

# Joint clinical and molecular subtyping of COPD with variational autoencoders

Enrico Maiorino    Margherita De Marzio    Zhonghui Xu    Jeong H. Yun  
Robert P. Chase    Craig P. Hersh    Scott T. Weiss    Edwin K. Silverman  
Peter J. Castaldi\*    Kimberly Glass<sup>†</sup>\*

Channing Division of Network Medicine  
Brigham and Women’s Hospital  
Harvard Medical School

## 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 classifications that do not fully capture the disease’s complexity. Here, we bridge this gap by introducing a subtyping pipeline that integrates clinical and gene expression data with variational autoencoders. We apply this methodology to the COPDGene study, a large study of current and former smoking individuals with and without COPD. Our approach generates a set of vector embeddings, called Personalized Integrated Profiles (PIPs), that recapitulate the joint clinical and molecular state of the subjects in the study. Prediction experiments show that the PIPs have a predictive accuracy comparable to or better than other embedding approaches. Using trajectory learning approaches, we analyze the main trajectories of variation in the PIP space and identify five well-separated subtypes with distinct clinical phenotypes, expression signatures, and disease outcomes. Notably, these subtypes are more robust to data resampling compared to those identified using traditional clustering approaches. Overall, our findings provide new avenues to establish fine-grained associations between the clinical characteristics, molecular processes, and disease outcomes of COPD.

## 1 Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a complex chronic respiratory disease that is among the leading causes of death worldwide [1]. The disease manifests as a spectrum of conditions, including persistent airflow obstruction, lung inflammation, chronic bronchitis, and emphysema. 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 aimed at dissecting 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 divided into those that characterize the observable phenotypes of the disease (clinical subtyping), and those that focus on disentangling the disease processes, often referred to as “endotypes”, underlying various COPD manifestations (molecular

\*These authors contributed equally as co-last authors

<sup>†</sup>Corresponding author. Email: [kimberly.glass@channing.harvard.edu](mailto:kimberly.glass@channing.harvard.edu)

39 subtyping) [5]. Applications of the former type leverage clinical data including demographics, disease  
40 symptoms, spirometry measurements, or chest imaging [6, 7, 8, 9], whereas applications of the latter  
41 type leverage measurements from various omics assays (transcriptomics, proteomics, epigenomics,  
42 etc.) [10, 11, 5, 12]. Although both approaches offer valuable insights into different aspects of the  
43 disease, subtype classifications from these applications are exclusively defined within either the clinical  
44 or molecular domain, with the analysis in the other domain performed primarily for validation or  
45 post-hoc examination [13, 6, 8, 11, 5]. Consequently, these domain-specific classifications cannot  
46 capture disease mechanisms arising from the interaction between molecular processes and clinical or  
47 lifestyle factors [14, 15, 16, 17], potentially leading to inconsistent subtypes that are not reproducible  
48 across different patient cohorts [18].

49 Multi-omics data integration is a widely researched subject [19, 20, 21], and multiple methods  
50 have been developed for this purpose, such as MOFA [22], iCluster [23] and SNF [24]. In contrast,  
51 the simultaneous integration of both clinical and omics data for disease subtyping has received  
52 comparatively less attention, with its applications in COPD being confined to specific domains [25].  
53 One of the challenges in integrating omics and non-omics data is their inherent complexity. Data  
54 heterogeneity and bias, already present in multi-omics studies [26], are exacerbated when including  
55 clinical data, which is typically composed of complex data structures with heterogeneous correlation  
56 patterns and significant variation in terms of scales, sparsity, and noise [27]. To alleviate these  
57 issues and to account for potential nonlinear interactions between variables from different domains,  
58 specialized integrative methodologies based on autoencoder neural networks have been proposed in  
59 several disease contexts, including COPD [10] and cancer [28, 29, 30, 31]. However, a comprehensive  
60 subtyping analysis that integrates both the clinical and molecular domains of COPD has not been  
61 performed.

62 In this work, we propose a joint subtyping approach to integrate clinical and gene expression  
63 data extracted from the COPDGene Study [32] (see Fig. 1 (a)), a large study of current and  
64 former smokers with and without COPD. Building on recent developments in multi-modal learning  
65 [33, 34], we developed an integrative method based on variational autoencoders (VAEs). A VAE is  
66 an unsupervised neural network architecture designed to compress the input data and generate a  
67 set of compact encodings [35]. We trained the VAE with clinical, imaging, and transcriptomic data  
68 from COPDGene, generating a set of personalized integrated profiles (PIPs) that encode the joint  
69 clinical and molecular configuration of every individual in the population. By performing multiple  
70 outcome prediction experiments, we demonstrate that the generated PIPs are highly informative of  
71 the individual’s disease state and enable accurate prediction of future disease outcomes. Next, we map  
72 the continuous trajectories in the VAE space using a recently proposed trajectory learning technique  
73 [36]. Through this approach, we identify several well-separated disease states, each exhibiting distinct  
74 clinical and molecular characteristics (joint subtypes). Finally, we show that these joint subtypes are  
75 characterized by different disease progression patterns and mortality and that they are robust to  
76 resampling noise.

## 77 2 Results

78 COPDGene is an ongoing longitudinal multi-center study of current or ex-smoking individuals with  
79 and without COPD who have undergone extensive clinical, physiological, and radiological profiling  
80 at three time points across 10 years (Phase 1, 2, 3, see Fig. 1 (b)). Additionally, periodic long-term  
81 follow-up (LFU) surveys have been conducted every 6 months throughout the duration of the study.  
82 10,198 individuals were enrolled at baseline. In this work we considered all of the subjects with clinical  
83 and blood gene expression data in Phase 2 (five-year follow-up) of the study that were available at the  
84 time of analysis (3,628 subjects, see Fig. 1 (c)). Clinical data includes demographics, lifestyle factors  
85 (e.g., smoking habits), spirometry measurements, medical and medication history, chest CT imaging  
86 measures, respiratory symptoms, and complete blood counts (CBC). Gene expression data consists of  
87 whole blood RNA-seq profiling. Both of these data modalities have been used extensively for COPD  
88 subtyping [6, 7, 8, 9, 11, 5], and they are among the most widely-used read-outs of the phenotypic

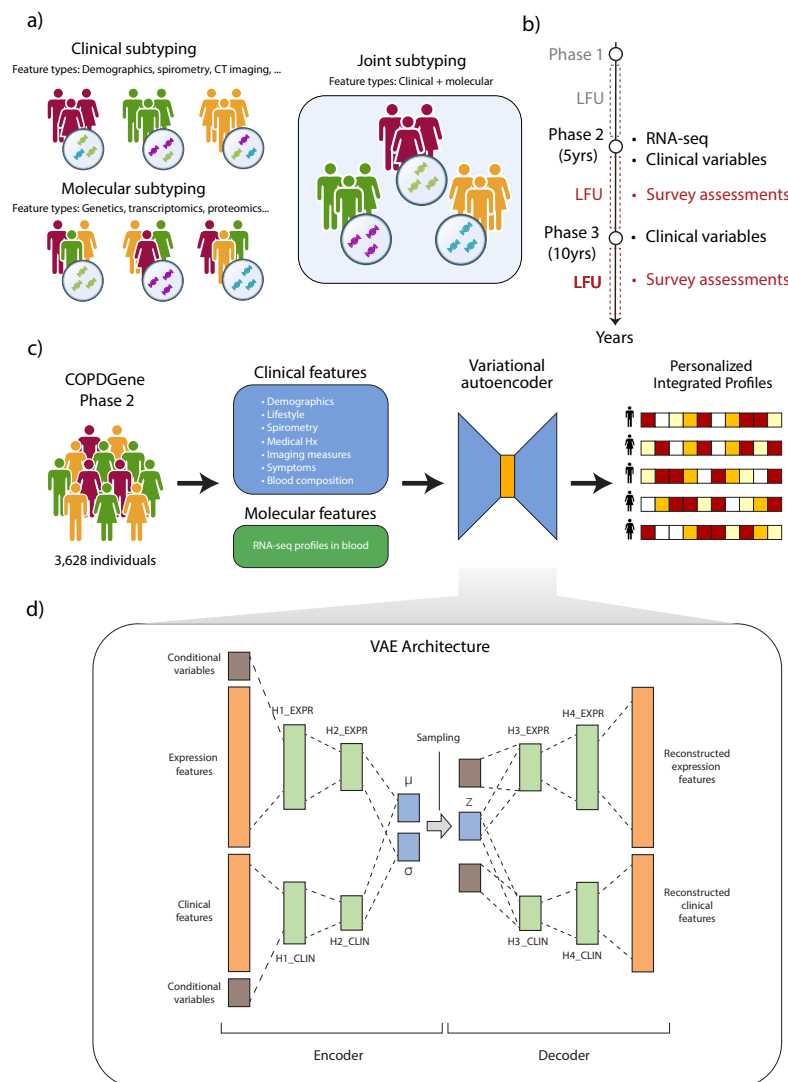


Figure 1: Overall organization of this work. (a) Joint subtyping aims at generating subtypes based on both clinical and molecular features. (b) COPDGene study design. Data from study stages preceding Phase 2 (gray) are not used in this work. On the right of the timeline are described the types of data collected at each study stage and considered in this study. Phase 2 and Phase 3 of the study are spaced approximately 5 years apart. LFU (Long-term Follow-Up) data consists of survey assessments conducted approximately every 6 months in between study phases. (c) Starting from the COPDGene population, an array of clinical and molecular features is extracted and used for training the VAE. The autoencoder produces a set of personalized integrated profiles (PIPs), one per individual in the population. (d) 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.

89 and molecular manifestations of COPD. To merge these two data types in a single representation of  
90 disease state, we designed a data integration scheme based on Variational Autoencoders (VAEs).

91 VAEs are probabilistic unsupervised neural network models designed to compress the input to  
92 generate low dimensional representations [37]. VAEs exploit statistical dependencies between input  
93 variables to construct a small set of variables (latent code) that preserve most of the input information.  
94 In contrast to linear techniques such as Principal Component Analysis (PCA), VAEs can capture  
95 nonlinear relationships between variables [38]. Drawing inspiration from other recently proposed  
96 architectures for multi-modal data integration [29, 30], we modified the standard VAE architecture so  
97 as to process and merge two data types (see Fig. 1 (d) and Methods). Our model implicitly performs  
98 a 2-step dimensionality reduction. The hidden layers in the encoder network (H1/2-expr, H1/2-clin in  
99 Fig. 1 (d)) process the two data types separately to obtain a data-type-specific representation of the  
100 input features. These representations are then coupled to generate a joint latent representation that  
101 encodes both the clinical and molecular information, which we refer to as “Personalized Integrated  
102 Profile” (PIP). Given the probabilistic nature of the VAE model, it is possible to correct for potential  
103 confounding factors by including them as a set of conditional variables [39]. These conditional  
104 variables have a similar role as regression covariates in linear regression modeling. Therefore, we set  
105 the age, sex, and race of each individual as conditional variables in both data modalities to regress  
106 out their effect on the learned representation.

107 The data processing and integration pipeline consists of several steps (see Methods for further  
108 details). In brief, we designed two separate processing pipelines for the two data types, consisting of  
109 feature selection and normalization operations. The resulting set of features selected for training  
110 the VAE is summarized in Supplementary Table 1. Next, we split the dataset into training and  
111 validation sets (80%/20%). We trained the VAE on the training set, and used the validation set for  
112 hyperparameter selection. We performed hyperparameter optimization to determine the optimal  
113 network depth, layer size, learning rate, and number of components of PIP vectors (latent components).  
114 Our primary objective was to select parameter configurations that resulted in high reconstruction  
115 accuracy of the validation set data, with a preference for configurations with fewer latent components.  
116 We determined that the optimal number of components in terms of reconstruction quality and latent  
117 vector size is 30. After training the network with the optimal parameters, we then used the Encoder  
118 subnetwork to generate the PIPs of the full dataset. The generated PIPs are the starting point of  
119 the subsequent analysis to identify joint subtypes of COPD.

## 120 2.1 Predicting future disease states with Personalized Integrated Profiles

121 The PIPs generated by the VAE contain information on both the expression and clinical features of  
122 a subject. To test the VAE’s performance in compressing and integrating different data modalities  
123 without sacrificing important information, we set up a prediction task of several prospective COPD  
124 outcomes. The COPD outcomes included all-cause mortality at 3 and 5 years after the Phase 2 and  
125 other clinical measurements collected in Phase 3 (P3) of the study, approximately five years after  
126 Phase 2 (P2) visits (see Methods for further details). These variables are extracted exclusively from  
127 data collected after Phase 2 of the study and thus were not used to train the VAE. We perform the  
128 classification using a Random Forest classifier, and define the PIP vectors as input variables and the  
129 Phase 3 outcomes as target variables (see Methods for details). For comparison, we evaluated the  
130 performance of the same classifier trained using other types of embedding as input: PCA of clinical  
131 variables (Clin PCA); PCA of expression variables (Expr PCA); PCA of concatenated expression  
132 and clinical variables (Expr + Clin PCA); Canonical Correlation Analysis (CCA) scores of the  
133 expression (Expr CCA) and clinical variables (Clin CCA); and factors calculated by applying the  
134 integrative method Multi Omics Factor Analysis (MOFA) [22] to expression and clinical variables.  
135 The performance metrics for the prediction, presented in Table 1, were derived from a 5-fold stratified  
136 validation repeated 3 times and then averaged (additional measures are reported in Supplementary  
137 Table 2). The VAE-based PIPs consistently achieve either the highest or second-highest scores across  
138 most prediction tasks. In the cases where the PIPs do not outperform other embeddings, no other  
139 alternative embedding scheme emerges as a clear leader in performance. This finding indicates that

140 despite encoding a broader range of information compared to domain-specific alternatives (e.g. Clin  
 141 PCA), the PIPs retain substantial information about an individual’s disease state and its likely  
 142 outcomes.

	F1-score						
	Clin PCA	Expr PCA	Expr+Clin PCA	Expr CCA	Clin CCA	MOFA	VAE
$\Delta$ FEV <sub>1</sub> % of pred.	<u>0.75 (0.03)</u>	0.74 (0.04)	0.73 (0.03)	<b>0.75 (0.04)</b> [n.s.]	0.74 (0.04)	0.74 (0.04)	0.73 (0.04)
Inc. chronic bronchitis	<u>0.02 (0.04)</u>	0.01 (0.03)	0.04 (0.06)	0.03 (0.05)	0.04 (0.05)	<u>0.07 (0.07)</u>	<b>0.12 (0.08)</b> **
Exacerbations (P3)	<u>0.37 (0.09)</u>	0.09 (0.06)	0.35 (0.06)	0.23 (0.07)	0.23 (0.08)	<u>0.25 (0.09)</u>	<b>0.41 (0.08)</b> *
$\Delta$ Exac. Freq. (P3>P2)	0.10 (0.08)	0.04 (0.05)	<u>0.14 (0.07)</u>	0.12 (0.07)	0.12 (0.08)	0.09 (0.06)	<b>0.20 (0.11)</b> **
Sev. Exacerbations (P3)	<u>0.11 (0.07)</u>	0.04 (0.04)	0.11 (0.08)	0.04 (0.05)	0.03 (0.04)	0.06 (0.06)	<b>0.15 (0.08)</b> *
$\Delta$ Sev. Exacerbations (P3>P2)	<u>0.08 (0.06)</u>	0.03 (0.05)	0.06 (0.08)	0.00 (0.02)	0.01 (0.03)	0.04 (0.06)	<b>0.11 (0.07)</b> [n.s.]
$\Delta$ MMRC (P3>P2)	<b>0.29 (0.05)</b> *	0.18 (0.05)	0.18 (0.05)	0.24 (0.04)	0.22 (0.05)	0.20 (0.05)	0.26 (0.04)
$\Delta$ SF-36 (P3<P2)	0.58 (0.03)	0.59 (0.03)	0.60 (0.03)	<u>0.60 (0.02)</u>	<b>0.61 (0.02)</b> *	0.58 (0.03)	0.59 (0.03)
Mortality (3yr)	0.08 (0.04)	0.12 (0.05)	0.11 (0.05)	<u>0.13 (0.06)</u>	<u>0.13 (0.06)</u>	0.10 (0.03)	<b>0.17 (0.05)</b> *
Mortality (5yr)	0.23 (0.05)	0.19 (0.04)	0.23 (0.03)	0.25 (0.05)	<u>0.25 (0.04)</u>	0.20 (0.04)	<b>0.30 (0.06)</b> **

Table 1: Prediction performance (F1-score) of COPD outcomes. Best performances and second-best are respectively displayed in bold or underlined. Abbreviations:  $\Delta$  = change from P2 to P3, P2 = phase 2 of COPDGene, P3 = phase 3 of COPDGene, (P3>P2) = value at P3 larger than value at P2, pred.=predicted, inc. chronic bronchitis = chronic bronchitis not present at P2 but present at P3, exac. freq.=exacerbations frequency

## 143 2.2 Summarizing COPD heterogeneity with principal graphs

144 The PIPs generated by the VAE are vectors distributed in a 30-dimensional space of variables that  
 145 implicitly describe the joint molecular and clinical characterization of every individual analyzed in our  
 146 COPDGene dataset. The geometry of the distribution of these generated vectors in the VAE space is  
 147 therefore informative of the patterns of variability of COPD features in the population, including the  
 148 presence, or lack thereof, of separate clusters. Growing evidence suggests that COPD manifestations  
 149 may form a continuous spectrum of disease states[3]. Under such circumstances, characterizations  
 150 based on discrete clusters may impose arbitrary boundaries within subpopulations that may impact  
 151 the robustness of the subtypes. To overcome this issue, we analyzed the trajectories of continuous  
 152 variation of COPD in the VAE space. We used the eLPiGraph method [36], an algorithm to fit  
 153 a branching network structure to a set of points in a multidimensional space. eLPiGraph and its  
 154 adaptations have been previously used to identify trajectories in several contexts, including the  
 155 clinical domain of COPD [40] and the molecular domain of cancer [41]. In this study, we extend its  
 156 application to the joint domain of clinical and molecular features of COPD.

157 In brief, eLPiGraph produces a tree-like network, called *principal graph*, that is embedded in the  
 158 VAE space. eLPiGraph optimizes the coordinates of the nodes of the principal graph to minimize  
 159 their distance from the data points. In this way, the principal graph traverses the main axes of  
 160 variation of the data points, approximating their intrinsic geometry. This procedure results in a  
 161 mapping between each data point and its corresponding projection onto the tree branches, providing  
 162 information on their relative positioning along the main axes of variation.

163 We applied eLPiGraph to construct the principal graph of the population and associated each  
 164 subject to their closest network branch in the space. The fitted principal graph is composed of 5  
 165 terminal branches, i.e., tree branches connected to the remaining graph only through one endpoint,  
 166 and 2 non-terminal branches, i.e., those connected at both their endpoints (see Fig. 2 (a)). Since  
 167 we were interested in identifying subtypes with distinct disease features and minimal overlap, we  
 168 restricted the analysis exclusively to the individuals within the terminal branches. Further, to  
 169 maximize the separation between different branches, we selected only the 50% of data points that lie  
 170 on the most extreme ends of each branch (see Methods and Figs. 2 (b) and Supplementary Figs. 1  
 171 (a-f)). Overall, 1,552 individuals were selected as members of any of the 5 branches.

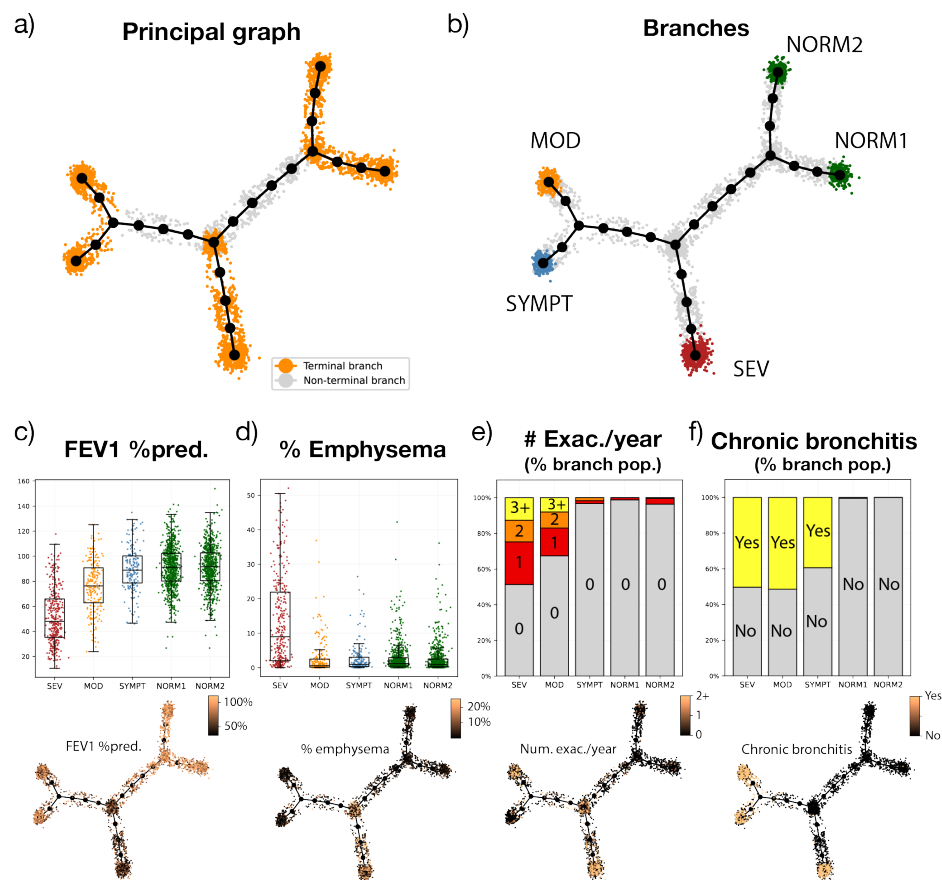


Figure 2: Features of the principal graph constructed from the COPDGene PIPs. (a) Principal graph layout. Large black dots and edges represent the fitted graph structure. Smaller dots represent individuals in the population, where the position is determined according to their proximity to the graph nodes. 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 members 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). All the graph layouts are generated with the Kamada-Kawai algorithm. Abbreviations: Pred.=predicted, Exac.=Exacerbations

### 172 2.3 The joint subtypes have distinct clinical characteristics

173 Of the 5 terminal branches, 3 are composed mainly of individuals with GOLD stage 2-4 corresponding  
 174 to moderate-to-severe COPD (bottom left branches in Fig. 2 (b)), whereas 2 branches (referred  
 175 to as NORM1 and NORM2) are composed mainly of individuals with preserved lung function or  
 176 mild COPD, i.e. GOLD stage equal to 0-1 (top right branches in Fig. 2 (b)). The branches are  
 177 characterized by different phenotypic profiles (see Supplementary Table 3).

178 The NORM1 and NORM2 branches differ primarily in mean age (NORM1:  $67 \pm 8.2$  yrs, NORM2:  
 179  $61 \pm 8.5$  yrs), current smoking status (NORM1: 18%, NORM2: 44%), and blood cell type composition  
 180 (predominantly neutrophilic in NORM1 and lymphocytic in NORM2).

181 The SEV branch consists of older individuals ( $68.7 \pm 8.3$  yrs) with the most severe disease  
 182 manifestations - low lung function (52% of predicted FEV<sub>1</sub>), frequent and severe exacerbations,

183 and high prevalence of chronic bronchitis (Figs. 2 (c-f)). These individuals also have a marked  
184 degree of emphysema, air trapping, and thickened airway walls. Consistent with the severity of  
185 their condition, self-reported metrics such as the mMRC dyspnoea scale and SGRQ scores indicate a  
186 severely compromised quality of life. Despite the high average number of pack years, the SEV branch  
187 is characterized by the smallest proportion of current smokers (31%) compared to other branches.  
188 This trend likely reflects the tendency of subjects with severe COPD to stop smoking.

189 The MOD branch consists of younger individuals ( $59.7 \pm 6.3$  yrs) with moderately impaired lung  
190 function and low percentage of emphysema and moderate airway wall thickening. However, compared  
191 to other branches (excluding SEV), MOD subjects are affected by a higher average frequency of  
192 exacerbations, and the severity of their respiratory symptoms is similar to those observed in the  
193 SEV branch, with an average mMRC score of  $2.4 \pm 1.3$  compared to  $2.7 \pm 1.2$  in the SEV branch and  
194 an SGRQ total score of  $46.2 \pm 19.7$  versus  $48.7 \pm 18.6$ . In line with previous studies [42], we defined  
195 frequent exacerbators as those individuals who experience two or more exacerbations in a year. The  
196 MOD branch has the second-largest proportion of frequent exacerbators (17%) after SEV (25%),  
197 and this subgroup has a substantial proportion of subjects in GOLD spirometric stage 2 or PRISm  
198 (preserved ratio impaired spirometry) group [43] (Supplementary Figure 2). Subjects in the MOD  
199 branch tend to have a phenotype that has been associated previously with airway-predominant  
200 COPD, with the highest average BMI and lowest amount of emphysema of all branches, as well as  
201 thick airways.

202 Individuals in the SYMPT branch are similar in most aspects to NORM1 and NORM2 branches  
203 (mild airway obstruction, low percent emphysema, infrequent exacerbations), yet this group has a  
204 larger proportion of current smokers and airway inflammation symptoms, such as cough, phlegm,  
205 and chronic bronchitis.

206 We examined the relation between the five branches and other previously proposed COPD  
207 subtypes. As emphysema is a common feature of COPD, recent works have distinguished classified  
208 phenotypes as emphysema-predominant (EPD, defined in individuals with GOLD>1 as CT-quantified  
209 densitometric emphysema  $\geq 10\%$  at -950 Hounsfield units), non-emphysema-predominant (NEPD,  
210 CT emph.  $< 5\%$ ) and intermediate emphysema (IE, CT emph. between 5% and 10%) [44]. The  
211 breakdown of each branch into separate classes shows that the SEV branch contains the largest fraction  
212 of EPD individuals (46%, see Supplementary Figure 3), a smaller fraction of NEPD individuals  
213 (23%), and negligible components of other states. Comparatively, the MOD branch is composed  
214 of a substantial proportion of NEPD individuals (22% overall, 62% within GOLD>1 stages) and  
215 only 8% of EPD phenotypes. Furthermore, MOD has the largest proportion (20%) of individuals  
216 in preserved FEV<sub>1</sub>/FVC ratio and impaired spirometry (PRISm). The remaining three branches  
217 (SYMPT, NORM1, NORM2) are composed mainly of individuals without significant COPD features  
218 and therefore contain negligible proportions of subjects with GOLD stage  $\geq 2$ .

## 219 2.4 The joint subtypes have distinct transcriptomic signatures and path- 220 way activations

221 We analyzed the transcriptomic differences between the SEV, MOD, and SYMPT branches using the  
222 combined NORM1 and NORM2 branches as the reference group (see Methods). Through differential  
223 expression (DE) analysis, we identified a set of DE genes for each contrast (see Supplementary Table  
224 4). For each set, we performed gene set enrichment analysis (GSEA) [45] using the 50 hallmark  
225 pathways of MSigDB [46] to find the over- or under-expressed biological pathways in each branch. The  
226 results are shown in Fig. 3, where it is evident that the SEV and MOD groups differ markedly in the  
227 expression of multiple biological pathways. Among the most significant pathways (FDR  $p\text{-adj.} < 0.05$ ),  
228 Interferon Alpha (IFN- $\alpha$ ) Response is highly over-expressed in the SEV and SYMPT branches and  
229 under-expressed in the MOD branch. The Oxidative Phosphorylation pathway is upregulated in both  
230 the SEV and MOD branches, while the Reactive Oxygen Species (ROS) pathway is upregulated only  
231 in the SEV branch and downregulated in the MOD branch. The majority of GSEA leading genes  
232 in the MOD branch are antioxidant agents. Among these, differential expression analysis reveals

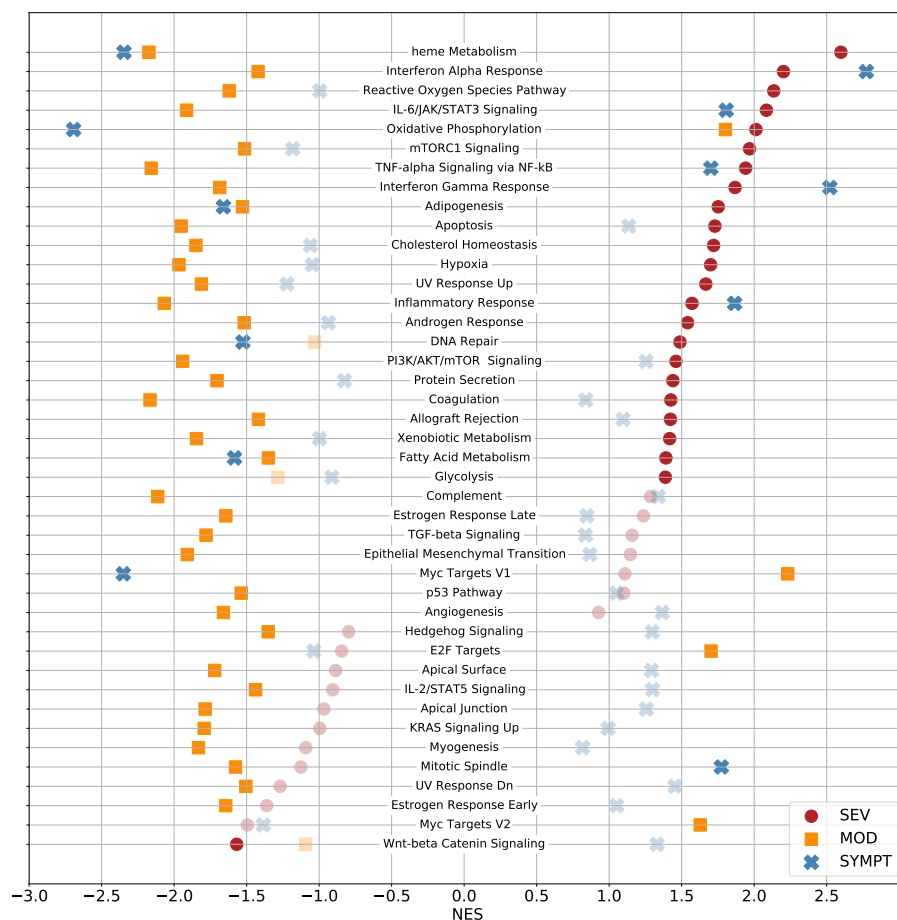


Figure 3: Gene set enrichment analysis of the differentially expressed genes between each branch and the NORM1 and NORM2 branches. Opaque points represent pathways that are significant (FDR p.adj.<0.05), while transparent points are non-significant. NES=normalized enrichment score.

233 downregulation of the antioxidant enzymes GPX3, G6PD, GSR, and TXNRD2 [47, 48].

## 234 2.5 The joint subtypes are associated to distinct disease outcomes and 235 risks

236 Next, we examined the associations between the joint subtypes and a set of COPD-related clinical  
237 outcomes. We collected data from the LFU dataset, in which COPDGene participants self-reported  
238 health updates via a survey every 6 months during the whole duration of the study (see Methods).  
239 Given that respiratory exacerbations are associated with COPD progression, we examined their  
240 temporal patterns among subjects from various branches. The temporal exacerbation patterns of the  
241 branches closely mirror the cross-sectional behavior observed during Phase 2 (Fig. 4 (a)), with the  
242 individuals in the SEV and MOD branches experiencing a higher rate of exacerbations throughout  
243 the entire time period. Interestingly, individuals who reported zero exacerbations in Phase 2, yet were  
244 classified in the SEV (163) and MOD (124) branches, demonstrated a significantly higher likelihood  
245 of experiencing one or more exacerbation events following Phase 2 — 42% and 50%, respectively —  
246 compared to the NORM1 and NORM2 branches ( 8%) and the SYMPT branch (16%). This finding  
247 suggests that branch membership can provide insights into the potential for future exacerbations,  
248 even when the present data does not explicitly indicate it. Moreover, to address potential confounding



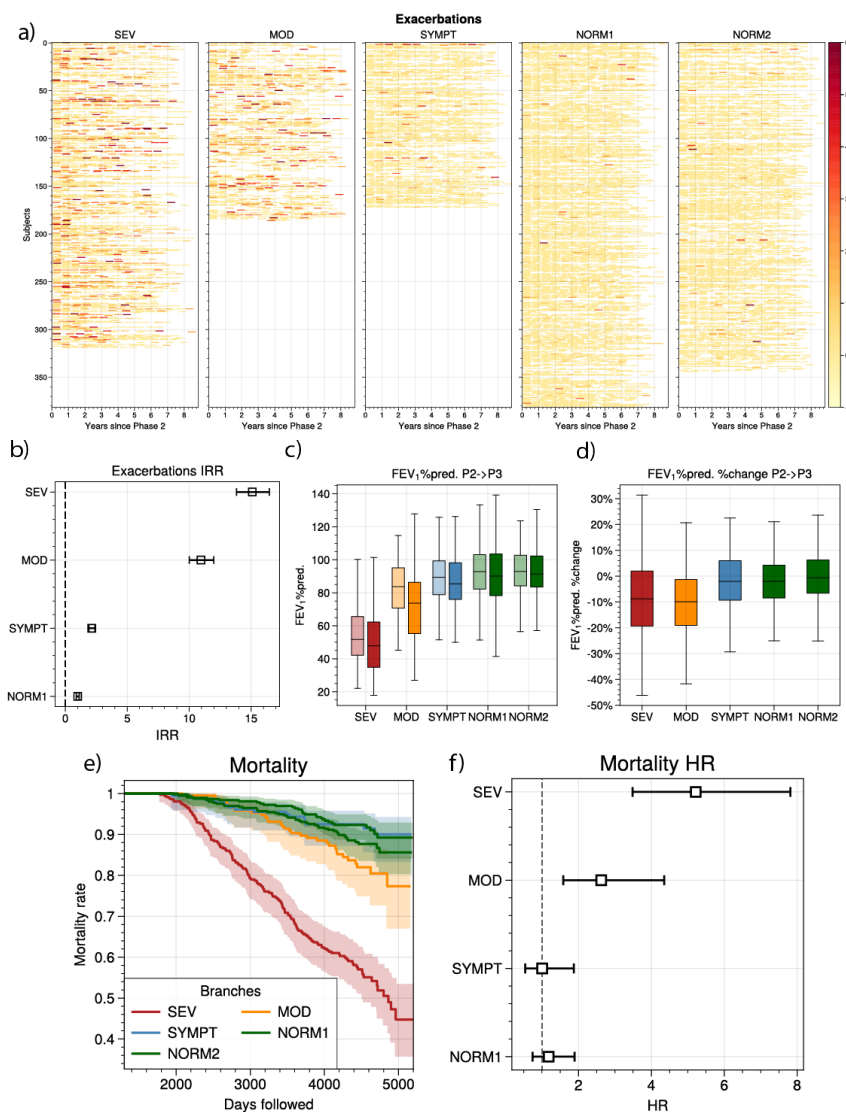


Figure 4: Prospective analysis of the branches. (a) Temporal trends of exacerbations among COPDGen subjects, categorized by their respective branches. Each row represents a subject, and different colored areas within these rows denote the number of exacerbations experienced by that subject during a 6-month 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<sub>1</sub> % of predicted levels categorized by branch (color), and by study phase (light=P2, dark=P3). (d) Distribution of relative changes in FEV<sub>1</sub> % of predicted categorized by branch. (e) Kaplan Meier curves of mortality, categorized by branch. (f) Hazard ratio (HR) of mortality between each branch and the reference group, NORM2.

249 factors, we used a Poisson regression model to analyze the incidence of exacerbations in the LFU data,  
 250 adjusting for the age, sex, and race of each participant (see Methods). The calculated incidence rate  
 251 ratios of each branch, using NORM2 as the reference, demonstrate that these insights are consistent  
 252 even when adjusting for demographic differences between the subpopulations (Fig 4 (b)).

253 Next, we assessed the change in FEV<sub>1</sub> % of predicted values over the five years between the Phase  
254 2 and Phase 3 visits (Figs 4 (c) and (d)). The SEV and MOD branches showed the largest decline in  
255 FEV<sub>1</sub> % of predicted, whereas the other three branches had a smaller decrease within the selected  
256 time frame. Finally, we estimated the mortality risk associated with each branch(see Fig. 4 (e) and  
257 (f) and Methods). The SEV and MOD branches demonstrate respectively a 5-fold and 2-fold average  
258 hazard ratio compared to the NORM2 branch, while the other branches do not exhibit increased risk.

## 259 2.6 The joint subtypes are robust to retraining and data resampling

260 To quantify the stability of the VAE space with respect to resampling of the training data, we  
261 performed two different robustness tests. In the first test, we evaluated the stability of the PIPs  
262 to random re-samplings of the training data. We retrained the VAE 100 times with different  
263 train/test splits (80%/20%, selected randomly) and generated 100 new sets of embeddings of the  
264 whole population (“resampled” embeddings). Each of these sets of embeddings correspond to the  
265 PIPs that would be generated by the VAE under scenarios where different subsets of the data were to  
266 be held out. In the ideal case, the resampled embeddings should provide similar information as the  
267 original embeddings, indicating that the identified patterns of variability are general and robust to  
268 noise. Therefore, we measured the overall similarity between the original and each resampled set of  
269 embeddings using the distance correlation measure (dCorr) [49]. dCorr is an extension of the Pearson  
270 correlation to multivariate settings and it ranges between 0, indicating statistical independence, and  
271 1, suggesting a linear relationship between the two variables (see Methods for further details). We  
272 measured a distribution of correlation values of  $0.900 \pm 0.006$ , indicating a strong similarity between  
273 the generated profiles.

274 As a second test, we tested the stability of the branch assignments to random resamplings of the  
275 training data. For each of the resampled embeddings described above, we constructed a principal  
276 graph (using the same settings) and assigned each point to a branch of the new graph (resampled  
277 branches). To measure the robustness of each of the original branches (SEV, MOD, SYMPT, NORM1,  
278 and NORM2), we defined the cluster purity measure as the proportion of subjects of each branch that  
279 were classified within the same branch in a resampled graph (see Methods). A high cluster purity  
280 indicates that individuals within the same cluster tend to be classified in the same cluster also in the  
281 resampled configurations. For each branch, we evaluated its purity against each resampled branch  
282 classification. In this way we obtained a set of 100 purity values per branch. For comparison, we  
283 repeated a similar operation by running the k-means algorithm on both the original VAE embeddings  
284 and each resampled embedding. For each embedding, we set a number of clusters equal to the number  
285 of branches identified within that embedding. Finally, we evaluated the purity values of each of the  
286 clusters retrieved in the original VAE embedding. Since the selection of non-terminal branches in  
287 the trajectory analysis has no equivalent operation in the clustering case, for the comparison we  
288 selected the 5 clusters with the highest median purity values. Additionally, to emulate the selection  
289 of the 50% highest confidence points along the trajectories, for each cluster we selected the 50% data  
290 points that are closest to the cluster centroid (‘core set’ of the cluster). The identified branches yield  
291 high purity values, approximately 90% for all branches (Supplementary Figure 1 (g), red boxes),  
292 outperforming both the standard k-means clusters (light blue boxes) and their core sets (dark blue  
293 boxes). This finding indicates that the branches are more robust to random resampling compared to  
294 the k-means clusters, suggesting that the principal graph description better reflects the continuous  
295 nature of the underlying data distribution.

## 296 3 Discussion

297 In this study, we proposed a novel approach to COPD subtyping, bridging the often-separate realms  
298 of clinical phenotyping and omics-driven profiling. Our main contributions are: (1) we showed that  
299 by integrating clinical and molecular data through variational autoencoders we retain domain-specific  
300 information while simultaneously capturing variability across multiple domains; (2) we showed that

301 trajectory analysis in the joint clinical and molecular space of COPD features identifies more robust  
302 subgroups compared to standard clustering approaches; (3) we identified 5 joint subtypes with  
303 distinctive clinical and transcriptomic features and disease outcomes.

304 The rationale of this work is that when clinical features aren't explicitly included in the discovery  
305 phase of subtyping, it can be challenging to ensure that the derived molecular subtypes hold clinical  
306 relevance [10]. Adding to previous work in multi-omic integration [19, 50], we designed a variational  
307 autoencoder architecture to integrate blood gene expression and clinical variables in COPD. While  
308 previous studies have focused on finding associations between clinical and molecular variables in specific  
309 contexts, such as CT imaging data [51], our integrative methodology describes this connection at a  
310 larger scale. Instead of focusing on a specific feature domain, we generated a comprehensive description  
311 of the COPD heterogeneity across multiple domains, including transcriptomics, demographics, lung  
312 function, lifestyle, CT measures, medical history, comorbidities, and symptoms.

313 Linear multi-omics integration methods, such as MOFA [22], assume a linear relationship among  
314 the features. In contrast, variational autoencoders offer flexibility in capturing complex, non linear  
315 interactions between features coming from different domains. The VAE architecture at the core of  
316 our methodology is designed to perform an implicit 2-step integration process. The first step consists  
317 in integrating features within the same domain (clinical or molecular). Then, the domain-specific  
318 higher-level features are subsequently integrated together to encode the Personalized Integrated  
319 Profile of a subject.

320 Several recent works have suggested that COPD manifestations are usually distributed as a  
321 continuum rather than discrete subgroups [18, 3], possibly stemming from the superposition of  
322 multiple endotypes [52]. In the absence of clearly-separated subpopulations, subjects with intermediate  
323 or hybrid COPD conditions are only loosely associated to their subtypes, and even slight noise  
324 perturbations can cause these subject to "switch" to adjacent groups. To address this issue, we used  
325 the elPiGraph trajectory learning algorithm [36] to map the main trajectories of the VAE latent space.  
326 Trajectory analysis yields an explicit linear ordering of individuals along each branch, generalizing  
327 the previously proposed concepts of "treatable traits" [52] and "disease axes" [53, 3]. We selected  
328 subjects with more extreme conditions to define subgroups, reducing the susceptibility to group  
329 switching. This approach improves the subtypes' robustness compared to standard clustering-based  
330 classifications, since in the latter the points with the highest confidence are those nearest to the  
331 centroid, and therefore reflect more average, rather than extreme, features.

332 Our analysis identified five joint subtypes with distinct phenotypic characteristics, severity, disease  
333 outcomes, and transcriptomics signatures. Three of these five subtypes present multiple COPD-like  
334 features. The subtype with the most severe COPD manifestations, SEV, includes the largest fraction  
335 of individuals with emphysema. The SEV group has the largest overlap with the established definition  
336 of emphysema-predominant phenotypes ( $>10\%$  CT-quantified emphysema), a subtype previously  
337 associated to larger annual  $FEV_1$  loss and higher risk of mortality [44]. Conversely, MOD individuals  
338 with moderate-to-severe COPD ( $GOLD \geq 2$ ) have large overlap with the non-emphysema-predominant  
339 subtype of COPD ( $<5\%$  CT-quantified emphysema). Individuals within the MOD subtype, despite  
340 their relatively younger average age and lower disease severity, experience frequent exacerbations  
341 and lose lung function at a comparable rate to the SEV group, consistent with prior observations  
342 linking exacerbations to loss of lung function [54, 55, 56, 57]. Notably, the fraction of frequent  
343 exacerbators in MOD ( $\geq 2$  exac. per year, as defined in [42]) is composed predominantly (83%) of  
344 mildly obstructed individuals ( $GOLD \leq 2$  or PRISm). The association between frequent exacerbations  
345 and milder airflow obstruction in certain COPD subgroups has been demonstrated previously, both  
346 within COPDGene (Phase 1) [54] and within the ECLIPSE study [42]. Individuals in the MOD  
347 subtype may be representative of earlier-stage COPD mechanisms where frequent exacerbations  
348 precede the development of severe airflow obstruction. An alternative possibility is that they represent  
349 a distinct trajectory of COPD characterized by airway inflammation and frequent exacerbations  
350 without emphysematous lung destruction.

351 Subjects in the SYMPT subtype are mostly current smokers ( $\sim 70\%$ ) with only slight spirometric  
352 abnormalities and relatively favorable disease progression, yet they suffer from multiple respiratory

353 symptoms, including chronic bronchitis, coughing and wheezing. SYMPT individuals are reminis-  
354 cent of the previously investigated subtype of symptomatic smokers with preserved lung function  
355 ( $FEV_1:FVC \geq 0.70$ ) [58, 59]. In agreement with the findings in [58], using LFU measurements we  
356 found that subjects within this subtype experience higher exacerbation frequency compared to the  
357 NORM2 baseline. However, in our prospective analysis we did not observe a substantial difference in  
358 loss of lung function between the SYMPT and the NORM1 and NORM2 subtypes.

359 Besides recapitulating previously observed patterns of phenotypic variation, the joint subtypes  
360 display marked transcriptomic differences. We measured a distinct pattern where multiple biological  
361 pathways exhibit an opposite directionality of activation between the SEV and MOD branches.  
362 Immune pathways, which are among the top enriched processes, have been consistently associated to  
363 COPD [11, 60, 61]. Among the most significant differences, we found several pathways related to  
364 immune response and regulation, including Interferon Alpha response, IL6/JAK/STAT3 Signaling  
365 and TNF- $\alpha$  Signaling via NF- $\kappa$ B. COPD patients who experience frequent exacerbations have been  
366 reported to exhibit reduced IFN- $\alpha$  levels in response to viral infection compared to individuals  
367 with lower exacerbation rates [62]. As such, downregulation of the IFN- $\alpha$  in the MOD branch  
368 might indicate compromised antiviral immunity, potentially leading to a higher susceptibility to  
369 exacerbations. ROS overproduction is known to suppress the activity of these enzymes [63, 64, 65, 66],  
370 and prolonged depletion of antioxidant capacity has been observed several days after the onset of  
371 exacerbation in COPD patients. Furthermore, previous reports have supported the role of the  
372 IL6/JAK/STAT3 signaling pathway in pulmonary inflammation and COPD severity [67, 68].

373 One limitation of this study is that we rely exclusively on expression data in blood for inferring  
374 the molecular processes associated to each subtype. The integration of diverse omics types and  
375 tissues [19], especially from the lung and airways, will be crucial for delineating more detailed  
376 disease subtypes that capture the diversity of COPD processes and their relationship with its clinical  
377 manifestations. Another limitation of our study is the lack of replication of our results in independent  
378 cohorts. Independent validation of this analysis is challenging since there are currently no other  
379 cohorts that have collected the data required in enough subjects with advanced COPD. In the future,  
380 with the additional generation of multi-omics data in NIH-funded studies such as SPIROMICS [69]  
381 through the Trans-Omics in Precision Medicine (TOPMed) program [70], it will be possible to pursue  
382 independent validation. In the meantime, we have pursued the next-best option, namely an extensive  
383 robustness analysis using resampling approaches. A crucial objective left for future work is the  
384 generation of distilled models that can reproduce our results in reduced datasets. Finally, cell type  
385 proportions play a dual role in gene expression analyses, serving both as a potential manifestation of  
386 the disease and a potential confounder [71]. This dual nature hinders the interpretability of gene  
387 expression results. Leveraging single cell RNA-seq data offers a way forward [72, 73]. By providing a  
388 higher resolution view of individual cell populations and their associated gene expression patterns, it  
389 becomes possible to discern between disease-associated shifts in cell populations and gene expression  
390 changes within specific cell types. Therefore, another promising future direction for this work is its  
391 application to single cell RNA-seq datasets.

## 392 4 Methods

### 393 4.1 Processing of clinical and phenotypic data

394 In the initial phase of preparing the input data for the VAE, we chose all the subjects who had  
395 available clinical data and RNA-seq expression profiles, obtaining 3,628 samples. We classified each  
396 feature as either numerical or categorical, and we devised different processing strategies to handle  
397 the two groups of features. Among the categorical features, we excluded those with more than 10%  
398 of missing values across all subjects or where the most frequent category was present in more than  
399 80% of subjects. The latter criterion was devised to avoid including features that are not sufficiently  
400 informative of patient heterogeneity. Similarly, we excluded the numerical features with more than  
401 10% missing values or with constant values. We then imputed the remaining missing values in the

402 categorical features by considering the most frequent category for each feature across all subjects.  
403 The numerical missing values were imputed through KNN imputation with  $k$  equal to 10. As further  
404 selection, we selected one representative variable among the groups of redundant categorical variables,  
405 i.e., those with high similarity values (adjusted Rand score  $> 0.95$ ). The resulting set of clinical  
406 features selected for training the VAE is summarized in Supplementary Table 1.

## 407 4.2 Processing of expression data and differential expression analysis

408 From the raw read counts matrix we removed low expressed transcripts by selecting transcripts with  
409 at least 1 CPM in more than 10 samples. Next, we processed the data with the DeSeq2 algorithm  
410 [74], and removed batch effects with the function “removeBatchEffect” of the package limma [75].  
411 Finally, to make the computations more manageable, and to perform a preliminary feature selection,  
412 we selected only the genes that were loosely associated to at least one clinical feature in the dataset.  
413 We assessed the significance of the relationships between each gene and each clinical feature, using  
414 Spearman’s correlation for numerical features and the Kruskal Wallis test for categorical ones. Genes  
415 with at least one FDR-adjusted  $p$ -value below 0.001 were retained, yielding 5,979 transcripts.

416 To perform differential expression (DE) analysis we executed the DE pipeline of DeSeq2 starting  
417 from the raw data. In brief, after basic data filtering we set up a design matrix with covariates  
418 including sequencing batch, age, sex, race, and white blood cell proportions. As contrasts we choose  
419 the membership to each of the three COPD branches (SEV, MOD, SYMPT) against the joined  
420 population of the reference branches NORM1 and NORM2. From the three contrasts we obtained  
421 three DE summary statistics. Next, we performed Gene Set Enrichment Analysis [76] with the  
422 GSEAPy python package [77]. Specifically, we ranked the genes by their negative log  $p$ -value score  
423 multiplied by the sign of their log fold change (logFC). In this way, the genes that are significantly  
424 upregulated in the contrast (low  $p$ -value, positive logFC) appear at the top of the ranking, while the  
425 genes that are downregulated (low  $p$ -value, negative logFC) will appear at the bottom, providing a  
426 coarse measure of their general state of differential activation.

## 427 4.3 Conditional Variational Autoencoder design

428 In order to find a shared latent space for expression and clinical features, we designed a conditional  
429 VAE architecture with an “X” shape, similar to the X-VAE model described in [29], shown in Fig. 1  
430 (d). The network takes as input a vector of concatenated RNA-seq read counts and clinical features  
431 that have been normalized to be in the unit range.

432 The architecture consists of four subnetworks: two encoders  $\text{Encoder}_{\phi_1}$  and  $\text{Encoder}_{\phi_2}$ , and two  
433 decoders  $\text{Decoder}_{\theta_1}$  and  $\text{Decoder}_{\theta_2}$ . Each data mode  $x_1$  and  $x_2$ , along with the conditional variable  
434  $c$ , is separately passed through an encoder network.

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

$$436 \quad h_l^{(i)} = \text{ReLU}(\text{BatchNorm}(W_l^{(i)} h_{l-1}^{(i)} + b_l^{(i)})) \quad i \in \{1, 2\} \quad (1)$$

436 with

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

437 Here,  $h_l^{(i)}$  denotes the  $l$ -th layer’s activations, BatchNorm is a batch normalization transformation,  
438  $W_l^{(i)} \in \phi_i$  and  $b_l^{(i)} \in \phi_i$  are the weight matrix and bias terms of the  $l$ -th layer, ReLU is the activation  
439 function, and  $[\cdot; \cdot]$  is the concatenation operator.

440 These encoders transform their inputs into latent representations:

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

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

441 The two latent representations  $h_1$  and  $h_2$  are then concatenated to form a shared representation  
 442  $h = [h_1; h_2]$ . Two separate linear layers are used to project this representation into the parameters of  
 443 a Gaussian distribution, forming the mean vector  $\mu(h)$  and the vector of standard deviations  $\sigma(h)$  of  
 444 the latent distribution of the latent vector  $z$ .

$$\mu(h), \sigma(h) = \text{Linear}_\mu(h), \text{Linear}_\sigma(h) \quad (5)$$

445 For the decoder part, the sampled latent variable  $z$  is concatenated with the conditional variable  
 446  $c$  and passed through the decoder networks to generate the reconstructed data  $\hat{x}_1$  and  $\hat{x}_2$ .

$$\hat{x}_1 = \text{Decoder}_{\theta_1}(z, c) \quad (6)$$

$$\hat{x}_2 = \text{Decoder}_{\theta_2}(z, c). \quad (7)$$

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

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

448 with

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

449 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.  
 450 Furthermore, the last layers of each decoder do not include a ReLU activation function to allow for  
 451 negative outputs.

### 452 4.3.1 The Evidence Lower Bound and Maximum Mean Discrepancy Loss

453 The training objective of our multimodal VAE is to maximize the Evidence Lower Bound (ELBO),  
 454 which in this case is a convex combination of the data-type-specific ELBO terms, governed by a  
 455 parameter  $\alpha_1$ . The reconstruction loss for each mode is calculated differently due to the nature  
 456 of their data. For the numerical data mode  $x_1$ , the Mean Squared Error (MSE) is used. For the  
 457 mixed data mode  $x_2$ , which contains numerical and categorical data, the loss is a convex combination  
 458 (controlled by a parameter  $\alpha_2$ ) of the MSE for the numerical part and the Categorical Cross-Entropy  
 459 for the categorical part. Furthermore, instead of the standard Kullback-Leibler (KL) divergence used  
 460 in the original formulation of VAE [35], we employ the Maximum Mean Discrepancy (MMD) measure  
 461 to regularize the model as proposed in the InfoVAE formulation [78]. Similarly to KL divergence,  
 462 MMD is a distance measure between two probability distributions. However, MMD only depends on  
 463 a set of statistics evaluated from the two distributions, and therefore it does not require the explicit  
 464 evaluation of their analytical form. In the Gaussian case, MMD calculation is done by embedding  
 465 the distributions in a Reproducing Kernel Hilbert Space (RKHS) identified by their average and  
 466 standard deviation, and by computing the distance between these statistics [79]. The ELBO  $\mathcal{L}_{\text{MMD}}$   
 467 has the form

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

468 Here,  $x_{2_{\text{num}}}$  represents the numerical part and  $x_{2_{\text{cat}}}$  represents the categorical part of the sec-  
 469 ond data mode  $x_2$ . The expected values  $\mathbb{E}_{q_\phi(z|x_1, x_2, c)}$  are approximated by a single sampling of  
 470  $z \sim \mathcal{N}(\mu(h), \sigma(h))$ .  $\text{Decoder}_{\theta_{2_{\text{num}}}}(z, c)$  and  $\text{Decoder}_{\theta_{2_{\text{cat}}}}(z, c)$  are the reconstructed numerical and

471 categorical parts of the data respectively, and CCE denotes the average Categorical Cross-Entropy  
472 loss across all the categorical variables.  $MMD(q_\phi(z|x), p_\theta(z))$  measures the dissimilarity between the  
473 approximate posterior  $q_\phi(z|x)$  and the prior  $p_\theta(z)$  in terms of their mean embeddings in the RKHS.

#### 474 4.3.2 Optimization

475 Optimization was performed with the ADAM optimizer [80]. In all our tests the full dataset is split  
476 in 80% training samples, 10% validation samples and 10% test samples. To identify the optimal  
477 hyperparameters (number and size of hidden layers, the learning rate, and training batch size), we  
478 performed hyperparameter optimization with ASHA (Asynchronous Successive Halving Algorithm), a  
479 scheme for parallel optimization equipped with greedy early stopping strategies to rule out inefficient  
480 hyperparameter configurations [81]. In the final optimal configuration the encoder subnetwork for  
481 processing the expression data has 2 hidden layers of dimensions 1024 and 512 neurons, while the  
482 encoder subnetworks for processing the clinical data has 1 hidden layer of dimension 64 neurons. The  
483 decoder networks are constrained to mirror the encoder structure in reverse. The mini-batch size is  
484 128.

#### 485 4.4 Prediction of future outcomes from PIPs

486 The prediction task was performed by training a set of random forest (RF) classifiers using multiple  
487 input embeddings and output disease outcomes, as listed in the main text.

488 The embeddings of MOFA [22] were calculated by running the mofapy python package, feeding  
489 the expression and clinical data as input and by setting 30 as the number of hidden factors. The  
490 MOFA embeddings were defined as the estimated means of the factors obtained after fitting the  
491 model.

492 For ease of comparison of performances, we considered a set of binary outcomes. COPD outcomes  
493 that are measured as continuous variables were transformed to binary variables by thresholding. The  
494 selection of thresholds was guided by practical considerations to ensure a balanced representation  
495 of positive and negative examples. The considered target variables are the following: (1)  $\Delta FEV_1$   
496 % of pred. (P3<P2): subjects with more than 10% decrease of  $FEV_1$  percent predicted between  
497 Phase 2 and Phase 3 (positive class) compared to subjects with more than 10% increase (negative  
498 class) ( $N=331$ ); (2) inc. bronchitis (P3): incident chronic bronchitis in Phase 3, restricted to  
499 individuals without chronic bronchitis in Phase 2 ( $N=1,087$ ); (3) exacerbations (P3): frequency  
500 of exacerbations in Phase 3 greater than 0 ( $N=1,251$ ); (4)  $\Delta$  exacerbation frequency (P3>P2):  
501 frequency of exacerbations in Phase 3 greater than Phase 2 ( $N=1,251$ ); (5) severe exacerbations  
502 (P3): presence of severe exacerbations in Phase 3 ( $N=1,250$ ); (6)  $\Delta$  severe exacerbations (P3>P2):  
503 presence of severe exacerbations in Phase 3 in subjects who did not experience severe exacerbations  
504 in Phase 2 ( $N=1,164$ ); (7)  $\Delta$ mMRC (P3>P2): increased mMRC dyspnea score in Phase 3 compared  
505 to Phase 2 ( $N=1,183$ ); (8)  $\Delta$ SF-36 (P3<P2): decrease of SF-36 score between Phase 2 and Phase  
506 3 ( $N=1,251$ ); (9,10) mortality (3/5yr): all-cause mortality at 3 and 5 years ( $N=3,361/3,347$ ). We  
507 set up each RF classifier with 100 decision trees. For a more robust performance assessment that  
508 is not too sensitive to a specific train/test split of the dataset, we conducted a stratified 5-fold  
509 cross-validation, repeated 3 times. For each split, we performed three steps: (1) we normalized the  
510 whole dataset according to the statistics obtained from the current training set; (2) since most clinical  
511 outcomes have highly unbalanced class distributions, we performed SMOTE oversampling [82] of  
512 the minority class, using the *imblearn* python package [83]; (3) we trained the classifier with the  
513 resulting data. The values shown in Table 1 are the summary statistics obtained by the 5-fold splits  
514 repeated 3 times, for a total of 15 performance values for each prediction. The significance values are  
515 obtained by performing t-tests between the performance values obtained by the best embedding and  
516 the second-best embedding.

## 517 4.5 Construction of principal graph

518 To build the principal graph describing the distance relationships between observations in the  
519 embedding space, we used the elPiGraph method [36]. We used default parameters, except for the  
520 maximum number of nodes which was set to 30 to increase resolution. Also we set the “collapse”  
521 argument to True, in order to merge the small and noisy branches within their main branch. All the  
522 2D embeddings shown in Fig. 2 are evaluated with a modified version of the elPiGraph function  
523 “visualize\_elfree\_with\_data”. In brief, this function first produces a 2D embedding of the principal  
524 graph using the Kamada-Kawai layout algorithm [84]. Next, it distributes all the data points across  
525 the branches according to their calculated projections. Finally, to improve clarity each point is  
526 scattered randomly in the direction orthogonal to the branch by an extent controlled by a fixed  
527 parameter.

## 528 4.6 Processing of LFU and mortality data

529 To visualize the trends in Fig. 4 (a), we considered all the long-term follow-up (LFU) survey data  
530 that were compiled after Phase 2 of the COPDGene study (as of August 2022). Since the time points  
531 refer to the time the survey was compiled, we considered as the interval range of each data point the  
532 6 months prior to the compile date, unless another survey was compiled by the same subject less  
533 than 6 months earlier. In that case, the time interval is the time span occurring between the two  
534 surveys. To analyze the risk of increased exacerbations over time, we set up a Poisson regression  
535 model, controlling for the age, sex, and race covariates. The model was fit through the *glmfit* R  
536 function, using a log link function. We also tested an alternative mixed effect model where subject  
537 identity was included as a random effect, obtaining similar results. To estimate mortality at 3 and 5  
538 years, we considered the COPDGene all-cause mortality data as of October 2022. To implement the  
539 Cox proportional hazard model of mortality we used the *lifelines* python package [85].

## 540 4.7 Evaluation of distance correlation (dCorr)

541 The similarity between two sets of  $N$  vectors embedded in two spaces can be estimated by modeling  
542 each vector set as the  $N$  realizations of a multivariate random variable. From this standpoint, the  
543 similarity between the two sets is equivalent to the level of statistical dependency between the two  
544 variables. dCorr is an extension of the Pearson correlation to multivariate settings and it ranges  
545 between 0 (statistical independence) and 1 (linear dependency) [49]. Furthermore, dCorr is invariant  
546 to rigid transformations applied to either of the two spaces (e.g. rotations). This makes it an ideal  
547 tool for assessing the similarity between the two sets of vector embeddings. dCorr is estimated as  
548 follows [86]:

549 let  $X$  and  $Y$  be two  $d$ -dimensional vector sets. Define  $a_{ij}$  and  $b_{ij}$  to be the Euclidean distances  
550 between the  $i$ th and  $j$ th elements of  $X$  and  $Y$ , respectively. We then form the centered distance  
551 matrices  $A$  and  $B$  as follows:

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

552 where

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

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



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

554 The distance covariance correlation was evaluated with the python package *hyppo* [87].

## 555 4.8 Evaluation of the branch purity with respect to data resamplings

Let us consider a set of  $n$  data points with branch labels  $Y = \{y_1, y_2, \dots, y_n\}$ , where each  $y_i$  belongs to one of  $K$  branches. In a resampled embedding we produce a new branch assignment  $Y' = \{y'_1, y'_2, \dots, y'_n\}$ , 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

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

556 where  $N_k = |\{i : y_i = k\}|$  and  $|\cdot|$  denotes the cardinality of a set. This value measures the number of  
557 items in the  $k$ -th branch that belong to the most common resampled branch. Purity is a fraction  
558 between 0 and 1, where higher values indicate stronger alignment between the original and resampled  
559 branch assignments.

## 560 5 Data Availability

561 All COPDGene data for the analysis are available in dbGaP accession number phs000179.v6.p2. The  
562 data generated in this study has been deposited in the Zenodo repository, accessible via the DOI:  
563 10.5281/zenodo.10431493.

## 564 6 Code Availability

565 The code used to reproduce the analyses in this study is available in the GitHub repository accessible  
566 via the following URL: [https://github.com/reemagit/joint\\_subtyping\\_vae](https://github.com/reemagit/joint_subtyping_vae).

## 567 References

- 568 [1] Rafael Lozano, Mohsen Naghavi, Kyle Foreman, Stephen Lim, Kenji Shibuya, Victor Aboyans,  
569 Jerry Abraham, Timothy Adair, Rakesh Aggarwal, Stephanie Y Ahn, et al. Global and regional  
570 mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for  
571 the global burden of disease study 2010. *The lancet*, 380(9859):2095–2128, 2012.
- 572 [2] Rosa Faner and Alvar Agusti. Multilevel, dynamic chronic obstructive pulmonary disease  
573 heterogeneity. a challenge for personalized medicine. *Annals of the American Thoracic Society*,  
574 13(Supplement 5):S466–S470, 2016.
- 575 [3] Peter J Castaldi, Adel Boueiz, Jeong Yun, Raul San Jose Estepar, James C Ross, George Washko,  
576 Michael H Cho, Craig P Hersh, Gregory L Kinney, Kendra A Young, et al. Machine learning  
577 characterization of copd subtypes: insights from the copdgene study. *Chest*, 157(5):1147–1157,  
578 2020.
- 579 [4] Suchi Saria and Anna Goldenberg. Subtyping: What it is and its role in precision medicine.  
580 *IEEE Intelligent Systems*, 30(4):70–75, 2015.

- 581 [5] Kristina L Buschur, Craig Riley, Aabida Saferali, Peter Castaldi, Grace Zhang, Francois Aguet,  
582 Kristin G Ardlie, Peter Durda, W Craig Johnson, Silva Kasela, et al. Distinct copd subtypes in  
583 former smokers revealed by gene network perturbation analysis. *Respiratory Research*, 24(1):  
584 1–12, 2023.
- 585 [6] James C Ross, Peter J Castaldi, Michael H Cho, Craig P Hersh, Farbod N Rahaghi, Gonzalo V  
586 Sánchez-Ferrero, Margaret M Parker, Augusto A Litonjua, David Sparrow, Jennifer G Dy,  
587 et al. Longitudinal modeling of lung function trajectories in smokers with and without chronic  
588 obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 198  
589 (8):1033–1042, 2018.
- 590 [7] Pierre-Régis Burgel, Jean-Louis Paillasseur, Wim Janssens, Jacques Piquet, Gerben Ter Riet,  
591 Judith Garcia-Aymerich, Borja Cosio, Per Bakke, Milo A Puhan, Arnulf Langhammer, et al.  
592 A simple algorithm for the identification of clinical copd phenotypes. *European Respiratory  
593 Journal*, 50(5), 2017.
- 594 [8] Peter J Castaldi, Jennifer Dy, James Ross, Yale Chang, George R Washko, Douglas Curran-  
595 Everett, Andre Williams, David A Lynch, Barry J Make, James D Crapo, et al. Cluster analysis  
596 in the copdgene study identifies subtypes of smokers with distinct patterns of airway disease  
597 and emphysema. *Thorax*, 69(5):416–423, 2014.
- 598 [9] Judith Garcia-Aymerich, Federico P Gómez, Marta Benet, Eva Farrero, Xavier Basagana,  
599 Angel Gayete, Carles Paré, Xavier Freixa, Jaume Ferrer, Antoni Ferrer, et al. Identification  
600 and prospective validation of clinically relevant chronic obstructive pulmonary disease (copd)  
601 subtypes. *Thorax*, 66(5):430–437, 2011.
- 602 [10] Lucas A Gillenwater, Shahab Helmi, Evan Stene, Katherine A Pratte, Yonghua Zhuang, Ronald P  
603 Schuyler, Leslie Lange, Peter J Castaldi, Craig P Hersh, Farnoush Banaei-Kashani, et al. Multi-  
604 omics subtyping pipeline for chronic obstructive pulmonary disease. *Plos one*, 16(8):e0255337,  
605 2021.
- 606 [11] Yale Chang, Kimberly Glass, Yang-Yu Liu, Edwin K Silverman, James D Crapo, Ruth Tal-  
607 Singer, Russ Bowler, Jennifer Dy, Michael Cho, and Peter Castaldi. Copd subtypes identified by  
608 network-based clustering of blood gene expression. *Genomics*, 107(2-3):51–58, 2016.
- 609 [12] Zili Zhang, Jian Wang, Yuanyuan Li, Fei Liu, Lingdan Chen, Shunping He, Fanjie Lin, Xinguang  
610 Wei, Yaowei Fang, Qiongqiong Li, et al. Proteomics and metabolomics profiling reveal panels of  
611 circulating diagnostic biomarkers and molecular subtypes in stable copd. *Respiratory Research*,  
612 24(1):73, 2023.
- 613 [13] Andrew Gregory, Zhonghui Xu, Katherine Pratte, Sool Lee, Congjian Liu, Robert Chase, Jeong  
614 Yun, Aabida Saferali, Craig P Hersh, Russell Bowler, et al. Clustering-based copd subtypes have  
615 distinct longitudinal outcomes and multi-omics biomarkers. *BMJ Open Respiratory Research*, 9  
616 (1):e001182, 2022.
- 617 [14] Margaret F Ragland, Christopher J Benway, Sharon M Lutz, Russell P Bowler, Julian Hecker,  
618 John E Hokanson, James D Crapo, Peter J Castaldi, Dawn L DeMeo, Craig P Hersh, et al.  
619 Genetic advances in chronic obstructive pulmonary disease. insights from copdgene. *American  
620 journal of respiratory and critical care medicine*, 200(6):677–690, 2019.
- 621 [15] Jian Wang, Margaret R Spitz, Christopher I Amos, Anna V Wilkinson, Xifeng Wu, and Sanjay  
622 Shete. Mediating effects of smoking and chronic obstructive pulmonary disease on the relation  
623 between the chrna5-a3 genetic locus and lung cancer risk. *Cancer*, 116(14):3458–3462, 2010.
- 624 [16] Ivy Reichert Vital da Silva, Cintia Laura Pereira de Araujo, Gilson Pires Dorneles, Alessandra  
625 Peres, Andreia Luciana Bard, Gustavo Reinaldo, Paulo José Zimmermann Teixeira, Pedro Dal Lago,  
626 and Viviane Rostirola Elsner. Exercise-modulated epigenetic markers and inflammatory response  
627 in copd individuals: A pilot study. *Respiratory physiology & neurobiology*, 242:89–95, 2017.

- 628 [17] NM Siafakas and EG Tzortzaki. Few smokers develop copd. why? *Respiratory medicine*, 96(8):  
629 615–624, 2002.
- 630 [18] Peter J Castaldi, Marta Benet, Hans Petersen, Nicholas Rafaels, James Finigan, Matteo Paoletti,  
631 H Marike Boezen, Judith M Vonk, Russell Bowler, Massimo Pistolesi, et al. Do copd subtypes  
632 really exist? copd heterogeneity and clustering in 10 independent cohorts. *Thorax*, 72(11):  
633 998–1006, 2017.
- 634 [19] Chuan-Xing Li, Craig E Wheelock, C Magnus Sköld, and Åsa M Wheelock. Integration of  
635 multi-omics datasets enables molecular classification of copd. *European Respiratory Journal*, 51  
636 (5), 2018.
- 637 [20] Enrico Maiorino and Joseph Loscalzo. Phenomics and robust multiomics data for cardiovascular  
638 disease subtyping. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 43(7):1111–1123, 2023.
- 639 [21] Laura Cantini, Pooya Zakeri, Celine Hernandez, Aurelien Naldi, Denis Thieffry, Elisabeth Remy,  
640 and Anaïs Baudot. Benchmarking joint multi-omics dimensionality reduction approaches for the  
641 study of cancer. *Nature communications*, 12(1):124, 2021.
- 642 [22] Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C  
643 Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-omics factor analysis—a  
644 framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, 14  
645 (6):e8124, 2018.
- 646 [23] Ronglai Shen, Adam B Olshen, and Marc Ladanyi. Integrative clustering of multiple genomic  
647 data types using a joint latent variable model with application to breast and lung cancer subtype  
648 analysis. *Bioinformatics*, 25(22):2906–2912, 2009.
- 649 [24] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin  
650 Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a  
651 genomic scale. *Nature methods*, 11(3):333–337, 2014.
- 652 [25] Junxiang Chen, Xu Zhonghui, Li Sun, Ke Yu, Craig P Hersh, Adel Boueiz, John Hokanson,  
653 Frank C Sciruba, Edwin K Silverman, Peter J Castaldi, et al. Deep learning integration of chest  
654 ct imaging and gene expression identifies novel aspects of copd. *medRxiv*, pages 2022–09, 2022.
- 655 [26] Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika.  
656 Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology  
657 insights*, 14:1177932219899051, 2020.
- 658 [27] Evangelina López de Maturana, Lola Alonso, Pablo Alarcón, Isabel Adoración Martín-Antoniano,  
659 Silvia Pineda, Lucas Piorno, M Luz Calle, and Núria Malats. Challenges in the integration of  
660 omics and non-omics data. *Genes*, 10(3):238, 2019.
- 661 [28] Dongjin Leng, Linyi Zheng, Yuqi Wen, Yunhao Zhang, Lianlian Wu, Jing Wang, Meihong  
662 Wang, Zhongnan Zhang, Song He, and Xiaochen Bo. A benchmark study of deep learning-based  
663 multi-omics data fusion methods for cancer. *Genome Biology*, 23(1):1–32, 2022.
- 664 [29] Nikola Simidjievski, Cristian Bodnar, Ifrah Tariq, Paul Scherer, Helena Andres Terre, Zohreh  
665 Shams, Mateja Jamnik, and Pietro Liò. Variational autoencoders for cancer data integration:  
666 design principles and computational practice. *Frontiers in genetics*, 10:1205, 2019.
- 667 [30] Xiaoyu Zhang, Jingqing Zhang, Kai Sun, Xian Yang, Chengliang Dai, and Yike Guo. Integrated  
668 multi-omics analysis using variational autoencoders: application to pan-cancer classification.  
669 In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages  
670 765–769. IEEE, 2019.

- 671 [31] Tianwei Yu. Aime: Autoencoder-based integrative multi-omics data embedding that allows for  
672 confounder adjustments. *PLoS Computational Biology*, 18(1):e1009826, 2022.
- 673 [32] Elizabeth A Regan, John E Hokanson, James R Murphy, Barry Make, David A Lynch, Terri H  
674 Beaty, Douglas Curran-Everett, Edwin K Silverman, and James D Crapo. Genetic epidemiology  
675 of copd (copdgene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 7  
676 (1):32–43, 2011.
- 677 [33] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. Multimodal deep  
678 learning for biomedical data fusion: a review. *Briefings in Bioinformatics*, 23(2):bbab569, 2022.
- 679 [34] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng.  
680 Multimodal deep learning. In *Proceedings of the 28th international conference on machine  
681 learning (ICML-11)*, pages 689–696, 2011.
- 682 [35] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint  
683 arXiv:1312.6114*, 2013.
- 684 [36] Sergey E Golovenkin, Jonathan Bac, Alexander Chervov, Evgeny M Mirkes, Yuliya V Orlova,  
685 Emmanuel Barillot, Alexander N Gorban, and Andrei Zinovyev. Trajectories, bifurcations, and  
686 pseudo-time in large clinical datasets: Applications to myocardial infarction and diabetes data.  
687 *GigaScience*, 9(11):giaa128, 2020.
- 688 [37] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations  
689 and Trends® in Machine Learning*, 12(4):307–392, 2019.
- 690 [38] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
- 691 [39] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational  
692 fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.
- 693 [40] AJ Bell, S Ram, WW Labaki, S Murray, E Kazerooni, S Galban, FJ Martinez, C Hatt, JM Wang,  
694 E Mirkes, et al. Clinical trajectory analysis with longitudinal validation in copd: A copdgene  
695 study. In *D97. IMPACT OF COPD ACROSS THE LIFESPAN*, pages A6589–A6589. American  
696 Thoracic Society, 2023.
- 697 [41] Alexander Chervov and Andrei Zinovyev. Clinical trajectories estimated from bulk tumoral  
698 molecular profiles using elastic principal trees. In *2021 International Joint Conference on Neural  
699 Networks (IJCNN)*, pages 1–9. IEEE, 2021.
- 700 [42] John R. Hurst, Jørgen Vestbo, Antonio Anzueto, Nicholas Locantore, Hana Müllerova, Ruth  
701 Tal-Singer, Bruce Miller, David A. Lomas, Alvar Agusti, William MacNee, Peter Calverley,  
702 Stephen Rennard, Emiel F.M. Wouters, and Jadwiga A. Wedzicha. Susceptibility to exacerbation  
703 in chronic obstructive pulmonary disease. *New England Journal of Medicine*, 363(12):1128–1138,  
704 2010. doi:10.1056/NEJMoa0909883. URL <https://doi.org/10.1056/NEJMoa0909883>. PMID:  
705 20843247.
- 706 [43] Emily S Wan, Peter J Castaldi, Michael H Cho, John E Hokanson, Elizabeth A Regan, Barry J  
707 Make, Terri H Beaty, MeiLan K Han, Jeffrey L Curtis, Douglas Curran-Everett, et al. Epi-  
708 demiology, genetics, and subtyping of preserved ratio impaired spirometry (prism) in copdgene.  
709 *Respiratory research*, 15(1):1–13, 2014.
- 710 [44] Peter J Castaldi, Zhonghui Xu, Kendra A Young, John E Hokanson, David A Lynch, Stephen M  
711 Humphries, James C Ross, Michael H Cho, Craig P Hersh, James D Crapo, et al. Copd  
712 heterogeneity and progression: Emphysema-predominant and non-emphysema-predominant  
713 disease. *American Journal of Epidemiology*, page kwad114, 2023.

- 714 [45] Jing Shi and Michael G Walker. Gene set enrichment analysis (gsea) for interpreting gene  
715 expression profiles. *Current Bioinformatics*, 2(2):133–137, 2007.
- 716 [46] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo,  
717 and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):  
718 1739–1740, 2011.
- 719 [47] Sharon Mumby and Ian M Adcock. Recent evidence from omic analysis for redox signalling and  
720 mitochondrial oxidative stress in copd. *Journal of Inflammation*, 19(1):10, 2022.
- 721 [48] Barry R Imhoff and Jason M Hansen. Extracellular redox status regulates nrf2 activation  
722 through mitochondrial reactive oxygen species. *Biochemical Journal*, 424(3):491–500, 2009.
- 723 [49] Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by  
724 correlation of distances. 2007.
- 725 [50] Yonghua Zhuang, Fuyong Xing, Debashis Ghosh, Brian D Hobbs, Craig P Hersh, Farnoush  
726 Banaei-Kashani, Russell P Bowler, and Katerina Kechris. Deep learning on graphs for multi-omics  
727 classification of copd. *Plos one*, 18(4):e0284563, 2023.
- 728 [51] Junxiang Chen, Zhonghui Xu, Li Sun, Ke Yu, Craig P Hersh, Adel Boueiz, John E Hokanson,  
729 Frank C Sciruba, Edwin K Silverman, Peter J Castaldi, et al. Deep learning integration of chest  
730 computed tomography imaging and gene expression identifies novel aspects of copd. *Chronic  
731 Obstructive Pulmonary Diseases (Miami, Fla.)*, 2023.
- 732 [52] Alvar Agusti, Elisabeth Bel, Mike Thomas, Claus Vogelmeier, Guy Brusselle, Stephen Holgate,  
733 Marc Humbert, Paul Jones, Peter G Gibson, Jørgen Vestbo, et al. Treatable traits: toward  
734 precision medicine of chronic airway diseases. *European Respiratory Journal*, 47(2):410–419,  
735 2016.
- 736 [53] Junxiang Chen, Michael Cho, Edwin K Silverman, John E Hokanson, Greg L Kinney, James D  
737 Crapo, Stephen Rennard, Jennifer Dy, and Peter Castaldi. Turning subtypes into disease axes  
738 to improve prediction of copd progression. *Thorax*, 74(9):906–909, 2019.
- 739 [54] Mark T Dransfield, Ken M Kunisaki, Matthew J Strand, Antonio Anzueto, Surya P Bhatt,  
740 Russell P Bowler, Gerard J Criner, Jeffrey L Curtis, Nicola A Hanania, Hrudaya Nath, et al.  
741 Acute exacerbations and lung function loss in smokers with and without chronic obstructive  
742 pulmonary disease. *American journal of respiratory and critical care medicine*, 195(3):324–330,  
743 2017.
- 744 [55] David MG Halpin, Marc Decramer, Bartolome R Celli, Achim Mueller, Norbert Metzdorf, and  
745 Donald P Tashkin. Effect of a single exacerbation on decline in lung function in copd. *Respiratory  
746 medicine*, 128:85–91, 2017.
- 747 [56] David MG Halpin, Marc Decramer, Bartolome Celli, Steven Kesten, Dacheng Liu, and Donald P  
748 Tashkin. Exacerbation frequency and course of copd. *International journal of chronic obstructive  
749 pulmonary disease*, pages 653–661, 2012.
- 750 [57] GC Donaldson, Terence AR Seemungal, A Bhowmik, and JA1746193 Wedzicha. Relationship  
751 between exacerbation frequency and lung function decline in chronic obstructive pulmonary  
752 disease. *Thorax*, 57(10):847–852, 2002.
- 753 [58] Prescott G Woodruff, R Graham Barr, Eugene Bleecker, Stephanie A Christenson, David Couper,  
754 Jeffrey L Curtis, Natalia A Gouskova, Nadia N Hansel, Eric A Hoffman, Richard E Kanner,  
755 et al. Clinical significance of symptoms in smokers with preserved pulmonary function. *New  
756 England Journal of Medicine*, 374(19):1811–1821, 2016.

- 757 [59] Elizabeth A Regan, David A Lynch, Douglas Curran-Everett, Jeffrey L Curtis, John HM  
758 Austin, Philippe A Grenier, Hans-Ulrich Kauczor, William C Bailey, Dawn L DeMeo, Richard H  
759 Casaburi, et al. Clinical and radiologic disease in smokers with normal spirometry. *JAMA*  
760 *internal medicine*, 175(9):1539–1549, 2015.
- 761 [60] Brian D Hobbs and Craig P Hersh. Integrative genomics of chronic obstructive pulmonary  
762 disease. *Biochemical and biophysical research communications*, 452(2):276–286, 2014.
- 763 [61] Timothy M Bahr, Grant J Hughes, Michael Armstrong, Rick Reisdorph, Christopher D Coldren,  
764 Michael G Edwards, Christina Schnell, Ross Kedl, Daniel J LaFlamme, Nichole Reisdorph, et al.  
765 Peripheral blood mononuclear cell gene expression in chronic obstructive pulmonary disease.  
766 *American journal of respiratory cell and molecular biology*, 49(2):316–323, 2013.
- 767 [62] Aran Singanayagam, Su-Ling Loo, Maria Calderazzo, Lydia J Finney, Maria-Belen Trujillo Tor-  
768 ralbo, Eteri Bakhsoliani, Jason Girkin, Punnam Veerati, Prabuddha S Pathinayake, Kristy S  
769 Nichol, et al. Antiviral immunity is impaired in copd patients with frequent exacerbations.  
770 *American Journal of Physiology-Lung Cellular and Molecular Physiology*, 317(6):L893–L903,  
771 2019.
- 772 [63] Yessica D Torres-Ramos, Araceli Montoya-Estrada, Alberto M Guzman-Grenfell, Javier Mancilla-  
773 Ramirez, Beatriz Cardenas-Gonzalez, Salvador Blanco-Jimenez, Jose D Sepulveda-Sanchez,  
774 Alejandra Ramirez-Venegas, and Juan J Hicks. Urban pm<sub>2.5</sub> induces ros generation and rbc  
775 damage in copd patients. *Front Biosci (Elite Ed)*, 3:808–817, 2011.
- 776 [64] Aravind T Reddy, Sowmya P Lakshmi, Asoka Banno, and Raju C Reddy. Role of gpx3 in  
777 ppar $\gamma$ -induced protection against copd-associated oxidative stress. *Free Radical Biology and*  
778 *Medicine*, 126:350–357, 2018.
- 779 [65] I Rahman and IM Adcock. Oxidative stress and redox regulation of lung inflammation in copd.  
780 *European respiratory journal*, 28(1):219–242, 2006.
- 781 [66] Isaac K Sundar, Sangwoon Chung, Jae-Woong Hwang, Gnanapragasam Arunachalam, Suzanne  
782 Cook, Hongwei Yao, Witold Mazur, Vuokko L Kinnula, Aron B Fisher, and Irfan Rahman. Perox-  
783 iredoxin 6 differentially regulates acute and chronic cigarette smoke-mediated lung inflammatory  
784 response and injury. *Experimental lung research*, 36(8):451–462, 2010.
- 785 [67] Saleela M Ruwanpura, Louise McLeod, Alistair Miller, Jessica Jones, Ross Vlahos, Georg Ramm,  
786 Anthony Longano, Philip G Bardin, Steven Bozinovski, Gary P Anderson, et al. Deregulated  
787 stat3 signaling dissociates pulmonary inflammation from emphysema in gp130 mutant mice.  
788 *American Journal of Physiology-Lung Cellular and Molecular Physiology*, 302(7):L627–L639,  
789 2012.
- 790 [68] Liang Yew-Booth, Mark A Birrell, Ming Sum Lau, Katie Baker, Victoria Jones, Iain Kilty, and  
791 Maria G Belvisi. Jak–stat pathway activation in copd. *European Respiratory Journal*, 46(3):  
792 843–845, 2015.
- 793 [69] David Couper, Lisa M LaVange, MeiLan Han, R Graham Barr, Eugene Bleeker, Eric A Hoffman,  
794 Richard Kanner, Eric Kleerup, Fernando J Martinez, Prescott G Woodruff, et al. Design of the  
795 subpopulations and intermediate outcomes in copd study (spiromics). *Thorax*, 69(5):492–495,  
796 2014.
- 797 [70] Daniel Taliun, Daniel N Harris, Michael D Kessler, Jedidiah Carlson, Zachary A Szpiech, Raul  
798 Torres, Sarah A Gagliano Taliun, André Corvelo, Stephanie M Gogarten, Hyun Min Kang,  
799 et al. Sequencing of 53,831 diverse genomes from the nhlbi topmed program. *Nature*, 590(7845):  
800 290–299, 2021.

- 801 [71] Yohannes Tesfaigzi, Jeffrey L Curtis, Irina Petrache, Francesca Polverino, Farrah Kheradmand,  
802 Ian M Adcock, and Stephen I Rennard. Does copd originate from different cell types? *American*  
803 *Journal of Respiratory Cell and Molecular Biology*, (ja), 2023.
- 804 [72] Maor Sauler, John E McDonough, Taylor S Adams, Neeharika Kothapalli, Thomas Barnthaler,  
805 Rhiannon B Werder, Jonas C Schupp, Jessica Nouws, Matthew J Robertson, Cristian Coarfa,  
806 et al. Characterization of the copd alveolar niche using single-cell rna sequencing. *Nature*  
807 *communications*, 13(1):494, 2022.
- 808 [73] Qiqing Huang, Yuanyuan Wang, Lili Zhang, Wei Qian, Shaoran Shen, Jingshen Wang, Shuang-  
809 shuang Wu, Wei Xu, Bo Chen, Mingyan Lin, et al. Single-cell transcriptomics highlights  
810 immunological dysregulations of monocytes in the pathobiology of copd. *Respiratory Research*,  
811 23(1):367, 2022.
- 812 [74] Michael Love, Simon Anders, and Wolfgang Huber. Differential analysis of count data—the  
813 *deseq2* package. *Genome Biol*, 15(550):10–1186, 2014.
- 814 [75] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K  
815 Smyth. *limma* powers differential expression analyses for rna-sequencing and microarray studies.  
816 *Nucleic acids research*, 43(7):e47–e47, 2015.
- 817 [76] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert,  
818 Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander,  
819 et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide  
820 expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.
- 821 [77] Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. Gseapy: a comprehensive package for performing  
822 gene set enrichment analysis in python. *Bioinformatics*, 39(1):btac757, 2023.
- 823 [78] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Balancing learning and inference  
824 in variational autoencoders. In *Proceedings of the aaai conference on artificial intelligence*,  
825 volume 33, pages 5885–5892, 2019.
- 826 [79] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A  
827 kernel method for the two-sample-problem. *Advances in neural information processing systems*,  
828 19, 2006.
- 829 [80] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint*  
830 *arXiv:1412.6980*, 2014.
- 831 [81] Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Jonathan Ben-Tzur, Moritz  
832 Hardt, Benjamin Recht, and Ameet Talwalkar. A system for massively parallel hyperparameter  
833 tuning. *Proceedings of Machine Learning and Systems*, 2:230–246, 2020.
- 834 [82] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote:  
835 synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:  
836 321–357, 2002.
- 837 [83] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python  
838 toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine*  
839 *Learning Research*, 18(17):1–5, 2017. URL <http://jmlr.org/papers/v18/16-365.html>.
- 840 [84] Tomihisa Kamada, Satoru Kawai, et al. An algorithm for drawing general undirected graphs.  
841 *Information processing letters*, 31(1):7–15, 1989.
- 842 [85] Cameron Davidson-Pilon. lifelines: survival analysis in python. *Journal of Open Source Software*,  
843 4(40):1317, 2019. doi:10.21105/joss.01317. URL <https://doi.org/10.21105/joss.01317>.

844 [86] Gábor J Székely and Maria L Rizzo. Partial distance correlation with methods for dissimilarities.  
845 2014.

846 [87] Sambit Panda, Satish Palaniappan, Junhao Xiong, Eric W. Bridgeford, Ronak Mehta, Cencheng  
847 Shen, and Joshua T. Vogelstein. hyppo: A comprehensive multivariate hypothesis testing python  
848 package, 2020.

## 849 7 Author contributions

850 EM, PJC and KG conceptualized the study and designed the project. EM led the data analysis and  
851 code implementation. MDM, ZX, RTC, and JHY contributed to the data processing steps. MDM,  
852 PJC, and KG contributed to the data analysis. CPH, STW, EKS, PJC provided domain expertise  
853 for the analysis and interpretation of the results. EM was the lead author of the manuscript. All  
854 authors contributed to writing and revising the manuscript.

## 855 8 Competing interests

856 JHY received consulting fees from Bridge BioTherapeutics. PJC received consulting fees from Verona  
857 Pharmaceuticals and research support from Bayer and Sanofi, both outside of this work. In the past  
858 three years, EKS received grant support from Bayer and Northpond Laboratories. STW receives  
859 royalties from UpToDate and is on the Board of Histolix, a digital pathology company.

## 860 9 Acknowledgments

861 This work was supported by grants from the NHLBI (EM: K01HL166705; MDM: K25HL168157; STW:  
862 PO1HL132825; KG: R01HL155749). The COPDGene study (NCT00608764) is supported by grants  
863 from the NHLBI (U01HL089897 and U01HL089856) and by NIH contract 75N92023D00011 and by  
864 the COPD Foundation through contributions made to an Industry Advisory Committee that has  
865 included AstraZeneca, Bayer Pharmaceuticals, Boehringer-Ingelheim, Genentech, GlaxoSmithKline,  
866 Novartis, Pfizer and Sunovion.