Nguyen *BMC Bioinformatics* (2024) 25:149
https://doi.org/10.1186/s12859-024-05755-0

BMC Bioinformatics

RESEARCH Open Access

Biomarker discovery with quantum neural networks: a case-study in *CTLA4*-activation pathways

Phuong-Nam Nguyen[1*]

*Correspondence: nam.nguyenphuong@phenikaa-uni.edu.vn

[1] Faculty of Computer Science, PHENIKAA University, Yen Nghia, Ha Dong, Hanoi 12116, Vietnam

Abstract

Background: Biomarker discovery is a challenging task due to the massive search space. Quantum computing and quantum Artificial Intelligence (quantum AI) can be used to address the computational problem of biomarker discovery from genetic data.

Method: We propose a Quantum Neural Networks architecture to discover genetic biomarkers for input activation pathways. The Maximum Relevance-Minimum Redundancy criteria score biomarker candidate sets. Our proposed model is economical since the neural solution can be delivered on constrained hardware.

Results: We demonstrate the proof of concept on four activation pathways associated with *CTLA4*, including (1) *CTLA4*-activation stand-alone, (2) *CTLA4-CD8A-CD8B* co-activation, (3) *CTLA4-CD2* co-activation, and (4) *CTLA4-CD2-CD48-CD53-CD58-CD84* co-activation.

Conclusion: The model indicates new genetic biomarkers associated with the mutational activation of *CLTA4*-associated pathways, including 20 genes: *CLIC4*, *CPE*, *ETS2*, *FAM107A*, *GPR116*, *HYOU1*, *LCN2*, *MACF1*, *MT1G*, *NAPA*, *NDUFS5*, *PAK1*, *PFN1*, *PGAP3*, *PPM1G*, *PSMD8*, *RNF213*, *SLC25A3*, *UBA1*, and *WLS*. We open source the implementation at: https://github.com/namnguyen0510/Biomarker-Discovery-with-Quantum-Neural-Networks.

Keywords: Quantum algorithm, Quantum computing, Biomarker identification, Bioinformatics

Introduction

A biomarker, a molecular marker or signature molecule, refers to a biological substance or characteristic found in body fluids, tissues, or blood that indicates the presence of a condition, disease, or abnormal process. Biomarkers can be measured to assess how well the body responds to treatment for a particular disease or condition [11]. Biomarkers play a crucial role in drug discovery and development by providing essential information on the safety and effectiveness of drugs. These measurable indicators can be categorized into diagnostic, prognostic, or predictive biomarkers, and they are utilized to choose patients for clinical trials or track patient response and treatment efficacy. The

Next-Generation Sequencing (NGS) technology has revolutionized the field of oncology by enabling the comprehensive and precise identification of *genetic biomarkers*, paving the way for personalized cancer therapies and improved patient outcomes.

The study [22] introduces a new metric, the Intelligent Gene (I-Gene) score, to measure the importance of individual biomarkers for predicting complex traits. Their Machine learning (ML) pipeline combines classical statistical methods and state-of-the-art algorithms for biomarker discovery. Another research group develops an autoencoder-based biomarker identification method by reversing the learning mechanism of the trained encoders [3]. It provides an explainable post hoc methodology for identifying influential genes likely to become biomarkers. In [27], a Deep learning (DL) pipeline predicts the status of five biomarkers in LGG using slide-level biomarker status labels and whole slide images stained with hematoxylin and eosin. The research assesses the performance of several state-of-the-art Random Forest (RF) based decision approaches, including the Boruta method, permutation-based feature selection with and without correction, and the backward elimination-based feature selection method [1]. A review offers tips to overcome common challenges in biomarker signature development, including supervised and unsupervised ML, feature selection, and hypothesis testing [25]. Another systematic review examines the current state of the art and computational methods, including feature selection strategies, ML and DL approaches, and accessible tools to uncover markers in single and multi-omics data [24]. Despite its valuable role, genetic biomarker discovery is a challenging task for classical-computational platforms due to the massive search space ("Problem statement" section).

Quantum computing is an emerging technology that utilizes the principles of quantum mechanics to solve problems beyond classical computers' capabilities. Quantum Machine Learning and Quantum Neural Networks are an advanced class of machine intelligence on quantum hardware, which promises more powerful models for myriad learning tasks ("Quantum neural networks" section). A comprehensive review discusses quantum computing technology and its status in solving molecular biology problems, especially in the next-generation computational biology scenario [59]. The review covers the basic concept of quantum computing, the functioning of quantum systems, quantum computing components, and quantum algorithms. HypaCADD, a hybrid classical-quantum workflow for finding ligands binding to proteins, is introduced in [42]. While accounting for genetic mutations, it combines classical docking and molecular dynamics with QML to infer the impact of mutations. The study found that the QML models can perform on par with, if not better than, classical baselines. Another systematic review presents the recent progress in quantum computing and simulation within the field of biological sciences. It discusses quantum computing components, such as quantum hardware, quantum processors, quantum annealing, and quantum algorithms [58]. A review comments on recently developed Quantum Computing (QC) bio-computing algorithms, focusing on multi-scale modeling and genomic analyses [47]. The research group highlights the possible advantages over the classical counterparts and describes some hybrid classical/quantum approaches. In a non-conventional track, scientists at Oak Ridge National Laboratory used their expertise in quantum biology, artificial intelligence, and bioengineering to improve how CRISPR Cas9 genome editing tools work on organisms like microbes that can be modified to produce

renewable fuels and chemicals [41]. The cross-section of the genetic editing technology with quantum methods shows promising new insights into biomedical science.

This work uses a class of Quantum Artificial Intelligence (AI) models to discover genetic biomarkers in biomedical research. We adopt the neural architecture proposed in a recent work that addresses the body dynamics modeling problem [55]. Here, we make a non-trivial adaptation of the proposed game theory to tackle a different class of problems. The main contribution of this study is summarized as follows:

1. The proposed quantum AI model is a general, cost-efficient, cost-effective algorithm for biomarker discovery from genetic data despite the extensive problem complexity.
2. The model outcomes suggest novel biomarkers for the mutational activation of the notable target in immuno-therapy - *CLTA4*, including 20 genes: *CLIC4*, *CPE*, *ETS2*, *FAM107A*, *GPR116*, *HYOU1*, *LCN2*, *MACF1*, *MT1G*, *NAPA*, *NDUFS5*, *PAK1*, *PFN1*, *PGAP3*, *PPM1G*, *PSMD8*, *RNF213*, *SLC25A3*, *UBA1* and *WLS*.

We organize the article as follows: "Preliminary" section formalizes the biomarker identification problem as a combinatory optimization problem and the preliminary for QNN models; "Method" section introduces our proposed model architecture and the scoring algorithm; "Results" section reports the in *silico* discovery for genetic biomarkers of four immunotherapy pathways with posthoc validation using literature mining over clinical research; "Conclusion" section concludes our research by suggesting several further research direction.

Preliminary

Problem statement

The learning task involves identifying the best combinations of genetic biomarkers from a given set of genes (represented as $\mathcal{G}$) to select those that are (1) most relevant to a specific pathway (represented as $\boldsymbol{Y}$) and (2) optimal for machine learning algorithms. This criterion is commonly referred to as minimizing redundancy and maximizing relevancy for selected feature sets, as proposed in [62] (See Additional file 1: Appendix A).

In this context, it is worth noting that each individual has unique patterns of mutational alterations that lead to distinct sets of genetic biomarkers. Considering all the possibilities, the number of candidate biomarker sets is the sum of all possible combinations, which can be expressed as

$$\sum_{i=0}^{N} \binom{N}{k} = 2^N \tag{1}$$

Since the human genome contains around 20,000 to 25,000 genes, this results in a massive search space of 3.2019×10^{6577} candidate biomarker set for any biomarker identification algorithm from genetic databases. This search space grows exponentially with the number of input genes, making it an incredibly challenging problem for conventional computing methods. Quantum computing holds great promise for advancing genomic research, particularly in quantifying biomarkers. In this context, the logarithmic scaling complexity of quantum algorithms becomes evident, as exemplified by the requirement of $\log_2 20,000$ to $\log_2 25,000$ noise-tolerant qubits for quantifying genome sets,

approximately equivalent to 15 qubits. The salient advantage of employing quantum hardware for biomarker quantification lies in the $O(\log N)$ scaling complexity, providing a significant computational advantage as the genomic dataset size increases. For instance, in the presence of four multimodality datasets encompassing DNA Methylation, RNA, mRNA, and Protein, each with a homogeneous number of features denoted as M, the required number of qubits scales only to $N = \log_2 4M = 2 + \log_2 M$, illustrating a scalable problem complexity. However, the current state-of-the-art quantum hardware faces limitations in facilitating multimodality analysis, primarily due to the constraints imposed by the limited and noisy qubits available, hindering their effectiveness in realizing the full potential of quantum computational power in genomics.

Quantum neural networks

QNNs represent input data using wavefunction representations, typically using qubits or "qurons" [71], in the form of:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle, \alpha \text{ and } \beta \in \mathbb{C}. \tag{2}$$

This distinguishes QNNs from classical ANNs as they capture physical events discretely [48] and offer insight into representation learning [8].

QNNs have been proposed to offer two advantages over classical ANNs [72]. Quantum feature maps can represent exponentially larger data sets than classical neural networks, with an n-qubit system representing 2^n bits [57]. Secondly, quantum feature maps inherit wavefunction uncertainty from quantum mechanics, allowing measurements of quantum states to return values of 0 or 1 with probability values of $\mathbb{P}(0) = |\alpha|^2$ and $\mathbb{P}(1) = |\beta|^2$. In our previous papers, we discuss the role of epistemic uncertainty estimation from quantum maps [54] and the effect of entanglement layouts on the classifier's performance.

The typical design for QNNs [7, 28, 38, 66, 73] involves a stack of identical ansatz structures, given by a global unitary transformation:

$$U_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathcal{U}^{(l)}(\boldsymbol{\theta} l)\mathcal{V}(\boldsymbol{x}) \ldots \mathcal{V}(\boldsymbol{x})\mathcal{U}^{(0)}(\boldsymbol{\theta} 0), \tag{3}$$

where $\boldsymbol{x}$ represents input features, $\boldsymbol{\theta}$ represents model weights, $\mathcal{U}^{(l)}(.)$ represents identical-parameterized circuits serving as variational (trainable neural) blocks, and $\mathcal{V}(.)$ represents feature-embedding blocks. From our perspective, QNNs are advanced mathematical models in the language of representation theory (See Additional file 1: Appendix B); thus, we are using mathematics to enable cancer discoveries.

Method

Model architecture

We depict the ansatz structure of our QNN model in Fig. 1. Each neural block in our proposed model includes (1) a parameterized RZ-rotation based on time-domain variables $t \in [0, 2\pi]$, given by

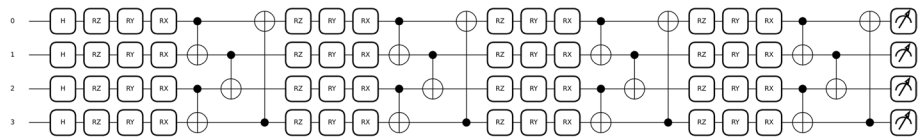

Fig. 1 The Ansatz circuit of our proposed QNN models using four qubits with four neural blocks. Here, the number of evaluated genes is $2^4 = 16$ genes. We extend the architecture to 11-qubit ansatz in the numerical result, scoring $2^{11} = 2048$ genes

$$\boldsymbol{R}_Z(t) = \begin{bmatrix} e^{-i\frac{t}{2}} & 0 \\ 0 & e^{i\frac{t}{2}} \end{bmatrix} \tag{4}$$

and (2) trainable RX-rotation and RY-rotation gates

$$\boldsymbol{R}_X(\alpha) = \begin{bmatrix} \cos(\alpha/2) & -i\sin(\alpha/2) \\ -i\sin(\alpha/2) & \cos(\alpha/2) \end{bmatrix} \tag{5}$$

and

$$\boldsymbol{R}_Y(\beta) = \begin{bmatrix} \cos(\beta/2) & -\sin(\beta/2) \\ \sin(\beta/2) & \cos(\beta/2) \end{bmatrix} \tag{6}$$

The parameterized wavefunction of the 1-layer model is given

$$|\psi_\Theta\rangle = |\psi_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{t})\rangle = \mathrm{CNOT}(\Lambda)\boldsymbol{R}_X(\boldsymbol{\alpha})\boldsymbol{R}_Y(\boldsymbol{\beta})\boldsymbol{R}_Z(\boldsymbol{t})H|0\rangle^{\otimes n}, \tag{7}$$

n is the number of qubits, and Λ is the architecture parameter of the entanglement layout constructed by CNOT gates, illustrated in Fig. 1.

Algorithm

Computations of gene scores (GSCORE©)

We train the proposed QNNs to learn the optimal sampling distribution based on mRMR criteria [56] (See Additional file 1: Appendix A). In other words, the search algorithm will assign higher probabilities for the more important markers. Thus, the search algorithm will more often select more important genetic biomarkers. Specifically, the output electronic wavefunction of the parameterized QML model is given by

$$|\psi_\Theta\rangle = \mathcal{U}_n(\Theta), \tag{8}$$

with n is the number of prepared qubits and Θ is the model parameter. The sampled probability density is given by

$$p(\Theta) = |\Psi(\Theta)|^2 = \Psi^*\Psi \tag{9}$$

where $|\Psi^*\rangle$ is the conjugate-transpose of the output wavefunction. In other words, we compute the wavefunction's probability amplitude or square modulus.

We further use the Softmax function with tunable temperature to normalize our density. Besides, using the Softmax function also allows us to control the conservative of the search engine as the low temperature will encourage the model confidence. In contrast,

high temperatures encourage less conservative predictions. Thus, the biomarker score is the sampling probability of each gene given by

$$\text{GSCORE}^{©} = \hat{p}(\Theta) = \text{SoftMax}\left(\frac{\sqrt{p(\Theta)}}{\text{temp}}\right), \tag{10}$$

where temp is the function temperature. Finally, we select the candidate marker set by parameterized thresholding with $\bar{p} = 1$ if $\hat{p} \geq \tau$, otherwise 0. We interpret the GSCORE© that the more important genes have a higher chance to be selected by the sampler, resulting in a higher probability over the output of quantum ansatzes.

Objective function

We adopt the efficient loss function of the Quadratic Programming Feature Selection (QPFS) method [65], given as

$$\min_{\boldsymbol{\Theta}} \left(\lambda \boldsymbol{p}^{\intercal} \boldsymbol{H} \boldsymbol{p} - \boldsymbol{p} \boldsymbol{F}\right), \sum_{i=1}^{n} p_i = 1, p_i > 0, \tag{11}$$

where $\boldsymbol{F}_{n\times 1}$ is the relevancy to target variables and $\boldsymbol{H}_{n\times n}$ is the pairwise redundancy computed from the feature set. Both natural and normalized quantum distributions $p(\Theta)$ and $\hat{p}(\Theta)$ satisfy the conditions for p_i in QPFS. We consider $\lambda = 1$ for further analysis, i.e., the balanced loss between redundant-relevant criteria.

Pseudo-code

We implement the proposed model using the quantum simulation package Pennylane [9] with Pytorch 3.7 [60]. The mutual information criteria are computed by Scikit-learn [61], and model optimization is conducted by Optuna [2] with Tree Parzen Estimators [10]. The pseudo-code is given in the following algorithm:

```
import pennylane as qml
import torch
import numpy as np
q = 11 #Number of Qubits
dev = qml.device("default.qubit", wires=q)
device = torch.device("cuda:0"
    if torch.cuda.is_available() else "CPU")
@qml.qnode(dev, interface="torch")
def quantum_sampler(a,b,n):
  t = torch.tensor(np.linspace(0,np.pi, n))
  for i in range(q):
      qml.Hadamard(wires=i)
  for k, dt in enumerate(t):
      for i in range(q):
          qml.RZ(dt,wires=i)
          qml.RY(a[k,i],wires=i)
          qml.RX(b[k,i],wires=i)
      for i in range(0, q - 1, 2):
          qml.CNOT(wires=[i, i + 1])
      for i in range(1, q - 1, 2):
          qml.CNOT(wires=[i, i + 1])
      qml.CNOT(wires = [q-1,0])
  return qml.state()
```

Listing 1: Model architecture of the duality game model developed for biomarker identification task

Hyper-parameter and training protocol

We report the hyper-parameters for our search engine powered by proposed QNNs in Table 1. Noteworthy, we adopt the Tree Parzen Estimator with Sequential Model-based Optimization [10] (SMBO) to train our model. We learned from our previous works - CSNAS [53] and BayesianQNN [54] that the used optimization is a cost-efficient algorithm, which enables effective search on a massive search space.

Results

Case-study

CTLA4-activation Pathways

CTLA4 - The gene is part of the immunoglobulin superfamily and produces a protein that transmits a signal to inhibit T cells. Mutations in this gene have been linked to

Table 1 Hyper-parameter of our model optimization

Parameters	Range	Role
Number of Qubits	11	Sampler for $2^{11} = 2048$ genes
$\boldsymbol{\alpha}$	$[-2\pi, 2\pi]$	$\boldsymbol{R}_X$ rotation
$\boldsymbol{\beta}$	$[-2\pi, 2\pi]$	$\boldsymbol{R}_Y$ rotation
ϵ	$[\pi/96, \pi/24]$	Standard Deviation of Model Weight
τ	[0.5, 0.7]	Threshold Value
temp	$[1/10^3, 2 \times 10^3]$	SoftMax Temperature

various autoimmune diseases, including insulin-dependent diabetes mellitus, Graves disease, Hashimoto thyroiditis, celiac disease, systemic lupus erythematosus, and thyroid-associated orbitopathy [31]. *CD8A* and *CD8B* encode the alpha and beta chains of the CD8 antigen, respectively. The CD8 antigen is a cell surface glycoprotein found on most cytotoxic T lymphocytes that mediate efficient cell-cell interactions within the immune system [14]. The CD8 antigen acts as a coreceptor with the T-cell receptor on the T lymphocyte to recognize antigens displayed by an antigen-presenting cell in class I MHC molecules. Moreover, CD8+ T cells play a significant role in the response to immunotherapies that target CTLA4 [40]. Following a common preclinical combination treatment protocol, the study used a radiolabeled antibody to detect changes in CD8a+ infiltration in murine colon tumors. The results showed that the treatment effectively inhibited tumor growth and increased the overall survival of mice. *CD2* and *CD28* are both important co-receptors involved in T-cell activation [30, 74]. They are part of the immunoglobulin superfamily and are known to regulate T-cell activation in a coordinated manner. *CD2* enhances adhesion between T cells and target cells and delivers an activation signal [29]. On the other hand, *CD28* is a costimulatory receptor that can strongly enhance TCR signaling responses. It is believed that *CD28* and *CD2* may function together to facilitate interactions of the T cell and antigen-presenting cells (APCs), allowing for efficient signal transduction through the TCR [30]. The precise mechanisms of *CTLA4*'s inhibitory role are not fully understood, but it is believed that *CTLA4* can compete with *CD28* for ligand binding, acting as an antagonist of *CD28*-mediated costimulation [68]. This interaction is thought to occur at the immune synapse between T cells and antigen-presenting cells (APCs), where *CTLA4* has been shown to recruit *CD80*, thereby limiting its interactions with *CD28*. Studying the coactivation of *CTLA4* and *CD2* could provide valuable insights into T-cell activation and immune regulation. However, limited research addresses the coactivation of *CTLA4* and *CD2*. This suggests further investigation is needed to establish a connection and understand its implications for immunotherapy and autoimmune disease treatment. Understanding these interactions could potentially lead to the development of more effective therapeutic strategies. In this study, we aim to broaden the scope of immunological research by studying the coactivation of *CTLA4*, *CD2*, and the *CD2*-associated genes, including *CD48*, *CD53*, *CD58*, and *CD84*. *CD48*, a member of the CD2 subfamily of the immunoglobulin superfamily, is found on the surface of lymphocytes and other immune cells and participates in activation and differentiation pathways in these cells. *CD53*, another member of the tetraspanin superfamily, regulates various cellular processes such as adhesion, migration, signaling, and cell fusion. *CD58*, also known as lymphocyte function-associated antigen 3 (LFA-3), is a cell adhesion molecule that strengthens the adhesion and recognition between T cells and antigen-presenting cells, facilitating signal transduction necessary for an immune response. Lastly, *CD84*, a member of the signaling lymphocyte activation molecule (SLAM) family, forms homophilic dimers by self-association and is reported as an important survival receptor in chronic lymphocytic leukemia. By studying the coactivation of these molecules, we hope to gain a deeper understanding of the complex interactions and signaling pathways involved in immune regulation. This could lead to the development of more effective therapeutic strategies for various immune-related diseases.

We will investigate four pathways regarding the co-occurrence of mutational activation, including:

1. **Pathway 1**: *CTLA4*-activation stand-alone.
2. **Pathway 2**: *CTLA4-CD8A-CD8B* activated simultaneously.
3. **Pathway 3**: *CTLA4-CD2* activated simultaneously.
4. **Pathway 4**: *CTLA4-CD2-CD48-CD53-CD58-CD84* activated simultaneously.

Advanced ML or Quantum AI has yet to study these mutational activation pathways to extend our knowledge. We summarize the biological meaning of quantified targets in Additional file 1: Appendix C.

Datasets

We use The Cancer Genome Atlas (TCGA [79]) with RNA expression data and Copy Number of Variation (CNV). We score 2048 genetic biomarkers based on its expression, which is equivalent to 2^{11} dimensional embedding generated by 11-qubit system ("Pseudo-code" section). The evaluated cohort includes 9, 136 patients, considered Big Data in the context of a cancer genetic study. The expression set $\boldsymbol{X}$ is normalized using Min-Max normalization, and the mutational signals are created from CNV with $\boldsymbol{Y} = 1$ if CNV $\neq 0$, otherwise $\boldsymbol{Y} = 0$. The evaluated expressions are continuous values, while the activation signals are binary.

Experimental settings

All experiments were carried out using Python 3.7.0, numpy 1.21.5, sci-kit-learn 1.0.2, and PyTorch 1.11 on an Intel i9 processor (2.3 GHz, eight cores), 16GB DDR4 memory and GeForce GTX 1060 Mobile GPU with 6GB memory. Besides, we made our implementation available at: https://github.com/namnguyen0510/Biomarker-Discovery-with-Quantum-Neural-Networks; and all experimental history available at https://tinyurl.com/55x77w4h. We train our sampler for 600 trials with paralleled computing using ten CPU workers for each pathway.

Quantum AI-driven genetic biomarkers for *CTLA4*-activation pathways

We summarize the case-study of *CTLA4*-activation pathways in Fig. 2A. The full reports for each pathway is given in Additional file 1: Figs. S3, S4, S5 and S6 in Additional file 1: Appendix D. We only consider the top 20 genetic biomarkers for further analysis. Of note, the discovered biomarker sets are distinctive for each studied pathway. Specifically, only one genetic biomarker *MACF1* is founded as the genetic biomarker for *CTLA4* and *CTLA4-CD8A-CD8B* activation pathways (Fig. 2B). Similarly, *HSPA1B* is addressed as a common genetic biomarker for the pathways *CTLA4-CD8A-CD8B* and *CTLA4-CD2*. Apart from *MACF1* and *HSPA1B*, the remaining genetic biomarkers are distinctively associated with the quantified targets.

Convergence analysis

We construct an end-to-end explainable quantum AI, illustrated in the results in Additional file 1: Figs. S1, S2, S3 and S4. Of note, we are addressing a complex problem

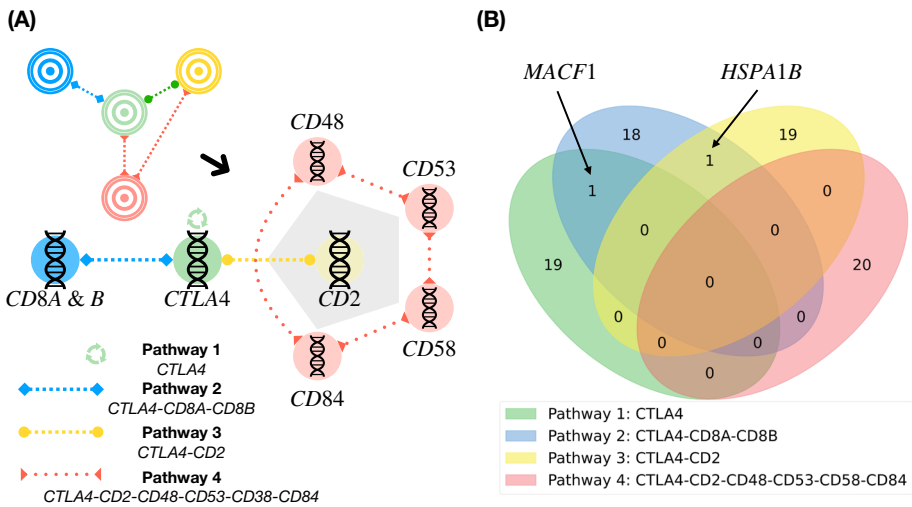

Fig. 2 **A** Activation Pathways Analyzed in Our Case-study: We expanded the scope of immunological research by studying the coactivation of CTLA4, CD2, and associated genes, including CD48, CD53, CD58, and CD84. These molecules play significant roles in T-cell immune regulation, and their interactions could potentially lead to the development of more effective therapeutic strategies for various immune-related diseases. **B** Venn Diagram of Discovered-Genetic Biomarker Sets regarding The Quantified Targeted Pathways: *MACF1*, a protein facilitating actin-microtubule interactions at the cell periphery, is a common genetic biomarker for both *CTLA4* and *CTLA4-CD8A-CD8B* pathways. On the other hand, *HSPA1B*, a member of the heat shock protein 70 family that stabilizes existing proteins against aggregation, is the common genetic biomarker for *CTLA4-CD8A-CD8B* and *CTLA4-CD2* pathways. These findings highlight the potential of these biomarkers in understanding immune regulation and developing therapeutic strategies, which has not yet been well-studied, discussed in "*CTLA4*-activation Pathways" section

beyond the capacity of classical computers; thus, our proposed algorithm will likely suggest a sub-optimal solution. We show in these results of Additional file 1: Figs. S1, S2, S3 and S4 that the proposed algorithm can effectively score genes and sampling biomarker sets as the sampling loss is reducing. Training beyond 600 trials does not guarantee better solutions as most of the best loss is found by trial 400th. We average top-50% models to have a more robust inference of GSCORE©, which shows that higher GSCORE© tends to have higher score variation since the markers are more frequently sampled by the quantum model. However, the score variation is extremely small with under $200\mu = 2 \times 10^{-6}$, indicating the well-convergence of neural solutions.

Furthermore, the output panels in Additional file 1: Figs. S1, S2, S3 and S4 also report the landscape of hyper-optimization protocol, in which model configuration with the lower score is in a darker color (purple) and model configuration with a higher score is in brighter color (yellow). This analysis significantly reduces the cost of model deployment on actual quantum computers, as we can select the optimal design of quantum ansatz circuits based on analytical solutions using classical simulations.

Significance of discovered genetic biomarkers

Statistical significance

Using the cBioPortal [16] databases from nine projects [12, 36, 50, 52, 64, 67, 70, 81, 86] with a total of 73, 717 samples, we validated the statistical significance of the top-5 genetic biomarkers for each pathway (see Section 8.4). We found that only a small proportion of the total samples contained mutations in certain biomarkers, with the highest

mutated biomarkers being *PAK1* (1.5%, $n = 858$) and *RNF213* (0.3%, $n = 62$). However, the remaining biomarkers accounted for an extremely small proportion of all mutations, making it insufficient to use mutational profiles alone to study the relationship between these biomarkers and targets. We also found co-occurrence of mutations in certain biomarkers with significant statistical evidence regarding the *CTLA4*-activation pathway. Specifically, *UBA1*, *HYOU1*, and *RNF213* mutations were co-occurring in this pathway. Additionally, mutations in *MAFC1*, *WLS*, *PSMD8*, and *PAK1* were co-occurring in the *CTLA4-CD8A-CD8B* pathways, while mutations in *NAPA*, *PGAP3*, *CLIC4*, and *PPM1G* were co-occurring in the activation of *CTLA4-CD*2. Lastly, the extended activation pathway four was associated only with the coactivation of *LCN2* and *FAM107A* mutations.

Clinical significance

We perform literature mining from the PubMed.gov library regarding 19 top-5 genetic biomarkers, excluding CPE due to similar abbreviations (see Section 8.4). The analysis is on publications over the three years from 2020 to 2023, up to 12/05/2023. Additional file 1: Table S2 shows that 12 over 19 biomarkers are rarely known in clinical-associated literature but addressed as significant genetic biomarkers by our proposed model.

Pathway 1: CTLA4

Hypoxia upregulated protein 1 is a protein that in humans is encoded by the *HYOU1* gene. The protein encoded by this gene belongs to the heat shock protein 70 family. This gene uses alternative transcription start sites. Using a type of bone-forming cell called MC3T3-E1, the study [92] demonstrated that high levels of glucose decrease the ability of the cells to survive and cause them to undergo programmed cell death. The high glucose levels also cause endoplasmic reticulum stress (ERS) by increasing the movement of calcium and producing a protein called binding immunoglobulin protein (BiP) in the endoplasmic reticulum. This results in the activation of eukaryotic initiation factor 2 alpha (eIF2alpha) downstream of a protein called PKR-like ER kinase (PERK). This, in turn, leads to the activation of a transcription factor called *ATF4* and an increase in the production of a protein called C/EBP-homologous protein (CHOP), which is involved in the regulation of apoptosis in response to ER stress, as well as other proteins like *DNAJC3*, *HYOU1*, and *CALR*. Besides, the discovered *HYOU1* and *HSPA1A* (in the *HSPA1B* family) and *DNAJB11, CALR, ERP29, GANAB, HSP90B1, HSPA5, LMAN1, PDIA4* and *TXNDC5* were involved in the endoplasmic reticulum (ER) stress [21]. The analysis of how proteins interact with each other revealed certain genes, such as *PTBP1*, *NUP98*, and *HYOU1*, that are linked to breast cancer brain metastasis [4]. *HYOU1* plays a role in supporting the growth, spreading, and metabolic activity of papillary thyroid cancer by increasing the stability of *LDHB* mRNA [77].

Apart from its significant protective function in the formation and progression of tumors, *HYOU1* can be a promising target for treating cancer. It may be an immune-stimulating additive because it can trigger an antitumor immune response. Additionally, it can be a molecular target for treating various endoplasmic reticulum-related ailments [63]. The study [43] found that the secretion of certain substances in response to a communication between lung cancer cells and endothelial cells (ECs) led to an increase in the expression of *HYOU1* in lung cancer spheroids. Additionally, direct interaction

between ECs and lung cancer cells caused an upregulation of *HYOU1* in multicellular tumor spheroids (MCTSs). When inhibiting *HYOU1* expression, it reduced the malignant behavior and stemness of the cancer cells, facilitated apoptosis, and made the MCTSs more sensitive to chemotherapy drugs in lung cancer.

ETS2 is responsible for producing a transcription factor that controls the activity of genes related to both development and apoptosis. The protein it produces is not only a proto-oncogene but has also been found to play a role in regulating telomerase. A non-functional copy of this gene, known as a pseudogene, is also on the X chromosome. Due to alternative splicing, various transcript variants of this gene are generated. The transcription factor *ETS2* controls the expression of genes responsible for various biological processes such as development, differentiation, angiogenesis, proliferation, and apoptosis. The transcription factor *ETS2* has been shown to downregulate the expression of cytokine genes in resting T-cells. The research [20] have investigated whether *ETS2* also regulates the expression of lymphotropic factors (LFs) that are involved in T-cell activation/differentiation and the kinase *CDK10*, which controls Ets-2 degradation and repression activity. In vitro experiments demonstrated that Ets-2 overexpression increased the expression of certain LFs while decreasing *CDK10* levels in both stimulated and unstimulated T-cells. Cyclin-dependent kinase 10 (CDK10) is a serine/threonine kinase related to CDC2 and plays a crucial role in various cellular processes such as cell proliferation, regulation of transcription, and cell cycle regulation. CDK10 has been identified as a candidate tumor suppressor in hepatocellular carcinoma, biliary tract cancers, and gastric cancer, but as a candidate oncogene in colorectal cancer (CRC) [6]. A study on CDK10's role in colorectal cancer revealed that it enhances cell growth, reduces chemosensitivity, and inhibits apoptosis by increasing the expression of *BCL-2* [80]. This effect depends on its kinase activity, as colorectal cancer cell lines with a kinase-defective mutation exhibit an exaggerated apoptotic response and reduced proliferating capacity.CDK10 is a serine/threonine kinase that regulates various cellular processes. It is a candidate tumor suppressor in hepatocellular carcinoma, biliary tract cancers, and gastric cancer but an oncogene in colorectal cancer (CRC). CDK10 promotes cell growth, reduces chemosensitivity, and inhibits apoptosis by increasing the expression of *BCL-2* in colorectal cancer. Kinase-defective mutations exaggerate apoptotic response and reduce proliferating capacity. The relation between the identified-genetic biomarker *ETS2* and *CDK10* could be used to develop new therapeutic applications of cancer treatment.

The relationship between *CELF1* and *ETS2* in colorectal cancer (CRC) and chemoresistance to oxaliplatin (L-OHP) is studied in [76]. *CELF1* was overexpressed in human CRC tissues and positively correlated with *ETS2* expression. Overexpression of *CELF1* increased CRC cell proliferation, migration, invasion, and L-OHP resistance, while knockdown of *CELF1* improved the response of CRC cells to L-OHP. Similarly, overexpression of *ETS2* increased malignant behavior and L-OHP resistance in CRC cells. The study concluded that *CELF1* regulates *ETS2*, resulting in CRC tumorigenesis and L-OHP resistance, and may be a promising target for overcoming chemoresistance in CRC.

GPR116 or *ADGRF5* is the probable G-Protein coupled receptor 116. *GPR116* has been reported to be involved in cancer progression and predicts poor prognosis in other types of cancer. The study [91] shows that *GPR116* expression is upregulated in gastric

cancer (GC) tissues and is positively correlated with tumor invasion and poor prognosis. *GPR116* may be a novel prognostic marker and a potential therapeutic target for GC treatment. mRNA and protein expression of *GPR116* in GC tissues and found that it was significantly upregulated, positively correlated with tumor node metastasis (TNM) staging and tumor invasion, and contributed to poor overall survival in GC patients [37]. *GPR116* overexpression was also found to be an independent prognostic indicator in GC patients. Enrichment analysis revealed that *GPR116* co-expression genes were mainly involved in various pathways.

The effect of *GPR116* receptor on NK cells concerning pancreatic cancer is studied in [32], which found that *GPR116* mice were able to efficiently eliminate pancreatic cancer through enhancing the proportion and function of NK cells in the tumor. The expression of *GPR116* receptor decreased upon NK cells' activation, and *GPR116* with NK cells showed higher cytotoxicity and antitumor activity in vitro and in vivo. Downregulation of *GPR116* receptor also promoted the antitumor activity of NKG2D-CAR-NK92 cells against pancreatic cancer. These findings suggest that downregulation of *GPR116* receptor could enhance the antitumor efficiency of CAR NK cell therapy.

Pathway 2: CTLA4-CD8A-CD8B

The 26 S proteasome (*PSMD2*) is a complex enzyme comprising a 20 S core and a 19 S regulator arranged in a precise structure. The 20 S core consists of four rings with 28 different subunits. Two rings contain seven alpha subunits each, while the others contain seven beta subunits each. The 19 S regulator consists of a base with six ATPase subunits, two non-ATPase subunits, and a lid with up to ten non-ATPase subunits. Proteasomes are found in high concentrations throughout eukaryotic cells and break down peptides through an ATP/ubiquitin-dependent process outside lysosomes. The immunoproteasome, a modified version, plays a critical role in processing class I MHC peptides. This gene codes for one of the non-ATPase subunits in the lid of the 19 S regulator. Besides its involvement in proteasome function, this subunit may also participate in the TNF signaling pathway as it interacts with the tumor necrosis factor type 1 receptor. A non-functional copy of this gene has been found on chromosome 1. Multiple transcript variants of this gene are produced through alternative splicing. *PSMD2* and *PSMD8* were significantly over-expressed in bladder urothelial carcinoma (BLCA) more than other cancers [69]. Besides, *PSMD8* with *AUNIP*, *FANCI*, *LASP1*, and *XPO5* are potential targets for the creation of an mRNA vaccine to combat mesothelioma [89].

PAK1 is responsible for producing a member of the PAK protein family, which are serine/threonine p21-activating kinases. PAK proteins connect RhoGTPases to reorganize the cytoskeleton and nuclear signaling. They act as targets for small GTP-binding proteins such as CDC42 and RAC. This particular family member specifically regulates the movement and shape of cells. Different isoforms of this gene have been identified through alternative splicing, resulting in various transcript variants.

Regarding its mechanism, ipomoea batatas polysaccharides specifically encourage the degradation of *PAK1* through ubiquitination and inhibit its downstream Akt1/mTOR signaling pathway, thereby resulting in an increased level of autophagic flux [15]. *PAK1* is a serine/threonine kinase gene overexpressed in some human breast carcinomas with poor prognosis, and aberrant *PAK1* expression is an early event in the development of

some breast cancers [26]. The structure of the actin cytoskeleton and protrusions in SW620 cells is related to their ability to move. Ce6-PDT treatment inhibits the migration of SW620 cells by reducing the activity of the *RAC1*/*PAK1*/*LIMK1*/cofilin signaling pathway, and this inhibition is improved by decreasing the expression of the *RAC1* gene [82].

The increased presence of *WLS* (Wnt Ligand Secretion Mediato) is an important indication of a negative outcome in breast cancer. It may have a vital function in the hormone receptor-positive (HR+) subtype [90]. Reducing the expression of *SNHG17* in lung adenocarcinoma cells hindered cell growth, migration, and invasion while increasing apoptosis. *SNHG17* acted as a sponge for miR-485-5p, resulting in increased expression of WLS. Therefore, *SNHG17* accelerates lung adenocarcinoma progression by upregulating *WLS* expression through sponging miR-485-5p [44].

NDUFS5 belongs to the NADH dehydrogenase (ubiquinone) iron-sulfur protein family. The protein it encodes is a component of the NADH - ubiquinone oxidoreductase (complex I), the initial enzyme complex in the electron transport chain situated in the inner membrane of mitochondria. Through alternative splicing, multiple transcript variants of this gene are generated. Additionally, pseudogenes of this gene have been discovered on chromosomes 1, 4, and 17.

Machine learning algorithms have been used to classify some cancer types, but not lung adenocarcinoma, in which *NDUFS5, P2RY2, PRPF18, CCL24, ZNF813, MYL6, FLJ41941, POU5F1B, and SUV420H1* were associated with alive without disease [23]. The study identified *MACF1* with *FTSJ3*, *STAT1*, *STX2*, *CDX2* and *RASSF4* that can be used as a signature to predict the overall survival of pancreatic cancer patients [84]. The poor response of low-grade serous ovarian carcinoma (LGSOC) to chemotherapy calls for a thorough genomic analysis to identify new treatment options. This study, 71 LGSOC samples were analyzed for 127 candidate genes using whole exome sequencing and immunohistochemistry to assess key protein expression. Mutations in *KRAS*, *BRAF*, and *NRAS* genes were found in 47% of cases. Several new genetic biomarkers were identified, including *USP9X*, *MACF1*, *ARID1A*, *NF2*, *DOT1L*, and *ASH1L* [17]. To improve the treatment of glioblastomas and enhance patient survival, *MACF1* can be used as a specific diagnostic marker that enhances the effectiveness of radiation therapy while minimizing damage to normal tissues [13]. This approach could potentially lead to the development of new combination radiation therapies that target translational regulatory processes, which are often involved in poor patient outcomes. Another study presented data indicating that reduced *MACF1* expression inhibited melanoma metastasis in mice by blocking the epithelial-to-mesenchymal transition process. Therefore, *MACF1* could be a potential target for melanoma treatment [78].

MACF1 is responsible for producing a substantial protein that consists of multiple spectrin and leucine-rich repeat (LRR) domains. The encoded protein belongs to a family of proteins that connect various cytoskeletal components. Specifically, this protein plays a role in enabling interactions between actin and microtubules at the outer edges of cells, and it also connects the microtubule network to cellular junctions. Multiple transcript variants of this gene are produced through alternative splicing, although the complete structure of some of these variants has yet to be determined [31]. A study included 695 patients with hepatocellular carcinoma (HCC), divided into a training group of 495

patients and a validation group of 200 patients [35]. A nomogram was developed using T stage, age, and the mutation status of *DOCK2*, *EYS*, *MACF1*, and *TP53*. The nomogram was found to have good accuracy in predicting outcomes and was consistent with the actual data. The study also found that T-cell exclusion may be a potential mechanism for malignant progression in the high-risk group. In contrast, the low-risk group may benefit from immunotherapy and *CTLA4* blocker treatment. In conclusion, the study developed a nomogram based on mutant genes and clinical parameters and identified the underlying association between these risk factors and immune-related processes.

Pathway 3: CTLA4-CD2

Fucoxanthin is a natural pigment present in brown seaweeds, and its derivative, fucoxanthin (FxOH), has been shown to effectively induce apoptosis (programmed cell death) in various cancer cells. The role of Chloride intracellular channel 4 (*CLIC4*), which plays a crucial role in cancer development and apoptosis, in FxOH-induced apoptosis was also investigated [85]. Treatment with FxOH induced apoptosis in human CRC DLD-1 cells. FxOH treatment downregulated *CLIC4*, integrin beta1, *NHERF2*, and *pSMAD2* (Ser(465/467)) compared to control cells, without affecting *RAB35* expression. *CLIC4* knockdown suppressed cell growth and apoptosis, and apoptosis induction by FxOH was reduced with *CLIC4* knockdown. The expression levels of *CLIC4* and *GAS2L1* were found to be higher in circulating tumor cells (CTCs) from pancreatic cancer patients compared to peripheral blood mononuclear cells [93]. Besides, the overexpression of *CLIC4* was associated with unfavorable outcomes in multiple cohorts of CN-AML patients [33].

The role of actin-binding proteins, including profiling, fascin, and ezrin, in the metastasis of non-small cell lung cancer (NSCLC) [39]. The study collected tumor and adjacent normal lung tissue samples from 46 NSCLC patients and used real-time PCR and Western blotting to determine the levels of *PFN1*, *FSCN1*, and *EZR* mRNAs and proteins. The results showed that patients with lymphatic metastasis had higher expression levels of the profilin, fascin, and ezrin mRNAs and profilin and fascin proteins. In contrast, mRNA and protein expression levels increased in patients with distant metastasis. The activation of AKT signaling in the progression of colorectal cancer (CRC) can be influenced by the lncHCP5/miR-299-3p/*PFN1* [5]. The loss of *PFN1* leads to the activation of several signaling pathways, including AKT, NF-(k)B, and WNT. On the other hand, overexpression of PFN1 in cells with high levels of SH3BGRL can counteract SH3BGRL-induced metastasis and tumor growth by upregulating *PTEN* and inhibiting the PI3K-AKT pathway [88].

PPM1G codes for a large protein that contains spectrin and leucine-rich repeat (LRR) domains. It belongs to a family of proteins that act as bridges between different cytoskeletal elements. Specifically, this protein facilitates the interaction between actin and microtubules at the periphery of cells and links the microtubule network to cellular junctions. The level of *PPM1G* expression in LIHC may be influenced by promoter methylation, CNVs, and kinases and could be linked to immune infiltration. High *PPM1G* expression was found to be related to mRNA splicing and the cell cycle according to GO terms. These findings suggest that *PPM1G* could be a prognostic indicator for liver hepatocellular carcinoma patients and may play a role in the tumor immune

microenvironment [45]. Besides, the irc-*PGAP3* plays a significant role in the growth and advancement of triple-negative breast cancer (TNBC), thus making it a potential target for the treatment of TNBC patients.

Pathway 4: CTLA4-CD2-CD48-CD53-CD58-CD84

MT1G (Metallothionein 1 G) and *MT1H* have the potential to suppress tumor growth and are regulated by DNA methylation in their promoter regions. In addition, they are associated with serum copper levels and may be linked to the survival rate of patients with hepatocellular carcinoma [75]. Besides, *MT1G*, *CXCL8*, *IL1B*, *CXCL5*, *CXCL11*, and *GZMB* are over-expressed in colorectal cancer tissues compared to normal tissues [49]. Three genes (*SLC7A11*, *HMOX1*, and *MT1G*) were identified as differentially expressed genes (DEGs) associated with renal cancer prognosis using survival analysis screening [18]. *SLC7A11* and *HMOX1* were found to be upregulated in renal cancer tissues, while *MT1G* was downregulated. The combination of receiver operating characteristic (ROC) curves, Kaplan-Meier analysis, and Cox regression analysis revealed that high expression of *SLC7A11* was a prognostic risk factor for four different types of renal cancers, low expression of *HMOX1* was a poor prognostic marker for patients, and increased expression of *MT1G* increased the prognostic risk for three additional classes of renal cancer patients, except for those with renal papillary cell carcinoma.

FAM107A (Family With Sequence Similarity 107 Member A) with *ADAM12*, *CEP55*, *LRFN4*, *INHBA*, *ADH1B*, *DPT*, and *LOC100506388* were analyzed and evaluated as potential prognostic genes for gastric cancer [34]. Among these genes, *LRFN4*, *DPT*, and *LOC100506388* were identified as having a potential prognostic role in gastric cancer, as determined through a nomogram. Besides, the interaction pairs of *HCG22/EGOT-hsa-miR-1275-FAM107A* and *HCG22/EGOT-hsa-miR-1246*-Glycerol-3-phosphate dehydrogenase 1 are likely to have a significant role in laryngeal squamous cell carcinoma.

LCN2 produces a protein classified as a lipocalin family member. Lipocalins are known for their ability to transport small hydrophobic molecules like lipids, steroid hormones, and retinoids. The specific protein encoded by this gene is called neutrophil gelatinase-associated lipocalin (NGAL), and it plays a significant role in innate immunity. NGAL sequesters iron-containing siderophores, which helps limit bacterial growth and infection [31]. Besides, *LCN2* is an innate immune protein that regulates immune responses by promoting sterile inflammation. *LCN2* is a biomarker associated with radioresistance and recurrence in nasopharyngeal carcinoma (NPC) [87]. *LCN2* expression was upregulated in radioresistant NPC tissues and associated with NPC recurrence. Knocking down *LCN2* enhances the radiosensitivity of NPC cells, while ectopic expression of *LCN2* confers additional radioresistance. *LCN2* may interact with *HIF-1A* and facilitate the development of a radioresistant phenotype. *LCN2* is a promising target for predicting and overcoming radioresistance in NPC. Moreover, the downregulation of the immune response, influenced by specific metastasis-evaluation genes (*BAMBI*, *F13A1*, *LCN2*) and their associated immune-prognostic genes (*SLIT2*, *CDKN2A*, *CLU*), was found to increase the risk of post-operative recurrence [46]. Higher *LCN2* expression was associated with poor clinical outcomes and correlated with increased infiltration of various immune cells. *LCN2* may serve as a genetic biomarker for immune infiltration and poor prognosis in cancers, suggesting potential therapeutic targets for cancer treatment [83].

Discussion

Complexity in biomarker identification for *CTLA4* activation pathway

Identifying the T cell gene CTLA4 and its genetic biomarker presents unique challenges, particularly when detecting certain molecular biomarkers expressed in the bloodstream. The detection of CTLA4 in T cells is complex due to several factors. First, CTLA4 is predominantly found in intracellular compartments before activation and only becomes increasingly detectable on the cell surface upon activation. This necessitates precise timing for effective detection. Second, the proportion of CTLA4-positive T-cell subgroups in the peripheral blood and tumor tissues could be higher, making their detection difficult and costly. Despite the challenges, identifying genetic biomarkers associated with CTLA4 has several advantages over detecting molecular biomarkers in the bloodstream. Genetic biomarkers can provide information about genetic susceptibility, genetic responses to environmental exposures, subclinical or clinical disease markers, or indicators of response to therapy [19]. They can help identify high-risk individuals reliably and promptly so that they can either be treated before the onset of the disease or as soon as possible. When a genetic biomarker is identified in a cancer through molecular or genetic testing, it tells the physician what makes the cancer grow and thrive. That information allows physicians to decide the most effective treatment for the patient [51].

Model adaptation for further applications

The further improvement of the proposed quantum neural network can consider the intricate nature of epigenetic modifications, which is of utmost importance. Epigenetic modifications are chemical alterations that occur to the DNA molecule itself or to the proteins tightly bound to it, and they play a crucial role in regulating gene expression and cellular differentiation. These modifications can include phosphorylation, acetylation, or methylation of various amino acids, and they not only add a layer of complexity but also enrich the genomic landscape with many potential biomarkers. To enhance the accuracy and reliability of our quantum neural network model, we need to integrate and analyze these diverse epigenetic markers. Doing so can unravel complex biological interactions and pathways associated with immune responses. This will allow us to identify novel and clinically relevant biomarkers for immunotherapy that may not be discovered through traditional methods.

Furthermore, addressing the complexities of epigenetic modifications will facilitate a more nuanced understanding of the model's adaptability. This understanding will offer insights into its potential applications and limitations in real-world clinical settings. By taking a comprehensive approach, we ensure that our model is robust and versatile, accommodating the vast array of biomarkers introduced by epigenetic modifications. In turn, this contributes to the advancement of personalized medicine in immuno-therapy, where we can tailor treatments to an individual's unique genetic makeup, epigenetic modifications, and immune system response.

Conclusion

To this end, a new framework based on the quantum AI model, used for genetic biomarker discovery in biomedical research ("Method" section), has been introduced. The proof of concept is demonstrated in four targeted pathways associating with therapeutic-target *CTLA4* in Sect. 4. Our model found clinical-relevant and notably potential biomarkers/targets for cancer treatment, which are extensively validated through statistical methods and literature mining (Sect. 5.2).

We suggest several research directions that can be extended from the study. First, deploying the proposed quantum AI models on real quantum computers is worth investigating in the future, in which the effect of noise should be addressed toward the model's efficiency and effectiveness. Second, extension to other pathway activation is possible as the proposed algorithm is generalized. Finally, in *vivo* and in *vitro*, validations of the discovered in *silico* biomarkers will translate the findings to therapeutic solutions for cancer treatment and prevention.

Abbreviations

AI	Artificial intelligence
ML	Machine learning
QNN	Quantum neural networks
QML	Quantum machine learning
mRMR	Maximum relevance-minimum redundancy
GSCORE	© Gene score
temp	Temperature
SMBO	Sequential model-based optimization
CNV	Copy number variation
TCGA	The cancer genome atlats
Genetic Biomarker(s)	Biomarker(s) identified from genomic data

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05755-0.

Additional file 1. Supplementary Materials.

Acknowledgements
The author would like to thank colleagues for stimulating discussion.

Author contributions
Phuong-Nam Nguyen conceptualized the algorithm and performed numerical analysis.

Availability of data and materials
The implementation is made open source at: https://github.com/namnguyen0510/Biomarker-Discovery-with-Quantum-Neural-Networks. The particular TCGA data utilized in the analyses discussed in this context are stored on Zenodo: - Gene expression (identifier: EB++AdjustPANCAN_IlluminaHiSeq_RNASeqV2.geneExp): https://figshare.com/articles/dataset/TCGA_PanCanAtlas_Gene_Expression_Data/6146519 - Copy number data (identifier: pancan_GISTIC_threshold): https://figshare.com/articles/dataset/TCGA_PanCanAtlas_Copy_Number_Data/6144122 The processed data and the experimental history are available at: https://drive.google.com/drive/folders/1QsL0N5E2xppWK2pmvfLI621PeXV2OQUA?usp=share_link. The evaluation for statistical significance (Sect. 5.1) by cBioPortal is reported at: https://drive.google.com/drive/folders/1ubbLEg35Pz9i_Wjr4uKvMuS1KNw5rrsS?usp=sharing. The literature mining for clinical significance (Sect. 5.2) is reported at: https://drive.google.com/drive/folders/1NHDuJQoah8ZmlP8zWlPD6aYwwWTxqoYe?usp=sharing.

Declarations

Competing interests
The author declares that they have no known competing financial interests reported in this paper.

Received: 25 December 2023 Accepted: 20 March 2024
Published online: 12 April 2024

References

1. Acharjee A, Larkman J, Xu Y, Cardoso VR, Gkoutos GV. A random forest based biomarker discovery and power analysis framework for diagnostics research. BMC Med Genom. 2020;13(1):1–14.
2. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 2623–2631;2019.
3. Al Abir F, Shovan S, Hasan MAM, Sayeed A, Shin J. Biomarker identification by reversing the learning mechanism of an autoencoder and recursive feature elimination. Mol Omics. 2022;18(7):652–61.
4. An M, Zang X, Wang J, Kang J, Tan X, Fu B. Comprehensive analysis of differentially expressed long noncoding rnas, mirnas and mrnas in breast cancer brain metastasis. Epigenomics. 2021;13(14):1113–28.
5. Bai N, Ma Y, Zhao J, Li B. Knockdown of lncrna hcp5 suppresses the progression of colorectal cancer by mir-299-3p/pfn1/akt axis. Cancer Manag Res. 2020;12:4747.
6. Bazzi ZA, Tai IT. Cdk10 in gastrointestinal cancers: dual roles as a tumor suppressor and oncogene. Front Oncol. 2021;11: 655479.
7. Benedetti M, Lloyd E, Sack S, Fiorentini M. Parameterized quantum circuits as machine learning models. Quant Sci Technol. 2019;4(4): 043001.
8. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013;35(8):1798–828.
9. Bergholm V, Izaac J, Schuld M, Gogolin C, Alam MS, Ahmed S, Arrazola JM, Blank C, Delgado A, Jahangiri S, et al. Pennylane: Automatic differentiation of hybrid quantum-classical computations. 2018; arXiv:1811.04968.
10. Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. Adv Neural Inf Process Syst. 2011;24:1.
11. Biomarker. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/biomarker, n.d. Accessed 10 May 2023.
12. Bolton KL, Ptashkin RN, Gao T, Braunstein L, Devlin SM, Kelly D, Patel M, Berthon A, Syed A, Yabe M, et al. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. Nat Genet. 2020;52(11):1219–26.
13. Bonner K, Borlay D, Kutten O, Quick QA. Inhibition of the spectraplakin protein microtubule actin crosslinking factor 1 sensitizes glioblastomas to radiation. Brain Tumor Res Treatm. 2020;8(1):43.
14. Borràs DM, Verbandt S, Ausserhofer M, Sturm G, Lim J, Verge GA, Vanmeerbeek I, Laureano RS, Govaerts J, Sprooten J, et al. Single cell dynamics of tumor specificity vs bystander activity in cd8+ t cells define the diverse immune landscapes in colorectal cancer. Cell Discov. 2023;9(1):114.
15. Bu H, Tan S, Yuan B, Huang X, Jiang J, Wu Y, Jiang J, Li R. Therapeutic potential of ibp as an autophagy inducer for treating lung cancer via blocking pak1/akt/mtor signaling. Mol Therapy-Oncolyt. 2021;20:82–93.
16. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, et al. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2(5):401–4.
17. Cheasley D, Nigam A, Zethoven M, Hunter S, Etemadmoghadam D, Semple T, Allan P, Carey MS, Fernandez ML, Dawson A, et al. Genomic analysis of low-grade serous ovarian carcinoma to identify key drivers and therapeutic vulnerabilities. J Pathol. 2021;253(1):41–54.
18. Chen L, Wang C, Wang Y, Hong T, Zhang G, Cui X. Functions, roles, and biological processes of ferroptosis-related genes in renal cancer: a pan-renal cancer analysis. Front Oncol. 2022;11: 697697.
19. Chen X-H, Huang S, Kerr D, et al. Biomarkers in clinical medicine. IARC Sci Publ. 2011;163:303–22.
20. Davoulou P, Aggeletopoulou I, Panagoulias I, Georgakopoulos T, Mouzaki A. Transcription factor ets-2 regulates the expression of key lymphotropic factors. Mol Biol Rep. 2020;47:7871–81.
21. de Seny D, Bianchi E, Baiwir D, Cobraiville G, Collin C, Deliège M, Kaiser M-J, Mazzucchelli G, Hauzeur J-P, Delvenne P, et al. Proteins involved in the endoplasmic reticulum stress are modulated in synovitis of osteoarthritis, chronic pyrophosphate arthropathy and rheumatoid arthritis, and correlate with the histological inflammatory score. Sci Rep. 2020;10(1):14159.
22. DeGroat W, Mendhe D, Bhusari A, Abdelhalim H, Zeeshan S, Ahmed Z. Intelligenes: a novel machine learning pipeline for biomarker discovery and predictive analysis using multi-genomic profiles. Bioinformatics. 2023;39(12):btad755.
23. Deng F, Shen L, Wang H, Zhang L. Classify multicategory outcome in patients with lung adenocarcinoma using clinical, transcriptomic and clinico-transcriptomic data: machine learning versus multinomial models. Am J Cancer Res. 2020;10(12):4624.
24. Dhillon A, Singh A, Bhalla VK. A systematic review on biomarker identification for cancer diagnosis and prognosis in multi-omics: from computational needs to machine learning and deep learning. Arch Comput Methods Eng. 2023;30(2):917–49.
25. Diaz-Uriarte R, Gómez de Lope E, Giugno R, Fröhlich H, Nazarov PV, Nepomuceno-Chamorro IA, Rauschenberger A, Glaab E. Ten quick tips for biomarker discovery and validation analyses using machine learning. PLoS Comput Biol. 2022;18(8): e1010357.
26. Duderstadt EL, McQuaide SA, Sanders MA, Samuelson DJ. Chemical carcinogen-induced rat mammary carcinogenesis is a potential model of p21-activated kinase positive female breast cancer. Physiol Genom. 2021;53(2):61–8.
27. Fang Z, Liu Y, Wang Y, Zhang X, Chen Y, Cai C, Lin Y, Han Y, Wang Z, Zeng S, et al. Deep learning predicts biomarker status and discovers related histomorphology characteristics for low-grade glioma. 2023; arXiv:2310.07464.
28. Farhi E, Neven H. Classification with quantum neural networks on near term processors. 2018; arXiv:1802.06002.
29. Feldhaus A, Evans L, Sutherland R, Jones L. A cd2/cd28 chimeric receptor triggers the cd28 signaling pathway in ctll. 2 cells. Gene Therapy. 1997;4(8):833–8.
30. Green JM, Karpitskiy V, Kimzey SL, Shaw AS. Coordinate regulation of t cell activation by cd2 and cd28. J Immunol. 2000;164(7):3591–5.

31. Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, Ainscough BJ, Ramirez CA, Rieke DT, Kujan L, et al. Civic is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. 2017;49(2):170–4.
32. Guo D, Jin C, Gao Y, Lin H, Zhang L, Zhou Y, Yao J, Duan Y, Ren Y, Hui X, et al. Gpr116 receptor regulates the anti-tumor function of nk cells via hif1α/nf-κb signaling pathway as a potential immune checkpoint. 2023.
33. Huang S, Huang Z, Chen P, Feng C. Aberrant chloride intracellular channel 4 expression is associated with adverse outcome in cytogenetically normal acute myeloid leukemia. Front Oncol. 2020;10:1648.
34. Huang S, Ma L, Lan B, Liu N, Nong W, Huang Z. Comprehensive analysis of prognostic genes in gastric cancer. Aging (Albany NY). 2021;13(20):23637.
35. Huang T, Yan T, Chen G, Zhang C. Development and validation of a gene mutation-associated nomogram for hepatocellular carcinoma patients from four countries. Front Genet. 2021;1825.
36. Hyman DM, Piha-Paul SA, Won H, Rodon J, Saura C, Shapiro GI, Juric D, Quinn DI, Moreno V, Doger B, et al. Her kinase inhibition in patients with her2-and her3-mutant cancers. Nature. 2018;554(7691):189–94.
37. Kang H, Fichna J, Matlawska-Wasowska K, Jacenik D. The expression pattern of adhesion g protein-coupled receptor f5 is related to cell adhesion and metastatic pathways in colorectal cancer-comprehensive study based on in silico analysis. Cells. 2022;11(23):3876.
38. Killoran N, Bromley TR, Arrazola JM, Schuld M, Quesada N, Lloyd S. Continuous-variable quantum neural networks. Phys Rev Res. 2019;1(3): 033063.
39. Kolegova E, Kakurina G, Kostromitskiy D, Dobrodeev AY, Kondakova I. Increases in mrna and protein levels of the genes for the actin-binding proteins profilin, fascin, and ezrin promote metastasis in non-small cell lung cancer. Mol Biol. 2020;54:249–55.
40. Kristensen LK, Christensen C, Alfsen MZ, Cold S, Nielsen CH, Kjaer A. Monitoring cd8a+ t cell responses to radiotherapy and ctla-4 blockade using [64 cu] nota-cd8a pet imaging. Mol Imag Biol. 2020;22:1021–30.
41. O. R. N. Laboratory. Scientists use quantum biology, ai to sharpen genome editing tool; 2023.
42. Lau B, Emani PS, Chapman J, Yao L, Lam T, Merrill P, Warrell J, Gerstein MB, Lam HY. Insights from incorporating quantum computing into drug design workflows. Bioinformatics. 2023;39(1):789.
43. Lee M, Song Y, Choi I, Lee S-Y, Kim S, Kim S-H, Kim J, Seo HR. Expression of hyou1 via reciprocal crosstalk between nsclc cells and huvecs control cancer progression and chemoresistance in tumor spheroids. Mol Cells. 2021;44(1):50.
44. Li W, Zheng Y, Mao B, Wang F, Zhong Y, Cheng D. Snhg17 upregulates wls expression to accelerate lung adenocarcinoma progression by sponging mir-485-5p. Biochem Biophys Res Commun. 2020;533(4):1435–41.
45. Lin Y-R, Yang W-J, Yang G-W. Prognostic and immunological potential of ppm1g in hepatocellular carcinoma. Aging (Albany NY). 2021;13(9):12929.
46. Luo Z, Chen X, Zhang Y, Huang Z, Zhao H, Zhao J, Li Z, Zhou J, Liu J, Cai J, et al. Development of a metastasis-related immune prognostic model of metastatic colorectal cancer and its usefulness to immunotherapy. Front Cell Dev Biol. 2021;8: 577125.
47. Marchetti L, Nifosì R, Martelli PL, Da Pozzo E, Cappello V, Banterle F, Trincavelli ML, Martini C, D'Elia M. Quantum computing algorithms: getting closer to critical problems in computational biology. Briefings Bioinf. 2022;23(6):437.
48. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943;5(4):115–33.
49. Meng J, Su R, Liao Y, Li Y, Li L. Identification of 10 hub genes related to the progression of colorectal cancer by co-expression analysis. PeerJ. 2020;8: e9633.
50. Miao D, Margolis CA, Vokes NI, Liu D, Taylor-Weiner A, Wankowicz SM, Adeegbe D, Keliher D, Schilling B, Tracy A, et al. Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors. Nat Genet. 2018;50(9):1271–81.
51. National Cancer Institute. Biomarker testing for cancer treatment.
52. Nguyen B, Fong C, Luthra A, Smith SA, DiNatale RG, Nandakumar S, Walch H, Chatila WK, Madupuri R, Kundra R, et al. Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. Cell. 2022;185(3):563–75.
53. Nguyen N, Chang JM. Csnas: contrastive self-supervised learning neural architecture search via sequential model-based optimization. IEEE Trans Artif Intell. 2021;3(4):609–24.
54. Nguyen N, Chen K-C. Bayesian quantum neural networks. IEEE Access. 2022.
55. Nguyen P-N. The duality game: a quantum algorithm for body dynamics modeling. Quant Inf Process. 2024;23(1):21.
56. Nguyen XV, Chan J, Romano S, Bailey J. Effective global approaches for mutual information based feature selection. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014; pp. 512–521.
57. Nielsen MA, Chuang I. Quantum computation and quantum information, 2002.
58. Pal S, Bhattacharya M, Dash S, Lee S-S, Chakraborty C. Future potential of quantum computing and simulations in biological science. Mol Biotechnol. 2023;1:1–18.
59. Pal S, Bhattacharya M, Lee S-S, Chakraborty C. Quantum computing in the next-generation computational biology landscape: From protein folding to molecular dynamics. Mol Biotechnol. 2024;66(2):163–78.
60. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al. Pytorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst. 2019;32:1.
61. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
62. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27(8):1226–38.
63. Rao S, Oyang L, Liang J, Yi P, Han Y, Luo X, Xia L, Lin J, Tan S, Hu J, et al. Biological function of hyou1 in tumors and other diseases. Onco Targets Ther. 2021;14:1727.

64. Robinson DR, Wu Y-M, Lonigro RJ, Vats P, Cobain E, Everett J, Cao X, Rabban E, Kumar-Sinha C, Raymond V, et al. Integrative clinical genomics of metastatic cancer. Nature. 2017;548(7667):297–303.
65. Rodriguez-Lujan I, Elkan C, Santa Cruz Fernández C, Huerta R, et al. Quadratic programming feature selection. J Mach Learn Res. 2010.
66. Romero J, Aspuru-Guzik A. Variational quantum generators: generative adversarial quantum machine learning for continuous distributions. Adv Quant Technol. 2021;4(1):2000003.
67. Rosen EY, Goldman DA, Hechtman JF, Benayed R, Schram AM, Cocco E, Shifman S, Gong Y, Kundra R, Solomon JP, et al. Trk fusions are enriched in cancers with uncommon histologies and the absence of canonical driver mutationslandscape and outcome of trk fusion-positive cancers. Clin Cancer Res. 2020;26(7):1624–32.
68. Rowshanravan B, Halliday N, Sansom DM. Ctla-4: a moving target in immunotherapy. Blood J Am Soc Hematol. 2018;131(1):58–67.
69. Salah Fararjeh A, Al-Khader A, Al-Saleem M, Abu Qauod R. The prognostic significance of proteasome 26s subunit, non-atpase (psmd) genes for bladder urothelial carcinoma patients. Cancer Inf. 2021;20:11769351211067692.
70. Samstein RM, Lee C-H, Shoushtari AN, Hellmann MD, Shen R, Janjigian YY, Barron DA, Zehir A, Jordan EJ, Omuro A, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. Nat Genet. 2019;51(2):202–6.
71. Schuld M, Killoran N. Quantum machine learning in feature hilbert spaces. Phys Rev Lett. 2019;122(4): 040504.
72. Schuld M, Killoran N. Is quantum advantage the right goal for quantum machine learning? 2022; arXiv:2203.01340.
73. Schuld M, Sweke R, Meyer JJ. Effect of data encoding on the expressive power of variational quantum-machine-learning models. Phys Rev A. 2021;103(3): 032430.
74. Skånland SS, Taskén K. Carboxyl-terminal src kinase binds cd28 upon activation and mutes downstream signaling. J Immunol. 2019;203(4):1055–63.
75. Udali S, De Santis D, Mazzi F, Moruzzi S, Ruzzenente A, Castagna A, Pattini P, Beschin G, Franceschi A, Guglielmi A, et al. Trace elements status and metallothioneins dna methylation influence human hepatocellular carcinoma survival rate. Front Oncol. 2021;10: 596040.
76. Wang H, Huang R, Guo W, Qin X, Yang Z, Yuan Z, Wei Y, Mo C, Zeng Z, Luo J, et al. Rna-binding protein celf1 enhances cell migration, invasion, and chemoresistance by targeting ets2 in colorectal cancer. Clin Sci. 2020;134(14):1973–90.
77. Wang J-M, Jiang J-Y, Zhang D-L, Du X, Wu T, Du Z-X. Hyou1 facilitates proliferation, invasion and glycolysis of papillary thyroid cancer via stabilizing ldhb mrna. J Cell Mol Med. 2021;25(10):4814–25.
78. Wang X, Jian X, Dou J, Wei Z, Zhao F. Decreasing microtubule actin cross-linking factor 1 inhibits melanoma metastasis by decreasing epithelial to mesenchymal transition. Cancer Manag Res. 2020;12:663.
79. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45(10):1113–20.
80. Weiswald L-B, Hasan MR, Wong JC, Pasiliao CC, Rahman M, Ren J, Yin Y, Gusscott S, Vacher S, Weng AP, et al. Inactivation of the kinase domain of cdk10 prevents tumor growth in a preclinical model of colorectal cancer, and is accompanied by downregulation of bcl-2. Mol Cancer Ther. 2017;16(10):2292–303.
81. Wu L, Yao H, Chen H, Wang A, Guo K, Gou W, Yu Y, Li X, Yao M, Yuan S, et al. Landscape of somatic alterations in large-scale solid tumors from an Asian population. Nat Commun. 2022;13(1):1–11.
82. Wufuer R, Ma H-X, Luo M-Y, Xu K-Y, Kang L. Downregulation of rac1/pak1/limk1/cofilin signaling pathway in colon cancer sw620 cells treated with chlorin e6 photodynamic therapy. Photodiagn Photodyn Ther. 2021;33: 102143.
83. Xu W-X, Zhang J, Hua Y-T, Yang S-J, Wang D-D, Tang J-H. An integrative pan-cancer analysis revealing lcn2 as an oncogenic immune protein in tumor microenvironment. Front Oncol. 2020;10: 605097.
84. Yang J, Shi W, Zhu S, Yang C. Construction of a 6-gene prognostic signature to assess prognosis of patients with pancreatic cancer. Medicine. 2020;99(37):1.
85. Yokoyama R, Kojima H, Takai R, Ohta T, Maeda H, Miyashita K, Mutoh M, Terasaki M. Effects of clic4 on fucoxanthinol-induced apoptosis in human colorectal cancer cells. Nutr Cancer. 2021;73(5):889–98.
86. Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, Srinivasan P, Gao J, Chakravarty D, Devlin SM, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med. 2017;23(6):703–13.
87. Zhang M-X, Wang L, Zeng L, Tu Z-W. Lcn2 is a potential biomarker for radioresistance and recurrence in nasopharyngeal carcinoma. Front Oncol. 2021;10: 605777.
88. Zhang S, Guo X, Liu X, Zhong Z, Yang S, Wang H. Adaptor sh3bgrl promotes breast cancer metastasis through pfn1 degradation by translational stub1 upregulation. Oncogene. 2021;40(38):5677–90.
89. Zhang S, Li S, Wei Y, Xiong Y, Liu Q, Hu Z, Zeng Z, Tang F, Ouyang Y. Identification of potential antigens for developing mrna vaccine for immunologically cold mesothelioma. Front Cell Dev Biol. 2022;10:1.
90. Zheng D, Jiang C, Yan N, Miao Y, Wang K, Gao G, Jiao Y, Zhang X, He M, Yang Z. Wntless (wls): a prognostic index for progression and patient survival of breast cancer. Onco Targets Ther. 2020;13:12649.
91. Zheng T, Sun M, Liu L, Lan Y, Wang L, Lin F. Gpr116 overexpression correlates with poor prognosis in gastric cancer. Medicine. 2021;100(48):1.
92. Zhou R, Ma Y, Tao Z, Qiu S, Gong Z, Tao L, Zhu Y. Melatonin inhibits glucose-induced apoptosis in osteoblastic cell line through perk-eif2*a*-atf4 pathway. Front Pharmacol. 2020;11: 602307.
93. Zhu L, Kan K-J, Grün JL, Hissa B, Yang C, Győrffy B, Loges S, Reißfelder C, Schölch S. Gas2l1 is a potential biomarker of circulating tumor cells in pancreatic cancer. Cancers. 2020;12(12):3774.

Publisher's Note

Phuong-Nam Nguyen is a current Ph.D. student in Electrical Engineering at the University of South Florida. He earned a Bachelor's degree in Mathematics and Education from the Hanoi University of Education and a Master's in Statistics from the University of South Florida. He is currently a lecturer and research fellow at Phenikaa University. His research interests are applying classical and quantum machine intelligence to biological problems, including oncology research, molecular dynamics modeling, large genome data analysis, and biomarker identification.