Evaluating unsupervised disentangled representation learning for genomic discovery and disease risk prediction

Taedong Yun [1]

Abstract

High-dimensional clinical data have become invaluable resources for genetic studies, due to their accessibility in biobank-scale datasets and the development of high performance modeling techniques especially using deep learning. Recent work has shown that low dimensional embeddings of these clinical data learned by variational autoencoders (VAE) can be used for genome-wide association studies and polygenic risk prediction. In this work, we consider multiple unsupervised learning methods for learning disentangled representations, namely autoencoders, VAE, β-VAE, and FactorVAE, in the context of genetic association studies. Using spirograms from UK Biobank as a running example, we observed improvements in the number of genome-wide significant loci, heritability, and performance of polygenic risk scores for asthma and chronic obstructive pulmonary disease by using FactorVAE or β-VAE, compared to standard VAE or non-variational autoencoders. FactorVAEs performed effectively across multiple values of the regularization hyperparameter, while β-VAEs were much more sensitive to the hyperparameter values.

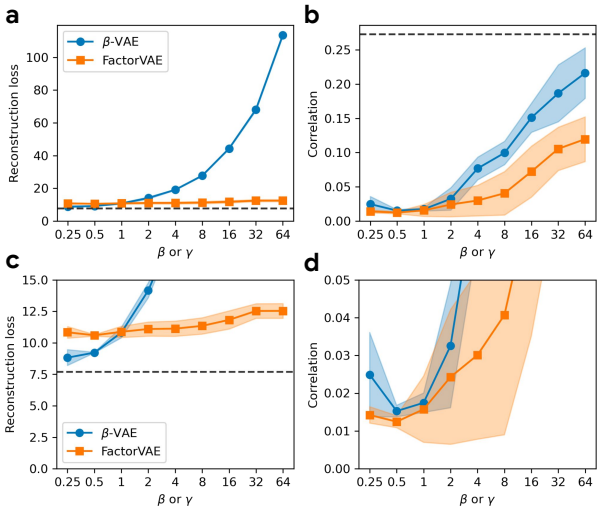

Figure 1. Reconstruction error and correlation between coordinates, as measured by the mean-squared error and the mean absolute values of correlations between all pairs of coordinates. We vary β or γ parameters in β-VAE and FactorVAE, which controls the strength of "regularization". Each setup is trained with 10 different random seeds. The data points are the mean of the 10 trials and the shaded regions are 95% confidence intervals. The horizontal dashed line is from a non-variational autoencoder. **(a)** Reconstruction error in the validation set. **(b)** Mean of absolute values of correlations between all pairs of coordinates. **(c)** Same as (a), zoomed in. **(d)** Same as (b), zoomed in.

1. Introduction

Large-scale biobank projects such as UK Biobank with deep phenotyping and genotyping data enabled new frontiers in the study of human genetics (Bycroft et al., 2018). High performance analysis methods using deep learning are well-suited for utilizing the high-dimensional clinical data (HDCD) (e.g. time series, images, videos) available in these datasets (Bai et al., 2020; Alipanahi et al., 2021; Aung et al., 2022; Pirruccello et al., 2022; Cosentino et al., 2023).

A recent study (Yun et al., 2023) demonstrated the potential for using HDCD in the context of genome-wide association studies (GWAS) to discover novel genetic insights about biological function, without any disease or trait labels. Using deep unsupervised representation learning techniques, specifically variational autoencoders (VAE) (Kingma & Welling, 2014), the authors demonstrated new associations between genotypes and lung function captured by spirograms, a graphical representation of widely-used clinical pulmonary function test results. VAEs generate latent representations whose coordinates are relatively *disentangled*, in which separable biological function can be better captured.

In this work, we extend this line of investigation by con-

[1]Google Research, Cambridge, MA 02142, USA. Correspondence to: Taedong Yun <tedyun@google.com>.

The 2023 ICML Workshop on Computational Biology. Honolulu, Hawaii, USA, 2023.

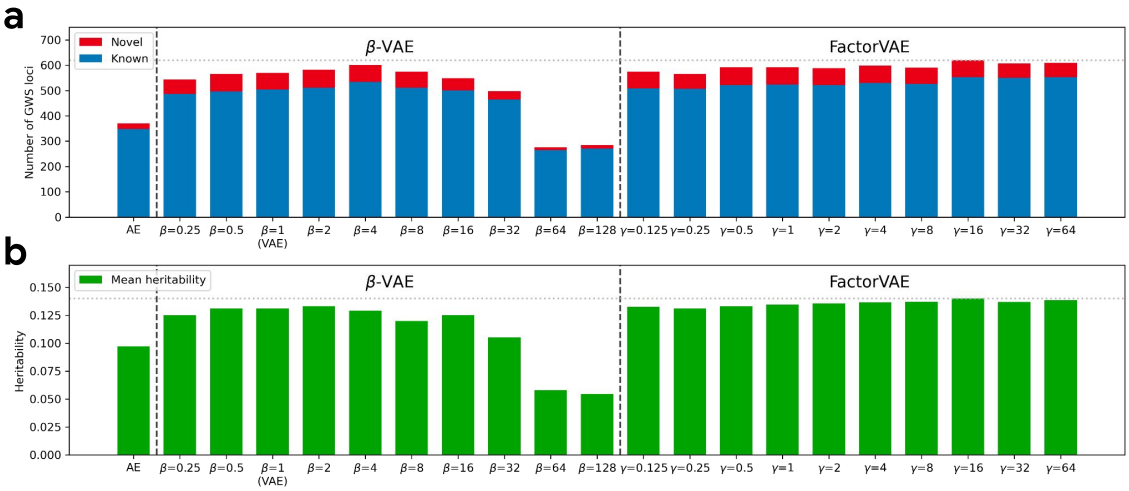

Figure 2. Genome-wide significant loci and heritability from GWAS on unsupervised representations. **(a)** Number of genome-wide significant loci. "Known" and "novel" is with respect to all loci from (Shrine et al., 2023) and GWAS Catalog lung function search. **(b)** Mean heritability (h^2_g) of all coordinates of the learned representation estimated by LD Score regression. Horizontal dotted line indicates the highest value. AE = (non-variational) autoencoder.

sidering several modifications of VAE introduced to further increase the effect of disentanglement, namely β-VAE (Higgins et al., 2017) and FactorVAE (Kim & Mnih, 2018), in addition to standard VAE and (non-variational) autoencoders. We perform comprehensive evaluation of these methods in terms of reconstruction performance, correlation between coordinates (as a measure of disentanglement), and their performance in the context of GWAS for genomic discovery and polygenic risk prediction.

2. Background

In representation learning of high-dimensional data, representations with "disentangled" coordinates are generally preferred so that independent factors of variations in the data can be separately captured by each coordinate (Bengio et al., 2013). One of the most widely used methods for unsupervised learning of disentangled representation is VAE, in which the Kullback–Leibler (KL) divergence term $D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}))$ in the loss function implicitly encourages the learned approximate posterior distribution $q(\mathbf{z}|\mathbf{x})$ to have independent coordinates when used with a factorized Gaussian prior distribution for $p(\mathbf{z})$. Several extensions or modifications to the VAE have been proposed to further enhance the disentanglement effect (Higgins et al., 2017; Burgess et al., 2018; Kim & Mnih, 2018; Kumar et al., 2018; Chen et al., 2018). See (Locatello et al., 2019) for empirical comparison of these methods.

In this work, we focus our attention to two of these methods: β-VAE (Higgins et al., 2017) and FactorVAE (Kim & Mnih, 2018). β-VAE amplifies the KL divergence term in the VAE loss function by a factor of $\beta > 1$, strengthening the "regularization" effect. With $\beta = 1$, β-VAE reduces to the standard VAE. On the other hand, FactorVAE adds an additional loss term to explicitly penalize the *total correlation* (TC; a measure of dependence for multiple random variables) (Watanabe, 1960) of the latent coordinates, where an additional discriminator is jointly trained to approximate the TC using the density ratio trick (Nguyen et al., 2010; Sugiyama et al., 2012). FactorVAE also has a hyperparameter γ that controls the strength of the TC loss term.

Meanwhile, a recent work (Yun et al., 2023) explored utilizing the disentangled representation of high-dimensional clinical data (HDCD) for genetic studies, specifically in the context of genome-wide association studies (GWAS) and polygenic risk scores (PRS). In their REGLE (REpresentation learning for Genetic discovery on Low-dimensional Embeddings) framework, they used VAEs to learn unsupervised disentangled representations of HDCD, e.g. spirograms, and performed GWAS on each coordinate to discover novel lung function loci in addition to recovering known lung function loci all without any disease labels. Moreover, using a (small) number of samples with paired genetics and disease labels (but not HDCD), the PRSs of the latent embeddings can be combined into a PRS specific to the given lung disease *post hoc*.

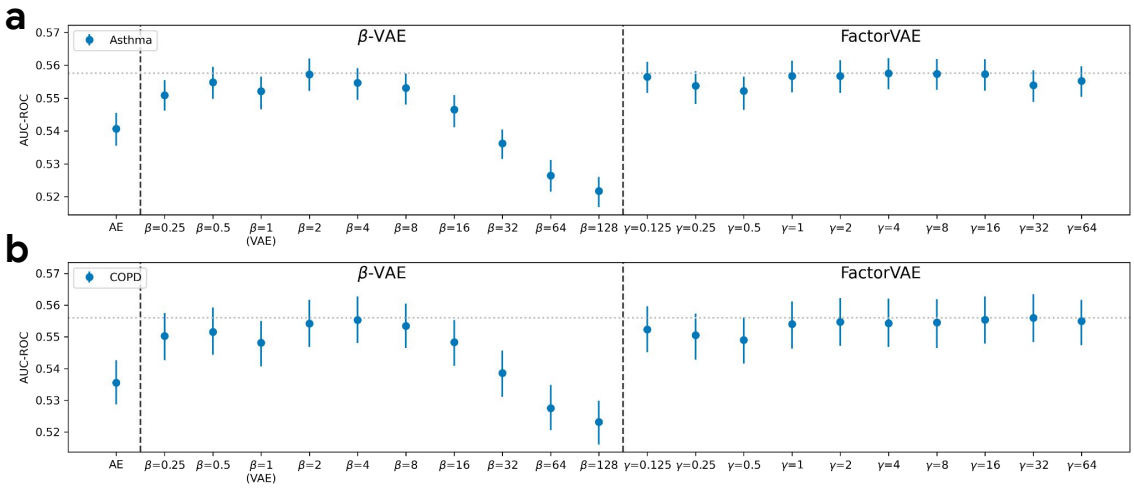

Figure 3. Polygenic risk score performance for asthma and COPD. We vary β or γ parameters in β-VAE and FactorVAE, which controls the strength of "regularization". Horizontal dotted line indicates the highest value. AE = (non-variational) autoencoder. **(a)** PRS performance in terms of AUC-ROC for asthma classification. **(b)** PRS performance in terms of AUC-ROC for COPD classification.

3. Experiments

We used spirograms in UK Biobank (Sudlow et al., 2015; Bycroft et al., 2018), following the preprocessing steps and the dataset splits (train, validation, and PRS evaluation sets) described in (Yun et al., 2023; Cosentino et al., 2023).

We compared convolutional (non-variational) autoencoders, VAEs, β-VAEs, and FactorVAEs with identical encoder and decoder architecture (Appendix A). We varied the β parameter in β-VAE and the γ parameter in FactorVAE such that $\beta \in \{2^i \mid -2 \leq i \leq 7,\ i \in \mathbb{Z}\}$ and $\gamma \in \{2^i \mid -3 \leq i \leq 6,\ i \in \mathbb{Z}\}$. Roughly speaking, the β and γ values control the extent to which the loss function in β-VAE and FactorVAE penalizes the entangled coordinates, higher values implying more penalization. The latent dimension was fixed to 5 to match (Yun et al., 2023) for direct comparison.

Two main metrics were computed to compare the models, first without using the genetic data: the reconstruction error and the pairwise correlations of latent coordinates, both in the validation set (Figure 1). The reconstruct error was defined by the mean-squared error between the reconstructed spirogram and the original input spirogram. We also computed "mean absolute correlation" in the validation set defined by $\frac{\sum_{i<j} |\mathrm{Corr}(\mathbf{z}_i, \mathbf{z}_j)|}{n(n-1)/2}$ for the latent vector $[\mathbf{z}_i]_{i=1}^n$. We trained each model 10 times with different random seeds to compute 95% confidence intervals ($\pm 1.96 \times$ standard error). As also observed in (Kim & Mnih, 2018), FactorVAEs retained good reconstruction across different values of the regularization hyperparameter γ, while β-VAE's reconstruction error significantly increased when β was high (Figure 1a, c). Notably, the pairwise correlation metric also increased with very high values of β or γ (Figure 1b, d), implying less disentangled coordinates. For all remaining genetic analysis, we chose the trained model with the median reconstruction error in the validation set (among identical models trained with different random seeds), as a "typical" example of the representation generated in a similar setup.

We performed GWAS on all 5 coordinates of the learned representations using linear mixed model association testing implemented by BOLT-LMM (Loh et al., 2015). Using the stratified linkage disequilibrium score regression (LDSC) (Bulik-Sullivan et al., 2015), we observed high heritability (h_g^2 up to 27%) of learned coordinates, especially from FactorVAEs (Figure 2b, Table 1). The intercept term from LDSC was close to 1, indicating minimal confounding (Table 1).

Similarly to (Yun et al., 2023), we compared our genome-wide significant (GWS) loci to the largest known lung function GWAS (Shrine et al., 2023) and all lung function-related loci found in the NHGRI-EBI GWAS Catalog (Sollis et al., 2023) (Appendix B). The independent GWS loci (linkage disequilibrium $R^2 \leq 0.1$ and $P \leq 5 \times 10^{-8}$) were defined by merging GWAS hits within 250kb together. We found that the encodings from FactorVAE generally produced more "known" and "novel" GWS loci compared to a plain VAE (i.e. β-VAE with $\beta = 1$), while the number of GWS loci varied significantly for β-VAE across different values of β (Figure 2a).

Finally, using the REGLE framework introduced in (Yun

Table 1. Estimated heritability and intercept from LD Score regression. The second column represents the β values for β-VAE and the γ values for FactorVAE. The min, max, and mean is taken across five learned latent coordinates from each model. h_g^2 = heritability; b = intercept.

MODEL	β OR γ	MIN(h_g^2)	MAX(h_g^2)	MEAN(h_g^2)	MIN(b)	MAX(b)	MEAN(b)
AE	-	0.0531	0.1598	0.0971	1.0005	1.0382	1.0143
β-VAE	0.25	0.0390	0.2239	0.1251	1.0053	1.0444	1.0275
	0.5	0.0410	0.2463	0.1311	1.0046	1.0466	1.0300
	1	0.0417	0.2515	0.1310	1.0051	1.0447	1.0283
	2	0.0411	0.2628	0.1333	1.0026	1.0549	1.0303
	4	0.0387	0.2582	0.1292	1.0031	1.0541	1.0262
	8	0.0113	0.2536	0.1200	1.0013	1.0557	1.0277
	16	0.0388	0.2428	0.1252	0.9874	1.0549	1.0156
	32	0.0347	0.1931	0.1052	1.0022	1.0548	1.0263
	64	0.0159	0.1866	0.0580	0.9966	1.0446	1.0146
	128	0.0165	0.1865	0.0545	0.9979	1.0449	1.0101
FACTORVAE	0.125	0.0409	0.2507	0.1325	1.0048	1.0461	1.0300
	0.25	0.0421	0.2526	0.1310	1.0023	1.0476	1.0291
	0.5	0.0414	0.2583	0.1331	1.0037	1.0543	1.0306
	1	0.0418	0.2597	0.1346	1.0026	1.0499	1.0306
	2	0.0408	0.2589	0.1356	1.0042	1.0532	1.0313
	4	0.0397	0.2547	0.1367	1.0055	1.0489	1.0308
	8	0.0407	0.2661	0.1372	1.0019	1.0583	1.0331
	16	0.0394	0.2657	0.1401	1.0034	1.0651	1.0331
	32	0.0382	0.2638	0.1370	1.0040	1.0614	1.0328
	64	0.0389	0.2647	0.1386	1.0041	1.0611	1.0349

et al., 2023), we generated PRSs specifically for two common lung-related diseases, asthma and chronic obstructive pulmonary disease (COPD), as a linear combination of 5 PRSs on learned latent coordinates. The weights of each latent coordinate PRS were learned using asthma and COPD labels in UK Biobank. As discussed in (Yun et al., 2023), one can use a small number of labels (as few as hundreds) to learn the weights. Comparing PRS prediction performance in terms of AUC-ROC in a held-out PRS evaluation set, we observed that PRSs generated by FactorVAEs consistently outperform PRS generated by a plain VAE or a non-variational autoencoder (Figure 3). β-VAE also outperformed the plain VAE with some values of β, but the performance was highly dependent on which β value was chosen.

4. Discussion

In this work, we performed comprehensive evaluation of multiple VAE-based methods for generating unsupervised disentangled representation, in the context of genomic discovery and polygenic risk prediction using high-dimensional clinical data (HDCD). To the best of our knowledge, this is the first such evaluation.

We studied how the different loss functions and the regularization hyperparameters in two types of VAE extensions (β-VAE and FactorVAE) affect the ability to capture genetic components of the biological function encoded in spirograms, in addition to the reconstruction quality and the measure of disentanglement between coordinates. FactorVAEs were able to control the reconstruction error better, showed consistent ability to capture highly heritable coordinates, and discovered more genome-wide significant loci and more of the known loci than a standard VAE. While β-VAE can also perform better than a standard VAE for certain values of β, the performance was highly sensitive to the choice of β. We observed the same patterns in PRS performance evaluation of asthma and chronic obstructive pulmonary disease based on the REGLE framework in (Yun et al., 2023).

We believe that unsupervised learning of HDCD is a promising direction for genetic studies and that high-performance representation learning methods will enhance our understanding of human genetics.

Software and Data

Code available in `regle` directory at github.com/Google-Health/genomics-research.

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 65275. We thank Farhad Hormozdiari, Cory Y. McLean, and Zachary R. McCaw for helpful discussions and comments.

References

Alipanahi, B., Hormozdiari, F., Behsaz, B., Cosentino, J., McCaw, Z. R., Schorsch, E., Sculley, D., Dorfman, E. H., Foster, P. J., Peng, L. H., Phene, S., Hammel, N., Carroll, A., Khawaja, A. P., and McLean, C. Y. Large-scale machine-learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology. *Am. J. Hum. Genet.*, 108(7):1217–1230, July 2021.

Aung, N., Vargas, J. D., Yang, C., Fung, K., Sanghvi, M. M., Piechnik, S. K., Neubauer, S., Manichaikul, A., Rotter, J. I., Taylor, K. D., Lima, J. A. C., Bluemke, D. A., Kawut, S. M., Petersen, S. E., and Munroe, P. B. Genome-wide association analysis reveals insights into the genetic architecture of right ventricular structure and function. *Nat. Genet.*, pp. 1–9, June 2022.

Bai, W., Suzuki, H., Huang, J., Francis, C., Wang, S., Tarroni, G., Guitton, F., Aung, N., Fung, K., Petersen, S. E., Piechnik, S. K., Neubauer, S., Evangelou, E., Dehghan, A., O'Regan, D. P., Wilkins, M. R., Guo, Y., Matthews, P. M., and Rueckert, D. A population-based phenome-wide association study of cardiac and aortic structure and function. *Nat. Med.*, 26(10):1654–1662, October 2020.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, 47(3):291–295, March 2015.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in β-VAE. April 2018. doi: 10.48550/arXiv.1804.03599. URL https://arxiv.org/abs/1804.03599.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., and Marchini, J. The UK biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, October 2018.

Chen, R. T. Q., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/1ee3dfcd8a0645a25a35977997223d22-Paper.pdf.

Cosentino, J., Behsaz, B., Alipanahi, B., McCaw, Z. R., Hill, D., Schwantes-An, T.-H., Lai, D., Carroll, A., Hobbs, B. D., Cho, M. H., McLean, C. Y., and Hormozdiari, F. Inference of chronic obstructive pulmonary disease with deep learning on raw spirograms identifies new genetic loci and improves risk models. *Nat. Genet.*, April 2023.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Kim, H. and Mnih, A. Disentangling by factorising. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2649–2658. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18b.html.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. URL https://openreview.net/forum?id=33X9fd2-9FyZd.

Kumar, A., Sattigeri, P., and Balakrishnan, A. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1kG7GZAW.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 2019.

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., Patterson, N., and Price, A. L. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.*, 47(3):284–290, March 2015.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inf. Theory*, 56(11):5847–5861, November 2010.

Pirruccello, J. P., Di Achille, P., Nauffal, V., Nekoui, M., Friedman, S. F., Klarqvist, M. D. R., Chaffin, M. D., Weng, L.-C., Cunningham, J. W., Khurshid, S., Roselli, C., Lin, H., Koyama, S., Ito, K., Kamatani, Y., Komuro, I., Jurgens, S. J., Benjamin, E. J., Batra, P., Natarajan, P., Ng, K., Hoffmann, U., Lubitz, S. A., Ho, J. E., Lindsay, M. E., Philippakis, A. A., and Ellinor, P. T. Genetic analysis of right heart structure and function in 40,000 people. *Nat. Genet.*, 54(6):792–803, June 2022.

Shrine, N., Izquierdo, A. G., Chen, J., Packer, R., Hall, R. J., Guyatt, A. L., Batini, C., Thompson, R. J., Pavuluri, C., Malik, V., Hobbs, B. D., Moll, M., Kim, W., Tal-Singer, R., Bakke, P., Fawcett, K. A., John, C., Coley, K., Piga, N. N., Pozarickij, A., Lin, K., Millwood, I. Y., Chen, Z., Li, L., China Kadoorie Biobank Collaborative Group, Wijnant, S. R. A., Lahousse, L., Brusselle, G., Uitterlinden, A. G., Manichaikul, A., Oelsner, E. C., Rich, S. S., Barr, R. G., Kerr, S. M., Vitart, V., Brown, M. R., Wielscher, M., Imboden, M., Jeong, A., Bartz, T. M., Gharib, S. A., Flexeder, C., Karrasch, S., Gieger, C., Peters, A., Stubbe, B., Hu, X., Ortega, V. E., Meyers, D. A., Bleecker, E. R., Gabriel, S. B., Gupta, N., Smith, A. V., Luan, J., Zhao, J.-H., Hansen, A. F., Langhammer, A., Willer, C., Bhatta, L., Porteous, D., Smith, B. H., Campbell, A., Sofer, T., Lee, J., Daviglus, M. L., Yu, B., Lim, E., Xu, H., O'Connor, G. T., Thareja, G., Albagha, O. M. E., Qatar Genome Program Research (QGPR) Consortium, Suhre, K., Granell, R., Faquih, T. O., Hiemstra, P. S., Slats, A. M., Mullin, B. H., Hui, J., James, A., Beilby, J., Patasova, K., Hysi, P., Koskela, J. T., Wyss, A. B., Jin, J., Sikdar, S., Lee, M., May-Wilson, S., Pirastu, N., Kentistou, K. A., Joshi, P. K., Timmers, P. R. H. J., Williams, A. T., Free, R. C., Wang, X., Morrison, J. L., Gilliland, F. D., Chen, Z., Wang, C. A., Foong, R. E., Harris, S. E., Taylor, A., Redmond, P., Cook, J. P., Mahajan, A., Lind, L., Palviainen, T., Lehtimäki, T., Raitakari, O. T., Kaprio, J., Rantanen, T., Pietiläinen, K. H., Cox, S. R., Pennell, C. E., Hall, G. L., Gauderman, W. J., Brightling, C., Wilson, J. F., Vasankari, T., Laitinen, T., Salomaa, V., Mook-Kanamori, D. O., Timpson, N. J., Zeggini, E., Dupuis, J., Hayward, C., Brumpton, B., Langenberg, C., Weiss, S., Homuth, G., Schmidt, C. O., Probst-Hensch, N., Jarvelin, M.-R., Morrison, A. C., Polasek, O., Rudan, I., Lee, J.-H., Sayers, I., Rawlins, E. L., Dudbridge, F., Silverman, E. K., Strachan, D. P., Walters, R. G., Morris, A. P., London, S. J., Cho, M. H., Wain, L. V., Hall, I. P., and Tobin, M. D. Multi-ancestry genome-wide association analyses improve resolution of genes and pathways influencing lung function and chronic obstructive pulmonary disease risk. *Nat. Genet.*, 55(3):410–422, March 2023.

Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., Ibrahim, A., Ji, Y., John, S., Lewis, E., MacArthur, J. A. L., McMahon, A., Osumi-Sutherland, D., Panoutsopoulou, K., Pendlington, Z., Ramachandran, S., Stefancsik, R., Stewart, J., Whetzel, P., Wilson, R., Hindorff, L., Cunningham, F., Lambert, S. A., Inouye, M., Parkinson, H., and Harris, L. W. The NHGRI-EBI GWAS catalog: knowledgebase and deposition resource. *Nucleic Acids Res.*, 51(D1):D977–D985, January 2023.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.*, 12(3):e1001779, March 2015.

Sugiyama, M., Suzuki, T., and Kanamori, T. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Ann. Inst. Stat. Math.*, 64(5):1009–1044, October 2012.

Watanabe, S. Information theoretical analysis of multivariate correlation. *IBM J. Res. Dev.*, 4(1):66–82, January 1960.

Yun, T., Cosentino, J., Behsaz, B., McCaw, Z. R., Hill, D., Luben, R., Lai, D., Bates, J., Yang, H., Schwantes-An, T.-H., Khawaja, A. P., Carroll, A., Hobbs, B. D., Cho, M. H., McLean, C. Y., and Hormozdiari, F. Unsupervised representation learning improves genomic discovery for lung function and respiratory disease prediction. *medRxiv*, 2023. doi: 10.1101/2023.04.28.23289285. URL https://www.medrxiv.org/content/early/2023/04/29/2023.04.28.23289285.

A. Model architecture and training

The following model architecture was used for all VAE, β-VAE, and FactorVAE experiments. The input shape is (1000, 2) representing volume-time and flow-time spirograms. FactorVAE has an additional discriminator trained jointly with the encoder and the decoder. Same architecture was used for (non-variational) autoencoders, except they do not have the final Gaussian sampling step (unique to VAE) in the encoder part. All models were implemented in TensorFlow 2 and Keras, and trained with Adam optimizer for 100 epochs, with learning rate of 0.0001 and batch size of 32.

A.1. Encoder

Layer	Activation	Note
Conv1D	ReLU	filter=8, kernel=10, padding=same
MaxPooling1D	–	size=2
Conv1D	ReLU	filter=16, kernel=10, padding=same
MaxPooling1D	–	size=2
Conv1D	ReLU	filter=32, kernel=10, padding=same
MaxPooling1D	–	size=2
Flatten	–	–
Dense	ReLU	size=64
Dense	ReLU	size=64
Dense	ReLU	size=64
Dense and Sampling	–	size=5 for mean and variance each, and then sample from Gaussian.

A.2. Decoder

Layer	Activation	Note
Dense	ReLU	size=64
Dense	ReLU	size=64
Dense	ReLU	size=64
Dense	ReLU	size=4000
Reshape	–	shape=(125, 32)
UpSampling1D	–	size=2
Conv1DTranspose	ReLU	filter=16, kernel=10, padding=same
UpSampling1D	–	size=2
Conv1DTranspose	ReLU	filter=8, kernel=10, padding=same
UpSampling1D	–	size=2
Conv1DTranspose	ReLU	filter=2, kernel=10, padding=same

A.3. Discriminator for FactorVAE

Layer	Activation	Note
Dense	LeakyReLU	size=1000
Dense	LeakyReLU	size=1000
Dense	LeakyReLU	size=1000
Dense	LeakyReLU	size=1000
Dense	LeakyReLU	size=1000
Dense	LeakyReLU	size=1000
Dense	–	size=2
Softmax	–	–

B. GWAS Catalog lung function search

We used the following case-insensitive keywords to search for previously known lung-related GWAS loci in GWAS Catalog version `v1.0.2-associations_e106_r2022-07-09`: “asthma”, “chronic obstructive pulmonary disease”, “copd”, “expiratory flow”, “fev1”, “forced expiratory”, “forced vital capacity”, “lung function”.