doi:10.20944/preprints202507.1009.v1

Article

From Gene Networks to Therapeutics: A Causal Inference and Deep Learning Approach for Drug Discovery

Sudhir Ghandikota [1] and Anil G Jegga [1,2,*]

[1] Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center; Sudhir.ghandikota@cchmc.org, anil.jegga@cchmc.org

[2] Department of Pediatrics, University of Cincinnati College of Medicine; anil.jegga@cchmc.org

* Correspondence: anil.jegga@cchmc.org; Tel.: 513-636-0261

Abstract

Background/Objectives: Drug discovery is a lengthy and expensive process, taking an average of 10 years and more than $2 billion from target discovery to drug approval. It is even more challenging in complex diseases due to the disease heterogeneity and limited knowledge about the underlying mechanisms. We present a novel computational framework that integrates network analysis, statistical mediation, and deep learning to identify causal target genes and repurposable small-molecule candidates. **Methods**: We applied weighted gene co-expression network analysis (WGCNA) and bidirectional mediation analysis (causal WGCNA) to transcriptomic data from idiopathic pulmonary fibrosis (IPF) patients to identify genes causally linked to the disease phenotype. These genes were used as a phenotypic signature for deep learning-based compound screening using the DeepCE model. **Results**: Using RNA-seq data from 103 IPF patients and 103 controls, we identified seven significantly correlated modules and 145 causal genes. Five of these genes (*ITM2C, PRTFDC1, CRABP2, CPNE7,* and *NMNAT2*) were predictive of disease severity in IPF. Our compound screening identified several promising candidates, such as Telaglenastat (GLS1 inhibitor), Merestinib (MET kinase inhibitor), and Cilostazol (PDE3 inhibitor), with significant inverse correlation with the IPF-specific gene signature. **Conclusions**: This study demonstrates the utility of combining causal inference and deep learning for drug discovery. Our framework identified novel gene targets and therapeutic candidates for IPF, offering a scalable strategy for phenotype-driven drug discovery and repurposing.

Keywords: drug repurposing; drug repositioning; drug discovery; deep learning; mediation analysis; network analysis; idiopathic pulmonary fibrosis

1. Introduction

Target selection and engagement is a crucial first step in drug discovery, and a significant percentage of the clinical trial failures are linked to incorrect target selection and poor target engagement [1]. It is even more challenging in complex diseases because of the multifactorial nature and heterogeneity of these diseases; they are often driven by multiple genes, biological pathways, and environmental factors. The interconnected nature of these biological mechanisms makes it even more challenging to identify robust druggable targets. Additionally, disease progression can alter the relevance and importance of the selected target over time, making the target selection step even more critical in complex diseases.

Conventional drug discovery approaches are mostly based on a one-drug-one-gene paradigm [2, 3], where compound screening is performed against a single, specific target [4]. Understandably, these approaches have limited success in complex diseases [3], as target-based screening methods often yield a poor correlation between modulating a single protein by a chemical compound and organism-level responses [2]. Furthermore, compound screening using single-gene targets often fails to recognize off-target interactions that can lead to unintended side effects [5]. On the other hand, phenotype-based compound screening represents a systematic approach by possibly capturing the full spectrum of a compound's biological activity [6]. They often begin with a phenotype-specific gene signature, which is queried against a reference database of small-molecule perturbations, typically using correlation measures. Moreover, deep learning (DL) based frameworks can help mitigate the typical issues of noise and incompleteness associated with large-scale transcriptomics data.

We present a bioinformatic methodology that combines network analysis, statistical mediation models, and DL-based small-molecule compound screening frameworks for therapeutic drug discovery (See Materials and Methods). Our methodology starts with analyzing one (or more) transcriptomic datasets, converting them into gene co-expression networks, and applying a network analysis framework [7, 8] to identify significantly correlated gene modules. Next, we employ statistical mediation techniques to examine the relationship between the phenotype and the candidate modules and identify candidate causal genes [9]. Finally, we use the significant mediator genes as the phenotypic signature and apply a mechanism-driven compound screening method to identify potential therapeutic candidates that can modulate this causal signature [2] (Figure 1).

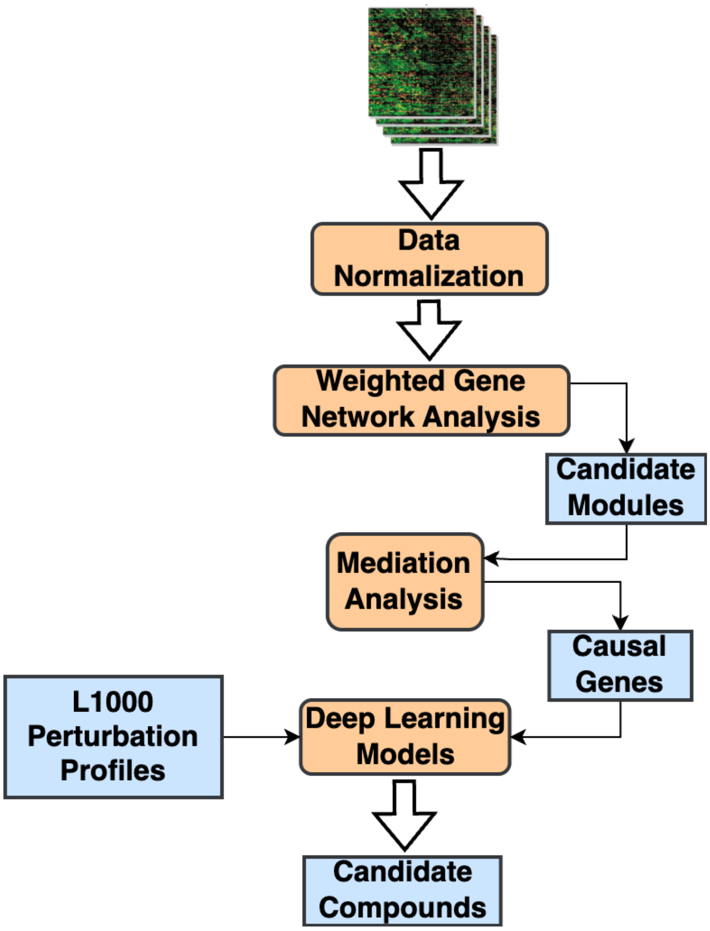

Figure 1. Overall workflow of our proposed compound screening methodology.

To demonstrate the utility of our methodology, we used idiopathic pulmonary fibrosis (IPF) as a disease case study. IPF is a severe fibrotic lung disease characterized by chronic, progressive scarring and destruction of the lung parenchyma, often requiring a lung transplant in advanced cases [10]. Two FDA-approved antifibrotic drugs, nintedanib and pirfenidone, are currently used as the first-line therapy for IPF. Even though both of them are effective in managing the symptoms of lung function decline in IPF patients, these drugs do not cure the disease and do not reverse the lung fibrosis [11, 12]. Owing to the complexity, heterogeneity, and multifactorial nature of IPF pathogenesis, combination therapies are hypothesized to be more effective than monotherapies [13].

In this study, we applied our methodology and identified several novel candidate causal genes potentially important in the progression of IPF. Several of these were found to be significantly correlated with the forced vital capacity (FVC) and diffusing capacity of lungs for carbon monoxide (DLCO) lung function traits. Furthermore, we ran machine learning models to evaluate whether these mediator genes can predict the phenotype status using independent study cohorts. Finally, we ran a DL-based compound screening method to identify small-molecule candidates that can potentially modulate the IPF causal signature.

2. Results

2.1. Transcriptomic Datasets

In our IPF case study, we implemented a network analysis framework on a gene co-expression network, layered with a statistical mediation analysis. For constructing the gene network, we used RNA-seq data from whole lung tissues of 103 IPF samples and 103 healthy controls [14]. Additionally, we used two other IPF RNA-seq datasets, GSE124685 [15] and GSE213001 [16], to validate the biomarker candidates identified in the network analysis step (Table 1). All three publicly available datasets were downloaded from NCBI's Gene Expression Omnibus (GEO) repository [17]. For compound screening, we used the Library of Integrated Network-Based Cellular Signatures (LINCS) [18], a small-molecule gene expression database of drug perturbation profiles.

Table 1. List of IPF transcriptomic datasets used.

GEO Accession ID	# IPF	# Controls	Reference
GSE150910	103	103	[14]
GSE124685	49 (19=mild (IPF1); 16=moderate (IPF2); 14=advanced (IPF3))	35	[15]
GSE213001	61 (22=Advanced, 27=Severe)	40	[16]

2.2. Correlated Gene Modules

Gene expression profiles from whole lung tissues (GSE150910) from both IPF and control samples were used for constructing gene co-expression networks and running the WGCNA algorithm [7]. Before this, we normalized the RNA-seq counts using the *voom* (variance modeling at the observational level) method [19]. We identified sixteen non-overlapping gene modules (Figure 2), out of which seven modules were significantly (Student asymptotic p-value < 0.05) correlated/upregulated in IPF (Supplementary File 1). The *greenyellow* module (486 genes) was the most significant module (Figure 2) that contained genes predominantly involved in extracellular matrix organization (ECM) (GO:0030198) and collagen fibril organization (GO:0030199). Furthermore, five out of the seven correlated modules (*greenyellow, brown, yellow, red,* and *purple)* were strongly associated with lung function (FVC and DLCO) (Figure 2). Finally, we used these seven significantly correlated modules to run the mediation analysis to identify potential candidate causal genes.

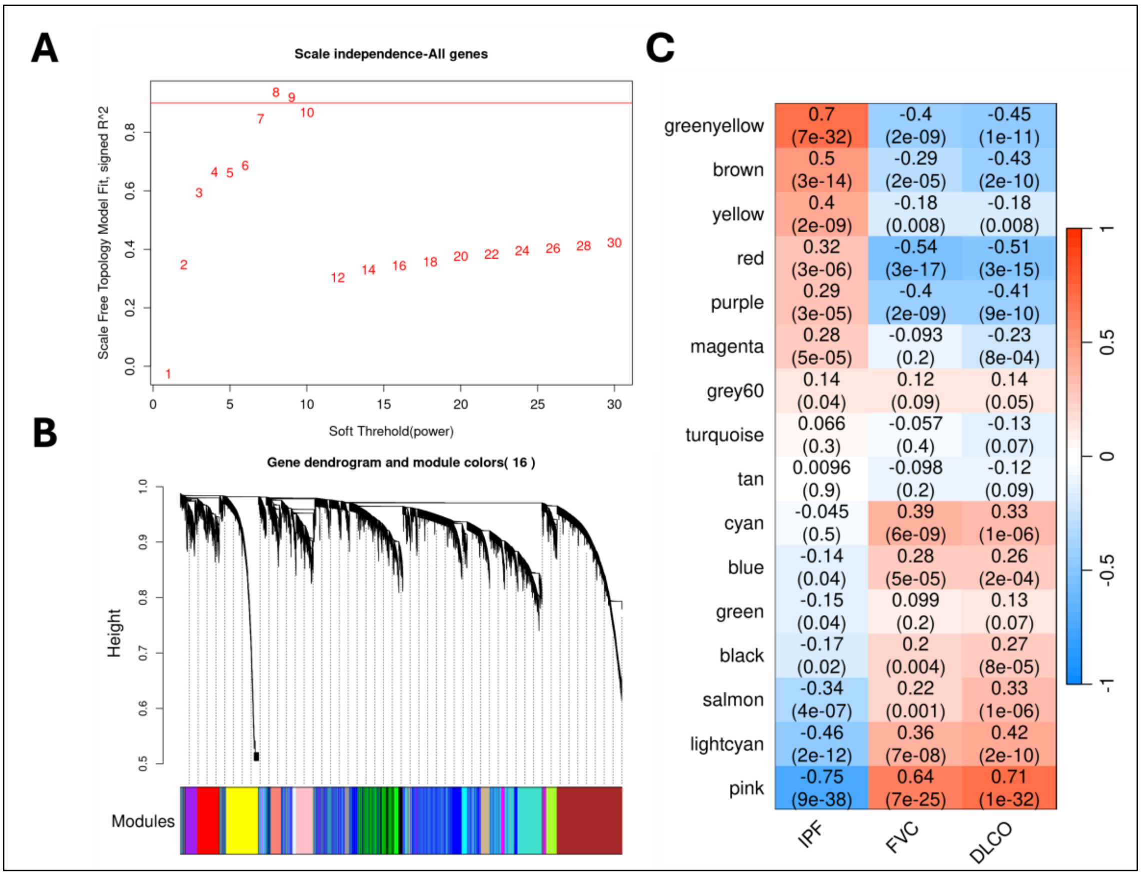

Figure 2. **WGCNA Steps: (a).** Scatter plot of the WGCNA scale-free analysis experiments on the IPF transcriptomic data. **(b).** Hierarchical dendrogram of gene modules constructed based on the topological overlap matrix (TOM). **(c).** Correlation heatmap of gene modules with the phenotypic status and the lung-function traits (FVC & DLCO) from the transcriptomic cohort.

2.3. Candidate Causal Genes in IPF

We implemented mediation analysis among the phenotype, the module, and the module genes using the CWGCNA framework [9]. Specifically, we applied bidirectional mediation models (See Materials and Methods) for each of the seven candidate WGCNA modules and identified 145 unique mediator genes that are potentially drivers of the disease progression in IPF (Supplementary File 2). The majority of them (114 out of 145 genes) were found to be significantly upregulated ($log_2FC >$ 0.58 and adjusted p-value < 0.05) in IPF. Similarly, 101 out of the 145 genes were significantly associated (adjusted p-value < 0.05) with both FVC and DLCO lung function traits (Supplementary File 2). We also found that 35/145 genes are part of the druggable genome collection, indicating their potential as therapeutic targets. This gene list was manually compiled from multiple sources, including DrugBank [20], DGIDb (v5.0) [21], CLUE-Drug repurposing hub [22]. All mediation analysis models were adjusted for any significant confounding factors (age and smoking status) identified before training the models (See Materials and Methods).

2.4. Candidate Genes Enriched in Pro-Fibrotic Niches

To identify potential disease-driving mediator candidates localized in the lung tissue, we utilized a spatial transcriptomics study in IPF [23]. In this study, Mayr et al., identified three disease-associated niches (fibrotic niche, the airway macrophage niche, and the immune niche) with unique cellular compositions and localizations. Intersecting the niche markers ($log_2FC > 0.25$ and adjusted p-value < 0.05) with the significant mediator candidates, we identified four genes (CRABP2, MKI67, PRDX4, PLPP5) found in all three IPF niches (Supplementary File 3). In one of our earlier studies, we demonstrated that CRABP2, a retinol-binding protein, is dysregulated in multiple lung cells of IPF and is strongly correlated with decline in lung function [24]. Additionally, we identified 24 genes (p-value < 0.05) belonging to the fibrotic niche, consisting of myofibroblasts and aberrant basaloid cells, located around the airways [23]. Similarly, we identified 20 genes (not statistically significant; p-value = 0.3) from the airway macrophage niche in the lumen containing SPP1$^+$ (Osteopontin) macrophages (Supplementary File 3). Our mediator gene list also included 36 genes (p-value < 0.05) from the immune niche within IPF tissue, predominantly characterized by lymphoid cells surrounded by remodeled endothelial cells. In total, our mediation analysis identified 54 unique candidates that are marker genes in at least one of the three pro-fibrotic niches (Supplementary File 3).

2.5. Mediator Genes Associated with IPF Severity

Next, to identify candidate causal genes associated with severe IPF phenotype, we intersected the significant mediator genes with the upregulated progressive stage and end-stage markers ($log_2FC > 0.58$ and adjusted p-value < 0.05) from a lung RNA-seq dataset (GSE124685) [15]. The study used quantitative micro-CT imaging and tissue histology on IPF lung samples to categorize three distinct stages of fibrotic remodeling (early-stage or IPF1, progressive-stage or IPF2, and end-stage or IPF3). We identified 40 mediator genes (p-value = 4.69E-13) that were upregulated in end-stage IPF and 38 candidate genes (p-value = 8.07E-14) overexpressed in progressive-stage IPF samples. There were 30 mediator genes upregulated in both IPF2 and IPF3 samples. Additionally, we also found 40 candidate genes (p-value = 2.88E-16) that were up in early-stage IPF (Supplementary File 4).

2.6. Biomarker Candidates in IPF

To assess the predictive ability of the mediator candidate genes and identify potential biomarker candidates in IPF, we ran generalized linear models (GLMs) via penalized maximum likelihood [25, 26]. We used an independent IPF cohort (Table 1) consisting of 61 IPF cases (including 22 advanced and 27 severe cases) and 40 control samples to train the GLM models. These models were evaluated using the area under the precision-recall curves (AUPRC) and receiver operating characteristic (AUROC) curves (See Materials and Methods). We identified several candidate genes (ITM2C, PRTFDC1, CRABP2, CPNE7, FAM83D, NMNAT2, P4HA3, PDGFD, PAPP2) with high area under the curve (AUC >0.9) scores (Table 2). Among them, the integral membrane protein 2C (ITM2C) gene performed the best with high AUPRC and AUROC scores (Figure 4). The ITM2C gene is a known marker of plasmablast cells in the lung [27], which are implicated in the progression of IPF [28].

We also trained GLM models to distinguish severe and advanced IPF samples from controls (Figure 4). Based on the AUC scores, we identified seven gene candidates (ITM2C, CPNE7, PRTFDC1, CRABP2, NMNAT2, LAX1, PAPPA2) with strong predictive performance in classifying severe IPF samples (Table 2). Similarly, we found nine additional gene candidates (ITM2C, CRABP2, PRTFDC1, FAM83D, NMNAT2, MYOF, P4HA3, CDH3, CPNE7) highly capable of distinguishing advanced IPF samples from the healthy controls (Table 2 and Supplementary File 5). Five of the mediator candidates (ITM2C, PRTFDC1, CRABP2, CPNE7, and NMNAT2) were strong predictive markers across both sets of models (severe and advanced).

Table 2. Best performing causal mediator genes in categorizing the IPF samples (including severe and advanced IPF) from healthy controls in an independent validation cohort (GSE213001). The table lists the AUPRC and AUROC scores averaged across 100 different experiments (both mean and standard deviation). The full list of AUC scores can be found in Supplementary File 5.

	Gene	AUPRC	AUROC
IPF vs Controls	ITM2C	0.993 ± 0.006	0.987 ± 0.01
	PRTFDC1	0.989 ± 0.007	0.983 ± 0.011
	CRABP2	0.984 ± 0.01	0.971 ± 0.018

	CPNE7	0.969 ± 0.018	0.955 ± 0.024
	FAM83D	0.965 ± 0.018	0.943 ± 0.029
	NMNAT2	0.964 ± 0.006	0.945 ± 0.026
	P4HA3	0.95 ± 0.03	0.933 ± 0.034
	PDGFD	0.945 ± 0.031	0.932 ± 0.036
	PAPPA2	0.943 ± 0.023	0.902 ± 0.04
Severe IPF vs Controls	ITM2C	0.996 ± 0.004	0.997 ± 0.003
	CPNE7	0.97 ± 0.036	0.98 ± 0.026
	PRTFDC1	0.969 ± 0.024	0.975 ± 0.02
	CRABP2	0.96 ± 0.022	0.958 ± 0.025
	NMNAT2	0.943 ± 0.018	0.956 ± 0.014
	LAX1	0.943 ± 0.039	0.957 ± 0.029
	PAPPA2	0.92 ± 0.042	0.927 ± 0.045
Advanced IPF vs Controls	ITM2C	0.996 ± 0.006	0.998 ± 0.004
	CRABP2	0.991 ± 0.013	0.994 ± 0.008
	PRTFDC1	0.983 ± 0.017	0.99 ± 0.01
	FAM83D	0.976 ± 0.02	0.986 ± 0.011
	NMNAT2	0.947 ± 0.033	0.968 ± 0.021
	MYOF	0.936 ± 0.033	0.955 ± 0.026
	P4HA3	0.927 ± 0.058	0.965 ± 0.022
	CDH3	0.923 ± 0.047	0.949 ± 0.029
	CPNE7	0.903 ± 0.058	0.936 ± 0.038

2.7. Small-Molecule Compounds Targeting the Causal Genes

Finally, we used a modified version of the DeepCE [2] model (See Materials and Methods) to identify compounds targeting the potentially causal genes in IPF. Specifically, we used 145 significant mediator genes (Section 2.3) as the query signature and implemented compound screening by computing Spearman's rank-order correlation between the drug perturbation gene expression profiles and the IPF-specific signature. We identified 126 small-molecule compounds with significant negative correlation scores (p-value < 0.05) in any of the seven different cell lines used in the model (Supplementary File 6). Of these, 23 compounds (Table 3) exhibited significant negative correlations in multiple (two or more) cell lines. Among them, the compound *Telaglenastat* (CB-839), a glutaminase 1 (GLS1) inhibitor, was significantly negatively correlated with our causal signature in five out of the seven cell lines. This compound has been shown to be therapeutically efficacious in treating both bleomycin- and transforming growth factor–β1–induced pulmonary fibrosis in mice [29]. GLS1 inhibition is also shown to regulate collagen production in human lung myofibroblasts by promoting its stabilization [29, 30]. Similarly, the kinase inhibitor *Merestinib* targets the tyrosine kinase receptor MET (hepatocyte growth factor) and is being investigated as a potential anti-cancer therapy candidate [31]. The MET-signaling pathway has been implicated in the epithelial-mesenchymal transition (EMT) and fibroblast activation in IPF [32, 33]. Consequently, MET has been proposed as a functional marker and an actionable target in IPF [34]. Our analysis also highlighted *Cilostazol,* a Phosphodiesterase 3 (PDE 3) inhibitor approved for use as an antiplatelet agent and vasodilator, for the symptomatic relief of intermittent claudication. Other studies have demonstrated its antifibrotic effects in a rat model of liver fibrosis [35] and cardiac fibrosis in mice [36]. We also identified the ROCK inhibitor *Fasudil*, which is known to ameliorate experimental pulmonary fibrosis by inhibiting mechanosensitive signaling in myofibroblasts [37] (Supplementary File 6).

Table 3. List of top compounds with significant negative correlation (Spearman rank-order) with the causal IPF signature. The correlations in each of the seven distinct cell lines, along with the significance p-values, are listed here. Supplementary File 6 has the full list of drugs.

Drug	A375	HA1E	HELA	HT29	MCF7	PC3	YAPC
1,4-Bis((3,4-dimethoxyphenyl)sulfonyl)-1,4-diazepane	-0.25 (0.051)	**-0.27 (0.036)**	**-0.34 (0.006)**	**-0.28 (0.027)**	**-0.33 (0.009)**	-0.25 (0.053)	**-0.28 (0.03)**
CB-839 (Telaglenastat)	**-0.29 (0.023)**	0.04 (0.76)	**-0.29 (0.02)**	**-0.26 (0.044)**	**-0.25 (0.049)**	**-0.28 (0.03)**	-0.18 (0.16)
[2-(4-Amino-1,2,5-oxadiazol-3-yl)-1-ethylimidazo[4,5-c]pyridin-7-yl]-[(3S)-3-aminopyrrolidin-1-yl]methanone	**-0.31 (0.014)**	-0.18 (0.17)	**-0.26 (0.041)**	**-0.32 (0.014)**	**-0.31 (0.015)**	**-0.29 (0.02)**	-0.19 (0.13)
Aminofurazanyl-azabenzimidazole 6n	**-0.35 (0.005)**	**-0.27 (0.03)**	**-0.3 (0.02)**	-0.23 (0.07)	**-0.29 (0.023)**	-0.25 (0.053)	-0.19 (0.15)
[2-(4-Amino-furazan-3-yl)-1-ethyl-1H-imidazo[4,5-c]pyridin-7-ylmethyl]-piperidin-4-yl-amine	**-0.28 (0.025)**	-0.1 (0.42)	-0.25 (0.052)	**-0.3 (0.02)**	**-0.28 (0.026)**	**-0.3 (0.02)**	-0.09 (0.51)
Pentamidine	**-0.31 (0.015)**	-0.16 (0.24)	-0.21 (0.098)	-0.21 (0.11)	**-0.26 (0.046)**	**-0.26 (0.044)**	**-0.27 (0.034)**
RHC-80267	-0.22 (0.089)	0.08 (0.52)	-0.18 (0.16)	**-0.35 (0.006)**	**-0.29 (0.023)**	-0.23 (0.075)	**-0.26 (0.039)**
RK-682	**-0.27 (0.033)**	0.12 (0.35)	0.019 (0.88)	0.22 (0.097)	-0.18 (0.17)	-0.22 (0.091)	**-0.29 (0.025)**
Merestinib	-0.08 (0.54)	-0.09 (0.48)	**-0.33 (0.008)**	-0.06 (0.64)	**-0.28 (0.03)**	-0.07 (0.57)	-0.05 (0.73)
LY-255283	-0.25 (0.053)	0.17 (0.2)	-0.18 (0.16)	**-0.26 (0.047)**	-0.2 (0.12)	**-0.28 (0.028)**	-0.07 (0.58)
Cilostazol	**-0.31 (0.015)**	0.018 (0.89)	-0.15 (0.25)	-0.16 (0.22)	-0.11 (0.41)	-0.06 (0.67)	**-0.28 (0.033)**
M2-PK-activator	-0.19 (0.15)	**-0.26 (0.04)**	-0.21 (0.11)	0.25 (0.05)	-0.22 (0.089)	-0.23 (0.07)	**-0.29 (0.02)**
NNC-711	-0.11 (0.4)	-0.03 (0.79)	-0.2 (0.11)	**-0.3 (0.018)**	**-0.27 (0.038)**	0.1 (0.44)	0.009 (0.95)
CID 11973736	-0.23 (0.077)	-0.2 (0.12)	**-0.33 (0.009)**	-0.24 (0.06)	**-0.32 (0.014)**	-0.18 (0.16)	-0.07 (0.6)
Azeloyl diethyl salicylate	-0.15 (0.26)	0.1 (0.45)	**-0.34 (0.008)**	**-0.32 (0.011)**	-0.16 (0.22)	-0.19 (0.13)	0.02 (0.88)
Cetrimonium	**-0.31 (0.015)**	-0.04 (0.78)	-0.15 (0.26)	0.23 (0.07)	-0.2 (0.13)	**-0.28 (0.027)**	-0.24 (0.06)
Decamethonium	**-0.26 (0.043)**	-0.04 (0.75)	-0.06 (0.65)	**-0.26 (0.047)**	-0.23 (0.072)	-0.23 (0.07)	-0.24 (0.056)
Gemcadiol	**-0.29 (0.021)**	-0.004 (0.98)	-0.06 (0.63)	**-0.25 (0.049)**	-0.14 (0.29)	-0.25 (0.056)	0.24 (0.066)
Clofilium	**-0.28 (0.026)**	0.23 (0.079)	-0.19 (0.14)	-0.25 (0.051)	-0.14 (0.29)	**-0.27 (0.03)**	-0.06 (0.64)

1-[[4,5-Bis(4-methoxyphenyl)-2-thiazolyl]carbonyl]-4-methylpiperazine	-0.13 (0.33)	-0.15 (0.24)	**-0.29 (0.025)**	-0.17 (0.19)	**-0.28 (0.029)**	-0.17 (0.19)	0.04 (0.78)
Zindotrine	-0.03 (0.82)	0.17 (0.18)	0.05 (0.68)	**-0.28 (0.03)**	-0.19 (0.15)	0.04 (0.74)	**-0.31 (0.015)**
Eliglustat	-0.04 (0.78)	-0.01 (0.94)	**-0.26 (0.04)**	**-0.27 (0.036)**	-0.1 (0.44)	-0.16 (0.22)	-0.01 (0.93)
3-(Azepan-1-ylsulfonyl)-N-(3-bromophenyl)benzamide	-0.02 (0.87)	-0.25 (0.054)	**-0.26 (0.04)**	-0.08 (0.54)	-0.11 (0.43)	-0.08 (0.53)	**-0.25 (0.047)**

3. Discussion

In this study, we presented a therapeutic target and compound discovery approach by combining unsupervised network analysis, statistical mediation analysis, and Deep Learning (DL) models. We applied this framework to IPF and identified 16 non-overlapping gene modules, including seven candidate modules significantly upregulated in the phenotype. From these modules, we identified 145 mediator gene candidates with significant mediation effects, after adjusting for confounders. These genes are hypothesized to *cause* the sample group differences by being involved in high-level pathways or biological processes (modules) [9]. Indeed, several of them were found to be localized in pro-fibrotic niches in the lung tissue. We also identified 35 known drug targets among these mediator genes. Additionally, machine learning models using these mediator genes demonstrated their predictive ability to categorize IPF samples from healthy controls and also for classifying severe and advanced IPF cases from controls with strong performance metrics.

The significant mediator genes were used as the phenotype signature for compound screening to identify small-molecule candidates that can potentially reverse the IPF phenotypic effects. By using Spearman rank-order correlations between the IPF signature and drug perturbation profiles, we identified more than 120 compounds with significant negative correlations with the IPF signature across different cell lines. Several of our identified candidate drugs have been reported to have therapeutic (antifibrotic) effects in *in vitro* and animal models. Further validation studies centered around some of the other small-molecule candidates are warranted to identify new therapeutic strategies for attenuating or reversing lung fibrosis.

Phenotype-based compound screening is generally considered systematic and comprehensive as it utilizes drug-induced gene expression profiles as readouts [2]. In comparison to previous methods, our methodology has the advantage of combining network analysis with causal inference, using bidirectional mediation models, to generate phenotype-specific signatures. This gives us the ability to potentially encode high-level causal mechanisms in these signatures. Consequently, the screened molecules are hypothesized to be safe and efficacious. Moreover, we used a customized DL model [2] for predicting robust drug perturbation profiles across multiple cell lines from otherwise noisy expression data (See Materials and Methods). Our customized model offers higher coverage by predicting expression profiles in the best inferred genes (BING) space instead of just the 978 landmark genes [18]. This enables capturing more comprehensive phenotypic signatures for compound screening. Our framework, including the experiments in the IPF use case, can be freely accessed from https://github.com/SudhirGhandikota/Causal_compound_screening.

Our approach does hold certain limitations. Firstly, our mechanism-driven compound screening approach is highly dependent on the input signature and the quality of the gene expression profiles used in the analysis. Irrelevant target genes within the signature can lead to false-positive small-molecule hits. Secondly, our mediation models assume linear relationships between the phenotype and the modules and therefore cannot identify non-linear effects. Also, unmeasured confounding effects can impact the statistical mediation models, which in turn can influence the causal phenotype signatures used in the compound screening step. Hence, proper attention should be given to test and

doi:10.20944/preprints202507.1009.v1

adjust for all possible confounding variables during the mediation analysis step. Finally, noisy and unreliable perturbation profiles can lead to unsuccessful compound hits. To manage this issue, we utilized a DL model that employs an effective data augmentation technique, which can extract useful information from noisy or unreliable expression profiles [2].

As a future direction, we plan to develop compound screening approaches that can leverage single-cell expression profiles and spatially resolved cellular niches towards drug discovery applications. Using such frameworks, we can also include cell-cell interaction information to potentially capture and adjust for off-target activity of the identified drug compounds.

4. Materials and Methods

4.1. WGCNA and Candidate Correlated Modules

The weighted gene co-expression network analysis (WGCNA) framework is useful for identifying modules of highly co-expressed genes. Briefly, WGCNA uses pairwise Pearson correlations for constructing gene networks and converts them into topological overlap matrices that encode relative gene interconnectedness within the network [8]. Then, average linkage hierarchical clustering with adaptive branch pruning is implemented to identify modules of strongly co-expressed genes (Figure 2B) [38]. We used signed co-expression networks with the power parameter $(\beta = 8)$ chosen based on the scale-free topology criterion (Figure 2A) [8]. Then, module eigengenes are identified to compute the module-level correlations with the phenotype status and any available clinical variables. These eigengenes are generally considered to be a module representative summarizing the expression patterns within a module and are identified through principal component analysis (PCA). Additionally, to assess the significance of the module-trait correlations, we calculate the Student asymptotic p-values for each correlation distribution (Figure 2C). Candidate modules were filtered using a significance threshold of p-value < 0.05.

4.2. Mediation Analysis and Candidate Causal Genes

We implemented statistical mediation analysis to reveal the causal relationships between the modules and the phenotype. Specifically, we used the CWGCNA package [9] to construct mediation models using the WGCNA module, the phenotype status, and the module genes. CWGCNA uses bidirectional models to test two causal directions: (a) the forward direction of module driving the phenotype via the gene (module -> gene -> phenotype), and (b) the reverse direction of phenotype causing the module changes via the gene (phenotype -> gene -> module). These tests, conducted using the inverse probability weighting (IPW) method [39], help clarify the cause and the mediators of a given phenotypic condition. In both models, the module genes are used as the potential mediator and tested for their mediating effect. For each gene, the proportion mediated was computed by combining both the natural and the indirect effects, and 95% confidence intervals (CIs) were estimated via bootstrapping (100 times). Additionally, CWGCNA applies a bagging-SMOTE (synthetic minority oversampling technique) framework to prevent the sample size differences from influencing the mediation analysis step [9].

We applied mediation analysis models using the seven modules that showed positive correlations with the IPF phenotype (Figure 3). Before this, type-III ANOVA analysis was implemented to test for any significant confounding effects. Both gender and smoking status were determined to be significant confounding variables for the phenotype status. Next, we tested the mediation effects of each module gene from the seven correlated modules, in both the forward and reverse directions. In total, we identified 156 genes with significant mediation effects uniquely in the forward direction (causal to phenotype). These genes were used in all the downstream analyses, including the phenotype-based compound screening step.

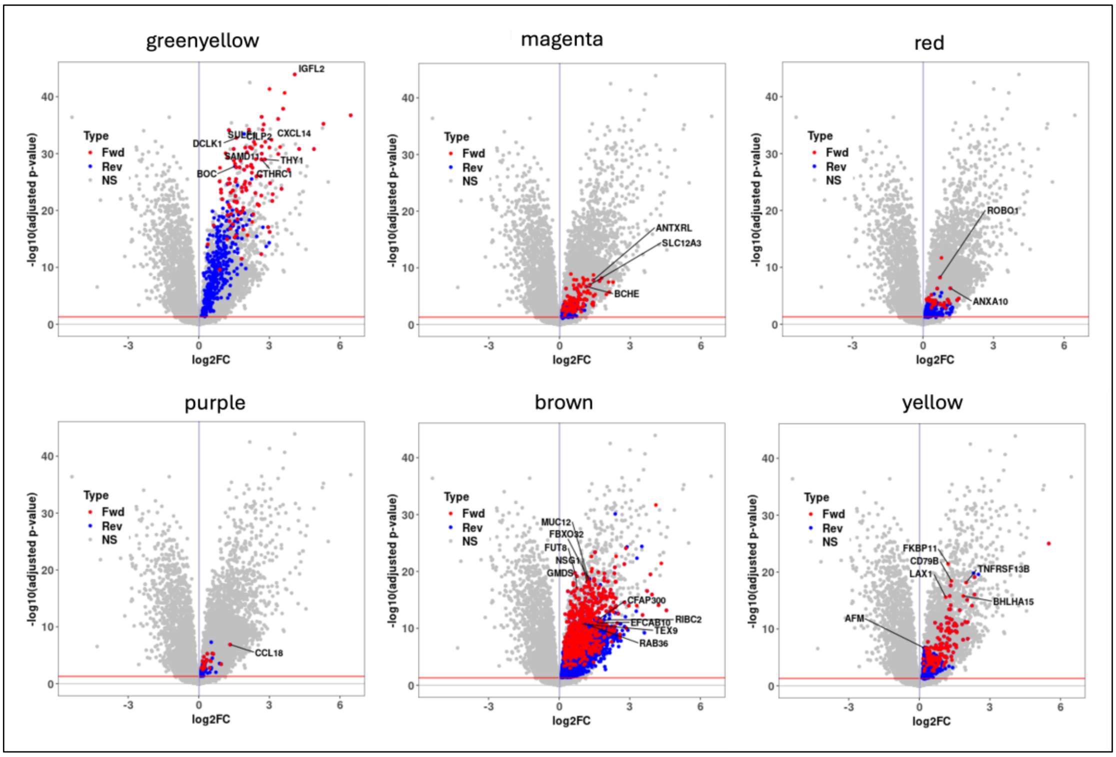

Figure 3. Volcano plots of the six candidate modules with significant mediator genes (red -> forward; blue -> reverse) highlighted in each of them.

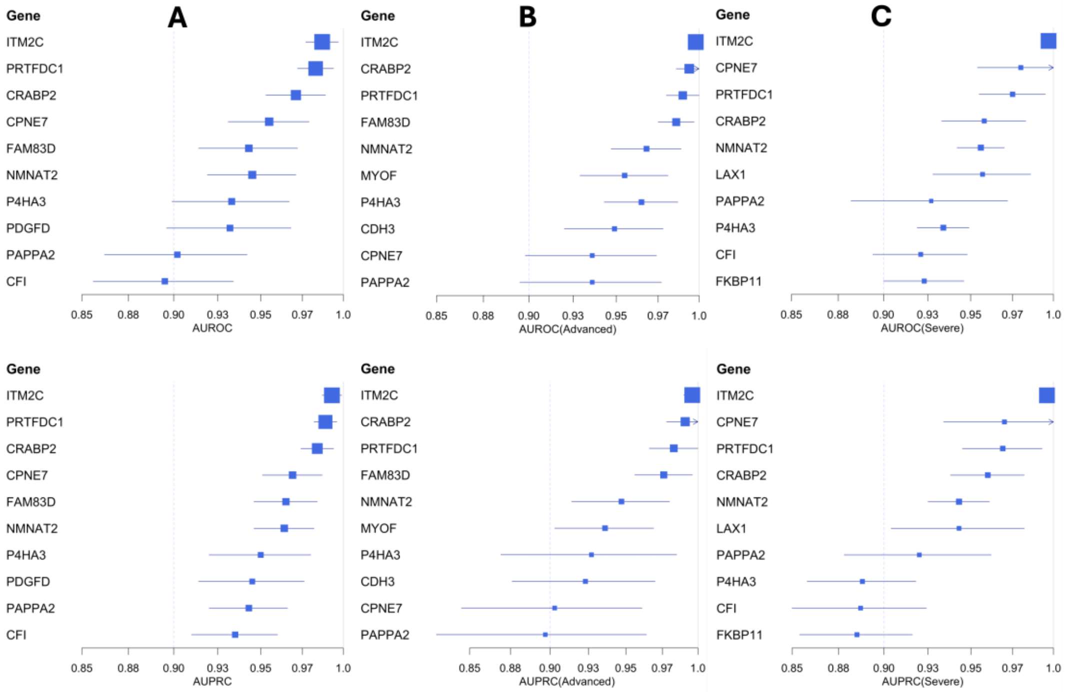

Figure 4. Forest plots of the top-performing biomarker candidates in classifying IPF samples (Panel A), including both advanced IPF (Panel B) and severe IPF (Panel C) samples from healthy controls. These biomarker genes are highlighted based on the AUPRC and AUROC scores.

4.3. Machine Learning Models for Biomarker Analysis

For identifying a subset of biomarker candidate genes, we used generalized linear models (GLMs) that were trained with the elastic net penalty [25]. We implemented a leave-one-out crossvalidation (LOOCV) approach to perform parameter tuning. In each training experiment, 70% of all samples were randomly picked for training the model and performing parameter tuning, and the remaining 30% test samples were used for model evaluation. The area under the curve (AUC) values of both precision-recall curves and the receiver operating characteristic (ROC) curves were computed to quantify the performance of these trained models. Furthermore, we conducted 100 distinct randomized experiments with separate training and validation data splits to compute robust evaluation metrics by averaging them across all the runs. Finally, in our IPF case study, we followed the same framework for models in both the training dataset (GSE150910) and the independent transcriptomic cohort (GSE213001).

4.4. DeepCE Model for Small-Molecule Screening

For screening small-molecule candidates, we followed the mechanism-driven approach used in [2]. Specifically, we used a modified version of the DeepCE model trained to predict the small-molecule perturbation profiles. It uses a graph neural network (GNN) to model the chemical substructure information [40] encoded in the form of a compound graph. Additionally, DeepCE also employs a network embedding model [41] for encoding gene-gene interaction networks [42]. It is trained using the expression profiles from the level 5 LINCS L1000 dataset (based on 978 landmark genes) [18]. DeepCE also includes experimental information, such as the cell line information and the drug dosage, in the model training step. Furthermore, a data augmentation technique is used, wherein a high-quality training set is first generated to train the model. The trained model is then used to predict and filter the gene expression profiles associated with the unreliable experiments in the L1000 data [2].

To increase the coverage of the small-molecule perturbation profiles, we trained a new DeepCE model using the level 5 L1000 data of 10,174 BING genes (978 landmark + 9,196 well-inferred genes with high fidelity). We followed the same procedure to train the new model by selecting the gene expression profiles from the seven most popular cell lines (MCF7, A375, HT29, PC3, HA1E, YAPC, HELA) and six most popular dosages (0.04 μm, 0.12 μm, 0.37 μm, 1.11 μm, 3.33 μm, 10.0 μm). We reused the high-quality signatures to train the new model and predicted the perturbation levels of approximately 2000 small molecules and nearly 6000 compounds from the BROAD Drug Repurposing Hub [22] across all seven cell lines. The model training was conducted on a workstation with Intel(R) Xeon(R) W-2133 CPU 3.60GHz, 64 GB RAM, and Nvidia Quadro RTX 8000 GPU. For screening the candidate compounds, we use Spearman rank-order correlations between the causal phenotype signature (Section 4.2) and the predicted perturbation profiles. We also computed the p-values for these correlations using a hypothesis test whose null hypothesis is that the drug and the signature have no ordinal correlation. Finally, compounds that demonstrated significant negative correlations with the phenotype signature across multiple cell lines are shortlisted as likely drug repurposing candidates.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/doi/s1, Figure S1: title; Table S1: title; Video S1: title.

Author Contributions: Conceptualization, S.G. and A.G.J.; methodology, S.G.; formal analysis, S.G.; investigation, S.G. and A.G.J.; resources, S.G. and A.G.J.; data curation, S.G. and A.G.J.; writing—original draft preparation, S.G.; writing—review and editing, S.G. and A.G.J.; visualization, S.G.; supervision, A.G.J.. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All data supporting reported results can be found in the manuscript and Supplementary Material.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sun, D., et al., *Why 90% of clinical drug development fails and how to improve it?* Acta Pharmaceutica Sinica B, 2022. **12**(7): p. 3049-3062.
2. Pham, T.H., et al., *A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing.* Nat Mach Intell, 2021. **3**(3): p. 247-257.
3. Xie, L., et al., *Towards Structural Systems Pharmacology to Study Complex Diseases and Personalized Medicine.* PLOS Computational Biology, 2014. **10**(5): p. e1003554.
4. Zhavoronkov, A., et al., *Deep learning enables rapid identification of potent DDR1 kinase inhibitors.* Nature Biotechnology, 2019. **37**(9): p. 1038-1040.
5. Lounkine, E., et al., *Large-scale prediction and testing of drug activity on side-effect targets.* Nature, 2012. **486**(7403): p. 361-367.
6. Ghandikota, S.K. and A.G. Jegga, *Chapter Seven - Application of artificial intelligence and machine learning in drug repurposing*, in *Progress in Molecular Biology and Translational Science*, V. Singh, Editor. 2024, Academic Press. p. 171-211.
7. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**(1): p. 559.
8. Zhang, B. and S. Horvath, *A General Framework for Weighted Gene Co-Expression Network Analysis.* 2005. **4**(1).
9. Liu, Y., *CWGCNA: an R package to perform causal inference from the WGCNA framework.* NAR Genom Bioinform, 2024. **6**(2): p. lqae042.
10. Kaur, A., S.K. Mathai, and D.A. Schwartz, *Genetics in Idiopathic Pulmonary Fibrosis Pathogenesis, Prognosis, and Treatment.* Frontiers in Medicine, 2017. **4**.
11. Maher, T.M., *Interstitial Lung Disease: A Review.* JAMA, 2024. **331**(19): p. 1655-1665.
12. Wang, Y., et al., *Pan-transcriptome-based candidate therapeutic discovery for idiopathic pulmonary fibrosis.* Therapeutic Advances in Respiratory Disease, 2020. **14**: p. 1753466620971143.
13. Kolb, M., F. Bonella, and L. Wollin, *Therapeutic targets in idiopathic pulmonary fibrosis.* Respiratory Medicine, 2017. **131**: p. 49-57.
14. Furusawa, H., et al., *Chronic Hypersensitivity Pneumonitis, an Interstitial Lung Disease with Distinct Molecular Signatures.* Am J Respir Crit Care Med, 2020. **202**(10): p. 1430-1444.
15. McDonough, J.E., et al., *Transcriptional regulatory model of fibrosis progression in the human lung.* JCI Insight, 2019. **4**(22).
16. Jaffar, J., et al., *Matrix metalloproteinase-7 is increased in lung bases but not apices in idiopathic pulmonary fibrosis.* ERJ Open Res, 2022. **8**(4).
17. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets--update.* Nucleic Acids Res, 2013. **41**(Database issue): p. D991-5.
18. Subramanian, A., et al., *A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles.* Cell, 2017. **171**(6): p. 1437-1452 e17.
19. Law, C.W., et al., *voom: precision weights unlock linear model analysis tools for RNA-seq read counts.* Genome Biology, 2014. **15**(2): p. R29.
20. Wishart, D.S., et al., *DrugBank: a comprehensive resource for in silico drug discovery and exploration.* Nucleic Acids Res, 2006. **34**(Database issue): p. D668-72.
21. Cannon, M., et al., *DGIdb 5.0: rebuilding the drug-gene interaction database for precision medicine and drug discovery platforms.* Nucleic Acids Res, 2024. **52**(D1): p. D1227-D1235.
22. Corsello, S.M., et al., *The Drug Repurposing Hub: a next-generation drug library and information resource.* Nat Med, 2017. **23**(4): p. 405-408.
23. Mayr, C.H., et al., *Spatial transcriptomic characterization of pathologic niches in IPF.* Science Advances. **10**(32): p. eadl5473.
24. Ghandikota, S., et al. *Consensus Gene Co-Expression Network Analysis Identifies Novel Genes Associated with Severity of Fibrotic Lung Disease*. International Journal of Molecular Sciences, 2022. **23**, DOI: 10.3390/ijms23105447.
25. Friedman, J.H., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Journal of Statistical Software, 2010. **33**(1): p. 1 - 22.

26. Tay, J.K., B. Narasimhan, and T. Hastie, *Elastic Net Regularization Paths for All Generalized Linear Models.* Journal of Statistical Software, 2023. **106**(1): p. 1 - 31.
27. Zhang, T., et al., *Single Cell RNA-Seq Identifies Cell Subpopulations Contributing to Idiopathic Pulmonary Fibrosis in Humans.* J Cell Mol Med, 2025. **29**(3): p. e70402.
28. Prele, C.M., et al., *Plasma cell but not CD20-mediated B-cell depletion protects from bleomycin-induced lung fibrosis.* Eur Respir J, 2022. **60**(5).
29. Cui, H., et al., *Inhibition of Glutaminase 1 Attenuates Experimental Pulmonary Fibrosis.* Am J Respir Cell Mol Biol, 2019. **61**(4): p. 492-500.
30. Ge, J., et al., *Glutaminolysis Promotes Collagen Translation and Stability via alpha-Ketoglutarate-mediated mTOR Activation and Proline Hydroxylation.* Am J Respir Cell Mol Biol, 2018. **58**(3): p. 378-390.
31. He, A.R., et al., *First-in-Human Phase I Study of Merestinib, an Oral Multikinase Inhibitor, in Patients with Advanced Cancer.* Oncologist, 2019. **24**(9): p. e930-e942.
32. Crestani, B., et al., *Hepatocyte growth factor and lung fibrosis.* Proc Am Thorac Soc, 2012. **9**(3): p. 158-63.
33. Stella, G.M., et al., *Ockham's razor for the MET-driven invasive growth linking idiopathic pulmonary fibrosis and cancer.* J Transl Med, 2016. **14**(1): p. 256.
34. Stella, G.M., et al., *The oncogenic landscape of the idiopathic pulmonary fibrosis: a narrative review.* Translational Lung Cancer Research, 2022. **11**(3): p. 472-496.
35. El Awdan, S.A., et al., *Regression of fibrosis by cilostazol in a rat model of thioacetamide-induced liver fibrosis: Up regulation of hepatic cAMP, and modulation of inflammatory, oxidative stress and apoptotic biomarkers.* PLoS One, 2019. **14**(5): p. e0216301.
36. Hada, Y., et al., *Cilostazol Attenuates AngII-Induced Cardiac Fibrosis in apoE Deficient Mice.* Int J Mol Sci, 2022. **23**(16).
37. Zhou, Y., et al., *Inhibition of mechanosensitive signaling in myofibroblasts ameliorates experimental pulmonary fibrosis.* The Journal of Clinical Investigation, 2013. **123**(3): p. 1096-1108.
38. Langfelder, P., B. Zhang, and S. Horvath, *Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R.* Bioinformatics, 2008. **24**(5): p. 719-20.
39. VanderWeele, T.J., *Mediation Analysis: A Practitioner's Guide.* Annu Rev Public Health, 2016. **37**: p. 17-32.
40. Weininger, D., *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules.* Journal of Chemical Information and Computer Sciences, 1988. **28**(1): p. 31-36.
41. Grover, A. and J. Leskovec. *node2vec: Scalable feature learning for networks*. in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 2016.
42. 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, 2023. **51**(D1): p. D638-D646.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.