Research Paper

AI-driven toolset for IPF and aging research associates lung fibrosis with accelerated aging

Fedor Galkin[1], Shan Chen[2], Alex Aliper[1], Alex Zhavoronkov[1,2,3,4], Feng Ren[2]

[1]Insilico Medicine AI Ltd., Abu Dhabi, UAE
[2]Insilico Medicine Shanghai, Ltd., Shanghai, China
[3]Insilico Medicine US, Inc., NY, USA
[4]Insilico Medicine Hong Kong, Ltd., Hong Kong, China

Correspondence to: Alex Zhavoronkov; **email:** alex@insilico.com
Keywords: aging, IPF, generative AI, transformer, proteomics
Received: January 17, 2025 **Accepted:** July 15, 2025 **Published:**

ABSTRACT

Idiopathic pulmonary fibrosis (IPF) is a condition predominantly affecting the elderly and leading to a decline in lung function. Our study investigates the aging-related mechanisms in IPF using artificial intelligence (AI) approaches. We developed a pathway-aware proteomic aging clock using UK Biobank data and applied it alongside a specialized version of Precious3GPT (ipf-P3GPT) to demonstrate an AI-driven mode of IPF research. The aging clock shows great performance in cross-validation (R^2=0.84) and its utility is validated in an independent dataset to show that severe cases of COVID-19 are associated with an increased aging rate. Computational analysis using ipf-P3GPT revealed distinct but overlapping molecular signatures between aging and IPF, suggesting that IPF represents a dysregulation rather than mere acceleration of normal aging processes. Our findings establish novel connections between aging biology and IPF pathogenesis while demonstrating the potential of AI-guided approaches in therapeutic development for age-related diseases.

INTRODUCTION

Idiopathic pulmonary fibrosis (IPF) is a chronic, progressive lung disease characterized by the excessive accumulation of extracellular matrix components, leading to declining lung function and ultimately respiratory failure. Predominantly affecting individuals over the age of 60, the correlation between aging and IPF underscores the importance of understanding the aging-related mechanisms contributing to its pathogenesis. Both aging and fibrotic diseases pose major healthcare challenges. As such, global deaths to fibrotic diseases have been on the rise over the last three decades and constitute 18%, according to the latest estimates [1]. Similarly, the population aging problem has been a mainstay of biomedical and political debates for decades [2, 3]. Identifying the mechanisms shared by aging and fibrosis is crucial for developing targeted therapies that can potentially benefit the global population.

Current treatments for IPF are limited and primarily focus on slowing progression rather than addressing underlying causes. Lung transplantation remains the only way to improve a patient's survival rate, and prior anti-fibrotic therapies may be considered a way to help a patient outlast the long waiting period [4, 5]. This scarcity of effective therapies stems partly from an incomplete understanding of the molecular and cellular processes driving fibrosis in the aging lung. However, novel strategies to fight IPF are emerging with many of them focusing on its aging-related nature [6].

Recent advancements in biomedical research, notably in artificial intelligence (AI), offer a new vector to developing IPF therapies. AI-driven approaches can analyze vast amounts of biological data to identify novel biomarkers, therapeutic targets, and actionable insights. The existing AI pipelines have been successfully used to analyze the aging footprints in

proteomic, transcriptomic, epigenetic and other types of omics biodata [7–9]. Today, these research technologies have matured enough to be used in more practical applications as testified by a number of granted deep aging clock patents [10–13]. Even more AI applications are actively used in biomedical research settings not focused on aging, such as target discovery, drug candidate design, clinical trial design, and others [14–16]. These AI models serve to open new classes of drugs with promising anti-aging potential, such as HIF-PH inhibitors for the treatment of irritable bowel disease [17, 18]. Models such as those from the Precious and Nach0 lineups hold the promise of enabling a fully digital mode of clinical research by emulating real-life experimental settings based on existing data [19–22]. Simultaneously, LLM-inspired genomic AI models allow scientists to discover new mechanistic and evolutionary theories of aging and explain the behavior of living systems in environments previously considered too complex to model [23–27].

The rapid pace of AI innovation and hardware improvements promotes an optimistic outlook on the future of clinical therapeutics. By extracting actionable knowledge from a diverse and vast range of biomedical studies, more connections between seemingly disparate pathological processes can be built to find the cures for conditions beyond the reach of contemporary medicine. IPF is such a disease with an unclear etiology and a paucity of life-extending clinical countermeasures, apart from lung transplantation [28]. The existing body of evidence suggests that the onset of IPF has a strong an aging-related component, suggesting that this disease may be a specific instance of the general aging process [29]. Various authors highlight that IPF is characterized by the accumulation of senescent cells, insufficient autophagy, proinflammatory environment, and mTOR deregulation, which are commonly considered hallmarks of aging [30, 31]. Other authors suggest that the activation of embryonic pathways is essential for lung regeneration, while their deregulation is common in older individuals which leads to aberrant tissue maintenance [32].

In this study, we aim to explore the similarities between aging and IPF using AI models, such as Precious3GPT (P3GPT) and aging clocks (Figure 1). By identifying age-related biomarkers and therapeutic targets and utilizing AI to predict disease status, we offer new avenues for developing novel anti-fibrotic treatments. This study represents a significant step forward in understanding and addressing the complexities of IPF and aging mechanisms with the potential to improve outcomes for patients suffering from this challenging condition.

RESULTS

Fibrosis-aware aging clock

We developed a proteomic aging clock using data from 55,319 UK Biobank participants aged 50-85 years. Overall, the clock demonstrated robust performance with a mean absolute error (MAE) of 2.68 years and R^2=0.84 in five-fold randomized cross-validation, indicating strong predictive capacity (Figure 2 and Supplementary File 5). To assess performance stability

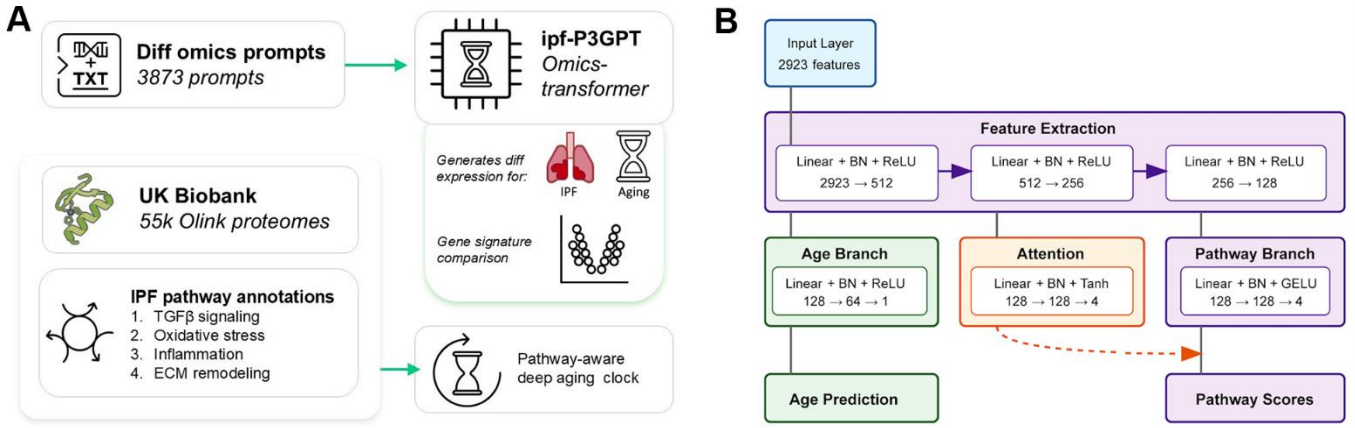

Figure 1. Deep learning models for studying IPF and fibrotic diseases. (**A**) This paper presents two deep learning models: an omics-transformer that generates differential gene expression profiles from text prompts and a pathway-aware aging clock trained on UK Biobank proteomics data. The models focus on IPF-relevant biological pathways including TGF-β signaling, oxidative stress, inflammation, and ECM remodeling. (**B**) Architecture of the pathway-aware proteomic aging clock. The neural network processes protein measurements through feature extraction layers that branch into age prediction and pathway-specific attention mechanisms, enabling interpretable aging predictions with pathway awareness.

across different age ranges, we stratified the cohort into four age brackets. The distribution of samples was weighted toward older participants, with 16.3% aged 50-59 (n=9,020), 29.4% aged 60-69 (n=16,278), 42.7% aged 70-79 (n=23,645), and 11.5% aged 80-89 (n=6,376).

Analysis of prediction errors revealed significant age-dependent variations in clock performance (ANOVA $p<0.0001$, Figure 2). The clock showed comparable MAE values for participants aged 50-59 (2.52 years), 60-69 (2.75 years), and 70-79 (2.32 years), but performance declined notably in the oldest age group (80-89 years, MAE=4.08 years). This pattern suggests that while the clock maintains strong overall predictive capacity, biological age assessment becomes more challenging in advanced age, possibly reflecting increased heterogeneity in aging mechanisms. Examining prediction biases across age brackets revealed a systematic pattern in error direction. The clock tended to overestimate ages in younger participants (mean error +2.25 years in ages 50-59), showed minimal bias in middle age groups (mean error +0.19 years in ages 60-69), and progressively underestimated ages in older participants (mean error -0.84 years in ages 70-79 and -4.07 years in ages 80-89).

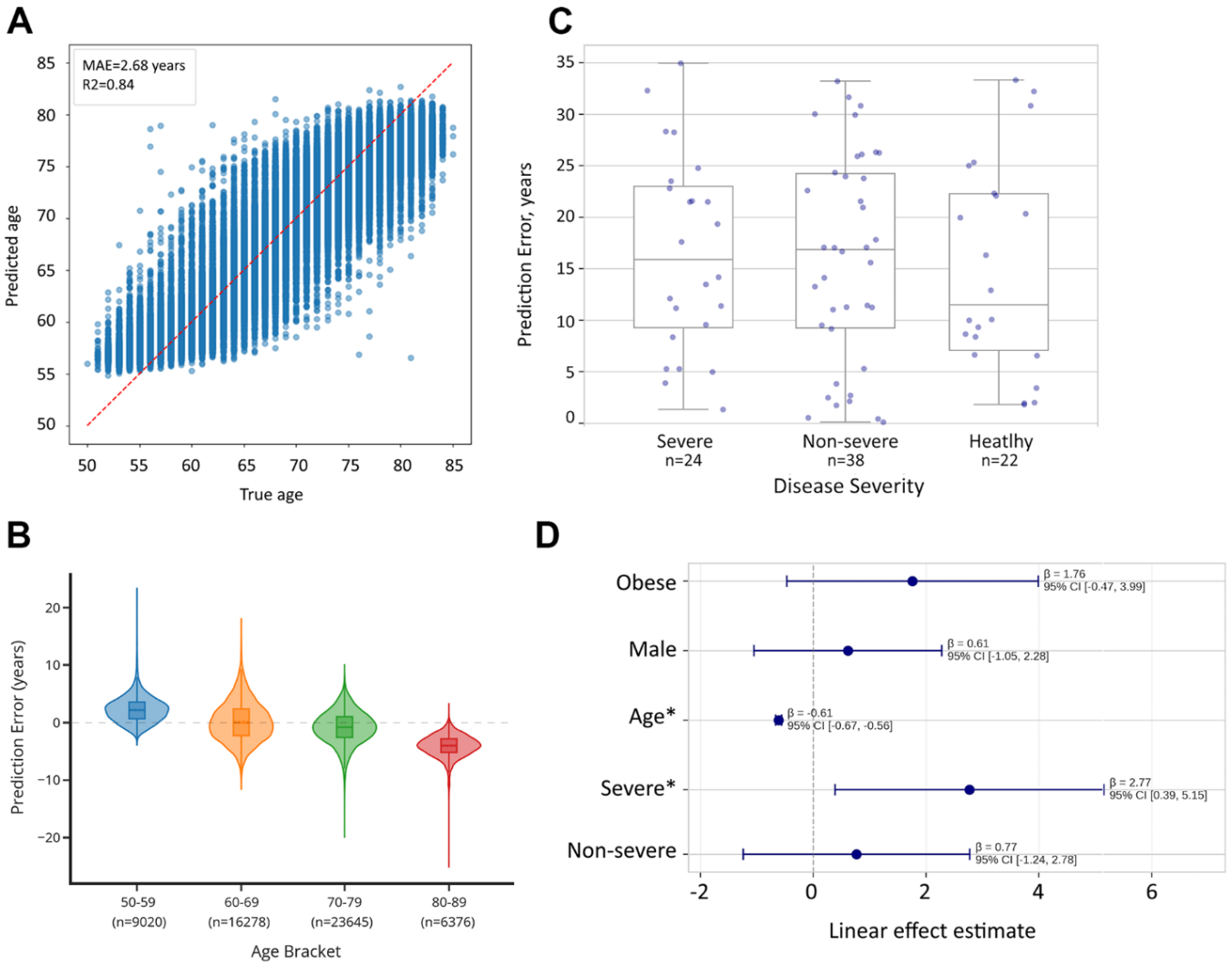

Figure 2. Comparison of biological age predictions between healthy controls and cases of severe COVID-19 infection. (**A**) Proteomic aging clock shows R2=0.84 in the task of age prediction in CV within the UK Biobank dataset (N = 55,319). (**B**) Proteomic aging clock's error depends on the age group and is skewed toward the mean of the total sample. (**C**) Biological age acceleration (difference between predicted and chronological age) across severity groups. compared to healthy controls. Error bars represent standard error of the mean. (**D**) Linear regression analysis reveals that patients with severe cases, which are likely to develop lung fibrosis, showed significantly higher biological age predictions (+2.77 years, p=0.026).

After confirming the clock's accuracy in CV, we applied it to the Olink Explore 1536 dataset from [33] featuring healthy, moderate, and severe cases of COVID-19 (N=84) with unidentified fibrosis status. Due to a large number of missing values in this dataset and non-uniform age distribution across outcome groups, a direct comparison of prediction errors between them was not feasible. Hence, we applied a linear regression method to assess the effect of disease severity on the pace of aging. Our analysis identified that the patient with a severe case of the infection, and thus likely to develop lung fibrosis, had significantly higher biological age compared to healthy controls (+2.77 years, p=0.026), suggesting that the trained clock carries biological relevance in fibrotic cases (Supplementary File 6).

ipf-Precious3GPT analysis

After exploring the function of the aging clock, we set out to identify the genes contributing to the progression of both IPF and aging by using ipf-P3GPT (Figure 3). ipf-P3GPT is an abridged version of the full-scale P3GPT that was trained on a data collection enriched in fibrotic disease from human and *in vitro* studies. Using this model, we generated two distinct gene expression profiles: one representing the classical IPF transcriptomic response and another modeling the aging process in lung tissue from 30 to 70 years old. The model assigned attention scores to each gene, indicating their relative importance in the respective biological processes.

Analysis of differentially expressed genes revealed distinct transcriptional signatures for IPF (n=96 genes) and aging-associated (n=93 genes) processes in lung tissue (Table 1 and Supplementary File 7). The overlap between these signatures was limited to 15 genes (15.6% of IPF signature), suggesting substantial divergence in the underlying molecular programs. Among the overlapping genes, 46.7% (7/15) showed concordant directional changes, while 53.3% (8/15) exhibited opposing regulation between IPF and aging conditions. This initial observation prompted us to investigate the specific molecular pathways affected in each condition.

Both aging and IPF signatures demonstrated significant involvement of ECM-associated genes, albeit with distinct regulatory patterns. IPF signature included multiple collagen types (COL1A1↓, COL3A1↓, COL5A1↑, COL15A1↑) and matrix-modifying enzymes (MMP1↑, MMP13↑). The aging signature, in contrast, showed a different matrix remodeling profile, characterized by changes in structural proteins (ELN↓, MFAP4↓, MFAP5↓) and matrix-associated factors (POSTN↓). Notably, COL1A1 showed opposing regulation between IPF (downregulated) and aging (upregulated), suggesting divergent matrix reorganization mechanisms.

Further examination revealed that the IPF signature was enriched for TGF-β pathway components (TGFBR1↓, GDF15↓, BGN↑) and inflammatory mediators (CXCL8↑, IL1B↓, CCL8↓). While the aging signature shared some inflammatory mediators (IL6↑, IL1RL1↑), it exhibited a distinct growth factor profile (FGF7↓, CSF3↑). Both conditions showed involvement of matrix-associated growth factors, though through different molecular effectors.

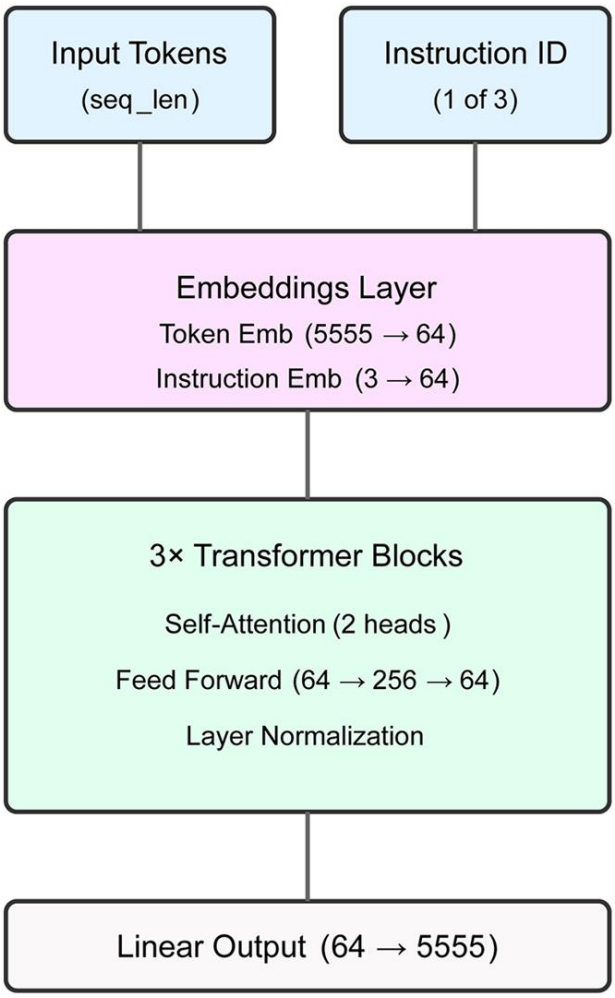

Figure 3. ipf-P3GPT model architecture. The transformer model features a 64-dimensional embedding layer, three transformer blocks with dual attention heads, and a vocabulary-sized output layer. The model processes disease, age group, and compound treatment comparisons using a custom XML-aware tokenizer, achieving 72.2-75.9% validation accuracy across instruction types.

Table 1A. Shared genes between IPF and aging-associated lung fibrosis.

	IPF		Aging	
Gene	**Direction**	**Score**	**Direction**	**Score**
Concordant genes				
IL1RL1	↑	48.8	↑	16.9
ASPN	↑	43.9	↑	9.4
PAPPA	↑	34.7	↑	12.5
F2R	↓	34.4	↓	5.6
COL15A1	↑	32.3	↑	3.4
ADAMTS1	↓	23.3	↓	2.4
IL6	↑	19.2	↑	3.6
CD83	↓	16.1	↓	3.3
Discordant genes				
COL1A1	↓	51.3	↑	10.0
SERPINA3	↓	43.2	↑	9.2
FOSB	↑	41.3	↓	7.7
EDNRB	↓	37.9	↑	14.2
APOH	↓	36.1	↑	9.4
NPTX2	↑	27.5	↓	9.2
GPC5	↓	24.8	↑	5.0
VIPR1	↓	23.4	↑	12.5
QSOX1	↑	21.9	↓	4.3

Table 1B. Key unique genes in each condition (Top 15 by attention score).

IPF-specific			Aging-specific		
Gene	**Direction**	**Attention**	**Gene**	**Direction**	**Attention**
AKT3	↓	100.0	IL10RA	↑	100.0
TUBB3	↑	88.3	MFAP4	↓	61.7
CFB	↓	86.0	FGF7	↓	57.4
PKN3	↓	70.7	S100A3	↑	49.6
CCL8	↓	63.5	SEMA3G	↓	45.7
GDF15	↓	60.3	IL11	↓	36.1
APOC1	↑	56.0	FASLG	↑	30.4
MET	↑	55.6	NTRK2	↑	25.5
BCHE	↓	55.6	VIT	↑	23.2
SFRP1	↑	52.7	MMP3	↑	20.4
AGBL2	↓	52.5	SCGB3A1	↑	19.5
DAPK2	↓	51.1	MARCO	↓	19.5
MRC1	↓	50.7	TAGLN3	↑	18.1
NRGN	↓	49.7	GZMA	↓	17.4
VEGFA	↑	47.9	FNDC1	↑	15.6

Values represent relative attention scores among the gene tokens attending to the fibrosis diagnosis by the ipf-P3GPT model. ↑ indicates upregulation, ↓ indicates downregulation.

DISCUSSION

Our study presents several significant findings regarding the intersection of aging biology and IPF progression, while introducing novel AI-driven approaches for understanding disease mechanisms and therapeutic interventions. To further the community's understanding of IPF and other fibrotic diseases, we present two deep learning models: a proteomic aging clock developed with UKB data and an abridged version of P3GPT, a transformer-based model for generating and analyzing omics-level data.

The proteomic clock we developed shows an accuracy (R^2=0.84) below another recently published aging clock ProtAge (R^2=0.94) in a different subsection of UKB [34]. Unfortunately, a direct comparison was not feasible since the authors had not deposited the weights of the model used. This performance difference can be attributed to several factors. First, our model was deliberately constrained to accommodate pathway-specific internal representations relevant to IPF pathogenesis. The integration of IPF-relevant pathway attention mechanisms introduced a trade-off between age prediction accuracy and capturing fibrosis-related biological signals. Additionally, while the UK Biobank dataset is extensive, the effective sample size for training neural architectures may be insufficient to fully capitalize on the model's capacity. We maintained a conservative approach to parameter count, particularly in the attention mechanism layers, to ensure generalization to independent datasets. This methodological choice prioritized biological relevance for fibrosis applications rather than maximizing chronological age prediction performance, which distinguishes our approach from general-purpose aging clocks.

The four hallmark pathways added to the model's attention block and secondary output head were selected based on their reported importance in IPF development [4, 35–37]. . These specific pathways consistently emerge as central to the fibrotic process in multiple tissues and represent key mechanistic nodes where aging and fibrosis intersect [38]. While numerous additional pathways play roles in IPF, these four were prioritized based on their documented centrality to disease progression and their known modulation during aging [39]. This focused approach also allowed us to maintain a lower number of model parameters given the sample size constraints while capturing the most biologically relevant signals. Yet, in future iterations we shall consider including more pathways in the model to expand its field of applications Additional pathways of interest for future development include cellular senescence, mitochondrial dysfunction, and autophagy pathways, which may provide complementary insights into age-related aspects of IPF pathogenesis. We then inspected an Olink proteomic dataset using this clock to find out if its prediction error carries biological signal. Olink platforms are rapidly gaining popularity and we have located multiple open access datasets generated with them, such as those featured in [40–44] . These datasets, however, lack sufficient chronological age annotation, describe non-fibrotic diseases, or were generated with Olink platforms measuring <1000 protein quantities. The datasets generated with the most complete Olink Explore 3072 platform remain rare and require a lengthy access procedure. Thus, we focused on the COVID-19 dataset from [33] as the most fitting for our purposes. While COVID-19 is not equivalent to IPF, we chose this dataset because severe COVID-19 cases frequently develop pulmonary fibrosis, making it biologically relevant for assessing our model's performance in fibrosis-associated conditions [45]. We acknowledge that direct validation on IPF-specific datasets would be optimal; however, comprehensive Olink proteomic data from IPF patients is currently limited in publicly accessible repositories. The COVID-19 dataset represents the most suitable available alternative given the established association between severe COVID-19 and pulmonary fibrotic changes. The application of the aging clock to this dataset revealed that more severe cases of these respiratory diseases are associated with significantly ($p<0.05$) higher age predictions. We anticipate that this trend shall be also observed and clearer in other pulmonary and fibrotic diseases. Future work will include validation against the recently completed phase-2 clinical trial of rentosertib in IPF patients, which will provide a more direct assessment of our model's utility in IPF-specific contexts [46, 47].To further investigate the similarities shared by aging and IPF, we used generative AI in the form of ipf-P3GPT which was instructed to generate transcriptomic lung IPF and aging (from 30 to 80 years) signatures. Both these signatures were identified by the model as the cases of fibrosis, and both contained genes from the key IPF-related pathways highlighted above (ECM remodeling, inflammatory signaling, TGF-β pathways). Yet, at a gene level, the two cases were quite dissimilar with only eight genes showing concordant expression in the emulated processes. Such genes include known contributors to IPF, such as COL15A1 [48]. Some known drivers of fibrosis are only present in the IPF signature: MMP1, MMP13, AKT3, IL6 [49–52]. And some key ECM components, such as COL1A1 are reported upregulated in IPF and down-regulated in normal aging, as also recorded in literature [53]. The predominance of non-overlapping genes indicates that IPF involves distinct pathological mechanisms beyond normal aging, including aberrant wound healing responses (MMP1, MMP13), altered growth factor signaling (TGFBR1, GDF15), and dysregulated

inflammatory mediators (CXCL8, IL1B). While aging may contribute to IPF susceptibility, the disease involves distinct pathological processes that diverge significantly from normal age-related changes.

This limited overlap provides critical insights into the relationship between aging and IPF pathogenesis. The 15 overlapping genes represent core processes affected in both conditions, including extracellular matrix remodeling (COL15A1, ASPN), inflammatory signaling (IL1RL1, IL6), and tissue repair mechanisms (PAPPA, ADAMTS1). Some of these genes represent important hubs in the intersection of IPF and aging. ADAMTS1 is an ECM protease with an established connection to cardiovascular disease and oncology [54, 55]. While ipfP3GPT suggests that this gene's expression goes down in IPF and old age, results obtained in rat models of lung fibrosis indicate that its higher expression is associated with alleviated protein deposition [56]. Such proteins require further investigation as the potential anti-IPF, anti-aging dual targets [57]. The combination of concordant and discordant gene expression presents both challenges and opportunities for therapeutic development. For discordantly regulated genes, interventions must carefully consider the potentially opposing effects on aging versus pathological fibrosis. Some of these genes may represent compensatory responses rather than primary drivers of pathology, further complicating therapeutic targeting.

The differential regulation of key fibrotic pathways between IPF and aging generations suggests that IPF may represent dysregulation rather than mere acceleration of normal aging processes. The shared pathway involvement but divergent gene regulation indicates that IPF might arise when normal age-related tissue maintenance mechanisms become pathologically altered. This hypothesis is supported by the opposing regulation of critical ECM components and the partial overlap in inflammatory mediators, suggesting that IPF may hijack normal aging-associated, compensatory tissue remodeling programs, driving them toward a pathological state. Even among genes whose expression, according to ipf-P3GPT, is concordant in aging and IPF, the attention scores assigned by the model vary, indicating a different level of involvement in the accompanying fibrotic processes.

The insights provided by ipf-P3GPT demonstrate the value of creating focused, application-specific AI models. While maintaining core capabilities of the original P3GPT architecture, this streamlined version offers reduced computational overhead, rapid inference times, and domain-specific knowledge concentration. These characteristics make it particularly suitable for exploring disease mechanisms and therapeutic responses in the context of IPF. An important extension to the previously reported P3GPT omics data representation is the addition of a fibrosis-specific tag to all prompts in training, which would allow ipf-P3GPT to be used directly as a tool for assessing the clinical significance of external omics signatures or its own generations. By exploring the gene tokens attending to the fibrotic status, we were able to identify key contributors to the aging and IPF progression.

Future studies incorporating tissue biopsies, rather than blood proteomes, and continuous follow-up would be valuable for validating our findings. As for the AI arm of experiments, a wider set of generations need to be explored, including those representing other fibrotic diseases, to identify more reliable patterns of multi-omic expression.

Our research project opens several promising avenues for investigation. We incentivize other scientists to use the data and models demonstrated here to gain a deeper understanding of the aging and fibrotic processes in their own studies. The particular use cases for the presented models may include indication expansion and the development of novel therapies for age-related diseases.

Beyond IPF, our approach holds potential for investigating other fibrotic conditions such as liver cirrhosis, NAFLD, kidney fibrosis, and systemic sclerosis, where aging-related mechanisms may similarly contribute to pathogenesis. The pathway-aware architecture of our aging clock and the specialized knowledge embedded in ipf-P3GPT could be adapted to these conditions through transfer learning approaches, potentially accelerating biomarker discovery and therapeutic development across the spectrum of fibrotic diseases. Furthermore, these tools enable a more personalized approach to patient stratification and treatment selection by identifying individual variations in aging-associated molecular patterns. This could lead to tailored therapeutic strategies based on a patient's specific aging and fibrotic signatures rather than conventional clinical parameters alone. The utility of the presented approach may also transfer to non-fibrotic diseases by training an array of specialized small-scale P3GPT-like models, each acting as a knowledge source for their respective diseases. Such small-scale AI models can be used to streamline drug development for rare and dangerous conditions and augment existing AI workflows already in use.

While the presented AI-driven approach offers a new tool for IPF and aging research, several limitations should be acknowledged. The most significant limitation

is the lack of experimental validation for the computational findings generated by ipf-P3GPT and our proteomic aging clock. Direct validation through wet-lab experiments, such as targeted gene expression assays or proteomic validation in clinical IPF samples, would strengthen the biological relevance of our findings. Additionally, the COVID-19 dataset represents an indirect validation setting only tangentially related to lung fibrosis. Future studies should incorporate dedicated IPF patient cohorts to directly validate the predictive capacity of our aging clock in this specific condition. The recently concluded phase-2 trial of rentosertib in IPF patients offers a convenient opportunity for further exploration [46]. Furthermore, the ipf-P3GPT model, while enhanced with fibrosis-specific data, is still constrained by the availability and quality of existing public datasets, which may not capture the full heterogeneity of IPF presentations. Finally, our approach is focused primarily on transcriptomic and proteomic data, potentially missing important epigenetic or other factors that contribute to the aging-IPF relationship. Integration of multi-omic data in the following iterations could provide a more comprehensive understanding of these biological processes.

As with any AI application in drug discovery, significant ethical considerations must be addressed. Empirical validation through rigorous experimental testing remains essential, as computational predictions alone are insufficient for making decisions in a process involving human health and well-being [58, 59]. The high-stakes nature of drug discovery for fatal conditions like IPF necessitates maintaining human oversight throughout the development pipeline, ensuring that responsibility for decisions remains with trained experts rather than automated systems. This "human-in-the-loop" approach helps address both accountability concerns and the interpretability challenges of complex AI models. These considerations underscore the importance of developing AI tools as supplements to, rather than replacements for, rigorous scientific investigation when addressing complex age-related diseases. Envisioning clinical utility in the presented clock and ipfP3GPT, it is important to maintain the necessary validation across diverse patient populations to ensure equitable performance and avoid reinforcing existing healthcare disparities [60]. Additionally, the interpretability of complex AI models remains challenging, potentially creating tensions between model performance and clinical explainability. Any future clinical applications would need to be sufficiently transparent for healthcare providers and patients.

The successful application of both our aging clock and ipf-P3GPT demonstrates the growing importance of AI tools in therapeutic development. These approaches not only enhance our understanding of disease mechanisms but also provide frameworks for identifying new therapeutic opportunities [19, 20]. The pathway-aware architecture of our aging clock in particular represents a step toward more biologically informed AI models that could accelerate the development of targeted geroprotective interventions.

CONCLUSION

This study underscores the value of integrating aging biology into therapeutic development for age-related diseases. The combination of targeted therapeutic intervention with AI-driven analysis provides a powerful approach for understanding disease mechanisms and identifying effective treatments. While further research is needed to fully characterize the research utility of ipf-P3GPT and the proteomic clock, our findings suggest they are promising new tools for the study of IPF and potentially other age-related fibrotic conditions.

MATERIALS AND METHODS

Data collection

For aging clock training, we used the UK Biobank collection of 55319 proteomic Olink NPX profiles annotated with age and gender. The validation set containing 84 Olink samples was obtained from [33] and the corresponding Dryad repository (https://doi.org/10.5061/dryad.9cnp5hqmn).

For training ipf-P3GPT, we used a filtered collection of prompts used in training the original P3GPT model that contains 3873 prompts, including 672 prompts with the disease2diff2disease instruction and 3201 with the age_group2diff2age_group instruction (Supplementary File 1). The disease2diff2disease prompts were further enriched to include differentially expressed genes from *in vivo* and *in vitro* datasets available via Gene Expression Omnibus (GEO). The included studies feature a variety of settings involving fibrotic diseases and their models, such as IPF, liver cirrhosis, NASH, alcoholic liver disease, chronic kidney disease, TGF-β cell models, keloid scarring (Supplementary File 2). The full list of the added studies and their differentially expressed features obtained from them are available in the Supplementary Files. Differential gene analysis for the added transcriptomic studies was carried out using Limma [61, 62].

All the prompts were modified to contain only the gene names represented in the Olink 3072 platform [63]. The prompts were filtered to keep only those with >100 significantly differently expressed genes (absolute log2 fold change > 1, q-value<0.05).

All prompts were further extended with XML-like tags denoting EFO identifiers of samples' tissue and its hierarchical ancestors, EFO identifiers of the condition studied in a dataset and its hierarchical ancestors, as well as the EFO identifier of the tissue primarily affected by the condition of interest. Terms from EFO release version 3.73.0 were used for the mapping of GEO-specified metadata to prompt tags. Additionally, we extended each prompt with a binary <is_fibrosis> tag that serves as a direct way to assess whether an observed omics signature is associated with fibrotic changes.

Proteomic aging clock

We developed a multi-task neural network architecture that combines biological age prediction with pathway activity inference (Figure 1). The network consists of three main components: a shared feature extraction backbone, an age prediction branch, and a pathway prediction branch modulated by an attention mechanism.

The shared feature extraction backbone processes 2923 protein measurements through three fully connected layers (2923→512→256→128) with batch normalization and ReLU activation. This shared representation feeds into both the age prediction and pathway branches. The age prediction branch consists of two fully connected layers (128→64→1) with batch normalization and ReLU activation, followed by a softplus activation and scale-shift layer to constrain outputs to a biologically plausible age range.

The pathway branch processes the shared features through a fully connected layer with batch normalization and GELU activation (128→128), followed by a final linear layer with tanh activation (128→4) to predict pathway scores. An attention mechanism, implemented as two fully connected layers (128→128→4) with batch normalization and tanh-softmax activation, modulates the pathway predictions.

The attention mechanism is initialized using weights derived from RNA study evidence described in the previous section. These initial weights are calculated by combining differential expression data across multiple studies, with each study weighted based on tissue relevance, disease type, and comparison setting. The attention weights can be updated during training while maintaining biological plausibility through a KL-divergence regularization term.

The model is trained using a combined loss function that incorporates age prediction error, pathway prediction error, and attention regularization. The age prediction component uses relative error to account for age-dependent uncertainty. The pathway prediction component uses mean squared error between predicted and directly calculated pathway scores. The attention regularization term uses KL-divergence to maintain consistency with the biologically derived initial weights while allowing for data-driven adaptation.

Training is performed using the Adam optimizer with cosine annealing learning rate scheduling and early stopping based on validation loss. Gradient clipping is applied to ensure stable training. The model implementation uses PyTorch and includes batch normalization and dropout (p=0.2) in the feature extraction and age prediction branches to prevent overfitting. At inference, the missing protein values imputed with sample means. See Supplementary File 3 for the package used to train the clock.

ipf-P3GPT training

We trained ipf-P3GPT, a lightweight transformer model, to learn differential gene expression patterns in pulmonary fibrosis. The model architecture comprises a 64-dimensional embedding layer, three transformer blocks with two attention heads each, and a vocabulary-sized output layer (Figure 3 and Supplementary File 4). The model was implemented in PyTorch 2.0.

The transformer blocks consist of multi-head self-attention layers followed by feed-forward networks (FFN). Each FFN expands the 64-dimensional input to 256 dimensions through a linear transformation, applies GELU activation, and projects back to 64 dimensions. Layer normalization is applied after both attention and FFN components. The model processes three instruction types: disease comparisons, age group comparisons, and compound treatment comparisons.

Training utilized a custom XML-aware tokenizer with a vocabulary size of 5,555 tokens. Input sequences were padded or truncated to 512 tokens. We employed mixed-precision training with gradient scaling using an AdamW optimizer (learning rate=1e-4) and batch size of 32. The training data was split 90:10 for training and validation, with weighted random sampling to balance instruction types.

The model was trained for 100 epochs, achieving a final validation loss of 2.03. Instruction-specific accuracies reached 72.2% for age group comparisons and 75.9% for disease comparisons. Model weights were saved at the best validation loss checkpoint.

All hyperparameters and architectural choices were empirically determined through ablation studies, with the final configuration optimizing for both model performance and computational efficiency.

Statistical tests

Treatment effects were analyzed using Mann-Whitney's U tests to compare healthy versus afflicted groups. For the linear effects assessment in the validation COVID-19 dataset, the Ordinary Least Squares module from the statsmodels for Python package.

Pathway analysis

We analyzed four key molecular pathways relevant to IPF pathogenesis: TGF-β signaling, ECM remodeling, inflammation, and oxidative stress. For each patient, pathway scores were calculated as aggregated standardized values of proteins known to participate in respective pathways, with protein-pathway assignments based on curated hallmark sets from MSigDB [64–66].

The biological aging clock used in this analysis was trained using a pathway-aware architecture that incorporated prior knowledge of molecular mechanisms involved in IPF and aging. The selection of the four specific pathways was based on comprehensive literature review identifying them as central to both IPF pathogenesis and aging-related tissue changes. The number of pathways was deliberately limited to four to maintain model parsimony, given the parameter-intensive nature of the attention mechanism and the constraints of the available training dataset size. This focused approach ensured robust training while still capturing the most biologically relevant signals for IPF research.

The pathway attention weights were initialized using a curated database of aging-related pathway signatures and were further refined during model training on the UK Biobank proteomic dataset.

Manuscript preparation

The initial manuscript draft was created using DORA (Draft Outline Research Assistant, https://dora.insilico.com/), an AI-powered scientific writing platform. DORA is Insilico Medicine's LLM-based assistant for automated scientific writing, leveraging an ensemble of over 50 specialized AI agents powered by large language models. These agents work in concert to gather relevant literature, analyze data, and generate high-quality scientific content. All agents are empowered with Retrieval-Augmented Generation (RAG) technology, which enables comprehensive data collection while maintaining scientific accuracy via online fact-checking and PubMed citation linking. After DORA generated the initial draft, all authors collaborated to critically review, expand, and refine the manuscript, ensuring scientific rigor and the accuracy of the presented statement and references.

Data and code availability

All the scripts and materials featured in this study are available as an OSF repository [67]. The XML-formatted prompts are found in Supplementary File 1, while Supplementary Files 3–5 contain the code base for this project. Supplementary File 7 contains an ipfP3GPT usage example in Python Notebook Format. To improve reproducibility, the key dependencies used in this project are provided in Supplementary File 8.

Abbreviations

AI: Artificial intelligence; CV: Cross validation; IPF: Idiopathic pulmonary fibrosis; LLM: Large language model; MAE: Mean absolute error; OSF: Opens Science Framework; P3GPT: Precious3GPT; R^2: Coefficient of determination; USPTO: United States Patent and Trademark Office.

AUTHOR CONTRIBUTIONS

AA — Supervision, Methodology, Writing (review and editing); AZ — Conceptualization, Funding acquisition, Writing (review and editing); FG — Data curation, Investigation, Software, Visualization, Writing (review and editing); FR — Supervision, Methodology, Writing (review and editing); SC —Formal analysis, Investigation, Writing (review and editing).

ACKNOWLEDGMENTS

The authors thank A.Pogorelskaya for her expertise and assistance in preparing ipfP3GPT training data.

CONFLICTS OF INTEREST

All authors are affiliated with Insilico Medicine, a commercial company developing and using generative artificial intelligence and other next-generation AI technologies and robotics for drug discovery, drug development, and aging research. Insilico Medicine has developed a portfolio of multiple therapeutic programs targeting fibrotic diseases, cancer, immunological diseases, and age-related diseases, utilizing its generative AI platform and a range of deep aging clocks and AI life models.

Insilico Medicine is a company developing an AI-based end-to-end integrated pipeline for drug discovery and development and is engaged in aging and IPF research. Insilico Medicine holds USPTO

patents for transcriptomic and proteomic aging clocks (US10665326B2, US10325673B2).

FUNDING

This research received no grants from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

1. Mutsaers HA, Merrild C, Nørregaard R, Plana-Ripoll O. The impact of fibrotic diseases on global mortality from 1990 to 2019. J Transl Med. 2023; 21:818. https://doi.org/10.1186/s12967-023-04690-7 PMID:37974206

2. Lutz W, Sanderson W, Scherbov S. The coming acceleration of global population ageing. Nature. 2008; 451:716–19. https://doi.org/10.1038/nature06516 PMID:18204438

3. Mendoza-Núñez VM, Mendoza-Soto AB. Is Aging a Disease? A Critical Review Within the Framework of Ageism. Cureus. 2024; 16:e54834. https://doi.org/10.7759/cureus.54834 PMID:38405657

4. Mei Q, Liu Z, Zuo H, Yang Z, Qu J. Idiopathic Pulmonary Fibrosis: An Update on Pathogenesis. Front Pharmacol. Frontiers; 2022; 12. https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2021.797292/full

5. Kumar A, Kapnadak SG, Girgis RE, Raghu G. Lung transplantation in idiopathic pulmonary fibrosis. Expert Rev Respir Med. 2018; 12:375–85. https://doi.org/10.1080/17476348.2018.1462704 PMID:29621919

6. Gupta N, Paryani M, Patel S, Bariya A, Srivastava A, Pathak Y, Butani S. Therapeutic Strategies for Idiopathic Pulmonary Fibrosis - Thriving Present and Promising Tomorrow. J Clin Pharmacol. 2024; 64:779–98. https://doi.org/10.1002/jcph.2408 PMID:38346921

7. Galkin F, Mamoshina P, Aliper A, Putin E, Moskalev V, Gladyshev VN, Zhavoronkov A. Human Gut Microbiome Aging Clock Based on Taxonomic Profiling and Deep Learning. iScience. 2020; 23:101199. https://doi.org/10.1016/j.isci.2020.101199 PMID:32534441

8. Mamoshina P, Volosnikova M, Ozerov IV, Putin E, Skibina E, Cortese F, Zhavoronkov A. Machine Learning on Human Muscle Transcriptomic Data for Biomarker Discovery and Tissue-Specific Drug Target Identification. Front Genet. 2018; 9:242. https://doi.org/10.3389/fgene.2018.00242 PMID:30050560

9. Galkin F, Mamoshina P, Kochetov K, Sidorenko D, Zhavoronkov A. DeepMAge: A Methylation Aging Clock Developed with Deep Learning. Aging Dis. 2021; 12:1252–62. https://doi.org/10.14336/AD.2020.1202 PMID:34341706

10. Aliper AM, Galkin F, Zavoronkovs A. Aging markers of human microbiome and microbiomic aging clock (US20200075127A1). 2020. https://patents.google.com/patent/US20200075127A1

11. Galkin F, Kochetov KS, Mamoshina P, Zavoronkovs A. Methylation data signatures of aging and methods of determining a methylation aging clock (US20220005552A1). 2022. https://patents.google.com/patent/US20220005552A1

12. Aliper AM, Putin E, Zavoronkovs A. Deep transcriptomic markers of human biological aging and methods of determining a biological aging clock (US10325673B2). 2019. https://patents.google.com/patent/US10325673B2

13. Aliper AM, Putin E, Zavoronkovs A. Deep proteome markers of human biological aging and methods of determining a biological aging clock (US10665326B2). 2020. https://patents.google.com/patent/US10665326B2

14. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, Terentiev VA, Polykovskiy DA, Kuznetsov MD, Asadulaev A, Volkov Y, Zholus A, Shayakhmetov RR, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol. 2019; 37:1038–40. https://doi.org/10.1038/s41587-019-0224-x PMID:31477924

15. Ren F, Aliper A, Chen J, Zhao H, Rao S, Kuppe C, Ozerov IV, Zhang M, Witte K, Kruse C, Aladinskiy V, Ivanenkov Y, Polykovskiy D, et al. A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models. Nat Biotechnol. 2025; 43:63–75. https://doi.org/10.1038/s41587-024-02143-0 PMID:38459338

16. Remmel HL, Hammer SS, Singh H, Shneyderman A, Veviorskiy A, Alawi KM, Korzinkin M, Zhavoronkov A, Quay SC. Comparative analysis of Endoxifen, Tamoxifen and Fulvestrant: A Bioinformatics Approach to Uncover Mechanisms of Action in Breast Cancer. bioRxiv; 2024. p. 2024.10.02.616224. https://www.biorxiv.org/content/10.1101/2024.10.02.616224v1

17. Galkin F, Pulous FE, Fu Y, Zhang M, Pun FW, Ren F, Zhavoronkov A. Roles of hypoxia-inducible factor-prolyl hydroxylases in aging and disease. Ageing Res Rev. 2024; 102:102551.

https://doi.org/10.1016/j.arr.2024.102551
PMID:39447706

18. Fu Y, Ding X, Zhang M, Feng C, Yan Z, Wang F, Xu J, Lin X, Ding X, Wang L, Fan Y, Li T, Yin Y, et al. Intestinal mucosal barrier repair and immune regulation with an AI-developed gut-restricted PHD inhibitor. Nat Biotechnol. 2024. [Epub ahead of print]. https://doi.org/10.1038/s41587-024-02503-w PMID:39663371

19. Urban A, Sidorenko D, Zagirova D, Kozlova E, Kalashnikov A, Pushkov S, Naumov V, Sarkisova V, Leung GH, Leung HW, Pun FW, Ozerov IV, Aliper A, et al. Precious1GPT: multimodal transformer-based transfer learning for aging clock development and feature importance analysis for aging and age-related disease target discovery. Aging (Albany NY). 2023; 15:4649–66. https://doi.org/10.18632/aging.204788 PMID:37315204

20. Sidorenko D, Pushkov S, Sakip A, Leung GH, Lok SW, Urban A, Zagirova D, Veviorskiy A, Tihonova N, Kalashnikov A, Kozlova E, Naumov V, Pun FW, et al. Precious2GPT: the combination of multiomics pretrained transformer and conditional diffusion for artificial multi-omics multi-species multi-tissue sample generation. NPJ Aging. 2024; 10:37. https://doi.org/10.1038/s41514-024-00163-3 PMID:39117678

21. Galkin F, Naumov V, Pushkov S, Sidorenko D, Urban A, Zagirova D, Alawi KM, Aliper A, Gumerov R, Kalashnikov A, Mukba S, Pogorelskaya A, Ren F, et al. Precious3GPT: Multimodal Multi-Species Multi-Omics Multi-Tissue Transformer for Aging Research and Drug Discovery. bioRxiv; 2024. 2024.07.25.605062. https://www.biorxiv.org/content/10.1101/2024.07.25.605062v1

22. Livne M, Miftahutdinov Z, Tutubalina E, Kuznetsov M, Polykovskiy D, Brundyn A, Jhunjhunwala A, Costa A, Aliper A, Aspuru-Guzik A, Zhavoronkov A. nach0: multimodal natural and chemical languages foundation model. Chem Sci. 2024; 15:8380–89. https://doi.org/10.1039/d4sc00966e PMID:38846388

23. Camillo LP de L, Sehgal R, Armstrong J, Higgins-Chen AT, Horvath S, Wang B. CpGPT: a Foundation Model for DNA Methylation. bioRxiv; 2024. 2024.10.24.619766. https://www.biorxiv.org/content/10.1101/2024.10.24.619766v1

24. Ying K, Song J, Cui H, Zhang Y, Li S, Chen X, Liu H, Eames A, McCartney DL, Marioni RE, Poganik JR, Moqri M, Wang B, et al. MethylGPT: a foundation model for the DNA methylome. bioRxiv; 2024. 2024.10.30.621013. https://www.biorxiv.org/content/10.1101/2024.10.30.621013v1

25. Sehgal R, Markov Y, Qin C, Meer M, Hadley C, Shadyab AH, Casanova R, Manson JE, Bhatti P, Crimmins EM, Hägg S, Assimes TL, Whitsel EA, et al. Systems Age: A single blood methylation test to quantify aging heterogeneity across 11 physiological systems. bioRxiv. 2024 May 28:2023.07.13.548904. https://doi.org/10.1101/2023.07.13.548904 PMID:37503069

26. Nguyen E, Poli M, Durrant MG, Kang B, Katrekar D, Li DB, Bartie LJ, Thomas AW, King SH, Brixi G, Sullivan J, Ng MY, Lewis A, et al. Sequence modeling and design from molecular to genome scale with Evo. Science. 2024; 386:eado9336. https://doi.org/10.1126/science.ado9336 PMID:39541441

27. Denisov KA, Gruber J, Fedichev PO. Discovery of Thermodynamic Control Variables that Independently Regulate Healthspan and Maximum Lifespan. bioRxiv; 2024. 2024.12.01.626230. https://www.biorxiv.org/content/10.1101/2024.12.01.626230v1

28. Kapnadak SG, Raghu G. Lung transplantation for interstitial lung disease. Eur Respir Rev. 2021; 30:210017 https://doi.org/10.1183/16000617.0017-2021 PMID:34348979

29. Collard HR. The age of idiopathic pulmonary fibrosis. Am J Respir Crit Care Med. 2010; 181:771–72. https://doi.org/10.1164/rccm.201001-0049ED PMID:20382799

30. Selman M, Pardo A. Revealing the pathogenic and aging-related mechanisms of the enigmatic idiopathic pulmonary fibrosis. an integral model. Am J Respir Crit Care Med. 2014; 189:1161–72. https://doi.org/10.1164/rccm.201312-2221PP PMID:24641682

31. Murtha LA, Morten M, Schuliga MJ, Mabotuwana NS, Hardy SA, Waters DW, Burgess JK, Ngo DT, Sverdlov AL, Knight DA, Boyle AJ. The Role of Pathological Aging in Cardiac and Pulmonary Fibrosis. Aging Dis. 2019; 10:419–28. https://doi.org/10.14336/AD.2018.0601 PMID:31011486

32. Selman M, López-Otín C, Pardo A. Age-driven developmental drift in the pathogenesis of idiopathic pulmonary fibrosis. Eur Respir J. 2016; 48:538–52. https://doi.org/10.1183/13993003.00398-2016 PMID:27390284

33. Feyaerts D, Hédou J, Gillard J, Chen H, Tsai ES, Peterson LS, Ando K, Manohar M, Do E, Dhondalay GKR, Fitzpatrick J, Artandi M, Chang I, et al. Integrated plasma proteomic and single-cell immune signaling

network signatures demarcate mild, moderate, and severe COVID-19. Cell Rep Med. 2022; 3:100680. https://doi.org/10.1016/j.xcrm.2022.100680 PMID:35839768

34. Argentieri MA, Xiao S, Bennett D, Winchester L, Nevado-Holgado AJ, Ghose U, Albukhari A, Yao P, Mazidi M, Lv J, Millwood I, Fry H, Rodosthenous RS, et al. Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations. Nat Med. 2024; 30:2450–60. https://doi.org/10.1038/s41591-024-03164-7 PMID:39117878

35. Fois AG, Paliogiannis P, Sotgia S, Mangoni AA, Zinellu E, Pirina P, Carru C, Zinellu A. Evaluation of oxidative stress biomarkers in idiopathic pulmonary fibrosis and therapeutic applications: a systematic review. Respir Res. 2018; 19:51. https://doi.org/10.1186/s12931-018-0754-7 PMID:29587761

36. Tomos IP, Tzouvelekis A, Aidinis V, Manali ED, Bouros E, Bouros D, Papiris SA. Extracellular matrix remodeling in idiopathic pulmonary fibrosis. It is the 'bed' that counts and not 'the sleepers'. Expert Rev Respir Med. 2017; 11:299–309. https://doi.org/10.1080/17476348.2017.1300533 PMID:28274188

37. Heukels P, Moor CC, von der Thüsen JH, Wijsenbeek MS, Kool M. Inflammation and immunity in IPF pathogenesis and treatment. Respir Med. 2019; 147:79–91. https://doi.org/10.1016/j.rmed.2018.12.015 PMID:30704705

38. Torres-Machorro AL, García-Vicente Á, Espina-Ordoñez M, Luis-García E, Negreros M, Herrera I, Becerril C, Toscano F, Cisneros J, Maldonado M. Update of Aging Hallmarks in Idiopathic Pulmonary Fibrosis. Cells. 2025; 14:222. https://doi.org/10.3390/cells14030222 PMID:39937013

39. Ren LL, Miao H, Wang YN, Liu F, Li P, Zhao YY. TGF-β as A Master Regulator of Aging-Associated Tissue Fibrosis. Aging Dis. 2023; 14:1633–50. https://doi.org/10.14336/AD.2023.0222 PMID:37196129

40. Ebihara T, Matsumoto H, Matsubara T, Togami Y, Nakao S, Matsuura H, Kojima T, Sugihara F, Okuzaki D, Hirata H, Yamamura H, Ogura H. Cytokine Elevation in Severe COVID-19 From Longitudinal Proteomics Analysis: Comparison With Sepsis. Front Immunol. 2022; 12:798338. https://doi.org/10.3389/fimmu.2021.798338 PMID:35095877

41. Carlyle BC, Kitchen RR, Mattingly Z, Celia AM, Trombetta BA, Das S, Hyman BT, Kivisäkk P, Arnold SE. Technical Performance Evaluation of Olink Proximity Extension Assay for Blood-Based Biomarker Discovery in Longitudinal Studies of Alzheimer's Disease. Front Neurol. 2022; 13:889647. https://doi.org/10.3389/fneur.2022.889647 PMID:35734478

42. Taleb S, Stephan N, Chennakkandathil S, Sohail MU, Yousef S, Sarwath H, Al-Noubi M, Suhre K, Hssain AA, Schmidt F. Olink and NULISAseq Proteomic Technologies Applied to a COVID-19-Induced Acute Respiratory Distress Syndrome (ARDS) Case-Control Study Revealed High Similarity and Complementarity and Shed Light on the Cytokine Storm. Rochester, NY: Social Science Research Network; 2024. https://papers.ssrn.com/abstract=4696495

43. Comparative analysis between Olink-PEA and Alamar-NULISA proteomic technologies applied to a critically ill COVID-19 cohort. figshare; 2024. https://figshare.com/articles/dataset/Comparative_analysis_between_Olink-PEA_and_Alamar-NULISA_proteomic_technologies_applied_to_a_critically_ill_COVID-19_cohort/28044881/1

44. Jia X, Song E, Liu Y, Chen J, Wan P, Hu Y, Ye D, Chakrabarti S, Mahajan H, George J, Yan S, Yu Y, Zhang G, et al. Identification and multicentric validation of soluble CDCP1 as a robust serological biomarker for risk stratification of NASH in obese Chinese. Cell Rep Med. 2023; 4:101257. https://doi.org/10.1016/j.xcrm.2023.101257 PMID:37918406

45. Alrajhi NN. Post-COVID-19 pulmonary fibrosis: An ongoing concern. Ann Thorac Med. 2023; 18:173–181. https://doi.org/10.4103/atm.atm_7_23 PMID:38058790

46. Tang Q, Xiao D, Veviorskiy A, Xin Y, Lok SW, Pulous FE, Zhang P, Zhu Y, Ma Y, Hu X, Gu S, Zong C, Mukba S, et al. AI-Driven Robotics Laboratory Identifies Pharmacological TNIK Inhibition as a Potent Senomorphic Agent. Aging Dis. 2025. [Epub ahead of print]. https://doi.org/10.14336/AD.2024.1492 PMID:39965245

47. Xu Z, Ren F, Wang P, Cao J, Tan C, Ma D, Zhao L, Dai J, Ding Y, Fang H, Li H, Liu H, Luo F, et al. A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: a randomized phase 2a trial. Nat Med. Nature Publishing Group; 2025; 1–9. https://doi.org/10.1038/s41591-025-03743-2

48. Adams TS, Schupp JC, Poli S, Ayaub EA, Neumark N, Ahangari F, Chu SG, Raby BA, Deluliis G, Januszyk M, Duan Q, Arnett HA, Siddiqui A, et al. Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell

populations in idiopathic pulmonary fibrosis. Sci Adv. 2020; 6:eaba1983. https://doi.org/10.1126/sciadv.aba1983 PMID:32832599

49. Nkyimbeng T, Ruppert C, Shiomi T, Dahal B, Lang G, Seeger W, Okada Y, D'Armiento J, Günther A. Pivotal role of matrix metalloproteinase 13 in extracellular matrix turnover in idiopathic pulmonary fibrosis. PLoS One. 2013; 8:e73279. https://doi.org/10.1371/journal.pone.0073279 PMID:24023851

50. Li Y, Zhao J, Yin Y, Li K, Zhang C, Zheng Y. The Role of IL-6 in Fibrotic Diseases: Molecular and Cellular Mechanisms. Int J Biol Sci. 2022; 18:5405–14. https://doi.org/10.7150/ijbs.75876 PMID:36147459

51. Geroldinger-Simić M, Bayati S, Pohjanen E, Sepp N, Nilsson P, Pin E. Autoantibodies against PIP4K2B and AKT3 Are Associated with Skin and Lung Fibrosis in Patients with Systemic Sclerosis. Int J Mol Sci. 2023; 24:5629. https://doi.org/10.3390/ijms24065629 PMID:36982700

52. Ye Z, Hu Y. TGF-β1: Gentlemanly orchestrator in idiopathic pulmonary fibrosis (Review). Int J Mol Med. 2021; 48:132. https://doi.org/10.3892/ijmm.2021.4965 PMID:34013369

53. Koloko Ngassie ML, De Vries M, Borghuis T, Timens W, Sin DD, Nickle D, Joubert P, Horvatovich P, Marko-Varga G, Teske JJ, Vonk JM, Gosens R, Prakash YS, et al. Age-associated differences in the human lung extracellular matrix. Am J Physiol Lung Cell Mol Physiol. 2023; 324:L799–814. https://doi.org/10.1152/ajplung.00334.2022 PMID:37039368

54. Toba H, de Castro Brás LE, Baicu CF, Zile MR, Lindsey ML, Bradshaw AD. Increased ADAMTS1 mediates SPARC-dependent collagen deposition in the aging myocardium. Am J Physiol Endocrinol Metab. 2016; 310:E1027–35. https://doi.org/10.1152/ajpendo.00040.2016 PMID:27143554

55. Chien MH, Yang YC, Ho KH, Ding YF, Chen LH, Chiu WK, Chen JQ, Tung MC, Hsiao M, Lee WJ. Cyclic increase in the ADAMTS1-L1CAM-EGFR axis promotes the EMT and cervical lymph node metastasis of oral squamous cell carcinoma. Cell Death Dis. 2024; 15:82. https://doi.org/10.1038/s41419-024-06452-9 PMID:38263290

56. Liu H, He Y, Jiang Z, Shen S, Mei J, Tang M. Prodigiosin Alleviates Pulmonary Fibrosis Through Inhibiting miRNA-410 and TGF-β1/ADAMTS-1 Signaling Pathway. Cell Physiol Biochem. 2018; 49:501–11. https://doi.org/10.1159/000492989 PMID:30157485

57. Pun FW, Leung GH, Leung HW, Liu BH, Long X, Ozerov IV, Wang J, Ren F, Aliper A, Izumchenko E, Moskalev A, de Magalhães JP, Zhavoronkov A. Hallmarks of aging-based dual-purpose disease and age-associated targets predicted using PandaOmics AI-powered discovery engine. Aging (Albany NY). 2022; 14:2475–506. https://doi.org/10.18632/aging.203960 PMID:35347083

58. Blanco-González A, Cabezón A, Seco-González A, Conde-Torres D, Antelo-Riveiro P, Piñeiro Á, Garcia-Fandino R. The Role of AI in Drug Discovery: Challenges, Opportunities, and Strategies. Pharmaceuticals (Basel). 2023; 16:891. https://doi.org/10.3390/ph16060891 PMID:37375838

59. Hicks MT, Humphries J, Slater J. ChatGPT is bullshit. Ethics Inf Technol. 2024; 26:38.

60. DISPARITIES IN HEALTHCARE. 2021 National Healthcare Quality and Disparities Report. Agency for Healthcare Research and Quality (US); 2021. https://www.ncbi.nlm.nih.gov/books/NBK578532/

61. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43:e47. https://doi.org/10.1093/nar/gkv007 PMID:25605792

62. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014; 15:R29. https://doi.org/10.1186/gb-2014-15-2-r29 PMID:24485249

63. Galkin F. Precious-3 GPT. OSF; 2024. https://osf.io/qrt3u/

64. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, Houstis N, Daly MJ, Patterson N, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003; 34:267–73. https://doi.org/10.1038/ng1180 PMID:12808457

65. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005; 102:15545–50. https://doi.org/10.1073/pnas.0506580102 PMID:16199517

66. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures

Database (MSigDB) hallmark gene set collection. Cell Syst. 2015; 1:417–25.
https://doi.org/10.1016/j.cels.2015.12.004
PMID:26771021

67. Galkin F. ipfP3GPT — Supplementary Files. US: Open Science Framework; 2025. https://osf.io/457w8/

SUPPLEMENTARY MATERIALS

Supplementary Files

Please browse Full Text version to see the data of Supplementary Files 1–8.

Supplementary File 1. Prompts used in ipfP3GPT training. The prompts are XML-formatted lists describing experimental metadata with associated lists of differentially expressed genes.

Supplementary File 2. Fibrosis-enriched study set. This file contains metadata for 161 fibrosis-related GEO datasets, and differential expression results for the most representative datasets used to train ipfP3GPT.

Supplementary File 3. Package used to train proteomic clock. The Python codebase used to train and use the presented proteomic clock.

Supplementary File 4. ipfP3GPTpython package. The Python codebase used to train and use ipfP3GPT.

Supplementary File 5. Code used to analyze the proteomic clock. Demonstration Python Notebook with the proteomic clock inference examples.

Supplementary File 6. Age predictions for the Olink data generated in (Feyaerts et al., 2022)

Supplementary File 7. Code used to invoke ipfP3GPT. This demonstration Python Notebook allows the reader to reproduce key findings from the study.

Supplementary File 8. Requirements.