IMPERIAL

Machine Learning-Driven Hit Discovery, Validation, and Phytochemical Recipe Enrichment

Luís Artur Domingues Rita

Supervisor: Dr. Kirill Veselkov

Co-Supervisor: Prof. Michael Bronstein

A thesis submitted for the degree of Doctor of Philosophy

of Imperial College London

Division of Cancer

Department of Surgery and Cancer

Faculty of Medicine

Imperial College London, United Kingdom

2024

REPÚBLICA PORTUGUESA | XXIV GOVERNO

COMPETE 2020 PORTUGAL 2020 UNIÃO EUROPEIA Fundo Social Europeu

fct Fundação para a Ciência e a Tecnologia

Statement of Originality

I certify that this thesis, and the research to which it refers, are the products of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline. Contributions by others include Dr Guadalupe Gonzalez's design of the graph deep learning model used to predict drug-drug and drug-food synergies. All other work presented in this thesis is my original contribution to the field.

Copyright Statement

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC). Under this license, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the license terms clear to others by naming the license and linking to the license text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this license or permitted under UK Copyright Law.

General Abstract

Bioactive phytochemicals show promise for disease prevention and treatment, but their interactions with drugs and disease pathways remain poorly understood. We address this problem by employing machine learning to identify therapeutic phytochemicals, predict their putative pharmacodynamic (PD) interactions with drugs, and discover phytochemically enriched recipes by using existing drugs as a reference. First, we developed a network-based machine learning approach to identify bioactive phytochemicals in extra virgin olive oil (EVOO) that could influence the protein network associated with Alzheimer's disease (AD). The model achieved a balanced classification accuracy of 70.3±2.6%, effectively distinguishing late-stage experimental AD drugs from other clinically approved drugs. The EVOO analysis highlighted phytochemicals such as quercetin, genistein, and luteolin with high predicted activity against AD, providing novel insights into potential mechanisms for prevention or treatment. Second, we utilised a novel network machine learning approach based on the DreamLab DRUGS network propagation method to predict synergies in PD drug-drug and phytochemical-drug interactions. The model was calibrated using an *in vitro* PD drug-drug interactions dataset, achieving a Pearson correlation score of 68.9%. Experimental validation was conducted in cancer cell lines. The network machine learning model demonstrated consistency with predicted outcomes in 7/11 tested drug-drug combinations and 5/9 food-drug combinations, indicating its effectiveness in identifying clinically relevant PD interactions. Third, we applied large language models (LLMs) to optimize ingredient substitutions in recipes, aiming to enhance phytochemical content. The application of LLMs in ingredient substitution improved Hit@1 accuracy from a baseline of 34.5±0.1% to 38.0±0.3% on the original GISMo dataset, and from 40.2±0.4% to 54.5±0.3% on a refined dataset. This work demonstrates that machine learning can significantly advance the identification of bioactive phytochemicals, predict their interactions with drugs and disease processes, and optimize dietary choices. These findings establish a computational foundation for integrating nutrition and pharmacology, though clinical validation remains essential.

Table of Contents

Statement of Originality 2

Copyright Statement 2

General Abstract 3

Table of Contents 4

List of Figures 8

List of Tables 9

Acronyms 10

1. General Introduction 12

1.1 Network Propagation in Biological Networks 12

1.1.1 From Simple Metrics to Diffusion-Based Algorithms: The Evolution of Network Propagation Methods 13

1.1.2 Medical Applications of Network Propagation: From Gene Prioritization to Precision Medicine 14

1.2 Graph Neural Networks in Biological Networks 17

1.2.1 Evolution of Graph Neural Network Architectures for Biological Data Analysis 17

1.2.2 Medical Applications of Graph Neural Networks 20

1.3 Language Models in Medical Research 21

1.3.1 From N-Grams to Transformers: The Evolution of Language Models 21

1.3.2 Medical Applications of Language Models: From Literature Mining to Drug Safety 24

1.4 Molecular Representation of Disease: Alzheimer's Disease, Cancer, and COVID-19 25

1.4.1 Gene/Protein Targets and Pathways in Alzheimer's Disease 26

1.4.2 Gene/Protein Targets and Pathways in Cancer 27

1.4.3 Gene/Protein Targets and Pathways in COVID-19 28

1.5 The Intersection of Diet and Disease: Implications for Risk, Prevention, and Treatment 29

1.6 Gut Microbiota as a Mediator of Phytochemical–Disease Interactions 31

2. General Methodology 33

2.1 Basic Concepts and Definitions 33
2.2 Biological Networks 33
2.3 Random Walks for Phytochemical Discovery 34
2.4 Representation Learning on Graphs and Validation of Compound Synergies 38
2.4.1 Computational Predictions of Compound Synergies 38
2.4.2 Compounds Response Curve and Synergies 40
2.5 Large Language Models for Recipe Personalization 44
2.6 Datasets Used in the Dissertation 46
3. Network Machine Learning-Driven Discovery of Phytochemicals and Drugs 52
3.1 Abstract 52
3.2 Introduction 52
3.3 Methodology 55
3.3.1 Alzheimer's Disease Protein-Encoding Gene Selection 58
3.3.2 Extra Virgin Olive Oil Phytochemical Selection 59
3.3.3 Network Propagation 60
3.3.4 Model Training and Application 61
3.3.5 Pathway Analysis 62
3.4 Results 62
3.5 Discussion 68
3.6 Conclusion 73
4. Experimental Validation of Pharmacodynamic Drug-Drug and Food-Drug Synergies 74
4.1 Abstract 74
4.2 Introduction 75
4.2.1 Quantifying Individual Drug/Food Responses 78
4.2.2 Quantifying Synergies in Drug-Drug and Food-Drug Interactions 78
4.3 Methodology 80
4.3.1 HCT116 and A549 Cell Passaging 81
4.3.2 HCT116 and A549 Cell Seeding 82

4.3.3 Drug/Food Treatments 82
4.3.4 Cell Viability Assay and Absorption Readings 85
4.3.5 Drug/Food IC50 Estimation 86
4.3.6 Drug-Drug and Food-Drug Treatments 87
4.3.7 Synergy Estimation 88
4.4 Results 89
4.5 Discussion 98
4.6 Conclusion 101
5. Ingredient Substitution using Large Language Models 103
5.1 Abstract 103
5.2 Introduction 103
5.3 Methodology 106
5.3.1 Recipe and Ingredient Substitution Dataset 106
5.3.2 GISMo Benchmark 107
5.3.3 Incorporation in GISMo of Food Category Feature 108
5.3.4 Dataset Filtration Based on Substitution Validity 108
5.3.5 Fine-tuning Language Models for Substitution Predictions 109
5.3.6 Evaluation of Ingredient Substitution Accuracy 110
5.3.7 Phytochemically Enriched Recipe Generation 113
5.4 Results 114
5.5 Discussion 116
5.6 Conclusion 120
6. General Discussion 121
7. General Conclusion 124
8. Appendix to Chapter 3 126
8.1 List of Dysregulated Genes in Alzheimer's Disease 126
8.2 Comprehensive List of Phytochemicals in Extra Virgin Olive Products 129

8.3 Predicted Extra Virgin Olive Oil Marker Phytochemicals Influencing Alzheimer's Disease Onset and Progression 129
8.4 Pathways Targeted by Phytochemicals in Extra Virgin Olive Oil Associated with Alzheimer's Disease 129
8.4.1 Alzheimer's Disease Pathway 129
8.4.2 Olfactory Transduction Pathway 131
8.4.3 Insulin Signalling Pathway 132
8.4.4 Phosphatidylinositol Signalling System Pathway 134
8.4.5 Vascular Smooth Muscle Contraction Pathway 136
8.5 Human Genomics (Published) 138
9. Appendix to Chapter 4 140
9.1 Food-Drug Computational Synergies in A549 140
9.2 Food-Drug Computational Synergies in HCT116 141
10. Appendix to Chapter 5 142
10.1 Classification of Ingredients in the Recipe1MSubs Dataset 142
10.2 MDPI MAKE (Published) 143
References 145
Funding 171
Acknowledgements 171

List of Figures

Figure 1 Overview of Phytochemical Discovery Chapter 3.

Figure 2 Overview of Drug-Food Synergy Chapter 4.

Figure 3 Overview of Recipe Enrichment Chapter 5.

Figure 4 Network Propagation of Protein Perturbation Profiles for Phytochemicals, Drugs and AD.

Figure 5 Machine Learning–Driven Prediction of Drugs and EVOO Phytochemicals with Anti-AD Profile Similarity.

Figure 6 Predicted EVOO Phytochemicals Targeting Dysregulated Pathways in AD.

Figure 7 Drug-Drug and Drug-Food Computational Predictions.

Figure 8 Drug and Food Monotherapy Treatments.

Figure 9 Drug-Drug and Drug-Food Combinatorial Treatments.

Figure 10 Computational Predictions of Drug-Food Combinations for A549 and HCT116 Cell Lines.

Figure 11 Dose-Response Curves for the Monotherapies of Thymoquinone, Selenomethionine, Tartaric Acid, Galangin, and Linoelaidic Acid in A549 Cancer Cell Line Assays.

Figure 12 Dose-response Curves for Sunitinib, Gemcitabine, Dactolisib, MK-8776, Parthenolide, Carboplatin, Doxorubicin, and Mitomycin in A549 Cancer Cell Line Assays.

Figure 13 Predicted and Experimentally Measured Synergy Values for Drug-Drug and Drug-Food Combinations in A549 and HCT116 Cell Lines.

Figure 14 Experimental Validation of Drug-Drug Combination in A549 Cell Line.

Figure 15 Experimental Validation of Drug-Food Combination in A549 Cell Line.

Figure 16 Ingredient Substitution Overview.

Figure 17 Ingredient Substitution Model Fine-Tuning and Evaluation.

Figure 18 Phytochemically Enriched Recipe Generation.

Figure 19 Phytochemically Enriched Recipes.

List of Tables

Table 1 Comparison of Traditional Language Models and LLMs.

Table 2 Summary of the Datasets Used in the Dissertation.

Table 3 Top FDA-Approved Drugs Not Undergoing AD Clinical Trials Ranked by Our Model.

Table 4 Top Ten EVOO Phytochemicals Predicted to Influence AD Onset or Progression.

Table 5 Predicted Drug-Drug Combinations Tested in A549 and HCT116 Cell Lines.

Table 6 Predicted Drug-Food Combinations Tested in A549 Cell Line.

Table 7 Monotherapies for Phytochemicals and Drugs in A549 Cell Line.

Table 8 Computational and Experimental Drug-Drug Synergies in A549 and HCT116 Cell Lines.

Table 9 Computational and Experimental Food-Drug Synergies in A549 and HCT116 Cell Lines.

Table 10 Training Configurations for DaVinci-002, GPT-3.5-Turbo-1106, and TinyLlama-1.1B.

Table 11 Performance Metrics of Five Filtering Runs for the GISMo Model.

Table 12 Performance of Fine-Tuned Models.

Table 13 Dysregulated genes in AD.

Table 14 Key Proteins/Genes and Complexes Involved in the AD Pathway.

Table 15 Key Proteins/Genes and Complexes Involved in the Olfactory Transduction Pathway.

Table 16 Key Proteins/Genes and Complexes Involved in the Insulin Signalling Pathway.

Table 17 Key Proteins/Genes and Complexes Involved in the Phosphatidylinositol Signalling System Pathway.

Table 18 Key Proteins/Genes and Complexes Involved in the Vascular Smooth Muscle Contraction Pathway.

Table 19 Top 20 Predicted Drug-Food Synergies in A549.

Table 20 Top 20 Predicted Drug-Food Synergies in HCT116.

Acronyms

A549: A type of human lung carcinoma cell line

ACE2: Angiotensin-Converting Enzyme 2

AD: Alzheimer's Disease

AI: Artificial Intelligence

ARDS: Acute Respiratory Distress Syndrome

BERT: Bidirectional Encoder Representations from Transformers

ClinGen: Clinical Genome Resource

CLM: Causal Language Modelling

COVID-19: Coronavirus Disease 2019

CTD: Comparative Toxicogenomics Database

EVOO: Extra Virgin Olive Oil

FDA: Food and Drug Administration

GAT: Graph Attention Network

GCN: Graph Convolutional Network

GDSC: Genomics of Drug Sensitivity in Cancer

GIN: Graph Isomorphism Network

GISMo: Graph-based Ingredient Substitution Model

GNN: Graph Neural Network

GSEA: Gene Set Enrichment Analysis

HetGNN: Heterogeneous GNN

HCT116: A type of human colorectal carcinoma cell line

HMM: Hidden Markov Model

HPRD: Human Protein Reference Database

IC40: 40% maximal inhibitory concentration

IC50: Half maximal inhibitory concentration

KEGG: Kyoto Encyclopedia of Genes and Genomes

LLM: Large Language Model

LSTM: Long Short-Term Memory

MAOIs: Monoamine Oxidase Inhibitors

ML: Machine Learning

MLM: Masked Language Modelling

MRF: Markov Random Fields

MRR: Mean Reciprocal Rank

NLP: Natural Language Processing

Panther: Protein ANalysis THrough Evolutionary Relationships

PD: Pharmacodynamic

PK: Pharmacokinetic

PPI: Protein-Protein Interaction

PsyGeNET: Psychiatric Disorders Gene Association Network

RecNN: Recursive Neural Network

RNN: Recurrent Neural Network

RW: Random Walk

RWR: Random Walk with Restarts

STRING: Search Tool for the Retrieval of Interacting Genes/Proteins

STITCH: Search Tool for Interactions of Chemicals

TGN: Temporal Graph Networks

TF-IDF: Term Frequency–Inverse Document Frequency

UniProt: Universal Protein Resource

VSMC: Vascular Smooth Muscle Contraction

ZIP: Zero Interaction Potency

1. General Introduction

Machine learning (ML), a branch of artificial intelligence (AI) that enables systems to learn from data, recognize complex patterns, and continually improve with minimal human intervention, forms the backbone of this thesis's integrative approach to systems biology. ML paradigms—including supervised learning (training on labelled examples), unsupervised learning (discovering intrinsic data structure), semi-supervised learning (combining labelled and unlabelled data), reinforcement learning (learning via rewards and penalties), and self-supervised learning (deriving training signals from the data itself)—are all well established in AI [1], yet this work relies exclusively on supervised learning. Ranging from supervised network propagation models, to supervised graph neural networks and supervised fine-tuning of language models.

1.1 Network Propagation in Biological Networks

Network propagation methods have emerged as powerful tools in systems biology and computational medicine, offering systematic approaches to uncover the complex interactions within biological networks [2–4]. These methods facilitate a deeper understanding of how biological entities such as genes, proteins, and metabolites interact and influence each other, which is crucial for identifying the mechanisms underlying various diseases and biological processes [5,6]. In biological systems, interactions are rarely linear or isolated; instead, they form intricate networks where the function of one component can significantly impact others. Network propagation leverages this interconnectedness by simulating the spread of information, influence, or perturbations through the network [7,8]. By doing so, it captures both local and global topological properties, enabling the identification of key nodes and subnetworks that play pivotal roles in biological functions and disease states [9]. A node (or vertex) is a fundamental unit in a network, representing an entity such as a protein, gene, or metabolite. A network is a collection of nodes linked by edges, which signify relationships or interactions - e.g., physical binding between proteins.

1.1.1 From Simple Metrics to Diffusion-Based Algorithms: The Evolution of Network Propagation Methods

The evolution of network propagation methods began with early network analysis techniques that focused on simple metrics such as shortest paths and direct neighbourhood connectivity [10]. These methods considered immediate interactions between nodes but were limited in their ability to capture the influence of more distant nodes within the network. Shortest path algorithms calculated the minimum number of steps between nodes, identifying the most direct connections but failing to account for the overall network topology. Neighbourhood analysis focused on immediate neighbours, providing insights into local interactions but not considering the broader network context [11].

To overcome the limitations of early methods, diffusion-based algorithms were introduced, marking a significant advancement in network analysis [12]. These methods model the spread of information or influence through the network in a way analogous to physical diffusion processes [2]. One diffusion-based method is the Heat Kernel Diffusion, which models the transfer of *heat* (information) across the network over time, using the heat kernel as a fundamental solution to the heat equation on graphs [13]. This method considers the entire network structure and accounts for both local and global topology, controlled by a diffusion time parameter that influences the spread [14]. It has been applied in identifying functionally related genes and analysing disease modules in biological networks [15,16].

In the early 2000s, probabilistic models like Markov Random Fields (MRF) began to be applied to biological networks [17]. MRFs provide a framework to model the dependencies between nodes using local conditional probabilities [18]. They are effective in capturing complex dependencies and interactions within biological data [19]. Applications include genomic sequence analysis, where sequential data is prevalent [20].

In the mid-2010s network embedding techniques were introduced aiming to map nodes to a low-dimensional space while preserving the network's structural properties [21]. Methods such as

DeepWalk and Node2Vec use random walks (RW) to generate sequences of nodes and apply word embedding techniques from natural language processing to learn vector representations of nodes [22]. These embeddings facilitate the use of advanced machine learning algorithms on network data and have been used in predicting protein functions and clustering nodes based on similarities [23]. This represented a shift towards integrating network analysis with machine learning and data mining techniques [24].

The most recent evolution in network propagation methods involves integrating them with deep learning, particularly with the advent of Graph Neural Networks (GNNs) [25]. GNNs combine network topology with node features and learn representations through message passing between nodes [26]. This integration allows for the modelling of complex, nonlinear relationships within the network data and has applications in predicting molecular interactions and classifying diseases based on network patterns [27]. The use of deep learning architectures in network analysis enabled the handling of large-scale, high-dimensional biological data and uncovering patterns that were previously undetectable [28,29].

1.1.2 Medical Applications of Network Propagation: From Gene Prioritization to Precision Medicine

Network propagation methods have found extensive applications in medicine due to their ability to integrate and analyse complex biological data [7,30]. They enable us to prioritize genes and proteins for further study, predict protein functions, identify novel drug targets, integrate diverse omics data, discover biomarkers, understand disease pathways, and contribute to the development of precision medicine [31,32].

In disease gene prioritization, identifying genes associated with diseases is fundamental for understanding disease mechanisms and developing therapeutic interventions [33]. Network propagation helps in this process by ranking candidate genes based on their connectivity and influence within biological networks [2]. Methods like Random Walk with Restarts (RWR) rank genes based on

proximity to known disease genes, Heat Kernel Diffusion identifies genes that influence or are influenced by disease modules, and GeneRank integrates gene expression with network topology for prioritization [34,35]. For example, in neurodegenerative diseases, these methods help identify genes that may contribute to disease susceptibility or progression, guiding further genetic studies and potential therapeutic targets [36].

Protein function prediction is important for uncovering biological processes [37]. Network propagation leverages protein interaction networks to predict functions of uncharacterized proteins [34]. Methods such as label propagation assign functions based on network neighbours, and network embedding learns representations to predict functions [31]. This accelerates the annotation of proteins in recently sequenced genomes and facilitates the discovery of novel protein functions, contributing to a better understanding of the proteome and helping in the identification of new drug targets [38].

In drug target identification and repositioning, network propagation predicts potential drug targets by analysing the network's structure and dynamics [39]. Drug repositioning (or repurposing) refers to finding new therapeutic uses for existing drugs by mapping their known targets onto disease networks and drug target identification is the process of discovering which proteins or genes, when modulated by a compound, can alter disease progression; in network terms, targets are often nodes central to a disease module [40]. Heterogeneous network propagation integrates drug-target interactions with disease networks, and network flow models identify critical nodes for intervention. Drug repositioning benefits from analysing connectivity between drugs and diseases, allowing existing drugs to be repurposed for new therapeutic uses. This approach not only helps in identifying novel drug targets but also in predicting off-target effects, where drugs interact with unintended proteins, helping anticipate side effects and optimizing therapeutic strategies [41]. Drug repositioning not only finds new indications but also highlights unexpected protein interactions: by projecting an existing drug’s known targets onto a disease-specific network, one can identify adjacent proteins (via network propagation) that the drug may affect off-target [42]. These adjacent nodes become candidate targets for further study.

Integrative omics analysis with network propagation methods provides a framework for integrating diverse data types generated by high-throughput technologies [43]. Dynamic network propagation incorporates temporal omics data, and multi-scale propagation analyses data at different biological scales [7]. This offers a complete view of biological processes, enhancing confidence in findings across omics layers. For example, combining gene expression data with protein-protein interaction networks can reveal how changes in gene expression influence protein interactions and contribute to disease progression [8].

Biomarker discovery is critical for disease detection, prognosis, and monitoring treatment efficacy. Network propagation identifies molecules central in disease networks [2]. Methods like diffusion kernels measure the influence of potential biomarkers, and graph neural networks learn complex patterns associated with biomarkers [44]. This improves disease diagnosis and prognosis by identifying biologically relevant biomarkers that reflect the underlying disease mechanisms, thereby enhancing the potential for early detection and personalized treatment strategies [45].

Understanding **pathways involved in diseases** is essential for identifying intervention points. Network propagation methods like Heat Kernel Diffusion analyse pathway dynamics, and Markov Random Fields model dependencies within pathways [46]. This enabled the identification on how perturbations in these pathways lead to diseases and pinpointing key nodes within pathways for therapeutic targeting [47]. This approach helps in understanding how diseases develop and progress, and in identifying potential points of therapeutic intervention to halt or reverse disease progression [48].

Precision medicine aims to tailor treatments based on individual characteristics, including genetic makeup and molecular profiles. Personalized network propagation uses patient-specific data for target identification, enhancing the potential for more effective and less toxic treatments [49]. For example, a patient's tumor RNA-seq profile can weight the initial network nodes—upregulating highly expressed oncogenes—so that the propagation ranks drugs whose targets most effectively perturb that individual's dysregulated network. This tailoring can improve efficacy and reduce toxicity compared to

a one-size-fits-all approach [50]. By integrating individual omics data with biological networks, these methods facilitate the identification of personalized therapeutic targets and treatment strategies, aligning with the goals of precision medicine to provide customized healthcare solutions [51].

1.2 Graph Neural Networks in Biological Networks

Unlike traditional neural networks that operate on fixed-size input data, GNNs can handle variable-sized graph inputs, making them suitable for modelling complex biological networks such as protein-protein interaction networks, gene regulatory networks, and metabolic pathways [52,53]. GNNs leverage the structure of graphs to learn representations of nodes, edges, or entire graphs, capturing the relationships and patterns in biological data [27,54].

1.2.1 Evolution of Graph Neural Network Architectures for Biological Data Analysis

The initial attempts to apply neural networks to graphs introduced **recursive neural networks** (RecNNs) and **graph neural networks** (GNNs) that operated on directed acyclic graphs [55,56]. In 2005, it was formalized the concept of GNNs by defining a framework where node representations are learned through an iterative process until a stable equilibrium is reached [25,57]. Concretely, each node has a vector that gets updated by combining the vectors of its neighbours using learnable functions (like summing or attention). These updates are repeated until they settle into a stable state. The update rules are trained using backpropagation on a task like node classification, so that the final vectors reflect both the graph's structure and the nodes' features [58]. These early GNNs employed a contraction mapping to ensure convergence, allowing them to handle graphs of arbitrary size and shape. However, they faced challenges with scalability and computational efficiency, limiting their applicability to large-scale networks commonly found in biological data [26].

A significant advancement came with the introduction of spectral methods for graph convolutions. In 2013, it was proposed **spectral graph convolutional networks** that utilized the eigen decomposition of the graph Laplacian to perform convolution operations in the spectral domain [28].

This approach enabled the extension of convolutional neural networks (CNNs) to graph-structured data by leveraging the mathematical properties of graphs [59]. Despite this breakthrough, spectral methods were computationally intensive and not scalable to large graphs due to the need of computing the eigenvectors of the Laplacian [60].

To overcome the limitations of spectral methods, researchers developed spatial-based GNNs that defined convolution operations directly on the graph structure without relying on spectral decomposition. In 2016, Graph Convolutional Networks (GCNs) simplified the convolution operation using a first-order approximation of spectral convolutions. GCNs efficiently aggregate information from a node's immediate neighbours, significantly improving scalability and enabling the processing of large graphs with millions of nodes. This model demonstrated higher performance in semi-supervised node classification tasks and became a foundation for subsequent GNN architectures [25].

The message-passing framework generalized the way GNNs aggregate information, where nodes update their representations by receiving *messages* from their neighbours. In 2017, this paradigm was formalized, unifying various GNN models under a common framework. This approach enhanced the flexibility of GNNs in modelling complex interactions within graphs [26].

Graph Attention Networks (GATs) incorporated attention mechanisms into GNNs. By assigning learned weights to neighbouring nodes during the aggregation process, GATs allowed the model to focus on the most relevant neighbours, improving performance on tasks where the importance of neighbouring nodes varies. This attention mechanism enabled GNNs to capture more patterns in the data, which is particularly beneficial in biological networks where certain interactions may be more important than others [28].

GraphSAGE, an inductive framework that generates node embeddings by sampling and aggregating features from a node's local neighbourhood. Unlike transductive models, which require retraining when new nodes are added, GraphSAGE can generalize to unseen nodes, making it suitable for dynamic graphs where nodes and edges may change over time. This property is essential in

biological networks that evolve, such as protein-protein interaction networks that expand with new discoveries [54].

While early GNNs were relatively shallow due to issues like over-smoothing, recent developments have focused on building deeper architectures. In 2018, it was introduced the **Graph Isomorphism Network** (GIN), which demonstrated that incorporating multi-layer perceptrons in the aggregation functions could increase the representational power of GNNs to match the Weisfeiler-Lehman graph isomorphism test [60]. A perceptron is a basic neural computing unit that computes a weighted sum of its inputs plus a bias and applies a non-linear activation function; stacking several perceptrons into multiple layers (an MLP) enables the model to learn more complex, non-linear mappings [61]. This allowed GNNs to distinguish between complex graph structures more effectively.

Hierarchical GNNs, such as DiffPool introduced pooling mechanisms to create hierarchical representations of graphs. By learning cluster assignments during training, these models can capture graph representations at multiple scales, analogous to pooling operations in CNNs. Hierarchical representations are particularly useful in bioinformatics for modelling biological structures at different organizational levels, from molecular interactions to cellular networks [59].

Recent research has expanded GNNs to handle temporal and heterogeneous graphs. **Temporal Graph Networks** (TGNs) model dynamic graphs where nodes and edges change over time, capturing the evolution of biological processes [62]. **Heterogeneous GNNs** (HetGNNs) manage graphs with multiple types of nodes and relations, reflecting the diversity in biological systems that include genes, proteins, metabolites, and their various interactions [63].

The integration of transformer architectures into GNNs has further advanced the field. By leveraging self-attention mechanisms, Graph Transformers can capture global dependencies in graphs more effectively. This approach is particularly beneficial for large and complex biological networks where long-range interactions play a critical role [64].

Despite their power, GNNs also face challenges such as over-smoothing (node features converge to similar values after many layers), over-squashing (distant information compressed into

limited capacity), and scalability issues on large graphs. Benchmarking and generalization can be problematic when models trained on one network topology fail to transfer to another without careful cross-validation [65].

1.2.2 Medical Applications of Graph Neural Networks

GNNs have been successfully applied to various bioinformatics tasks. In node classification, GNNs predict the function of genes or proteins based on their connectivity patterns in biological networks. By analysing the local neighbourhood of a gene or protein, GNNs can infer its role in cellular processes, helping in the annotation of uncharacterized genes and proteins [66].

For link prediction, GNNs identify potential interactions between proteins or genes, allowing the discovery of new biological pathways and molecular mechanisms. By predicting missing links in incomplete networks, GNNs suggest novel interactions that can be experimentally validated, accelerating the discovery process [67].

In graph classification, GNNs classify entire biological networks, such as differentiating between healthy and diseased tissue samples based on their network properties. For example, in cancer genomics, where GNNs can help identify network patterns associated with different cancer types. This way assists in diagnosis and prognosis, enabling tailored therapeutic strategies [4].

In the field of drug discovery, GNNs model the molecular structures of compounds as graphs, where nodes represent atoms and edges represent chemical bonds. By learning representations of these molecular graphs, GNNs can predict properties such as bioactivity, toxicity, and solubility, facilitating the identification of promising drug candidates. This approach accelerates the drug discovery process by screening large chemical libraries and prioritizing compounds for experimental validation, significantly reducing time and resource investment [68].

Pathway analysis benefits from GNNs by modelling the complex interactions within signalling pathways. By integrating multi-omics data, GNNs can identify key regulatory nodes and interactions that drive specific biological processes, providing an overview into disease mechanisms and therapeutic targets [69].

1.3 Language Models in Medical Research

Language models have become central in medical research, especially within the field of NLP applied to biomedical texts. These models enable machines to understand, interpret, and generate human language specific to the medical domain, facilitating tasks such as extracting clinical information from patient records, analysing biomedical literature, and assisting in diagnostic decision-making [70,71]. By predicting the likelihood of sequences of words within medical contexts, language models support applications ranging from automated reporting and coding to the summarization of clinical trial results [72,73]. Over the years, language modelling in medicine has evolved significantly, moving from traditional statistical approaches to advanced neural network-based models that can handle the complexity and specificity of medical language [74].

1.3.1 From N-Grams to Transformers: The Evolution of Language Models

One of the earliest approaches was the N-gram model, which predicts the probability of a word based on the previous $N-1$ words. This method is computationally efficient but suffers from the curse of dimensionality as N increases, leading to sparse data issues [75]. Despite its limitations, the N-gram model was widely used in early NLP applications such as text prediction [76].

Hidden Markov Models (HMMs) represent a step forward by incorporating states and transitions to model sequences of observations, which in the context of language are sequences of words. In an HMM, the model posits a series of latent *hidden* states that evolve according to a first-order Markov chain, where each state emits an observed word based on an emission probability distribution, and transitions between states are governed by a transition probability matrix [77]. HMMs are particularly useful in applications like part-of-speech tagging and speech recognition. However, HMMs assume that the probability of a word depends only on a finite state, limiting their ability to capture long-range dependencies in text [78].

Statistical language models, including Maximum Entropy models, addressed some limitations of N-gram models and HMMs by using statistical methods to predict the probability distribution of words. These models could incorporate various features and were more flexible in capturing language patterns. However, they still could not handle long-term dependencies and required extensive feature engineering [76].

The introduction of **neural networks** revolutionized language modelling. Early neural network-based models, such as feedforward neural networks, could learn continuous representations of words (word embeddings) and capture more complex patterns in text [79]. These models led to the development of Word2Vec, which provided a method for generating dense vector representations of words that capture semantic relationships [80].

Recurrent Neural Networks (RNNs) marked an advancement by introducing the concept of sequential processing, where the output from previous steps is fed as input to the current step [81]. This architecture allowed RNNs to capture temporal dependencies in text. However, RNNs cannot maintain context over long sequences due to issues like vanishing and exploding gradients [82].

To address these limitations, **Long Short-Term Memory** (LSTM) networks and **Gated Recurrent Units** (GRUs) were developed [83]. These models introduced gating mechanisms to better manage long-range dependencies and mitigate gradient issues. LSTMs and GRUs became the standard for many NLP tasks, such as machine translation and text generation, due to their ability to capture long-term dependencies more effectively [84].

The introduction of **attention mechanisms** represented a shift in language modelling [85]. Attention mechanisms allowed models to weigh the importance of different parts of the input sequence when making predictions. This led to the development of the Transformer model, which relies entirely on attention mechanisms and eschews recurrence [74]. Transformers process input sequences in parallel, significantly improving training efficiency and enabling the capture of long-range dependencies more effectively. The Transformer's architecture, consisting of an encoder and a decoder with multiple

layers of self-attention and feedforward neural networks, has become the foundation for modern language models such as BERT [86], RoBERTa [87] and GPT-3 [88].

Language Models can also be categorized by architecture—autoregressive (e.g. GPT), autoencoding (e.g. BERT), or encoder-decoder (e.g. T5)—and by their training objective, including causal language modelling (CLM), masked language modelling (MLM), and instruction-following fine-tuning. These distinctions shape how models learn context and generate outputs [89].

Most recently, **Large Language Models** (LLMs) have transformed NLP by demonstrating capabilities in generating and understanding human language [90]. They are characterized by their large scale, often involving billions of parameters, and are trained on vast, diverse text corpora (Table 1). This extensive training allows them to capture fine language patterns and context, enabling deep contextual understanding and generation [91].

However, Large Language Models present unique challenges in medical contexts: they may hallucinate plausible sounding but incorrect statements, perpetuate biases learned from training data, lack transparent reasoning (explainability), and raise regulatory and legal concerns around patient privacy, liability, and clinical decision-making [92].

Table 1 Comparison of Traditional Language Models and Large Language Models. The table compares traditional Language Models and LLMs across various dimensions. Language Models, often smaller with fewer parameters, are trained on specific datasets and can include n-grams, RNNs, LSTMs, or small transformers. They have limited capabilities, contextual understanding, and transfer learning effectiveness, making them suitable for specific applications like spell-checking and autocomplete. In contrast, LLMs are large-scale models with billions of parameters, trained on vast, diverse text corpora using transformer architecture. They offer deep contextual understanding, versatile capabilities across many tasks, and highly effective transfer learning, enabling wide-ranging applications in natural language processing.

Dimension	Language Models	Large Language Models
Scale	Smaller, fewer parameters	Large, often billions of parameters
Training Data	Trained on smaller, more specific datasets	Trained on vast, diverse text corpora
Architecture	Can include n-grams, RNNs, LSTMs, or small transformers	Primarily transformer-based
Capabilities	Limited to specific tasks or domains	Versatile, capable of many tasks with little tuning
Contextual Handling	Limited contextual understanding	Deep contextual understanding and generation
Transfer Learning	Less effective	Highly effective
Applications	Specific applications like spell-checking, autocomplete	Wide-ranging applications across NLP tasks

1.3.2 Medical Applications of Language Models: From Literature Mining to Drug Safety

The integration of language models into the medical field has resulted in significant advancements in medical research. Their ability to process and understand large volumes of textual data has made them indispensable in various medical applications [93,94].

The **exponential growth of biomedical research publications** presents a challenge for healthcare professionals to stay updated [95]. Language models can automatically summarize research articles, clinical trial results, and guidelines, compiling complex information into concise summaries

[96,97]. This helps clinicians and researchers in quickly assimilating new knowledge, thereby facilitating evidence-based practice. Tools leveraging transformer-based models, like BERT, have been developed to generate abstracts and highlight key findings from extensive medical texts [98,99].

Language models play a crucial role in **drug discovery** by mining scientific literature to identify potential drug targets and repurpose existing medications. They can detect patterns and relationships within vast datasets that might be overlooked by human researchers [100].

Additionally, in **pharmacovigilance**, language models help monitor and analyse reports of adverse drug reactions from various sources, including social media and clinical records, contributing to drug safety surveillance [101,102].

In **genomics**, language models help interpret genetic information by annotating genomic sequences and predicting the effects of genetic variations. They can process and correlate vast amounts of genomic data with phenotypic information, helping in the understanding of genetic diseases and the development of personalized therapies [103,104].

In **text generation**, LLMs can produce human-like text for creative writing, content creation, and dialogue systems. In machine translation, LLMs leverage their ability to understand and generate text in multiple languages, translating text between languages with high accuracy [105]. For summarization, LLMs can condense long documents into concise summaries, facilitating information retrieval and knowledge management [106].

1.4 Molecular Representation of Disease: Alzheimer's Disease, Cancer, and COVID-19

Diseases such as AD, cancer and COVID-19 present significant global health challenges, impacting millions of people and leading to substantial morbidity and mortality ('2022 Alzheimer's Disease Facts and Figures', 2022; Johns Hopkins University, 2021; World Health Organization, 2020a). These diseases

are characterized by complex pathophysiological mechanisms that require holistic approach for their prevention, diagnosis, and treatment [110,111].

1.4.1 Gene/Protein Targets and Pathways in Alzheimer's Disease

AD is a progressive neurodegenerative disorder marked by cognitive decline, memory loss, and impaired daily functioning ('Alzheimer's Disease Facts and Figures', 2021). The hallmark features of AD include amyloid-beta plaques and neurofibrillary tangles in the brain [113]. Amyloid-beta plaques are extracellular deposits that disrupt cell communication and trigger immune responses, leading to inflammation and neuronal death [114]. Neurofibrillary tangles are intracellular aggregates of hyperphosphorylated tau protein that destabilize microtubules essential for nutrient and molecular transport within neurons [115].

The pathogenesis of AD involves complex interactions among multiple pathways, including amyloid processing, tau phosphorylation, neuroinflammation, mitochondrial dysfunction, and oxidative stress[116,117]. Genetic factors, such as mutations in the APP, PSEN1, and PSEN2 genes, contribute to early-onset familial AD, while the APOE ε4 allele is a major risk factor for late-onset sporadic AD [118]. Environmental factors and lifestyle choices also influence disease risk [119].

Despite extensive research, effective treatments that can alter the disease course remain inexistent [120]. Current medications provide symptomatic relief but do not prevent or reverse neurocognitive decline [121].

Network-based methods are particularly relevant for studying diseases like AD, as they can model the interactions between multiple molecular entities and identify key nodes and pathways critical for disease progression[7,122]. Using techniques like RWR, we can simulate the spread of perturbations through the protein-protein interaction network, starting from known AD-related proteins [34]. This approach helps identify proteins most affected by the disease and those that could

be potential therapeutic targets. Understanding these interactions at a molecular level can lead to the development of novel therapeutic agents and strategies.

1.4.2 Gene/Protein Targets and Pathways in Cancer

Cancer is a disease characterized by the uncontrolled proliferation of abnormal cells, which can invade and destroy normal body tissues and metastasize to other parts of the body [123]. At the molecular level, cancer involves numerous disrupted pathways, including those regulating the cell cycle, apoptosis, DNA repair, and signal transduction [124,125]. Mutations in oncogenes and tumour suppressor genes lead to aberrant signalling pathways that promote cell survival, proliferation, angiogenesis, and metastasis [126,127].

The clinical management of cancer depends on understanding these molecular disruptions and targeting them effectively. Traditional treatments such as surgery, radiation therapy, and chemotherapy target rapidly dividing cells but often come with significant side effects and challenges, including resistance to treatment [128]. Advances in targeted therapy and immunotherapy have revolutionized cancer treatment by focusing on specific molecular changes unique to cancer cells [129]. Targeted therapies inhibit specific enzymes or growth factor receptors involved in cancer cell proliferation and survival, such as tyrosine kinase inhibitors that block signalling pathways [130]. Immunotherapies enhance the body's immune response against cancer cells by targeting regulatory proteins like PD-1 and CTLA-4, enabling immune cells to recognize and destroy cancer cells more effectively [131,132]. Despite these advancements, cancer treatment remains a challenge due to the heterogeneity of the disease and the ability of cancer cells to develop resistance [133].

Ongoing research into reliable biomarkers and novel therapeutic agents is crucial [134]. Biomarkers can aid in early detection, prognosis, and the development of personalized treatment strategies [134,135]. Understanding the molecular representation of cancer through advanced

computational techniques can help identify novel therapeutic targets, predict treatment responses, and improve patient outcomes [135].

1.4.3 Gene/Protein Targets and Pathways in COVID-19

COVID-19, caused by the SARS-CoV-2 virus, has led to a global pandemic with severe health, social, and economic impacts [136]. The virus primarily targets the respiratory system, causing symptoms ranging from mild cough and fever to severe acute respiratory distress syndrome, pneumonia, and multi-organ failure [137]. COVID-19 also affects cardiovascular, neurological, and gastrointestinal systems [138,139]. The heterogeneous nature of COVID-19 and the emergence of new variants present ongoing challenges for public health and clinical management [140,141].

The molecular representation of COVID-19 involves understanding the virus's mechanisms of infection, including how it interacts with host cells, evades the immune system, and causes widespread tissue damage [142]. The SARS-CoV-2 virus uses its spike protein to bind to the angiotensin-converting enzyme 2 (ACE2) receptor on host cells, facilitating viral entry [143]. Once inside the host cell, the virus takes over the cellular machinery to replicate and produce viral proteins, leading to cell damage and an immune response [144].

The host's immune response to the virus involves complex interactions among various cells and signalling pathways, which can lead to severe inflammatory responses and tissue damage, known as cytokine storms. Understanding these interactions is crucial for developing effective treatments and managing the disease [145].

Therapeutic strategies for COVID-19 have included antiviral drugs, monoclonal antibodies, and supportive therapies [146]. Antiviral drugs like remdesivir target viral replication enzymes [147], while monoclonal antibodies neutralize the virus by binding to the spike protein, reducing viral load [148].

Supportive therapies, such as corticosteroids, manage severe inflammatory responses by suppressing the immune system ('Dexamethasone in Hospitalized Patients with Covid-19', 2021).

Advanced computational techniques and experimental validations are essential in uncovering the molecular mechanisms underlying COVID-19 [150]. By leveraging network-based machine learning, researchers aim to identify key molecular targets and pathways that can inform the development of novel therapeutic approaches, vaccines, and diagnostic tools, ultimately improving disease management and patient outcomes [151,152].

1.5 The Intersection of Diet and Disease: Implications for Risk, Prevention, and Treatment

The intricate relationship between diet and diseases such as AD, cancer, and COVID-19 underscores the important role of nutrition in the development, prevention, and management of these conditions [153,154]. Dietary habits influence disease risk through various mechanisms, including modulation of inflammation, oxidative stress, immune function, and hormonal balance [155,156].

Epidemiological studies have consistently linked specific dietary patterns to the risk of developing certain diseases [157,158]. Diets high in processed foods, red meats, saturated fats, and added sugars are associated with an increased risk of conditions such as cardiovascular diseases, type 2 diabetes, and certain cancers [159–161]. In contrast, diets rich in fruits, vegetables, whole grains, legumes, nuts, and lean proteins—often exemplified by Mediterranean or plant-based diets—are correlated with a reduced risk of many diseases [162–164]. These healthier diets are abundant in antioxidants, fibres, and phytochemicals, which contribute to their protective effects [165].

Antioxidants such as vitamins C and E, selenium, and carotenoids protect cells from oxidative stress, a known contributor to the development of diseases [166,167]. Phytochemicals, particularly flavonoids found in fruits, vegetables, and teas, have anti-inflammatory and antioxidant properties [168]. Cruciferous vegetables like broccoli and Brussels sprouts are rich in glucosinolates, which metabolize into bioactive compounds such as sulforaphane, indoles, and isothiocyanates [169,170].

These compounds modulate detoxification enzymes, alter hormone metabolism, and exert anti-inflammatory effects [171].

Beyond general dietary patterns, specific phytochemicals have garnered attention for their disease-preventing properties. For example, omega-3 fatty acids found in fatty fish have anti-inflammatory effects and are associated with cardiovascular health [155,172]. Resveratrol in grapes and berries has been studied for its potential neuroprotective and anti-cancer properties [173,174].

Emerging research suggests that certain dietary approaches may influence disease progression and enhance the efficacy of conventional treatments [175,176]. Calorie restriction and intermittent fasting have been observed to improve metabolic parameters and may enhance the response to therapies [177,178]. Diets like the ketogenic diet are being explored for their potential to improve metabolic health and possibly impede disease progression in conditions such as epilepsy and certain cancers [179,180].

Nutrition plays a critical role during the treatment of diseases. Adequate nutrition can affect a patient's strength, tolerance to treatment side effects, and overall quality of life [181]. Malnutrition is a common challenge among patients with diseases, leading to poorer outcomes [182,183]. Therefore, a balanced diet, customized to meet the individual needs of patients, is crucial for maintaining body weight and energy, rebuilding tissue, and supporting the immune system during and after treatment [184].

While the role of diet in disease risk reduction and management is well-supported, several challenges exist in translating dietary interventions into therapeutic outcomes [185]. Determining the optimal intake levels of phytochemicals that confer protective or therapeutic effects is complex due to variability in natural concentrations and bioavailability [186]. Many bioactive compounds have low bioavailability, meaning only a small portion reaches target tissues [187]. Human absorption and metabolism can vary widely among individuals [188]. Genetic variability influences individual responses to dietary interventions, affecting efficacy and necessitating personalized nutrition approaches. Much of the current evidence is based on *in vitro* studies or animal models [189,190]. There is a need for

extensive human clinical trials to validate the therapeutic potential of specific dietary components. Interactions between phytochemicals and pharmaceutical agents are not fully understood, which may affect drug efficacy and safety.

1.6 Gut Microbiota as a Mediator of Phytochemical–Disease Interactions

Diet profoundly shapes the gut microbiota: diets rich in fermentable fibers - such as inulin, resistant starches, and non-digestible oligosaccharides-serve as substrates for short-chain fatty acid - producing bacteria (e.g., Bifidobacterium, Faecalibacterium, Roseburia), whose fermentation products acetate, propionate, and butyrate strengthen intestinal barrier integrity by upregulating tight-junction proteins and mucin production, modulate mucosal immunity through regulatory T-cell induction, and attenuate systemic inflammation - whereas high-fat, low-fiber *Western* diets reduce overall microbial diversity, promote expansion of pro-inflammatory pathobionts, and contribute to colorectal carcinogenesis and neuroinflammation [191].

Gut microbes also play an important role in determining phytochemical bioactivity by enzymatically transforming parent compounds into more potent, bioavailable metabolites: for instance, only 30-50 % of individuals harbour equol-producing bacteria (e.g., Adlercreutzia equolifaciens, Slackia isoflavoniconvertens) that convert soy isoflavones into the estrogen-receptor–agonist equol, which exhibits higher receptor affinity and longer systemic half-life; similarly, ellagitannin-rich foods yield urolithins A and B via bacterial decarboxylation of ellagic acid, metabolites that readily cross the blood–brain barrier, exert anti-inflammatory effects in microglia, and ameliorate amyloid-β pathology in AD models [192].

Conversely, many phytochemicals actively reshape the microbiota itself: epigallocatechin-3-gallate from green tea promotes Akkermansia muciniphila bloom and enhances bile-acid–mediated signaling, while high-dose dietary flavonoids exhibit antimicrobial activity against

Clostridium species, together fostering a bidirectional diet-microbiota interaction that modulates host metabolic and immune pathways relevant to disease prevention and therapy [193].

By joining computational discovery, experimental validation, and nutritional delivery into a single translational pipeline, this thesis takes phytochemicals from *in silico* findings to practical dietary recommendations. First, network-based machine learning identified EVOO compounds predicted to modulate AD-related protein networks. Next, a GNN-driven synergy model suggested pharmacodynamic interactions between these phytochemicals and approved drugs, which we then confirmed through *in vitro* assays in human cancer cell lines. Finally, LLMs leveraged those experimentally supported bioactive compounds to generate phytochemically enriched recipes tailored for AD, cancer, and COVID-19 pathways.

2. General Methodology

All the mathematical and computational foundations—covering random walks, network diffusion, graph embeddings, synergy scoring metrics, and transformer architectures—are explained in this chapter, ensuring that each subsequent application chapter can focus on its domain-specific methods.

2.1 Basic Concepts and Definitions

Networks were utilized to represent data at various levels of abstraction, ranging from molecular interactions to population dependencies, encompassing social and health interactions. Network analysis tools have been widely applied in biology and medicine for tasks such as drug target identification, gene function prediction, and early disease diagnosis. Networks commonly used in systems biology include Protein-Protein Interaction (PPI) networks, biochemical networks, transcriptional regulation networks, and signal transduction and metabolic networks. This dissertation leverages PPI networks for various applications.

2.2 Biological Networks

Mathematically, the PPI networks used in this dissertation were represented as $G = (V, E, X)$, where V is the set of nodes, $|V| = N$ represents the number of nodes, E is the set of edges representing the connections between the nodes, and $X = \mathbb{R}^{Nxd}$ represents the d-dimensional node attributes. An edge $e = \{(v_i, v_j) | v_i, v_j \in V\}$ defines the connection between nodes v_i and v_j. Nodes v_i and v_j are considered neighbors, and the set of neighbors of a node v_i is termed its neighborhood, denoted as $N(v_i) = \{v_j : (v_j, v_i) \in E\}$.

Graphs are commonly represented using an adjacency matrix. Given a graph $G = (V, E, X)$, the adjacency matrix is an NxN matrix $A = (a_{ij})$, where $a_{ij} = 1$ if $(v_i, v_j) \in E$ or $a_{ij} = 0$ otherwise. For a weighted graph, $a_{ij} = w_{ij}$ if $(v_i, v_j) \in E$ or $a_{ij} = 0$ otherwise.

PPI networks depict physical interactions between proteins within a cell. The fundamental assumption is that proteins involved in the same biological processes are closely connected within the PPI network. These networks are constructed using large-scale, high-throughput techniques like pull-down assays, tandem affinity purification, and mass spectrometry. Protein-protein interactions are collected and stored in public databases to facilitate their use by the scientific community. Some of these databases include STRING and STITCH. Proteins whose interactions result in a particular product or change within a cell are considered to belong to the same biological pathway. Biological pathways are classified into several types, including metabolic pathways, signalling pathways, and regulatory networks. These pathways are compiled in various databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), Panther, and Reactome.

2.3 Random Walks for Phytochemical Discovery

Graphs are a powerful way to represent biological systems due to their ability to capture complex relationships. However, extracting meaningful patterns from biological data on graphs can be challenging because such data is often sparse and noisy. To address this, network propagation has been widely applied to graphs in various biological applications, including drug and disease protein target prediction, disease stratification, and drug repositioning. Network propagation effectively amplifies biological signals within networks, aiding in the discovery of significant patterns that characterize biological systems.

In network propagation, the process begins with a vector p_0 representing initial scores on nodes (e.g., proteins or genes) based on prior knowledge (e.g., targeted proteins for drugs or altered proteins for diseases). For instance, $p_0(v_i) = 1$ for known disease- or drug-associated proteins and $p_0(v_i) = 0$ for all other proteins. At each time point k, the information at each node v_i depends on the sum of the information at its neighboring nodes $N(v_i)$ from the previous time point $k - 1$. The information from each neighboring node is weighted in proportion to the corresponding edge weight:

$$p_k(v_i) = \sum_{v_j \in N(v_i)} p_{k-1}(v_j) w(v_i, v_j) \quad (1)$$

where $w(v_i, v_j)$ is the normalized weight of the interaction between v_i and v_j. Running this process for k steps provide $p_k(v_i)$, which ranks each node based on its proximity to the initial set of nodes set to 1.

In matrix notation, network propagation is expressed as:

$$p_k = W p_{k-1} \quad (2)$$

where W is a normalized version of the adjacency matrix of the network. Repeated iterations give $p_k = W^k p_0$. If W is a stochastic matrix (its columns sum to 1), this process is equivalent to a RW on the network. For small k, the diffused distribution is close to the initial distribution, while for large k, it reflects the network topology.

To prevent over-diffusion of information, RWR includes a restart probability parameter α, which controls the amount of information propagated at each step. Iterative updates using RWR are formulated as:

$$p_k = \alpha p_0 + (1 - \alpha) W p_{k-1} \quad (3)$$

where $\alpha \in [0,1]$ is the restart probability. When the network is connected and the eigenvalues of W are at most 1 in absolute value, the process converges to a steady-state distribution:

$$p = \alpha (I - (1 - \alpha) W)^{-1} p_0 \quad (4)$$

with $W = AD^{-1}$. Network propagation has yielded state-of-the-art results in biological studies (Figure 1).

However, traditional network propagation algorithms have limitations. They fail to learn propagation specific to the task and often require a grid search to optimize parameters, such as the

restart probability in RWR. Recent advances in deep learning on graphs have introduced differentiable and learnable propagation or diffusion rules on graphs, which have been successfully applied in numerous biomedical applications. The next section will review representation learning approaches on graphs, focusing on traditional methods and recent approaches using random walks and graph convolutional operators to compute node and (sub)graph embeddings.

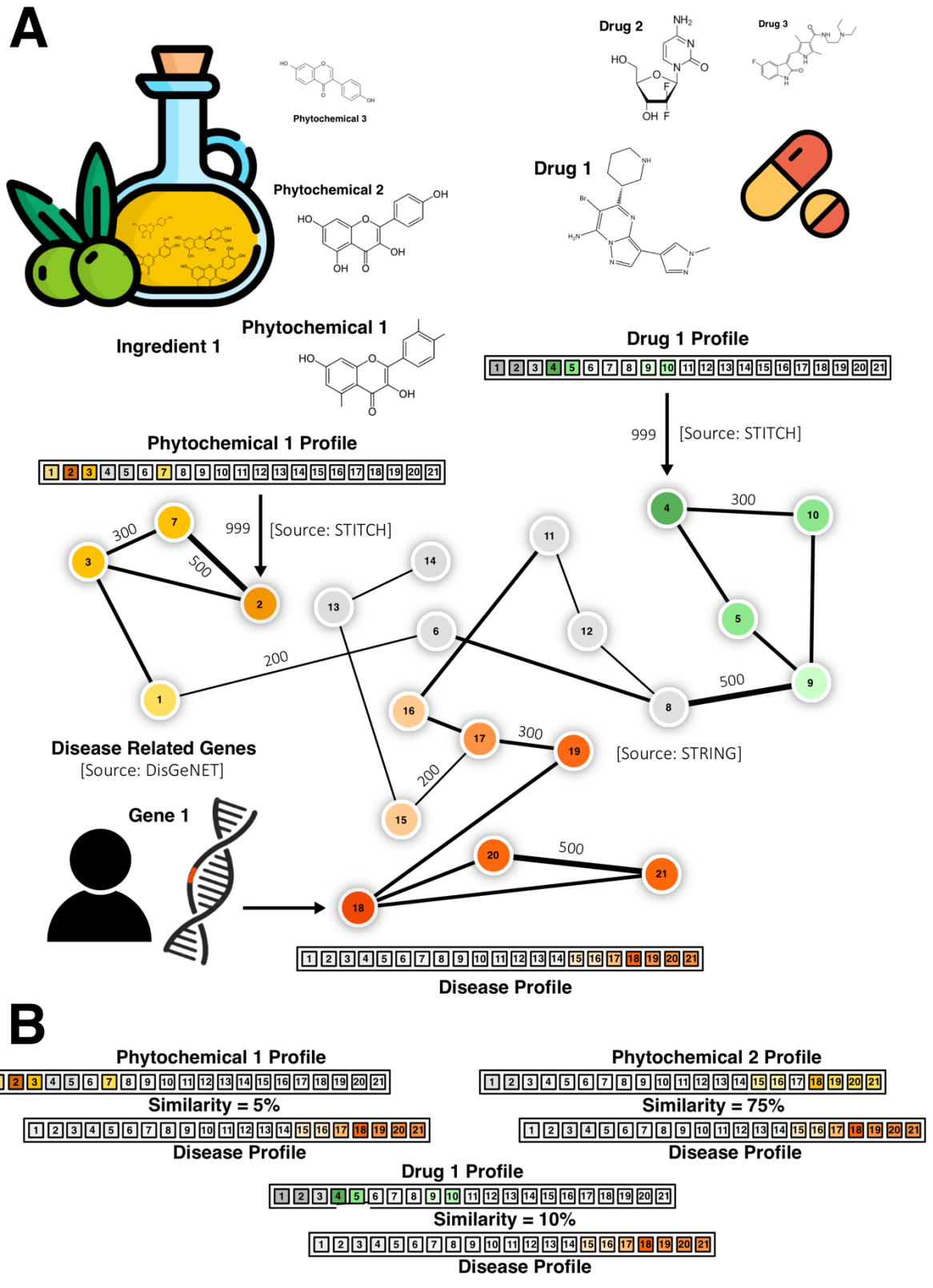


Figure 1 Overview of Phytochemical Discovery Chapter 3. Machine learning-driven discovery of phytochemicals using the RWR algorithm on biological networks, integrating data from STRING, STITCH, and DisGeNET to identify phytochemicals targeting AD biological network. The phytochemical profiles are compared with disease profiles to determine their similarity, guiding the selection of compounds for further study.

2.4 Representation Learning on Graphs and Validation of Compound Synergies

2.4.1 Computational Predictions of Compound Synergies

Hamilton et al. provided an insightful review of representation learning algorithms on graphs, and this section draws heavily from their work. The primary goal of machine learning on graphs is to integrate the graph structure into the learning model. Traditional methods achieve this by extracting hand-engineered features from the graph. However, contemporary approaches aim to learn representations that encode the graph structure directly and can adapt during the learning process. This is accomplished through a mapping that embeds nodes or entire (sub)graphs as points in a low-dimensional vector space $\mathbb{R}^d$. The resulting space reflects the geometric relationships inherent in the original graph structure. These learned embeddings, which capture the structural information of the graph, are then used for downstream machine learning tasks, thereby implicitly incorporating the graph's structural information [194].

Representation learning diverges from traditional approaches by treating the task of capturing structural information as a machine learning problem. Instead of using hand-engineered statistics, representation learning employs a data-driven approach to learn embeddings that encode graph structure [194]. A theoretical framework was introduced to unify all representation learning approaches, which is briefly outlined here. Let $G = (V, E)$ be an undirected graph with a binary adjacency matrix A and real-valued node attributes $X \in \mathbb{R}^{m \times |V|}$. The objective of graph representation learning is to use X and A to map each node, or a (sub)graph, to a vector $z \in \mathbb{R}^d$ where $d \ll |V|$.

Node embedding methods involve two key mapping functions: an encoder, which maps nodes to vector embeddings $z_i \in \mathbb{R}^d$, and a decoder, which transforms embeddings to user-specified graph statistics. Most approaches use a pairwise decoder that, when applied to embeddings of two nodes z_i and z_j, outputs a measure of proximity between v_i and v_j in the original graph. Formally:

$$ENC: V \rightarrow \mathbb{R}^d \text{ (5)}$$

$$DEC: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^+ \text{ (6)}$$

The aim is to optimize the encoder and decoder mappings to minimize reconstruction error:

$$DEC(ENC(v_i), ENC(v_j)) = DEC(z_i, z_j) \approx s_G(v_i, v_j) \text{ (7)}$$

where $s_G(v_i, v_j)$ is a user-defined, graph-based proximity measure between nodes, defined over the graph G. In practice, most methods optimize this task by minimizing an empirical loss, L, over a set of training node pairs, D:

$$L = \sum_{(v_i, v_j) \in D} \quad l(DEC(z_i, z_j), s_G(v_i, v_j)) \text{ (8)}$$

where $l: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is a user-specified loss function measuring the discrepancy between estimated and true values of pairwise node distances.

In direct encoding approaches, the embedding matrix is optimized directly, making the encoding function a simple embedding lookup:

$$ENC(v_i) = Zv_i \text{ (9)}$$

where Z is a matrix containing the embeddings, and v_i is a one-hot indicator vector indicating which column of Z corresponds to each node (i.e., a binary vector of length $|V|$ with a single *1* at the position for node v_i and *0*s elsewhere). Random-walk approaches fall into this category as they optimize node embeddings so that nodes co-occurring on short random walks over the graph have similar embeddings. Notable algorithms include DeepWalk [21] and node2vec [22]. Despite their expressiveness, these methods have drawbacks: no parameter sharing between nodes, leading to statistical inefficiency; computational inefficiency as the number of parameters grows with $O(|V|)$; failure to leverage node attributes; and transductiveness, meaning they can only generate embeddings for observed nodes.

Generalized encoder-decoder architectures utilize more complex encoders that leverage graph structure and node attributes to overcome the limitations of direct-encoding methods. This category includes neighbourhood aggregation and convolutional encoders. These methods compute node embeddings by aggregating embeddings from neighbouring nodes in the graph. Initially, node embeddings are set to the input node attributes. At each iteration, the encoder aggregates information from neighbours for each node. New node embeddings are the result of combining the aggregated neighbourhood information with the node's previous embeddings.

Embeddings are processed through a multi-layer neural network either before or after aggregation. This iterative process repeats for as many layers as the encoder has. With each additional layer, node embeddings incorporate information from an additional hop in the graph; for instance, a 1-layer encoder aggregates information from the one-hop neighbourhood, a 2-layer encoder from the two-hop neighbourhood, and so forth. The dimensionality of the embeddings remains constrained as the process iterates because the encoder compresses all the neighbourhood information into a low-dimensional vector. After K iterations, the process concludes, and the final embedding vectors are output as the node representations [194].

Several established approaches in deep learning on graphs follow this procedure, including Graph Convolutional Networks (GCN) [25], the GraphSAGE algorithm [54], and the ChebNet algorithm [195]. The trainable parameters in the encoder dictate how to aggregate information from a node's local neighbourhood and are shared across nodes. This parameter sharing enhances efficiency (i.e., the parameter dimensions are independent of the graph size), provides regularization, and enables the generation of embeddings for nodes not observed during training [194].

2.4.2 Compounds Response Curve and Synergies

Each compound response such as food or drug was described by the Hill curve as in Equation 10.

$$Y = \frac{100}{1+10^{(logIC50-X)xHill\ Slope}} \quad (10)$$

where Y is the response, X is the logarithm of the drug concentration, *IC50* is the concentration of the drug at which the response is 50% of the maximum and *Hill Slope* indicates the steepness of the curve.

Equation 10 can also be described in term of values directly obtained through experimental values as in Equation 11.

$$E(c) = E_{max} \frac{c^h}{EC_{50}^h + c^h} \quad (11)$$

where $E(c)$ is the observed effect (e.g., percentage inhibition) at concentration c. E_{max} is the maximum observed inhibition, expressed as a percentage. It represents the theoretical saturation point of the drug's effect. EC_{50} is the half-maximal effective concentration at which the drug reaches 50% of its maximum effect, providing an indication of drug potency—the lower the more potent the drug, and h is the Hill coefficient that describes the steepness of the curve. A value greater than 1 means that the effect increases rapidly with concentration. A coefficient close to 1 suggests a standard sigmoidal response, while values less than 1 suggest a more gradual response

The curve fitting process involved **estimating the parameters E_{max}, EC_{50}, and h** from the experimental data using non-linear least squares regression. In this analysis, the Levenberg-Marquardt algorithm was employed to minimize the sum of squared residuals between the experimental data points and the predicted values from the Hill equation. The optimization procedure iteratively adjusted the parameters until the best fit was achieved. First, experimental data were first pre-processed by removing outliers, missing, infinite values. Second, the algorithm started with initial guesses for E_{max}, EC_{50}, and h. These values were provided within predefined bounds. E_{max} was constrained between 0 and 100%, assuming a range between no inhibition and complete inhibition. EC_{50} was bounded between 0 and an upper limit based on experimental concentrations. The Hill coefficient h was restricted between 1 and 3 to model typical sigmoidal curve shapes. Thirdly, the fitting algorithm minimized the error between predicted and observed responses by adjusting the parameters. The

output of this curve fitting were the previous estimated parameter values plus the covariance matrix, which contains the variances and covariances of these.

The **standard deviations associated with E_{max}, EC_{50}, and h** were calculated from the covariance matrix generated during the fitting process. The diagonal elements of the covariance matrix represent the variances of the fitted parameters, and the standard deviations for each parameter are given by the square root of the corresponding diagonal element as indicated in Equations 12, 13 and 14.

$$Std(E_{max}) = \sqrt{Cov(E_{max}, E_{max})} \quad (12)$$

$$Std(EC_{50}) = \sqrt{Cov(EC_{50}, EC_{50})} \quad (13)$$

$$Std(h) = \sqrt{Cov(h, h)} \quad (14)$$

In addition to the fitted parameters, *IC40* and *IC50* were calculated. The following equation was used to calculate the IC_x value (where x is the desired inhibition percentage, such as 50 or 40):

$$IC_x = EC_{50}\left(\frac{xE_{max}}{E_{max} - xE_{max}}\right)^{1/h} \quad (15)$$

Where x is 0.5 for IC50 or 0.4 for IC40.

To calculate the **standard deviation associated with IC50 and IC40**, error propagation was performed using the partial derivatives of the IC_x formula with respect to E_{max}, EC_{50}, and h. This allows the incorporation of the uncertainties of the fitted parameters into the final uncertainty of the IC_x values. The variance of the IC_x is given by:

$$Var(IC_x) = \left(\frac{\partial IC_x}{\partial E_{max}}\right)^2 Var(E_{max}) + \left(\frac{\partial IC_x}{\partial EC_{50}}\right)^2 Var(EC_{50}) + \left(\frac{\partial IC_x}{\partial h}\right)^2 Var(h) \quad (16)$$

And the standard deviation by the square root of the variance:

$$Std(IC_x) = \sqrt{Var(IC_x)} \text{ (17)}$$

After the monotherapies, combinatorial treatments were performed, allowing us to estimate whether the **ZIP score** of each combination was aligned with the previous computational predictions. The ZIP score was calculated by comparing the observed combination response to the expected response under the assumption of no interaction. The expected response is derived from the dose-response curves of each compound individually as in Equation 18.

$$E_i(d) = \frac{E_{max,i} \cdot d^h}{EC_{50,i}^h + d^h} \text{ (18)}$$

where $E_i(d)$ is the effect of drug i at dose d, $E_{max,i}$ is the maximum effect of drug i, $EC_{50,i}$ is the concentration of drug i that produces 50% of its maximum effect and h is the Hill coefficient, representing the steepness of the dose-response curve.

The **ZIP score** ($E_{expected}$) was calculated by comparing the observed combination response to the expected response under the assumption of no interaction. For a given combination of two compound concentrations, ZIP Score was calculated following Equation 21. For the overall synergy, all values were averaged across all tested combinations.

$$E_{expected}(d_1, d_2) = E_1(d_1) + E_2(d_2) - E_1(d_1) \cdot E_2(d_2) \text{ (21)}$$

The **standard deviations for the overall ZIP score** associated to each replica was obtained using Equation 22 (Figure 2).

$$S_{ZIP} = \frac{E_{observed} - E_{expected}}{\sigma_{expected}} \text{ (22)}$$

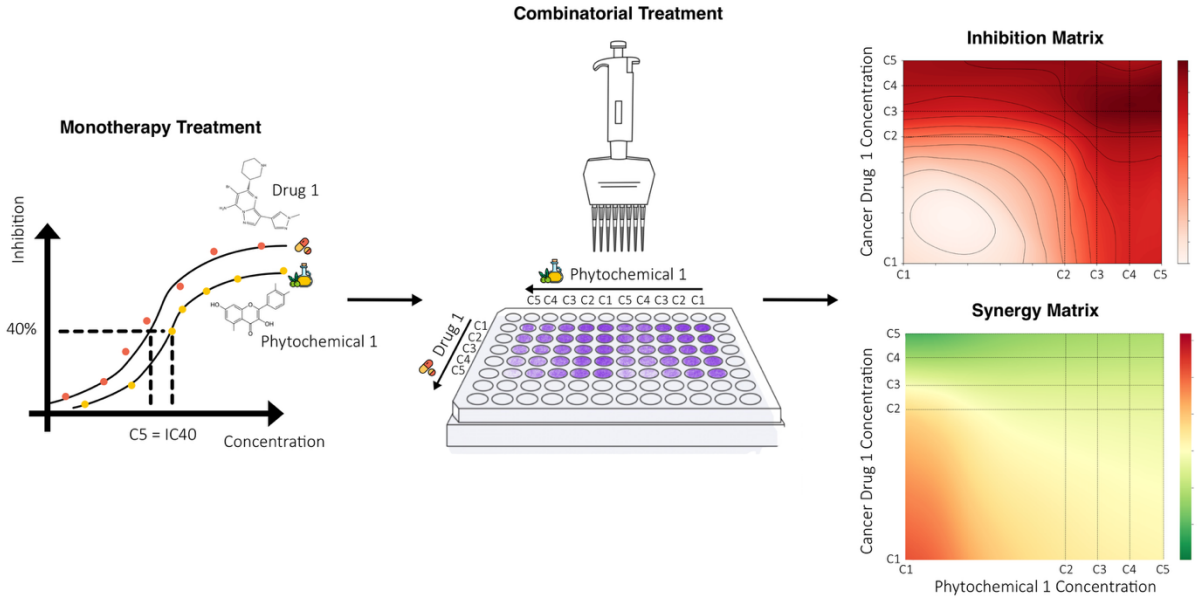


Figure 2 Overview of Drug-Food Synergy Chapter 4. Experimental validation of drug-food synergies, where the selected phytochemicals undergo monotherapy assays followed by combination studies with cancer drugs. Inhibition and synergy matrices visualize the effect of various concentrations of cancer drugs and phytochemicals on cell viability, validated through lab experiments.

2.5 Large Language Models for Recipe Personalization

Initially, a dataset of recipes, including ingredient lists, preparation instructions, was gathered. Plus, the information on the phytochemical content of several ingredients is compiled. Then, the LLMs were fine-tuned on this dataset to learn the relationships between ingredients and their phytochemical properties. Next, the fine-tuned LLMs generate potential ingredient substitutions that enhance the phytochemical profile of the recipes by suggesting a substitute to every ingredient and only keeping the recipes at the end that were phytochemically enriched.

The core architecture of LLMs is the Transformer. The self-attention mechanism is the foundation of the Transformer model, allowing it to weigh the importance of different words in a recipe relative to each other. Given this input sequence, the self-attention mechanism computes a set of attention weights that determine how much focus each word should receive when encoding a particular word. This mechanism can be described as follows:

1. **Input Embeddings**: Each word in the recipe is transformed into a vector using an embedding matrix.

2. **Query, Key, and Value Matrices**: Three matrices W_Q, W_K, and W_V are learned, which are used to project the input embeddings into queries Q, keys K, and values V, respectively:

$$Q = W_Q X, K = W_K X, V = W_V X \quad (21)$$

where X is the input embedding matrix.

3. **Scaled Dot-Product Attention**: The attention scores are computed using the dot product of the queries and keys, scaled by the square root of the dimensionality of the keys:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (22)$$

where d_k is the dimensionality of the keys.

4. **Multi-Head Attention**: To capture different aspects of the relationships between words, the Transformer employs multiple attention heads:

$$MultiHead(Q, K, V) = Concat(head_1, head_2, \ldots, head_h)W_O \quad (23)$$

where each $head_i = Attention(QW_{Q_i}, KW_{K_i}, VW_{V_i})$ and W_O is a learned output projection matrix.

5. **Position-Wise Feed-Forward Networks**: After the multi-head attention layer, the output is passed through a position-wise feed-forward network:

$$FFN(x) = ReLU(xW_1 + b_1)W_2 + b_2 \quad (24)$$

where W_1, W_2, b_1, and b_2 are learned parameters.

6. **Positional Encoding**: Since Transformers do not inherently capture the order of words, positional encodings are added to the input embeddings to provide information about the position of each word in the sequence

OpenAI's DaVinci, GPT-3.5 and Meta's TinyLlama were benchmarked against the current state of the art Meta's GISMO for the ingredient substitution task. We describe in detail how each model was employed for the specific task of ingredient substitution in *Section 5.3* (Figure 3).

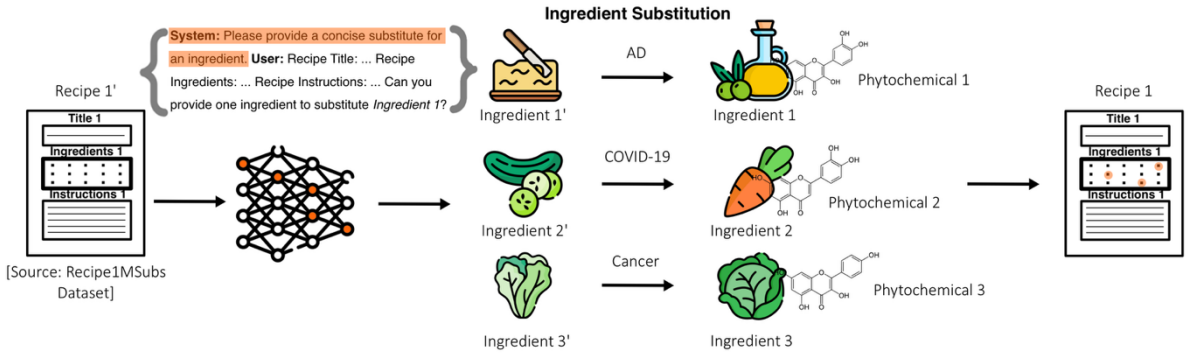


Figure 3 Overview of Recipe Enrichment Chapter 5. LLMs are employed to generate phytochemically enriched recipes by optimizing ingredient substitutions. Ingredients are substituted based on their potential health benefits against diseases like AD, COVID-19, and cancer, resulting in health-enhancing recipes. The LLM is fine-tuned on a recipe substitution dataset to integrate bioactive compounds.

2.6 Datasets Used in the Dissertation

To build the PPI network, we used **STRING** (Search Tool for the Retrieval of Interacting Genes/Proteins) [196] and **STITCH** (Search Tool for Interactions of Chemicals) [197] databases. STRING is a database with protein-protein interactions. It includes direct (physical) and indirect (functional) associations obtained from various sources such as computational prediction, knowledge transfer between organisms, and interactions aggregated from other (primary) databases. We used the version 11.0, which contains over 24.6 million proteins from 5,090 organisms and features more than 1 billion interactions. The confidence scores provided by STRING allowed us to filter and weigh the interactions within the network. STRING was used to build the PPI network necessary for the RWR leading to phytochemical discovery. STITCH is a database that integrates information about interactions between chemicals and proteins. The interactions are derived from experimental data and computational predictions. The version used in this dissertation is STITCH v5.0, which includes information on interactions for over 9 million proteins and chemicals. STITCH was used to explore the chemical-protein interaction networks that complement the protein-protein interaction data from STRING, providing information on the targets of the phytochemicals. We restricted our analyses to interactions with high confidence to maximize reliability, and we discuss remaining limitations—such as incomplete coverage,

potential false positives from text-mining errors, and biases toward well-studied proteins—in Chapter 3.

For obtaining drug target information we relied on **DrugBank** [198] and **DrugCentral** [199] databases. DrugBank combines drug data with drug-target information. It includes information on drug actions, mechanisms, chemical properties, and drug-target interactions. The version used is DrugBank 5.1.8, featuring over 13,000 drug entries, including FDA-approved drugs, experimental drugs, and nutraceuticals, with corresponding target, enzyme, transporter, and carrier data. DrugBank was utilized to identify FDA-approved drugs and their targets, which were essential for training the machine learning models to predict interactions and validate phytochemical efficacy. DrugCentral is a curated database of approved drugs and drug targets, providing drug information, including pharmacological actions, mechanisms, and targets. The version used is DrugCentral 2021, containing information on over 4,500 drugs and their associated targets. DrugCentral data were used alongside DrugBank to enrich the dataset of FDA-approved drugs, providing larger training data for our models.

For obtaining the targets within our PPI for the phytochemicals, we used **FoodDB** as a database of food components and their associated health effects, including information on phytochemicals found in various food items and their biological activities. The version used is FoodDB 2020, which includes information on thousands of phytochemicals from diverse food sources, with their chemical properties and potential health benefits. FoodDB was used for identifying and quantifying the phytochemicals present in EVOO and other foods, which were subsequently analysed for their potential health benefits [200].

EVOO Phytochemical Dataset includes the phytochemical content in EVOO, including respective concentrations. The concentration range values were obtained experimentally, providing a detailed profile of various phytochemicals present in EVOO, drupes, and olive leaves. This dataset includes information on the mass-to-charge ratio (m/z), retention time (RT), molecular formula, and the names of identified phytochemicals. Additionally, it provides qualitative formulas for each phytochemical and records their minimum and maximum concentrations across different sources

(EVOO, drupes, and leaves). This dataset was used identifying the phytochemicals present at significant concentration in EVOO.

For obtaining the dysregulated genes, **DisGeNET** is a platform integrating information on gene-disease associations from various sources, including scientific literature and curated databases. The version used was DisGeNET v7.0, which comprises over 1.1 million gene-disease associations, covering more than 21,000 genes and 30,000 diseases. DisGeNET was used to identify genes associated with AD. **UniProt** is a comprehensive resource for protein sequence and annotation data. The version used is UniProt 2021, which contains over 564,000 reviewed (Swiss-Prot) protein sequences and millions of unreviewed (TrEMBL) entries. UniProt provides high-quality, manually curated protein information, including function, classification, and sequence data [201]. **Comparative Toxicogenomics Database** (CTD) provides information on chemical-gene/protein interactions and their effects on diseases. The version used is CTD 2021, which includes over 46 million toxicogenomic connections among chemicals, genes, phenotypes, diseases, and exposure events [202]. **Orphanet** is a reference portal for information on rare diseases and orphan drugs. The version used is Orphanet 2017, containing detailed information on over 6,000 rare diseases, including clinical signs, epidemiology, and genetic data [203]. **Clinical Genome Resource (**ClinGen) provides curated information on clinically relevant genes and variants. The version used is ClinGen 2015, which includes expert-curated gene-disease validity classifications, dosage sensitivity, and variant interpretations [204]. **Genomics England PanelApp** is a crowdsourcing platform for gene panels used in disease diagnosis. The version used is Genomics England PanelApp 2019, which contains gene panels for various diseases, curated by experts and the scientific community [205]. **Cancer Genome Interpreter** interprets the biological and clinical relevance of tumor genomic alterations. The version used is Cancer Genome Interpreter 2018, which includes a database of cancer driver genes and mutations, along with their implications for targeted therapies [206]. **Psychiatric Disorders Gene Association Network** (PsyGeNET) is a database of gene-disease associations focused on psychiatric disorders. The version used is PsyGeNET 2015, containing curated associations between

genes and psychiatric disorders from expert literature [207]. This refinement process ensured that only the most relevant and well-characterized genes were included in our analyses.

For evaluating the pathways involved in the mechanisms of disease, KEGG (Kyoto Encyclopedia of Genes and Genomes) is a collection of databases including genomes, biological pathways, diseases, drugs, and chemical substances. The version used is KEGG 2021, containing information on metabolic pathways, disease pathways, and drug action pathways. KEGG pathways were used in the Gene Set Enrichment Analysis (GSEA) to understand the biological processes and pathways influenced by the identified phytochemicals [208].

For the compound synergy validation, GDSC (Genomics of Drug Sensitivity in Cancer) dataset. It is accessible through the CancerRxGene database and provided data on the sensitivity of cancer cell lines to various anti-cancer drugs. The dataset included drug response data across a wide range of cancer types and cell lines, making it one of the most complete resources for studying drug efficacy. This dataset was used to validate the computational predictions involving drug and phytochemical interactions [209].

Recipe1M is a large-scale dataset of culinary recipes, and Recipe1MSubs is a subset focused on ingredient substitutions. The version used is Recipe1M and Recipe1MSubs 2021, which included over one million recipes with associated ingredients, instructions, and nutrient information. Recipe1MSubs contains 70,520 pairs of ingredient substitutes. These datasets were used to train LLMs for generating phytochemically enriched recipes [210].

We compiled in Table 2 how all datasets were used across the three main parts of this dissertation including molecule discovery, experimental validation and recipe enrichment.

Table 2 Summary of the Datasets Used in the Dissertation. Overview of the datasets, categorizing them into types such as protein-protein interaction networks, drug and phytochemical data, disease-related information, pathway details, experimental validation data, and custom datasets. Each dataset entry includes the name, type, a brief description, version, key features (e.g., number of entries, interaction types), and its specific application in the dissertation. These datasets were used to construct machine learning models, performing network analyses, validating findings experimentally, and generating phytochemically enriched recipes.

Dataset	Type	Description	Version	Key Features	Application in Dissertation	Reference
STRING	PPI Network	Known and predicted protein-protein interactions, including direct and indirect associations	v11.0	24.6M proteins, 5,090 organisms, 1B interactions	Constructing PPI network for network propagation	[196]
STITCH	Chemical-Protein	Interactions between chemicals and proteins	v5.0	Interactions for over 9M proteins and chemicals	Exploring chemical-protein interaction networks	[197]
DrugBank	Drug and Phytochemical	Detailed drug data with drug-target information, including actions, mechanisms, chemical properties, interactions	5.1.8	13,000 drug entries, FDA-approved drugs, experimental drugs, nutraceuticals	Identifying FDA-approved drugs and their targets for training ML models	[198]
DrugCentral	Drug and Phytochemical	Curated database of approved drugs and drug targets, providing drug information	2021	4,500 drugs and associated targets	Enriching the dataset of FDA-approved drugs alongside DrugBank	[199]
FoodDB	Phytochemical	Information on food components and associated health effects, including phytochemicals	2020	Thousands of phytochemicals, detailed chemical properties, potential health benefits	Identifying and cataloguing phytochemicals in EVOO and other foods	[200]
DisGeNET	Disease-Related	Gene-disease associations from scientific literature and curated databases	v7.0	1.1M gene-disease associations, 21,000 genes, 30,000 diseases	Identifying genes associated with AD, for disease perturbation profiles	[211]
UniProt	Disease-Related	Comprehensive resource for protein sequence and annotation data	2021	Over 564,000 reviewed protein sequences	Refining genes associated with AD	[201]
CTD	Disease-Related	Information on chemical-gene/protein interactions and	2021	Over 46M toxicogenomic connections	Refining genes associated with AD	[202]

		their effects on diseases				
Orphanet	Disease-Related	Reference portal for information on rare diseases and drugs	2017	Information on over 6,000 rare diseases	Refining genes associated with AD	[203]
ClinGen	Disease-Related	Curated information on clinically relevant genes and variants	2015	Expert-curated gene-disease validity classifications	Refining genes associated with AD	[204]
Genomics England PanelApp	Disease-Related	Crowdsourced gene panels for disease diagnosis	2019	Gene panels for various diseases	Refining genes associated with AD	[205]
Cancer Genome Interpreter	Disease-Related	Interprets the biological and clinical relevance of tumor genomic alterations	2018	Database of cancer driver genes and mutations	Refining genes associated with AD	[206]
PsyGeNET	Disease-Related	Database of gene-disease associations focused on psychiatric disorders	2015	Associations between genes and psychiatric disorders	Refining genes associated with AD	[207]
KEGG	Pathway	Genomes, biological pathways, diseases, drugs, and chemical substances	2021	Metabolic pathways, disease pathways, drug action pathways	GSEA to understand biological processes influenced by phytochemicals	[208]
GDSC	Drug Sensitivity	Sensitivity of cancer cell lines to various anti-cancer drugs	2016	Drug response data across various cancer types and cell lines	Correlating drug sensitivity with molecular features of cell lines, aiding in validation of predictions	[209]
Recipe1MSubs	Ingredient Substitutions	Large-scale dataset of recipes and ingredient substitutions	2023	1M+ recipes, ingredients, instructions, nutrient information, 70,520 ingredient substitute pairs	Training LLMs for generating phytochemically enriched recipes, ensuring nutritional enhancement without compromising culinary quality	[210]
EVOO Phytochemical	Phytochemical	Catalog of phytochemical content in EVOO, including concentration ranges	2023	m/z, RT, molecular formula, phytochemical names, qualitative formulas, min/max concentrations in EVOO, drupes, leaves	Analysing phytochemical profiles and understanding their potential health benefits, used in ML models to predict therapeutic efficacy and generate enriched recipes	Internal Experimental Data

3. Network Machine Learning-Driven Discovery of Phytochemicals and Drugs

3.1 Abstract

Alzheimer's Disease (AD) presents a significant challenge, imposing severe human, social, and economic burdens. Prior research indicates that Extra Virgin Olive Oil (EVOO) may contribute to the prevention of cognitive decline. This dissertation introduces a network-based machine learning approach to identify bioactive phytochemicals in EVOO with the greatest potential to influence the protein network associated with the development and progression of AD. Our method achieved a balanced classification accuracy of 70.3±2.6% in a fivefold cross-validation setting, distinguishing late-stage experimental AD drugs from other clinically approved drugs. Utilizing this calibrated machine learning algorithm, we predicted the likelihood of existing drugs and known EVOO phytochemicals to exhibit similar actions to those affecting AD protein networks. The analysis highlighted ten EVOO phytochemicals with the highest predicted activity against AD: quercetin, genistein, luteolin, palmitoleate, stearic acid, apigenin, epicatechin, kaempferol, squalene, and daidzein. This *in silico* study presents an integrative framework combining artificial intelligence, analytical chemistry, and omics technologies to identify potential therapeutic agents. It provides novel insights into the mechanisms by which EVOO constituents may aid in the treatment or prevention of AD, laying a foundation for future clinical investigations.

3.2 Introduction

AD is a neurodegenerative disorder marked by a progressive decline in cognitive functions and memory. The disease is characterized by the pathological accumulation of misfolded proteins, such as amyloid beta peptides and phosphorylated tau protein, and the loss of cells in brain areas crucial for language and memory. Despite extensive research, the exact pathogenesis of AD remains elusive, and no treatments currently exist to reverse or prevent its neurocognitive decline [212]. Consequently, AD

imposes a significant burden on affected individuals, their caregivers, and the healthcare systems that support them. Given the profound human, social, and economic impact of AD, the limited understanding of its pathophysiology, and the lack of effective treatments, there is a pressing need to explore novel and unconventional therapies for its prevention and treatment [213].

Recent years have seen growing interest in the potential neuroprotective effects of key components of the Mediterranean diet. It has been associated with a lower incidence of dementia and cognitive decline in large epidemiological studies [214–216]. Beyond classic antioxidants and monounsaturated fats, this diet is also rich in the polyamine spermidine, found in high concentrations in wheat germ, soybeans, nuts, and various fruits and vegetables, and further produced by our gut microbiota. Dietary spermidine was demonstrated to mimic many of the health-span-extending effects of caloric restriction by enhancing autophagy and promoting protein deacetylation across multiple tissues. In animal models, spermidine supplementation improves cardiac function, delays cognitive decline, and reduces neuroinflammation and protein aggregation—hallmarks of AD pathology. At the cellular level, spermidine stimulates the clearance of damaged organelles and misfolded proteins via autophagy, maintains mitochondrial homeostasis, and modulates epigenetic regulators through histone deacetylase inhibition. Epidemiological studies link higher dietary spermidine intake to reduced all-cause mortality and lower incidence of neurodegenerative diseases in elderly cohorts. In the context of AD, enhanced autophagy facilitates the removal of amyloid-β and hyperphosphorylated tau, suggesting that spermidine—whether ingested directly or produced by a healthy microbiome—may be a practical dietary lever for preserving brain health without the compliance challenges of long-term caloric restriction [217]. Moreover, several interventional clinical studies have shown promising results regarding the use of olive oil and EVOO phytochemicals to improve cognitive outcomes in individuals with AD [218–220]. However, identifying the specific phytochemicals responsible for these effects remains challenging due to the complex nature of EVOO [216,219,221], and the time and cost associated with isolating enough individual constituents and examining their effects on AD progression.

Various methods have been used to identify bioactive molecules in food that may have therapeutic potential, particularly for complex diseases like AD. One prominent computational approach involves investigating bioactive phytochemicals based on their structural similarity to approved or experimentally validated drugs. This method assumes that similarities in the chemical structures of phytochemicals and drug molecules will result in similar biological effects [222]. However, this method has limitations, as even minor changes in a molecule's chemical structure can lead to significant differences in biological effects. Thus, a subtly modified molecule may not possess the same therapeutic potential as the original compound or may even cause undesirable side effects.

Another approach involves exploring bioactive phytochemicals by examining their individual protein targets, often referred to as the *one disease–one target–one drug* approach [223]. This reductionist strategy links a specific disease to a single target molecule (such as a protein or enzyme), which is then modulated by a single drug or phytochemical to treat the disease. The underlying idea is that a disease can be understood as a consequence of a single molecular dysfunction and correcting this dysfunction can lead to disease management or cure. This approach has proven successful in some cases, such as the development of targeted therapies for certain types of cancer. For example, pembrolizumab (Keytruda®) and nivolumab (Opdivo®) target the programmed cell death protein 1 (PD-1) pathway, helping the immune system combat cancer cells in various types of cancer, including melanoma, non-small cell lung cancer, and head and neck cancer [224]. However, the *one disease–one target–one drug/phytochemical* approach may be less effective for complex and multifactorial diseases involving multiple molecular pathways and genetic factors.

Recent advances in AI, combined with the growth of large-scale, multi-source data on food, drugs, and diseases, offer an opportunity to identify molecules within foods that may prevent or reverse disease using more holistic and systems-based approaches [225]. Among these, network machine learning [2,226] and graph neural networks [42] have shown promising results in predicting bioactive molecules within foods based on their ability to target disease networks, such as dysregulated genes

and protein pathways. An example of this approach is the virus–host interaction network responsible for COVID-19 [227].

In the present study, we hypothesize that a successful prevention or treatment strategy for AD should focus on the modulation of multiple biochemical networks involved in its pathogenesis, including amyloid beta production and aggregation, tau hyperphosphorylation, and neuroinflammation [208]. Building on our previous research on bioactive compounds in food against cancer [42] and COVID-19 [227], we utilized network-based machine learning methods and human protein–protein interaction data to identify EVOO-based bioactive molecules that could target AD.

3.3 Methodology

Our approach involves modelling the interactions between disease-causing proteins and those affected by drugs or EVOO phytochemicals, considering the broader network interactions. Human biological processes can be viewed as a complex network of interacting genes and proteins, with humans possessing over 20,000 protein-encoding genes and millions of protein-protein interactions. This intricate network is referred to as the interactome [228].

In a healthy state, this system can be represented as a baseline or ground level. When disease-causing proteins malfunction, they disrupt their biochemical and signalling processes, which can be represented as a deviation from the baseline. This disturbance impacts other proteins through protein-protein interactions, leading to a chain reaction known as *network propagation* [2]. Consequently, a single protein dysfunction can influence many other proteins and pathways in a cascade of perturbations. The impact on different proteins varies based on their proximity in the interaction network to the dysregulated protein.

To model these perturbations for over 20,000 proteins, we used known dysfunctional proteins linked to AD as starting points and employed a network propagation technique called *random walk with restarts* [229]. The resulting disruption levels for over 20,000 proteins form what we refer to as the disease perturbation profile of AD. Similarly, drugs or phytochemicals can interact with specific

proteins, and their effects can propagate through the network. The overall levels of protein disruption caused by these interactions are termed as drug or phytochemical perturbation profiles (see Figure 4).

The methodology utilized in this study posits that drugs and EVOO phytochemicals with anti-AD properties influence the proteins and pathways involved in the onset and progression of AD. However, these drugs and phytochemicals do not need to target the exact same genes or proteins as the disease; instead, they should affect common network regions connected to these targets. To exert an anti-AD effect, a drug or phytochemical perturbation profile should be similar (though not necessarily identical) to the AD disease perturbation profile. Conversely, a drug that does not affect AD (non-anti-AD) would be expected to have a perturbation profile with minimal overlap with the AD disease perturbation profile.

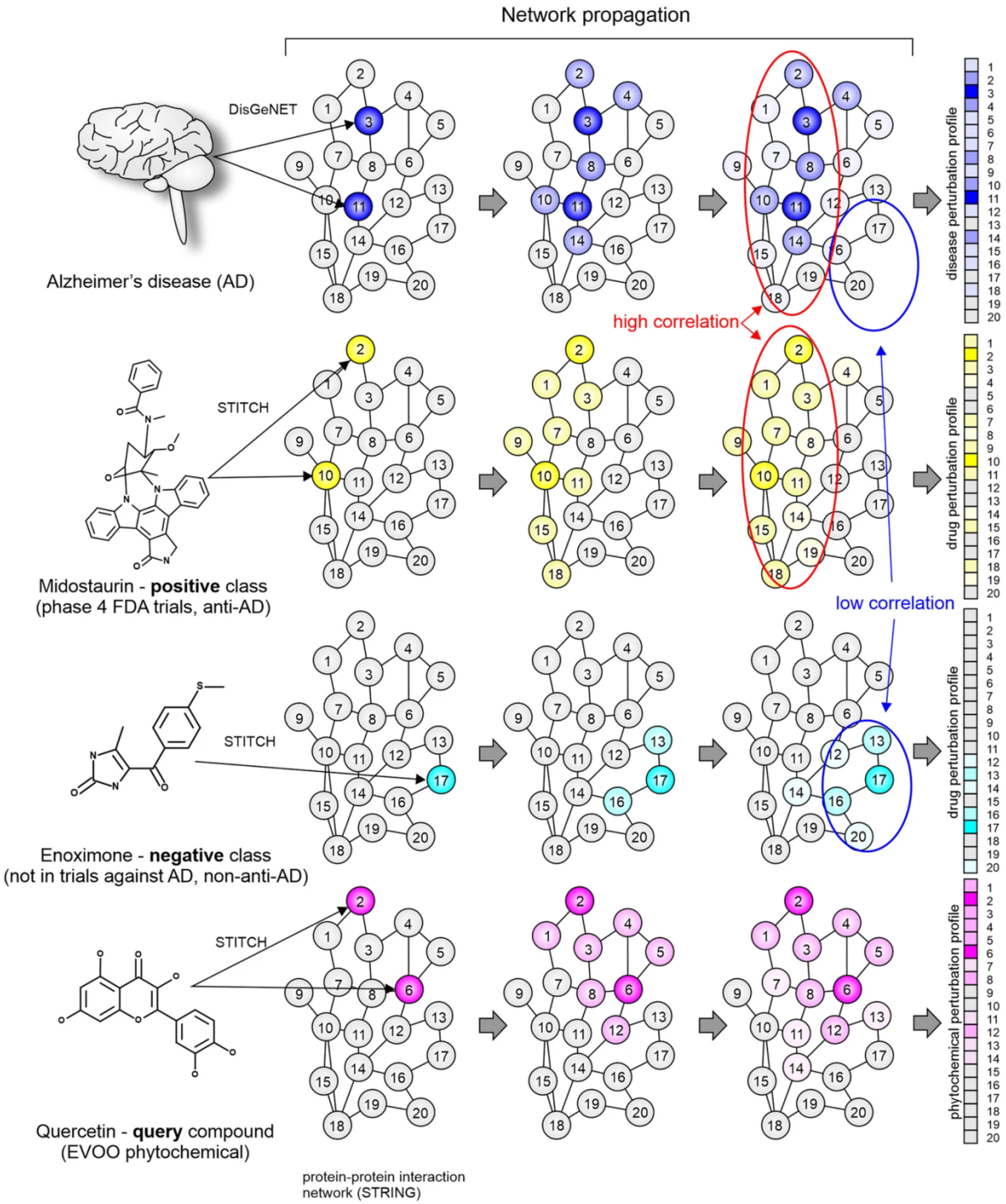


Figure 4 Network Propagation of Protein Perturbation Profiles for Phytochemicals, Drugs and AD. The core principles of network propagation used to generate protein perturbation profiles for diseases, drugs, and EVOO phytochemicals are illustrated. The human interactome is depicted as a graph with interconnected nodes (numbered circles), where each node represents a protein, and the edges represent protein-protein interactions. Unique protein identifiers (numbers) allow for quick visual cross-referencing. Proteins directly affected by drugs or EVOO compounds were identified using the STITCH database, and proteins directly dysregulated in AD were identified using the DisGeNET database (indicated by arrows on the left). The extent of protein perturbation is indicated by colour intensity: blue for AD, yellow for an anti-AD drug in late-stage

clinical trials (e.g., midostaurin), cyan for a non-anti-AD drug (e.g., enoximone), and magenta for the query EVOO compound (e.g., quercetin). The network propagation section shows how perturbations spread through protein-protein connections, with colour intensity reflecting the level of perturbation. Highly correlated profile regions are marked with red ovals, while poorly correlated regions are shown in blue ovals.

To determine the necessary level of similarity for identifying unique drugs or phytochemicals with anti-AD properties, we trained a machine learning predictive model using known anti-AD drugs (e.g., midostaurin) as the positive class and non-anti-AD drugs (e.g., enoximone) as the negative class (Figure 4, 5 and *Section 3.3.4*). This approach allows us to identify potential candidates by comparing their perturbation profiles to the AD disease profile.

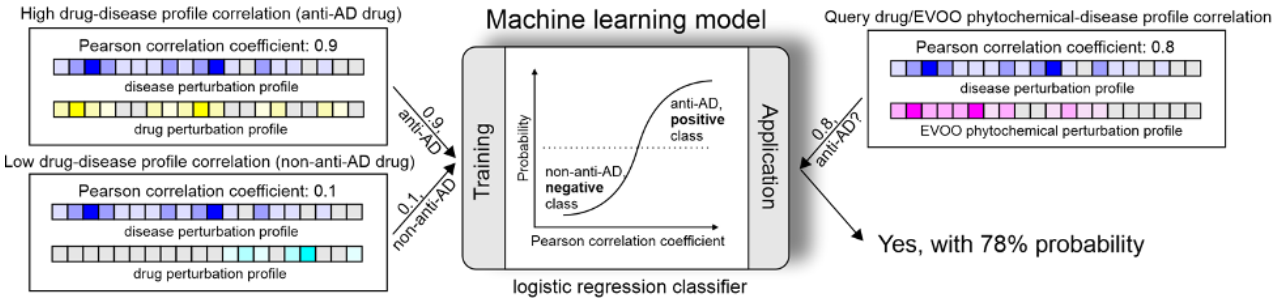


Figure 5 Machine Learning–Driven Prediction of Drugs and EVOO Phytochemicals with Anti-AD Profile Similarity. Machine learning-based approach for predicting EVOO phytochemicals that target the AD interactome similarly to advanced-stage anti-AD drugs in clinical trials is represented. A logistic regression classifier was trained to differentiate between anti-AD drugs (positive class) and non-anti-AD drugs (negative class) based on profile correlations. This classifier was then used to predict the likelihood of EVOO phytochemicals exhibiting anti-AD properties [230]. The extent of protein perturbation is indicated by colour intensity: blue for AD, yellow for an anti-AD drug in late-stage clinical trials, cyan for a non-anti-AD drug, and magenta for the query EVOO compound.

3.3.1 Alzheimer's Disease Protein-Encoding Gene Selection

In this study, we identified proteins involved in AD using the comprehensive public dataset DisGeNET [231]. From this dataset, 3,397 human protein-encoding genes were initially identified and then manually refined using expert-curated databases to ensure high-quality data. The databases used for refinement included UniProt [201], CTD [202], Orphanet [203], Clinical Genome Resource [204], Genomics England PanelApp [205], Cancer Genome Interpreter [206], and the Psychiatric Disorders Gene Association Network [207].

Our focus was on proteins that influence the onset or progression of AD, deliberately excluding those marked as therapeutic to avoid circularity in our analysis. This selection process led to the identification of 101 *key* protein-encoding genes strongly associated with AD risk and progression. This way genes with low Gene–Disease Association score backed primarily by text mining or lacking disease relevance annotations were removed. These *key* genes were then mapped onto the STRING database [42], which provides a list of known protein–protein interactions for various organisms, along with their confidence scores (0–999). This step resulted in the identification of 73 *key* AD proteins (*Appendix 8.1*). This reduction in number is because not all 101 curated AD-associated genes had high-confidence or experimentally validated protein–protein interaction data available in the STRING database. Genes without mapped interactions or with incomplete linkage in the interactome were excluded from downstream propagation to ensure robust and interpretable network modelling. Protein–protein interactions were subsequently filtered by thresholding the confidence scores as part of the machine learning process (described below in the *Section 3.3.3*).

3.3.2 Extra Virgin Olive Oil Phytochemical Selection

We have developed a comprehensive list of 782 phytochemicals found in extra virgin olive products, including EVOO, table olives, and olive leaves (*Appendix 8.2*). This list encompasses plant-based chemicals while excluding metabolic intermediates, such as glucose [232]. From these 782 phytochemicals, we selected only those unique to EVOO or found in high concentrations [233]. This selection process involved an extensive literature review to confirm the presence of these chemicals in the olive oil matrix, resulting in the identification of 117 EVOO *marker* phytochemicals.

To ensure consistency and avoid ambiguity in compound identification, we standardized all phytochemical entries using their InChIKeys. Specifically, we considered only the first 14 characters of each InChIKey (the connectivity layer), thereby treating structural and stereoisomers as equivalent. This was done under the rationale that such isomers often possess similar physicochemical properties and target overlapping molecular pathways. Consequently, phytochemicals differing only by stereochemistry or isotopic labelling were grouped and analysed as a single compound.

The DrugBank [198] and DrugCentral [199] databases were utilized to compile a list of experimental and FDA-approved drugs, while the FoodDB [200] database was used to compile a list of known phytochemicals, as described previously [42]. The STITCH database [197] was then employed to identify human proteins with which these molecules could interact, narrowing down the list to 67 EVOO marker phytochemicals. The STITCH database provides a confidence score (0–999) for each chemical–protein interaction, which quantifies the likelihood of a true interaction based on available evidence. For this study, we applied a fixed confidence cutoff of 200, meaning that only interactions with a score of 200 or higher were considered. This threshold was chosen to ensure that the interactions studied were more likely to be biologically relevant and supported by experimental evidence rather than solely predicted computationally.

As with phytochemicals, all FDA-approved and experimental drug entries were standardized using InChIKeys. Only the connectivity block of each InChIKey was used to unify compounds that share core atomic connectivity, effectively grouping stereoisomers, conformers, and isotopic variants. This ensured coherent learning signals across the dataset and minimized overrepresentation of structurally redundant molecules during training and prediction.

3.3.3 Network Propagation

After compiling the lists of drugs and EVOO phytochemicals, along with their impacted genes/proteins, and identifying the dysfunctional proteins associated with AD, we utilized our previously developed network propagator tool (used in cancer [42] and COVID-19 [227] projects) to conduct a RWR [229]. This step enabled us to generate drug, EVOO phytochemical, and disease perturbation profiles based on the STRING protein–protein network, which consists of 1,048,574 connections between 20,256 human proteins (matched between the STITCH and STRING databases).

The random walk propagation was executed independently for drugs/EVOO phytochemicals and the AD interactomes, using a range of restart probability values and several protein–protein connection thresholds (STRING protein connection thresholds of 400, 600, and 800; restart probability parameter c ranging from 0.0001 to 1.0). This range was selected based on our previous experience

with random walk propagation in applications to cancer and COVID-19 [42,227]. Pearson correlation coefficients were calculated to assess the extent of overlap between the AD perturbation profile and each of the drug or EVOO phytochemical perturbation profiles. For each molecule–disease pair, a total of 2,160 Pearson correlation coefficients were calculated with different propagation settings, allowing for a comprehensive assessment of molecule–disease perturbation similarities.

3.3.4 Model Training and Application

Leveraging molecule–disease perturbation similarities, we trained machine learning models to predict which FDA-approved drugs, among those available in DrugBank and DrugCentral (N = 1,802), are in advanced stages (3 and 4) of FDA clinical trials (N = 32) for the treatment of AD. These 32 drugs, approved for other disease states, are being re-evaluated as potential therapeutic agents for AD. A logistic regression classifier, based on the scikit-learn Python library for machine learning [234], was fitted to the calculated Pearson correlation coefficients for each propagation setting from the previous step. The best settings were selected using fivefold stratified cross-validation. This approach is akin to the one we used to predict phytochemicals with anticancer properties [42]. The overall predictive capacity of this approach was validated using 25 repeats of fivefold stratified cross-validation, with the optimal predictive model/propagation setting being selected by the internal fivefold stratified cross-validation.

A final set of models was identified that exhibited individual balanced accuracies higher than the cross-validation one, resulting in 64 models. These models were then used to predict the probabilities of EVOO phytochemicals exhibiting anti-AD properties, based on the correlation between their perturbation profiles and the perturbation profile of AD. Predicted probabilities were averaged among the selected models to provide a more robust ensemble-based probability prediction, referred to as the correlation probability (%).

3.3.5 Pathway Analysis

Next, we determined the biological pathways more strongly perturbed in common between each top-scoring molecule and Alzheimer's Disease. To do this, we ran GSEA Prerank against the KEGG 2023 gene sets for each of the ten FDA-approved drugs and ten EVOO phytochemicals with highest correlation probability.

First, for each molecule we built a pre-ranked gene list by elementwise multiplying its RWR perturbation profile with the AD perturbation profile and sorting genes by the resulting scores (descending). That ranked list was then supplied to GSEA Prerank together with the full collection of KEGG 2023 pathways. GSEA performed 1,000 random permutations of gene labels to generate a null distribution of enrichment scores and computed a nominal p-value for each pathway (the fraction of permuted scores ≥ the observed score). Finally, we ranked pathways by nominal p-value and retained only those with $p < 0.001$ for reporting.

KEGG pathways are curated gene-set definitions spanning metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, and human diseases [196]. Here, they allow us to interpret which network modules—rather than individual genes—are jointly influenced by AD perturbations and each candidate molecule.

All data and Python scripts for the modelling, analysis, and figures are available in this Bitbucket repository.

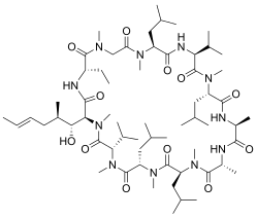
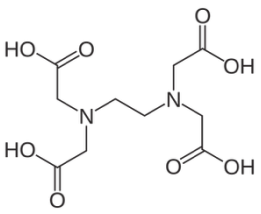
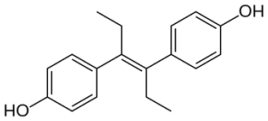
3.4 Results

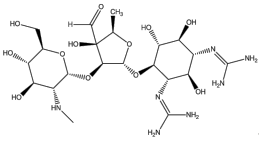
Our model allowed us to predict with 70.3±2.6% balanced classification accuracy which previously FDA-approved drugs were in phase 3 and 4 trials for AD specifically, compared to all other FDA-approved drugs not in AD trials. Here, balanced accuracy is defined as the arithmetic mean of sensitivity (true positive rate – 40.6%) and specificity (true negative rate – ~100%) across the two classes. The resulting 64 models were used to score the EVOO phytochemicals, and the probabilities of these phytochemicals being similar to the drugs in FDA phase 3 and 4 trials were averaged to produce the final consensus

prediction. EVOO phytochemicals with the highest probability of being like compounds in FDA trials were considered most likely to be biologically active.

Using our model, FDA-approved drugs for other diseases (i.e., non-anti-AD drugs) were re-evaluated and ranked according to their correlation probability of affecting AD (Table 3). From all the re-evaluated drugs, Cyclosporine, Edetic Acid, Diethylstilbesterol, Nitric Oxide and Streptomycin were ranked the highest.

Table 3 Top FDA-Approved Drugs Not Undergoing AD Clinical Trials Ranked by Our Model.

Rank [a]	Approved Drug	Correlation Probability (%) [b]	Putative Mechanism of Action (Literature) [c]	Putative Pathways Affected [d]	Molecular Structure
1	Cyclosporine	78.8	Inhibits calcineurin, while reducing inflammation and oxidative stress (↑) [235]	**Alzheimer's Disease**, Olfactory Transduction, Insulin Signalling Pathway	
2	Edetic Acid	78.4	Chelates metals such as aluminium, which may be involved in Alzheimer's pathology (↓) [236]	**Alzheimer's Disease**, Olfactory Transduction	
3	Diethylstilbesterol	78.0	Binds to and activates oestrogen receptors, which have neuroprotective effects (↑) [237]	**Alzheimer's Disease**, Olfactory Transduction	
4	Nitric Oxide	77.9	Regulates cerebral blood flow and modulates neurotransmitter release, improving cognitive function (↑) [238]	**Alzheimer's Disease**, Olfactory Transduction, Long Term Potentiation, Insulin Signalling Pathway, GnRH Signalling Pathway, Vascular Smooth Muscle Contraction,	

				Neurotrophins Signalling Pathway	
5	Streptomycin	77.6	Inhibits protein synthesis by binding to the bacterial ribosome, potentially reducing the risk of bacterial infections that may exacerbate Alzheimer's pathology (↑) [239]	**Alzheimer's Disease**, Olfactory Transduction, Phosphatidylinositol Signalling System, Insulin Signalling Pathway	
Mean		**77.7**			

a. Drugs are ranked from highest to lowest based on their correlation probability with AD.

b. Logistic regression was used to calculate the correlation probability, predicting the likelihood of a drug impacting AD.

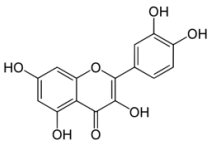
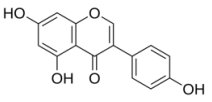
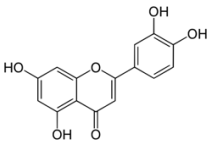
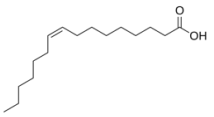
c. Arrows at the end of each identified mechanism of action indicate the overall evidence found positive (↑), equivocal (≈), or negative (↓) towards preventing or halting the onset and/or progression of AD.

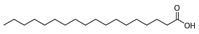
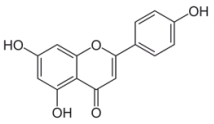
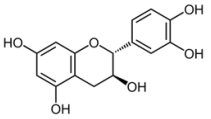
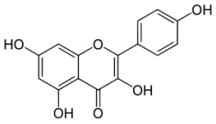
d. Only statistically significant KEGG 2023 pathways (GSEA nominal p-value < 0.001) are listed for each compound, ordered by increasing p-value. The Alzheimer's Disease pathway corresponds to KEGG hsa05010. Other relevant pathways include hsa04740 (Olfactory Transduction), hsa04910 (Insulin Signalling Pathway), hsa04720 (Long-Term Potentiation), hsa04912 (GnRH Signalling Pathway), hsa04270 (Vascular Smooth Muscle Contraction), hsa04722 (Neurotrophins Signalling Pathway), and hsa04070 (Phosphatidylinositol Signalling System).

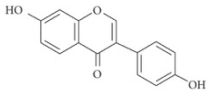
The most common pathways targeted by these drugs included *Alzheimer's disease* (10), *Olfactory Transduction* (10), *Insulin Signalling* (5), *Phosphatidylinositol Signalling System* (2), and *Long-Term Potentiation* (1). The numbers in parentheses indicate the frequency with which each pathway was identified for these drugs with a statistically significant p value for involvement with AD.

Of the 67 EVOO marker phytochemicals (*Appendix 8.3*), Quercetin, Genistein, Luteolin, Palmitoleate, Stearic Acid, Apigenin, Epicatechin, Kaempferol, Squalene, and Daidzein had the highest probability of interfering with AD (Table 4).

Table 4 Top Ten EVOO Phytochemicals Predicted to Influence AD Onset or Progression.

Rank [a]	EVOO Marker	Correlation Probability (%) [b]	Putative Mechanism of Action (Literature) [c]	Putative Pathways Affected [d]	Molecular Structure
1	Quercetin	78.0	Decreases oxidative stress, modulates cytokines, inhibits amyloid beta aggregation, and decreases tau phosphorylation [240–242]. (↑)	**Alzheimer's Disease**, Olfactory Transduction, Phosphatidylinositol Signalling System, Insulin Signalling Pathway	
2	Genistein	75.5	Has antioxidant activity and ameliorates amyloid-beta induced neurotoxicity induced by amyloid beta [243–245]. (↑)	**Alzheimer's Disease**, Olfactory Transduction, Insulin Signalling Pathway, Vascular Smooth Muscle Contraction	
3	Luteolin	73.9	Modulates insulin resistance systemically and in the central nervous system and alters amyloid beta deposition [242]. (↑)	Olfactory Transduction	
4	Palmitoleate	69.5	Preserves endothelial function, prevents endoplasmic reticulum stress, reduces lipogenesis and fatty acid desaturation in adipocytes, and suppresses cytokine production [246]. Decreases NF-KB nuclear translocation *via* the stimulation of PPARγ and the phosphorylation of AMPK, which increases MGL2, IL-10, TGFß1, and MRC1 [243]. (≈)	**Alzheimer's Disease**, Olfactory Transduction, Insulin Signalling Pathway, Vascular Smooth Muscle Contraction	

5	Stearic Acid	67.5	No consistent evidence.	**Alzheimer's Disease**, Olfactory Transduction	
6	Apigenin	67.4	Has anti-inflammatory and antioxidative activity and downregulates BACE1 enzyme [247–249]. (↑)	**Alzheimer's disease**, Olfactory Transduction, Long-Term Potentiation, Phosphatidylinositol Signalling System, Glioma, Insulin Signalling Pathway, GnRH Signalling Pathway, Neurotrophins Signalling Pathway	
7	Epicatechin	66.4	Reduces amyloid beta concentrations in brain and serum and sequesters by-products of oxidation [250,251]. (↑)	**Alzheimer's disease**, Olfactory Transduction, Long-Term Potentiation, Phosphatidylinositol Signalling System, Insulin Signalling Pathway, Calcium Signalling Pathway	
8	Kaempferol	65.1	Attenuates oxidative stress and inflammation through multiple mechanisms and delays memory loss [252,253]. (↑)	**Alzheimer's disease**, Olfactory Transduction, Phosphatidylinositol Signalling System, Insulin Signalling Pathway, Vascular Smooth Muscle Contraction	
9	Squalene	63.7	No consistent evidence.	**Alzheimer's disease**, Olfactory Transduction,	

				Maturity Onset Diabetes of the Young	
10	Daidzein	62.1	Decreases amyloid beta aggregation and amyloid beta induced neuronal death and increase hippocampal neuronal cell proliferation [254,255]. (↑)	**Alzheimer's disease**, Olfactory Transduction, Glioma, Phosphatidylinositol Signalling System, Long-Term Potentiation, Insulin Signalling Pathway, mTOR Signalling Pathway	
Mean		**68.9**			

a. Phytochemicals are ranked from highest to lowest based on their correlation probability with AD.

b. Logistic regression was used to calculate the correlation probability, predicting the likelihood of an EVOO marker phytochemical impacting AD.

c. Arrows at the end of each identified mechanism of action indicate the overall evidence found positive (↑), equivocal (≈), or negative (↓) towards preventing or halting the onset and/or progression of AD.

d. Statistically significant KEGG 2023 pathways (GSEA nominal p-value < 0.001) associated with each EVOO phytochemical are shown, ordered by increasing p-value. The Alzheimer's Disease pathway corresponds to KEGG hsa05010. Additional enriched pathways include hsa04740 (Olfactory Transduction), hsa04910 (Insulin Signalling Pathway), hsa04270 (Vascular Smooth Muscle Contraction), hsa04720 (Long-Term Potentiation), hsa04070 (Phosphatidylinositol Signalling System), hsa04912 (GnRH Signalling Pathway), hsa04722 (Neurotrophins Signalling Pathway), hsa04020 (Calcium Signalling Pathway), hsa05214 (Glioma), hsa04150 (mTOR Signalling Pathway), and hsa04950 (Maturity Onset Diabetes of the Young).

The most common pathways predicted to be affected by these phytochemicals include *Olfactory Transductio*n (10), *Alzheimer's disease* (9), Insulin Signalling Pathway (7), *Phosphatidylinositol Signalling System* (5), and *Vascular Smooth Muscle Contraction* (3). The numbers in parentheses indicate the frequency with which each pathway was identified for these EVOO phytochemicals with a statistically significant p value for involvement with AD. Examples of select EVOO phytochemicals and their connections with the most common AD pathways and involved proteins in our analysis are depicted in Figure 6.

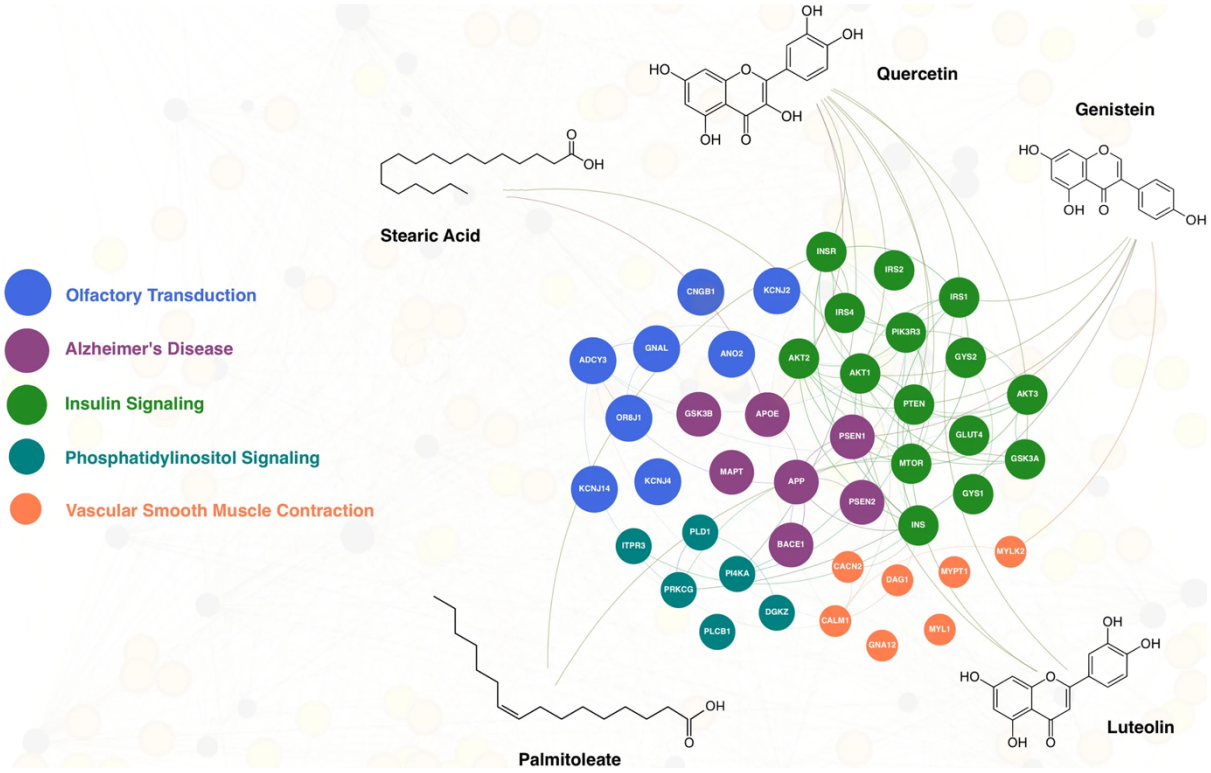


Figure 6 Predicted EVOO Phytochemicals Targeting Dysregulated Pathways in AD. The targeted pathways include *Olfactory Transduction*, *Alzheimer's Disease*, *Insulin Signalling*, *Phosphatidylinositol Signalling* and *Vascular Smooth Muscle Contraction*. The key protein-encoding genes associated with each specific pathway are color-coded for easy identification. The size of each gene/protein node is proportional to the frequency with which the related pathway is targeted by the model's predicted EVOO phytochemicals listed in Table 4. Chemical structures of the highest ranked phytochemicals—Quercetin, Genistein, Luteolin, Stearic Acid, and Palmitoleate—are shown. Only experimentally validated protein-protein interactions with a STRING score greater than 800 are emphasized [230].

3.5 Discussion

In this study, we utilized machine learning to identify FDA-approved drugs and EVOO phytochemicals likely to interact with genes and proteins associated with AD pathophysiology. The model was initially trained to distinguish between FDA-approved drugs in phases 3 or 4 clinical trials for AD treatment (positive class) and FDA-approved drugs used for other diseases unrelated to AD (negative class). This trained model was then employed to predict which EVOO phytochemicals might influence the development or progression of AD.

The correlation probability scores assigned to the EVOO phytochemicals were comparable to those of the FDA-approved phase 3 or 4 AD drugs, suggesting that these phytochemicals may interact with the same biological pathways as the approved medications.

Machine learning applications for identifying drug targets using existing databases are still in their early stages, and the accuracy of predictions largely depends on the quality of the databases employed, which are continuously updated as new information becomes available. The results obtained in our study using network propagation (RWR) support the validity of this approach.

The top-ranked FDA-approved drugs not currently undergoing AD clinical trials, as identified by our logistic regression model (Table 3), display mechanistic features that plausibly intersect with AD pathophysiology, though their translational value varies substantially. Cyclosporine, the highest-ranked candidate (78.8%), is supported by literature showing its ability to inhibit calcineurin and reduce mitochondrial dysfunction, oxidative stress, and neuroinflammation—mechanisms directly implicated in AD progression. However, its immunosuppressive profile may limit chronic use unless repurposed at sub-immunosuppressive doses [256]. Edetic acid (EDTA) emerged as a metal-chelating agent potentially mitigating aluminum-related neurotoxicity, yet its systemic toxicity and use in acute heavy metal poisoning make it an unlikely direct therapeutic candidate without extensive reformulation [257]. Diethylstilbestrol, a synthetic estrogen, shares a structural pharmacophore with several neuroprotective flavonoids (e.g., Genistein) due to its dual phenolic rings connected via a short linker. Despite suggestive estrogen-receptor-mediated neuroprotection and some links to AD pathophysiology, its well-established carcinogenicity and endocrine-disrupting effects preclude repurposing [258]. Nitric oxide (NO), another high-ranking compound, plays a dual role in the CNS, with neuroprotective functions via vasodilation and neurotransmitter modulation, but also neurotoxic effects at high concentrations. Thus, targeting NO signalling pathways may hold promise only under tightly controlled pharmacodynamic conditions [259]. Streptomycin, though primarily an antibiotic, has been hypothesized to attenuate AD-related inflammation indirectly by limiting bacterial triggers, yet its ototoxicity and limited BBB penetration constrain its practical applicability [260].

Several identified EVOO phytochemicals, such as quercetin, have previously demonstrated positive effects in preventing the development or progression of AD (see discussion below). For instance, quercetin—the EVOO phytochemical with the highest correlation probability in our study—

has been evaluated as a neuroprotective agent in AD [241]. It is believed to confer protection against AD by decreasing oxidative stress, modulating cytokines, inhibiting amyloid-beta aggregation, and reducing tau phosphorylation [240–242]. Similarly, genistein, apigenin, catechin (also known as epicatechin), and kaempferol have been identified as potential therapeutic agents for AD [243–245,247–253]. Genistein improved cognitive performance in a 12-month, double-blind, placebo-controlled trial of prodromal AD patients receiving 120 mg/day, showing significant gains in verbal memory and Rey Copy scores [261]. Epicatechin protected neurons from Aβ-induced toxicity and oxidative damage in rodent hippocampal cultures [262–264]. Kaempferol, studied in both *in vitro* and *in vivo* models, exerts anti-amyloidogenic and anti-inflammatory effects, improving learning and memory in AD mouse models [265]. In contrast, the flavonoids luteolin and daidzein are novel agents that have not yet been investigated as therapeutics in AD. Luteolin alleviated cognitive deficits and reduced amyloid pathology in transgenic AD mice through anti-inflammatory and antioxidant mechanisms [266]. Palmitoleate and stearic acid have undergone limited evaluation regarding their biological actions in the central nervous system [267–270]. Palmitoleate, a monounsaturated fatty acid, activates PPAR-α—a regulator of lipid metabolism implicated in synaptic plasticity and cognitive function—and has been shown to reduce AD-like pathology and improve memory in rodent studies [271,272]. Squalene, by contrast, lacks classical drug-like pharmacophores—being a hydrophobic isoprenoid—and although its metabolism has been linked to AD pathogenesis [273], it remains untested as a direct therapeutic candidate [274–276].

We identified eleven different KEGG pathways targeted by the ten EVOO marker phytochemicals most likely to interact with AD (Table 4). The pathways with the highest number of occurrences were Olfactory Transduction, Alzheimer's disease, Insulin Signalling Pathway, Phosphatidylinositol Signalling System, and Vascular Smooth Muscle Contraction. The Alzheimer's disease pathway encompasses key pathological events such as the accumulation of beta-amyloid peptides, formation of neurofibrillary tangles, oxidative stress, inflammation, and neuronal apoptosis [208]. While Olfactory Transduction is not directly linked to AD, evidence suggests that changes in

olfactory function may serve as an early marker of the disease. Individuals affected by AD often exhibit a reduced sense of smell [277,278], and olfactory dysfunction may precede cognitive symptoms by several years [279]. Additionally, the accumulation of amyloid-beta protein—a hallmark of AD—may disrupt olfactory function by interfering with the olfactory transduction pathway [280]. Thus, Olfactory Transduction in our analysis likely reflects an early consequence of AD pathology rather than a causal driver, although interventions that preserve olfactory signalling could still mitigate early symptoms.

The Insulin Signalling Pathway is implicated in several aspects of AD pathology, including amyloid-beta and tau metabolism, neuroinflammation, and synaptic plasticity [281]. The Phosphatidylinositol Signalling System plays a crucial role in regulating various cellular processes such as cell growth, proliferation, and survival [282]. In AD, this system has been involved in the regulation of amyloid-beta production, neuroinflammation, and synaptic dysfunction; thus, its dysregulation may contribute to the pathological processes underlying AD [282]. Vascular Smooth Muscle Contraction is essential for regulating blood flow and pressure. It is implicated in AD progression, as cerebrovascular dysfunction can lead to impaired blood flow, reduced cerebral oxygen supply, and exacerbation of AD symptoms and neuronal damage [283]. Amyloid-beta accumulation can also disrupt blood vessels and impair vascular smooth muscle function, limiting their responsiveness to changes in blood flow and pressure, thereby increasing the risk of hypoperfusion and neurovascular damage [284].

The analytical approach employed in this study has several limitations, many stemming from the nature of the datasets underlying our inputs. Our analytical approach relies on the STRING and STITCH resources to seed both disease and molecule - perturbation profiles. STRING v9.1 achieves approximately 80% precision and 60% recall against curated protein–protein interaction gold standards, and STITCH shows comparable performance when validated against known ligand–binding assays [285,286]. To mitigate noise from computational predictions, we filtered STRING interactions at confidence - score thresholds of 400, 600, and 800 and applied a STITCH cutoff of ≥200, then reran random - walk propagation across this grid of thresholds and restart probabilities. This confirmed that our top candidates and model performance were stable across settings. Nonetheless, both databases

remain incomplete and include interactions of varying reproducibility; some AD-relevant proteins or true chemical–protein edges may still be missing, while spurious links could propagate through the network. As STRING and STITCH continue to incorporate new experimental data and refined confidence metrics, future iterations of our pipeline—ideally incorporating high-throughput validation datasets—should further enhance the robustness and interpretability of our predictions.

Additionally to the database completeness and interaction confidence, it is also important to recognize that even a fully accurate interactome may overestimate the impact of single‐node perturbations. In many cases, propagation of a signal through a network is constrained by the availability of essential cofactors or substrates, and biological systems often employ redundant pathways or compensatory mechanisms that buffer against the loss or modulation of individual proteins. Indeed, studies have shown that as many as 90% of gene‐knockout mouse models fail to exhibit overt phenotypes precisely because of such buffering effects [287]. This intrinsic resilience means that our network‐based predictions may identify putative targets whose actual *in vivo* influence is muted.

Moreover, we analysed only a subset of AD expert-curated genes; it is possible that the 73 genes selected do not fully characterize the disease process. Our analysis also does not determine whether the identified EVOO phytochemicals are necessarily effective against AD or whether their effects are anti-AD or pro-AD, although they are more likely to be as effective as the phase 3 and 4 FDA-approved drugs currently undergoing AD clinical trials than any other molecules found in EVOO. It is important to note that the number of positive class drugs (i.e., FDA-approved drugs undergoing AD trials) was much smaller (32) than the number of negative class drugs (i.e., FDA-approved non-anti-AD drugs) used for model training (1,745). Such an unbalanced dataset increases the likelihood of bias toward the negative class, potentially leading to underestimation of EVOO phytochemicals that may influence AD. With the identification of more anti-AD drugs and updated databases in the future, it should be possible to further enhance the predictive capacity of the current model.

We hope that our *in silico* findings will inspire further studies to experimentally address the existing limitations, validate our results, and fully evaluate the role of EVOO in the prevention of AD.

3.6 Conclusion

Diet and lifestyle are well-established influencers of health, and machine learning offers a novel, cost-effective approach for evaluating the potential health benefits of individual phytochemicals found in EVOO. We present a methodology that combines AI, analytical chemistry, and omics studies to investigate the interactions of phytochemicals with disease-related pathways. This integrated approach can lead to the discovery of novel therapeutic entities within natural products that contain a diverse mixture of phytochemicals. We identified several individual EVOO phytochemicals with a high likelihood of interfering with AD, including quercetin and genistein, which have already shown promising effects on AD pathogenesis. Others, such as luteolin, warrant further *in vitro* and *in vivo* studies to validate their potential. The full predictive utility of our machine learning approach will be realized through these follow-up studies. While the current study provides insights into how EVOO may help treat or prevent AD, this same methodology can be extended to identify EVOO phytochemicals or other food constituents that may treat other diseases, such as hypertension or dyslipidaemia.

4. Experimental Validation of Pharmacodynamic Drug-Drug and Food-Drug Synergies

4.1 Abstract

In oncology - specifically in the treatment of human lung and colorectal cancers - pharmacodynamic (PD) interactions between drugs and food-derived molecules can significantly influence the efficacy and safety of medications, leading to substantial clinical outcomes. While computational methods exist for identifying PD drug-drug interactions, PD drug-food interactions remain largely unexplored due to a lack of comprehensive *in vitro* datasets and the complexity introduced by the vast number of food-derived compounds. Building on the network-propagation developed in Chapter 3 to predict EVOO phytochemical interactions with AD molecular networks, this study integrates computational predictions with experimental validation to explore their interactions with drugs. We employed a GNN-based model trained on experimentally validated drug–drug PD interaction datasets to predict synergistic combinations. The model learns the molecular and topological features of drug targets and their positions in protein–protein interaction networks, enabling it to infer potential synergy by propagating pharmacodynamic influence across cancer molecular networks. This architecture was repurposed to identify candidate food-derived molecules likely to interact synergistically with drugs based on shared target neighbourhoods and inferred biological effects. Then, performed monotherapy assays and IC50 calculations to determine the efficacy of each of the compounds from the previous combinations. Finally, tested experimentally using human lung (A549) and colorectal (HCT116) cancer cell lines some of the combinatorial predictions. Experimental results were consistent with the computational predictions in 7 out of 11, and 5 out of 9 tested drug-drug and food-drug combinations, respectively. These findings demonstrate that models trained on experimental PD drug-drug interaction

datasets can effectively identify PD drug-drug and food-drug interactions. This work underscores the potential of network machine learning approaches to enhance our understanding of clinically relevant PD interactions, which may lead to improved drug efficacy and reduced toxicity.

4.2 Introduction

The intake of food can significantly influence drug activity by altering PK or PD properties, impacting the safety and efficacy of oral pharmacotherapy [288,289]. PK food-drug interactions modify the absorption, distribution, metabolism, and excretion of drugs due to changes in metabolic activity induced by food intake or dietary supplements [290]. These alterations can occur at the cellular level, such as modifications in cell metabolic processes and membrane transporters, or at the organ level, affecting gastric emptying kinetics, luminal bile salt concentrations, or hepatic perfusion [291]. On the other hand, PD food-drug interactions alter the pharmacological effects of drugs, resulting in either attenuation or enhancement of drug efficacy and toxicity [292]. The mechanisms and outcomes of each PD interaction are specific to the involved food molecule and drug, and their discovery has often been serendipitous [293].

Some food molecules, such as naringenin - a major flavonoid found in grapefruit - are known to modulate drug bioavailability through complex, sometimes opposing mechanisms. On one hand, naringenin can increase the bioavailability of drugs metabolized by cytochrome P450 3A4 (CYP3A4) by acting as a potent CYP3A4 inhibitor [294]. Inhibition of this metabolic enzyme reduces the first-pass metabolism of many drugs, leading to higher systemic concentrations. For example, co-administration of grapefruit juice has been shown to increase the plasma concentration of drugs such as felodipine, nifedipine, and simvastatin, all of which are CYP3A4 substrates [294]. On the other hand, naringin - the glycosylated precursor of naringenin - can decrease drug absorption by inhibiting intestinal uptake transporters, particularly the organic anion transporting polypeptide 1A2 (OATP1A2) [295]. OATP1A2 facilitates the active uptake of various drugs across the intestinal epithelium; its inhibition can lead to a significant reduction in the systemic availability of drugs such as fexofenadine and atenolol [295]. This dual activity - inhibition of efflux metabolism and inhibition of uptake transport - means that the net

effect of naringenin or grapefruit-derived compounds on drug bioavailability is highly context-dependent, varying with the specific drug, its metabolic pathway, transporter involvement, and site of absorption. These contradictory effects underscore the importance of characterizing food-drug interactions in a mechanistically resolved, drug-specific manner rather than assuming unidirectional outcomes. But there are also known food molecules for having the potential to increase drug bioavailability in some cases. For instance, curcumin, a polyphenolic compound extracted from turmeric (*Curcuma longa*), has been shown to enhance the efficacy of chemotherapy drugs such as doxorubicin. This effect is partially explained by curcumin's ability to inhibit several cytochrome P450 (CYP) enzymes involved in drug metabolism. Specifically, curcuminoids have been shown to inhibit CYP1A2, CYP2B6, CYP2C9, CYP2D6, and CYP3A4 - the latter being the primary isoform responsible for the oxidative metabolism of doxorubicin in the liver. By inhibiting CYP3A4, curcumin reduces the metabolic clearance of doxorubicin, leading to increased systemic exposure and prolonged circulation time, thereby enhancing its cytotoxic efficacy against tumor cells [296]. Similarly, epigallocatechin gallate (EGCG) from green tea can enhance the therapeutic effects of anticancer drugs like doxorubicin and cisplatin [297,298]. Quercetin, found in onions and apples, is another food molecule that can potentially enhance drug efficacy through its antioxidant properties [299,300]. Resveratrol, a compound found in grapes and red wine, which has been reported to similarly enhance the effects of the same agents [301,302]. Not limited to chemotherapy drugs, some foods are also known for their positive interactions with anti-inflammatory and antibiotic drugs. Foods rich in omega-3 fatty acids, such as fish oil, have been found to enhance the effects of anti-inflammatory drugs [172,303]. Garlic, containing allicin, has shown potential in enhancing the efficacy of antibiotics and reducing their required dosage [304,305]. While these interactions hold promise for improving therapeutic outcomes, they also highlight the complexity of drug-food interactions. Understanding these interactions can lead to more effective and safer treatment protocols, integrating dietary recommendations with pharmacotherapy [306,307].

Despite the potential significance, the specificity of PD interactions and the vast number of food molecules have limited large-scale studies of PD food-drug interactions. However, recent efforts to

compile comprehensive food chemical composition databases have enabled *in silico* studies of these interactions [308,309]. For instance, food molecules that affect the pharmacokinetics or pharmacodynamics of drugs were identified, noting that genes targeted by drugs for cancer, cardiovascular, and neurological diseases are also highly targeted by food molecules [308]. A food-disease interaction network was constructed using publicly available gene expression data, identifying beneficial and harmful food-disease pairs [309]. In these studies, interactions were assumed to exist if both agents affected the same set of targets or genes. However, the real-world implications of these interactions, especially in the context of drug-binding mechanisms and overall effects, remain unclear [198]. Interactions can arise not only from overlapping targeted proteins but also from affected proteins participating in the same or similar biological functions within a PPI network [310].

In oncology, the gut microbiome both metabolizes dietary components and modulates chemotherapeutic responses, adding a critical layer to PD food-drug interactions. Commensal bacteria convert polyphenols and fibers into bioactive metabolites - such as short-chain fatty acids and urolithins - that can interact with drug targets or signalling pathways, potentially enhancing drug efficacy or altering toxicity [192]. For example, microbial conversion of dietary ellagitannins into urolithin A has been shown to enhance anti-inflammatory signalling and potentiate the efficacy of chemotherapeutics via NF-κB and PI3K/Akt modulation [311]. Conversely, species like Eggerthella lenta can inactivate drugs - digoxin is reduced to inactive dihydrodigoxin in an arginine-dependent process - demonstrating that microbiome-mediated transformations can both boost and blunt drug activity [312].

To discover context-specific PD food-drug interactions, it is crucial to characterize the effects of food molecules and drugs in terms of affected gene modules and biological functions, rather than individual genes, and relate these modules to their interaction contexts [7]. Techniques such as network propagation have traditionally been used to uncover affected protein modules from an initial set of targeted proteins [2]. These techniques propagate information across a graph, typically a PPI network, using pre-defined rules, as seen in the DreamLab DRUGS project [313]. Alternatively, learnable propagation algorithms like GNNs have been employed to optimize information propagation for specific

tasks, achieving state-of-the-art performance in bioinformatics applications, including the modelling of drug and food molecule effects on the genome [314–316].

4.2.1 Quantifying Individual Drug/Food Responses

Quantifying PD response of drugs and food-derived molecules individually is essential for understanding their impact on cellular systems and their potential interactions [317]. This process involves fitting dose-response data to sigmoidal curves, described by the Hill equation [318,319]. The Hill equation describes the relationship between the concentration of a drug or food molecule and the biological response it elicits [320]. As one of the parameters, the Hill coefficient indicates the steepness of the curve, reflecting how the response changes with incremental increases in concentration [321]. A higher Hill coefficient suggests cooperative binding, where the binding of one molecule increases the likelihood of subsequent molecules binding. In contrast, a coefficient closer to 1 implies non-cooperative binding [322]. The Emax parameter represents the maximum effect that the drug or food molecule can achieve, providing a measure of its efficacy [323]. EC50 refers to the concentration of the drug or food molecule that produces 50% of the maximal effect (Emax), and it is a key indicator of the compound's potency. Finally, the IC50 is calculated with the previous parameters and represents the concentration of the compound that results in a 50% reduction in cell viability or other biological responses [324]. IC50 value is particularly important as it allows for comparisons between different compounds or treatments [325]. Lower IC50 values indicate higher potency, meaning that a lower concentration of the drug or food molecule is required to achieve the same level of inhibition or effect [326]. For

4.2.2 Quantifying Synergies in Drug-Drug and Food-Drug Interactions

Drug synergy is a critical concept in pharmacology, especially when considering combination therapies for diseases such as cancer. Synergistic interactions between drugs can lead to enhanced therapeutic effects, allowing for lower doses of each drug and reducing potential side effects. However, it is important to recognize that there are several methods available.

The **Bliss Independence model** assumes that the combined effect of two drugs can be predicted by their individual effects if they act independently. The expected combined effect is calculated as the product of the individual effects of the drugs. The Bliss synergy score is then the difference between the observed combined effect and the expected combined effect. Positive scores indicate synergy, negative scores indicate antagonism, and scores around zero indicate additivity [327].

The **Loewe Additivity model** assumes that a drug cannot interact with itself and, therefore, the combined effect of two drugs can be described by their dose-response curves. This model involves constructing isobolograms to determine if the combined effect deviates from the expected additivity. An isobologram is a two-dimensional plot in which the dose of Drug A is on the x-axis and the dose of Drug B is on the y-axis for all combinations that achieve the same specified effect; the straight *isobole* connecting the individual-drug doses defines theoretical additivity, so points below it indicates synergy and points above indicate antagonism. Synergy is indicated when the data points fall below the line of additivity, antagonism is indicated when points fall above the line, and additivity is indicated when points fall on the line [328].

The **HSA model** compares the effect of a drug combination to the effect of the most potent drug alone. The HSA score is calculated by subtracting the highest single agent effect from the combination effect. Positive HSA scores indicate synergy, negative scores indicate antagonism, and scores of zero indicate additivity. This model is useful in identifying combinations where the combined effect exceeds the effect of the most potent single agent [329].

The **Combination Index (CI) method** quantifies drug interactions based on the dose-effect relationship. The CI is calculated using the median-effect equation, which incorporates the doses of drugs required to achieve a certain effect level when used alone and in combination. A CI less than 1 indicates synergy, a CI equal to 1 indicates additivity, and a CI greater than 1 indicates antagonism. This method is widely used due to its comprehensive mathematical framework for evaluating drug interactions [330].

The **ZIP score** quantifies the interaction between two drugs based on their combined effect compared to their individual effects [329]. A ZIP score greater than zero indicates synergy (the combined effect is greater than the expected additive effect), a score less than zero indicates antagonism (the combined effect is less than the expected additive effect), and a score of zero indicates an additive interaction (the combined effect is equal to the expected additive effect). ZIP scores are extensively used in preclinical and clinical research to evaluate the efficacy of drug combinations.

Using the GNN model trained on drug-drug combinations [316], we uncover synergistic and antagonistic drug-drug and food molecule-drug interactions in the cancer cell lines HCT116 and A549. We perform *in vitro* validation of selected predicted food molecule-drug interacting pairs**.** Our work highlights the significant impact that machine learning can have on PD food-drug interaction discovery by narrowing the search space of lead molecules and reducing the cost of screening efforts.

4.3 Methodology

The experimental validation of the predictions made by our computational model (Figure 7, *Appendix 9.1* and *9.2*) involved culturing both A549 and HCT116 cell lines, treating them with the drugs and food molecules, and assessing cell viability post-treatment.

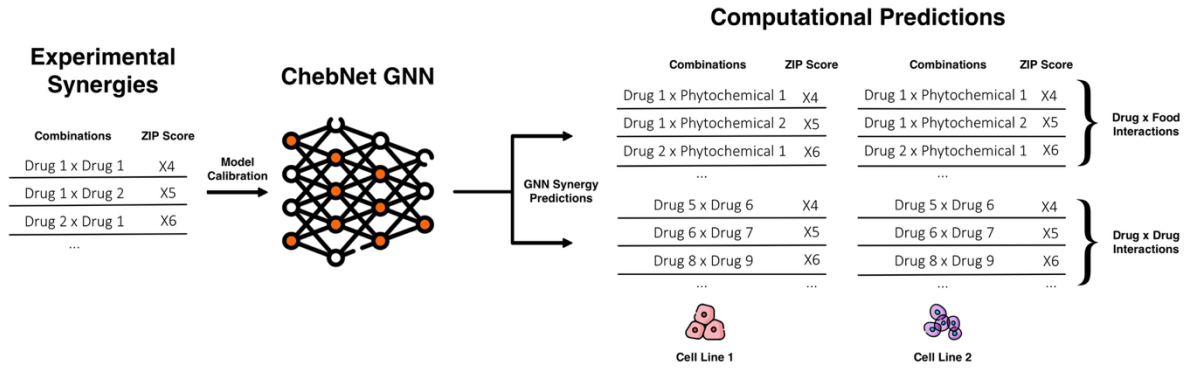


Figure 7 Drug-Drug and Drug-Food Computational Predictions. Computational predictions using a GNN to estimate synergy between various drug-drug and drug-food combinations. The GNN was retrained to predict interactions, generating synergy ZIP scores for different pairings, which were later experimentally validated in two cell lines to explore potential therapeutic synergies.

Cell maintenance through **cell passaging** and **seeding** ensured experimental estimation of drug/food interactions. Cell passaging is the process of subculturing cells to maintain optimal growth

conditions, ensures that cells remain in the logarithmic growth phase, where they are most responsive to treatments [331]. This minimizes variability in drug responses due to differences in cell confluence or stress from overgrowth. Cell seeding, where cells are plated at a specific density in multi-well plates, ensures that the number of cells in each well is consistent, providing a standardized platform for treatment [332]. This uniformity is essential for generating reproducible dose-response curves, which are sensitive to changes in cell number and health [333]. After seeding, cells are treated with increasing concentrations of drugs or food molecules to generate dose-response curves. The response to these treatments is quantified using cell viability assays, such as crystal violet staining, which reflect the proportion of viable cells after treatment [334,335]. These viability data are used to construct dose-response curves, which can be fitted using the Hill equation to estimate key pharmacodynamic parameters such as IC50 [336].

4.3.1 HCT116 and A549 Cell Passaging

A549 and HCT116 cell lines were passaged weekly in T-75 Flasks. Trypsin, McCoy medium, and PBS were warmed in a water bath for 15 minutes before aspirating cell medium from the T-flask using a lab aspirator. Cells were then washed with 2 mL of PBS to inactivate trypsin inhibitors present in the medium. After aspirating the PBS, 1 mL of trypsin was added to detach the cells. The cells were placed in an incubator for 2 to 3 minutes until they assumed a circular shape (confirmed under the microscope), indicating they were in suspension. To inactivate the trypsin, 6 mL of McCoy medium was added to the flask. The solution, now containing suspended cells, was transferred to a new 15 mL Falcon tube. The tube was centrifuged for 5 minutes at 1,000 RPM. The supernatant was removed using a lab aspirator, and 1 mL of medium was added to resuspend the cell pellet. The resuspension was performed by pipetting up and down to ensure a homogeneous solution without bubbles. The cells were then passaged at a 1:3 ratio by adding 1/3 mL of the cell suspension to a new T-flask, followed by the addition of 10 mL of fresh medium. Both flasks were placed in the incubator for 48h at 37°C, before proceeding for the cell seeding.

To minimize phenotypic drift and variability in drug response, all experiments were conducted using A549 and HCT116 cells between passages 5 and 20. Lower passage numbers were prioritized whenever possible, and cells outside this range were discarded to preserve experimental consistency and cellular integrity [337].

4.3.2 HCT116 and A549 Cell Seeding

Similarly to the previous cell passaging protocol, we started by aspirating the medium from the T-flasks with the attached cancer cells. The cells were washed with 2 mL of PBS, and the PBS was aspirated. 1 mL of trypsin was added to detach the cells, which were then incubated for 2 to 3 minutes until becoming detached (confirmed under the microscope). Detached cells were resuspended in 6 mL of medium to inactivate the trypsin. The cell suspension was transferred to a 15 mL Falcon tube and centrifuged for 5 minutes at 1,000 RPM. The supernatant was aspirated, and 10 mL of medium was added to resuspend the cells. The cell suspension was pipetted until making it homogeneous. A 10 µL sample of the suspension was used for cell counting to determine the required seeding density. This sample was initially mixed with the same volume of Trypan Blue in an Eppendorf tube to make the viable cells visible and after cells were counted using a haemocytometer from each cell concentration c_1 was retrieved (equation below). The volume V_1 needed to be extracted from the 15 mL Falcon to achieve 6,000 cells/well density was calculated and the remaining of the volume was filled with medium. This final solution was deposited in the lab boat and our electronic multi-channel micropipette used to transfer 100 µL of cell suspension to each well from a 96-well plate. Each plate contains 50 wells that were filled with medium plus cells and 10 wells filled only with medium as the control. The plates were left overnight at 37°C in the incubator to allow cell adherence.

4.3.3 Drug/Food Treatments

Monotherapy experiments were conducted for each drug and food molecule to establish their individual IC50 values under our experimental conditions. We used five food molecules - Galangin, Thymoquinone, Linoelaidic Acid, Selenomethionine, and Tartaric Acid. And nine cancer drugs -

Sunitinib, Gemcitabine, Dactolisib, MK-8776, Parthenolide, Carboplatin, Doxorubicin, Mitomycin-C and Etoposide. The selection of compounds used in the monotherapy and synergy assays was based on a combination of practical and experimental design criteria. While the compounds used in this chapter differ from those in Chapter 3, they were selected to test the generalizability of our GNN model across diverse chemical classes, including both FDA-approved chemotherapeutics and food-derived molecules with known bioactivity. Compounds were prioritized based on their (i) documented pharmacological relevance in oncology, (ii) predicted synergy with selected food molecules from Chapter 3 based on network-based propagation, and (iii) availability, solubility, and cost constraints in our experimental setting. Several high-confidence hits from Chapter 3 were excluded from wet-lab testing due to commercial unavailability or formulation instability. The goal was to test a representative subset of drug-food interactions across distinct pharmacological mechanisms.

For every compound that was received in solid form, we dissolved it in DMSO to create a stock solution. All compounds were fully solubilized with no visible precipitate, and solubility was confirmed via visual inspection and vortexing. These stock solutions were then serially diluted in culture medium to prepare working solutions. The final volume per well in the 96-well plates was 150 µL, consisting of 100 µL of cell suspension + 50 µL of compound solution. All compound solutions were prepared such that the final concentration of DMSO per well did not exceed 0.5% v/v, aligning with cytotoxicity standards recommended in cell-based assays. For instance, if the highest concentration of the compound stock required 100% DMSO (v/v), it was diluted 200 times in the working solution added to each well (1:3 dilution into 150 µL). Multiple intermediate dilution steps ensured that the cumulative DMSO content remained below the threshold. Before addition to each plate, all compound solutions were vortexed to ensure homogeneous distribution and were in true solution at the time of plating.

The range of intervals tested for each compound differed based on their literature values specific to the HCT116 and A549 cell lines. Whenever these values were unavailable, the mean IC50 value across a wide range of cell lines (retrieved from GDSC database) was chosen as the midpoint. A range of 10 concentrations prepared through serial dilutions were tested (starting with 0), with at least

5 technical replicates considered for every concentration in each 96-well plate. Technical replicates are repeated measurements performed on the same biological sample (e.g., multiple wells treated with the same compound concentration within the same plate), used to assess intra-experimental variability. Biological replicates, on the other hand, represent independent experiments conducted on different days using separately cultured cell batches, and are used to evaluate reproducibility across experimental runs. Each well contained 50µL of drugs or 50µL of food molecules at the determined concentrations added to 100µL of cells and medium. Before adding each 50µL, we ensured the solution of each compound was well homogenized using a vortex. Cells were then incubated for 72 hours, and cell viability was assessed using the crystal violet assay and the absorbance readings were made followed by an estimation of each compound's IC50 values (Figure 8).

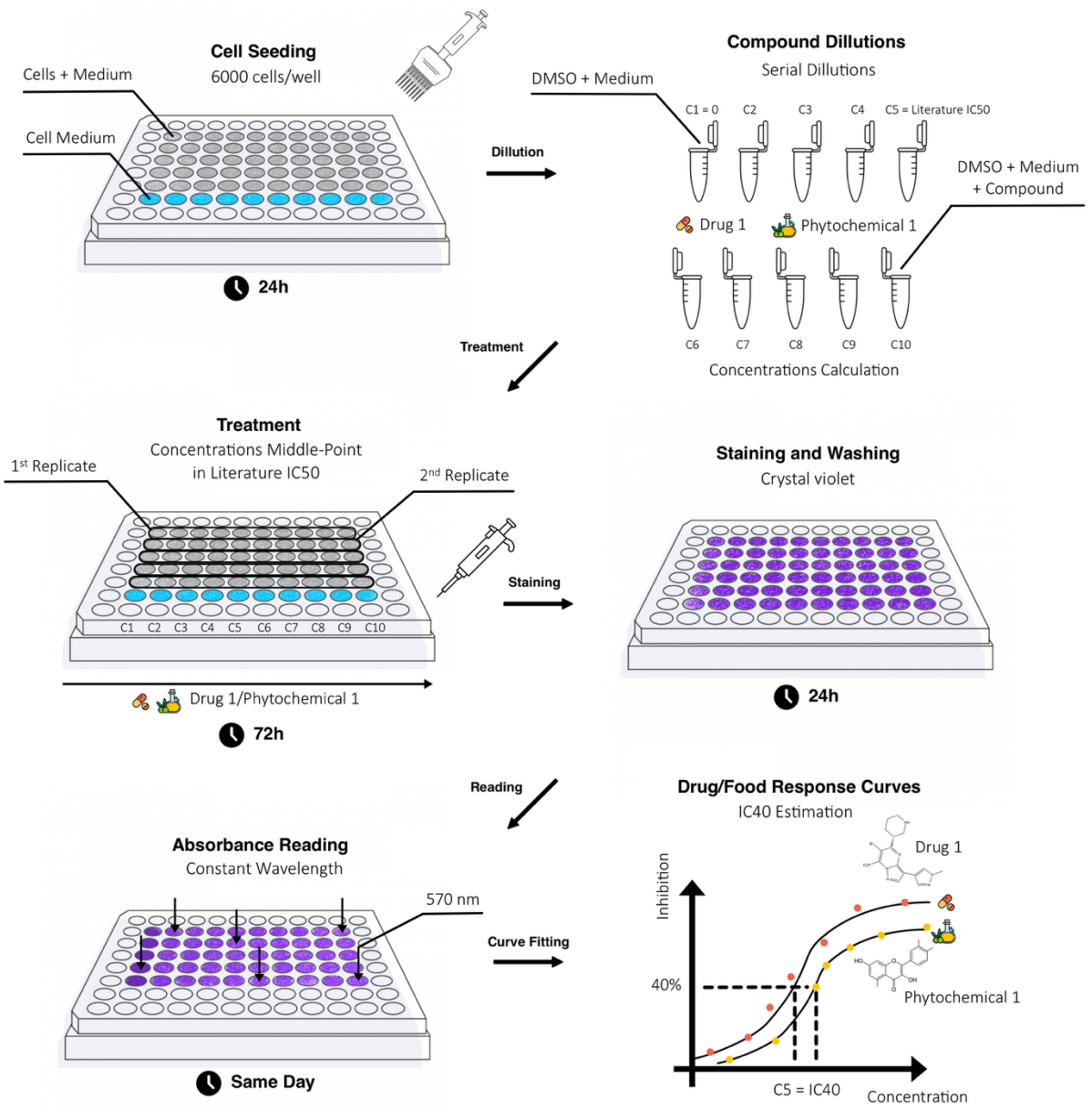


Figure 8 Drug and Food Monotherapy Treatments. Monotherapy experiments, where cells were seeded in 96-well plates (6,000 cells/well) and treated with drugs or food-extracted compounds based on their respective IC50 values for 72 hours. Cell viability was assessed using crystal violet staining, followed by absorbance readings at 570 nm to establish dose-response curves and determine the IC50 values of each compound.

4.3.4 Cell Viability Assay and Absorption Readings

Cell viability post-treatment was assessed using a crystal violet assay. The medium was aspirated from the wells, and the cells were washed twice with PBS. 50 µL of crystal violet solution (0.5% m/v in 25% methanol) was added to each well, and the plates were incubated at room temperature for 20 minutes on a bench rocker set to 20 oscillations per minute. The plates were then washed four times with PBS

and left to dry for at least 2 hours at room temperature. After drying, 200 µL of ethanol was added to each well, and the plates were rocked for another 20 minutes. Absorbance of the treatment and control wells was measured at 570 nm using a Tecan Sunrise plate reader powered by the software i-control 2.0 software to quantify cell viability. The plate definition used was Creimer 96 Flat Transparent. Hill curves for each compound and respective IC50s were then calculated.

To assess baseline cell viability and quantify the magnitude of drug-induced cytotoxicity, untreated wells (containing cells, medium, and vehicle only) were included in every 96-well plate and served as negative controls. These control wells were used to normalize the absorbance readouts of treated samples. Viability values were expressed as a percentage relative to the mean of these untreated controls (set to 100%). This allowed direct comparison of treatment effect sizes across conditions and replicates.

4.3.5 Drug/Food IC50 Estimation

Experimental data were first pre-processed by removing missing or infinite values. Then, the algorithm starts with initial guesses for E_{max}, EC_{50}, and h. These initial values were provided within predefined bounds to ensure biologically plausible results. E_{max} was constrained between 0 and 100%, assuming a range between no inhibition and complete inhibition. EC_{50} was bounded between 0 and an upper limit based on experimental concentrations. The Hill coefficient h was restricted between 1 and 3 to model typical sigmoidal curve shapes.

The fitting algorithm minimized the error between predicted and observed responses by adjusting the previous parameters. The output of the fitting to the Hill Curve consisted of E_{max}, EC_{50}, and h which were estimated from the experimental data using non-linear least squares regression. Which were used to estimate IC40 and IC50 values. In this analysis, the Levenberg-Marquardt algorithm was employed to minimize the sum of squared residuals between the experimental data points and the predicted values from the Hill equation. Plus, the covariance matrix, which contains the variances and covariances of these estimated parameters for the calculation of the standard deviations associated

with each parameter. In addition to the fitted parameters, variance of IC40 and IC50 values were calculated through error propagation using the partial derivatives of the IC_x with respect to E_{max}, EC_{50}, and h. And the resulting standard deviation calculated from the previous.

With the IC40 values for each compound, concentrations range to be used in the combinatorial treatments was chosen. For more details on the mathematical basis for calculating E_{max}, EC_{50}, h, IC_{40} and IC_{50} refer to *Chapter 2.4* in the *General Methodology*.

4.3.6 Drug-Drug and Food-Drug Treatments

Serial dilutions were performed to achieve the desired concentrations ranging from 0 until the IC40 value of the same compound obtained from the monotherapies. Five concentrations were tested for each compound, a total of 25 combinations and 2 technical replicas per 96-well plate (layout in Figure 9). For each of the plates previously prepared, we added 50µL of drugs and 50µL of food molecules/drugs at the determined concentration to 100µL of cells and medium. To achieve the desired final concentration of each compound in the well, drug and food molecule stock solutions were prepared at four times the target concentration. This accounts for the subsequent four-fold dilution that occurs when 50 µL of each compound is added to 100 µL of cell suspension, resulting in a final volume of 200 µL per well. For example, to reach a final concentration of 10 µM in the well, a 40 µM working solution was prepared and added at 50 µL. Both compounds were previously mixed and homogenized with a vortex before pipetting in each well. Plates were returned to the incubator at 37°C for 72 hours and cell viability was assessed using the crystal violet assay and absorbance readings as detailed before. Therefore, synergy scores could be estimated. Table 5 and 6 summarize the combinations tested along with the number of technical and biological replicates for drug-drug and drug-food combinations, respectively.

Table 5 Predicted Drug-Drug Combinations Tested in A549 and HCT116 Cell Lines. List of the drug combinations that were predicted using a GNN [316] and subsequently tested in two cancer cell lines: A549 and HCT116. For each combination, the table specifies the two molecules tested, the number of technical replicates (the repeated measurements within the same biological sample), and the number of biological replicates (independent experiments conducted on distinct biological samples).

Cell line	Molecule 1	Molecule 2	Technical replicates	Biological replicates
A549	Temozolomide	Rucaparib	3	2
	Sunitinib	Bosutinib	3	2
	Midostaurin	Sunitinib	3	3
	Mitomycin-C	Sorafenib	3	3
	Sunitinib	Rucaparib	3	2
HCT116	Sunitinib	Nintedanib	4	3
	Bosutinib	Nintedanib	4	3
	Midostaurin	Bosutinib	4	2
	Sorafenib	Bosutinib	4	2
	Sunitinib	Rucaparib	4	2
	Sunitinib	Bosutinib	4	2

Table 6 Predicted Drug-Food Combinations Tested in A549. List of the drug-food combinations that were predicted using a GNN [316] and subsequently tested in the cancer cell lines A549. For each combination, the table specifies the two molecules tested, the number of technical replicates, and the number of biological replicates.

Cell line	Molecule 1	Molecule 2	Technical Replicates	Biological Replicates
A549	Gemcitabine	Thymoquinone	1	1
	MK-8776	D-Tartaric acid	3	2
	MK-8776	Thymoquinone	2	2
	Gemcitabine	D-Tartaric acid	3	2
	Sunitinib	Linoelaidic acid	2	2
	BEZ-235	Galangin	2	2
	Carboplatin	Parthenolide	2	2
	Mitomycin-C	Selenomethionine	2	2
	Doxorubicin	Selenomethionine	4	2

4.3.7 Synergy Estimation

Before calculating the synergy values, the inhibition matrices generated from the experimental data were first interpolated to fill in missing data points and then smoothed using a Gaussian filter to reduce

noise. Average ZIP synergy values were computed across technical and biological replicates. These synergy values were used to determine the degree of synergistic or antagonistic interactions between the tested compounds, ultimately resulting in an overall synergy score for each drug-drug or drug-food combination (Figure 9). For more details on the mathematical basis for calculating ZIP Score, refer to *Chapter 2.4* in the *General Methodology*.

The processed data and source code have been made publicly available for further research and validation in this Bitbucket repository.

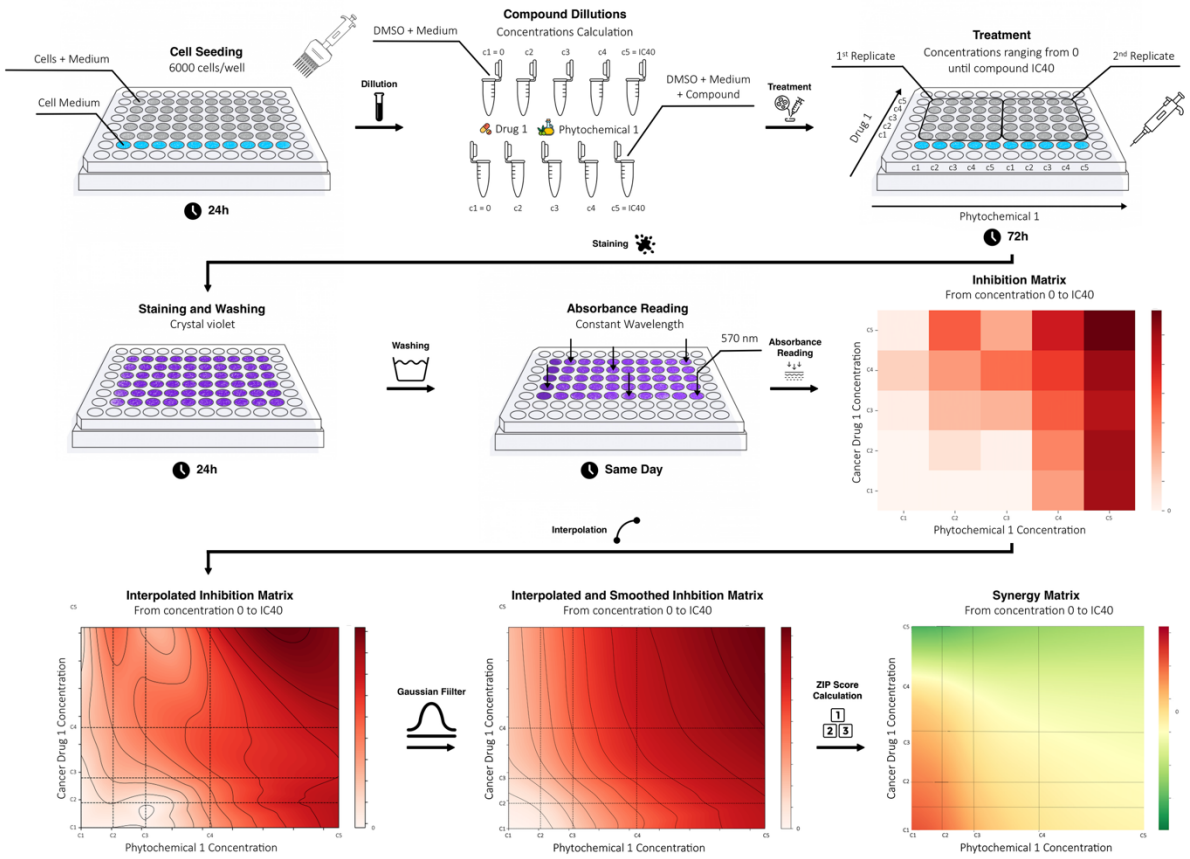


Figure 9 Drug-Drug and Drug-Food Treatments. Experimental validation of drug-drug and drug-food combinations, where cells were treated with combinations of drugs or food molecules for 72 hours. After treatment, crystal violet staining and absorbance measurements were performed to assess cell viability. The resulting data were used to generate inhibition matrices, which were then smoothed through Gaussian interpolation to produce inhibition and synergy matrices. ZIP scores were calculated to quantify synergistic or antagonistic interactions between the combinations.

4.4 Results

The graph neural network used in this study shortlisted drug-drug and food-drug combinations for experimental validation in A549 and HCT116 cancer cell lines. Figure 10 illustrates the computational

predictions for all 780 combinations of food molecules and drugs across the two cell lines, displaying their predicted synergy scores and associated standard deviations as estimated by the model.

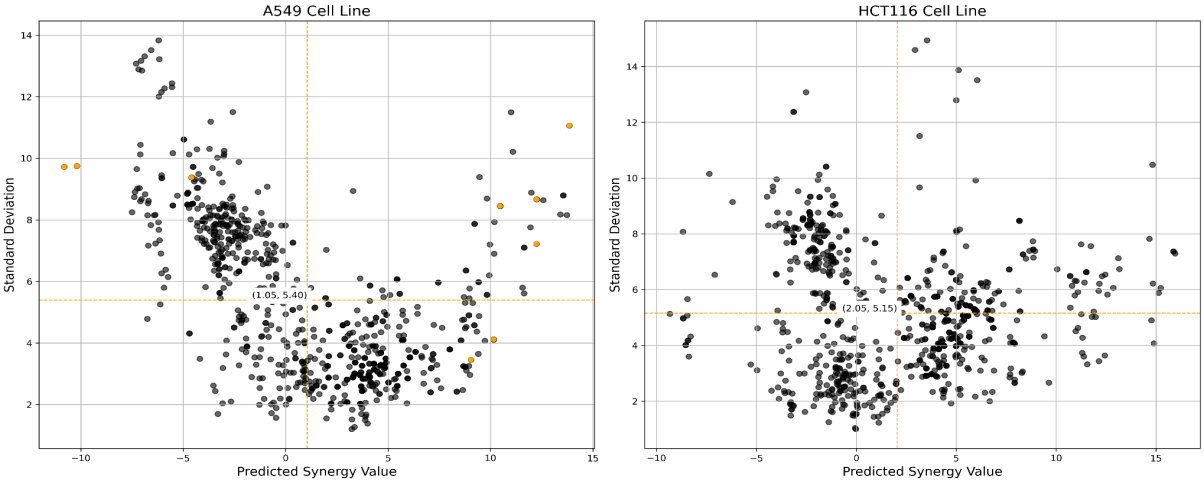


Figure 10 Computational Predictions of Drug-Food Combinations for A549 and HCT116 Cell Lines. Computationally predicted synergy scores for food-drug combinations in the A549 (left panel) and HCT116 (right panel) cancer cell lines. The X-axis represents the predicted synergy values, where positive values indicate potential synergy and negative values antagonism, while the Y-axis shows the standard deviation of these predictions, reflecting the uncertainty of the computational model. Each dot represents a food-drug combination, with specific combinations of interest highlighted and labelled. The orange points represent combinations that were selected for experimental validation based on their predicted synergy and low standard deviation. In A549, 780 drug-food combinations were analysed, while 780 combinations were assessed in HCT116.

Monotherapy treatments were performed, and the dose-response curves were generated for five food molecules—Thymoquinone, Selenomethionine, Tartaric Acid, Galangin, and Linoelaidic Acid—to assess their inhibitory effects on the A549 cancer cell line. Figure 11 presents these curves. On the side of each curve, Emax, EC_{50}, Hill coefficient, IC40, and IC50 were extracted.

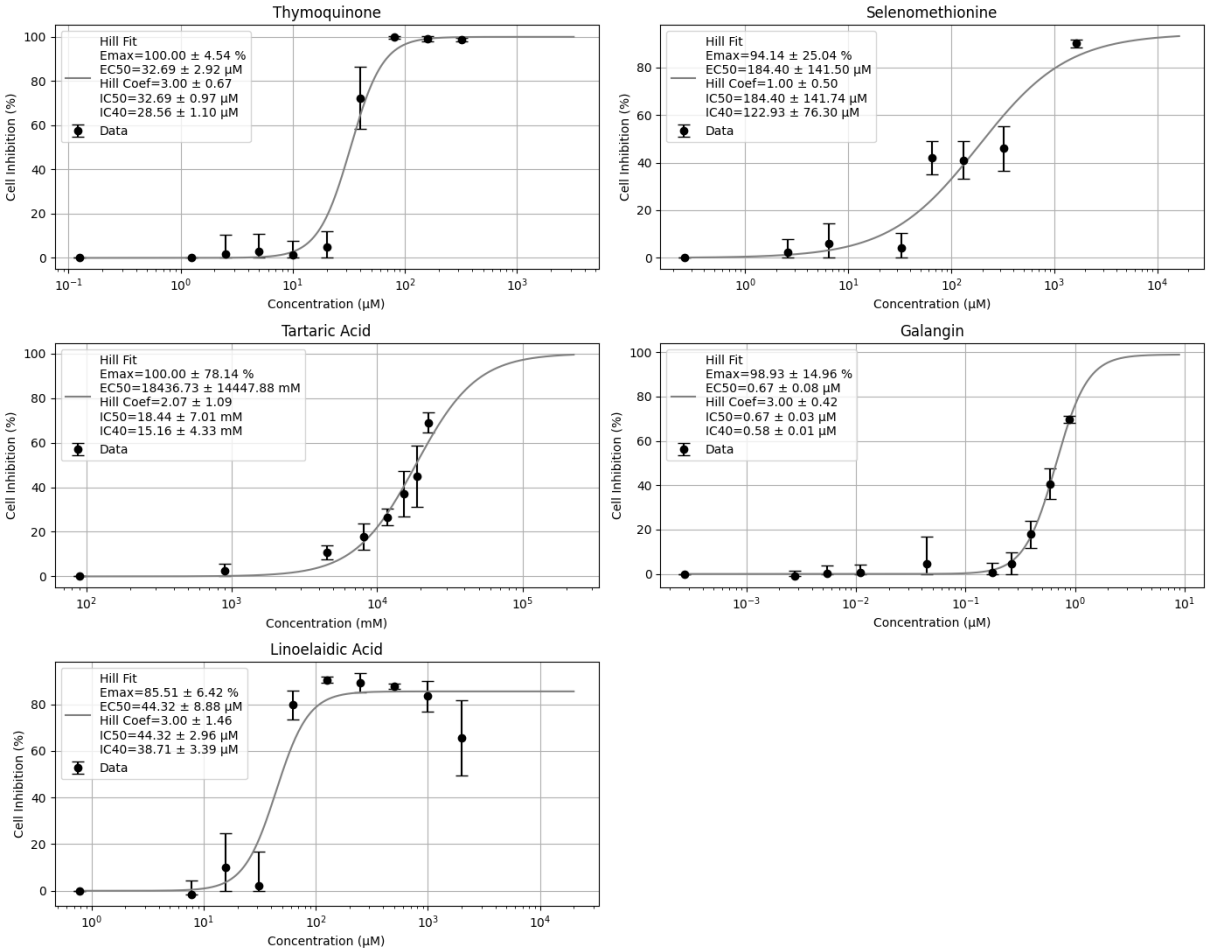


Figure 11 Dose-response curves for the monotherapies of Thymoquinone, Selenomethionine, Tartaric Acid, Galangin, and Linoelaidic Acid in A549 cancer cell line assays. Dose-response curves for five food molecules, showing their inhibitory effects on cancer cells. The X-axis represents the concentration of each food molecule, and the Y-axis indicates the percentage of cell inhibition. The curves are fitted to a Hill Fit curve, and the data points include error bars representing the standard deviations. Hill fit parameters Emax, EC50, Hill Coefficient, IC50, and IC40 were included.

Similarly, dose-response curves were obtained for nine drugs—Doxorubicin, Gemcitabine, Mitomycin-C, Dactolisib, Etoposide, Sunitinib, Carboplatin, MK-8776, and Parthenolide—to evaluate their inhibitory effects on cell viability across a range of concentrations. Figure 12 displays these curves, from which the same pharmacodynamic parameters were derived.

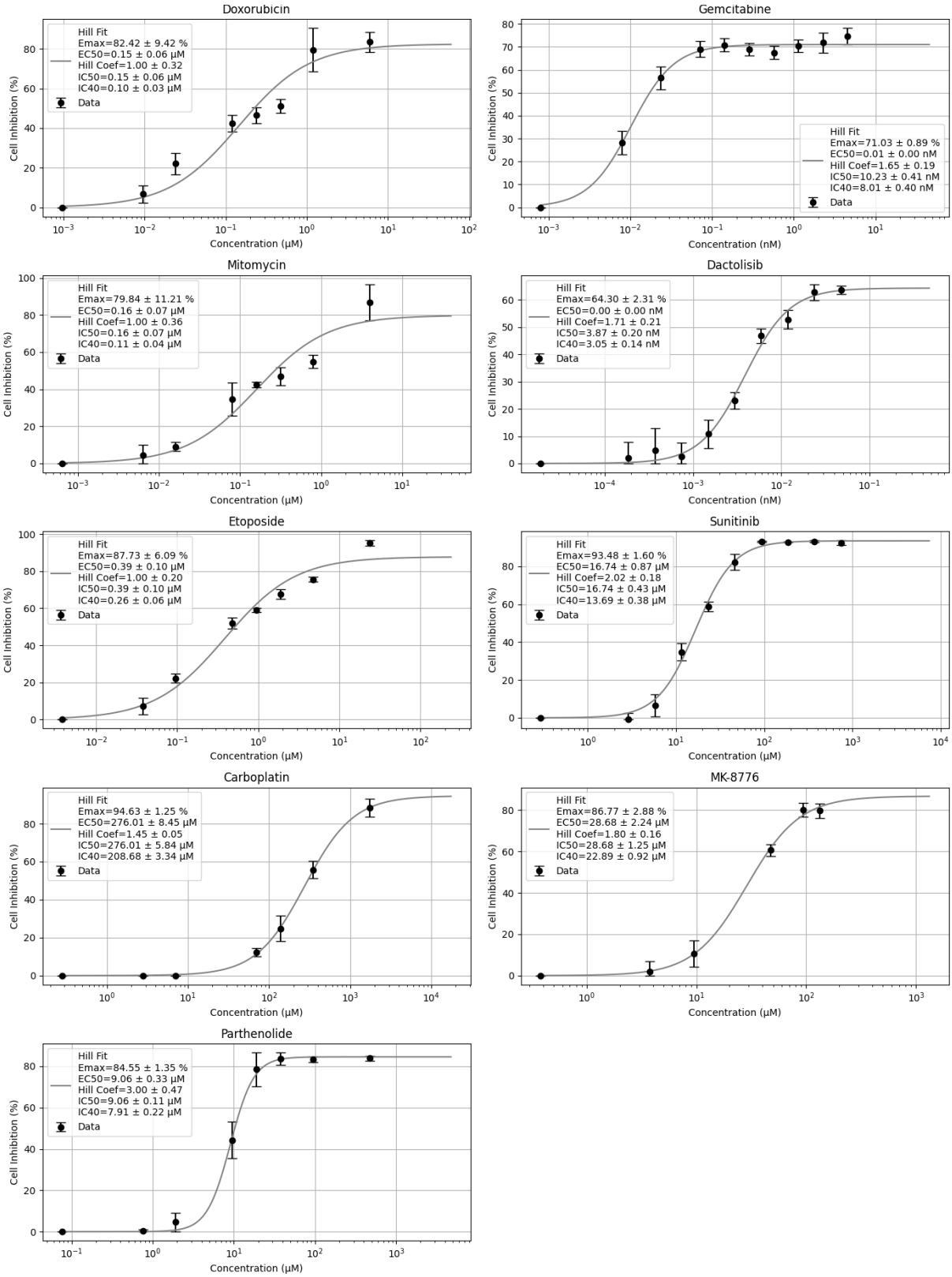


Figure 12 Dose-response curves for Sunitinib, Gemcitabine, Dactolisib, MK-8776, Parthenolide, Carboplatin, Doxorubicin, and Mitomycin in A549 cancer cell line assays. Dose-response curves for nine drugs, showing their inhibitory effects. The X-axis represents the concentration of each food molecule, and the Y-axis indicates the percentage of cell inhibition. The curves are fitted to a Hill Fit curve, and the data points include error bars representing the standard deviations. Hill fit parameters Emax, EC50, Hill Coefficient, IC40, and IC50 were reported.

The experimentally estimated IC50 values for each food molecule and drug were compared with those reported in the literature, as shown in Table 7.

Table 7 Monotherapies for Food Molecules and Drugs in A549 Cell Line. IC50 values for various food molecules and drugs tested as monotherapies in the A549 cell line are presented. The experimentally estimated IC50 values, along with standard deviations, are shown alongside literature-reported IC50 values for comparison where available.

Molecule Type	Molecule	IC50 Experimentally Estimated (µM)	IC50 in Literature (µM)	Reference
Food	Galangin	$6.7 \times 10^{2} \pm 3.0 \times 10^{1}$	220	[338]
	Thymoquinone	33 ± 1	10	[339]
	Linoelaidic Acid	$44.3 \times 10^{3} \pm 3.0 \times 10^{3}$	-	-
	Selenomethionine	184 ± 142	65	[340]
	Tartaric Acid	$18.4 \times 10^{3} \pm 7.0 \times 10^{3}$	-	-
Drug	Sunitinib	17 ± 1	2.4	[341]
	Gemcitabine	$10.2 \times 10^{-3} \pm 0.4 \times 10^{-3}$	6.6×10^{-3}	[342]
	Dactolisib	3.9 ± 0.2	0.4	[343]
	MK-8776	29 ± 1	-	-
	Parthenolide	9.1 ± 0.1	2.6	[344]
	Etoposide	0.4 ± 0.1	3.5	[345]
	Carboplatin	276 ± 6	5	[346]
	Doxorubicin	0.15 ± 0.06	70	[347]
	Mitomycin-C	0.16 ± 0.07	300	[348]
	Nintedanib	-	5	[349]
	Bosutinib	-	8.4	[209]
	Midostaurin	-	0.3	[209]
	Sorafenib	-	10.8	[209]
	Rucaparib	-	192.0	[209]

To evaluate potential synergistic effects, eleven drug-drug combinations and nine food-drug combinations were tested on A549 and HCT116 cell lines. The computationally and experimentally predicted ZIP synergy scores for each combinatorial treatment are shown in Tables 8 and 9, respectively.

Table 8 Computational and Experimental Drug-Drug Synergies in A549 and HCT116 Cell Lines. Computational Synergy column lists the synergy scores obtained from the predictions [316]. Experimental Synergy column provides experimentally obtained synergy values, calculated from various assay results across different dates. Positive values indicate synergy, while negative values suggest antagonism.

Cell line	Drug 1	Drug 2	Computational Synergy [316]	Experimental Synergy
A549	Rucaparib	Temozolomide	7.1 ± 3.2	3.5 ± 6.8
	Sunitinib	Bosutinib	6.9 ± 4.6	4.3 ± 4.9
	Midostaurin	Sunitinib	2.4 ± 4.8	4.2 ± 5.9
	Sorafenib	Mitomycin	-12.0 ± 4.4	-3.4 ± 5.3
	Sunitinib	Rucaparib	4.1 ± 2.9	2.3 ± 3.8
HCT116	Sunitinib	Nintedanib	18.1 ± 6.7	8.9 ± 11.0
	Bosutinib	Nintedanib	18.4 ± 7.7	7.7 ± 12.9
	Midostaurin	Bosutinib	19.0 ± 9.6	3.9 ± 3.3
	Sorafenib	Bosutinib	9.3 ± 3.4	5.0 ± 6.4
	Sunitinib	Rucaparib	11.6 ± 3.7	1.6 ± 3.1
	Sunitinib	Bosutinib	20.0 ± 5.9	3.8 ± 3.9

Table 9 Computational and Experimental Food-Drug Synergies in A549 Cell Lines. The Predicted Synergy (Computationally) column lists the synergy scores derived from computational predictions [316], where higher values suggest greater synergistic potential. The Predicted Synergy (Experimentally Recalculated) column provides experimentally obtained synergy values, recalculated from various assay results across different dates. Positive values indicate synergy, while negative values suggest antagonism.

Cell line	Drug	Food Molecule	Computational Synergy	Experimental Synergy
A549	Gemcitabine	Thymoquinone	13.9 ± 11.1	10.9 ± 13.4
	MK-8776	D-Tartaric acid	12.3 ± 8.7	-1.6 ± 5.7
	MK-8776	Thymoquinone	12.2 ± 7.2	1.7 ± 3.7
	Gemcitabine	D-Tartaric acid	10.5 ± 8.5	0.9 ± 4.8
	Sunitinib	Linoelaidic acid	10.2 ± 4.1	-3.6 ± 12.1
	BEZ-235	Galangin	9.0 ± 3.5	4.8 ± 9.9
	Carboplatin	Parthenolide	-4.6 ± 9.4	-2.3 ± 5.8
	Mitomycin-C	Selenomethionine	-10.8 ± 9.7	-2.4 ± 6.8
	Doxorubicin	Selenomethionine	-10.2 ± 9.8	-4.1 ± 7.0

Figure 13 compares the computationally predicted synergy scores with the experimentally measured synergy values for both drug-drug and food-drug combinations across the two cell lines. For the drug-drug treatments in A549 cell line, the computational predictions - Rucaparib + Temozolomide,

Sunitinib + Bosutinib, Midostaurin + Sunitinib and Sunitinib + Rucaparib. For the drug-drug treatments in HCT116 cell line, three out of six were aligned with the predictions from the model – Sunitinib + Nintedanib, Bosutinib + Nintedanib and Sorafenib + Bosutinib. For the drug-food treatments in A549 cell line, five out of nine treatments were aligned with the predictions – Gemcitabine + Thymoquinone, BEZ-235 + Galangin, Carboplatin + Parthenolide, Mitomycin-C + Selenomethionine and Doxorubicin + Selenomethionine.

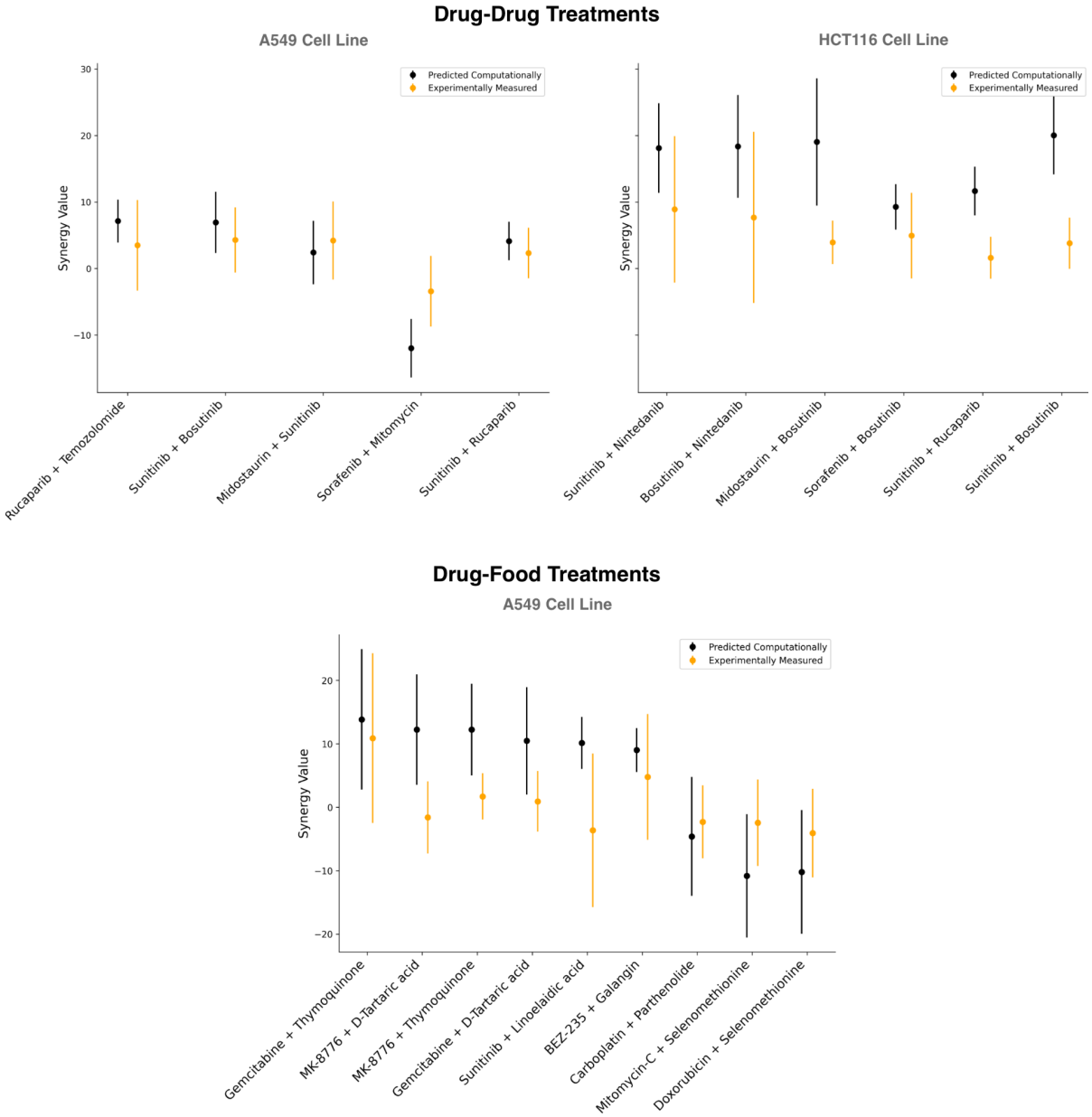


Figure 13 Predicted and Experimentally Measured Synergy Values for Drug-Drug and Drug-Food Combinations in A549 and HCT116 Cell Lines. Computationally predicted (black) [316] and experimentally measured (orange) synergy values for 11 drug-drug and 9 drug-food combinations in A549 and HCT116 cell lines. The Y-axis represents synergy values, with positive values indicating synergy and negative values suggesting antagonism, and error bars showing variability. In A549 cells, 5 drug-drug combinations were tested, with 4 out of 5 experimental results closely matching computational predictions. In HCT116 cells

(right panel), 6 drug-drug combinations were evaluated, with 4 showing consistency between predicted and experimental results.

As a representative example, Figure 14 shows the analysis for the combinations of Rucaparib with Temozolomide (drug-drug). This includes the calculated inhibition matrices, interpolated and smoothed inhibition matrices, and the resulting ZIP synergy scores.

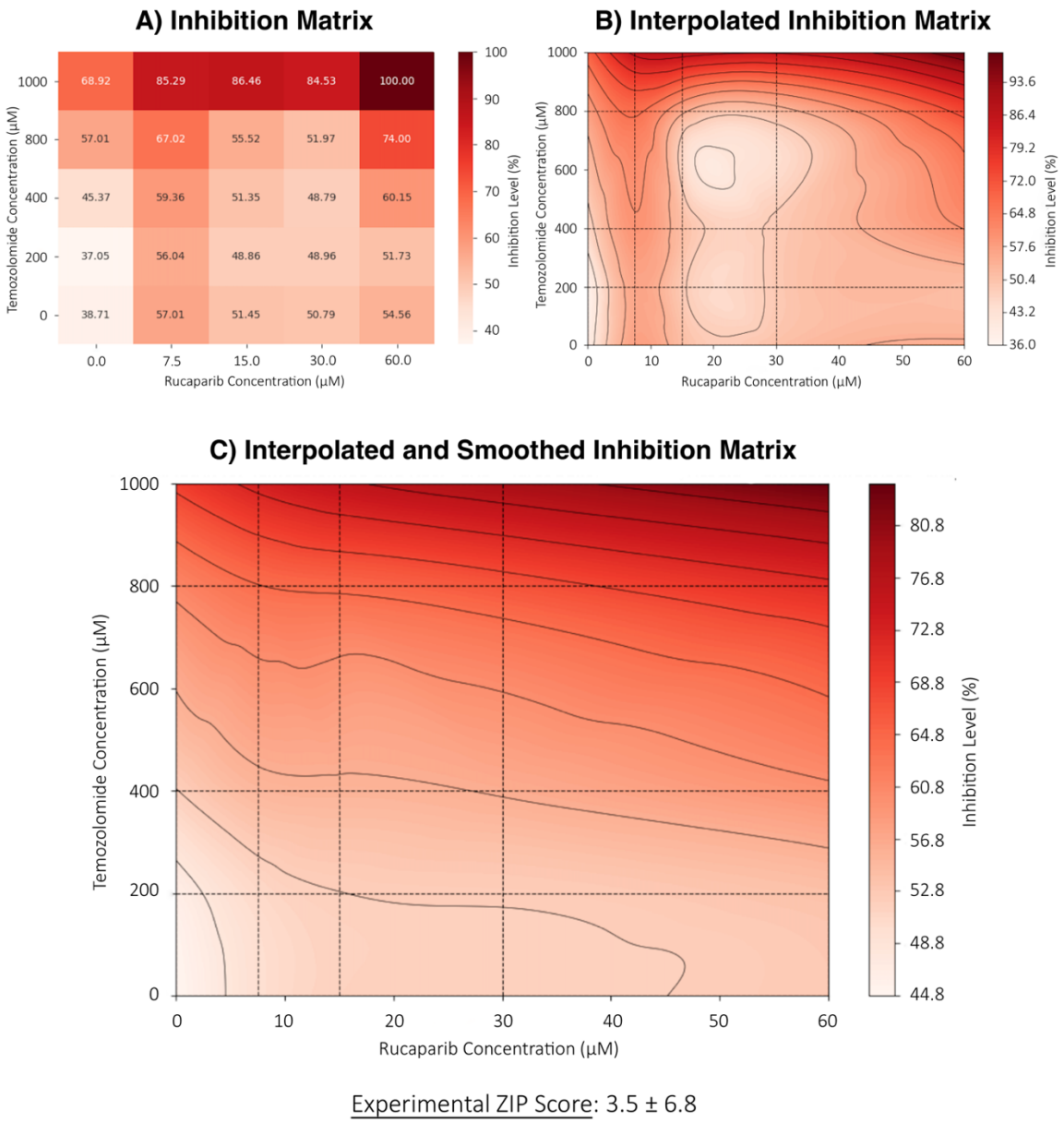


Figure 14 Experimental Validation of Drug-Drug Combination in A549 Cell Line. Smoothed, interpolated heatmaps showing cancer cell inhibition for various concentrations of drug-drug combinations in the A549 cell line. The X-axis and Y-axis represent the concentrations of the two compounds being tested, with colour gradients from light red to dark red indicating increasing levels of cancer cell inhibition. **A)** Shows the raw inhibition matrix for drug-drug combination, highlighting the observed inhibition at specific concentration pairings. **B)** Presents the interpolated inhibition matrices, with contour lines providing additional detail on areas of equal inhibition, showing how the combination effects vary across different drug concentrations. **C)** Further smooths the data to enhance visualization of trends across a wider range of concentrations. The mean ZIP synergy score is shown at the bottom of the figure, reflecting the average synergy observed across all concentrations tested.

And Figure 15 shows the interactions between Gemcitabine with Thymoquinone (food-drug) in the A549 cell line.

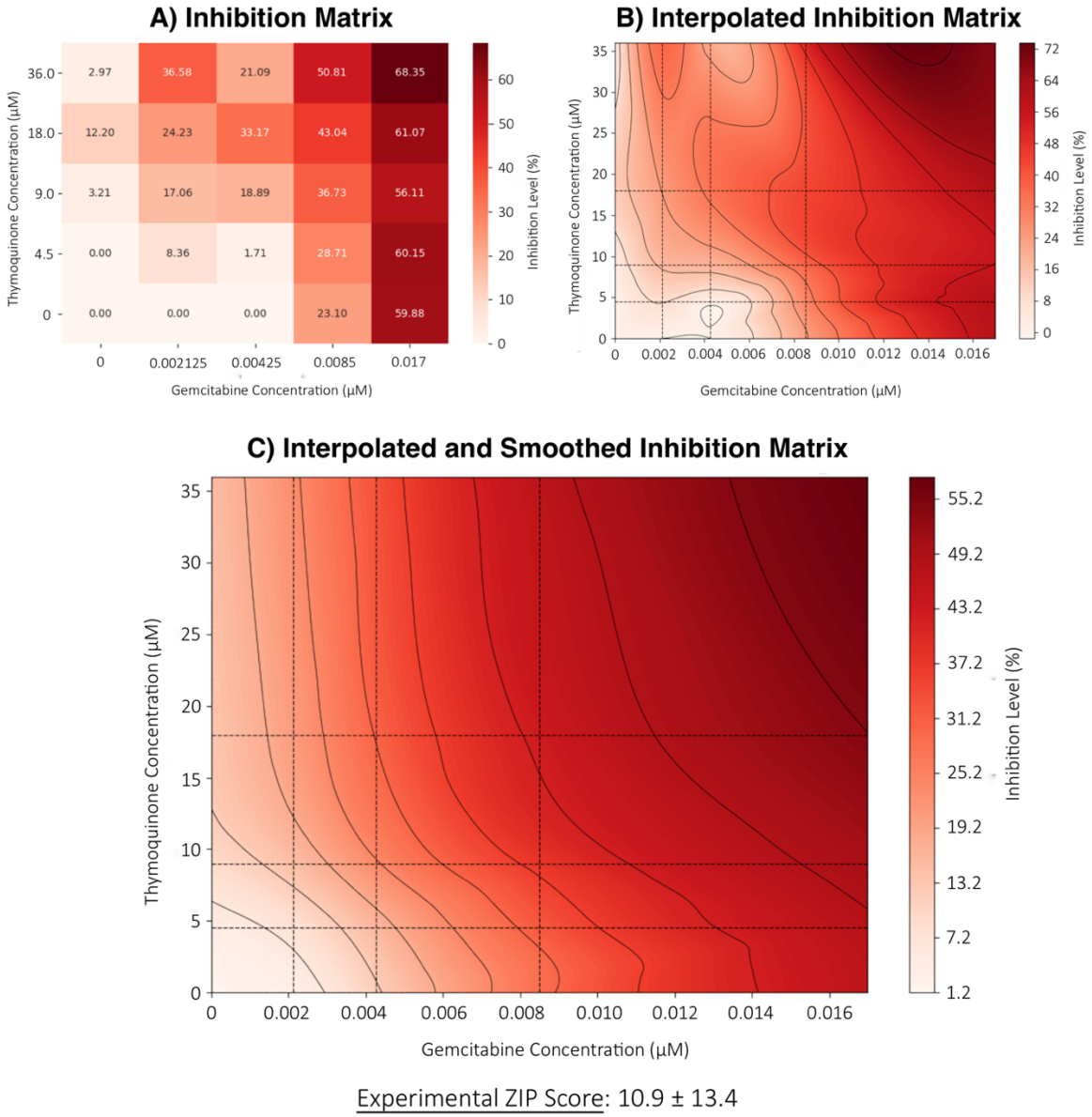


Figure 15 Experimental Validation of Drug-Food Combination in A549 Cell Line. Smoothed, interpolated heatmaps showing cancer cell inhibition for various concentrations of drug-food combination in the A549 cell line. The X-axis and Y-axis represent the concentrations of the two compounds being tested, with colour gradients from light red to dark red indicating increasing levels of cancer cell inhibition. **Panel A** shows the raw inhibition matrix for drug-food combination, highlighting the observed inhibition at specific concentration pairings. **Panel B** presents the interpolated inhibition matrices, with contour lines providing additional detail on areas of equal inhibition, showing how the combination effects vary across different compound concentrations. **Panel C** further smooths the data to enhance visualization of trends across a wider range of concentrations. The mean ZIP synergy score is shown at the bottom of the figure, reflecting the average synergy observed across all concentrations tested.

4.5 Discussion

Standard deviations predicted by the GNN for drug-food synergies are generally higher than those for drug-drug synergies, reflecting greater variability in the interaction effects between drugs and food molecules. This increased variability can be attributed to several factors. First, food molecules are typically less studied in the context of pharmacodynamics, leading to less predictable interactions with drugs. Additionally, the biological pathways and mechanisms through which food molecules exert their effects are often more diverse and less understood compared to drugs, which are designed with specific targets in mind.

We chose the A549 and HCT116 cancer cell lines to test our computational predictions because they are well-characterized biologically and very common cancer types, enhancing the validity and applicability of our findings. A549 cells are derived from human lung carcinoma, while HCT116 cells originate from human colorectal carcinoma, representing two distinct types of solid tumours with different genetic backgrounds and molecular pathways. This selection allows us to evaluate the predictive accuracy of our model across diverse cellular contexts, thereby assessing its generalizability. Both cell lines are extensively used in cancer research due to their consistent growth characteristics and responsiveness to chemotherapeutic agents, which facilitates reproducible and reliable experiments.

In comparing our experimentally determined IC50 values with those reported in the literature for various molecules, we observed a few discrepancies. For several compounds, including Galangin, Thymoquinone, Selenomethionine, Sunitinib, Dactolisib, Parthenolide, and Carboplatin, our experimentally estimated IC50 values were higher than the literature values. For instance, Galangin exhibited an experimental IC50 of 670 ± 30 µM, whereas the literature reports an IC50 of 220 µM in A549 cells. Similarly, Sunitinib showed an experimental IC50 of 16.7 ± 0.4 µM, compared to a literature value of 2.4 µM. For accuracy purposes, IC40 values were also computationally derived from each dose–response curve rather than being calculated as 80% of the IC50, although this had a negligible impact

on our experimental measurements. These differences suggest that under our experimental conditions, the cells were less sensitive to these compounds. These discrepancies can be attributed to several methodological and biological factors. First, differences in cell viability assays can significantly affect IC50 estimations. Our study used crystal violet staining, which measures total biomass, while many published IC50 values rely on metabolic assays such as MTT, WST-1, or CellTiter-Glo, which detect early cytotoxic effects. These assays often yield lower IC50 values, particularly for drugs that impair metabolism before causing cell detachment or lysis. Second, incubation time plays a critical role: we used a 72-hour treatment window, while some literature reports use shorter exposures (e.g., 24–48 hours), which can capture acute toxicity more sensitively. Third, cell-related factors such as passage number, cell density at seeding, and epigenetic drift between laboratories can influence drug response due to changes in proliferation rate, receptor expression, or efflux transporter activity. Fourth, compound-specific properties, including purity, solubility, degradation in culture media, and batch variability, may affect effective concentrations. Fifth, In our monotherapy assays, the dose–response data around each compound's IC50 exhibited high variability and a relatively flat transition region, which could have compromised the confidence of the Hill - curve fit and resulted in large standard deviations for the Hill coefficient and the derived IC40, IC50, and EC50 values [350].

To maintain experimental integrity, we ensured that the DMSO concentration did not exceed 0.5% to avoid cytotoxic effects. A549 and HCT116 cells were cultured in RPMI-1640 medium supplemented with 10% fetal calf serum (FCS) and penicillin-streptomycin-glutamine (PSG). Cells were plated at a density of 6,000 cells per well in 96-well plates. Prior to seeding, cells and media were warmed in a water bath for 15 minutes. All cell culture procedures were performed under sterile conditions in a laminar flow hood. Equipment and surfaces inside the hood were sterilized with 70% ethanol, and pipettes were cleaned with alcohol. To prevent contamination, pipettes used for aspirating liquids had cotton filters. The container for cell culture liquid waste was emptied when it reached half

capacity, and disinfectant was added before sealing it. Cells were incubated under standard conditions, and McCoy medium was stored in the refrigerator.

We observed consistency with predicted outcomes in 7 out of 11 drug-drug combinations and 5 out of 9 tested food-drug combinations aligned with the accuracy of 68.9% of our computational model. Worth noting that that the accuracy was different between cell lines. The predictions for the A549 lung cancer cell line demonstrated a higher degree of concordance with experimental synergy measurements compared to those for the HCT116 colorectal cancer cell line. This closer alignment in A549 can be attributed to several factors. Firstly, the A549 cells may have molecular and genetic profiles that are more similar to the training data used for our GNN-based model, leading to more accurate predictions. Lung cancer cells like A549 often exhibit specific pathways and receptor expressions that are well-characterized and widely represented in pharmacogenomic databases, enhancing the model's predictive capability for this cell line. Secondly, the heterogeneity inherent in HCT116 cells, including differences in gene expression, metabolic activity, and drug efflux mechanisms, can result in more variable responses to drug combinations, making computational predictions less precise [351]. Additionally, A549 cells may have more consistent drug uptake and intracellular processing pathways that align closely with the assumptions embedded in our computational model. These factors suggest that while our model is broadly applicable, its predictive accuracy can vary between cell lines due to biological differences, emphasizing the need for incorporating diverse and representative datasets in model training to enhance its generalizability across different cancer types. For this reason, food-drug combinatorial treatments were only performed in A549.

Although the agreement between computational and experimental synergy predictions was encouraging, a critical evaluation of statistical significance is necessary. Specifically, in both drug-drug (Table 8) and food-drug (Table 9) synergy analyses, the experimentally obtained ZIP synergy scores were accompanied by relatively large standard deviations, often overlapping zero. This overlap implies that, despite positive mean synergy values in combinations such as Rucaparib + Temozolomide and Midostaurin + Bosutinib (drug-drug, Table 8), or Gemcitabine + Thymoquinone and BEZ-235 + Galangin

(food-drug, Table 9), these interactions did not reach statistical significance under standard 95% confidence intervals. This limitation arises from the small number of biological replicates (n = 2–3) for each combination, which constrains statistical power and increases uncertainty. While technical replicates were used to improve precision within plates, biological variability—stemming from separate experimental runs—has a more profound influence on confidence intervals. Additionally, inter-plate variation and compound-specific kinetics may have contributed to the variability in synergy estimates. Therefore, while the directionality and magnitude of several predictions are promising, these results should be interpreted with caution, and further validation with increased biological replication is necessary to establish significance.

However, there are inherent limitations to our study that must be acknowledged. Reliance on *in vitro* data for model training and validation of the GNN, while a robust starting point, does not fully encapsulate the complexity of *in vivo* interactions. Biological systems in living organisms present additional layers of complexity that are not entirely replicated *in vitro*. Furthermore, the observed high variance in prediction scores for new drug-drug and food-drug combinations suggests that, while the model is effective, caution must be exercised in interpreting exact prediction scores without further experimental validation.

4.6 Conclusion

This study demonstrates the significant potential of using GNNs to predict PD interactions between food molecules and drugs, effectively addressing a gap in pharmacological research concerning the effects of food-drug interactions. Using a deep learning model derived from the DreamLab DRUGS network propagation approach, we achieved a relatively strong correlation in predicting PD drug-drug interactions. Experimental validation in human lung (A549) and colorectal (HCT116) cancer cell lines showed consistency with predicted results in many drug-drug and food-drug combinations, highlighting the model's promising robustness and applicability across different interaction types. The model's ability to propagate information across a multiscale interactome—including protein-protein, protein-

biological function, and biological function-biological function interactions—provides comprehensive insights into the mechanisms underlying PD interactions, guiding further experimental validation and therapeutic exploitation. These findings open possibilities for optimizing drug efficacy and reducing toxicity through dietary recommendations tailored to individual pharmacotherapy regimens, particularly in cancer treatment. While the study has limitations, such as reliance on *in vitro* data and variability in prediction scores requiring cautious interpretation and experimental validation, future research incorporating larger and more diverse datasets—including *in vivo* data—and integrating active learning components could refine the model further. Ultimately, this work underscores the power of deep learning and network-based approaches in predicting and understanding PD interactions between food molecules and drugs, providing a foundation for integrating dietary recommendations with pharmacotherapy to enhance drug efficacy and safety, thereby advancing personalized medicine and therapeutic outcomes.

5. Ingredient Substitution using Large Language Models

5.1 Abstract

In the emerging field of computational gastronomy, aligning culinary practices with scientifically supported nutritional goals is increasingly important. We explore how Large Language Models (LLMs) can be applied to optimize ingredient substitutions in recipes, specifically to enhance the phytochemical content of meals. Phytochemicals are bioactive compounds found in plants, which, based on preclinical studies, may offer potential health benefits, such as acting synergistically with approved drugs. We fine-tuned models of OpenAI's DaVinci, GPT-3.5 and Meta's TinyLlama on Recipe1MSubs dataset to generate ingredient substitutions and, simultaneously, enhance phytochemical content and a recipe dataset enriched on them. Our approach improved the accuracy of ingredient substitution tasks, from the baseline 34.5 ± 0.1% to 38.0 ± 0.3% on Recipe1MSubs dataset, and from 40.2 ± 0.4% to 54.5 ± 0.3% on a refined version of the same dataset. These substitutions led to the creation of 1,951 phytochemically enriched ingredient pairings and 1,639 unique recipes. While this approach demonstrates potential in optimizing ingredient substitutions, caution must be taken when drawing conclusions about health benefits, as the claims are based on preclinical evidence. This research represents a step forward in using AI to promote healthier eating practices, providing potential pathways for integrating computational methods with nutritional science.

5.2 Introduction

In recent years, computational gastronomy emerged as an interdisciplinary field, combining computational techniques with culinary science to enhance food preparation and consumption. One of the primary goals of this field is to develop methods for ingredient substitution, aimed at improving nutritional content, preserving flavour integrity, and aligning meals with specific dietary needs. A key

focus has been the integration of phytochemically enriched ingredients into diets, which has shown *in silico* potential to target biological networks of diseases like cancer [42], AD [230], and COVID-19 [227].

Phytochemicals, bioactive compounds found in plants, have gathered significant attention due to their antioxidant, anti-inflammatory, and anti-carcinogenic properties. Preclinical studies suggest that these compounds may play a role in disease prevention and treatment. For instance, brassinolide, a phytochemical present in tea, has shown potential to inhibit tumour growth and induce apoptosis in cancer cells [352]. In the context of AD, quercetin, found in EVOO, has been linked to improved brain health by exhibiting antioxidant and anti-inflammatory effects [353]. Moreover, genistein, a phytochemical in blackcurrant, has been investigated for its immune-supporting properties, including its potential to modulate inflammation and interfere with viral replication, making it relevant in the study of COVID-19 [354].

Initial attempts at ingredient substitution used statistical methods, such as Term Frequency–Inverse Document Frequency (TF-IDF), to identify potential substitutes based on the occurrence patterns within large recipe datasets. These methods analysed the co-occurrences of ingredients across a large corpus of recipes to identify statistically significant substitutes [355–357]. Later, co-occurrence-based methods further refined this approach by constructing ingredient networks to map their relationships across recipes, suggesting substitutes based on their mutual presence in culinary contexts [358–362]. The introduction of language model-based methods marked a significant evolution, utilising natural language processing techniques such as word2vec [80], BERT [86], and R-BERT [363] to capture semantic relationships between ingredients [80]. This approach proved effective in improving ingredient substitution tasks through learned embeddings [45], although language models require substantial computational resources and may not always capture the full culinary context.

More recently, GNNs have been utilised to combine the relational information encoded in ingredient graphs with the specific context of given recipes, leading to a deeper understanding of ingredient interactions [45]. Large-scale graphs, such as FlavorGraph, have been introduced to explore ingredient substitutions and food pairings, demonstrating the potential of GNN-based methods in

culinary applications [364]. Building on this approach, GISMo was introduced by Meta, a GNN-based model that incorporates both recipe-specific contexts and ingredient relationships from FlavorGraph. By constructing a benchmark dataset, Recipe1MSubs, which includes ingredient substitution pairs extracted from user comments, GISMo significantly outperformed previous methods in ranking valid ingredient substitutions. With a performance improvement of 14% in the top substitute ingredient prediction (Hit@1) over existing models [210].

The latest stage in this evolving field is represented by LLMs, which promise to overcome the limitations of previous approaches by leveraging their capacity for understanding human-like text [88,90]. The introduction of LLMs, such as GPT-3 developed by OpenAI [88], addresses the limitations inherent in previous methods for ingredient substitution by capturing the full culinary context [210]. LLMs, trained on extensive and diverse culinary datasets, can potentially offer more contextually aware and accurate ingredient substitutions by leveraging their understanding of both the syntax and semantics of culinary texts [210]. This capacity for high-level language comprehension and manipulation allows for considering factors such as ingredient compatibility. Additionally, LLMs can be fine-tuned for specific tasks such as ingredient substitution [90].

Recognizing the limitations of statistical, co-occurrence, language and GNN-based methods, our research proposes a novel approach leveraging the capabilities of LLMs, such as OpenAI's GPT-3.5 [90], DaVinci [88], and Meta's TinyLlama [365], for ingredient substitution. By fine-tuning these models on a dataset of recipes and ingredient substitutes, we benchmark them against the current state of the art GISMo considering only the top substitute predicted (Hit@1 metric) [210]. Plus, with the phytochemically enriched ingredient substitutions, we produce a new set of recipes targeting cancer, AD and COVID-19 disease biological networks (Figure 16).

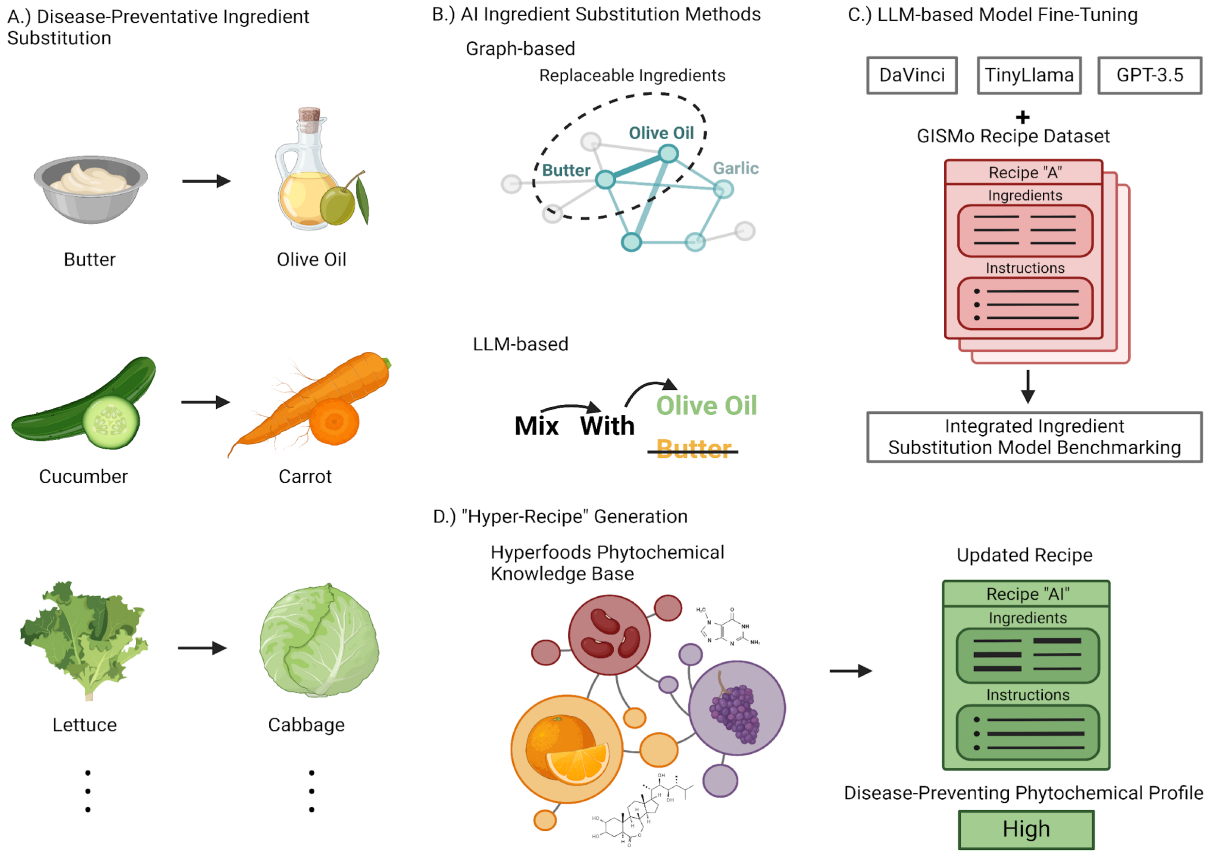


Figure 16. Ingredient Substitution Overview. **A)** Disease-Preventative Ingredient Substitution: it illustrates the process of substituting ingredients to enhance the phytochemical content of recipes focused on different disease-specific biological networks. Examples include substituting butter with olive oil (AD), cucumber with carrot (COVID-19), and lettuce with cabbage (cancer). **B)** Ingredient Substitution Methods: it compares graph-based approaches, which rely on ingredient co-occurrence and relational data, with LLM-based approaches, which utilise advanced language models for more context-sensitive substitutions. **C)** LLM-Based Model Fine-Tuning: it details the fine-tuning of LLMs such as DaVinci, TinyLlama, and GPT-3.5 for the ingredient substitution task using the Recipe1MSubs dataset. And it includes the benchmark against the GISMo model to evaluate accuracy. **D)** Phytochemically Enriched Recipe Generation: it details the generation of recipes with enhanced phytochemical profile. These recipes are generated using the fine-tuned model with the best performance [366].

5.3 Methodology

5.3.1 Recipe and Ingredient Substitution Dataset

Recipe1MSubs dataset [210] contains ingredient substitutes from a subset of recipes from Recipe1M [367]. This dataset contains 70,520 ingredient substitutions, with 49,044 designated for training, 10,729 for validation, and 10,747 for testing. Each recipe is structured with a recipe title, followed by the list of ingredients, with associated quantities, and, finally, the cooking instructions. By adopting this format, we optimize the GISMo and LLM models' capacity to analyse and understand the recipes, thereby improving the accuracy of ingredient substitutions. The original GISMo model was trained using a

methodology focused on ingredient context and co-occurrence. This benchmark approach aimed to identify potential ingredient substitutions within the dataset by leveraging statistical relationships between ingredients based on their appearance in recipes.

5.3.2 GISMo Benchmark

We reproduced the machine learning architecture and parameters as previously described in the reference study GISMo [210] to serve as a baseline for comparison with new models developed using LLMs. This benchmark setup includes a learning rate of 5e-05, weight decay of 0.0001, 300 hidden units per layer, an embedding dimension of 300, and a dropout rate of 0.25, across two layers, and involves 400 training rounds with regular negative sampling and random embedding initialization. Unlike data augmentation, ingredient sets are included to enhance the model's context sensitivity without altering the original dataset's composition. Average pooling is utilised for contextual embedding.

After implementing the two-layer graph architecture and hyperparameters from, we trained our GISMo replication on the original Recipe1MSubs training split and evaluated it on the held-out test split using exactly the four metrics reported in: mean reciprocal rank (MRR), Hit@1, Hit@3, and Hit@10 (MRR—the average of the reciprocal of the rank at which the first correct substitute appears—and the fraction of cases where the correct substitute is found within the top 1, 3, or 10 predictions) [210]. We directly compared our test set values to those published to confirm parity. To ensure robustness, we repeated the full training and evaluation pipeline with multiple random seeds and verified that the observed variability in MRR, Hit@1, Hit@3, and Hit@10 aligned with the ranges in original study.

Before fine-tuning the LLM's, we start by using them to augment the capabilities of the current GISMo model by introducing several enhancements. Each ingredient was classified in one of 23 food categories and the Recipe1MSubs dataset was curated using LLM, before retraining with GISMo.

5.3.3 Incorporation in GISMo of Food Category Feature

Using GPT-4-0613, the latest of OpenAI's language models, we categorize ingredients into culinary groups. This process involved a Python script utilising the *pandas* library for dataset manipulation and the *openai* library for API interactions. We queried GPT-4-0613 for each ingredient, requesting its classification into one of 23 categories ranging from common food groups like *Fruits* and *Vegetables* to more specialized ones such as *Confectioneries* and *Aquatic foods* (*Appendix 10.1*). By setting the temperature parameter to 0, the script prioritises reproducibility to ensure consistency in GPT-4-0613's responses. This approach processed a CSV file of ingredients, appending a category column with the GPT-4-0613-determined categories to the dataset. The augmented dataset now serves as a resource for ingredient substitution models, enabling more contextually relevant substitutions.

5.3.4 Dataset Filtration Based on Substitution Validity

To enhance the ingredient substitution model's accuracy, we used GPT-3.5-Turbo with an asynchronous Python script for evaluating the validity of the ingredient substitutions. This process involved sending prompts to GPT-3.5-Turbo, asking if one ingredient could feasibly substitute another within a specific recipe, and classifying responses into *Correct*, *Potential*, or *Incorrect*. By processing substitutions in multiple batches using the *aiohttp* library for asynchronous HTTP requests, we efficiently assessed the 70,520 substitutions. Substitutions categorized as *Correct* were considered suitable and retained, *Potential* indicated possible suitability requiring further consideration, and *Incorrect* were excluded, leading to their removal from the dataset. The final results, saved into a *JSON* file, formed a filtered dataset for retraining the model, ensuring it is based on accurate substitution data. Key settings included a prediction temperature of 0.5, a limit of 10 output tokens, and five runs to evaluate prediction stability. Plus, a batch size of 100 substitutions with respective recipes to avoid reaching the maximum number of requests per second. GISMo and LLM were re-trained and fine-tuned, respectively, with this curated Recipe1MSubs dataset.

5.3.5 Fine-tuning Language Models for Substitution Predictions

We fine-tuned GPT-3.5-Turbo-1106, DaVinci-002, and TinyLlama-1.1B with consistent specifications to ensure comparability. Key settings included a prediction temperature of 0.5, a limit of 10 output tokens, and five runs to evaluate prediction stability, all conducted over a single epoch. And training data was provided in prompt completion format for DaVinci-002 and TinyLlama-1.1B, and chat completion format for GPT-3.5-Turbo-1106. Number of epochs, training steps and batch sizes are in Table 10.

Table 10. Training Configurations for DaVinci-002, GPT-3.5-Turbo-1106, and TinyLlama-1.1B. Training configurations for DaVinci-002, GPT-3.5-Turbo-1106, and TinyLlama-1.1B, as well as their filtered dataset variants, are detailed below. This table includes the number of epochs, training steps, and batch sizes used for each model. Models highlighted with an (*) had their training parameters manually optimized to enhance performance.

Recipe1MSubs Dataset	Fine Tuned Model	Epochs	Steps	Batch Size
Filtered	GPT-3.5-Turbo-1106	1	1554	20
	DaVinci-002	1	1554	20
	TinyLlama-1.1B-1.1B*	1	970	8
Original	GPT-3.5-Turbo-1106	1	1533	32
	DaVinci-002	1	1533	32
	TinyLlama-1.1B-1.1B*	1	1532	8

For the fine-tuning of TinyLlama-1.1B models, we encapsulated the hyperparameters within the *TrainingArguments* setup. This configuration specified an output directory, a per-device train batch size of 8 (due to memory constraints) and applied gradient accumulation over 4 steps to efficiently balance between computational demand and memory constraints. The model optimisation was conducted using *paged_adamw_32bit* with a *learning rate* set at 5e-4, and a cosine learning rate scheduler was employed for optimal learning rate adjustments throughout the training phase. A save strategy based on epochs was utilised, coupled with logging and evaluation intervals set at 25 and 50 steps respectively, aligning with an evaluation strategy that triggers at specified steps to closely monitor the model's performance. The training was streamlined to complete within 1 epoch to ensure quick adaptation while preventing overfitting, without setting a maximum step limit and avoiding mixed

precision training to maintain computational accuracy. The *SFTTrainer* was used in the training process, directly interfacing with the training and validation datasets, and was configured with *peft_config* for tailored pre-fine-tuning adjustments. Text preprocessing was managed using a specified *dataset_text_field* and tokenizer, with packing disabled and a maximum sequence length of 512 to standardize input data handling. This approach aimed at enhancing the model's learning efficiency, prioritizing a balance between computational resource optimisation and achieving high-quality model training.

Building upon the filtration methodology outlined above, we randomly chose one of the previously filtered datasets and the original one to fine-tune TinyLlama-1.1B, DaVinci-002 and GPT-3.5-Turbo-1106, and retrain GISMo model. Accuracy of each model was determined after several uniformization steps and using Hit@1 metric as detailed below.

5.3.6 Evaluation of Ingredient Substitution Accuracy

To measure the accuracy of ingredient substitution predictions generated by LLMs, we developed a specialized algorithm to standardize and process ingredient names before comparing them to the ground truth provided by Recipe1MSubs, a curated list of one-to-one original–substitute pairs extracted from Recipe1M and partitioned into training, validation, and test splits. This ground truth consists of exactly one validated substitute per original ingredient, which allows unambiguous evaluation of single-recommendation accuracy. We begin by extracting predictions from the model output, where each line contains an original ingredient, its corresponding ground truth substitute, and the predicted substitution. To ensure consistency across ingredient names, several preprocessing steps are applied, including converting all text to lowercase, removing numeric values, and applying predefined rules to replace or eliminate special characters. After preprocessing, a clustering mechanism is employed to group similar ingredients, accounting for variations in lexical forms such as singular and plural versions or different types of the same ingredient (e.g., basmati rice and long grain rice). This clustering strategy avoids the need for manually constructing synonym libraries by using unsupervised grouping based on string similarity (Levenshtein distance) that collapse commonly interchangeable terms into the same

conceptual entity. Each ingredient is assigned a unique cluster identifier to ensure that similar ingredients are treated as equivalent during comparison.

Once the LLM-predicted ingredient names are uniformized and categorized, we compare them against the ground truth using the Hit@1 metric. Because Recipe1MSubs defines exactly one correct substitute per ingredient and our fine-tuned LLMs output only a single recommendation, we report Hit@1 exclusively for these models. This metric assesses the model's precision by determining whether the first predicted substitution matches the ground truth or falls within the same ingredient cluster. For example, if the ground truth substitution is barley and the model predicts basmati rice, both ingredients would be considered correct because they belong to the same grain cluster (Figure 17).

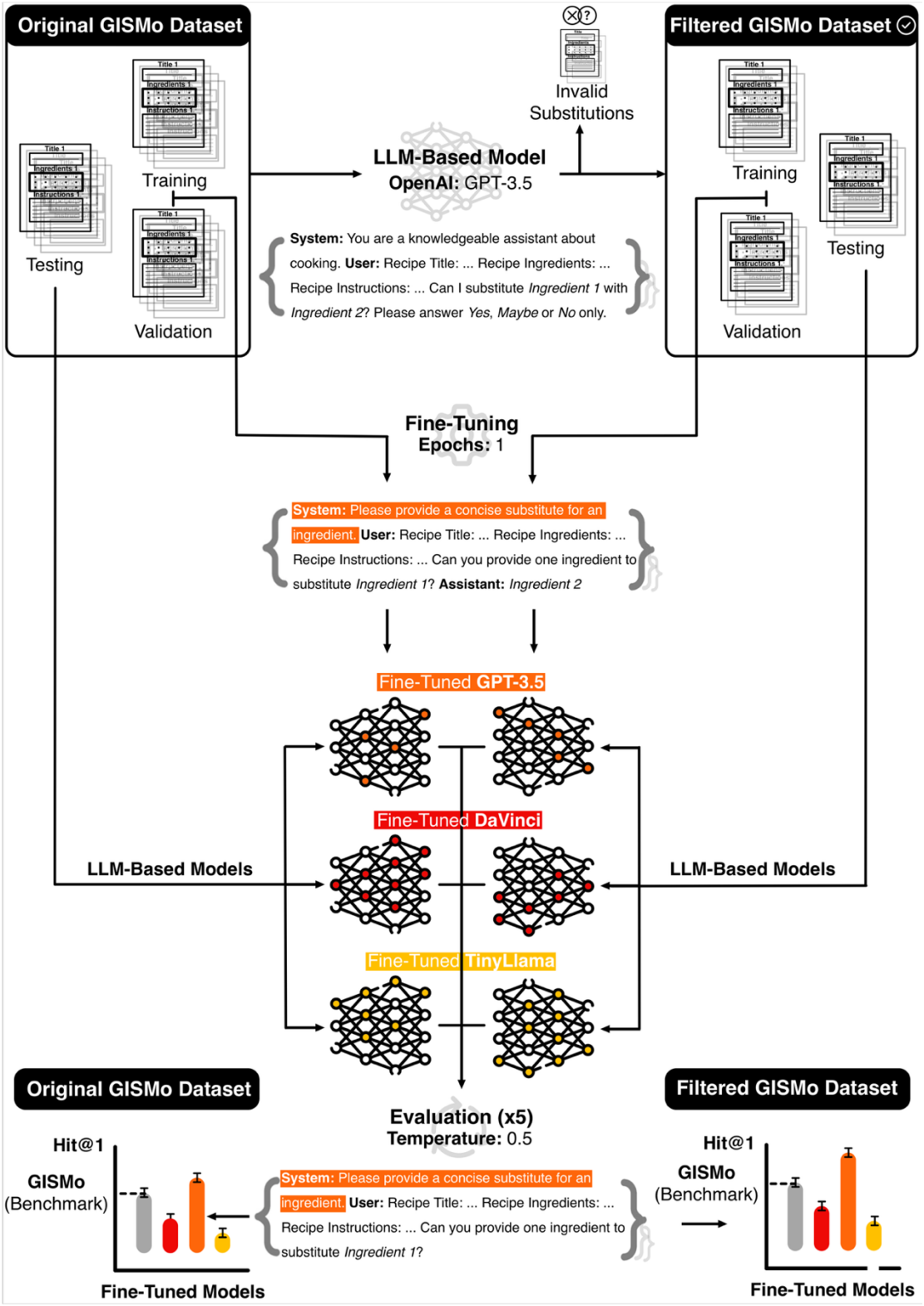


Figure 17 Ingredient Substitution Model Fine-Tuning and Evaluation. This figure illustrates the methodology for optimizing ingredient substitutions using LLMs. This details the fine-tuning of three LLMs—OpenAI's GPT-3.5, DaVinci, and Meta's TinyLlama—and benchmarked against the GISMo ingredient substitution model both using the original and filtered (invalid substitutions were discarded) Recipe1MSubs dataset. The process involved prompt engineering to optimize model responses for accurately recommending a single ingredient, with workflow diagrams showing the application of prompts and evaluation of responses for a step that was repeated five times with a temperature parameter of 0.5.

5.3.7 Phytochemically Enriched Recipe Generation

Finally, we integrated phytochemically enriched ingredients based on their ability to target molecular networks responsible for disease development in cancer [42], AD [230], and COVID-19 [227]. Every ingredient from every recipe in Recipe1MSubs was substituted using the best performer model and the recipes with phytochemically enriched ingredients at the end were kept as part of a new original dataset. Then, they were ranked based on their cumulative phytochemical profile. Only salads were considered given the lower number of cooking processes involved in their preparation and, consequently, higher chances of phytochemical preservation [368] (Figure 18).

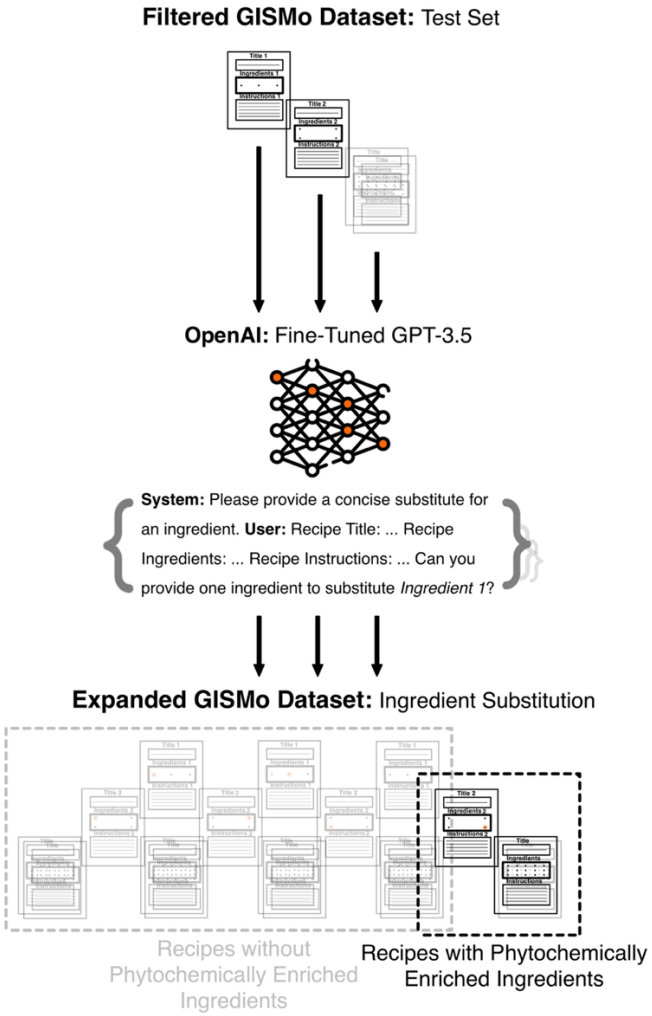


Figure 18 Phytochemically Enriched Recipe Generation. This figure focuses on applying the best-performing LLM to generate new recipes enriched with phytochemically beneficial ingredients. It illustrates model selection based on performance metrics, recipe generation incorporating phytochemically rich ingredients, and final selection based on phytochemical content analysis [366].

Data on the categories each ingredient from Recipe1M belongs to, as well as the script used to retrieve these categories, was made available. The repository can be accessed in Bitbucket. However, due to the use of proprietary models by OpenAI and limitations set by third-party data sources, some restrictions apply to the direct sharing of model weights and specific algorithmic implementations [366]

5.4 Results

Filtering the original Recipe1MSubs dataset through five separate runs with GPT-3.5-Turbo-1106 resulted in five filtered sets, with an average of 31,819 ± 67 training samples, 7,094 ± 25 validation samples, and 7,085 ± 21 testing samples, totalling 45,998 ± 85 (Table 11). From these, a randomly selected dataset with 44,615 ingredient substitutions, divided into 31,063 for training, 6,831 for validation, and 6,721 for testing, was used to rerun GISMo and fine-tune LLMs as detailed below.

Table 11 Performance metrics of five filtering runs for the GISMo model. Performance evaluation of the GISMo model, after being trained, validated, and tested on five datasets generated by filtering the original dataset through the GPT-3.5-Turbo-1106. The table below details the Hit@1 accuracy metric, and the number of recipes used in each phase—training, validation, and testing—across all runs.

Filtering Run	Hit@1 (%)	Training	Validation	Testing	Total
Run 1	40.3	31,733	7,082	7,080	45,895
Run 2	40.8	31,795	7,097	7,056	45,948
Run 3	40.2	31,797	7,083	7,096	45,976
Run 4	40.0	31,908	7,073	7,113	46,094
Run 5	39.9	31,860	7,136	7,081	46,077
Final	40.2 ± 0.4	31,819 ± 67	7,094 ± 25	7,085 ± 21	45,998 ± 85

Running the GISMo model five times with both the original and filtered datasets yielded Hit@1 accuracies of 34.5 ± 0.1% and 40.2 ± 0.4%, respectively. Integrating food categories (*Appendix 10.1*) as additional node features into our computational model with the original dataset resulted in a Hit@1 value of 34.6%.

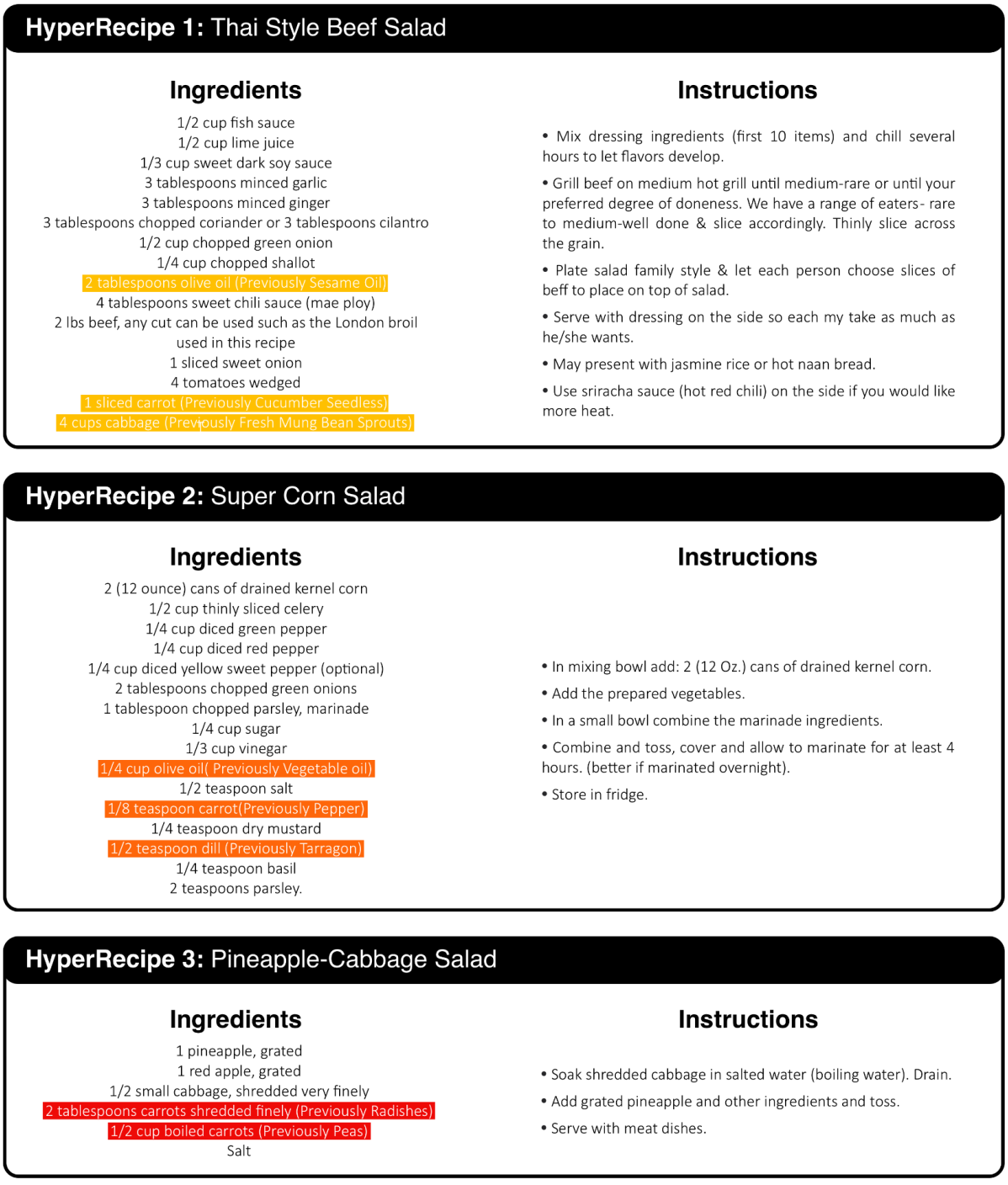
Subsequent experiments with fine-tuned LLMs on the original GISMo ingredient substitution dataset yielded Hit@1 values of 20.1 ± 0.3% for DaVinci-002, 20.9 ± 0.3% for TinyLlama-1.1B, and 38.0

± 0.3% for GPT-3.5-Turbo-1106. Using the filtered dataset in combination with these fine-tuned models resulted in improved Hit@1 values of 29.4 ± 0.3%, 34.5 ± 0.3%, and 54.5 ± 0.3%, respectively (Table 12).

Table 12 Performance of Fine-Tuned Models. Performance evaluation of fine-tuned models DaVinci-002, TinyLlama-1.1B, GPT-3.5-Turbo-1106, and GISMo, tested on both the original and filtered Recipe1MSubs datasets. The displayed Hit@1 accuracies are presented as mean ± standard deviation. The models are ordered in the table from best to worst performance based on their Hit@1 accuracy.

GISMo Dataset	Fine Tuned Model	Run 1	Run 2	Run 3	Run 4	Run 5	Final Hit@1 (%)
Filtered	GPT-3.5-Turbo-1106	54.1	54.8	54.7	54.4	54.4	54.5 ± 0.3
	GISMo	40.3	40.8	40.2	40.0	39.9	40.2 ± 0.4
	TinyLlama-1.1B	35.1	34.2	34.5	34.5	34.4	34.5 ± 0.3
	DaVinci-002	29.0	29.6	29.6	29.3	29.7	29.4 ± 0.3
Original	GPT-3.5-Turbo-1106	38.1	38.3	38.0	38.3	37.6	38.0 ± 0.3
	GISMo	34.6	34.4	34.7	34.5	34.5	34.5 ± 0.1
	TinyLlama-1.1B	20.4	20.8	20.4	20.2	20.2	20.4 ± 0.3
	DaVinci-002	20.4	19.7	19.8	20.3	20.3	20.1 ± 0.3

This process yielded 1,951 phytochemically enriched ingredient substitutions, resulting in 1,639 unique recipes, each incorporating at least one phytochemically rich ingredient. For example, the *Watercress Salad* was enriched with phytochemicals predicted *in silico* to target COVID-19; the *Kale and Quinoa Salad* was enhanced with phytochemicals targeting both COVID-19 and AD; and the *Thai-Style Beef Salad* was fortified with phytochemicals targeting COVID-19, AD, and cancer disease networks (Figure 19).



Source: Salvador et al, 2017

Figure 19 Phytochemically Enriched Recipes. Three examples showing the original and substituted ingredients considering the entire context of each recipe [366].

5.5 Discussion

We started by enhancing the GISMo model through the incorporation of an additional node feature—food categories for each ingredient, classified into one of the 23 categories utilised in FooDB, based on

classifications retrieved via the GPT-4. However, contrary to our expectations, this modification did not yield any improvements in the model's performance. This outcome may be attributed to several factors. Firstly, including this additional categorical information might have led to overfitting the model to the training data, compromising its ability to generalize effectively to unseen data in the test set. Another potential reason for the lack of observed benefit from the food category feature could be that part of the value of ingredient categorization might have been indirectly achieved by the model's consideration of ingredient co-occurrence in recipes alongside the presence of flavour molecules. These inherent features within the training data might already provide a substantial basis for the model to make effective substitution predictions without the need for explicit categorical labels. This suggests that the categorical information, in the way it was implemented, may not significantly enhance the model's predictive capabilities for ingredient substitution beyond what was achieved through analysing ingredient relationships and flavour profiles embedded in the recipe data.

While GISMo independently showcased a remarkable threefold enhancement in performance compared to prior methods [369], our approach to refine the GISMo model through the preliminary filtration of the original dataset via GPT-3.5's API further accentuated it in 17%. This filtration process involved the exclusion of *Potential* and *Incorrect* substitutions from the dataset, thereby ensuring a higher quality of data for model training. The filtration step encompassed five different datasets, and although only one was randomly selected to rerun GISMo, the improved results are generalizable across all due to their almost perfect ingredient substitute similarity across training, validation, and testing datasets. To demonstrate the consistency of our filtration process, here are some examples of substitutions that were consistently classified across the five runs:

- *Correct Substitutions*: Orange Juice to Pineapple Juice, Carrot to Red Pepper, Black Bean to Chickpea, Basil to Dried Oregano, Onion to Shallot.
- *Potential Substitutions*: Lemon to Orange, Apple to Peach, Apple to Apricot, Water to Wine, Blueberry to Strawberry.

- *Incorrect Substitutions*: Seedless Watermelon to Lime, Fresh Cilantro to Ground Coriander, Horseradish to Honey, Carrot to Seasoning Salt, Clove to Garlic.

Using GISMo original dataset, our experiments explored the potential benefits of fine-tuning DaVinci, TinyLlama and GPT-3.5. The first two models did not demonstrate any performance enhancements over the initial method. In contrast, the fine-tuned model leveraging the GPT-3.5 showed a 4% improvement in performance over the GISMo model. Building upon this Recipe1MSubs filtered dataset, we fine-tuned the same three models. Again, the GPT-3.5 model was the only one showing an increase in performance of 20% when compared with current state of the art. The findings of this study highlight the importance of data quality and model compatibility in the development of ingredient substitution algorithms. The higher performance achieved through the combination of GPT-3.5's advanced language processing capabilities and the GISMo model's framework highlights the potential of leveraging state-of-the-art AI technologies to enhance existing computational models without necessarily replacing them.

We specifically selected examples of recipes with ingredients phytochemically enriched targeting COVID-19; COVID-19 and AD; and COVID-19, AD and cancer molecular networks. In the previous order, those are *Watercress Salad*, *Kale and Quinoa Salad* and *Thai-Style Beef Salad* (Figure 19). We only considered salads in this analysis due to their minimal food processing steps. This choice was made because fewer processing steps generally help preserve the phytochemicals that are critical to the health benefits discussed. Salads undergo minimal thermal processing, which helps maintain the integrity of essential nutrients and active compounds compared to more extensively cooked dishes [368].

Fine-tuned models GISMo and TinyLlama, along with the scripts used for filtering the Recipe1MSubs dataset and fine-tuning the GISMo, DaVinci, GPT-3.5, and TinyLlama models, are hosted on the Bitbucket repository.

One inherent limitation is the diversity of the training datasets used to fine-tune the LLMs. Although these datasets are extensive, they may not fully capture the vast diversity of global cuisines and dietary

preferences, potentially impacting the model's ability to generalize across different culinary traditions and suggest culturally and regionally appropriate substitutions. Additionally, the methodology primarily focuses on textual data, which might not capture the full spectrum of culinary contexts, including taste profiles, textures, and the interplay of flavours. LLMs, while proficient in parsing and generating text, have limited capacity to understand and replicate the sensory experiences of cooking and eating. Additionally, the fine-tuning process, especially when using a limited set of high-quality substitutions, poses a risk of overfitting, where models may become overly specialized to the training data and less capable of generalizing to unseen recipes or ingredients. Furthermore, the reliance on the Hit@1 metric, while providing a clear measure of the model's ability to suggest the correct first substitution, does not capture the overall utility and flexibility of the model in providing a range of suitable alternatives. Moreover, the computational resources required for fine-tuning and deploying LLMs may also limit the accessibility of these advanced tools to researchers and practitioners with limited resources. Finally, while our machine-learning pipeline effectively enriches recipes with phytochemicals predicted to interact with COVID-19, AD and cancer networks, *in silico* enrichment does not guarantee that a single salad serving delivers bioactive doses. For example, 100g of raw watercress contains only 5-10mg of quercetin glycosides, and human absorption studies report around 20% bioavailability of quercetin glycosides, yielding plasma maximum concentration values of ~0.3–0.5 µM—near but not assured to reach around 0.5–1µM needed for antiviral or anti-inflammatory effects *in vitro* [370]. Quantitative HPLC–MS/MS analyses are therefore essential to confirm that theoretical enrichments translate into physiologically relevant concentrations. To establish actual biochemical activity, we propose two complementary validation pipelines. First, a cell-based assay workflow applying ethanolic extracts of each enriched salad to models such as SARS-CoV-2 pseudovirus entry in human airway epithelial cells, Aβ aggregation or tau phosphorylation in neuroblastoma cultures, and cytotoxicity assays in cancer cell lines (e.g., HCT116, A549) with downstream readouts including NF-κB/NLRP3 reporter assays, ELISA for cytokines, and viability assays (MTT, CellTiter-Glo). Second, a controlled human dietary crossover study in which volunteers consume one enriched salad per day, with pre- and

post-intervention plasma and urine sampled for targeted metabolomics of parent compounds (e.g., quercetin, kaempferol) and microbial metabolites (equol, urolithins), alongside inflammatory markers (CRP, IL-1β) and PBMC transcriptomics to directly link intake to in vivo pathway modulation [371]. Implementing these experimental pipelines would close the loop between our computational recipe enrichment and real-world biochemical activity, providing the rigorous evidence of efficacy that is essential for translational application.

5.6 Conclusion

By integrating state-of-the-art LLMs like OpenAI's GPT-3.5 and DaVinci, and Meta's TinyLlama, we have developed a method that enhances the accuracy and relevance of ingredient substitutions, producing recipes with phytochemically enriched ingredients. The fine-tuning of these models on Recipe1MSubs dataset allowed us to exceed the capabilities of traditional statistical and network-based methods, achieving a significant improvement in the Hit@1 metric in the original unfiltered and filtered datasets using fine-tuned GPT-3.5. Furthermore, by incorporating specific *in silico* phytochemical data, our approach led to the generation of phytochemically enriched ingredient substitutions and recipes targeting at least one of the following disease networks - COVID-19, AD and cancer. As we continue to refine the models and expand our datasets, we anticipate that incorporating domain-specific knowledge, such as clinical and biochemical data, will be crucial for enhancing the accuracy and relevance of ingredient substitutions. Future research should focus on rigorous validation of these substitutions through clinical trials and controlled dietary studies to assess their efficacy in improving health outcomes. These developments hold promise for revolutionizing personalized nutrition and optimizing dietary practices in a scientifically robust and clinically validated manner.

6. General Discussion

This study integrates machine learning and experimental validation to understand the complex interactions between drugs, phytochemicals, and disease pathophysiology. We explored how dietary components and pharmaceuticals can influence disease networks from AD, cancer, and COVID-19.

Initially, we used machine learning to identify FDA-approved drugs and EVOO phytochemicals likely to interact with genes and proteins associated with AD pathophysiology. By training a model to distinguish between FDA-approved drugs in phase 3 or 4 trials for AD and those approved for other conditions, we predicted which EVOO phytochemicals might influence the development or progression of AD. The correlation probability scores for the EVOO phytochemicals were comparable to those for FDA-approved drugs, suggesting that these phytochemicals interact with similar pathways involved in AD. Notably, compounds such as quercetin, genistein, apigenin, catechin, and kaempferol were identified, many of which have been previously shown to have neuroprotective effects in AD.

Building upon molecule discovery, we used a graph deep learning model based on the DreamLab DRUGS network propagation approach to predict PD interactions between phytochemicals and drugs. This model has a Pearson correlation score of 68.9% in predicting PD drug-drug interactions, highlighting the robustness of network-based approaches in capturing complex biological interactions. Experimental validation in human lung and colorectal cancer cell lines showed consistency with predicted outcomes in most tested combinations, confirming the model's reliability and applicability across different types of interactions. The higher variability observed in drug-phytochemical synergies underscores the complexity and less-understood nature of phytochemical interactions compared to drug-drug interactions.

Furthermore, we explored the enhancement of ingredient substitution models to generate phytochemically enriched recipes targeting specific disease molecular networks. By refining the GISMo model through dataset filtration based on substitution validity using GPT-3.5's API, we achieved a 17% improvement in performance. Fine-tuning GPT-3.5 on the filtered dataset led to a 20% improvement

compared to the current state-of-the-art, emphasizing the importance of data quality and model compatibility. The generated recipes, such as *Watercress Salad*, *Kale and Quinoa Salad*, and *Thai-Style Beef Salad*, were specifically designed to be rich in phytochemicals targeting AD, cancer, and COVID-19 molecular networks, demonstrating a practical application of our computational approach.

The integration of these methodologies highlights the power of combining computational predictions with experimental validation. Machine learning and deep learning approaches allowed for the identification of potential therapeutic agents among dietary phytochemicals, prediction of PD interactions between phytochemicals and drugs, and optimization of ingredient substitution for health-promoting recipes. These findings suggest that dietary components can play a significant role in disease prevention and management, potentially enhancing drug efficacy and minimizing toxicity through tailored dietary recommendations.

However, several limitations need to be acknowledged. The accuracy of our predictions is dependent on the quality and completeness of existing databases, which may lack critical interactions or omit important aspects of disease processes due to the unknown pathogenesis of conditions like AD and cancer. The reliance on *in vitro* data for model training and validation does not fully capture the complexity of *in vivo* interactions, and the models may not fully generalize across different culinary traditions or capture the sensory experiences of cooking and eating. Additionally, the computational resources required for fine-tuning and training may limit accessibility. As well, the concentration of phytochemicals in foods can vary widely based on cultivation practices, environmental conditions, storage, and preparation methods. This variability introduces uncertainty in estimating the actual intake of phytochemicals through dietary interventions, complicating the assessment of their health benefits. Moreover, ethical and regulatory considerations regarding the use of phytochemicals as therapeutic agents need to be addressed. While many phytochemicals are generally recognized as safe, their effects in combination with pharmaceuticals are not fully understood. Potential interactions could lead to adverse effects or diminished therapeutic efficacy. Rigorous clinical trials are necessary to evaluate the safety, efficacy, and appropriate dosing of phytochemicals, especially when used alongside

conventional medications. Lastly, the predictive models may not account for individual differences in genetics, metabolism, and microbiome composition. Personalized responses to drugs and dietary components are influenced by a multitude of factors unique to everyone. Our models, which are based on generalized data, may not accurately predict outcomes at the individual level.

7. General Conclusion

We considered three nutritional levels in this work: molecular, ingredient and recipe. By harnessing advanced machine learning techniques—including network-based machine learning, deep learning, and large language models—this research identified and optimized bioactive phytochemicals in food targeting cancer, COVID-19 and AD disease networks.

Firstly, a network-based machine learning model was developed to identify bioactive phytochemicals in EVOO with potential therapeutic effects against AD. This model ranked ten EVOO phytochemicals likely to influence AD pathways, showing how dietary phytochemicals can modulate disease-related protein networks.

Secondly, a novel graph deep learning model was introduced to predict PD interactions between drugs and phytochemicals. Experimentally validated in lung and colorectal cancer cell lines, this model demonstrated that systems trained on drug-drug interaction data can effectively predict phytochemical-drug interactions.

Lastly, we applied large language models to optimize ingredient substitutions in recipes, resulting in the creation of phytochemically enriched recipes. By fine-tuning these models on extensive datasets, improvements in substitution accuracy were achieved, leading to the generation of numerous health-promoting dietary options.

The methodologies and findings presented have profound implications for both the scientific community and practical health applications. Integrating AI-driven approaches with nutritional science provides a powerful tool for identifying and optimizing bioactive compounds in food, paving the way for personalized medicine and nutrition. These innovations have the potential to revolutionize dietary planning, enhance therapeutic outcomes, and contribute significantly to the emerging field of computational gastronomy.

Future research should extend these models to a broader spectrum of diseases and dietary components, enabling a more comprehensive understanding of the interplay between diet and health.

Conducting large-scale clinical trials will be imperative to validate the therapeutic potential of the identified phytochemicals and their interactions with conventional drugs. Furthermore, improving the accessibility and user-friendliness of AI-driven tools for personalized nutrition and therapeutic optimization will be essential for widespread adoption.

8. Appendix to Chapter 3

8.1 List of Dysregulated Genes in Alzheimer's Disease

Table 13 Dysregulated genes in AD. List of genes including both the full gene name and its corresponding symbol.

Gene Full Name	Gene Symbol
24-Dehydrocholesterol Reductase	DHCR24
Acetylcholinesterase	ACHE
Activity Regulated Cytoskeleton Associated Protein	ARC
Alpha-2-Macroglobulin	A2M
Amyloid Beta Precursor Protein	APP
Amyloid Beta Precursor Protein Binding Family B Member 2	APBB2
Angiotensin I Converting Enzyme	ACE
Apolipoprotein E	APOE
ATP Binding Cassette Subfamily A Member 7	ABCA7
ATP Synthase F1 Subunit Alpha	ATP5F1A
Autocrine Motility Factor Receptor	AMFR
Beta-Secretase 1	BACE1
Bleomycin Hydrolase	BLMH
Bridging Integrator 1	BIN1
Butyrylcholinesterase	BCHE
Calmodulin 1	CALM1
Caspase 3	CASP3
CD2 Associated Protein	CD2AP

CD33 Molecule	CD33
Cholinergic Receptor Nicotinic Alpha 7 Subunit	CHRNA7
Cholinergic Receptor Nicotinic Beta 2 Subunit	CHRNB2
Clusterin	CLU
Coagulation Factor II (Thrombin)	F2
Complement C3b/C4b Receptor 1 (Knops Blood Group)	CR1
Cytochrome P450 Family 2 Subfamily D Member 6	CYP2D6
Cytochrome P450 Family 46 Subfamily A Member 1	CYP46A1
Dihydropyrimidinase Like 2	DPYSL2
Enolase 1	ENO1
EPH Receptor A1	EPHA1
Estrogen Receptor 1	ESR1
Eukaryotic Translation Initiation Factor 2 Subunit Alpha	EIF2S1
GATA Binding Protein 1	GATA1
Glyceraldehyde-3-Phosphate Dehydrogenase, Spermatogenic	GAPDHS
Glycogen Synthase Kinase 3 Beta	GSK3B
Heme Oxygenase 1	HMOX1
Homeostatic Iron Regulator	HFE
Insulin	INS
Insulin Degrading Enzyme	IDE
Insulin Like Growth Factor 1	IGF1
Insulin Like Growth Factor 1 Receptor	IGF1R
Insulin Like Growth Factor 2	IGF2
Insulin Like Growth Factor 2 Receptor	IGF2R
Insulin Receptor	INSR

Interleukin 1 Beta	IL1B
Methylenetetrahydrofolate Reductase	MTHFR
Microtubule Associated Protein Tau	MAPT
Monoamine Oxidase B	MAOB
Myeloperoxidase	MPO
Nicastrin	NCSTN
Nitric Oxide Synthase 3	NOS3
PAX Interacting Protein 1	PAXIP1
Peptide YY	PYY
Peroxisome Proliferator Activated Receptor Gamma	PPARG
Phosphatidylinositol Binding Clathrin Assembly Protein	PICALM
Plasminogen Activator, Urokinase	PLAU
Presenilin 1	PSEN1
Presenilin 2	PSEN2
Prion Protein	PRNP
Protocadherin 11 X-Linked	PCDH11X
Reelin	RELN
Solute Carrier Family 2 Member 4	SLC2A4
Solute Carrier Family 30 Member 4	SLC30A4
Solute Carrier Family 30 Member 6	SLC30A6
Sortilin Related Receptor 1	SORL1
Superoxide Dismutase 2	SOD2
Transcription Factor A, Mitochondrial	TFAM
Transferrin	TF
Triggering Receptor Expressed On Myeloid Cells 2	TREM2

Triosephosphate Isomerase 1	TPI1
Tripeptidyl Peptidase 1	TPP1
Tumor Necrosis Factor	TNF
Vascular Endothelial Growth Factor A	VEGFA
Visinin Like 1	VSNL1

8.2 Comprehensive List of Phytochemicals in Extra Virgin Olive Products

List of 782 phytochemicals present in extra virgin olive products, including EVOO, table olives, and olive leaves. It includes both EVOO markers and non-markers. Accessible here.

8.3 Predicted Extra Virgin Olive Oil Marker Phytochemicals Influencing Alzheimer's Disease Onset and Progression

67 EVOO marker phytochemicals predicted and ranked according to the highest probability of influencing AD onset and/or progression. Accessible here.

8.4 Pathways Targeted by Phytochemicals in Extra Virgin Olive Oil Associated with Alzheimer's Disease

Examination of the pathways targeted by phytochemicals in EVOO, which are associated with AD and have been highly predicted by our network propagation model.

8.4.1 Alzheimer's Disease Pathway

One of the most important markers of AD (as illustrated in the KEGG pathway dataset) include amyloid-beta production and aggregation, tau hyperphosphorylation, and neuroinflammation.

Amyloid-Beta Production and Aggregation: AD is characterized by the accumulation of amyloid-beta (Aβ) peptides, which form extracellular plaques. Aβ peptides are generated from the amyloid precursor protein (APP) by sequential cleavage via beta-secretase (BACE1) and gamma-secretase complex

(composed of presenilin, nicastrin, APH1, and PEN2). Mutations in APP or presenilin genes can lead to increased Aβ production, contributing to AD [113,372].

Tau Hyperphosphorylation: Intracellular neurofibrillary tangles, composed of hyperphosphorylated tau protein, are another marker of AD. Tau is a microtubule-associated protein, and its hyperphosphorylation destabilizes microtubules, leading to impaired axonal transport and neuronal dysfunction. Dysregulation of kinases, such as glycogen synthase kinase-3β (GSK-3β) and cyclin-dependent kinase 5 (CDK5), and phosphatases contributes to tau hyperphosphorylation in AD [373,374].

Neuroinflammation: In AD, activation of microglia and astrocytes leads to chronic neuroinflammation. Microglia, the resident immune cells of the brain, can become overactivated and release pro-inflammatory cytokines, such as interleukin-1β (IL-1β), tumour necrosis factor-alpha (TNF-α), and interleukin-6 (IL-6), which contribute to neuronal dysfunction and death [375,376].

Table 14 Key Proteins/Genes and Complexes Involved in the AD Pathway. List of the primary proteins and genes associated with the AD pathway, including their symbols and functions.

Protein/Gene	Symbol	Function
Amyloid precursor protein	APP	Cleavage of APP by β-secretase and γ-secretase produces amyloid-β (Aβ) peptides, which aggregate to form amyloid plaques, a hallmark of AD.
Presenilin 1	PSEN1	Catalytic component of the γ-secretase complex, responsible for the cleavage of APP to generate Aβ peptides.
Presenilin 2	PSEN2	Catalytic component of the γ-secretase complex, responsible for the cleavage of APP to generate Aβ peptides.
Beta-site APP cleaving enzyme 1	BACE1	Also known as β-secretase, BACE1 is an enzyme involved in the cleavage of APP, producing Aβ peptides.
Apolipoprotein E	APOE	APOE is involved in lipid metabolism and is associated with an increased risk of developing AD, especially the APOE ε4 allele.

Tau protein	MAPT	Hyperphosphorylated tau protein forms neurofibrillary tangles, another hallmark of AD, leading to neuronal dysfunction and cell death.
Glycogen synthase kinase 3 beta	GSK3B	Involved in tau protein hyperphosphorylation, contributing to the formation of neurofibrillary tangles.

8.4.2 Olfactory Transduction Pathway

The connection between olfactory transduction and AD is not direct, but there is evidence to suggest that olfactory dysfunction may be an early symptom of AD and other neurodegenerative diseases. The reason behind might be the presence of pathological hallmarks in the olfactory system, as well as the proximity of the olfactory bulb to brain regions affected by AD.

Olfactory Dysfunction: Individuals with AD often experience a decline in their ability to smell. Olfactory dysfunction can precede cognitive impairment and memory loss and may serve as an early sign for AD [377,378].

Presence of AD pathology in Olfactory Regions: The olfactory bulb and olfactory cortex are among the first regions in the brain to be affected by AD pathology, such as amyloid-beta plaques and neurofibrillary tangles. This may contribute to the observed olfactory dysfunction in AD patients [379].

Proximity to Brain Regions Affected by AD: The olfactory bulb is anatomically close to the entorhinal cortex and hippocampus, which are critical for memory formation and are among the first regions to be affected in AD [380].

Table 15 Key Proteins/Genes and Complexes Involved in the Olfactory Transduction Pathway. Proteins/genes involved in the olfactory transduction pathway, including their symbols and functions.

Protein/Gene	Symbol	Function
Odorant molecules	-	Bind to olfactory receptors on the surface of olfactory receptor neurons in the nose.

Olfactory receptors	ORs (OR8J1)	G-protein coupled receptors (GPCRs) that bind odorant molecules and activate the downstream signalling cascade.
G-protein (Golf)	GNAL	Activates adenylate cyclase upon binding to an activated olfactory receptor.
Adenylate cyclase 3	ADCY3	Converts ATP to cyclic AMP (cAMP) in response to G-protein activation.
Cyclic AMP (cAMP)	-	Second messenger molecule that activates cyclic nucleotide-gated ion channels.
Cyclic nucleotide-gated ion channels	CNGA2, CNGA4, CNGB1	Non-selective cation channels that open in response to cAMP, allowing the influx of Na+ and Ca2+ ions, leading to membrane depolarization.
Ca^{2+}-activated Cl^- channels	ANO2	Activated by the influx of Ca^{2+} ions, mediating Cl^- efflux to amplify the receptor potential.
Inward rectifier K^+ channels	KCNJ2, KCNJ4, KCNJ14	Help to repolarize the membrane potential, terminating the olfactory response.

8.4.3 Insulin Signalling Pathway

The connection between the insulin signalling pathway and AD has been increasingly studied. AD has even been referred to as *type 3 diabetes* due to the involvement of insulin resistance and glucose metabolism dysfunction in the brain. The main connections between the insulin signalling pathway and AD are as follows.

Insulin resistance: Insulin resistance in the brain impairs insulin signalling, leading to reduced glucose uptake and energy metabolism. This can contribute to neuronal dysfunction and synaptic loss, which are characteristic of AD [281].

Amyloid-beta (Aβ) Metabolism: Impaired insulin signalling may affect the clearance and accumulation of Aβ peptides, which are the main components of amyloid plaques in AD. Insulin has been shown to regulate the expression and activity of enzymes responsible for Aβ production and degradation [381].

Tau Hyperphosphorylation: Insulin signalling can modulate tau phosphorylation through the activation of protein kinase B (Akt). Impaired insulin signalling may lead to the dysregulation of Akt and other

kinases, promoting tau hyperphosphorylation and the formation of neurofibrillary tangles, another hallmark of AD [382].

Inflammation: Impaired insulin signalling can contribute to chronic inflammation in the brain, which has been implicated in the progression of AD. Inflammation can exacerbate Aβ accumulation and tau hyperphosphorylation, leading to further neuronal dysfunction and cell death [383].

Neuronal Survival and Synaptic Plasticity: Insulin signalling plays a crucial role in maintaining neuronal survival and promoting synaptic plasticity. Impaired insulin signalling in the brain can lead to reduced neurotrophic support, increased oxidative stress, and impaired synaptic function, all of which contribute to the cognitive decline observed in AD [384].

Table 16 Key Proteins/Genes and Complexes Involved in the Insulin Signalling Pathway. This table details the critical proteins and genes involved in the insulin signalling pathway, including their symbols and functions.

Protein/Gene	Symbol	Function
Insulin	INS	Hormone that binds to the insulin receptor, initiating the insulin signalling cascade.
Insulin receptor	INSR	Receptor tyrosine kinase that is activated upon insulin binding, autophosphorylates, and triggers downstream signalling.
Insulin receptor substrate 1-4	IRS1, IRS2, IRS3, IRS4	Adaptor proteins that, when phosphorylated by the insulin receptor, serve as docking sites for downstream signalling molecules.
Phosphoinositide 3-kinase	PIK3R3	A lipid kinase that generates phosphatidylinositol (3,4,5)-trisphosphate (PIP3) upon activation by IRS proteins.
Phosphatase and tensin homolog	PTEN	A lipid phosphatase that negatively regulates the insulin signalling pathway by dephosphorylating PIP3.
Protein kinase B (Akt)	AKT1, AKT2, AKT3	Serine/threonine kinase activated by PIP3, responsible for multiple downstream effects, including glucose uptake and glycogen synthesis.

Glycogen synthase kinase-3	GSK3A, GSK3B	A kinase that phosphorylates and inhibits glycogen synthase; Akt inactivates GSK-3, allowing glycogen synthesis.
Mammalian target of rapamycin	mTOR	A serine/threonine kinase that regulates protein synthesis, cell growth, and metabolism; activated by Akt.
Glucose transporter type 4	GLUT4	A glucose transporter that is translocated to the cell membrane in response to insulin signalling, allowing glucose uptake.
Glycogen synthase	GYS1, GYS2	Enzyme responsible for glycogen synthesis; its activity is regulated by the insulin signalling pathway.

8.4.4 Phosphatidylinositol Signalling System Pathway

The phosphatidylinositol signalling system has been implicated in AD through several mechanisms. Dysfunction in the phosphatidylinositol signalling pathway may contribute to the development and progression of AD. The main connections between the phosphatidylinositol signalling system and AD are as follows [385].

Aβ Production and Clearance: Phosphatidylinositol signalling, particularly PI3K/Akt/mTOR pathway, has been implicated in regulating amyloid precursor protein (APP) processing and Aβ production. Dysregulation of this pathway can affect APP metabolism, leading to increased Aβ production and accumulation of amyloid plaques in the brain [386].

Tau Phosphorylation: The PI3K/Akt/GSK-3β signalling axis is involved in the regulation of tau phosphorylation. Impairment of the phosphatidylinositol signalling pathway can lead to dysregulation of Akt and GSK-3β, promoting tau hyperphosphorylation and the formation of neurofibrillary tangles, another pathological hallmark of AD [387].

Synaptic Function and Plasticity: Phosphatidylinositol signalling plays a role in the regulation of synaptic function and plasticity. Alterations in phosphatidylinositol signalling can impair synaptic transmission, leading to synaptic loss and cognitive decline in AD [388].

Neuroinflammation: Dysregulation of the phosphatidylinositol signalling pathway can contribute to neuroinflammation, a process that has been implicated in the progression of AD. Inflammatory responses can exacerbate Aβ accumulation and tau hyperphosphorylation, leading to further neuronal dysfunction and cell death [389].

Neuronal survival and apoptosis: Phosphatidylinositol signalling, particularly the PI3K/Akt pathway, is crucial for neuronal survival and protection against apoptosis. Dysfunction in this pathway can lead to increased neuronal vulnerability and cell death, contributing to neurodegeneration in AD [390,391].

Table 17 Key Proteins/Genes and Complexes Involved in the Phosphatidylinositol Signalling System Pathway. Proteins and genes involved in the phosphatidylinositol signalling system, presenting their symbols and functions.

Protein/Gene	Symbol	Function
Phosphatidylinositol kinase	PI4KA	Enzymes responsible for the phosphorylation of phosphatidylinositol (PI) to generate PI(4)P and PI(4,5)P2.
Phosphoinositide 3-kinase	PIK3R3	A lipid kinase that generates phosphatidylinositol (3,4,5)-trisphosphate (PIP3) by phosphorylating PI(4,5)P2.
Phospholipase C	PLCB1	Enzyme that hydrolyses PI(4,5)P2 to generate inositol 1,4,5-trisphosphate (IP3) and diacylglycerol (DAG), initiating downstream signalling.
Inositol 1,4,5-trisphosphate receptor	ITPR3	Calcium channel that releases calcium from the endoplasmic reticulum upon binding IP3, initiating calcium signalling.
Diacylglycerol kinase	DGKZ	Enzyme that phosphorylates DAG to generate phosphatidic acid (PA), modulating DAG signalling.
Protein kinase C	PRKCG	A family of serine/threonine kinases activated by DAG and calcium, involved in the regulation of numerous cellular processes.

Phosphatase and tensin homolog	PTEN	A lipid phosphatase that negatively regulates the phosphatidylinositol signalling system by dephosphorylating PIP3.
Protein kinase B (Akt)	AKT1, AKT2, AKT3	Serine/threonine kinase activated by PIP3, involved in cell survival, growth, and metabolism.
Phospholipase D	PLD1	Enzyme that hydrolyses PI to generate PA and inositol, involved in various cellular processes, such as vesicle trafficking and signal transduction.

8.4.5 Vascular Smooth Muscle Contraction Pathway

While the vascular smooth muscle contraction pathway is not directly involved in the classic AD pathways, it has been associated with AD. The main connections between this pathway and AD are the following [392].

Cerebral Blood Flow and Vascular Dysfunction: Vascular smooth muscle cells regulate cerebral blood flow by controlling the diameter of blood vessels. In AD, vascular dysfunction and impaired cerebral blood flow have been observed, which can contribute to neuronal dysfunction, reduced glucose and oxygen supply, and cognitive decline. The dysfunction of proteins involved in vascular smooth muscle contraction can contribute to this impaired blood flow regulation in AD [393].

Blood-Brain Barrier (BBB) Integrity: Vascular smooth muscle cells are a component of the neurovascular unit, which maintains the integrity of the blood-brain barrier. BBB dysfunction has been observed in AD, which may lead to increased infiltration of peripheral immune cells and neurotoxic substances into the brain, exacerbating inflammation and neuronal damage. Abnormal vascular smooth muscle contraction may contribute to this BBB dysfunction in AD [394].

Cerebral amyloid angiopathy (CAA): CAA is characterized by the accumulation of amyloid-beta (Aβ) in the walls of cerebral blood vessels. This accumulation can lead to vessel wall thickening, reduced vessel elasticity, and impaired vascular smooth muscle function, contributing to cerebral blood flow dysregulation in AD. Aβ deposition in the vascular smooth muscle cells can disrupt the vascular smooth muscle contraction pathway and contribute to vascular dysfunction in AD [395].

Hypertension and Cardiovascular Risk Factors: The dysfunction of the vascular smooth muscle contraction pathway can contribute to hypertension, which is a known risk factor for AD. Long-term hypertension can lead to vascular damage, inflammation, and impaired cerebral blood flow regulation, all of which may contribute to the development or progression of AD [396].

Table 18 Key Proteins/Genes and Complexes Involved in the Vascular Smooth Muscle Contraction Pathway. List of important proteins/genes involved in the regulation of vascular smooth muscle contraction, presenting their symbols and functions.

Protein/Gene	Symbol	Role in AD Pathway
G protein-coupled receptors	GNA12	Receptors that bind various ligands, initiating intracellular signalling cascades that modulate smooth muscle contraction.
Heterotrimeric G proteins	GNA12	Transduce signals from GPCRs to downstream effectors, such as phospholipase C, adenylate cyclase, and ion channels.
Phospholipase C	PLCB1	Enzyme that hydrolyses phosphatidylinositol 4,5-bisphosphate (PIP2) to generate inositol 1,4,5-trisphosphate (IP3) and diacylglycerol (DAG).
Inositol 1,4,5-trisphosphate receptor	ITPR3	Calcium channel that releases calcium from the sarcoplasmic reticulum upon binding IP3, increasing intracellular calcium levels.
Diacylglycerol	DAG1	Activates protein kinase C (PKC), which phosphorylates various proteins involved in smooth muscle contraction.
Protein kinase C	PRKCG	Serine/threonine kinase that is activated by DAG and calcium, regulating proteins involved in smooth muscle contraction.
Voltage-gated calcium channels	CACNA1C	Channels that allow calcium influx in response to membrane depolarization, contributing to smooth muscle contraction.
Calcium/calmodulin complex	CALM1	Calcium-binding protein that, upon calcium binding, activates myosin light chain kinase (MLCK) and other target proteins.

Myosin light chain kinase	MYLK2	Enzyme that phosphorylates myosin light chain (MLC) upon activation by calcium/calmodulin, promoting smooth muscle contraction.
Myosin light chain phosphatase	PPP1R12A	Enzyme that dephosphorylates MLC, promoting smooth muscle relaxation.
Myosin light chain	MYL1	Regulatory protein of myosin that, when phosphorylated, initiates the cross-bridge cycle and muscle contraction.

8.5 Human Genomics (Published)

Rita, L., Neumann, N.R., Laponogov, I. et al. Alzheimer's disease: using gene/protein network machine learning for molecule discovery in olive oil. Hum Genomics 17, 57 (2023). https://doi.org/10.1186/s40246-023-00503-6

Rita *et al. Human Genomics* (2023) 17:57
https://doi.org/10.1186/s40246-023-00503-6

Human Genomics

RESEARCH Open Access

Alzheimer's disease: using gene/protein network machine learning for molecule discovery in olive oil

Luís Rita[1†], Natalie R. Neumann[2†], Ivan Laponogov[1], Guadalupe Gonzalez[3,4], Dennis Veselkov[1], Domenico Pratico[5], Reza Aalizadeh[7], Nikolaos S. Thomaidis[7], David C. Thompson[6], Vasilis Vasiliou[6*] and Kirill Veselkov[1,6*]

Abstract

Alzheimer's disease (AD) poses a profound human, social, and economic burden. Previous studies suggest that extra virgin olive oil (EVOO) may be helpful in preventing cognitive decline. Here, we present a network machine learning method for identifying bioactive phytochemicals in EVOO with the highest potential to impact the protein network linked to the development and progression of the AD. A balanced classification accuracy of 70.3 ± 2.6% was achieved in fivefold cross-validation settings for predicting late-stage experimental drugs targeting AD from other clinically approved drugs. The calibrated machine learning algorithm was then used to predict the likelihood of existing drugs and known EVOO phytochemicals to be similar in action to the drugs impacting AD protein networks. These analyses identified the following ten EVOO phytochemicals with the highest likelihood of being active against AD: quercetin, genistein, luteolin, palmitoleate, stearic acid, apigenin, epicatechin, kaempferol, squalene, and daidzein (in the order from the highest to the lowest likelihood). This in silico study presents a framework that brings together artificial intelligence, analytical chemistry, and omics studies to identify unique therapeutic agents. It provides new insights into how EVOO constituents may help treat or prevent AD and potentially provide a basis for consideration in future clinical studies.

Keywords Olive oil, Alzheimer's disease, Network propagation, Nutrition

†Luís Rita and Natalie Neumann contributed equally to this work.

*Correspondence:
Vasilis Vasiliou
vasilis.vasiliou@yale.edu
Kirill Veselkov
kirill.veselkov04@imperial.ac.uk
[1] Division of Cancer, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London, UK
[2] Department of Emergency Medicine, Yale School of Medicine, New Haven, CT, USA
[3] Department of Computing, Faculty of Engineering, Imperial College London, London, UK
[4] Prescient Design, Genentech | Roche, Basel, Switzerland
[5] Alzheimer's Center at Temple, Lewis Katz School of Medicine, Temple University, Philadelphia, PA, USA
[6] Department of Environmental Health Sciences, Yale University, New Haven, CT, USA
[7] Laboratory of Analytical Chemistry, Department of Chemistry, National and Kapodistrian University of Athens, Panepistimiopolis Zografou, 15771 Athens, Greece

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

9. Appendix to Chapter 4

9.1 Food-Drug Computational Synergies in A549

Table 19 Top 20 Predicted Drug-Food Synergies in A549. Top 20 synergies between drugs and phytochemicals as predicted by the GNN model. The table includes the ranking of each drug-phytochemical pair, along with their average synergy score and standard deviation.

Ranking	Drug	Phytochemical	Average Synergy	Standard Deviation
1	**Gemcitabine**	**Thymoquinone**	**13.8**	**11.1**
2	AZD1775	5-Phenylvaleric acid	13.7	8.2
3	AZD1775	D-Tartaric acid	13.6	8.8
4	AZD1775	Metatartaric acid	13.6	8.8
5	AZD1775	Thymoquinone	13.4	8.2
6	MK-8776	5-Phenylvaleric acid	12.6	8.6
7	**MK-8776**	**D-Tartaric acid**	**12.2**	**8.7**
8	MK-8776	Metatartaric acid	12.2	8.7
9	**MK-8776**	**Thymoquinone**	**12.2**	**7.2**
10	AZD1775	dUTP	12.0	8.9
11	AZD1775	dTDP	11.9	7.8
12	AZD1775	Deoxycytidine	11.6	5.6
13	AZD1775	fumarate	11.6	7.1
14	AZD1775	Maleic acid	11.6	7.1
15	AZD1775	Chondroitin sulfate	11.6	5.8
16	Gemcitabine	5-Phenylvaleric acid	11.1	10.2
17	Gemcitabine	13S-Hydroxyoctadecadienoic acid	11.0	11.5

18	**Gemcitabine**	**D-Tartaric acid**	**10.5**	**8.4**
19	Gemcitabine	Metatartaric acid	10.5	8.4
20	Gemcitabine	Deoxycytidine	10.2	7.9

9.2 Food-Drug Computational Synergies in HCT116

Table 20 Top Predicted Drug-Food Synergies in HCT116. Top 20 synergies between drugs and phytochemicals as predicted by the GNN model. The table includes the ranking of each drug-phytochemical pair, along with their average synergy score and standard deviation.

Ranking	Drug	Phytochemical	Average Synergy	Standard Deviation
1	Sunitinib	Cholic acid	15.9	7.3
2	Sunitinib	Ursocholic acid	15.9	7.4
3	Sunitinib	3b,7b,12a-Trihydroxy-5a-Cholanoic acid	15.9	7.4
4	Sunitinib	Pentadecanoic acid	15.2	6.1
5	Sunitinib	Centoic acid	15.1	5.9
6	Sunitinib	3,4,5,4'-Tetramethoxystilbene	14.9	4.1
7	Sunitinib	propanoate	14.8	6.2
8	Sunitinib	Lipoxin A4	14.8	10.5
9	Sunitinib	Propranolol	14.8	4.9
10	Sunitinib	epsilon-Polylysine	14.7	7.8
11	AZD1775	5-Phenylvaleric acid	13.2	6.7
12	MK-8776	5-Phenylvaleric acid	13.1	7.1
13	**MK-8776**	**Thymoquinone**	**12.9**	**6.1**
14	AZD1775	Thymoquinone	12.6	5.8
15	AZD1775	7alpha,24-Dihydroxy-4-cholesten-3-one	12.5	6.5

16	AZD1775	7alpha,26-Dihydroxy-4-cholesten-3-one	12.3	6.5
17	AZD1775	7alpha,27-Dihydroxycholesterol	12.2	6.2
18	AZD1775	Thiosulfate	12.1	6.0
19	AZD1775	27-Hydroxycholesterol	12.0	5.0
20	AZD1775	Molybdopterin	11.9	5.2

10. Appendix to Chapter 5

10.1 Classification of Ingredients in the Recipe1MSubs Dataset

All ingredients in the Recipe1MSubs dataset were classified into one of the following 23 food categories as identified in the FooDB database [200]:

1. Herbs and Spices
2. Fats and Oils
3. Unclassified
4. Baby Foods
5. Snack Foods
6. Dishes
7. Baking Goods
8. Confectioneries
9. Eggs
10. Milk and Milk Products
11. Animal Foods
12. Aquatic Foods
13. Beverages
14. Cocoa and Cocoa Products
15. Soy
16. Coffee and Coffee Products
17. Gourds
18. Teas
19. Pulses

20. Cereals and Cereal Products
21. Nuts
22. Fruits
23. Vegetables

10.2 MDPI MAKE (Published)

Rita, L., Southern, J., Laponogov, I., Higgins, K., & Veselkov, K. (2024). Optimizing Ingredient Substitution Using Large Language Models to Enhance Phytochemical Content in Recipes. *Machine Learning and Knowledge Extraction*, *6*(4), 2738-2752. https://doi.org/10.3390/make6040131

machine learning & knowledge extraction

MDPI

Article

Optimizing Ingredient Substitution Using Large Language Models to Enhance Phytochemical Content in Recipes

Luís Rita [1], Joshua Southern [2], Ivan Laponogov [1], Kyle Higgins [1,3] and Kirill Veselkov [1,4,*]

[1] Division of Cancer, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London W12 0NN, UK; l.rita19@imperial.ac.uk (L.R.); i.laponogov@imperial.ac.uk (I.L.); kyle.higgins@childrens.harvard.edu (K.H.)
[2] Department of Computing, Faculty of Engineering, Imperial College London, London SW7 2AZ, UK; joshua.southern17@imperial.ac.uk
[3] Department of Neurobiology, Boston Children's Hospital and Harvard Medical School, Boston, MA 02115, USA
[4] Department of Environmental Health Sciences, Yale University, New Haven, CT 06520, USA
* Correspondence: kirill.veselkov04@imperial.ac.uk

Abstract: In the emerging field of computational gastronomy, aligning culinary practices with scientifically supported nutritional goals is increasingly important. This study explores how large language models (LLMs) can be applied to optimize ingredient substitutions in recipes, specifically to enhance the phytochemical content of meals. Phytochemicals are bioactive compounds found in plants, which, based on preclinical studies, may offer potential health benefits. We fine-tuned models, including OpenAI's GPT-3.5-Turbo, DaVinci-002, and Meta's TinyLlama-1.1B, using an ingredient substitution dataset. These models were used to predict substitutions that enhance the phytochemical content and to create a corresponding enriched recipe dataset. Our approach improved the top ingredient prediction accuracy on substitution tasks, from the baseline 34.53 ± 0.10% to 38.03 ± 0.28% on the original substitution dataset and from 40.24 ± 0.36% to 54.46 ± 0.29% on a refined version of the same dataset. These substitutions led to the creation of 1951 phytochemically enriched ingredient pairings and 1639 unique recipes. While this approach demonstrates potential in optimizing ingredient substitutions, caution must be taken when drawing conclusions about health benefits, as the claims are based on preclinical evidence. This research represents a step forward in using AI to promote healthier eating practices, providing potential pathways for integrating computational methods with nutritional science.

Keywords: ingredient substitution; nutritional optimization; large language models

check for updates

Citation: Rita, L.; Southern, J.; Laponogov, I.; Higgins, K.; Veselkov, K. Optimizing Ingredient Substitution Using Large Language Models to Enhance Phytochemical Content in Recipes. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 2738–2752. https://doi.org/10.3390/make6040131

Received: 29 September 2024
Revised: 16 November 2024
Accepted: 18 November 2024
Published: 26 November 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

In recent years, computational gastronomy has emerged as an interdisciplinary field that applies computational techniques such as data mining and machine learning to the study of food and cooking. The aim is to understand and model the complex interactions between ingredients, cooking methods, and the human perception of taste and nutrition. One of the primary goals of this field is to develop methods for ingredient substitution, aimed at improving the nutritional content, preserving the flavor integrity, and aligning meals with specific dietary needs. A key focus has been the integration of phytochemically enriched ingredients into diets, which has shown in silico potential to target biological networks of chronic diseases like cancer [1], Alzheimer's disease (AD) [2], and COVID-19 [3].

Phytochemicals, bioactive compounds found in plants, have gathered significant attention due to their antioxidant, anti-inflammatory, and anti-carcinogenic properties. Preclinical studies suggest that these compounds may play a role in disease prevention and treatment. For instance, brassinolide, a phytochemical present in tea, has shown potential to inhibit tumor growth and induce apoptosis in cancer cells [4]. In the context of AD,

References

1. Murphy K. Machine Learning: A Probabilistic Perspective. The MIT Press; 2012.

2. Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. Nat Rev Genet. 2017;18:551–62.

3. Khatri P, Sirota M, Butte AJ. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Comput Biol. 2012;8:e1002375.

4. Navlakha S, Kingsford C. The power of protein interaction networks for associating genes with diseases. Bioinformatics. 2010;26:1057–63.

5. Yeger-Lotem E, Riva L, Su LJ, Gitler AD, Cashikar AG, King OD, et al. Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. Nat Genet. 2009;41:316–23.

6. Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge J V, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet. 2015;47:106–14.

7. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12:56–68.

8. van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression analysis for functional classification and gene–disease predictions. Brief Bioinform. 2017;bbw139.

9. Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, et al. Identification of influential spreaders in complex networks. Nat Phys. 2010;6:888–93.

10. Newman M. Networks. Oxford University Press; 2010.

11. Albert R, Jeong H, Barabási A-L. Error and attack tolerance of complex networks. Nature. 2000;406:378–82.

12. Zhou T, Lü L, Zhang Y-C. Predicting missing links via local information. Eur Phys J B. 2009;71:623–30.

13. Kloster K, Gleich DF. Heat kernel based community detection. 2014;

14. Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007;23:1274–81.

15. Jeong H, Mason SP, Barabási A-L, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001;411:41–2.

16. Ma H, Schadt EE, Kaplan LM, Zhao H. COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method. Bioinformatics. 2011;27:1290–8.

17. Schulz MH, Devanny WE, Gitter A, Zhong S, Ernst J, Bar-Joseph Z. DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data. BMC Syst Biol. 2012;6:104.
18. Li Y, Patra JC. Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. Bioinformatics. 2010;26:1219–24.
19. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Mol Syst Biol. 2007;3.
20. Kolar M, Xing EP. Estimating networks with jumps. Electron J Stat. 2012;6.
21. Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online Learning of Social Representations. 2014;
22. Grover A, Leskovec J. node2vec. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2016. p. 855–64.
23. Zhang C, Butepage J, Kjellstrom H, Mandt S. Advances in Variational Inference. IEEE Trans Pattern Anal Mach Intell. 2019;41:2008–26.
24. Wang D, Cui P, Zhu W. Structural Deep Network Embedding. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2016. p. 1225–34.
25. Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. 2016;
26. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural Message Passing for Quantum Chemistry. 2017;
27. Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, et al. Graph neural networks: A review of methods and applications. AI Open. 2020;1:57–81.
28. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph Attention Networks. 2017;
29. Battaglia PW, Hamrick JB, Bapst V, Sanchez-Gonzalez A, Zambaldi V, Malinowski M, et al. Relational inductive biases, deep learning, and graph networks. 2018;
30. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. Proceedings of the National Academy of Sciences. 2007;104:8685–90.
31. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. 2011;21:1109–21.
32. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc. 2019;14:482–517.
33. Wang Z, Clark NR, Ma'ayan A. Drug-induced adverse events prediction with the LINCS L1000 data. Bioinformatics. 2016;32:2338–45.
34. Köhler S, Bauer S, Horn D, Robinson PN. Walking the Interactome for Prioritization of Candidate Disease Genes. The American Journal of Human Genetics. 2008;82:949–58.

35. Chiang AP, Butte AJ. Systematic Evaluation of Drug–Disease Relationships to Identify Leads for Novel Drug Uses. Clin Pharmacol Ther. 2009;86:507–10.
36. Schadt EE. Molecular networks as sensors and drivers of common human diseases. Nature. 2009;461:218–23.
37. Zhou X, Menche J, Barabási A-L, Sharma A. Human symptoms–disease network. Nat Commun. 2014;5:4212.
38. Ideker T, Sharan R. Protein networks in disease. Genome Res. 2008;18:644–52.
39. Butina D, Segall MD, Frankcombe K. Predicting ADME properties in silico: methods and models. Drug Discov Today. 2002;7:S83–8.
40. Cheng F, Desai RJ, Handy DE, Wang R, Schneeweiss S, Barabási A-L, et al. Network-based approach to prediction and population-based validation of in silico drug repurposing. Nat Commun. 2018;9:2691.
41. Weile J, Sun S, Cote AG, Knapp J, Verby M, Mellor JC, et al. A framework for exhaustively mapping functional missense variants. Mol Syst Biol. 2017;13.
42. Veselkov K, Gonzalez G, Aljifri S, Galea D, Mirnezami R, Youssef J, et al. HyperFoods: Machine intelligent mapping of cancer-beating molecules in foods. Sci Rep. 2019;9:9237.
43. Oany AR, Mia M, Pervin T, Alyami SA, Moni MA. Integrative Systems Biology Approaches to Identify Potential Biomarkers and Pathways of Cervical Cancer. J Pers Med. 2021;11:363.
44. Cho H, Berger B, Peng J. Compact Integration of Multi-Network Topology for Functional Analysis of Genes. Cell Syst. 2016;3:540-548.e5.
45. Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A Comprehensive Survey on Graph Neural Networks. 2019;
46. Sleator DD, Endre Tarjan R. A data structure for dynamic trees. J Comput Syst Sci. 1983;26:362–91.
47. Würtz P, Havulinna AS, Soininen P, Tynkkynen T, Prieto-Merino D, Tillin T, et al. Metabolite Profiling and Cardiovascular Event Risk. Circulation. 2015;131:774–85.
48. Himmelstein DS, Baranzini SE. Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. PLoS Comput Biol. 2015;11:e1004259.
49. Demir Karaman E, Işık Z. Multi-Omics Data Analysis Identifies Prognostic Biomarkers across Cancers. Medical Sciences. 2023;11:44.
50. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nat Methods. 2013;10:1108–15.
51. Silverman EK, Loscalzo J. Network medicine approaches to the genetics of complex diseases. Discov Med. 2012;14:143–52.
52. Li N, Yang Z, Wang J, Lin H. Drug–target interaction prediction using knowledge graph embedding. iScience. 2024;27:109393.

53. Alaimo S, Pulvirenti A. Network-Based Drug Repositioning: Approaches, Resources, and Research Directions. 2019. p. 97–113.
54. Hamilton WL, Ying R, Leskovec J. Inductive Representation Learning on Large Graphs. 2017;
55. Gori M, Monfardini G, Scarselli F. A new model for learning in graph domains. Proceedings 2005 IEEE International Joint Conference on Neural Networks, 2005. IEEE; p. 729–34.
56. Scarselli F, Gori M, Ah Chung Tsoi, Hagenbuchner M, Monfardini G. The Graph Neural Network Model. IEEE Trans Neural Netw. 2009;20:61–80.
57. Bruna J, Zaremba W, Szlam A, LeCun Y. Spectral Networks and Locally Connected Networks on Graphs. 2013;
58. Scarselli F, Gori M, Ah Chung Tsoi, Hagenbuchner M, Monfardini G. The Graph Neural Network Model. IEEE Trans Neural Netw. 2009;20:61–80.
59. Ying R, You J, Morris C, Ren X, Hamilton WL, Leskovec J. Hierarchical Graph Representation Learning with Differentiable Pooling. 2018;
60. Xu K, Hu W, Leskovec J, Jegelka S. How Powerful are Graph Neural Networks? 2018;
61. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323:533–6.
62. Rossi E, Chamberlain B, Frasca F, Eynard D, Monti F, Bronstein M. Temporal Graph Networks for Deep Learning on Dynamic Graphs. 2020;
63. Zhang C, Song D, Huang C, Swami A, Chawla N V. Heterogeneous Graph Neural Network. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA: ACM; 2019. p. 793–803.
64. Yun S, Jeong M, Kim R, Kang J, Kim HJ. Graph Transformer Networks. 2019;
65. Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A Comprehensive Survey on Graph Neural Networks. IEEE Trans Neural Netw Learn Syst. 2021;32:4–24.
66. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics. 2018;34:i457–66.
67. Wang Y-B, You Z-H, Li X, Jiang T-H, Chen X, Zhou X, et al. Predicting protein–protein interactions from protein sequences by a stacked sparse autoencoder deep neural network. Mol Biosyst. 2017;13:1336–44.
68. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A Deep Learning Approach to Antibiotic Discovery. Cell. 2020;180:688-702.e13.
69. Gao H, Ji S. Graph U-Nets. 2019;
70. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: A literature review. J Biomed Inform. 2018;77:34–49.

71. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep learning in clinical natural language processing: a methodical review. Journal of the American Medical Informatics Association. 2020;27:457–70.
72. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19:1236–46.
73. Allot A, Chen Q, Kim S, Vera Alvarez R, Comeau DC, Wilbur WJ, et al. LitSense: making sense of biomedical literature at sentence level. Nucleic Acids Res. 2019;47:W594–9.
74. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. 2017;
75. Chen SF, Goodman J. An empirical study of smoothing techniques for language modeling. Proceedings of the 34th annual meeting on Association for Computational Linguistics -. Morristown, NJ, USA: Association for Computational Linguistics; 1996. p. 310–8.
76. Rosenfeld R. Two decades of statistical language modeling: where do we go from here? Proceedings of the IEEE. 2000;88:1270–8.
77. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. 1989;77:257–86.
78. Brants T. TnT. Proceedings of the sixth conference on Applied natural language processing -. Morristown, NJ, USA: Association for Computational Linguistics; 2000. p. 224–31.
79. Mikolov T, Karafiát M, Burget L, Černocký J, Khudanpur S. Recurrent neural network based language model. Interspeech 2010. ISCA: ISCA; 2010. p. 1045–8.
80. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. 2013;
81. Elman JL. Finding Structure in Time. Cogn Sci. 1990;14:179–211.
82. Hochreiter S. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 1998;06:107–16.
83. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997;9:1735–80.
84. Cho K, van Merrienboer B, Bahdanau D, Bengio Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. 2014;
85. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. 2014;
86. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018;
87. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019;

88. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners. 2020;
89. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. 2023;
90. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report. 2023;
91. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the Opportunities and Risks of Foundation Models. 2021;
92. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44–56.
93. Prabhakar SK, Won D-O. Medical Text Classification Using Hybrid Deep Learning Models with Multihead Attention. Comput Intell Neurosci. 2021;2021.
94. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25:24–9.
95. Hunter L, Cohen KB. Biomedical Language Processing: What's Beyond PubMed? Mol Cell. 2006;21:589–94.
96. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform. 2009;42:760–72.
97. Cohen KB. Biomedical Natural Language Processing and Text Mining. Methods in Biomedical Informatics. Elsevier; 2014. p. 141–77.
98. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36:1234–40.
99. Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. 2019;
100. Mullenbach J, Wiegreffe S, Duke J, Sun J, Eisenstein J. Explainable Prediction of Medical Codes from Clinical Text. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics; 2018. p. 1101–11.
101. Harpaz R, Callahan A, Tamang S, Low Y, Odgers D, Finlayson S, et al. Text Mining for Adverse Drug Events: the Promise, Challenges, and State of the Art. Drug Saf. 2014;37:777–90.
102. Yang CC, Yang H, Jiang L, Zhang M. Social media mining for drug safety signal detection. Proceedings of the 2012 international workshop on Smart health and wellbeing. New York, NY, USA: ACM; 2012. p. 33–40.
103. Ji Y, Zhou Z, Liu H, Davuluri R V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics. 2021;37:2112–20.

104. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18:1196–203.

105. Johnson M, Schuster M, Le Q V., Krikun M, Wu Y, Chen Z, et al. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. Trans Assoc Comput Linguist. 2017;5:339–51.

106. Dong C, Li Y, Gong H, Chen M, Li J, Shen Y, et al. A Survey of Natural Language Generation. 2021;

107. 2022 Alzheimer's disease facts and figures. Alzheimer's & Dementia. 2022;18:700–89.

108. World Health Organization. Cancer. https://www.who.int/news-room/fact-sheets/detail/cancer. 2020.

109. Johns Hopkins University. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE). https://coronavirus.jhu.edu/map.html. 2021.

110. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011;144:646–74.

111. DeTure MA, Dickson DW. The neuropathological diagnosis of Alzheimer's disease. Mol Neurodegener. 2019;14:32.

112. Alzheimer's disease facts and figures. Alzheimer's & Dementia. 2021;17:327–406.

113. Hardy J, Selkoe DJ. The Amyloid Hypothesis of Alzheimer's Disease: Progress and Problems on the Road to Therapeutics. Science (1979). 2002;297:353–6.

114. Selkoe DJ. Soluble oligomers of the amyloid β-protein impair synaptic plasticity and behavior. Behavioural Brain Research. 2008;192:106–13.

115. Wang Y, Mandelkow E. Tau in physiology and pathology. Nat Rev Neurosci. 2016;17:22–35.

116. Querfurth HW, LaFerla FM. Alzheimer's Disease. New England Journal of Medicine. 2010;362:329–44.

117. Heneka MT, Carson MJ, Khoury J El, Landreth GE, Brosseron F, Feinstein DL, et al. Neuroinflammation in Alzheimer's disease. Lancet Neurol. 2015;14:388–405.

118. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, et al. Gene Dose of Apolipoprotein E Type 4 Allele and the Risk of Alzheimer's Disease in Late Onset Families. Science (1979). 1993;261:921–3.

119. Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. The Lancet. 2020;396:413–46.

120. Long JM, Holtzman DM. Alzheimer Disease: An Update on Pathobiology and Treatment Strategies. Cell. 2019;179:312–39.

121. Yiannopoulou KG, Papageorgiou SG. Current and Future Treatments in Alzheimer Disease: An Update. J Cent Nerv Syst Dis. 2020;12:117957352090739.

122. Guney E, Oliva B. Exploiting Protein-Protein Interaction Networks for Genome-Wide Disease-Gene Prioritization. PLoS One. 2012;7:e43557.
123. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin. 2020;70:7–30.
124. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med. 2004;10:789–99.
125. Hanahan D, Weinberg RA. The Hallmarks of Cancer. Cell. 2000;100:57–70.
126. Croce CM. Oncogenes and Cancer. New England Journal of Medicine. 2008;358:502–11.
127. Levine AJ. p53, the Cellular Gatekeeper for Growth and Division. Cell. 1997;88:323–31.
128. Tannock IF, Hickman JA. Limits to Personalized Cancer Medicine. New England Journal of Medicine. 2016;375:1289–94.
129. Druker BJ, Sawyers CL, Kantarjian H, Resta DJ, Reese SF, Ford JM, et al. Activity of a Specific Inhibitor of the BCR-ABL Tyrosine Kinase in the Blast Crisis of Chronic Myeloid Leukemia and Acute Lymphoblastic Leukemia with the Philadelphia Chromosome. New England Journal of Medicine. 2001;344:1038–42.
130. Cohen P, Alessi DR. Kinase Drug Discovery – What's Next in the Field? ACS Chem Biol. 2013;8:96–104.
131. Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. Nat Rev Cancer. 2012;12:252–64.
132. Sharma P, Allison JP. The future of immune checkpoint therapy. Science (1979). 2015;348:56–61.
133. Holohan C, Van Schaeybroeck S, Longley DB, Johnston PG. Cancer drug resistance: an evolving paradigm. Nat Rev Cancer. 2013;13:714–26.
134. Henry NL, Hayes DF. Cancer biomarkers. Mol Oncol. 2012;6:140–6.
135. Jameson JL, Longo DL. Precision Medicine — Personalized, Problematic, and Promising. New England Journal of Medicine. 2015;372:2229–34.
136. World Health Organization. Coronavirus Disease (COVID-19) Pandemic. https://www.who.int/emergencies/diseases/novel-coronavirus-2019. 2020.
137. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet. 2020;395:497–506.
138. Gupta A, Madhavan M V., Sehgal K, Nair N, Mahajan S, Sehrawat TS, et al. Extrapulmonary manifestations of COVID-19. Nat Med. 2020;26:1017–32.
139. Mao L, Jin H, Wang M, Hu Y, Chen S, He Q, et al. Neurologic Manifestations of Hospitalized Patients With Coronavirus Disease 2019 in Wuhan, China. JAMA Neurol. 2020;77:683.
140. Callaway E. The coronavirus is mutating — does it matter? Nature. 2020;585:174–7.
141. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 2020;5:1403–7.

142. Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell. 2020;181:281-292.e6.
143. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell. 2020;181:271-280.e8.
144. V'kovski P, Kratzel A, Steiner S, Stalder H, Thiel V. Coronavirus biology and replication: implications for SARS-CoV-2. Nat Rev Microbiol. 2021;19:155–70.
145. Mehta P, McAuley DF, Brown M, Sanchez E, Tattersall RS, Manson JJ. COVID-19: consider cytokine storm syndromes and immunosuppression. The Lancet. 2020;395:1033–4.
146. Sanders JM, Monogue ML, Jodlowski TZ, Cutrell JB. Pharmacologic Treatments for Coronavirus Disease 2019 (COVID-19). JAMA. 2020;
147. Beigel JH, Tomashek KM, Dodd LE, Mehta AK, Zingman BS, Kalil AC, et al. Remdesivir for the Treatment of Covid-19 — Final Report. New England Journal of Medicine. 2020;383:1813–26.
148. Chen P, Nirula A, Heller B, Gottlieb RL, Boscia J, Morris J, et al. SARS-CoV-2 Neutralizing Antibody LY-CoV555 in Outpatients with Covid-19. New England Journal of Medicine. 2021;384:229–37.
149. Dexamethasone in Hospitalized Patients with Covid-19. New England Journal of Medicine. 2021;384:693–704.
150. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science (1979). 2020;367:1260–3.
151. Morselli Gysi D, do Valle Í, Zitnik M, Ameli A, Gan X, Varol O, et al. Network medicine framework for identifying drug-repurposing opportunities for COVID-19. Proceedings of the National Academy of Sciences. 2021;118.
152. Zhou Y, Hou Y, Shen J, Huang Y, Martin W, Cheng F. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discov. 2020;6:14.
153. de Lorgeril M, Salen P. The mediterranean diet: Rationale and evidence for its benefit. Curr Atheroscler Rep. 2008;10:518–22.
154. Butler MJ, Barrientos RM. The impact of nutrition on COVID-19 susceptibility and long-term consequences. Brain Behav Immun. 2020;87:53–4.
155. Calder PC. Omega-3 polyunsaturated fatty acids and inflammatory processes: nutrition or pharmacology? Br J Clin Pharmacol. 2013;75:645–62.
156. Serhan CN, Levy BD. Resolvins in inflammation: emergence of the pro-resolving superfamily of mediators. Journal of Clinical Investigation. 2018;128:2657–69.
157. Hu FB. Dietary pattern analysis: a new direction in nutritional epidemiology. Curr Opin Lipidol. 2002;13:3–9.

158. Schwingshackl L, Hoffmann G. Adherence to Mediterranean diet and risk of cancer: an updated systematic review and meta-analysis of observational studies. Cancer Med. 2015;4:1933–47.
159. Micha R, Wallace SK, Mozaffarian D. Red and Processed Meat Consumption and Risk of Incident Coronary Heart Disease, Stroke, and Diabetes Mellitus. Circulation. 2010;121:2271–83.
160. Chan DSM, Lau R, Aune D, Vieira R, Greenwood DC, Kampman E, et al. Red and Processed Meat and Colorectal Cancer Incidence: Meta-Analysis of Prospective Studies. PLoS One. 2011;6:e20456.
161. Johnson RJ, Segal MS, Sautin Y, Nakagawa T, Feig DI, Kang D-H, et al. Potential role of sugar (fructose) in the epidemic of hypertension, obesity and the metabolic syndrome, diabetes, kidney disease, and cardiovascular disease. Am J Clin Nutr. 2007;86:899–906.
162. Estruch R, Ros E, Salas-Salvadó J, Covas M-I, Corella D, Arós F, et al. Primary Prevention of Cardiovascular Disease with a Mediterranean Diet. New England Journal of Medicine. 2013;368:1279–90.
163. Satija A, Hu FB. Plant-based diets and cardiovascular health. Trends Cardiovasc Med. 2018;28:437–41.
164. Sofi F, Abbate R, Gensini GF, Casini A. Accruing evidence on benefits of adherence to the Mediterranean diet on health: an updated systematic review and meta-analysis. Am J Clin Nutr. 2010;92:1189–96.
165. Liu RH. Dietary Bioactive Compounds and Their Health Implications. J Food Sci. 2013;78.
166. Sies H. Oxidative stress: oxidants and antioxidants. Exp Physiol. 1997;82:291–5.
167. Frei B. Ascorbic acid protects lipids in human plasma and low-density lipoprotein against oxidative damage. Am J Clin Nutr. 1991;54:1113S-1118S.
168. Wang X, Ouyang Y, Liu J, Zhu M, Zhao G, Bao W, et al. Fruit and vegetable consumption and mortality from all causes, cardiovascular disease, and cancer: systematic review and dose-response meta-analysis of prospective cohort studies. BMJ. 2014;349:g4490–g4490.
169. Fahey JW, Zhang Y, Talalay P. Broccoli sprouts: An exceptionally rich source of inducers of enzymes that protect against chemical carcinogens. Proceedings of the National Academy of Sciences. 1997;94:10367–72.
170. Herr I, Büchler MW. Dietary constituents of broccoli and other cruciferous vegetables: Implications for prevention and therapy of cancer. Cancer Treat Rev. 2010;36:377–83.
171. ZHANG Y, TANG L. Discovery and development of sulforaphane as a cancer chemopreventive phytochemical. Acta Pharmacol Sin. 2007;28:1343–54.
172. Swanson D, Block R, Mousa SA. Omega-3 Fatty Acids EPA and DHA: Health Benefits Throughout Life. Advances in Nutrition. 2012;3:1–7.

173. Baur JA, Sinclair DA. Therapeutic potential of resveratrol: the in vivo evidence. Nat Rev Drug Discov. 2006;5:493–506.
174. Smoliga JM, Baur JA, Hausenblas HA. Resveratrol and health – A comprehensive review of human clinical trials. Mol Nutr Food Res. 2011;55:1129–41.
175. Mattson MP, Longo VD, Harvie M. Impact of intermittent fasting on health and disease processes. Ageing Res Rev. 2017;39:46–58.
176. Longo VD, Panda S. Fasting, Circadian Rhythms, and Time-Restricted Feeding in Healthy Lifespan. Cell Metab. 2016;23:1048–59.
177. Fontana L, Partridge L. Promoting Health and Longevity through Diet: From Model Organisms to Humans. Cell. 2015;161:106–18.
178. de Cabo R, Mattson MP. Effects of Intermittent Fasting on Health, Aging, and Disease. New England Journal of Medicine. 2019;381:2541–51.
179. Paoli A, Rubini A, Volek JS, Grimaldi KA. Beyond weight loss: a review of the therapeutic uses of very-low-carbohydrate (ketogenic) diets. Eur J Clin Nutr. 2013;67:789–96.
180. Neal EG, Chaffe H, Schwartz RH, Lawson MS, Edwards N, Fitzsimmons G, et al. The ketogenic diet for the treatment of childhood epilepsy: a randomised controlled trial. Lancet Neurol. 2008;7:500–6.
181. Arends J, Bachmann P, Baracos V, Barthelemy N, Bertz H, Bozzetti F, et al. ESPEN guidelines on nutrition in cancer patients. Clinical Nutrition. 2017;36:11–48.
182. Muscaritoli M, Lucia S, Farcomeni A, Lorusso V, Saracino V, Barone C, et al. Prevalence of malnutrition in patients at first medical oncology visit: the PreMiO study. Oncotarget. 2017;8:79884–96.
183. Arends J, Baracos V, Bertz H, Bozzetti F, Calder PC, Deutz NEP, et al. ESPEN expert group recommendations for action against cancer-related malnutrition. Clinical Nutrition. 2017;36:1187–96.
184. Baldwin C, Weekes CE. Dietary counselling with or without oral nutritional supplements in the management of malnourished patients: a systematic review and meta-analysis of randomised controlled trials. Journal of Human Nutrition and Dietetics. 2012;25:411–26.
185. Bouayed J, Bohn T. Exogenous Antioxidants—Double-Edged Swords in Cellular Redox State: Health Beneficial Effects at Physiologic Doses versus Deleterious Effects at High Doses. Oxid Med Cell Longev. 2010;3:228–37.
186. Manach C, Williamson G, Morand C, Scalbert A, Rémésy C. Bioavailability and bioefficacy of polyphenols in humans. I. Review of 97 bioavailability studies. Am J Clin Nutr. 2005;81:230S-242S.
187. Thilakarathna S, Rupasinghe H. Flavonoid Bioavailability and Attempts for Bioavailability Enhancement. Nutrients. 2013;5:3367–87.

188. Manach C, Scalbert A, Morand C, Rémésy C, Jiménez L. Polyphenols: food sources and bioavailability. Am J Clin Nutr. 2004;79:727–47.
189. Kim Y-I. Role of Folate in Colon Cancer Development and Progression. J Nutr. 2003;133:3731S-3739S.
190. Freedman ND, Park Y, Subar AF, Hollenbeck AR, Leitzmann MF, Schatzkin A, et al. Fruit and vegetable intake and head and neck cancer risk in a large United States prospective cohort study. Int J Cancer. 2008;122:2330–6.
191. Koh A, De Vadder F, Kovatcheva-Datchary P, Bäckhed F. From Dietary Fiber to Host Physiology: Short-Chain Fatty Acids as Key Bacterial Metabolites. Cell. 2016;165:1332–45.
192. Tomás-Barberán FA, Selma M V., Espín JC. Interactions of gut microbiota with dietary polyphenols and consequences to human health. Curr Opin Clin Nutr Metab Care. 2016;19:471–6.
193. Sheng L, Jena PK, Liu H-X, Hu Y, Nagar N, Bronner DN, et al. Obesity treatment by epigallocatechin-3-gallate-regulated bile acid signaling and its enriched Akkermansia muciniphila. FASEB J. 2018;32:fj201800370R.
194. Hamilton WL, Ying R, Leskovec J. Representation Learning on Graphs: Methods and Applications. 2017;
195. Defferrard M, Bresson X, Vandergheynst P. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. 2016;
196. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2023;51:D638–46.
197. Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P, Kuhn M. STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Res. 2016;44:D380-4.
198. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46:D1074–82.
199. Avram S, Wilson TB, Curpan R, Halip L, Borota A, Bora A, et al. DrugCentral 2023 extends human clinical data and integrates veterinary drugs. Nucleic Acids Res. 2023;51:D1276–87.
200. Harrington RA, Adhikari V, Rayner M, Scarborough P. Nutrient composition databases in the age of big data: foodDB, a comprehensive, real-time database infrastructure. BMJ Open. 2019;9:e026652.
201. Bateman A, Martin M-J, Orchard S, Magrane M, Agivetova R, Ahmad S, et al. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49:D480–9.
202. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, Wiegers J, Wiegers TC, et al. Comparative Toxicogenomics Database (CTD): update 2021. Nucleic Acids Res. 2021;49:D1138–43.

203. Pavan S, Rommel K, Mateo Marquina ME, Höhn S, Lanneau V, Rath A. Clinical Practice Guidelines for Rare Diseases: The Orphanet Database. PLoS One. 2017;12:e0170365.
204. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen — The Clinical Genome Resource. New England Journal of Medicine. 2015;372:2235–42.
205. Martin AR, Williams E, Foulger RE, Leigh S, Daugherty LC, Niblock O, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat Genet. 2019;51:1560–5.
206. Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med. 2018;10:25.
207. Gutiérrez-Sacristán A, Grosdidier S, Valverde O, Torrens M, Bravo À, Piñero J, et al. PsyGeNET: a knowledge platform on psychiatric disorders and their genes. Bioinformatics. 2015;31:3075–7.
208. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2023;51:D587–92.
209. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell. 2016;166:740–54.
210. Fatemi B, Duval Q, Girdhar R, Drozdzal M, Romero-Soriano A. Learning to Substitute Ingredients in Recipes. 2023;
211. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2020;48:D845–55.
212. Perneczky R. Dementia treatment versus prevention. Dialogues Clin Neurosci. 2019;21:43–51.
213. Shah H, Albanese E, Duggan C, Rudan I, Langa KM, Carrillo MC, et al. Research priorities to reduce the global burden of dementia by 2025. Lancet Neurol. 2016;15:1285–94.
214. Scarmeas N, Stern Y, Tang M-X, Mayeux R, Luchsinger JA. Mediterranean diet and risk for Alzheimer's disease. Ann Neurol. 2006;59:912–21.
215. Román GC, Jackson RE, Reis J, Román AN, Toledo JB, Toledo E. Extra-virgin olive oil for potential prevention of Alzheimer disease. Rev Neurol (Paris). 2019;175:705–23.
216. Vassallo N, editor. Natural Compounds as Therapeutic Agents for Amyloidogenic Diseases. Cham: Springer International Publishing; 2015.
217. Madeo F, Eisenberg T, Pietrocola F, Kroemer G. Spermidine in health and disease. Science (1979). 2018;359.
218. Martínez-Lapiscina EH, Clavero P, Toledo E, Estruch R, Salas-Salvadó J, San Julián B, et al. Mediterranean diet improves cognition: the PREDIMED-NAVARRA randomised trial. J Neurol Neurosurg Psychiatry. 2013;84:1318–25.

219. Kaddoumi A, Denney TS, Deshpande G, Robinson JL, Beyers RJ, Redden DT, et al. Extra-Virgin Olive Oil Enhances the Blood-Brain Barrier Function in Mild Cognitive Impairment: A Randomized Controlled Trial. Nutrients. 2022;14.
220. Valls-Pedret C, Sala-Vila A, Serra-Mir M, Corella D, de la Torre R, Martínez-González MÁ, et al. Mediterranean Diet and Age-Related Cognitive Decline: A Randomized Clinical Trial. JAMA Intern Med. 2015;175:1094–103.
221. Rigacci S. Olive Oil Phenols as Promising Multi-targeting Agents Against Alzheimer's Disease. Adv Exp Med Biol. 2015;863:1–20.
222. Bolognesi ML, Banzi R, Bartolini M, Cavalli A, Tarozzi A, Andrisano V, et al. Novel Class of Quinone-Bearing Polyamines as Multi-Target-Directed Ligands To Combat Alzheimer's Disease. J Med Chem. 2007;50:4882–97.
223. Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, et al. A comprehensive map of molecular drug targets. Nat Rev Drug Discov. 2017;16:19–34.
224. Gong J, Chehrazi-Raffle A, Reddi S, Salgia R. Development of PD-1 and PD-L1 inhibitors as a form of cancer immunotherapy: a comprehensive review of registration trials and future considerations. J Immunother Cancer. 2018;6:8.
225. David L, Thakkar A, Mercado R, Engkvist O. Molecular representations in AI-driven drug discovery: a review and practical guide. J Cheminform. 2020;12:56.
226. Ruiz C, Zitnik M, Leskovec J. Identification of disease treatment mechanisms through the multiscale interactome. Nat Commun. 2021;12:1796.
227. Laponogov I, Gonzalez G, Shepherd M, Qureshi A, Veselkov D, Charkoftaki G, et al. Network machine learning maps phytochemically rich "Hyperfoods" to fight COVID-19. Hum Genomics. 2021;15:1.
228. Vidal M, Cusick ME, Barabási A-L. Interactome Networks and Human Disease. Cell. 2011;144:986–98.
229. Valdeolivas A, Tichit L, Navarro C, Perrin S, Odelin G, Levy N, et al. Random walk with restart on multiplex and heterogeneous biological networks. Bioinformatics. 2019;35:497–505.
230. Rita L, Neumann NR, Laponogov I, Gonzalez G, Veselkov D, Pratico D, et al. Alzheimer's disease: using gene/protein network machine learning for molecule discovery in olive oil. Hum Genomics. 2023;17:57.
231. Piñero J, Saüch J, Sanz F, Furlong LI. The DisGeNET cytoscape app: Exploring and visualizing disease genomics data. Comput Struct Biotechnol J. 2021;19:2960–7.

232. Kalogiouri NP, Aalizadeh R, Thomaidis NS. Application of an advanced and wide scope non-target screening workflow with LC-ESI-QTOF-MS and chemometrics for the classification of the Greek olive oil varieties. Food Chem. 2018;256:53–61.
233. Kalogiouri NP, Aalizadeh R, Dasenaki ME, Thomaidis NS. Authentication of Greek PDO Kalamata Table Olives: A Novel Non-Target High Resolution Mass Spectrometric Approach. Molecules. 2020;25:2919.
234. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. 2012;
235. Laye MJ, Tran V, Jones DP, Kapahi P, Promislow DEL. The effects of age and dietary restriction on the tissue-specific metabolome of *<scp>D</scp> rosophila* . Aging Cell. 2015;14:797–808.
236. Flaten TP. Aluminium as a risk factor in Alzheimer's disease, with emphasis on drinking water. Brain Res Bull. 2001;55:187–96.
237. Tounekti O, Zhang Y, Klaiman G, Goodyer CG, LeBlanc A. Proteasomal degradation of caspase-6 in 17beta-estradiol-treated neurons. J Neurochem. 2004;89:561–8.
238. Chesik D, Glazenburg L, De Keyser J, Wilczak N. Enhanced proliferation of astrocytes from ? $_2$ -adrenergic receptor knockout mice is influenced by the IGF system. J Neurochem. 2006;0:070209222715079-???
239. Ifteni P, Grudnikoff E, Koppel J, Kremen N, Correll CU, Kane JM, et al. Haloperidol and sudden cardiac death in dementia: autopsy findings in psychiatric inpatients. Int J Geriatr Psychiatry. 2015;30:1224–9.
240. Ansari MA, Abdul HM, Joshi G, Opii WO, Butterfield DA. Protective effect of quercetin in primary neurons against Abeta(1-42): relevance to Alzheimer's disease. J Nutr Biochem. 2009;20:269–75.
241. Khan H, Ullah H, Aschner M, Cheang WS, Akkol EK. Neuroprotective Effects of Quercetin in Alzheimer's Disease. Biomolecules. 2019;10.
242. Zaplatic E, Bule M, Shah SZA, Uddin MS, Niaz K. Molecular mechanisms underlying protective role of quercetin in attenuating Alzheimer's disease. Life Sci. 2019;224:109–19.
243. Uddin MS, Kabir MT. Emerging Signal Regulating Potential of Genistein Against Alzheimer's Disease: A Promising Molecule of Interest. Front Cell Dev Biol. 2019;7:197.
244. Devi KP, Shanmuganathan B, Manayi A, Nabavi SF, Nabavi SM. Molecular and Therapeutic Targets of Genistein in Alzheimer's Disease. Mol Neurobiol. 2017;54:7028–41.
245. Bagheri M, Joghataei M-T, Mohseni S, Roghani M. Genistein ameliorates learning and memory deficits in amyloid β(1-40) rat model of Alzheimer's disease. Neurobiol Learn Mem. 2011;95:270–6.
246. Frigolet ME, Gutiérrez-Aguilar R. The Role of the Novel Lipokine Palmitoleic Acid in Health and Disease. Advances in Nutrition. 2017;8:173S-181S.

247. Zhao L, Wang J-L, Liu R, Li X-X, Li J-F, Zhang L. Neuroprotective, Anti-Amyloidogenic and Neurotrophic Effects of Apigenin in an Alzheimer's Disease Mouse Model. Molecules. 2013;18:9949–65.
248. Venigalla M, Sonego S, Gyengesi E, Münch G. Curcumin and Apigenin - novel and promising therapeutics against chronic neuroinflammation in Alzheimer's disease. Neural Regen Res. 2015;10:1181.
249. Balez R, Steiner N, Engel M, Muñoz SS, Lum JS, Wu Y, et al. Neuroprotective effects of apigenin against inflammation, neuronal excitability and apoptosis in an induced pluripotent stem cell model of Alzheimer's disease. Sci Rep. 2016;6:31450.
250. George RC, Lew J, Graves DJ. Interaction of cinnamaldehyde and epicatechin with tau: implications of beneficial effects in modulating Alzheimer's disease pathogenesis. J Alzheimers Dis. 2013;36:21–40.
251. Zeng Y-Q, Wang Y-J, Zhou X-F. Effects of (âˆ’)Epicatechin on the Pathology of APP/PS1 Transgenic Mice. Front Neurol. 2014;5.
252. Beg T, Jyoti S, Naz F, Rahul, Ali F, Ali SK, et al. Protective Effect of Kaempferol on the Transgenic Drosophila Model of Alzheimer's Disease. CNS Neurol Disord Drug Targets. 2018;17:421–9.
253. Silva dos Santos J, Gonçalves Cirino JP, de Oliveira Carvalho P, Ortega MM. The Pharmacological Action of Kaempferol in Central Nervous System Diseases: A Review. Front Pharmacol. 2021;11.
254. Pan M, Han H, Zhong C, Geng Q. Effects of genistein and daidzein on hippocampus neuronal cell proliferation and BDNF expression in H19-7 neural cell line. J Nutr Health Aging. 2012;16:389–94.
255. Choi RCY, Zhu JTT, Yung AWY, Lee PSC, Xu SL, Guo AJY, et al. Synergistic Action of Flavonoids, Baicalein, and Daidzein in Estrogenic and Neuroprotective Effects: A Development of Potential Health Products and Therapeutic Drugs against Alzheimer's Disease. Evidence-Based Complementary and Alternative Medicine. 2013;2013:1–10.
256. Reese LC, Taglialatela G. Neuroimmunomodulation by calcineurin in aging and Alzheimer's disease. Aging Dis. 2010;1:245–53.
257. Flora SJS, Pachauri V. Chelation in Metal Intoxication. Int J Environ Res Public Health. 2010;7:2745–88.
258. Kuiper GG, Lemmen JG, Carlsson B, Corton JC, Safe SH, van der Saag PT, et al. Interaction of estrogenic chemicals and phytoestrogens with estrogen receptor beta. Endocrinology. 1998;139:4252–63.
259. Azargoonjahromi A. Dual role of nitric oxide in Alzheimer's disease. Nitric Oxide. 2023;134–135:23–37.

260. Minter MR, Zhang C, Leone V, Ringus DL, Zhang X, Oyler-Castrillo P, et al. Antibiotic-induced perturbations in gut microbial diversity influences neuro-inflammation and amyloidosis in a murine model of Alzheimer's disease. Sci Rep. 2016;6:30028.
261. Viña J, Escudero J, Baquero M, Cebrián M, Carbonell-Asíns JA, Muñoz JE, et al. Genistein effect on cognition in prodromal Alzheimer's disease patients. The GENIAL clinical trial. Alzheimers Res Ther. 2022;14:164.
262. Cox CJ, Choudhry F, Peacey E, Perkinton MS, Richardson JC, Howlett DR, et al. Dietary (−)-epicatechin as a potent inhibitor of βγ-secretase amyloid precursor protein processing. Neurobiol Aging. 2015;36:178–87.
263. Hole KL, Staniaszek LE, Menon Balan G, Mason JM, Brown JT, Williams RJ. Oral (−)-Epicatechin Inhibits Progressive Tau Pathology in rTg4510 Mice Independent of Direct Actions at GSK3β. Front Neurosci. 2021;15.
264. Ayuda-Durán B, Garzón-García L, González-Manzano S, Santos-Buelga C, González-Paramás AM. Insights into the Neuroprotective Potential of Epicatechin: Effects against Aβ-Induced Toxicity in Caenorhabditis elegans. Antioxidants. 2024;13:79.
265. Nejabati HR, Roshangar L. Kaempferol as a potential neuroprotector in Alzheimer's disease. J Food Biochem. 2022;46.
266. Mahwish, Imran M, Naeem H, Hussain M, Alsagaby SA, Al Abdulmonem W, et al. Antioxidative and Anticancer Potential of Luteolin: A Comprehensive Approach Against Wide Range of Human Malignancies. Food Sci Nutr. 2025;13.
267. Snigdha S, Astarita G, Piomelli D, Cotman CW. Effects of diet and behavioral enrichment on free fatty acids in the aged canine brain. Neuroscience. 2012;202:326–33.
268. Hosseini M, Poljak A, Braidy N, Crawford J, Sachdev P. Blood fatty acids in Alzheimer's disease and mild cognitive impairment: A meta-analysis and systematic review. Ageing Res Rev. 2020;60:101043.
269. Otoki Y, Kato S, Nakagawa K, Harvey DJ, Jin L-W, Dugger BN, et al. Lipidomic Analysis of Postmortem Prefrontal Cortex Phospholipids Reveals Changes in Choline Plasmalogen Containing Docosahexaenoic Acid and Stearic Acid Between Cases With and Without Alzheimer's Disease. Neuromolecular Med. 2021;23:161–75.
270. Fraser T, Tayler H, Love S. Fatty acid composition of frontal, temporal and parietal neocortex in the normal human brain and in Alzheimer's disease. Neurochem Res. 2010;35:503–13.
271. Wójtowicz S, Strosznajder AK, Jeżyna M, Strosznajder JB. The Novel Role of PPAR Alpha in the Brain: Promising Target in Therapy of Alzheimer's Disease and Other Neurodegenerative Disorders. Neurochem Res. 2020;45:972–88.

272. Souza CO, Teixeira AAS, Biondo LA, Silveira LS, de Souza Breda CN, Braga TT, et al. Palmitoleic acid reduces high fat diet-induced liver inflammation by promoting PPAR-γ-independent M2a polarization of myeloid cells. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids. 2020;1865:158776.
273. Yellamma K; PK. SQUALENE: A POTENTIAL ANTI-ALZHEIMER'S COMPOUND IN ALBINO RAT. Int J Pharm Sci Res. 2022;14.
274. Varma VR, Büşra Lüleci H, Oommen AM, Varma S, Blackshear CT, Griswold ME, et al. Abnormal brain cholesterol homeostasis in Alzheimer's disease—a targeted metabolomic and transcriptomic study. NPJ Aging Mech Dis. 2021;7:11.
275. Zarrouk A, Debbabi M, Bezine M, Karym EM, Badreddine A, Rouaud O, et al. Lipid Biomarkers in Alzheimer's Disease. Curr Alzheimer Res. 2018;15:303–12.
276. Hottman DA, Li L. Protein Prenylation and Synaptic Plasticity: Implications for Alzheimer's Disease. Mol Neurobiol. 2014;50:177–85.
277. Devanand DP, Michaels-Marston KS, Liu X, Pelton GH, Padilla M, Marder K, et al. Olfactory Deficits in Patients With Mild Cognitive Impairment Predict Alzheimer's Disease at Follow-Up. American Journal of Psychiatry. 2000;157:1399–405.
278. MURPHY C, BACON AW, BONDI MW, SALMON DP. Apolipoprotein E Status is Associated with Odor Identification Deficits in Nondemented Older Personsa. Ann N Y Acad Sci. 1998;855:744–50.
279. Devanand DP, Lee S, Manly J, Andrews H, Schupf N, Doty RL, et al. Olfactory deficits predict cognitive decline and Alzheimer dementia in an urban community. Neurology. 2015;84:182–9.
280. Wesson DW, Levy E, Nixon RA, Wilson DA. Olfactory Dysfunction Correlates with Amyloid-β Burden in an Alzheimer's Disease Mouse Model. The Journal of Neuroscience. 2010;30:505–14.
281. de la Monte SM, Wands JR. Alzheimer's Disease is Type 3 Diabetes—Evidence Reviewed. J Diabetes Sci Technol. 2008;2:1101–13.
282. Berman DE, Dall'Armi C, Voronov S V, McIntire LBJ, Zhang H, Moore AZ, et al. Oligomeric amyloid-β peptide disrupts phosphatidylinositol-4,5-bisphosphate metabolism. Nat Neurosci. 2008;11:547–54.
283. Iadecola C. Neurovascular regulation in the normal brain and in Alzheimer's disease. Nat Rev Neurosci. 2004;5:347–60.
284. Sweeney MD, Sagare AP, Zlokovic B V. Blood–brain barrier breakdown in Alzheimer disease and other neurodegenerative disorders. Nat Rev Neurol. 2018;14:133–50.
285. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2012;41:D808–15.

286. Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P. STITCH: interaction networks of chemicals and proteins. Nucleic Acids Res. 2007;36:D684–8.
287. Zambrowicz BP, Sands AT. Knockouts model the 100 best-selling drugs—will they model the next 100? Nat Rev Drug Discov. 2003;2:38–51.
288. Roberts AG, Gibbs ME. Mechanisms and the clinical relevance of complex drug–drug interactions. Clin Pharmacol. 2018;Volume 10:123–34.
289. Gibbs MA, Hosea NA. Factors Affecting the Clinical Development of Cytochrome P450 3A Substrates. Clin Pharmacokinet. 2003;42:969–84.
290. Bailey DG, Dresser G, Arnold JMO. Grapefruit–medication interactions: Forbidden fruit or avoidable consequences? Can Med Assoc J. 2013;185:309–16.
291. Singh BN. Effects of Food on Clinical Pharmacokinetics. Clin Pharmacokinet. 1999;37:213–55.
292. Hu Z, Yang X, Ho PCL, Chan SY, Heng PWS, Chan E, et al. Herb-Drug Interactions. Drugs. 2005;65:1239–82.
293. Budisan L, Gulei D, Zanoaga O, Irimie A, Chira S, Braicu C, et al. Dietary Intervention by Phytochemicals and Their Role in Modulating Coding and Non-Coding Genes in Cancer. Int J Mol Sci. 2017;18:1178.
294. Bailey DG, Arnold JMO, Spence JD. Grapefruit juice-drug interactions. Br J Clin Pharmacol. 2004;58.
295. Bailey DG, Dresser GK, Leake BF, Kim RB. Naringin is a Major and Selective Clinical Inhibitor of Organic Anion-Transporting Polypeptide 1A2 (OATP1A2) in Grapefruit Juice. Clin Pharmacol Ther. 2007;81:495–502.
296. Volak LP, Ghirmai S, Cashman JR, Court MH. Curcuminoids Inhibit Multiple Human Cytochromes P450, UDP-Glucuronosyltransferase, and Sulfotransferase Enzymes, whereas Piperine Is a Relatively Selective CYP3A4 Inhibitor. Drug Metabolism and Disposition. 2008;36:1594–605.
297. Suganuma M, Okabe S, Kai Y, Sueoka N, Sueoka E, Fujiki H. Synergistic effects of (--)-epigallocatechin gallate with (--)-epicatechin, sulindac, or tamoxifen on cancer-preventive activity in the human lung cancer cell line PC-9. Cancer Res. 1999;59:44–7.
298. Cao J, Han J, Xiao H, Qiao J, Han M. Effect of Tea Polyphenol Compounds on Anticancer Drugs in Terms of Anti-Tumor Activity, Toxicology, and Pharmacokinetics. Nutrients. 2016;8:762.
299. D'Andrea G. Quercetin: A flavonol with multifaceted therapeutic applications? Fitoterapia. 2015;106:256–71.
300. Maruszewska A, Tarasiuk J. Quercetin Triggers Induction of Apoptotic and Lysosomal Death of Sensitive and Multidrug Resistant Leukaemia HL60 Cells. Nutr Cancer. 2021;73:484–501.

301. Chen L, Sun X, Wang Z, Chen M, He Y, Zhang H, et al. Resveratrol protects against doxorubicin-induced cardiotoxicity by attenuating ferroptosis through modulating the MAPK signaling pathway. Toxicol Appl Pharmacol. 2024;482:116794.

302. Fukui M, Yamabe N, Zhu BT. Resveratrol attenuates the anticancer efficacy of paclitaxel in human breast cancer cells in vitro and in vivo. Eur J Cancer. 2010;46:1882–91.

303. Cleland L, James M, Proudman S. Fish oil: what the prescriber needs to know. Arthritis Res Ther. 2006;8:202.

304. Ankri S, Mirelman D. Antimicrobial properties of allicin from garlic. Microbes Infect. 1999;1:125–9.

305. Cutler RR, Wilson P. Antibacterial activity of a new, stable, aqueous extract of allicin against methicillin-resistant *Staphylococcus aureus*. Br J Biomed Sci. 2004;61:71–4.

306. Izzo AA, Ernst E. Interactions Between Herbal Medicines and Prescribed Drugs. Drugs. 2009;69:1777–98.

307. Rombolà L, Scuteri D, Marilisa S, Watanabe C, Morrone LA, Bagetta G, et al. Pharmacokinetic Interactions between Herbal Medicines and Drugs: Their Mechanisms and Clinical Relevance. Life. 2020;10:106.

308. Atreya R V, Sun J, Zhao Z. Exploring drug-target interaction networks of illicit drugs. BMC Genomics. 2013;14:S1.

309. Mierziak J, Kostyn K, Boba A, Czemplik M, Kulma A, Wojtasik W. Influence of the Bioactive Diet Components on the Gene Expression Regulation. Nutrients. 2021;13:3673.

310. Yıldırım MA, Goh K-I, Cusick ME, Barabási A-L, Vidal M. Drug—target network. Nat Biotechnol. 2007;25:1119–26.

311. Komatsu W, Kishi H, Yagasaki K, Ohhira S. Urolithin A attenuates pro-inflammatory mediator production by suppressing PI3-K/Akt/NF-κB and JNK/AP-1 signaling pathways in lipopolysaccharide-stimulated RAW264 macrophages: Possible involvement of NADPH oxidase-derived reactive oxygen species. Eur J Pharmacol. 2018;833:411–24.

312. Haiser HJ, Gootenberg DB, Chatman K, Sirasani G, Balskus EP, Turnbaugh PJ. Predicting and Manipulating Cardiac Drug Inactivation by the Human Gut Bacterium *Eggerthella lenta*. Science (1979). 2013;341:295–8.

313. DreamLab App. The DRUGS Project. https://www.vodafone.co.uk/mobile/dreamlab. 2018.

314. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics. 2018;34:i457–66.

315. Gaudelet T, Day B, Jamasb AR, Soman J, Regep C, Liu G, et al. Utilizing graph machine learning within drug discovery and development. Brief Bioinform. 2021;22.

316. Gonzalez G, LI, BJ, CD, R-JM, RL, QJ, GM, WY, ZE, ZY, WX, BM, VK. Leveraging food-drug pharmacodynamic interactions to enhance drug activity. In Preparation. 2024;
317. Motulsky H, Christopoulos A. Fitting Models to Biological Data Using Linear and Nonlinear Regression. Oxford University PressNew York, NY; 2004.
318. Sebaugh JL. Guidelines for accurate EC50/IC50 estimation. Pharm Stat. 2011;10:128–34.
319. Yuan JS, Reed A, Chen F, Stewart CN. Statistical analysis of real-time PCR data. BMC Bioinformatics. 2006;7:85.
320. Weiss JN. The Hill equation revisited: uses and misuses. The FASEB Journal. 1997;11:835–41.
321. Atkins P, & de PJ. Physical Chemistry. 10th ed. Oxford University Press; 2014.
322. Nelson DL, Cox MM. Lehninger Biochemie. Berlin, Heidelberg: Springer Berlin Heidelberg; 2001.
323. A Pharmacology Primer. Elsevier; 2014.
324. Yung-Chi C, Prusoff WH. Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. Biochem Pharmacol. 1973;22:3099–108.
325. Tallarida RJ. Drug synergism: its detection and applications. J Pharmacol Exp Ther. 2001;298:865–72.
326. Greco WR, Bravo G, Parsons JC. The search for synergy: a critical review from a response surface perspective. Pharmacol Rev. 1995;47:331–85.
327. Liu Q, Yin X, Languino LR, Altieri DC. Evaluation of Drug Combination Effect Using a Bliss Independence Dose–Response Surface Model. Stat Biopharm Res. 2018;10:112–22.
328. Lederer S, Dijkstra TMH, Heskes T. Additive Dose Response Models: Explicit Formulation and the Loewe Additivity Consistency Condition. Front Pharmacol. 2018;9.
329. Yadav B, Wennerberg K, Aittokallio T, Tang J. Searching for Drug Synergy in Complex Dose–Response Landscapes Using an Interaction Potency Model. Comput Struct Biotechnol J. 2015;13:504–13.
330. Chou T-C. The combination index (CI < 1) as the definition of synergism and of synergy claims. Synergy. 2018;7:49–50.
331. Leal-Esteban LC, Fajas L. Cell cycle regulators in cancer cell metabolism. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease. 2020;1866:165715.
332. Phelan MC. Basic Techniques in Mammalian Cell Tissue Culture. Curr Protoc Cell Biol. 2007;36.
333. Ghasemi M, Turnbull T, Sebastian S, Kempson I. The MTT Assay: Utility, Limitations, Pitfalls, and Interpretation in Bulk and Single-Cell Analysis. Int J Mol Sci. 2021;22:12827.
334. Mosmann T. Rapid colorimetric assay for cellular growth and survival: Application to proliferation and cytotoxicity assays. J Immunol Methods. 1983;65:55–63.

335. Kamiloglu S, Sari G, Ozdal T, Capanoglu E. Guidelines for cell viability assays. Food Front. 2020;1:332–49.
336. Motulsky HJ, Brown RE. Detecting outliers when fitting data with nonlinear regression – a new method based on robust nonlinear regression and the false discovery rate. BMC Bioinformatics. 2006;7:123.
337. Hughes P, Marshall D, Reid Y, Parkes H, Gelber C. The Costs of using Unauthenticated, Over-Passaged Cell Lines: How Much more Data do we Need? Biotechniques. 2007;43:575–86.
338. Zhao X, Zhang J. Inhibitory effect of galangin on DNA topoisomerases in lung cancer cells]. Zhong Nan Da Xue Xue Bao Yi Xue Ban. 2015;40:479–85.
339. Acharya BR, Chatterjee A, Ganguli A, Bhattacharya S, Chakrabarti G. Thymoquinone inhibits microtubule polymerization by tubulin binding and causes mitotic arrest following apoptosis in A549 cells. Biochimie. 2014;97:78–91.
340. Redman C. Involvement of polyamines in selenomethionine induced apoptosis and mitotic alterations in human tumor cells. Carcinogenesis. 1997;18:1195–202.
341. Selleckchem. Sunitinib. https://www.selleckchem.com/products/sunitinib.html. 2024.
342. Effendi WI, Nagano T, Tachihara M, Umezawa K, Kiriu T, Dokuni R, et al. Synergistic interaction of gemcitabine and paclitaxel by modulating acetylation and polymerization of tubulin in non-small cell lung cancer cell lines . Cancer Manag Res. 2019;Volume 11:3669–79.
343. Abdelwahab M, Saeed H, Elnikhely N, Nematalla H. Synergistic effect of Dactolisib/Lys05 combination on autophagy in A549 cells. Acta Biochim Pol. 2023;
344. Xin Y, Yin F, Qi S, Shen L, Xu Y, Luo L, et al. Parthenolide reverses doxorubicin resistance in human lung carcinoma A549 cells by attenuating NF-κB activation and HSP70 up-regulation. Toxicol Lett. 2013;221:73–82.
345. Davou G, Chuwang N, Essien U, Choji T, Echeonwu B, Lugos M. Cytotoxicity analysis of etoposide and cisplatin on cell lines from human lung cancer and normal human lung. International Research Journal of Medicine and Medical Sciences. 2019;7:40–7.
346. Liu Y, He C, Huang X. Metformin partially reverses the carboplatin-resistance in NSCLC by inhibiting glucose metabolism. Oncotarget. 2017;8:75206–16.
347. Kashkin KN, Musatkina EA, Komelkov A V, Favorskaya IA, Trushkin E V, Shleptsova VA, et al. Expression profiling and putative mechanisms of resistance to doxorubicin of human lung cancer cells. Dokl Biochem Biophys. 2010;430:20–3.
348. An Q, Han C, Zhou Y, Li F, Li D, Zhang X, et al. In vitro effects of mitomycin C on the proliferation of the non-small-cell lung cancer line A549. Int J Clin Exp Med. 2015;8:20516–23.
349. Selleckchem. Nintedanib. https://www.selleckchem.com/products/BIBF1120.html. 2024.

350. Fotakis G, Timbrell JA. In vitro cytotoxicity assays: Comparison of LDH, neutral red, MTT and protein assay in hepatoma cell lines following exposure to cadmium chloride. Toxicol Lett. 2006;160:171–7.
351. Duldulao MP, Lee W, Le M, Chen Z, Li W, Wang J, et al. Gene expression variations in microsatellite stable and unstable colon cancer cells. J Surg Res. 2012;174:1–6.
352. Ma F, An Z, Yue Q, Zhao C, Zhang S, Sun X, et al. Effects of brassinosteroids on cancer cells: A review. J Biochem Mol Toxicol. 2022;36.
353. Elreedy HA, Elfiky AM, Mahmoud AA, Ibrahim KS, Ghazy MA. Neuroprotective effect of quercetin through targeting key genes involved in aluminum chloride induced Alzheimer's disease in rats. Egyptian Journal of Basic and Applied Sciences. 2023;10:174–84.
354. Liu J, Zhang L, Gao J, Zhang B, Liu X, Yang N, et al. Discovery of genistein derivatives as potential SARS-CoV-2 main protease inhibitors by virtual screening, molecular dynamics simulations and ADMET analysis. Front Pharmacol. 2022;13.
355. Shirai SS, Seneviratne O, Gordon ME, Chen C-H, McGuinness DL. Identifying Ingredient Substitutions Using a Knowledge Graph of Food. Front Artif Intell. 2021;3.
356. Yamanishi R, Shino N, Nishihara Y, Fukumoto J, Kaizaki A. Alternative-ingredient Recommendation Based on Co-occurrence Relation on Recipe Database. Procedia Comput Sci. 2015;60:986–93.
357. Boscarino C, Nedović V, Koenderink NJJP, Top JL. Automatic extraction of ingredient's substitutes. Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication. New York, NY, USA: ACM; 2014. p. 559–64.
358. Ahn Y-Y, Ahnert SE, Bagrow JP, Barabási A-L. Flavor network and the principles of food pairing. Sci Rep. 2011;1:196.
359. Achananuparp P, Weber I. Extracting Food Substitutes From Food Diary via Distributional Similarity. 2016;
360. Kazama M, Sugimoto M, Hosokawa C, Matsushima K, Varshney LR, Ishikawa Y. A neural network system for transformation of regional cuisine style. 2017;
361. Lawo D, Böhm L, Stevens G. Veganaizer: AI-assisted Ingredient Substitution. 2020;
362. Morales-Garzon A, Gomez-Romero J, Martin-Bautista MJ. A Word Embedding-Based Method for Unsupervised Adaptation of Cooking Recipes. IEEE Access. 2021;9:27389–404.
363. Wu S, He Y. Enriching Pre-trained Language Model with Entity Information for Relation Classification. 2019;
364. Park D, Kim K, Kim S, Spranger M, Kang J. FlavorGraph: a large-scale food-chemical graph for generating food representations and recommending food pairings. Sci Rep. 2021;11:931.
365. Zhang P, Zeng G, Wang T, Lu W. TinyLlama: An Open-Source Small Language Model. 2024;

366. Rita L, Southern J, Laponogov I, Higgins K, Veselkov K. Optimizing Ingredient Substitution Using Large Language Models to Enhance Phytochemical Content in Recipes. 2024;
367. Salvador A, Hynes N, Aytar Y, Marin J, Ofli F, Weber I, et al. Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2017. p. 3068–76.
368. Nicoli MC, Anese M, Parpinel M. Influence of processing on the antioxidant properties of fruit and vegetables. Trends Food Sci Technol. 1999;10:94–100.
369. Hasan MdI, Hossain MdA, Rahman MH, Sohel M, Ahsan A, Soikot MdSH, et al. Galangin for COVID-19 and Mucormycosis co-infection: a potential therapeutic strategy of targeting critical host signal pathways triggered by SARS-CoV-2 and Mucormycosis. Network Modeling Analysis in Health Informatics and Bioinformatics. 2023;12:26.
370. Hollman PCH, Bijsman MNCP, van Gameren Y, Cnossen EPJ, de Vries JHM, Katan MB. The sugar moiety is a major determinant of the absorption of dietary flavonoid glycosides in man. Free Radic Res. 1999;31:569–73.
371. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res. 2018;46:D608–17.
372. Selkoe DJ. Alzheimer's Disease: Genes, Proteins, and Therapy. Physiol Rev. 2001;81:741–66.
373. Lee VM-Y, Goedert M, Trojanowski JQ. Neurodegenerative Tauopathies. Annu Rev Neurosci. 2001;24:1121–59.
374. Iqbal K, Liu F, Gong C-X, Grundke-Iqbal I. Tau in Alzheimer Disease and Related Tauopathies. Curr Alzheimer Res. 2010;7:656–64.
375. Heneka MT, Carson MJ, Khoury J El, Landreth GE, Brosseron F, Feinstein DL, et al. Neuroinflammation in Alzheimer's disease. Lancet Neurol. 2015;14:388–405.
376. Wyss-Coray T, Rogers J. Inflammation in Alzheimer Disease--A Brief Review of the Basic Science and Clinical Literature. Cold Spring Harb Perspect Med. 2012;2:a006346–a006346.
377. Murphy C. Olfactory and other sensory impairments in Alzheimer disease. Nat Rev Neurol. 2019;15:11–24.
378. Wilson RS, Arnold SE, Schneider JA, Tang Y, Bennett DA. The relationship between cerebral Alzheimer's disease pathology and odour identification in old age. J Neurol Neurosurg Psychiatry. 2007;78:30–5.
379. Attems J, Walker L, Jellinger KA. Olfactory bulb involvement in neurodegenerative diseases. Acta Neuropathol. 2014;127:459–75.
380. Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol. 1991;82:239–59.

381. Zhao W-Q, Alkon DL. Role of insulin and insulin receptor in learning and memory. Mol Cell Endocrinol. 2001;177:125–34.
382. Lesort M, Jope RS, Johnson GVW. Insulin Transiently Increases Tau Phosphorylation. J Neurochem. 1999;72:576–84.
383. Craft S. Insulin Resistance and Alzheimers Disease Pathogenesis: Potential Mechanisms and Implications for Treatment. Curr Alzheimer Res. 2007;4:147–52.
384. Zhao W, Chen H, Xu H, Moore E, Meiri N, Quon MJ, et al. Brain Insulin Receptors and Spatial Memory. Journal of Biological Chemistry. 1999;274:34893–902.
385. Ando K, Erneux C, Homa M, Houben S, de Fisenne M-A, Brion J-P, et al. Dysregulation of Phosphoinositide 5-Phosphatases and Phosphoinositides in Alzheimer's Disease. Front Neurosci. 2021;15.
386. Razani E, Pourbagheri-Sigaroodi A, Safaroghli-Azar A, Zoghi A, Shanaki-Bavarsad M, Bashash D. The PI3K/Akt signaling axis in Alzheimer's disease: a valuable target to stimulate or suppress? Cell Stress Chaperones. 2021;26:871–87.
387. Kitagishi Y, Nakanishi A, Ogura Y, Matsuda S. Dietary regulation of PI3K/AKT/GSK-3β pathway in Alzheimer's disease. Alzheimers Res Ther. 2014;6:35.
388. Frere SG, Chang-Ileto B, Di Paolo G. Role of Phosphoinositides at the Neuronal Synapse. 2012. p. 131–75.
389. Kinney JW, Bemiller SM, Murtishaw AS, Leisgang AM, Salazar AM, Lamb BT. Inflammation as a central mechanism in Alzheimer's disease. Alzheimer's & Dementia: Translational Research & Clinical Interventions. 2018;4:575–90.
390. Long H-Z, Cheng Y, Zhou Z-W, Luo H-Y, Wen D-D, Gao L-C. PI3K/AKT Signal Pathway: A Target of Natural Products in the Prevention and Treatment of Alzheimer's Disease and Parkinson's Disease. Front Pharmacol. 2021;12.
391. Morrison RS, Kinoshita Y, Johnson MD, Ghatan S, Ho JT, Garden G. Neuronal Survival and Cell Death Signaling Pathways. 2003. p. 41–86.
392. Aguilar-Pineda JA, Vera-Lopez KJ, Shrivastava P, Chávez-Fumagalli MA, Nieto-Montesinos R, Alvarez-Fernandez KL, et al. Vascular smooth muscle cell dysfunction contribute to neuroinflammation and Tau hyperphosphorylation in Alzheimer disease. iScience. 2021;24:102993.
393. Kisler K, Nelson AR, Montagne A, Zlokovic B V. Cerebral blood flow regulation and neurovascular dysfunction in Alzheimer disease. Nat Rev Neurosci. 2017;18:419–34.
394. Sharma C, Woo H, Kim SR. Addressing Blood–Brain Barrier Impairment in Alzheimer's Disease. Biomedicines. 2022;10:742.

395. Weller RO, Massey A, Newman TA, Hutchings M, Kuo Y-M, Roher AE. Cerebral Amyloid Angiopathy. Am J Pathol. 1998;153:725–33.
396. Iadecola C, Davisson RL. Hypertension and Cerebrovascular Dysfunction. Cell Metab. 2008;7:476–84.

Funding

Funding was provided by Fundação para a Ciência e a Tecnologia (2021.05460.BD), Sam Goodenough (Goodenough College) and Mecklenburgh Square (Goodenough College).

DOI: https://doi.org/10.54499/2021.05460.BD

Acknowledgements

I am forever grateful to my supervisor, Kirill V., for making this four-year journey one of the most rewarding experiences of my life. ☺ Hopefully, the beginning of even greater ones with him and all those who made this adventure so special: David T., Dennis V., Domenico P., Donald C., Guada G., Hamideh K., Ivan L, Jerusa B., Josh S., Kyle H., Natalie N., Nikolaos T., Reza A., Vasilis V. and Wei W.

★Codrin ★Gatis

Mum, Dad, Sister, Grandma and Grandpa, Thank you!

Mãe, Pai, Irmã, Avó e Avô, Obrigado!

* GenAI was used for grammar and spell checking of the dissertation.