JOINT CLINICAL AND MOLECULAR SUBTYPING OF COPD WITH VARIATIONAL AUTOENCODERS

PRELIMINARY DRAFT

Enrico Maiorino*
Channing Division of Network Medicine
Brigham and Women's Hospital
Harvard Medical School

Margherita De Marzio
Channing Division of Network Medicine
Brigham and Women's Hospital
Harvard Medical School

Scott T. Weiss
Channing Division of Network Medicine
Brigham and Women's Hospital
Harvard Medical School

Edwin K. Silverman
Channing Division of Network Medicine
Brigham and Women's Hospital
Harvard Medical School

Peter J. Castaldi
Channing Division of Network Medicine
Brigham and Women's Hospital
Harvard Medical School

Kimberly Glass
Channing Division of Network Medicine
Brigham and Women's Hospital
Harvard Medical School

August 18, 2023

ABSTRACT

Chronic Obstructive Pulmonary Disease (COPD) is a complex, heterogeneous disease. Traditional subtyping methods generally focus on either the clinical manifestations or the molecular endotypes of the disease, resulting in domain-specific classifications that may not capture its full complexity. Here, we introduce an integrative approach based on variational autoencoders to integrate clinical and blood gene expression data from the COPDGene cohort study. We generate Personalized Integrated Profiles (PIPs) that recapitulate the joint clinical and molecular state of each individual in the population. Through prediction experiments we show that the PIPs encode the complex disease state of each individual in a compact representation, with an accuracy comparable or better than other embedding approaches. Through these profiles we study the space of continuous variation of COPD features by using graph-based trajectory learning techniques, and delineate five well-separated subtypes. The identified subtypes exhibit distinct phenotypes, expression signatures, and disease outcomes. Overall, our findings show that integrating clinical and molecular data is beneficial for gaining a more comprehensive understanding of COPD heterogeneity.

Keywords

1 Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a complex chronic respiratory disease estimated to be among the leading causes of death worldwide [1]. The disease encompasses a spectrum of conditions, including persistent airflow obstruction, lung inflammation, chronic bronchitis, and emphysema, that lead to a progressive impairment of lung function. COPD susceptibility has been attributed to a combination of physical, environmental, and genetic factors, resulting in significant phenotypic variation across individuals.

*

This heterogeneity has prompted substantial research efforts to dissect the various manifestations of the disease, to understand their etiological origins, and to predict their outcomes [2]. A practical goal of these studies has been to delineate distinct COPD subtypes by employing advanced clustering and machine learning techniques trained with large datasets of clinical and genomic data extracted from human cohort studies [3, 4].

Current COPD subtyping approaches can be roughly divided in those that aim at characterizing the observable phenotypes of the disease (clinical subtyping), and those that focus on disentangling the disease processes, often referred to as "endotypes", underlying a COPD manifestation (molecular subtyping) [5]. Applications of the former approach are based on clinical evidence such as includes demographics, disease symptoms, spirometry measurements, or chest imaging data [6, 7, 8, 9], whereas applications of the latter approach are based on measurements from multi-omics assays (transcriptomics, proteomics, epigenomics, etc.) [10, 11, 5, 12]. Although both types of approaches offer valuable insights into different aspects of the disease, these applications are exclusively defined within either the clinical or molecular domain, with the analysis in the other domain performed primarily for validation or post-hoc examination [13, 6, 8, 11, 5]. Consequently, these domain-specific classifications cannot capture disease mechanisms arising from the interaction between genetic, clinical, and lifestyle factors [14, 15, 16, 17], potentially leading to inconsistent subtypes that are not reproducible across different patient cohorts [18].

While several methods exist for omics data integration, the simultaneous integration of both clinical and omics data for disease subtyping is relatively unexplored, with its applications in COPD being confined to specific domains [19]. One of the challenges in integrating omics and non-omics data derive is due to their inherent complexity. Data heterogeneity and bias, already present in multi-omics studies [20], are exacerbated when including clinical data, which is typically composed of complex data structures with heterogeneous correlation patterns and significant variation in terms of scales, sparsity, and noise [21]. To alleviate these issues, and to account for potential nonlinear interactions across variables from different domains, specialized integrative methodologies based on autoencoder neural networks have been proposed in several disease contexts, including COPD [10] and cancer [22, 23, 24, 25]. However, a comprehensive subtyping that integrates both the clinical and molecular domains of COPD has not been conducted to date.

Here, we perform a joint subtyping in a comprehensive space encompassing different types of clinical variables together with gene expression data extracted from the COPDGene cohort [26] (see Fig. 1 (a)), a large population of current and former smokers with COPD. Building on recent developments on multi-modal learning [27, 28], we propose an integrative method based on variational autoencoders (VAEs). A VAE is an unsupervised neural network architecture designed to compress the input data and generate a set of compact encodings. We trained the VAE with clinical and expression data from COPDGene, generating a set of personalized integrated profiles (PIPs) that encode the joint clinical and molecular configuration of each individual in the population. By performing multiple outcome prediction experiments, we demonstrate that the generated PIPs are highly informative of the disease state and enable accurate prediction of future disease outcomes. Next, we map the continuous trajectories in the VAE space using a recently proposed trajectory learning technique [29]. Through this approach, we identify several well-separated classes of disease states, each exhibiting distinct clinical and molecular characteristics (joint subtypes). Finally we show that these joint subtypes are characterized by different mortality outcomes and that they are robust to resampling noise.

2 Results

COPDGene is an ongoing longitudinal multi-center study of 10,198 individuals with and without COPD who have undergone extensive profiling at three time points across 10 years (Phase 1, 2, 3). This study includes measurements of phenotypic variables and multi-omics data, making it one of the largest and most comprehensive cohort studies of COPD to date. In this work we considered all the subjects with clinical and blood gene expression data in Phase 2 of the study that were available at the time of analysis (3,628 subjects, see Fig. 1 (b)). Clinical data includes demographics, lifestyle factors (e.g. smoking behavior), spirometry measurements, medical and medication history, chest imaging measures, symptoms, and complete blood counts (CBC). Expression data consists of whole blood RNA-seq samples. Both these data modalities have been used extensively for COPD subtyping [6, 7, 8, 9, 11, 5], and they are among the most widely-used read-outs of the phenotypic and molecular manifestations of COPD. To merge these two data types in a single representation of disease state, we designed a data integration scheme based on Variational Autoencoders (VAEs).

Variational Autoencoders (VAEs) are probabilistic unsupervised neural network models designed to compress the input to generate low dimensional representations [30]. VAEs exploit statistical dependencies between input variables to construct a small set of variables (latent code) that preserve most of the input information. In contrast to linear techniques such as Principal Component Analysis (PCA), VAEs can capture nonlinear relationships between variables. As such, in situations where nonlinear behaviors are prevalent, VAEs produce more informative representations for downstream analysis tasks such as clustering and visualization [31]. Drawing inspiration from other recently proposed

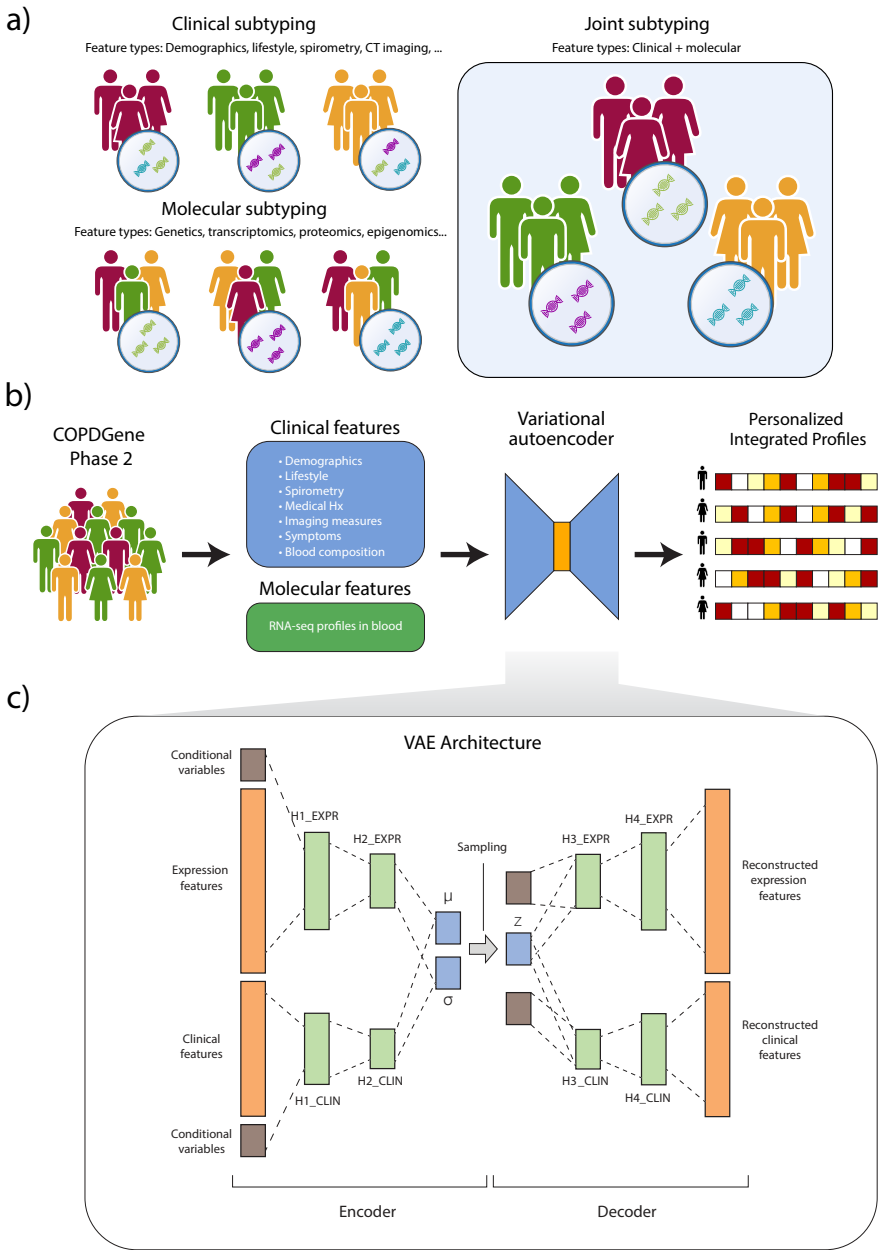

Figure 1: Overall organization of this work. (a) Joint subtyping aims at generating subtypes that are uniform in terms of both clinical and molecular features; (b) starting from the COPDGene population, an array of clinical and molecular features are extracted and used for training the VAE. The autoencoders produces a set of personalized integrated profiles (PIPs), one per individual in the population; (c) architecture of the VAE. Expression and clinical features are concatenated with conditional variables (age, sex, race) and processed separately in two encoder subnetworks. The two encodings are subsequently merged in a shared latent representation (PIP). Next, the latent representation is concatenated to the conditional variables and processed with two separate decoder subnetworks to reconstruct the original data.

architectures for multi-modal data integration [23, 24], we modified the standard VAE architecture so as to process and merge two data types (see Fig. 1 (c) and Methods). Our model implicitly performs a 2-step dimensionality reduction. The hidden layers in the encoder network (H1/2-expr, H1/2-clin in Fig. 1 (c)) process the two data types separately to obtain a data-type-specific representation of the input features. These representations are then coupled to generate a joint latent representation that encodes both the clinical and molecular information, which we refer to as "personalized integrated profile" (PIP). Furthermore, given the probabilistic nature of the VAE model, it is possible to correct for

potential confounding factors by including them as a set of conditional variables. These conditional variables have a similar role as covariates in linear regression modeling. Therefore, we set the age, gender and race of each individual as conditional variables in both data modalities to regress out their effect on the learned representation.

The data processing and integration pipeline consists of several steps (see Methods). In brief, we designed two separate processing pipelines for the two data types, consisting of feature selection and normalization operations. Next, we split the dataset into training and validation sets (80%/20%). We trained the VAE on the training set, and used the validation set for hyperparameter selection. We performed hyperparameter optimization for finding the optimal architecture parameters, including layer size, learning rate and number of components of PIP vectors. We determined that the optimal number of components in terms of reconstruction quality and latent vector size is 30. After training the network with the optimal parameters, we then used the Encoder subnetwork to generate the PIPs of the full dataset. The generated PIPs are the starting point of the subsequent analysis to identify joint subtypes of COPD.

2.1 The Personalized Integrated Profiles allow accurate prediction of future COPD states

The PIPs generated by VAE contain information on both the expression and clinical features of a patient. To test the VAE's performance in compressing and integrating different data modalities without sacrificing important information, we set up a prediction task of several prospective COPD outcomes. The COPD outcomes included all-cause mortality at 3 and 5 years and other clinical measurements collected in Phase 3 (P3) of the study, approximately five years after Phase 2 (P2) visits (see Methods for further details). Note that these variables are extracted from Phase 3 and long-term follow-up (LFU) data and therefore they are not encoded in the VAE training data. We perform the classification using an off-the-shelf classification model, the Balanced Random Forest classifier, and define the PIP vectors as input variables and the Phase 3 outcomes as target variables (see Methods for details). For comparison, we evaluated the performance of the same classifier trained using other integrated representations as input: PCA of clinical variables (Clin PCA); PCA of expression variables (Expr PCA); PCA of concatenated expression and clinical variables (Expr + Clin PCA); CCA of expression variables (Expr CCA); CCA of clinical variables (Clin CCA); and factors calculated by applying the integrative method MOFA [32] to expression and clinical variables (MOFA). The performance values of the prediction are reported in Table 1 and Supplementary Table FIG are calculated through a stratified 5-fold validation executed 3 times and averaged. The VAE-based PIPs are among the highest scoring or second-highest scoring across most prediction tasks. The prediction performance of the VAE representations is comparable to the prediction performance obtained by using the clinical variables alone, which are the most relevant variables for predicting the disease outcomes in Phase 3. This finding suggests that that the PIPs preserve substantial information about an individual's disease state, allowing for accurate prediction of future COPD outcomes.

	F1-score						
	Clin PCA	Expr PCA	Expr+Clin PCA	Expr CCA	Clin CCA	MOFA	VAE
Δ FEV_1%pred. (P3<P2)	0.75 (0.04)	0.74 (0.04)	0.73 (0.04)	**0.75 (0.04)** [n.s.]	0.74 (0.04)	0.74 (0.04)	0.73 (0.04)
Inc. bronchitis (P3)	0.02 (0.04)	0.02 (0.04)	0.06 (0.07)	0.03 (0.05)	0.04 (0.05)	0.07 (0.07)	**0.12 (0.08)** **
Exacerbations (P3)	0.37 (0.08)	0.10 (0.06)	0.34 (0.07)	0.23 (0.07)	0.23 (0.08)	0.25 (0.09)	**0.41 (0.08)** *
Δ Exac. Freq. (P3>P2)	0.10 (0.07)	0.04 (0.05)	0.17 (0.09)	0.12 (0.07)	0.12 (0.08)	0.09 (0.06)	**0.20 (0.11)** [n.s.]
Sev. Exacerbations (P3)	0.10 (0.07)	0.03 (0.04)	0.12 (0.08)	0.04 (0.05)	0.03 (0.04)	0.06 (0.06)	**0.15 (0.08)** [n.s.]
Δ Sev. Exacerbations (P3>P2)	0.07 (0.07)	0.03 (0.05)	0.05 (0.08)	0.00 (0.02)	0.01 (0.03)	0.04 (0.06)	**0.11 (0.07)** *
Δ MMRC (P3>P2)	**0.29 (0.05)** **	0.18 (0.05)	0.18 (0.05)	0.24 (0.04)	0.22 (0.05)	0.20 (0.05)	0.26 (0.04)
Δ SF-36 (P3<P2)	0.27 (0.05)	0.22 (0.04)	0.23 (0.05)	0.27 (0.05)	0.26 (0.04)	0.24 (0.04)	**0.28 (0.05)** [n.s.]
Δ Emph.% (P3>P2)	0.86 (0.02)	0.89 (0.02)	0.88 (0.01)	0.89 (0.02)	**0.90 (0.02)** [n.s.]	0.87 (0.02)	0.88 (0.02)
Δ %GasTrap (P3>P2)	0.85 (0.02)	0.87 (0.02)	0.86 (0.02)	0.88 (0.02)	**0.88 (0.02)** [n.s.]	0.85 (0.03)	0.86 (0.02)
Mortality (3yr)	0.08 (0.05)	0.13 (0.05)	0.10 (0.04)	0.13 (0.06)	0.13 (0.06)	0.10 (0.03)	**0.17 (0.05)** *
Mortality (5yr)	0.23 (0.05)	0.18 (0.04)	0.23 (0.05)	0.25 (0.05)	0.25 (0.04)	0.20 (0.04)	**0.30 (0.06)** **

Table 1: Prediction performance of COPD outcomes. Best performances and second-best are respectively displayed in bold or underlined. Abbreviations: pred.=predicted, inc.=incident, exac.=exacerbations, freq.=frequency, emph.=emphysema, GasTrap=gas trapping.

2.2 Summarizing cohort heterogeneity with principal graphs reveals COPD subtypes with distinct clinical and transcriptomic signatures

The PIPs generated through VAE are vectors distributed in a 30-dimensional space of variables that implicitly describe the joint molecular and clinical characterization of each individual in COPDGene. The whole population is thus distributed according to a specific geometry. Growing evidence suggests that COPD manifestations may form a continuous spectrum of disease states[3]. The lack of clear separation between different COPD subtypes is problematic

when one seeks a robust characterization of different disease mechanisms. To overcome this issue, we analyzed the intrinsic trajectories of continuous variation within individuals in the COPDGene population. To do so, we used elPiGraph [29], an algorithm to fit a branching network structure, called *principal graph*, to a set of points in a multidimensional space. In brief, elPiGraph constructs a tree-like graph whose nodes and edges minimize the distance from the points, resulting in a mapping between each data point and a position along the tree branches. Through these projections, it is therefore possible to approximate the underlying geometry and proximity relations between points in a multidimensional space. We applied elPiGraph to construct the principal graph of the COPDGene population, and associated each subject to their closest network branch in the space. The fitted principal graph is composed by 5 terminal branches, i.e. tree branches connected to the remaining graph only through one endpoint, and 2 non-terminal branches, i.e. those connected in both their endpoints (see Fig. 2 (a)).

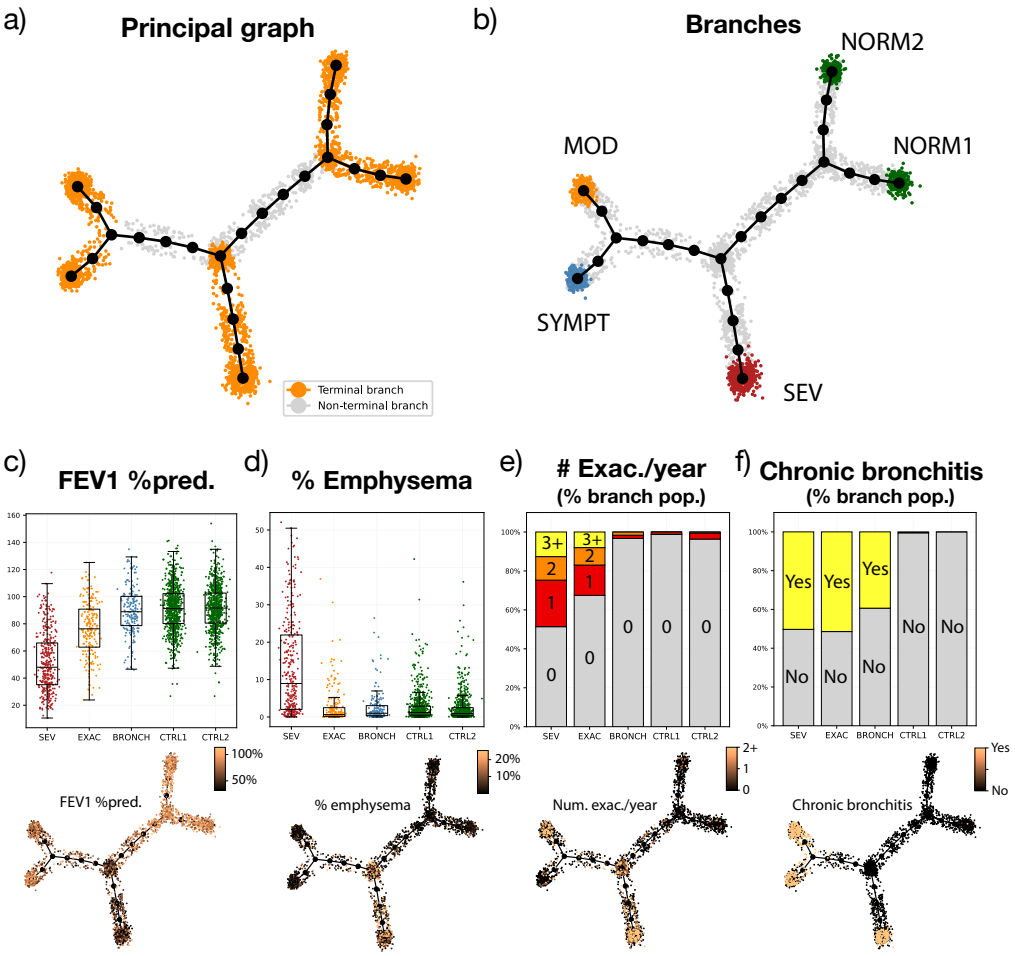

Figure 2: Features of the principal graph constructed from the COPDGene PIPs. (a) Principal graph layout. Large black dots and edges represent the fitted graph structure. Smaller dots represent individuals in the population, where the position is determined according to their proximity to the graph nodes. Points highlighted in orange are members of the terminal branches, while gray points are in non-terminal branches; (b) colored dots represent the individuals selected as meember of each branch after thresholding. Gray dots are individuals that are not assigned to any branch; (c-f) Distribution of selected features in each branch (top) and across the principal graph (bottom). Abbreviations: Pred.=predicted,Exac.=Exacerbations

Of the 5 terminal branches, 3 are composed mainly of individuals with GOLD stage 2-4 corresponding to moderate-to-severe COPD (bottom left branches in Fig. 2 (b)), whereas 2 branches are mainly associated to individuals with mild COPD or without COPD, i.e. GOLD stage equal to 0-1 (top right branches in Fig. 2 (b)). Subjects in non-terminal branches are composed of intermediate disease states that combine the features of the adjacent terminal branches. Since we were interested in identifying subtypes with distinct disease features and minimal overlap, we restricted the analysis

exclusively to the individuals within the terminal branches. Furthermore, in order to maximize the separation between different branches, we selected only the 50% data points that lie on the most extreme ends of each branch (Fig. 2 (c)). This threshold was chosen by calculating the pointwise purity of the obtained clusters for different thresholds and choosing the lowest threshold value that attained sufficient purity values for all the branches (see Supplementary Figures REF (a-f) and Methods).

2.3 Clinical characteristics of the joint subtypes

The 5 identified branches are characterized by different phenotypic and molecular features. The average values of a set of relevant clinical characteristics of each branch are summarized in Table 2. Owing to their distinct characteristics, we

	SEV	MOD	SYMPT	NORM1	NORM2
			Demographics		
Age	68.7 (8.3)	59.7 (6.3)	63.5 (8.0)	67.6 (8.2)	61.9 (8.5)
Sex	42.03%	64.15%	42.08%	42.36%	54.26%
Race	15.66%	48.58%	24.04%	10.84%	42.38%
BMI	29.2 (6.7)	30.2 (7.0)	28.1 (5.0)	27.6 (5.2)	27.1 (5.0)
			Spirometry		
FEV_1%	52.0 (22.1)	76.3 (20.6)	89.8 (17.7)	92.5 (16.4)	91.4 (15.9)
FEV_1/FVC	0.5 (0.2)	0.7 (0.1)	0.7 (0.1)	0.7 (0.1)	0.8 (0.1)
FVC%	73.8 (18.4)	85.6 (17.4)	94.2 (15.3)	95.1 (14.4)	93.4 (14.4)
FEF2575	0.7 (0.7)	1.5 (0.9)	2.1 (1.1)	2.0 (1.0)	2.2 (1.1)
BDR FEV_1	8.9 (12.4)	6.5 (14.1)	4.4 (6.9)	4.6 (6.8)	4.0 (7.6)
			CT imaging		
% Emph.	13.3 (13.0)	2.8 (5.1)	2.7 (4.3)	3.0 (4.1)	2.5 (4.2)
Pi10	2.8 (0.6)	2.5 (0.6)	2.1 (0.5)	2.0 (0.4)	2.0 (0.4)
AWT	1.2 (0.2)	1.1 (0.2)	1.0 (0.2)	1.0 (0.2)	0.9 (0.2)
% Gas trap.	38.0 (21.9)	15.2 (14.8)	13.9 (11.8)	15.1 (12.3)	12.7 (11.7)
			Symptoms		
MMRC	2.7 (1.2)	2.4 (1.3)	0.6 (0.9)	0.1 (0.5)	0.2 (0.6)
Exac.Freq.	0.9 (1.3)	0.6 (1.1)	0.0 (0.3)	0.0 (0.1)	0.0 (0.2)
Chron.Bronch.	50.27%	51.42%	39.34%	0.49%	0.00%
Dist Walk.	945.9 (438.1)	1207.4 (372.1)	1467.8 (354.4)	1527.8 (390.6)	1517.6 (317.8)
			SGRQ		
SGRQ Symptom	55.9 (19.3)	56.2 (21.7)	32.0 (13.8)	6.8 (9.5)	7.3 (10.3)
SGRQ Active	66.9 (21.9)	62.0 (25.4)	23.9 (20.2)	10.6 (15.2)	10.1 (15.8)
SGRQ Impact	36.2 (20.5)	34.2 (20.0)	8.9 (8.7)	2.0 (3.9)	2.8 (6.6)
SGRQ Total	48.7 (18.6)	46.2 (19.7)	16.9 (10.6)	5.4 (6.9)	5.8 (8.6)
			Smoking		
Smoker	31.32%	75.00%	69.95%	17.98%	44.16%
ATS PackYears	56.3 (28.1)	45.0 (25.0)	47.2 (24.5)	37.6 (21.0)	36.4 (19.5)
			Complete blood counts		
Neutr.%	70.3 (8.5)	51.4 (9.6)	59.4 (8.2)	66.3 (7.1)	49.2 (9.5)
Lymph.%	18.8 (6.7)	37.8 (8.7)	29.3 (8.2)	22.1 (5.9)	39.8 (9.1)
Eosin.%	2.5 (2.1)	2.7 (1.9)	2.5 (1.8)	2.5 (1.8)	2.6 (1.7)
Platelets	241.5 (74.7)	237.9 (65.2)	235.0 (65.5)	232.5 (62.9)	234.0 (61.7)

Table 2: Average values of selected clinical characteristics, categorized by branch. The values in parenthesis describe the standard deviations. Abbreviations: BMI=body mass index; FEV_1=forced expiratory volume in 1 second; FVC=forced vital capacity; FEF2575=forced mid-expiratory flow (25-75%); BDR=Bronchodilator response; emph.=emphysema; AWT=airway wall thickness; gas trap.=gas trapping; mMRC=Modified Medical Research Council Dyspnea Scale; exac.freq.=exacerbation frequency; chron. bronch.=chronic bronchitis; dist walk.=distance walked (ft) in a 6 minute walk test; SGRQ=St George's Respiratory Questionnaire; neutr.=neutrophils; lymph=lymphocytes; eosin.=eosinophils.

defined brief identifiers for each branch (Fig. 2 (a) and (b)). Two branches, "NORM1" and "NORM2" are composed mainly of subjects that do not show substantial COPD features. They mainly differ in terms of their average age (67 for

NORM1 against 61 for NORM2), current smoking status (18% for NORM1 and 44% for NORM2), and blood cell type composition (prevalently neutrophilic in NORM1 and lymphocytic in NORM2). The "SEV" branch consists of older individuals suffering from the most severe form of the disease, with greatly compromised lung function (as low as 52% of predicted FEV_1), frequent and severe exacerbations, and high prevalence of chronic bronchitis (Figs. 2 (c-f)). Additionally, subjects categorized in SEV had a marked degree of emphysema, air trapping, and thickened airway walls. Consistent with the severity of their condition, self-reported metrics such as the mMRC dyspnoea scale and SGRQ scores indicate a severely compromised quality of life. Despite the high average number of pack years, the SEV branch is characterized by the smallest proportion of current smokers (31%) compared to other branches. Note that this trend is likely influenced by the lifestyle modifications that are typically recommended to individuals suffering from severe respiratory conditions. The "MOD" branch is composed by younger individuals with moderately impaired lung function and low percentage of emphysema and moderate airway wall thickening. However, compared to other branches (excluding SEV), MOD subjects are affected by a higher average frequency of exacerbations, and the severity of their symptoms is similar to those observed in the SEV branch, with an average mMRC score of approximately 2.4 compared to about 2.7 in the SEV branch, and an SGRQ total score of approximately 46.2 versus 48.7. Furthermore, the MOD branch is associated to the highest average BMI compared to the other branches. Individuals in the "SYMPT" branch are instead similar in most aspects to NORM1 and NORM2 branches (mild airway obstruction, low percent emphysema, infrequent exacerbations), yet this group has a larger proportion of current smokers with airway inflammation symptoms, such as cough, phlegm, and chronic bronchitis.
Next, we examined the relation between the five branches and other previously proposed COPD subtypes. As emphysema represents a hallmark feature of COPD manifestations, recent works have distinguished between emphysema-predominant (EPD, defined in individuals with GOLD>1 as CT-quantified densitometric emphysema >10% at -950 Hounsfield units), non-emphysema-predominant (NEPD, CT emph. <5%) phenotypes and intermediate emphysema (IE, CT emph. between 5% and 10%) [33]. The breakdown of each branch into separate classes shows that the SEV branch contains the largest fraction of EPD individuals (46%, see Supplementary Figure TODO[2]), a smaller fraction of NEPD individuals (23%), and negligible parts of other states. Comparatively, the MOD branch is composed by a substantial proportion of NEPD individuals (22% overall, 62% within GOLD>1 stages) and only 8% of EPD phenotypes. Furthermore, MOD has the largest proportion (20%) of individuals with preserved FEV/FVC ratio and impaired spirometry (PRISm) [34]. As seen above, the remaining three branches (SYMPT,NORM1,NORM2) are composed mainly of individuals without significant COPD features and therefore contain negligible proportions of GOLD>1 cases.

2.4 The joint subtypes exhibit distinct transcriptomic signatures and pathway activations

We analyzed the transcriptomic differences between the SEV,MOD and SYMPT branches compared to a subpopulation composed of the NORM1 and NORM2 branches (see Methods). Through differential expression (DE) analysis, we identified a set of DE genes for each contrast (see Supplementary Table REF). For each set, we performed gene set enrichment analysis (GSEA) [35] using the 50 hallmark pathways of MSigDB [36], to find the over- or under-expressed biological pathways in each branch compared to NORM1/2. The results are shown in Fig. 3. Among the most significant pathways (FDR p-adj.<0.05), Interferon Alpha (IFN-α) Response is highly over-expressed in the SEV and SYMPT branches and under-expressed in the MOD branch. Interestingly, COPD patients who experience frequent exacerbations exhibit reduced IFN-α levels in response to viral infection [37], in contrast to individuals with lower exacerbation rates. As such, downregulation of the IFN-α in the MOD branch might indicate compromised airway immunity, potentially leading to a higher susceptibility to exacerbations. Additionally, the Oxidative Phosporylation pathway is upregulated in both the SEV and MOD branches, while the Reactive Oxygen Species (ROS) pathway is upregulated only in the SEV branch and downregulated in the MOD branch. The majority of GSEA leading genes in the MOD branch are antioxidant agents. Among these, differential expression analysis reveals downregulation of the antioxidant enzymes GPX3, G6PD, GSR, and TXNRD2 [38, 39]. ROS overproduction is known to suppress the activity of these enzymes [40, 41, 42, 43] and prolonged depletion of antioxidant capacity has been observed several days after the onset of exacerbation in COPD patients. In agreement with these reports, downregulation of these genes in the MOD branch might reflect loss in antioxidant defenses. The IL6/JAK/STAT3 signaling pathway was also statistically significant. Previous reports have supported the role of this pathway in pulmonary inflammation and COPD severity [44, 45]. Notably, increased gene expression levels of IL6ST have been observed in COPD patients with exacerbation history compared to COPD patients without known exacerbation history [46]. In support of these findings, IL6ST is highly upregulated in the MOD branch, supporting the specificity of our subtypes.

[2] ref, Fig. epd_nepd.pdf

Figure 3: Gene set enrichment analysis of the differentially expressed genes between each branch and the NORM1 and NORM2 branches. Opaque points represent pathways that are significant (FDR p.adj.<0.05), while transparent points are non significant. NES=normalized enrichment score.

2.5 The joint subtypes are associated to distinct disease outcomes and risks

Next, we evaluated the key clinical outcomes correlated with each subtype. We collected and synthesized data from Phase 3 as well the Long-Term Follow-Up (LFU) dataset, in which COPDGene participants self-reported health updates via a survey every six months during the whole duration of the study (see Methods). Given that exacerbations represent a key indicator of COPD progression, we examined their temporal patterns among subjects from various branches. The temporal exacerbation patterns of the branches closely mirror the cross-sectional behavior observed during Phase 2 (Fig. 4 (a)), with the individuals in the SEV and MOD branches experiencing frequent exacerbations throughout the entire time period. Interestingly, individuals who reported zero exacerbations in Phase 2, yet were classified in the SEV (163) and MOD (124) branches, demonstrated a significantly higher likelihood of experiencing one or more exacerbation events following Phase 2 — 42% and 50%, respectively — compared to the NORM1 and NORM2 branches (8%) and the SYMPT branch (16%). This finding suggests that branch membership can provide insights into the potential of future exacerbations, even when the present data does not explicitly indicate it. Moreover, to address potential confounding factors, we used a Poisson regression model to analyze the incidence of exacerbations in the LFU data, adjusting for age, sex, and race of each participant (see Methods). The calculated incidence rate ratios of each branch, using NORM2 as the reference, demonstrate that these insights are consistent even when adjusting for demographic differences between the subpopulations (Fig 4 (b)).

Next, we assessed the change in FEV1 and FEV1 % of predicted values over the five years between the Phase 2 and Phase 3 visits (Figs 4 (c) and (d)). While the SEV branch remained the most impacted in terms of (post-bronchodilator)

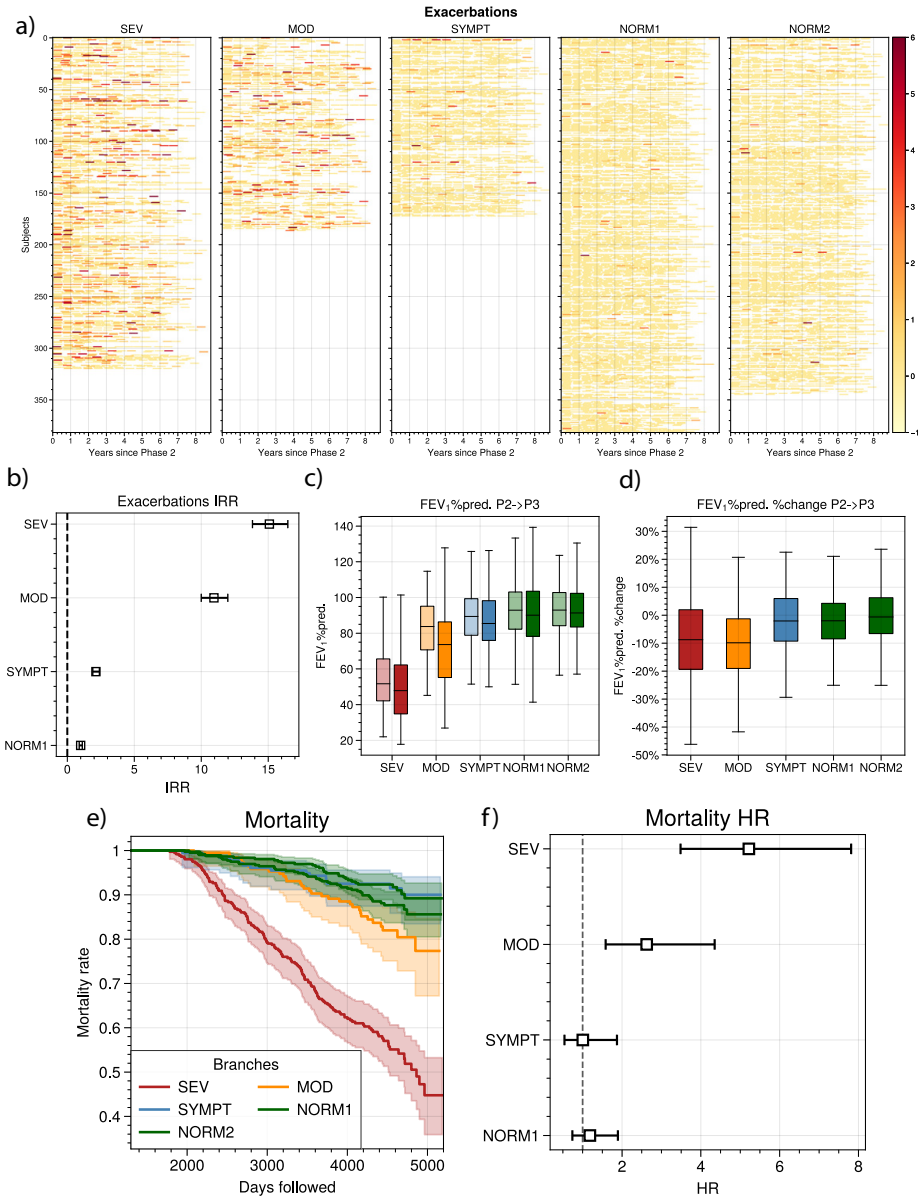

Figure 4: Prospective analysis of the branches. (a) Temporal trends of exacerbations among the population subjects, categorized by their respective branches. Each row represents a subject, and different colored areas within these rows denote the number of exacerbations experienced by that subject during a 6-month timeframe. The color intensity corresponds to the number of exacerbations, with darker or larger areas indicating a higher number of exacerbations; (b) incidence rate ratio (IRR) of exacerbations between each branch and the reference group, NORM2; (c) Distribution of FEV_1 % predicted levels categorized by branch (color), and by study phase (light=P2, dark=P3); (d) Distribution of relative changes in FEV_1 % predicted categorized by branch; (d) Kaplan Meier curves of mortality, categorized by branch; (d) hazard ratio (HR) of mortality between each branch and the reference group, NORM2.

FEV_1 % predicted, the MOD branch is characterized by a comparable percentage change between Phase 2 and Phase 3. Conversely, the other three branches do not display significant FEV_1 decline within the selected time frame.

Finally, we estimated the mortality risk associated to each branch(see Fig. 4 (e) and (f) and Methods). The SEV and MOD branches demonstrate respectively a 5-fold and 2-fold average hazard ratio compared to the NORM2 branch, while the other branches do not exhibit increased risk.

2.6 The joint subtypes are robust to retraining and data resampling

In order quantify the stability of the VAE space with respect to resampling of the training data, we performed two different robustness tests. In the first test, we evaluated the stability of the VAE vector embeddings (the PIPs) to random re-samplings of the training data. We retrained the VAE 100 times with random train/validation/test splits (80%/10%/10%) and generated 100 embeddings of the COPDGene population ("resampled" embeddings). Next, we estimated the similarity of the original embeddings with each resampled set of embeddings generated with different data splits. In order to estimate the similarity between two sets of vectors, we used the distance correlation measure (dCorr) [47]. dCorr is an extension of the Pearson correlation to multivariate settings and it ranges between 0, indicating statistical independence, and 1, suggesting a linear relationship between the two variables (see Methods for further details). We measured a distribution of correlation values of 0.900 ± 0.006, indicating a strong similarity between the generated spaces.

Next, we tested the stability of the branch assignments to random resamplings. For each of the resampled embeddings, we constructed a principal graph (using same settings) and assigned each point to a branch of the new graph. Next, we calculated the cluster purity of each of the original branches (SEV, MOD, SYMPT, NORM1 and NORM2), i.e. the proportion of subjects that were classified within the same branch in the resampled graph (see Methods). For comparison, we performed k-means clustering executed with the same number of clusters K and K' respectively in the original and resampled embeddings. To provide a fair comparison between the branch and the k-means assignments, for each cluster we selected a set of high confidence points, consisting of the 50% data points that are closest to the cluster centroid. Finally, we calculated the purity of the k-means clusters as the purity of their high-confidence sets. As shown in Supplementary Figure FIG (g), the branches identified through the principal graph projection have high purity values (around 90% for all branches) that are on average higher than those of the k-means clusters. This finding indicates that the branches are more robust to random resamplings compared to the k-means clusters, suggesting that the principal graph description better reflects the continuous nature of the underlying data distribution.

3 Discussion

In this work, we proposed an integrative methodology based on variational autoencoders to characterize the clinical and molecular features of different COPD subtypes. We constructed personalized profiles that encode both types of information in a single description, and mapped the geometry of their distribution via principal graph fitting. We delineated five joint subtypes with distinct phenotypic characteristics, severity, disease outcomes and transcriptomics signatures. Three of these five subpopulations present multiple COPD features or symptoms. The subtype with the most severe COPD manifestations, SEV, includes the largest fraction of individuals with emphysema. Furthermore, despite the relatively younger average age and lower disease severity, individuals within the MOD subtype experience frequent exacerbations, coupled with unfavorable disease prognoses. These trends suggest that MOD encompasses conditions with a high risk of swift deterioration in health conditions. Conversely, subjects in the SYMPT subtype display only slight spirometric abnormalities and relatively favorable disease progression, yet they suffer from a relatively adverse symptomatology. At the molecular level, the SEV and MOD branches show divergent expression patterns in pathways associated to COPD-associated processes, including immune response, cellular respiration, and response to reactive oxygen species. Furthermore, the SEV and MOD branches delineate groups that are reminiscent of the classical emphysema-predominant (EPD) and non-emphysema-predominant (NEPD) subtypes [33], lending credibility to the hypothesis of distinct molecular being associated with these two conditions. Among the limitations of this study, we rely exclusively on expression data in blood for inferring the molecular processes associated to each subtypes. The integration of diverse omics types and tissues, in particular lung, will be crucial for delineating more detailed disease subtypes that capture the diversity of COPD processes and their relation to its clinical manifestations. Furthermore, this analysis represents an unsupervised endeavor aimed at uncovering the most distinctive features of the joint disease subtypes. The integration of supervised objectives to guide discovery in this setting is a promising direction left for future work. Overall, these findings highlight that joining the clinical and molecular features of COPD is an important step for identifying robust subtypes with clearly delineated mechanistic underpinnings and prognoses.

4 Methods

4.1 Processing of clinical and phenotypic data

In the initial phase of preparing the input data for the VAE, we chose all the subjects who had available clinical data and RNA-seq expression profiles, obtaining 3,628 samples. We classified each feature as either numerical or categorical, and we devised different processing strategies to handle the two groups of features. Among the categorical features, we selected all the ones with at most 10% of missing values across all subjects and where the most frequent category was

present in at most 80% of subjects. The latter criterion was devised to avoid including features that are not sufficiently informative of patient heterogeneity. Similarly, we considered all the numerical features with at most 10% missing values and with standard deviation greater than 0, to avoid constant values. We then imputed the remaining missing values in the categorical features by considering the most frequent category for each feature across all subjects. The numerical missing values were imputed through KNN imputation with k equal to 10. As further selection, we removed redundant categorical variables, classified as those that had high values of adjusted Rand score similarity with other variables (> 0.95).

4.2 Processing of expression data and differential expression analysis

From the raw read counts matrix we removed low expressed transcripts by selecting transcripts with at least 1 CPM in more than 10 samples. Next, we processed the data with the DeSeq2 algorithm [48], removing the batches resulting from library construction with the function "removeBatchEffect" of the package limma [49]. To perform differential expression (DE) analysis we executed the DE pipeline of DeSeq2 starting from the raw data. In brief, after basic data filtering we set up a design matrix with covariates including sequencing batch, age, sex, race, and cell composition. As contrasts we choose the membership to each of the three COPD branches (SEV,MOD,SYMPT) against the joined population of the reference branches NORM1 and NORM2. From the three contrasts we obtained three DE summary statistics. Next, we performed Gene Set Enrichment Analysis [50] with the GSEApy python package [51]. Specifically, we ranked the genes by their negative log p-value score multiplied by the sign of their log fold change (logFC). In this way, the genes that are significantly upregulated in the contrast (low p-value, positive logFC) appear at the top of the ranking, while the genes that are downregulated (low p-value, negative logFC) will appear at the bottom, providing a coarse measure of their general state of differential activation.

4.3 Conditional Variational Autoencoder design

In order to find a shared latent space for expression and clinical features, we designed a conditional VAE architecture with an "X" shape, similar to the X-VAE model described in [23], shown in Fig. 1 (c). The network takes as input a vector of concatenated RNA-seq read counts and clinical features that have been normalized to be in the unit range.

The architecture consists of four subnetworks: two encoders Encoder_{ϕ_1} and Encoder_{ϕ_2}, and two decoders $\text{Decoder}_{\theta_1}$ and $\text{Decoder}_{\theta_2}$. Each data mode x_1 and x_2, along with the conditional variable c, is separately passed through an encoder network.

Each encoder network, $\text{Encoder}_{\phi_i}(x_1, c)$ is composed of L layers where each layer l is defined as:

$$h_l^{(i)} = \text{ReLU}(\text{BatchNorm}(W_l^{(i)} h_{l-1}^{(i)} + b_l^{(i)})) \qquad i \in \{1, 2\} \tag{1}$$

with

$$h_0^{(i)} = \text{ReLU}(\text{BatchNorm}(W_0^{(i)} [x_i; c] + b_0^{(i)})) \qquad i \in \{1, 2\} \tag{2}$$

Here, h_l denotes the l-th layer's activations, BatchNorm is a batch normalization transformation, $W_l^{(i)} \in \phi_i$ and $b_l^{(i)} \in \phi_i$ are the weight matrix and bias terms of the l-th layer, ReLU is the activation function, and $[\cdot;\cdot]$ is the concatenation operator.
These encoders transform their inputs into latent representations:

$$h_1 = \text{Encoder}_{\phi_1}(x_1, c) \tag{3}$$

$$h_2 = \text{Encoder}_{\phi_2}(x_2, c) \tag{4}$$

The two latent representations h_1 and h_2 are then summed to form a shared representation $h = h_1 + h_2$. Two separate linear layers are used to project this representation into the parameters of a Gaussian distribution, forming the mean $\mu(h)$ and standard deviation $\Sigma(h)$ of the latent distribution of the latent vector z.

$$\mu(h), \Sigma(h) = \text{Linear}_\mu(h), \text{Linear}_\Sigma(h) \tag{5}$$

For the decoder part, the sampled latent variable z is concatenated with the conditional variable c and passed through the decoder networks to generate the reconstructed data $\hat{x_1}$ and $\hat{x_2}$.

$$\hat{x_1} = \text{Decoder}_{\theta_1}(z, c) \tag{6}$$

$$\hat{x_2} = \text{Decoder}_{\theta_2}(z, c). \tag{7}$$

Similarly, a decoder $\text{Decoder}_{\theta_i}(z, c)$ is composed of L layers

$$h_l^{(i)} = \text{ReLU}(\text{BatchNorm}(V_l^{(i)} h_{l-1}^{(i)} + d_l^{(i)})) \qquad i \in \{1, 2\} \tag{8}$$

with

$$h_0^{(i)} = \text{ReLU}(\text{BatchNorm}(V_0^{(i)}[z; c] + d_0^{(i)})) \qquad i \in \{1, 2\}. \tag{9}$$

As before, $V_l^{(i)} \in \theta_i$ and $d_l^{(i)} \in \theta_i$ are the weight matrix and bias terms of the l-th layer. Furthermore, the last layers of each decoder do not include a ReLU activation function to allow for negative outputs.

4.3.1 The Evidence Lower Bound (ELBO) and Maximum Mean Discrepancy (MMD) Loss

Instead of the usual Kullback-Leibler (KL) divergence, we employ the Maximum Mean Discrepancy (MMD) measure to optimize the latent space. The MMD is a distance measure between probability distributions in a Reproducing Kernel Hilbert Space (RKHS), and it does not require explicit computation of the distributions. The training objective of our multimodal VAE is to maximize the ELBO, which in this case is a convex combination of the ELBOs for each data mode, governed by a parameter α_1. The reconstruction loss for each mode is calculated differently due to the nature of their data. For the numerical data mode x_1, the Mean Squared Error (MSE) is used. For the mixed data mode x_2, which contains numerical and categorical data, the loss is a convex combination (controlled by a parameter α_2) of the MSE for the numerical part and the Categorical Cross-Entropy for the categorical part:

$$\begin{aligned}
\mathcal{L}_{\text{MMD}}(x_1, x_2, c, \theta, \phi) = &\alpha_1 \cdot \mathbb{E}_{q_\phi(z|x_1,x_2,c)}[\text{MSE}(x_1, \text{Decoder}_{\theta_1}(z, c))] \\
&+ (1 - \alpha_1) \cdot \Big(\mathbb{E}_{q_\phi(z|x_1,x_2,c)}[\alpha_2 \cdot \text{MSE}(x_{2_{\text{num}}}, \text{Decoder}_{\theta_{2_{\text{num}}}}(z, c)) \\
&+ (1 - \alpha_2) \cdot \text{CCE}(x_{2_{\text{cat}}}, \text{Decoder}_{\theta_{2_{\text{cat}}}}(z, c))]\Big) \\
&- \gamma \cdot \text{MMD}(q_\phi(z|x_1, x_2, c), p_\theta(z|c))
\end{aligned} \tag{10}$$

Here, $x_{2_{num}}$ represents the numerical part and $x_{2_{cat}}$ represents the categorical part of the second data mode x_2. $\text{Decoder}_{\theta_{2_{\text{num}}}}(z, c)$ and $\text{Decoder}_{\theta_{2_{\text{cat}}}}(z, c)$ are the reconstructed numerical and categorical parts of the data respectively, and CCE denotes the average Categorical Cross-Entropy loss across all the categorical variables. $\text{MMD}(q_\phi(z|x), p_\theta(z))$ measures the dissimilarity between the approximate posterior $q_\phi(z|x)$ and the prior $p_\theta(z)$ in terms of their mean embeddings in a RKHS.

4.3.2 Optimization

The parameters of the VAE, θ and ϕ, are typically optimized using stochastic gradient descent (SGD) or one of its variants. Due to the presence of the expectation over $q_\phi(z|x)$, the derivative is not directly computable. To tackle this problem, the reparameterization trick is used:

$$z = \mu(x, c) + \epsilon \odot \sqrt{\Sigma(x, c)}, \quad \epsilon \sim \mathcal{N}(0, I) \tag{11}$$

This trick allows the gradients to flow back through the network, as the randomness of the sampling process is isolated to the ϵ term. It results in low-variance and unbiased gradient estimates, thus enabling effective optimization.
Optimization was performed with the ADAM optimizer. In all our tests the full dataset is split in 80% training samples, 10% validation samples and 10% test samples. To identify the optimal hyperparameters, including number and size of hidden layers, the learning rate and training batch size we performed hyperparameter optimization with ASHA

(Asynchronous Successive Halving Algorithm), a scheme for parallel optimization equipped with greedy early stopping strategies to rule out inefficient hyperparameter configurations [52]. The final optimal configuration is an encoder subnetwork with 1 hidden layer of 32 neurons per each data type, while the decoder network is always constrained to mirror the encoder structure in reverse. The optimal mini-batch size is 32.

4.4 Prediction of future outcomes from PIPs

The prediction task was performed by training an array off-the shelf random forest (RF) classifiers with multiple input embeddings and output disease outcomes, as listed in the main text.

The embeddings of MOFA [32] were calculated by running the mofapy python package, feeding the expression and clinical data as input and by setting 30 as the number of hidden factors. The MOFA embeddings were defined as the expected values of the factors obtained after fitting the model.

For ease of comparison of performances, we transformed each disease outcome to a binary variable by selecting appropriate thresholds as follows. The considered target variables are the following: (1) ΔFEV$_1$ %pred. (P3<P2): more than 10% decrease of FEV$_1$ percent predicted between Phase 2 and Phase 3; (2) inc. bronchitis (P3): incident chronic bronchitis in Phase 3, restricted to individuals without chronic bronchitis in Phase 2; (3) exacerbations (P3): frequency of exacerbations in Phase 3 greater than 0; (4) Δ exacerbation frequency (P3>P2): frequency of exacerbations in Phase 3 greater than Phase 2; (5) severe exacerbations (P3): presence of severe exacerbations in Phase 3; (6) Δ severe exacerbations (P3>P2): presence of severe exacerbations in Phase 3 in subjects who did not experience severe exacerbations in Phase 2; (7) ΔmMRC (P3>P2): increased mMRC dyspnea score in Phase 3 compared to Phase 2; (8) ΔSF-36 (P3<P2): decrease of SF-36 score between Phase 2 and Phase 3 greater than 5 points; (9) Δ% emphysema (P3>P2): increase percent emphysema greater than 2.5% between Phase 2 and Phase 3; (10) Δ% gas trapping : increase in percent gas trapping greater than 5%; (11,12) mortality (3/5yr): all-cause mortality at 3 and 5 years. We set up each RF classifier with 100 decision trees. We executed a repeated (n=3) and stratified 5-fold cross-validation. For each split, we performed three steps: (1) we normalized the whole dataset according to the statistics obtained from the current training set; (2) since most clinical outcomes have highly unbalanced class distributions, we performed SMOTE oversampling [53] of the minority class, using the *imblearn* python package [54]; (3) we trained the classifier with the resulting data. The values shown in Table 1 are the summary statistics obtained by the 5-fold splits repeated 3 times, for a total of 15 performance values for each prediction. The significance values are obtained by performing t-test between the performance values obtained by the best embedding and the second-best embedding.

4.5 Construction of principal graph

To build the principal graph approximating data, we used the elPiGraph method [29]. We used default parameters, except for the maximum number of nodes, that was set to 30 to increase resolution. Also we set the "collapse" argument to True, in order to merge the small and noisy branches within their main branch. All the 2D embeddings shown in Figs. 2 and REF are evaluated with a modified version of the elPiGraph function "visualize_eltree_with_data". In brief, this function first produces a 2D embedding of the principal graph. Next, it distributes all the data points across the branches according to their calculated projections. Finally, to improve clarity each point is scattered randomly in the direction orthogonal to the branch by an extent controlled by a fixed parameter.

4.6 Processing of LFU and mortality data

To visualize the trends in Fig. 4 (a), we considered all the long-term follow-up (LFU) survey data that were compiled after Phase 2 of the COPDGene study (as of august 2022). Since the time points refer to the time the survey was compiled, we considered as the interval range of each data point the 6 months prior the compile date, unless another survey was compiled by the same subject less than 6 months earlier. In that case, the time interval is the time span occurring between the two surveys. To analyze the risk of increased exacerbations over time, we set up a Poisson regression model, controlling for the age, sex, and race covariates. The model was fit through the *glmfit* R function, using a log link function. We also tested an alternative mixed effect model where subject identity was included as a random effect, obtaining similar results. To estimate mortality at 3, and 5 years, we considered the COPDGene all-cause mortality data as of october 2022. To implement the Cox proportional hazard model of mortality we used the *lifelines* python package [55].

4.7 Evaluation of distance correlation (dCorr)

The similarity between two sets of N vectors embedded in two spaces can be estimated by modeling each vector set as the N realizations of a multivariate random variable. From this standpoint, the similarity between the two sets

is equivalent to the level of statistical dependency between the two variables. dCorr is an extension of the Pearson correlation to multivariate settings and it ranges between 0 (statistical independence), and 1 (linear dependency) [47]. Furthermore, dCorr is invariant to rigid transformations applied to either of the two spaces (e.g. rotations). This makes it an ideal tool for assessing the similarity between the two sets of vector embeddings. dCorr is estimated as follows [56]: let X and Y be two d-dimensional vector sets. Define a_{ij} and b_{ij} to be the Euclidean distances between the ith and jth elements of X and Y, respectively. We then form the centered distance matrices A and B as follows:

$$A_{ij} = a_{ij} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..}, \quad B_{ij} = b_{ij} - \bar{b}_{i.} - \bar{b}_{.j} + \bar{b}_{..},$$

where

$$\bar{a}_{i.} = \frac{1}{n}\sum_{k=1}^{n} a_{ik}, \quad \bar{a}_{.j} = \frac{1}{n}\sum_{k=1}^{n} a_{kj}, \quad \bar{a}_{..} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl},$$

$$\bar{b}_{i.} = \frac{1}{n}\sum_{k=1}^{n} b_{ik}, \quad \bar{b}_{.j} = \frac{1}{n}\sum_{k=1}^{n} b_{kj}, \quad \bar{b}_{..} = \frac{1}{n^2}\sum_{k,l=1}^{n} b_{kl},$$

The distance covariance (dCov) and distance correlation (dCorr) are defined as

$$\text{dCov}(X,Y) = \sqrt{\frac{1}{n^2}\sum_{i,j=1}^{n} A_{ij}B_{ij}},$$

$$\text{dCorr}(X,Y) = \frac{\text{dCov}(X,Y)}{\sqrt{\text{dCov}(X,X)\cdot\text{dCov}(Y,Y)}}.$$

The distance covariance correlation was evaluated with the python package *hyppo* [57].

4.8 Evaluation of the branch purity with respect to data resamplings

Let us consider a set of n data points with branch labels $Y = \{y_1, y_2, ..., y_n\}$, where each y_i belongs to one of K branches. In a resampled embedding we produce a new branch assignment $Y' = \{y'_1, y'_2, ..., y'_n\}$, where each y'_i belongs to one of K' branches, with K not necessarily equal to K'. The purity of branch k in the original labeling Y with respect to the resampled branch labels Y' is defined as

$$\text{Purity}(k; Y, Y') = \frac{1}{N_k}\max_{k'=1,...,K'} |\{i : y_i = k\} \cap \{i : y'_i = k'\}|$$

where $N_k = |\{i : y_i = k\}|$ and $|.|$ denotes the cardinality of a set. This values measures the number of items in the k-th branch that belong to the most common resampled branch. Purity is a fraction between 0 and 1, where 1 indicates that all the data points in the original branch are assigned to the same branch in the resampled graph.

References

[1] Rafael Lozano, Mohsen Naghavi, Kyle Foreman, Stephen Lim, Kenji Shibuya, Victor Aboyans, Jerry Abraham, Timothy Adair, Rakesh Aggarwal, Stephanie Y Ahn, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. *The lancet*, 380(9859):2095–2128, 2012.

[2] Rosa Faner and Alvar Agusti. Multilevel, dynamic chronic obstructive pulmonary disease heterogeneity. a challenge for personalized medicine. *Annals of the American Thoracic Society*, 13(Supplement 5):S466–S470, 2016.

[3] Peter J Castaldi, Adel Boueiz, Jeong Yun, Raul San Jose Estepar, James C Ross, George Washko, Michael H Cho, Craig P Hersh, Gregory L Kinney, Kendra A Young, et al. Machine learning characterization of copd subtypes: insights from the copdgene study. *Chest*, 157(5):1147–1157, 2020.

[4] Suchi Saria and Anna Goldenberg. Subtyping: What it is and its role in precision medicine. *IEEE Intelligent Systems*, 30(4):70–75, 2015.

[5] Kristina L Buschur, Craig Riley, Aabida Saferali, Peter Castaldi, Grace Zhang, Francois Aguet, Kristin G Ardlie, Peter Durda, W Craig Johnson, Silva Kasela, et al. Distinct copd subtypes in former smokers revealed by gene network perturbation analysis. *Respiratory Research*, 24(1):1–12, 2023.

[6] James C Ross, Peter J Castaldi, Michael H Cho, Craig P Hersh, Farbod N Rahaghi, Gonzalo V Sánchez-Ferrero, Margaret M Parker, Augusto A Litonjua, David Sparrow, Jennifer G Dy, et al. Longitudinal modeling of lung function trajectories in smokers with and without chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 198(8):1033–1042, 2018.

[7] Pierre-Régis Burgel, Jean-Louis Paillasseur, Wim Janssens, Jacques Piquet, Gerben Ter Riet, Judith Garcia-Aymerich, Borja Cosio, Per Bakke, Milo A Puhan, Arnulf Langhammer, et al. A simple algorithm for the identification of clinical copd phenotypes. *European Respiratory Journal*, 50(5), 2017.

[8] Peter J Castaldi, Jennifer Dy, James Ross, Yale Chang, George R Washko, Douglas Curran-Everett, Andre Williams, David A Lynch, Barry J Make, James D Crapo, et al. Cluster analysis in the copdgene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. *Thorax*, 69(5):416–423, 2014.

[9] Judith Garcia-Aymerich, Federico P Gómez, Marta Benet, Eva Farrero, Xavier Basagana, Angel Gayete, Carles Paré, Xavier Freixa, Jaume Ferrer, Antoni Ferrer, et al. Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (copd) subtypes. *Thorax*, 66(5):430–437, 2011.

[10] Lucas A Gillenwater, Shahab Helmi, Evan Stene, Katherine A Pratte, Yonghua Zhuang, Ronald P Schuyler, Leslie Lange, Peter J Castaldi, Craig P Hersh, Farnoush Banaei-Kashani, et al. Multi-omics subtyping pipeline for chronic obstructive pulmonary disease. *Plos one*, 16(8):e0255337, 2021.

[11] Yale Chang, Kimberly Glass, Yang-Yu Liu, Edwin K Silverman, James D Crapo, Ruth Tal-Singer, Russ Bowler, Jennifer Dy, Michael Cho, and Peter Castaldi. Copd subtypes identified by network-based clustering of blood gene expression. *Genomics*, 107(2-3):51–58, 2016.

[12] Zili Zhang, Jian Wang, Yuanyuan Li, Fei Liu, Lingdan Chen, Shunping He, Fanjie Lin, Xinguang Wei, Yaowei Fang, Qiongqiong Li, et al. Proteomics and metabolomics profiling reveal panels of circulating diagnostic biomarkers and molecular subtypes in stable copd. *Respiratory Research*, 24(1):73, 2023.

[13] Andrew Gregory, Zhonghui Xu, Katherine Pratte, Sool Lee, Congjian Liu, Robert Chase, Jeong Yun, Aabida Saferali, Craig P Hersh, Russell Bowler, et al. Clustering-based copd subtypes have distinct longitudinal outcomes and multi-omics biomarkers. *BMJ Open Respiratory Research*, 9(1):e001182, 2022.

[14] Margaret F Ragland, Christopher J Benway, Sharon M Lutz, Russell P Bowler, Julian Hecker, John E Hokanson, James D Crapo, Peter J Castaldi, Dawn L DeMeo, Craig P Hersh, et al. Genetic advances in chronic obstructive pulmonary disease. insights from copdgene. *American journal of respiratory and critical care medicine*, 200(6): 677–690, 2019.

[15] Jian Wang, Margaret R Spitz, Christopher I Amos, Anna V Wilkinson, Xifeng Wu, and Sanjay Shete. Mediating effects of smoking and chronic obstructive pulmonary disease on the relation between the chrna5-a3 genetic locus and lung cancer risk. *Cancer*, 116(14):3458–3462, 2010.

[16] Ivy Reichert Vital da Silva, Cintia Laura Pereira de Araujo, Gilson Pires Dorneles, Alessandra Peres, Andreia Luciana Bard, Gustavo Reinaldo, Paulo José Zimermann Teixeira, Pedro Dal Lago, and Viviane Rostirola Elsner. Exercise-modulated epigenetic markers and inflammatory response in copd individuals: A pilot study. *Respiratory physiology & neurobiology*, 242:89–95, 2017.

[17] NM Siafakas and EG Tzortzaki. Few smokers develop copd. why? *Respiratory medicine*, 96(8):615–624, 2002.

[18] Peter J Castaldi, Marta Benet, Hans Petersen, Nicholas Rafaels, James Finigan, Matteo Paoletti, H Marike Boezen, Judith M Vonk, Russell Bowler, Massimo Pistolesi, et al. Do copd subtypes really exist? copd heterogeneity and clustering in 10 independent cohorts. *Thorax*, 72(11):998–1006, 2017.

[19] Junxiang Chen, Xu Zhonghui, Li Sun, Ke Yu, Craig P Hersh, Adel Boueiz, John Hokanson, Frank C Sciurba, Edwin K Silverman, Peter J Castaldi, et al. Deep learning integration of chest ct imaging and gene expression identifies novel aspects of copd. *medRxiv*, pages 2022–09, 2022.

[20] Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, 14:1177932219899051, 2020.

[21] Evangelina López de Maturana, Lola Alonso, Pablo Alarcón, Isabel Adoración Martín-Antoniano, Silvia Pineda, Lucas Piorno, M Luz Calle, and Núria Malats. Challenges in the integration of omics and non-omics data. *Genes*, 10(3):238, 2019.

[22] Dongjin Leng, Linyi Zheng, Yuqi Wen, Yunhao Zhang, Lianlian Wu, Jing Wang, Meihong Wang, Zhongnan Zhang, Song He, and Xiaochen Bo. A benchmark study of deep learning-based multi-omics data fusion methods for cancer. *Genome Biology*, 23(1):1–32, 2022.

[23] Nikola Simidjievski, Cristian Bodnar, Ifrah Tariq, Paul Scherer, Helena Andres Terre, Zohreh Shams, Mateja Jamnik, and Pietro Liò. Variational autoencoders for cancer data integration: design principles and computational practice. *Frontiers in genetics*, 10:1205, 2019.

[24] Xiaoyu Zhang, Jingqing Zhang, Kai Sun, Xian Yang, Chengliang Dai, and Yike Guo. Integrated multi-omics analysis using variational autoencoders: application to pan-cancer classification. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 765–769. IEEE, 2019.

[25] Tianwei Yu. Aime: Autoencoder-based integrative multi-omics data embedding that allows for confounder adjustments. *PLoS Computational Biology*, 18(1):e1009826, 2022.

[26] Elizabeth A Regan, John E Hokanson, James R Murphy, Barry Make, David A Lynch, Terri H Beaty, Douglas Curran-Everett, Edwin K Silverman, and James D Crapo. Genetic epidemiology of copd (copdgene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 7(1):32–43, 2011.

[27] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. Multimodal deep learning for biomedical data fusion: a review. *Briefings in Bioinformatics*, 23(2):bbab569, 2022.

[28] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[29] Sergey E Golovenkin, Jonathan Bac, Alexander Chervov, Evgeny M Mirkes, Yuliya V Orlova, Emmanuel Barillot, Alexander N Gorban, and Andrei Zinovyev. Trajectories, bifurcations, and pseudo-time in large clinical datasets: Applications to myocardial infarction and diabetes data. *GigaScience*, 9(11):giaa128, 2020.

[30] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[31] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[32] Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, 14(6):e8124, 2018.

[33] Peter J Castaldi, Zhonghui Xu, Kendra A Young, John E Hokanson, David A Lynch, Stephen M Humphries, James C Ross, Michael H Cho, Craig P Hersh, James D Crapo, et al. Copd heterogeneity and progression: Emphysema-predominant and non-emphysema-predominant disease. *American Journal of Epidemiology*, page kwad114, 2023.

[34] Emily S Wan, Peter J Castaldi, Michael H Cho, John E Hokanson, Elizabeth A Regan, Barry J Make, Terri H Beaty, MeiLan K Han, Jeffrey L Curtis, Douglas Curran-Everett, et al. Epidemiology, genetics, and subtyping of preserved ratio impaired spirometry (prism) in copdgene. *Respiratory research*, 15(1):1–13, 2014.

[35] Jing Shi and Michael G Walker. Gene set enrichment analysis (gsea) for interpreting gene expression profiles. *Current Bioinformatics*, 2(2):133–137, 2007.

[36] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[37] Aran Singanayagam, Su-Ling Loo, Maria Calderazzo, Lydia J Finney, Maria-Belen Trujillo Torralbo, Eteri Bakhsoliani, Jason Girkin, Punnam Veerati, Prabuddha S Pathinayake, Kristy S Nichol, et al. Antiviral immunity is impaired in copd patients with frequent exacerbations. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, 317(6):L893–L903, 2019.

[38] Sharon Mumby and Ian M Adcock. Recent evidence from omic analysis for redox signalling and mitochondrial oxidative stress in copd. *Journal of Inflammation*, 19(1):10, 2022.

[39] Barry R Imhoff and Jason M Hansen. Extracellular redox status regulates nrf2 activation through mitochondrial reactive oxygen species. *Biochemical Journal*, 424(3):491–500, 2009.

[40] Yessica D Torres-Ramos, Araceli Montoya-Estrada, Alberto M Guzman-Grenfell, Javier Mancilla-Ramirez, Beatriz Cardenas-Gonzalez, Salvador Blanco-Jimenez, Jose D Sepulveda-Sanchez, Alejandra Ramirez-Venegas, and Juan J Hicks. Urban pm2. 5 induces ros generation and rbc damage in copd patients. *Front Biosci (Elite Ed)*, 3:808–817, 2011.

[41] Aravind T Reddy, Sowmya P Lakshmi, Asoka Banno, and Raju C Reddy. Role of gpx3 in pparγ-induced protection against copd-associated oxidative stress. *Free Radical Biology and Medicine*, 126:350–357, 2018.

[42] I Rahman and IM Adcock. Oxidative stress and redox regulation of lung inflammation in copd. *European respiratory journal*, 28(1):219–242, 2006.

[43] Isaac K Sundar, Sangwoon Chung, Jae-Woong Hwang, Gnanapragasam Arunachalam, Suzanne Cook, Hongwei Yao, Witold Mazur, Vuokko L Kinnula, Aron B Fisher, and Irfan Rahman. Peroxiredoxin 6 differentially regulates acute and chronic cigarette smoke–mediated lung inflammatory response and injury. *Experimental lung research*, 36(8):451–462, 2010.

[44] Saleela M Ruwanpura, Louise McLeod, Alistair Miller, Jessica Jones, Ross Vlahos, Georg Ramm, Anthony Longano, Philip G Bardin, Steven Bozinovski, Gary P Anderson, et al. Deregulated stat3 signaling dissociates pulmonary inflammation from emphysema in gp130 mutant mice. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, 302(7):L627–L639, 2012.

[45] Liang Yew-Booth, Mark A Birrell, Ming Sum Lau, Katie Baker, Victoria Jones, Iain Kilty, and Maria G Belvisi. Jak–stat pathway activation in copd. *European Respiratory Journal*, 46(3):843–845, 2015.

[46] JM Kiszałkiewicz, S Majewski, WJ Piotrowski, P Górski, D Pastuszak-Lewandoska, M Migdalska-Sęk, and E Brzeziańska-Lasota. Evaluation of selected il6/stat3 pathway molecules and mirna expression in chronic obstructive pulmonary disease. *Scientific Reports*, 11(1):22756, 2021.

[47] Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. 2007.

[48] Michael Love, Simon Anders, and Wolfgang Huber. Differential analysis of count data–the deseq2 package. *Genome Biol*, 15(550):10–1186, 2014.

[49] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7): e47–e47, 2015.

[50] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[51] Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics*, 39(1):btac757, 2023.

[52] Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Jonathan Ben-Tzur, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. A system for massively parallel hyperparameter tuning. *Proceedings of Machine Learning and Systems*, 2:230–246, 2020.

[53] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[54] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017. URL `http://jmlr.org/papers/v18/16-365.html`.

[55] Cameron Davidson-Pilon. lifelines: survival analysis in python. *Journal of Open Source Software*, 4(40):1317, 2019. doi:10.21105/joss.01317. URL `https://doi.org/10.21105/joss.01317`.

[56] Gábor J Székely and Maria L Rizzo. Partial distance correlation with methods for dissimilarities. 2014.

[57] Sambit Panda, Satish Palaniappan, Junhao Xiong, Eric W. Bridgeford, Ronak Mehta, Cencheng Shen, and Joshua T. Vogelstein. hyppo: A comprehensive multivariate hypothesis testing python package, 2020.