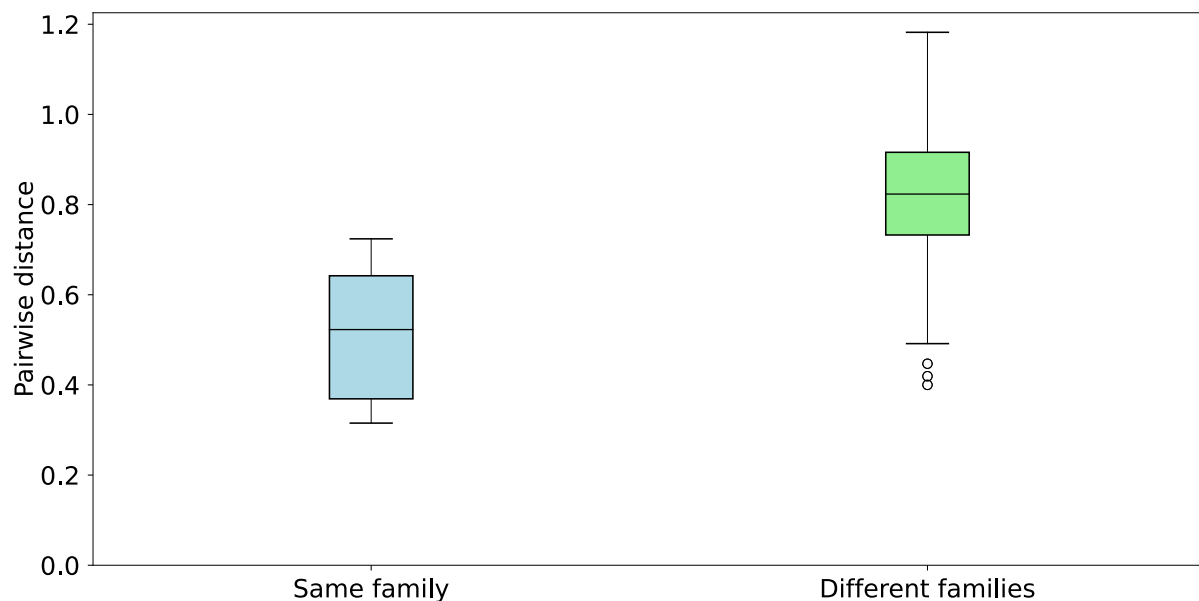
In the discovery mode for novel phenotypes (second use case), all cases in the gallery are without diagnosis. For the performance readout, only the correct disease gene of a match is revealed. As an example, for individuals of the *TMEM94* study (shown in bold in the table), eight out of ten subjects 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 Figure 11 and Appendix A.1 Supplementary Figure 6. The accuracy of connected subjects corresponds to the accuracy of using Enc-F2G on the F2G-rare test set (shown in Table 3), but in discovery mode in a gallery of almost the same size as F2G rare gallery set.

<sup>a</sup> Number of families (subjects) matched by a photo from another family in the top-10 or top-30 rank.



**Figure 11: 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 (Stephen et al. 2018). 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.





**Figure 12: 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 1<sup>st</sup> to the 3<sup>rd</sup> quartiles. The total number of data points (n) for the same family is 28, and n is 928 for the different families.

In this scenario, we matched 30 of 91 subjects and connected 26 of 79 families when using the top-10 criterion (Table 4, Figure 11 and Appendix A.1 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 (Appendix A.1 Supplementary Table 4). Hence, using the encoder trained with facial dysmorphic individuals improves the matching considerably.

As an example, in a study of *TMEM94* (Stephen et al. 2018), 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 (Figure 11). 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

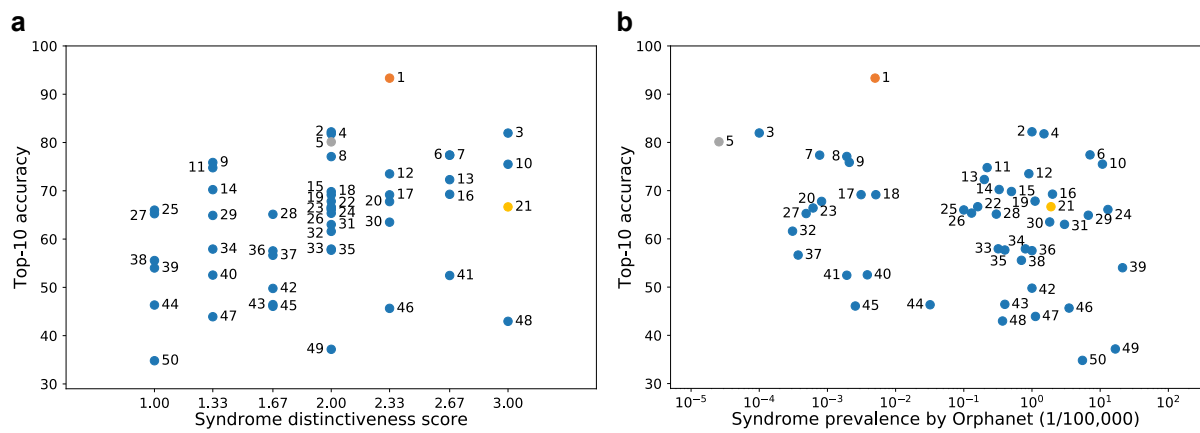
sister, F-2-9, matched families 1, 4, 5, and 6. The six families were recruited at five different institutes in India, Qatar, the United States (NIH Undiagnosed Diseases Network), and Switzerland, indicating that GestaltMatcher can also connect patients of different ancestries. However, a more systematic 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 (Figure 12). 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 (Alvi, Zisserman, and Nellåker 2019) and will also diminish over time when more diverse training data become available (Lumaka et al. 2017).

#### **4.4.5 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 (Schuurs-Hoeijmakers et al. 2012), 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 195/299 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 (Figure 13a and Appendix A.1 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 (Appendix A.1 Supplementary Fig. 7 and Supplementary Table 6).

The correlation for GestaltMatcher accuracy and disease prevalence was not significant ( $P = 0.130$ ; Figure 13b). 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 novel phenotypes in the real world.

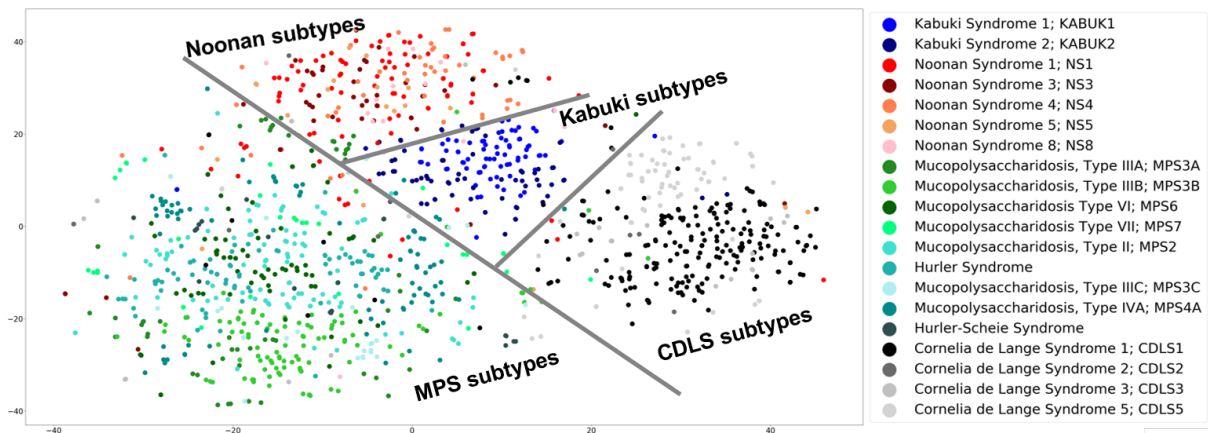


**Figure 13: Correlation among syndrome prevalence, distinctiveness score, and top-10 accuracy.** **a**, Distribution of top-10 accuracy and distinctiveness score. The Spearman rank correlation coefficient was 0.400 ( $P = 0.004$ ). **b**, Distribution of top-10 accuracy and prevalence. The Spearman rank correlation coefficient was  $-0.217$  ( $P = 0.130$ ). The details of each syndrome can be found in Appendix A.1 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.

#### 4.4.6 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 lab, it becomes a research case, and they may compare the patient’s condition to known disorders. For example, a potentially novel 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 (Appendix A.1 Supplementary Table 7).

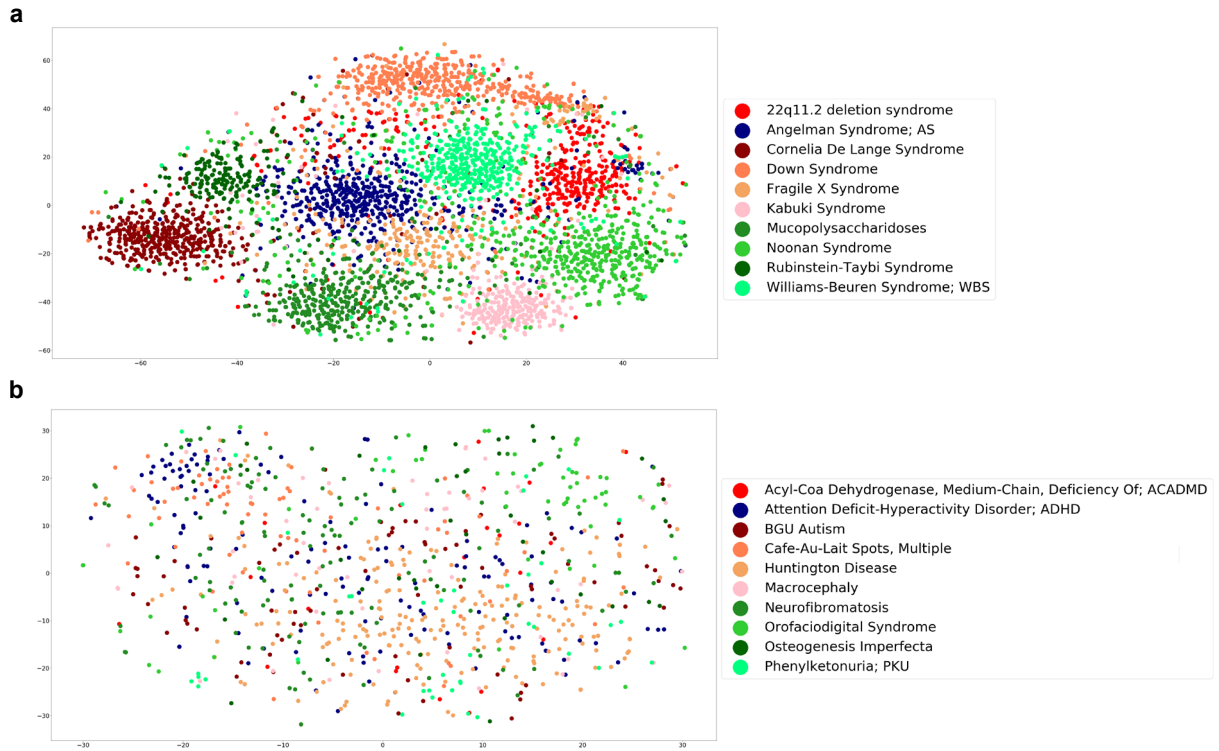
If a novel disease gene has been identified and the similarities of the patients to known phenotypes outweigh the differences, OMIM groups them into a phenotypic series. On the gene or protein level, such phenotypic series often correspond to molecular-pathway diseases, such as GPI-anchor deficiencies for hyperphosphatasia with mental retardation syndrome (HPMRS) or cohesinopathies for CdLS. For our cluster analysis, we sampled 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*-SNE (van der Maaten and Hinton 2008) projection of the FPDs into two dimensions yielded the best visualization results (Figure 14). 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.



**Figure 14: Hierarchical clustering of four phenotypic series, Kabuki syndrome, Noonan syndrome, mucopolysaccharidosis, and Cornelia de Lange syndrome, using a *t*-SNE projection of the Facial Phenotype Descriptors.**

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 non-distinct syndromes (syndromes without facial dysmorphism) into 2D space. In addition, we calculated the Silhouette index (Rousseeuw 1987) 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 non-distinct syndromes created

no clear clusters (Figure 15). Moreover, the Silhouette index of the frequent 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.

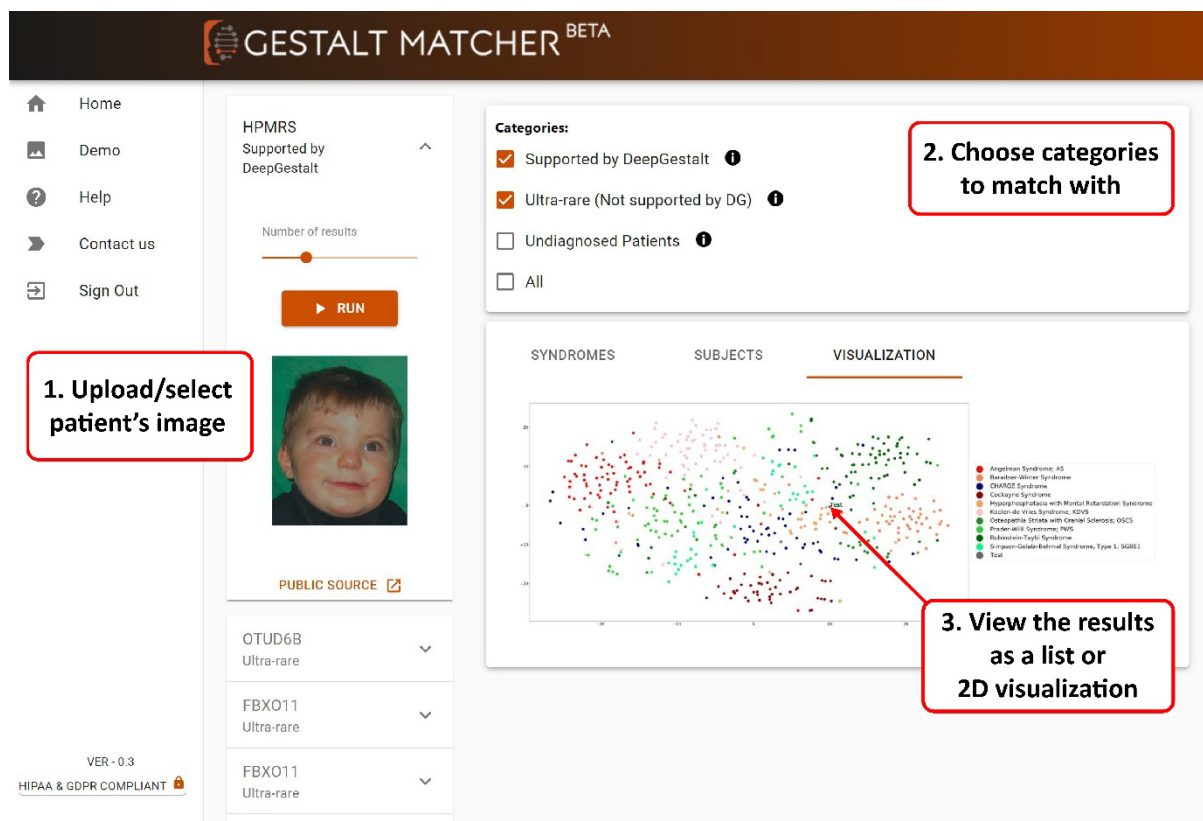


**Figure 15: *t*-SNE visualization of Facial Phenotype Descriptors of (a) ten syndromes with and (b) ten syndromes without facial dysmorphism.**

#### 4.4.7 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 novel disease gene *PSMC3* in a demonstration on the GestaltMatcher web service (Figure 16, [www.gestaltmatcher.org](http://www.gestaltmatcher.org)). Ebstein *et al.* (Ebstein et al. 2021) 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 *PACSI*, 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 22 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.



**Figure 16: 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](http://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, e.g., *PSMC3* in the online demo).

## 4.5 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 novel 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 (Figure 13a), whereas syndrome prevalence did not (Figure 13b). 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 (Appendix A.1 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.

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 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 (Richards et al. 2015; Tavtigian et al. 2018). (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 novel 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 (Philippakis et al. 2015), 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 HPO 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 novel disease caused by the same *de novo* mutation in *LEMD2* (Marbach et al. 2019), our approach has successfully been applied to other ultra-rare disorders (Figure 6). We matched 40 of 79 different families in 15 GeneMatcher publications by top-30 rank (Figure 11 and Appendix A.1 Supplementary Fig. 6), and 11 candidate genes are currently under evaluation. This result shows the power and potential of GestaltMatcher to identify novel 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.

## **4.6 Methods**

### **4.6.1 Study approval**

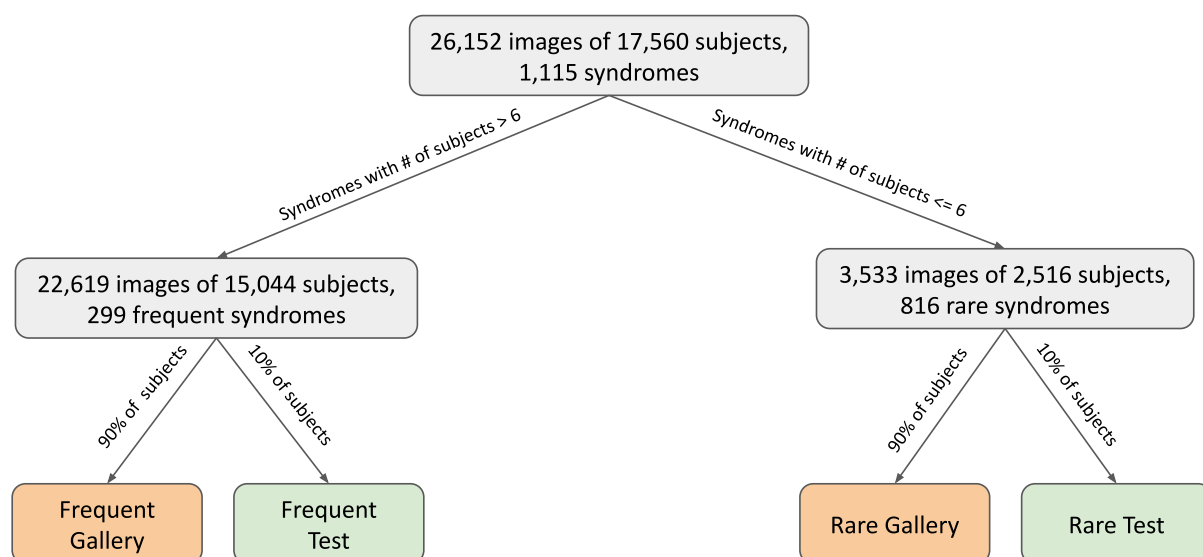
This study is governed by the approval of the following Institutional Review Boards: Charité–Universitätsmedizin Berlin, Germany (EA2/190/16); UKB Universitätsklinikum Bonn, Germany (Lfd.Nr.386/17). The authors have obtained written informed consent from the patients or their guardians, including permission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

### **4.6.2 Face2Gene datasets**

We collected images of individuals 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 26,152 images from 17,560 individuals with a total of 1,115 syndromes (Appendix A.1 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. Figure 17 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 “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 (Figure 6). 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.



**Figure 17: 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.

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%

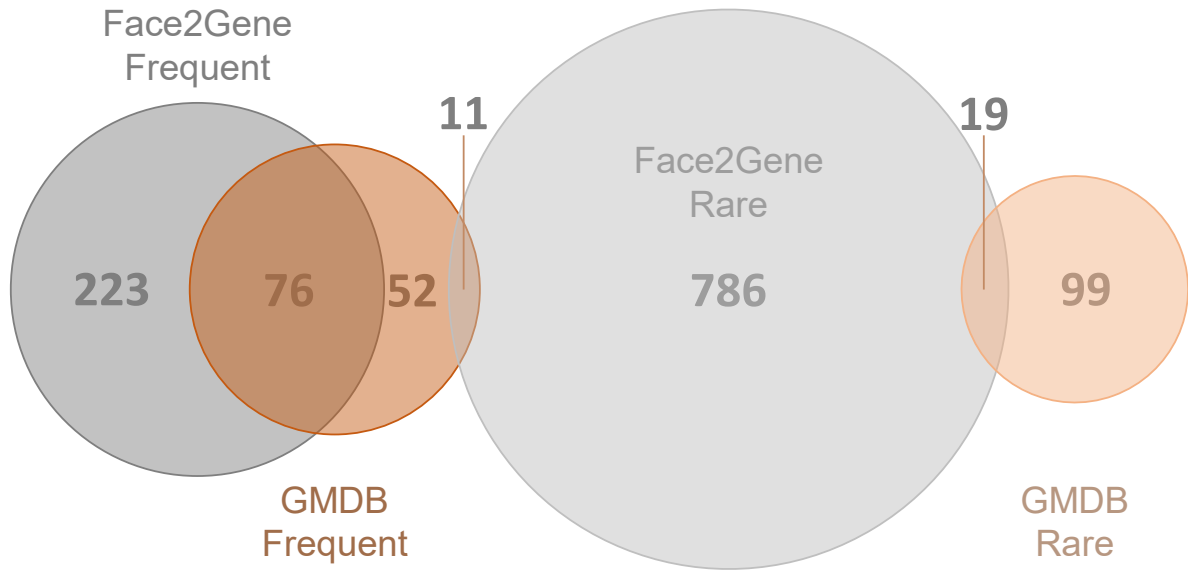
of subjects were used to validate the DeepGestalt training; the 90% then became the frequent gallery and the 10% were assigned to the frequent test set. For the rare dataset, we performed 10-fold cross-validation. In each syndrome, 90% and 10% of subjects were assigned to the gallery and test set, respectively. 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.

### **4.6.3 GMDB datasets**

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 GestaltMatcher Database ([www.gestaltmatcher.org](http://www.gestaltmatcher.org)).



**Figure 18: Venn diagram of numbers of syndromes in the Face2Gene and GMDB datasets.**

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 (Appendix A.1 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 Face2Gene 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 Appendix A.1 Supplementary Figure 8. Of the 3,693 individuals in GMDB, 963 are also in the Face2Gene dataset. To use the GMDB rare set as the test set for both the GMDB frequent set and the Face2Gene frequent set, we made sure that no syndrome was in both the GMDB rare set and the Face2Gene frequent set (Figure 18).

#### 4.6.4 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 deep convolutional network (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 (facial phenotype descriptor, FPD), resulting in a vector of size 320.

DeepGestalt was first trained on images of healthy individuals from CASIA-WebFace (Yi et al. 2014), and later fine-tuned on a dataset with patient images (Face2Gene or GMDB). The encoder without fine-tuning on patient images was called Enc-healthy. The encoder later trained on 299 frequent syndromes in the Face2Gene 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 Face2Gene and GMDB datasets. The summary of all the encoders used in this study is shown in Appendix A.1 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.

### 4.6.5 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 7. 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.

### 4.6.6 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 (percent 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.

### 4.6.7 London Medical Dataset validation analysis

We compiled 323 images of patients diagnosed with 91 frequent syndromes from the LMD publication test set (Winter and Baraitser 1987; Gurovich et al. 2019) 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).

#### 4.6.8 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.

#### 4.6.9 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 novel 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 4. The 15 genes were not among the Face2Gene frequent syndromes, so we can consider them each as a novel 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.

#### **4.6.10 Syndrome facial distinctiveness score**

To evaluate the importance of the facial gestalt for clinical diagnosis of the patient, we asked three dysmorphologists (co-authors 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.

#### **4.6.11 Syndrome prevalence**

The prevalence of each syndrome was collected from Orphanet ([www.orpha.net](http://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 (Nguengang Wakap et al. 2020).

#### **4.6.12 Unseen syndromes correlation analysis**

To investigate the influence of prevalence and distinctiveness score on the performance of novel 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 Appendix A.1 Supplementary Figure 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 Face2Gene 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 Appendix A.1 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.

#### **4.6.13 Analysis of number of training syndromes and subjects**

In this analysis, we evaluated the influence of training with additional syndromes and subjects to the novel 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 9<sup>th</sup> position in the first list, but in the 20<sup>th</sup> 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 ten subjects, where  $X = 10$  to 240, incremented at an interval of ten 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 ten 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 Figure 8. We also used the models described above to encode the GMDB dataset, tested them with the GMDB rare set, and report the results in Appendix A.1 Supplementary Figure 2.

Because the GMDB dataset is smaller than Face2Gene 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 Appendix A.1 Supplementary Figure 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 influence of adding more syndromes and adding more subjects. The results are shown in Figure 9 and Appendix A.1 Supplementary Figures 4 and 5.

## 4.7 Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through individual grants to P.M.K. (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 (EXC2151–390873048, ImmunoSensation2). A. Schmidt received additional support by the BONFOR program of the Medical Faculty of the University of Bonn (2020-1A-15). We also acknowledge support from the TRANSLATE-NAMSE project. We are also grateful for the editorial assistance provided by Dr. Naomi Ruff.

## 4.8 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](http://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.

## 4.9 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 are 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](http://www.cbsr.ia.ac.cn/english/casia-webFace/casia-webfAce_AgreEmeNtS.pdf), and requests should be sent to [cbsr-request@authenmetric.com](mailto:cbsr-request@authenmetric.com). For GMDB, please contact [info@gestaltmatcher.org](mailto: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.

## **Chapter 5   GestaltMatcher Database: medical imaging data for deep learning on rare disorders**

### **5.1 Introduction**

The next-generation phenotyping (NGP) technology has rapidly progressed in the last decade. NGP applications such as Face2Gene or GestaltMatcher are increasingly used by geneticists and pediatricians in the diagnostic workup of patients with facial dysmorphism (Gurovich et al. 2019; T.-C. Hsieh et al. 2022). The high performance of these tools is achieved by extending deep convolutional neural networks (DCNNs) that were developed for related but non-clinical tasks, such as, e.g., intra-person face verification (Yi et al. 2014). Training such a network complemented by an additional layer, which can address the medical classification problem, becomes feasible after knowledge transfer with much fewer data (Gurovich et al. 2019; T.-C. Hsieh et al. 2022). However, despite the increasing interest and technological advances, properly labeled training data is currently the biggest bottleneck in developing NGP applications (Hennekam and Biesecker 2012).

Furthermore, the existing data are often siloed, so curation is often done repeatedly (Nellåker et al. 2019). In addition, one of the key challenges in collecting data is that curation usually needs to be done by clinicians and not computer scientists. Hence, each research group will spend tremendous effort and time collecting their datasets from publication or their patients, raising the threshold for entering this research field. The non-transparency and non-shareability of data also increase the difficulty in benchmarking the methods proposed by different research teams that delay the development of novel approaches.

Human and artificial neural networks used in a professional context in Medical Genetics need to be trained on image data for pattern recognition. For residents in syndromology, there is no simple way to get exposure to, e.g., many characteristic portraits of dysmorphic patients at a glance. For the dysmorphology analysis of rare disorders, London Medical

Database (LMD) (Winter and Baraitser 1987) is one of the most well-known databases. LMD was an essential resource for clinicians and NGP technology, but people now often question its outdated resource, and many syndromes in the database are based on differential diagnoses only.

To solve the problem of lacking a public medical image dataset, Minerva & Me has been proposed in 2019 to enable global collaboration on patient data (Nellåker et al. 2019). However, Minerva & Me is currently not publicly available. Hence, the requirement for collecting and sharing medical images is still urgent to be solved.

In addition, patient recruitment is usually time-consuming due to acquiring a consent form signed by the patient or their guardian. A faster and easier way for the patient recruiting process is needed.

Therefore we designed GestaltMatcher DB, a web framework that addresses the needs of human syndromologist first and yields data curation as a by-product. We also implemented infinitive scrolling to display photos collections to make medical imaging data from the scientific literature explorable. Besides, the electrical consent allows the patient to easily sign the document with the link provided by the clinicians. With this incentive, we could curate hundreds of case reports with a community-driven effort in a short time.

Currently, GestaltMatcher DB focuses on medical imaging data of patients with rare monogenic diseases and is currently mainly populated by but not limited to photos of facial portraits. Additional data under curation are X-rays documenting skeletal malformations and photos of the fundus of the eye documenting retinal diseases. By that means, GestaltMatcher DB provides training data not only for detecting a typical facial gestalt but also for characteristic phenotypes of the bone or the retina.

## 5.2 Methods

We first built an online platform by Ruby on Rails to allow users to input the images and other patient data. For the back-end, we set up a database by MySQL to store the patient data. We asked five medical students to collect and curate the patients from the publications,

and we later involved the other around 30 clinician-scientists as the beta user to contribute the patients from their papers or their unpublished patients with proper consent.

The patient data has the following categories: general information, publication or consent, group, files (images and documents), phenotypic information, and genomic information. The database schema is shown in Figure 19.

**General information** includes patient label, sex, ethnicity, and note. The patient label is the name that the clinician can use to identify the patient, and the actual name is not allowed for the patient to protect the patient's privacy. We encourage the clinicians to fill up sex and ethnicity to analyze the confounder effect.

**Ethical, Legal, and Social Aspects (ELSA)** include the PubMed ID, DOI, the numbering used in the publication, and the corresponding author of the publication. When the patient is not published yet, consent from the patient in paper or electric form is required. With this information, we can source back to the patient described in the original paper and handle legal issues such as reusing the images for another publication.

**Group** is the patient group where clinicians or the patient's family can contact each other. For example, Sirius e.V. is a patient group for Smith Magenis Syndrom (<https://smith-magenis.de/>). Another patient group is the Kontaktgruppe "Eltern kleinwüchsiger Kinder" (<http://www.kleinwuchs-elterngruppe.de/>). When the patient is recruited from the patient group, we will create the patient, set the group\_id in the patient, and send the electronic consent form by an invitation link to the patient or the patient's guardian.

**Images** have several types: frontal photo, profile, limbs, hand X-ray, and funduscopy image. The age of when the photo was taken is recorded. Moreover, we ask the clinician to grade how well this medical photography can contribute to the diagnosis based on the dysmorphism. There are three levels: significant, supportive, not supportive.

**Documents** include medical reports and lab results that curators or clinicians can review. However, the actual name or any identifier that can be traced back to the patient should be removed.

**Phenotypic information** contains disorder and clinical descriptions. The disorders are based on the disorder list from OMIM. The clinician can choose whether the disorder is Differentially diagnosed, Clinically diagnosed, or Molecularly diagnosed. We use HPO terms for the clinical description, and the clinician can select whether the HPO term is present or absent.

**Genomic information** contains gene and disease-causing mutation, and the clinician can specify the gene that is diagnosed as the disease-causing gene. When the disease-causing mutation is known, we store the mutation in the HGVS code and the zygosity.

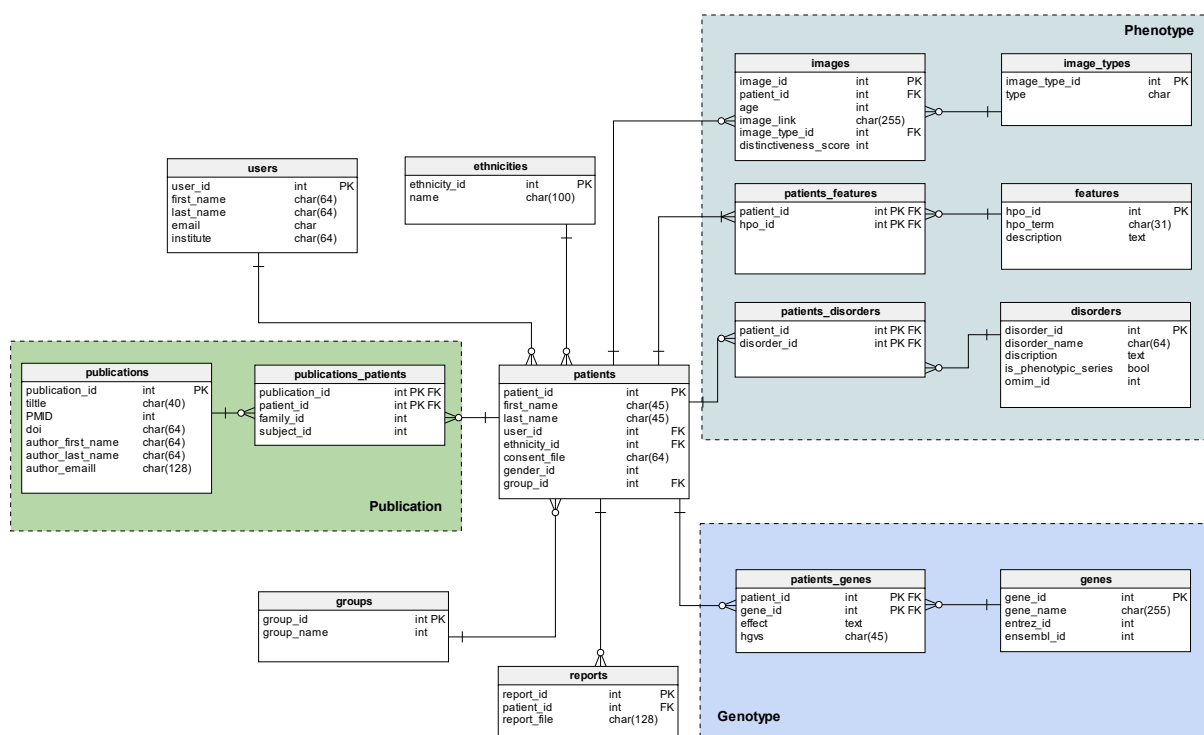


Figure 19: GMDB database schema.

## 5.3 Online platform and database

The GestaltMatcher Database (GMDB) is hosted physically in the university hospital of Bonn and guarded by Arbeitsgemeinschaft für Gen-Diagnostik e.V. (AGD) that is a non-profit organization for the genome research. The status of GMDB will be reported together with other genome research talks in the annual meeting. All the AGD members have free



access to GMDB. Until January 2022, we collected 5925 patients with 9173 images from 1702 publications, and the patients covered 620 rare disorders.

### 5.3.1 GUI for patient data annotation

Users can input the patient data such as general information, publication, phenotypic and genomic data. Here we take a patient curated from the publication (Goldenberg et al. 2016) as an example.

Figure 20 shows the patient curated by Prof. Shahida Moosa. The information listed in the ELSA region, such as the PubMed ID, DOI allows us to trace back to the original publication.









Moreover, this patient has two photos listed in the original publication, and here we have one frontal image and one profile image in GMDB (Figure 21). The corresponding age when the photo was taken is also stored. The score column is for the distinctiveness score with three categories (supportive, important, and key).

Patient Information		
<b>Case ID:</b> 4135	<b>Clinicians Reference:</b>	<b>User:</b> Prof. Shahida Moosa
<b>Ethnicity:</b> Caucasian	<b>Ethnicity note:</b>	<b>Gender:</b> male
<b>Group:</b> None		
<b>Note:</b> -		
Ethical, Legal, and Social Aspects (ELSA)		
<b>PubMed:</b> <a href="#">29258554</a>	<b>DOI:</b> <a href="#">10.1186/s13023-017-0736-8</a>	<b>Consent obtained:</b> <input type="checkbox"/>
<b>Family numbering:</b> -	<b>Subject numbering:</b> 1	
<b>Corresponding author or clinician that obtained informed consent</b> Mustafa Tekin	<b>Email:</b> mtekin@med.miami.edu	<button>Add another patient</button>

**Figure 20: Patient information in GMDB.**

In the end, the phenotypic and molecular information is stored, as shown in Figure 22. OMIM ID is used to identify disorders, and we present the HPO terms for the clinical features. The molecular information stores the information from the genetic testing. The

disease-causing gene of this patient is *ANKRD11*, and it is diagnosed by microarray. If the mutation information is given, it could be stored in the form of the HGVS code.

Photos and Documents									
Upload		Gallery photo: 8399							
ID	Image	Type	Age	Age note	Which person	Score	Private	Updated date	
8399		Frontal face	20.0	-	index	Supportive	N	2022-01-17	  
8400		Profile	20.0	-	index	Supportive	N	2022-01-17	  

**Figure 21: Patient photos in GMDB.** The images are from the publication (Goldenberg et al. 2016).

Diagnosed disorders		
OMIM	Disorder	Diagnosed
<a href="#">148050</a>	KBG SYNDROME	clinically diagnosed

Molecular Information		
Gene	Test	HGVS
<a href="#">ANKRD11</a> <a href="#">29123</a>	microarray none	


  

Phenotypic Information		
HPO	Description	Status
HP:0000311	Round face	Present
HP:0000316	Hypertelorism	Present
HP:0000426	Prominent nasal bridge	Present

**Figure 22: Phenotypic and genomic information in GMDB.**

### 5.3.2 Digital consent for easier patient recruitment

When the patient is not published yet, we require consent from patients for using their data. If the clinicians already had the consent from patients in paper form, they can upload the consent after scanning the consent. However, it is usually time-consuming to acquire consent in paper form because the clinician does not frequently meet with patients. We provided the digital consent that allows the patient to sign the consent on their page to smooth this recruiting process. We first created a new patient in GMDB, and there will be an invitation link shown at the top of the patient's page (Figure 23). We then sent this link to the patient, and the patient can sign the document directly on the webpage (Figure 24). In this way, we can also recruit the patients from the patient group on the internet.

Patient access link: <https://db.gestaltmatcher.org/patients/5564?token=AW6J6R2cb2ojDcfqtKMd> [Invite Patient](#) 

Patient Information		
Case ID: 5564	Clinicians Reference:	User: Mr. Tzung-Chien Hsieh
Ethnicity:	Ethnicity note:	Gender: unknown
Group: None		
Note: -		

**Figure 23: Patient invitation link.** The patient can access the patient's page by the link.

[Home](#)
[Downloads](#)
[About](#)
[Sign up](#)
[Login](#)

Sehr geehrte Damen und Herren,

vielen Dank für Ihr Interesse an der GestaltMatcher Forschungsstudie und Datenbank (GMDB)! Ihre Teilnahme ist freiwillig. Bevor Sie uns Daten/Fotos/weitere Informationen zur Verfügung stellen, vergewissern Sie sich bitte, dass Sie die folgenden Studieninformationen gelesen haben und die notwendige Zustimmung erteilen.

### Informationen zur Studie

Sie oder Ihr Kind haben eine genetische Störung (oder es besteht der Verdacht darauf). Um eine Differentialdiagnose zu stellen, können Merkmale, z.B. eine bestimmte Gesichtsform, wichtige Informationen liefern. Zunehmend werden computergestützte Methoden der Bildanalyse zur Beurteilung medizinischer Aufnahmen eingesetzt. Diese Next-Generation-Phenotyping (NGP)-Ansätze dienen dazu, anhand von Fotografien Muster zu erkennen und Ähnlichkeiten zu Individuen mit bereits gesicherten genetischen Erkrankungen zu berechnen. NGP kann dem Arzt Anhaltspunkte für eine genetische Störung liefern und ihn bei der Auswahl eines bestimmten molekularen Tests unterstützen. NGP kann auch zur Interpretation der Ergebnisse molekulargenetischer Untersuchungen verwendet werden.

In unserer Forschungsstudie soll die Qualität der derzeit verfügbaren Bildanalyseverfahren (z. B. GestaltMatcher) untersucht werden. Außerdem werden wir untersuchen, ob ein neues Protokoll zur Bewertung von Mutationen, genannt PEDIA (Priorisierung von Exomdaten durch Bildanalyse), die diagnostische Ausbeute verbessert.

Wenn unsere Forschungsstudie zeigt, dass NGP helfen kann, schneller die richtige Diagnose zu finden, können in Zukunft unnötige Untersuchungen vermieden werden und die Patienten erhalten früher eine angemessene medizinische Versorgung für ihre spezifische Erkrankung. Unsere Studie zielt darauf ab, die Diagnose genetischer Erkrankungen zu verbessern und wird langfristig auch zu deren erfolgreicher Behandlung beitragen. Ein direkter therapeutischer Nutzen für Sie oder ihr Kind ist durch die Teilnahme an diesem Projekt jedoch nicht zu erwarten.

Im Rahmen der Routinediagnostik werden die klinischen Merkmale anhand von Fotos und medizinisch-genetischer Terminologie dokumentiert. Diese phänotypischen Merkmale, die medizinischen Fotografien und die Testergebnisse, sofern vorhanden, werden uns von Ihnen oder Ihrem Arzt in pseudonymisierter Form übermittelt. Pseudonymisierung bedeutet, dass andere Nutzer der Plattform weder den Namen noch die Adresse des Studienteilnehmers kennen. Die klinischen Daten können nur über einen Token, also einen alphanumerischen Code, der nur Ihnen und dem einreichenden Arzt bekannt ist, einer Person zugeordnet werden. Alle Daten der Forschungsstudie werden auf einem Server am Institut für Genomische Statistik und Bioinformatik (IGSB) in Bonn gespeichert, der den aktuellen Datenschutzbestimmungen entspricht. Wir beabsichtigen, die Ergebnisse dieses Forschungsprojekts in wissenschaftlichen Fachzeitschriften und auf Konferenzen zu veröffentlichen. Diese Veröffentlichungen werden keine persönlichen Informationen enthalten, die eine Identifizierung Ihrer Person ermöglichen. Die Teilnahme an der Studie ist kostenlos, und es besteht kein Anspruch auf eine Vergütung, andere finanzielle Vorteile oder Gewinne, die auf der Grundlage dieser Forschung erzielt werden können.

### Einwilligungserklärung

Ich habe das Informationsmaterial über GestaltMatcher und PEDIA gelesen, bzw. es wurde mir vorgelesen. Ich hatte die Möglichkeit, Fragen dazu zu stellen, und alle Fragen, die ich gestellt habe, wurden zu meiner Zufriedenheit beantwortet. Ich willige [für mein Kind] freiwillig in die Teilnahme an dieser Studie ein. Ich bin darüber informiert worden, dass ich meine Einwilligung jederzeit und ohne Angabe von Gründen widerrufen kann. Im Falle eines Widerrufs der Einwilligung werden alle zuzuordnenden Daten gelöscht, und diese Entscheidung wird sich in keiner Weise negativ auf mich auswirken.

Ich bin damit einverstanden, dass diese Fotos in der GMDB gespeichert und in medizinischen Publikationen, einschließlich medizinischer Fachzeitschriften, Lehrbüchern und elektronischen Publikationen, verwendet werden. Ich verstehe, dass die Bilder von Ärzten und Wissenschaftlern, die die Plattform nutzen, gesehen werden können. Obwohl diese Fotos ohne identifizierende Informationen, wie z. B. meinen Namen, verwendet werden, verstehe ich, dass mich möglicherweise jemand erkennt. Ich bin auch damit einverstanden, dass meine Bilder für Trainingszwecke der menschlichen und künstlichen Intelligenz gezeigt werden. Dies umfasst die Ausbildung von Medizinstudenten und Assistenzärzten, sowie das maschinelle Lernen.

Wenn Sie noch weitere Fragen haben, nehmen Sie bitte Kontakt mit uns auf: [info@gestaltmatcher.org](mailto:info@gestaltmatcher.org)

Name:

First name / Vorname:

Date of birth / Geburtsdatum:

Gender / Geschlecht:

Place / Ort:  Date / Datum:

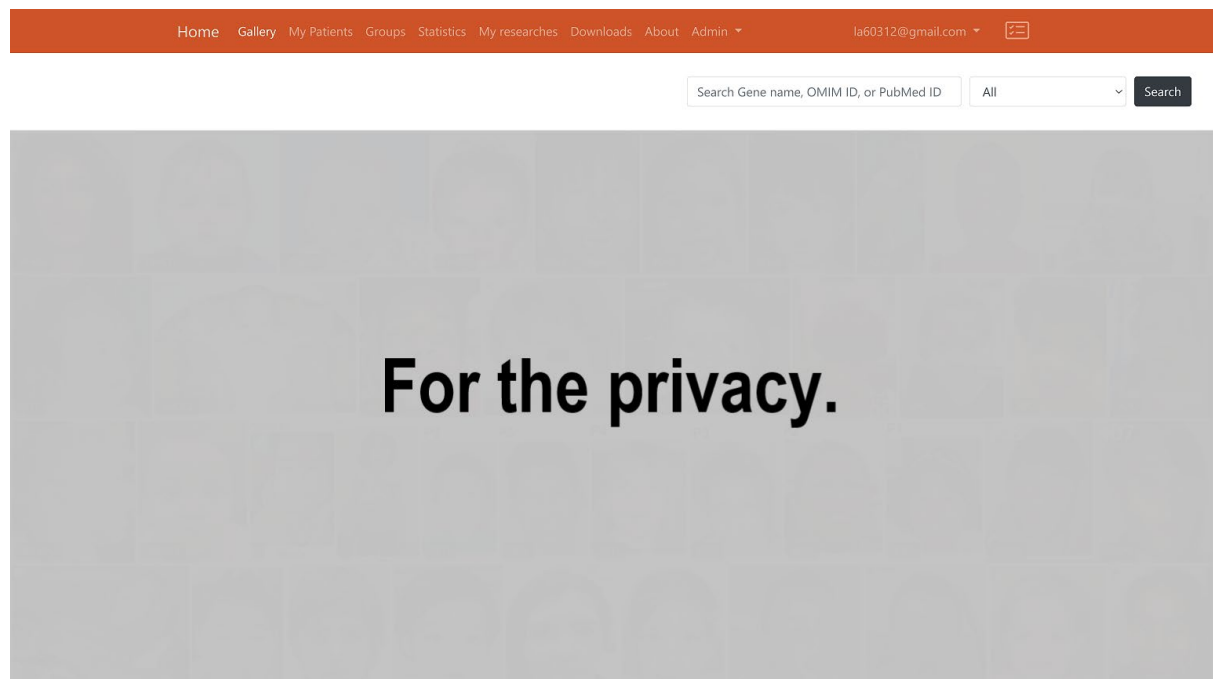
[Signature / Unterschrift](#)

[Submit consent](#)

Figure 24: Digital consent.

### 5.3.3 Visualize patients in gallery view

The first usage of this platform is the alternative and the up-to-date online resources to LMD. The gallery view of patient images in the platform provides the visualization of all the images of the particular query (gene, disorder, phenotypic series, or PMID). The clinician can see and learn the dysmorphism quickly by the gallery view (Figure 25).



**Figure 25: Gallery view in GMDB.** Patient images are blocked to protect privacy.

### 5.3.4 Training data for next-generation phenotyping

The second usage of this platform is that the patient data can be used as a resource for developing next-generation phenotyping. We exported each photo with the following information, frontal image, image\_id, patient\_id, and disorder\_id for the first version of the release. The disorder\_id can be treated as the label for training and testing. The current release can be found in the download section on the platform (Figure 26). The benchmarking of the GMDB dataset is shown in Table 5. It can be used to reproduce the GestaltMatcher publication results and develop the NGP approach with more advanced architecture.

# Downloads

## GestaltMatcher DB

- [GestaltMatcher GitHub repository](#) for cropping photos, training model, and encoding photos
- [GestaltMatcher analysis demo code](#)
- [GMDB photos](#)
- [GMDB-rare encodings](#)
- [GMDB metadata](#)

**Figure 26: Download section of GMDB dataset.**

**Table 5: Benchmarking on GMDB dataset.**

Test set	Model	Gallery	Test	Top-1	Top-5	Top-10	Top-30
GMDB-frequent	Softmax		360	25.13	47.23	59.79	77.26
GMDB-frequent	Enc-GMDB	3438	360	19.87	39.78	53.37	74.29
GMDB-frequent	Enc-healthy	3438	360	16.86	35.52	44.18	65.00
GMDB-rare	Enc-GMDB	369.2	138.8	15.13	34.76	46.15	68.51
GMDB-rare	Enc-healthy	369.2	138.8	12.18	29.03	40.36	61.46
GMDB-frequent	Enc-GMDB	3812	360	18.97	38.76	51.05	69.83
GMDB-frequent	Enc-healthy	3812	360	15.18	34.72	42.53	62.37
GMDB-rare	Enc-GMDB	3807.2	138.8	8.28	15.03	20.21	34.51
GMDB-rare	Enc-healthy	3807.2	138.8	6.61	13.33	16.60	28.13

## 5.4 Discussion

In GestaltMatcher Database (GMDB), we implemented the platform for users to upload patient data such as general information, publication, images, phenotypic and genomic data. This database can be seen as an up-to-date LMD that enables clinicians to check and learn

the dysmorphism of a disorder by the gallery view. Moreover, the images and patient data can be exported as the resource for developing the next-generation phenotyping approaches.

The size and the cleanliness of the dataset are the key factors to push computer vision research forward. However, like most of the research groups who would like to develop the next-generation phenotyping approaches for facial dysmorphism, we encountered tremendous difficulties in data curation at the beginning. In the PEDIA study (T. C. Hsieh et al. 2019), we collected the patient metadata (HPO terms, gene, and disorder) and frontal photos from publication by storing them in the Face2Gene platform. The user interface of Face2Gene provided the convenience of data curation. However, Face2Gene is a commercial platform that is not open to the research community. It is tough to reuse the data, and therefore, it is not beneficial to the entire research community.

In addition to the internal curators, our user interface enables the external users to contribute the patient data. That facilitates global collaboration.

Moreover, with images, HPO terms, and genomic data in GMDB, it can not only connect with many next-generation phenotyping approaches such as GestaltMatcher or PEDIA in the back-end but also connect with GeneMatcher (Sobreira et al. 2015), MatchMaker Exchange (Philippakis et al. 2015) or MyGene2 to strengthen the patient match network.

We believe this form of data curation can be a good example for the global collaboration of medical images dataset.

## Chapter 6 Prioritization of exome data by image analysis

### 6.1 Summary

In the previous chapters, we introduced NGP technology that helped reduce the search space to diagnose the patient by facial analysis. Although this kind of tool, such as Face2Gene, is already widely used by thousands of clinicians in their daily diagnostic workup, there are still several limitations in the diagnostic process. To diagnose the patient with a rare genetic disorder, we need to identify the disease-causing mutations. In the past, clinicians performed gene panels to analyze the variants among the selected genes. Face2Gene could suggest a shortlist of candidate disorders, but clinicians should convert the disorder to a gene by themselves. Moreover, in recent years, exome sequencing has been more commonly used in the diagnostic setting. Hence, the automatic workflow that combines facial analysis and variant analysis is vital for the current diagnostic process.

On the other hand, the existed variant prioritization tools only integrate the feature analysis that analyzes Human Phenotype Ontology (HPO) with variant analysis. However, in most cases, the “facial gestalt” presented in certain diseases that cannot be described by HPO terminology was only referred to as “characteristic.” Therefore, the approaches that could capture the facial gestalt were also needed for the current variant prioritization pipeline.

We then proposed this work, Prioritization of Exome Data by Image Analysis (PEDIA), to integrate the facial analysis, feature analysis, and variant analysis into an automatic variant prioritization pipeline. Sections 6.2 to 6.7 will present the work “PEDIA: prioritization of exome data by image analysis” published in *Genetics in Medicine* in 2019 (T. C. Hsieh et al. 2019). Section 6.8 will further introduce how to implement this approach to an existing variants analysis platform.



## 6.2 Abstract

**Purpose:**

Phenotype information is crucial for the interpretation of genomic variants. So far it has only been accessible for bioinformatics workflows after encoding into clinical terms by expert dysmorphologists.

**Methods:**

Here, we introduce an approach driven by artificial intelligence that uses portrait photographs for the interpretation of clinical exome data. We measured the value added by computer-assisted image analysis to the diagnostic yield on a cohort consisting of 679 individuals with 105 different monogenic disorders. For each case in the cohort we compiled frontal photos, clinical features, and the disease-causing variants, and simulated multiple exomes of different ethnic backgrounds.

**Results:**

The additional use of similarity scores from computer-assisted analysis of frontal photos improved the top 1 accuracy rate by more than 20–89% and the top 10 accuracy rate by more than 5–99% for the disease-causing gene.

**Conclusion:**

Image analysis by deep-learning algorithms can be used to quantify the phenotypic similarity (PP4 criterion of the American College of Medical Genetics and Genomics guidelines) and to advance the performance of bioinformatics pipelines for exome analysis.

## 6.3 Introduction

Worldwide, more than half a million children born per year have a rare genetic disorder that is suitable for diagnostic evaluation by exome sequencing. This test's unprecedented diagnostic yield is contrasted by the time requirement for variant interpretation. Making phenotypic information—the observable, clinical presentation—computer-readable is key to solving this problem and important for providing clinicians with a much-needed tool for diagnosing genetic syndromes (Biesecker and Green 2014).

To date, the most advanced exome prioritization algorithms combine deleteriousness scores for variants with semantic similarity searches of the clinical description of a patient (Pengelly et al. 2017). The Human Phenotype Ontology (HPO) has become the lingua franca for this purpose (Robinson et al. 2008). However, a facial gestalt for which no term exists and that is simply described as "characteristic" for a certain disease is not suitable for these computational approaches.

Beyond language, capturing indicative patterns through deep-learning approaches has recently gained attention in assessing facial dysmorphism (Ferry et al. 2014; Gurovich et al. 2019). Artificial neural networks measure the similarities of patient photos to hundreds of disease entities. We hypothesized that results of this next-generation phenotyping tool could be used similarly to deleteriousness scores on the molecular level. This would enable us to transition from the dichotomous PP4 criterion “matching phenotype” in the American College of Medical Genetics and Genomics (ACMG) guidelines for variant interpretation to a quantifiable one (Richards et al. 2015; Tavtigian et al. 2018).

We therefore developed an approach to interpret sequence variants integrating results from the next-generation phenotyping tool DeepGestalt. By this means the clinical presentation of an individual is not only assessed by a human expert clinician, but also by using an artificial intelligence approach on the basis of frontal photographs. In short, we call this approach prioritization of exome data by image analysis (PEDIA).

## 6.4 Materials and methods

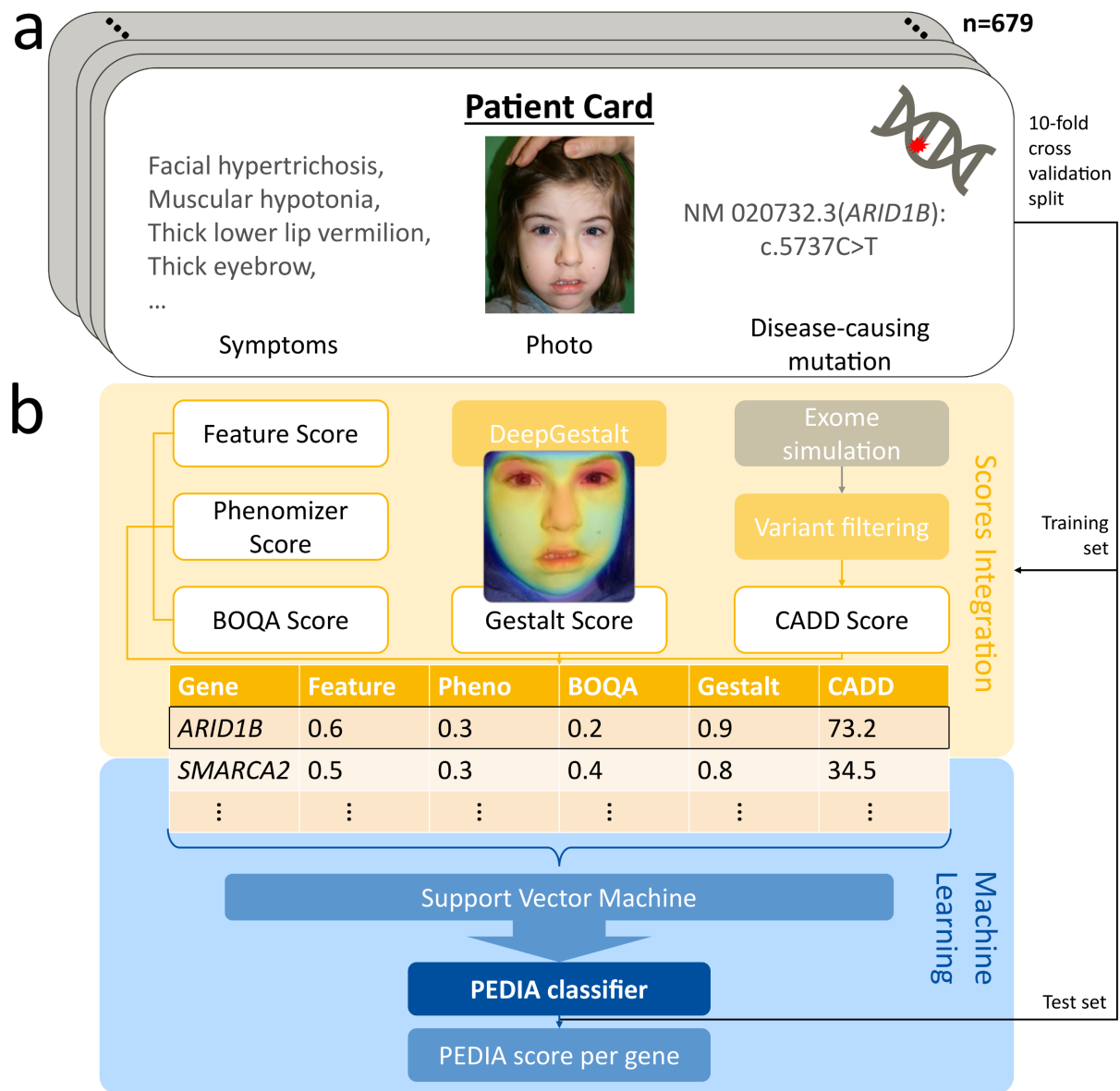
We compiled a cohort comprising 679 individuals with frontal facial photographs and clinical features documented in HPO terminology (Robinson et al. 2008). The diagnoses of all individuals have previously been confirmed molecularly and are suitable for analysis by exome sequencing. In total, the cohort covers 105 different monogenic syndromes linked to 181 different genes. Of the individuals in this cohort, 446 were published and 233 have not been previously reported (see PMID column in Supplementary Table 1 of Appendix A.2).

The study was approved by the ethics committees of the Charité–Universitätsmedizin Berlin and of the University Hospital Bonn. Written informed consent was given by the patients or

their guardians, including permission to publish photographs. Easy to understand, transparent information with both text and illustrations about the pattern recognition in our algorithm that processes personal data in the form of 2D portrait photographs can be found at <https://www.pedia-study.org/documents>. Through technical and organizational measures (privacy by design), we process the photos and the data obtained from them in the least identifiable manner necessary for achieving the purpose. This respects the data minimization principle of data being adequate, relevant, and limited.

In addition to the PEDIA data set, we analyzed a subset of the DeepGestalt study. By removing disorders that are confirmed by tests other than exome sequencing, such as Down syndrome (Supplementary Table 2 of Appendix A.2), we ended up with 260 of 329 cases from the DeepGestalt set (Gurovich et al. 2019).

The facial images were analyzed with DeepGestalt, a deep convolutional neural network trained on more than 17,000 patient images (Gurovich et al. 2019). The results of this analysis are gestalt scores that quantify the similarity to 216 different rare phenotypes per individual. These vectors can also be used to identify duplicates in the DeepGestalt training set and test set without the need to access the original photos. To avoid overfitting, we excluded all cases of the PEDIA cohort from a DeepGestalt model that we used for benchmarking. It is noteworthy that the version of DeepGestalt available at Face2Gene will not yield the same results when photos of the PEDIA cohort are reanalyzed because it is built as a framework that aims to learn from every solved case.



**Figure 27: Prioritization of exome data by image analysis (PEDIA): cohort and classification approach.** (a) Clinical features, facial photograph, and pathogenic variant of one individual of the PEDIA cohort. In total the cohort consists of 679 cases with monogenic disorders that are suitable for a diagnostic workup by exome sequencing. (b) Clinical features, images, and exome variants were evaluated separately and integrated to a single score by a machine learning approach. The disease-causing gene is shown at the top of the list.

In addition to the image analysis, we performed semantic similarity searches with the annotated HPO terms by three different tools: Feature Match (FDNA), Phenomizer, and Bayesian Ontology Querying for Accurate Comparisons (BOQA) (Köhler et al. 2009; Bauer et al. 2012). HPO terms for all published cases as well as the clinical notes in the electronic health records were independently extracted by two data curators. All terms that did not occur in both lists were revisited by a third curator (see Figure 27a and Supplementary Table 1 of Appendix A.2). The similarity scores from image analysis as well as semantic similarity searches were mapped to genes by *mim2gene* and *morbidmap* from (McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University 2019). If there were several syndromes linked to a gene, the highest gestalt and feature scores were selected for this gene.

Exome sequencing data was not available for the vast majority of cases. Therefore, we spiked in the disease-causing variant of each case into randomly selected exomes of healthy individuals of different ethnicities from the 1000 Genomes Project (1000 Genomes Project Consortium et al. 2015). All sequence variants were then filtered as described by Wright et al. and scored for deleteriousness with CADD (Kircher et al. 2014; Wright et al. 2018). Per gene, the variant with the highest CADD score was used, regardless of the genotype. This heuristic was chosen to maximize the sensitivity also for compound heterozygous cases where the second hit in a recessive disease gene achieves only a relatively low CADD score.

For each case this procedure resulted in a table with rows for genes and the five different scores in the columns (Figure 27b). All five scores per line as well as the Boolean label disease gene “true” or “false” (i.e., the vector) were used to train a classifier that yields a single value per gene, the PEDIA score, that can be used for prioritization (Figure 27b). A detailed description of preprocessing and filtering, as well as all the annotated data, can be found in our code repository.

We used a support vector machine (SVM) to prioritize the genes based on the five scores for each case. To benchmark our approach, we performed tenfold cross-validation. First, we split the PEDIA cohort into ten groups, ensuring that a certain disease gene was included only in one of ten groups. By this means, we avoided overfitting, in case the same disease-causing variant occurred in two different individuals (Supplementary Fig. 1 of Appendix A.2). We used a linear kernel on the five scores to train the SVM and selected the

hyperparameter  $C$  in the range from  $2^{-6}$  to  $2^{12}$  by performing internal fivefold cross-validation on the training set. The  $C$  with the highest top 1 accuracy was selected for training a linear SVM. We further benchmarked the performance of each case in the test set with this model. The distance of each gene to the hyperplane—defined as the PEDIA score—was used to rank the genes for the case. If the disease-causing gene was at the first position, we called it a top 1 match, or if it was among the first ten genes, we considered it a top 10 match.

For the 260 cases from the DeepGestalt publication test set, where exome diagnostics would be applicable, we randomly selected cases from the PEDIA cohort with the same diagnosis and added the CADD and the feature scores per case (see column C in Supplemental Table 1 of Appendix A.2). The cases in the PEDIA cohort with the same pathogenic variant as already assigned to the DeepGestalt test set were removed from the training set. Then we trained the classifier on the PEDIA cohort and tested it on the DeepGestalt publication test set. The experiment was repeated ten times with random selection. By this means we studied how the publicly available portraits of the DeepGestalt test set would improve the performance when used in exome analysis with the PEDIA approach. However, it has to be emphasized that both approaches solve different multiclass classification problems (MCPs), the first tool operating on phenotypes and the second on genes. The difficulty of the task is not only characterized by the number of classes and the distinguishability of the different entities but also by the information available for the classification. For both MCPs the maximum number of classes can be estimated from OMIM by querying with the HPO term “abnormal facial shape”, yielding around 700 disorders and genes with disease-causing variants. As there is additional and nonredundant information available from the molecular level for PEDIA, it achieves better top 1 and top 10 accuracies.

## 6.5 Results

The performance of a prioritization tool can be assessed by the proportion of cases for which the correct diagnosis or disease gene is placed at the first position or among the first ten suggestions (top 1 and top 10 accuracy). The composition of the test set has an influence on the accuracy because some disease phenotypes are easier to recognize, and some gene variants are more readily identified as deleterious. The setup of the PEDIA cohort, which is

comprehensively documented in the Supplementary Appendix A.2, therefore aims at emulating the whole spectrum of cases that could be analyzed with DeepGestalt and diagnosed by exome sequencing.

When only CADD scores are used for variant ranking, the disease-causing gene is in the top 10 in less than 45% of all tested cases. The top 10 accuracy increases up to 63–94%, when different semantic similarity scores based on HPO feature annotations are included (Supplementary Table 3 of Appendix A.2).

The additional information from frontal photos of cases pushes the correct disease gene to the top 10 in 99% of all PEDIA cases (Figure 28a). Particularly striking is the performance gain for the top 1 accuracy rate from 36–74% without DeepGestalt scores to 86–89% including the scores from image analysis (Supplementary Table 3 of Appendix A.2).

The distribution of the PEDIA scores does not differ using exomes with different ethnic backgrounds (Supplementary Fig. 2 of Appendix A.2).

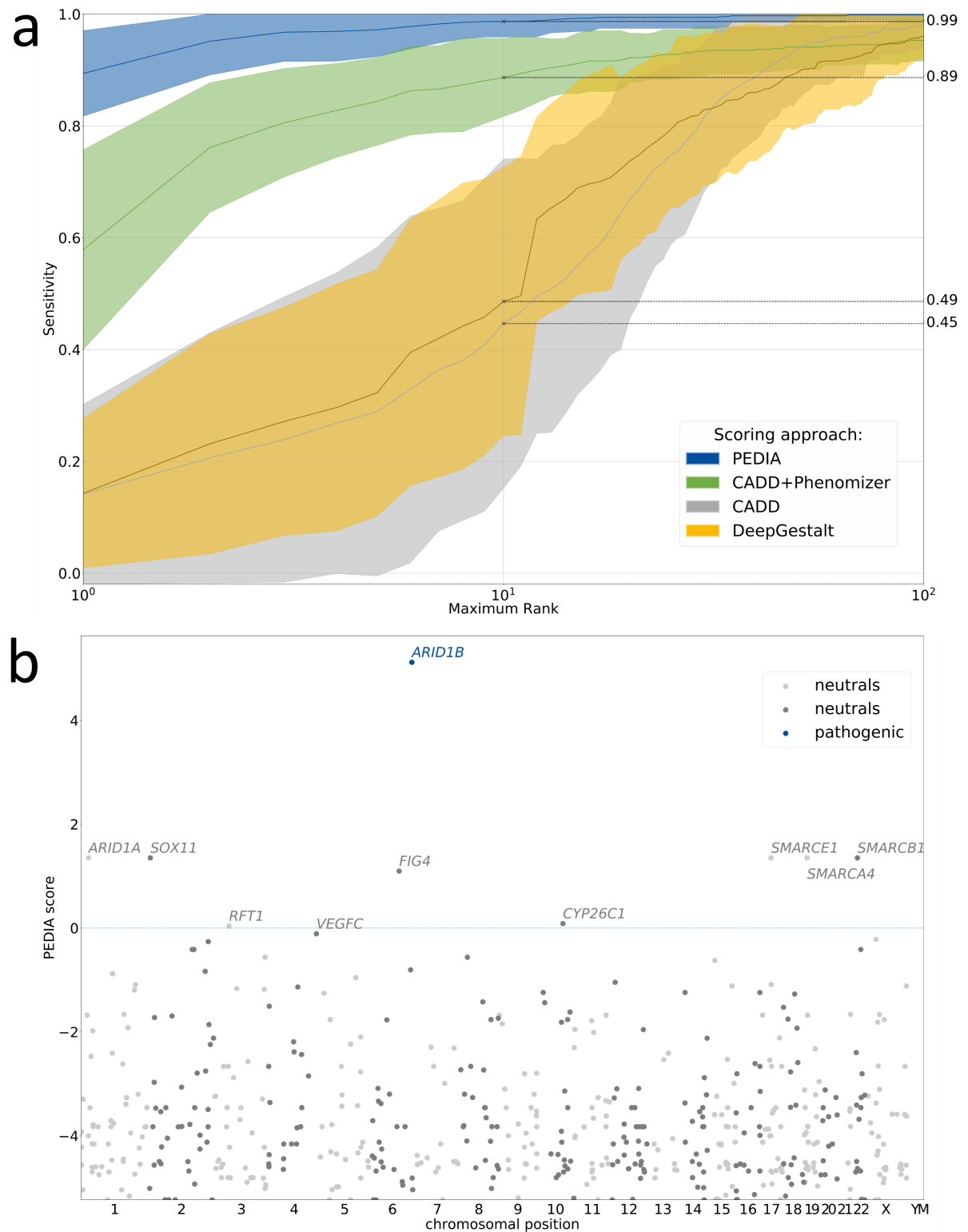
Although the top 10 accuracies of DeepGestalt scoring on the phenotype level and PEDIA scoring on the gene level cannot be compared directly, both approaches operate on a similar number of classes (Figure 28). Adding suitable molecular information to 260 cases from the DeepGestalt publication test set confirms our results in the PEDIA cohort by achieving a top 10 accuracy rate of 99% (Supplementary Table 2 of Appendix A.2).

The value of a frontal photograph is demonstrated by a case with Coffin–Siris syndrome (shown in Figure 27): the characteristic facial features are relatively mild, so the correct diagnosis is only listed as the third suggestion by DeepGestalt. Among all the variants encountered in the exome, the disease-causing gene *ARID1B* would only achieve rank 27, if scored by the molecular information alone. However, combined with the phenotypic information, the PEDIA approach lists this gene as the first candidate (Figure 28b).

Although the diagnosis of the illustrated case could be molecularly confirmed by a directed single-gene test in other instances where the facial gestalt is more indicative, syndromic disorders often puzzle clinicians due to their high phenotypic variability. In the Deciphering Developmental Disorders (DDD) project many syndromes were diagnosed only after exome

sequencing (Deciphering Developmental Disorders Study 2017). Still, the top 10 accuracy rate of 49% that DeepGestalt can achieve for phenotypes linked to genes is impressive (Figure 28a). The contribution from the different sources of evidence to the PEDIA score is also reflected by the relative weight of the deleteriousness of the pathogenic variant (0.44), all feature-based scores combined (0.25), and the results from image analysis by DeepGestalt (0.31) that can be derived from a linear SVM model. The information contained in a frontal photograph of a patient therefore goes beyond what clinical terms can capture. The top 1 and top 10 accuracies are reported for all combinations of scores in the Supplementary Table 3 of Appendix A.2.





**Figure 28: Performance readout and visualization of test results for a representative prioritization of exome data by image analysis (PEDIA) case.** (a) For each case the exome variants are ordered according to four different scoring approaches, solely by a

molecular deleteriousness score (CADD), by a score from image analysis (DeepGestalt), by a combination of a molecular deleteriousness score and a clinical feature–based semantic similarity score (CADD+Phenomizer), or the PEDIA score that includes all three levels of evidence. The sensitivity of the prioritization approach depends on the number of genes that are considered in an ordered list. The top 1 and top 10 accuracy rates correspond to the intersection of the curves at maximum rank 1 and 10. Note that for benchmarking DeepGestalt on the gene level, syndrome similarity scores first have to be mapped to the gene level, resulting in a lower performance compared with the readout on a phenotype level, due to heterogeneity. The area under the curve is largest for PEDIA scoring. (b) The disease-causing gene of the case depicted in Figure 27 achieves the highest PEDIA score and molecularly confirms the diagnosis of Coffin–Siris syndrome. Other genes associated with similar phenotypes, such as Nicolaides–Baraitser syndrome, also achieved high scores for gestalt but not for variant deleteriousness.

## 6.6 Discussion

The guidelines for variant classification in the laboratory follow a qualitative heuristic that combines distinct types of evidence (functional, population, phenotype, etc.). Interestingly, it is also compatible with Bayesian statistics (Tavtigian et al. 2018) and the advantage of such a framework is that continuous evidence types can be integrated into the classification system. While *in silico* predictions about a variant’s pathogenicity have a relatively long history in bioinformatics and machine learning, the quantification of phenotypic raw data such as facial images with artificial intelligence systems has just begun: the PEDIA approach uses scores from DeepGestalt for gene prioritization in combination with quantitative scores from the molecular level in Mendelian disorders identifiable by exome sequencing.

Interestingly, the ethnicity, which affects the number of variant calls or the deleterious variant load, had minor influence on the performance of PEDIA. Although the total number of variants detected by reference-guided sequencing in individuals of African descent is considerably higher than in individuals of European or Asian descent, the distribution of the CADD scores for rare variants is comparable (Supplementary Figs. 3, 4 of Appendix A.2). That means the rank that a gene achieves due to the molecular score and the corresponding scores from the phenotypic information is hardly affected by the background population (Supplementary Fig. 2 of Appendix A.2).

With regard to the routine use in the laboratory we have learned three important lessons from specific subgroups or cases achieving lower PEDIA ranks:

1. Although DeepGestalt, the convolutional neural network used for image analysis, has been pretrained on real-world uncontrolled 2D images, patient photographs that were not frontal, of low resolution, had poor lightening and contrast, or contained artifacts such as glasses, yielded lower gestalt scores for the searched disorder. In one use case envisioned for PEDIA, the human expert in the lab will only receive the similarity scores from DeepGestalt, but not the original photograph. In this setting it is not clear whether low scores originate from a low-quality photograph or whether there is little dysmorphic signal indicative of a syndromic disorder. This potential problem could be addressed by providing gestalt scores from additional photographs.
2. Particularly rare diseases or recently described disorders, for which the classifier's representation is based on a smaller training set, show a lower performance, even if experienced dysmorphologists would consider them highly distinguishable. In a recent publication by Duddin-Byth et al. the machine learning approach showed the lowest accuracy for the disorder with the smallest number of training cases; however, so did humans (Dudding-Byth et al. 2017).
3. Disease-causing variants in genes that interact in a molecular pathway often result in highly similar phenotypes that are organized as series in OMIM and modeled as a single entity by DeepGestalt. Often there are subtle gene-specific differences in the gestalt and modeling the entire phenotypic series by a single class is not the theoretical optimum achievable with more cases (Jean T. Pantel et al. 2018; Knaus et al. 2018). This will especially diminish the performance of genes less frequently mutated in a molecular pathway. This is exemplified in the PEDIA cohort by Hyperphosphatasia with Mental Retardation Syndrome (HPMRS), where the least frequently mutated gene, PGAP2, shows the lowest performance. Likewise, this applies to microdeletion syndromes that can also be caused by pathogenic variants in single genes, such as Smith–Magenis syndrome, or an atypical clinical presentation with Kabuki syndrome (see e.g., case IDs 246245 and 204233 in Supplementary Table 1 of Appendix A.2) (Badalato et al. 2017).

It is noteworthy that these shortcomings are mainly due to the limited training data for these particular genes and that they will most likely be overcome by more molecularly confirmed cases. DeepGestalt and PEDIA are therefore built as frameworks that will be improved continuously with additional data. In general, the use of artificial intelligence in medical sciences raises new or exacerbates existing ethical and legal issues as repositories of combined genotype and phenotype data become crucial for the machine learning community (Hallowell, Parker, and Nellåker 2019; Mascalzoni et al. 2019). Sharing portrait photos of individuals with rare diseases can be accomplished within the scope of even the most elaborate data privacy laws, such as the European Union General Data Protection Regulation 2016/679 (GDPR). The GDPR not only ensures the protection of individuals, but also the free movement of personal data, *inter alia*, for scientific research purposes (Bentzen and Høstmælingen 2019).

The interpretation of genetic variants is greatly facilitated by sequencing additional family members. Analogously, we hypothesize that the signal-to-noise ratio of next-generation phenotyping technologies can further be improved by including unaffected siblings or parents in the analysis.

We include and strive to include a wide variety of ethnicities, but European backgrounds are currently best represented, leading to best performance for this population. As the data set expands further, the algorithm will improve for currently underrepresented ethnicities.

Assistance with diagnosis of rare genetic disorders is highly valuable to clinicians, and by extension to the patients themselves and their families. Especially in inconclusive cases with findings of unknown clinical significance, additional evidence from computer-assisted analysis of medical imaging data could be a decisive factor (Wright et al. 2018).

In conclusion, the PEDIA study documents that exome variant interpretation benefits from computer-assisted image analysis of facial photographs. By including similarity scores from DeepGestalt, we improved the top 10 accuracy rate significantly compared with state-of-the-art algorithms. Artificial intelligence-driven pattern recognition of frontal facial patient photographs is therefore an example of next-generation phenotyping technology that has

proven its clinical value for the interpretation of next-generation sequencing data (Hennekam and Biesecker 2012).

## 6.7 Code availability

All training data as well as the classifier are available at <https://github.com/PEDIA-Charite/PEDIA-workflow>. The trained PEDIA model is provided as a service that is ready to use at <https://pedia-study.org>.

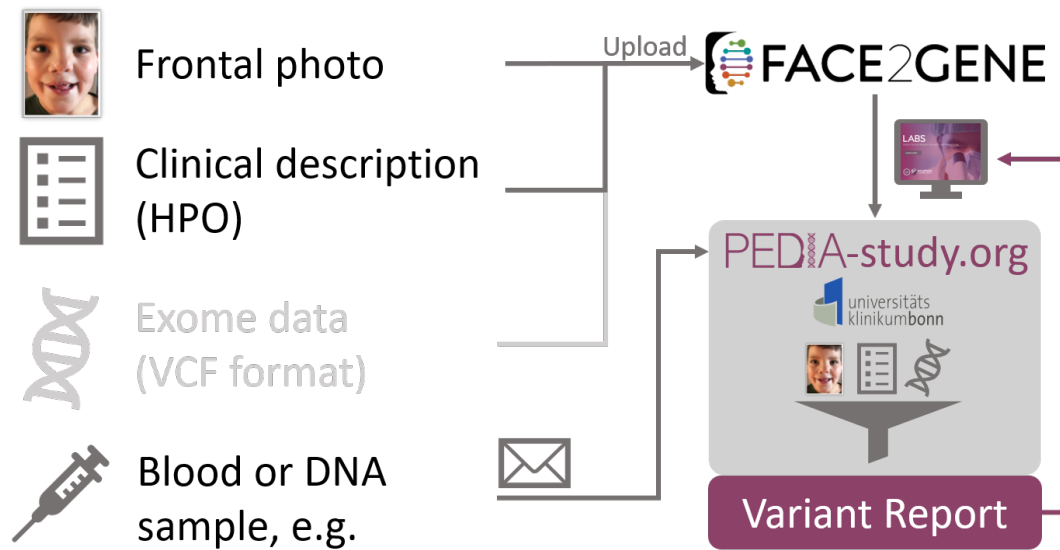
## 6.8 PEDIA in variants analysis platform

In the previous section, we introduced how PEDIA improved the exome diagnosis. As the first method, which integrates facial analysis into exome data analysis, we also implemented the PEDIA platform, a web platform for clinicians, and REST API for developers to integrate into their existing variants analysis platform. This section will introduce the PEDIA platform and demonstrate integrating the PEDIA workflow into GeneTalk (<https://www.gene-talk.de>), a variant analysis platform.

### 6.8.1 PEDIA platform

PEDIA workflow takes five scores as input, and Face2Gene generates the gestalt and feature match scores. Although we could use Phenomizer and Boqa for the feature analysis, there is no local version of DeepGestalt for facial analysis. So we need to send the photo to Face2Gene and obtain the gestalt scores. Therefore, we developed the PEDIA platform ([www.pedia-study.org](http://www.pedia-study.org)) to enable users to analyze their patients' exome data with the PEDIA scores. The PEDIA platform is an online platform that provides a VCF viewer for reviewing variants, PEDIA scores, and other annotations from external databases.

We developed the PEDIA API to receive data from Face2Gene and perform the PEDIA service. As an example of integrating the PEDIA service into the variants analysis platform, we launched the PEDIA service in the PEDIA platform and integrated it with Face2Gene LAB. The whole workflow is shown in Figure 29. The user first uploads the patient's HPO



**Figure 29: The flowchart of PEDIA platform.**

terms and facial photo to Face2Gene, and later uploads a VCF file of exome data or sends the sample to the University hospital of Bonn (UKB). Once the exome data is finished, UKB will upload the VCF file. In the end, the user submits the data to the PEDIA platform. Once the PEDIA analysis is finished, the user can visualize the PEDIA results on the PEDIA platform (Figure 30 and Figure 31).

In addition to visualizing PEDIA scores in a Manhattan plot and a list of genes, users can review the variants with PEDIA scores and CADD scores in a VCF viewer (Figure 32). The user can significantly narrow the search space by sorting PEDIA scores and clinical significance in the VCF viewer. After pressing the review button of a variant, the user can review the variant with more detail from external annotation databases such as ClinVar, Ensembl, ExAC, gnomAD, and Mutation Taster. Once the disease-causing mutation is found, the user can select the variant classification on the variant's page (Figure 33). The classification will be stored in the database and later submitted to ClinVar. These steps are the classical way to diagnose a patient with exome data, and the PEDIA platform is an example of how variants analysis platform integrates with PEDIA analysis.

PEDIA result

VCF Viewer

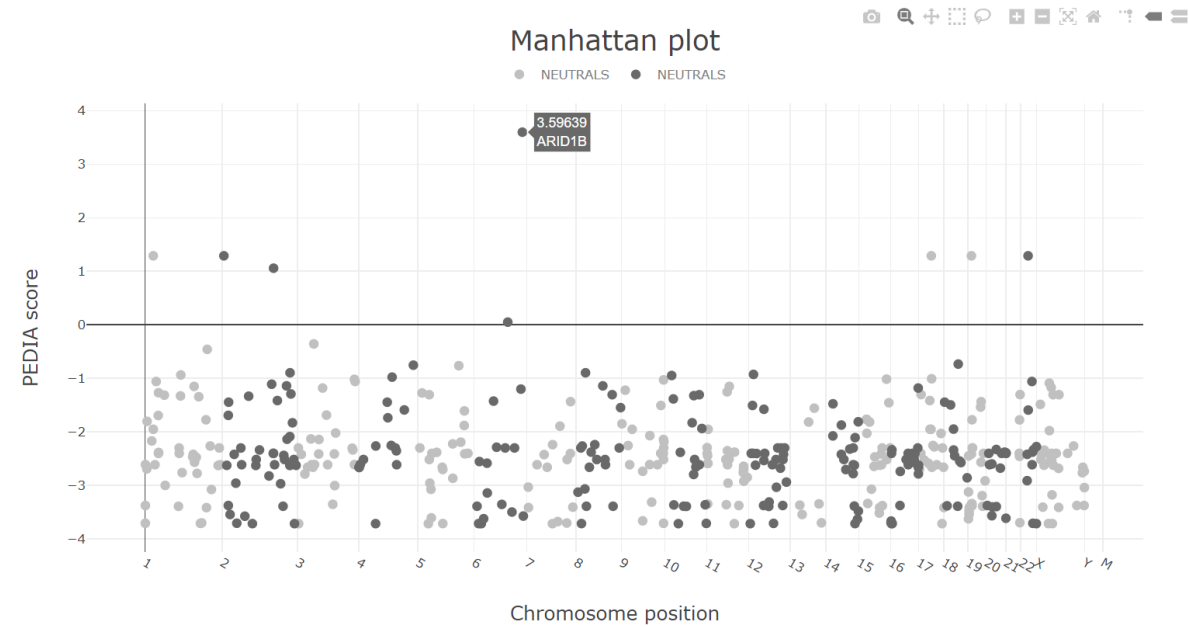


Figure 30: PEDIA Manhattan plot in PEDIA platform.

Rank	Gene	Gene ID	PEDIA Score	Gestalt Score	CADD Score	Feature Score	Phenomizer Score	Boqa Score
1	ARID1B	57492	3.59639	0.808158	22.7	0.687555	0.9981	0.0
2	SMARCA4	6597	1.28693	0.808158	0.0	0.687555	0.9981	0.0
3	ARID1A	8289	1.28693	0.808158	0.0	0.687555	0.9981	0.0
4	SMARCB1	6598	1.28693	0.808158	0.0	0.687555	0.9981	0.0
5	SOX11	6664	1.28693	0.808158	0.0	0.687555	0.9981	0.0
6	SMARCE1	6605	1.28693	0.808158	0.0	0.687555	0.9981	0.0
7	SCN9A	6335	1.05534	0.239703	34.0	0.443294	0.0	0.0
8	FIG4	9896	0.0479031	0.0	37.0	0.0	0.0	0.0
9	RFT1	91869	-0.35905	0.0	33.0	0.0	0.0	0.0
10	CACNA1S	779	-0.460788	0.0	32.0	0.0	0.0	0.0
11	LOC101928	155336	-0.735133	0.0	28.0	0.0	0.0	0.0

Figure 31: List of genes with PEDIA scores in the PEDIA platform.

VCF Viewer

File: 542847\_annotated.vcf.gz

Variants: 2740

Load all variants

Chrom. Pos	ID	Gene	Ref	Genotype	PEDIA	CADD	Effect	HGVS	Significance
5:131728202	rs11568514	SLC22A5	T	T/G	-0.766	29	missense_variant	NM_003060.3:c.1...	★★★★★
2:167145152	rs200391162	SCN9A	G	G/A	1.055	34	missense_variant	NM_002977.3:c.1...	★★★★☆
2:167099130	rs73019664	SCN9A	A	G/A	1.055	16.68	missense_variant	NM_002977.3:c.3...	★★★★☆
1:201010661	rs201310235	CACNA1S	C	C/T	-0.461	14.85	missense_variant	NM_000069.2:c.5...	★★★★☆
1:116269620	rs142036299	CASQ2	G	A/G	-0.939	27.3	missense_variant	NM_001232.3:c.7...	★★★★☆
1:116269619	rs28730716	CASQ2	T	C/T	-0.939	25.3	missense_variant	NM_001232.3:c.7...	★★★★☆
1:116311366	rs577401188	CASQ2	GCACA...	GCACACA/G	-0.939	6.11	5_prime_UTR_exo...	NM_001232.3:c.-2...	★★★★☆
1:116311183	rs12067472	CASQ2	C	C/T	-0.939	1.04	5_prime_UTR_exo...	NM_001232.3:c.-2...	★★★★☆
9:139391608	rs376422513	NOTCH1	C	C/T	-1.031	13.6	missense_variant	NM_017617.3:c.6...	★★★★☆
22:36688078	rs530533580	MYH9	C	C/T	-1.061	26.1	missense_variant	NM_002473.4:c.4...	★★★★☆
8:87638297	rs115246141	CNGB3	A	A/T	-1.142	19.44	missense_variant	NM_019098.4:c.1...	★★★★☆
5:13737585	rs529225111	DNAH5	G	T/G	-1.275	24	missense_variant	NM_001369.2:c.1...	★★★★☆
2:85884913	rs113824447	SFTPB	C	T/C	-1.336	2.11	3_prime_UTR_exo...	NM_000542.3:c.*1...	★★★★☆
1:173795885	rs114714497	DARS2	G	A/G	-1.346	23.3	missense_variant	NM_018122.4:c.1...	★★★★☆
2:179419216	rs115070904	TTN	G	G/A	-1.417	15.99	synonymous_vari...	NM_001256850.1:...	★★★★☆
2:179416989	rs114026724	TTN	A	A/G	-1.417	7.46	missense_variant	NM_001256850.1:...	★★★★☆
2:179614082	rs140064945	TTN	G	T/G	-1.417	6.44	missense_variant	NM_133379.4:c.1...	★★★★☆
6:64430897	rs61754905	FVS	T	C/C	-1.427	15.4	synonymous_vari...	NM_001142800.1:...	★★★★☆

Figure 32: Variants sorted by clinical significance in VCF viewer.

Mutation Review

Position: chr5:131728202

Ref: T

Genotype: T/G

Gene: SLC22A5

OMIM: SLC22A5

CADD score: 29

Classification: ★★★★★

- Annotations
- dbSNP
- Ensembl
- ExAC
- gnomAD
- Mutation Taster

benign

benign  
likely benign  
uncertain significance  
likely pathogenic  
pathogenic

Submit

Annotations:

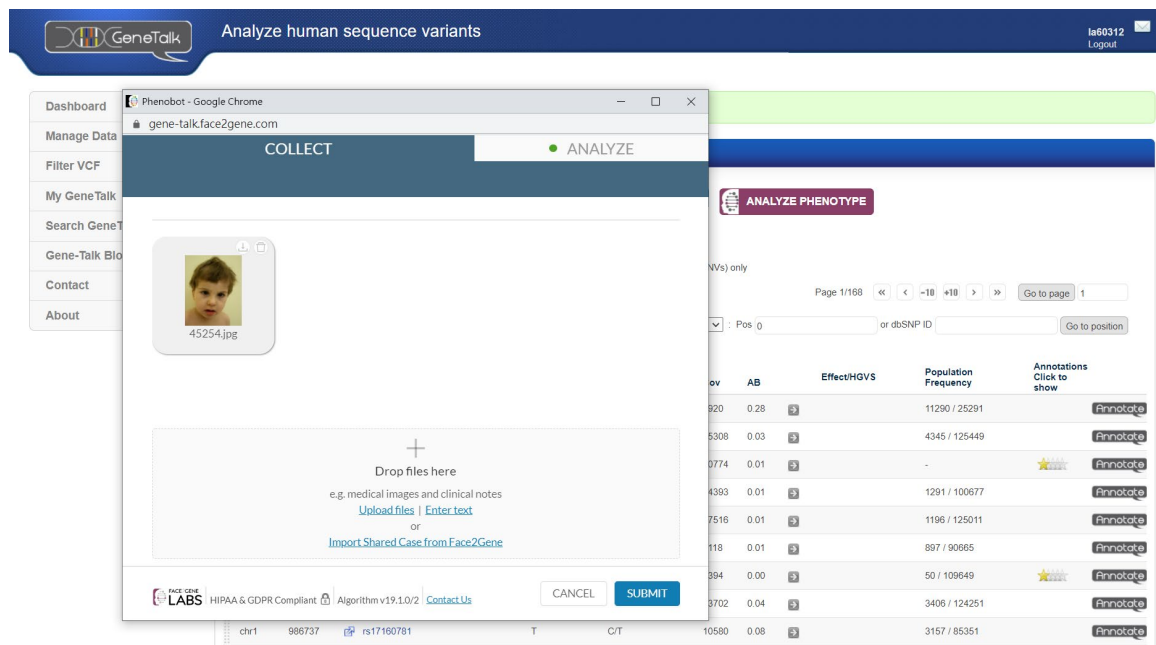
ID	User	Clinical Significance	Review Status	Submission	Link
513588	Clinvar	not provided	no assertion provided	SCV000043061.2	<a href="#">Open in ClinVar</a>
513589	Clinvar	Uncertain significance	criteria provided, single submitter	SCV000111944.7	<a href="#">Open in ClinVar</a>
513590	Clinvar	Pathogenic	criteria provided, single submitter	SCV000239169.10	<a href="#">Open in ClinVar</a>
513591	Clinvar	Uncertain significance	criteria provided, single submitter	SCV000452738.2	<a href="#">Open in ClinVar</a>

Figure 33: Variant annotation from external databases.



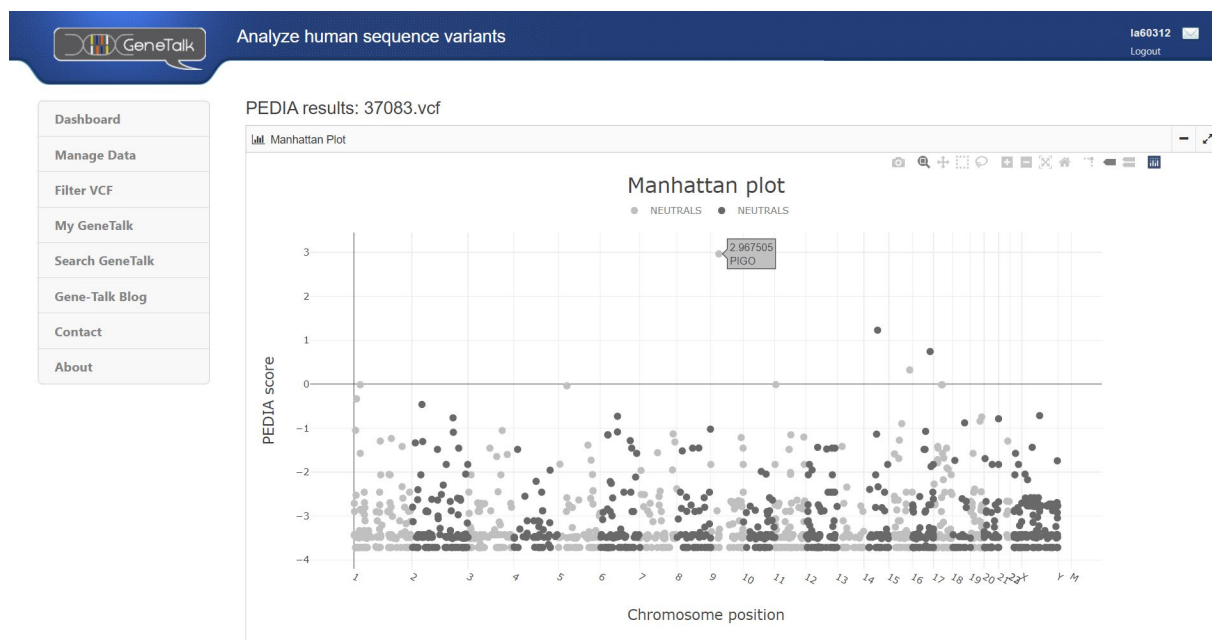
## 6.8.2 PEDIA in GeneTalk

The other online variants analysis platform can also use PEDIA API. Here we take GeneTalk (<https://www.gene-talk.de>) as an example. To obtain the gestalt score without uploading the patient to the Face2Gene website, Face2Gene developed PhenoBot, a widget that can run facial analysis on any platform. PhenoBot can be easily embedded as a browser plugin into any online platform. With PhenoBot, the user can upload the patient's photo to it, and the results will be shown in the widget after analysis. The online platform will obtain the results in JSON format to perform any other downstream analysis such as the PEDIA pipeline. The screenshot of PhenoBot is shown in Figure 34. PEDIA API and PhenoBot were integrated into GeneTalk (Kamphans and Krawitz 2012), demonstrating how a variants analysis platform obtains the gestalt scores and deploys PEDIA service. After submitting the photo and HPO terms to the PEDIA platform, GeneTalk will receive the PEDIA results that can be used to prioritize the variants in GeneTalk (Figure 35 and Figure 36).



**Figure 34: PhenoBot in GeneTalk.**

Using PEDIA API requires data transmitted to the server hosted in the University hospital of Bonn that might violate the data protection policy in some hospitals. To avoid the data transfer outside the hospital, we released the PEDIA workflow in a docker image (<https://hub.docker.com/repository/docker/la60312/pedia>). Therefore, the developer can deploy the PEDIA workflow locally in their server without sending data to the PEDIA platform.



**Figure 35: PEDIA Manhattan plot in GeneTalk.**

PEDIA Scores

Rank	Gene Name	Entrez ID	PEDIA Score	F2G Feature Score	F2G Gestalt Score	CADD Score	Boqa Score	Pheno Score	Class	VCF
1	<a href="#">PIGO</a>	84720	2.97	0.58	0.77	29.30			Neutral	
2	<a href="#">KIAA0586</a>	9786	1.23	0.60	0.30	26.40		0.54	Neutral	
3	<a href="#">TMEM231</a>	79583	0.74	0.60	0.30	21.90		0.52	Neutral	
4	<a href="#">KIF7</a>	374654	0.32	0.60	0.30	23.30			Neutral	
5	<a href="#">PGAP2</a>	27315	-0.01	0.58	0.77				Neutral	
6	<a href="#">PIGW</a>	284098	-0.01	0.58	0.77				Neutral	
7	<a href="#">PIGV</a>	55650	-0.01	0.58	0.77				Neutral	
8	<a href="#">PGAP3</a>	93210	-0.01	0.58	0.77				Neutral	
9	<a href="#">CPANE1</a>	65250	-0.04	0.60	0.30	13.96		0.54	Neutral	

Search in PEDIA results

Show/Hide column:

Rank

Gene Name

Entrez ID

PEDIA Score

F2G Feature Score

F2G Gestalt Score

CADD Score

Boqa Score

Pheno Score

Class

VCF

Figure 36: List of genes with PEDIA scores in GeneTalk.

## **Chapter 7 Discussion and the future of next-generation phenotyping**

The previous chapters introduced how to utilize the current next-generation phenotyping to analyze rare genetic disorders. Although we have seen that the NGP approaches can be used in the daily diagnosing workup and exploring the novel gene-phenotype association, further improvement to the existing approach is still needed. Because the network architecture used in the previous chapters is all based on the work published in 2014, updating the architectures to a more advanced one might improve the performance. Besides, the methods for aggregating the models trained by different facial crops or training splits were not well studied yet. The following sections will first introduce the possible improvements by updating the DeepGestalt method.

Moreover, most of the current databases are biased in ethnicity and age. The majority of the patients collected are from Caucasians, and the photos are taken at a younger age. The model trained on the biased dataset might influence prediction performance (Lumaka et al. 2017). Hence, the algorithm for removing bias while training on a biased dataset is an urgent need.

In addition to improving the algorithm, enlarging the currently available dataset is also an important issue. It is difficult to collect data because of the rareness of disorders and patients' concerns. Patients might be deterred due to privacy leaking or data abuse, not to mention we are collecting the frontal images that are easily recognizing the identity. To tackle this difficulty, many researchers have utilized generative adversarial networks (GANs) to synthesize medical images. Therefore, the synthesis of faces with facial dysmorphism by GANs could be one solution to enlarging the training dataset for the deep learning approach.

In the end, with the experience of COVID-19, global collaboration becomes more and more critical. It not only helps ease the pandemic, but will also contribute to the rare disorders. The most well-known global collaboration platform, MatchMaker Exchange (Philippakis et al. 2015), integrated several patient databases that enable clinicians to find similar patients

worldwide. However, this kind of platform only focuses on genomic data such as the disease-causing mutations, and the automatic matching by medical images is not yet integrated into the matching databases. In addition, looking for a way to share the de-identified medical images across different sites is also very important to tackle data rareness. I envision that these further improvements could strengthen the NGP approaches for diagnosing rare disorders.

## 7.1 Modernizing DeepGestalt approach

To make a fair comparison to DeepGestalt, the network architecture and the pre-trained dataset, CASIA-WebFace, used in this thesis are the same as the one introduced in the study presented in 2014 (Yi et al. 2014). Since then, lots of approaches have been proposed with different architectures and loss functions (Schroff, Kalenichenko, and Philbin 2015; W. Liu et al. 2017; H. Wang et al. 2018; Jiankang Deng et al. 2019; Jiankang Deng, Guo, Liu, et al. 2020; Jiankang Deng et al. 2021). In addition to the architecture, many face datasets were proposed after CASIA-WebFace, such as IMDB-WIKI, VGGFace2, FairFace (Rothe, Timofte, and Van Gool 2018; Cao et al. 2017; Kärkkäinen and Joo 2021). Therefore, it will be necessary to benchmark the performance with different pre-trained datasets and architectures combinations.

Moreover, the original DeepGestalt publication proposed a method that aggregates the results from different facial regions. For simplifying the experiment setting, only the whole face was taken into analysis in GestaltMatcher. The aggregation of the different facial regions was not yet well discussed. For example, we can concatenate the feature vectors derived from different facial regions and calculate the cosine distance to quantify the similarity. Besides, we found that the results sometimes differed among different models. So the same method can also be applied to the feature vectors generated from different models to stabilize the prediction.

In the end, some facial regions that DeepGestalt was not analyzing are sometimes crucial for making the diagnosis. For example, the patients with Waardenburg syndrome (MIM: 193500) present white forelock, and this feature is one of the hints for clinicians to make

Waardenburg syndrome the diagnosis. However, this kind of hair feature is not yet included in the DeepGestalt. Additionally, the profile image is sometimes also vital for clinicians to diagnose. Therefore, the facial cropper that can include more regions such as hair or profile would be helpful for future NGP approaches.

## 7.2 Bias removal

The next-generation phenotyping technology for syndromology, such as GestaltMatcher, has enabled matching patients with ultra-rare phenotypes by the facial representations learned from thousands of patient photos. However, the currently available patient photos are unbalanced in ethnicity and age. For example, most of the photos are from Caucasians and taken at an early age. It results in biased models when training on an unbalanced dataset. The model might learn the ethnicity instead of facial dysmorphic features to classify the disorders. Therefore, I will demonstrate how to remove these biases when training on an unbalanced dataset.

To prove the method can remove the ethnic bias while training on an unbalanced dataset, I first collected an unbalanced dataset on purpose. The dataset contained 167 images of Cornelia de Lange syndrome (CdLS) and 199 images of Williams-Beuren syndrome (WBS). In CdLS, 142 of 167 images were European, and the rest 25 images were Non-European. The ethnic distribution in WBS was the opposite. Only 25 images were European, and the rest 174 images were Non-European. The dataset was further divided into an unbalanced training set and a balanced test set. The training set contained 117 European of CdLS and 149 Non-European of WBS. The test set had a balanced distribution between European and Non-European. Both categories in CdLS and WBS had 25 images (Table 6). In this way, the training set was biased on ethnicity while the test set was balanced. I later utilized deep convolutional neural networks for training a model on images, and the joint learning and unlearning (JLU) technique proposed by Alvi et al. was used to remove the ethnicity bias during the model training (Alvi, Zisserman, and Nellåker 2019).

Figure 37 shows the results before and after applying JLU. The results before applying JLU was taken as the baseline because I want to prove that the results are improved after applying

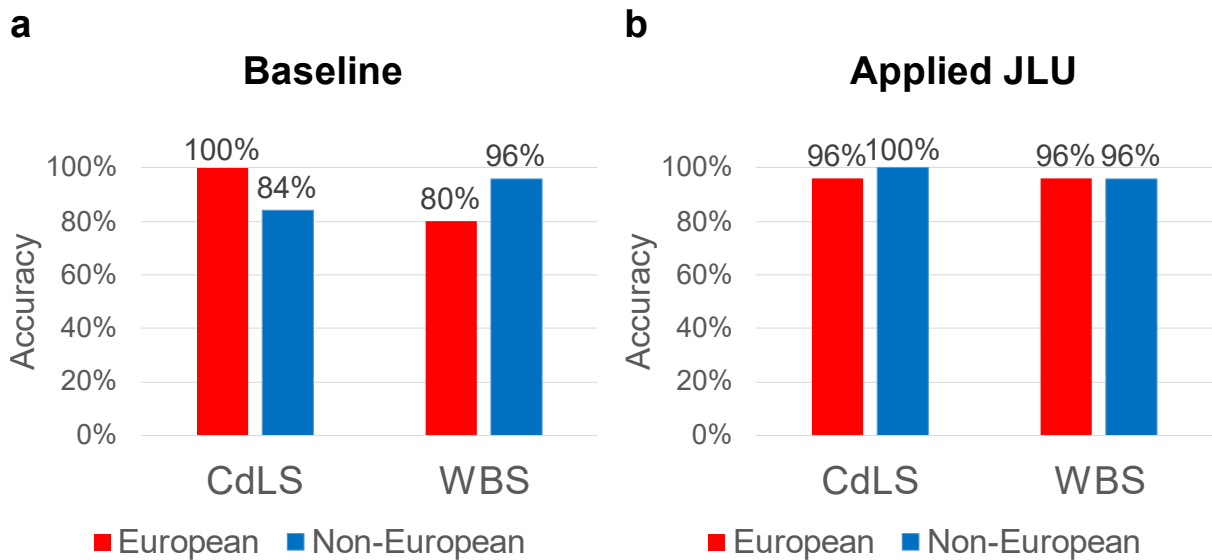
JLU. In CdLS, the accuracy of Non-European was 84% that is lower than European (100%). In WBS, the accuracy of European was 80% that is lower than Non-European (96%). We can see that the classes not included in the training performed worse than the classes in the training set. However, after applying JLU, the accuracies of European and Non-European were balanced. European CdLS had 96%, and Non-European CdLS had 100%. In WBS, both European and Non-European had 96%.

**Table 6: Ethnicity distribution in training and test set.**

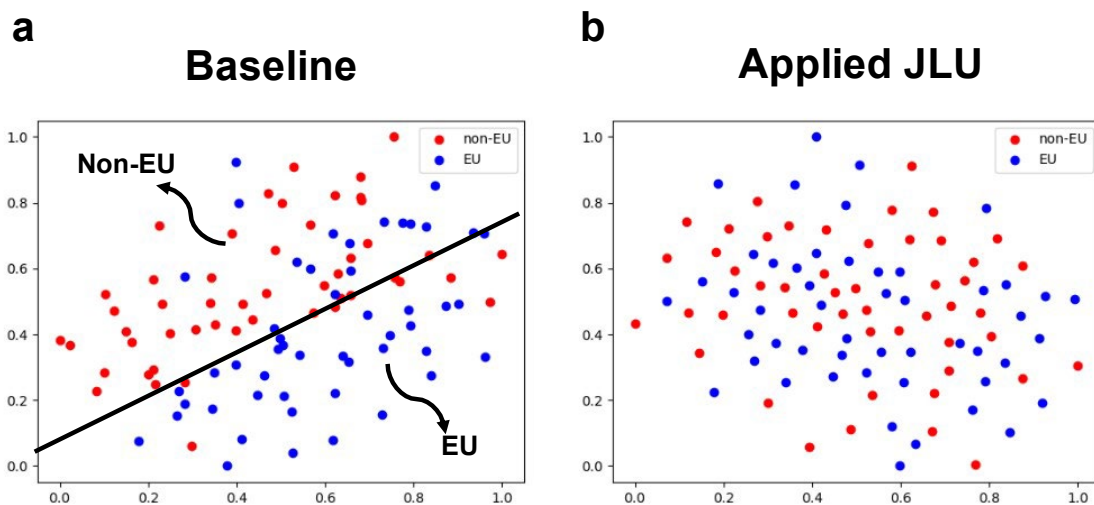
		European	Non-European
Training set	CdLS	117	0
	WBS	0	149
Test set	CdLS	25	25
	WBS	25	25

Moreover, *t*-SNE (van der Maaten and Hinton 2008) was used to visualize the distribution of patients in a two-dimensional space by projecting the features extracted from the layer before softmax. In Figure 38, we can see that the patients can be easily separated by ethnicity (European and Non-European). However, after applying JLU, we could no longer separate the patients by ethnicity.

The results proved that the adversarial networks such as JLU unlearned the bias and better generalized facial dysmorphic features. With this method, we could improve the disease classification on the patients of the minority class in this society. However, the results were only based on two disorders. More data with labels such as sex, age, and ethnicity is needed for further comprehensive analysis.



**Figure 37: Results of baseline and after applied JLU.** a) The baseline results. b) The results after applying JLU.



**Figure 38: *t*-SNE visualization of baseline and after applied JLU.** a) The baseline results. b) The results after applying JLU. Both figures contain the 100 patients in the test set (CdLS: 50 and WBS: 50).



## 7.3 Synthesizing faces with facial dysmorphism

Although we proposed GestaltMatcher DB and will dedicate collecting medical images, a large and balanced dataset is still challenging to acquire. As discussed in the previous section, the majority of data with rare disorders is from Caucasian patients and at a younger age. Despite the algorithm for bias removal, the lack of balanced data is a significant problem. Moreover, data sharing is a big concern to the patients. Patients are worried about privacy leaking and the abuse of their data, especially to the frontal images that easily recognize the actual identity. Therefore, synthesizing the face with a rare genetic disorder given by the chosen configuration (age and ethnicity) that cannot be traced back to the original patient is essential to tackle the issue of data rareness.

With the rapid development of generative adversarial networks (GANs), GANs have shown their capability to generate fake faces that are hard to distinguish from real faces and synthesize the images with the given style (Goodfellow et al. 2014; Mescheder, Nowozin, and Geiger 2017; Z. Zhang, Song, and Qi 2017; Karras et al. 2018, 2020). Beyond normal image generation, synthesizing medical images has become a hot topic in recent years. GANs have been utilized on the synthesis of retinal images (Costa et al. 2018; Zhao et al. 2018; Y.-C. Liu et al. 2019; Diaz-Pinto et al. 2019), skin lesions (Izadi et al. 2018; Bissoto et al. 2018; Ali, Mohamed, and Mahdy 2019; Bissoto, Valle, and Avila 2021), chest radiographs (Chuquicusma et al. 2018; Han et al. 2020, 2021; DuMont Schütte et al. 2021), and many other types of medical images. Hence, we envision that GANs can be utilized to synthesize faces with facial dysmorphism.

The frontal images from the GMDB can be taken as the training data. As a proof of concept, we will start with synthesizing the face with a chosen disorder such as Kabuki syndrome (MIM: 147920) or Cornelia de Lange syndrome (MIM: 122470). Moving forward, age, sex, and ethnicity will be considered. In the end, the model can generate the face with a specific disorder, age, sex, and ethnicity, and the dysmorphologists will validate the simulated faces to evaluate the performance.

The capability to simulate the characteristic facial gestalt of a genetic disorder is also important besides its value of augmenting and enlarging the training set. Facial portraits are

still the best teaching material for students and residents in medical genetics. However, since individuals are re-identifiable by their faces, many patients no longer consent to publish their photos. Therefore, this simulation approach is also of interest for the education of the next generation of physicians.

## 7.4 Enabling global collaboration

With the experience of COVID-19, global collaboration is more and more critical for public health (Jit et al. 2021; Moshtagh, Mirlashari, and Amiri 2021; Cai, Fry, and Wagner 2021; Vervoort, Ma, and Luc 2021). It is not only crucial for the pandemic, especially for the rare disorders, because there is sometimes only one patient with an ultra-rare disorder. For facilitating the diagnosis of rare disorders, many online platforms host the data collected worldwide, such as GeneMatcher, MatchMaker Exchange, MyGene2, and ClinVar (Sobreira et al. 2015; Philippakis et al. 2015; Chong et al. 2016; Landrum et al. 2018, 2020). However, most of these platforms focus on genomic data such as disease-causing mutations, and the phenotypic information is mainly collected in the form of HPO terminology. Although MyGene2 supports uploading photos, there is no automatic matching by facial photo analysis. The Minerva Initiative has been proposed to enable the collaboration on identifiable data such as facial images (Nellåker et al. 2019), but it is currently unavailable. Therefore, it is crucial to support patient matching by facial photo analysis and connect medical image databases such as GMDB to other databases.

To begin with, GMDB can first collaborate with the institutions in Germany to continue the development of a database for Next-Generation Phenotyping that includes medical images and the relevant molecular information so that it is compliant with the German and European regulations on general data protection. With the experience of the TRANSLATE-NAMSE project (<https://translate-namse.charite.de>), GMDB has established connections with ten German university hospitals (Charité, Bonn, Heidelberg, Munich, Tuebingen, Essen, Bochum, Luebeck, Dresden, and Hamburg). The TRANSLATE-NAMSE project has collected more than one thousand patients with exome sequencing data and facial photos in the last three years. Over the next few years, many innovative concepts from TRANSLATE-

---

NAMSE will also be used within genomDE and the Modellvorhaben Genomsequenzierung. In the end, I anticipate GMDB as a German-based rare disorders database with medical images and sequencing data that can also be a valuable contribution on the international level.

In addition to sharing the data directly, the decentralized learning method such as swarm learning could be considered as a solution to facilitate global collaboration (Warnat-Herresthal et al. 2021). Abiding the legal regulation is usually the most challenging task when collaborating among different institutes. This decentralized way that avoids transferring the medical data can train the model on the data in different sites locally and only transfer the model. The scientists in different hospitals would be more willing to join the network since it could enable collaboration and avoid the time-consuming paperwork to allow others to access their data.

In the end, the collaborations on both data and algorithm levels still have lots of room to improve. I hope GMDB could become a German-based rare disorder database and further connect with other patient platforms such as MatchMaker Exchange to enhance the diagnosis of rare disorders globally.

## Chapter 8 Conclusion

This thesis presented how GestaltMatcher, GMDB, and PEDIA tackle the current limitations on the algorithm, dataset, and availability for clinical diagnostic settings. We showed that GestaltMatcher could support ultra-rare disorders and novel diseases and analyze the patients' similarities to explore the novel gene-phenotype relationship. Moreover, the medical images in GMDB could be the open-access resource to the research community for deep learning purposes and the easy-visualized reference material for clinician-scientists. In addition, PEDIA integrated the facial analysis into the exome prioritization pipeline and can be easily integrated into the existing variants analysis platform. In the end, we envision that both GestaltMatcher and GMDB can be integrated into the patient match platforms to enable global collaboration and further improve the diagnosis of rare Mendelian disorders.

## Bibliography

- 1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74.
- Ali, Ibrahim Saad, Mamdouh Farouk Mohamed, and Yousef Bassyouni Mahdy. 2019. "Data Augmentation for Skin Lesion Using Self-Attention Based Progressive Generative Adversarial Network." *ArXiv [Eess.IV]*. arXiv. <http://arxiv.org/abs/1910.11960>.
- Alvi, Mohsan, Andrew Zisserman, and Christoffer Nellåker. 2019. "Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings." In *Computer Vision – ECCV 2018 Workshops*, 556–72. Springer International Publishing.
- Au, P. Y. Billie, Jing You, Oana Caluseriu, Jeremy Schwartzentruber, Jacek Majewski, Francois P. Bernier, Marcia Ferguson, et al. 2015. "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." *Human Mutation* 36 (10): 1009–14.
- Badalato, Lauren, Sali M. K. Farhan, Allison A. Dilliot, Care4Rare Canada Consortium, Dennis E. Bulman, Robert A. Hegele, and Sharan L. Goobie. 2017. "KMT2D p.Gln3575His Segregating in a Family with Autosomal Dominant Choanal Atresia Strengthens the Kabuki/CHARGE Connection." *American Journal of Medical Genetics. Part A* 173 (1): 183–89.
- Baird, P. A., T. W. Anderson, H. B. Newcombe, and R. B. Lowry. 1988. "Genetic Disorders in Children and Young Adults: A Population Study." *American Journal of Human Genetics* 42 (5): 677–93.
- Balak, Chris, Marianne Benard, Elise Schaefer, Sumaiya Iqbal, Keri Ramsey, Michèle Ernoul-Lange, Francesca Mattioli, et al. 2019. "Rare De Novo Missense Variants in RNA Helicase DDX6 Cause Intellectual Disability and Dysmorphic Features and Lead to P-Body Defects and RNA Dysregulation." *American Journal of Human Genetics* 105 (3): 509–25.

- Bauer, Sebastian, Sebastian Köhler, Marcel H. Schulz, and Peter N. Robinson. 2012. "Bayesian Ontology Querying for Accurate and Noise-Tolerant Semantic Searches." *Bioinformatics* 28 (19): 2502–8.
- Bentzen, Heidi Beate, and Njål Høstmælingen. 2019. "Balancing Protection and Free Movement of Personal Data: The New European Union General Data Protection Regulation." *Annals of Internal Medicine* 170 (5): 335–37.
- Biesecker, Leslie G., and Robert C. Green. 2014. "Diagnostic Clinical Genome and Exome Sequencing." *The New England Journal of Medicine* 370 (25): 2418–25.
- Bissoto, Alceu, Fábio Perez, Eduardo Valle, and Sandra Avila. 2018. "Skin Lesion Synthesis with Generative Adversarial Networks." In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, 294–302. Springer International Publishing.
- Bissoto, Alceu, Eduardo Valle, and Sandra Avila. 2021. "GAN-Based Data Augmentation and Anonymization for Skin-Lesion Analysis: A Critical Review." In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE. <https://doi.org/10.1109/cvprw53098.2021.00204>.
- Cai, X., C. V. Fry, and C. S. Wagner. 2021. "International Collaboration during the COVID-19 Crisis: Autumn 2020 Developments." *Scientometrics*, February, 1–10.
- Cao, Qiong, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. 2017. "VGGFace2: A Dataset for Recognising Faces across Pose and Age." *ArXiv [Cs.CV]*. arXiv. <http://arxiv.org/abs/1710.08092>.
- Cerrolaza, J. J., A. R. Porras, A. Mansoor, Q. Zhao, M. Summar, and M. G. Linguraru. 2016. "Identification of Dysmorphic Syndromes Using Landmark-Specific Local Texture Descriptors." In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 1080–83.
- Chong, Jessica X., Joon-Ho Yu, Peter Lorentzen, Karen M. Park, Seema M. Jamal, Holly K. Tabor, Anita Rauch, et al. 2016. "Gene Discovery for Mendelian Conditions via Social Networking: De Novo Variants in KDM1A Cause Developmental Delay and Distinctive Facial Features." *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 18 (8): 788–95.
- Chuquicusma, Maria J. M., Sarfaraz Hussein, Jeremy Burt, and Ulas Bagci. 2018. "How to Fool Radiologists with Generative Adversarial Networks? A Visual Turing Test for Lung Cancer Diagnosis." In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. <https://doi.org/10.1109/isbi.2018.8363564>.

- Cipriani, Valentina, Nikolas Pontikos, Gavin Arno, Panagiotis I. Sergouniotis, Eva Lenassi, Penpitcha Thawong, Daniel Danis, et al. 2020. "An Improved Phenotype-Driven Tool for Rare Mendelian Variant Prioritization: Benchmarking Exomiser on Real Patient Whole-Exome Data." *Genes* 11 (4). <https://doi.org/10.3390/genes11040460>.
- Costa, Pedro, Adrian Galdran, Maria Ines Meyer, Meindert Niemeijer, Michael Abramoff, Ana Maria Mendonca, and Aurelio Campilho. 2018. "End-to-End Adversarial Retinal Image Synthesis." *IEEE Transactions on Medical Imaging* 37 (3): 781–91.
- Deciphering Developmental Disorders Study. 2017. "Prevalence and Architecture of de Novo Mutations in Developmental Disorders." *Nature* 542 (7642): 433–38.
- Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. "ImageNet: A Large-Scale Hierarchical Image Database." In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–55.
- Deng, Jiankang, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. 2020. "Sub-Center ArcFace: Boosting Face Recognition by Large-Scale Noisy Web Faces." In *Computer Vision – ECCV 2020*, 741–57. Springer International Publishing.
- Deng, Jiankang, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. "RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild." In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5202–11.
- Deng, Jiankang, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition." In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:4685–94. IEEE Computer Society.
- Deng, Jiankang, Jia Guo, Jing Yang, Alexandros Lattas, and Stefanos Zafeiriou. 2021. "Variational Prototype Learning for Deep Face Recognition." In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11901–10.
- Diaz-Pinto, Andres, Adrian Colomer, Valery Naranjo, Sandra Morales, Yanwu Xu, and Alejandro F. Frangi. 2019. "Retinal Image Synthesis and Semi-Supervised Learning for Glaucoma Assessment." *IEEE Transactions on Medical Imaging* 38 (9): 2211–18.
- Diets, Illja J., Roos van der Donk, Kristina Baltrunaite, Esmé Waanders, Margot R. F. Reijnders, Alexander J. M. Dingemans, Rolph Pfundt, et al. 2019. "De Novo and Inherited Pathogenic Variants in KDM3B Cause Intellectual Disability, Short

- Stature, and Facial Dysmorphism.” *American Journal of Human Genetics* 104 (4): 758–66.
- Donk, Roos van der, Sandra Jansen, Janneke H. M. Schuurs-Hoeijmakers, David A. Koolen, Lia C. M. J. Goltstein, Alexander Hoischen, Han G. Brunner, et al. 2019. “Next-Generation Phenotyping Using Computer Vision Algorithms in Rare Genomic Neurodevelopmental Disorders.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 21 (8): 1719–25.
- Dudding-Byth, Tracy, Anne Baxter, Elizabeth G. Holliday, Anna Hackett, Sheridan O’Donnell, Susan M. White, John Attia, et al. 2017. “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 Biotechnology* 17 (1): 1–9.
- DuMont Schütte, August, Jürgen Hetzel, Sergios Gatidis, Tobias Hepp, Benedikt Dietz, Stefan Bauer, and Patrick Schwab. 2021. “Overcoming Barriers to Data Sharing with Medical Image Generation: A Comprehensive Evaluation.” *NPJ Digital Medicine* 4 (1): 141.
- Ebstein, Frédéric, Sébastien Küry, Victoria Most, Cory Rosenfelt, Marie-Pier Scott-Boyer, Geeske M. van Woerden, Thomas Besnard, et al. 2021. “De Novo Variants in the PSMC3 Proteasome AAA-ATPase Subunit Gene Cause Neurodevelopmental Disorders Associated with Type I Interferonopathies.” *BioRxiv*. <https://doi.org/10.1101/2021.12.07.21266342>.
- Ferreira, Carlos R. 2019. “The Burden of Rare Diseases.” *American Journal of Medical Genetics. Part A* 179 (6): 885–92.
- Ferry, Quentin, Julia Steinberg, Caleb Webber, David R. FitzPatrick, Chris P. Ponting, Andrew Zisserman, and Christoffer Nellåker. 2014. “Diagnostically Relevant Facial Gestalt Information from Ordinary Photos.” *ELife* 3 (June): e02020.
- Goldenberg, Alice, Florence Riccardi, Aude Tessier, Rolph Pfundt, Tiffany Busa, Pierre Cacciagli, Yline Capri, et al. 2016. “Clinical and Molecular Findings in 39 Patients with KBG Syndrome Caused by Deletion or Mutation of ANKRD11.” *American Journal of Medical Genetics. Part A* 170 (11): 2847–59.
- Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. “Generative Adversarial Nets.” In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 2672–80. NIPS’14. Cambridge, MA, USA: MIT Press.



- Guo, Lily, Jiyeon Park, Edward Yi, Elaine Marchi, Yana Kibalnyk, Anastassia Voronova, Tzung-Chien Hsieh, Peter M. Krawitz, and Gholson J. Lyon. 2021. “KBG Syndrome: Prospective Videoconferencing and Use of AI-Driven Facial Phenotyping in 25 New Patients.” *BioRxiv*. <https://doi.org/10.1101/2021.11.18.21266480>.
- Guo, Yandong, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. “MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition.” In *Computer Vision – ECCV 2016*, 87–102. Springer International Publishing.
- Gurovich, Yaron, Yair Hanani, Omri Bar, Guy Nadav, Nicole Fleischer, Dekel Gelbman, Lina Basel-Salmon, et al. 2019. “Identifying Facial Phenotypes of Genetic Disorders Using Deep Learning.” *Nature Medicine*. Nature Publishing Group. <https://doi.org/10.1038/s41591-018-0279-0>.
- Hallowell, Nina, Michael Parker, and Christoffer Nellåker. 2019. “Big Data Phenotyping in Rare Diseases: Some Ethical Issues.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 21 (2): 272–74.
- Han, Tianyu, Sven Nebelung, Christoph Haarbuerger, Nicolas Horst, Sebastian Reinartz, Dorit Merhof, Fabian Kiessling, Volkmar Schulz, and Daniel Truhn. 2020. “Breaking Medical Data Sharing Boundaries by Using Synthesized Radiographs.” *Science Advances* 6 (49). <https://doi.org/10.1126/sciadv.abb7973>.
- Han, Tianyu, Sven Nebelung, Federico Pedersoli, Markus Zimmermann, Maximilian Schulze-Hagen, Michael Ho, Christoph Haarbuerger, et al. 2021. “Advancing Diagnostic Performance and Clinical Usability of Neural Networks via Adversarial Training and Dual Batch Normalization.” *Nature Communications* 12 (1): 4315.
- Harms, Frederike Leonie, Katta M. Girisha, Andrew A. Hardigan, Fanny Kortüm, Anju Shukla, Malik Alawi, Ashwin Dalal, et al. 2017. “Mutations in EBF3 Disturb Transcriptional Profiles and Cause Intellectual Disability, Ataxia, and Facial Dysmorphism.” *American Journal of Human Genetics* 100 (1): 117–27.
- Hart, T. C., and P. S. Hart. 2009. “Genetic Studies of Craniofacial Anomalies: Clinical Implications and Applications.” *Orthodontics & Craniofacial Research* 12 (3): 212–20.
- Hennekam, Raoul C. M., and Leslie G. Biesecker. 2012. “Next-Generation Sequencing Demands next-Generation Phenotyping.” *Human Mutation* 33 (5): 884–86.
- Hong, Dian, Ying-Yi Zheng, Ying Xin, Ling Sun, Hang Yang, Min-Yin Lin, Cong Liu, et al. 2021. “Genetic Syndromes Screening by Facial Recognition Technology: VGG-

- 16 Screening Model Construction and Evaluation.” *Orphanet Journal of Rare Diseases* 16 (1): 344.
- Hoyer, Juliane, Arif B. Ekici, Sabine Ende, Bernt Popp, Christiane Zweier, Antje Wiesener, Eva Wohlleber, et al. 2012. “Haploinsufficiency of ARID1B, a Member of the SWI/SNF-a Chromatin-Remodeling Complex, Is a Frequent Cause of Intellectual Disability.” *American Journal of Human Genetics* 90 (3): 565–72.
- Hsieh, Tzung Chien, Martin A. Mensah, Jean T. Pantel, Dione Aguilar, Omri Bar, Allan Bayat, Luis Becerra-Solano, et al. 2019. “PEDIA: Prioritization of Exome Data by Image Analysis.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 21 (12): 2807–14.
- Hsieh, Tzung-Chien, Aviram Bar-Haim, Shahida Moosa, Nadja Ehmke, Karen W. Gripp, Jean Tori Pantel, Magdalena Danyel, et al. 2022. “GestaltMatcher Facilitates Rare Disease Matching Using Facial Phenotype Descriptors.” *Nature Genetics*, February. <https://doi.org/10.1038/s41588-021-01010-x>.
- Huang, Gary B., Marwan A. Mattar, Honglak Lee, and Erik Learned-Miller. 2012. “Learning to Align from Scratch.” In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, 764–72. NIPS’12. Red Hook, NY, USA: Curran Associates Inc.
- Huang, Gary B., Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. “Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.” University of Massachusetts, Amherst. <http://www.cs.umass.edu/lfw/>.
- Izadi, Saeed, Zahra Mirikharaji, Jeremy Kawahara, and Ghassan Hamarneh. 2018. “Generative Adversarial Networks to Segment Skin Lesions.” In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 881–84.
- Jansen, Sandra, Ilse M. van der Werf, A. Micheil Innes, Alexandra Afenjar, Pankaj B. Agrawal, Ilse J. Anderson, Paldeep S. Atwal, et al. 2019. “De Novo Variants in FBXO11 Cause a Syndromic Form of Intellectual Disability with Behavioral Problems and Dysmorphisms.” *European Journal of Human Genetics: EJHG* 27 (5): 738–46.
- Jit, Mark, Aparna Ananthakrishnan, Martin McKee, Olivier J. Wouters, Philippe Beutels, and Yot Teerawattananon. 2021. “Multi-Country Collaboration in Responding to Global Infectious Disease Threats: Lessons for Europe from the COVID-19 Pandemic.” *The Lancet Regional Health. Europe* 9 (October): 100221.

- Kamphans, Tom, and Peter M. Krawitz. 2012. "GeneTalk: An Expert Exchange Platform for Assessing Rare Sequence Variants in Personal Genomes." *Bioinformatics* 28 (19): 2515–16.
- Kanca, Oguz, Jonathan C. Andrews, Pei-Tseng Lee, Chirag Patel, Stephen R. Braddock, Anne M. Slavotinek, Julie S. Cohen, et al. 2019. "De Novo Variants in WDR37 Are Associated with Epilepsy, Colobomas, Dysmorphism, Developmental Delay, Intellectual Disability, and Cerebellar Hypoplasia." *American Journal of Human Genetics* 105 (2): 413–24.
- Kärkkäinen, Kimmo, and Jungseock Joo. 2021. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation." In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1547–57.
- Karras, Tero, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. "Progressive Growing of GANs for Improved Quality, Stability, and Variation." <https://openreview.net/pdf?id=Hk99zCeAb>.
- Karras, Tero, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. "Analyzing and Improving the Image Quality of StyleGAN." In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. <https://doi.org/10.1109/cvpr42600.2020.00813>.
- Kircher, Martin, Daniela M. Witten, Preti Jain, Brian J. O’Roak, Gregory M. Cooper, and Jay Shendure. 2014. "A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants." *Nature Genetics* 46 (3): 310–15.
- Kline, Antonie D., Joanna F. Moss, Angelo Selicorni, Anne-Marie Bisgaard, Matthew A. Deardorff, Peter M. Gillett, Stacey L. Ishman, et al. 2018. "Diagnosis and Management of Cornelia de Lange Syndrome: First International Consensus Statement." *Nature Reviews. Genetics* 19 (10): 649–66.
- Knaus, Alexej, Jean Tori Pantel, Manuela Pendziwiat, Nurulhuda Hajjir, Max Zhao, Tzung-Chien Hsieh, Max Schubach, et al. 2018. "Characterization of Glycosylphosphatidylinositol Biosynthesis Defects by Clinical Features, Flow Cytometry, and Automated Image Analysis." *Genome Medicine* 10 (1): 3.
- Köhler, Sebastian, Marcel H. Schulz, Peter Krawitz, Sebastian Bauer, Sandra Dölken, Claus E. Ott, Christine Mundlos, Denise Horn, Stefan Mundlos, and Peter N. Robinson. 2009. "Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies." *American Journal of Human Genetics* 85 (4): 457–64.
- Kuru, Kaya, Mahesan Niranjana, Yusuf Tunca, Erhan Osvank, and Tayyaba Azim. 2014. "Biomedical Visual Data Analysis to Build an Intelligent Diagnostic Decision

- Support System in Medical Genetics.” *Artificial Intelligence in Medicine* 62 (2): 105–18.
- Landrum, Melissa J., Shanmuga Chitipiralla, Garth R. Brown, Chao Chen, Baoshan Gu, Jennifer Hart, Douglas Hoffman, et al. 2020. “ClinVar: Improvements to Accessing Data.” *Nucleic Acids Research* 48 (D1): D835–44.
- Landrum, Melissa J., Jennifer M. Lee, Mark Benson, Garth R. Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, et al. 2018. “ClinVar: Improving Access to Variant Interpretations and Supporting Evidence.” *Nucleic Acids Research* 46 (D1): D1062–67.
- Li, Haoxiang, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. 2015. “A Convolutional Neural Network Cascade for Face Detection.” In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5325–34.
- Liehr, T., N. Acquarola, K. Pyle, S. St-Pierre, M. Rinholm, O. Bar, K. Wilhelm, and I. Schreyer. 2018. “Next Generation Phenotyping in Emanuel and Pallister-Killian Syndrome Using Computer-Aided Facial Dysmorphology Analysis of 2D Photos.” *Clinical Genetics* 93 (2): 378–81.
- Liu, Weiyang, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. “SphereFace: Deep Hypersphere Embedding for Face Recognition.” *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-Janua: 6738–46.
- Liu, Yi-Chieh, Hao-Hsiang Yang, Chao-Han Huck Yang, Jia-Hong Huang, Meng Tian, Hiromasa Morikawa, Yi-Chang James Tsai, and Jesper Tegner. 2019. “Synthesizing New Retinal Symptom Images by Multiple Generative Models.” *ArXiv [Cs.CV]*. arXiv. <http://arxiv.org/abs/1902.04147>.
- Lumaka, A., N. Cosemans, A. Lulebo Mampasi, G. Mubungu, N. Mvuama, T. Lubala, S. Mbuyi-Musanzayi, et al. 2017. “Facial Dysmorphism Is Influenced by Ethnic Background of the Patient and of the Evaluator.” *Clinical Genetics* 92 (2): 166–71.
- Maaten, Laurens van der, and Geoffrey Hinton. 2008. “Visualizing Data Using T-SNE.” *Journal of Machine Learning Research: JMLR* 9 (86): 2579–2605.
- Marbach, Felix, Cecilie F. Rustad, Angelika Riess, Dejan Đukić, Tzung Chien Hsieh, Itamar Jobani, Trine Prescott, et al. 2019. “The Discovery of a LEMD2-Associated Nuclear Envelopathy with Early Progeroid Appearance Suggests Advanced Applications for AI-Driven Facial Phenotyping.” *American Journal of Human Genetics* 104 (4): 749–57.

- Mascalzoni, Deborah, Heidi Beate Bentzen, Isabelle Budin-Ljøsne, Lee Andrew Bygrave, Jessica Bell, Edward S. Dove, Christian Fuchsberger, et al. 2019. "Are Requirements to Deposit Data in Research Repositories Compatible With the European Union's General Data Protection Regulation?" *Annals of Internal Medicine* 170 (5): 332–34.
- McKusick, V. A. 1969. "On Lumpers and Splitters, or the Nosology of Genetic Disease." *Perspectives in Biology and Medicine* 12 (2): 298–312.
- McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University. 2019. "Online Mendelian Inheritance in Man (OMIM)." 2019. <https://omim.org/>.
- Mescheder, Lars, Sebastian Nowozin, and Andreas Geiger. 2017. "The Numerics of GANs." In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 1823–33. NIPS'17. Red Hook, NY, USA: Curran Associates Inc.
- Morimoto, Marie, Helen Waller-Evans, Zineb Ammous, Xiaofei Song, Kevin A. Strauss, Davut Pehlivan, Claudia Gonzaga-Jauregui, et al. 2018. "Bi-Allelic CCDC47 Variants Cause a Disorder Characterized by Woolly Hair, Liver Dysfunction, Dysmorphic Features, and Global Developmental Delay." *American Journal of Human Genetics* 103 (5): 794–807.
- Moshtagh, Mozghan, Jila Mirlashari, and Rana Amiri. 2021. "Global Collaboration and Social Practices to Mitigate Impacts of COVID-19 in the World: A Lived Experience of Infecting." *Qualitative Social Work : QSW : Research and Practice* 20 (1–2): 366–74.
- Nellåker, Christoffer, Fowzan S. Alkuraya, Gareth Baynam, Raphael A. Bernier, Francois P. J. Bernier, Vanessa Boulanger, Michael Brudno, et al. 2019. "Enabling Global Clinical Collaborations on Identifiable Patient Data: The Minerva Initiative." *Frontiers in Genetics* 10 (July): 611.
- Nguengang Wakap, Stéphanie, Deborah M. Lambert, Annie Olry, Charlotte Rodwell, Charlotte Gueydan, Valérie Lanneau, Daniel Murphy, Yann Le Cam, and Ana Rath. 2020. "Estimating Cumulative Point Prevalence of Rare Diseases: Analysis of the Orphanet Database." *European Journal of Human Genetics: EJHG* 28 (2): 165–73.
- Olson, Heather E., Nolwenn Jean-Marçais, Edward Yang, Delphine Heron, Katrina Tatton-Brown, Paul A. van der Zwaag, Emilia K. Bijlsma, et al. 2018. "A Recurrent De Novo PACS2 Heterozygous Missense Variant Causes Neonatal-Onset Developmental Epileptic Encephalopathy, Facial Dysmorphism, and Cerebellar Dysgenesis." *American Journal of Human Genetics* 102 (5): 995–1007.

- Oti, Martin, Martijn A. Huynen, and Han G. Brunner. 2008. "Phenome Connections." *Trends in Genetics: TIG* 24 (3): 103–6.
- Pantel, Jean T., Max Zhao, Martin A. Mensah, Nurulhuda Hajjir, Tzung-Chien Hsieh, Yair Hanani, Nicole Fleischer, et al. 2018. "Advances in Computer-Assisted Syndrome Recognition by the Example of Inborn Errors of Metabolism." *Journal of Inherited Metabolic Disease*, April. <https://doi.org/10.1007/s10545-018-0174-3>.
- Pantel, Jean Tori, Nurulhuda Hajjir, Magdalena Danyel, Jonas Elsner, Angela Teresa Abad-Perez, Peter Hansen, Stefan Mundlos, et al. 2020. "Efficiency of Computer-Aided Facial Phenotyping (DeepGestalt) in Individuals With and Without a Genetic Syndrome: Diagnostic Accuracy Study." *Journal of Medical Internet Research* 22 (10): e19263.
- Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman. 2015. "Deep Face Recognition." In *Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association. <https://doi.org/10.5244/c.29.41>.
- Pengelly, Reuben J., Thahmina Alom, Zijian Zhang, David Hunt, Sarah Ennis, and Andrew Collins. 2017. "Evaluating Phenotype-Driven Approaches for Genetic Diagnoses from Exomes in a Clinical Setting." *Scientific Reports* 7 (1): 13509.
- Philippakis, Anthony A., Danielle R. Azzariti, Sergi Beltran, Anthony J. Brookes, Catherine A. Brownstein, Michael Brudno, Han G. Brunner, et al. 2015. "The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery." *Human Mutation* 36 (10): 915–21.
- Porras, Antonio R., Kenneth Rosenbaum, Carlos Tor-Diez, Marshall Summar, and Marius George Linguraru. 2021. "Development and Evaluation of a Machine Learning-Based Point-of-Care Screening Tool for Genetic Syndromes in Children: A Multinational Retrospective Study." *The Lancet. Digital Health*, September. [https://doi.org/10.1016/S2589-7500\(21\)00137-0](https://doi.org/10.1016/S2589-7500(21)00137-0).
- Richards, Sue, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, et al. 2015. "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." *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 17 (5): 405–24.
- Robinson, Peter N., Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. 2008. "The Human Phenotype Ontology: A Tool for Annotating

- and Analyzing Human Hereditary Disease.” *American Journal of Human Genetics*, 610–15.
- Rothe, Rasmus, Radu Timofte, and Luc Van Gool. 2018. “Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks.” *International Journal of Computer Vision* 126 (2): 144–57.
- Rousseeuw, Peter J. 1987. “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.” *Journal of Computational and Applied Mathematics* 20: 53–65.
- Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, et al. 2014. “ImageNet Large Scale Visual Recognition Challenge.” *ArXiv [Cs.CV]*. arXiv. <http://arxiv.org/abs/1409.0575>.
- Santiago-Sim, Teresa, Lindsay C. Burrage, Frédéric Ebstein, Mari J. Tokita, Marcus Miller, Weimin Bi, Alicia A. Braxton, et al. 2017. “Biallelic Variants in OTUD6B Cause an Intellectual Disability Syndrome Associated with Seizures and Dysmorphic Features.” *American Journal of Human Genetics* 100 (4): 676–88.
- Schroff, Florian, Dmitry Kalenichenko, and James Philbin. 2015. “FaceNet: A Unified Embedding for Face Recognition and Clustering.” In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June-2015:815–23. IEEE Computer Society.
- Schuurs-Hoeijmakers, Janneke H. M., Edwin C. Oh, Lisenka E. L. M. Vissers, Mariëlle E. M. Swinkels, Christian Gilissen, Michèl A. Willemsen, Maureen Holvoet, et al. 2012. “Recurrent de Novo Mutations in PACS1 Cause Defective Cranial-Neural-Crest Migration and Define a Recognizable Intellectual-Disability Syndrome.” *American Journal of Human Genetics* 91 (6): 1122–27.
- Shukla, P., T. Gupta, A. Saini, P. Singh, and R. Balasubramanian. 2017. “A Deep Learning Frame-Work for Recognizing Developmental Disorders.” In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 705–14.
- Smedley, Damian, Julius O. B. Jacobsen, Marten Jäger, Sebastian Köhler, Manuel Holtgrewe, Max Schubach, Enrico Siragusa, et al. 2015. “Next-Generation Diagnostics and Disease-Gene Discovery with the Exomiser.” *Nature Protocols* 10 (12): 2004–15.
- Sobreira, Nara, François Schiettecatte, David Valle, and Ada Hamosh. 2015. “GeneMatcher: A Matching Tool for Connecting Investigators with an Interest in the Same Gene.” *Human Mutation* 36 (10): 928–30.

- Stankiewicz, Paweł, Tahir N. Khan, Przemysław Szafranski, Leah Slattery, Haley Streff, Francesco Vetrini, Jonathan A. Bernstein, et al. 2017. "Haploinsufficiency of the Chromatin Remodeler BPTF Causes Syndromic Developmental and Speech Delay, Postnatal Microcephaly, and Dysmorphic Features." *American Journal of Human Genetics* 101 (4): 503–15.
- Stephen, Joshi, Sateesh Maddirevula, Sheela Nampoothiri, John D. Burke, Matthew Herzog, Anju Shukla, Katharina Steindl, et al. 2018. "Bi-Allelic TMEM94 Truncating Variants Are Associated with Neurodevelopmental Delay, Congenital Heart Defects, and Distinct Facial Dysmorphism." *American Journal of Human Genetics* 103 (6): 948–67.
- Stevens, Servi J. C., Anthonie J. van Essen, Conny M. A. van Ravenswaaij, Abdallah F. Elias, Jaclyn A. Haven, Stefan H. Lelieveld, Rolph Pfundt, et al. 2016. "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 Medicine* 8 (1): 131.
- Sun, Yi, Xiaogang Wang, and Xiaoou Tang. 2014. "Deep Learning Face Representation from Predicting 10,000 Classes." In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. <https://doi.org/10.1109/cvpr.2014.244>.
- Taigman, Yaniv, Ming Yang, Marc'aurelio Ranzato, and Lior Wolf. 2014. "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*, 1701–8. IEEE Computer Society.
- Tanaka, Akemi J., Megan T. Cho, Kyle Retterer, Julie R. Jones, Catherine Nowak, Jessica Douglas, Yong-Hui Jiang, et al. 2016. "De Novo Pathogenic Variants in CHAMP1 Are Associated with Global Developmental Delay, Intellectual Disability, and Dysmorphic Facial Features." *Cold Spring Harbor Molecular Case Studies* 2 (1): a000661.
- Tavtigian, Sean V., Marc S. Greenblatt, Steven M. Harrison, Robert L. Nussbaum, Snehit A. Prabhu, Kenneth M. Boucher, and Leslie G. Biesecker. 2018. "Modeling the ACMG/AMP Variant Classification Guidelines as a Bayesian Classification Framework." *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 20 (9): 1054–60.
- Vervoort, Dominique, Xiya Ma, and Jessica G. Y. Luc. 2021. "COVID-19 Pandemic: A Time for Collaboration and a Unified Global Health Front." *International Journal for Quality in Health Care: Journal of the International Society for Quality in Health Care / ISQua* 33 (1). <https://doi.org/10.1093/intqhc/mzaa065>.



- Wang, Hao, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. "CosFace: Large Margin Cosine Loss for Deep Face Recognition." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 5265–74.
- Wang, Kuan, and Jiebo Luo. 2016. "Detecting Visually Observable Disease Symptoms from Faces." *EURASIP Journal on Bioinformatics & Systems Biology* 2016 (1): 13.
- Warnat-Herresthal, Stefanie, Hartmut Schultze, Krishnaprasad Lingadahalli Shastry, Sathyanarayanan Manamohan, Saikat Mukherjee, Vishesh Garg, Ravi Sarveswara, et al. 2021. "Swarm Learning for Decentralized and Confidential Clinical Machine Learning." *Nature* 594 (7862): 265–70.
- Weiss, Karin, Paulien A. Terhal, Lior Cohen, Michael Bruccoleri, Melita Irving, Ariel F. Martinez, Jill A. Rosenfeld, et al. 2016. "De Novo Mutations in CHD4, an ATP-Dependent Chromatin Remodeler Gene, Cause an Intellectual Disability Syndrome with Distinctive Dysmorphisms." *American Journal of Human Genetics* 99 (4): 934–41.
- Winter, R. M., and M. Baraitser. 1987. "The London Dysmorphology Database." *Journal of Medical Genetics* 24 (8): 509–10.
- Wright, Caroline F., Jeremy F. McRae, Stephen Clayton, Giuseppe Gallone, Stuart Aitken, Tomas W. FitzGerald, Philip Jones, et al. 2018. "Making New Genetic Diagnoses with Old Data: Iterative Reanalysis and Reporting from Genome-Wide Data in 1,133 Families with Developmental Disorders." *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 20 (10): 1216–23.
- Yi, Dong, Zhen Lei, Shengcai Liao, and Stan Z. Li. 2014. "Learning Face Representation from Scratch." *ArXiv [Cs.CV]*. arXiv. <http://arxiv.org/abs/1411.7923>.
- Zhang, Kaipeng, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks." *IEEE Signal Processing Letters* 23 (10): 1499–1503.
- Zhang, Zhifei, Yang Song, and Hairong Qi. 2017. "Age Progression/Regression by Conditional Adversarial Autoencoder." In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. <https://doi.org/10.1109/cvpr.2017.463>.
- Zhao, He, Huiqi Li, Sebastian Maurer-Stroh, and Li Cheng. 2018. "Synthesizing Retinal and Neuronal Images with Generative Adversarial Nets." *Medical Image Analysis* 49 (October): 14–26.



## Appendix

### A.1 Supplementary information of GestaltMatcher

Please find the Supplementary Information and Supplementary Table 8 in the following links from the journal.

- Supplementary Information:  
[https://static-content.springer.com/esm/art%3A10.1038%2Fs41588-021-01010-x/MediaObjects/41588\\_2021\\_1010\\_MOESM1\\_ESM.pdf](https://static-content.springer.com/esm/art%3A10.1038%2Fs41588-021-01010-x/MediaObjects/41588_2021_1010_MOESM1_ESM.pdf)
- Supplementary Table 8:  
[https://static-content.springer.com/esm/art%3A10.1038%2Fs41588-021-01010-x/MediaObjects/41588\\_2021\\_1010\\_MOESM4\\_ESM.xlsx](https://static-content.springer.com/esm/art%3A10.1038%2Fs41588-021-01010-x/MediaObjects/41588_2021_1010_MOESM4_ESM.xlsx)

### A.2 Supplementary information of PEDIA

Please find the Supplementary Materials and Supplementary Table 1 in the following links from the journal.

- Supplementary Material:  
[https://static-content.springer.com/esm/art%3A10.1038%2Fs41436-019-0566-2/MediaObjects/41436\\_2019\\_566\\_MOESM1\\_ESM.pdf](https://static-content.springer.com/esm/art%3A10.1038%2Fs41436-019-0566-2/MediaObjects/41436_2019_566_MOESM1_ESM.pdf)
- Supplementary Table 1:  
[https://static-content.springer.com/esm/art%3A10.1038%2Fs41436-019-0566-2/MediaObjects/41436\\_2019\\_566\\_MOESM2\\_ESM.xlsx](https://static-content.springer.com/esm/art%3A10.1038%2Fs41436-019-0566-2/MediaObjects/41436_2019_566_MOESM2_ESM.xlsx)