UNIVERSIDADE SÃO FRANCISCO
Programa de Pós-Graduação *Stricto Sensu* em Ciências da Saúde

PEDRO HENRIQUE GODOY SANCHES

DESENVOLVIMENTO DE *PIPELINE* PARA ANÁLISES MULTI-ÔMICAS DE CARCINOMA PAPILAR DE TIREOIDE

Bragança Paulista
2026

PEDRO HENRIQUE GODOY SANCHES – R.A.: 202127476

DESENVOLVIMENTO DE *PIPELINE* PARA ANÁLISES MULTI-ÔMICAS DE CARCINOMA PAPILAR DE TIREOIDE

Tese apresentada ao Programa de Pós-Graduação *Stricto Sensu* em Ciências da Saúde da Universidade São Francisco, como requisito para obtenção do Título de Doutor em Ciências da Saúde.

Área de Concentração: Ciências da Saúde.

Orientador: Prof.(a) Dr.(a) Lucas Miguel de Carvalho

Bragança Paulista
2026

SANCHES, Pedro Henrique Godoy. "Desenvolvimento de pipeline para análises multi-ômicas de carcinoma papilar de tireoide". Tese defendida e aprovada no programa de Pós-Graduação *Stricto Sensu* em Ciências da Saúde da Universidade São Francisco em 29 de julho de 2026 pela Banca examinadora constituída pelos membros:

Prof(a). Dr(a). Lucas Miguel de Carvalho
Orientador(a) e Presidente
Universidade São Francisco

Prof(a). Dr(a). Alessandra Sussulini
(por videoconferência)
Universidade Estadual de Campinas

Prof(a). Dr(a). Fernando Augusto de Lima Marson
Universidade São Francisco

Prof(a). Dr(a). Leticia Ferreira Alves
(por videoconferência)
Josep Carreras Leukaemia Research Institute

Prof(a). Dr(a). Tais Mendes de Camargo
Universidade São Francisco

WK 270 Sanches, Pedro Henrique Godoy
S191d Desenvolvimento de *pipeline* para análises multi-ômicas de carcinoma papilar de tireoide / Pedro Henrique Godoy Sanches. -- Bragança Paulista, 2026.
70 p.

Tese (Doutorado) – Programa de Pós-Graduação *Stricto Sensu* em Ciências da Saúde da Universidade São Francisco.
Orientação de: Lucas Miguel de Carvalho

1. Carcinoma papilar de tireoide. 2. Biologia computacional. 3. Metabolômica. 4. Transcriptômica. I. Carvalho, Lucas Miguel de. II. Título.

Sistema de Bibliotecas da Universidade São Francisco – USF
Ficha catalográfica elaborada por: Denise Isabel Arten - CRB-8/5823

AGRADECIMENTOS

Agradeço imensamente a minha família, especialmente pais e avós, por ter compreendido meu gosto por fazer ciência, por ter me educado e ajudado de diferentes formas durante todo meu percurso acadêmico. Eles foram o pilar fundamental que viabilizou a minha dedicação à pesquisa e à ciência.

Manifesto minha mais profunda gratidão aos amigos que tive a honra de conhecer e que, ao longo de toda esta trajetória, tornaram-se presenças indispensáveis em minha vida. Em especial, à Andréia de Melo Porcari, cuja amizade, orientação e apoio incondicional foram de valor imensurável. Sua presença foi, sem dúvida, uma das mais significativas em todo este percurso, contribuindo de maneira decisiva não apenas para a condução desta pesquisa, mas também para o meu amadurecimento pessoal e acadêmico. Agradeço, igualmente, ao Álex Ap. Rosini Silva, cuja amizade levo comigo com imenso apreço, e cuja parceria, especialmente nos momentos mais difíceis, foi fundamental para que eu pudesse prosseguir com este estudo.

Registro também minha sincera gratidão a Danilo C. Oliveira, cuja convivência, solidariedade e amizade enriquecem minha trajetória e permanecem como parte importante da minha história. Não poderia deixar de agradecer, ainda, à Cíntia, cuja presença e apoio igualmente contribuíram de forma relevante para este caminho, especialmente nos últimos meses. A todos eles, meu mais profundo reconhecimento, por sua amizade, seu incentivo e por permanecerem, de diferentes formas, como referências inesquecíveis em minha vida.

Especialmente, à minha companheira de vida, Jaqueline Rodrigues Carreira, expresso minha mais profunda gratidão por sua presença constante, por sua paciência admirável e pela confiança sempre depositada em mim. Sua parceria foi absolutamente essencial ao longo de toda esta trajetória, não apenas pelo companheirismo e pelas valiosas opiniões, mas, sobretudo, por sua disponibilidade incondicional em permanecer ao meu lado em todos os momentos, sejam eles de tranquilidade ou de grande dificuldade. Sua compreensão, seu apoio e sua força foram fundamentais para que eu pudesse seguir com serenidade, equilíbrio e perseverança, tornando este percurso mais leve e significativo.

Ao Prof. Dr. Lucas Miguel de Carvalho, cuja notável dedicação, competência e proatividade têm se destacado de maneira exemplar desde seu início no grupo, e que, com grande generosidade intelectual e humildade, aceitou orientar-me. Registro minha profunda gratidão por seu apoio constante, por sua postura sempre acolhedora nos momentos mais desafiadores e por sua visão

positiva, firme e construtiva, qualidades que não apenas favoreceram o desenvolvimento deste trabalho, mas também inspiram confiança, disciplina e perseverança. Sua condução segura, senso de responsabilidade e compromisso com a excelência foram fundamentais para o percurso até o momento.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão do apoio financeiro, por meio dos processos de número #88887.639447/2021-00, #88887.824612/2023-00 e também pelo processo #88881.846560/2023-01, cujo suporte foi fundamental e sem o qual não teria sido possível dar início a este projeto, registro meus mais sinceros agradecimentos.

Pelo processo número *2022/14179-1*, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP); deixo meus mais sinceros agradecimentos por esse imprescindível apoio financeiro concedido durante a realização deste doutorado. Tal suporte foi fundamental para a viabilização desta pesquisa, permitindo-me dedicar-me integralmente ao desenvolvimento do trabalho e à consolidação de uma trajetória acadêmica mais consistente e aprofundada.

Para além do auxílio material, reconheço na CAPES e FAPESP, instituições de enorme relevância para o fortalecimento da ciência em nosso país, ao oferecer aos seus bolsistas a oportunidade concreta de se dedicarem com excelência à produção do conhecimento. Essa contribuição transcende o âmbito acadêmico, uma vez que favorece também um expressivo desenvolvimento pessoal, ético e intelectual, indispensável à formação de pesquisadores comprometidos com a ciência e com a sociedade.

RESUMO

O Carcinoma Papilar de Tireoide (PTC, do inglês *Papillary Thyroid Carcinoma*) é a malignidade endócrina mais comum. Ele acomete principalmente mulheres, numa proporção de 3:1, entre seus 45 a 50 anos. O principal fator é a radiação ionizante, que perturba o bom funcionamento da via MAPK (*Mitogen-activated protein kinase*), causando rearranjo em proteína RET/PTC, ou dano nos genes *RAS* e *BRAF*, principais fatores relacionados ao desenvolvimento do PTC. Entretanto, a descrição de um perfil molecular ainda é necessária. As ciências ômicas amparam a pesquisa moderna em oncologia, permitindo a caracterização molecular profunda de tumores, a descoberta de biomarcadores com potencial diagnóstico e prognóstico, elucidação de alvos terapêuticos e o desenvolvimento de terapias de precisão. Neste sentido, o objetivo deste trabalho foi desenvolver e padronizar um fluxo de análise de dados ômicos, aplicados a caracterização molecular do PTC. Para isso, dados públicos independentes foram utilizados na construção de cada *pipeline* de metabolômica e transcriptômica. Além do desenvolvimento dos fluxos de análise, a melhor forma de integração de dados ômicos foi definida, a fim de relacionar as informações metabolômicas e transcriptômicas para que os dados possam revelar novos *insights* sobre o PTC. Devido a limitação de dados públicos de PTC, o fluxo de análise de dados metabolômicos foi desenvolvido, mas não pode ser totalmente aplicado, entretanto, os metabólitos foram minerados em bases de dados da literatura científica atual, com o objetivo de construir um conjunto de dados metabolômicos que possa servir como uma camada ômica adicional para a integração dos dados. Ademais, a *pipeline* de transcriptômica identificou onze genes candidatos, que demonstraram forte poder discriminatório entre tecidos tumorais e não tumorais. Foi observado que na análise de sobrevida que quatro dos onze genes se destacaram em sua significância na predição da sobrevida dos pacientes com PTC. Além disso, o mapeamento de interação *gene–droga* revelou genes como potenciais alvos terapêuticos, enfatizando o potencial translacional dos genes selecionados. Diferente de outros estudos que abordam o PTC, esses estudos dependem de conjunto de dados únicos, e de plataformas analíticas diversas. Nosso fluxo de dados multi-camadas reduz ruído, aprimora reprodutibilidade e provê uma interpretação biológica mais clara. A junção das camadas ômicas foi possível devido às análises de redes gene-metabólito que, com as interações obtidas observou-se um efeito importante no metabolismo bioenergético, onde o Efeito Warburg representa um importante fator na etiologia do câncer. Além dele, metabólitos como ácido araquidônico e ácido oleico foram nós centrais da rede, evidenciando uma mudança no status inflamatório da doença. Este trabalho representa uma das primeiras abordagem integrativas de ômicas no contexto do PTC, o que provê uma visão holística dos dados coletados. Também inova em sua capacidade de utilizar dados públicos previamente publicados, para extrair informação biológica relevante, ao descrever marcadores e processos patológicos associados ao PTC.

Palavras-chave: Carcinoma Papilar de Tireoide. Biologia Computacional. Metabolômica. Transcriptômica.

ABSTRACT

Papillary Thyroid Carcinoma (PTC) is the most common endocrine malignancy. It primarily affects women at a ratio of 3:1, between the ages of 45 and 50. The main causative factor is ionizing radiation, which disrupts the MAPK (mitogen-activated protein kinase) pathway, leading to RET/PTC protein rearrangements or damage to the RAS and BRAF genes, which are key factors in PTC development. However, a comprehensive molecular profile has yet to be fully defined. "Omics" sciences support modern oncology research by enabling in-depth molecular characterization of tumors, the discovery of biomarkers with diagnostic and prognostic potential, the identification of therapeutic targets, and the development of precision therapies. Accordingly, the aim of this study was to develop and standardize an omics data analysis workflow applied to the molecular characterization of PTC. Public datasets were used to construct the metabolomics and transcriptomics pipelines. In addition to developing these analysis workflows, the optimal method for integrating omics data was determined to link metabolomic and transcriptomic information, thereby revealing new insights into PTC. Due to limitations in available public PTC data, the metabolomics analysis workflow was developed but could not be fully implemented; however, metabolites were mined from current scientific literature databases to construct a metabolomics dataset that could serve as an additional omics layer for data integration. Furthermore, the transcriptomics pipeline identified eleven candidate genes that demonstrated strong discriminatory power between tumor and non-tumor tissues. Survival analysis revealed that four of these eleven genes were particularly significant in predicting survival outcomes for PTC patients. Finally, gene-drug interaction mapping identified specific genes as potential therapeutic targets, highlighting the translational potential of the selected genes. Unlike other studies addressing PTC, these rely on unique datasets and diverse analytical platforms. Our multi-layered data workflow reduces noise, enhances reproducibility, and provides a clearer biological interpretation. Integrating these omics layers was made possible through gene-metabolite network analyses; the resulting interaction data revealed a significant impact on bioenergetic metabolism, highlighting the Warburg effect as a key factor in PTC etiology. Additionally, metabolites such as arachidonic acid and oleic acid emerged as central network nodes, indicating a shift in the disease's inflammatory status. This work represents one of the first integrative omics approaches in the context of PTC, offering a holistic view of the collected data. It also innovates by leveraging previously published public data to extract relevant biological insights, characterizing markers and pathological processes associated with PTC.

Keywords: *Papillary Thyroid Carcinoma. Computational Biology. Metabolomics. Transcriptomics*.

LISTA DE SÍMBOLOS E ABREVIAÇÕES

ABCA6 - do inglês, *ATP-binding cassette transporter*

ABHD17C - do inglês, *Abhydrolase Domain Containing 17C*

ADK - do inglês, *Adenosine Kinase*

AKT - do inglês, *Serine/Threonine Kinase*

AR - do inglês, *Androgen Receptor*

ATA - do inglês, *American Thyroid Association*

ATIC - do inglês, *5-Aminoimidazole-4-Carboxamide Ribonucleotide Formyltransferase/IMP Cyclohydrolase*

BRAF - B-RAF Proto-Oncogene

cDNA - do inglês, *complementary DNA*

DDA - do inglês, *Data-Dependent Acquisition*

DIA - do inglês, *Data-Inependent Acquisition*

DNA - Ácido Desoxirribonucleico

DPP4 - do inglês, *Dipeptidyl Peptidase 4*

ERK - do inglês, *Extracellular Signal-Regulated Kinase*

GEO - Gene Expression Omnibus

GLT1D1 - do inglês, *Glycosyltransferase 1 Domain Containing 1*

IA - Inteligência Artificial

IGSF10 - do inglês, *Immunoglobulin Superfamily Member 10*

KRAS - KRAS proto-oncogene

LASSO - do inglês, *Least Absolute Shrinkage and Selection Operator*

m/z - Razão massa/carca

MAPK - do inglês, *Mitogen-Activated Protein Kinase*

MEK - do inglês, *Mitogen-Activated Protein Kinase Kinase*

mRNA - RNA mensageiro

MS - do inglês, *Mass Spectrometry*

MS/MS - Perfil de fragmentação MS2

NCBI - do inglês, *National Center for Biotechnology Information*

ncRNA - do inglês, *non-coding RNA*

NT5E - do inglês, *5'-Nucleotidase Ecto*

PC - do inglês, *Principal Component*

PCA - do inglês, *Principal Component Analysis*

PDGFC - do inglês, *Platelet Derived Growth Factor C*

PDK1 - do inglês, *Pyruvate Dehydrogenase Kinase 1*

PDZK1IP1 - do inglês, *PDZK1 Interacting Protein 1*

PI3K - do inglês, *Phosphoinositide 3-Kinases*

PTC - do inglês, *Papillary Thyroid Carcinoma*

RAS - Família de proteínas RAS

RET - Receptor Tirosina Cinase

RF - do inglês, *Random Forest*

RHOU - do inglês, *Ras Homolog Family Member U*

RNA - Ácido Ribonucleico

RNA-seq - Sequenciamento de Ácidos Ribonucleicos

RPS16P5 - do inglês, *Ribosomal Protein S16 Pseudogene 5*

scRNA-Seq - do inglês, *single-cell RNA Sequencing*

siRNA - do inglês, *small interfering RNA*

SLC6A14 - do inglês, *Solute Carrier Family 6 Member 14*

TENM1 - do inglês, *Teneurin Transmembrane Protein 1*

TIMP4 - do inglês, *TIMP Metallopeptidase Inhibitor 4*

TP53 - do inglês, *Tumor Suppressor Transcription Factor*

LISTA DE FIGURAS E GRÁFICOS

FIGURA 1. Cascata MAPK, principal mecanismo de carcinogênese para o Câncer de Tireóide.....13

FIGURA 2. Representação de dados metabolômicos em suas três dimensões.....19

FIGURA 3. Níveis de anotação metabolômica.....21

FIGURA 4. Métodos de integração Ômica agrupados em um diagrama de Venn, onde mostra as intersecções e seus respectivos métodos.....24

SUMÁRIO

1. INTRODUÇÃO 12
1.1. Carcinoma papilar de tireoide 12
1.2. Multi-ômicas 14
1.3. Dados ômicos 16
1.4. Uso de banco de dados públicos 22
1.5. Aprendizado de máquina nas ciências ômicas 22
1.6. Integração multi-ômica 23
2. OBJETIVOS 26
2.1. Gerais 26
2.2. Específicos 26
3. CAPÍTULO 1 – PUBLICADO (A1) 27
4. CAPÍTULO 2 – PUBLICADO (A1) 49
5. LIMITAÇÕES 75
6. CONCLUSÃO 75
7. DECLARAÇÃO DE USO DE INTELIGÊNCIA ARTIFICIAL 76
REFERÊNCIAS 77

1. INTRODUÇÃO

1.1. Carcinoma papilar de tireoide

1.1.1. Epidemiologia

O Carcinoma Papilar de Tireoide (PTC, *Papillary Thyroid Carcinoma*) é a malignidade endócrina mais comum atualmente, acometendo principalmente pessoas com idade entre 45 e 50 anos (LIMAIEM; REHMAN; MAZZONI, 2026). Predominante no sexo feminino, apresenta proporção de 3:1 (feminino-masculino), além de ser mais comum em locais com maior Índice de Desenvolvimento Humano (LIMAIEM; REHMAN; MAZZONI, 2026; WILTSHIRE et al., 2016). A literatura científica reporta fatores de risco como exposição à radiação ionizante na região de cabeça/pescoço, como o principal fator de risco de PTC (BOGOVIĆ CRNČIĆ, 2020).

Nos últimos anos, houve um aumento significativo na incidência, entretanto, a literatura discute os motivos, e declara que este aumento se deve graças aos avanços científicos e tecnológicos, como o aumento do reconhecimento de lesões microtumorais, técnicas diagnósticas mais robustas, biópsia de agulha e novas subdivisão histopatológicas (CELIK et al., 2019).

1.1.2. Fisiopatologia

Existem vários subtipos de carcinoma de tireoide, incluindo o papilar, folicular, medular e anaplásico, gerados de forma independente a partir de células de origem específicas. O mais frequente, representando mais de 80% dos casos, é o carcinoma papilar de tireoide (PTC, do inglês *Papillary Thyroid Carcinoma*) (LIN et al., 2017; SHANK; ARE; WENOS, 2022).

O PTC é iniciado com alterações genéticas constitutivas, que desencadeiam fenômenos cuja ativação leva à proliferação celular, sobrevivência do tumor e diferenciação celular. Neste sentido, o desenvolvimento do PTC pode começar com rearranjos de proteínas RET em RET/PTC ou, em mutações nos genes RAS e BRAF, levando a mudanças de expressão que culminam na diferenciação celular do epitélio folicular da tireoide, um processo chamado de cascata MAPK (JEON; TKACIK; ECK, 2024; LIU; YU, 2026; PRETE et al., 2020), como visto na **FIGURA 1**.

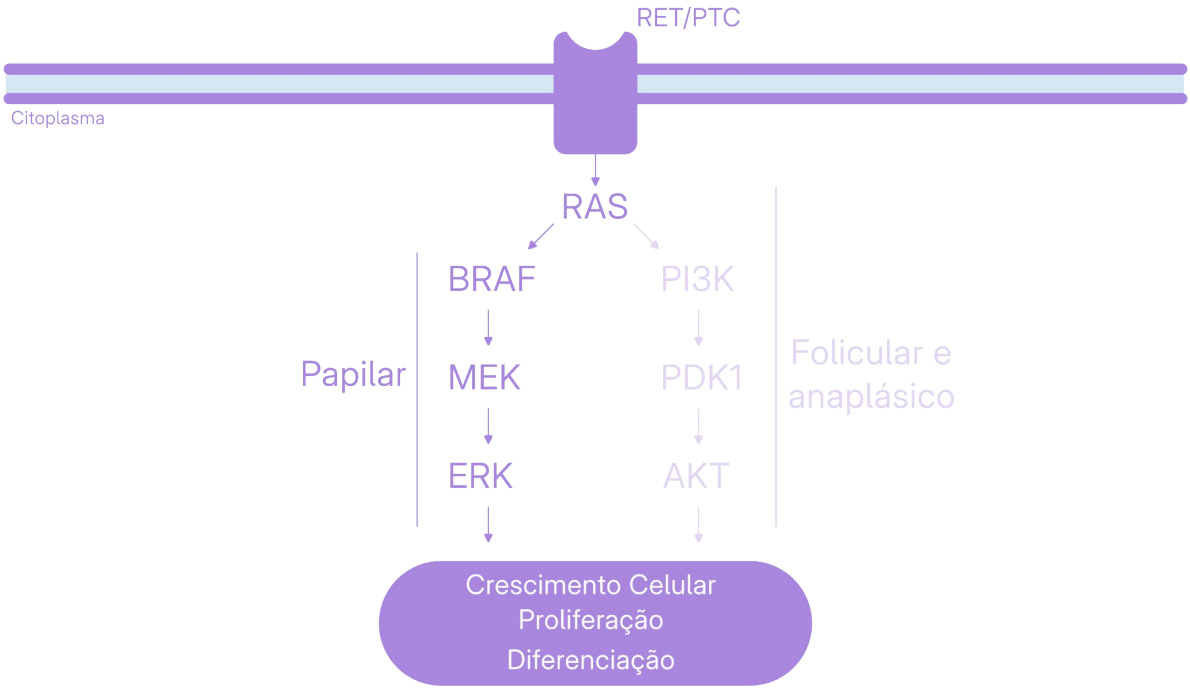


FIGURA 1. Cascata MAPK, principal mecanismo de carcinogênese para o Câncer de Tireóide. RET: *Receptor Tyrosine Kinase*; PTC: *Papillary Thyroid Carcinoma*; RAS: Família de proteínas RAS; BRAF: B-RAF Proto-Oncogene; MEK: *Mitogen-Activated Protein Kinase Kinase*; ERK: *Extracellular Signal-Regulated Kinase*; PI3K: *Phosphoinositide 3-Kinases*; PDK1: *Pyruvate Dehydrogenase Kinase 1*; AKT: *Serine/Threonine Kinase*. Fonte: (PRETE et al., 2020).

Casos de metástases acontecem principalmente na região cervical, nos linfonodos (30 – 80%), levando ao transporte de componentes tumorais via sistema linfático e circulatório. Entretanto, metástases distantes são menos frequentes, acometendo principalmente pulmões (46 – 53%) e ossos (28 – 38%) (BAINS et al., 2025; LIU; YU, 2026; TORAIH et al., 2021).

1.1.3. Diagnóstico e tratamento

O diagnóstico e tratamento de PTC são abordados como um passo a passo, onde, no Brasil, o paciente inicia a suspeita via autopalpação, para então o clínico em anamnese confirma o nódulo, que só será mais preocupante quando maior que 1 centímetro. Então o próximo passo é tomado, e a primeira linha de exame a ser realizada é a ultrassonografia focada na detecção dos nódulos suspeitos, onde será avaliada as características morfológicas do nódulo, assim como ecogenicidade, presença de halo, calcificação e aspectos dos linfonodos cervicais. A partir da confirmação da malignidade do nódulo por citopatologia via biópsia por agulha fina, a tireoidectomia é indicada

como primeira frente de tratamento. A radioiodoterapia é indicada após a tireoidectomia, com o fim de destruir tecido tumoral remanescente; é importante ressaltar que nenhuma abordagem sozinha tem alto valor preditivo, sendo a abordagem como um todo, um grande *framework* de tratamento (PCDT - CARCINOMA DIFERENCIADO DA TIREOIDE., [s.d.]).

A *American Thyroid Association* (ATA) desenvolveu um sistema de classificação que é mundialmente empregado, cujos pacientes são estratificados em quatro níveis, baseando-se no risco de recorrência clínica após a operação. Ela toma por base diferentes fontes de informação, como dados clínicos e patológicos pós-operatórios (clínica, bioquímica e imagem), dividindo os pacientes em risco baixo, baixo-intermediário, intermediário-alto e alto risco (RINGEL et al., 2025). Este novo modelo de classificação tem como incumbência estimar o risco de persistência ou recorrência de forma mais precisa para guiar o tratamento da doença (ULUDAG et al., 2025).

A literatura científica nota que a testagem molecular pode melhorar a confiança no diagnóstico, especialmente em casos de resultado indeterminado ou casos complexos. Como a definição do subtipo histológico é realizada pela análise dos núcleos celulares, algumas confusões podem acontecer, já que as características histológicas analisadas podem sofrer sobreposição de subtipo a subtipo, gerando classificações incorretas (CZARNIECKA et al., 2021; VOLPI; RAMIREZ-ORTEGA; CARRILLO, 2023).

1.2. Multi-ômicas

Muitos avanços vêm sendo feitos no que diz respeito ao manejo do paciente, como formas mais acuradas de imageamento não invasiva, definição de perfil molecular e patologia assistida por Inteligência Artificial (IA) (PARK et al., 2021).

As ciências ômicas (genômica, transcriptômica, proteômica, metabolômica) amparam a pesquisa moderna em oncologia, permitindo a caracterização molecular profunda de tumores, a descoberta de biomarcadores, de alvos terapêuticos e o desenvolvimento de terapias de precisão (CORREA-AGUILA; ALONSO-PUPO; HERNÁNDEZ-RODRÍGUEZ, 2022). Ao integrar múltiplas camadas ômicas com computação avançada e IA, é possível classificar melhor os cânceres, prever o prognóstico e compreender a resposta e a resistência ao tratamento (HSU et al., 2025).

As tecnologias ômicas de alto rendimento (*high-throughput*) permitem a análise de mutações de DNA, expressão de RNA e macromoléculas (proteínas, lipídeos e metabólitos). A multiômica na oncologia de precisão enfatiza uma visão em várias camadas, para capturar a heterogeneidade molecular do câncer, indo muito além do que marcadores únicos ou a histologia podem fornecer (CHEN et al., 2023a; JEONG; YOON, 2024).

Portanto, as ciências ômicas englobam as etapas do metabolismo, sendo as etapas iniciais (*upstream*) representadas pela genômica e transcriptômica, enquanto as etapas finais (*downstream*) são representadas pela proteômica, metabolômica e lipidômica.

1.2.1. Metabolismo upstream

A genômica (sequenciamento do genoma/exoma, painéis genéticos) identifica mutações fundamentais, alterações no número de cópias e variantes estruturais em genes carcinogênicos (KRAS, BRAF e TP53), esclarecendo vias oncogênicas e potenciais alvos terapêuticos. A transcriptômica (*bulk* e *RNA-seq*) revela vias desreguladas, assinaturas imunológicas e subtipos baseados na expressão que podem ser relacionados com o prognóstico e a resposta a medicamentos (CORREA-AGUILA; ALONSO-PUPO; HERNÁNDEZ-RODRÍGUEZ, 2022; HSU et al., 2025).

A combinação da genômica e da transcriptômica ajuda a refinar os esquemas de classificação molecular usados para a estratificação de pacientes em estudos clínicos (CHEN et al., 2023a; CORREA-AGUILA; ALONSO-PUPO; HERNÁNDEZ-RODRÍGUEZ, 2022; HSU et al., 2025). Dados integrados de expressão e mutação são amplamente usados para definir subtipos em cânceres de mama (SAMMUT et al., 2022), pulmão (YAN et al., 2025) e colorretal (DU et al., 2023; ZHOU; YANG; LI, 2025).

1.2.2. Metabolismo downstream

A proteômica captura a abundância de proteínas, modificações pós-traducionais e estados de ativação de vias, que podem divergir dos níveis de RNA e estão mais próximos do fenótipo funcional (GONÇALVES et al., 2022). Assinaturas proteômicas derivadas de tumores e fluidos biológicos estão sendo investigadas como biomarcadores diagnósticos e prognósticos e para monitorar o envolvimento do alvo do medicamento em ensaios de tratamento (CHEN et al., 2023a; JEONG; YOON, 2024).

A metabolômica monitora uma molécula-alvo, ou traça o perfil de moléculas de até 1500 Da, assim como vias metabólicas reprogramadas, revelando vulnerabilidades como glicólise alterada, dependência de glutamina e alterações no metabolismo lipídico (CHEN et al., 2023a; HSU et al., 2025; JEONG; YOON, 2024). Ela adiciona mais uma camada de dados, fornecendo informação para detecção precoce, classificação tumoral, mapeamento de microambiente e alvos terapêuticos (LIESENFELD et al., 2013; SURI et al., 2023).

A metabolômica utiliza alta tecnologia analítica (espectrometria de massas, plataformas de imageamento) para fazer suas medições, fornecendo informações diretas do fenótipo, podendo capturar consequências funcionais de alterações *upstream*, complementando a genômica e proteômica (SURI et al., 2023; ZHOU et al., 2012).

1.3. Dados ômicos

1.3.1. Dados transcriptômicos

A transcriptômica é o estudo da expressão de todas as moléculas de RNA em uma população celular específica, oferecendo uma visão global e dinâmica das alterações moleculares induzidas por doenças, fatores ambientais ou agentes patogênicos (CHEN et al., 2023a). Diferente do genoma, que é majoritariamente estático, o transcriptoma reflete o estado funcional de uma célula ou tecido em um determinado momento (CHEN et al., 2023a; DE CARVALHO et al., 2019). O termo "transcriptoma" surgiu na década de 1990, período no qual se desenvolveram os primeiros métodos de sequenciamento em larga escala, como a análise serial da expressão gênica (SAGE) baseada no sequenciamento de Sanger (LOWE et al., 2017).

Ao longo dos anos, houve uma transição tecnológica importante: microarrays e o RNA-Seq tornaram-se as abordagens dominantes. Os microarrays utilizam a hibridização de fragmentos de RNA a sondas complementares fixadas em chips para quantificar simultaneamente milhares de transcritos (LOWE et al., 2017). Posteriormente, o RNA-Seq (sequenciamento de RNA de alto rendimento) ganhou popularidade massiva a partir de 2008, permitindo sequenciar diretamente os cDNAs gerados a partir do RNA e quantificar a abundância gênica com base no número de leituras (*read counts*) obtidas para cada transcrito (CHEN et al., 2023a; LOWE et al., 2017).

1.3.1.1. Tipos de dados e Expressão Gênica

Os dados gerados por análises transcriptômicas cobrem um amplo espectro de moléculas funcionais. Embora o RNA mensageiro (mRNA), responsável por codificar proteínas, seja o alvo mais tradicional, ele representa apenas cerca de 1% a 4% do transcriptoma total (CHEN et al., 2023a). O restante é composto por RNAs não-codificantes (ncRNAs), que desempenham papéis cruciais como catalisadores ou na regulação da expressão gênica (DE CARVALHO et al., 2019). Entre eles destacam-se: (a) RNAs não-codificantes longos e curtos (como microRNAs e siRNAs). (b) RNAs circulares, associadas a diversas patologias, como doenças cardiovasculares e câncer (CHEN et al., 2023a).

Atualmente, com a evolução metodológica, destaca-se também o transcriptoma de célula única (scRNA-seq), capaz de identificar perfis de expressão em tipos celulares específicos, avançando no entendimento de cenários complexos como microambientes tumorais ou doenças neurodegenerativas (CHEN et al., 2023a).

1.3.1.2. Pipeline geral de análise de dados

Experimentos em transcriptômica exigem infraestrutura computacional e ferramentas de bioinformática e bioestatística robustas para transformar dados brutos em informações biológicas interpretáveis (LOWE et al., 2017; MANZONI et al., 2018). De forma geral, o pipeline de bioinformática para RNA-Seq segue as seguintes etapas:

A) Controle de Qualidade: Os arquivos iniciais (geralmente em formato FASTQ) contêm milhões de sequências curtas (*reads*). É necessário avaliar e filtrar leituras de baixa qualidade e remover adaptadores sequenciais.
B) Alinhamento/Mapeamento: As *reads* filtradas devem ser alinhadas contra um genoma e transcriptoma de referência (composto por milhões ou bilhões de pares de bases). Em cenários onde não há genoma de referência disponível, realiza-se a montagem de novo (LOWE et al., 2017).
C) Quantificação: Conta-se o número de sequências mapeadas sobre cada região gênica anotada para determinar os níveis brutos de abundância de RNA.
D) Normalização e Análise de expressão diferencial: Como o tamanho das bibliotecas de sequenciamento e o tamanho dos genes variam, os dados precisam ser

normalizados para permitir comparações estatísticas válidas entre diferentes condições experimentais (por exemplo, tecido tumoral *vs.* saudável).

1.3.1.3. A Importância de Gráficos de Redução de Dimensionalidade

Devido à natureza de alta dimensionalidade dos dados transcriptômicos (onde milhares de genes são quantificados simultaneamente para um conjunto de amostras), técnicas de redução de dimensionalidade são indispensáveis. A Análise de Componentes Principais (PCA) é uma das abordagens estatísticas e visuais mais fundamentais no pipeline transcriptômico. O PCA transforma as milhares de variáveis originais (os níveis de expressão de cada gene) em um novo conjunto de variáveis ortogonais chamadas componentes principais (PC1, PC2, etc), ordenadas pela quantidade de variação que explicam nos dados. Plots bidimensionais de PCA permitem:

A) Controle de qualidade global: Identificar se amostras replicadas biologicamente agrupam-se de forma consistente e detectar rapidamente outliers.
B) Visualização de Efeito de Lote (*Batch Effect*): Observar se o agrupamento das amostras ocorre devido a variáveis biológicas reais ou devido a artefatos técnicos (como o dia de processamento ou lote de reagentes). Correções precoces desses problemas evitam vieses nas análises downstream (CHEN et al., 2023a).

1.3.2. Dados metabolômicos

A metabolômica geralmente estuda amostras complexas e, portanto, é necessária uma abordagem eficiente para a separação e ordenação de compostos, tornando-a uma mistura mais simples, sendo várias pequenas misturas, ordenadas em função de características físico-químicas e distribuídas em função do tempo, gerando um gráfico de *intensidade relativa de sinal* por *tempo*, chamado de "cromatograma" (CHEN et al., 2023b).

Após a separação dos compostos, a metabolômica utiliza a Espectrometria de Massas (MS) para detectar as moléculas. O espectrômetro de massas ioniza as moléculas, separa os íons resultantes dessa ionização pela sua razão massa-carga (*m/z*) usando campos elétricos e/ou magnéticos, para finalmente detectá-los e construindo, desta forma, um espectro de massas (ALSALEH et al., 2019; VADAKEDATH et al., 2022).

Os espectrômetros de massas modernos compartilham os mesmos componentes: (i) fonte de ionização, que cria uma nuvem de íons em fase gasosa a partir das amostras; (ii) analisador de massas, que separa os íons resultantes; e (iii) um detector de íons, responsável pela captura da abundância desses íons, para gerar um espectro de massas (*intensidade de sinal* por *m/z*) (ALSALEH et al., 2019; URBAN, 2016).

Portanto, a metabolômica possui dados extremamente complexos, com informações físico-químicas e espectrométricas, que se organizam num dado (no mínimo) tridimensional (*tempo* por *intensidade de sinal* por *m/z*; **FIGURA 2**). Para lidar com esses dados tridimensionais, métodos estatísticos convencionais podem não ser os mais indicados, já que os mesmos teriam dificuldades com dados em diferentes escalas e relações entre variáveis (GE et al., 2025), sendo necessário lançar mão de métodos mais complexos, como métodos de redução de dimensionalidade e modelos de aprendizado de máquina (LEE; PARK, 2022).

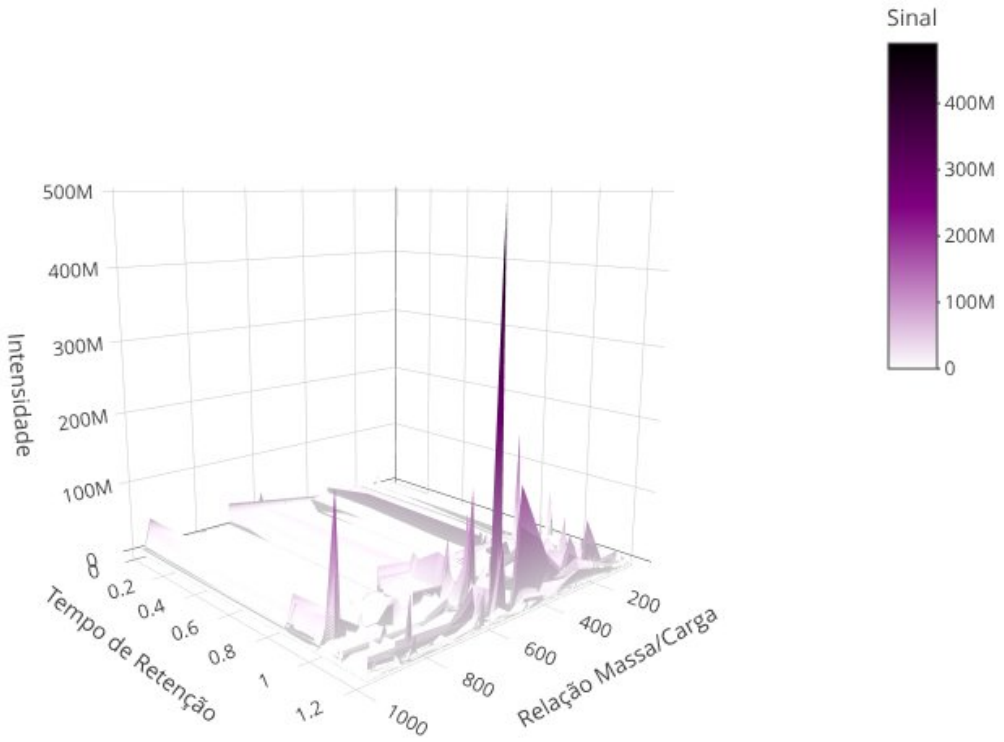


FIGURA 2. Representação de dados metabolômicos em suas três dimensões; x: Tempo de Retenção, y: Quantidade total de sinais, ou Intensidade, e z: Relação Massa/Carga. Fonte: Próprio Autor.

1.3.2.1. Tipos de aquisição de dados

Existem dois braços de análises, as direcionadas (*target*) e as não-direcionadas (*untargeted*), que diferem basicamente se as moléculas sendo medidas são conhecidas/anotadas (*target*) ou não (*untargeted*). Isso cria um *trade-off* entre a acurácia quantitativa e a abrangência de descobrimento de moléculas (ROBERTS et al., 2012; SOUZA; PATTI, 2021).

Independente da escolha de análise, *target* ou *untarget*, existem duas formas de adquirir os dados em um espectrômetro de massas: aquisição dependente de dados (*data-dependent acquisition*, DDA); e aquisição independente de dados (*data-independent acquisition*, DIA) (GUO; HUAN, 2020; HOLEWINSKI et al., 2016).

No DDA, o instrumento faz uma varredura completa para identificar as massas, para então, em tempo real, filtrar um número *n* de íons precursores mais intensos, para fragmentá-los na próxima etapa de MS/MS. Essa abordagem dos "top-*n* íons" produz espectros MS/MS limpos e simples, que podem ser comparados a bancos de dados, proporcionando anotação de alta qualidade, mas com viés em favor de espécies de alta abundância e suscetíveis a omitir moléculas de baixa abundância (GUO; HUAN, 2020).

No DIA, por sua vez, o instrumento percorre uma série de janelas de massas (*m/z*) pré-definidas, geralmente amplas (de 50 a 1500 Da), e fragmenta todos os íons precursores dentro de cada janela, coletando espectros de fragmentação para praticamente todos os íons detectáveis na faixa de massa. As análises por DIA oferecem uma cobertura muito mais abrangente e reprodutível entre as corridas, mas produzem dados MS/MS altamente complexos que exigem algoritmos de deconvolução especializados e, frequentemente, bibliotecas espectrais (HOLEWINSKI et al., 2016; LI; SMITH; ZHU, 2021; VAN DER LAAN et al., 2020).

1.3.2.2. Anotação metabolômica

A anotação de íons é o processo de atribuição de uma identidade plausível, com nível de confiança estabelecido e explicitado, caso contrário, o íon permanece anônimo e a informação é parcialmente perdida. É importante ressaltar que esse é um processo de múltiplos passos, que combina evidência experimental (massa acurada, fragmentação, tempo de retenção, *collision cross-section*) com bibliotecas externas de compostos, para reportar com os níveis de anotação (ALSEEKH et al., 2021).

Como observado na **FIGURA 3**, a anotação de um íon é medida em níveis de 1 a 5, cumulativos e decrescentes (SCHRIMPE-RUTLEDGE et al., 2016), onde o *nível 5* representa as evidências experimentais exclusivamente, enquanto o *nível 4* os íons já possuem, além das evidências experimentais, informação estrutural, como sua fórmula molecular.

Uma anotação onde são utilizadas evidências experimentais, informação estrutural e padrão isotópico, concede um *nível 3* de anotação. Alternativamente, o *nível 2* é concedido quando o íon possui estrutura predita (nível 3) e corresponde a bibliotecas utilizando dados de fragmentação de compostos. Finalmente, o *nível 1* é o maior nível de confiança na anotação de compostos, é quando a estrutura é confirmada com pelo menos duas informações ortogonais de um padrão de referência, sob as mesmas condições analíticas (p. ex.: tempo de retenção e dados de fragmentação) (SCHRIMPE-RUTLEDGE et al., 2016).

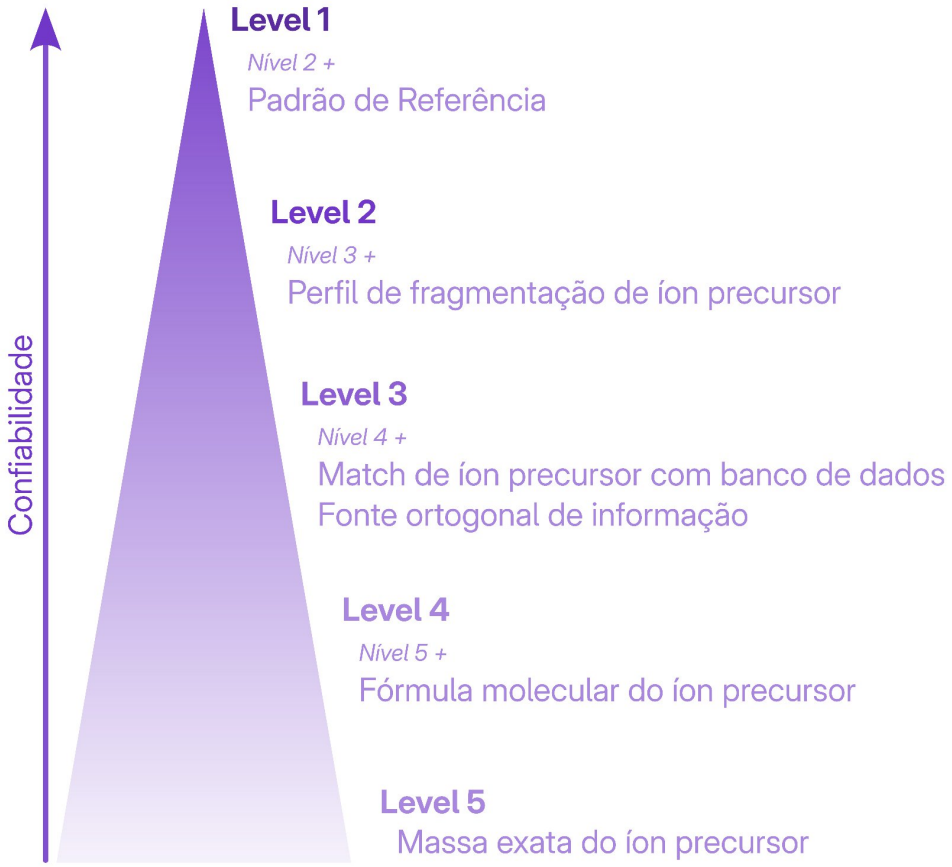


FIGURA 3. Níveis de anotação metabolômica. Níveis de 1 a 5, cumulativos e decrescentes, onde o nível de importância de evidência cresce. Fonte: Adaptado de (SCHRIMPE-RUTLEDGE et al., 2016).

O processo de anotação é importante para atribuir uma identidade para as moléculas detectadas em um estudo *untarget*, dando a ela um sentido químico e biológico. No entanto, a integração entre softwares de naturezas distintas, como um sistema proprietário e um de código aberto, frequentemente, exige compatibilizações técnicas. Essas compatibilizações envolvem adequação de formatos, convenções e metadados para assegurar interoperabilidade. Nesse contexto, *softwares*, como o SDF2PQI, são um exemplo de camada intermediadora para tornar a anotação consistente (SANCHES et al., 2023).

1.4. Uso de banco de dados públicos

O avanço exponencial das tecnologias de obtenção de dados biológicos em larga escala promoveu uma mudança de paradigma na pesquisa biomédica, consolidando a necessidade de ferramentas de compartilhamento e armazenamento (MANZONI et al., 2018). Bancos de dados públicos de livre acesso, como o Gene Expression Omnibus (GEO) do *National Center for Biotechnology Information* (NCBI), servem como repositórios centrais onde pesquisadores do mundo todo depositam os dados brutos e processados de seus estudos de *microarrays* e RNA-*Seq*. O uso desses repositórios abertos possibilita que pesquisadores possam aproveitar dados existentes para responder a novas perguntas científicas sem o custo financeiro associado ao sequenciamento inicial.

No entanto, a integração desses dados públicos impõe desafios metodológicos severos. Devido à heterogeneidade nas plataformas de sequenciamento, reagentes e protocolos de manipulação entre diferentes laboratórios, o desenho de pipeline de bioinformática integrativos deve, obrigatoriamente, prever etapas robustas de normalização e correção de efeito de lote (*batch effect*). Sem esse refinamento estatístico, artefatos técnicos podem mascarar a verdadeira variação biológica. Superados esses desafios, a exploração combinada de dados próprios com grandes consórcios públicos (como o The Cancer Genome Atlas - TCGA) mitiga as barreiras de comunicação entre equipes multidisciplinares e potencializa a descoberta de novos biomarcadores e alvos terapêuticos (CHEN et al., 2023a).

1.5. Aprendizado de máquina nas ciências ômicas

Com o avanço das tecnologias nas ciências ômicas, os dados coletados se tornaram mais volumosos e complexos, necessitando de técnicas mais avançadas para lidar com dados muitas vezes não lineares e multidimensionais (GE et al., 2025), já que os métodos estatísticos clássicos

não são poderosos o suficiente para lidar com aquele tipo de dado, fazendo as técnicas de aprendizado de máquina ficarem cada vez mais populares nos estudos em saúde (REEL et al., 2021).

Em PTC, uma série de modelos foram utilizados, incluindo modelos baseados em árvores de decisão, regressão logística e até mesmo redes neurais (LEE; PARK, 2022), entretanto dois se destacam quando se trata de multi-ômica: (i) modelos baseados em árvores de decisão, e (ii) métodos avançados baseados em regressão.

O algoritmo baseado em árvores de decisão, chamado de Floresta Aleatória (RF, do inglês, *Random Forest*) é uma abordagem não paramétrica, que acomoda diferentes tipos de variáveis de resposta, como categórica ou contínua, ou mesmo diferentes escalas. Por isso, é adequado para contextos de alta dimensionalidade, onde o número de variáveis é maior que o número de observações, por exemplo. Desta forma, também se faz adequada para dados ômicos, que frequentemente, apresentam alta dimensionalidade (HU; SZYMCZAK, 2023).

Já que a seleção de variáveis é extremamente importante para métodos ômicos, a regressão por LASSO (do inglês, *Least Absolute Shrinkage and Selection Operator*) é um ótimo exemplo de método assim, onde, ele zera o coeficiente de uma variável independente insignificante, efetivamente tirando-a do modelo, além de incorporar fatores que seletivamente evitam o *overfitting* do modelo. Em comparação com os métodos de regressão tradicionais, o LASSO pode destacar as variáveis mais relevantes para uma condição estudada (AHMED et al., 2024; GOU et al., 2025), e fundamenta métodos ainda mais complexos, como as Redes Elásticas (AHMED et al., 2024).

1.6. Integração multi-ômica

A genômica, transcriptômica, proteômica, metabolômica e lipidômica são as tecnologias ômicas mais usadas, cada uma provendo diferentes camadas para o entendimento de sistemas biológicos, cada uma provendo diferentes observações, com diferentes classes de moléculas, tipos de dados e interpretações próprias (REEL et al., 2021).

Com o advento das tecnologias de *high-throughput*, o campo das ciências ômicas dá passos significativos em direção à caracterização de sistemas biológicos, em diferentes níveis de complexidade. Entretanto, analisar cada camada separadamente pode não prover o entendimento

completo do objeto de estudo, fazendo com que a integração dessas abordagens seja necessária, de forma a ter uma visão mais completa de sistemas biológicos (SANCHES et al., 2024).

As integrações combinadas procuram explicar cada conjunto de dados ômicos através de técnicas abrangentes e *frameworks* integrados (MORABITO et al., 2025), visando combinar os conjuntos de dados em um único, para prosseguir com as análises pertinentes. Técnicas baseadas em aprendizado de máquina podem ser utilizadas para minerar informações escondidas no dado experimental, mas também para garantir uma melhor integração dos dados

Estratégias baseadas em correlação, quando aplicadas em diferentes tipos de dados ômicos, buscam quantificar a relação entre as diferentes moléculas (co-expressão, co-regulação ou interações funcionais), com o objetivo de encontrar padrões e desta forma, descrever fenômenos biológicos (REEL et al., 2021). Existem inúmeras tentativas de junção de dados ômicos, e em uma revisão recente pelo grupo, essas tentativas foram agrupadas de acordo com a **FIGURA 4**, abaixo.

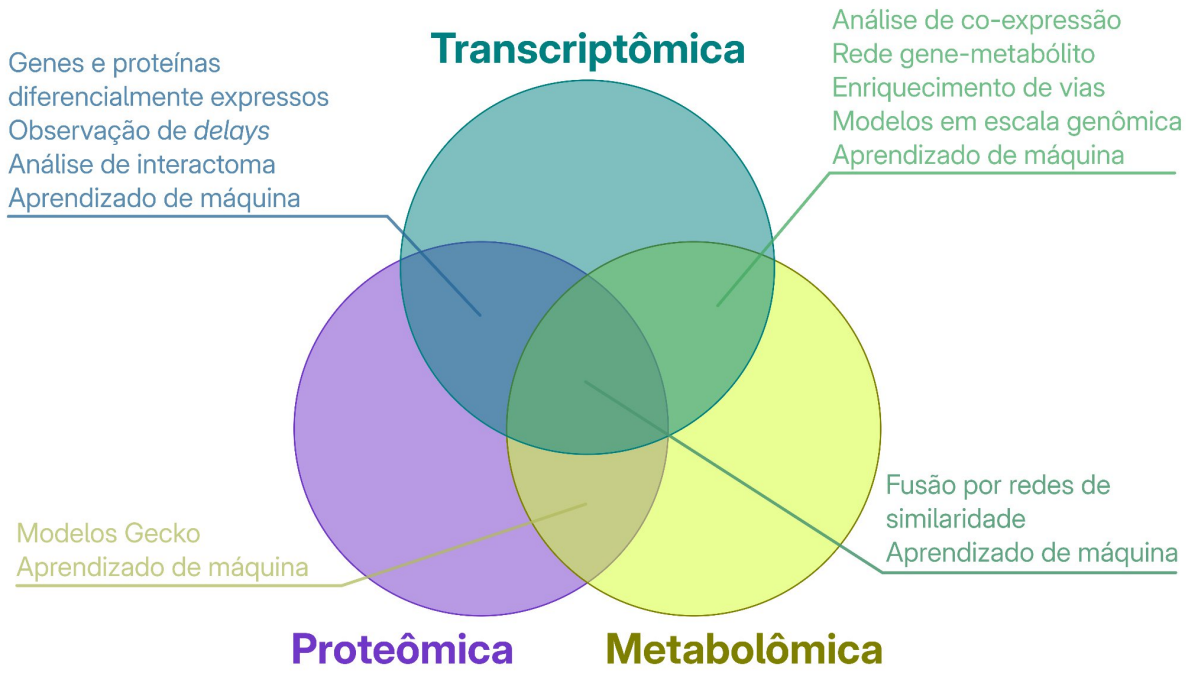


FIGURA 4. Métodos de integração Ômica agrupados em um diagrama de Venn, onde mostra as intersecções e seus respectivos métodos. Fonte: Próprio Autor.

As abordagens multi-ômicas, portanto, vem ganhando popularidade devido ao seu potencial de oferecer uma visão mais holística dos mecanismos biológicos, que as ômicas isoladas não

conseguem (AHMED et al., 2024). Apesar dos avanços impressionantes, ainda existem desafios como a heterogeneidade dos dados, os efeitos de lote, o tamanho limitado das amostras, *pipeline* padronizada e validação externa robusta antes da implantação clínica.

A interpretação de modelos complexos de IA e aprendizado profundo continua difícil, e a integração de ômicas espaciais, temporais e de *single-cell* representam grandes demandas computacionais e de infraestruturas (CORREA-AGUILA; ALONSO-PUPO; HERNÁNDEZ-RODRÍGUEZ, 2022; JEONG; YOON, 2024; WEKESA; KIMWELE, 2023).

As direções futuras, segundo literatura recente, incluem uma adoção mais ampla da multi-ômica espacial e de *single-cell*, uma integração mais estreita da ômica convencional com a de imageamento e o desenvolvimento de modelos de IA cujos clínicos possam confiar para *apoiar* as decisões em oncologia de precisão (JEONG; YOON, 2024; WEI et al., 2023).

À medida que o compartilhamento e a padronização de dados melhorarem, a multi-ômica tem o potencial de informar de forma mais rotineira a avaliação de riscos, a detecção precoce, a adaptação do tratamento em tempo real e a prevenção da resistência em todo o continuum do tratamento do câncer (CHEN et al., 2023a; HSU et al., 2025).

2. OBJETIVOS

2.1. Gerais

Desenvolver uma *pipeline* de análise e integração de dados para análises multi-ômicas, focadas na descrição molecular do carcinoma papilar de tireoide.

2.2. Específicos

- Desenvolver um método de tratamento e modelagem de dados metabolômicos;
- Criar uma *pipeline* de tratamento e análise de dados transcriptômicos;
- Avaliar os métodos de integração de dados multi-ômicos e aprendizado de máquina disponíveis na literatura;
- Desenvolver modelos de aprendizado de máquina para dados de Carcinoma Papilar de Tireoide; e
- Integrar dados ômicos públicos de Carcinoma Papilar de Tireoide à *pipeline* de análise.

3. CAPÍTULO 1 – PUBLICADO (A1)

Integrating Molecular Perspectives: Strategies for Comprehensive Multi-Omics Integrative Data Analysis and Machine Learning Applications in Transcriptomics, Proteomics, and Metabolomics.

Com a introdução de tecnologias de alto rendimento (*high-throughput*), a área das ômicas avançou significativamente na caracterização de sistemas biológicos em diversos níveis de complexidade. A transcriptômica, a proteômica e a metabolômica são as três tecnologias ômicas mais utilizadas, fornecendo, cada uma, informações valiosas sobre diferentes camadas de um sistema biológico. No entanto, a análise isolada de cada conjunto de dados ômicos pode não oferecer uma compreensão abrangente do objeto de estudo. Por isso, a integração de dados multi-ômicos tornou-se cada vez mais importante na pesquisa em bioinformática. Neste artigo, revisamos estratégias para integrar dados de transcriptômica, proteômica e metabolômica, incluindo análise de coexpressão, redes gene-metabólitos, modelos baseados em restrições, análise de enriquecimento de vias e análise de interactoma. Discutimos, ainda, abordagens combinatórias de integração ômica, estratégias baseadas em correlação e técnicas de aprendizado de máquina que utilizam um ou mais tipos de dados ômicos. Ao apresentar esses métodos, buscamos proporcionar aos pesquisadores uma melhor compreensão de como integrar dados ômicos para obter uma visão mais abrangente de um sistema biológico, facilitando a identificação de padrões e interações complexas que poderiam passar despercebidos em análises de uma única abordagem ômica.

biology

MDPI

Review

Integrating Molecular Perspectives: Strategies for Comprehensive Multi-Omics Integrative Data Analysis and Machine Learning Applications in Transcriptomics, Proteomics, and Metabolomics

Pedro H. Godoy Sanches [1,†], Nicolly Clemente de Melo [2,†], Andreia M. Porcari [1] and Lucas Miguel de Carvalho [3,*]

[1] MS4Life Laboratory of Mass Spectrometry, Health Sciences Postgraduate Program, São Francisco University, Bragança Paulista 12916-900, SP, Brazil
[2] Graduate Program in Biomedicine, São Francisco University, Bragança Paulista 12916-900, SP, Brazil
[3] Post Graduate Program in Health Sciences, São Francisco University, Bragança Paulista 12916-900, SP, Brazil
* Correspondence: lucas.miguel@usf.edu.br
† These authors contributed equally to this work.

Simple Summary: Recent high-throughput technologies such as transcriptomics, proteomics, and metabolomics have allowed progress in understanding biological systems at different levels of detail. Even so, it is necessary to integrate multiple omics data sets to achieve a comprehensive understanding of the subject under study. In this article, we review the methods used for integrating transcriptomics, proteomics, and metabolomics data and summarize them in three approaches: combined omics integration, correlation-based integration strategies, and machine learning integrative approaches. Our goal is to showcase the uses and limitations of each approach, allowing researchers to choose the more appropriate tool for each scenario to extract a comprehensive view of a biological system.

Abstract: With the advent of high-throughput technologies, the field of omics has made significant strides in characterizing biological systems at various levels of complexity. Transcriptomics, proteomics, and metabolomics are the three most widely used omics technologies, each providing unique insights into different layers of a biological system. However, analyzing each omics data set separately may not provide a comprehensive understanding of the subject under study. Therefore, integrating multi-omics data has become increasingly important in bioinformatics research. In this article, we review strategies for integrating transcriptomics, proteomics, and metabolomics data, including co-expression analysis, metabolite–gene networks, constraint-based models, pathway enrichment analysis, and interactome analysis. We discuss combined omics integration approaches, correlation-based strategies, and machine learning techniques that utilize one or more types of omics data. By presenting these methods, we aim to provide researchers with a better understanding of how to integrate omics data to gain a more comprehensive view of a biological system, facilitating the identification of complex patterns and interactions that might be missed by single-omics analyses.

Keywords: omics data; transcriptomics; metabolomics; proteomics; omics integration; multi-omics

check for updates

Citation: Sanches, P.H.G.; de Melo, N.C.; Porcari, A.M.; de Carvalho, L.M. Integrating Molecular Perspectives: Strategies for Comprehensive Multi-Omics Integrative Data Analysis and Machine Learning Applications in Transcriptomics, Proteomics, and Metabolomics. *Biology* **2024**, *13*, 848. https://doi.org/10.3390/biology13110848

Academic Editors: Dario Di Silvestre, Min Zhao and Ruifeng Hu

Received: 30 June 2024
Revised: 19 July 2024
Accepted: 25 July 2024
Published: 22 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Omics data result from the use of large-scale instruments used in biology. They can contain measurements of different biomolecules, their functions, and interactions [1]. They have become an essential tool in modern biology and biomedicine [2,3]. Transcriptomics, proteomics, and metabolomics are three major omics fields that provide different types of biological information.

As an indirect measure of DNA activity [4], transcriptomics measures the expression levels of a set of RNA transcripts (mRNA, non-coding RNA, etc.) in a cell or tissue,

i.e., the transcriptome. Produced according to the instructions from mRNA, proteins and enzymes, typically > 2 kDa, are the functional products of genes and play several roles in cellular processes [5], being the macromolecule responsible for direct interactions among cells and tissues, besides maintaining the cellular structure [6]. Proteomics then focuses on the identification and quantification of a set of proteins, the *proteome* [7]. Smaller molecules (≤1.5 kDa), referred to as metabolites, are intermediate or end products of metabolic reactions and regulators of metabolism [8], but are not analyzed with the instrumental methodologies used in proteomics. Metabolomics comprehensively analyzes these molecules, trying to describe and quantify the molecular composition of a sample (the metabolome). Additionally, a big branch of metabolomics is the study of the lipidic composition of a sample, its lipidome, and for that, the "Lipidomics" term is used [8]. By indirectly measuring how a gene is acting, transcriptomics covers the upstream processes of metabolism, while proteomics is the intermediate step, defining protein structure and biocommunication. Metabolomics focuses on the regulators and the ultimate mediators of a metabolic process (usually smaller molecules not more than 1.5 kDa) [9]. Together, these omics technologies offer a comprehensive and streamlined view of biological processes.

Integrating multiple omics data sets is a challenging but necessary task to fully understand complex biological systems. Data integration can provide novel biological insights and reveal previously unknown relationships between different molecular components. Moreover, it can help identify biomarkers and therapeutic targets for various diseases. Several methods have been developed for integrating omics data, including correlation-based approaches, machine learning algorithms, and network-based analyses [4,10–12]

In this article, we will review and discuss different methods for integrating transcriptomics, proteomics, and metabolomics data. We will discuss the strengths and limitations of each method and provide examples of their applications in various biological contexts. Also, we will cite strategies and articles that use these omics in machine learning-based studies. By doing so, we hope to contribute to the development of effective strategies for omics data integration and pave the way for new discoveries in biomedical and biotechnology research.

2. Methods for Integrating Multi-Omics Data

Integrating omics data from several domains is critical for gaining complete knowledge of biological systems. To uncover critical regulatory pathways and networks, transcriptomics data can be combined with proteomics or metabolomics data. Many methodologies for integrating transcriptomics and proteomics data have been developed, including correlation-based approaches and pathway and co-expression analysis. Merging proteomics data with metabolomics data is also a potential strategy for biomarker development and disease diagnosis, since it can uncover alterations in metabolic pathways linked to disease states. Co-expression analysis and network-based techniques have been utilized successfully in the integration of transcriptomics and metabolomics. Overall, integrating omics data can be a significant tool for deciphering complicated biological systems and discovering novel treatment targets [3,10,12–16].

We divide the methods of omics integration into three major approaches: combined omics integration, correlation-based integration strategies, and machine learning integrative approaches (Figure 1). Combined omics integration approaches attempt to explain what occurs within each type of omics data in an integrated manner, generating independent data sets. Correlation-based strategies apply correlations between the generated omics data and create data structures to represent these relationships, such as networks. Finally, machine learning strategies utilize one or more types of omics data, potentially incorporating additional information inherent to these data sets, to comprehensively understand responses at the classification and regression levels, particularly in relation to diseases. These methods enable a comprehensive view of biological systems, facilitating the identification of complex patterns and interactions that might be missed by single-omics analyses. By leveraging these integrative approaches, researchers can achieve deeper insights into the

molecular mechanisms underlying health and disease, ultimately aiding in the discovery of novel biomarkers and therapeutic targets.

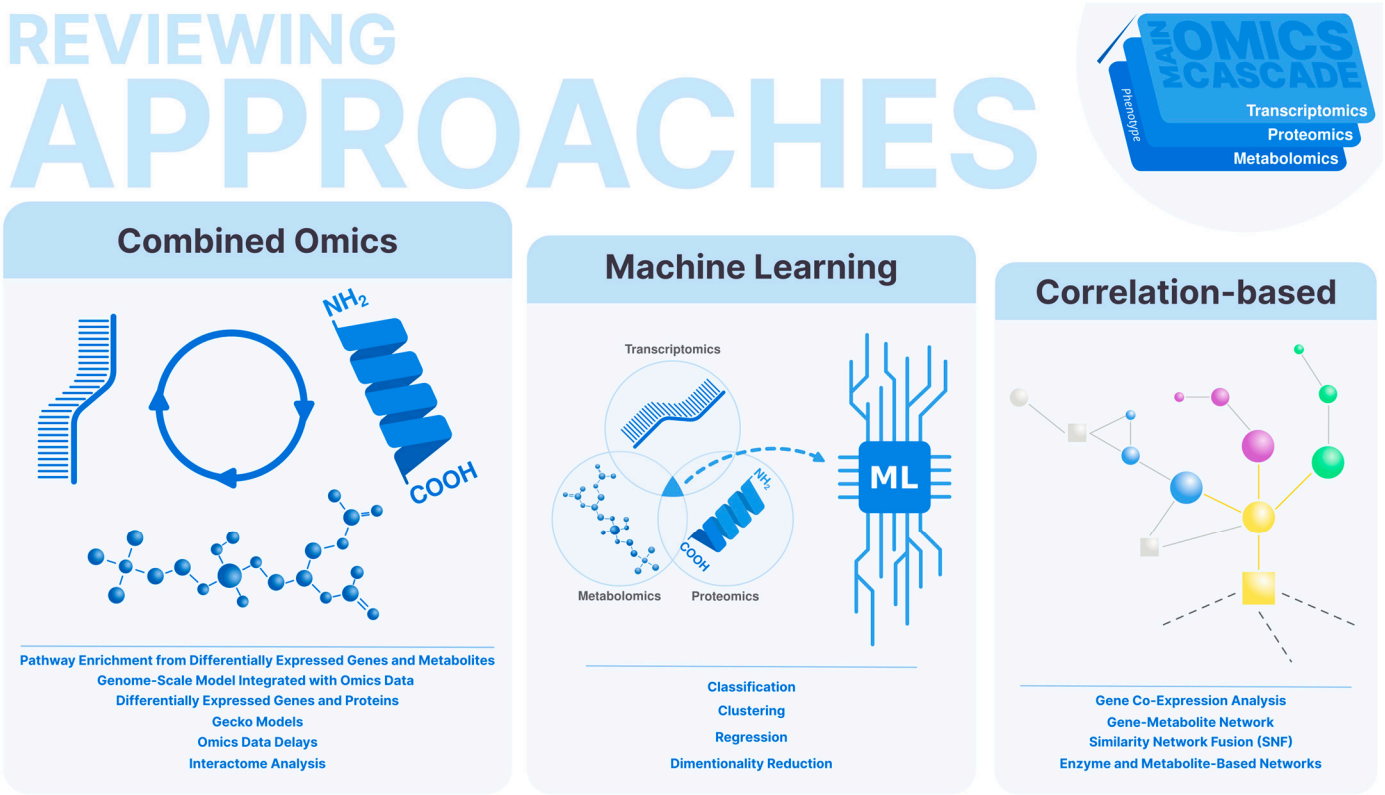


Figure 1. Strategies for integrating omics data. Methods are based on correlation-based approaches, which identify associations between different types of data; machine learning algorithms, which can predict outcomes and identify patterns across data sets; and combined individual approaches, which map the interactions and relationships between molecular components.

In this article, we focus on non-approximate strategies for cell-to-cell communication, such as single-cell RNA-seq (scRNA-seq), due to their level of resolution and ability to detect communication between individual cells. Bulk analysis assumes that cells are identical and can model the exchange between cells and the environment [17]. However, we will provide excellent recent reviews that present strategies for integrating multi-omics data from single-cell data [17–25].

2.1. Correlation-Based Methods

Correlation-based strategies involve applying statistical correlations between different types of generated omics data to uncover and quantify relationships between various molecular components. These methods, summarized in Table 1, then create data structures, such as networks, to visually and analytically represent these relationships. By mapping out these correlations, researchers can identify patterns of co-expression, co-regulation, and functional interactions that occur across different omics layers. This approach allows for the detection of complex interdependencies and the construction of interaction networks that highlight key molecules and pathways involved in biological processes.

Table 1. Summary of methods and strategies for integrating transcriptomics, proteomics, and metabolomics data using the correlation-based approach.

Integration Approach	Strategy or Method	Possible Omics Data	Main Idea
Correlation-based	Gene co-expression analysis	Transcriptomics and metabolomics	Identify co-expressed gene modules with metabolite similarity patterns under the same biological conditions
	Gene–metabolite network	Transcriptomics and metabolomics	Perform a correlation network of genes and metabolites
	Similarity Network Fusion	Transcriptomics, proteomics, and metabolomics	Builds a similarity network for each omics data separately, and subsequently, all networks are merged, and the edges with high associations in each omics network are highlighted
	Enzyme and metabolite-based network	Proteomics and metabolomics	Identify a network of protein–metabolite or enzyme–metabolite interactions using genome-scale models or pathways databases

2.1.1. Gene Co-Expression Analysis Integrated with Metabolomics Data

Co-expression analysis is a powerful approach for identifying genes with the same expression pattern that may participate in the same biological pathways or have the same biological function [26,27]. One strategy for integrating transcriptomics and metabolomics data is to perform a co-expression analysis on transcriptomics data and identify gene modules that are co-expressed. These modules can then be linked to metabolites identified from metabolomics data to identify metabolic pathways that are co-regulated with the identified gene modules [28–34].

To further understand the relationship between co-expressed genes and metabolites, the correlation between metabolite intensity patterns and the eigengenes of each co-expression module can be calculated. Eigengenes are representative expression profiles for each module that summarize the overall expression pattern of the genes within the module. By correlating these eigengenes with metabolite intensity patterns, it is possible to identify which metabolites are most strongly associated with each co-expression module [35–39]. Additionally, you can use the normalized metabolomics data directly in Weighted Correlation Network Analysis (WGCNA) [27], conducting a module–sample relationship analysis (in this case, module–metabolite relationship), and identify relationships between module eigengenes and metabolite intensity.

This approach can provide important insights into the regulation of metabolic pathways and the formation of specific metabolites. For example, if a particular co-expression module is strongly correlated with the production of a specific metabolite, it suggests that the genes within the module are involved in regulating the metabolic pathway leading to that metabolite. By combining transcriptomics and metabolomics data, it is possible to identify key genes and metabolic pathways involved in specific biological processes or disease states and to develop targeted interventions to modulate these processes, taking into account that the biological conditions of both omics analyses must be the same.

2.1.2. Gene–Metabolite Network

A gene–metabolite network is a visualization of the interactions between genes and metabolites in a biological system. Generating and analyzing these networks involves collecting gene expression and metabolite abundance data, integrating the data, constructing the network, analyzing it, and interpreting the results. Gene–metabolite networks can help

identify key regulatory nodes and pathways that are involved in metabolic processes and can be used to generate hypotheses about the underlying biology.

To generate a gene–metabolite network, researchers must first collect gene expression and metabolite abundance data from the same biological samples. These data are then integrated using Pearson correlation coefficient (PCC) analysis or other statistical methods to identify genes and metabolites that are co-regulated or co-expressed [40–42]. For example, Nikiforova et al., 2005, exhibited a systematic procedure to construct a gene–metabolite network based on the profiles of transcripts and metabolites [43]. Also, gene–metabolite networks are constructed using visualization software, such as Cytoscape [44] or igraph [45], with genes and metabolites represented as nodes in the network and connected with edges that represent the strength and direction of their interactions.

Once the network is constructed, it can be analyzed using network analysis tools to identify key regulatory nodes and pathways that are involved in metabolic processes [46–50]. Furthermore, a gene–metabolite network could be constructed with genes and metabolites specifically deregulated, to focus on the process that is modulated by each biological condition. Validation and interpretation of the network can then be performed by comparing it to known metabolic pathways and regulatory networks and using pathway analysis, gene ontology analysis, and functional enrichment analysis to identify enriched pathways and processes. Afterwards, you may select genes present in metabolic pathways or biological processes and obtain the mRNA levels of these genes by real-time qRT-PCR.

One interesting way of constructing a metabolite–gene network is based on the corto package developed by Mercatelli D. et al. (2020) [51]. First, the metabolite and gene data are combined into a single matrix, with only the metabolites designated as "centroids" or hubs in the resulting co-occurrence network. To determine significant edges in the network, you may set the minimum Pearson correlation coefficient *p*-value to 0.05 or a chosen cut-off. To test the significance of each edge in the network, it is necessary to conduct 100 or more bootstraps. The R code snippet below describes the implementation of this method, with "full_matrix" as the input matrix, and "metabolites" as a vector containing the metabolite names, as proposed by Cavicchioli M.V. et al. (2022) [52].

2.1.3. Similarity Network Fusion (SNF)

Similarity Network Fusion (SNF) [53] is a computational approach for integrating various data types [54,55]. In essence, SNF merges diverse measurements, such as mRNA expression, protein abundance, miRNA expression, metabolomics, clinical data, questionnaires, and image data, among others, for a given set of samples, like patients. Essentially, it starts by creating a sample similarity network for each data type and then combines these networks iteratively using a unique network fusion technique. By operating in the sample network space, SNF effectively circumvents issues related to different scales, collection biases, and noise in various data types.

For example, let us say that we have transcriptomics data from N patients (treated vs. control) and metabolomics data from the same patients. The SNF algorithm constructs a similarity network for each omics data set separately, where each node represents a patient, and the edge intensity connecting two patients indicates the level of association based on that omics data set. Subsequently, both networks are merged, and edges with high associations in each omics network are highlighted. Clustering can be performed to identify phenotypic associations among patients and cancer types, for instance. Furthermore, patient information (such as weight, survival time, among others) can be added to the nodes to enhance information visualization. Looking at the final network and examining each edge and its source (whether it is supported by one or more omics datasets), we can see how each omics dataset supports each group of patients. This allows for comparisons between different diseases. We highlight that there are limitations in the SNF method, such as dependence on pre-processing, noise in the input data, and computational complexity.

2.1.4. Enzyme and Metabolite-Based Network

In this approach, we utilize the available metabolomics and proteomics data to identify a network of protein–metabolite or enzyme–metabolite interactions using genome-scale models or pathways databases and then combine it with omics data such as fold changes to visualize the enriched pathways. There are two main strategies to define such networks, based on either proteomics or metabolomics data in the first step.

Let us start with the proteomics-based strategy. First, we identify the reactions in which the identified proteins participate in the genome-scale model, which can be in the GPR in AND or OR operator. Then, we determine the metabolites that are consumed (met-c) and produced (met-p) in those reactions. This information is used to construct a protein–metabolite network, where the protein is derived from the identification of the GPR, and the metabolites are the reactions that the protein is involved in (met-c and met-p). Additional details such as compartments, reactions ID, and fold changes can be included to enhance the network.

The second strategy relies on metabolomics data, which require standardization of the metabolite names or IDs in the database. Using the identified metabolites, along with the genome-scale model, we identify all the reactions in which those metabolites participate and annotate the proteins belonging to the GPR, in AND or OR operator. Then, we assemble the protein–metabolite network using this interaction information.

Both strategies are based on omics data with the addition of genome-scale models as the foundation for assembling the network structure. Although the associations from the Kyoto Encyclopedia of Genes and Genomes (KEGG) can be used directly, we expand to a whole genome-scale model. The first strategy yields a larger network due to the greater number of identifications that proteomics data provide compared to metabolomics data.

MetaBridge [56] is a powerful web-based tool that aims to integrate metabolomics with proteomics data. The tool achieves this by utilizing data from two key databases: the MetaCyc metabolic pathway database [57] and KEGG. MetaBridge maps metabolite compounds to interacting upstream or downstream enzymes in enzymatic reactions and metabolic pathways and generates a list of enzymes that can be integrated with proteomics or transcriptomics data using protein–protein interaction (PPI) networks. The resulting PPI network can be used for integrative multi-omics analyses, allowing users to identify key proteins and pathways that are differentially regulated across the data sets and also integrate the fold change per protein and select specific submodules from it. By providing a user-friendly interface and detailed protocols, MetaBridge makes it easy to perform integrative multi-omics analyses and gain insights into complex biological systems.

2.2. Combined Omics Approaches

Combined omics integration approaches, as summarized in Table 2, seek to explain the phenomena occurring within each type of omics data through a comprehensive and integrated framework. These methods generate independent data sets for each omics layer, such as transcriptomics, proteomics, and metabolomics, and then combine them to provide a full view of a biological system. This integrated analysis can reveal insights that would not be apparent when examining each omics data set in isolation, thus enabling the identification of novel interactions, pathways, and regulatory mechanisms that drive biological functions and disease states. Here, we will discuss different strategies and the necessary care for integrating omics data.

Table 2. Summary of methods and strategies for integrating transcriptomics, proteomics, and metabolomics data using the combined omics approach.

Integration Approach	Strategy or Method	Possible Omics Data	Main Idea
Combined omics	Pathway enrichment from differentially expressed genes and metabolites	Transcriptomics and metabolomics	Identify pathways enriched in both types of omics data and perform a post-analysis with these results
	Integrating genome-scale models with omics data	Transcriptomics and metabolomics	Integrate metabolic and transcriptomic data to create content-specific models and perform specific metabolic simulations
	Gecko models	Proteomics and metabolomics	Integrate proteomics data into an enzyme model, which can be validated with metabolomics data under the same biological conditions.
	Differentially expressed genes and proteins	Transcriptomics and proteomics	Identify similarities between the lists of differentials in the two omics data sets
	Observing delays between omics data	Transcriptomics and proteomics	Identify whether there is a temporal delay in the acquisition of omics data based on gene expression and protein abundance
	Interactome analysis	Transcriptomics and proteomics	Identify functional relationships between different proteins and genes using interactome databases and fold-change values

2.2.1. Pathway Enrichment from Differentially Expressed Genes and Metabolites

One strategy for integrating transcriptomics and metabolomics data is to perform KEGG enrichment analysis on differentially expressed genes (DEGs) identified from transcriptomics data and then link the enriched pathways to the metabolites identified from metabolomics data. This approach can reveal the metabolic pathways that are most affected by changes in gene expression and provide insights into the underlying biological mechanisms. Additionally, fold-change information from DEGs can be integrated with metabolomics data to identify metabolites that are significantly changed in abundance and may play a key role in the observed changes in gene expression.

DEG pathway enrichment may be performed using the clusterProfiler [58] or GSEApy [59] packages, for example. One way to convert a gene name to Entrez Gene ID, which provides unique integer identifiers for genes and other loci, or KEGG ID to perform the enrichment through the clusterProfiler is through the genekitr package using the *transId()* function with the argument *transTo* = *"entrez"*. The MetaboAnalyst platform [60] may be used to enhance metabolic pathways with identified metabolites. Furthermore, information from gene expression, such as fold changes, and metabolite intensity analysis can be integrated into metabolic pathway figures using the Pathview package [61].

Additionally, it is possible to analyze the correlation between the abundance of metabolites and the expression of genes or transcripts across various biological conditions using integrated pathway analysis. To perform this analysis, it is necessary to include multiple time points or a large number of biological conditions. The association cut-off for this analysis is based on a p-value $< \alpha$ and Pearson coefficients $> \beta$, which can be plotted to determine the extent to which relevant metabolites are correlated with relevant mRNA transcripts. To carry out this analysis, Multi-Omics Factor Analysis (MOFA) [62] is a suitable approach for determining the degree to which changes in metabolite abundance and mRNA expression variables are related [63–65]. In addition, correlation analysis can be applied to

determine the degree of relationship between metabolite intensities/concentrations and gene expression, but this must be carried out with normalized (post-processing) data. Data normalization in the post-processing of metabolomics data is usually provided, especially if the statistical analyses are conducted in MetaboAnalyst.

Also, if a multivariate analysis is chosen, such as Principal Component Analysis (PCA), it is important to understand the factors that can lead to failure of the analysis of variance. We point out that data integration through concatenation can become complex when the data sets to be merged differ significantly in size. Not only do metabolomics and transcriptomics data sets differ significantly in size, but they are also generated using vastly different technologies [66]. This means that the data sets have distinct structures, unique patterns of expected values, dissimilar distributions of underlying noise, and varying levels of variance. With that in mind, the best proposal is to carry out a PCA biplot for the metabolites and another for the genes, identify the most correlated metabolites and genes either with each biological condition or with each principal component (PC1 or PC2), and then associate this result with the pathway that both belong to.

2.2.2. Integrating Genome-Scale Models with Metabolomics and Transcriptomics Data

Integrating transcriptomics data, genome-scale models (GEMs), and metabolomics data can provide a comprehensive understanding of cellular metabolism and its regulation [67]. One approach to integrating these data types is to generate condition-specific models (CSMs) that incorporate the transcriptomics data and use them to simulate metabolic fluxes in different situations [68–72]. Here, we describe a strategy for generating CSMs using GEMs and transcriptomics data and then integrating these models with metabolomics data.

The first step in this strategy is to generate a GEM that represents the metabolic network of the organism of interest. The GEM should include all the reactions and metabolites involved in cellular metabolism and should be curated and validated using experimental data. Once the GEM is generated, it can be used to simulate metabolic fluxes under different conditions.

Next, transcriptomics data can be used to generate CSMs that incorporate the expression levels of genes under different conditions. This can be achieved using constraint-based modeling techniques such as flux balance analysis (FBA) [73] or parsimonious FBA (pFBA) [74]. CSMs can be generated by constraining the fluxes through the reactions in the GEM based on the expression levels of the corresponding genes. MEWpy [75] is a package that covers a wide range of metabolic and regulatory modeling approaches, as well as phenotype simulation and Computational Strain Optimization (CSO) algorithms. This makes it a useful tool for generating transcriptome-based simulations with FBA or pFBA.

Once the CSMs are generated, they can be used to simulate metabolic fluxes in different conditions and predict the metabolic phenotypes of a cell. However, to fully understand the regulation of cellular metabolism, it is important to also integrate metabolomics data into the models.

One way to integrate metabolomics data into CSMs is to use the former to constrain the fluxes through the exchange reactions that correspond to the measured metabolites. It is essential to normalize the exchange reactions based on metabolite concentrations in a metabolic model. For example, if an organism cannot consume the entire concentration of a metabolite in 24 h, you can estimate the upper exchange flux as follows: [Concentration]/(1 gDW $\times$ 24 h). Moreover, one possibility is using linear programming techniques to find the flux distributions that are consistent with both the transcriptomics and the metabolomics data, and this approximation is used to insert the exchange reaction fluxes. We highlight that there are limitations in the method, such as noise and data variability, differences in temporal scales between the omics data, incomplete annotations in the metabolic model, and limitations in experimental validation.

2.2.3. Gecko Models

In recent years, the integration of omics data into genome-scale metabolic models (GEMs) has become a powerful approach for exploring the relationship between genotype and phenotype for different organisms [76–78]. GEMs allow for the prediction of metabolic behavior and can be used to design experiments and engineer biological systems. One such tool that has enabled the integration of proteomics data into GEMs is the GECKO toolbox [79]. GECKO models take into account enzyme and proteomics constraints to study phenotypes that are constrained by protein limitations [80–83]. With the GECKO toolbox, it is possible to generate enzyme-constrained models (ecModels) for a variety of organisms, including budding yeasts such as *Saccharomyces cerevisiae* and humans, as well as build your own model. These models can be used to study the long-term adaptation of organisms to stress factors and nutrient-limited conditions.

The GECKO models simplify the process of limiting the metabolic fluxes in any GEM that contains enzymatic data, reducing the variability of constraint-based modeling results and improving predictions. This approach is executed by representing enzymes as entities with limited capacities in the corresponding reactions in the model, thereby extending the genome-scale modeling. In traditional genome-scale modeling, a stoichiometric matrix is defined that represents the whole metabolism, with columns indicating each reaction's stoichiometry, and rows indicating the mass balance for each metabolite. With GECKO, this approach is expanded by adding new rows to the matrix to represent the enzymes and new columns to represent each enzyme usage. Kinetic information, in the form of k_{cat} values, is included as stoichiometric coefficients to convert the metabolic flux in mmol/gDWh to the required enzyme usage in mmol/gDW. The protein level is included as an upper bound for each enzyme usage, ensuring that the desired constraint on each flux is respected.

To create your own GECKO model, k_{cat} values of reactions, the molecular weight of proteins, and protein activity information will be required and can be directly changed and included in your GEM, because different metabolic groups have different k_{cat} values and molecular weight distribution [80]. All molecular and enzymatic parameters could be automatically retrieved from the BRENDA database [84] and/or the UNIPROT database [85].

The GECKO toolbox is dependent on MATLAB and other packages. There is an option in Python to work with GECKO models using the MEWpy package. Both require ecModels and normalized proteomics data.

2.2.4. Strategies for Integrating Proteomics and Transcriptomics Data

The integration of proteomics and transcriptomics data has become a crucial part of modern Systems Biology research. The combination of these two omics data types can provide a more comprehensive understanding of biological systems. Here, we will elaborate on the three strategies for integrating proteomics and transcriptomics data.

One of the challenges in integrating transcriptomics and proteomics data is the difficulty in obtaining the same sets of differentially expressed genes and differentially expressed proteins. This is often due to differences in the timing of sample collection for RNA and protein analysis. RNA samples are typically collected at the transcription stage, whereas protein samples are collected at the translation stage. As a result, there can be significant differences in the expression patterns of genes and proteins between these two stages. Furthermore, RNA and protein stability can also differ, which can further complicate the comparison between these two types of data.

To overcome this challenge, it is important to carefully plan the experimental design and sample collection protocols. Ideally, the samples for both RNA and protein analysis should be collected at the same time point and under the same conditions. If this is not possible, researchers can try to account for the differences between RNA and protein data by using statistical methods to normalize the data or by applying machine learning algorithms to identify patterns of expression that are consistent across both data sets.

Differentially Expressed Genes and Proteins

The first strategy for integrating proteomics and transcriptomics data is Venn diagram analysis or Jaccard index calculation. In this strategy, differentially expressed genes (DEGs) and differentially abundant proteins (DAPs) are identified in the same biological conditions from both types of omics data. A Venn diagram is then used to identify overlapping genes and proteins, which can provide insights into the mechanisms underlying a certain biological condition. This strategy can be particularly useful for identifying key pathways or processes that are regulated by both proteins and transcripts. However, it is important to ensure that the transcription and translation processes are aligned at the time of collection to avoid false positives.

There are articles that show that there is not a great intersection between DEGs and DAPs [86–95], and others showing a successful intersection [96–100]. This discrepancy is based on the following: (i) induced and repressed proteins behaving differently, revealing regulatory and kinetic differences in protein synthesis and turnover; (ii) taking into account the transcription–translation delay when comparing protein and mRNA levels during dynamic adaptation; (iii) protein variation being mainly influenced by mRNA concentration in a new steady state [99]. Furthermore, if there is a significant overlap between DEGs and DAPs or if there is a GO/KEGG enrichment common to both, using the intersection list is the most appropriate approach. If there are many differences, the most common strategy is to construct a Venn diagram of the GO processes and KEGG pathways to identify the similarities and differences.

Observing Delays between Omics Data

The second strategy for integrating proteomics and transcriptomics data is scatter plot analysis. This strategy involves plotting the log of the genes fold change by the log of the proteins fold change. By observing the scatter plot, we can identify whether there is a correlation between the proteomics and the transcriptomics data. A positive correlation in the scatter plot suggests agreement in data extraction and provides a better understanding of the mechanisms underlying the biological condition. Scatter plot analysis can also be useful for identifying genes or proteins that do not have a direct correlation between their expression and protein levels based on the problems as cited before.

When we scatter plot the log fold change from a differential genes expression test (logFCt) versus the log fold change of the protein levels (logFCp), we can observe different patterns that can provide insights into the agreement or disagreement between the transcriptomics and the proteomics data. In this case, the logFCt and logFCp values increase or decrease together (Figure 2A), meaning that the genes and proteins are co-regulated in the biological system. This pattern suggests that there is a good agreement between the transcriptomics and the proteomics data and the protein abundance changes can be explained by changes in gene expression. If you see a concentration of points above (Figure 2B) and below (Figure 2C) the 45-degree line in the scatterplot of logFCt versus logFCp, it means that there is a disagreement between the changes in gene expression and protein levels for some genes/proteins. In other words, the gene expression and protein levels of these genes/proteins do not show a consistent pattern across the biological condition studied. This could be due to various factors such as post-transcriptional regulation, protein stability, or differences in the sampling and processing methods between the two omics (transcriptomics and proteomics) data sets.

The R package ReactomeGSA [101] includes a function named "plot_correlations()" which generates a comparative scatter plot of transcriptomics and proteomics data, allowing for a quick assessment of the similarity between the two data sets at the pathway level.

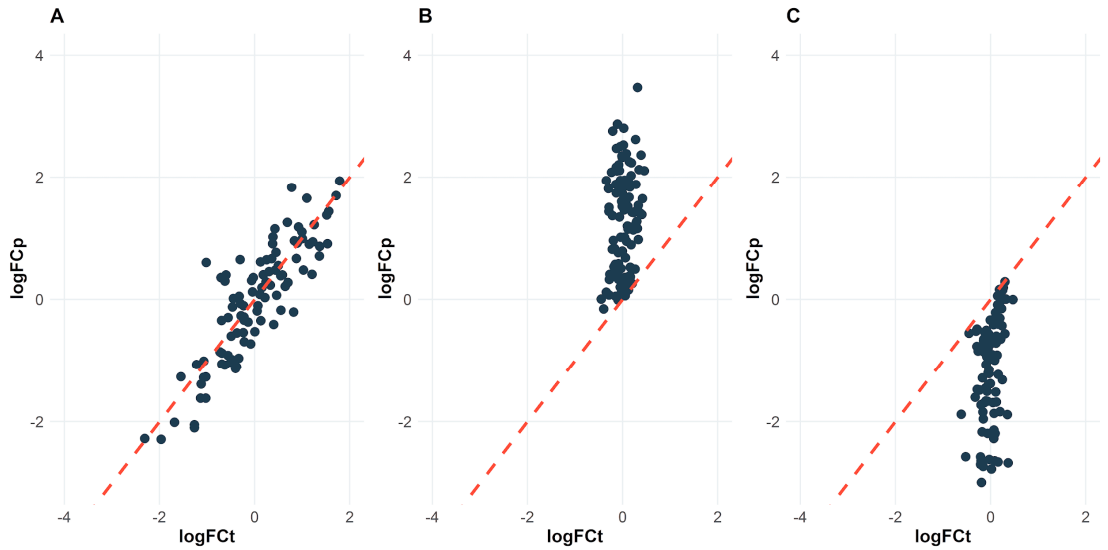


Figure 2. Scatter plot between the log fold change from a differential gene expression test (logFCt) and the log fold change of the protein levels (logFCp) in three scenarios: (**A**) high association between transcriptomics and proteomics data; (**B**,**C**) disagreement between the changes in gene expression and protein levels for some genes/proteins. The red dashed 45-degree line indicates the theoretical correspondence where changes in gene expression at the RNA-Seq level would be equally reflected at the protein level.

Interactome Analysis

The third strategy for integrating proteomics and transcriptomics data is interactome analysis. With this strategy, we could generate an interactome from differentially abundant proteins (DAPs) or differentially expressed genes (DEGs), identify submodules and hubs, and apply the fold change in gene expression to the interactome. This approach can help to identify functional relationships between different proteins and genes, providing a more comprehensive understanding of a biological system. Interactome analysis can also help to identify potential targets for further analysis, such as drugs or therapies.

First, from the list of DEGs, for example, we can generate an interactome using a protein–protein interaction (PPI) database, such as BioGRID [102] or STRING [103]. From the generated interactome, we can extract network metrics and identify hub genes and also submodules. With the application of the fold change in gene expression in the network, we can identify submodules with a predominance of a biological condition and apply biological enrichment to support such predominance.

Second, we can integrate this type of analysis based on co-expression analysis as well. Once the co-expression modules have been identified, it is possible to construct an interactome network that represents the interactions between the proteins encoded by these genes. The interactome can then be visualized and analyzed using network analysis tools such as Cytoscape, which allows for the identification of subnetworks, central nodes, and pathways that are enriched for the genes of interest, as well as of potential drug targets or biomarkers for diseases.

2.3. Machine Learning Methods Based on Omics Data

In this section, we will explore various methods that utilize omics data such as transcriptomics, proteomics, and metabolomics data for supervised and unsupervised machine learning algorithms (Figure 3). Supervised learning methods involve machine learning algorithms that use known data–outcome pairs as examples, whereas unsupervised methods operate on data sets without an outcome variable or prior knowledge of relationships between observations, dealing with unlabeled data. These strategies employ different methodologies and can involve the use of one, two, or three types of omics data. Additionally, numerous studies and reviews have extended this discussion to cover other

strategies and provide valuable tools to apply these methodologies with a greater emphasis on machine learning and multi-omics data [4,6,104–108].

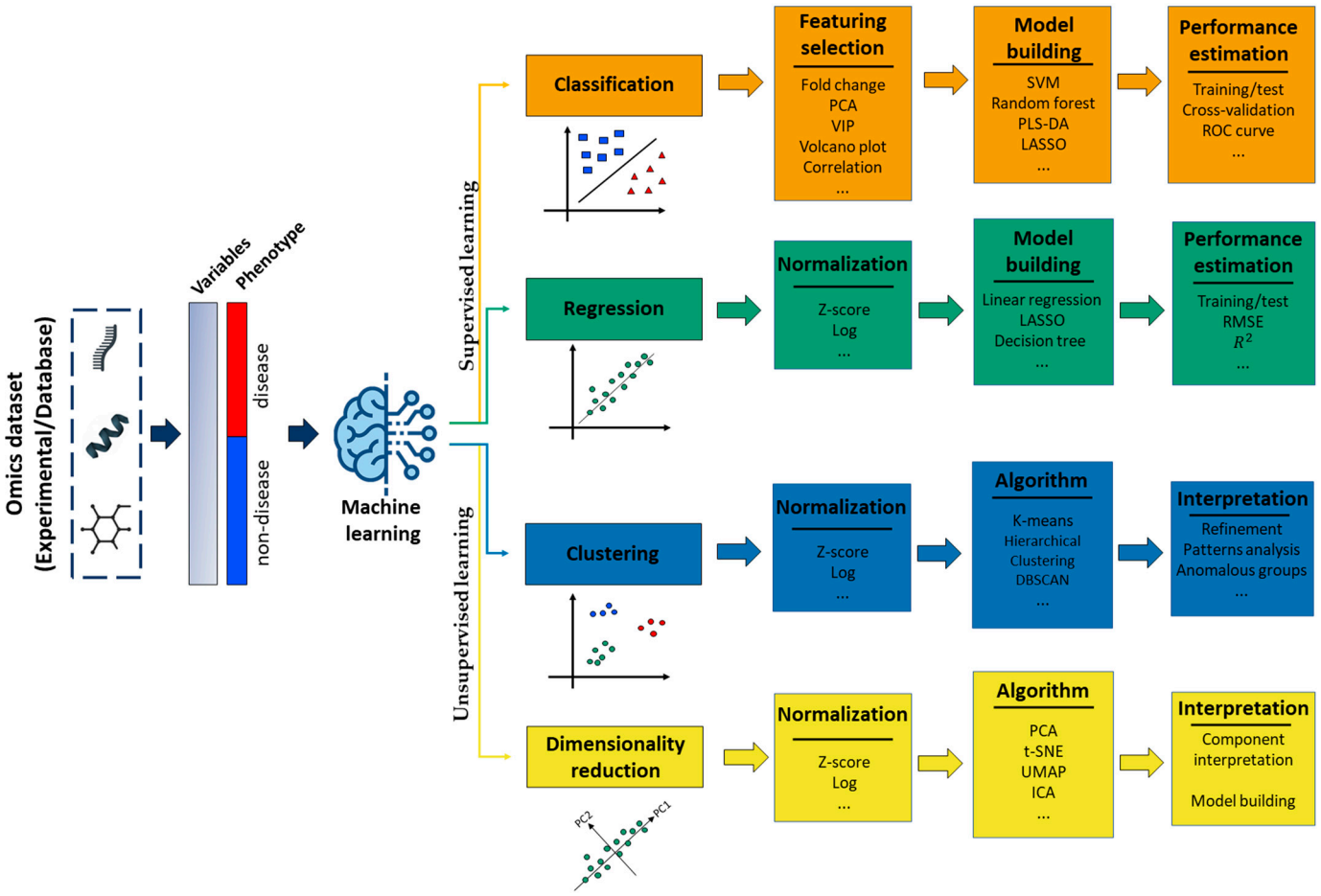


Figure 3. The pipeline illustrates the differences between supervised and unsupervised learning strategies applied to omics data. Legend: PCA: Principal Component Analysis; t-SNE: t-Distributed Stochastic Neighbor Embedding; UMAP: Uniform Manifold Approximation and Projection; ICA: Independent Component Analysis; SVM: Support Vector Machines; PLS-DA: Partial Least-Squares Discriminant Analysis; LASSO: Least Absolute Shrinkage and Selection Operator; RMSE: Root-Mean-Square Error; VIP: Variable Importance in Projection; ROC Curve: Receiver Operating Characteristic Curve.

2.3.1. Transcriptomics Data

Gene expression data can provide insights into the complex interplay between genes and cellular processes. With the advent of high-throughput technologies such as microarrays and RNA sequencing, it is now possible to generate large-scale gene expression data sets for a wide range of biological systems. In recent years, machine learning algorithms have emerged as a powerful tool for analyzing gene expression data. By leveraging the computational power of machine learning, researchers can uncover complex patterns and relationships within data that would be difficult or impossible to detect using traditional statistical methods, such as unsupervised machine learning methods.

In the literature, numerous studies have demonstrated the successful application of classification methods to predict cancer and cell types by utilizing gene expression data from microarray or bulk RNA-Seq data [109–115] and single-cell transcriptomics [116–122]. The process of employing gene expression data commences with the selection of suitable data sets, which can be accessed from databases like GEO [123], TCGA [124], and SRA [124]. Subsequently, pre-processing is carried out on the corresponding metadata for the clas-

sification algorithms, such as data from patients with or without cancer. Following this, the data set may or may not undergo a feature selection process followed by the application of a sampling technique to reach the final classification model, but a cross-validation step is essential. Each of these steps uses different data sets, pre-processing steps, model training and prediction algorithms, along with different k-fold cross-validation values, leading to varying values of accuracy, sensitivity (measuring the proportion of true positives accurately identified), and specificity (measuring the proportion of true negatives accurately identified).

It should be noted that the accuracy of machine learning classifiers using RNA-Seq data is dependent on various factors such as the type of sequencer used, the library preparation method, and the sample preparation technique. As a result, these classifiers exhibit varying levels of accuracy, with better performance observed at the transcript level compared to the gene level [125].

Biomedical researchers need to confirm the biological significance of the list or cluster of genes associated with a particular condition or developmental process that have been identified through comprehensive data analysis. To achieve this, they must evaluate the false-positive rate and conduct an autonomous biological validation. Northern blots and PCR-based methods are commonly employed to verify gene expression data, and these methods have the advantages of being able to screen through a large number of candidates relatively rapidly and perform quantitative measurements. Additionally, in situ hybridization and immunohistochemistry are used to determine the precise tissue in which the candidate genes are expressed [126]. Although these methods are not typically quantitative or high-throughput, they can be used to screen a large number of candidate genes and, in some cases, be performed in a quantitative manner.

2.3.2. Proteomics Data

With the advent of high-throughput technologies, such as mass spectrometry (MS) and protein microarrays, it is now possible to identify and quantify thousands of proteins in a single experiment. However, the sheer volume and complexity of proteomics data presents a challenge for traditional statistical and computational methods. Machine learning algorithms offer a promising solution for analyzing and interpreting these data, enabling researchers to extract meaningful information about protein function, interactions, and disease mechanisms. In this section, we will discuss how machine learning algorithms can be applied to proteomics data, including feature selection, classification, and clustering methods.

According to the literature, the machine learning algorithms used for proteomics data are based on retention time prediction, MS/MS spectrum prediction, the identification of peptides, biomarker identification, bias reduction during data processing, secondary structure prediction, protein toxicity prediction, protein function, and protein interactions [127–142]. Considering a comprehensive and reliable data integration, processing proteomics data and utilizing machine learning algorithms to accurately identify and analyze proteins require a consolidated pipeline. Due to the complexity of proteomics data, a comprehensive approach to data processing, normalization, quality control, feature extraction, and statistical analysis is necessary [7]. Without a consolidated pipeline, the analysis of proteomics data can be prone to errors and inconsistencies, leading to inaccurate results.

Identifying a biomarker from machine learning models based on proteomics data poses several challenges, including the need for validation in independent tests and the demonstration of clinical utility. These challenges are crucial in the translation of a promising biomarker candidate into a clinical tool. Validating the findings in independent tests is essential to ensure that the identified biomarker is reliable and reproducible. Demonstrating its clinical utility is also necessary to prove that the biomarker can effectively diagnose, monitor, or predict disease outcomes. These challenges highlight the importance of rigorous testing and validation before the implementation of any biomarker-based clinical assay.

2.3.3. Metabolomics Data

Machine learning has been applied to metabolomics data to identify biomarkers associated with diseases, understand metabolic pathways, contribute to the development of biotechnologies, and predict drug responses. PCA is one of the most widely used techniques to analyze metabolomics data, where metabolites are reduced into principal components that represent the majority of the variability in the data. PCA can identify metabolites that are significantly associated with different biological conditions, such as healthy and disease states, and can be used to cluster samples based on their metabolic profiles [143–148].

The variable in projection (VIP) score from the partial least-squares–discriminant analysis (PLS-DA) is another important method in metabolomics data analysis that can be used to identify relevant metabolites associated with a particular biological condition [149–155]. The VIP score is a measure of the contribution of each metabolite to the separation between two groups and is calculated by applying a supervised learning algorithm to the data, such as the PLS-DA algorithm. The MetaboAnalyst online platform is an example of a tool that allows researchers to apply VIP score analysis to their metabolomics data, as well as to perform other statistical and machine learning analyses.

Metabolomics data have been widely used in machine learning models to predict different types of cancer as well, as they provide valuable information about metabolic pathways altered in cancer cells. For instance, recent studies used metabolomics data in combination with machine learning algorithms to distinguish between different types of cancer, including lung, breast, non-Hodgkin's lymphoma, and ovarian cancer, and non-cancer conditions, such as coronavirus disease (COVID-19), type-2 diabetes, acute myocardial ischemia, schizophrenia, and autism in relation to gestational age [156]. Moreover, the identification of biomarkers in several of these cited studies has the potential to improve disease diagnosis, treatment, and monitoring. It allows for the discovery of complex relationships between metabolites and biological conditions that may not be easily detected through traditional methods.

Transcriptomic data are still used to predict metabolite concentrations using machine learning models. Auslander N. et al. (2016) [157] demonstrated that the levels of a wide range of metabolites in breast cancer can be successfully predicted from the transcriptome. The authors developed a Support Vector Machine (SVM) classifier to identify reaction-gene–metabolite (RGM) triplets where the gene and metabolite involved in the same reaction showed a significant association, whether positive or negative.

Selecting the suitable ML algorithm (including linear regression, logistic regression, support vector machines, k-nearest neighbors, decision trees, random forests, neural networks, and deep learning) plays a critical role in the achievement of a metabolomics study. It is crucial for researchers to be knowledgeable about the advantages of various ML approaches and to choose the most appropriate one based on their requirements to obtain accurate and easily understandable results [156,158].

3. Conclusions

In conclusion, integrating multiple omics data types is a powerful approach that can provide a more comprehensive understanding of biological systems. Transcriptomics, proteomics, and metabolomics offer complementary views of biological processes at the RNA, protein, and metabolite levels, respectively. By combining these data types, researchers can gain insights into complex biological phenomena that may not be possible with any single omics data type alone.

There are several strategies for integrating omics data, including co-expression analysis, pathway analysis, and network analysis. Each strategy has its strengths and weaknesses, and the choice of the approach will depend on the specific research question being addressed.

Overall, the integration of omics data is a rapidly evolving field, with new methods and tools being developed to address the challenges of analyzing and interpreting large

and complex data sets. As technology continues to advance, the integration of omics data is likely to become even more important for understanding molecular mechanisms in biology.

In this article, we reviewed several methods for integrating omics data and provided examples of their application in various biological contexts. Moreover, we explored the applications of omics data in machine learning studies. We hope that this review will inspire further research in this field and lead to new insights into the complex interplay between genes, proteins, and metabolites in living systems.

Author Contributions: P.H.G.S.: conceptualization, visualization, writing—original draft, writing—review and editing. N.C.d.M.: conceptualization, writing—original draft, writing—review and editing. A.M.P.: conceptualization, writing—review and editing. L.M.d.C.: conceptualization, visualization, supervision, writing—original draft, writing—review and editing, project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financed by the São Paulo Research Foundation (FAPESP) through grant 2022/14179-1.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors have not declared any conflicts of interest.

References

1. Hasin, Y.; Seldin, M.; Lusis, A. Multi-Omics Approaches to Disease. *Genome Biol.* **2017**, *18*, 83. [CrossRef]
2. Hasanzad, M.; Sarhangi, N.; Ehsani Chimeh, S.; Ayati, N.; Afzali, M.; Khatami, F.; Nikfar, S.; Aghaei Meybodi, H.R. Precision Medicine Journey through Omics Approach. *J. Diabetes Metab. Disord.* **2022**, *21*, 881–888. [CrossRef] [PubMed]
3. Karczewski, K.J.; Snyder, M.P. Integrative Omics for Health and Disease. *Nat. Rev. Genet.* **2018**, *19*, 299–310. [CrossRef] [PubMed]
4. Picard, M.; Scott-Boyer, M.P.; Bodein, A.; Périn, O.; Droit, A. Integration Strategies of Multi-Omics Data for Machine Learning Analysis. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 3735–3746. [CrossRef] [PubMed]
5. Rozanova, S.; Barkovits, K.; Nikolov, M.; Schmidt, C.; Urlaub, H.; Marcus, K. Quantitative Mass Spectrometry-Based Proteomics: An Overview. *Methods Mol. Biol.* **2021**, *2228*, 85–116. [CrossRef] [PubMed]
6. Reel, P.S.; Reel, S.; Pearson, E.; Trucco, E.; Jefferson, E. Using Machine Learning Approaches for Multi-Omics Data Analysis: A Review. *Biotechnol. Adv.* **2021**, *49*, 107739. [CrossRef] [PubMed]
7. de Carvalho, L.M.; Borelli, G.; Camargo, A.P.; de Assis, M.A.; de Ferraz, S.M.F.; Fiamenghi, M.B.; José, J.; Mofatto, L.S.; Nagamatsu, S.T.; Persinoti, G.F.; et al. Bioinformatics Applied to Biotechnology: A Review towards Bioenergy Research. *Biomass Bioenergy* **2019**, *123*, 195–224. [CrossRef]
8. Wang, R.; Li, B.; Lam, S.M.; Shui, G. Integration of Lipidomics and Metabolomics for In-Depth Understanding of Cellular Mechanism and Disease Progression. *J. Genet. Genom.* **2020**, *47*, 69–83. [CrossRef] [PubMed]
9. Idle, J.R.; Gonzalez, F.J. Metabolomics. *Cell Metab.* **2007**, *6*, 348–351. [CrossRef]
10. Athieniti, E.; Spyrou, G.M. A Guide to Multi-Omics Data Collection and Integration for Translational Medicine. *Comput. Struct. Biotechnol. J.* **2023**, *21*, 134–149. [CrossRef]
11. Budzinski, I.G.F.; de Moraes, F.E.; Cataldi, T.R.; Franceschini, L.M.; Labate, C.A. Network Analyses and Data Integration of Proteomics and Metabolomics from Leaves of Two Contrasting Varieties of Sugarcane in Response to Drought. *Front. Plant Sci.* **2019**, *10*, 446557. [CrossRef] [PubMed]
12. Dimitrakopoulos, C.; Hindupur, S.K.; Hafliger, L.; Behr, J.; Montazeri, H.; Hall, M.N.; Beerenwinkel, N. Network-Based Integration of Multi-Omics Data for Prioritizing Cancer Genes. *Bioinformatics* **2018**, *34*, 2441–2448. [CrossRef] [PubMed]
13. Misra, B.B.; Langefeld, C.; Olivier, M.; Cox, L.A. Integrated Omics: Tools, Advances and Future Approaches. *J. Mol. Endocrinol.* **2019**, *62*, R21–R45. [CrossRef] [PubMed]
14. Cambiaghi, A.; Ferrario, M.; Masseroli, M. Analysis of Metabolomic Data: Tools, Current Strategies and Future Challenges for Omics Data Integration. *Brief. Bioinform.* **2017**, *18*, 498–510. [CrossRef] [PubMed]
15. Bersanelli, M.; Mosca, E.; Remondini, D.; Giampieri, E.; Sala, C.; Castellani, G.; Milanesi, L. Methods for the Integration of Multi-Omics Data: Mathematical Aspects. *BMC Bioinform.* **2016**, *17*, 167–177. [CrossRef] [PubMed]
16. Fukushima, A.; Kusano, M.; Redestig, H.; Arita, M.; Saito, K. Integrated Omics Approaches in Plant Systems Biology. *Curr. Opin. Chem. Biol.* **2009**, *13*, 532–538. [CrossRef] [PubMed]
17. Hrovatin, K.; Fischer, D.S.; Theis, F.J. Toward Modeling Metabolic State from Single-Cell Transcriptomics. *Mol. Metab.* **2022**, *57*, 101396. [CrossRef] [PubMed]

18. Sun, Y.; Liu, Z.; Fu, Y.; Yang, Y.; Lu, J.; Pan, M.; Wen, T.; Xie, X.; Bai, Y.; Ge, Q. Single-Cell Multi-Omics Sequencing and Its Application in Tumor Heterogeneity. *Brief. Funct. Genom.* **2023**, *22*, 313–328. [CrossRef] [PubMed]
19. Dimitriu, M.A.; Lazar-Contes, I.; Roszkowski, M.; Mansuy, I.M. Single-Cell Multiomics Techniques: From Conception to Applications. *Front. Cell Dev. Biol.* **2022**, *10*, 854317. [CrossRef]
20. Flynn, E.; Almonte-Loya, A.; Fragiadakis, G.K. Single-cell multiomics. *Annu. Rev. Biomed. Data Sci.* **2023**, *6*, 313–337. [CrossRef]
21. Adossa, N.; Khan, S.; Rytkönen, K.T.; Elo, L.L. Computational Strategies for Single-Cell Multi-Omics Integration. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 2588–2596. [CrossRef]
22. Baysoy, A.; Bai, Z.; Satija, R.; Fan, R. The Technological Landscape and Applications of Single-Cell Multi-Omics. *Nat. Rev. Mol. Cell Biol.* **2023**, *24*, 695–713. [CrossRef] [PubMed]
23. Chen, G.; Yu, R.; Chen, X. Editorial: Integrative Analysis of Single-Cell and/or Bulk Multi-Omics Sequencing Data. *Front. Genet.* **2023**, *13*, 1121999. [CrossRef] [PubMed]
24. Huang, Y.; Mohanty, V.; Dede, M.; Tsai, K.; Daher, M.; Li, L.; Rezvani, K.; Chen, K. Characterizing Cancer Metabolism from Bulk and Single-Cell RNA-Seq Data Using METAFlux. *Nat. Commun.* **2023**, *14*, 4883. [CrossRef] [PubMed]
25. Cheng, C.; Chen, W.; Jin, H.; Chen, X. A Review of Single-Cell RNA-Seq Annotation, Integration, and Cell-Cell Communication. *Cells* **2023**, *12*, 1970. [CrossRef] [PubMed]
26. Serin, E.A.R.; Nijveen, H.; Hilhorst, H.W.M.; Ligterink, W. Learning from Co-Expression Networks: Possibilities and Challenges. *Front. Plant Sci.* **2016**, *7*, 185898. [CrossRef]
27. Zeng, F.; Shi, M.; Xiao, H.; Chi, X. WGCNA-Based Identification of Hub Genes and Key Pathways Involved in Nonalcoholic Fatty Liver Disease. *BioMed Res. Int.* **2021**, *2021*, 5633211. [CrossRef]
28. Choi, H.; Na, K.J. Integrative Analysis of Imaging and Transcriptomic Data of the Immune Landscape Associated with Tumor Metabolism in Lung Adenocarcinoma: Clinical and Prognostic Implications. *Theranostics* **2018**, *8*, 1956–1965. [CrossRef]
29. Liu, P.; Luo, J.; Zheng, Q.; Chen, Q.; Zhai, N.; Xu, S.; Xu, Y.; Jin, L.; Xu, G.; Lu, X.; et al. Integrating Transcriptome and Metabolome Reveals Molecular Networks Involved in Genetic and Environmental Variation in Tobacco. *DNA Res.* **2020**, *27*, dsaa006. [CrossRef]
30. Amiri, F.; Moghadam, A.; Tahmasebi, A.; Niazi, A. Identification of Key Genes Involved in Secondary Metabolite Biosynthesis in Digitalis Purpurea. *PLoS ONE* **2023**, *18*, e0277293. [CrossRef]
31. Moghadam, A.; Foroozan, E.; Tahmasebi, A.; Taghizadeh, M.S.; Bolhassani, M.; Jafari, M. System Network Analysis of Rosmarinus Officinalis Transcriptome and Metabolome—Key Genes in Biosynthesis of Secondary Metabolites. *PLoS ONE* **2023**, *18*, e0282316. [CrossRef]
32. Ponsuksili, S.; Trakooljul, N.; Hadlich, F.; Methling, K.; Lalk, M.; Murani, E.; Wimmers, K. Genetic Regulation of Liver Metabolites and Transcripts Linking to Biochemical-Clinical Parameters. *Front. Genet.* **2019**, *10*, 419414. [CrossRef] [PubMed]
33. Zhu, Y.; Mordaunt, C.E.; Durbin-Johnson, B.P.; Caudill, M.A.; Malysheva, O.V.; Miller, J.W.; Green, R.; James, S.J.; Melnyk, S.B.; Fallin, M.D.; et al. Expression Changes in Epigenetic Gene Pathways Associated with One-Carbon Nutritional Metabolites in Maternal Blood from Pregnancies Resulting in Autism and Non-Typical Neurodevelopment. *Autism Res.* **2021**, *14*, 11–28. [CrossRef] [PubMed]
34. Hoang, L.T.; Domingo-Sabugo, C.; Starren, E.S.; Willis-Owen, S.A.G.; Morris-Rosendahl, D.J.; Nicholson, A.G.; Cookson, W.O.C.M.; Moffatt, M.F. Metabolomic, Transcriptomic and Genetic Integrative Analysis Reveals Important Roles of Adenosine Diphosphate in Haemostasis and Platelet Activation in Non-Small-Cell Lung Cancer. *Mol. Oncol.* **2019**, *13*, 2406–2421. [CrossRef] [PubMed]
35. Wang, Z.; Zhang, X.; He, S.; Rehman, A.; Jia, Y.; Li, H.; Pan, Z.; Geng, X.; Gao, Q.; Wang, L.; et al. Transcriptome Co-Expression Network and Metabolome Analysis Identifies Key Genes and Regulators of Proanthocyanidins Biosynthesis in Brown Cotton. *Front. Plant Sci.* **2022**, *12*, 822198. [CrossRef] [PubMed]
36. Green, R.E.; Lord, J.; Scelsi, M.A.; Xu, J.; Wong, A.; Naomi-James, S.; Handy, A.; Gilchrist, L.; Williams, D.M.; Parker, T.D.; et al. Investigating Associations between Blood Metabolites, Later Life Brain Imaging Measures, and Genetic Risk for Alzheimer's Disease. *Alzheimer's Res. Ther.* **2023**, *15*, 38. [CrossRef] [PubMed]
37. Green, R.; Lord, J.; Xu, J.; Maddock, J.; Kim, M.; Dobson, R.; Legido-Quigley, C.; Wong, A.; Richards, M.; Proitsi, P. Metabolic Correlates of Late Midlife Cognitive Outcomes: Findings from the 1946 British Birth Cohort. *Brain Commun.* **2022**, *4*, fcab291. [CrossRef] [PubMed]
38. Carson, C.; Lawson, H.A. Genetic Background and Diet Affect Brown Adipose Gene Coexpression Networks Associated with Metabolic Phenotypes. *Physiol. Genom.* **2020**, *52*, 223–233. [CrossRef] [PubMed]
39. Zhao, X.; Ge, W.; Miao, Z. Integrative Metabolomic and Transcriptomic Analyses Reveals the Accumulation Patterns of Key Metabolites Associated with Flavonoids and Terpenoids of *Gynostemma pentaphyllum* (Thunb.) Makino. *Sci. Rep.* **2024**, *14*, 8644. [CrossRef]
40. Xie, Z.; Wang, J.; Wang, W.; Wang, Y.; Xu, J.; Li, Z.; Zhao, X.; Fu, B. Integrated Analysis of the Transcriptome and Metabolome Revealed the Molecular Mechanisms Underlying the Enhanced Salt Tolerance of Rice Due to the Application of Exogenous Melatonin. *Front. Plant Sci.* **2021**, *11*, 618680. [CrossRef]
41. Zhou, Z.; Liu, J.; Meng, W.; Sun, Z.; Tan, Y.; Liu, Y.; Tan, M.; Wang, B.; Yang, J. Integrated Analysis of Transcriptome and Metabolome Reveals Molecular Mechanisms of Rice with Different Salinity Tolerances. *Plants* **2023**, *12*, 3359. [CrossRef] [PubMed]

42. Zhao, D.; Zhang, Y.; Ren, H.; Shi, Y.; Dong, D.; Li, Z.; Cui, G.; Shen, Y.; Mou, Z.; Kennelly, E.J.; et al. Multi-Omics Analysis Reveals the Evolutionary Origin of Diterpenoid Alkaloid Biosynthesis Pathways in Aconitum. *J. Integr. Plant Biol.* **2023**, *65*, 2320–2335. [CrossRef] [PubMed]
43. Nikiforova, V.J.; Daub, C.O.; Hesse, H.; Willmitzer, L.; Hoefgen, R. Integrative Gene-Metabolite Network with Implemented Causality Deciphers Informational Fluxes of Sulphur Stress Response. *J. Exp. Bot.* **2005**, *56*, 1887–1896. [CrossRef] [PubMed]
44. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13*, 2498. [CrossRef]
45. Csárdi, G.; Nepusz, T.; Müller, K.; Horvát, S.; Traag, V.; Zanini, F.; Noom, D. Igraph for R: R Interface of the Igraph Library for Graph Theory and Network Analysis, Version 1.4. *Zenodo* **2023**. [CrossRef]
46. Rai, A.; Rai, M.; Kamochi, H.; Mori, T.; Nakabayashi, R.; Nakamura, M.; Suzuki, H.; Saito, K.; Yamazaki, M. Multiomics-Based Characterization of Specialized Metabolites Biosynthesis in Cornus Officinalis. *DNA Res.* **2020**, *27*, dsaa009. [CrossRef] [PubMed]
47. Saito, K.; Hirai, M.Y.; Yonekura-Sakakibara, K. Decoding Genes with Coexpression Networks and Metabolomics—"Majority Report by Precogs". *Trends Plant Sci.* **2008**, *13*, 36–43. [CrossRef] [PubMed]
48. Cline, M.S.; Smoot, M.; Cerami, E.; Kuchinsky, A.; Landys, N.; Workman, C.; Christmas, R.; Avila-Campilo, I.; Creech, M.; Gross, B.; et al. Integration of Biological Networks and Gene Expression Data Using Cytoscape. *Nat. Protoc.* **2007**, *2*, 2366–2382. [CrossRef] [PubMed]
49. Liu, X.L.; Ming, Y.N.; Zhang, J.Y.; Chen, X.Y.; Zeng, M.D.; Mao, Y.M. Gene-Metabolite Network Analysis in Different Nonalcoholic Fatty Liver Disease Phenotypes. *Exp. Mol. Med.* **2017**, *49*, e283. [CrossRef]
50. Mounet, F.; Moing, A.; Garcia, V.; Petit, J.; Maucourt, M.; Deborde, C.; Bernillon, S.; Le Gall, G.; Colquhoun, I.; Defernez, M.; et al. Gene and Metabolite Regulatory Network Analysis of Early Developing Fruit Tissues Highlights New Candidate Genes for the Control of Tomato Fruit Composition and Development. *Plant Physiol.* **2009**, *149*, 1505–1528. [CrossRef]
51. Mercatelli, D.; Lopez-Garcia, G.; Giorgi, F.M. Corto: A Lightweight R Package for Gene Network Inference and Master Regulator Analysis. *Bioinformatics* **2020**, *36*, 3916–3917. [CrossRef] [PubMed]
52. Cavicchioli, M.V.; Santorsola, M.; Balboni, N.; Mercatelli, D.; Giorgi, F.M. Prediction of Metabolic Profiles from Transcriptomics Data in Human Cancer Cell Lines. *Int. J. Mol. Sci.* **2022**, *23*, 3867. [CrossRef] [PubMed]
53. Wang, B.; Mezlini, A.M.; Demir, F.; Fiume, M.; Tu, Z.; Brudno, M.; Haibe-Kains, B.; Goldenberg, A. Similarity Network Fusion for Aggregating Data Types on a Genomic Scale. *Nat. Methods* **2014**, *11*, 333–337. [CrossRef] [PubMed]
54. Miranda, J.; Paules, C.; Noell, G.; Youssef, L.; Paternina-Caicedo, A.; Crovetto, F.; Cañellas, N.; Garcia-Martín, M.L.; Amigó, N.; Eixarch, E.; et al. Similarity Network Fusion to Identify Phenotypes of Small-for-Gestational-Age Fetuses. *iScience* **2023**, *26*, 107620. [CrossRef] [PubMed]
55. Raphael, B.J.; Hruban, R.H.; Aguirre, A.J.; Moffitt, R.A.; Yeh, J.J.; Stewart, C.; Robertson, A.G.; Cherniack, A.D.; Gupta, M.; Getz, G.; et al. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell* **2017**, *32*, 185–203.e13. [CrossRef]
56. Blimkie, T.; Lee, A.H.Y.; Hancock, R.E.W. MetaBridge: An Integrative Multi-Omics Tool for Metabolite-Enzyme Mapping. *Curr. Protoc. Bioinform.* **2020**, *70*, e98. [CrossRef]
57. Caspi, R.; Billington, R.; Keseler, I.M.; Kothari, A.; Krummenacker, M.; Midford, P.E.; Ong, W.K.; Paley, S.; Subhraveti, P.; Karp, P.D. The MetaCyc Database of Metabolic Pathways and Enzymes—A 2019 Update. *Nucleic Acids Res.* **2020**, *48*, D445–D453. [CrossRef] [PubMed]
58. Wu, T.; Hu, E.; Xu, S.; Chen, M.; Guo, P.; Dai, Z.; Feng, T.; Zhou, L.; Tang, W.; Zhan, L.; et al. ClusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data. *Innovation* **2021**, *2*, 100141. [CrossRef]
59. Fang, Z.; Liu, X.; Peltz, G. GSEApy: A Comprehensive Package for Performing Gene Set Enrichment Analysis in Python. *Bioinformatics* **2023**, *39*, btac757. [CrossRef] [PubMed]
60. Pang, Z.; Chong, J.; Zhou, G.; De Lima Morais, D.A.; Chang, L.; Barrette, M.; Gauthier, C.; Jacques, P.É.; Li, S.; Xia, J. MetaboAnalyst 5.0: Narrowing the Gap between Raw Spectra and Functional Insights. *Nucleic Acids Res.* **2021**, *49*, W388–W396. [CrossRef]
61. Luo, W.; Brouwer, C. Pathview: An R/Bioconductor Package for Pathway-Based Data Integration and Visualization. *Bioinformatics* **2013**, *29*, 1830–1831. [CrossRef] [PubMed]
62. Argelaguet, R.; Velten, B.; Arnol, D.; Dietrich, S.; Zenz, T.; Marioni, J.C.; Buettner, F.; Huber, W.; Stegle, O. Multi-Omics Factor Analysis—A Framework for Unsupervised Integration of Multi-Omics Data Sets. *Mol. Syst. Biol.* **2018**, *14*, 8124. [CrossRef] [PubMed]
63. Dutta, N.K.; Tornheim, J.A.; Fukutani, K.F.; Paradkar, M.; Tiburcio, R.T.; Kinikar, A.; Valvi, C.; Kulkarni, V.; Pradhan, N.; Shivakumar, S.V.B.Y.; et al. Integration of Metabolomics and Transcriptomics Reveals Novel Biomarkers in the Blood for Tuberculosis Diagnosis in Children. *Sci. Rep.* **2020**, *10*, 19527. [CrossRef]
64. Clark, C.; Dayon, L.; Masoodi, M.; Bowman, G.L.; Popp, J. An Integrative Multi-Omics Approach Reveals New Central Nervous System Pathway Alterations in Alzheimer's Disease. *Alzheimer's Res. Ther.* **2021**, *13*, 71. [CrossRef]
65. Wang, H.; Liu, C.; Xie, X.; Niu, M.; Wang, Y.; Cheng, X.; Zhang, B.; Zhang, D.; Liu, M.; Sun, R.; et al. Multi-Omics Blood Atlas Reveals Unique Features of Immune and Platelet Responses to SARS-CoV-2 Omicron Breakthrough Infection. *Immunity* **2023**, *56*, 1410–1428.e8. [CrossRef] [PubMed]
66. Cavill, R.; Jennen, D.; Kleinjans, J.; Briedé, J.J. Transcriptomic and Metabolomic Data Integration. *Brief. Bioinform.* **2016**, *17*, 891–901. [CrossRef]

67. Gu, C.; Kim, G.B.; Kim, W.J.; Kim, H.U.; Lee, S.Y. Current Status and Applications of Genome-Scale Metabolic Models. *Genome Biol.* **2019**, *20*, 121. [CrossRef]
68. Larsson, I.; Uhlén, M.; Zhang, C.; Mardinoglu, A. Genome-Scale Metabolic Modeling of Glioblastoma Reveals Promising Targets for Drug Development. *Front. Genet.* **2020**, *11*, 459644. [CrossRef] [PubMed]
69. Karakitsou, E.; Foguet, C.; Contreras Mostazo, M.G.; Kurrle, N.; Schnütgen, F.; Michaelis, M.; Cinatl, J.; Marin, S.; Cascante, M. Genome-Scale Integration of Transcriptome and Metabolome Unveils Squalene Synthase and Dihydrofolate Reductase as Targets against AML Cells Resistant to Chemotherapy. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4059–4066. [CrossRef]
70. Sen, P.; Orešič, M. Integrating Omics Data in Genome-Scale Metabolic Modeling: A Methodological Perspective for Precision Medicine. *Metabolites* **2023**, *13*, 855. [CrossRef]
71. Kim, M.K.; Lun, D.S. Methods for Integration of Transcriptomic Data in Genome-Scale Metabolic Models. *Comput. Struct. Biotechnol. J.* **2014**, *11*, 59–65. [CrossRef] [PubMed]
72. Roshanzamir, F.; Robinson, J.L.; Cook, D.; Karimi-Jafari, M.H.; Nielsen, J. Metastatic Triple Negative Breast Cancer Adapts Its Metabolism to Destination Tissues While Retaining Key Metabolic Signatures. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2205456119. [CrossRef] [PubMed]
73. Orth, J.D.; Thiele, I.; Palsson, B.O. What Is Flux Balance Analysis? *Nat. Biotechnol.* **2010**, *28*, 245. [CrossRef]
74. Lewis, N.E.; Hixson, K.K.; Conrad, T.M.; Lerman, J.A.; Charusanti, P.; Polpitiya, A.D.; Adkins, J.N.; Schramm, G.; Purvine, S.O.; Lopez-Ferrer, D.; et al. Omic Data from Evolved *E. coli* Are Consistent with Computed Optimal Growth from Genome-Scale Models. *Mol. Syst. Biol.* **2010**, *6*, 390. [CrossRef] [PubMed]
75. Pereira, V.; Cruz, F.; Rocha, M. MEWpy: A Computational Strain Optimization Workbench in Python. *Bioinformatics* **2021**, *37*, 2494–2496. [CrossRef] [PubMed]
76. Lu, H.; Li, F.; Sánchez, B.J.; Zhu, Z.; Li, G.; Domenzain, I.; Marcišauskas, S.; Anton, P.M.; Lappa, D.; Lieven, C.; et al. A Consensus *S. cerevisiae* Metabolic Model Yeast8 and Its Ecosystem for Comprehensively Probing Cellular Metabolism. *Nat. Commun.* **2019**, *10*, 3586. [CrossRef]
77. Robinson, J.L.; Kocabaş, P.; Wang, H.; Cholley, P.E.; Cook, D.; Nilsson, A.; Anton, M.; Ferreira, R.; Domenzain, I.; Billa, V.; et al. An Atlas of Human Metabolism. *Sci. Signal.* **2020**, *13*, eaaz1482. [CrossRef] [PubMed]
78. Monk, J.M.; Lloyd, C.J.; Brunk, E.; Mih, N.; Sastry, A.; King, Z.; Takeuchi, R.; Nomura, W.; Zhang, Z.; Mori, H.; et al. IML1515, a Knowledgebase That Computes Escherichia Coli Traits. *Nat. Biotechnol.* **2017**, *35*, 904–908. [CrossRef] [PubMed]
79. Domenzain, I.; Sánchez, B.; Anton, M.; Kerkhoven, E.J.; Millán-Oropeza, A.; Henry, C.; Siewers, V.; Morrissey, J.P.; Sonnenschein, N.; Nielsen, J. Reconstruction of a Catalogue of Genome-Scale Metabolic Models with Enzymatic Constraints Using GECKO 2.0. *Nat. Commun.* **2022**, *13*, 3766. [CrossRef]
80. Sánchez, B.J.; Zhang, C.; Nilsson, A.; Lahtvee, P.; Kerkhoven, E.J.; Nielsen, J. Improving the Phenotype Predictions of a Yeast Genome-scale Metabolic Model by Incorporating Enzymatic Constraints. *Mol. Syst. Biol.* **2017**, *13*, 935. [CrossRef]
81. Zhou, J.; Zhuang, Y.; Xia, J. Integration of Enzyme Constraints in a Genome-Scale Metabolic Model of Aspergillus Niger Improves Phenotype Predictions. *Microb. Cell Factories* **2021**, *20*, 125. [CrossRef] [PubMed]
82. Arend, M.; Zimmer, D.; Xu, R.; Sommer, F.; Mühlhaus, T.; Nikoloski, Z. Proteomics and Constraint-Based Modelling Reveal Enzyme Kinetic Properties of Chlamydomonas Reinhardtii on a Genome Scale. *Nat. Commun.* **2023**, *14*, 4781. [CrossRef] [PubMed]
83. Wu, K.; Mao, Z.; Mao, Y.; Niu, J.; Cai, J.; Yuan, Q.; Yun, L.; Liao, X.; Wang, Z.; Ma, H. EcBSU1: A Genome-Scale Enzyme-Constrained Model of Bacillus Subtilis Based on the ECMpy Workflow. *Microorganisms* **2023**, *11*, 178. [CrossRef] [PubMed]
84. Placzek, S.; Schomburg, I.; Chang, A.; Jeske, L.; Ulbrich, M.; Tillack, J.; Schomburg, D. BRENDA in 2017: New Perspectives and New Tools in BRENDA. *Nucleic Acids Res.* **2017**, *45*, D380–D388. [CrossRef] [PubMed]
85. Bateman, A.; Martin, M.J.; Orchard, S.; Magrane, M.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E.H.; Britto, R.; Bye-A-Jee, H.; Cukura, A.; et al. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51*, D523–D531. [CrossRef]
86. Chen, J.; Luo, P.; Wang, C.; Yang, C.; Bai, Y.; He, X.; Zhang, Q.; Zhang, J.; Yang, J.; Wang, S.; et al. Integrated Single-Cell Transcriptomics and Proteomics Reveal Cellular-Specific Responses and Microenvironment Remodeling in Aristolochic Acid Nephropathy. *JCI Insight* **2022**, *7*, e157360. [CrossRef] [PubMed]
87. Alsagaby, S.A. Integration of Proteomics and Transcriptomics Data Sets Identifies Prognostic Markers in Chronic Lymphocytic Leukemia. *Majmaah J. Health Sci.* **2019**, *7*, 1. [CrossRef]
88. Higdon, R.; Kala, J.; Wilkins, D.; Yan, J.F.; Sethi, M.K.; Lin, L.; Liu, S.; Montague, E.; Janko, I.; Choiniere, J.; et al. Integrated Proteomic and Transcriptomic-Based Approaches to Identifying Signature Biomarkers and Pathways for Elucidation of Daoy and UW228 Subtypes. *Proteomes* **2017**, *5*, 5. [CrossRef]
89. Gygi, S.P.; Rochon, Y.; Franza, B.R.; Aebersold, R. Correlation between Protein and MRNA Abundance in Yeast. *Mol. Cell. Biol.* **1999**, *19*, 1720–1730. [CrossRef]
90. Yue, X.; Huan, P.; Hu, Y.; Liu, B. Integrated Transcriptomic and Proteomic Analyses Reveal Potential Mechanisms Linking Thermal Stress and Depressed Disease Resistance in the Turbot Scophthalmus Maximus. *Sci. Rep.* **2018**, *8*, 1896. [CrossRef]
91. Li, G.; Zhang, B.; Zhang, H.; Xu, A.; Qian, H. Integration of Transcriptomic and Proteomic Analyses Reveals New Insights into the Regulation of Immune Pathways in Midgut of Samia Ricini upon SariNPV Infection. *Insects* **2022**, *13*, 294. [CrossRef] [PubMed]
92. Sun, Z.; Liu, Y.; He, X.; Di, R.; Wang, X.; Ren, C.; Zhang, Z.; Chu, M. Integrative Proteomics and Transcriptomics Profiles of the Oviduct Reveal the Prolificacy-Related Candidate Biomarkers of Goats (*Capra hircus*) in Estrous Periods. *Int. J. Mol. Sci.* **2022**, *23*, 14888. [CrossRef] [PubMed]

93. Zhang, Q.; Li, Y.; Sun, L.; Chu, S.; Xu, H.; Zhou, X. Integration of Transcriptomic and Proteomic Analyses of *Rhododendron chrysanthum* Pall. in Response to Cold Stress in the Changbai Mountains. *Mol. Biol. Rep.* **2023**, *50*, 3607–3616. [CrossRef] [PubMed]
94. Miao, J.; Yang, Z.; Guo, W.; Liu, L.; Song, P.; Ding, C.; Guan, W. Integrative Analysis of the Proteome and Transcriptome in Gastric Cancer Identified LRP1B as a Potential Biomarker. *Biomark. Med.* **2022**, *16*, 1101–1111. [CrossRef] [PubMed]
95. Colak, D.; Alaiya, A.A.; Kaya, N.; Muiya, N.P.; AlHarazi, O.; Shinwari, Z.; Andres, E.; Dzimiri, N. Integrated Left Ventricular Global Transcriptome and Proteome Profiling in Human End-Stage Dilated Cardiomyopathy. *PLoS ONE* **2016**, *11*, e0162669. [CrossRef] [PubMed]
96. Du, Y.; Clair, G.C.; Al Alam, D.; Danopoulos, S.; Schnell, D.; Kitzmiller, J.A.; Misra, R.S.; Bhattacharya, S.; Warburton, D.; Mariani, T.J.; et al. Integration of Transcriptomic and Proteomic Data Identifies Biological Functions in Cell Populations from Human Infant Lung. *Am. J. Physiol. Cell. Mol. Physiol.* **2019**, *317*, L347. [CrossRef] [PubMed]
97. Peng, Z.; He, S.; Gong, W.; Xu, F.; Pan, Z.; Jia, Y.; Geng, X.; Du, X. Integration of Proteomic and Transcriptomic Profiles Reveals Multiple Levels of Genetic Regulation of Salt Tolerance in Cotton. *BMC Plant Biol.* **2018**, *18*, 128. [CrossRef] [PubMed]
98. Zheng, W.; Zhang, Y.; Sun, C.; Ge, S.; Tan, Y.; Shen, H.; Yang, P. A Multi-Omics Study of Human Testis and Epididymis. *Molecules* **2021**, *26*, 3345. [CrossRef]
99. Liu, Y.; Beyer, A.; Aebersold, R. On the Dependency of Cellular Protein Levels on MRNA Abundance. *Cell* **2016**, *165*, 535–550. [CrossRef]
100. Zhang, G.; Zhong, F.; Chen, L.; Qin, P.; Li, J.; Zhi, F.; Tian, L.; Zhou, D.; Lin, P.; Chen, H.; et al. Integrated Proteomic and Transcriptomic Analyses Reveal the Roles of Brucella Homolog of BAX Inhibitor 1 in Cell Division and Membrane Homeostasis of Brucella Suis S2. *Front. Microbiol.* **2021**, *12*, 632095. [CrossRef]
101. Griss, J.; Viteri, G.; Sidiropoulos, K.; Nguyen, V.; Fabregat, A.; Hermjakob, H. ReactomeGSA-Efficient Multi-Omics Comparative Pathway Analysis. *Mol. Cell. Proteom.* **2020**, *19*, 2115–2124. [CrossRef] [PubMed]
102. Stark, C.; Breitkreutz, B.J.; Reguly, T.; Boucher, L.; Breitkreutz, A.; Tyers, M. BioGRID: A General Repository for Interaction Datasets. *Nucleic Acids Res.* **2006**, *34*, D535. [CrossRef]
103. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.; et al. STRING V11: Protein–Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets. *Nucleic Acids Res.* **2019**, *47*, D607–D613. [CrossRef]
104. Wang, X.W.; Wang, T.; Schaub, D.P.; Chen, C.; Sun, Z.; Ke, S.; Hecker, J.; Maaser-Hecker, A.; Zeleznik, O.A.; Zeleznik, R.; et al. Benchmarking Omics-Based Prediction of Asthma Development in Children. *Respir. Res.* **2023**, *24*, 62. [CrossRef] [PubMed]
105. Arjmand, B.; Hamidpour, S.K.; Tayanloo-Beik, A.; Goodarzi, P.; Aghayan, H.R.; Adibi, H.; Larijani, B. Machine Learning: A New Prospect in Multi-Omics Data Analysis of Cancer. *Front. Genet.* **2022**, *13*, 824451. [CrossRef]
106. Wekesa, J.S.; Kimwele, M. A Review of Multi-Omics Data Integration through Deep Learning Approaches for Disease Diagnosis, Prognosis, and Treatment. *Front. Genet.* **2023**, *14*, 1199087. [CrossRef]
107. Li, R.; Li, L.; Xu, Y.; Yang, J. Machine Learning Meets Omics: Applications and Perspectives. *Brief. Bioinform.* **2022**, *23*, bbab460. [CrossRef] [PubMed]
108. Leite, D.M.C.; Brochet, X.; Resch, G.; Que, Y.A.; Neves, A.; Peña-Reyes, C. Computational Prediction of Inter-Species Relationships through Omics Data Analysis and Machine Learning. *BMC Bioinform.* **2018**, *19*, 151–159. [CrossRef]
109. Alharbi, F.; Vakanski, A. Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review. *Bioengineering* **2023**, *10*, 173. [CrossRef]
110. Bashiri, A.; Ghazisaeedi, M.; Safdari, R.; Shahmoradi, L.; Ehtesham, H. Improving the Prediction of Survival in Cancer Patients by Using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review. *Iran. J. Public Health* **2017**, *46*, 165.
111. Hijazi, H.; Chan, C. A Classification Framework Applied to Cancer Gene Expression Profiles. *J. Healthc. Eng.* **2013**, *4*, 255–283. [CrossRef]
112. Khalsan, M.; MacHado, L.R.; Al-Shamery, E.S.; Ajit, S.; Anthony, K.; Mu, M.; Agyeman, M.O. A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction. *IEEE Access* **2022**, *10*, 27522–27534. [CrossRef]
113. Ravindran, U.; Gunavathi, C. A Survey on Gene Expression Data Analysis Using Deep Learning Methods for Cancer Diagnosis. *Prog. Biophys. Mol. Biol.* **2023**, *177*, 1–13. [CrossRef]
114. Tabares-Soto, R.; Orozco-Arias, S.; Romero-Cano, V.; Bucheli, V.S.; Rodríguez-Sotelo, J.L.; Jiménez-Varón, C.F. A Comparative Study of Machine Learning and Deep Learning Algorithms to Classify Cancer Types Based on Microarray Gene Expression Data. *PeerJ Comput. Sci.* **2020**, *2020*, e270. [CrossRef] [PubMed]
115. Yuan, F.; Lu, L.; Zou, Q. Analysis of Gene Expression Profiles of Lung Cancer Subtypes with Machine Learning Algorithms. *Biochim. Biophys. Acta (BBA)-Mol. Basis Dis.* **2020**, *1866*, 165822. [CrossRef]
116. He, B.; Thomson, M.; Subramaniam, M.; Perez, R.; Ye, C.J.; Zou, J. CloudPred: Predicting Patient Phenotypes from Single-Cell RNA-Seq. *Pac. Symp. Biocomput.* **2022**, *2021*, 337–348. [CrossRef]
117. Ma, Y.; Chen, J.; Wang, T.; Zhang, L.; Xu, X.; Qiu, Y.; Xiang, A.P.; Huang, W. Accurate Machine Learning Model to Diagnose Chronic Autoimmune Diseases Utilizing Information From B Cells and Monocytes. *Front. Immunol.* **2022**, *13*, 870531. [CrossRef] [PubMed]
118. Galdos, F.X.; Xu, S.; Goodyer, W.R.; Duan, L.; Huang, Y.V.; Lee, S.; Zhu, H.; Lee, C.; Wei, N.; Lee, D.; et al. DevCellPy Is a Machine Learning-Enabled Pipeline for Automated Annotation of Complex Multilayered Single-Cell Transcriptomic Data. *Nat. Commun.* **2022**, *13*, 5271. [CrossRef]

119. Liu, J.; Fan, Z.; Zhao, W.; Zhou, X. Machine Intelligence in Single-Cell Data Analysis: Advances and New Challenges. *Front. Genet.* **2021**, *12*, 655536. [CrossRef]
120. Patil, A.R.; Schug, J.; Liu, C.; Lahori, D.; Descamps, H.C.; Naji, A.; Kaestner, K.H.; Faryabi, R.B.; Vahedi, G. Modeling Type 1 Diabetes Progression Using Machine Learning and Single-Cell Transcriptomic Measurements in Human Islets. *Cell Rep. Med.* **2024**, *5*, 101535. [CrossRef]
121. Hu, Y.; Hase, T.; Li, H.P.; Prabhakar, S.; Kitano, H.; Ng, S.K.; Ghosh, S.; Wee, L.J.K. A Machine Learning Approach for the Identification of Key Markers Involved in Brain Development from Single-Cell Transcriptomic Data. *BMC Genom.* **2016**, *17*, 1025. [CrossRef] [PubMed]
122. Vrahatis, A.G.; Tasoulis, S.K.; Maglogiannis, I.; Plagianakos, V.P. Recent Machine Learning Approaches for Single-Cell RNA-Seq Data Analysis. *Stud. Comput. Intell.* **2020**, *891*, 65–79. [CrossRef]
123. Barrett, T.; Wilhite, S.E.; Ledoux, P.; Evangelista, C.; Kim, I.F.; Tomashevsky, M.; Marshall, K.A.; Phillippy, K.H.; Sherman, P.M.; Holko, M.; et al. NCBI GEO: Archive for Functional Genomics Data Sets—Update. *Nucleic Acids Res.* **2013**, *41*, D991–D995. [CrossRef] [PubMed]
124. Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.M.; Ozenberger, B.A.; Ellrott, K.; Sander, C.; Stuart, J.M.; Chang, K.; Creighton, C.J.; et al. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* **2013**, *45*, 1113–1120. [CrossRef] [PubMed]
125. Johnson, N.T.; Dhroso, A.; Hughes, K.J.; Korkin, D. Biological Classification with RNA-Seq Data: Can Alternatively Spliced Transcript Expression Enhance Machine Learning Classifiers? *RNA* **2018**, *24*, 1119–1132. [CrossRef] [PubMed]
126. Kuo, W.P.; Kim, E.Y.; Trimarchi, J.; Jenssen, T.K.; Vinterbo, S.A.; Ohno-Machado, L. A Primer on Gene Expression and Microarrays for Machine Learning Researchers. *J. Biomed. Inform.* **2004**, *37*, 293–303. [CrossRef] [PubMed]
127. Kelchtermans, P.; Bittremieux, W.; De Grave, K.; Degroeve, S.; Ramon, J.; Laukens, K.; Valkenborg, D.; Barsnes, H.; Martens, L. Machine Learning Applications in Proteomics Research: How the Past Can Boost the Future. *Proteomics* **2014**, *14*, 353–366. [CrossRef] [PubMed]
128. Swan, A.L.; Mobasheri, A.; Allaway, D.; Liddell, S.; Bacardit, J. Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology. *OMICS* **2013**, *17*, 595–610. [CrossRef] [PubMed]
129. Barla, A.; Jurman, G.; Riccadonna, S.; Merler, S.; Chierici, M.; Furlanello, C. Machine Learning Methods for Predictive Proteomics. *Brief. Bioinform.* **2008**, *9*, 119–128. [CrossRef]
130. Neely, B.A.; Dorfer, V.; Martens, L.; Bludau, I.; Bouwmeester, R.; Degroeve, S.; Deutsch, E.W.; Gessulat, S.; Käll, L.; Palczynski, P.; et al. Toward an Integrated Machine Learning Model of a Proteomics Experiment. *J. Proteome Res.* **2023**, *22*, 681–696. [CrossRef]
131. Desaire, H.; Go, E.P.; Hua, D. Advances, Obstacles, and Opportunities for Machine Learning in Proteomics. *Cell Rep. Phys. Sci.* **2022**, *3*, 101069. [CrossRef] [PubMed]
132. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef] [PubMed]
133. Sonsare, P.M.; Gunavathi, C. Investigation of Machine Learning Techniques on Proteomics: A Comprehensive Survey. *Prog. Biophys. Mol. Biol.* **2019**, *149*, 54–69. [CrossRef] [PubMed]
134. Vishnoi, S.; Matre, H.; Garg, P.; Pandey, S.K. Artificial Intelligence and Machine Learning for Protein Toxicity Prediction Using Proteomics Data. *Chem. Biol. Drug Des.* **2020**, *96*, 902–920. [CrossRef]
135. Mann, M.; Kumar, C.; Zeng, W.F.; Strauss, M.T. Artificial Intelligence for Proteomics and Biomarker Discovery. *Cell Syst.* **2021**, *12*, 759–770. [CrossRef] [PubMed]
136. Li, J.; Zhou, K.; Mu, B. Machine Learning for Mass Spectrometry Data Analysis in Proteomics. *Curr. Proteom.* **2020**, *18*, 620–634. [CrossRef]
137. Wen, B.; Zeng, W.F.; Liao, Y.; Shi, Z.; Savage, S.R.; Jiang, W.; Zhang, B. Deep Learning in Proteomics. *Proteomics* **2020**, *20*, 1900335. [CrossRef] [PubMed]
138. Hindson, J. Proteomics and Machine-Learning Models for Alcohol-Related Liver Disease Biomarkers. *Nat. Rev. Gastroenterol. Hepatol.* **2022**, *19*, 488. [CrossRef] [PubMed]
139. Cox, J. Prediction of Peptide Mass Spectral Libraries with Machine Learning. *Nat. Biotechnol.* **2022**, *41*, 33–43. [CrossRef]
140. Sengupta, A.; Naresh, G.; Mishra, A.; Parashar, D.; Narad, P. Proteome Analysis Using Machine Learning Approaches and Its Applications to Diseases. *Adv. Protein Chem. Struct. Biol.* **2021**, *127*, 161–216. [CrossRef]
141. Tilocca, B.; Britti, D.; Urbani, A.; Roncada, P. Computational Immune Proteomics Approach to Target COVID-19. *J. Proteome Res.* **2020**, *19*, 4233–4241. [CrossRef]
142. Bernardes, J.S.; Pedreira, C.E. A Review of Protein Function Prediction Under Machine Learning Perspective. *Recent Pat. Biotechnol.* **2013**, *7*, 122–141. [CrossRef] [PubMed]
143. Worley, B.; Powers, R. Multivariate Analysis in Metabolomics. *Curr. Metabolomics* **2013**, *1*, 92. [CrossRef]
144. Sinha, N.; Viswan, A.; Singh, C.; Rai, R.K.; Azim, A.; Baronia, A.K. Metabolomics Based Predictive Biomarker Model of ARDS: A Systemic Measure of Clinical Hypoxemia. *PLoS ONE* **2017**, *12*, e0187545. [CrossRef]
145. Martín-Blázquez, A.; Díaz, C.; González-Flores, E.; Franco-Rivas, D.; Jiménez-Luna, C.; Melguizo, C.; Prados, J.; Genilloud, O.; Vicente, F.; Caba, O.; et al. Untargeted LC-HRMS-Based Metabolomics to Identify Novel Biomarkers of Metastatic Colorectal Cancer. *Sci. Rep.* **2019**, *9*, 20198. [CrossRef]

146. Goldberg, E.; Ievari-Shariati, S.; Kidane, B.; Kim, J.; Banerji, S.; Qing, G.; Srinathan, S.; Murphy, L.; Aliani, M. Comparative Metabolomics Studies of Blood Collected in Streck and Heparin Tubes from Lung Cancer Patients. *PLoS ONE* **2021**, *16*, e0249648. [CrossRef] [PubMed]
147. French, C.D.; Willoughby, R.E.; Pan, A.; Wong, S.J.; Foley, J.F.; Wheat, L.J.; Fernandez, J.; Encarnacion, R.; Ondrush, J.M.; Fatteh, N.; et al. NMR Metabolomics of Cerebrospinal Fluid Differentiates Inflammatory Diseases of the Central Nervous System. *PLoS Negl. Trop. Dis.* **2018**, *12*, e0007045. [CrossRef] [PubMed]
148. Collakova, E.; Aghamirzaie, D.; Fang, Y.; Klumas, C.; Tabataba, F.; Kakumanu, A.; Myers, E.; Heath, L.S.; Grene, R. Metabolic and Transcriptional Reprogramming in Developing Soybean (Glycine Max) Embryos. *Metabolites* **2013**, *3*, 347–372. [CrossRef]
149. You, J.; Zhang, Y.; Liu, A.; Li, D.; Wang, X.; Dossa, K.; Zhou, R.; Yu, J.; Zhang, Y.; Wang, L.; et al. Transcriptomic and Metabolomic Profiling of Drought-Tolerant and Susceptible Sesame Genotypes in Response to Drought Stress. *BMC Plant Biol.* **2019**, *19*, 267. [CrossRef]
150. Alreshidi, M.M. Selected Metabolites Profiling of Staphylococcus Aureus Following Exposure to Low Temperature and Elevated Sodium Chloride. *Front. Microbiol.* **2020**, *11*, 512774. [CrossRef]
151. Broughton-Neiswanger, L.E.; Rivera-Velez, S.M.; Suarez, M.A.; Slovak, J.E.; Piñeyro, P.E.; Hwang, J.K.; Villarino, N.F. Urinary Chemical Fingerprint Left behind by Repeated NSAID Administration: Discovery of Putative Biomarkers Using Artificial Intelligence. *PLoS ONE* **2020**, *15*, e0228989. [CrossRef] [PubMed]
152. Xu, J.; Pan, T.; Qi, X.; Tan, R.; Wang, X.; Liu, Z.; Tao, Z.; Qu, H.; Zhang, Y.; Chen, H.; et al. Increased Mortality of Acute Respiratory Distress Syndrome Was Associated with High Levels of Plasma Phenylalanine. *Respir. Res.* **2020**, *21*, 99. [CrossRef] [PubMed]
153. Monteleone, A.M.; Troisi, J.; Serena, G.; Fasano, A.; Grave, R.D.; Cascino, G.; Marciello, F.; Calugi, S.; Scala, G.; Corrivetti, G.; et al. The Gut Microbiome and Metabolomics Profiles of Restricting and Binge-Purging Type Anorexia Nervosa. *Nutrients* **2021**, *13*, 507. [CrossRef] [PubMed]
154. Caterino, M.; Costanzo, M.; Fedele, R.; Cevenini, A.; Gelzo, M.; Di Minno, A.; Andolfo, I.; Capasso, M.; Russo, R.; Annunziata, A.; et al. The Serum Metabolome of Moderate and Severe COVID-19 Patients Reflects Possible Liver Alterations Involving Carbon and Nitrogen Metabolism. *Int. J. Mol. Sci.* **2021**, *22*, 9548. [CrossRef] [PubMed]
155. Silva, A.A.R.; Cardoso, M.R.; Cardoso De Oliveira, D.; Godoy, P.; Cecília, M.; Talarico, R.; Marrero Gutiérrez, J.; Rodrigues Peres, R.M.; De Carvalho, L.M.; Angelo Da, N.; et al. Plasma Metabolome Signatures to Predict Responsiveness to Neoadjuvant Chemotherapy in Breast Cancer. *Cancers* **2024**, *16*, 2473. [CrossRef] [PubMed]
156. Galal, A.; Talal, M.; Moustafa, A. Applications of Machine Learning in Metabolomics: Disease Modeling and Classification. *Front. Genet.* **2022**, *13*, 1017340. [CrossRef] [PubMed]
157. Auslander, N.; Yizhak, K.; Weinstock, A.; Budhu, A.; Tang, W.; Wang, X.W.; Ambs, S.; Ruppin, E. A Joint Analysis of Transcriptomic and Metabolomic Data Uncovers Enhanced Enzyme-Metabolite Coupling in Breast Cancer. *Sci. Rep.* **2016**, *6*, 29662. [CrossRef]
158. Ghosh, T.; Zhang, W.; Ghosh, D.; Kechris, K. Predictive Modeling for Metabolomics Data. *Methods Mol. Biol.* **2020**, *2104*, 313. [CrossRef]

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

4. CAPÍTULO 2 – PUBLICADO (A1)

Exploratory machine learning and omics integration in the search for biomarkers of papillary thyroid cancer.

O Carcinoma Papilar de Tireoide (PTC) está entre as neoplasias malignas endócrinas mais comuns em todo o mundo e, embora geralmente associado a um prognóstico favorável, um subgrupo de pacientes desenvolve doença agressiva com maior risco de recorrência. Isso ressalta a necessidade de uma melhor caracterização molecular. Este estudo visa integrar dados multi-ômicos e aprendizado de máquina para identificar biomarcadores clinicamente relevantes no PTC. Esses modelos foram validados utilizando conjuntos de dados externos independentes. Além disso, genes diferencialmente expressos (DEGs) foram integrados a dados de metabolômica provenientes da literatura, permitindo a construção de uma rede de interação gene-metabólito. Essa nova abordagem, que combina integração de dados e aprendizado de máquina, fornece novos insights sobre a biologia do câncer de tireoide, apoiando avanços na medicina de precisão. Afim de tornar o estudo aberto para reprodução, os scripts estão publicamente disponíveis através do link: github.com/pedrohgodoys.

biology

MDPI

Article

Exploratory Machine Learning and Omics Integration in the Search for Biomarkers of Papillary Thyroid Cancer

Pedro Henrique Godoy Sanches, Nicolly Clemente de Melo, Danilo Cardoso de Oliveira and Lucas Miguel de Carvalho *

Health Sciences Postgraduate Program, São Francisco University, Bragança Paulista Campus, Av. São Francisco de Assis, 218, Bragança Paulista 12916-900, SP, Brazil
* Correspondence: lucas.miguel@usf.edu.br

Simple Summary

Papillary thyroid carcinoma (PTC) is among the most common endocrine malignancies worldwide, and although generally associated with a favorable prognosis, a subset of patients develops aggressive disease with higher recurrence risk. This highlights the need for improved molecular characterization. This study aims to integrate multi-omics data and machine learning (ML) to identify clinically relevant biomarkers in papillary thyroid carcinoma. These models were validated using independent external datasets. Additionally, differentially expressed genes (DEGs) were integrated with metabolomics data from the literature, enabling the construction of a metabolite–gene interaction network. This novel pipeline combining data integration and machine learning provides new insights into thyroid cancer biology, supporting advances in precision medicine.

Abstract

Papillary thyroid carcinoma (PTC) is among the most common endocrine malignancies worldwide, and although generally associated with a favorable prognosis, a subset of patients develops aggressive disease with higher recurrence risk. This highlights the need for improved molecular characterization. Data integration approaches combined with computational methods offer new opportunities to refine diagnosis and uncover disease mechanisms. This study aims to integrate omics data and apply machine learning (ML) to identify clinically relevant biomarkers in papillary thyroid carcinoma. We selected 11 genes from the differentially expressed genes (DEGs)–LASSO intersection approach. Genes were validated using an independent external dataset (AUC = 91%, Sens. = 92%, Spec. = 97%, and Acc. = 95%). DEGs were integrated with metabolomics data from the literature, enabling the construction of a metabolite–gene interaction network, highlighting norepinephrine, arachidonic acid, and glutamic acid as representative metabolites, while the main genes were *SLC6A14*, *ADK*, *ATIC*, *NT5E*, and *AR*. We identified potential drug–gene interactions and performed survival analysis to assess the relevance of the possible biomarkers. This novel pipeline combining integration and machine learning provides new insights into thyroid cancer biology and identifies promising diagnostic markers, supporting advances in precision medicine.

Keywords: papillary thyroid cancer; papillary thyroid carcinoma; bioinformatics; transcriptomics; metabolomics

Check for updates

Academic Editor: Sushil Shakyawar

Received: 11 May 2026
Revised: 19 June 2026
Accepted: 23 June 2026
Published: 25 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Thyroid carcinoma (TC) ranks as the seventh most common malignancy worldwide, exhibiting a notable female predominance (3:1) and a higher incidence in developed countries [1]. Despite its high incidence, TC is generally associated with a favorable prognosis; however, a subset of patients presents with aggressive disease and increased risk of recurrence, highlighting the need for improved molecular characterization [2,3]. The histological landscape of TC is diverse: papillary thyroid carcinoma (PTC) accounts for approximately 80% of cases, followed by the follicular type (FTC) at 15%, while medullary (MTC) and anaplastic (ATC) carcinomas comprise 4% and 1% of occurrences, respectively [4].

Several risk factors have been proposed, including genetic, environmental, and hormonal influences, such as iodine intake, TSH level, autoimmune thyroid disease, gender, estrogen, obesity, lifestyle changes, and environmental pollutants, but up to now, only childhood exposure to ionizing radiation has been fully recognized as a risk factor [5,6]. Clinically, PTC is often characterized by slow tumor progression and nonspecific symptoms [3]. The main issue caused by PTC is excessive growth in size, compressing the surrounding regions, and causing mild symptoms. Thus, neck pain is the most reported symptom, and nerve compression can be an issue; respiratory difficulties are rarer [4].

Therapeutic strategies are currently guided by the American Thyroid Association's (ATA) risk stratification, which classifies recurrence probability from low to high. Management typically involves surgical intervention via thyroidectomy, often complemented by adjuvant radioactive iodine therapy [7]. However, these approaches may lead to overtreatment in low-risk patients and remain insufficient for accurately predicting disease progression in more aggressive cases. In this context, recent advances in clinical management, including high-resolution imaging and molecular profiling, are reshaping the standard of care [8].

Omics sciences have emerged as a cornerstone of modern oncology, enabling the comprehensive characterization of molecular alterations at multiple biological levels [9], including transcriptomics [10], proteomics [11], and metabolomics [12]. These approaches have facilitated the identification of novel biomarkers and therapeutic targets across different cancer types [13]. In thyroid cancer, bioinformatics-driven studies have increasingly applied high-throughput data analysis to uncover gene signatures, regulatory networks, and diagnostic markers associated with PTC progression and prognosis [14–16]. Furthermore, the integration of omics layers with advanced computational approaches and artificial intelligence have shown great potential to improve tumor classification, risk prediction, and personalized treatment strategies [17].

Although several studies have explored transcriptomic alterations in PTC, most rely on single-layer analyses, which may overlook complex biological interactions underlying tumor development. Here, we advance beyond traditional approaches by integrating gene expression profiles with literature metabolomics data, enabling a systems-level understanding of the disease. In addition, the application of machine learning models allows for robust feature selection in high-dimensional datasets, reducing noise while preserving biologically relevant signals. The novelty of this strategy enhances biomarker discovery and improves model generalizability across independent cohorts, addressing a critical gap in translational oncology.

Additionally, as shown in a recent review [18], little has been done to provide clear biomarking features for PTC. While some decision tree models have been used to find patterns in lncRNAs and miRNAs, not much has been done to oversee the basic diagnostic criteria. In addition, no attempt has been made to integrate different omics layers. Our approach aims to identify clinically relevant biomarkers with medical value while also

uncovering potential molecular mechanisms that could guide precision medicine strategies in papillary thyroid carcinoma.

2. Materials and Methods

We implemented an integrative multistep analytical framework combining transcriptomic profiling, machine learning, and clinical outcome assessment. Initially, we searched for repositories with publicly available experimental data on PTC for both metabolomics and transcriptomics. Transcriptomics studies were found in Gene Expression Omnibus (GEO) under the ID GSE29265 for the training set and GSE33630 for the validation set. The search for metabolomics data was conducted in Metabolomics Workbench [19] and Metabolights [20]. Unfortunately, there were no experimental data from metabolomics studies; thus, a manual curation was considered. To obtain publicly available data, a manual search in the PUBMED database was performed—the terms used were "Metabolomics" and "Papillary Thyroid Carcinoma".

Firstly, the main inclusion criteria for a metabolite were the presence of an official name together with its fold-change. Secondly, the developed algorithm for transcriptomics data was applied. It consisted of: (i) data handling—log2 normalization, remove near zero variance; (ii) differential expression analysis—(using limma package) fit linear model, Bayes empirical moderation, Wilcoxon analysis (FDR < 0.05) and fold-change ($|\log2FC| > 0.58$); (iii) Least Absolute Shrinkage and Selection Operator (*LASSO*) model—cross-validation (best λ), fit LASSO model, ROC curve to evaluate; and (iv) intersection—intersects limma's DEGs with LASSO-selected features to find a final list of genes ($n = 11$).

Finally, a network of interactions among genes and literature-gathered metabolites was built using MetaboAnalyst v6.0 [21] and exported to CytoScape v3.10.3 [22] where parameters of quality were calculated. Additionally, overall survival analysis and disease-free survival analysis were performed using the selected genes ($n = 11$). Afterward, in silico matching of the intersected genes with FDA-approved drugs was conducted with additional enrichment analysis. For both metabolomics and transcriptomics, enrichment analysis was also explored (Figure 1).

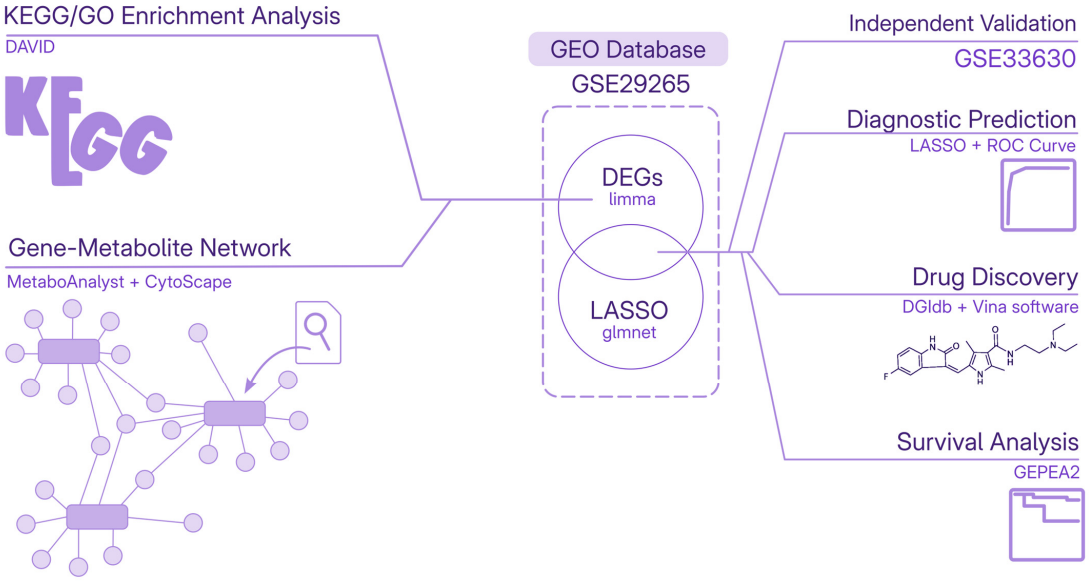


Figure 1. The whole pipeline developed to associate transcriptomics data with metabolomics data. Arrows indicate the direction of the analytical workflow, and boxes represent the main processing and analysis steps, including metabolite selection, gene identification, network construction, survival analysis, drug discovery, and enrichment analyses.

All analyses were conducted using fully reproducible results, which are provided in the Supplementary Material.

2.1. Datasets and Preprocessing

Gene expression datasets were obtained from the Gene Expression Omnibus (GEO) database. The dataset GSE29265 was used as the training cohort, comprising 20 papillary thyroid cancer (PTC) samples and 20 normal thyroid samples. The dataset GSE33630, containing 49 PTC and 45 normal samples, was used as an independent validation cohort. Expression data were assessed for scale, and log2 transformation was applied when gene expression value exceeded 50. Genes with near-zero variance ($<1 \times 10^{-6}$) in the training dataset were excluded. Principal component analysis (PCA) was performed to evaluate sample distribution and data quality.

2.2. Differential Expression Analysis

Differential expression analysis between PTC and normal samples in the training dataset (GSE29265) was performed using the limma package (v3.66.0) [23]. A linear model was fitted to the data, followed by empirical Bayes moderation. Genes were considered differentially expressed based on the following criteria: adjusted *p*-value (false discovery rate, FDR) < 0.05 and absolute log2 fold change ($|\log2FC|$) > 0.58.

2.3. LASSO Feature Selection

To identify predictive features, a LASSO logistic regression model was implemented using the glmnet package (v5.0) [24]. To reduce dimensionality and computational burden, the top 5000 most variable genes were selected. A binomial model was applied, and 5-fold cross-validation was used to determine the optimal regularization parameter (λ). Genes with non-zero coefficients in the final LASSO model were selected as relevant features.

2.4. Intersection Strategy and External Validation

To increase robustness, an intersection strategy was applied by overlapping the list of differentially expressed genes (DEGs) with the genes selected from LASSO. The resulting intersecting genes were considered candidate biomarkers. These genes were used to build a predictive model, which was evaluated in the independent validation dataset (GSE33630). Only genes present in both datasets were retained.

Prediction probabilities were calculated using the LASSO-derived coefficients. Model performance was assessed by a receiver operating characteristic (ROC) curve analysis, and the area under the curve (AUC), sensitivity, specificity, and accuracy were calculated. The optimal classification threshold was determined based on ROC curve coordinates. Additionally, statistical differences in gene expression between PTC and normal samples in the validation dataset were assessed using the Wilcoxon rank-sum test (*p*-value < 0.05).

2.5. Survival and Progression-Free Interval Analysis

Overall survival and progression-free interval analysis were performed using the GEPIA3 platform [25], based on The Cancer Genome Atlas (TCGA) thyroid carcinoma dataset. Genes identified as statistically significant were evaluated for their association with patient outcomes.

Overall survival and disease-free survival analyses were conducted by stratifying patients into high- and low-expression groups according to gene expression levels. Kaplan–Meier survival curves were generated, and statistical significance was assessed using the *p*-value ($p < 0.05$).

2.6. Drug–Gene Interaction Analysis

In order to generate hypotheses and find possible links between genes and drugs, we here use the DGIdb [26] as a publicly available drug–gene interaction platform, where the 11 genes found in the Venn diagram were imported into the platform. We only selected drugs that were approved by the FDA.

2.7. Functional Enrichment Analysis

To investigate the biological functions and pathways associated with the initial 1428 differentially expressed genes (DEGs), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and biological process in Gene Ontology (GO) enrichment analyses were performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) platform [27], considering p-adj < 0.05 as a significance cutoff.

2.8. Metabolites Associated with PTC

To investigate metabolite associations with PTC, datasets from published metabolomics studies focusing on papillary thyroid cancer were collected from the PUBMED database using the terms "Papillary Thyroid Carcinoma" AND "Metabolomics". At the time of the systematic search, 58 articles were found and filtered by title (n = 33), considering the correct study design (in this case, PTC vs. nontumoral). Additionally, filtering by the abstract and text (n = 28) was performed, and finally, by data availability (n = 7).

Metabolomics is a younger field compared to transcriptomics; thus, it is expected from the articles not to have much publicly available data. In this case, a manual curation was necessary, searching for the presence of a link, corresponding author email, public repository upload, or other ways of data availability. In the case of manual collection of data, the presence of a reported p-value and fold-change were mandatory (n = 3).

2.9. Integration Strategy

An integrative approach was applied to combine transcriptomic and metabolomic data [28]. Differentially expressed genes identified in the training dataset were integrated with PTC-associated metabolites. Gene–metabolite interaction networks were constructed using the MetaboAnalyst platform [21].

Network topology analysis was performed to identify key nodes using CytoHubba in Cytoscape software [22], and the top three genes and top three metabolites with the highest degree of centrality were selected as hub components. This integrative strategy aimed to uncover potential molecular mechanisms linking gene expression and metabolic alterations in papillary thyroid cancer.

2.10. Molecular Docking

The molecular docking approach was applied to analyze the receptor–ligand relationship. The genes used for docking were selected from the integration between DEGs and LASSO. Given the challenges and costs of antibody production, only small molecules were selected to evaluate drug repurposing in this study, due to their potential to complement and synergize with therapeutics compared with antibody-based immunotherapy [29].

The Protein Data Bank (PDB) [30] was used to curate experimentally resolved 3D protein structures based on the presence of small-molecule ligand information. Homo sapiens were the source organism, refinement resolution between 2 and 3 Å, and 100% amino acid sequence identity. The amino acid sequence was retrieved from KEGG by searching *Homo sapiens* species in the "Gene" section. Protein and ligand preparation followed these steps: AutoDock Vina v1.2.x [31] was used to perform molecular docking, and the AutoDockTools (ADT) v1.5.7 [32] graphical interface to prepare the protein at a pH

of 7.4 with Kollman charges. In order to prepare the ligands, we combined RDKit [33] and OpenBabel [34] to convert molecules to "pdbqt" files and correct hydrogen states.

In cases where the gene did not have an inhibitor in the PDB file, a blind docking approach was performed. The predicted structure was retrieved from UniProt [35] by AlphaFold [36]. The KVFinder-web server [37] was used to predict cavities and potential active site regions. Discovery Studio (DS) v24.1.0.23298 [38] was utilized to identify the coordinates of active sites of the proteins in docking and blind docking.

The generated files were visualized using UCSF ChimeraX v1.8 [39], and amino acid interactions were analyzed in DS. Uniprot [40] was used to confirm amino acid positions in the predicted binding site using gene symbols for searching. The results with the best binding energy rankings were qualified, with more negative values indicating higher affinity between the ligand and the protein, representing a more favorable and stable binding to the target.

3. Results

3.1. Combined Differential Expression and Feature Selection Pipeline

After preprocessing of the GSE29265 dataset, a total of 23,520 genes were retained for downstream analysis following probe collapsing and variance filtering. Principal component analysis (PCA) revealed a clear separation between PTC and normal samples, indicating strong transcriptomic differences between the groups (Figure 2A). Differential expression analysis identified 1428 differentially expressed genes (DEGs) (adjusted p-value < 0.05 and $|\log2FC| > 0.58$), of which 679 were upregulated and 749 were downregulated in PTC samples (Figure 2B and Supplementary Table S1).

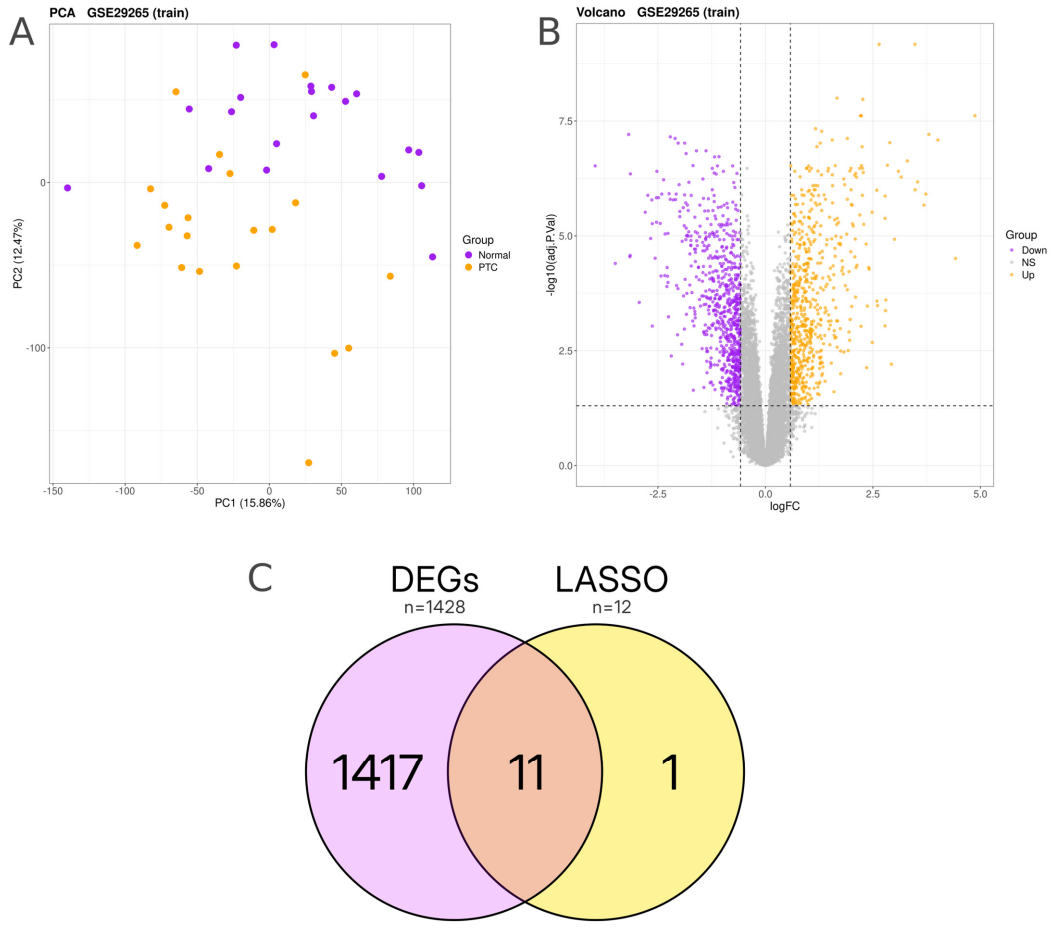


Figure 2. In (**A**), principal component analysis displays nontumoral tissue in purple, while papillary thyroid carcinoma is in orange. Likewise, (**B**) shows a volcano plot, where the purple dots represent

the downregulated variables while orange the upregulated ones; the nonsignificant (NS) features are displayed in gray; dashed line represent the significance limits. The Venn diagram in (**C**) shows the left circle, the differentially expressed genes (DEGs), intersecting with the LASSO model (Least Absolute Shrinkage and Selection Operator).

The LASSO model, applied to the training dataset, selected 12 genes with non-zero coefficients under the optimal $\lambda = 0.098$ determined by cross-validation (Supplementary Table S2). These genes represent features with the highest predictive power for distinguishing PTC from normal samples.

The intersection between DEGs and LASSO-selected genes resulted in 11 genes (Figure 2C) which were considered robust candidate biomarkers, namely *DPP4*, *GLT1D1*, *TENM1*, *TIMP4*, *PDZK1IP1*, *ABHD17C*, *IGSF10*, *RHOU*, *PDGFC*, *ABCA6*, and *RPS16P5*.

3.2. Validation of the Selected Biomarkers

The predictive performance of the 11 intersecting genes was evaluated in the independent GSE33630 dataset. The model achieved an area under the curve (AUC) of 91%, indicating strong discrimination between PTC and normal samples. Using the optimal threshold derived from ROC analysis, the model achieved an area under the curve (AUC) of 91%, sensitivity of 92%, specificity of 97%, and overall accuracy of 95%.

Statistical analysis demonstrated that the selected genes maintained consistent expression patterns in the validation dataset, with 11 out of the 11 genes meeting the Wilcoxon significance threshold ($p < 0.05$) and showing significant differences between PTC and normal samples (Figure 3). These findings support the reproducibility and biological relevance of the identified biomarkers.

There were four downregulated genes (*ABCA6*, *IGSF10*, *RPS16P5*, and *TIMP4*) and seven upregulated genes (*ABHD17C*, *DPP4*, *GLT1D1*, *PDGFC*, *PDZK1IP1*, *RHOU*, and *TENM1*). There was no strong evidence in the literature on the function of *GLT1D1*, *IGSF10*, and *RPS16P5*. However, *DPP4*, *TENM1*, *TIMP4*, *PDZK1IP1*, *RHOU*, *PDGFC*, and *ABCA6* return some relationships with PTC, organized individually in Supplementary Table S3. The individual statistical and prediction values are displayed in Table 1.

Table 1. Relevant statistical significance and predictive values of the eleven genes selected by this approach.

Genes	Average Expression	log2FC [1]	Regulation	*p*-Value		Effect Size *	AUC [2]	Sens.	Spec.	CI [3]	
				Raw	Adjusted					Lower	Upper
ABCA6	4.16	−1.15	Down	3.15×10^{-8}	4.28×10^{-6}	0.6444	0.95	0.92	0.98	0.90	1.00
ABHD17C	5.92	1.16	Up	3.47×10^{-9}	9.93×10^{-7}	0.6772	0.96	0.94	0.98	0.91	1.00
DPP4	6.51	3.47	Up	4.32×10^{-14}	6.79×10^{-10}	0.7779	0.96	0.94	0.98	0.90	1.00
GLT1D1	5.38	1.67	Up	1.28×10^{-12}	1.01×10^{-8}	0.7850	0.86	0.754	0.87	0.79	0.94
IGSF10	4.52	−1.2	Down	3.50×10^{-9}	9.93×10^{-7}	0.7818	0.95	0.92	0.96	0.9	1.00
PDGFC	7.13	0.90	Up	2.76×10^{-8}	3.92×10^{-6}	0.3587	0.89	0.85	0.91	0.8	0.96
PDZK1IP1	4.98	2.27	Up	9.95×10^{-10}	4.18×10^{-7}	0.7810	0.95	0.96	0.91	0.91	1.00
RHOU	8.08	1.06	Up	8.30×10^{-9}	1.65×10^{-6}	0.6741	0.90	0.85	0.90	0.82	0.96
RPS16P5	6.50	−1.73	Down	1.52×10^{-7}	1.26×10^{-5}	0.3493	0.71	0.65	0.8	0.60	0.81
TENM1	4.81	2.23	Up	8.27×10^{-12}	2.43×10^{-8}	0.7857	0.87	0.71	0.98	0.79	0.95
TIMP4	3.93	−1.23	Down	6.52×10^{-10}	3.39×10^{-7}	0.6202	0.70	0.67	0.70	0.60	0.81

[1]: Log_2 fold-change; [2]: Area under the curve; Sens.: Sensitivity; Spec.: Specificity. [3]: Confidence intervals. *: Small effect sizes were considered when the effect was <0.3, moderate effects were between 0.3 and <0.5, and finally, a large effect size was defined as values > 0.5.



Figure 3. All the boxplots for the selected genes. Wilcoxon's adjusted *p*-value is shown above the boxes. The *y*-axis expresses the log2 (Expression), while the *x*-axis displays the compared groups. Purple boxes show the nontumoral ("normal") tissues and the orange ones the papillary thyroid carcinoma. At the top of the plot's panel is displayed the Wilcoxon's significance value corrected by the false discovery rate (*ABCA6* ($p = 4.3 \times 10^{-10}$), *ABHD17C* ($p = 5.3 \times 10^{-11}$), *DPP4* ($p = 4.8 \times 10^{-14}$), *GLT1D1* ($p = 2.8 \times 10^{-14}$), *IGSF10* ($p = 3.6 \times 10^{-14}$), *PDGFC* ($p = 5.1 \times 10^{-4}$), *PDZ1IP1* ($p = 3.8 \times 10^{-14}$), *RHOU* ($p = 6.5 \times 10^{-11}$), *RPS16P5* ($p = 7.2 \times 10^{-4}$), *TENM1* ($p = 2.7 \times 10^{-14}$), and *TIMP4* ($p = 1.9 \times 10^{-9}$)).

3.3. Overall Survival and Progression-Free Interval Analyses

Overall survival analysis using TCGA data revealed that four of the selected genes were significantly associated with patient outcomes (*GLTD1*, *PDZK1IP1*, *TENM1*, and *TIMP4*). Kaplan–Meier curves demonstrated that altered expression levels of these genes were linked to differences in both overall survival (OS) (Figure 4A–D; Supplementary Table S4) and progression-free intervals (Figure 4E; Supplementary Table S5). Specifically, patients stratified into high- and low-expression groups showed significantly distinct

survival probabilities ($p < 0.05$), indicating that these genes may have prognostic value in PTC.

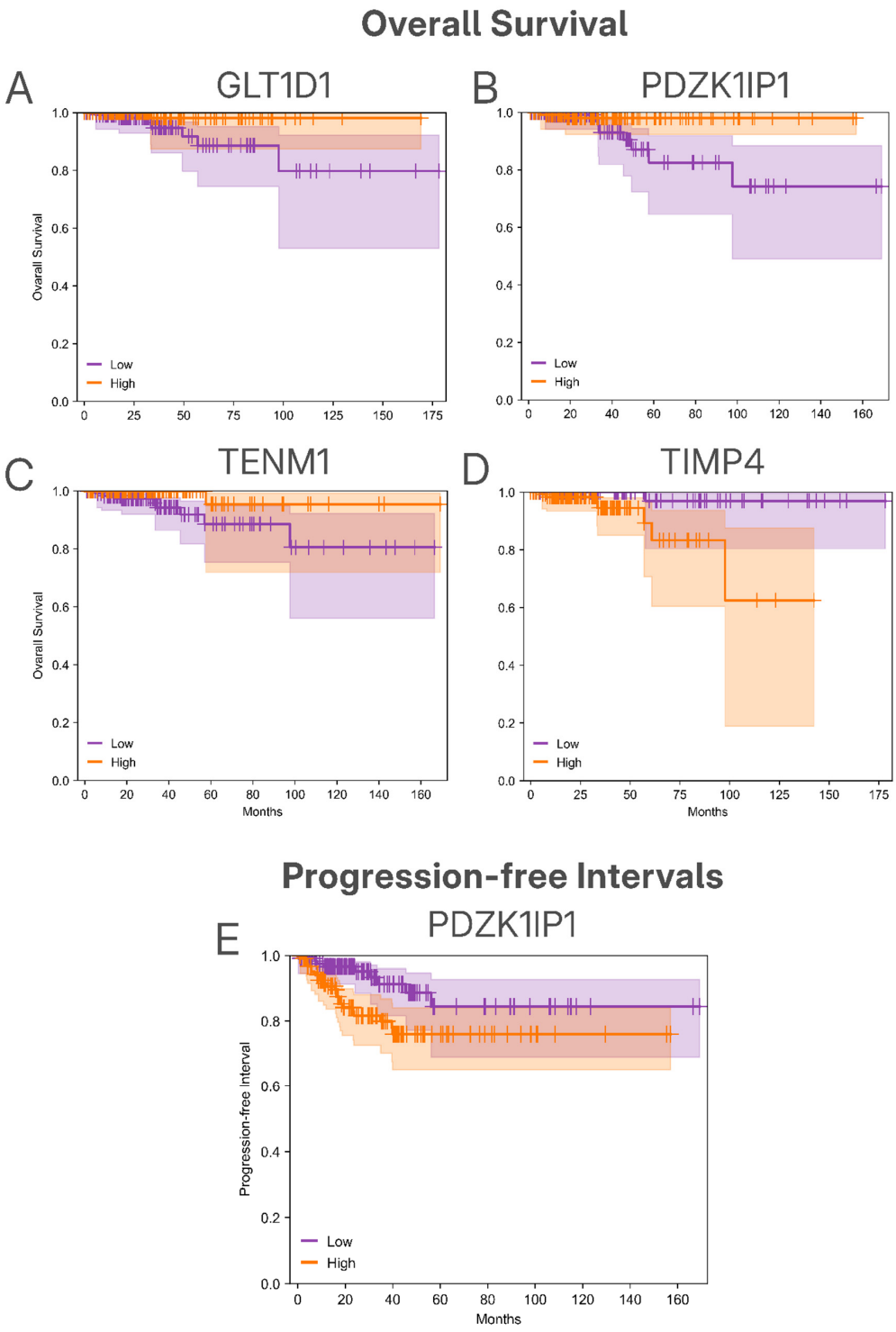


Figure 4. Overall survival curves in (**A**) are displayed for the *GLT1D1* gene ($p = 0.0494$), in (**B**), the *PDZK1IP1* gene ($p = 0.0398$), in (**C**), the *TENM1* gene ($p = 0.036$), and (**D**), the *TIMP4* gene ($p = 0.00507$). The progression-free intervals in (**E**) show the gene *PDZK1IP1* with $p = 0.0145$. The orange lines represent the high group and the purple one the low group. Confidence intervals are shown around the line as a painted box.

3.4. Enrichment Analysis of DEGs

Functional enrichment analysis of DEGs revealed significant involvement of these genes in PTC (Supplementary Table S6). Several pathways were returned, including cytoskeleton in muscle cells (FDR = 0.0000617), cell adhesion molecules (FDR = 0.00525), cytokine–cytokine receptor interaction (FDR = 0.00382), but also transcriptional misregulation in cancer (FDR = 0.00292), the p53 signaling pathway (FDR = 0.00292), and extracellular matrix–receptor interaction (FDR = 0.00292). Table 2 compiles the GO biological process enrichment analysis.

Table 2. Biological process enrichment analysis.

Name	*p*-Value	False Discovery Rate
Cell adhesion	4.54×10^{-21}	2.09×10^{-17}
Extracellular matrix organization	1.40×10^{-6}	3.22×10^{-3}
Cell–cell adhesion	4.87×10^{-7}	7.48×10^{-3}
Homophilic cell–cell adhesion	9.61×10^{-7}	1.11×10^{-2}
Inflammatory response	1.54×10^{-5}	1.42×10^{-2}
Positive regulation of apoptotic process	3.30×10^{-5}	2.14×10^{-2}
Receptor-mediated endocytosis	3.66×10^{-5}	2.14×10^{-2}
Nervous system development	3.70×10^{-5}	2.14×10^{-2}
Ossification	1.02×10^{-4}	5.20×10^{-2}
Positive regulation of phosphatidylinositol 3-kinase/protein kinase B signal transduction	1.39×10^{-4}	6.39×10^{-2}

3.5. Molecular Docking of the Drug Interactions

The gene–drug interaction analysis performed using DGIdb identified two genes with FDA-approved associated drugs: *PDGFC* and *DPP4* (Supplementary Table S7). In total, 22 drugs were retrieved for these genes. However, when restricting the analysis to compounds classified as antineoplastic agents, only a limited subset of interactions remained. After these selections, we performed molecular docking to confirm this promising interaction only for small molecules (see Section 2.10).

For PDGFC, a single approved antineoplastic agent was identified, Sunitinib, with an interaction score of 1.0876. Cavities were predicted using KVFinder that revealed two pockets named KAK and KAD due to its homodimeric structure. After that, Discovery Studio was used to determine the docking-grid coordinates (KAD: X = −8.597717, Y = 16.791843, Z = 3.858157; KAK: X = 3.858157, Y = −12.895208, Z = −14.675940).

Blind docking analysis revealed that Sunitinib binds favorably to both predicted cavities (KAK and KAD) of PDGFC, with similar binding affinities (−7.212 and −7.205 kcal/mol, respectively), indicating strong affinity across the identified binding sites (Figure 5).

For *DPP4*, several approved antineoplastic agents were identified. These include Nintedanib esylate (interaction score: 0.0409), Masitinib (0.0602), Quizartinib (0.0539), and Olaratumab (0.5118). Among these, Olaratumab showed the highest interaction score for *DPP4* but was not selected for docking tests, as it is an antibody, and thus was not included in the analysis.

The binding site was set at X = −1.775483, Y = −5.364172, Z = 72.346828 after identifying the native ligand coordinates in the PDB file. Molecular docking analysis demonstrated that Masitinib, Quizartinib, and Nintedanib esylate exhibit strong binding affinities to DPP4 (−9.342, −9.627, and −9.688 kcal/mol, respectively), comparable to the reference ligand PZF (−9.89 kcal/mol), with interactions predominantly mediated by Van der Waals forces and conventional hydrogen bonds within the active site. The active site of DPP4 can be defined by key amino acid residues at positions 630, 670, and 740, of which at least one is also involved in the interactions observed for the analyzed candidates (Figure 6).

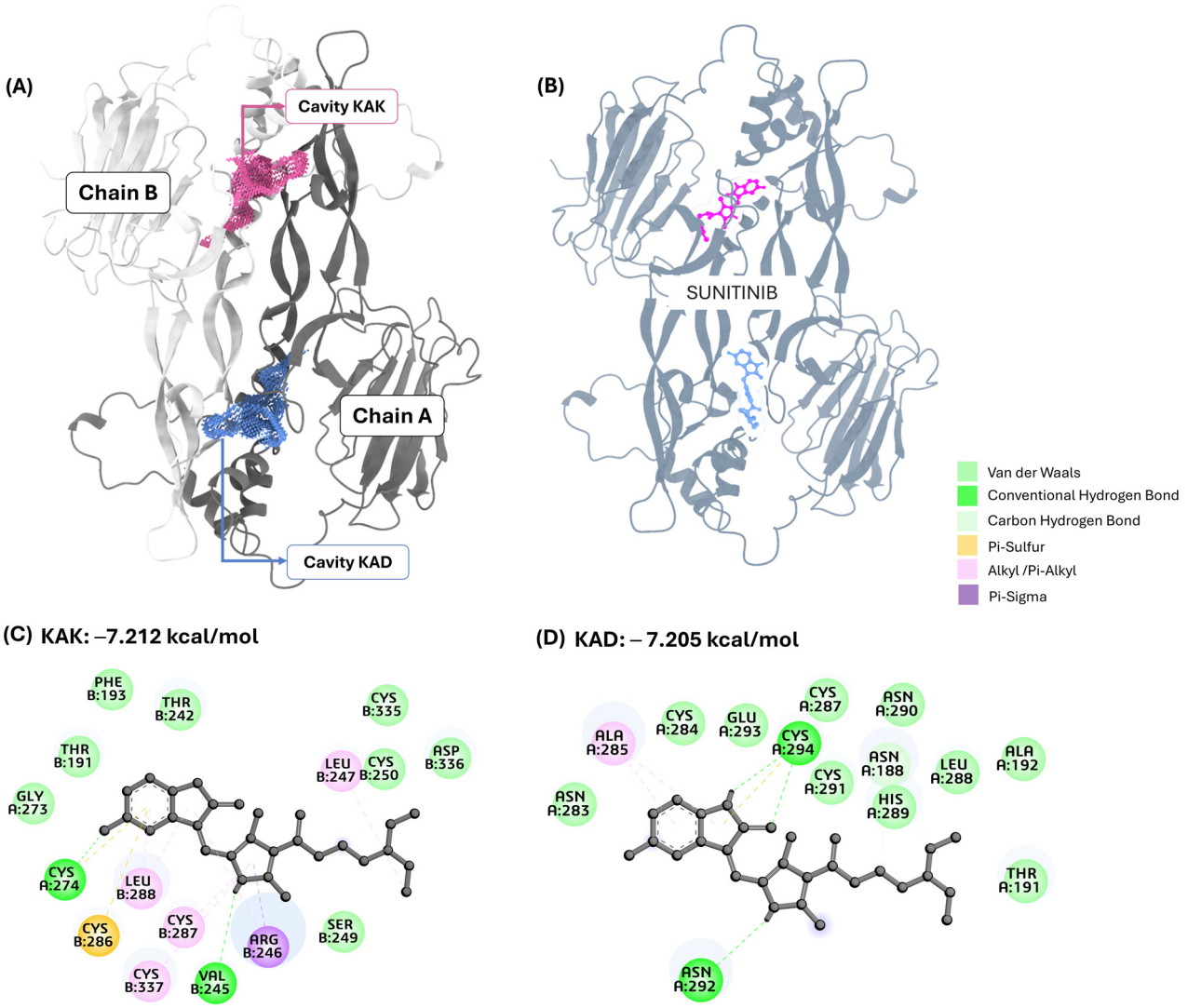


Figure 5. Blind docking of Sunitinib with PDFCG after cavity prediction. (**A**) 3D mapping of KAK (pink) and KAD (blue) cavities predicted by KVFinder-web in PDFCG predicted by AlphaFold constructed in two chains. (**B**) Predicted binding poses of Sunitinib in both protein pockets. (**C**) 2D interaction map for the KAK site (−7.212 kcal/mol); (**D**) 2D interaction map for the KAD site (−7.205 kcal/mol). Each interaction type color is represented in the figure.

3.6. Data Integration and Network Analysis of PTC

A total of 58 metabolomics studies focusing on PTC were curated, from which 101 metabolites reported as significantly altered were compiled (Supplementary Table S8). These metabolites were integrated with the DEGs identified in the transcriptomic analysis.

A gene–metabolite interaction network was constructed, revealing complex relationships between deregulated genes and altered metabolites (Supplementary Figure S1). Network topology analysis identified key hub components, with the top three genes (*SLC6A14*, 8; *ADK*, 8; *ATIC*, 5) and top three metabolites (norepinephrine, 24; arachidonic acid, 23; glutamic acid, 21) showing the highest degree of centrality.

For a more profound description of the interactions, Table 3 summarizes the connections made from the genes to the metabolites.

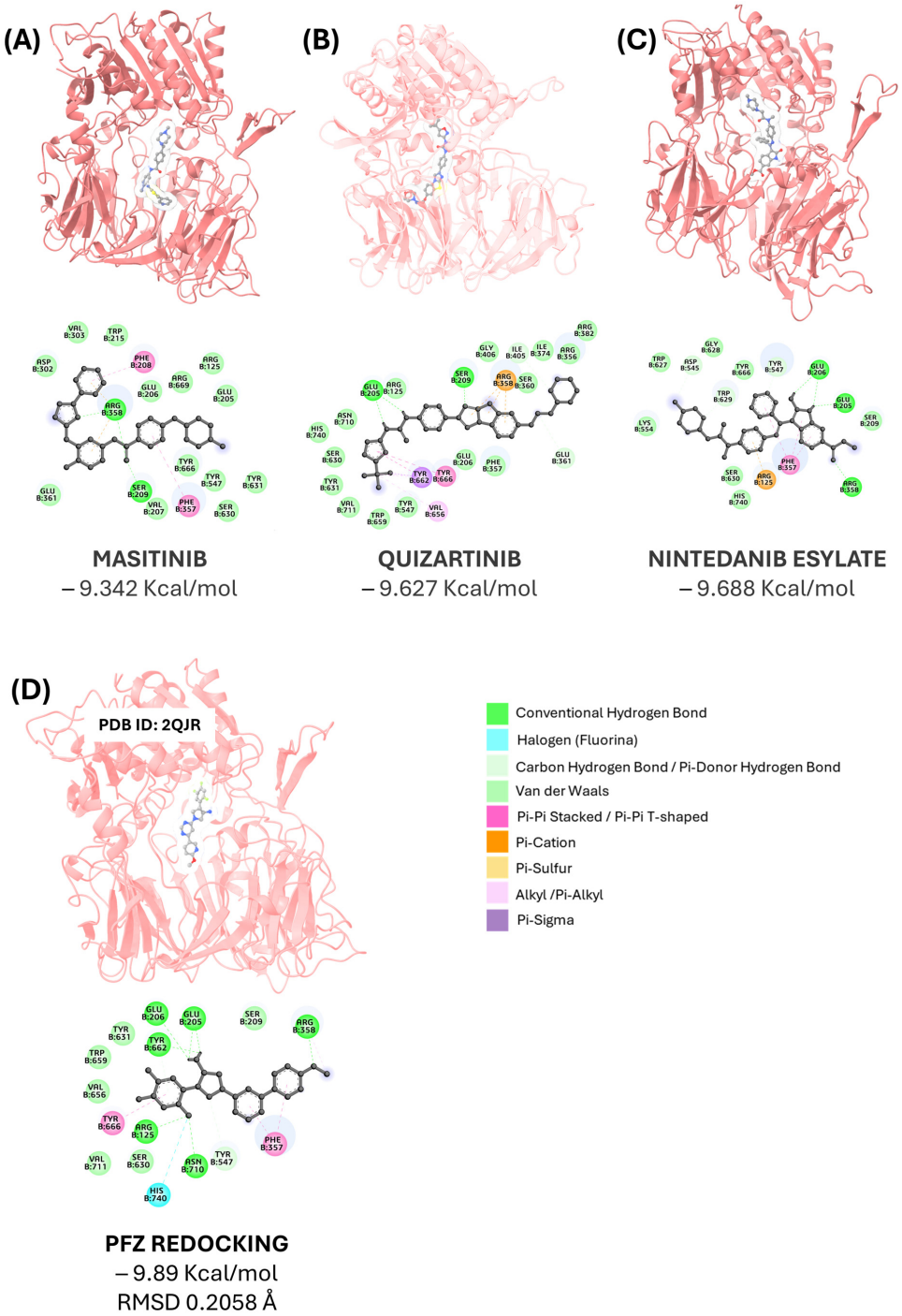


Figure 6. Molecular docking results between selected drugs and the DPP4 protein. (**A**) Docking result of Masitinib with a binding energy of −9.342 kcal/mol; the 3D structure is positioned within the active site, featuring a prevalence of Van der Waals and conventional hydrogen bond interactions. (**B**) Quizartinib presented a binding energy of −9.627 kcal/mol; transparency was applied to demonstrate a minor deviation of the molecule from the active site, with interactions primarily consisting of Van der Waals and conventional hydrogen bonds. (**C**) Docking result for Nintedanib esylate with −9.688 kcal/mol of energy; the molecule is in the active site, showing a predominance of Van der Waals and conventional hydrogen bond types. (**D**) Redocking results for (3R,4S)-1-[6-(6-methoxypyridin-3-yl)pyrimidin-4-yl]-4-(2,4,5-trifluorophenyl)pyrrolidin-3-amine (PZF) showed a binding affinity of −9.89 kcal/mol; interactions included Van der Waals, conventional hydrogen bonds, and exclusive halogen bond types. Each interaction type color is represented in the figure.

Table 3. Connections of the top three genes based on the degree.

Gene	Connects to
SLC6A14	Glycine, cysteine, alanine, lysine, methionine, isoleucine, beta-alanine, and leucine
ADK	Aspartic acid, adenine, guanosine, hypoxanthine, aminolevulinic acid, inosine, creatine, and guanidoacetic acid
ATIC	Aspartic acid, adenine, hypoxanthine, citric acid, and inosine

Additionally, the top three metabolites' networks, shown in Figure 7, display several DEGs connected to the main metabolites, but some highlight the interaction of the nodes: *SYT1* (FDR = 0.0046) connects norepinephrine to glutamic acid, while *CYSLTR2* (FDR = 3.9195×10^{-6}), *PPARG* (FDR = 0.0007), and *SLC4A4* (FDR = 3.0020×10^{-5}) link norepinephrine to the arachidonic acid. Finally, *CYCS* (FDR = 0.0031) and *JUN* (FDR = 0.0104) connect glutamic acid to arachidonic acid, rendering those genes as potential regulators of the interaction among these metabolites.

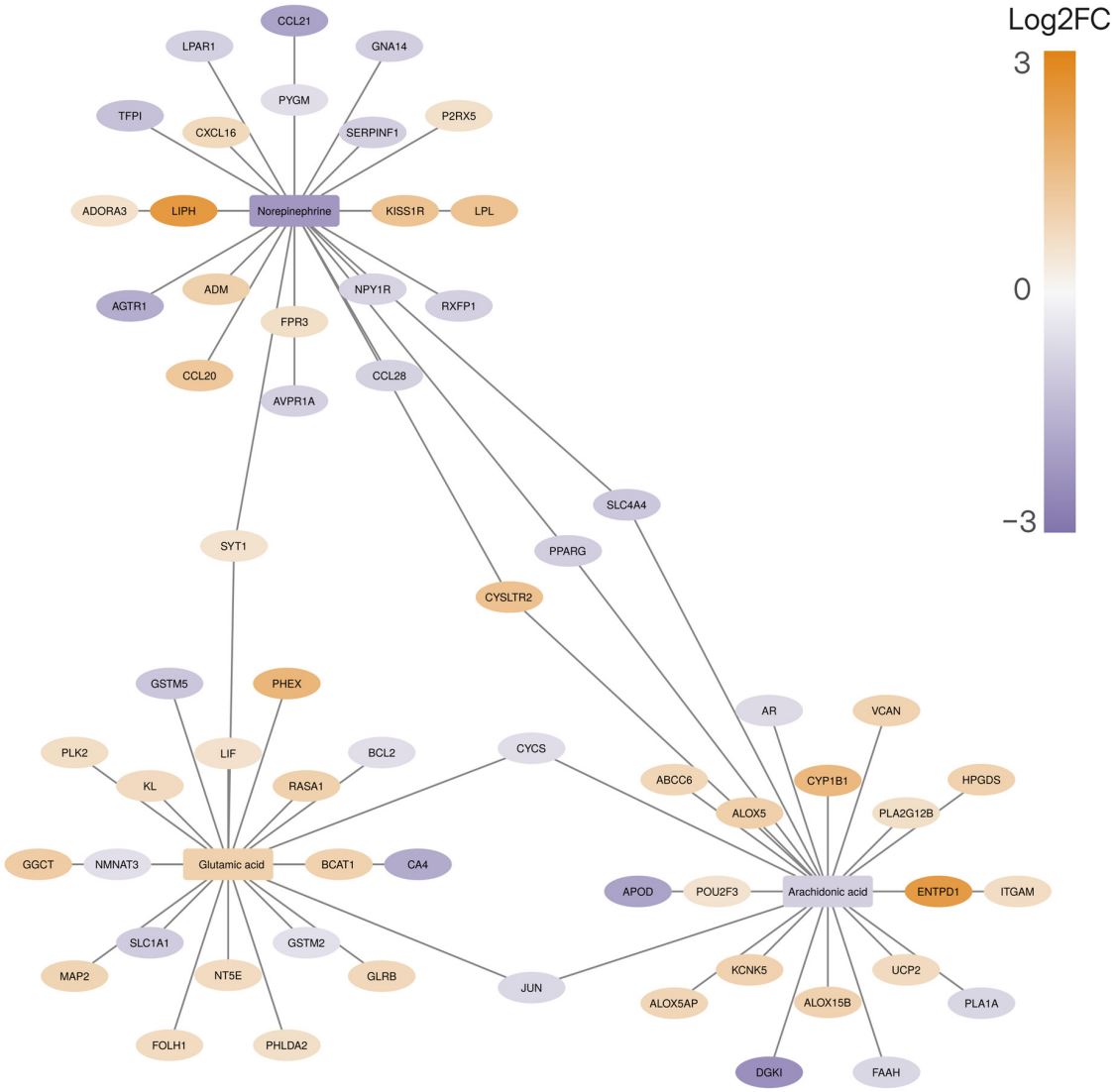


Figure 7. Gene–metabolite network analysis, showing the top three metabolites. The squared shapes represent metabolites, while the circles represent genes. The scale at the right shows the log2FC scale, where each component is then colorized accordingly.

4. Discussion

4.1. Biological Significance of the Selected Biomarkers

It was possible to select variables beyond DEGs, intersecting them with advanced machine learning models (namely, LASSO). Some genes (*DPP4*, *GLT1D1*, *PDZK1IP1*, and *TENM1*) displayed a distinct separate boxplot (Figure 3), indicating more considerable differences between PTC and normal tissue, thus indicating that a hyperexpression status was active and could contribute to the disease's carcinogenesis. Particularly, *PDZK1IP1* showed a considerable separation in the boxplot and was shown as significant in the survival analysis, indicating that it could represent a prognostic feature, discussed further in more detail.

Other work [41] has found ABCA6 as a hub gene in their analysis, reaching AUC = 89.5%. They used several datasets, including the one used here as an independent data validation (*GSE33630*), thus explaining why it appeared. Here, we selected *ABCA6* by the intersection approach; nonetheless, the gene did not show up as a hub gene. Their technique was network-based, and as both works used different approaches, the *ABCA6* gene could actually have an important role in PTC by being selected.

DPP4 was identified among the genes selected by the LASSO model and showed increased expression in PTC compared with normal thyroid tissues. The marked differential expression observed in our dataset supports its potential relevance in PTC biology. It functions as a peptidase involved in chemokine processing, signal transduction, and glucose metabolism [42]. Previous studies have reported that elevated *DPP4* expression is associated with poor prognosis in thyroid cancer and that its inhibition may affect pathways related to tumor progression and resistance to programmed cell death [43].

PDGFC was identified among the selected biomarkers and showed increased expression in PTC samples. Although it was not associated with survival in our cohort, previous studies have reported an inverse correlation between *PDGFC* expression and lymphocyte infiltration in PTC tissues [44]. This observation may suggest a role in influencing the tumor microenvironment and could help explain its selection among the most informative genes. Additionally, in this work, we found the gene to be associated with an antineoplastic agent, namely, Sunitinib.

TENM1 was upregulated in PTC samples compared with normal thyroid tissues in our analysis. Furthermore, *TENM1* expression was significantly associated with overall survival, supporting its potential prognostic relevance. This expression pattern is consistent with previous studies reporting increased *TENM1* expression in thyroid cancer. Prior evidence has suggested a possible involvement of *TENM1* in angiogenesis and tumor progression, which may provide a biological explanation for the dysregulation observed in our dataset [45].

PDZK1IP1 exhibited expression differences between PTC and normal tissues and was significantly associated with both overall survival and progression-free intervals in our cohort, highlighting its potential prognostic utility. Previous enrichment analyses have linked *PDZK1IP1* to neutrophil activation and other immune-related pathways in PTC [46].

GLT1D1 was identified among the biomarkers selected by the integrative DEG–LASSO approach and was significantly associated with overall survival in our cohort, suggesting potential prognostic relevance in PTC. Although the specific role of *GLT1D1* in thyroid cancer remains poorly characterized, previous studies have linked this gene to tumor immune evasion mechanisms. In lymphoma models, *GLT1D1* overexpression promoted tumor growth by increasing *PD-L1* glycosylation, thereby facilitating immune escape, whereas *GLT1D1* downregulation reduced glycosylated *PD-L1* levels and enhanced cytotoxic T-cell activity [47].

Additionally, post-translational modifications have been implicated in *PD-L1* regulation in other malignancies, with succinylation being associated with increased *PD-L1* expression in prostate cancer [48]. Together, these findings suggest that *GLT1D1* may contribute to pathways involved in *PD-L1* regulation and tumor–immune interactions.

IGSF10 was identified among the candidate biomarkers selected, and its expression is reduced in PTC samples. Although the biological role of *IGSF10* in thyroid cancer remains poorly understood, accumulating evidence suggests that it may be involved in tumor progression and immune-related processes. A pan-cancer analysis demonstrated that *IGSF10* expression was associated with patient prognosis and showed strong correlations with immune cell infiltration, immune checkpoints, and several immunomodulatory factors [49]. In addition, as we observed in PTC samples, reduced *IGSF10* expression has been reported in breast cancer, where lower expression levels were associated with shorter overall and relapse-free survival [50].

RPS16P5 had a decreased expression in thyroid tissues in our validation analysis. It is a retrocopy (processed pseudogene) belonging to the ribosomal protein S16 gene family, whose biological function remains largely uncharacterized. Previous transcriptomic analyses have reported subtype-specific expression patterns of retrogenes across different cancers.

In breast cancer, *RPS16P5* was found to be differentially expressed in the triple-negative subtype among a set of retrogenes exhibiting distinct regulation patterns associated with tumor subtype specificity [51]. Although functional evidence for *RPS16P5* is still lacking, retrogenes have been increasingly recognized as potential regulators of gene expression and tumor biology, and their differential expression has been associated with malignant phenotypes in several cancer types [51]. However, the specific role of RPS16P5 in thyroid cancer remains unknown and requires further investigation.

4.2. Survival Analyses of the Selected Genes

The most noticeable difference ($p = 0.00507$) lies in *TIMP4* (Figure 4D), which, according to Kaplan–Meier survival curves, the higher the expression, the worse the outcome. It regulates carcinogenesis through enriching the stem cell population in cervical cancer cells, such as the thyroid itself, giving it volume [52]. Additionally, *TENM1* ($p = 0.036$) (Figure 4C) was already reported in PTC associated with the MAPK signaling pathway [45], reported to be responsible for cell proliferation, differentiation, migration, senescence, and apoptosis [53].

Changes in the RAS-MAPK pathway have been reported in human cancer, including PTC, with abnormal behavior mainly in the RAS or RAF proteins (associated with the development of thyroid carcinomas) [54]. Additionally, the *BRAF* gene is also a component of the MAPK cascade of reactions and has also been reported in PTC. Strategies focusing on using the RAS-MAPK pathway as a therapeutic target have already been discussed in the literature [55].

The only progression-free interval curve that returned significant value was the *PDZK1IP1* curve, with $p = 0.0398$. It acts as an oncogene and modulates the tumor immune microenvironment on PTC [46], but also can be related to the proliferation, migration, invasion, apoptosis, and cell cycle arrest of PTC cells [56].

Additionally, the gene was overexpressed in the tumoral group (Figure 3), indicating its potential as a biomarker. However, this observation should be taken with a grain of salt due to a conflicting result: the overall survival curve indicates that the lower the expression, the worse the outcome, whereas in the progression-free interval curve, the higher the expression, the worse the outcome.

4.3. Biological Interpretation of Enriched Pathways

The pathway returned spans basic cellular function and organization to regulatory functions. The cytoskeleton in muscle cells and cell adhesion molecules provide support to cancer to thrive, favoring cell–cell adhesion, which can considerably increase its size, leading to symptoms like local compression.

On the other hand, it can modulate inflammatory response, such as through cytokine–cytokine receptor interaction, as well as through tumor-related pathways (e.g., transcriptional misregulation in cancer and the p53 signaling pathway) and signaling mechanisms (extracellular matrix–receptor interaction), as compiled in Supplementary Table S6.

The enrichment analysis reveals a profound reorganization of cellular architecture and its interface with the tumor microenvironment. The enrichment of pathways related to ECM–receptor interaction, cell adhesion, and extracellular matrix organization indicates that the DEGs are strongly associated with the Epithelial–Mesenchymal Transition (EMT). It is a process where epithelial cells lose their tight cell–cell contacts and polarity and acquire mesenchymal traits that promote movement and invasion [57]. EMT contributes to cancer dissemination by downregulating epithelial markers and upregulating mesenchymal/ECM-remodeling genes, enabling invasion and survival in the microenvironment [58].

This process is fundamental to thyroid cancer progression, as it enables follicular cells to lose their polarity and cohesion, which is partly mediated by changes in homophilic cell–cell adhesion, and acquire a migratory and invasive phenotype [59,60]. Furthermore, the enrichment of angiogenesis and ephrin receptor signaling corroborates the tumor's capacity to remodel the local vasculature, facilitating metastatic dissemination to cervical lymph nodes—a common mark of aggressive PTC [61–63].

Simultaneously, the prominence of the inflammatory response reflects the intimate relationship between the immune microenvironment and thyroid oncogenesis [64]. The tumor appropriates these mechanisms to recruit immune cells that, rather than providing surveillance, secrete cytokines that promote proliferation and further tissue remodeling [64–66]. This chronic inflammatory state is often linked to the hydrogen peroxide catabolic process. While normal thyroid cells produce H_2O_2 for hormone synthesis, the dysregulation of this system in cancer leads to oxidative stress that activates survival cascades like PI3K/Akt signaling [67,68].

This is coupled with a marked loss of follicular identity, where thyroid hormone generation is downregulated as the cell redirects its energy toward cell population proliferation and fatty acid transport [69]. The metabolic shift promotes rapid growth and marks the transition toward radioiodine-refractory tumors, representing a significant challenge in clinical management.

4.4. Drug Interaction with Selected Biomarkers

Several proteins have already been described as therapeutic molecular targets for thyroid cancer, with response rates between 6% and 89%. Among these targets are *PDGFR*, *RET*, *KIT*, and others [70]. In this context, many studies have used molecular docking to identify new compounds targeting these receptors, as well as to evaluate potential ligands for novel proposed therapeutic targets. Studies that evaluated the affinity of new compounds or validated docking protocols through redocking of the RET protein, for example, reported binding energies around −8 kcal/mol, guiding the authors to conclude that the investigated compounds, potential anticancer agents, may inhibit pRET tyrosine kinase activity, although further in vitro and in vivo validation is needed [71].

Another study assessing the interaction of compounds with the DPP4 protein reported docking scores between −7 and −8 kcal/mol; the authors concluded that molecular docking analysis demonstrated promising binding affinities of sitagliptin toward the evaluated

oncogenes, exceeding those observed for the reference inhibitors, making it a potential therapeutic option [72].

Overall, these reports support the results obtained in our study in an exploratory and hypothesis-generating context. Binding energies ranging from −7 to −9 kcal/mol, combined with favorable interactions between the ligand and key amino acid residues of the target protein, suggest that the investigated compounds in our research are promising candidates for further research and development as potential therapies for thyroid cancer.

Sunitinib is approved for the treatment of imatinib-resistant gastrointestinal stromal tumors (GISTs), renal carcinoma, and pancreatic neuroendocrine tumors, and as an oral tyrosine kinase inhibitor (TKI), it inhibits several receptors in the body, including important receptors for PTC, such as PDGF-R, VEGFR, FLT3, and RET, reducing tumor vascularization, which can reduce tumor size [73].

Nintedanib esylate is a triple angiokinase inhibitor that simultaneously blocks VEGFR 1–3, FGFR 1–3, and PDGFRα/β receptors, conferring particular biological relevance in the context of PTC. In PTC, these three receptors are frequently overexpressed, contributing to both tumor angiogenesis and the maintenance of the neoplastic microenvironment, including pericytes and smooth muscle cells involved in tumor vascularization [74].

Based on this molecular rationale, Nintedanib was evaluated in the randomized, double-blind, placebo-controlled phase II study EORTC-1209 (NCT01788982) [74,75]. Although the study did not demonstrate a clinically significant improvement in progression-free survival over the placebo, the triple inhibition profile of Nintedanib distinguishes it from the more selective approved inhibitors for this indication, such as Lenvatinib and Sorafenib, and may be of interest in specific molecular subgroups harboring *FGFR* or *PDGFR* amplification, targets that remain largely unexplored in standard therapies for advanced PTC. Importantly, Nintedanib esylate has been evaluated across multiple tumor types, where it has demonstrated favorable clinical outcomes [76,77].

Masitinib is an orally administered tyrosine kinase inhibitor primarily targeting the c-Kit receptor, while also inhibiting *PDGFRα/β*, *FAK*, *Lck*, and, to a lesser extent, *FGFR3* and *CSF1R* [78]. In the context of PTC, two of these targets have well-established and directly relevant molecular justifications. *PDGFRα*, in particular, plays a documented role in the metastatic progression of papillary thyroid carcinoma [79]. Members of the PDGF ligand family, including *PDGFA*, *PDGFB*, and *PDGFC*, are significantly upregulated in PTC compared to non-neoplastic thyroid tissue, further reinforcing the role of this pathway in tumor biology [44].

The loss of *c-KIT* expression is important in PTC [80], and masitinib demonstrated antitumor activity in c-KIT–negative hepatocellular carcinoma cells by inducing apoptosis through oxidative stress and JNK pathway activation, highlighting a promising alternative mechanism for cancer therapy [81]. The antineoplastic activity of masitinib in PTC identified by the DGIdb appears to be predominantly mediated through *PDGFRα* inhibition, a target with established relevance to PTC aggressiveness, making it a rational candidate for investigation in tumors overexpressing this receptor, particularly those with advanced or refractory lymph node disease.

Quizartinib is a second-generation inhibitor of class III receptor tyrosine kinases, acting on FLT3, KIT, and PDGFRα/β, with an excellent pharmacokinetic profile and high potency against mutant isoforms of these receptors [70].

In PTC specifically, FLT3 and c-KIT are part of a set of receptors with documented roles in thyroid neoplastic development, being targeted by multikinase inhibitors that have demonstrated antitumor activity in this context [82,83]. Recently, a Phase 3 clinical trial compared quizartinib plus standard chemotherapy versus placebo in patients with *FLT3-ITD*-positive acute myeloid leukemia and found that adding quizartinib significantly

improved overall survival without a meaningful increase in toxicity, suggesting it is an effective and generally well-tolerated treatment option for this subtype of the disease [84].

The identification of quizartinib by the DGIdb as an antineoplastic agent relevant to PTC is therefore grounded in the shared molecular targets (*FLT3*, *KIT*, and *PDGFR*) whose expression has been documented in thyroid tumors, positioning this drug as a promising candidate for repurposing strategies, particularly in PTC with increased expression profiles of these kinases or in settings of resistance to currently approved targeted therapies.

4.5. Data Integration in PTCNetwork Analysis Insights on PTC

These hub nodes likely represent critical regulators linking transcriptional and metabolic alterations in PTC. Norepinephrine is the biggest hub in the network, comprising 24 interactions. It can be related to promoting the Warburg effect, serving cancer as an alternative bioenergetics system [85], where cells oxidize glucose into lactic acid even with oxygen bioavailable for conventional glycolysis [86].

Additionally, arachidonic acid presented as an important member of the network, with 23 links. Inflammation is a well-known process fundamentally related to cancer progression [87]. Arachidonic acid is a well-known inflammatory and immune precursor molecule [88], and can be involved in several different processes in cancer once its enzymes (e.g., COX, LOX) are well studied [88], which were reported to be altered in other solid tumors [89].

Glutamic acid, on the other hand, is also important, with 21 connections. It is used as a carbon substrate for lipid synthesis and as a nitrogen donor [90]. In other work [91], it appears upregulated in PTC patients when compared to healthy volunteers, such as in this study, and it supposedly has effects in thyroid-stimulating hormone (TSH) production [92].

Additionally, the genes that link these hubs (norepinephrine, arachidonic acid, and glutamic acid), organized in Supplementary Table S9, show the tendency of centrality importance, but the relationships among them are poorly described and need more exploration.

The *SLC6A14* gene is presented as the most important of the genes in the network; it was found to be upregulated (as in this work) in most solid tumors [93], and its high concentration is an unfavorable prognostic indicator in other tumors. SLC6A14 is a solute transporter that supplies amino acids to the cell, supporting the growth and proliferation of cancer cells [94].

Other work using the same dataset found *ADK* in their analysis as a suggestive biomarker for PTC; here, we confirm this finding. In their work, ADK was found to participate in succinate metabolism and m6A modification of RNA, which could predict PTC [95]. Also, ADK was reported to convert extracellular adenosine to AMP [96], causing global hypermethylation [97,98].

Thus, the integrative network can provide novel insights into disease mechanisms and highlight potential targets for further investigation.

4.6. Limitations and Future Perspectives

The main limitation of this study concerns metabolomics data availability. A better approach would be to retrieve the metabolomics data, reprocess, and annotate the features. Unfortunately, this step of the process, besides being developed in-house, is only possible with the publication of metabolomics data in public repositories; only a few articles published the actual final feature table, making viable a sort of mining and integration of data by using a molecular network strategy.

On the other hand, the transcriptome datasets used have a reasonable but limited sample size, leading to unreliable findings. Due to sampling, it represents European people the most.

Hereafter, this pipeline will be tested with other cancers, integrating upstream with downstream data, preferably with an in vivo study design. The best scenario would be to perform an in-house RNA-seq and untargeted metabolomics analysis of the same samples to further explore and confirm these results. Additionally, a better way to integrate omics is not necessarily with molecular networks, even if it works, but with a well-designed experiment focused on multi-omics data merge. Arguably, molecular networking is still necessary as not everyone who has conducted studies have uploaded their data publicly to be systematically reviewed.

5. Conclusions

In this work, a methodology was demonstrated using publicly available data to intersect genes found in differential analysis (from 23,520 to 12) with the LASSO model (11 selected), achieving 11 selected genes. The intersection between DEGs and LASSO was a distinct technique to select the most relevant genes, as shown in the independent validation set curve (AUC = 91%, Sens. = 92%, Spec. = 97%, and Acc. = 95%). In addition, we performed survival analysis to assess the prognostic relevance of the 11 selected biomarkers, and four of them returned statistical differences, namely, *GLT1D1*, *PDZK1IP1*, *TENM1*, and *TIMP4*.

Merging different data types was possible due to network analysis, and with those interactions among metabolites and genes, biological discussion and insights were generated. Bioenergetics seems to be the main mechanism highlighted by this data, pointing to the Warburg effect as an important player behind PTC biology. In addition, inflammatory molecules, such as arachidonic acid and oleic acids, can regulate the inflammation status of the cancer, and once these molecules are a hub node in the network, the role of inflammation should be considered when describing PTC.

Drug–gene interactions were evaluated, and the observed result showed PDGFC-sunitinib and DPP4-nintedanib esylate, -masitinib, -quizartinib, and -olaratumab as ligand pairs. Afterward, molecular docking was performed to evaluate those findings; the binding energy demonstrates that the actual binding of the drug with the gene can happen, serving as a pointer regarding where to go in terms of therapeutic strategies.

Generally, this work has shown how well DEG–LASSO intersection performed, using multiple variables and restricting them to the most important ones. Complementary analyses were performed to further explore the potential of the genes as biomarkers. The network has shown a multi-layer integration of data types, bringing to light newer discussions in the PTC clinical context, such as the extent of the Warburg effect and inflammatory status, and novel competitors to biomarkers.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/biology15131004/s1, Figure S1: Complete gene-metabolite network displaying in squared shapes the metabolites, and in oval the genes. Colors are mapped to the fold–change values; Table S1: All the Differential Expressed Genes and their statistical values; Table S2: LASSO coefficients different than zero; Table S3: Genes and metabolites and their main function in cancer; Table S4: Overall Survival Analysis results of the selected genes; Table S5: Progression-free Intervals from the selected genes; Table S6: Enrichment Analysis of the DEGs (FDR < 0.05); Table S7: Drug-gene interaction table; Table S8: Metabolites table comprising all used metabolites with respective *p*-values, fold-change and PLS-DA's VIP score; Table S9: Genes that link the hub metabolites and its function; Table S10: Over Representation Analysis of the metabolites; Table S11: Input list used to perform some of the analysis; Table S12: Pathways in which the selected genes play a row based on KEGG; Table S13: Prediction evaluation of each of the selected genes; Table S14: All scores generated in the network analysis, ranking the metabolites and genes.

Author Contributions: Conceptualization, P.H.G.S. and L.M.d.C.; methodology, P.H.G.S. and L.M.d.C.; software, P.H.G.S.; validation, P.H.G.S., N.C.d.M. and D.C.d.O.; formal analysis, P.H.G.S.; investigation, P.H.G.S. and L.M.d.C.; resources, L.M.d.C.; data curation, P.H.G.S., N.C.d.M., D.C.d.O. and L.M.d.C.; writing—original draft preparation, P.H.G.S., N.C.d.M. and L.M.d.C.; writing—review and editing, P.H.G.S., N.C.d.M., D.C.d.O. and L.M.d.C.; visualization, P.H.G.S., N.C.d.M. and D.C.d.O.; supervision, L.M.d.C.; project administration, P.H.G.S. and L.M.d.C.; funding acquisition, P.H.G.S., N.C.d.M. and L.M.d.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the São Paulo Research Foundation (FAPESP) grant (#2024/21955-3 from N.C.M; #2022/14179-1 from P.H.G.S.).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data supporting the findings in this study are included within the Supplementary File. Gene expression data analyzed during the current study are available in the Gene Expression Omnibus (GEO) repository under persistent accession numbers GSE29265 and GSE33630. Metabolites were collected from [99–101].

Acknowledgments: The authors acknowledge the São Francisco University for project maintenance, and gratefully thank the São Paulo Research Foundation (FAPESP) for the funding.

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

Abbreviations

The following abbreviations are used in this manuscript:

TC	Thyroid Carcinoma
PTC	Papillary Thyroid Carcinoma
FTC	Follicular Thyroid Carcinoma
MTC	Medullary Thyroid Carcinoma
TSH	Thyrostimulating Hormone
ATA	American Thyroid Association
GEO	Gene Expression Omnibus
FDR	False Discovery Rate
LASSO	Least Absolute Shrinkage and Selection Operator
ROC	Receiver Operating Characteristic
DEG	Differentially Expressed Genes
PCA	Principal Component Analysis
AUC	Area Under the Curve
TCGA	The Cancer Genome Atlas
DAVID	The Database for Annotation, Visualization, and Integrated Discovery
DPP4	Dipeptidyl Peptidase 4
GLT1D1	Glycosyltransferase 1 Domain Containing 1
TENM1	Teneurin Transmembrane Protein 1
TIMP4	TIMP Metallopeptidase Inhibitor 4
PDZK1IP1	PDZK1 Interacting Protein 1
ABHD17C	Abhydrolase Domain Containing 17C, Depalmitoylase
IGSF10	Immunoglobulin Superfamily Member 10
RHOU	Ras Homolog Family Member U
PDGFC	Platelet-Derived Growth Factor C
ABCA6	ATP Binding Cassette Subfamily A Member 6
RPS16P5	Ribosomal Protein S16 Pseudogene 5
RAS	RAS GTPase
MAPK	Mitogen-activated Protein Kinase
COX	Cyclooxygenase

LOX	Lipooxygenase
iPLA2β	Calcium-independent phospholipase A2β
SLC6A14	Solute Carrier Family 6 Member 14
ADK	Adenosine Kinase
RNA	Ribonucleic acid
AMP	Adenosine monophosphate
SYT1	Synaptotagmin 1
CYSLTR2	Cysteinyl Leukotriene Receptor 2
PPARG	Peroxisome Proliferator Activated Receptor Gamma
SLC4A4	Solute Carrier Family 4 Member 4
CYCS	Cytochrome C, Somatic
JUN	Jun Proto-Oncogene, AP-1 Transcription Factor Subunit

References

1. Wiltshire, J.J.; Drake, T.M.; Uttley, L.; Balasubramanian, S.P. Systematic Review of Trends in the Incidence Rates of Thyroid Cancer. *Thyroid* **2016**, *26*, 1541–1552. [CrossRef] [PubMed]
2. Coca-Pelaz, A.; Rodrigo, J.P.; Zafereo, M.; Nixon, I.; Pace-Asciak, P.; Randolph, G.W.; Suárez, C.; Ferlito, A. Molecular Diagnostics and Personalized Therapeutics in Differentiated Thyroid Carcinoma: A Clinically Oriented Review. *Diagnostics* **2025**, *15*, 2493. [CrossRef] [PubMed]
3. Li, Y.-D.; Ye, Q.-Y.; Chen, Y.-X.; Hu, X.-R. Thyroid Cancer: Pathogenesis, Clinicopathology, Diagnosis, and Management. *MedComm* **2025**, *6*, e70449. [CrossRef] [PubMed]
4. Limaiem, F.; Rehman, A.; Mazzoni, T. Papillary Thyroid Carcinoma. In *StatPearls*; StatPearls Publishing: Treasure Island, FL, USA, 2026.
5. Bogović Crnčić, T.; Ilić Tomaš, M.; Girotto, N.; Grbac Ivanković, S. Risk Factors for Thyroid Cancer: What Do We Know So Far? *Acta Clin. Croat.* **2020**, *59*, 66–72. [CrossRef] [PubMed]
6. Furukawa, K.; Preston, D.; Funamoto, S.; Yonehara, S.; Ito, M.; Tokuoka, S.; Sugiyama, H.; Soda, M.; Ozasa, K.; Mabuchi, K. Long-Term Trend of Thyroid Cancer Risk among Japanese Atomic-Bomb Survivors: 60 Years after Exposure. *Int. J. Cancer* **2013**, *132*, 1222–1226. [CrossRef] [PubMed]
7. Ringel, M.D.; Sosa, J.A.; Baloch, Z.; Bischoff, L.; Bloom, G.; Brent, G.A.; Brock, P.L.; Chou, R.; Flavell, R.R.; Goldner, W.; et al. 2025 American Thyroid Association Management Guidelines for Adult Patients with Differentiated Thyroid Cancer. *Thyroid* **2025**, *35*, 841–985. [CrossRef] [PubMed]
8. Park, K.S.; Kim, S.H.; Oh, J.H.; Kim, S.Y. Highly Accurate Diagnosis of Papillary Thyroid Carcinomas Based on Personalized Pathways Coupled with Machine Learning. *Brief. Bioinform.* **2021**, *22*, bbaa336. [CrossRef] [PubMed]
9. Yan, L.; Su, P.; Sun, X. Role of Multi-omics in Advancing the Understanding and Treatment of Prostate Cancer (Review). *Mol. Med. Rep.* **2025**, *31*, 130. [CrossRef] [PubMed]
10. Supplitt, S.; Karpinski, P.; Sasiadek, M.; Laczmanska, I. Current Achievements and Applications of Transcriptomics in Personalized Cancer Medicine. *Int. J. Mol. Sci.* **2021**, *22*, 1422. [CrossRef] [PubMed]
11. Jaiswal, I.; Kulkarni, R.; Singh, G.; Rindhe, V.; Patil, K. Proteomics in Personalized Cancer Therapy: Advances, Applications, and Future Perspectives. *Int. J. Sci. Res. Technol.* **2025**, *2*, 127–146. [CrossRef]
12. Zeng, Z.; Chen, C.-X. Metabonomic Analysis of Tumor Microenvironments: A Mini-Review. *Front. Oncol.* **2023**, *13*, 1164266. [CrossRef] [PubMed]
13. Correa-Aguila, R.; Alonso-Pupo, N.; Hernández-Rodríguez, E.W. Multi-Omics Data Integration Approaches for Precision Oncology. *Mol. Omics* **2022**, *18*, 469–479. [CrossRef] [PubMed]
14. Ren, H.; Liu, X.; Li, F.; He, X.; Zhao, N. Identification of a Six Gene Prognosis Signature for Papillary Thyroid Cancer Using Multi-Omics Methods and Bioinformatics Analysis. *Front. Oncol.* **2021**, *11*, 624421. [CrossRef] [PubMed]
15. Alnwisser, B.; Alshehri, S.; Qattan, A.; Zou, M.; Aboussekhra, A.; Alzahrani, A.S.; Shi, Y. Identification of Novel Prognostic Biomarkers for Thyroid Cancer by Integrated Transcriptome Analysis of Metastasis-Associated Genes. *Front. Oncol.* **2025**, *15*, 1536270. [CrossRef] [PubMed]
16. Al-Madhagi, H. Multidimensional Transcriptomic and Immune Profiling of Papillary Thyroid Cancer: Bridging Molecular Networks, Disease Stage, Survival Outcomes, and Drug Discovery. *Discov. Oncol.* **2025**, *16*, 1760. [CrossRef] [PubMed]
17. Hsu, C.-Y.; Askar, S.; Alshkarchy, S.S.; Nayak, P.P.; Attabi, K.A.L.; Khan, M.A.; Mayan, J.A.; Sharma, M.K.; Islomov, S.; Soleimani Samarkhazan, H. AI-Driven Multi-Omics Integration in Precision Oncology: Bridging the Data Deluge to Clinical Decisions. *Clin. Exp. Med.* **2025**, *26*, 29. [CrossRef] [PubMed]

18. Lee, K.-S.; Park, H.; Lee, K.-S.; Park, H. Machine Learning on Thyroid Disease: A Review. *Front. Biosci.-Landmark* **2022**, *27*, 101. [CrossRef] [PubMed]
19. Fiehn, O. FatB Gene Project—PR000001. Available online: http://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Project&ProjectID=PR000001 (accessed on 8 May 2026).
20. Yurekten, O.; Payne, T.; Tejera, N.; Amaladoss, F.X.; Martin, C.; Williams, M.; O'Donovan, C. MetaboLights: Open Data Repository for Metabolomics. *Nucleic Acids Res.* **2024**, *52*, D640–D646. [CrossRef] [PubMed]
21. Pang, Z.; Lu, Y.; Zhou, G.; Hui, F.; Xu, L.; Viau, C.; Spigelman, A.F.; MacDonald, P.E.; Wishart, D.S.; Li, S.; et al. MetaboAnalyst 6.0: Towards a Unified Platform for Metabolomics Data Processing, Analysis and Interpretation. *Nucleic Acids Res.* **2024**, *52*, W398–W406. [CrossRef] [PubMed]
22. Otasek, D.; Morris, J.H.; Bouças, J.; Pico, A.R.; Demchak, B. Cytoscape Automation: Empowering Workflow-Based Network Analysis. *Genome Biol.* **2019**, *20*, 185. [CrossRef] [PubMed]
23. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res.* **2015**, *43*, e47. [CrossRef] [PubMed]
24. Friedman, J.H.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [CrossRef]
25. Tang, Z.; Kang, B.; Li, C.; Chen, T.; Zhang, Z. GEPIA2: An Enhanced Web Server for Large-Scale Expression Profiling and Interactive Analysis. *Nucleic Acids Res.* **2019**, *47*, W556–W560. [CrossRef] [PubMed]
26. Cannon, M.; Stevenson, J.; Stahl, K.; Basu, R.; Coffman, A.; Kiwala, S.; McMichael, J.F.; Kuzma, K.; Morrissey, D.; Cotto, K.; et al. DGIdb 5.0: Rebuilding the Drug–Gene Interaction Database for Precision Medicine and Drug Discovery Platforms. *Nucleic Acids Res.* **2024**, *52*, D1227–D1235. [CrossRef] [PubMed]
27. Sherman, B.T.; Panzade, G.; Imamichi, T.; Chang, W. DAVID Ortholog: An Integrative Tool to Enhance Functional Analysis through Orthologs. *Bioinformatics* **2024**, *40*, btae615. [CrossRef] [PubMed]
28. Sanches, P.H.G.; de Melo, N.C.; Porcari, A.M.; de Carvalho, L.M. Integrating Molecular Perspectives: Strategies for Comprehensive Multi-Omics Integrative Data Analysis and Machine Learning Applications in Transcriptomics, Proteomics, and Metabolomics. *Biology* **2024**, *13*, 848. [CrossRef] [PubMed]
29. Cheng, B.; Yuan, W.-E.; Su, J.; Liu, Y.; Chen, J. Recent Advances in Small Molecule Based Cancer Immunotherapy. *Eur. J. Med. Chem.* **2018**, *157*, 582–598. [CrossRef] [PubMed]
30. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef] [PubMed]
31. Trott, O.; Olson, A.J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461. [CrossRef] [PubMed]
32. Morris, G.M.; Huey, R.; Lindstrom, W.; Sanner, M.F.; Belew, R.K.; Goodsell, D.S.; Olson, A.J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30*, 2785–2791. [CrossRef] [PubMed]
33. Landrum, G.; Tosco, P.; Kelley, B.; Rodriguez, R.; Cosgrove, D.; Vianello, R.; sriniker; Gedeck, P.; Jones, G.; Kawashima, E.; et al. Rdkit/Rdkit: 2026_03_2 (Q1 2026) Release 2026. Available online: https://zenodo.org/records/19922430 (accessed on 23 May 2026).
34. O'Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3*, 33. [CrossRef] [PubMed]
35. Ahmad, S.; Jose da Costa Gonzales, L.; Bowler-Barnett, E.H.; Rice, D.L.; Kim, M.; Wijerathne, S.; Luciani, A.; Kandasaamy, S.; Luo, J.; Watkins, X.; et al. The UniProt Website API: Facilitating Programmatic Access to Protein Knowledge. *Nucleic Acids Res.* **2025**, *53*, W547–W553. [CrossRef] [PubMed]
36. Varadi, M.; Bertoni, D.; Magana, P.; Paramval, U.; Pidruchna, I.; Radhakrishnan, M.; Tsenkov, M.; Nair, S.; Mirdita, M.; Yeo, J.; et al. AlphaFold Protein Structure Database in 2024: Providing Structure Coverage for over 214 Million Protein Sequences. *Nucleic Acids Res.* **2024**, *52*, D368–D375. [CrossRef] [PubMed]
37. Guerra, J.V.S.; Ribeiro-Filho, H.V.; Pereira, J.G.C.; Lopes-de-Oliveira, P.S. KVFinder-Web: A Web-Based Application for Detecting and Characterizing Biomolecular Cavities. *Nucleic Acids Res.* **2023**, *51*, W289–W297. [CrossRef] [PubMed]
38. Systèmes, D. Free Download: BIOVIA Discovery Studio Visualizer. Available online: https://discover.3ds.com/discovery-studio-visualizer-download (accessed on 8 May 2026).
39. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Meng, E.C.; Couch, G.S.; Croll, T.I.; Morris, J.H.; Ferrin, T.E. UCSF ChimeraX: Structure Visualization for Researchers, Educators, and Developers. *Protein Sci.* **2021**, *30*, 70–82. [CrossRef] [PubMed]
40. The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* **2025**, *53*, D609–D617. [CrossRef] [PubMed]
41. Liu, Y.; Chen, T.-Y.; Yang, Z.-Y.; Fang, W.; Wu, Q.; Zhang, C. Identification of Hub Genes in Papillary Thyroid Carcinoma: Robust Rank Aggregation and Weighted Gene Co-Expression Network Analysis. *J. Transl. Med.* **2020**, *18*, 170. [CrossRef] [PubMed]

42. Gao, X.; Le, Y.; Geng, C.; Jiang, Z.; Zhao, G.; Zhang, P. DPP4 Is a Potential Prognostic Marker of Thyroid Carcinoma and a Target for Immunotherapy. *Int. J. Endocrinol.* **2022**, *2022*, 5181386. [CrossRef] [PubMed]
43. Xie, Y.; Zhu, S.; Song, X.; Sun, X.; Fan, Y.; Liu, J.; Zhong, M.; Yuan, H.; Zhang, L.; Billiar, T.R.; et al. The Tumor Suppressor P53 Limits Ferroptosis by Blocking DPP4 Activity. *Cell Rep.* **2017**, *20*, 1692–1704. [CrossRef] [PubMed]
44. Bruland, O.; Fluge, Ø.; Akslen, L.A.; Eiken, H.G.; Lillehaug, J.R.; Varhaug, J.E.; Knappskog, P.M. Inverse Correlation between PDGFC Expression and Lymphocyte Infiltration in Human Papillary Thyroid Carcinomas. *BMC Cancer* **2009**, *9*, 425. [CrossRef] [PubMed]
45. Cheng, S.-P.; Chen, M.-J.; Chien, M.-N.; Lin, C.-H.; Lee, J.-J.; Liu, C.-L. Overexpression of Teneurin Transmembrane Protein 1 Is a Potential Marker of Disease Progression in Papillary Thyroid Carcinoma. *Clin. Exp. Med.* **2017**, *17*, 555–564. [CrossRef] [PubMed]
46. Geng, C.; Gao, X.; Le, Y.; Zhu, S.; Zhao, G.; Zhang, P. Pan-Cancer Analysis of PDZK1IP1 Reveals Its Role in Tumorigenesis and Tumor Immunity: Focused Validation in Thyroid Carcinoma. *Hereditas* **2026**, *163*, 45. [CrossRef] [PubMed]
47. Liu, X.; Zhang, Y.; Han, Y.; Lu, W.; Yang, J.; Tian, J.; Sun, P.; Yu, T.; Hu, Y.; Zhang, H.; et al. Overexpression of GLT1D1 Induces Immunosuppression through Glycosylation of PD-L1 and Predicts Poor Prognosis in B-Cell Lymphoma. *Mol. Oncol.* **2020**, *14*, 1028–1044. [CrossRef] [PubMed]
48. Wang, R.; He, S.; Long, J.; Wang, Y.; Jiang, X.; Chen, M.; Wang, J. Emerging Therapeutic Frontiers in Cancer: Insights into Posttranslational Modifications of PD-1/PD-L1 and Regulatory Pathways. *Exp. Hematol. Oncol.* **2024**, *13*, 46. [CrossRef] [PubMed]
49. Zhou, Y.; Gao, M.; Jing, Y.; Wang, X. Pan-Cancer Analyses Reveal IGSF10 as an Immunological and Prognostic Biomarker. *Front. Genet.* **2023**, *13*, 1032382. [CrossRef] [PubMed]
50. Wang, M.; Dai, M.; Wu, Y.; Yi, Z.; Li, Y.; Ren, G. Immunoglobulin Superfamily Member 10 Is a Novel Prognostic Biomarker for Breast Cancer. *PeerJ* **2020**, *8*, e10128. [CrossRef] [PubMed]
51. Staszak, K.; Makałowska, I. Cancer, Retrogenes, and Evolution. *Life* **2021**, *11*, 72. [CrossRef] [PubMed]
52. Lizarraga, F.; Espinosa, M.; Ceballos-Cancino, G.; Vazquez-Santillan, K.; Bahena-Ocampo, I.; Schwarz-Cruz y Celis, A.; Vega-Gordillo, M.; Garcia Lopez, P.; Maldonado, V.; Melendez-Zajgla, J. Tissue Inhibitor of Metalloproteinases-4 (TIMP-4) Regulates Stemness in Cervical Cancer Cells. *Mol. Carcinog.* **2016**, *55*, 1952–1961. [CrossRef] [PubMed]
53. Sun, Y.; Liu, W.-Z.; Liu, T.; Feng, X.; Yang, N.; Zhou, H.-F. Signaling Pathway of MAPK/ERK in Cell Proliferation, Differentiation, Migration, Senescence and Apoptosis. *J. Recept. Signal Transduct.* **2015**, *35*, 600–604. [CrossRef] [PubMed]
54. Czarniecka, A.; Oczko-Wojciechowska, M.; Barczyński, M. BRAF V600E Mutation in Prognostication of Papillary Thyroid Cancer (PTC) Recurrence. *Gland Surg.* **2016**, *5*, 495–505. [CrossRef] [PubMed]
55. Santarpia, L.; Lippman, S.M.; El-Naggar, A.K. Targeting the MAPK–RAS–RAF Signaling Pathway in Cancer Therapy. *Expert Opin. Ther. Targets* **2012**, *16*, 103–119. [CrossRef] [PubMed]
56. Zhang, W.; Zheng, D.; Jin, L.; Hirachan, S.; Bhandari, A.; Li, Y.; Chen, B.; Lu, Y.; Wen, J.; Lin, B.; et al. PDZK1IP1 Gene Promotes Proliferation, Migration, and Invasion in Papillary Thyroid Carcinoma. *Pathol.-Res. Pract.* **2022**, *238*, 154091. [CrossRef] [PubMed]
57. Huang, Z.; Zhang, Z.; Zhou, C.; Liu, L.; Huang, C. Epithelial–Mesenchymal Transition: The History, Regulatory Mechanism, and Cancer Therapeutic Opportunities. *MedComm* **2022**, *3*, e144. [CrossRef] [PubMed]
58. Rapanotti, M.C.; Cugini, E.; Scioli, M.G.; Cenci, T.; Anzillotti, S.; Puzzuoli, M.; Terrinoni, A.; Ferlosio, A.; De Luca, A.; Orlandi, A. The Clinical Relevance of Epithelial-to-Mesenchymal Transition Hallmarks: A Cut-Off-Based Approach in Healthy and Cancerous Cell Lines. *Int. J. Mol. Sci.* **2025**, *26*, 3617. [CrossRef] [PubMed]
59. Vasko, V.; Espinosa, A.V.; Scouten, W.; He, H.; Auer, H.; Liyanarachchi, S.; Larin, A.; Savchenko, V.; Francis, G.L.; de la Chapelle, A.; et al. Gene Expression and Functional Evidence of Epithelial-to-Mesenchymal Transition in Papillary Thyroid Carcinoma Invasion. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 2803–2808. [CrossRef] [PubMed]
60. Shakib, H.; Rajabi, S.; Dehghan, M.H.; Mashayekhi, F.J.; Safari-Alighiarloo, N.; Hedayati, M. Epithelial-to-Mesenchymal Transition in Thyroid Cancer: A Comprehensive Review. *Endocrine* **2019**, *66*, 435–455. [CrossRef] [PubMed]
61. Gucciardo, E.; Sugiyama, N.; Lehti, K. Eph- and Ephrin-Dependent Mechanisms in Tumor and Stem Cell Dynamics. *Cell. Mol. Life Sci.* **2014**, *71*, 3685–3710. [CrossRef] [PubMed]
62. Pasquale, E.B. Eph Receptors and Ephrins in Cancer Progression. *Nat. Rev. Cancer* **2024**, *24*, 5–27. [CrossRef] [PubMed]
63. Rudno-Rudzińska, J.; Kielan, W.; Frejlich, E.; Kotulski, K.; Hap, W.; Kurnol, K.; Dzierżek, P.; Zawadzki, M.; Hałoń, A. A Review on Eph/Ephrin, Angiogenesis and Lymphangiogenesis in Gastric, Colorectal and Pancreatic Cancers. *Chin. J. Cancer Res.* **2017**, *29*, 303–312. [CrossRef] [PubMed]
64. Cunha, L.L.; Marcello, M.A.; Ward, L.S. The Role of the Inflammatory Microenvironment in Thyroid Carcinogenesis. *Endocr. Relat. Cancer* **2014**, *21*, R85–R103. [CrossRef] [PubMed]
65. Ferrari, S.M.; Fallahi, P.; Galdiero, M.R.; Ruffilli, I.; Elia, G.; Ragusa, F.; Paparo, S.R.; Patrizio, A.; Mazzi, V.; Varricchi, G.; et al. Immune and Inflammatory Cells in Thyroid Cancer Microenvironment. *Int. J. Mol. Sci.* **2019**, *20*, 4413. [CrossRef] [PubMed]
66. Song, M.; Liu, Q.; Sun, W.; Zhang, H. Crosstalk between Thyroid Carcinoma and Tumor-Correlated Immune Cells in the Tumor Microenvironment. *Cancers* **2023**, *15*, 2863. [CrossRef] [PubMed]
67. Xing, M. Oxidative Stress: A New Risk Factor for Thyroid Cancer. *Endocr. Relat. Cancer* **2012**, *19*, C7–C11. [CrossRef] [PubMed]

68. Szanto, I.; Pusztaszeri, M.; Mavromati, M. H_2O_2 Metabolism in Normal Thyroid Cells and in Thyroid Tumorigenesis: Focus on NADPH Oxidases. *Antioxidants* **2019**, *8*, 126. [CrossRef] [PubMed]
69. Ping, P.; Ma, Y.; Xu, X.; Li, J. Reprogramming of Fatty Acid Metabolism in Thyroid Cancer: Potential Targets and Mechanisms. *Chin. J. Cancer Res.* **2025**, *37*, 227–249. [CrossRef] [PubMed]
70. Gil-Bernabé, S.; García-DeLaFuente, L.; García-Rostán, G. The Revolution of Targeted Therapies in Thyroid Cancer Treatment: Present and Future Promising Anti-Cancer Drugs. *Int. J. Mol. Sci.* **2025**, *26*, 3663. [CrossRef] [PubMed]
71. Talukder, M.E.K.; Aktaruzzaman, M.; Siddiquee, N.H.; Islam, S.; Wani, T.A.; Alkahtani, H.M.; Zargar, S.; Raihan, M.O.; Rahman, M.M.; Pokhrel, S.; et al. Cheminformatics-Based Identification of Phosphorylated RET Tyrosine Kinase Inhibitors for Human Cancer. *Front. Chem.* **2024**, *12*, 1407331. [CrossRef] [PubMed]
72. Cheng, S.-Y.; Wu, A.T.H.; Batiha, G.E.-S.; Ho, C.-L.; Lee, J.-C.; Lukman, H.Y.; Alorabi, M.; AlRasheedi, A.N.; Chen, J.-H. Identification of DPP4/CTNNB1/MET as a Theranostic Signature of Thyroid Cancer and Evaluation of the Therapeutic Potential of Sitagliptin. *Biology* **2022**, *11*, 324. [CrossRef] [PubMed]
73. Ferrari, S.M.; Centanni, M.; Virili, C.; Miccoli, M.; Ferrari, P.; Ruffilli, I.; Ragusa, F.; Antonelli, A.; Fallahi, P. Sunitinib in the Treatment of Thyroid Cancer. *Curr. Med. Chem.* **2019**, *26*, 963–972. [CrossRef] [PubMed]
74. Leboulleux, S.; Kapiteijn, E.; Litière, S.; Schöffski, P.; Godbert, Y.; Rodien, P.; Jarzab, B.; Salvatore, D.; Zanetta, S.; Capdevila, J.; et al. Safety and Efficacy of Nintedanib as Second-Line Therapy for Patients with Differentiated or Medullary Thyroid Cancer Progressing after First-Line Therapy. A Randomized Phase II Study of the EORTC Endocrine Task Force (Protocol 1209-EnTF). *Front. Endocrinol.* **2024**, *15*, 1403687. [CrossRef] [PubMed]
75. Awasthi, N.; Schwarz, R.E. Profile of Nintedanib in the Treatment of Solid Tumors: The Evidence to Date. *OncoTargets Ther.* **2015**, *8*, 3691–3701. [CrossRef] [PubMed]
76. Gabasa, M.; Ikemori, R.; Hilberg, F.; Reguart, N.; Alcaraz, J. Nintedanib Selectively Inhibits the Activation and Tumour-Promoting Effects of Fibroblasts from Lung Adenocarcinoma Patients. *Br. J. Cancer* **2017**, *117*, 1128–1138. [CrossRef] [PubMed]
77. Pan, L.; Meng, F.; Wang, W.; Wang, X.; Shen, H.; Bao, P.; Kang, J.; Kong, D. Nintedanib in an Elderly Non-Small-Cell Lung Cancer Patient with Severe Steroid-Refractory Checkpoint Inhibitor-Related Pneumonitis: A Case Report and Literature Review. *Front. Immunol.* **2023**, *13*, 1072612. [CrossRef] [PubMed]
78. Dubreuil, P.; Letard, S.; Ciufolini, M.; Gros, L.; Humbert, M.; Castéran, N.; Borge, L.; Hajem, B.; Lermet, A.; Sippl, W.; et al. Masitinib (AB1010), a Potent and Selective Tyrosine Kinase Inhibitor Targeting KIT. *PLoS ONE* **2009**, *4*, e7258. [CrossRef] [PubMed]
79. Zhang, J.; Wang, P.; Dykstra, M.; Gelebart, P.; Williams, D.; Ingham, R.; Adewuyi, E.E.; Lai, R.; McMullen, T. Platelet-Derived Growth Factor Receptor-α Promotes Lymphatic Metastases in Papillary Thyroid Cancer. *J. Pathol.* **2012**, *228*, 241–250. [CrossRef] [PubMed]
80. Franceschi, S.; Lessi, F.; Panebianco, F.; Tantillo, E.; Ferla, M.L.; Menicagli, M.; Aretini, P.; Apollo, A.; Naccarato, A.G.; Marchetti, I.; et al. Loss of C-KIT Expression in Thyroid Cancer Cells. *PLoS ONE* **2017**, *12*, e0173913. [CrossRef] [PubMed]
81. Semba, Y.; Yamamoto, S.; Takahashi, S.; Shinomiya, T.; Nagahara, Y. C-Kit Inhibitor Masitinib Induces Reactive Oxygen Species-Dependent Apoptosis in c-Kit-Negative HepG2 Cells. *Eur. J. Pharmacol.* **2022**, *931*, 175183. [CrossRef] [PubMed]
82. Kampa-Schittenhelm, K.M.; Heinrich, M.C.; Akmut, F.; Döhner, H.; Döhner, K.; Schittenhelm, M.M. Quizartinib (AC220) Is a Potent Second Generation Class III Tyrosine Kinase Inhibitor That Displays a Distinct Inhibition Profile against Mutant-FLT3, -PDGFRA and -KIT Isoforms. *Mol. Cancer* **2013**, *12*, 19. [CrossRef] [PubMed]
83. Stjepanovic, N.; Capdevila, J. Multikinase Inhibitors in the Treatment of Thyroid Cancer: Specific Role of Lenvatinib. *Biol. Targets Ther.* **2014**, *8*, 129–139. [CrossRef] [PubMed]
84. Erba, H.P.; Montesinos, P.; Kim, H.-J.; Patkowska, E.; Vrhovac, R.; Žák, P.; Wang, P.-N.; Mitov, T.; Hanyok, J.; Kamel, Y.M.; et al. Quizartinib plus Chemotherapy in Newly Diagnosed Patients with FLT3-Internal-Tandem-Duplication-Positive Acute Myeloid Leukaemia (QuANTUM-First): A Randomised, Double-Blind, Placebo-Controlled, Phase 3 Trial. *Lancet* **2023**, *401*, 1571–1583. [CrossRef] [PubMed]
85. Fitzgerald, P.J. Norepinephrine Release May Play a Critical Role in the Warburg Effect: An Integrative Model of Tumorigenesis. *Neoplasma* **2020**, *67*, 947–957. [CrossRef] [PubMed]
86. Barba, I.; Carrillo-Bosch, L.; Seoane, J. Targeting the Warburg Effect in Cancer: Where Do We Stand? *Int. J. Mol. Sci.* **2024**, *25*, 3142. [CrossRef] [PubMed]
87. Tredicine, M.; Mucci, M.; Recchiuti, A.; Mattoscio, D. Immunoregulatory Mechanisms of the Arachidonic Acid Pathway in Cancer. *FEBS Lett.* **2025**, *599*, 927–951. [CrossRef] [PubMed]
88. Calder, P.C. Eicosanoids. *Essays Biochem.* **2020**, *64*, 423–441. [CrossRef] [PubMed]
89. Xia, Q.; Gao, W.; Yang, J.; Xing, Z.; Ji, Z. The Deregulation of Arachidonic Acid Metabolism in Ovarian Cancer. *Front. Oncol.* **2024**, *14*, 1381894. [CrossRef] [PubMed]
90. Jin, J.; Byun, J.-K.; Choi, Y.-K.; Park, K.-G. Targeting Glutamine Metabolism as a Therapeutic Strategy for Cancer. *Exp. Mol. Med.* **2023**, *55*, 706–715. [CrossRef] [PubMed]

91. Hellmann, A.; Turyn, J.; Zwara, A.; Korczynska, J.; Taciak, A.; Mika, A. Alterations in the Amino Acid Profile in Patients with Papillary Thyroid Carcinoma with and without Hashimoto's Thyroiditis. *Front. Endocrinol.* **2023**, *14*, 1199291. [CrossRef] [PubMed]
92. Aizawa, S.; Sakai, T.; Sakata, I. Glutamine and Glutamic Acid Enhance Thyroid-Stimulating Hormone β Subunit mRNA Expression in the Rat Pars Tuberalis. *J. Endocrinol.* **2012**, *212*, 383–394. [CrossRef] [PubMed]
93. Nałęcz, K.A. Amino Acid Transporter SLC6A14 (ATB0,+)—A Target in Combined Anti-Cancer Therapy. *Front. Cell Dev. Biol.* **2020**, *8*, 594464. [CrossRef] [PubMed]
94. Sniegowski, T.; Korac, K.; Bhutia, Y.D.; Ganapathy, V. SLC6A14 and SLC38A5 Drive the Glutaminolysis and Serine–Glycine–One-Carbon Pathways in Cancer. *Pharmaceuticals* **2021**, *14*, 216. [CrossRef] [PubMed]
95. Li, M.; Fu, X.; Zhou, T.; Han, H. Biomarkers Related to m6A and Succinic Acid Metabolism in Papillary Thyroid Carcinoma. *BMC Med. Genom.* **2024**, *17*, 199. [CrossRef] [PubMed]
96. Murugan, M.; Fedele, D.; Millner, D.; Alharfoush, E.; Vegunta, G.; Boison, D. Adenosine Kinase: An Epigenetic Modulator in Development and Disease. *Neurochem. Int.* **2021**, *147*, 105054. [CrossRef] [PubMed]
97. Luo, H.-Y.; Shen, H.-Y.; Perkins, R.S.; Wang, Y.-X. Adenosine Kinase on Deoxyribonucleic Acid Methylation: Adenosine Receptor-Independent Pathway in Cancer Therapy. *Front. Pharmacol.* **2022**, *13*, 908882. [CrossRef] [PubMed]
98. Iser, I.C.; Bertoni, A.P.S.; Beckenkamp, L.R.; Consolaro, M.E.L.; Maria-Engler, S.S.; Wink, M.R. Adenosinergic Signalling in Cervical Cancer Microenvironment. *Expert Rev. Mol. Med.* **2025**, *27*, e5. [CrossRef] [PubMed]
99. Chen, M.; Shen, M.; Li, Y.; Liu, C.; Zhou, K.; Hu, W.; Xu, B.; Xia, Y.; Tang, W. GC-MS-Based Metabolomic Analysis of Human Papillary Thyroid Carcinoma Tissue. *Int. J. Mol. Med.* **2015**, *36*, 1607–1614. [CrossRef] [PubMed]
100. Xu, Y.; Zheng, X.; Qiu, Y.; Jia, W.; Wang, J.; Yin, S. Distinct Metabolomic Profiles of Papillary Thyroid Carcinoma and Benign Thyroid Adenoma. *J. Proteome Res.* **2015**, *14*, 3315–3321. [CrossRef] [PubMed]
101. Shang, X.; Zhong, X.; Tian, X. Metabolomics of Papillary Thyroid Carcinoma Tissues: Potential Biomarkers for Diagnosis and Promising Targets for Therapy. *Tumor Biol.* **2016**, *37*, 11163–11175. [CrossRef] [PubMed]

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

5. LIMITAÇÕES

Este trabalho relata que a principal limitação para reproduzir os estudos foi a falta de disponibilidade de dados públicos em repositórios. Embora as revistas científicas peçam o depósito anonimamente, há escassez e falta de padronização entre plataformas, o que dispersa informações; mesmo com critérios de inclusão flexíveis, os autores encontraram dificuldade em obter dados publicados, e quando havia indicação de autor correspondente, não haviam respostas, forçando a equipe a realizar curadoria sistemática da literatura; apesar desses desafios, o estudo foi bem-sucedido ao integrar duas ômicas usando tecnologia aberta e dados públicos, produzindo resultados robustos e de impacto.

6. CONCLUSÃO

Foi possível utilizando a linguagem de programação R, desenvolver scripts para tratamento e modelagem de dados transcriptômicos e metabolômicos. Além disso, foi definida a melhor estratégia de integração ômica em função dos dados disponíveis e, apesar das limitações, houve a captura de dados públicos e associação dos mesmos na análise de redes.

A *pipeline* de transcriptômica mostrou que a abordagem utilizada foi robusta na sua intersecção de genes diferencialmente expressos com LASSO, evidenciando 11 genes, que foram selecionados, e bem sucedidos em prever o diagnóstico da doença. Além disso, foram avaliados possíveis interações deles com medicamentos, e mesmo com limitações, unificar dados transcriptômicos com metabolômicos numa rede de interação multi-ômica. A rede aponta para o Efeito Warburg como o principal fator associado à PTC. Adicionalmente, as moléculas inflamatórias presentes na rede direcionam o raciocínio bioquímico para cascatas de inflamação e regulação de células inflamatórias.

Este projeto usa bioinformática para organizar informações públicas disponíveis na literatura científica, oferecendo novas perspectivas sobre dados já existentes. Além disso, funciona como ferramenta complementar, amparando o corpo clínico com informações relevantes para sua prática.

As limitações inerentes ao projeto indicam a necessidade de um estudo de validação, com uma coorte brasileira, e com tecidos PTC e Controle, onde o mesmo tecido deverá ser analisado por transcriptômica e metabolômica, visando validar experimentalmente os achados deste estudo.

7. DECLARAÇÃO DE USO DE INTELIGÊNCIA ARTIFICIAL

Em conformidade com a Política de Integridadde na Atividade Científica do CNPq (Portaria nº 2.664/2026), declaro que ferramentas de Inteligência Artificial Generativas foram utilizadas neste trabalho nas seguintes condições:

Ao lado da plataforma PUBMED, a ferramenta *Perplexity* foi utilizada para encontrar artigos científicos, na etapa de contextualização na sessão de Introdução. Todo conteúdo gerado com o apoio dessas ferramentas foi submetido à revisão crítica pelo autor, que assume integral responsabilidade pela originalidade, veracidade e integridade do texto final.

REFERÊNCIAS

AHMED, Z.; WAN, S.; ZHANG, F.; ZHONG, W. Artificial intelligence for omics data analysis. **BMC Methods**, [s. l.], v. 1, n. 1, p. 4, 2024.

ALSALEH, M.; BARBERA, T. A.; ANDREWS, R. H.; SITHITHAWORN, P.; KHUNTIKEO, N.; LOILOME, W.; YONGVANIT, P.; COX, I. J.; SYMS, R. R. A.; HOLMES, E.; TAYLOR–ROBINSON, S. D. Mass Spectrometry: A Guide for the Clinician. **Journal of Clinical and Experimental Hepatology**, [s. l.], v. 9, n. 5, p. 597–606, 2019.

ALSEEKH, S.; AHARONI, A.; BROTMAN, Y.; CONTREPOIS, K.; D'AURIA, J.; EWALD, J.; C. EWALD, J.; FRASER, P. D.; GIAVALISCO, P.; HALL, R. D.; HEINEMANN, M.; LINK, H.; LUO, J.; NEUMANN, S.; NIELSEN, J.; PEREZ DE SOUZA, L.; SAITO, K.; SAUER, U.; SCHROEDER, F. C.; SCHUSTER, S.; SIUZDAK, G.; SKIRYCZ, A.; SUMNER, L. W.; SNYDER, M. P.; TANG, H.; TOHGE, T.; WANG, Y.; WEN, W.; WU, S.; XU, G.; ZAMBONI, N.; FERNIE, A. R. Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. **Nature Methods**, [s. l.], v. 18, n. 7, p. 747–756, 2021.

BAINS, L.; KATURI, U. K.; POLISETTY, C.; LAL, P.; SINGH, M. An uncommon site for papillary thyroid carcinoma metastasis: a case of breast involvement. **World Journal of Surgical Oncology**, [s. l.], v. 23, n. 1, p. 181, 2025.

BOGOVIĆ CRNČIĆ, T. Risk Factors for Thyroid Cancer: What Do We Know So Far? **Acta Clinica Croatica**, [s. l.], 2020. Disponível em: <https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=360857>. Acesso em: 12 jun. 2026.

CELIK, M.; BULBUL, B. Y.; AYTURK, S.; DURMUS, Y.; GURKAN, H.; CAN, N.; TASTEKIN, E.; USTUN, F.; SEZER, A.; GULDIKEN, S. The relation between BRAFV600E mutation and clinicopathological characteristics of papillary thyroid cancer. **Medicinski glasnik**, [s. l.], v. 17, n. 1, 2019. Disponível em: <https://medicinskiglasnik.ba/article/id/343/>. Acesso em: 12 jun. 2026.

CHEN, C.; WANG, J.; PAN, D.; WANG, X.; XU, Y.; YAN, J.; WANG, L.; YANG, X.; YANG, M.; LIU, G.-P. Applications of multi-omics analysis in human diseases. **MedComm**, [s. l.], v. 4, n. 4, p. e315, 2023. a.

CHEN, C.-J.; LEE, D.-Y.; YU, J.; LIN, Y.-N.; LIN, T.-M. Recent advances in LC-MS-based metabolomics for clinical biomarker discovery. **Mass Spectrometry Reviews**, [s. l.], v. 42, n. 6, p. 2349–2378, 2023. b.

CORREA-AGUILA, R.; ALONSO-PUPO, N.; HERNÁNDEZ-RODRÍGUEZ, E. W. Multi-omics data integration approaches for precision oncology. **Molecular Omics**, [s. l.], v. 18, n. 6, p. 469–479, 2022.

CZARNIECKA, A.; ZEMAN, M.; WOZNIAK, G.; MACIEJEWSKI, A.; STOBIECKA, E.; CHMIELIK, E.; OCZKO-WOJCIECHOWSKA, M.; KRAJEWSKA, J.; HANDKIEWICZ-JUNAK, D.; JARZAB, B. Therapeutic Strategy in Low-Risk Papillary Thyroid Carcinoma – Long-Term Results of the First Single-Center Prospective Non-Randomized Trial Between 2011 and 2015. **Frontiers in Endocrinology**, [s. l.], v. 12, p. 718833, 2021.

DE CARVALHO, L. M.; BORELLI, G.; CAMARGO, A. P.; DE ASSIS, M. A.; DE FERRAZ, S. M. F.; FIAMENGHI, M. B.; JOSÉ, J.; MOFATTO, L. S.; NAGAMATSU, S. T.; PERSINOTI, G.

F.; SILVA, N. V.; VASCONCELOS, A. A.; PEREIRA, G. A. G.; CARAZZOLLE, M. F. Bioinformatics applied to biotechnology: A review towards bioenergy research. **Biomass and Bioenergy**, [s. l.], v. 123, p. 195–224, 2019.

DU, M.; GU, D.; XIN, J.; PETERS, U.; SONG, M.; CAI, G.; LI, S.; BEN, S.; MENG, Y.; CHU, H.; CHEN, L.; WANG, Q.; ZHU, L.; FU, Z.; ZHANG, Z.; WANG, M. Integrated multi-omics approach to distinct molecular characterization and classification of early-onset colorectal cancer. **Cell Reports Medicine**, [s. l.], v. 4, n. 3, p. 100974, 2023.

GE, S.; SUN, S.; XU, H.; CHENG, Q.; REN, Z. Deep learning in single-cell and spatial transcriptomics data analysis: advances and challenges from a data science perspective. **Briefings in Bioinformatics**, [s. l.], v. 26, n. 2, p. bbaf136, 2025.

GONÇALVES, E.; POULOS, R. C.; CAI, Z.; BARTHORPE, S.; MANDA, S. S.; LUCAS, N.; BECK, A.; BUCIO-NOBLE, D.; DAUSMANN, M.; HALL, C.; HECKER, M.; KOH, J.; LIGHTFOOT, H.; MAHBOOB, S.; MALI, I.; MORRIS, J.; RICHARDSON, L.; SENEVIRATNE, A. J.; SHEPHERD, R.; SYKES, E.; THOMAS, F.; VALENTINI, S.; WILLIAMS, S. G.; WU, Y.; XAVIER, D.; MACKENZIE, K. L.; HAINS, P. G.; TULLY, B.; ROBINSON, P. J.; ZHONG, Q.; GARNETT, M. J.; REDDEL, R. R. Pan-cancer proteomic map of 949 human cell lines. **Cancer Cell**, [s. l.], v. 40, n. 8, p. 835- 849.e8, 2022.

GOU, D.; MIN, C.; PENG, X.; WU, H.; ZHANG, L.; CHEN, Y.; TAO, M. Associating factors of cognitive frailty among older people with chronic heart failure: Based on LASSO-logistic regression. **Journal of Advanced Nursing**, [s. l.], v. 81, n. 3, p. 1399–1411, 2025.

GUO, J.; HUAN, T. Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography–Mass Spectrometry Based Untargeted Metabolomics. **Analytical Chemistry**, [s. l.], v. 92, n. 12, p. 8072–8080, 2020.

HOLEWINSKI, R. J.; PARKER, S. J.; MATLOCK, A. D.; VENKATRAMAN, V.; VAN EYK, J. E. Methods for SWATH^{TM}: Data Independent Acquisition on TripleTOF Mass Spectrometers. Em: SECHI, S. (Ed.). **Quantitative Proteomics by Mass Spectrometry**. New York, NY: Springer, 2016. p. 265–279.

HSU, C.-Y.; ASKAR, S.; ALSHKARCHY, S. S.; NAYAK, P. P.; ATTABI, K. A. L.; KHAN, M. A.; MAYAN, J. A.; SHARMA, M. K.; ISLOMOV, S.; SOLEIMANI SAMARKHAZAN, H. AI-driven multi-omics integration in precision oncology: bridging the data deluge to clinical decisions. **Clinical and Experimental Medicine**, [s. l.], v. 26, n. 1, p. 29, 2025.

HU, J.; SZYMCZAK, S. A review on longitudinal data analysis with random forest. **Briefings in Bioinformatics**, [s. l.], v. 24, n. 2, p. bbad002, 2023.

JEON, H.; TKACIK, E.; ECK, M. J. Signaling from RAS to RAF: The Molecules and Their Mechanisms. **Annual Review of Biochemistry**, [s. l.], v. 93, n. Volume 93, 2024, p. 289–316, 2024.

JEONG, E.; YOON, S. Current advances in comprehensive omics data mining for oncology and cancer research. **Biochimica et Biophysica Acta (BBA) - Reviews on Cancer**, [s. l.], v. 1879, n. 1, p. 189030, 2024.

LEE, K.-S.; PARK, H. Machine learning on thyroid disease: a review. **Frontiers in Bioscience**, Landmark edition, v. 27, n. 3, p. 101, 2022.

LI, J.; SMITH, L. S.; ZHU, H.-J. Data-independent acquisition (DIA): An emerging proteomics technology for analysis of drug-metabolizing enzymes and transporters. **Drug Discovery Today: Technologies**, [s. l.], v. 39, p. 49–56, 2021.

LIESENFELD, D. B.; HABERMANN, N.; OWEN, R. W.; SCALBERT, A.; ULRICH, C. M. Review of Mass Spectrometry–Based Metabolomics in Cancer Research. **Cancer Epidemiology, Biomarkers & Prevention**, [s. l.], v. 22, n. 12, p. 2182–2201, 2013.

LIMAIEM, F.; REHMAN, A.; MAZZONI, T. Papillary Thyroid Carcinoma. Em: **StatPearls**. Treasure Island (FL): StatPearls Publishing, 2026.

LIN, J. S.; BOWLES, E. J. A.; WILLIAMS, S. B.; MORRISON, C. C. Introduction. Em: **Screening for Thyroid Cancer: A Systematic Evidence Review for the U.S. Preventive Services Task Force [Internet]**. [s.l.] : Agency for Healthcare Research and Quality (US), 2017.

LIU, P.; YU, X. Lymphatic metastasis of papillary thyroid carcinoma: mechanism and clinicopathological physiology. **Frontiers in Endocrinology**, [s. l.], v. 16, 2026. Disponível em: <https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2025.1725077/full>. Acesso em: 12 jun. 2026.

LOWE, R.; SHIRLEY, N.; BLEACKLEY, M.; DOLAN, S.; SHAFEE, T. Transcriptomics technologies. **PLoS Computational Biology**, [s. l.], v. 13, n. 5, p. e1005457, 2017.

MANZONI, C.; KIA, D. A.; VANDROVCOVA, J.; HARDY, J.; WOOD, N. W.; LEWIS, P. A.; FERRARI, R. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. **Briefings in Bioinformatics**, [s. l.], v. 19, n. 2, p. 286–302, 2018.

MORABITO, A.; DE SIMONE, G.; PASTORELLI, R.; BRUNELLI, L.; FERRARIO, M. Algorithms and tools for data-driven omics integration to achieve multilayer biological insights: a narrative review. **Journal of Translational Medicine**, [s. l.], v. 23, n. 1, p. 425, 2025.

PARK, K. S.; KIM, S. H.; OH, J. H.; KIM, S. Y. Highly accurate diagnosis of papillary thyroid carcinomas based on personalized pathways coupled with machine learning. **Briefings in Bioinformatics**, [s. l.], v. 22, n. 4, p. bbaa336, 2021.

PCDT - Carcinoma Diferenciado da Tireoide. [s.d.]. Disponível em: <https://www.gov.br/conitec/pt-br/midias/protocolos/pcdt_carcinomatireoide.pdf>. Acesso em: 26 jun. 2026.

PRETE, A.; BORGES DE SOUZA, P.; CENSI, S.; MUZZA, M.; NUCCI, N.; SPONZIELLO, M. Update on Fundamental Mechanisms of Thyroid Cancer. **Frontiers in Endocrinology**, [s. l.], v. 11, 2020. Disponível em: <https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2020.00102/full>. Acesso em: 12 jun. 2026.

REEL, P. S.; REEL, S.; PEARSON, E.; TRUCCO, E.; JEFFERSON, E. Using machine learning approaches for multi-omics data analysis: A review. **Biotechnology Advances**, [s. l.], v. 49, p. 107739, 2021.

RINGEL, M. D.; SOSA, J. A.; BALOCH, Z.; BISCHOFF, L.; BLOOM, G.; BRENT, G. A.; BROCK, P. L.; CHOU, R.; FLAVELL, R. R.; GOLDNER, W.; GRUBBS, E. G.; HAYMART, M.; LARSON, S. M.; LEUNG, A. M.; OSBORNE, J. R.; RIDGE, J. A.; ROBINSON, B.; STEWARD, D. L.; TUFANO, R. P.; WIRTH, L. J. 2025 American Thyroid Association

Management Guidelines for Adult Patients with Differentiated Thyroid Cancer. **Thyroid®**, [s. l.], v. 35, n. 8, p. 841–985, 2025.

ROBERTS, L. D.; SOUZA, A. L.; GERSZTEN, R. E.; CLISH, C. B. Targeted Metabolomics. **Current Protocols in Molecular Biology**, [s. l.], v. 98, n. 1, p. 30.2.1-30.2.24, 2012.

SAMMUT, S.-J.; CRISPIN-ORTUZAR, M.; CHIN, S.-F.; PROVENZANO, E.; BARDWELL, H. A.; MA, W.; COPE, W.; DARIUSH, A.; DAWSON, S.-J.; ABRAHAM, J. E.; DUNN, J.; HILLER, L.; THOMAS, J.; CAMERON, D. A.; BARTLETT, J. M. S.; HAYWARD, L.; PHAROAH, P. D.; MARKOWETZ, F.; RUEDA, O. M.; EARL, H. M.; CALDAS, C. Multi-omic machine learning predictor of breast cancer therapy response. **Nature**, [s. l.], v. 601, n. 7894, p. 623–629, 2022.

SANCHES, P. H. G.; DE MELO, N. C.; PORCARI, A. M.; DE CARVALHO, L. M. Integrating Molecular Perspectives: Strategies for Comprehensive Multi-Omics Integrative Data Analysis and Machine Learning Applications in Transcriptomics, Proteomics, and Metabolomics. **Biology**, [s. l.], v. 13, n. 11, p. 848, 2024.

SANCHES, P. H. G.; DE OLIVEIRA, D. C.; DOS REIS, I.; FERNANDES, A.; SILVA, A. A. R.; EBERLIN, M.; CARVALHO, P. O.; DUARTE, G.; PORCARI, A.. Fitting Structure-Data Files (.SDF) Libraries to Progenesis QI Identification Searches. **Journal of the Brazilian Chemical Society**, [N.p.], 2023.

SCHRIMPE-RUTLEDGE, A. C.; CODREANU, S. G.; SHERROD, S. D.; MCLEAN, J. A. Untargeted Metabolomics Strategies—Challenges and Emerging Directions. **Journal of the American Society for Mass Spectrometry**, [s. l.], v. 27, n. 12, p. 1897–1905, 2016.

SHANK, J. B.; ARE, C.; WENOS, C. D. Thyroid Cancer: Global Burden and Trends. **Indian Journal of Surgical Oncology**, [s. l.], v. 13, n. 1, p. 40–45, 2022.

SOUZA, A. L.; PATTI, G. J. A Protocol for Untargeted Metabolomic Analysis: From Sample Preparation to Data Processing. Em: WEISSIG, V.; EDEAS, M. (Eds.). **Mitochondrial Medicine : Volume 2: Assessing Mitochondria**. New York, NY: Springer US, 2021. p. 357–382.

SURI, G. S.; KAUR, G.; CARBONE, G. M.; SHINDE, D. Metabolomics in oncology. **Cancer Reports**, [s. l.], v. 6, n. 3, p. e1795, 2023.

TORAIH, E. A.; HUSSEIN, M. H.; ZERFAOUI, M.; ATTIA, A. S.; MARZOUK ELLYTHY, A.; MOSTAFA, A.; RUIZ, E. M. L.; SHAMA, M. A.; RUSSELL, J. O.; RANDOLPH, G. W.; KANDIL, E. Site-Specific Metastasis and Survival in Papillary Thyroid Cancer: The Importance of Brain and Multi-Organ Disease. **Cancers**, [s. l.], v. 13, n. 7, p. 1625, 2021.

ULUDAG, M.; UNLU, M. T.; CETINOGLU, I.; CALISKAN, O.; AYGUN, N. What Has Changed in the 2025 American Thyroid Association Management Guidelines for Adult Patients with Differentiated Thyroid Cancer? Part 2: Postoperative Initial Treatment. **The Medical Bulletin of Sisli Etfal Hospital**, [s. l.], v. 59, n. 3, p. 273–283, 2025.

URBAN, P. L. Quantitative mass spectrometry: an overview. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, [s. l.], v. 374, n. 2079, p. 20150382, 2016.

VADAKEDATH, S.; KANDI, V.; GODISHALA, V.; PINNELLI, V. B. K.; ALKAFAAS, S. S.; EIKAFAS, S. S. The Principle, Types, and Applications of Mass Spectrometry: A Comprehensive Review. **Biomedicine and Biotechnology**, [s. l.], v. 7, n. 1, p. 6–22, 2022.

VAN DER LAAN, T.; BOOM, I.; MALIEPAARD, J.; DUBBELMAN, A.-C.; HARMS, A. C.; HANKEMEIER, T. Data-Independent Acquisition for the Quantification and Identification of Metabolites in Plasma. **Metabolites**, [s. l.], v. 10, n. 12, p. 514, 2020.

VOLPI, E. M.; RAMIREZ-ORTEGA, M. C.; CARRILLO, J. F. Editorial: Recent advances in papillary thyroid carcinoma: diagnosis and predictive factors. **Frontiers in Endocrinology**, [s. l.], v. 14, 2023. Disponível em: <https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2023.1283397/full>. Acesso em: 12 jun. 2026.

WEI, L.; NIRAULA, D.; GATES, E. D. H.; FU, J.; LUO, Y.; NYFLOT, M. J.; BOWEN, S. R.; EL NAQA, I. M.; CUI, S. Artificial intelligence (AI) and machine learning (ML) in precision oncology: a review on enhancing discoverability through multiomics integration. **British Journal of Radiology**, [s. l.], v. 96, n. 1150, p. 20230211, 2023.

WEKESA, J. S.; KIMWELE, M. A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment. **Frontiers in Genetics**, [s. l.], v. 14, 2023. Disponível em: <https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2023.1199087/full>. Acesso em: 12 jun. 2026.

WILTSHIRE, J. J.; DRAKE, T. M.; UTTLEY, L.; BALASUBRAMANIAN, S. P. Systematic Review of Trends in the Incidence Rates of Thyroid Cancer. **Thyroid**, [s. l.], v. 26, n. 11, p. 1541–1552, 2016.

YAN, Y.; SUN, D.; HU, J.; CHEN, Y.; SUN, L.; YU, H.; XIONG, Y.; HUANG, Z.; XIA, H.; ZHU, X.; BIAN, D.; SUN, F.; HOU, L.; WU, C.; FAN, O. R.; HU, H.; ZENG, A.; ZHANG, L.; SUN, Y. E.; WANG, C.; ZHANG, P. Multi-omic profiling highlights factors associated with resistance to immuno-chemotherapy in non-small-cell lung cancer. **Nature Genetics**, [s. l.], v. 57, n. 1, p. 126–139, 2025.

ZHOU, B.; XIAO, J. F.; TULI, L.; RESSOM, H. W. LC-MS-based metabolomics. **Molecular BioSystems(MBS)**, [s. l.], v. 8, n. 2, p. 470–481, 2012.

ZHOU, K.; YANG, C.; LI, Y. Multi-omics in colorectal cancer liver metastasis: applications and research advances. **Cancer Biology & Medicine**, [s. l.], 2025. Disponível em: <https://www.cancerbiomed.org/content/early/2025/06/26/j.issn.2095-3941.2025.0066>. Acesso em: 12 jun. 2026.