RESEARCH ARTICLE

National Science Review
12: nwaf189, 2025
https://doi.org/10.1093/nsr/nwaf189
Advance access publication 14 May 2025

INFORMATION SCIENCE

sPGGM: a sample-perturbed Gaussian graphical model for identifying pre-disease stages and signaling molecules of disease progression

Jiayuan Zhong[1,†], Junxian Li[2,†], Xuerong Gu[3,†], Dandan Ding[4], Fei Ling[3], Pei Chen[2,*] and Rui Liu[2,*]

ABSTRACT

Complex disease progression typically involves sudden and non-linear transitions accompanied by devastating effects. Uncovering such critical states or pre-disease stages and discovering dynamic network biomarkers (signaling molecules) is vital for both comprehending disease progression and preventing or delaying disease deterioration. However, the detection of critical points using high-dimensional limited sample data or single-cell data proves notably challenging, as traditional statistical approaches often fail to deliver accurate results. In this study, based on optimal transport theory and Gaussian graphical models, we present an innovative computational framework, the sample-perturbed Gaussian graphical model (sPGGM), designed to analyze disease progression and identify pre-disease stages at the specific sample/cell level. Specifically, by employing population-level optimal transport and Gaussian graphical models, the proposed sPGGM effectively characterizes dynamic differences between the baseline distribution and the perturbed distribution relative to the specific case sample, thus enabling the identification of pre-disease stages and the discovery of signaling molecules during disease progression. The reliability and effectiveness of our method is demonstrated by conducting a simulated dataset and evaluating various data types, including four single-cell datasets, influenza infection data, and six distinct bulk tumour datasets. In comparison with existing single-sample methods, our proposed method exhibits improved capability in pinpointing critical point or pre-disease stages. Moreover, the effectiveness of computational results is highlighted through the analysis of the functional roles of signaling molecules.

Keywords: critical point, optimal transport, dynamic network biomarker (DNB), pre-disease stage, sample-perturbed Gaussian graphical model (sPGGM)

[1]School of Mathematics, Foshan University, Foshan 528000, China; [2]School of Mathematics, South China University of Technology, Guangzhou 510640, China; [3]School of Biology and Biological Engineering, South China University of Technology, Guangzhou 510640, China and [4]Department of Oncology, First People's Hospital of Foshan, Foshan 528000, China

*Corresponding authors. E-mails: chenpei@scut.edu.cn; scliurui@scut.edu.cn

†Equally contributed to this work.

Received 15 December 2024; **Revised** 16 April 2025; **Accepted** 30 April 2025

INTRODUCTION

Disease progression is inherently dynamic and prone to dramatic shifts over time, often triggered by subtle internal or external disturbances, leading to irreversible and severe consequences. Such a process marked by abrupt critical shifts can typically classified into three phases [1,2]: the normal stage, pre-disease stage and disease stage (Fig. 1a). The normal stage reflects a relatively healthy condition in disease progression, where the system maintains normality and exhibits high stability. The pre-disease stage marks the critical threshold preceding the appearance of disease symptoms [3,4]. That is, as this critical point approaches, the patient often experiences a catastrophic and irreversible transition, commonly resulting in deterioration. In contrast to the reversible normal stage, the irreversible deterioration of disease stage poses a serious threat to the life and health of patients. Therefore, grasping the dynamics of disease progression and unveiling the pre-disease stage plays a key role in facilitating early disease intervention and treatment [5,6]. However, the accurate detection of the pre-deterioration stage or critical point for complex diseases presents a considerable difficulty. There show only minor changes in gene expression patterns and clinical phenotypes between the normal stage and the pre-disease stage. Additionally,

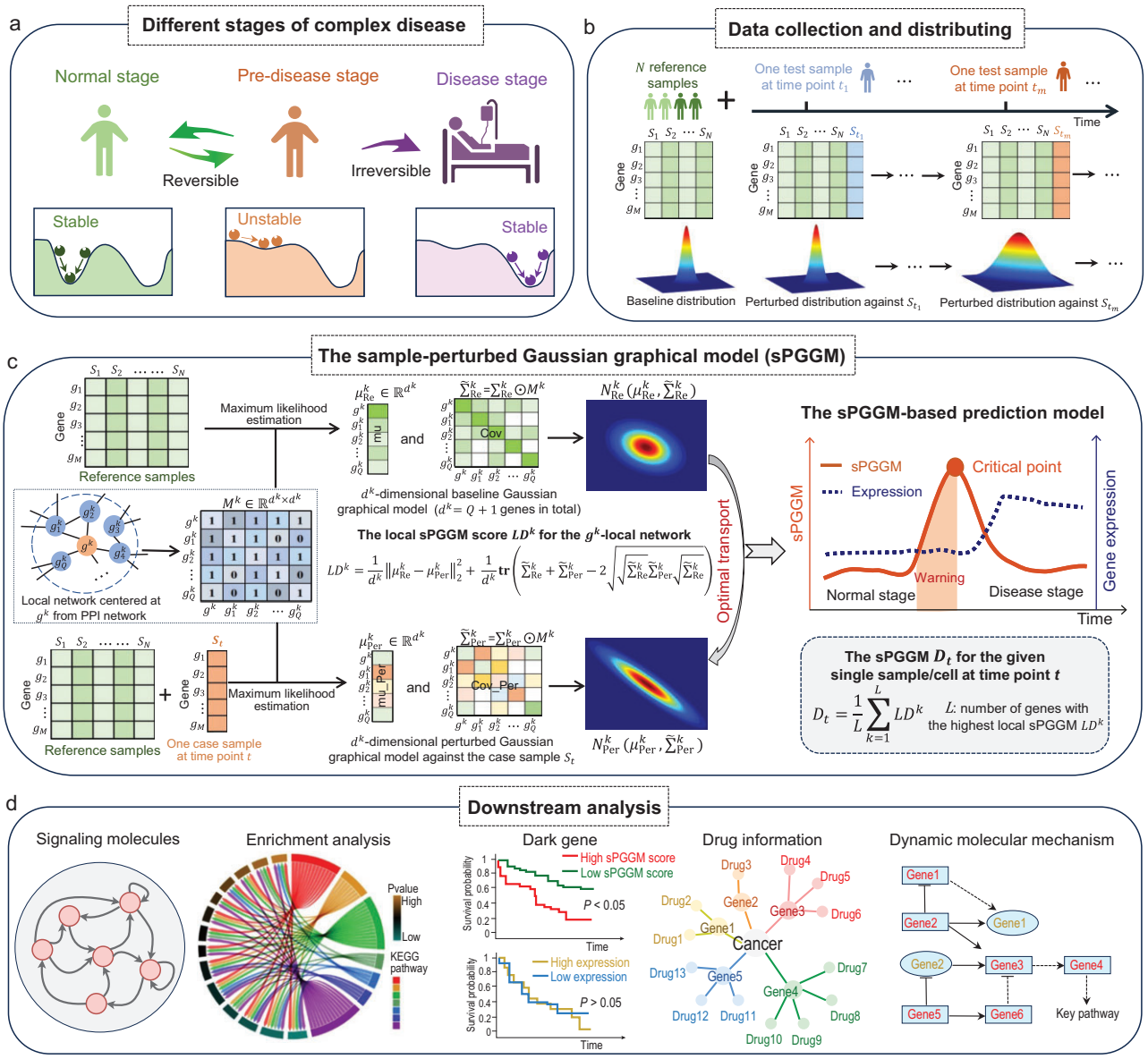

Figure 1. Schematic illustration of sample-perturbed Gaussian graphical model (sPGGM) for identifying pre-disease stages. (a) Disease progression can be classified into three states: the normal stage, pre-disease stage and disease stage, with the pre-disease stage representing a critical threshold just before the onset of disease symptoms. (b) The baseline distribution is fitted from reference samples, whereas the perturbed distribution is derived from mixed samples that combine a specific case sample with the reference group. (c) The proposed sPGGM constructs candidate detection stages at the single-sample level by utilizing a Gaussian graphical model embedded with prior knowledge of the PPI network and quantifies the distributional changes between the baseline and perturbed distributions through the application of optimal transport theory. Then the sPGGM score is used to measure the critical transitions of complex diseases, with a marked increase signaling the pre-disease stage. (d) In downstream analysis, we validate the results by identifying signaling molecules, performing functional analyses, investigating potential molecular regulatory mechanisms, and so on.

challenges such as data noise, patient heterogeneity, limited sample sizes, and model inaccuracies hinder the reliable identification of critical transitions.

The identification of critical transitions in disease progression have increasingly gained attention in recent studies. The computational approach, grounded in flux theory of non-equilibrium dynamical systems, has been devised to estimate various properties of state transitions in the system [7]. Despite its robust theoretical foundation, the high computational complexity of this method makes it challenging to use in large-scale, high-dimensional biological systems [8]. Recently, a new concept of dynamic network biomarkers (DNBs) has been introduced to pinpoint key transitions in complex biological systems [9,10]. Unlike conventional biomarkers that solely assess static molecular activity levels, DNBs can uncover the critical points and potential molecular mechanisms of biological processes. The application of the DNB theoretical

framework has shown effectiveness in analyzing critical states of complex diseases like diabetes, cancer and Alzheimer's disease [11–15]. However, existing DNB methods predominantly rely on multiple samples to estimate statistical conditions [9], which constrains their application in biological research due to the challenge of collecting multi-sample data from each time point in practical scenarios. In addition, while these methods mainly aim to detect early critical signals through traditional bulk omics analysis, they still face robustness problem in noisy and heterogeneous single-cell data. In summary, previous methods are limited in their ability to address specific issues in complex diseases and cannot effectively handle challenges such as small sample sizes, highly noisy data, and sample heterogeneity. Therefore, our aim is to design a novel single-sample approach that effectively overcomes these limitations, allowing for the identification of pre-disease stages and the prediction of the important molecules driving disease progression.

Single-cell data has been proven to provide unprecedented insights into the dynamic processes of cellular systems [16,17]. In recent years, there has been growing interest in characterizing transitions at the single-cell level. For instance, methods like MuTrans and QuanTC have been proposed to precisely dissect transition cells from single-cell data [18,19]. BioTIP has been developed to detect critical transition signals from single-cell transcriptomes [20]. Meanwhile, quantifying the critical properties of complex biological systems from a distributional perspective is gaining increasing attention. The multivariate distribution method has been used to detect critical states during complex biological processes [21]. The Gaussian distribution-based model has been proposed to accurately identify critical transitions in disease progression [22]. Additionally, the Kullback–Leibler divergence index has been employed to pinpoint the critical state of cancer by capturing dynamic distributional changes [23]. Such integration of gene expression profiles with distribution-based approaches is vital for more effectively characterizing the dynamic changes within biological systems. In this research, inspired by pioneer works, we propose a new and generalized method called sample-perturbed Gaussian graphical model (sPGGM) based on optimal transport theory and Gaussian graphical models, to identify the critical point or pre-disease stage and discover signaling molecules during disease progression from a sample-specific perspective. Specifically, to reduce irrelevant variables and improve actual biomolecular associations, our proposed sPGGM constructs candidate stages of detection at a single-sample level using a Gaussian graphical model embedded with prior knowledge of the protein-protein interaction network. Then, sPGGM captures the distributional changes between the baseline distribution (fitted from reference samples) and the perturbed distribution (fitted from mixed samples that combine a specific case sample with reference group) through optimal transport [24], and utilize the Wasserstein distance to quantify the relative differences between various detection stages (Fig. 1b and c). The critical properties of complex disease can be unveiled by sPGGM, with significant increases serving as a critical signal for disease prediction due to its sensitivity to distribution shifts and ability to measure the minimal 'effort' required to transition from the normal stage to the pre-disease stage. To demonstrate the robustness and effectiveness of sPGGM, we applied it to both simulated data and various real-world disease datasets, including an influenza dataset, two single-cell datasets, and six cancer datasets from the TCGA database: colon adenocarcinoma (COAD), thyroid carcinoma (THCA), kidney clear cell carcinoma (KIRC), uterine corpus endometrial carcinoma (UCEC), kidney renal papillary cell carcinoma (KIRP), and liver hepatocellular carcinoma (LIHC). The results indicate that the proposed sPGGM effectively handles real-world disease data, accurately detects pre-disease stages across various disease categories, and identifies signaling molecules at critical points. Moreover, it exhibits a better performance in capturing critical signals of complex diseases compared to other existing single-sample detection approaches [25–28]. In addition, we conducted functional analysis on the signaling molecules identified by sPGGM, uncovering potential molecular regulatory mechanisms in disease progression and understanding the biochemical basis of disease (Fig. 1d). In brief, our sPGGM provides a new single-sample way to identify the pre-disease state and discover signaling molecules leading to potential disease, which showcases exceptional effectiveness and robustness for both bulk and single-cell data analyses, offering a novel perspective for personalized disease prediction.

RESULTS

Performance of the sPGGM based on numerical simulation

To assess the validity of our proposed sPGGM, an 18-node modulated network is used to demonstrate how the algorithm captures critical signals or tipping points (Fig. 2a). Such a modulated network is represented by a framework of stochastic differential equations based on Michaelis-Menten or Hill

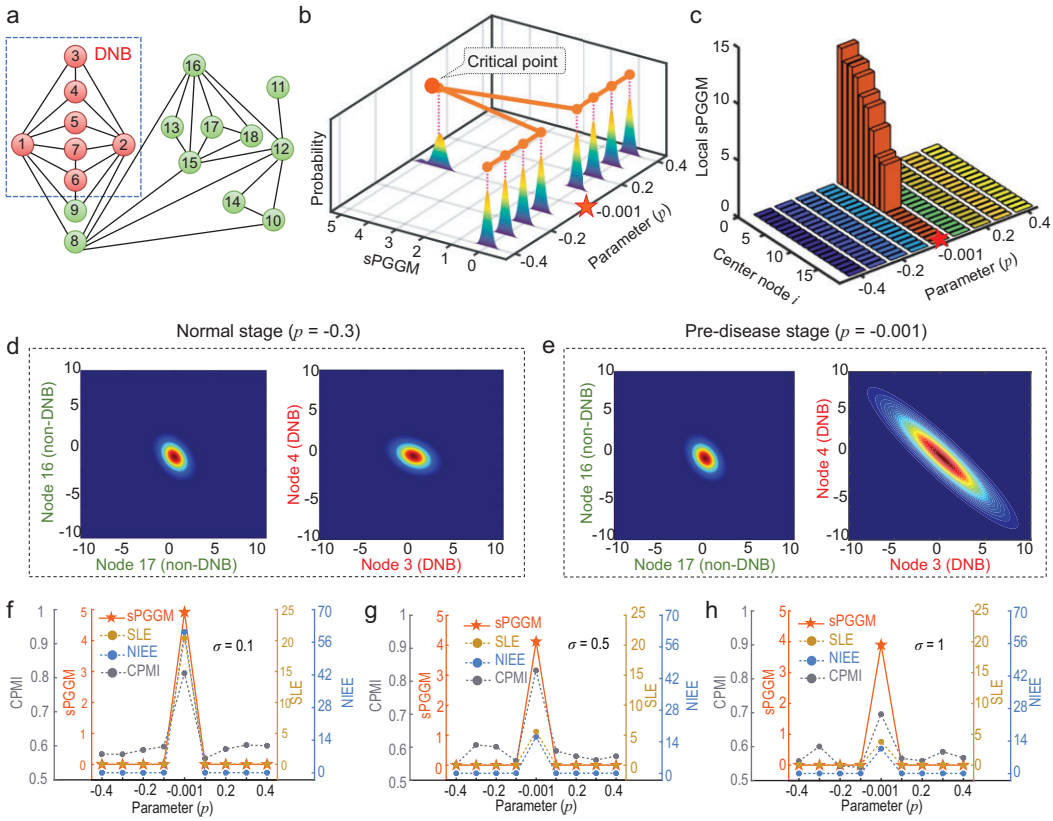

Figure 2. Performance assessment of the sPGGM based on numerical simulation. (a) The numerical simulation is performed using an 18-node graph, constructed from a gene modulated network to depict the relationships between the nodes. (b) The sPGGM curve reveals that a sharp increase in the sPGGM score signals the impending critical transition. (c) The sPGGM landscape shows that the scores for specific local networks with signaling molecules or DNBs exhibit a sharp increase as the system approaches the tipping point. (d, e) The comparison of the multivariate Gaussian distributions for nodes 3 and 4 (DNBs) versus nodes 16 and 17 (non-DNBs) is analyzed between the normal state ($p = -0.3$) and the critical point ($p = -0.001$). (f–h) A comparison of the resilience performance between sPGGM and previous single-sample methods is provided.

dynamics, which is commonly used to analyze gene regulation in biological processes, including transcription, translation, and complex nonlinear interactions [29,30]. The network's critical signal is governed by the equation parameter p, where $p = 0$ indicates the bifurcation marking the critical point (refer to Section A of the Supplementary Information for further details). Moreover nodes 1 to 7, functioning as signaling molecules or DNBs, are directly influenced by the parameter p, while the other nodes remain unaffected by p and act as irrelevant molecules. Simulated data is generated by adjusting the parameter p from -0.4 to 0.4, illustrating how effectively the sPGGM uncovers the critical transition near the bifurcation.

It can be seen from Fig. 2b that a notable rise in the sPGGM score signals the impending critical state as the system nears the bifurcation point ($p = 0$), while the score remains stable and low when the system is distant from the tipping point. Moreover, to reveal the specific dynamics of each node and pinpoint signaling molecules throughout the progression, we show the evolution of local sPGGM landscapes across different nodes in Fig. 2c. As the system is away from the tipping point, the sPGGM score for all local networks remains uniformly low, but a notable spike in the sPGGM score occurs in specific local networks containing DNBs when nearing the critical point. Additionally, Fig. 2d and e illustrates the transport of distributions from normal states to the critical point. As the system approaches the tipping point, the multivariate Gaussian distribution of signaling molecules becomes more divergent and fluctuates significantly, indicating a substantial rise in their variance. To demonstrate the resilience of sPGGM, we conducted a comparative analysis of sPGGM and other existing single-sample methods on samples subjected to varying levels of noise perturbation (Fig. 2f–h), highlighting our proposed method's superior sensitivity

and clarity in detecting critical signals. As the noise level increases, our sPGGM method demonstrates enhanced robustness and efficacy when subjected to high noise levels (Fig. 2f–h and Fig. S1). The simulation results show that the sPGGM effectively and accurately detects critical transitions. Besides, our proposed sPGGM can pinpoint signaling molecules and shed light on the key changes during system progression.

Identifying pre-disease stages for individual influenza infection

In this research, we utilized the sPGGM to analyze the time-series dataset related to influenza infection. This dataset consists of samples from 17 volunteers infected with the Wisconsin/H3N2 virus via intranasal administration, with gene expression data collected at 16 time points over a 132-hour span (−24 to 108 hours) (Fig. 3a). Among them, 9 volunteers (subjects 1, 5, 6, 7, 8, 10, 12, 13, and 15) with severe influenza-like symptoms were classified as the symptomatic group, while the remaining 8 volunteers showing no clinical symptoms were categorized as the asymptomatic group. For each participant, the gene expression data from the first four time points were considered as a reference group, indicating their relatively healthy state. The sample-specific sPGGM score (denoted as D_t in Eq. (S35)) was calculated for each of 17 participants by the algorithm detailed in the Materials and Methods section. A rapid increase in the sPGGM score acts as an early indicator of disease onset, particularly signaling the moment when clinical symptoms begin to manifest. Fig. 3b depicts the sPGGM score for all participants across each time point. The symptomatic group (indicated by red curves) show a marked increase in the sPGGM score prior to the appearance of symptoms, providing early signs of imminent critical transitions. In contrast, the asymptomatic group (shown by blue curves) exhibit consistent sPGGM scores with no significant changes. Moreover, Fig. 3c depicts the sPGGM score tailored for each of the nine symptomatic participants, highlighting the pre-disease stage that precedes the emergence of clinical symptoms in each case. Consequently, the identification of the pre-disease stages for each participant validates the effectiveness of our sPGGM from a sample-specific viewpoint.

Identifying pre-disease stages for cancer progression

To determine how well the proposed sPGGM unveils pre-disease stages of cancer progression, we applied this method to six tumour datasets (COAD, THCA, KIRC, UCEC, KIRP, and LIHC) sourced from the TCGA database. Using adjacent non-tumour samples as the reference group, we determined the sample-specific sPGGM score (as outlined in Eq. (S35)) for each individual case. The mean sPGGM score at each stage was applied as a quantitative indicator to evaluate the pre-disease state. The analytical findings revealed that the pre-disease stage was identified as stage II for THCA, KIRC, and LIHC, stage III for KIRP, and stage IIB for COAD and UCEC (Fig. 4a–f). In the COAD dataset, a significant shift ($P = 7.2\text{E}-9$) in the sPGGM score was observed around stage IIB (Fig. 4a), signaling the onset of lymph node metastasis at stages IIIA–IIIB [31]. For the THCA dataset, it is seen from Fig. 4b that the sPGGM score reaches its highest point at stage II ($P = 5.4\text{E}-4$), indicating an approaching critical shift. The literature reveals that stage III involves the sternothyroid muscle or nearby thyroid-related soft tissues, along with metastasis to regional lymph nodes [32]. When applied to the KIRC dataset, as illustrated in Fig. 4c, a notable increase ($P = 5.4\text{E}-62$) in the sPGGM score from stages I to II points to a critical deterioration event, that is, stage III is characterized by a rapid escalation of lipid levels around the kidney and tumour invasion into the renal vein [33]. In the UCEC dataset, there is a substantial increase ($P = 5.5\text{E}-10$) in the sPGGM score between stages IIA and IIB (Fig. 4d), implying the occurrence of tumour extension into surrounding tissues or metastasis to lymph nodes after stage IIB [34]. For the LIHC dataset, the sPGGM score shows a sudden rise ($P = 3.9\text{E}-2$) during stage II (Fig. 4e), after which direct invasion of nearby organs appears [35]. When applied to the KIRP dataset, the sPGGM score increased drastically before stage III ($P = 2.1\text{E}-7$), signaling that distant metastasis generally occurs at stage IV (Fig. 4f) [36]. Conversely, as depicted by the dark blue curve in Fig. 4a–f, the expression levels of highly expressed genes from specific samples does not effectively indicate a critical transition from the perspective of both accuracy and signal significance. Furthermore, compared to four other existing single-sample methods [25–28] (see Table 1 and Figs S2–S4), our proposed sPGGM demonstrates improved performance in identifying pre-disease stages throughout the progression of the disease.

To verify the determined pre-disease state, we utilized the Kaplan–Meier (log-rank) method to conduct a prognostic survival analysis on clinical samples taken from before and after the critical point. It can be observed from Fig. 4g–l that there shows a significant difference in prognosis between patients diagnosed before and after the critical stage, with

a

Subject		Symptom															
Symptomatic	S1	N	N	N	N	N	N	N	S	S	S	S	S	S	S	S	S
	S5	N	N	N	N	N	N	N	S	S	S	S	S	S	S	S	S
	S6	N	N	N	N	N	N	N	S	S	S	S	S	S	S	S	S
	S7	N	N	N	N	N	N	N	S	S	S	S	S	S	S	S	S
	S8	N	N	N	N	N	N	N	N	N	N	S	S	S	S	S	S
	S10	N	N	N	N	N	N	N	N	N	N	N	S	S	S	S	S
	S12	N	N	N	N	N	N	N	N	N	N	N	N	N	S	S	S
	S13	N	N	N	N	N	N	N	N	N	N	N	N	N	N	S	S
	S15	N	N	N	N	N	N	N	N	N	N	N	N	N	N	S	S
Asymptomatic	S2	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	S3	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	S4	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	S9	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	S11	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	S14	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	S16	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	S17	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	Hour	-24	0	5	12	21	29	36	45	53	60	69	77	84	93	101	108

N Non-symptom N or S Early warning signal S Symptom

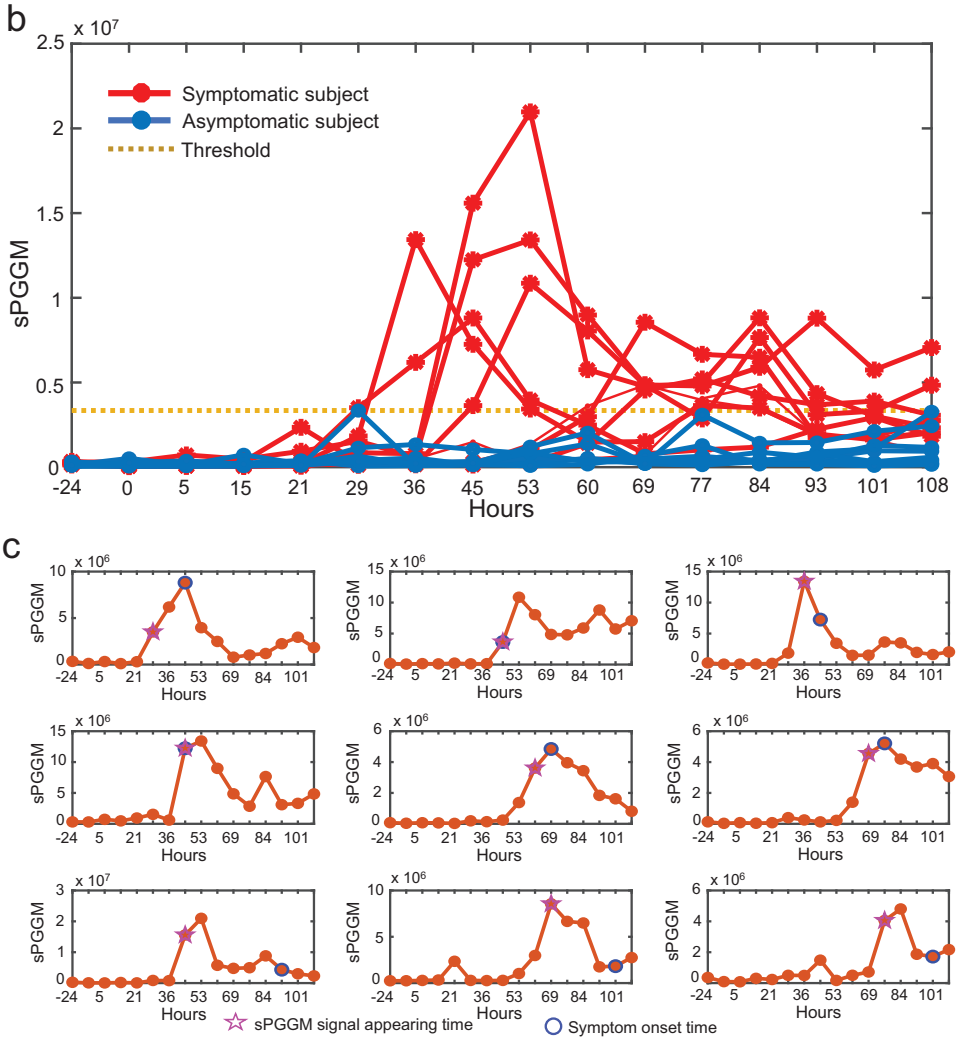

Figure 3. Identification of pre-disease stages for influenza infection based on sPGGM. (a) A temporal chart detailing the onset of influenza symptoms and the pre-disease stages determined by sPGGM for all participants. (b) The sPGGM curves for all 17 subjects are shown, with the red curve representing the nine symptomatic participants and the blue curve depicting the eight asymptomatic participants. (c) The curves for sample-specific sPGGM score of nine symptomatic individuals are displayed, where the blue circle indicates the onset of influenza symptoms (as clinically observed), and the pink box marks critical signals identified by sPGGM.

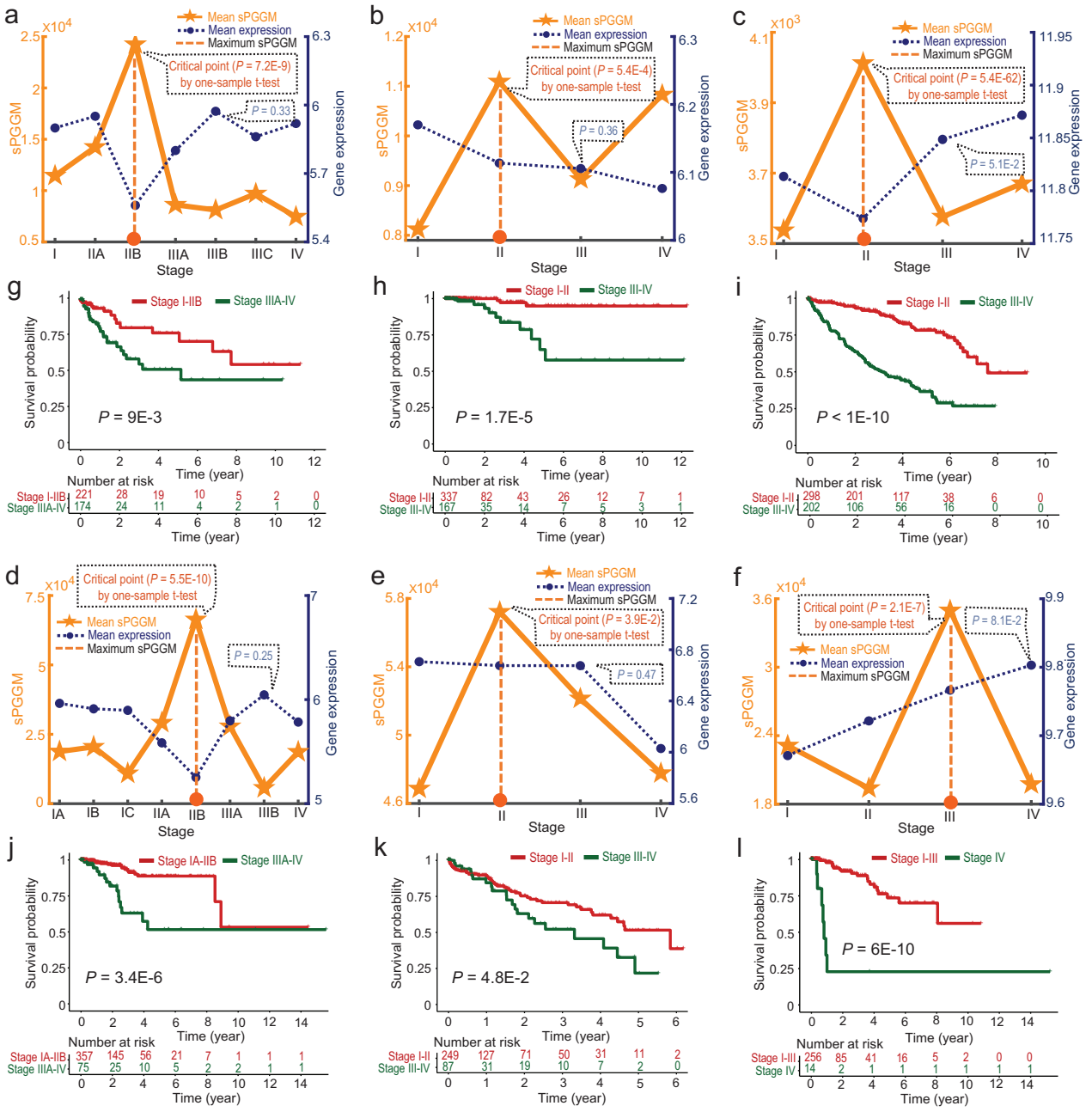

Figure 4. Identification of pre-disease stages for cancer progression based on sPGGM. The dynamic behavior of sPGGM and gene expression was evaluated across six tumour types: (a) COAD, (b) THCA, (c) KIRC, (d) UCEC, (e) LIHC, and (f) KIRP. The significant rise in the sPGGM score indicates an upcoming critical transition before disease deterioration. Survival durations before and after the critical stage was analyzed for the following tumour types: (g) COAD, (h) THCA, (i) KIRC, (j) UCEC, (k) LIHC, and (l) KIRP. Patients experience significantly longer survival times before reaching the critical point compared to after it.

p values below 0.05, indicating that those treated before critical transition have higher survival rates and longer survival times. Therefore, the sPGGM score can effectively signal the pre-disease states related to survival time before disease deterioration, which facilitates prompt medical intervention and follow-up care.

Functional analysis of the signaling molecules involved in cancer progression

In addition to detecting the early pre-disease stages of tumour progression, we also conduct a functional analysis of signaling molecules (the top 5% of genes exhibiting the highest sPGGM score) to gain insights into their role in disease development. The

Table 1. Comparison of the performance among different single-sample detection methods

Dataset	sPGGM	L-DNB	SLE	CPMI	NIEE
COAD	Stage IIB ($P = 7.2E-9$)	Stage IIIC ($P = 9.6E-11$)	Stage IIB ($P = 7.7E-9$)	Stage IIB ($P = 4.5E-5$)	Stage IIIC ($P = 1.2E-50$)
THCA	Stage II ($P = 5.4E-4$)	Stage III ($P = 1.4E-2$)	Stage III ($P = 1.5E-2$)	Stage III ($P = 2.1E-4$)	None
KIRC	Stage II ($P = 5.4E-62$)	Stage II ($P = 3.2E-2$)	None	Stage III ($P = 4.2E-2$)	Stage IV ($P = 1.1E-13$)
UCEC	Stage IIB ($P = 5.5E-10$)	Stage IIB ($P = 6.5E-257$)	Stage IIB ($P = 6.4E-4$)	Stage IIB ($P = 6.8E-84$)	Stage IIIA ($P = 3.2E-6$)
LIHC	Stage II ($P = 3.9E-2$)	None	None	Stage IV ($P = 2.6E-2$)	Stage III ($P = 3.7E-57$)
KIRP	Stage III ($P = 2.1E-7$)	Stage III ($P = 1.9E-11$)	Stage III ($P = 1.2E-308$)	Stage II ($P = 4.6E-2$)	Stage III ($P = 3.1E-3$)
T-cell exhaustion	Stage 5 ($P = 1.2E-129$)	Stage 7 ($P = 1.7E-2$)	Stage 7 ($P = 2.4E-2$)	Stage 5 ($P = 1.1E-2$)	None
GABAergic interneurons	D54 ($P = 1.8E-2$)	None	None	None	D125 ($P = 4.7E-4$)

None: represents the inability to detect the critical signal.

transport map derived from the normal to abnormal state via sPGGM enables us to describe the disease progression using Gaussian graphical distributions. In this study, we utilize principal component analysis (PCA) [37] to demonstrate the main transformation processes in the distribution of signaling molecules at different stages of the disease. As illustrated in Fig. 5a–c, the distribution transport process for three tumour datasets (LIHC, COAD and UCEC) reveals that the distribution becomes more concentrated as it moves away from critical points and gradually disperses as it nears them. The temporal evolution of the distribution transport process across all stages is presented in Fig. S5. This indicates that the sPGGM effectively captures these distribution state changes and identifies critical transitions across various diseases. Furthermore, functional enrichment analysis of the identified signaling molecules shows a significant enrichment in cancer-related pathways, including oxidative phosphorylation [38], chemical carcinogenesis—reactive oxygen species [39], and PI3K-Akt signaling pathway [40] (Fig. 5d–f). An additional functional analysis of both 'dark molecule' (non-differential genes sensitive to the sPGGM score) and differentially expressed genes (DEGs) among signaling molecules is given in Fig. S6. Additionally, we also discovered that certain 'dark molecules' involved in cancer-related pathways are essential for disease progression and serve as effective prognostic indicators, not at the gene expression level but at the sPGGM score level (Fig. 5g–i). Therefore, our sPGGM-based method can be viewed as a valuable complement to traditional differential expression analysis, helping to identify new biomarkers, drug targets, and prognostic indicators from a network-level perspective (since the sPGGM score is derived from network-based computations) rather than focusing solely on the gene level. Moreover, to enhance the validation of 'dark molecules' as prognostic indicators, Figs S7 and S8 present a comparison of survival analysis among the 'dark molecule', the top 5% most significant DEGs, and randomly selected genes of the same size. Besides, it can be seen from Fig. 5j–l that key signaling molecules targeted by specific cancer-related drugs can be identified based on the iGMDR database [41], which provides drug targets and effective drugs for early therapeutic intervention in patients with specific cancer.

Identifying pre-disease stages for complex diseases at single-cell level

To gain deeper insights into the pre-disease stages of complex diseases at the single-cell resolution, we applied the proposed sPGGM to two disease-associated single-cell data: CD8+ T-cell exhaustion dataset and GABAergic interneurons dataset. For the CD8+ T-cell exhaustion dataset, the brown-yellow curve in Fig. 6a indicates a marked increase in the sPGGM score at stage 5 ($P = 1.2E-129$), signaling an early warning of the impending critical transition for the cell subpopulation which began to exhibit exhaustion characteristics thereafter [42]. When applied to the GABAergic interneurons dataset, it is seen from Fig. 6b that the sPGGM score demonstrates a significant rise from day 24 to day 54 ($P = 1.8E-2$), after which there is a transition from neurogenesis to gliogenesis, with certain genes involved in astrocyte function [43]. In contrast, the

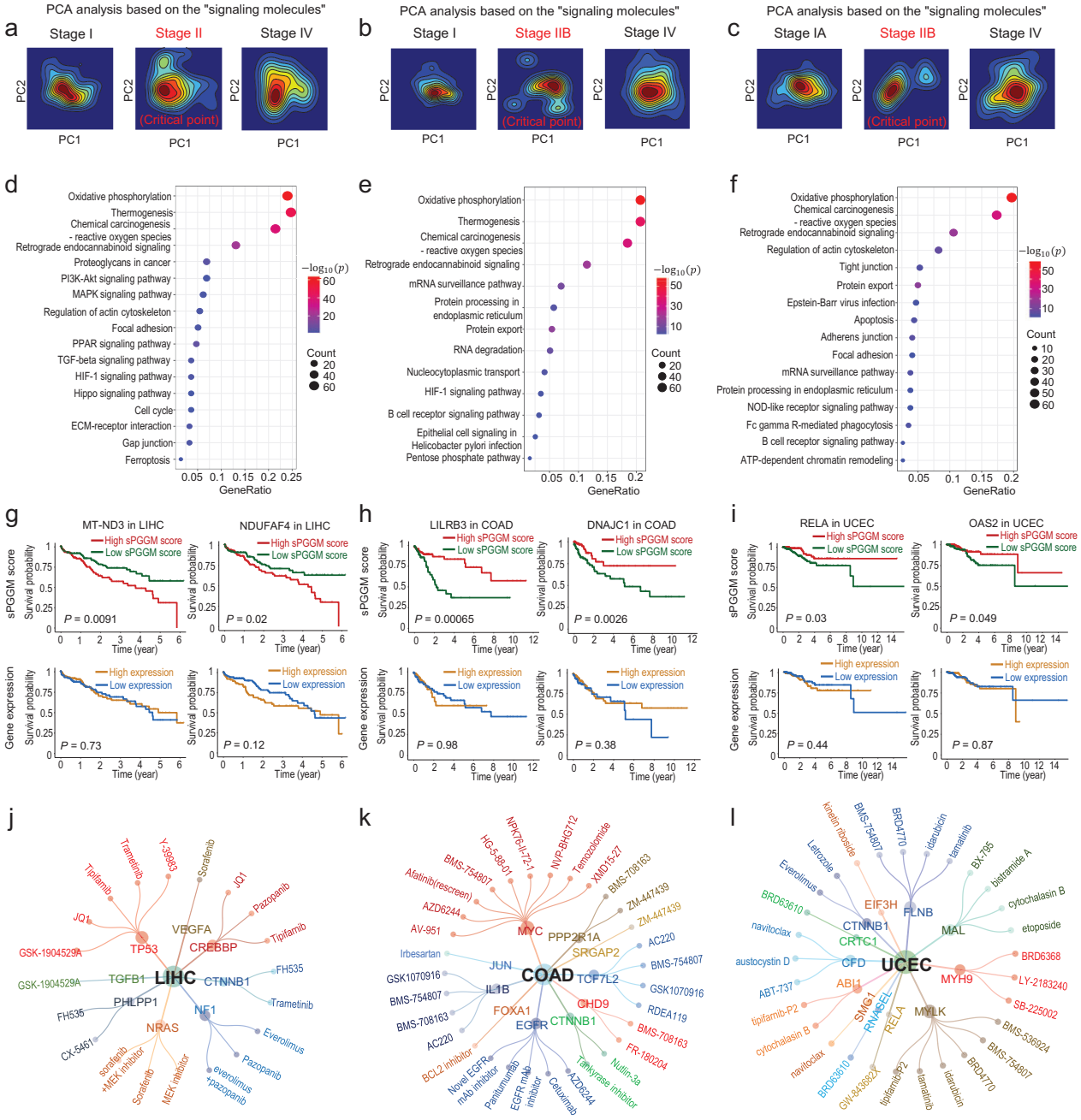

Figure 5. Functional analysis of the signaling molecules implicated in cancer development. PCA-based visualizations of the optimal transport map transitioning from the normal to abnormal states for three tumour types: (a) LIHC, (b) COAD, and (c) UCEC. The KEGG pathway enrichment analysis of signaling molecules in three tumour types: (d) LIHC, (e) COAD, and (f) UCEC. The results demonstrate that signaling molecules are mainly enriched in cancer-associated pathways. The survival analysis of 'dark molecule' (non-differential genes sensitive to the sPGGM score) for three tumour types: (g) LIHC, (h) COAD, and (i) UCEC. This survival analysis, based on the local sPGGM score rather than gene expression values, proves effective in prognosis and successfully distinguishes significant differences in survival times. The key signaling molecules targeted by specific cancer-related drugs identified for three tumour types: (j) LIHC, (k) COAD, and (l) UCEC.

dark blue curve shown in Fig. 6a and b reveals that the expression levels of highly expressed genes do not adequately signify a critical transition in terms of accuracy and signal significance. As illustrated in Fig. S9A, for the T-cell exhaustion dataset, sPGGM, MuTrans, and BioTIP detects the critical transition with significant increases, with the sPGGM providing an earlier warning signal. When applied to the GABAergic interneurons dataset, it is seen from Fig. S9B that sPGGM, MuTrans, QuanTC, and BioTIP

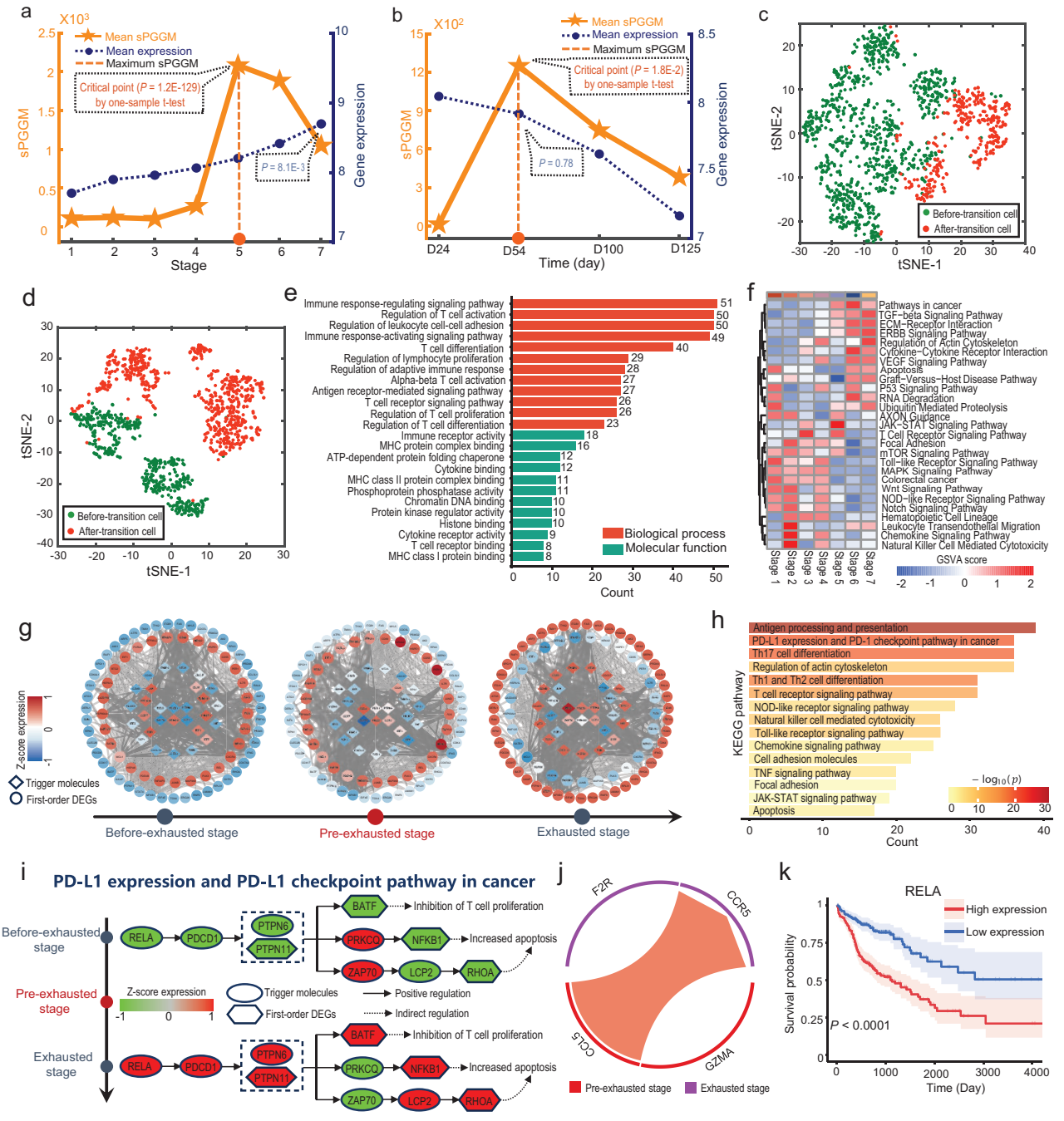

Figure 6. Identifying pre-disease stages for complex diseases at the single-cell level. The dynamic behavior of sPGGM and gene expression analyzed for (a) CD8+ T-cell exhaustion dataset and (b) GABAergic interneurons dataset. The cell clustering results based on t-distributed stochastic neighbor embedding (t-SNE) of the signaling molecules for (c) CD8+ T-cell exhaustion dataset and (d) GABAergic interneuron dataset. (e) Gene ontology (GO) analysis for functional enrichment indicated that the targeted signaling molecules are enriched in biological processes related to CD8+ T-cell exhaustion. (f) Functional analysis using GSVA indicates that targeted signaling molecules have distinct roles in the progression of CD8+ T-cell exhaustion. (g) The dynamic evolution of the regulatory network constructed from targeted signaling molecules and their neighboring differentially expressed genes (DEGs) was explored during the CD8+ T-cell exhaustion process. (h) KEGG pathway enrichment analysis was carried out for the first-order DEGs. (i) The functional analysis of targeted signaling molecules and first-order DEGs revealed the signaling mechanisms associated with the CD8+ T-cell exhaustion in the PD-L1 expression and PD-1 checkpoint pathway in cancer. (j) Cellular communication between the pre-exhausted CD8+ T-cell subset (stage 5) and the exhausted CD8+ T-cell subset (stage 7) occurs through *CCL5_CCR5* and *GZMA_F2R* receptor-ligand interactions. (k) Survival analysis of *RELA* based on the TCGA-COAD dataset.

consistently indicate the critical state at day 54, with the sPGGM showing the most statistically significant signal. Moreover, the sPGGM is a sample-perturbed critical point detection model, meaning that it can detect critical signals at the specific sample/cell level when given a set of reference samples. Therefore, the sPGGM has its own specific advantages in analyzing critical states. At the identified critical state, the top 5% of genes showing the highest local sPGGM value were chosen as signaling molecules for further analyses of functional roles and biological processes. It is seen from Fig. 6c–d that the cell clustering results based on signaling molecules clearly differentiate the stages before and after the critical transition, specifically around stage 5 for the CD8+ T-cell exhaustion dataset and day 54 for the GABAergic interneurons dataset.

Transcription factors (TFs) have been shown to play a pivotal role in regulating CD8+ T-cell exhaustion in colorectal cancer (CRC) by controlling the expression of related target genes [44]. To investigate the mechanism of signaling molecules, we conducted an analysis to explore the functions of signaling molecules regulated by upstream transcription factors. In this study, 67% of signaling molecules were found to be regulated by TFs (Fig. S10), suggesting their potential role in mediating CD8+ T-cell exhaustion in CRC. Moreover, we performed GO enrichment analysis on targeted signaling molecules. As shown in Fig. 6e, they are significantly enriched in biological processes associated with CD8+ T-cell exhaustion, such as the immune response-regulating signaling pathway, regulation of T-cell activation, and regulation of leukocyte cell-cell adhesion. Additionally, from a molecular function standpoint, they are primarily enriched in MHC protein complex binding, cytokine binding, and T-cell receptor binding, with disruptions in these functions directly impairing T-cell activation, reducing anti-tumour immunity, and celebrating T-cell exhaustion [45]. GSVA analysis of the targeted signaling molecules was conducted to further explore the dynamic changes in biological functions and pathways throughout the CD8+ T-cell exhaustion process [46]. As illustrated in Fig. 6f, the GSVA score for pathways such as Pathways in cancer, TGF-beta signaling pathway, ECM-receptor interaction, ERBB signaling pathway, and Regulation of actin cytoskeleton exhibits an upward trend as the exhaustion process advances. The upregulation of these pathways may promote the formation of an immunosuppressive microenvironment, thereby inhibiting CD8+ T-cell function and ultimately affecting the anti-tumour immune response.

To further uncover the molecular mechanisms behind tumour progression at the network level, functional analysis was performed on the PPI subnetworks of targeted signaling molecules, constructed from these molecules and their neighboring differentially expressed genes (DEGs) within the PPI network. As illustrated in Fig. 6g, a distinct shift in gene expression patterns within the networks emerges after the pre-exhaustion stage, with significant changes in the expression levels of targeted signaling molecules and their first-order DEGs, underscoring the crucial role of these reversed genes in CD8+ T-cell exhaustion. Moreover, the KEGG enrichment analysis shows that they are significantly enriched in cancer immunology-related pathways, such as the antigen processing and presentation, the PD-L1 expression and PD-1 checkpoint pathway in cancer, and the T-cell receptor signaling pathway (Fig. 6h). Notably, the PD-L1 expression and PD-1 checkpoint pathway in cancer represents a particularly significant immunological pathway related to CD8+ T-cell exhaustion. It is observed that the upregulation of *RELA*, acting as a targeted signaling molecule regulated by transcription factors (TFs), drives the high expression of *PDCD1* and *PTPN6/PTPN11* (Fig. 6i), which may play a critical role in CD8+ T-cell exhaustion. Specifically, the upstream signaling molecule *RELA* can promote the high expression of *PDCD1* (PD-1), an inhibitory receptor crucial for CD8+ T-cell exhaustion [47]. *PDCD1* subsequently drives the high expression of *PTPN6* and *PTPN11*, which mainly negatively regulate TCR signaling through dephosphorylation, leading to a reduction in T-cell activity [48]. The activation of *PTPN6/PTPN11* triggers several key downstream signaling pathways: first, *PTPN6/PTPN11* inhibit T-cell proliferation by activating *BATF* [49]. Second, they suppress the expression of *PRKCQ*, which reduces the activation of the NF-κB signaling pathway, leading to decreased T-cell survival signals and increased apoptosis of T-cells [50]. Additionally, they negatively regulate TCR signaling by inducing low expression of *ZAP70*, a critical molecule in T-cell receptor signaling, thereby weakening TCR-mediated cell activation [51]. Consequently, the low expression of *ZAP70* activates the high expression of *LCP2* and *RHOA* primarily involved in cytoskeletal remodelling and regulation. Its upregulation affects T-cell migration and the formation of immune synapses by modulating the cytoskeleton, thereby diminishing T-cell activation and further promoting T-cell apoptosis [52]. Thus, our findings suggest that the upregulation of *RELA* drives the high expression of key molecules such as *PDCD1* and *PTPN6/PTPN11* (Fig. S11), which may activate downstream signaling pathways that mediate CD8+ T-cell exhaustion. In terms of intercellular communication, the

pre-exhausted CD8+ T-cell subset (stage 5) transmits a robust exhaustion signal to the exhausted CD8+ T-cell subset (stage 7), with *CCL5_CCR5* and *GZMA_F2R* receptor-ligand interactions positively regulating this process (Fig. 6j), thereby promoting CD8+ T-cell exhaustion. Additionally, the prognosis analysis results indicate that high *RELA* expression in tumour tissues is associated with poor overall survival (Fig. 6k).

DISCUSSION

Identifying the pre-disease stages of complex diseases is crucial for preventing or delaying disease deterioration. However, traditional methods are not well-suited to capture the dynamics of disease progression and often fail to identify critical transitions in real biological datasets characterized by high data noise, patient heterogeneity, and small sample sizes. To address this challenge, based on our recently proposed DNB theory [53], along with the concepts of population-level optimal transport and Gaussian graphical models [24], we present a robust computational method called the sPGGM, which effectively reveals critical points or pre-disease stages and identifies signaling molecules involved in critical transitions from a sample-specific perspective. Our proposed sPGGM has been validated with simulated data and applied to the analysis of both scRNA-seq and bulk sequencing data across various diseases, including four single-cell datasets, influenza infection data, and six distinct tumour datasets (COAD, THCA, KIRC, UCEC, KIRP, and LIHC). The accurate prediction of pre-disease stages for these complex diseases at the specific sample/cell level highlights that our method is a valuable tool for health assessment and personalized precision medicine. Additionally, the strong performance of the sPGGM in identifying disease-related critical states was proven through comparisons with previous single-sample approaches on both single-cell and bulk data.

The advantages of our proposed sPGGM can be briefly summarized as follows. First, being a distribution-based model, the sPGGM exhibits a strong robustness and stability, as shown by its effective performance across bulk data of small sample sizes and high-noise single-cell data. Second, by introducing Gaussian graphical optimal transport to measure the dynamic differences between baseline and sample-perturbed distributions, the sPGGM outperforms existing single-sample methods in identifying pre-disease stages during disease progression. Third, given a set of reference samples, the sPGGM not only identifies critical signals toward a deteriorated stage at the sample-specific level but also highlights key signaling molecules associated with crucial biological processes, offering significant advancements in disease pathology analysis and personalized precision medicine. In particular, the trend of the signal curve, i.e. a sudden increase as it approaches the critical point, indicates that our proposed method is effective when the sizes of the reference samples fall within a specified range (Figs S12 and S13). Fourth, unlike traditional techniques that rely on differential equations for simulations, the sPGGM emphasizes data-driven insights by directly extracting information from the data, without the need for predefined parameters. Moreover, the sPGGM can demonstrate its scalability and effectiveness for analyzing large single-cell RNA-seq datasets (over 2 million cells across five time points and 356,213 cells from six age groups) (Fig. S14). However, a limitation of the sPGGM is its reliance on undirected networks, which overlook causal relationships between nodes, presenting a potential area for improvement in our future research.

METHODS

Theoretical background

Disease progression frequently exhibits abrupt shifts in temporal patterns and can be described as a time-varying nonlinear process in the context of dynamical systems, where a sudden state deterioration signifies a qualitative transition at a bifurcation point [54]. The dynamic progression of diseases generally is defined by three states (see Fig. 1a): (i) a normal stage characterized by minimal fluctuation and strong resilience; (ii) a pre-disease stage marked by inherent instability and considerable complexity, indicating a critical transition toward a disease deterioration state; and (iii) a subsequent disease stage associated with the onset or worsening of the disease. The key to detect pre-disease stages or critical points lies in developing an index that quantitatively measures dynamical changes in the state of disease systems. However, the difficulty in distinguishing between the normal stage and the pre-disease stage becomes more evident when compared to the disease state. Therefore, traditional statistical techniques may struggle to differentiate the pre-disease stage.

Based on our recently proposed theoretical concept of DNB [9,54], as the system nears a critical point, a set of molecules (DNB variables) with strong correlations and large fluctuations emerges, signaling an impending critical state transition from a network-level perspective. It is evident that the state shift or phase of a system can be characterized by a dynamic change in both the multivariate distribution and molecular associations of DNB

members. Therefore, by applying the Gaussian graphical model and optimal transport theory, we introduce the sPGGM to detect critical signals that mark the key transition from the normal stage to the disease stage, which identifies the pre-disease stage from a sample-specific perspective and addresses challenges such as small sample sizes, high noise levels, and sample heterogeneity. In our study, the Gaussian graphical model is defined by a graph structure paired with a Gaussian distribution, enabling the graph to depict the dependencies among molecules within the multivariate Gaussian distribution [55–57]. A comprehensive description of the Gaussian graphical model employed in our analysis is introduced in Section N of the online Supplementary Information.

A quantitative approach to identify the pre-disease stage based on the sPGGM

Using a set of reference samples/cells obtained from a relatively healthy condition, the proposed sPGGM was employed to identify the pre-disease stage or critical state from a sample-specific/cell-specific perspective. The details of this process are provided in Section N within the online Supplementary Information.

DATA AND CODE AVAILABILITY

Nine real datasets were employed in this study, which included the influenza infection dataset (GSE30550), GABAergic interneurons dataset (GSE93593) and CD8+ T-cell exhaustion dataset (GSE108989) sourced from the GEO database (http://www.ncbi.nlm.nih.gov/geo/), and COAD, THCA, KIRC, UCEC, KIRP, and LIHC datasets obtained from the TCGA database (http://cancergenome.nih.gov). The source code of the algorithm and related data are available at https://github.com/Junxian-Li-0/sPGGM_project.

SUPPLEMENTARY DATA

Supplementary data are available at *NSR* online.

FUNDING

This research was supported by the National Natural Science Foundation of China (T2341022, 12322119, 42450084, 12271180, and 12401630), the Educational Commission of Guangdong Province of China (2023KQNCX073), the Natural Science Foundation of Guangdong Province of China (2023A1515110558 and 2024A1515011797), and the Guangdong Provincial Key Laboratory of Mathematical and Neural Dynamical Systems (2024B1212010004).

AUTHOR CONTRIBUTIONS

R.L., P.C. and L.F. conceived the research. J.Y.Z., J.X.L. and X.R.G. performed the real data analysis. All authors wrote the paper. All authors read and approved the final manuscript.

Conflict of interest statement. None declared.

REFERENCES

1. Liu R, Wang X, Aihara K *et al.* Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers. *Med Res Rev* 2014; **34**: 455–78.
2. Liu J, Ding D, Zhong J *et al.* Identifying the critical states and dynamic network biomarkers of cancers based on network entropy. *J Transl Med* 2022; **20**: 254.
3. Achiron A, Grotto I, Balicer R *et al.* Microarray analysis identifies altered regulation of nuclear receptor family members in the pre-disease state of multiple sclerosis. *Neurobiol Dis* 2010; **38**: 201–9.
4. Tang S, Xue Y, Qin Z *et al.* Counteracting lineage-specific transcription factor network finely tunes lung adeno-to-squamous transdifferentiation through remodeling tumor immune microenvironment. *Natl Sci Rev* 2023; **10**: nwad028.
5. Scheffer M, Bascompte J, Brock WA *et al.* Early-warning signals for critical transitions. *Nature* 2009; **461**: 53–9.
6. Trefois C, Antony PM, Goncalves J *et al.* Critical transitions in chronic disease: transferring concepts from ecology to systems medicine. *Curr Opin Biotechnol* 2015; **34**: 48–55.
7. Yan H, Zhang F, Wang J. Thermodynamic and dynamical predictions for bifurcations and non-equilibrium phase transitions. *Commun Phys* 2023; **6**: 110.
8. Zhou JX, Aliyu MD, Aurell E *et al.* Quasi-potential landscape in complex multi-stable systems. *J R Soc Interface* 2012; **9**: 3539–53.
9. Liu R, Li M, Liu ZP *et al.* Identifying critical transitions and their leading biomolecular networks in complex diseases. *Sci Rep* 2012; **2**: 813.
10. Zhong J, Liu H, Chen P. The single-sample network module biomarkers (sNMB) method reveals the pre-deterioration stage of disease progression. *J Mol Cell Biol* 2022; **14**: mjac052.
11. Liu X, Liu R, Zhao XM *et al.* Detecting early-warning signals of type 1 diabetes and its leading biomolecular networks by dynamical network biomarkers. *BMC Med Genomics* 2013; **6**: S8.
12. Yang B, Li M, Tang W *et al.* Dynamic network biomarker indicates pulmonary metastasis at the tipping point of hepatocellular carcinoma. *Nat Commun* 2018; **9**: 678.
13. Guo WF, Zhang SW, Zeng T *et al.* Network control principles for identifying personalized driver genes in cancer. *Brief Bioinform* 2020; **21**: 1641–62.
14. Li M, Zeng T, Liu R *et al.* Detecting tissue-specific early warning signals for complex diseases based on dynamical network biomarkers: study of type 2 diabetes by cross-tissue analysis. *Briefings Bioinf* 2014; **15**: 229–43.

15. Jiang L, Sui D, Qiao K *et al.* Impaired functional criticality of human brain during Alzheimer's disease progression. *Sci Rep* 2018; **8**: 1324.
16. Sha Y, Qiu Y, Zhou P *et al.* Reconstructing growth and dynamic trajectories from single-cell transcriptomics data. *Nat Mach Intell* 2024; **6**: 25–39.
17. Jin S, Plikus MV, Nie Q. CellChat for systematic analysis of cell-cell communication from single-cell transcriptomics. *Nat Protoc* 2025; **20**: 180–219.
18. Zhou P, Wang S, Li T *et al.* Dissecting transition cells from single-cell transcriptome data through multiscale stochastic dynamics. *Nat Commun* 2021; **12**: 5609.
19. Sha Y, Wang S, Zhou P *et al.* Inference and multiscale model of epithelial-to-mesenchymal transition via single-cell transcriptomic data. *Nucleic Acids Res* 2020; **48**: 9505–20.
20. Yang XH, Goldstein A, Sun Y *et al.* Detecting critical transition signals from single-cell transcriptomes to infer lineage-determining transcription factors. *Nucleic Acids Res* 2022; **50**: e91.
21. Peng H, Zhong J, Chen P *et al.* Identifying the critical states of complex diseases by the dynamic change of multivariate distribution. *Brief Bioinform* 2022; **23**: bbac177.
22. Hua W, Cui R, Yang H *et al.* Uncovering critical transitions and molecule mechanisms in disease progressions using Gaussian graphical optimal transport. *Commun Biol* 2025; **8**: 575.
23. Zhong J, Liu R, Chen P. Identifying critical state of complex diseases by single-sample Kullback-Leibler divergence. *BMC Genomics* 2020; **21**: 87.
24. Peyré G and Cuturi M. Computational optimal transport: with applications to data science. *FNT in Machine Learning* 2019; **11**: 355–607.
25. Liu X, Chang X, Leng S *et al.* Detection for disease tipping points by landscape dynamic network biomarkers. *Natl Sci Rev* 2019; **6**: 775–85.
26. Liu R, Chen P, Chen L. Single-sample landscape entropy reveals the imminent phase transition during disease progression. *Bioinformatics* 2020; **36**: 1522–32.
27. Ren J, Li P, Yan J. CPMI: comprehensive neighborhood-based perturbed mutual information for identifying critical states of complex biological processes. *BMC Bioinf* 2024; **25**: 215.
28. Lyu C, Chen L, Liu X. Detecting tipping points of complex diseases by network information entropy. *Brief Bioinform* 2024; **25**: bbae311.
29. Ronen M, Rosenberg R, Alon U *et al.* Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate ex-pression kinetics. *Proc Natl Acad Sci USA* 2002; **99**: 10555–60.
30. Khanin R, Vinciotti V, Mersinias V *et al.* Statistical reconstruction of transcription factor activity using Michaelis–Menten kinetics. *Biometrics* 2007; **63**: 816–23.
31. Hari DM, Leung AM, Lee JH *et al.* AJCC Cancer Staging Manual 7th edition criteria for colon cancer: do the complex modifications improve prognostic assessment? *J Am Coll Surg* 2013; **217**: 181–90.
32. Shaha AR. TNM classification of thyroid carcinoma. *World J Surg* 2007; **31**: 879–87.
33. Su S and Shahriyari L. RGS5 plays a significant role in renal cell carcinoma. *R Soc Open Sci* 2020; **7**: 191422.
34. Rose PG. Endometrial carcinoma. *N Engl J Med* 1996; **335**: 640–9.
35. Maida M, Orlando E, Cammà C *et al.* Staging systems of hepatocellular carcinoma: A review of literature. *World J Gastroenterol* 2014; **20**: 4141–50.
36. Ficarra V, Galfano A, Mancini M *et al.* TNM staging system for renal-cell carcinoma: current status and future perspectives. *Lancet Oncol* 2007; **8**: 554–8.
37. Reich D, Price AL, Patterson N. Principal component analysis of genetic data. *Nat Genet* 2008; **40**: 491–2.
38. Solaini G, Sgarbi G, Baracca A. Oxidative phosphorylation in cancer cells. *BBA-Bioenergetics* 2011; **1807**: 534–42.
39. Waris G and Ahsan H. Reactive oxygen species: role in the development of cancer and various chronic conditions. *J Carcinog* 2006; **5**: 14.
40. Fresno Vara JA, Casado E, de Castro J *et al.* PI3K/Akt signalling pathway and cancer. *Cancer Treat Rev* 2004; **30**: 193–204.
41. Chen X, Guo Y, Chen X. iGMDR: integrated pharmacogenetic resource guide to cancer therapy and research. *Genom Proteom Bioinform* 2020; **18**: 150–60.
42. Hu J, Han C, Zhong J *et al.* Dynamic network biomarker of pre-exhausted $CD8^+$ T cells contributed to T cell exhaustion in colorectal cancer. *Front Immunol* 2021; **12**: 691142.
43. Close JL, Yao Z, Levi BP *et al.* Single-cell profiling of an in vitro model of human interneuron development reveals temporal dynamics of cell type production and maturation. *Neuron* 2017; **93**: 1035–1048.e5.
44. Ichiyama K, Long J, Kobayashi Y *et al.* Transcription factor Ikzf1 associates with Foxp3 to repress gene expression in Treg cells and limit autoimmunity and anti-tumor immunity. *Immunity* 2024; **57**: 2043–60.E10.
45. Corria-Osorio J, Carmona SJ, Stefanidis E *et al.* Orthogonal cytokine engineering enables novel synthetic effector states escaping canonical exhaustion in tumor-rejecting $CD8^+$ T cells. *Nat Immunol* 2023; **24**: 869–83.
46. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinf* 2013; **14**: 7.
47. Zhou C, Li W, Liang Z *et al.* Mutant KRAS-activated circATXN7 fosters tumor immunoescape by sensitizing tumor-specific T cells to activation-induced cell death. *Nat Commun* 2024; **15**: 499.
48. Tojjari A, Saeed A, Sadeghipour A *et al.* Overcoming immune checkpoint therapy resistance with SHP2 Inhibition in cancer and immune cells: A review of the literature and novel combinatorial approaches. *Cancers* 2023; **15**: 5384.
49. Zha H, Jiang Y, Wang X *et al.* Non-canonical PD-1 signaling in cancer and its potential implications in clinic. *J Immunother Cancer* 2021; **9**: e001230.
50. Song Y, Zhao M, Zhang H *et al.* Double-edged roles of protein tyrosine phosphatase SHP2 in cancer and its inhibitors in clinical trials. *Pharmacol Ther* 2022; **230**: 107966.
51. Liu Q, Qu J, Zhao M *et al.* Targeting SHP2 as a promising strategy for cancer immunotherapy. *Pharmacol Res* 2020; **152**: 104595.
52. Iyer VS, Boddul SV, Johnsson AK *et al.* Modulating T-cell activation with antisense oligonucleotides targeting lymphocyte cytosolic protein 2. *J Autoimmun* 2022; **131**: 102857.
53. Zhong J, Tang H, Huang Z *et al.* Uncovering the pre-deterioration state during disease progression based on sample-specific causality network entropy (SCNE). *Research* 2024; **7**: 0368.
54. Scheffer M, Carpenter S, Foley JA *et al.* Catastrophic shifts in ecosystems. *Nature* 2001; **413**: 591–6.
55. Uhler C. Gaussian graphical models: an algebraic and geometric perspective. arXiv: 1707.04345.
56. Bishop CM and Nasrabadi NM. *Pattern Recognition and Machine Learning*. New York: Springer, 2007.
57. Zhao H and Duan ZH. Cancer genetic network inference using Gaussian Graphical Models. *Bioinform Biol Insights* 2019; **13**: 1177932219839402.