Precision medicine for pandemics: stratification of COVID-19 molecular phenotypes defined by topological analysis of global blood gene expression.

Rebekah Penrice-Randal [1,2,*], Fabio Strazzeri [1], Benoit Ernst [3], Brice Van Eeckhout [4], Julien Guiot [3], Anna Julie Peired [5], Cosimo Nardi [5], Erika Parkinson [1], Monique Henket [3], Alicia Staderoli [3], Elora Guglielmi [3], Laurie Giltay [3], Rebecca Baker [7], Kit Howard [7], Catherine Hartley [2], Tessa Prince [2], Thomas Kleyntssens [4], Tommaso Manciulli [8], Ratko Djukanovic [9,10], Tristan Clark [9,10], Diana Baralle [10,11], Scott S Wagers [12], Xiaodan Xing [13], Yang Nan [13], Shiyi Wang [13], Guang Yang [13], Paul J Skipp [1,15], Julian A Hiscox [2,16,17], James P R Schofield [1].

Affiliations:

[1]TopMD Precision Medicine Ltd, Southampton, United Kingdom

[2]Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool, UK.

[3]Department of Respiratory Medicine, University Hospital of Liège, Liège, Belgium.

[4]Comunicare Solutions, Liège, Belgium.

[5]Department of Experimental and Clinical Biomedical Sciences "Mario Serio", University of Florence, Viale Morgagni 50, 50134 Florence, Italy.

[6]Department of Clinical and Experimental Medicine, Careggi University Hospital, Largo Brambilla 3, 50134 Florence, IT

[7]Clinical Data Interchange Standards Consortium (CDISC), Austin, TX, United States of America

[8]Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy.

[9]Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Southampton, United Kingdom

[10]National Institute for Health Research Southampton Biomedical Research Centre, Southampton, United Kingdom

[11]School of Human Development and Health, Faculty of Medicine, University of Southampton, Southampton, United Kingdom

[12]BioSci Consulting, Maasmechelen, Belgium

[13]Department of Bioengineering and Imperial-X, Imperial College London

[14]National Heart and Lung Institute, Imperial College London

[15]School of Biological Sciences, University of Southampton, Southampton, UK
[16]NIHR Health Protection Research Unit in Emerging and Zoonotic Infections, Liverpool, UK.
[17]A*STAR Infectious Diseases Laboratories (A*STAR ID Labs), Agency for Science, Technology and Research (A*STAR), Singapore.
***Correspondence:** Rebekah Penrice-Randal
R.Penrice-Randal@liverpool.ac.uk

Abstract:

Precision medicine offers a promising pathway for improving therapeutic responses to pandemics like COVID-19. This study utilises independent patient cohorts from Florence and Liège, collected under the DRAGON consortium, to classify molecular phenotypes associated with COVID-19 through topological analysis of whole blood gene expression. Samples from 173 patients were collected, and RNA was sequenced using the Novaseq platform. Molecular phenotypes were identified through topological analysis of gene expression in relation to the biological network using the TopMD algorithm. Clustering of patients' topological maps of differential pathway activation uncovered three distinct molecular phenotypes of COVID-19 in the Florence cohort, which were also found in the Liège cohort.

Cluster 1 was characterised by high activation of pathways linked to ESC pluripotency, NRF2, and TGF-β receptor signalling. Cluster 2 showed high activation of pathways, including focal adhesion-PI3K-Akt-mTOR signalling and type I interferon induction and signalling, while Cluster 3 displayed low IRF7-related pathway activation. TopMD was also used alongside the Drug-Gene Interaction Database (DGIdb), revealing pharmaceutical interventions that target mechanisms across multiple phenotypes and individuals.

The data demonstrate the usefulness of molecular phenotyping through topological analysis of blood gene expression and hold promise for guiding personalised therapeutic strategies, not only for COVID-19 but also for Disease X. Its potential for transferability across various diseases emphasises its value in pandemic response efforts, providing insights before large-scale clinical studies are undertaken.

65 **Introduction:**

66 The ongoing challenges of COVID-19, caused by the emergence of SARS-CoV-2, require a
67 thorough understanding of disease heterogeneity. Despite extensive research characterising
68 the host response to SARS-CoV-2 through pre-clinical and clinical [1,2]functional genomic data,
69 few approaches have utilised data that cover the full spectrum of symptom severity and
70 disease heterogeneity, and that deliver personalised medicine.

71 Examination of gene expression patterns in blood has been used in previous studies to
72 identify molecular phenotypes associated with different disease profiles in several emerging
73 viral infections, including Ebola virus (EBOV) [3] and SARS-CoV-2 [1,2,4-6], as well as more endemic
74 infections such as influenza virus[5]. Medical countermeasures focus on either reducing viral
75 load through anti-virals. These target viral biology or modulate the host response to infection,
76 thereby decreasing sequelae such as inflammation. With any emerging viral pathogen, direct-
77 acting antivirals take time to develop and undergo trials. Identifying therapeutics that can
78 modulate the host response to reduce symptomatology remains a priority. Being able to
79 characterise aberrations in host pathways that lead to disease quickly and integrating this
80 with therapeutics on the FDA-approved list will improve pandemic preparedness and rapid
81 response. Therefore, a deeper understanding of the host response can guide the selection of
82 host-directed medication countermeasures.

83 The field of digital health and precision medicine is rapidly evolving, with emerging
84 technologies and initiatives aimed at integrating diverse datasets to inform clinical decision-
85 making. In this study, we present a novel approach to analysing complex data collected by the
86 DRAGON international consortium, which enables the rapid identification of targets for
87 treatment by novel and/or repurposed drugs. Within DRAGON, efforts have been made to
88 harmonise data in digital healthcare, to identify biomarkers[7] and to propose guidelines for
89 the integration of clinical data from various modalities[8]. Additionally, an online platform has
90 been developed to host validated COVID-19 predictive models, facilitating their utilisation by
91 clinicians in real-time decision-making[9]. However, challenges persist, as evidenced by the
92 limited success of outcome prediction models for COVID-19 patients based on demographic
93 and comorbidity data, which highlights the need for more sophisticated approaches[10].

94 While omics data have been instrumental in advancing our understanding of SARS-CoV-2 and
95 COVID-19, their integration into digital health platforms for clinical decision-making remains
96 limited[11-13]. Traditional molecular phenotyping approaches often provide only shallow

insights. In previous work, using topological analysis, we demonstrated how gene expression data derived from whole blood at the time of admission could predict ICU admission[6]. However, the current study analysed the blood transcriptomes of patients with COVID-19 as part of the DRAGON-EU consortium and used TopMD. This algorithm considers all available data across a landscape of pathways to characterise molecular phenotypes of COVID-19 patients admitted to the hospital. Pathways were identified that correlated with clinical disease in the patient cohort. TopMD mapped pathways onto a database containing information on FDA-approved drugs and their known gene and pathway interactions to generate a list of potential therapeutics for modulating the severity of COVID-19. The ability to rapidly identify and therapeutically modulate host pathways responsible for disease, utilising pre-existing medical countermeasures, will be crucial in responding to the emergence of novel diseases and future pandemics.

This study describes an analysis of the blood transcriptomes of patients with COVID-19 admitted to hospitals in Liège and Florence between February and July 2021, as part of the DRAGON-EU consortium. Alongside collecting blood samples, demographic and clinical observations were recorded; additionally, CT scan data were obtained for a subset of these patients. We applied an unsupervised approach, in which we characterised the molecular phenotypes of patients within this cohort. We have previously reported the development of a gene signature in patients with COVID-19 that predicts admission to the ICU[6]. This predictive signature revealed the activation of pathways regulating epidermal growth factor receptor (EGFR) signalling, peroxisome proliferator-activated receptor alpha (PPAR-α) signalling and transforming growth factor beta (TGF-β) signalling. The observed molecular phenotype aligns with the mechanisms implicated in pulmonary fibrosis, which is also associated with increased severity of disease[14-16].

Methods

Study population and sample collection and ethics

Blood samples were collected between February and July 2021 from 132 COVID-19 patients admitted to Careggi University Hospital in Florence and 41 from a predefined cohort in Liège. All tested positive for SARS-CoV-2 via nasopharyngeal PCR, and samples were drawn on Day 0 of hospital admission. The study was approved by the ethics committees of the University

Hospital of Liège (2021/89) and UNIFI (#18085/OSS), with informed consent obtained from all participants.

Ethical Approval statement

All procedures adhered to the Declaration of Helsinki, relevant laws, and institutional guidelines, with approval from the appropriate ethics committees. Informed consent was obtained from all participants. Clinical data—including age, sex, BMI, and comorbidities—were extracted from electronic medical records and formatted according to the CDISC Study Data Tabulation Model (SDTM).

Chest CT analysis

Chest CT data were available for 109 of the 173 patients with RNA-seq data. Scans were acquired using a 128-detector multislice Spiral CT (Somatom Definition AS, Siemens Healthcare, Erlangen, Germany) with the following parameters: 150 mAs, 100 kV, 0.3 s rotation, 1.2 mm pitch, 0.465 mm pixel size, 128 × 0.6 mm collimation, 1 mm slice thickness and reconstruction, and Bf70 very sharp kernel. Axial images were captured from the lung apex to base during full inspiration with a breath hold. Coronal and sagittal 1-mm reconstructions were oriented on the tracheal plane. No intravenous contrast was used. Images were reviewed on a 24-inch, 3-megapixel Barco medical monitor (2048 × 1536 resolution). Native MSCT software was used for image analysis, including evaluation of scan quality (inspiratory level and motion artefacts). CT-derived metrics included:

- **CO-RADS score** (1–5): Reflects likelihood of COVID-19 lung involvement (1: very low; 2: low; 3: equivocal; 4: high; 5: very high).
- **Lobar involvement score** (0–5): 0 (0%), 1 (<5%), 2 (5–25%), 3 (26–50%), 4 (51–75%), 5 (>75%).
- **Dominant CT pattern**: Ground-glass opacities, consolidations, mixed, crazy-paving, reverse halo (per Fleischner Society).
- **Distribution**: Lower/upper lobes, peripheral, bronchocentric, dorsal, diffuse.
- **Additional COVID-19-related findings**: Pleural thickening, vascular enlargement, subpleural sign, halo sign, air bubble sign, perilobular pattern, subpleural sparing.
- **Non-typical findings**: Pleural/pericardial effusion, lymphadenopathy, cavitation, tree-in-bud, small nodules, isolated consolidation, atelectasis, septal thickening.

161 **RNA extraction**

162 Total RNA was extracted from PAXgene Blood RNA Tubes using the PAXgene Blood RNA Kit
163 (PreAnalytix), following the manufacturer's protocol, and stored at –80°C. RNA was processed
164 using the QIAseq FastSelect–rRNA/Globin Kit (Qiagen) to remove cytoplasmic/mitochondrial
165 rRNA, as well as globin mRNA, with fragmentation times of 7 or 15 minutes. Libraries were
166 prepared using the NEBNext Ultra II Directional RNA Library Prep Kit (New England Biolabs),
167 with either 11 or 13 amplification cycles, followed by AMPure XP bead purification. Libraries
168 were quantified by Qubit, assessed with the Agilent 2100 Bioanalyzer, pooled in equimolar
169 ratios, and sequenced as 150 bp paired-end reads on an Illumina® NovaSeq 6000.

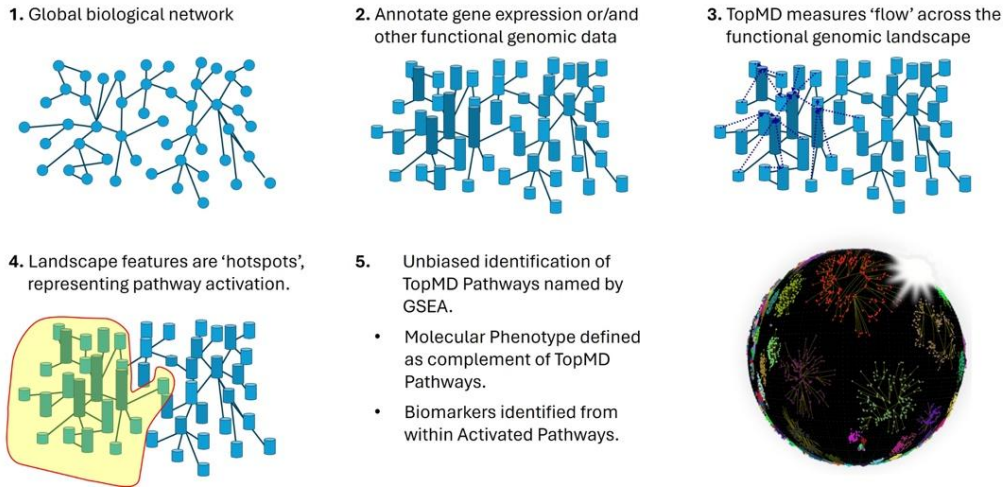
170 **Bioinformatics**

171 Raw FASTQ files were trimmed using fastp[17]. Trimmed reads were quantified with salmon
172 (v1.5.2) using -l A --validateMappings --seqBias --gcBias[18]. Resulting quantification files were
173 imported into RStudio (v4.1.1) with tximport, and normalised using the edgeR package
174 (v3.34.1)[19]. Sequencing data are available under BioProject ID: PRJNA1085259 on the SRA.

175

176 **TopMD molecular phenotype mapping**

177 Topological analysis was performed using TopMD (Patent: GB202306368D0), which maps
178 global gene expression onto a STRING-derived interaction network[20]. Each gene was assigned
179 to a network vertex, and TopMD identified topological features—hotspots of differential
180 expression—representing modulated gene pathways. Pathway activation was quantified as
181 topological volume: the sum of squared differential expression of genes in each cluster. The
182 resulting molecular phenotype is defined as the global profile of pathway activation volumes.
183



Drug interactions mapped by topological analysis

TopMD gene clusters were compared to the Drug-Gene Interaction Database (DGIdb)[21], which catalogues known or predicted drug–gene interactions. Overlap between TopMD gene groups and drug-associated genes was assessed using a binomial distribution test to calculate the probability of random overlap. Results included the raw and Bonferroni-adjusted p-values, TopMD volume (integrating pathway volume with drug-gene overlap significance), and an activation value (sum of Log2 fold-changes for genes shared between the TopMD cluster and the drug-associated set).

Regression

Regression analysis was performed using a Logistic regression model with the following optimisation problem:

$$\min_{w} C \sum_{i=1}^{n} s_i \left(-y_i \log(\hat{p}(X_i)) - (1-y_i)\log(1-\hat{p}(X_i))\right) + r(w)$$

Where X is the **pathway matrix** and y is the classification vector, with 0 indicating that the i-th sample belongs to the considered class and 1 otherwise. We used a regularisation parameter C value of 1. For the penalisation term *r(w)* for the regression weights w, we applied an ElasticNet penalisation with an **l1 ratio** parameter value of 0.5.

$$\frac{1-\rho}{2} w^T w + \rho \|w\|_1$$

The probability the i-th sample with **pathways values** equal to Xi is then:

$$\hat{p}(X_i) = \text{expit}(X_i w + w_0) = \frac{1}{1 + \exp(-X_i w - w_0)}$$

With w0 the intercept. The python module used was **scikit-learn (version 1.4.1)** and the algorithm used LogisticRegression function in the linear_model submodule.

A 70/30 stratified split was carried out separately for each cohort and repeated over 10 random iterations. For each split, regression models were trained using between 1 and 20 pathways, ranked by pathway volume within that split. Each model's performance was assessed by averaging AUC scores across training and test sets. The best model was selected using a max-min criterion—maximising the minimum AUC between training and test sets—to prevent models that perform poorly in training but seem effective on test data due to random variation.

Patient Clustering

Pathway volumes were visualised through PCA using PCAtools (v2.14.0), revealing three distinct clusters, which were confirmed by K-means clustering based on pathway activation relative to healthy controls. The top 10% of PCA loadings were extracted to identify pathways responsible for cluster separation. To compare pathway activity between clusters, the average volume of each pathway within each cluster was calculated relative to the overall COVID-19 patient average.

Logistic Regression Receiver Operating Characteristic (LRROC) analysis using patient clusters derived from the patient pathway volume matrix.

The area under the ROC curve (AUC) gauges a model’s capacity to differentiate between classes, with higher AUC scores signifying better discrimination and more clearly defined patient groups.

LRROC for Florence patients:

Logistic regression ROC (LRROC) analysis was conducted using the pathway volume matrix from Florence patients. The dataset was split into training and test sets to assess generalisation, and separate ROC curves were generated for each.

Cluster validation for Liège patients:

The LRROC model trained on Florence data was validated using Liège patient data. An ROC curve was generated to evaluate the model's ability to classify Liège patient clusters based on pathway volume.

246 **Results:**

247 RNA sequencing was conducted on peripheral blood samples from 173 patients recruited
248 through the DRAGON consortium, with 132 from Florence and 41 from Liège. *The aim was* to
249 evaluate the potential of combining blood transcriptomics with clinical data using the TopMD
250 machine learning approach. Patient details are listed in Supplementary Table 1. Ten patients
251 experienced fatal outcomes; however, due to limited statistical power, an unsupervised
252 analysis was utilised. CT scans, scored by clinicians, were available for 109 patients
253 (Supplementary Table 2). Most had high or very high CORADS scores, while 26% were
254 equivocal, 4.6% low, and 2.8% very low. CORADS ("COVID-19 Reporting and Data System") is
255 a CT-based classification to estimate the likelihood of COVID-19 infection, ranging from very
256 low to very high.

257

258 *Table 1: Characteristics of 132 patients from Florence included in the study, including lab*
259 *results at admission.*

Characteristic	N	N = 132[1]
Died	132	
N		127 (96%)
Y		5 (3.8%)
Age	132	60 (50, 68)
Sex	132	
F		40 (30%)
M		92 (70%)
Non-invasive ventilation	132	
N		123 (93%)
Y		9 (6.8%)
Continuous positive airway pressure	132	
N		129 (98%)
Y		3 (2.3%)
Tracheostomy	132	
N		131 (99%)

Y		1 (0.8%)
High flow nasal cannula oxygen therapy	132	
N		105 (80%)
Y		27 (20%)
Hypertension	132	
N		77 (58%)
Y		55 (42%)
Malnutrition	132	
N		131 (99%)
Y		1 (0.8%)
Cardiovascular disease	132	
N		119 (90%)
Y		13 (9.8%)
Respiratory disease	132	
N		118 (89%)
Y		14 (11%)
Cancer	132	
N		118 (89%)
Y		14 (11%)
Chronic kidney disease	132	
N		130 (98%)
Y		2 (1.5%)
Chronic hepatitis	132	
N		130 (98%)
Y		2 (1.5%)
Cerebrovascular disease	132	
N		125 (95%)
Y		7 (5.3%)
Chronic hematologic disease	132	
N		129 (98%)
Y		3 (2.3%)

Diastolic blood pressure (mmHg)	132	79 (70, 85)
Heart rate (BPM)	130	80 (75, 89)
Systolic blood pressure (mmHg)	132	125 (115, 140)
Temperature (°C)	131	36.50 (36.00, 37.20)
Weight (kg)	128	78 (70, 89)
Height (cm)	126	170 (165, 175)
Alanine aminotransferase (U/L)	128	27 (17, 39)
Aspartate aminotransferase (U/L)	64	31 (24, 46)
Bilirubin (mg/dL)	128	0.50 (0.30, 0.60)
Calcium (mg/dL)	96	4.50 (4.34, 4.63)
Creatinine (mg/dL)	130	0.83 (0.73, 0.95)
D-dimer (ng/mL)	91	728 (429, 1,091)
Direct bilirubin (mg/dL)	44	0.25 (0.17, 0.29)
Fibrinogen (mg/dL)	117	572 (446, 654)
Fraction of inspired oxygen (%)	127	28 (21, 36)
Hematocrit (%)	132	42.7 (39.7, 45.8)
Lactate dehydrogenase (U/L)	118	297 (247, 359)
Lactic acid (mg/dL)	102	9.0 (7.0, 11.9)
Leukocytes (10^9/L)	132	6.2 (4.6, 7.7)
Lymphocytes (10^9/L)	129	0.90 (0.68, 1.25)
Neutrophils (10^9/L)	129	4.67 (3.08, 6.16)
Oxygen saturation (%)	109	96.10 (94.20, 97.70)
Partial pressure oxygen (mmHg)	131	74 (65, 87)
Partial pressure carbon dioxide (mmHg)	127	36.2 (34.0, 39.0)
Platelets (10^9/L)	132	196 (156, 255)
Potassium (mmol/L)	128	3.85 (3.50, 4.10)
Procalcitonin (ug/L)	126	0.09 (0.06, 0.15)
Prothrombin time (seconds)	127	13.00 (12.30, 13.70)
Sodium (mmol/L)	129	137 (135, 140)
Urea nitrogen (g/L)	64	30 (30, 50)

[1] **n (%); Median (IQR)**

Patients form three clusters based on pathway activation

Gene expression (mRNA identity and abundance) was obtained from RNA sequencing using Salmon and Tximport in R, with values converted to log2 counts per million (CPM). Pathway activation was then determined using TopMD. Principal component analysis (PCA) of pathway activation identified three distinct patient clusters (Figure 1). Clinical features, demographics, and CT scan data were compared across these clusters, with significant differences summarised in Table 3. Lactic acid levels were slightly higher in clusters 1 and 2, and lower in cluster 3. Cluster 2 showed a higher prevalence of respiratory disease, increased fraction of inspired oxygen, and elevated direct bilirubin. Most COVID-19-related deaths occurred within cluster 2. CORADS scores did not distinguish the clusters at the molecular level.

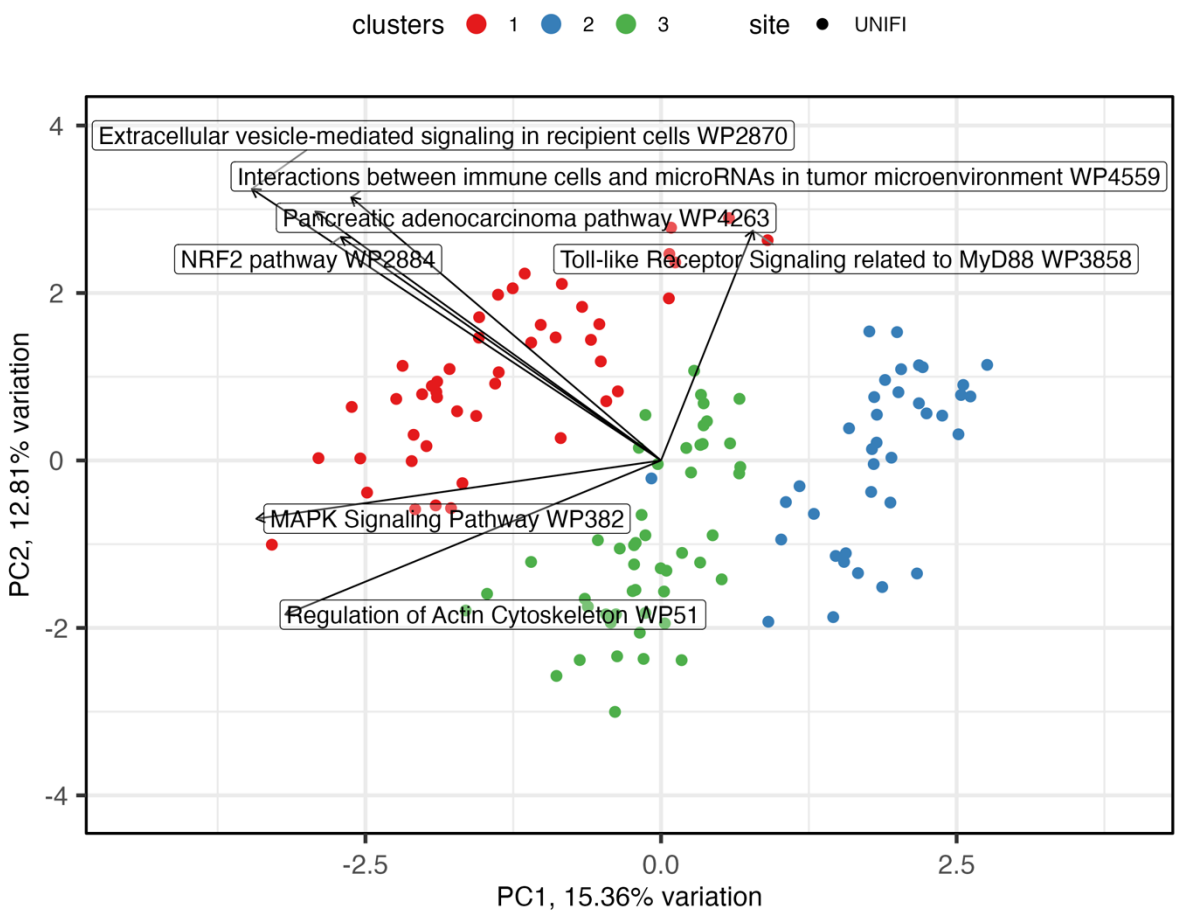


Figure 1:TopMD pathway volumes of each patient in the Florence cohort, plotted as a PCA plot. The data reveal three distinct clusters based on pathway activation, as determined by k-means.

274 *Table 2: Patient characteristics that differ between the three clusters in the Florence cohort (p*
275 *= <0.05).*

Characteristic	N	1, N = 46[1]	2, N = 37[1]	3, N = 49[1]	p-value[2]
Lactic acid (mg/dL)	102	10.0 (7.7, 13.0)	10.0 (7.2, 12.0)	8.0 (5.3, 9.7)	0.008
Fraction of inspired oxygen (%)	127	28 (21, 36)	32 (27, 40)	28 (21, 29)	0.019
Died	132				0.032
N		46 (100%)	33 (89%)	48 (98%)	
Y		0 (0%)	4 (11%)	1 (2.0%)	
Respiratory disease	132				0.042
N		44 (96%)	29 (78%)	45 (92%)	
Y		2 (4.3%)	8 (22%)	4 (8.2%)	
Direct bilirubin (mg/dL)	44	0.20 (0.17, 0.27)	0.28 (0.24, 0.32)	0.20 (0.17, 0.28)	0.047

[1] n (%); Median (IQR)

[2] Fisher's exact test; Kruskal-Wallis rank sum test; Pearson's Chi-squared test

276

277 **Molecular phenotypes and pathway activation**

278 Cluster 1 was characterised by high activation of pathways related to ESC pluripotency, NRF2,
279 and TGF-β receptor signalling (Figure 2). Cluster 2 showed elevated activity in focal adhesion–
280 PI3K–Akt–mTOR and type I interferon signalling pathways. Cluster 3 displayed low activation
281 of IRF7-related pathways.

282 **Classification performance and validation**

283 LRROC analysis was conducted on models trained with 70% of Florence patients, with testing
284 on the remaining 30%. AUCROC scores were 0.84, 0.85, and 0.72 for Clusters 1, 2, and 3,
285 respectively. Validation in the Liège cohort yielded AUCROC values of 0.76, 0.93, and 0.69 for
286 Clusters 1, 2, and 3, respectively (Supplementary Figures 1 and 2).

287

288 **Potential drug candidates for each cluster**

289 TopMD pathway activation profiles were mapped to the Drug-Gene Interaction Database
290 (Figure 3), identifying potential drug targets specific to each cluster (Supplementary Table 4).
291 This dual-purpose approach provides insights into both potential therapeutic strategies and
292 the underlying biology of each phenotype. While shared targets such as *ITGB2*, *GNAS*, and
293 *CXCR2* appeared across all clusters, unique targets emerged: *IFNAR1*, *TGFBR2*, and *CSF2RB* in
294 Cluster 1; *SERPING1* and *TLN1* in Cluster 2; and *SERPING1* (shared with Cluster 2) in Cluster 3.
295 These distinctions emphasise both common and cluster-specific therapeutic opportunities.

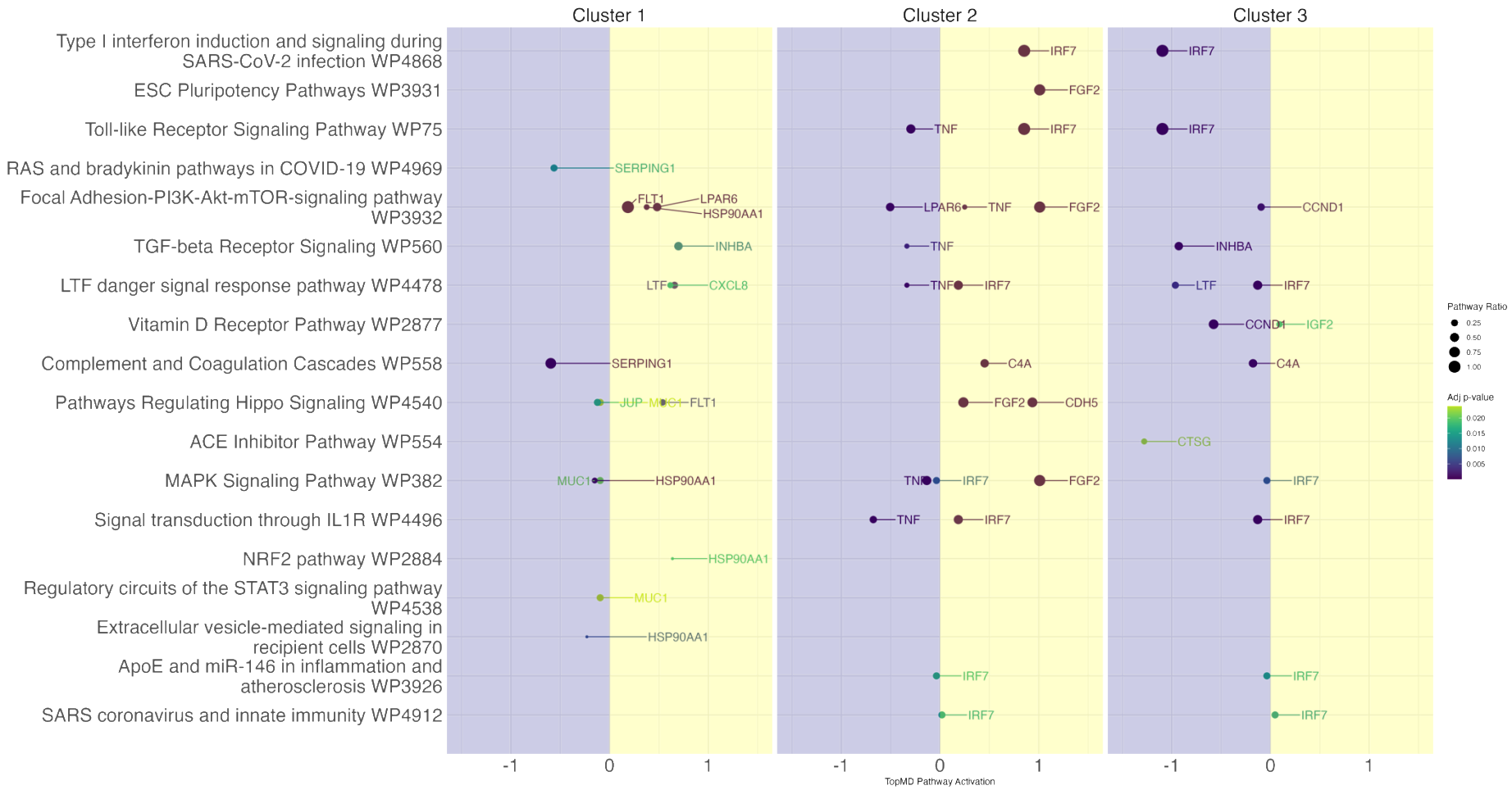
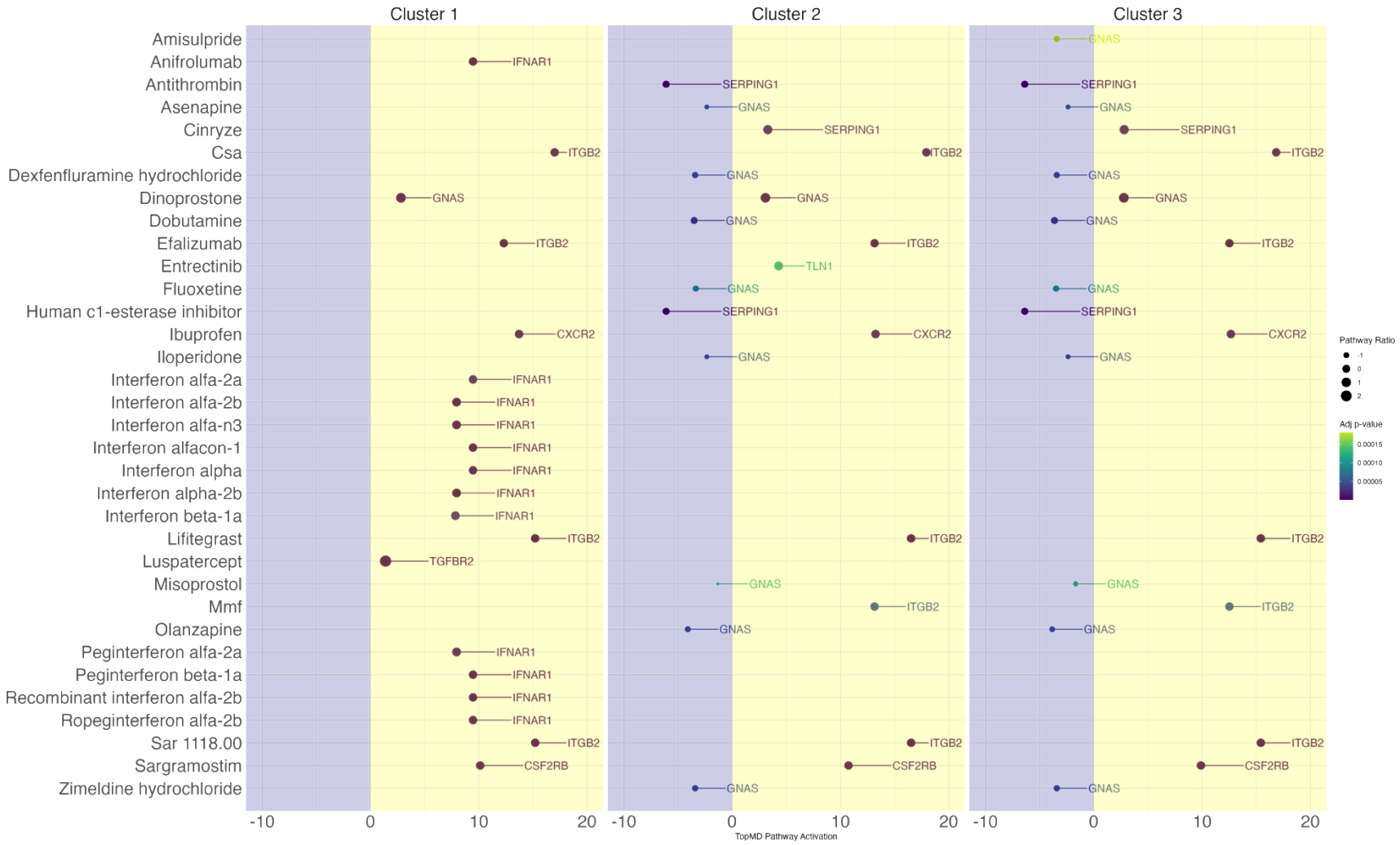


Figure 2: The average pathway volume for each cluster was considered in a TopMD enrichment analysis against the average pathway activation for the whole cohort to identify differentially activated pathways. The enrichment analysis was filtered by adjusted P value, then the top pathways were plotted. The pathways are annotated with the gene that leads the identified pathway. The dots are coloured by adjusted p-value, and the size represents the proportion of genes identified within that pathway from TopMD analysis.


Cluster 1
Cluster 2
Cluster 3
Amisulpride
Anifrolumab
Antithrombin
Asenapine
Cinryze
Csa
Dexfenfluramine hydrochloride
Dinoprostone
Dobutamine
Efalizumab
Entrectinib
Fluoxetine
Human c1-esterase inhibitor
Ibuprofen
Iloperidone
Interferon alfa-2a
Interferon alfa-2b
Interferon alfa-n3
Interferon alfacon-1
Interferon alpha
Interferon alpha-2b
Interferon beta-1a
Lifitegrast
Luspatercept
Misoprostol
Mmf
Olanzapine
Peginterferon alfa-2a
Peginterferon beta-1a
Recombinant interferon alfa-2b
Ropeginterferon alfa-2b
Sar 1118.00
Sargramostim
Zimeldine hydrochloride
IFNAR1
SERPING1
GNAS
ITGB2
TLN1
CXCR2
TGFBR2
CSF2RB
-10
0
10
20
TopMD Pathway Activation
Pathway Ratio
-1
0
1
2
Adj p-value
0.00015
0.00010
0.00005

301

302 *Figure 3: TopMD enrichment analysis was mapped against the Drug-Gene Interaction Database, using a healthy baseline, revealing approved*
303 *drugs that are known to target genes and their corresponding pathways. The top drug candidates are plotted based on adjusted p-value and*
304 *pathway volume.*

305 **Pathway activation in fatal cases: identifying potential intervention targets**

306 Pathway activation was investigated in the 10 deceased patients from the Florence and Liège
307 cohorts. As expected, these patients were elderly and had high rates of comorbidities—
308 cardiovascular (70%), respiratory (50%), malnutrition (40%), hypertension (90%),
309 cerebrovascular (30%), and chronic hepatitis (40%). Notably, all 10 exhibited strong activation
310 of the non-alcoholic fatty liver disease pathway, primarily driven by NDUFA9 and UQCRC2
311 (Figure 4).

312

313 Despite this shared pathway, analysis at the individual level revealed considerable
314 heterogeneity, reflecting the complex interaction between COVID-19, comorbidities, and
315 personal demographics. Enrichment analysis identified potential therapeutic targets based on
316 individual pathway profiles. All patients showed activation of druggable targets such as CXCR2
317 (Figure 5), while additional targets were found in specific individuals, including GNAS (multiple
318 patients), ITGB2 (patients 2 and 6), CSF2RB (multiple patients), SERPING1 (5 patients), PIK3CD
319 (patient 5), TGFBR2 (patient 9), and CUL4B (patient 10).

320

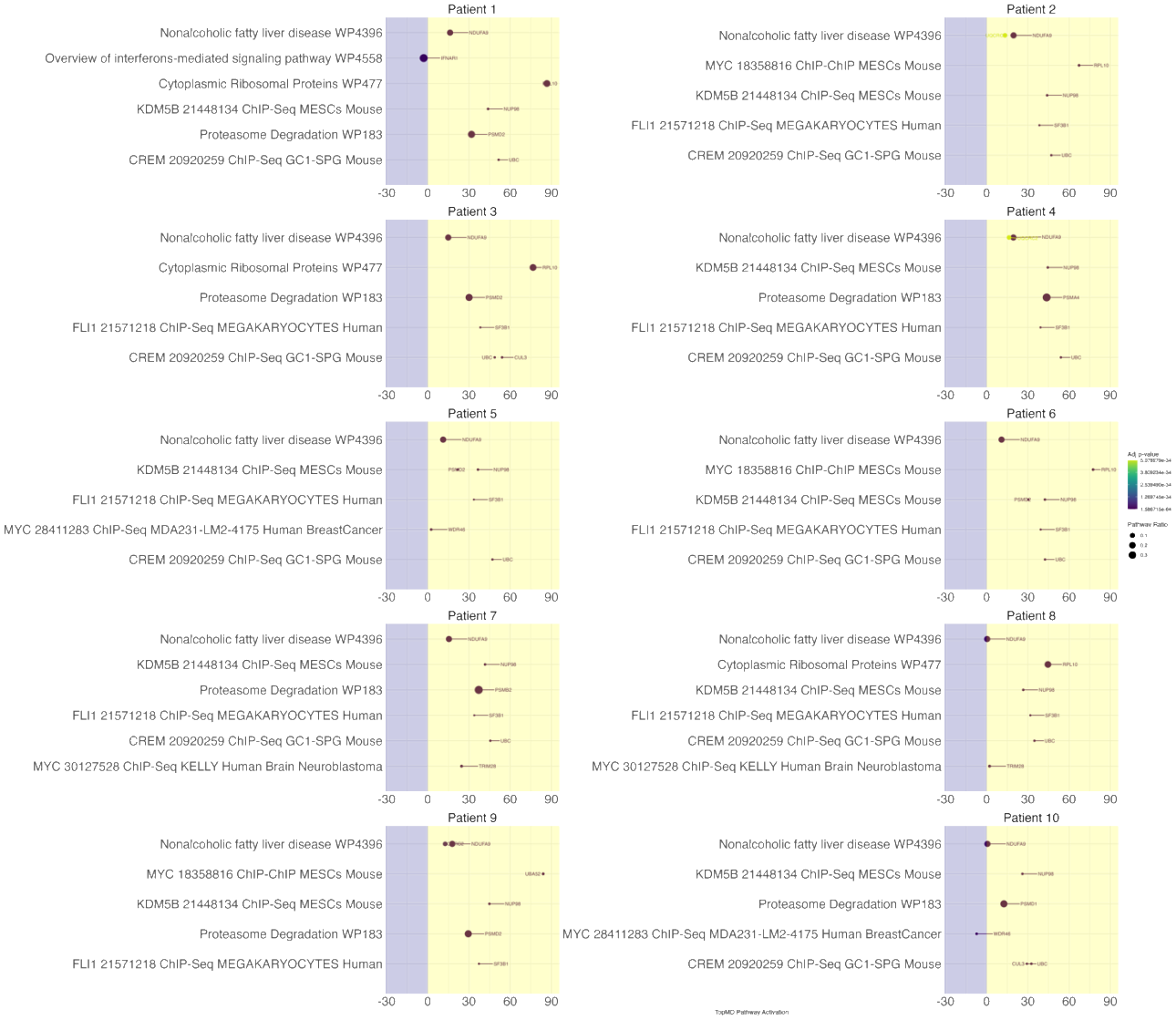


Figure 4: The top 6 pathways enriched in fatal cases within the Florence and Liège cohort using a healthy baseline.

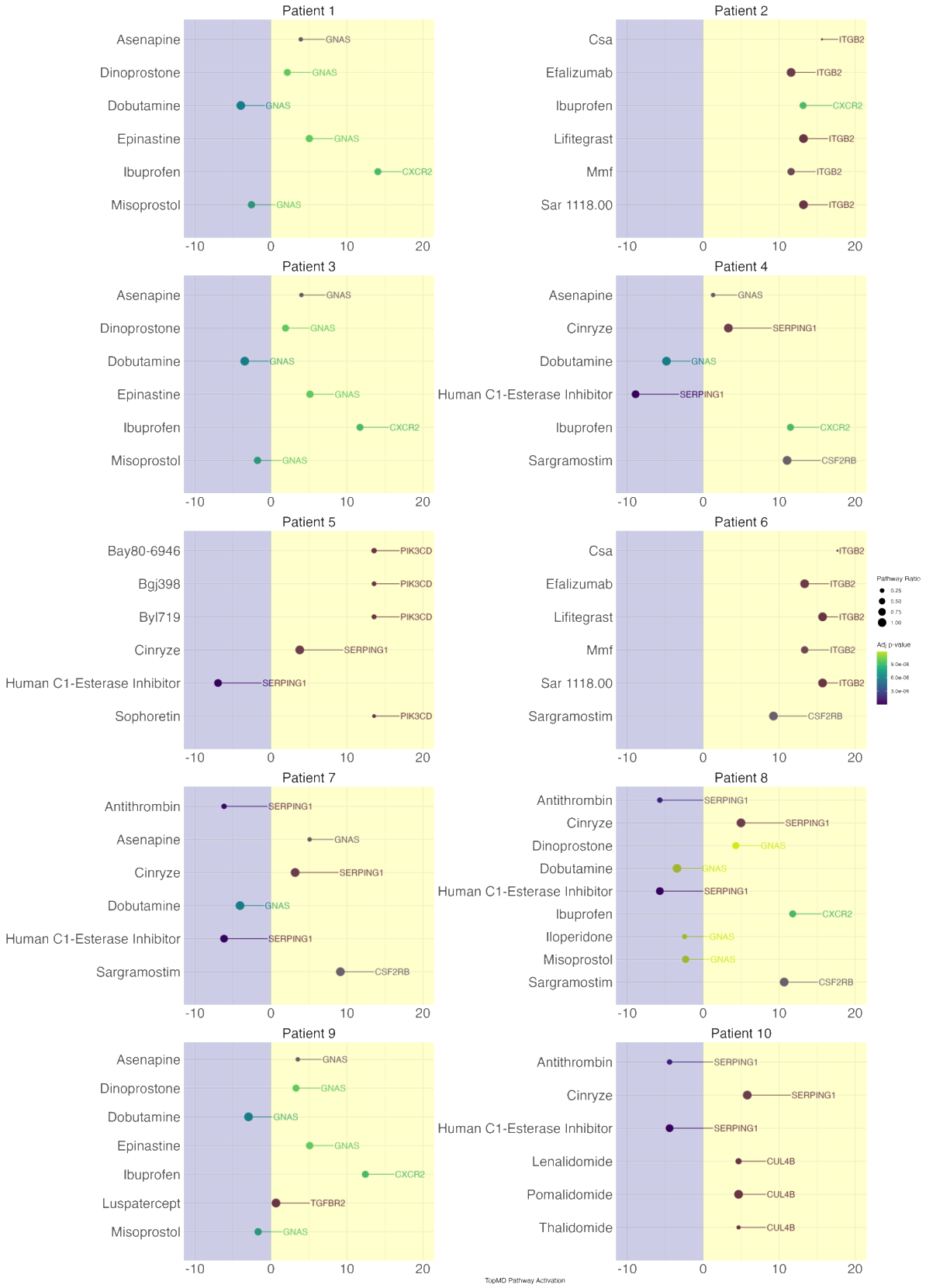


Figure 5: The top significant drug candidates and peak genes that could potentially modulate the phenotype of the 10 fatal cases patients in the Florence and Liège cohort.

340 **Discussion**

341 Traditional molecular phenotyping relies on data reduction and feature selection to remove
342 biological and technical 'noise' before pathway enrichment analysis. However, this can
343 obscure essential signals. TopMD topological analysis overcomes this limitation by capturing
344 global gene expression patterns, including low-abundance transcripts often discarded. These
345 transcripts form the 'foothills' of highly activated pathways, contributing to a comprehensive
346 view of the molecular phenotype. Understanding these phenotypes enables more precise
347 therapeutic selection, as drugs act at the molecular rather than clinical level[22].

348 There is increasing recognition of the need to integrate diverse data types—such as
349 biomarkers, clinical measures, and imaging—to refine outcome prediction. AI frameworks
350 applied to chest CT, for example, have achieved high diagnostic accuracy (85.2%) for COVID-
351 19 and facilitated rapid clinical decision-making[23]. Another AI tool, CACOVID-CT, quantified
352 lung damage severity and correlated with hospital stay, ICU admission, mechanical ventilation,
353 and mortality across 476 patients[24]. These tools enhance risk stratification and reduce
354 radiologist workload.

355 In our study, three molecular phenotypes of COVID-19 were identified via TopMD analysis of
356 blood transcriptomics. These clusters showed strong classification performance in the
357 Florence cohort (AUC: 0.84–0.85) and validated well in the Liège cohort (AUC: 0.69–0.93).
358 Cluster-specific clinical features emerged: clusters 1 and 2 had elevated lactic acid, a known
359 marker of severity[25]; cluster 2 also showed more respiratory disease, higher oxygen demand,
360 and raised direct bilirubin[26]. Most COVID-19 fatalities occurred in cluster 2. Importantly,
361 CORADS CT scores could not distinguish clusters, suggesting underlying molecular differences
362 despite similar radiological findings. Continuous CT scoring may offer improved granularity[23].

363 Cluster 1 was marked by high activation of ESC pluripotency, NRF2, and TGF-β receptor
364 signalling—pathways linked to tissue repair, antioxidant response, and inflammation
365 regulation[22,27]. Conversely, extracellular vesicle-mediated signalling and
366 complement/coagulation cascades were underactive, suggesting impaired intercellular
367 communication and immune dysregulation[28]. Cluster 2 showed increased PI3K-Akt-mTOR and
368 type I interferon pathway activation, associated with antiviral responses and tissue repair[29,30].
369 Cluster 3 exhibited low IRF7, TGF-β, and IL-1R signalling, as well as a reduced LTF danger signal

370 response. Diminished lactoferrin signalling may impair viral control and modulate
371 inflammation[31,32].

372 Drug-pathway mapping using the Drug-Gene Interaction Database revealed both shared and
373 unique therapeutic targets across clusters (Figure 3, Supplementary Table 4). For example,
374 ITGB2, GNAS, and CXCR2 were common targets, while IFNAR1, TGFBR2, and CSF2RB were
375 specific to Cluster 1. Cluster 2 included SERPING1, TLN1, and Cluster 3 shared SERPING1. This
376 offers potential for cluster-specific repurposing of existing therapies.

377 Cyclosporine A (CSA) emerged as a candidate across all clusters, having shown safety and
378 potential to reduce hyperinflammation in COVID-19[33,34]. Interferon-modulating drugs and
379 TGF-β inhibitors such as Luspatercept were identified in Cluster 1, aligning with reports of their
380 efficacy[22,35]. Lifitegrast, shown to inhibit SARS-CoV-2 in vitro[36], also matched this cluster. In
381 Cluster 2, SSRIs like fluoxetine and fluvoxamine were highlighted, in line with evidence
382 suggesting benefit in COVID-19 and long-COVID[37]. Cinryze (C1 esterase inhibitor) and
383 Asenapine were identified as cluster-specific candidates[38,39]. Amisulpride emerged as a target
384 in Cluster 3.

385 **Individual-level precision medicine in fatal cases**

386 Enrichment analysis of the 10 deceased patients (Florence and Liège) further demonstrated
387 TopMD's utility in individualised treatment guidance. While all patients shared CXCR2 and
388 GNAS activation (suggesting potential benefit from drugs like ibuprofen), other targets varied.
389 For example, ITGB2 was enriched in patients 2 and 6, CSF2RB in several patients (implicating
390 Sargramostim)[40], and SERPING1 in five patients (suggesting use of Cinryze or antithrombin).
391 Patient-specific targets included PIK3CD (patient 5; candidate: Sophoretin)[28], TGFBR2 (patient
392 9; candidate: Luspatercept), and CUL4B (patient 10; candidate: Thalidomide, Pomalidomide,
393 or Lenalidomide)[41,42].

394 This research identified three molecular phenotypes in hospitalised COVID-19 patients, which
395 were not reflected in CT imaging or routine clinical observations. These phenotypes
396 correspond to distinct molecular mechanisms and drug targets, supporting a biomarker-
397 based, stratified approach to treatment.

398 Topological analysis of global gene expression to define a patient's pathway activation map
399 could be used in future pandemics to assist treatment decisions before clinical trials are
400 completed and to identify and prioritise candidate therapeutics for rapid testing.

401

Funding:

TopMD, the University of Southampton, Imperial College London, CDISC, Comunicare Solutions, University Hospital of Liège (CHU Liège), and the University of Liverpool are members of the DRAGON consortium. The DRAGON project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 101005122. The JU receives support from the European Union's Horizon 2020 research and innovation program and EFPIA. This publication reflects the author's view. Neither IMI nor the European Union, EFPIA or the DRAGON consortium, are responsible for any use that may be made of the information contained therein. JAH, TP, RPR and CH were supported by the US Food and Drug Administration Medical Countermeasures Initiative (no 75F40120C00085) awarded to JAH and work was also supported by the MRC funded MR/Y004205/1 'The G2P2 virology consortium'. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. DB is funded by NIHR and MRC.

Acknowledgments:

The authors wish to thank the study participants and the hospital staff for their participation in this study.

Contributions:

RPR, FS: methodology, RPR, FS: visualisation, RPR, FS, BVE, TK: software, BE: project administration, RPR, FS, JG, MH, AS, EG, LG, AJP, CN: investigation, RPR, FS, JG, CH, TP, SW: formal analysis, JG, CN: clinical supervision, RPR, FS, AJP, ST, RB, KH, XX, YN, SW: data curation, RPR, FS, EP, PS, JAH, JPS: writing - original draft preparation, RPR, FB, BE, BE, JG, MH, AS, EG, LG, AJP, CN, ST, RB, KH, CH, TP, RD, TC, DB, SW, XX, YN, SW, SW, GY, PJS, JAH, JPRS: writing - review and editing, JPRS, JAH, PJS, GY, SW: funding acquisition

Conflict of interest disclosure statement

AS, BE, EG, JG, LG, MH, TM, AP,CH,DB, GY, JH, CN, RPR, SW, TK, TP, XX, YN, SW, KH have nothing to disclose. BVE declares he is Shareholder in Comunicare Solutions SA.TC declares the following potential conflicts of interest: Regarding support for attending meetings and/or travel, received support from Cepheid, Roche Diagnostics, Roche, Qiagen, and Janssen. TC has participated on Data Safety Monitoring Boards or Advisory Boards for Roche, Shionogi, GSK,

Sanofi, Seqirus, Cepheid, Roche Diagnostics, and Janssen. JPRS, PJS & FS report being Directors and shareholders in TopMD Precision Medicine Ltd. EP is a shareholder in TopMD Precision Medicine Ltd. RD reports receiving fees for lectures at symposia organised by Novartis, AstraZeneca and TEVA, consultation for TEVA and Novartis as member of advisory boards, and participation in a scientific discussion about asthma organised by GlaxoSmithKline. He is a co-founder and current consultant, and has shares in Synairgen, a University of Southampton spin out company.

Data availability:

Sequencing reads available under SRA bioproject: PRJNA1085259

References

1 Clark, J. J. *et al.* Sequential Infection with Influenza A Virus Followed by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Leads to More Severe Disease and Encephalitis in a Mouse Model of COVID-19. *Viruses* **16**, 863 (2024).

2 De Neck, S. *et al.* The Stereotypic Response of the Pulmonary Vasculature to Respiratory Viral Infections: Findings in Mouse Models of SARS-CoV-2, Influenza A and Gammaherpesvirus Infections. *Viruses* **15** (2023).

3 Liu, X. *et al.* Transcriptomic signatures differentiate survival from fatal outcomes in humans infected with Ebola virus. *Genome Biology* **18**, 4, doi:10.1186/s13059-016-1137-3 (2017).

4 Dorward, D. A. *et al.* Tissue-Specific Immunopathology in Fatal COVID-19. *Am J Respir Crit Care Med* **203**, 192-201, doi:10.1164/rccm.202008-3265OC (2021).

5 Legebeke, J. *et al.* Evaluating the Immune Response in Treatment-Naive Hospitalised Patients With Influenza and COVID-19. *Front Immunol* **13**, 853265, doi:10.3389/fimmu.2022.853265 (2022).

6 Penrice-Randal, R. *et al.* Blood gene expression predicts intensive care unit admission in hospitalised patients with COVID-19. *Front Immunol* **13**, 988685, doi:10.3389/fimmu.2022.988685 (2022).

7 Polese, B. *et al.* Circulating inflammatory cytokines predict severity disease in hospitalized COVID-19 patients: A prospective multicenter study of the European DRAGON consortium. *J Infect Public Health* **17**, 102589, doi:10.1016/j.jiph.2024.102589 (2024).

8 Nan, Y. *et al.* Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions. *Information Fusion* **82**, 99-122, doi:https://doi.org/10.1016/j.inffus.2022.01.001 (2022).

9 Halilaj, I. *et al.* Covid19Risk.ai: An Open Source Repository and Online Calculator of Prediction Models for Early Diagnosis and Prognosis of Covid-19. *BioMed* **1**, 41-49 (2021).

10 Chatterjee, A. *et al.* Can predicting COVID-19 mortality in a European cohort using only demographic and comorbidity data surpass age-based prediction: An externally validated study. *PLOS ONE* **16**, e0249920, doi:10.1371/journal.pone.0249920 (2021).

11 Cen, X. *et al.* Towards precision medicine: Omics approach for COVID-19. *Biosaf Health* **5**, 78-88, doi:10.1016/j.bsheal.2023.01.002 (2023).

12 Teodori, L., Osimani, B., Isidoro, C. & Ramakrishna, S. Mass versus personalized medicine against COVID-19 in the "system sciences" era. *Cytometry Part A* **101**, 995-999, doi:https://doi.org/10.1002/cyto.a.24662 (2022).

13 Wang, Z. & He, Y. Precision omics data integration and analysis with interoperable ontologies and their application for COVID-19 research. *Briefings in Functional Genomics* **20**, 235-248, doi:10.1093/bfgp/elab029 (2021).

14 Venkataraman, T., Coleman, C. M. & Frieman, M. B. Overactive Epidermal Growth Factor Receptor Signaling Leads to Increased Fibrosis after Severe Acute Respiratory Syndrome Coronavirus Infection. *Journal of virology* **91**, doi:10.1128/jvi.00182-17 (2017).

15 Venkataraman, T. & Frieman, M. B. The role of epidermal growth factor receptor (EGFR) signaling in SARS coronavirus-induced pulmonary fibrosis. *Antiviral Res* **143**, 142-150, doi:10.1016/j.antiviral.2017.03.022 (2017).

16 Vagapova, E. R., Lebedev, T. D. & Prassolov, V. S. Viral fibrotic scoring and drug screen based on MAPK activity uncovers EGFR as a key regulator of COVID-19 fibrosis. *Scientific Reports* **11**, 11234, doi:10.1038/s41598-021-90701-w (2021).

17 Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890, doi:10.1093/bioinformatics/bty560 (2018).

18 Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**, 417-419, doi:10.1038/nmeth.4197 (2017).

19 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).

20 Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* **51**, D638-d646, doi:10.1093/nar/gkac1000 (2023).

21 Freshour, S. L. *et al.* Integration of the Drug-Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. *Nucleic Acids Res* **49**, D1144-d1151, doi:10.1093/nar/gkaa1084 (2021).

22 Farahani, M. *et al.* Molecular pathways involved in COVID-19 and potential pathway-based therapeutic targets. *Biomedicine & Pharmacotherapy* **145**, 112420, doi:https://doi.org/10.1016/j.biopha.2021.112420 (2022).

23 Guiot, J. *et al.* Automatized lung disease quantification in patients with COVID-19 as a predictive tool to assess hospitalization severity. *Frontiers in Medicine* **9**, doi:10.3389/fmed.2022.930055 (2022).

24 Guiot, J. *et al.* Development and Validation of an Automated Radiomic CT Signature for Detecting COVID-19. *Diagnostics* **11**, 41 (2021).

25 Carpenè, G. *et al.* Blood lactate concentration in COVID-19: a systematic literature review. *Clin Chem Lab Med* **60**, 332-337, doi:10.1515/cclm-2021-1115 (2022).
26 Chen, W. *et al.* Effect of Direct Bilirubin Level on Clinical Outcome and Prognoses in Severely/Critically Ill Patients With COVID-19. *Front Med (Lausanne)* **9**, 843505, doi:10.3389/fmed.2022.843505 (2022).
27 Chen, J., Wu, W., Wang, W., Tang, Y. & Lan, H.-Y. Role of TGF-β Signaling in Coronavirus Disease 2019. *Integrative Medicine in Nephrology and Andrology* **9**, 9, doi:10.4103/2773-0387.348713 (2022).
28 Conway, E. M. *et al.* Understanding COVID-19-associated coagulopathy. *Nat Rev Immunol* **22**, 639-649, doi:10.1038/s41577-022-00762-9 (2022).
29 Fattahi, S., Khalifehzadeh-Esfahani, Z., Mohammad-Rezaei, M., Mafi, S. & Jafarinia, M. PI3K/Akt/mTOR pathway: a potential target for anti-SARS-CoV-2 therapy. *Immunol Res* **70**, 269-275, doi:10.1007/s12026-022-09268-x (2022).
30 Khezri, M. R., Varzandeh, R. & Ghasemnejad-Berenji, M. The probable role and therapeutic potential of the PI3K/AKT signaling pathway in SARS-CoV-2 induced coagulopathy. *Cell Mol Biol Lett* **27**, 6, doi:10.1186/s11658-022-00308-w (2022).
31 Puddu, P., Valenti, P. & Gessani, S. Immunomodulatory effects of lactoferrin on antigen presenting cells. *Biochimie* **91**, 11-18, doi:https://doi.org/10.1016/j.biochi.2008.05.005 (2009).
32 Zimecki, M., Actor, J. K. & Kruzel, M. L. The potential for Lactoferrin to reduce SARS-CoV-2 induced cytokine storm. *Int Immunopharmacol* **95**, 107571, doi:10.1016/j.intimp.2021.107571 (2021).
33 Blumberg, E. A. *et al.* A phase I trial of cyclosporine for hospitalized patients with COVID-19. *JCI Insight* **7**, doi:10.1172/jci.insight.155682 (2022).
34 Devaux, C. A., Melenotte, C., Piercecchi-Marti, M. D., Delteil, C. & Raoult, D. Cyclosporin A: A Repurposable Drug in the Treatment of COVID-19? *Front Med (Lausanne)* **8**, 663708, doi:10.3389/fmed.2021.663708 (2021).
35 Kamyshnyi, A. *et al.* Therapeutic Effectiveness of Interferon-α2b against COVID-19 with Community-Acquired Pneumonia: The Ukrainian Experience. *Int J Mol Sci* **24**, doi:10.3390/ijms24086887 (2023).
36 Shen, X.-R. *et al.* ACE2-independent infection of T lymphocytes by SARS-CoV-2. *Signal Transduction and Targeted Therapy* **7**, 83, doi:10.1038/s41392-022-00919-x (2022).
37 Hashimoto, K. Overview of the potential use of fluvoxamine for COVID-19 and long COVID. *Discov Ment Health* **3**, 9, doi:10.1007/s44192-023-00036-3 (2023).
38 Mansour, E. *et al.* Safety and Outcomes Associated with the Pharmacological Inhibition of the Kinin-Kallikrein System in Severe COVID-19. *Viruses* **13**, doi:10.3390/v13020309 (2021).
39 Ku, K. B. *et al.* Repurposing Screens of FDA-Approved Drugs Identify 29 Inhibitors of SARS-CoV-2. *J Microbiol Biotechnol* **30**, 1843-1853, doi:10.4014/jmb.2009.09009 (2020).
40 Actor, J. K., Hwang, S. A. & Kruzel, M. L. Lactoferrin as a natural immune modulator. *Curr Pharm Des* **15**, 1956-1973, doi:10.2174/138161209788453202 (2009).
41 Biernacka, A., Dobaczewski, M. & Frangogiannis, N. G. TGF-β signaling in fibrosis. *Growth Factors* **29**, 196-202, doi:10.3109/08977194.2011.595714 (2011).

42 Al Sbihi, A., Manasrah, N. & Sano, D. Can Lenalidomide Protect against Severe COVID-19 Symptoms in Multiple Myeloma Patients? A Case Series and Review of the Literature. *Eur J Case Rep Intern Med* **9**, 003216, doi:10.12890/2022_003216 (2022).