Machine Learning approaches for the characterization of COPD: supplemental document

1. SUPPLEMENTARY TEXT

To perform the literature review, we used the following query:

```
"((GeneSymbol) AND ((Chronic Obstructive Lung Disease[MeSH Terms]) OR
(Chronic Obstructive Pulmonary Diseases[MeSH Terms]) OR (COAD[MeSH Terms]) OR
(COPD[MeSH Terms]) OR (Chronic Obstructive Airway Disease[MeSH Terms]) OR
(Chronic Obstructive Pulmonary Disease[MeSH Terms]) OR(Airflow Obstruction, Chronic[MeSH Terms]) OR
(Airflow Obstructions, Chronic[MeSH Terms]) OR (Chronic Airflow Obstructions[MeSH Terms]) OR
(Chronic Airflow Obstruction[MeSH Terms])))"
```

2. SUPPLEMENTARY FIGURES

A. Supplementary Figure 1

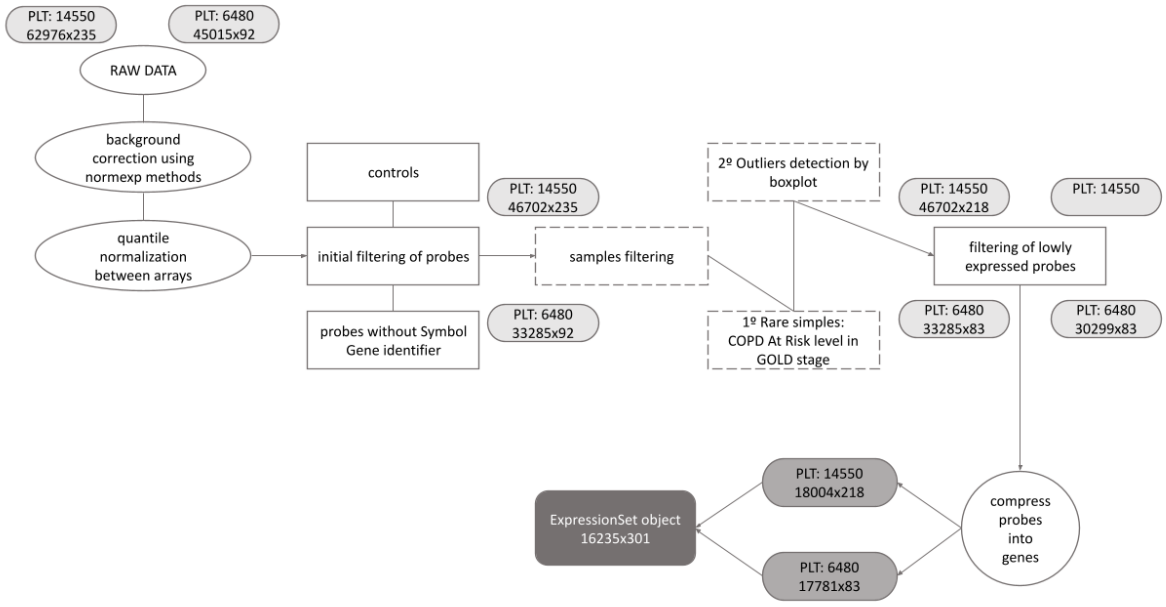

Fig. S1. Pipeline for processing microarray data. Collection of preprocessing steps applied to the two Agilent microarray platforms (PLT:14550 and PLT:6480) separately until their union. Ovals represent the first pre-processing steps that should be done over the microarray data (background correction using normexp method and quantile normalization between arrays). Rectangles show the filtering steps of the probes (continuous) and samples (dotted). First, filtering of control probes and those that have no correspondence with any GeneSymbol identifier (platform annotations were downloaded from GEO). After that, samples wrongly annotated (that is, COPD patients marked as being At Risk in the GOLD stage category) were deleted. Moreover, we used the Kolmogorov-Smirnov statistic to compare each array's intensity distribution and the distribution of the pooled data for obtaining the outliers samples. As this method bases its results on a simulated p-value, we generated 10000 randomizations and selected as outliers those samples that appear in at least 25% of the trials. Then, lowly expressed probes were also filtered, that is, probes with an expression count lower than half of the samples in the disease condition with fewer samples (> 42 in PLT:14550 and > 8 in PLT:6480). Finally, we compressed the probes into genes to obtain our final expression object. Note that light gray figures show the dimensions of each platform in terms of probes, the gray ones in terms of genes, and the dark gray figure shows the dimension of the final expression set object (once both platforms were joined).

B. Supplementary Figure 2

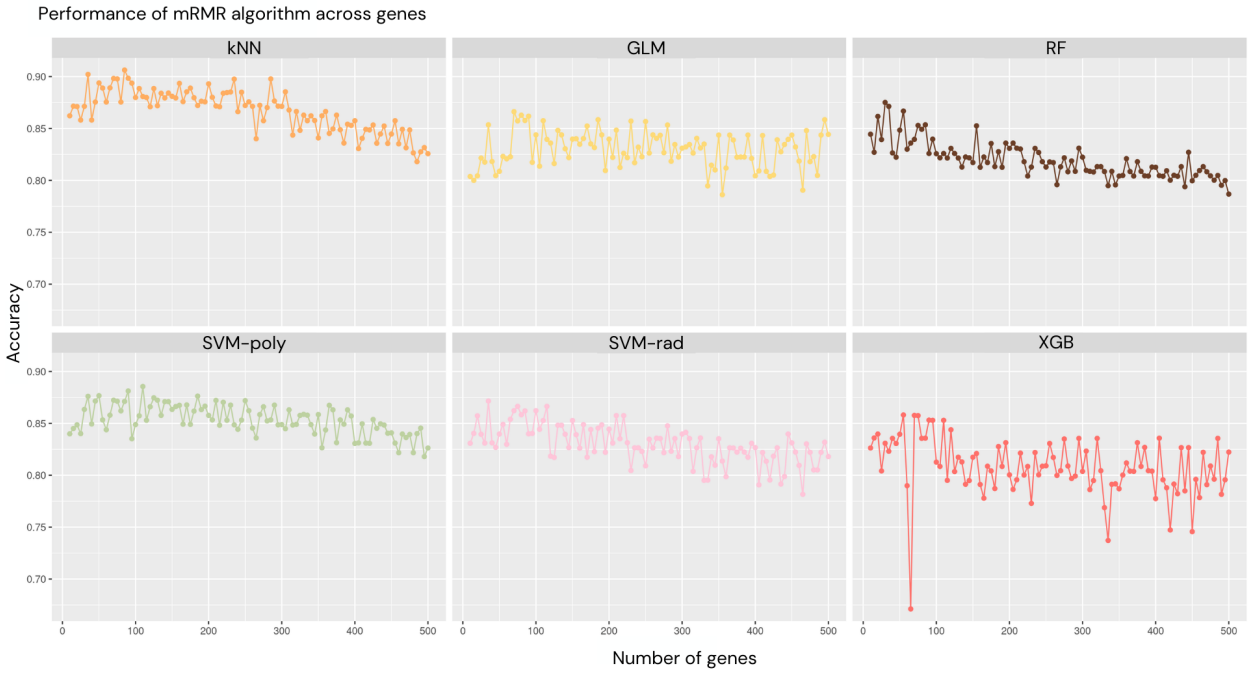

Fig. S2. mRMR accuracy performance across genes. These curves show how the prediction performance of the different classifiers (represented with different colors in the plot) built including as input the top k gene in the mRMR list, with $k \in [10, 500]$ behaves. The curves are based on the cross-validation accuracy achieved on the Bayes optimization tunning methodology.

C. Supplementary Figure 3

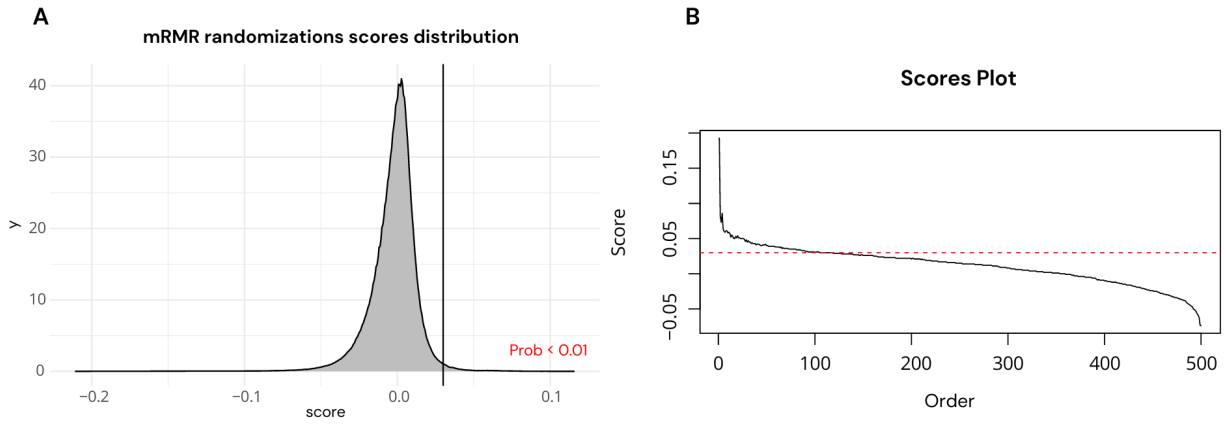

Fig. S3. mRMR randomization scores distribution and top 500 real scores. A shows the mRMR scores distribution obtained by running the mRMR algorithm on 1000 datasets with sample disease labels altered. The vertical line represents the threshold, a, for which the probability of a random variable, x, falling in the interval $[a, 1]$ is equal or less than 0.01. B shows the curve of the scores of the top 500 features the algorithm selects when applied to the real training data. The horizontal line represents the threshold $a = 0.03$ for identifying a gene as significant for distinguish between COPD and control samples.

D. Supplementary Figure 4

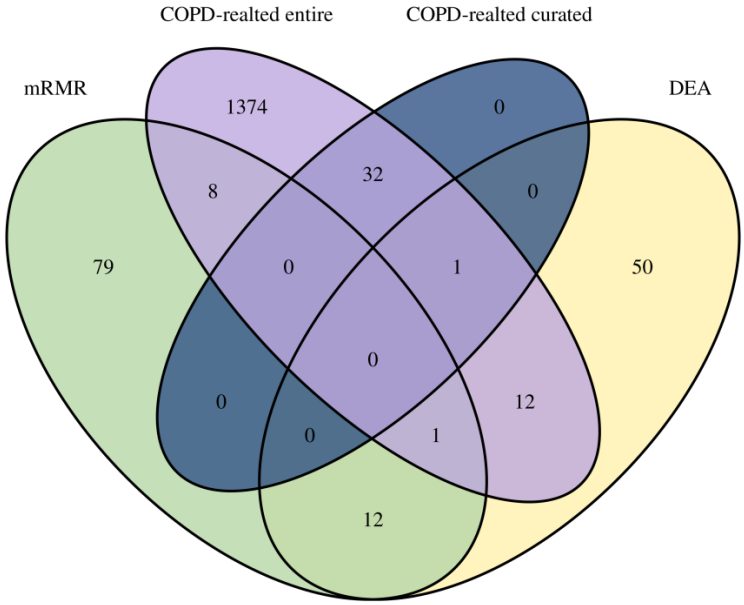

Fig. S4. Venn Diagram of the intersection among different "seed" gene lists.

E. Supplementary Figure 5

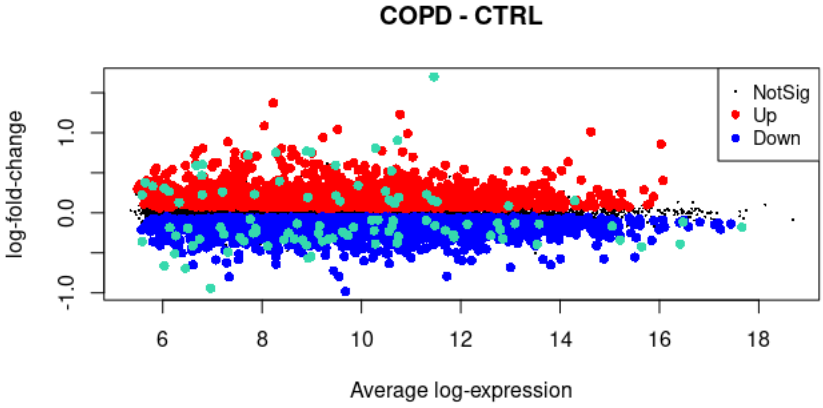

Fig. S5. Mean Difference (MD) plot of differentially expressed genes. MD displays log2 fold change versus average log2 expression values for all the genes (16235). Highlighted genes are significantly differentially expressed in COPD compared CTRL samples using fdr <0.05 cut-off. Upregulated genes are marked in red (1347), and downregulated ones in blue (2126). Green points correspond to the top 100 MRMR genes.

F. Supplementary Figure 6

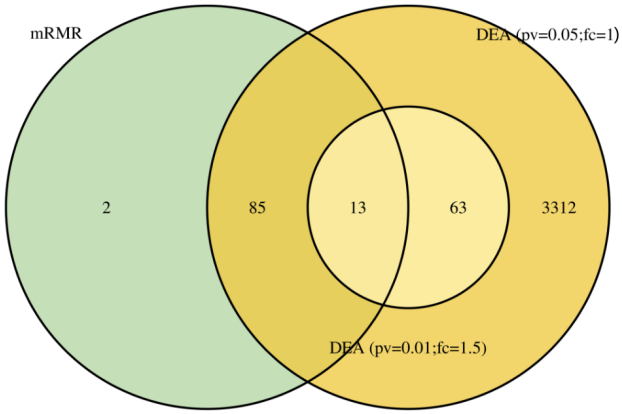

Fig. S6. Venn diagram showing the overlap between DEG considering different cut-offs and the top 100 mRMR genes. Most mRMR genes (98 out of 100) are consistent with the DEA gene list when the FDR and FC cut-offs are relaxed. This indicates that the mRMR method captures many genes identified as differentially expressed but with smaller fold changes or lower statistical significance.

G. Supplementary Figure 7

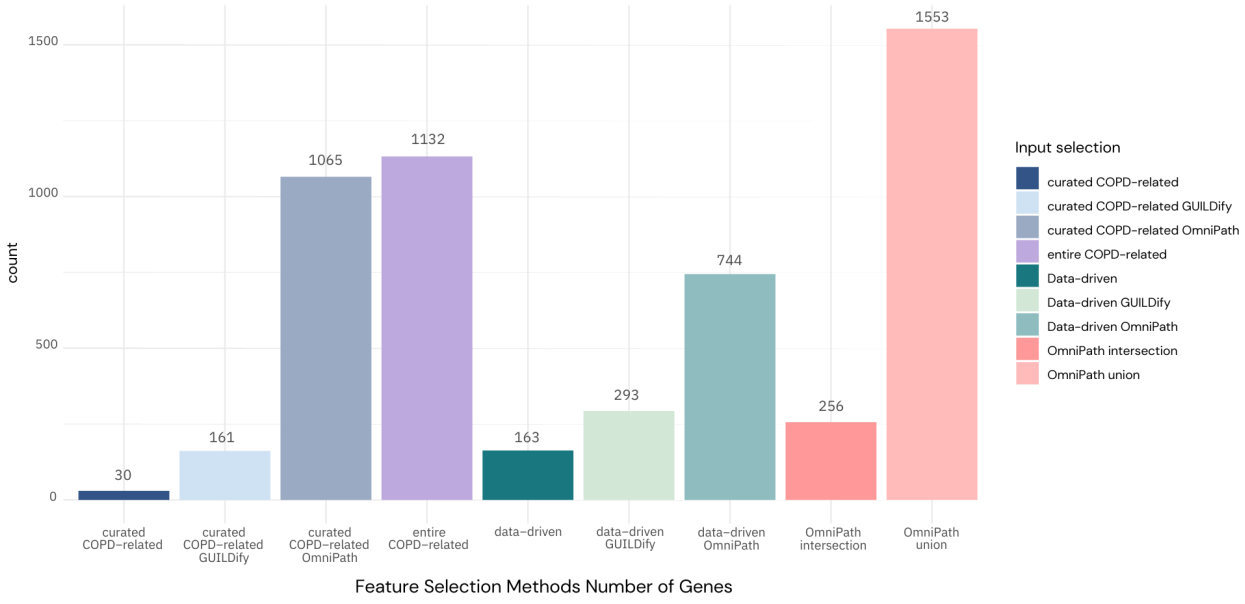

Fig. S7. Input gene sets. The barplot represents the nine different input sets and their respective sizes.

H. Supplementary Figure 8

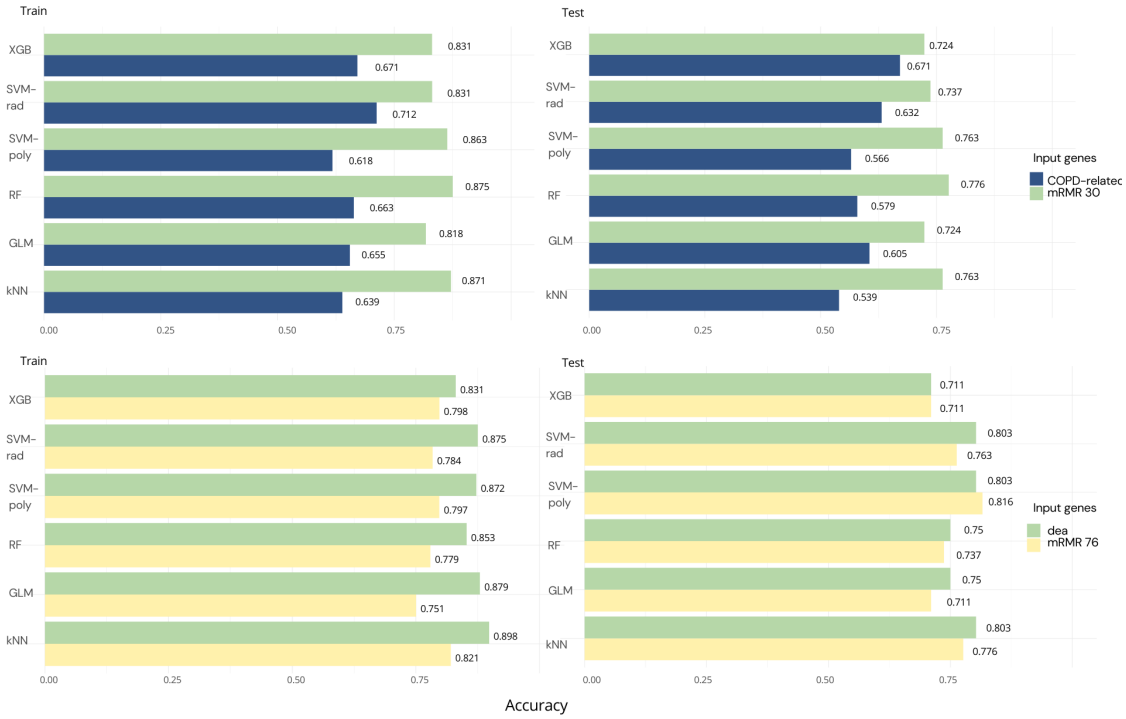

Fig. S8. Efficiency of mRMR genes. The top figures compare 30 COPD-related curated and top 30 mRMR genes cross-validation and test accuracies. Figures in the bottom contrast 76 DEA genes and the top 76 mRMR genes' cross-validation and test accuracies. The mRMR algorithm outperforms the accuracy of the DEA and COPD-related genes using the corresponding number of genes.

I. Supplementary Figure 9

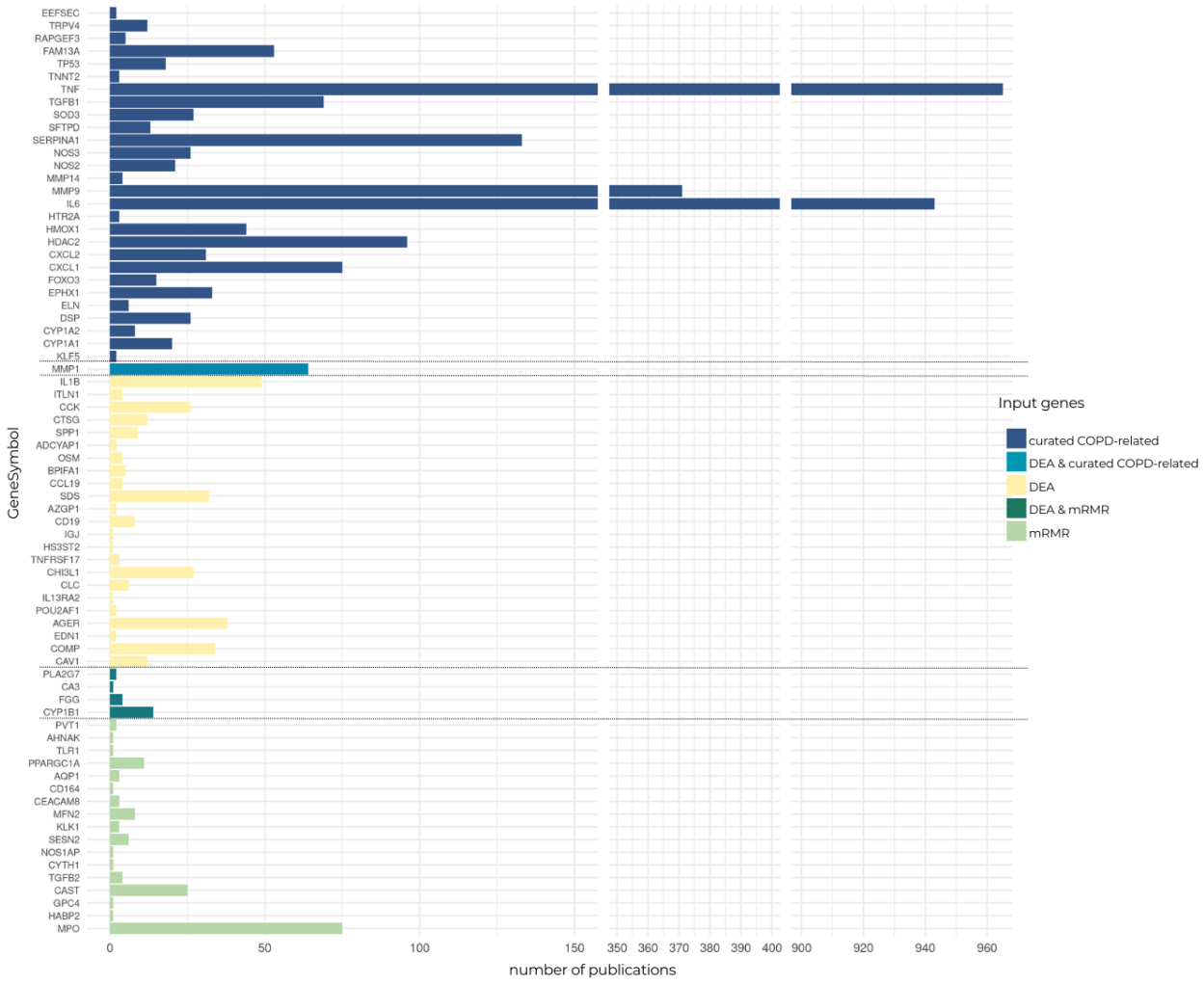

Fig. S9. Confirmation of the association of selected genes with COPD function by literature reviewing in PubMed Databank with the query 1. The figure shows the genes found to have publications cited along with COPD and the number of publications by a gene. Different colors represent the different seed gene sets (COPD-related curated, DEA, mRMR) and their intersections (COPD-related curated∪DEA, DEA∪mRMR). Dashed horizontal lines separate the different groups of genes as well.

J. Supplementary Figure 10

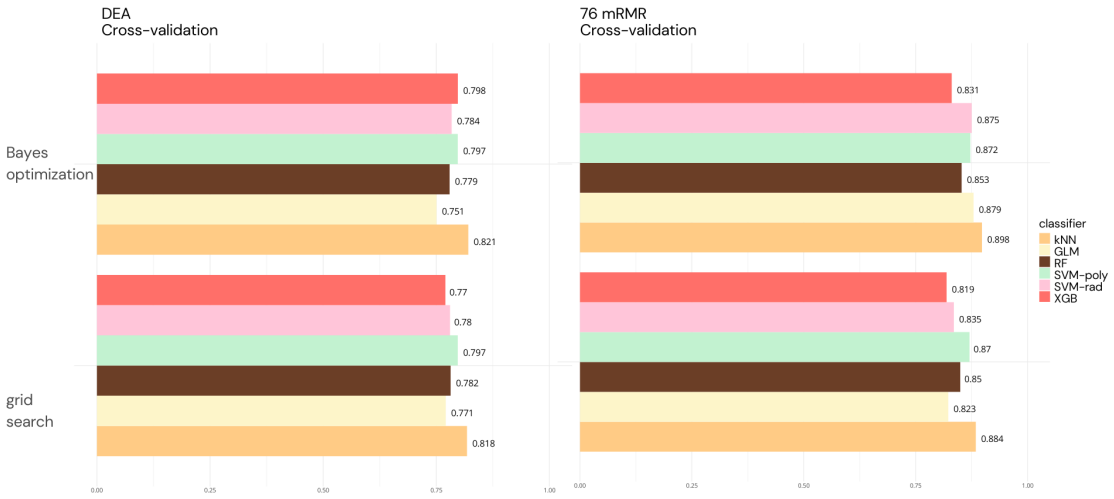

Fig. S10. Tunning methodologies comparison. Accuracy comparison between two tuning methodologies: Bayes optimization (top) and grid search (bottom). Cross-validation performance of DEA (left side) and 76 mRMR genes (right side) are shown.

K. Supplementary Figure 11

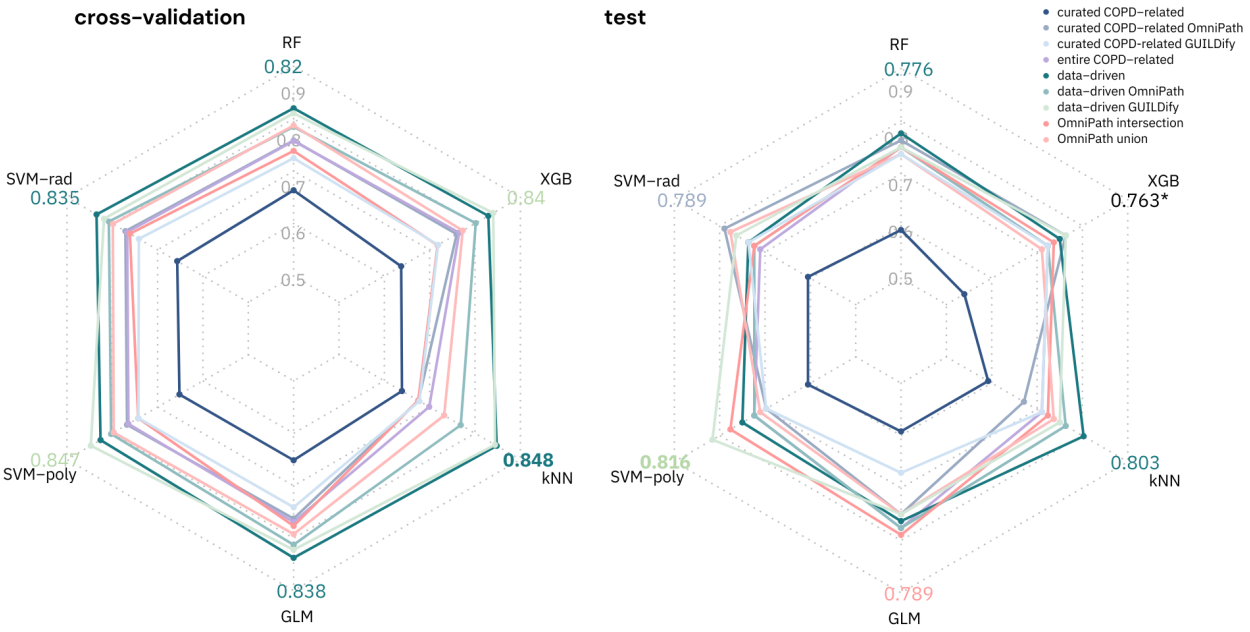

Fig. S11. Accuracy performance for the different ML models (vertices) colored by the input set of genes used. Left: cross-validation results. Right: test results. The highest performance accuracies are highlighted in bold (kNN on cross-validation and SVM-poly on test). *Black label means that the highest accuracy is achieved by more than one classifier (data-driven GUILDify and COPD-related curated OmniPath in this case).

L. Supplementary Figure 12

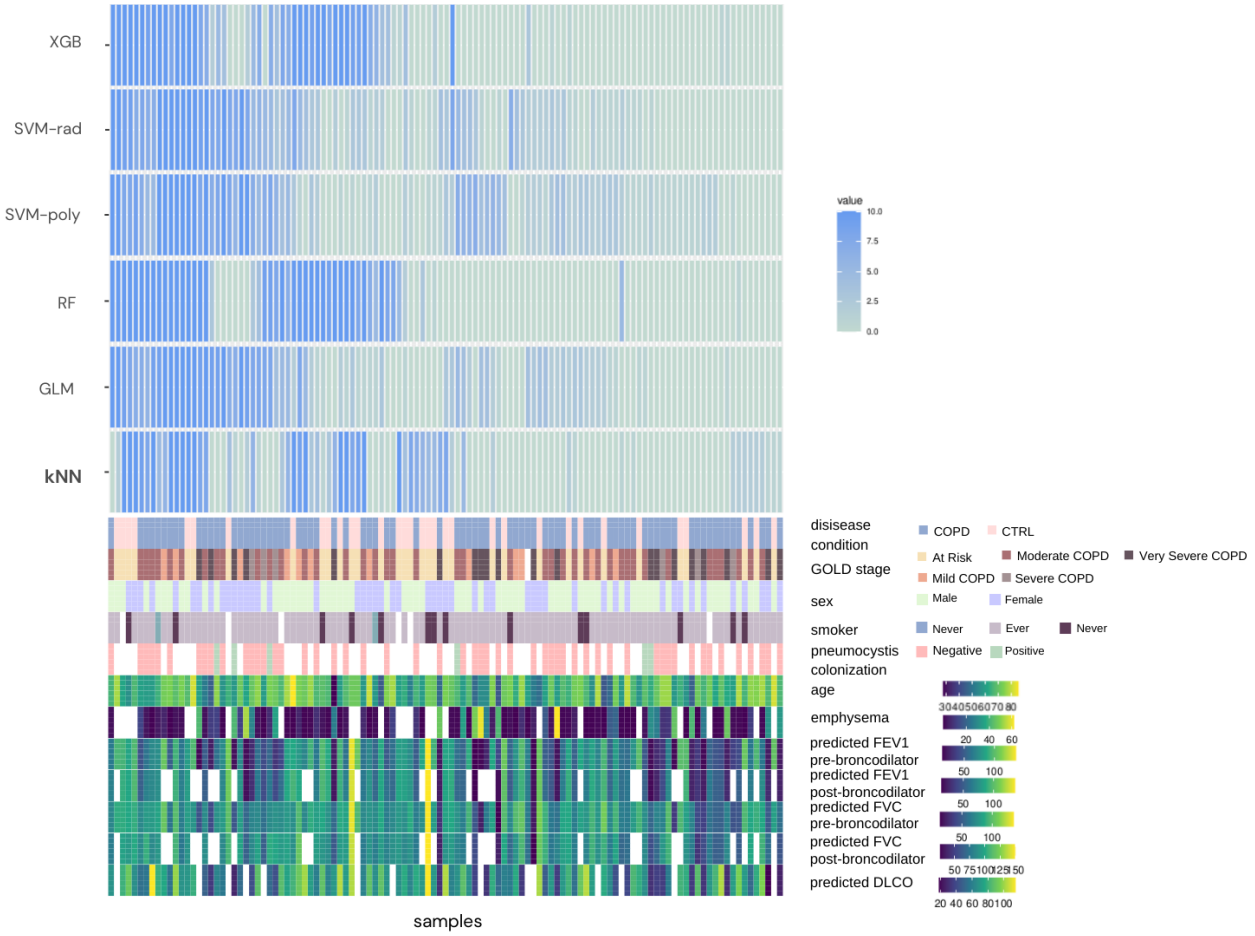

Fig. S12. Cross-validation misclassified samples and phenotypic variables representation. The heatmap illustrates the frequency (1 to 10) of misclassification for each sample across different methods using data-driven input. The best result, achieved with kNN using data-driven input, is highlighted. The sidebars display the phenotypic variables corresponding to the misclassified samples, where white space represents missing values. No pattern is observed among the misclassified samples.

M. Supplementary Figure 13

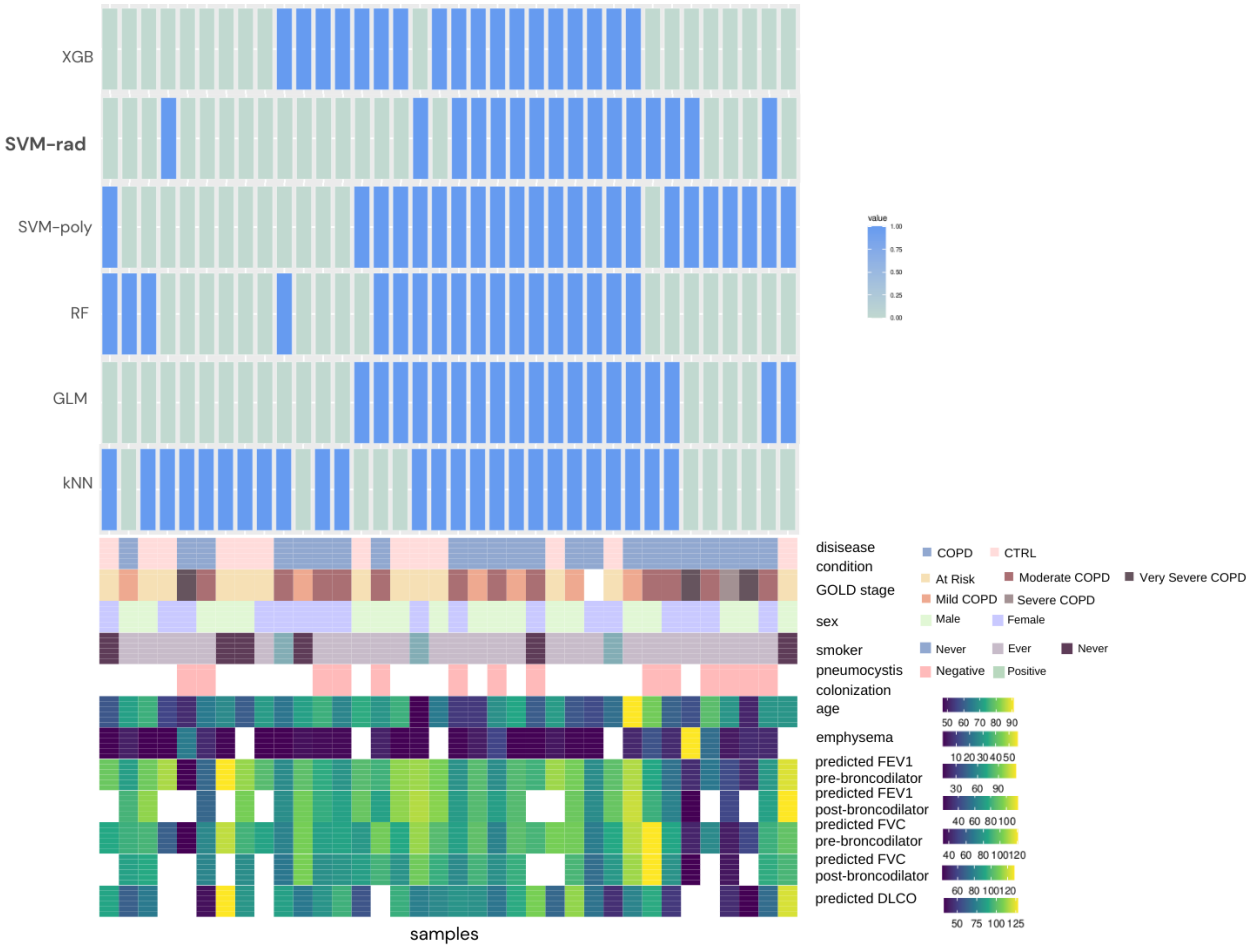

Fig. S13. Test misclassified samples and phenotypic variables representation. The heatmap illustrates all misclassified samples using curated COPD-related OmniPath input across different methods. The best result achieved with SVM-rad is highlighted. The sidebars display the phenotypic variables corresponding to the misclassified samples, where white space represents missing values. No pattern is observed among the misclassified samples.

N. Supplementary Figure 14

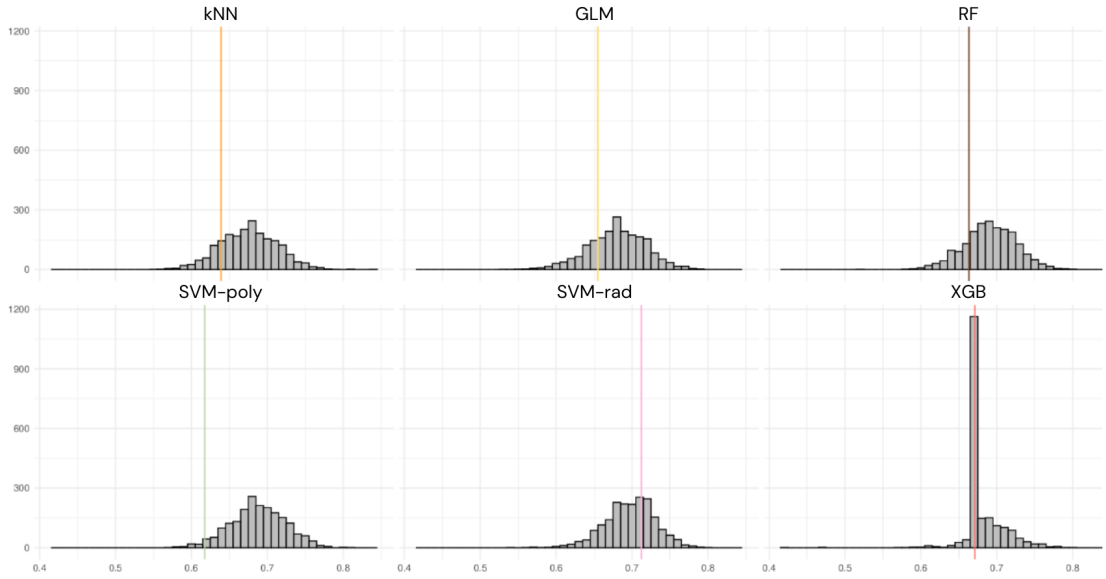

Fig. S14. Random performances' accuracy using 30 genes as input. The figure displays the distribution of accuracies from 1000 models that utilized 30 randomly selected genes as input. The vertical color lines indicate the performance of the 30 curated COPD-related genes for the different ML classifiers.

O. Supplementary Figure 15

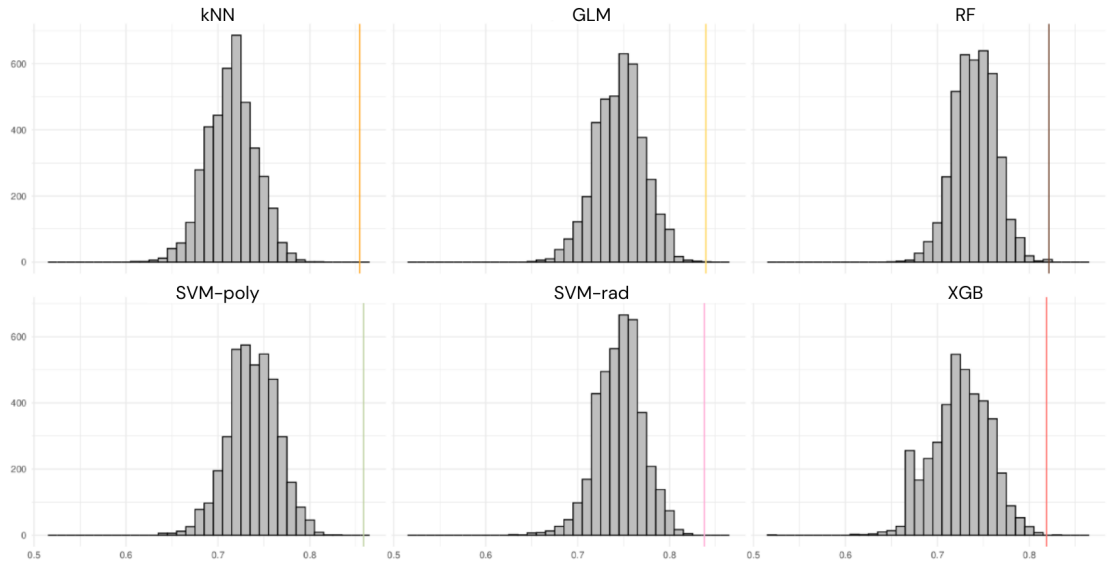

Fig. S15. Random performances' accuracy using 163 genes as input. The figure displays the distribution of accuracies from 1000 models that utilized 163 randomly selected genes as input. The vertical color lines indicate the performance of the 163 data-driven genes for the different ML classifiers.

3. SUPPLEMENTARY TABLES

A. Supplementary Table 1

Table S1. Collection of tunned hyperparameters of ML models.

Models	Tuning hyperparameters
RF	mtry (number of randomly selected predictors) min_n (minimal node size)
SVM-rad	cost (cost of predicting a sample within or on the wrong side of the margin) rbf_sigma (radial basis function)
SVM-poly	cost (cost of predicting a sample within or on the wrong side of the margin) degree (polynomial degree)
GLM	penalty (amount of regularization) mixture (proportion of Lasso Penalty)
kNN	neighbors (number of neighbors to consider) dist_power (parameter used in calculating Minkowski distance) weight_func (type of kernel function used to weight distances between samples)
XGB	tree_depth (tree depth) min_n (minimal node size) loss_reduction (minimum loss reduction) sample_size (proportion observations sampled) mtry (number of randomly selected predictors) learn_rate (learning rate)

B. Supplementary Table 2

Table S2: Ranking of cross-validation results for all classifiers and input sets. The table displays the performance rankings of different classifiers across all input sets based on cross-validation results.

Classifiers ranking normMCC in cross-validation

	classifier	metric	mean	sd	ML input
1	kNN	normMCC	0.834	0.147	data-driven
2	GLM	normMCC	0.833	0.154	data-driven
3	kNN	normMCC	0.828	0.153	data-driven GUILDify
4	SVM-rad	normMCC	0.827	0.159	data-driven
5	SVM-poly	normMCC	0.824	0.160	data-driven GUILDify
6	XGB	normMCC	0.824	0.150	data-driven GUILDify
7	SVM-poly	normMCC	0.821	0.149	data-driven
8	GLM	normMCC	0.818	0.146	data-driven GUILDify
9	XGB	normMCC	0.817	0.161	data-driven
10	SVM-rad	normMCC	0.811	0.167	data-driven GUILDify
11	RF	normMCC	0.808	0.156	data-driven
12	GLM	normMCC	0.806	0.168	data-driven OmniPath
13	SVM-rad	normMCC	0.799	0.160	data-driven OmniPath
14	RF	normMCC	0.798	0.174	data-driven GUILDify
15	SVM-poly	normMCC	0.795	0.175	data-driven OmniPath
16	XGB	normMCC	0.792	0.173	data-driven OmniPath
17	SVM-rad	normMCC	0.789	0.176	OmniPath union
18	GLM	normMCC	0.787	0.165	OmniPath union
19	SVM-poly	normMCC	0.786	0.175	OmniPath union
20	RF	normMCC	0.778	0.153	OmniPath union
21	RF	normMCC	0.775	0.171	data-driven OmniPath
22	GLM	normMCC	0.770	0.172	OmniPath intersection
23	XGB	normMCC	0.761	0.195	OmniPath union
24	GLM	normMCC	0.761	0.195	entire COPD-related
25	SVM-rad	normMCC	0.758	0.184	curated COPD-related OmniPath
26	SVM-rad	normMCC	0.758	0.183	entire COPD-related
27	SVM-rad	normMCC	0.757	0.173	OmniPath intersection
28	kNN	normMCC	0.756	0.154	data-driven OmniPath
29	SVM-poly	normMCC	0.756	0.175	entire COPD-related
30	SVM-poly	normMCC	0.755	0.185	curated COPD-related OmniPath
31	GLM	normMCC	0.751	0.192	curated COPD-related OmniPath
32	XGB	normMCC	0.751	0.189	entire COPD-related
33	RF	normMCC	0.748	0.183	curated COPD-related OmniPath
34	XGB	normMCC	0.747	0.169	curated COPD-related OmniPath
35	RF	normMCC	0.744	0.191	entire COPD-related

Continued on next page

Table S2: Ranking of cross-validation results for all classifiers and input sets. The table displays the performance rankings of different classifiers across all input sets based on cross-validation results. (Continued)

36	SVM-poly	normMCC	0.738	0.209	curated COPD-related GUILDify
37	SVM-rad	normMCC	0.738	0.190	curated COPD-related GUILDify
38	RF	normMCC	0.736	0.168	OmniPath intersection
39	SVM-poly	normMCC	0.734	0.190	OmniPath intersection
40	GLM	normMCC	0.734	0.208	curated COPD-related GUILDify
41	kNN	normMCC	0.724	0.178	OmniPath union
42	RF	normMCC	0.716	0.173	curated COPD-related GUILDify
43	XGB	normMCC	0.714	0.189	OmniPath intersection
44	XGB	normMCC	0.708	0.196	curated COPD-related GUILDify
45	kNN	normMCC	0.703	0.175	entire COPD-related
46	kNN	normMCC	0.686	0.153	curated COPD-related GUILDify
47	kNN	normMCC	0.683	0.184	OmniPath intersection
48	kNN	normMCC	0.681	0.176	curated COPD-related OmniPath
49	kNN	normMCC	0.659	0.168	curated COPD-related
50	SVM-rad	normMCC	0.656	0.167	curated COPD-related
51	SVM-poly	normMCC	0.655	0.186	curated COPD-related
52	GLM	normMCC	0.651	0.196	curated COPD-related
53	RF	normMCC	0.649	0.190	curated COPD-related
54	XGB	normMCC	0.633	0.181	curated COPD-related

C. Supplementary Table 3

Table S3: Ranking of test results for all classifiers and input sets. The table displays the performance rankings of different classifiers across all input sets based on test results.

Classifiers ranking normMCC in test

	classifier	metric	estimate	ML input
1	SVM-rad	normMCC	0.786601258	curated COPD-related OmniPath
2	SVM-poly	normMCC	0.7825304776	data-driven GUILDify
3	GLM	normMCC	0.779308454	OmniPath intersection
4	kNN	normMCC	0.7788480938	data-driven
5	GLM	normMCC	0.7686866453	data-driven OmniPath
6	SVM-rad	normMCC	0.7686866453	OmniPath union
7	RF	normMCC	0.7615982529	data-driven
8	XGB	normMCC	0.7583500315	curated COPD-related OmniPath
9	GLM	normMCC	0.7583500315	data-driven
10	SVM-rad	normMCC	0.7583500315	data-driven GUILDify
11	GLM	normMCC	0.7551477483	entire COPD-related
12	SVM-poly	normMCC	0.7493321396	OmniPath intersection
13	RF	normMCC	0.7482707813	entire COPD-related

Continued on next page

Table S3: Ranking of test results for all classifiers and input sets. The table displays the performance rankings of different classifiers across all input sets based on test results. (Continued)

14	SVM-poly	normMCC	0.7482707813	data-driven
15	GLM	normMCC	0.7482707813	data-driven GUILDify
16	RF	normMCC	0.7438809943	curated COPD-related OmniPath
17	XGB	normMCC	0.7438809943	data-driven GUILDify
18	XGB	normMCC	0.7403252772	data-driven
19	RF	normMCC	0.7403252772	data-driven OmniPath
20	kNN	normMCC	0.7375309217	data-driven OmniPath
21	GLM	normMCC	0.7329506056	curated COPD-related OmniPath
22	GLM	normMCC	0.7329506056	OmniPath union
23	RF	normMCC	0.7329506056	data-driven GUILDify
24	kNN	normMCC	0.7329506056	data-driven GUILDify
25	SVM-rad	normMCC	0.730098805	data-driven
26	RF	normMCC	0.730098805	curated COPD-related GUILDify
27	SVM-rad	normMCC	0.7287838454	data-driven OmniPath
28	SVM-poly	normMCC	0.7287838454	data-driven OmniPath
29	RF	normMCC	0.7261189626	OmniPath intersection
30	RF	normMCC	0.7223220311	OmniPath union
31	SVM-rad	normMCC	0.7223220311	curated COPD-related GUILDify
32	SVM-poly	normMCC	0.7201087225	entire COPD-related
33	SVM-rad	normMCC	0.7193303401	entire COPD-related
34	XGB	normMCC	0.7150552469	OmniPath intersection
35	XGB	normMCC	0.7119639091	entire COPD-related
36	XGB	normMCC	0.7119639091	data-driven OmniPath
37	SVM-rad	normMCC	0.7119639091	OmniPath intersection
38	SVM-poly	normMCC	0.7103300029	OmniPath union
39	kNN	normMCC	0.7082726558	OmniPath union
40	kNN	normMCC	0.7043029583	OmniPath intersection
41	kNN	normMCC	0.7018475786	curated COPD-related GUILDify
42	XGB	normMCC	0.6970901769	curated COPD-related GUILDify
43	XGB	normMCC	0.6938288197	OmniPath union
44	kNN	normMCC	0.6862294995	entire COPD-related
45	SVM-poly	normMCC	0.683602541	curated COPD-related OmniPath
46	SVM-poly	normMCC	0.675655311	curated COPD-related GUILDify
47	GLM	normMCC	0.6552411728	curated COPD-related GUILDify
48	kNN	normMCC	0.6470076637	curated COPD-related OmniPath
49	kNN	normMCC	0.6077498111	curated COPD-related
50	RF	normMCC	0.5980196059	curated COPD-related
51	SVM-poly	normMCC	0.5979568555	curated COPD-related
52	SVM-rad	normMCC	0.5888426727	curated COPD-related
53	GLM	normMCC	0.5885614886	curated COPD-related

Continued on next page

Table S3: Ranking of test results for all classifiers and input sets. The table displays the performance rankings of different classifiers across all input sets based on test results. (Continued)

54	XGB	normMCC	0.5615118652	curated COPD-related

D. Supplementary Table 4

Table S4. Ranking of ML Models based on normMCC metric using data-driven input. The table presents the ranking of ML models based on the normMCC metric and additional performance metrics, including MCC, accuracy, sensitivity, specificity, and rocAUC. The models are evaluated using data-driven input.

cross-validation	metrics (data-driven input)					
	MCC	normMCC	Accuracy	Sensitivity	Specificity	rocAUC
Ranking MCC						
kNN	**0.667**	**0.834**	0.848	0.876	0.802	0.913
GLM	**0.666**	**0.833**	0.838	0.823	0.876	0.922
SVM-rad	**0.654**	**0.827**	0.835	0.822	0.863	0.923
SVM-poly	**0.642**	**0.821**	0.826	0.811	0.866	0.912
XGB	**0.634**	**0.817**	0.829	0.837	0.818	0.911
RF	**0.616**	**0.808**	0.820	0.824	0.820	0.915

E. Supplementary Table 5

Table S5. Ranking of ML Models based on normMCC metric using curated COPD-related OmniPath input. The table presents the ranking of ML models based on the normMCC metric and additional performance metrics, including MCC, accuracy, sensitivity, specificity, and rocAUC. The models are evaluated using curated COPD-related OmniPath input.

test	metrics (curated COPD-related OmniPath input)					
	MCC	normMCC	Accuracy	Sensitivity	Specificity	rocAUC
Ranking MCC						
SVM-rad	**0.573**	**0.787**	0.789	0.765	0.84	0.828
XGB	**0.517**	**0.758**	0.763	0.745	0.8	0.817
RF	**0.488**	**0.744**	0.763	0.784	0.72	0.783
GLM	**0.466**	**0.733**	0.75	0.765	0.72	0.816
SVM-poly	**0.367**	**0.684**	0.697	0.706	0.68	0.801
kNN	**0.294**	**0.647**	0.671	0.706	0.6	0.729

F. Supplementary Table 6

Table S6. Enrichment of kNN misclassified samples. This table provides a summary of the enrichment analysis conducted on the misclassified samples generated by the kNN algorithm during cross-validation. Only samples misclassified in more than 5 folds are included in the analysis. Samples with p-value < 0.05 are considered significantly enriched. (*DLCO: Diffusing Capacity of the Lung for Carbon monoxide)

kNN cross-validation (missclassified in more than 5 folds)	
Variable	p-value
Disease condition	0.4851
GOLD stage	0.6062
Sex	0.5741
Smoker	0.6442
Age	0.482
Pneumocystis Colonization	0.07803
Emphysema	0.419
Predicted DLCO*	0.6186

G. Supplementary Table 7

Table S7. Enrichment of SVM-rad misclassified samples. This table provides a summary of the enrichment analysis conducted on the misclassified samples generated by the SVM-rad algorithm during test. Samples with p-value < 0.05 are considered significantly enriched. (*DLCO: Diffusing Capacity of the Lung for Carbon monoxide)

SVM-rad test	
Variable	p-value
Disease condition	0.6476
GOLD stage	0.08346
Sex	0.523
Smoker	0.6297
Age	0.216
Pneumocystis Colonization	1
Emphysema	0.6272
Predicted DLCO*	**0.0386**