# Improving deep facial phenotyping for ultra-rare disorder verification using model ensembles: supplementary materials

## LIST OF FIGURES



## LIST OF TABLES

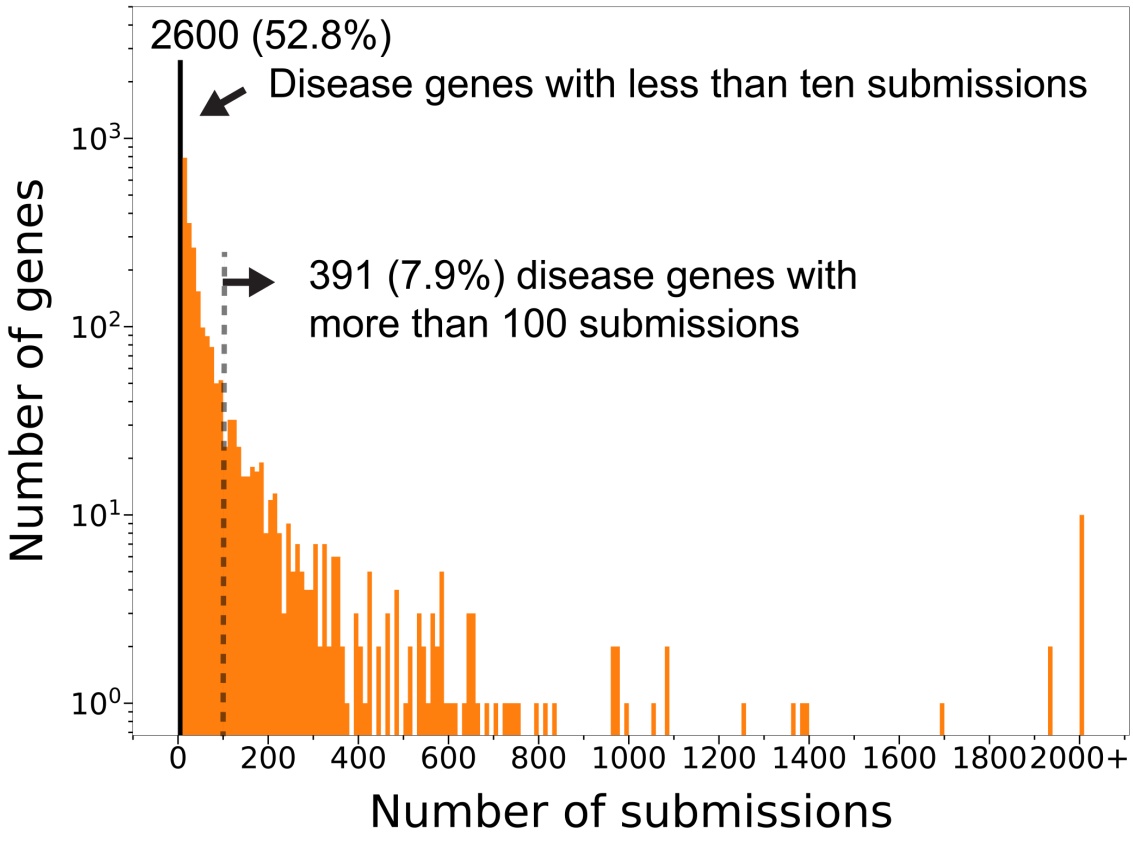


**Fig. S1.** Estimation of disease prevalence distribution by ClinVar submissions [1]. The X-axis is the number of pathogenic submissions for each gene in ClinVar. The Y-axis shows the number of genes with the respective number of submissions, and it is in log scale. The black bar indicates disease genes with less than ten submissions which cover 52.8% disease genes (2600/4920), and only 7.9% disease genes (391/4920) have more than 100 submissions. It shows an extremely imbalance distribution.

## 1. ESTIMATION OF DISORDER PREVALENCE

We utilized pathogenic variants submitted to ClinVar [1] to estimate disorder prevalence. The number of pathogenic submissions for each gene can be seen as the number of patients diagnosed with this disease gene. When discussing rare Mendelian disorders, “disorder" and “gene" is usually interchangeable because a disorder is caused by a disease-causing gene. Moreover, the researcher first finds a novel disease gene. Later, this gene will be reviewed and linked to a disorder. Therefore, we used the number of patients to estimate the overall disease prevalence, and it is shown in Figure S1. The prevalence of the disorders is extremely imbalance as shown in Figure S1. 52.8% of the disease genes have less than ten submissions, while only 7.9% of genes have more than 100 submissions.

## 2. PERFORMANCE WITH SEPERATE GALLERY

The main paper includes the performance of using GMDB's separate frequent and rare galleries, with just the top-1 and top-5 accuracy. In this supplemental section we extend those tables with top-10 and top-30 to allow for more complete overview, and comparison with other works. The performances are shown in Tables S1, S2, S3, and S4.

| Model | LFW | GMDB-Frequent | | | | GMDB-Rare | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Top-1 | Top-5 | Top-10 | Top-30 | Top-1 | Top-5 | Top-10 | Top-30 |
| GM-Hsieh2022 | 93.8% | 10.99% | 29.39% | 38.45% | 58.22% | 14.64% | 27.03% | 36.01% | 53.06% |
| GM-Hsieh2022* | - | **15.96%** | **33.83%** | **45.46%** | **69.64%** | **19.26%** | **36.28%** | **44.07%** | **60.73%** |
| ArcFace-r50 | 98.4% | 21.84% | 40.87% | 49.06% | 67.99% | **22.74%** | **37.35%** | **46.06%** | **61.49%** |
| ArcFace-r50* | - | **35.37%** | **53.25%** | **61.56%** | **77.34%** | 19.29% | 36.00% | 44.19% | 60.43% |

**Table S1.** Extended comparison of the performance of the GM-Hsieh2022 model and the ArcFace-r50 model on LFW and GMDB. Both have been pretrained on CASIA and models marked with (*) have been fine-tuned on GMDB. For each column, the best accuracy between the models before fine-tuning and after fine-tuning is boldfaced. This table is an extention of Table 3 in the main paper.

| Dataset | LFW | GMDB-Frequent | | | | GMDB-Rare | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Top-1 | Top-5 | Top-10 | Top-30 | Top-1 | Top-5 | Top-10 | Top-30 |
| VGG2 | 98.5% | 15.52% | 31.56% | 39.90% | 65.72% | 20.31% | 33.57% | 41.79% | 56.10% |
| CASIA | 98.4% | 21.84% | 40.87% | 49.06% | 67.99% | 22.74% | 37.35% | 46.06% | 61.49% |
| MS1MV2 | 99.0% | 29.14% | 48.86% | 58.98% | 78.69% | 29.04% | 44.74% | 53.65% | 69.81% |
| MS1MV3 | 98.9% | 31.54% | 49.36% | 60.50% | 76.67% | 29.52% | 46.36% | 53.40% | 69.69% |
| Glint360K | 99.0% | 32.43% | 53.14% | 63.18% | 78.12% | 33.00% | 47.62% | 56.49% | 71.08% |
| VGG2* | 85.8% | 27.50% | 49.92% | 58.13% | 70.90% | 17.56% | 33.41% | 42.11% | 56.22% |
| CASIA* | 75.7% | 35.37% | 53.25% | 61.56% | 77.34% | 19.29% | 36.00% | 44.19% | 60.43% |
| MS1MV2* | 84.1% | 39.98% | 59.81% | 68.82% | 83.06% | 21.86% | 39.89% | 48.83% | 65.27% |
| MS1MV3* | 76.4% | 45.06% | 64.64% | 68.28% | 82.17% | 24.31% | 40.28% | 48.56% | 64.50% |
| Glint360K* | 84.9% | 41.58% | 62.60% | 70.18% | 84.34% | 26.55% | 42.69% | 51.16% | 66.35% |

**Table S2.** Comparison of the performance of the ArcFace-r50 models trained on a variety of face recognition datasets. Models marked with (*) have been fine-tuned on GMDB. This table is an extention to Table 4 in the main paper.

| Model | LFW | GMDB-Frequent | | | | GMDB-Rare | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Top-1 | Top-5 | Top-10 | Top-30 | Top-1 | Top-5 | Top-10 | Top-30 |
| r50 | 84.9% | 41.58% | 62.60% | 70.18% | 84.34% | 26.55% | 42.69% | 51.16% | **66.35%** |
| r50-D/O | 86.2% | **46.95%** | **66.07%** | **74.78%** | **85.80%** | 28.85% | 45.36% | 50.61% | 63.72% |
| r50-D/O† | **87.6** % | 44.33% | 65.76% | 73.44% | 83.98% | **29.06%** | **46.35%** | **52.86%** | 66.02% |
| r100 | 91.0% | 47.96% | 68.87% | 77.06% | 84.48% | 26.03% | 42.22% | 49.67% | 66.22% |
| r100-D/O | 91.1 % | 48.37% | **71.78%** | 78.29% | **88.51%** | 28.02% | 44.32% | 52.58% | 67.63% |
| r100-D/O† | **93.0%** | **49.25%** | 69.95% | **78.38%** | 84.82% | **30.33%** | **47.85%** | **56.20%** | **70.19%** |

**Table S3.** Comparison of the performance of iResNet-50 and -100 fine-tuned on GMDB. D/O indicates an additional dropout layer and (†) indicates the use of $L_2$ weight decay on the feature layer. For each column, the best accuracy among the models (without regularizaton, D/O, and D/O†) is boldfaced. This table is an extention to Table 5 in the main paper.

| Model | Dataset | Loss | GMDB-Frequent | | | | GMDB-Rare | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Top-1 | Top-5 | Top-10 | Top-30 | Top-1 | Top-5 | Top-10 | Top-30 |
| GM-Hsieh2022 | CASIA* | CE | 15.96% | 33.83% | 45.46% | 69.64% | 19.26% | 36.28% | 44.07% | 60.73% |
| r50-D/O† | Glint360K* | CE | 44.33% | 65.76% | 73.44% | 83.98% | 29.06% | 46.35% | 52.86% | 66.02% |
| r50-D/O†+ TTA | Glint360K* | CE | 47.73% | 67.67% | 72.09% | 85.95% | 30.29% | 46.38% | 54.08% | 69.50% |
| r100-D/O | Glint360K* | CE | 48.37% | **71.78%** | 78.29% | 88.51% | 28.02% | 44.32% | 52.58% | 67.63% |
| r100-D/O + TTA | Glint360K* | CE | 51.16% | 69.58% | **79.21%** | 88.27% | 27.92% | 46.26% | 54.25% | 69.46% |
| r100 | Glint360K | ArcFace | 30.25% | 54.81% | 64.91% | 79.60% | 33.25% | 50.22% | 57.91% | 72.15% |
| r100 + TTA | Glint360K | ArcFace | 35.25% | 56.52% | 64.73% | 81.88% | 33.47% | 51.61% | 58.89% | 71.92% |
| Model ensemble | n/a | n/a | 52.06% | 70.70% | 77.84% | 89.97% | 34.93% | 52.78% | 61.65% | **76.81%** |
| Model ensemble + TTA | n/a | n/a | **52.99%** | 71.01% | 79.19% | **89.99%** | **35.98%** | **53.93%** | **62.43%** | 76.56% |

**Table S4.** Comparison of the performance of the GM-Hsieh2022 model, two ArcFace models fine-tuned on GMDB, one ArcFace face verifrication model, and finally our model ensemble using the three listed ArcFace models. TTA indicates the model was evaluated using test time augmentation, (*) indicates the model was fine-tuned on GMDB, (D/O) indicates and additional dropout layer and (†) indicates the use of $L_2$ weight decay on the feature layer. This table is an extention to Table 6 in the main paper.

## 3. PERFORMANCE WITH UNIFIED GALLERY

In the main paper, we presented the performance of using GMDB's separate frequent and rare galleries. When evaluating the frequent set, we used only the frequent set as gallery, and when evaluting the rare set, we used only the rare set as gallery. To simulate real-world scenarios, we combined the frequent gallery and rare gallery into a unified gallery. We reported the performance when using the unified gallery in Tables S5, S6, S7, and S8.

| Model | LFW | GMDB-Frequent | | | | GMDB-Rare | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Top-1 | Top-5 | Top-10 | Top-30 | Top-1 | Top-5 | Top-10 | Top-30 |
| GM-Hsieh2022 | 93.8% | 10.19% | 28.40% | 36.27% | 55.04% | 9.02% | 13.74% | 17.51% | 27.39% |
| GM-Hsieh2022* | - | **14.81%** | **31.72%** | **43.49%** | **64.97%** | **9.99%** | **17.87%** | **23.48%** | **37.19%** |
| ArcFace-r50 | 98.4% | 20.41% | 38.38% | 47.13% | 64.46% | **15.39%** | 24.37% | 28.42% | 38.10% |
| ArcFace-r50* | - | **32.84%** | **49.51%** | **56.56%** | **70.57%** | 11.73% | **25.34%** | **32.39%** | **45.82%** |

**Table S5.** Comparison of the performance of the GM-Hsieh2022 model and the ArcFace-r50 model on LFW and GMDB with **unified gallery (frequent + rare)**. Both have been pretrained on CASIA and models marked with (*) have been fine-tuned on GMDB. For each column, the best accuracy between the models before fine-tuning and after fine-tuning is boldfaced. This table is the supplement to Table 3 in the main paper, which provides the performance when using the unified gallery.

| Dataset | LFW | GMDB-Frequent | | | | GMDB-Rare | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Top-1 | Top-5 | Top-10 | Top-30 | Top-1 | Top-5 | Top-10 | Top-30 |
| VGG2 | 98.5% | 15.20% | 29.08% | 36.93% | 63.59% | 11.87% | 20.20% | 24.92% | 33.92% |
| CASIA | 98.4% | 20.41% | 38.38% | 47.13% | 64.46% | 15.39% | 24.37% | 28.42% | 38.10% |
| MS1MV2 | 99.0% | 28.16% | 47.79% | 57.32% | 74.42% | 18.40% | 29.89% | 35.33% | 45.98% |
| MS1MV3 | 98.9% | 30.40% | 48.40% | 58.88% | 74.47% | 18.62% | 31.30% | 37.21% | 46.87% |
| Glint360K | 99.0% | 30.95% | 51.57% | 60.49% | 76.13% | 21.60% | 32.93% | 38.11% | 48.60% |
| VGG2* | 85.8% | 25.18% | 43.79% | 53.63% | 66.95% | 10.01% | 22.82% | 30.37% | 43.67% |
| CASIA* | 75.7% | 32.84% | 49.51% | 56.56% | 70.57% | 11.73% | 25.34% | 32.39% | 45.82% |
| MS1MV2* | 84.1% | 37.94% | 54.15% | 64.46% | 76.92% | 13.51% | 28.13% | 35.93% | 50.51% |
| MS1MV3* | 76.4% | 43.09% | 60.71% | 67.32% | 74.93% | 15.58% | 28.95% | 36.77% | 50.51% |
| Glint360K* | 84.9% | 38.37% | 59.35% | 67.80% | 77.13% | 18.41% | 30.36% | 36.64% | 51.65% |

**Table S6.** Comparison of the performance of the ArcFace-r50 models trained on a variety of face recognition datasets with **unified gallery (frequent + rare)**. Models marked with (*) have been fine-tuned on GMDB. This table is the supplement to Table 4 in the main paper, which provides the performance when using the unified gallery.

| Model | LFW | GMDB-Frequent | | | | GMDB-Rare | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Top-1 | Top-5 | Top-10 | Top-30 | Top-1 | Top-5 | Top-10 | Top-30 |
| r50 | 84.9% | 38.37% | 59.35% | 67.80% | 77.13% | 18.41% | 30.36% | 36.64% | 51.65% |
| r50-D/O | 86.2% | **44.88%** | **64.03%** | **72.87%** | 81.03% | 17.76% | **30.73%** | 36.79% | 48.28% |
| r50-D/O† | **87.6%** | 42.19% | 63.90% | 70.73% | **81.24%** | **19.52%** | 30.38% | **37.10%** | **50.29%** |
| r100 | 91.0% | 45.23% | 65.23% | 72.93% | 82.68% | 17.53% | 30.98% | 38.42% | 50.20% |
| r100-D/O | 91.1 % | 46.86% | **68.84%** | **76.24%** | **84.19%** | 17.04% | 29.43% | 36.96% | 50.43% |
| r100-D/O† | **93.0%** | **47.98%** | 68.25% | 76.08% | 83.86% | **19.18%** | **31.51%** | **39.14%** | **51.93%** |

**Table S7.** Comparison of the performance of iResNet-50 and -100 fine-tuned on GMDB with **unified gallery (frequent + rare)**. D/O indicates an additional dropout layer and (†) indicates the use of $L_2$ weight decay on the feature layer. For each column, the best accuracy among the models (without regularizaton, D/O, and D/O†) is boldfaced. This table is the supplement to Table 5 in the main paper, which provides the performance when using the unified gallery.

| Model | Dataset | Loss | GMDB-Frequent | | | | GMDB-Rare | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Top-1 | Top-5 | Top-10 | Top-30 | Top-1 | Top-5 | Top-10 | Top-30 |
| GM-Hsieh2022 | CASIA* | CE | 14.81% | 31.72% | 43.49% | 64.97% | 9.99% | 17.87% | 23.48% | 37.19% |
| r50-D/O† | Glint360K* | CE | 42.19% | 63.90% | 70.73% | 81.24% | 19.52% | 30.38% | 37.10% | 50.29% |
| r50-D/O†+ TTA | Glint360K* | CE | 45.74% | 65.29% | 70.94% | 84.06% | 19.25% | 31.65% | 38.83% | 51.13% |
| r100-D/O | Glint360K* | CE | 46.86% | 68.84% | 76.24% | 84.19% | 17.04% | 29.43% | 36.96% | 50.43% |
| r100-D/O + TTA | Glint360K* | CE | 48.62% | 67.87% | **76.78%** | 85.66% | 17.08% | 29.79% | 38.28% | 52.49% |
| r100 | Glint360K | ArcFace | 29.65% | 53.80% | 62.40% | 77.49% | 21.79% | 35.15% | 40.25% | 51.18% |
| r100 + TTA | Glint360K | ArcFace | 34.12% | 53.84% | 62.03% | 78.24% | 22.35% | 34.41% | 39.79% | 51.89% |
| Model ensemble | n/a | n/a | 49.59% | **69.77%** | 75.44% | 87.79% | **24.11%** | 37.64% | 43.19% | 57.92% |
| Model ensemble + TTA | n/a | n/a | **50.79%** | 69.17% | 76.66% | **88.37%** | 24.05% | **38.44%** | **44.53%** | **57.95%** |

**Table S8.** Comparison of the performance of the GM-Hsieh2022 model, two ArcFace models fine-tuned on GMDB, one ArcFace face verifrication model, and finally our model ensemble using the three listed ArcFace models with **unified gallery (frequent + rare)**. TTA indicates the model was evaluated using test time augmentation, (*) indicates the model was fine-tuned on GMDB, (D/O) indicates and additional dropout layer and (†) indicates the use of $L_2$ weight decay on the feature layer. This table is the supplement to Table 6 in the main paper, which provides the performance when using the unified gallery.

## REFERENCES

1. M. J. Landrum, J. M. Lee, M. Benson, G. R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Karapetyan, K. Katz, C. Liu, Z. Maddipatla, A. Malheiro, K. McDaniel, M. Ovetsky, G. Riley, G. Zhou, J. B. Holmes, B. L. Kattman, and D. R. Maglott, "ClinVar: improving access to variant interpretations and supporting evidence," Nucleic Acids Res. **46**, D1062–D1067 (2018).