Imperial College
London

Imperial College of Science, Technology and Medicine

Department of Computing

Therapeutic lead discovery using graph deep learning

Guadalupe Gabriela González Pigorini

Supervisor: Michael Bronstein

Co-supervisor: Kirill Veselkov

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computing, 2023

Statement of originality

I hereby certify that the contents of this dissertation are entirely of my own aside from where appropriate references are stated.

Copyright statement

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 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.

Abstract

Novel drug lead identification is a key challenge in the development of new therapies against cancer. While natural compounds were initially preferred for their evolutionary honed properties, high-throughput screening shifted focus to synthetic compounds. This approach, however, has resulted in limited success so far. As a result, recent years have seen a revival of natural products as promising sources for lead discovery. We posit that the recent developments in computational network-based methods for drug development can further push the revival of natural product-based drug discovery. In this dissertation, we introduce novel approaches for therapeutic lead discovery using graph-based deep learning methods. The first part of this dissertation explores the use of graph-based deep learning methods to predict therapeutic leads from molecules found in foods, either as single agents or in combination with existing anticancer drugs. Our results show promise for the exploration of natural compounds and their relationships with drugs across other diseases. In the second part of this dissertation, we outline the limitations of the current lead discovery problem formulation and propose a novel problem formulation that overlaps with lead design, therefore expanding the search space. To solve this problem formulation, we introduce an algorithm that combines the rigor of causal inference with the representation power of graph neural networks. Our algorithm predicts perturbagens that shift cell states from an initial to a desired target state across various genetic and chemical perturbation datasets and opens the door for the development of new, more flexible methods, for lead discovery. By introducing a fresh problem formulation for lead discovery, this dissertation sets the stage for more nuanced and effective strategies to lead discovery in cancer research.

Acknowledgments

I would like to express my heartfelt gratitude to my advisors, Kirill Veselkov and Michael Bronstein, for their unwavering support. This dissertation truly would not have come to fruition without their guidance. I am especially grateful to Kirill, who has been an exceptional mentor from the beginning of our collaboration during my master's degree. I extend my sincere thanks to Ivan Laponogov for his support and constant availability in reviewing manuscripts and explaining complex concepts. Your assistance has been invaluable to me.

I also want to express my appreciation to all the members of Bronstein's lab who embarked on this journey alongside me: Fabrizio Frasca, Giorgos Bouritsas, Shunwang Gong, and Mehdi Bahri. Despite the challenges we faced, we have grown together, and I am grateful for the camaraderie and shared experiences. To Joshua Southern, thank you for being a great colleague and friend since joining our lab. It has been a pleasure collaborating with you. Jerusa Brignardello, I want to express my gratitude for sharing this journey with me and for your unwavering support. Your presence has brought joy to this experience. I am fortunate to have had the opportunity of working with an exceptional group of collaborators and coauthors. I am sincerely grateful to each of them: Isuru Herath, Valentina Giunchiglia, Chirag Shukla, Chirag Agarwal, Zhiqiang Zhong, Daniele Grattarola, Pia Borgas, Diego Galeano, and Dieter Galea.

Working with Marinka Zitnik has been an absolute delight. I want to thank her for being an outstanding researcher and mentor, and for her continued guidance and support. Collaborating with her has been one of the greatest joys of my journey. I am grateful to all the members of the Zitnik Lab, especially those with whom I had the opportunity to spend more time during my stay in Boston, including ManQing Liang and Yepeng Huang.

A special thank you goes to my husband, Matias Mikkola, for being a constant source of strength and support throughout my journey. Your unwavering belief in me has allowed me to build my confidence and push myself outside my comfort zone time and time again.

Lastly, I want to acknowledge and thank my parents, Wilexis Gonzalez and Zaira Pigorini. Emigrating from Venezuela was a sacrifice for all of us but for you more than anyone else; and this is the very reason I have had the opportunities that have led me to where I am today. Thank you from the bottom of my heart.

Acronyms

- AI: Artificial Intelligence
- ATC: Anatomical Therapeutic Chemical
- B.C.: Before Christ
- BioGRID: Biological General Repository for Interaction Datasets
- BioPlex: Biological Phase-Change Exchange
- CMap: Connectivity Map
- COVID-19: Coronavirus Disease 2019
- DL: Deep Learning
- FDA: Food and Drug Administration
- GNN: Graph Neural Network
- GO: Gene Ontology
- HPRD: Human Protein Reference Database
- HTS: High-throughput screening
- KEGG: Kyoto Encyclopedia of Genes and Genomes
- LINCS: Library of Integrated Network-based Cellular Signatures
- MAOI: Monoamine Oxidase Inhibitors
- ML: Machine Learning
- MLP: Multi-Layer Perceptron
- NCI-PID: National Cancer Institute-Pathway Interaction Database
- NCM: Neural Causal Model
- PD: Pharmacodynamic
- Panther: Protein ANalysis THrough Evolutionary Relationships

- PPI: Protein-protein interaction
- PK: Pharmacokinetic
- RWR: Random Walk with Restarts
- Reactome: REACtion pathways online knowledgebase
- RW: Random Walk
- SCM: Structural Causal Model
- SMILES: Simplified Molecular Input Line Entry System
- STITCH: Search Tool for Interactions of Chemicals
- STRING: Search Tool for the Retrieval of Interacting Genes/Proteins
- SVM: Support Vector Machine

Contents

Statement of originality 1

Copyright statement 1

Abstract 2

Acknowledgments 3

Acronyms 4

1 Introduction 22

1.1 A brief history of therapeutics discovery . . . 22

1.2 Drug discovery today . . . 23

1.3 Opportunities in modern drug discovery . . . 24

1.3.1 Going beyond the "one disease, one gene, one drug" paradigm . . . 24

1.3.2 Deep learning on graphs for drug discovery . . . 25

1.3.3 A return to natural products in drug discovery . . . 25

1.4 Cancer . . . 26

1.4.1 Lung cancer . . . 26

1.4.2 Colorectal cancer . . . 27

1.4.3 Breast cancer . . . 28

1.5 The intersection of diet and cancer: implications for risk, prevention, and treatment . . . 28

1.6 Research questions and contributions . . . 30

1.7 List of publications . . . 32

2 Methods 35

2.1 Preliminaries . . . 35

2.2 Basic concepts and definitions . . . 35

2.3 Random walks . . . 36

2.4 Representation learning on graphs . . . 38

2.5 Causal inference . . . 40

3 Predicting food molecules with anticancer potential 42

3.1 Abstract . . . 42
3.2 Background . . . 42
3.3 Related work . . . 45
3.4 Contributions and impact . . . 47
3.5 Problem definition . . . 47
3.6 Datasets . . . 48
3.7 Methods . . . 50
3.8 Results . . . 55
3.9 Summary . . . 60

4 Predicting food molecules to enhance drug activity 63

4.1 Abstract . . . 63
4.2 Background . . . 63
4.3 Related work . . . 66
4.4 Contributions and impact . . . 70
4.5 Problem definition . . . 71
4.6 Datasets . . . 72
4.7 Methods . . . 75
4.8 Results . . . 86
4.9 Summary . . . 98

5 A principled method for lead discovery 102

5.1 Abstract . . . 102
5.2 Background . . . 102
5.3 Related work . . . 105
5.4 Contributions and impact . . . 107
5.5 Problem definition . . . 107
5.6 Datasets . . . 109
5.7 Methods . . . 115

5.8 Experimental setup . . . 120
5.9 Results . . . 122
5.10 Summary . . . 132

6 Conclusions **135**

6.1 Summary of contributions . . . 135
6.2 Future directions . . . 137

7 Appendix to Chapter 3 **139**

7.1 Neratinib and Calcitriol over-represented pathways . . . 139
7.2 Over-represented pathways for top predicted anticancer molecules . . . 141

8 Appendix to Chapter 4 **145**

8.1 Drugs and cell lines used in training . . . 145
8.2 Synergy experiments results . . . 146
8.2.1 Drug-drug combinations tested in A549 . . . 146
8.2.2 Drug-drug combinations tested in HCT116 . . . 149
8.2.3 Food molecule-drug combinations tested in A549 . . . 151

List of Tables

3.1 Summary of results (%) for anticancer drug prediction. By learning how to propagate information across the interactome, our approach predicts which drugs treat cancer up to 67% more effectively than random-walk-based approaches. Reported values are averaged across five-fold cross-validation (see the *Methods* section for more details); the top results are highlighted in bold. *-P represents neural models with pathway pooling. ACC=balanced accuracy, F1=harmonic mean of precision and recall, AUPR=area under the precision-recall curve, ac = anticancer. 56

3.2 Anticancer likeness of food molecules was computed using the best-performing neural model, and 20 molecules were predicted as anticancer molecules in addition to those reported in our previous work[121]. Description of molecules was extracted from FooDB[149]. 59

4.1 Predicted combinations tested in A549 and HCT116 86

5.1 Table with number of healthy, diseased, and treated samples for lung cancer (A549), breat cancer (MCF7), and prostate cancer (PC3), across genetic and chemical perturbagens. 111

5.2 Our problem formulation is similar to conventional node and graph classification tasks albeit some major differences exist. 117

7.1 We compute embeddings of the top correctly (neratinib) and incorrectly (calcitriol) classified drugs as anticancer for our best-performing model (ChebNet) and obtain the top 25 over-represented pathways using GSEA[152]. 139

7.2 Top over-represented pathways in KEGG[103] using GSEA[152] computed from embeddings of food molecules with highest anticancer likeness 141

8.3 We extracted drug-drug interaction data of the Merck and Co's study[180], and filtered samples to include only those with available drug targets and cell line mutations. The final set contains 546 unique pairwise drug interactions between 37 drugs in 32 cell lines. This table contains such drugs and cell lines. 145

List of Figures

3.1 **(A) Overview of our approach.** Given a drug represented by its protein targets, our goal is to predict an anticancer score representing the drug's known anticancer therapeutic potential. We propagate protein target information using message passing operators across a protein-protein interaction (PPI) network to generate diffusion profiles capturing affected protein communities. Learned diffusion profiles are then fed to a multi-layer perceptron (MLP) to predict an anticancer score. **(B) Message passing scheme.** Graph neural networks (GNNs) follow a message-passing scheme to propagate information across a network. In a message passing scheme, information is propagated in the form of one-hop neighbor messages that are aggregated and transformed using a nonlinear function. **(C) Pathway pooling.** We pool protein-level information into biological pathways using learnable parameters. **(D) Relationship to canonical graph prediction tasks.** Our task resembles canonical graph and node prediction tasks, but our method has distinct properties: (i) in graph classification, a set of graphs is paired with a set of node attributes, and the task is to predict a label for each graph; (ii) in node classification, a single graph is paired with a single set of node attributes, and the task is to predict a label for each node; (iii) in our formulation, a single graph is paired with a set of node attributes, and the task is to predict a label for each graph-node attribute pair (constant graph structure, varying node attributes). 52

3.2 **(A) Computation of attribution recall score.** We compute protein attribution scores for each sample classified as anticancer in the test set and average protein attribution scores across samples. Using GSEA[152], we compute over-represented pathways from the Kyoto Encyclopedia of Genes (KEGG) in the top-attributed proteins. We measure the attribution recall score (ARS) as the number of cancer-related, over-represented pathways divided by the total number of cancer-related pathways in KEGG. **(B) Our model predicted that food molecules pterostilbene and mercenene have high anticancer likeness.** We obtained diffusion profiles for pterostilbene and mercenene and found that both molecules affect genes known to participate in cancer-related pathways, either directly or through downstream network effects. The detection of diffusion profiles indicating that pterostilbene and mercenene affect genes in cancer-related pathways suggests their potential role in modulating these pathways, which could be pivotal in understanding their anticancer properties. Further investigation into the specific genes and pathways affected, as well as the underlying mechanisms of action, would be essential to substantiate these findings and harness their therapeutic potential. 57

4.1 **Drug-drug interactions depend on network proximity, biological function overlap, and context. (A)** In 99% of the interacting drug combinations in the Merck Co's dataset [180] (synergy value > 10 or < -10), proteins targeted by drugs are closer to disease (cancer cell line) genes than expected by chance. The distance metric used is the Z-score of the shortest path distance between drug combination targets and disease genes (see *Methods*). **(B)** In 97% of interacting drug combinations in Merck and Co's dataset [180] (synergy value > 10 or < -10), drug targets affect the same biological functions as mutated genes of cell lines where they interact more often than is expected by random chance. The functional overlap metric used is the Z-score of the intersection of Gene Ontology (GO) term multisets (see *Methods*). **(C)** Histogram of distribution of standard deviation of synergy values across cell lines per combination, for all (drug, drug, cell line) triplets in Merck and Co's dataset [180]. This plot shows that there is great variability of drug-drug synergy values depending on the context (cell line) in which they interact. **(D)** Distribution of synergy values across cell lines for the combination of Metformin and Dasatinib in Merck and Co's dataset [180], which exhibits a highly context-specific behavior. 67

4.2 **Merck and Co's dataset[180], an unbiased oncology screen**. **(A)** Distribution of drug targets for the 37 drugs in Merck and Co's (see Supplementary Table 8.3). Drug targets were extracted from STITCH[90]. Only high confidence drug-target associations were kept. **(B)** Distribution of mutations for the 32 cell lines in Merck and Co's dataset (see Supplementary Table 8.3). Cell line mutations were extracted from COSMIC CLP[91]. Mutations were filtered to keep only those recorded as "pathogenic". **(C)** Distribution of ZIP synergy scores. The ZIP synergy score represents the deviation, in percentage, of a drug combination effect compared to the expected effect under the hypothesis that the drugs do not interact. Synergy scores of samples in Merck and Co's dataset distribute evenly around 1.6. **(D)** Distribution of ZIP synergy positive ratio values per combination across cell lines. For a given drug combination, the positive ratio represents the number of cell lines in which the combination had a positive synergy value divided by the total number of cell lines the given combination was tested in. Positive ratios of 1 and 0 represent combinations with positive and negative synergy value, respectively, in all cell lines in which the combination was tested. 72% of combinations have positive ratios within [0.2, 0.8], ensuring that a high proportion of combinations have balanced positive/negative synergy scores across cell lines. 76

4.3 **Study design for food molecule-drug interaction prediction. (A)** A graph deep learning model is trained with annotated drug-drug interactions. After experimental validation, the trained model is used to predict food molecule-drug interactions. **(B -C)** We introduce a graph neural network (GNN)-based model for the prediction of synergy scores. Drug targets and cancer cell line mutations are encoded onto the PPI as node (protein) features. A first GNN diffuses information across the PPI, generating transformed node features capturing the initial node features and PPI topology. Node features are *pooled* or *propagated* to node's biological functions (pathways), using a learnable node feature transformation process. A second GNN diffuses information across the pathway ontology network generating transformed node embeddings capturing the initial pathway features (diffused from the PPI) and the pathway ontology network topology. Drug and cell line protein and biological function embeddings are concatenated and multiplied. A final feature aggregation reduces the protein and biological function embeddings dimensionality to 1. This embedding is fed to a Multi-Layer Perceptron (MLP) to predict a synergy score. **(D)** Shown is the generic message-passing mechanism followed by GNNs to propagate information across the graph. For a given protein node u, in each layer, information from the 1-hop neighborhood $\mathcal{N}(u)$ of u is aggregated with the current node information to generate a transformed representation of node u. This process is repeated for each node in the graph. **(E)** Shown is the pooling or *propagation* process from proteins to their biological functions. **(F)** Shown is the message-passing mechanism followed by GNNs for a single node. For a given pathway node u, in each layer, information from the 1-hop neighborhood $\mathcal{N}(u)$ of u is aggregated with the current node information to generate a transformed representation of node u. 78

4.4 **Our model predicts drug-drug synergy accurately and is interpretable.**

(A) Mean Squared Error (MSE), and Pearson correlation on the drug-drug ZIP synergy prediction task. Average performance values on held-out test set for the 5-fold cross-validation splits used for model hyperparameter tuning. **(B)** Test Mean Squared Error across 5 folds for each cell line showing that model performance remains consistent across cell lines with only a handful of cell lines showing poor convergence. **(C)** For all samples in the test set, we obtain predictive importance scores of proteins and biological functions across 5 folds and average them (see *Methods*). Then, we draw a network with the proteins and biological functions most relevant for model predictions. **(D)** Shown are the top predictive proteins and biological functions in our proposed ChebNet-P model. Proteins and biological functions that most contribute to the predictive power of our proposed model are related to cancer and cell growth. Node size corresponds to the relative discriminating capacity of a given protein and biological function. Node color illustrates the node type. **(E)** Sorafenib and MK-2206 were predicted with a synergy score of 12.5. Their synergistic interaction could be explained by Sorafenib's targeting of multiple intracellular and cell surfaces kinases, which reduces blood flow to the tumor, and by MK-2206's inhibition of AKT, which helps sensitize tumors to Sorafenib. 88

4.5 **Experimental validation of drug-drug interactions in A549.** **(A)** We predicted synergy values for 17,766 new drug-drug combinations against A549 (see *Methods*). Shown are the combinations we subsequently tested experimentally. **(B)** Shown are predicted and experimentally measured synergy values for five combinations of which four were consistent with predicted results. **(C)** Shown is the synergy map of Rucaparib and Sunitinib computed using SynergyFinder. Rucaparib and Sunitinib were predicted with a synergy score of 3.85±3.42. Experimental testing showed a synergy of 4.84±1.59. **(D)** Network representation of the top-scored proteins in diffusion profiles computed by our model CheNet-P for the prediction of synergy score between Rucaparib and Sunitinib. Sunitinib's activity against A549 cells is driven by its effects on STK11/LKB1, a kinase that phosphorylates AMPK and triggers essential adaptations to subvert metabolic stress and promote tumor cell survival[217,218]; whereas Rucaparib acts as an adjuvant. Their combined effects lead to epithelial-mesenchymal transition (EMT) and metabolic disruption resulting in a synergistic apoptotic induction. We include in Supplementary files the embedding scores for all proteins and biological functions in our dataset. **(E)** Shown is the synergy map of Sunitinib and Bosutinib computed using SynergyFinder. Sunitinib and Bosutinib were predicted with a synergy score of 6.61±4.11. Experimental testing showed a synergy of 8.62±1.47. **(F)** Network representation of the top-scored proteins in diffusion profiles computed by our model CheNet-P for the prediction of synergy score between Sunitinib and Bosutinib. The additive interaction between Sunitinib and Bosutinib in A549 could be explained as a result of the the proapoptotic environment triggered through the inhibition of Fgr, FRK, EGFR, Pyk2 and MAP3K2. Sunitinib and Bosutinib also inhibit MAP3K7 function, promoting proteasomal degradation, and apoptosis induction. We include in Supplementary files the embedding scores for all proteins and biological functions in our dataset. 92

4.7 **Experimental validation of food molecule-drug interactions in A549.** **(A)** We predicted synergy scores for 122,655 new food molecule-drug combinations against lung cancer (cell line A549; see *Methods*). Shown are the combinations we subsequently tested experimentally. **(B)** Shown are predicted and experimentally measured synergy values for 3 combinations chosen based on availability and synergy scores. All combinations tested were consistent with predicted results. **(C)** Shown is the synergy map of Sunitinib and Brassinolide computed using SynergyFinder (see *Methods*). Sunitinib and Brassinolide were predicted with a synergy value of 4.40 ± 3.53. Experimental testing showed a synergy of 3.86 ± 3.34. **(D)** Network representation of the top-scored proteins in diffusion profiles computed by our model CheNet-P for the prediction of synergy score between Sunitinib and Brassinolide. Their additive interaction against A549 cells could be explained by their combined inhibition of different cell processes such as metabolic adaptation, cell growth, cell cycle progression, and cell survival. Sunitinib induces extrinsic apoptotic pathways through its inhibition of CSNK1D/CK1 and CSNK1E/CK1. Brassinolide induces reactive oxygen species (ROS) production which leads to apoptosis induction via intrinsic pathway induction.This activation of both extrinsic and intrinsic apoptotic signaling pathways converges in the activation of executor caspases 3/7. We include in Supplementary files the embedding scores for all proteins and biological functions in our dataset. 97

5.1 **Overview of datasets and PDGrapher**. **(A,B)** We train PDGrapher across 2 genetic (A) and 2 chemical (B) datasets, where each dataset is comprised by healthy, diseased, and treated gene expression samples for one cell line, and one noisy causal graph (PPI). **(C)** Given a paired diseased and treated gene expression samples, PDGrapher first predicts a candidate perturbagen to shift cell gene expression from diseased to treated state through a perturbagen discovery module f_p. Then PDGrapher takes the diseased gene expression sample and the predicted perturbagen and predicts the treated gene expression through a response prediction module f_r. **(D)** Our perturbagen discovery module f_p is optimized using 2 losses: a cross-entropy cycle loss to predicted a perturbagen $\mathcal{U}'$ which would shift diseased cell state to a state as close as possible to the treated state: $CE\left(\mathbf{x}^t,\ f_r(\mathbf{x}^d,\ f_p(\mathbf{x}^d,\ \mathbf{x}^t))\right)$ $(with\ f_r\ frozen)$, and a cross-entropy supervision loss that directly supervises the prediction of $\mathcal{U}'$: $CE\left(\mathcal{U}', f_p(\mathbf{x}^d,\ \mathbf{x}^t)\right)$. See *Methods* for more details. **(E)** PDGrapher has two modules, f_r and f_p, both of which are GNN-based models. Depicted is f_r, which takes as input a diseased gene expression and a perturbagen and represents perturbagens effects in the noisy causal graph as edge mutilations. Both f_r and f_p follow the standard message-passing framework, where node representations are updated by aggregating the information from neighbors in the graph. 123

5.2 **PDGrapher predicts genetic perturbagens to achieve a desired cell state. (A)** We evaluate PDGrapher's performance across two settings: randomly splitting samples between the training and test sets, and the leave-cell-out setting where we train in one cell line and evaluate PDGrapher's performance in the remaining cell line. **(B,C)** Performance metrics in the perturbagen discovery task for Genetic-PPI-Lung (B) and Genetic-PPI-Breast (C). PD accuracy = perturbagen discovery accuracy; Proportion non-zeros = proportion of samples from the test set in which the predicted perturbagen has an overlap of 0 with the ground truth perturbagen. **(D,E)** Ground-truth interventions (knocked-out genes) are at the top of predicted rankings by PDGrapher more often than would be expected with a *Random* predictor for both Chemical-PPI-Lung (D) and Genetic-PPI-Breast (E). **(F,G)** Interventions predicted by PDGrapher are closer in the network to ground-truth interventions compared to the *Random* baseline for Genetic-PPI-Breast (G), whereas the opposite is observed for Chemical-PPI-Lung (F). *There are dashes in performance tables where it was not possible to obtain a given metric given a model's output. 126

5.3 **PDGrapher predicts chemical perturbagens to achieve a desired cell state. (A, B)** Performance metrics in the perturbagen discovery task for Chemical-PPI-Lung (A) and Chemical-PPI-Breast (B). PD accuracy = perturbagen discovery accuracy; Proportion non-zeros = proportion of samples from the test set in which the predicted perturbagen has an overlap of 0 with the ground truth perturbagen. **(C,D)** Ground-truth interventions (targeted genes) are at the top of predicted rankings by PDGrapher more often than would be expected by chance, for both Chemical-PPI-Lung (C) and Chemical-PPI-Breast (D). **(E,F)** Interventions predicted by PDGrapher are closer in the network to ground-truth interventions compared to random, for both Chemical-PPI-Lung (E) and Chemical-PPI-Breast (F). *There are dashes in performance tables where it was not possible to obtain a given metric given a model's output. 128

5.4 **PDGrapher predicts genetic perturbagens to achieve a desired cell state for new cell lines. (A,B)** Performance metrics in the perturbagen discovery task in the leave-cell-out setting for Genetic-PPI-Lung (A) and Genetic-PPI-Breast (B). PD accuracy = perturbagen discovery accuracy; Proportion non-zeros = proportion of samples from the test set in which the predicted perturbagen has an overlap of 0 with the ground truth perturbagen. **(C,D)** Ground-truth interventions (knocked-out genes) are at the top of predicted rankings by PDGrapher more often than would be expected by chance, for both Genetic-PPI-Lung (C) and Genetic-PPI-Breast (D) **(E,F)** Interventions predicted by PDGrapher are closer in the network to ground-truth interventions compared to the *Random* baseline, for Genetic-PPI-Lung (D), whereas the opposite is observed for Genetic-PPI-Breast (F). *There are dashes in performance tables where it was not possible to obtain a given metric given a model's output. 129

5.5 **PDGrapher predicts chemical perturbagens to achieve a desired cell state for new cell lines. (A,B)** Performance metrics in the perturbagen discovery task in the leave-cell-out setting for Chemical-PPI-Lung (A) and Chemical-PPI-Breast (B). PD accuracy = perturbagen discovery accuracy; Proportion non-zeros = proportion of samples from the test set in which the predicted perturbagen has an overlap of 0 with the ground truth perturbagen. **(C,D)** Ground-truth interventions (targeted genes) are at the top of predicted rankings by PDGrapher more often than would be expected by chance, for both Chemical-PPI-Lung (C) and Chemical-PPI-Breast (D). **(E,F)** Interventions predicted by PDGrapher are closer in the network to ground-truth interventions compared to the *Random* baseline, for both Chemical-PPI-Lung (E) and Chemical-PPI-Breast (F). *There are dashes in performance tables where it was not possible to obtain a given metric given a model's output. . 131

5.6 **Ablation of PDGrapher's loss components.** **(A)** Performance metrics of PDGrapher in Genetic-PPI-Lung. **(B)** Performance metrics of PDGrapher in Genetic-PPI-Breast. **(C)** Performance metrics of PDGrapher in Chemical-PPI-Lung. **(D)** Performance metrics of PDGrapher in Genetic-PPI-Lung. 132

8.1 Synergy map for Midostaurin + Sunitinib 146

8.2 Synergy map for Sunitinib + Rucaparib 147

8.3 Synergy map for Sunitinib + Bosutinib 147

8.4 Synergy map for Temozolomide + Rucaparib 148

8.5 Synergy map for Mitomycin + Sorafenib 148

8.6 Synergy map for Sunitinib + Rucaparib 149

8.7 Synergy map for Bosutinib + Nintedanib 149

8.8 Synergy map for Sorafenib + Bosutinib 150

8.9 Synergy map for Sunitinib + Nintedanib 150

8.10 Synergy map for Midostaurin + Bosutinib 151

8.11 Synergy map for Sunitinib + Bosutinib 151

8.12 Synergy map for Dactosilib + Galangin. Replicate 1. 152

8.13 Synergy map for Dactosilib + Galangin. Replicate 2. 152

8.14 Synergy map for Dactosilib + Galangin. Replicate 3. 153

8.15 Synergy map for Sunitinib + Brassinolide 153

8.16 Synergy map for Sunitinib + Linoelaidic acid 154

1 Introduction

1.1 A brief history of therapeutics discovery

The study of therapeutics dates back to the early days of human civilization, when people sought to understand and harness the healing properties of natural compounds[1–3]. The earliest records of medicinal use of plants were found in Mesopotamia on clay tablets, in cuneiform (2600 B.C.), documenting oils from Cupressus sempervirens (Cypress) and Commiphora species (myrrh) that are still in use today to treat coughs, colds, and inflammation[2]. The medicinal use of natural products was studied across the globe. For example, the Egyptians and Greeks used a wide variety of plant-based remedies and treatments, and the Chinese and Indians developed complex systems of traditional medicine[1–3].

However, the beginning of drug research as we know it today is recent, dating back to the 1870s in central Europe[4]. At the University of Strasbourg, Paul Ehrlich observed the selective affinity of dyes for biological tissues, leading him to postulate the existence of chemoreceptors[4]. Ehrlich argued that certain chemoreceptors on foreign organisms or cancer cells would be different from analogous structures in host tissues, and that these differences could be exploited therapeutically. This marked the birth of chemotherapy.[4].

The year 1928 saw the breakthrough discovery of penicillin, a molecule with antibiotic activity produced by Penicillium chrysogenum[5], by Alexander Fleming. This propelled the discovery of countless antibiotics from natural sources, including streptomycin, gentamicin, and tetracycline[5]. Given the evident benefits of natural products as potential drug sources, pharmaceutical companies increased their natural product-based drug discovery efforts: during the second half of the 20th century, every pharmaceutical company developed a natural product-based discovery program focused on the treatment of a wide variety of diseases (e.g., antibacterial, antifungal, infectious diseases)[5]. During the 1970s, compactin and mevinolin were discovered. These molecules were able to inhibit cholesterol biosynthesis and enable the production of statins which are used to this day[5].

Despite these groundbreaking discoveries, the use of natural products as drug discovery sources began to decline in the late 20th century as synthetic drugs and pharmaceuticals became more prevalent. This shift was driven by several factors, including high costs, dif-

ficulty in sourcing/harvesting sufficient biological starting material, low economic returns, difficulties in isolation and/or purification procedures, and lack of infrastructure and insufficient capital investments[5]. Arguably, the introduction of combinatorial and computational chemistry techniques, in particular high-throughput screening (HTS) technologies, played the greatest role in the decline of natural product-based drug discovery efforts[5]. HTS technologies promised to revolutionize drug discovery by allowing for the simultaneous screening of thousands of single compounds and combinations of compounds at one time which could result in the rapid identification of potential therapeutics. However, HTS technologies were cumbersome and sometimes impossible to use for the screening of compounds extracted from natural sources due to their tendency to decompose, precipitate, interfere with reagents, and potential have unspecified effects[5]. Therefore, the introduction of HTS technologies brought with it a shift towards using known or synthetically generated compounds for drug discovery.

1.2 Drug discovery today

The current drug discovery pipeline comprises several phases[6]. The first phase, *target discovery and validation*, consists of discovering single molecular targets or sets of targets that are related to disease. In the second phase, *library design and hit discovery*, compounds with validated activity against the discovered targets are designed, developed, and tested. The third stage, *hit-to-lead and lead optimization*, consists of selecting the most promising hits and optimizing them for potency while minimizing off-target bioactivity. The final phase consists of *preclinical and clinical studies* to optimize dose, formulation, and method of administration and to test for efficacy and safety in model organisms and patients.

HTS technologies have greatly improved the efficiency of the first two phases, target and hit discovery. In spite of this, the expected benefits of HTS technologies have yet to materialize[2,5,7]. For example, from 1981 to 2010, only 36% of the 1,135 approved drugs were purely synthetic whereas more than half were molecules from natural sources, derivatives, or analogues[8]. One reason for this is the high costs associated with the acquisition of HTS equipment and infrastructure. In addition, HTS technologies often have a high failure rate, with a large number of compounds identified through screening failing to progress to clinical trials due to unexpected pharmacodynamic properties, poor pharmacokinetics, or unexpected toxicity[6]. This can result in a significant loss of resources and time for compa-

nies. Furthermore, HTS technologies often rely on existing targets, which limits their ability to identify novel compounds. This can result in the development of drugs that are similar to existing drugs rather than truly novel therapeutics.

Recent efforts to improve the efficacy of the drug discovery pipeline leverage artificial intelligence (AI) to either complement HTS technologies or substitute them altogether. Particularly, network-based machine learning algorithms have proved especially useful in the first three phases of the drug discovery pipeline[9,10].

1.3 Opportunities in modern drug discovery

1.3.1 Going beyond the "one disease, one gene, one drug" paradigm

The rise of HTS technologies was accompanied with the consolidation of the one disease, one gene, one drug paradigm in which researchers focused on identifying a single disease-associated gene against which a maximally selective compounds could be designed[11–13]. This paradigm was adopted to simplify compound screening, reduce unwanted side effects, and simplify registration[14]. However, the one disease, one gene, one drug paradigm proved to be overly simplistic: modern studies have shown that disease mechanisms are not driven by single-gene alterations; they are in fact complex subnetworks within the protein-protein interaction (PPI) network[14,15]. Knockout experiments have also revealed phenotypic robustness, showing that as few as 10% of druggable gene knockouts, in mice, translate into a phenotype that may be of value for drug target validation[9]. Due to this robustness, multitarget therapies appear as a more suitable alternative to circumvent redundant functions and alternative compensatory signaling routes[9].

Novel systems pharmacology and network-based approaches aim to go beyond the limitations of the one disease, one gene, one drug paradigm by characterizing disease and drug mechanisms in a broader sense through the use of heterogeneous high-throughput screens[16]. In network-based drug discovery, data on the interactions between proteins, biological pathways and diseases is leveraged to characterize diseases and find new therapeutics. Instead of disease genes, network-based efforts aim to leverage perturbations in network-causing diseases, and take into account redundancy and pleiotropy to design suitable therapies[9,14,15]. Many network-based approaches have been developed to understand disease mechanisms[17,18], decipher new drug-target interactions[19–22], reposition drugs[20,23,24], and predict drug side ef-

fects[25].

1.3.2 Deep learning on graphs for drug discovery

Graph deep learning is an umbrella term encompassing deep learning models operating on graph-structured data[26]. These models combine the strong representational and fitting capabilities of deep learning with the inductive biases encoded in graph-structured data. Using deep learning models that operate directly on graphs allows for the specification of predictive tasks within graphs themselves, for example, predicting a label for a node, an edge, a subgraph, or a whole graph. Despite its recent emergence, it has become an established and powerful approach for network-based drug discovery[27]. Some prominent examples include graph deep learning methods for target discovery, hit discovery and design, as well as drug side effect prediction[28–30].

1.3.3 A return to natural products in drug discovery

The 1950s saw the beginning of the "golden age of natural products" after the award of Nobel Prizes in physiology or medicine to Sir Alexander Fleming, Ernst B. Chain, and Sir Howard Florey in 1945 for the discovery of penicillin and to Selman A. Waksman in 1952 for the discovery of streptomycin[7]. However, the advent of HTS technologies and subsequent focus on synthetically generated compounds brought about the decline of natural product-based drug discovery during the late 20th century. This has correlated with the reduction in new leads and the substantial decline in drug approval[7]. There is evidence, however, that a return to natural products in drug discovery might be approaching. In 2015, the Nobel Prize in physiology or medicine was awarded to William C. Campbell and Satoshi Omura for the discovery of the microbial natural product avermectins and to Youyou Tu for the discovery of the natural plant product artemisinin[31]. Ivermectin, a derivative of avermectins, extracted from the bacteria Streptomyces avermitilis, has radically lowered the incidence of river blindness and elephantiasis, whereas artemisin, an extract from the plant Artemisia annua, has significantly reduced the mortality rates for malaria patients[7]. This, together with advances in microbial genomics and metagenomics, synthetic biology, and bioinformatics tools facilitating the analysis of natural compounds, signals that scientific interest is increasing in natural compounds as therapeutic sources[7].

Natural products have been generated as a result of centuries of adaptation to guarantee survival against the environment; therefore, they often have unique chemical structures and biological activities that are not found in synthetic compounds, which can make them valuable targets for drug discovery[2,5]. In addition, natural compounds typically have multiple targets, which makes them great potential candidates for network-based drug discovery approaches.

1.4 Cancer

Cancer, a complex and multifaceted disease, is characterized by uncontrolled cell growth and the ability of these cells to invade other tissues. It remains one of the leading causes of death globally, with its impact and clinical significance widely acknowledged[32–34]. The standard of care and existing drugs for cancer vary based on the type and stage of the disease, and treatment often involves a combination of surgery, radiation therapy, chemotherapy, and newer approaches like targeted therapy and immunotherapy[35–38]. In this dissertation, we specifically focus on lung cancer (cell line A549) and colorectal cancer (cell line HCT116) in Chapter 4, and on lung cancer (cell line A549) and breast cancer (cell line MCF7) in Chapter 5. The following subsections introduce the available clinical knowledge and rationale behind our choice.

1.4.1 Lung cancer

Lung cancer, one of the leading causes of cancer mortality worldwide, is primarily divided into Non-Small Cell Lung Cancer (NSCLC) and Small Cell Lung Cancer (SCLC). NSCLC, which accounts for approximately 85% of lung cancer cases, is further classified into adenocarcinoma, squamous cell carcinoma, and large cell carcinoma[39]. These subtypes differ in their cellular origin, genetic mutations, and response to treatment. The management of NSCLC often involves a combination of surgical resection, chemotherapy, radiation therapy, and increasingly, targeted therapies and immunotherapies[40,41]. Targeted treatments, especially effective in patients with mutations in genes such as EGFR, ALK, and ROS1, include tyrosine kinase inhibitors like efitinib, erlotinib, and crizotinib[40,41]. Immunotherapies such as pembrolizumab, nivolumab, and atezolizumab have become standard for advanced NSCLC, particularly for patients with high PD-L1 expression. Research continues to focus on iden-

tifying new biomarkers and developing novel targeted agents to improve outcomes for lung cancer patients[40,41].

SCLC, characterized by rapid growth and early spread, accounts for the remaining 15% of lung cancer cases[42]. Due to its aggressive nature, SCLC is typically treated with a combination of chemotherapy and radiation. Recently, immunotherapy drugs such as atezolizumab and durvalumab have been approved for use in SCLC, showing promising results in extending survival[43].

We chose cell line A549 in Chapter 4 and Chapter 5 because it is a well-established model for studying NSCLC, the predominant form of lung cancer. Research on A549 can provide insights into the the potential of dietary molecules to enhance existing treatments (Chapter 4), and potentially new treatments (Chapter 5). This is particularly relevant given the challenges in treating NSCLC effectively[44,45].

1.4.2 Colorectal cancer

Colorectal cancer, which includes cancers of the colon and rectum, is a major global health concern[37]. It typically begins as benign polyps that can evolve into cancer over time. Early detection through screening is vital for effective treatment. The treatment of colorectal cancer varies based on the stage of the disease and may include surgical resection, chemotherapy, radiation therapy, and targeted therapies[46]. In early stages, surgery is often the primary treatment, sometimes followed by adjuvant chemotherapy. For more advanced stages, chemotherapy regimens including drugs like 5-fluorouracil, oxaliplatin, and irinotecan are commonly used[47]. Targeted therapies, such as bevacizumab (anti-VEGF) and cetuximab or panitumumab (anti-EGFR), are employed in metastatic colorectal cancer, often in combination with chemotherapy[48].

Recent advances in molecular profiling have led to the identification of specific subgroups of colorectal cancer that benefit from particular treatments. For instance, tumors with high microsatellite instability (MSI-H) or mismatch repair deficiency (dMMR) are more responsive to immunotherapies like pembrolizumab and nivolumab[49,50]. The ongoing research in colorectal cancer focuses on better understanding the molecular pathways involved in the disease and developing new therapeutic agents[51,52].

We chose cell line HCT116, a valuable model for colorectal carcinoma, in Chapter 4. In

Chapter 4, we find molecules that synergize with drugs in colorectal cancer, which is specially relevant given the myriad of studies documenting the associations between diet and risk and progression of colorectal cancer[53–56].

1.4.3 Breast cancer

Breast cancer, the most common cancer among women globally, exhibits significant heterogeneity, with various subtypes requiring different treatment approaches[57]. The main subtypes of breast cancer are hormone receptor-positive, HER2-positive, and triple-negative breast cancer[58].

Hormone receptor-positive breast cancers, which express estrogen or progesterone receptors, are typically treated with hormone therapies such as tamoxifen or aromatase inhibitors like letrozole and anastrozole[57,58]. These therapies aim to block the effects of hormones on breast cancer growth. In recent years, CDK4/6 inhibitors like palbociclib have been used in combination with hormone therapy for advanced hormone receptor-positive breast cancers, showing improved survival rates[59]. HER2-positive breast cancers, characterized by the overexpression of the HER2 protein, are treated with targeted therapies such as trastuzumab, pertuzumab, and ado-trastuzumab emtansine[60]. Triple-negative breast cancer, which lacks hormone receptors and HER2, is more challenging to treat and is generally managed with chemotherapy[61]. However, recent advancements include the approval of PARP inhibitors, like olaparib and talazoparib, for patients with BRCA mutations[62].

We chose cell line MCF7, a model for hormone receptor-positive breast cancer, in Chapter 5, where we develop a model for the prediction of therapeutics that reverse disease effects. This can contribute to identifying new phenotypic-driven leads and understanding treatment responses.

1.5 The intersection of diet and cancer: implications for risk, prevention, and treatment

The complex relationship between diet and cancer encompasses the influence of dietary choices on cancer risk, the potential of specific food molecules in cancer prevention, and the role of nutrition in cancer treatment. This multifaceted interaction highlights the significance of diet in both the development and management of cancer[63].

Epidemiological studies have consistently linked certain dietary patterns to the risk of developing cancer. Diets that are high in processed meats, red meats, and saturated fats are associated with an increased risk of cancers such as colorectal, breast, and prostate cancer[64,65]. In contrast, diets rich in fruits, vegetables, whole grains, and lean proteins, often exemplified by Mediterranean or plant-based diets, have been correlated with a reduced risk of various cancer types[56,66,67]. The beneficial properties of these diets are attributed to their richness in antioxidants, fibers, and phytochemicals. Antioxidants like vitamins C and E, selenium, and carotenoids are particularly noted for their capacity to protect cells from oxidative stress, a known contributor to cancer development[68].

Beyond general dietary patterns, specific food molecules have garnered attention for their cancer-preventing properties. Phytochemicals, particularly flavonoids found in an array of fruits, vegetables, and teas, have shown promise in inhibiting tumor growth and inducing apoptosis in cancer cells[67,69–71]. Similarly, cruciferous vegetables such as broccoli and Brussels sprouts are rich in glucosinolates. These compounds, when metabolized, form biologically active agents like sulforaphane, indoles, and isothiocyanates, which exhibit chemopreventive effects[72–75]. They achieve this by modulating detoxification enzymes, altering hormone metabolism, and exerting anti-inflammatory and antioxidant effects, all of which are mechanisms implicated in cancer prevention.

The impact of diet extends into cancer progression and response to treatment. Emerging research suggests that certain dietary approaches may influence tumor growth rates and enhance the efficacy of conventional cancer therapies[76]. Calorie restriction, for example, has been observed to slow tumor growth and improve the response to chemotherapy and radiation therapy in some animal studies[77]. Additionally, diets like the ketogenic diet, which is high in fats and low in carbohydrates, are being explored for their potential to limit glucose availability to cancer cells, thereby possibly impeding their growth[78].

Nutrition's role is also critical during cancer treatment. Adequate nutrition can affect a patient's strength, tolerance to treatment side effects, and overall quality of life[79]. Malnutrition, a common challenge among cancer patients, can lead to poorer outcomes. 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 cancer therapy[79,80].

This exploration underscores the dynamic and evolving nature of the relationship between diet and cancer. While further research is necessary to fully understand and leverage this connection, current evidence strongly supports the role of diet in cancer risk reduction, potential therapeutic intervention, and the enhancement of patient outcomes during and after cancer treatment[81,82]. However, determining the effective dosage of food molecules that might translate into a protective or therapeutic effect is complex due to the variability in their natural concentrations, differences in human absorption and metabolism, and individual genetic factors[83,84]. This complexity is further compounded by significant research gaps. Many of these bioactive compounds have low bioavailability, meaning only a small portion may reach the tumor site[85]. Additionally, the majority of current evidence is based on in vitro studies or animal models, with a notable lack of large-scale human clinical trials. This situation is further complicated by potential interactions between different food molecules and between these molecules and conventional drugs, which are not yet fully understood[86,87]. Despite these challenges, the progressive strides in nutritional science and oncology offer a promising horizon, suggesting that with continued research and innovative approaches, we can enhance our understanding and utilization of food molecules in cancer prevention and treatment, paving the way for more personalized and effective dietary strategies.

1.6 Research questions and contributions

Having introduced the relevant background detailing the importance of network-based methods and natural products for drug discovery, we next introduce three research questions that we seek to address in this dissertation. We first develop methods to leverage existing natural compound data for hit identification (**Q1**). Then, we turn our focus to network-based methodological advancements for lead identification (**Q2**).

Q1. Can we build graph deep learning algorithms that leverage natural compounds for drug discovery?

Previous computational research, facilitated by growing databases on drugs, diseases, and foods[88–94], has explored relationships between plant-based compounds, drugs, and diseases[94,95] based on the overlap of targeted or differentially expressed genes. Our goal is to push this forward, following principles of network-based drug discovery, by studying relationships be-

tween plant-based compounds, drugs, and diseases in terms of affected biological pathways, rather than individual genes. Specifically, we address two sub-questions:

Q1.1. Can graph deep learning algorithms surpass traditional network-based algorithms in the prediction of plant-based compounds with therapeutic properties? In addition, are these methods able to capture inductive biases that exist in biological data?

We address this question in Chapter 3. We previously introduced a network-based machine learning algorithm to predict plant-based compounds with therapeutic properties against cancer based on the similarity of affected pathways compared to FDA-approved existing drugs. In Chapter 3, we introduce an optimized algorithm leveraging operators from the modern trend of deep learning on graphs[26]. We show increased performance, and interpretability of our deep learning model and predict an additional set of plant-based compounds as potentially anti-carcinogenic.

Q1.2. Can we build a graph deep learning algorithm to predict plant-based compounds that synergize with existing drugs?

We address this question in Chapter 4. We introduce a graph deep learning method to predict plant-based compounds with synergistic effects against cancer in combination with approved therapeutics. To do this, we first build and calibrate a graph deep learning-based algorithm to predict the synergy score of drug combinations against several cancer cell lines. We validate this model computationally and experimentally and then use it for the prediction of synergy between plant-based compounds and drugs against cancer.

Q2. Given the limitations of the current lead identification problem formulation, can we propose a novel formulation for network-based lead identification together with an algorithm to address it?

We address this question in Chapter 5. Current lead identification approaches are limited to compounds from curated libraries. We aim to expand the space for lead identification by introducing a novel formulation of the task. Instead of predicting which compounds, from a predefined list of candidates, would have therapeutic properties against a disease, we reformulate the task so that the goal is to predict what genes a lead compound should target to reverse the effects of disease. We introduce a causally grounded algorithm to address this,

and we evaluate our algorithm's performance in the target prediction task in various genetic and chemical perturbation datasets.

This dissertation is organized as follows: Chapter 2 introduces basic concepts and necessary mathematical formulations; Chapters 3, 4, and 5 address our research questions; and Chapter 6 contains a summary of the conclusions and contributions of this dissertation. In this dissertation, we work in the domain of computational drug discovery, focusing on two core objectives. The first is the development and application of advanced graph deep learning algorithms to identify potential therapeutic natural compounds. Simultaneously, we confront the limitations of traditional lead identification in drug discovery. However, the transition from these computational innovations to tangible clinical applications is not straightforward. The predictive models and hypotheses generated within this computational framework necessitate thorough experimental validation. This process encompasses a comprehensive series of laboratory experiments, preclinical studies, and clinical trials, all aimed at confirming the efficacy and safety of these predicted compounds and targets. The path from computational discovery to clinical application is complex and multi-layered, involving rigorous scientific validation and adherence to regulatory standards. Thus, while this dissertation contributes significantly to computational drug discovery, particularly in the exploration of natural compounds and novel therapeutic targets, the direct clinical translation of these findings remains a long-term and multi-stage endeavor.

1.7 List of publications

As a first author

1. **Gonzalez, G.**, Herath, I., Veselkov, K., Bronstein, M., Zitnik, M. (2023, under review). Combinatorial prediction of therapeutic perturbations using a causally-inspired neural network.

2. **Gonzalez, G.**, Laponogov, I., Brignardello, J., Chio, D., Reyna-Jeldes, M., Rita, L., Qiu, J., Gallota, M., Wang, Y., Zhang, E., Zhong, Y., Wang,X., Bronstein, M., Veselkov, K. (2023, in preparation). Leveraging food-drug pharmacodynamic interactions to enhance drug activity.

3. Southern, J.*,**Gonzalez, G.***, Borgas, P., Poynter L., Laponogov, I., Zhong, Y., Minezarmi, R., Veselkov, D., Bronstein, M., Veselkov, K. (2023). Genomic-driven nutritional interventions for radiotherapy-resistant rectal cancer patient. Scientific Reports, accepted manuscript. *Equal contribution.

4. **Gonzalez, G.**, Gong, S., Laponogov, I., Bronstein, M., Veselkov, K. (2021). Predicting anticancer hyperfoods with graph convolutional networks. Human Genomics, 15(1), 33. https://doi.org/10.1186/s40246-021-00333-4.

5. Laponogov, I.*, **Gonzalez, G.***, Shepherd, M., Qureshi, A., Veselkov, D., Charkoftaki, G., Vasiliou, V., Youssef, J., Mirnezami, R., Bronstein, M., Veselkov, K. (2021). Network machine learning maps phytochemically rich "Hyperfoods" to fight COVID-19. Human Genomics, 15(1), 1. https://doi.org/10.1186/s40246-020-00297-x. *Equal contribution.

6. **Gonzalez, G.***, Gong, S.*, Laponogov, I., Veselkov, K., Bronstein, M. (2019). Graph Attentional Autoencoder for Anticancer Hyperfood Prediction. 2019 NeurIPS Graph Representation Learning Workshop.https://arxiv.org/abs/2001.05724 *Equal contribution.

As a collaborator

1. Rita, L, Neumann, N., Laponogov, I., **Gonzalez, G.**, Veselkov, D., Pratico, D., Thompson, D., Aalizadeh, R., Thomaidis, N., Vasiliou, V., Veselkov, K. (2023). Alzheimer's disease: using gene/protein network machine learning for molecule discovery in olive oil. Human Genomics 17, 57 (2023). https://doi.org/10.1186/s40246-023-00503-6.

2. Giunchiglia, V., Shukla, C.V., **Gonzalez, G.**, Agarwal, C. (2022). Towards Training GNNs using Explanation Directed Message Passing. Learning on Graphs Conference. https://arxiv.org/abs/2211.16731.

3. Zhong, Z., **Gonzalez, G.**, Grattarola, D., Pang, J. (2022). Unsupervised Heterophilous Network Embedding via r-Ego Network Discrimination. https://doi.org/10.48550/arxiv.2203.10866.

4. Borgas, P., **Gonzalez, G.**, Veselkov, K., Mirnezami, R. (2021). Phytochemically rich dietary components and the risk of colorectal cancer: A systematic review and meta-

analysis of observational studies. World Journal of Clinical Oncology, 12(6), 482–499. https://doi.org/10.5306/wjco.v12.i6.482.

5. Frasca, F., Galeano, D., **Gonzalez, G.**, Laponogov, I., Veselkov, K., Paccanaro, A., Bronstein, M. M., Lugano, U., Holloway, R. (2019). Learning Interpretable Disease Self-Representations for Drug Repositioning. 2019 NeurIPS Graph Representation Learning Workshop. https://doi.org/10.48550/arxiv.1909.06609.

6. Veselkov, K., **Gonzalez, G.**, Aljifri, S., Galea, D., Mirnezami, R., Youssef, J., Bronstein, M., Laponogov, I. (2019). HyperFoods: Machine intelligent mapping of cancer-beating molecules in foods. Scientific Reports, 9(1), 9237. https://doi.org/10.1038/s41598-019-45349-y.

2 Methods

2.1 Preliminaries

Here we introduce the notation used throughout this dissertation. A calligraphic letter $\mathcal{X}$ indicates a set, an italic uppercase letter X denotes a graph, uppercase $\mathbf{X}$ denotes a matrix, lowercase $\mathbf{x}$ denotes a vector, and a monospaced letter $\mathtt{X}$ indicates a tuple. Uppercase letter X indicates a random variable, and lowercase letter x indicates its corresponding value; bold uppercase $\mathbf{X}$ denotes a set of random variables, and lowercase letter $\mathbf{x}$ indicates its corresponding values. We denote $P(\mathbf{X})$ as a probability distribution over a set of random variables $\mathbf{X}$ and $P(\mathbf{X} = \mathbf{x})$ as the probability of $\mathbf{X}$ being equal to the value of $\mathbf{x}$ under the distribution $P(\mathbf{X})$. For simplicity, $P(\mathbf{X} = \mathbf{x})$ is abbreviated as $P(\mathbf{x})$.

2.2 Basic concepts and definitions

In biology, networks can be used to represent data at different levels of abstraction, from molecular interactions to population dependencies, including social and health interactions. Tools from network analysis have been extensively used in biology and medicine in applications such as drug target identification, gene function prediction, and early diagnosis of diseases[96]. Networks used in systems biology include protein-protein interaction (PPI) networks, biochemical networks, transcriptional regulation networks, and signal transduction and metabolic networks[96]. In this dissertation, we leverage PPI networks for various applications. PPI networks represent physical interactions between proteins within a cell. The basic assumption is that proteins that participate in the same biological processes are tightly connected in the PPI. PPI networks are created using large-scale and high-throughput techniques, including 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. Well-known databases include the BioGRID database[97], BioPlex[98], the Human Protein Reference Database (HPRD)[99], STRING[100], and STITCH[90]. Proteins whose interactions lead to a certain product or a change in a cell are said to belong to the same biological *pathway*. There are several types of biological pathways, including metabolic pathways, signaling pathways, and regulatory networks[101]. These are collected

in a variety of pathway databases, including HumanCyc[102], Kyoto Encyclopedia of Genes and Genomes (KEGG[103]), National Cancer Institute-Pathway Interaction Database (NCI-PID[104]), Panther[105], and Reactome[106].

Mathematically, a network or graph can be represented as $G = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ where $\mathcal{V}$ is the set of nodes, $|\mathcal{V}| = N$ the number of nodes, $\mathcal{E}$ is a set of edges representing the connections between the nodes, and $\mathbf{X} = \mathbb{R}^{N \times d}$ represents the d-dimensional node attributes. $e = \{(v_i, v_j) | v_i, v_j \in \mathcal{V}\}$ is defined as the single edge between nodes v_i and v_j. In this case, v_i and v_j are said to be neighbors, and the set of neighbors of a node v_i is called its neighborhood $\mathcal{N}(v_i) = \{v_j : (v_j, v_i) \in \mathcal{E}\}$. Graphs are represented using an adjacency matrix. Given a graph $G = (\mathcal{V}, \mathcal{E}, \mathbf{X})$, the adjacency matrix representation consists of an $N \times N$ matrix $\mathbf{A} = (a_{ij})$ such that $a_{ij} = 1$ if $(v_i, v_j) \in \mathcal{E}$ or $a_{ij} = 0$ otherwise. For a weighted graph $a_{ij} = w_{ij}$ if $(v_i, v_j) \in \mathcal{E}$ or $a_{ij} = 0$ otherwise.

2.3 Random walks

Although using graphs to represent biological systems is effective in capturing complex relationships in them, it is still difficult to extract meaningful patterns from biological data lying on graphs because this data is often sparse and noisy. To overcome this limitation, network propagation has been effectively applied on graphs in numerous applications, including drug and disease protein target prediction[107,108], disease stratification[109,110], and drug repositioning[107,111–113]. Network propagation has been shown to amplify biological signals lying on networks and facilitate the finding of significant patterns characterizing biological systems[114].

In the framework of network propagation, the starting point is a vector $\mathbf{p}_0$ of scores on nodes (e.g., proteins or genes) representing prior knowledge (e.g., targeted proteins in the case of drugs, or altered proteins in the case of diseases). We could set $\mathbf{p}_0(v_i) = 1$ for known disease- or drug-associated proteins and $\mathbf{p}_0(v_i) = 0$ for all other proteins. At each time point k, the amount of information at each node v_i depends on the sum of the information at the neighboring nodes $\mathcal{N}(v_i)$ at time $k-1$, with information from each neighboring node being weighted in proportion to the corresponding edge weight according to:

$$\mathbf{p}_k(v_i) = \sum_{v_j \in \mathcal{N}(v_i)} \mathbf{p}_{k-1}(v_j)\ w(v_i, v_j)$$

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 gives $\mathbf{p}_k(v_i)$ – a ranking for each node, which is a measure of how close all other nodes are to the initial nodes set to 1.

Using matrix notation, network propagation has the form of:

$$\mathbf{p}_k = \mathbf{W}\ \mathbf{p}_{k-1}$$

where $\mathbf{W}$ is a normalized version of the adjacency matrix of the network of interest. Repeated iterations give $\mathbf{p}_k = \mathbf{W}^k\mathbf{p}_0$. If $\mathbf{W}$ is a stochastic matrix (its columns sum to 1), the process is equivalent to a random walk (RW) on the network. If k is small, the diffused distribution is close to the initial distribution, but when k is large, it reflects the network topology. To prevent the over-diffusion of information, RW with restarts (RWR) include a parameter called *restart probability*, which controls how much information is propagated on each step. Iterative updates using RWR have the form

$$\mathbf{p}_k = \alpha\ \mathbf{p}_0 + (1-\alpha)\ \mathbf{W}\ \mathbf{p}_{k-1}$$

where $\alpha \in [0, 1]$ is the restart probability. When the network is connected, and the eigenvalues of $\mathbf{W}$ are at most 1 in absolute value, the process can be shown to converge to a steady-state distribution[114]:

$$\mathbf{p} = \alpha(\mathbf{I} - (1-\alpha)\mathbf{W})^{-1}\mathbf{p}_0$$

with $\mathbf{W} = \mathbf{A}\mathbf{D}^{-1}$.

Using network propagation has yielded state-of-the-art results in biological studies[114]. However, traditional network propagation algorithms fail to learn a propagation that is specific to the task, and require a grid search to optimize the parameter controlling how the propagation spreads (for example, the restart probability in RWR). Recent methods, comprised under the term of deep learning on graphs, have defined differentiable and learnable propagation or diffusion rules on graphs, which have been successfully applied in numerous biomedical applications[27]. The following section will present a review of representation learning approaches on graphs that aim at learning node and (sub)graph representations that can be used in downstream machine learning tasks. We introduce traditional approaches,

and recent approaches using RW and graph convolutional operators to compute node and (sub)graphs embeddings.

2.4 Representation learning on graphs

Hamilton et al.[115] provided an excellent review of representation learning algorithms on graphs and the remainder of this subsection is heavily based on Hamilton et al.'s work. The primary objective of machine learning on graphs is to incorporate the graph structure into the machine learning model. Traditional approaches accomplish this by extracting hand-engineered features from the graph. However, recent approaches seek to directly learn representations that encode graph structure and can adapt during the learning process. This is achieved through a mapping that embeds nodes, or entire (sub)graphs, as points in a low-dimensional vector space ($\mathbb{R}^d$). The learned space reflects the geometric relationships of the original graph structure[115]. The use of these learned embeddings, which capture the structural information of the graph, for downstream machine learning tasks, implicitly includes the structural information of the graph[115].

Representation learning and traditional approaches differ in how they treat the task of capturing structural information about the graph. This task has traditionally been treated as a preprocessing step, using hand-engineered statistics to extract structural information. Representation learning approaches treat this as a machine learning task using a data-driven approach to learn embeddings that encode graph structure[115]. Hamilton et al. introduced a theoretical framework to unify all representation learning approaches which we will briefly introduce here. Let $G = (\mathcal{V}, \mathcal{E})$ be an undirected graph associated with a binary adjacency matrix $\mathbf{A}$, and real-valued node attributes $\mathbf{X} \in \mathbb{R}^{mx|\mathcal{V}|}$. The goal of graph representation learning is to use $\mathbf{X}$ and $\mathbf{A}$ to map each node, or a (sub)graph, to a vector $\mathbf{z} \in \mathbb{R}^d$ where $d << |\mathcal{V}|$.

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

$$ENC : \mathcal{V} \rightarrow \mathbb{R}^d$$

$$DEC : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^+$$

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

$$DEC(ENC(v_i), ENC(v_j)) = DEC(\mathbf{z}_i, \mathbf{z}_j) \approx s_G(v_i, v_j)$$

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 approaches optimize the task by minimizing an empirical loss, L, over a set of training node pairs, D:

$$L = \sum_{(v_i, v_j) \in D} l(DEC(\mathbf{z}_i, \mathbf{z}_j), s_G(v_i, v_j))$$

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.

Under this theoretical framework, different node embedding approaches differ in their definition of (i) a pairwise proximity function s_G, (ii) an encoder function ENC, (iii) a decoder function DEC, and (iv) a loss function l.

In **direct encoding approaches**, the embedding matrix is optimized directly, making the encoding function a simple embedding lookup: $ENC(v_i) = \mathbf{Z}\mathbf{v}_i$ where $\mathbf{Z}$ is a matrix with the embeddings, and $\mathbf{v}_i$ is a one-hot indicator vector indicating which column of $\mathbf{Z}$ corresponds to each node. Random-Walk approaches are included in this category given that they optimize node embeddings so that nodes that tend to co-occur on short RWs over the graph have similar embeddings. DeepWalk[116] and node2vec[117] are two well-known RW-based algorithms in this category. In spite of their expressiveness, these methods have three drawbacks: no parameters are shared between nodes in the encoder, which can be statistically inefficient, since parameter sharing can act as a form of regularization, and they are computationally inefficient since the number of parameters grows as $O(|\mathcal{V}|)$. Additionally, they fail to leverage node attributes and are transductive, that is, they can only generate embeddings for seen nodes.

Generalized encoder-decoder architectures use more complex encoders, that effectively leverage graph structure and node attributes to overcome the drawbacks of direct-encoding methods. This category includes neighborhood aggregation and convolutional encoders. They compute node embeddings by performing aggregation of embeddings from neighboring nodes in the graph. In this framework, node embeddings are initialized to be the input node attributes. At each iteration, the encoder aggregates information of neighbors for each node. New node embeddings are equal to the aggregated neighborhood combined with the node's previous embeddings from the last iteration.

Embeddings are fed through a multi-layer neural network either before or after aggregation. This iterative process repeats as many times as there are layers in the encoder. With each added layer, node embeddings aggregate information from an additional hop in the graph; that is, a 1-layer encoder aggregates information from the one-hop neighborhood, a 2-layer encoder from the two-hop neighborhood, and so on. The dimensionality of the embeddings remains constrained as the process iterates because the encoder is forced to compress all the neighborhood information into a low-dimensional vector. After K iterations, the process terminates, and the final embedding vectors are output as the node representations[115].

A number of established approaches in deep learning on graphs follow this basic procedure, including graph convolutional networks (GCN)[118], the GraphSAGE algorithm[119], and the ChebNet algorithm[120]. The trainable parameters in the encoder specify how to aggregate information from a node's local neighborhood and are shared across nodes. This parameter sharing increases efficiency (i.e., the parameter dimensions are independent of the size of the graph), provides regularization, and allows this approach to be used to generate embeddings for nodes that were not observed during training[115].

2.5 Causal inference

Structural Causal Model (SCM). An SCM $\mathtt{M}$ is a 4-tuple $< \mathbf{E}, \mathbf{V}, \mathcal{F}, P(\mathbf{E}) >$ where $\mathbf{E}$ is a set of exogenous variables determined by factors outside the model; $\mathbf{V}$ is a set $\{\mathsf{V}_1, \mathsf{V}_2, \dots, \mathsf{V}_n\}$ of endogenous variables of interest that are determined by other variables in the model, that is, in $\mathbf{E} \cup \mathbf{V}$; $\mathcal{F}$ is a set of functions $\{f_{\mathsf{V}_1}, f_{\mathsf{V}_2}, \dots, f_{\mathsf{V}_n}\}$ such that each f_i is a mapping from $\mathbf{E}_{\mathsf{V}_i} \cup \mathbf{Pa}_{\mathsf{V}_i}$ to V_i, where $\mathbf{E}_{\mathsf{V}_i} \subseteq \mathbf{E}$, $\mathbf{Pa}_{\mathsf{V}_i} \subseteq \mathbf{V} \backslash \mathsf{V}_i$, and the entire set $\mathcal{F}$ forms a mapping from

$\mathbf{E}$ to $\mathbf{V}$. That is, for each $i = 1, \ldots, n$, each $f_i \in \mathcal{F}$ is such that $\mathsf{v}_i \leftarrow f_{\mathsf{V}_i}(\mathbf{pa}_{\mathsf{V}_i}, \mathbf{e}_{\mathsf{V}_i})$; and $P(\mathbf{E})$ is a probability function defined over the domain of $\mathbf{E}$.

The SCM $\mathtt{M}$ is associated with causal graph G where every $\mathsf{V}_i \in \mathbf{V}$ is a node and there is a directed edge $(\mathsf{V}_j \to \mathsf{V}_i)$ for every $\mathsf{V}_i \in \mathbf{V}$ and $\mathsf{V}_j \in \mathbf{PA}_{\mathsf{V}_i}$. The exogenous $\mathbf{E}_{\mathsf{V}_i}$'s are assumed independent (Markovianity holds). We consider recursive SCMs (implies directed acyclic graphs).

Layers 1, 2 valuations. An SCM $\mathtt{M}$ induces layer $L_2(\mathtt{M})$, a set of distributions over $\mathbf{V}$, one for each intervention $\mathbf{x}$. For each $\mathbf{Y} \subseteq \mathbf{V}$,

$$P^{\mathtt{M}}(\mathbf{y}_{\mathbf{x}}) = \sum_{\{\mathbf{u} | \mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{y}\}} P(\mathbf{u}) \tag{1}$$

where $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$ is the solution for $\mathbf{Y}$ after evaluating $\mathcal{F}_{\mathbf{x}} := \{f_{\mathsf{V}_i} : \mathsf{V}_i \in \mathbf{V} \backslash \mathbf{X}\} \cup \{f_X \leftarrow \mathsf{x} : \mathsf{X} \in \mathbf{X}\}$. The distribution $P(\mathbf{V})$ when $\mathbf{X}$ is empty is defined as layer $L_1(\mathtt{M})$.

An external intervention forcing a set of variables $\mathbf{X}$ to take values $\mathbf{x}$ is modeled by replacing the original mechanisms f_{X} for each $\mathsf{X} \in \mathbf{X}$ with its corresponding value in $\mathbf{x}$. This operation is represented formally by the do-operator $do(\mathbf{X} = \mathbf{x})$, and graphically as the *mutilation* procedure.

Under markovianity assumptions (no hidden confounders), interventional probabilities $P^G(\mathbf{Y}|do(\mathbf{X}))$ in SCM $\mathtt{M}$ with causal graph G can be obtained as conditional probabilities $P^{G'}(\mathbf{Y}|\mathbf{X})$ obtained from SCM $\mathtt{M}$ with mutilated graph G' after intervention.

3 Predicting food molecules with anticancer potential

3.1 Abstract

Recent efforts in the field of nutritional science have allowed the discovery of molecules within foods with disease-fighting properties based on the commonality of bioactive food molecules to FDA-approved drugs. In our pioneering work in this field, we used an unsupervised network propagation algorithm to learn propagation profiles of drugs representing their effects on a human protein-protein interaction network[121]. Learned drug propagation profiles were fed to a supervised machine learning algorithm to predict anticancer therapeutics. Then, a set of bioactive molecules within foods was fed into the model, which predicted molecules with cancer-beating potential. The employed methodology consisted of disjoint unsupervised learning of propagation profiles and classification tasks, which can result in sub-optimal learned drug propagation profiles with respect to the classification task and costly computational resources. To overcome these technical limitations, we propose a graph neural network (GNN)-based model to perform the information propagation and classification task end-to-end. Our proposed GNN model outperforms the baseline approach in the anticancer therapeutic prediction task by up to 67%. Using attribution methods, we show that our proposed model captures knowledge of biological pathways to predict anticancer molecules based on the molecules' effects on cancer-related protein communities. Finally, we provide a list of additional food molecules with potential cancer-beating properties as predicted by our model.

3.2 Background

As lifespan increases, the prevalence of chronic diseases such as cancer has spiked in recent years[122,123]. It is estimated 1 in 2 people born after 1960 in the UK will be diagnosed with cancer during their lifetime[124]. Recent studies have indicated that dietary and lifestyle changes can prevent 30-40% of all cancer cases[125,126]. Specifically, high consumption of whole grains, fruits, and vegetables have been associated with lower risks of several types of cancer[127–129]. However, molecules and mechanisms underlying such cancer-preventing effects remain relatively unexplored.

Almost 25% of all anticancer drugs approved between 1981 and 2019 were related to natural products[130,131]; therefore, identifying therapeutic molecules in food represents an invaluable opportunity to reduce healthcare costs and potentially enhance health outcomes for chronic diseases such as cancer. Experimental approaches to elucidate therapeutic properties of a single molecule can take months or even years. For example, 20 years passed between the discovery of anticancer activity of camptothecin and the elucidation of its mechanism of action. It took and an additional 15 years for the FDA to approve two captothecin analogues, topotecan and irinotecan, for treating various types of cancer[132]. With thousands of molecules within a single food, experimental techniques are a time-consuming and expensive strategy.

Recent advances in machine learning (ML) methods together with growing databases on drugs, diseases, and foods[88–94] offer a unique opportunity to identify therapeutic molecules in silico. Ni et al. applied a text mining strategy to millions of PubMed abstracts to extract molecular constituents of foods and their health-promoting or detrimental associations with disease phenotypes[93]. Subsequently, Jensen et al. identified food molecules affecting pharmacokinetics or pharmacodynamics of drugs, and observed that genes targeted by drugs used to treat cancer, cardiovascular, and neurological diseases are also highly targeted by molecules within foods[94]. Zheng et al. built a food-disease interaction network based on publicly available gene expression data and identified beneficial and harmful food-disease pairs[95]. Due to the small size of existing diet-induced gene expression datasets, this correlation-driven analysis was restricted to a very limited number of foods. Nevertheless, intriguing diet-disease associations have been identified through this approach.

Existing studies have extracted relationships between foods, drugs, and diseases based on overlap of targeted or differentially expressed genes[93–95]. However, in the context of cancer, similarities at the genome level can only be fully captured by affected pathways (i.e., gene communities where members interact to affect a specific function), not individual genes, due to the polygenetic nature of cancer and the heterogeneity of cancer genotypes[133–135]. To uncover associated protein modules or pathways in silico, a set of methods under the umbrella term of *network propagation* has been widely used in cancer research[114]. Network propagation identifies affected pathways based on the assumption that genes underlying similar phenotypes tend to interact with one another[114].

We have previously leveraged network propagation to identify molecules within foods with anticancer properties at large scale[136]. In our previous work, we identified as potential anticancer molecules those that affect similar protein communities as approved anticancer drugs. We propagated the effects of each drug and food molecule across a protein-protein interaction (PPI) network to learn diffusion profiles of each: a real-valued vector with one value per protein representing that represents how each protein is affected by a drug or food molecule. Each diffusion profile was computed through random walks with restarts that started at the proteins targeted by a drug or a food molecule. After algorithm convergence, the diffusion profile measured how often each node was visited, thus representing the effect of the drug or food molecule on that protein. The resulting drug diffusion profiles were fed to a supervised ML algorithm, a support-vector machine (SVM), trained to classify drugs into anticancer and non-anticancer classes. We then used the trained classifier to predict anticancer molecules within foods.

Our approach[121] consisted of disjoint network propagation and prediction of anticancer class, which required testing hundreds of hyperparameter combinations and significant computational resources. Additionally, the random walk algorithm propagated or "convolved", the information across the network using a predefined propagation rule that was decoupled from the prediction task. An alternative to these predefined propagation operators on graphs are graph neural networks (GNNs), developed as part of the trend of deep learning (DL) on graphs[26,119]. Instead of convolving the information across the graph using a pre-defined propagation rule, GNNs have learnable parameters allowing them to propagate or convolve information across the graph to maximize performance in a supervised task. GNNs have been leveraged for several tasks in bioinformatics incorporating PPI networks and have achieved state-of-the-art performance in many cases[137–139], including in drug repositioning (see *Related work* section). Here, we leverage the capabilities of GNNs to overcome the limitations of our previous work[121]. We propose an end-to-end graph neural model that operates directly on graphs to (i) learn how to propagate information across the PPI network to capture effects of drugs and food molecules on proteins and (ii) identify anticancer molecules from the learned diffusion profiles.

The rest of this Chapter is organized as follows: subsection 3.3 introduces related work and

state of the art approaches for drug repositioning, subsection 3.4 states contributions and impact, subsections 3.6, 3.5, and 3.7 introduce the datasets used, our problem definition, and our approach, respectively, subsection 3.8 contains results, and subsection 3.9 includes a summary and discussion of contents in this Chapter.

3.3 Related work

Computational drug repositioning. Drug repositioning consists of finding new indication for existing drugs. In the early years of computational drug repositioning, ML methods relied on manually constructed descriptors of molecules and diseases[140]. In recent years, DL approaches have become the go-to for drug repositioning due to their ability to recover complex patterns from high-dimensional data without need for manual selection and tuning of features[140,141]. A detailed review of DL approaches for drug repositioning can be found in Pan et al.'s work[140]. DL approaches to drug repositioning typically use sequence-based or graph-based data representations[140]. **Sequence-based representations** can partly overcome the limitations of protein/target structural data availability[140]. For molecular compounds, traditional representations include the Simplified Molecular Input Line Entry System (SMILES[142]), and chemical fingerprints[143]. More recently, Mol2vec[144] was proposed to compute embeddings of atoms using Word2Vec[145], a natural language processing algorithm, by considering molecular substructures as words and compounds as sentences. On the other hand, protein sequences are composed of 20 standard aminoacid. Several protein sequence-based representations are used in drug repositioning studies, including a simple one-hot encoding of each aminoacid, or a two-dimensional distance map which calculates the distance between all possible aminoacid residue pairs in a three-dimensional protein structure[140]. **Network-based representations** have become more widespread in recent years due to the vast amounts of large-scale heterogeneous data generated. For both, molecular compounds and proteins, their topological structure can be represented as a graph or network, with nodes representing atoms/aminoacids and edges or links representing chemical associations between them[140].

Drug repositioning efforts aim at predicting unknown drug-target or drug-disease interactions, therefore, they can be classified into "target-centered" or "disease-centered" meth-

ods, respectively[140]. In the drug-target interaction prediction task, the goal was to encode the chemical structure of drugs to screen targeted proteins which would provide the polypharmacological interpretation[140]. However, a single target cannot fully describe the characteristics of the disease, therefore, identifying drug-disease associations has become essential for drug repositioning. Current disease-centered methods can be categorized into similarity-based and network-based approaches, according to Pan et al.[140]. **Similarity-based methods** leverage drug-drug, disease-disease, and drug-disease similarities to predict drug-disease associations. On the other hand, **Network-based methods** utilize network-based data representations to predict drug-disease associations. For example, Liu et al. built drug-drug and disease-disease graphs, with edges between known drug-disease associations and used a random walk-based algorithm to predict score of drug-disease associations[111]. For each queried drug-disease association (known or unknown), the method consisted on computing 2 independent random walks on the heterogeneous graph, one starting from the initial drug, and another starting from the initial disease. The final association score being the mean of the obtained scores after the random walk on the queried disease and drug, respectively. An important limitation of this method was its inability to generalize to drugs that do not have an association to at least one disease, and conversely, to diseases that do not have already a drug indicated. More recently, GNN-based methods have taken center stage in network-based drug repositioning. A prominent example is the work by Stokes et al.[146] who utilized a GNN-based model to compute molecule representations from 2D molecular structures to find new antibiotics. The authors identified Halicin as a new wide-spectrum antibiotic agent and validated its antibiotic effects experimentally.

Given the success of GNN models for drug repositioning[146–148], we leverage GNNs to introduce a method for the task of predicting food molecules with anticancer properties which can be seen as a repositioning task, i.e., finding potential anticancer indication of known food molecules. Given the limited data availability for food molecules, we leverage only protein targets, and use our GNN-based model to propagate the information of protein targets across a PPI network for the prediction task.

3.4 Contributions and impact

Contributions:

- We propose an end-to-end GNN-based model for the anticancer molecule prediction task and explore graph pooling based on biological pathways to integrate prior biological knowledge into the neural architecture.
- We use attribution methods to uncover proteins that causally influence model predictions.

Impact: Our end-to-end model to predict bioactive molecules within foods with anticancer properties has high practical value for the discovery of new potentially therapeutic molecules and prioritization for experimental testing. Our model outperforms the existing baseline and predicts additional food molecules with potential anticancer properties as those reported in our initial study[121].

This section is based on the following publications:

1. Veselkov, K., **Gonzalez, G.**, Aljifri, S., Galea, D., Mirnezami, R., Youssef, J., Bronstein, M., Laponogov, I. (2019). HyperFoods: Machine intelligent mapping of cancer-beating molecules in foods. Scientific Reports, 9(1), 9237. https://doi.org/10.1038/s41598-019-45349-y
2. **Gonzalez, G.***, Gong, S.*, Laponogov, I., Veselkov, K., Bronstein, M. (2019). Graph Attentional Autoencoder for Anticancer Hyperfood Prediction. 2019 NeurIPS Graph Representation Learning Workshop. *equal contribution.
3. **Gonzalez, G.**, Gong, S., Laponogov, I., Bronstein, M., Veselkov, K. (2021). Predicting anticancer hyperfoods with graph convolutional networks. Human Genomics, 15(1), 33. https://doi.org/10.1186/s40246-021-00333-4

3.5 Problem definition

Anticancer food molecule prediction. Given a set of bioactive molecules within foods, each represented by a binary vector $\mathbf{x} \in \{0,1\}^{|\mathcal{V}|}$ encoding its protein targets, our goal

is to predict a single score per molecule representing its anticancer therapeutic potential. Intuitively, a food molecule should be predicted to have high anticancer likelihood if its effects on the human interactome are similar to those of approved anticancer drugs. Hence, we pretrain a classifier to predict anticancer drugs and use the trained model to predict anticancer therapeutic likelihood of food molecules.

Anticancer drug molecule prediction. Given a set of drugs, each represented by a binary vector $\mathbf{x} \in \{0,1\}^{|\mathcal{V}|}$ encoding its protein targets, our goal is to predict a label $y \in \{0,1\}$ for each drug representing its known anticancer therapeutic potential ($y = 1$ if it is a known anticancer drug and $y = 0$ if the drug is not known to be associated with cancer, either for primary, secondary, or tertiary use).

3.6 Datasets

Protein-protein interactions (PPIs). For this work, we compiled a human PPI network of 177,848 physical interactions between 15,135 proteins from the STRING v10.5 database[100]. We consider only the subset of experimentally validated PPIs, removing isolated proteins (i.e., proteins without interactions to any other protein) and keeping the largest connected subgraph. The reliability of the PPI network from the STRING database is generally considered high within the scientific community. STRING integrates PPI data from several sources, including experimental results, computational prediction methods, and public text collections, assigning a confidence score between 0 and 999 to each interaction based on the evidence supporting it. Here, only experimentally validated PPIs were considered to ensure higher reliability. For experimentally validated interactions, we did not apply a confidence cutoff to prioritize coverage. STRING does not explicitly differentiate between positive or negative interactions but rather provides a framework to understand the functional relationships between proteins, therefore, the extracted interactions are undirected, and are all considered equal.

Drug protein targets. We map clinically approved drugs extracted from DrugBank v5[88] and DrugCentral[89] (Feb 2019) to their protein targets using STITCH[90] (Jan 2019), a database containing molecule-target connections. We filter drug-target pairs to only in-

clude proteins that are included in the PPI network. STITCH includes a confidence score ranging from 150 to 999 for each interaction based on the evidence supporting it. Here, we selected a confidence threshold of 700 to ensure the reliability of the drug-target associations, as advised in the STITCH documentation. By applying this stringent criterion, we aimed to include only the most reliable interactions in the analysis, significantly enhancing the data's accuracy. It is important to note, however, that while this approach prioritizes high-confidence data, it may exclude potentially relevant lower-confidence interactions. In the development of the machine learning model, we focused on this high-confidence subset, ensuring that the input data was of substantial reliability and relevance.

Food molecule protein targets. We extract molecules in foods from FooDB[149] (N = 7,793. Nov 2018) and map them to their protein targets using STITCH[90] (Jan 2019), a database containing molecule-target connections. We filter food molecule-target pairs to only include proteins that are included in the PPI network. Here, we also utilize a confidence threshold of 700 to extract molecule-target interactions.

Drug classification labels. We use the DrugCentral[89] database (Feb 2019) to identify drugs designed for primary use against cancer. RepoDB[150] (Nov 2017) was used to identify drugs that have been successfully repositioned for anti-cancer purposes (secondary or tertiary use). For our algorithm training, the "positive" (anticancer) class comprises drugs designed and tested specifically for cancer treatment (n = 209), and the "negative" class comprises drugs with no known association with cancer (n = 1,839). Drugs that have been repositioned for secondary/tertiary use in cancer treatment have been excluded to build clean positive and negative sets. Drugs extracted from different databases were matched using InChI keys[151].

Biological pathways. We extract pathways from the Kyoto Encyclopedia of Genes (KEGG[103], downloaded version 7.1 from the Molecular Signatures Database[152]). We build a pathway assignation matrix $\mathbf{P} \in \{0,1\}^{npathways \times 15{,}135}$, where $npathways = 186$ and $P_{ij} = 1$ if gene j participates in pathway i and 0 otherwise. Of the $15,135$ genes in the PPI network, $4,590$ participate in at least one pathway. While utilizing KEGG pathways, we acknowledge certain limitations. Firstly, KEGG pathways may not encompass all known gene-pathway associations, potentially leading to incomplete representation. Secondly, the static nature

of pathway data might not fully reflect dynamic biological processes. In our algorithms, these uncertainties are implicitly considered through the binary pathway assignation matrix, $\mathbf{P}$. This matrix represents known gene-pathway associations without making assumptions about unrepresented or dynamic interactions. By focusing on established associations, the algorithm prioritizes reliability, although it may not capture the entire complexity of gene-pathway dynamics.

Data samples. The compiled dataset contains 2,048 drugs and 7,793 food molecules. Each drug (food molecule) is represented by a graph $G = (\mathcal{V}, \mathcal{E})$ of PPIs, with $|\mathcal{V}| = 15,135$ nodes and $|\mathcal{E}| = 177,848$ edges, and a feature vector $\mathbf{x} \in \{0,1\}^{|\mathcal{V}|}$ (i.e., one binary feature per node equal to 1 if the protein is a drug [food molecule] target, and 0 otherwise).

3.7 Methods

Classification model. We represent molecules by the proteins they target; however, proteins do not act in isolation. There are complex relations between them, and perturbations in one protein can have downstream effects on interacting proteins. We leverage this contextual information in the form of a known PPI graph $G = (\mathcal{V}, \mathcal{E})$ that encodes physical interactions between proteins. For each molecule, we propagate information from the starting protein targets across the PPI network to generate diffusion profiles capturing affected protein communities. We rely on GNNs to propagate the information across the PPI in a learnable fashion, allowing for an end-to-end model for the anticancer molecule prediction task (Figure 3.1A).

To predict an anticancer label for each drug, we use a multi-layer GNN operating on graph G. For each drug, our model takes as input the graph G and its feature vector $\mathbf{x}$ and outputs a classification label indicating anticancer/non-anticancer category. To build our model, we stack K GNN layers to propagate information across graph G (Figure 3.1B). In practice, we use three variants of GNNs: GCN[118], ChebNet[120], and GraphSAGE[119]. For each of these, the *l-th* GNN layer computes an updated representation $\mathbf{H}^{(l)} \in \mathbb{R}^{|\mathcal{V}| \times d_l}$ of proteins in $\mathcal{V}$ where d_l is the dimensionality of representations in layer l. Inspired by the Jumping Knowledge framework[153], the final representation of drug i is obtained by concatenating representations

generated by all layers:

$$\mathbf{H} = \left[\mathbf{H}^{(1)}|\mathbf{H}^{(2)}|\ ...\ |\mathbf{H}^{(K)}\right] \in \mathbb{R}^{|V|\times(d_1+d_2+...+d_K)}$$

We aggregate protein features in $\mathbf{H}$ according to biological pathways (Figure 3.1C). Formally, given the drug embedding $\mathbf{H}$, pathway pooling can be formulated as follows:

$$\mathbf{Z} = \mathbf{PH}$$

where $\mathbf{P} \in \mathbb{R}^{npathways\times|V|}$ is an assignation matrix with $P_{ij} = 1$ if gene j is involved in pathway i and 0 otherwise. We then feed the embedding $\mathbf{Z}$ to a fully connected layer to reduce the dimensionality of drug representations by aggregating node features to a single dimension:

$$\mathbf{z} = \mathbf{ZW}_{fcp}$$

where $\mathbf{z} \in \mathbb{R}^{npathways\times1}$ is the transformed representation of drug i and $\mathbf{W}_{fcp} \in \mathbb{R}^{(d_1+d_2+...+d_K)\times1}$ is a learnable matrix. The transformed representation $\mathbf{z}$ of drug i is then fed to a two-layer multi-layer perceptron (MLP) to output a probability distribution for the classification task:

$$\mathbf{p}_i = Softmax(ReLU(\mathbf{z}^T\mathbf{W}_{l1} + \mathbf{b}_1)\mathbf{W}_{l2} + \mathbf{b}_2)$$

where $\mathbf{W}_{l1}$, $\mathbf{b}_1$, $\mathbf{W}_{l2}$, and $\mathbf{b}_2$ are learnable weight matrices.

Model training During training, we optimize hyperparameters using cross-entropy loss:

$$-\sum_{i=1}^{N}\sum_{c=0}^{1} y_{i,c}log(p_{i,c})$$

where $y_{i,c}$ is a binary indicator if class label c is the correct classification for drug i, $p_{i,c}$ is the predicted probability that drug i is of class c, and N is the number of drugs in the training set. We train the model in an end-to-end fashion and optimize all model parameters using backpropagation. We train the model for a maximum of 100 epochs (training iterations) with the Adam optimizer and early stopping with a window size of 20: the training stops if the validation loss does not decrease by at least $1e^{-4}$ for 20 consecutive epochs. We implement our model using PyTorch[154] and the Torch Geometric Library[155]. We perform hyperparameter search for the learning rate, L2 regularization on the weights of the neural

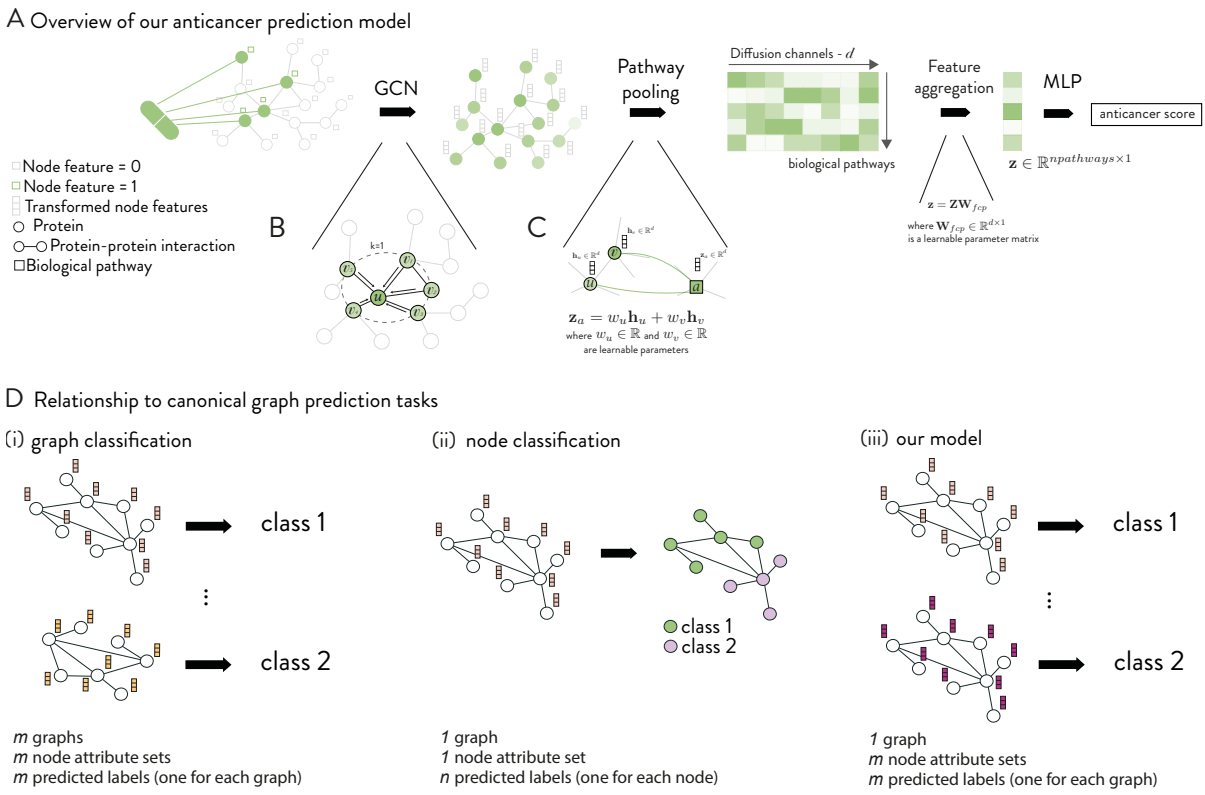


Figure 3.1: (A) Overview of our approach. Given a drug represented by its protein targets, our goal is to predict an anticancer score representing the drug's known anticancer therapeutic potential. We propagate protein target information using message passing operators across a protein-protein interaction (PPI) network to generate diffusion profiles capturing affected protein communities. Learned diffusion profiles are then fed to a multi-layer perceptron (MLP) to predict an anticancer score. **(B) Message passing scheme.** Graph neural networks (GNNs) follow a message-passing scheme to propagate information across a network. In a message passing scheme, information is propagated in the form of one-hop neighbor messages that are aggregated and transformed using a nonlinear function. **(C) Pathway pooling.** We pool protein-level information into biological pathways using learnable parameters. **(D) Relationship to canonical graph prediction tasks.** Our task resembles canonical graph and node prediction tasks, but our method has distinct properties: (i) in graph classification, a set of graphs is paired with a set of node attributes, and the task is to predict a label for each graph; (ii) in node classification, a single graph is paired with a single set of node attributes, and the task is to predict a label for each node; (iii) in our formulation, a single graph is paired with a set of node attributes, and the task is to predict a label for each graph-node attribute pair (constant graph structure, varying node attributes).

network, and number of dropout layers in the MLP. In addition to training the proposed model, we optimize an alternative architecture where pathway pooling is not used.

Attributing predictions to input features for interpretability. We assess whether the trained model has learned the top biological pathways (i.e., protein communities) responsible for the anticancer properties of drugs. Given a trained model and an input, an attribution method assigns scores to each input feature that reflect the contribution of that feature to the model prediction. Inspecting the attribution scores reveals which features, or in our case, proteins, were most relevant to the model's decision. Formally, suppose a function $F : \mathbb{R}^n \rightarrow$

$[0, 1]$ represents a deep neural network. The attribution at input $\mathbf{x} = (x_1, ..., x_n) \in \mathbb{R}^n$ is a vector $A_{F(\mathbf{x})} = (a_1, ..., a_n) \in \mathbb{R}^n$, where a_j is the contribution of x_j to the prediction $F(\mathbf{x})$.

In our case, given drug i and its feature vector $\mathbf{x}$ as input, $F(\mathbf{x})$ denotes the probability that the drug belongs to the anticancer category. To compute attributions to individual proteins, we use the integrated gradients method[156]. For extended definitions and comparisons with other attribution methods, we refer the reader to[156]. Integrated gradients provides attributions relative to a provided baseline input. Here, we use an input where all drug targets are set to zero. Integrated gradients are defined as the path integral of the gradients along the straight-line path from the baseline to the input. The integrated gradient along the j^{th} dimension for an input $\mathbf{x}$ and baseline $\mathbf{x}'$ is defined as follows:

$$a_j(\mathbf{x}) ::== (x_j - x'_j) \times \int_{\alpha=0}^{1} \frac{\delta F(\mathbf{x}' + \alpha \times (\mathbf{x} - \mathbf{x}'))}{\delta \mathbf{x}_j} \delta\alpha$$

Attribution recall score. We would like the attribution scores to recover protein targets whose corresponding genes are present in cancer-related pathways. This would mean that our model makes decisions based on feature values in genes relevant to the anticancer properties of drugs. To evaluate how well the model can recover cancer-related genes, we introduce a metric called *attribution recall score* that measures how well the attribution scores recover cancer-related genes (Figure 5.2A). We compute attribution scores for samples classified as anticancer in the test set across all splits and average them across all samples. With the average attribution scores, we use the PreRanked module of Gene Set Enrichment Analysis (GSEA)[152] to obtain over-represented pathways (from the Kyoto Encyclopedia of Genes [KEGG][103]) in the most positively attributed genes. This measures whether genes in KEGG pathways are over-represented in the most positively scored genes compared to what is expected by chance. We then filter over-represented pathways using a false discovery rate (FDR) of 25% (as advised in GSEA documentation[152]) and measure the recovery of anticancer pathways as the number of over-represented pathways that are related to cancer divided by the total number of anticancer pathways in KEGG.

Experimental setup. We perform five-fold cross-validation to assess model performance. In each split, 20% of the data is kept as the test set; from the remaining 80%, 10% is used as a validation set to perform early stopping. All splits are generated by stratifying samples

with respect to labels. The dataset is highly unbalanced with respect to the target label (only 10.2% of drugs are anticancer). To balance the positive/negative classes, we rescaled the contribution of each class to the loss function so it is inversely proportional to class frequencies of each class during training. Models were trained on NVIDIA Tesla V100 and GEForce RTX 2080 GPUs.

Here, we utilized five-fold cross-validation to balance computational efficiency with robust model performance assessment. This approach was preferred over ten-fold or leave-one-out cross-validation due to the high computational demands of deep learning models. Five-fold cross-validation provides a stable performance estimate with lower variance and ensures adequate representation of the minority class in our unbalanced dataset. The choice of cross-validation technique reflects a strategic compromise between the need for a reliable performance evaluation and managing the computational resources effectively.

We use our models that output, for each drug a probability of it being an approved anticancer drug. We evaluate the performance of our models against our previous work[121]. In previous work, we represented drug protein interactions as binary signals on the PPI network and applied a random walk algorithm to learn diffusion profiles of drugs. We fed the learned diffusion profiles to an SVM for the binary classification task of anticancer/non-anticancer drugs. In the present work, we use various metrics for the comparative analysis of performance. Balanced accuracy is the average of recall obtained on each class; F1 is the weighted average of precision and recall for the positive (anticancer) class; and AUPR is the area under the precision-recall curve, which represents the average precision across all recall values.

Hyperparameter settings for every method are determined using a validation set with a grid search over candidate hyperparameter values. For the baseline approach, the grid search for the restart probability is $[0.001, 0.01, 0.1, 0.2, ..., 0.9]$. For neural models, we tested learning rates of $\{5.10^{-4}, 5.10^{-3}\}$, L2 regularization of $\{1.10^{-5}, 1.10^{-4}, 5.10^{-4}\}$, number of convolutional layers of $\{1, 2, 3\}$, and number of hops for the ChebNet of $\{2, 4, 6\}$. All convolutional layers in our model have $d = 8$ hidden units. The first prediction layer has 32 hidden units, and the final prediction layer has two output units. We use a mini-batch size of 16.

Relationship to canonical graph prediction tasks. Because we rely on GNNs for

our anticancer molecule prediction task, we comment on the connection between our task and canonical graph prediction tasks. Under a graph representation learning lens, we can formulate our problem as follows: Given a graph $G = (\mathcal{V}, \mathcal{E})$, paired sets of node attributes $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m\}$, and graph labels $\mathcal{Y} = \{y_1, y_2, \dots, y_m\}$, where each $y = \{0, 1\}$, we aim to train a graph neural architecture that predicts the correct graph label y_i given node attributes $\mathbf{X}_i$ and the graph G. There are some major differences between our problem formulation and canonical graph prediction tasks, namely, graph and node classification (summarized in Figure 3.1D).

In node classification, a single graph G is paired with node attributes $\mathbf{X}$, and the task is to predict node labels $\mathbf{Y}$. Our formulation differs in that we have m paired sets of node attributes $\mathcal{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_m\}$ (instead of a single set) and graph labels $\mathcal{Y} = \{y_1, \dots, y_m\}$ (instead of node labels), but they are similar in that there is a single graph in which GNNs operate. In graph classification, a set of graphs $\mathcal{G} = \{G_1, \dots, G_m\}$ is paired with a set of node attributes $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m\}$, and the task is to predict a label for each graph $\mathbf{Y} = \{y_1, \dots, y_m\}$. Here, graphs have varying structure, and both the topological information and node attributes are predictive of graph labels. In our formulation, a single graph is used in combination with each of the node attributes $\mathbf{X}_i$, and the goal is to predict a label for each graph. Here, the topological information contained in graph G is not discriminate of samples per se, but rather, acts as a support to obtain enhanced sample representations to facilitate the prediction task.

3.8 Results

Learnable network diffusion outperforms random walks. By learning how to propagate information across the PPI network, our approach predicts which drugs treat cancer up to 67% more accurately than random-walk-based approaches (ChebNet vs. RWR: F1 67.99 vs. 51.84, +31.15%; balanced accuracy 83.46 vs. 81.13, +2.87%; precision anticancer 65.46 vs. 38.98, +67.93%) (Table 3.1). Surprisingly, pathway pooling leads to a significant decrease in performance for all neural models. We initially hypothesized that decreasing the number of genes contributing to the prediction would have a denoising effect. However, results indicate that such a hard regularization prevents the model from learning from other

Method	ACC	F1	AUPR	Precision ac	Recall ac
RWR	81.13 ± 3.79	51.84 ± 5.79	67.43 ± 8.14	38.98 ±5.38	**75.08±6.92**
GCN	80.52± 3.33	63.95± 3.90	66.45± 5.82	63.33± 5.72	65.51± 7.42
GraphSAGE	78.27± 6.11	59.93± 6.53	64.42± 9.96	61.04± 5.72	61.15± 13.48
ChebNet	**83.46± 2.52**	**67.99± 2.87**	**73.91± 3.49**	**65.46± 4.53**	71.27± 5.58
GCN-P	78.70± 5.36	57.43± 7.61	60.03± 8.48	52.77± 7.69	64.03± 11.05
GraphSAGE-P	77.09± 4.18	54.07± 4.88	60.55± 9.51	48.87± 4.06	61.64± 9.65
ChebNet-P	76.10± 2.67	55.71± 4.46	59.68± 9.53	53.72± 4.07	57.86± 4.96

Table 3.1: Summary of results (%) for anticancer drug prediction. By learning how to propagate information across the interactome, our approach predicts which drugs treat cancer up to 67% more effectively than random-walk-based approaches. Reported values are averaged across five-fold cross-validation (see the *Methods* section for more details); the top results are highlighted in bold. *-P represents neural models with pathway pooling. ACC=balanced accuracy, F1=harmonic mean of precision and recall, AUPR=area under the precision-recall curve, ac = anticancer.

equally relevant but experimentally understudied genes.

Our model identifies cancer-related proteins. Using attribution methods, we identify proteins that our best-performing model (ChebNet) attends to when identifying a drug as anticancer. Specifically, we compute protein attribution scores for each sample classified as anticancer in the test set and average the protein scores across samples. This results in a real-valued score per protein representing how important features of that protein are when predicting a drug as anticancer. We rank proteins based on this score and use GSEA[152] to obtain over-represented pathways in the top-scoring proteins. We measure the attribution recall score as the number of over-represented cancer-related pathways divided by the total number of cancer-related pathways in KEGG[103] (see the subsection "Attribution recall score" in the *Methods* section for more details; Figure 5.2A). The attribution recall score for the most positively attributed proteins is 85.29%, which indicates that the proteins most influential for the prediction of anticancer molecules are preferentially found in cancer-related pathways, adding to the biological plausibility of the model.

Furthermore, case studies validate the diffusion profiles that our model computes. We investigate two case studies of the top drugs correctly and incorrectly classified as anticancer that have experimentally validated pathways in the literature (Neratinib and Calcitriol, respectively; these drugs classified as anticancer with highest probability). For each of these drugs, represented by a vector $\mathbf{x}$, we compute its representation $\mathbf{z}$ with our best-performing model and obtain the top 25 over-represented pathways using GSEA (see subsection 7.1).

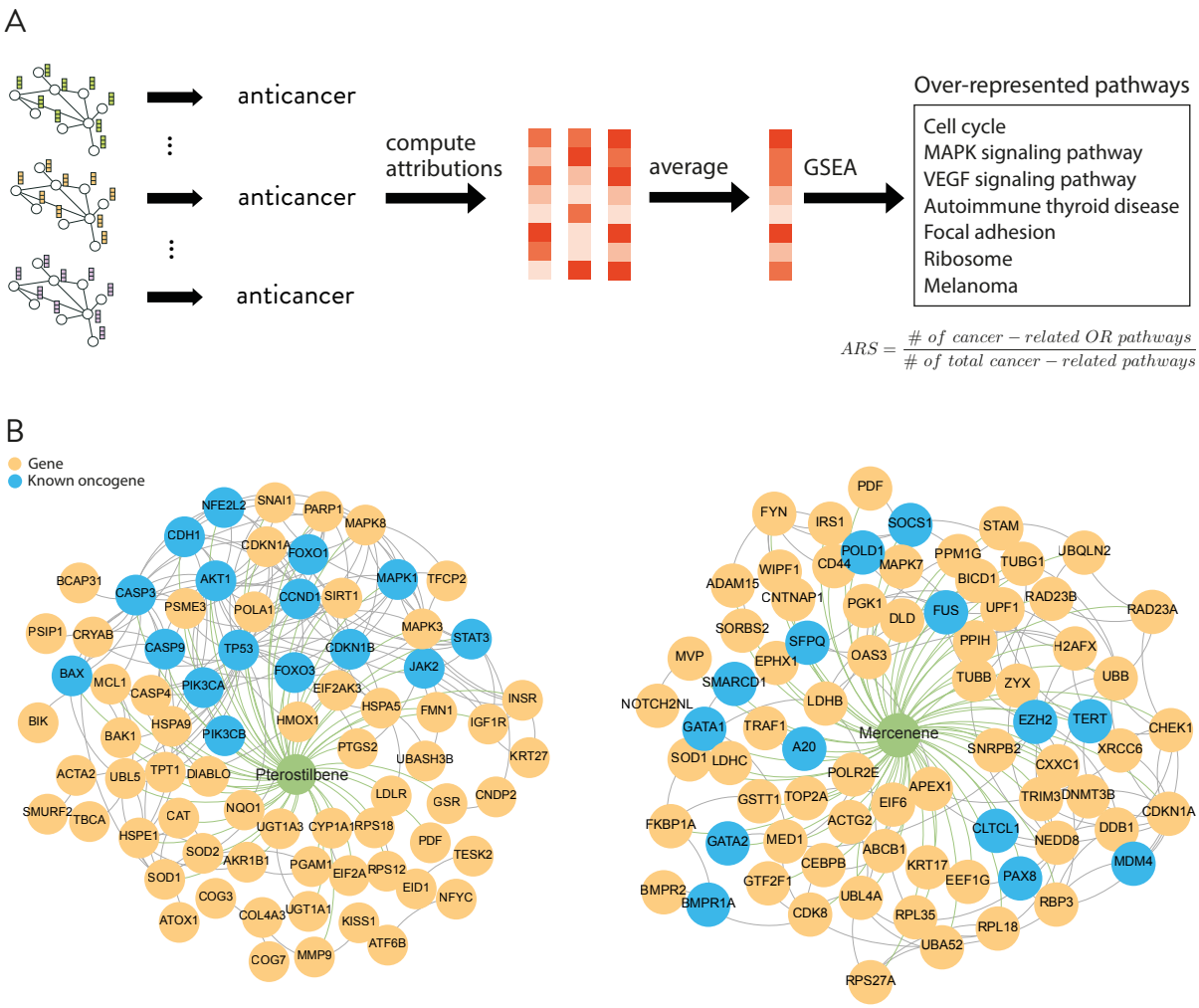


Figure 3.2: (A) Computation of attribution recall score. We compute protein attribution scores for each sample classified as anticancer in the test set and average protein attribution scores across samples. Using GSEA[152], we compute over-represented pathways from the Kyoto Encyclopedia of Genes (KEGG) in the top-attributed proteins. We measure the attribution recall score (ARS) as the number of cancer-related, over-represented pathways divided by the total number of cancer-related pathways in KEGG. **(B) Our model predicted that food molecules pterostilbene and mercenene have high anticancer likeness.** We obtained diffusion profiles for pterostilbene and mercenene and found that both molecules affect genes known to participate in cancer-related pathways, either directly or through downstream network effects. The detection of diffusion profiles indicating that pterostilbene and mercenene affect genes in cancer-related pathways suggests their potential role in modulating these pathways, which could be pivotal in understanding their anticancer properties. Further investigation into the specific genes and pathways affected, as well as the underlying mechanisms of action, would be essential to substantiate these findings and harness their therapeutic potential.

Neratinib is a protein kinase inhibitor used to treat breast cancer that over-expresses the HER2 receptor [88]. Existing findings indicate that Neratinib participates in a range of cancer-related pathways: EGFR tyrosine kinase inhibitor resistance, the ERBB signaling pathway, pathways in cancer, and breast cancer[103]. Two of these pathways were recovered as over-represented pathways using GSEA alongside various cancer-related pathways (see subsection

Our model predicts food molecules with potential anticancer properties. We compute anticancer likeness of food molecules using the five trained ChebNet models during the cross-validation procedure and average them to provide a more robust anticancer likeness measure. Of the top predictions (anticancer likeness >= 0.9), 20 molecules were found in addition to those reported in our previous work[121] (see Table 3.2). We obtained embeddings of these molecules using our best-performing model and over-represented pathways from the KEGG database using GSEA. Over-represented pathways of these molecules captured a wide range of cancer-related mechanisms and signaling pathways including the P53 signaling pathway; MAPK signaling pathway; ERBB signaling pathway; and those involved in apoptosis, cell growth, and cell proliferation (see subsection 7.2).

Of the 20 anticancer-predicted molecules, pterostilbene and mercenene have the most promise as cancer-preventing agents, as supported by the proteins they affect (Figure 5.2B) and substantial experimental evidence. Pterostilbene, found in grapes and blueberries, is chemically related to resveratrol, a well-studied molecule with antimicrobial, antioxidant, and anti-inflammatory activity which translate into chemopreventive effects[157]. Pterostilbene has demonstrated excellent pharmacological benefits for the prevention and treatment of various types of cancer at different stages of progression through apoptotic or nonapoptotic anticancer activities[158,159]. The antiproliferative properties of mercenene, a compound isolated from the common clam Mercenaria mercenaria, were documented as early as the second half of the 20th century[160,161]. In fact, mercenene or 1,3-Bis(2-chloroethyl)-1-nitrosourea, was introduced as the anticancer therapeutic Carmustine in 1962 and received FDA approval in 1977[162,163]. Carmustine is currently used in the treatment of various malignancies including brain tumors and multiple myeloma[88].

Table 3.2: Anticancer likeness of food molecules was computed using the best-performing neural model, and 20 molecules were predicted as anticancer molecules in addition to those reported in our previous work[121]. Description of molecules was extracted from FooDB[149].

ID	Name	Description
FDB001084	Pancreatin	Digestive enzyme. Used in replacement therapy. Used to prepare protein hydrolysates for pre- and postoperative diets.
FDB006967	Anthracene	Organic compounds containing a system of three linearly fused benzene rings. Anthracene can be found in sorrel. Anthracene is formally rated as an unfounded noncarcinogenic (IARC 3) potentially toxic compound.
FDB008856	2,2'-Bis(4-hydroxyphenyl) propane	Potential food contaminant due to its use in reusable polycarbonate food containers such as water carboys, baby bottles, and kitchen utensils.
FDB011663	Coumestrol	A natural organic compound in the class of phytochemicals known as coumestans. It has garnered research interest because of its estrogenic activity and its prevalence in some foods, such as soybeans and herbs like Pueraria mirifica. Coumestrol is a phytoestrogen, mimicking the biological activity of estrogens.
FDB011828	Genistein	A phenolic compound belonging to the isoflavonoid group. Isoflavonoids are found mainly in soybeans.
FDB012375	Pterostilbene	A member of the class of compounds known as stilbenes. Pterostilbene can be found in common grape and grape wine and is a stilbenoid chemically related to resveratrol.
FDB012974	Mercenene	Found in the common clam Mercenaria mercenaria and Mercenaria campechiensis
FDB014654	Myristicin	Natural organic compound present in the essential oil of nutmeg and to a lesser extent in other spices such as parsley and dill.
FDB016593	2,5-Dihydro-4,5-dimethyl-2-(1-methylpropyl) thiazole	Flavoring ingredient. Reported in hydrolyzed vegetable protein.
FDB020870	1-Methyl-6-phenyl-1H-imidazo[4,5-b]pyridin-2-amine	Food-related mutagen, reported to be the most abundant heterocyclic amine found in cooked meat and fish.
FDB022056	5a-Androstane-3a,17a-diol	Steroid compound.
FDB022182	Isourso-deoxycholic acid	Bile acid.
FDB022318	11alpha-Hydroxy-progesterone	Steroid hormone involved in the female menstrual cycle, pregnancy (supports gestation), and embryogenesis of humans and other species.
FDB023086	Dihydro-testosterone	Potent androgenic metabolite of testosterone.
FDB023772	Testosterone enanthate	Used in androgen substitution.
FDB024072	5b-Dihydro-testosterone	Intermediate in androgen and estrogen metabolism.
FDB028898	Methyl-arsonite	Found in the arsenate detoxification I pathway.
FDB030068	Platinum	Member of the class of compounds known as homogeneous transition metal compounds. Platinum can be found in food items such as white cabbage, sunburst squash (pattypan squash), potato, and broccoli.
FDB030278	17-α-hydroxy-pregnenolone	Belongs to gluco/mineralocorticoids, progestogins and derivatives class of compounds.
FDB030678	Androst-4-en-3,17-dione	Belongs to androgens and derivatives class of compounds.

3.9 Summary

Despite the documented benefits of fruits and vegetables against cancer in observational studies[127–129], the specific molecules and corresponding biological mechanisms responsible for the cancer-preventing properties of individual food ingredients remain largely unexplored. In our previous work, we introduced the first large-scale ML method to identify individual food molecules with potential cancer-beating properties and elucidate the biological mechanisms behind them[121]. To overcome limitations of our previous approach, here, we introduced a GNN-based model for the prediction of food molecules with anticancer properties. The model takes as input a graph structure and node features representing protein targets of a food molecule and outputs an anticancer likelihood score. The model outputs a high anticancer likelihood for a given food molecule if this molecule affects similar protein communities as those affected by anticancer drugs. Because it learns how to propagate information across the graph during training, our graph neural model outperforms the previous random walk baseline approach[121] by up to 67%. Moreover, using attribution methods, we demonstrate that our model identifies proteins and pathways relevant in cancer. Finally, we provided a list of additional food molecules with potential cancer-beating properties. A potential limitation of our work is the stochastic nature of DL models which can translate into unstable results. We envision this work going forward through incorporation of heterogeneous data sources such as gene expression (where available), and multi-scale knowledge of genes and associated biological functions.

A critical consideration in our exploration of food molecules with potential anticancer properties is the fundamental difference between preventive measures and treatments for existing cancer. Traditional cancer therapies, such as chemotherapy and radiotherapy, are often toxic and, when applied to healthy individuals, do not serve as preventive measures but rather pose a risk of inducing cancer due to their inherently damaging nature to cellular structures and DNA. This distinction is crucial when discussing food molecules that mimic the effects of anticancer therapies. While our model predicts food molecules with anticancer likelihood based on similarities to known anticancer drugs, it is essential to acknowledge that molecules effective in treating cancer may not be suitable or beneficial as preventive agents in healthy individuals. This potential limitation underscores the need for careful

evaluation and distinction between therapeutic and preventive applications of these food molecules. Our findings open avenues for further research to discern and validate the roles of these molecules in cancer prevention specifically, as their efficacy and safety profiles might differ significantly from those in a therapeutic context. This aspect warrants a thorough investigation to ensure that the translation of our findings into dietary recommendations does not inadvertently pose health risks.

It is also worth noting that we focused exclusively on the molecular constituents of fruits and vegetables, as our primary data source, FooDB, did not include molecular constituents of meats at the time of this study. While our current research focuses on fruits and vegetables, exploring the molecular constituents of meats presents a significant potential for future studies. Investigating meats could provide insights into how their molecular makeup contributes to increased cancer risks[164,165], thereby complementing our current understanding of dietary impacts on cancer. This could be an important area for future research, potentially leading to a more comprehensive understanding of diet-cancer relationships across different types of food.

A fascinating future direction worth noting is exploring dietary variability across different regions and cultures, and correlating it with cancer incidence rates, which could significantly bolster the hypothesis that certain food products possess anticancer properties. Such an approach would leverage epidemiological data to draw connections between dietary patterns and cancer prevalence. This could provide a broader, real-world context to our findings, enhancing the validity of our conclusions about the anticancer potential of specific food molecules. This global perspective could be instrumental in understanding the complex interactions between diet and cancer risk, and in identifying key dietary factors that could be targeted for cancer prevention strategies.

Results presented in this chapter situate within the computational domain, harnessing the power of graph neural networks to interrogate the vast landscape of bioactive food molecules and their potential anticancer properties. It is important to emphasize that while our computational models offer promising directions for identifying molecules within foods that might exhibit therapeutic effects, these findings are preliminary and serve primarily as hypotheses for further empirical investigation. This work does not advocate for the replacement of established cancer therapies with dietary changes but rather suggests that

certain dietary molecules may complement traditional therapeutics. The clinical translation of these computational insights would require rigorous and extensive validation through experimental studies, encompassing not only the verification of their anticancer efficacy but also assessments of safety and potential interactions with standard cancer treatments. The path from computational discovery to clinical application is a cautious and iterative process, demanding a convergence of computational predictions with traditional biomedical research to ultimately benefit patient care.

4 Predicting food molecules to enhance drug activity

4.1 Abstract

Pharmacological effects of drugs can be greatly influenced by their pharmacodynamic (PD) interactions with other drugs and nutrients, resulting in marked beneficial or harmful clinical effects in subjects. Computational approaches have emerged for PD drug-drug interactions identification; however, PD drug-food interactions remain largely unexplored due to the lack of *in vitro* PD drug-food datasets fueled by the large number of food molecules. Here, we introduce a novel graph deep learning model for the identification of PD food molecule-drug interactions. We calibrated the model using an *in vitro* PD drug-drug interactions dataset achieving a Pearson correlation score of 0.68 for the prediction of PD drug-drug interactions. We experimentally validated this model for the prediction of PD drug-drug interactions in a human lung (A549) and a colorectal (HCT116) cancer cell line, achieving consistency with the predicted results in 8/11 combinations tested. We transferred our proposed model for the prediction of food molecule-drug interactions in a human lung cancer cell line (A549), achieving consistency with predicted results in 3/3 of the combinations tested. Our results show that models trained with experimental PD drug-drug interactions datasets can be successfully utilized for the identification of PD food molecule-drug interactions. This work highlights the utility of deep learning approaches to expand our knowledge of clinically relevant food molecule-drug interactions that could result in an enhancement of drug activity with reduced toxicity.

4.2 Background

Food intake can affect drug activity by altering drug pharmacokinetic (PK) or pharmacodynamic (PD) features[166,167], influencing safety and effectiveness of oral pharmacotherapy[168]. PK food-drug interactions alter drug absorption, distribution, metabolism, and excretion. These alterations occur due to changes in metabolic activity caused by food intake and/or dietary supplement consumption, either at the cell level, such as alterations in cell metabolic processes and cell membrane transporters[169], or at the organ level, such as alterations in gastric emptying kinetics, luminal bile salt concentrations, or liver perfusion[166,167,170]. PD

food-drug interactions affect pharmacological effects of drugs and result in attenuation or enhancement of drug effects and toxicity. The mechanism and effects of each PD interaction are specific to the food molecule and drug involved[168], and their discovery is often serendipitous. A prominent example is the harmful interaction between tyramine, a constituent of cheese and raw sausages, and monoamine oxidase inhibitors (MAOI), which can lead to hypertensive crisis and serotonin syndrome[171]. This association was first noticed by a British pharmacist, G.E.F. Rowe, who observed in 1963 that his wife, who was prescribed a MAOI, would experience a severe headache after eating cheese[172,173]. Rowe wrote a letter to Barry Blackwell, then a resident psychiatrist at Maudsley Hospital in London[173] who later studied the association and finally confirmed it 1965[174].

The specificity of PD interactions combined with the massive number of molecules within foods has hindered the study of PD food-drug interactions at large scale. However, recent efforts to compile comprehensive databases of food chemical composition[93,149] have facilitated *in silico* studies on food-drug and food-disease interactions. For example, Jensen *et al.* identified food molecules that affect the pharmacokinetics or pharmacodynamics of drugs, observing that genes targeted by drugs used to treat cancer, cardiovascular, and neurological diseases are also highly targeted by molecules within foods[94]. Zheng *et al.* built a food-disease interaction network based on publicly available gene expression data and identified beneficial and harmful food-disease pairs[95]. Due to the small size of existing diet-induced gene expression datasets, this correlation-driven analysis was restricted to a very limited number of foods. However, intriguing diet-disease associations have been identified through this approach.

In both studies, interactions were discovered under the assumption that a food molecule-drug/disease interaction exists if both agents affect the same set of targets/genes, perhaps through target-binding competition in the case of drugs. However, even if a food molecule was found to be capable of binding to a drug target, it is unclear whether this interaction would affect the drug binding mechanism and the overall effect[175]. In reality, interactions can arise not only when there is an overlap of affected/targeted proteins, but also when affected proteins are close in a network of protein-protein interactions, i.e., when affected proteins participate in the same or similar biological functions[176] (Figure 4.1A-B). Furthermore, previous works have not taken into account the context in which these interactions might occur

or not, that is, a given pair of molecules could interact differently depending on where they are in an organism (Figure 4.1C-D).

To discover context-specific PD food-drug interactions, it is then key to characterize the effects of food molecules and drugs as affected gene modules and biological functions, rather than individual affected genes[176], and relate the affected gene modules to the context where they interact. Traditionally, techniques to uncover affected protein modules from an initial set of targeted proteins fall under the term *network propagation*[114]. This set of techniques has been used to model the effects of molecules on the genome, including food molecules in our previous works[121,177]. Network propagation techniques propagate or convolve information across a graph, typically a protein-protein interaction (PPI) network, using a predefined propagation rule. An alternative to predefined propagation rules is learnable propagation algorithms, such as graph neural networks (GNNs)[26], which learn how to propagate information across a graph to maximize the performance of a supervised task. GNNs have been used in a number of applications in bioinformatics incorporating PPI networks, achieving state-of-the-art performance[137–139], including the modeling of drug and food molecule effects on the genome in our previous work[178], detailed in Chapter 3.

We used a GNN and a multiscale interactome[179] composed of protein-protein, protein-biological function, and biological function-biological function interactions to learn protein modules and biological functions affected by drugs and food molecules, and to characterize the cellular contexts where they interact, to predict PD food molecule-drug interactions leading to an increase or decrease of drug activity in several types of cancer cell lines. Specifically, PD food molecule-drug interactions resulting in an increased or decreased drug activity are modeled as food molecule-drug synergistic and antagonistic interactions, respectively. We propose a graph neural model for this task and first train it utilizing publicly available experimental data to predict drug-drug synergistic/antagonistic interactions, achieving a Pearson correlation value of 0.68. We experimentally validated selected drug-drug combinations achieving consistency with predicted results in 8/11 combinations tested. We transfer this model, which has learned to predict cell line-specific PD drug-drug interactions, to predict PD food molecule-drug interactions in a lung cancer cell line (A549).

The rest of this Chapter is organized as follows: subsection 4.3 introduces related work and

state-of-the-art approaches for PD drug-drug interaction prediction, subsection 4.4 states the contributions and impact of this Chapter, subsections 4.6, 4.5, and 4.7 introduce the datasets used, our problem definition and our approach, respectively, subsection 4.8 contains results, and subsection 4.9 includes a summary and discussion of the contents of this Chapter.

4.3 Related work

Pharmacodynamic (PD) drug-drug interactions can be classified as additive, antagonistic, and synergistic, if the combined effect is equal, lower, or higher than what would be expected under a hypothesis of non-interaction, respectively. There are four commonly-used synergy models, each one differing in the definition of non-interaction. Suppose that one drug produced an effect y_1 at dose x_1 and the other drug produced an effect y_2 at dose x_2 while combining them produced y_c. In the Highest Single Agent (HSA) model, the expected effect is equal to the highest effect of individual drugs: $y_{HSA} = max(y_1, y_2)$. In the Bliss independence model, the expected effect is calculated based on the probability of independent events: $y_{BLISS} = y_1 + y_2 - y_1 y_2$. The Loewe additivity model states that the expected effect y_{LOEWE} must satisfy: $x_1/x^1_{LOEWE} + x_2/x^2_{LOEWE} = 1$ where x^1_{LOEWE} and x^2_{LOEWE} are the doses of drug 1 and drug 2 alone that produce y_{LOEWE}. The Zero Interaction Potency (ZIP) model assumes that two non-interacting drugs are expected to incur minimal changes in their dose–response curves and captures the drug interaction relationships by comparing the change in the potency of the dose–response curves between individual drugs and their combinations. We refer the reader to[181] for extended definitions of synergy models. In the context of cancer, researchers are often interested in synergistic combinations, those which would elicit a potent response in ideally much lower doses of each individual drug, thus reducing toxicity. Because of this, *synergy prediction* is the term widely used to refer to PD drug-drug interactions studies. *In silico* synergy prediction has been enabled by several publicly available datasets and databases including Merck and Co's dataset[180], NCI ALMANAC[182], DREAM datasets[183,184], DrugCombDB[185], and DrugComb[186]. These datasets contain synergy values of drug-drug combinations, that is, (drug, drug, cell line) and the corresponding synergy values.

In 2018, DeepSynergy[187] was introduced, the first work using deep learning (DL) for the task of predicting drug synergy. The authors modeled Loewe synergy in the Merck

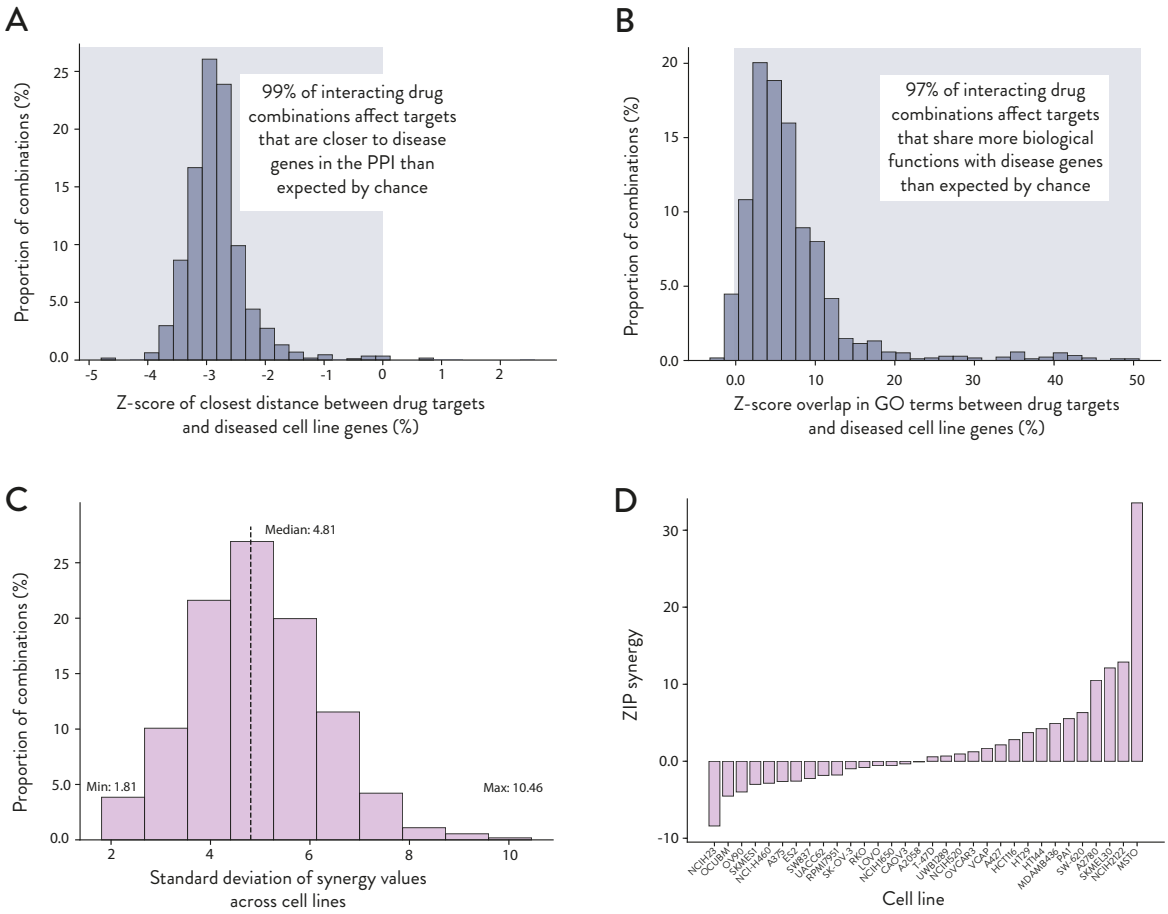


Figure 4.1: Drug-drug interactions depend on network proximity, biological function overlap, and context. (A) In 99% of the interacting drug combinations in the Merck Co's dataset [180] (synergy value > 10 or < -10), proteins targeted by drugs are closer to disease (cancer cell line) genes than expected by chance. The distance metric used is the Z-score of the shortest path distance between drug combination targets and disease genes (see *Methods*). **(B)** In 97% of interacting drug combinations in Merck and Co's dataset [180] (synergy value > 10 or < -10), drug targets affect the same biological functions as mutated genes of cell lines where they interact more often than is expected by random chance. The functional overlap metric used is the Z-score of the intersection of Gene Ontology (GO) term multisets (see *Methods*). **(C)** Histogram of distribution of standard deviation of synergy values across cell lines per combination, for all (drug, drug, cell line) triplets in Merck and Co's dataset [180]. This plot shows that there is great variability of drug-drug synergy values depending on the context (cell line) in which they interact. **(D)** Distribution of synergy values across cell lines for the combination of Metformin and Dasatinib in Merck and Co's dataset [180], which exhibits a highly context-specific behavior.

and Co's dataset[180] using a feed-forward neural network with drug fingerprints, physico-chemical properties, and toxicophore features, in addition to cell line gene expression as input. DeepSynergy achieved a Pearson correlation score of 73±4 in a leave-combination-out setting, outperforming a number of traditional machine learning (ML) algorithms including Gradient Boosting Machines, Random Forests, Support Vector Machines and Elastic Nets. This catalyzed the development of DL algorithms for drug synergy prediction. Following DeepSynergy, Zhang *et al.* proposed DeepSignalingSynergy[188] in 2020, a simplified version of DeepSynergy[187] achieving comparative performance. DeepSignalingSynergy[188] introduced knowledge of signaling pathways within the neural network to aggregate neuron outputs, each corresponding to a gene, according to existing knowledge of membership in gene signaling pathways. Since then, several DL-based approaches have emerged, leveraging various types of techniques and architectures such as Autoencoders[189], Tranformers[190], and GNNs[191–195]. Next, we will briefly review such DL approaches and justify the need for our work.

AuDDNsynergy[189], introduced in 2021, is an autoencoder-based approach that models Loewe synergy. It comprises three autoencoders that generate representations of cell line gene expression, copy number variation, and genetic mutation data. These representations are concatenated with drug chemical features and fed to a feed-forward neural network for the prediction of synergy scores. TranSynergy[190], introduced in 2021, uses a Transformer architecture that operates on drug targets, cell line gene expression, and gene dependency data extracted from DepMap[196]. Both AuDDNsynergy[189] and TranSynergy[190] achieved similar performance to DeepSynergy[187] under similar training settings in the Merck and Co's dataset[180]. GNN-based approaches for the prediction of drug-drug synergy appeared as early as 2020 with Jiang *et al.*'s work[191] who reformulated the problem of predicting drug-drug synergy as a link prediction (classification) task over a heterogeneous network composed of drug-drug, drug-protein, and protein-protein connections. Specifically, a drug-drug connection represented synergy between drugs, and a lack of connection represented a lack of synergy. Due to the lack of cell line-specific information in the heterogeneous network, Jiang *et al.* trained one separate model per cell line. Furthermore, given that the task was framed as a link prediction instead of a regression task, it is challenging to put their performance in the context of competing methods. In DeepDDS[192], introduced in 2021, Wang *et al.* leveraged GNNs to compute drug representations from their 2D structure, maintaining the

standard problem formulation used by previous non-GNN-based approaches.

More recently, DL approaches have focused on introducing custom architectures for the drug synergy prediction task, either tabular approaches such as MatchMaker[197], or GNN-based models such as SDCNet[193], PRODeepSyn[195], and HypergraphSynergy[194]. In SDCNet[193], Zhang *et al.* use the problem formulation introduced earlier by Jiang *et al.*[191] where the drug-drug synergy prediction is viewed as a link prediction task. They built individual drug-drug networks for each cell line and integrated them into a relational network in which nodes are drugs and edges between drugs have a label indicating the cell line in which the pair of drugs are synergistic. In the classification task, they achieved slightly superior performance compared to previously introduced methods including Jiang *et al.*'s work[191], DeepSynergy[187], and MatchMaker[197]. PRODeepSyn[195], on the other hand, utilizes a GNN-based model to compute cell line representations and adheres to the standard problem formulation as a regression task. Cell line representations are concatenated with chemical fingerprints of drugs in combinations and fed to a feed-forward neural network for the synergy prediction task. PRODeepSyn was tested on the same dataset and split as DeepSynergy, achieving comparative performance. In HypergraphSynergy[194], Liu *et al.* build a hypergraph with drugs and cell lines as nodes and (drug, drug, cell line) triplets representing hyperedges. A Hypergraph GNN[198] was used to compute representations of drugs and cell lines then fed to a feed-forward neural network for the prediction of synergy scores, achieving comparative performance to previously-introduced methods.

In spite of the substantial efforts to introduce more sophisticated and complex DL architectures, drug-drug synergy performance has remained comparable to the earliest DL approach DeepSynergy[187]. Therefore, rather than focusing on architecture sophistication, our goal is to introduce a simple and flexible approach that can be leveraged for the task of food molecule-drug synergy prediction. To solve this, we introduce and train a drug-drug synergy prediction model and use the trained model to predict food molecule-drug synergy. Food molecule data availability is limited, therefore, we build our drug-drug synergy prediction model leveraging protein target data only, in addition to disease genes (cancer cell line mutations).

We hypothesize that drugs will interact if they affect similar protein modules or biological functions as those dysregulated in disease (Figure 4.1A-B), therefore, our goal is to build

representations of drugs and disease that capture their effects on both proteins and biological functions. Moreover, we hypothesize that due to the sparsity of protein target and mutation data, a network-based approach propagating the information of protein targets and disease genes across the genome would successfully capture effects of drugs, food molecules, and diseases. We aim to capture their effects on protein and biological functions by using network propagation on a multiscale network containing protein-protein, protein-biological function, and biological function-biological function interactions. In a similar fashion to existing approaches, for a given (drug, drug, cell line) triplet, we compute representations of drugs and cell line and feed them to a feed-forward neural network to predict a synergy score.

4.4 Contributions and impact

Contributions

- We propose a graph neural model for the prediction of drug-drug synergy achieving a Pearson correlation score of 0.68. We perform in-vitro validation of selected predicted drug-drug interacting pairs, in one colorectal (HCT116) and one lung cancer (A549) cell line, showing consistency with predicted results in 8/11 combinations tested.
- We use attribution methods to uncover proteins that causally influence model predictions.
- Using the model trained on drug-drug combinations, we uncover synergistic and antagonistic food molecule-drug interactions in one lung cancer cell line (A549). We perform in-vitro validation of selected predicted food molecule-drug interacting pairs, showing consistency with predicted results in 3/3 combinations tested.

Impact. We introduce a simple yet powerful model for drug synergy prediction in cancer relying only on drug protein targets and disease genes. Our model, trained to predict drug-drug synergy, successfully predicts food molecule-drug synergy, as validated experimentally. 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.

This section is based on the following publication:

1. **Gonzalez, G.**, Laponogov, I., Brignardello, J., Chio, D., Reyna-Jeldes, M., Rita, L., Qiu, J., Gallota, M., Wang, Y., Zhang, E., Zhong, Y., Wang,X., Bronstein, M., Veselkov, K. (in preparation). Leveraging food-drug pharmacodynamic interactions to enhance drug activity

Contributions of collaborators:

- Jerusa Brignardello, Donald Chio, Luis Rita, Marco Gallota, Yuxi Zhing, Xiuqi Wang and myself carried out the wet-lab validation of selected drug-drug and food molecule-drug combinations
- Mauricio Reyna-Jeldes kindly provided putative mechanisms of action of drug-drug and food molecule-drug synergy included in the results

4.5 Problem definition

Food molecule-drug synergy score prediction. Given a food molecule and a drug, each one represented by a binary vector $\mathbf{x} \in \{0,1\}^{|\mathcal{V}|}$ encoding its protein targets, and a cancer cell line, represented by a binary vector $\mathbf{x} \in \{0,1\}^{|\mathcal{V}|}$ encoding its cancer-causing mutated genes (disease genes), our goal is to predict a single value representing the ZIP synergy score between food molecule and drug in the selected cell line. A food molecule-drug pair should be predicted with high (low) ZIP synergy score if a PD interaction between them increases (decreases) drug efficacy in the cell line of interest; and should be predicted with a ZIP synergy around zero if no PD interaction exists. We pre-train a regression model to predict drug-drug ZIP synergy scores from available in-vitro drug-drug interaction data and then use the trained model to predict food molecule-drug interactions (see Figure 4.3A).

Drug-drug synergy score prediction. Given a pair of drugs, each one represented by a binary vector $\mathbf{x} \in \{0,1\}^{|\mathcal{V}|}$ encoding its protein targets, and a cancer cell line, represented by a binary vector $\mathbf{x} \in \{0,1\}^{|\mathcal{V}|}$ encoding its cancer-causing mutated genes, our goal is to predict a single value representing the ZIP synergy score between drugs in the selected cell line.

4.6 Datasets

Drug-drug interactions. We extracted drug-drug interaction data of Merck and Co's study[180] from the DrugComb database[186] (accessed 30 Nov 2020). The dataset contains 583 unique drug-drug combinations, each tested against 39 human cancer cell lines derived from 7 different tissue types. Pairwise combinations were constructed from a set of 38 drugs, of which 22 were tested exhaustively in combination (the "exhaustive" set), while the remaining 16 (the "supplemental" set) were tested only in combination with those of the exhaustive set resulting in 22,737 (drug, drug, cell line) triplets. Drug targets and cell line mutations were extracted from STITCH[90] and COSMIC CLP[91] (see headers "Drug and food molecule protein targets" and "Cell line mutations" in this section).

We filtered combinations to keep only those with drug targets and cell line mutations available (see Supplementary Table 8.3 for a list of the remaining drugs and cell lines). The final dataset contains 546 unique pairwise drug combinations between 37 drugs, each tested in 32 cell lines, resulting in 17,472 (drug, drug, cell line) triplets.

We model ZIP synergy values in this study. Each drug-drug combination (i, j, c) of drugs i and j tested in cell line c in the dataset is associated to a real-valued synergy score s_{ijc} representing the degree of deviation, in percentage, of the measured effect of the given combination compared to the expected effect under the hypothesis of no interaction. Higher values indicate synergistic, lower values indicate antagonistic and intermediate values indicate additive PD interactions. We use thresholds of 10 and -10 when referring to synergistic and antagonistic drug-drug interactions, respectively, as advised in[199].

Synergy scores in the dataset are distributed evenly around 1.6, with a positive to negative ratio of 64:36 (Figure 4.2C). To assess the positive/negative balance of drug-drug synergy scores, for each pairwise drug combination, we compute the positive ratio as the number of cell lines in which the combination exhibits positive synergy divided by the total number of cell lines in which the combination was tested. A positive ratio equal to 1 or 0 means that the drug pair has a positive or negative synergy value, respectively, across all tested cell lines. Although the positive ratio is slightly biased to values higher than 0.5 (median positive ratio = 0.66, see Figure 4.2D), indicating that a higher proportion of drug combinations exhibit positive synergies in a bigger proportion of cell lines compared to negative synergies, 72%

of combinations have a positive ratio within [0.2, 0.8], and only 1% of combinations (n=6) have a positive ratio equal to 1. This shows that, although slightly biased to positive synergy values, a high proportion of drug combinations in the dataset has balanced positive/negative synergy scores across cell lines, which prevents the model from easily predicting the sign of synergy values by memorizing the combinations.

Protein-protein interactions (PPIs). We constructed a human protein-protein interaction network for protein sequences from COSMIC[91], NCBI Gene[200], STRING[100] and UniProt[201] databases. The databases were joined to construct a final unified set of 20,256 proteins. The list of corresponding genes was further populated with different gene IDs and synonyms (including Ensembl and HGNC). PPIs were obtained from the STRING database and supplemented with confidence scores (0-999) from STRING. For the purposes of this study, we build three PPI networks using confidence thresholds: 700, 800 and 900. The PPI network is a graph $G = (\mathcal{V}, E)$ of protein-protein interactions, with $|\mathcal{V}| = 20,256$ nodes and $|E| = \{739,413;\ 639,083;\ 434,069\}$ edges using confidence thresholds 700, 800, or 900. STRING integrates PPI data from several sources, including experimental results, computational prediction methods, and public text collections, assigning a confidence score between 0 and 999 to each interaction based on the evidence supporting it. We decided to perform our experiments with a range of confidence thresholds to account for the uncertainty in the PPI data.

Protein-biological function interactions. We obtain protein-biological function interactions from[179]. In brief, proteins were mapped to the biological functions they affect using the human version of the Gene Ontology[202,203], only allowing experimentally verified associations between genes and biological functions. Protein-coding genes were mapped to genes in the PPI by Hugo Gene ID (37,328 protein-biological function interactions between 8,391 genes and 6,383 biological functions). It is worth noting the incompleteness of the protein-biological function interaction data, which is biased towards well-studied proteins. Even though we do not address this uncertainty in our modeling, we acknowledge it as a potential source of noise.

Biological function-biological function interactions. We obtain biological function-

biological function interactions from[179]. In brief, a hierarchy of biological functions was constructed by using the Gene Ontology's Biological Processes[202,203], which represents a curated hierarchy of biological functions (pathways), where highly specific biological functions are children of more general biological functions according to numerous relationship types. Allowed relationships are of the following types: *regulates*, *positively regulates*, *negatively regulates*, *part of*, and *is a*. The pathway ontology network is represented by a graph $G = (\mathcal{B}, L)$ of biological function-biological function interactions, containing $|L| = 32,343$ interactions between with $|\mathcal{B}| = 9,798$ nodes. In the current implementation of our graph deep learning model, the relationship types, such as *positively regulates* and *negatively regulates*, provided by Gene Ontology annotations were not explicitly modeled. The graph representation of protein-biological function interactions and biological function-biological function interactions was constructed to capture the connectivity and hierarchical structure of biological functions as dictated by Gene Ontology, but without differentiation of the regulatory direction. This approach was chosen to prioritize the establishment of a foundational network of interactions upon which future models could iteratively incorporate the directionality of regulation. We acknowledge that the direction of change induced by these regulators is a critical aspect of pathway dynamics, and future work could extend the model to differentiate these regulatory effects for a more nuanced understanding of pathway modulation.

Drug and food molecule protein targets. We extracted drug information from DrugBank[88] and DrugCentral[89], and food molecule information from FooDB[149]. We obtained drug and food molecule protein targets from STITCH[90] (Figure 4.2A). STITCH includes a confidence score ranging from 150 to 999 for each interaction based on the evidence supporting it. Here, we filtered drug and food molecule protein targets using a confidence threshold of 700 to ensure the reliability of the drug-target associations, as advised in the STITCH documentation. Extracted targets were mapped to the constructed PPI network. Each drug (food molecule) i is represented by and a protein node feature vector $\mathbf{x}_i \in \{0,1\}^{|\mathcal{V}|}$, i.e. one binary feature per node: 1 if the protein is a drug (food molecule) target, 0 otherwise. Our dataset contains 3,315 food molecules.

Cancer cell line mutations. We obtained cell line mutations from COSMIC Cell Line Project (CLP)[91] (version 92). We used COSMIC ID to map cell lines from DrugComb

database to COSMIC. Mutations were filtered to keep only those recorded as "pathogenic" in COSMIC CLP (Figure 4.2B). Similarly to drugs, we represent each cell line c by a binary protein node feature vector $\mathbf{x}_c \in \{0,1\}^{|\mathcal{V}|}$, i.e. one binary feature per node: 1 if the corresponding gene is mutated in cell line c, 0 otherwise.

4.7 Methods

Network proximity between drug-drug combination targets and disease genes. Let $\mathcal{T}$ be the set of drug targets, $\mathcal{G}$ be the set of disease genes, and $spd(t, g)$ be the shortest path distance between nodes in $\mathcal{T}$ and $\mathcal{G}$. We measure the closest distance between $\mathcal{T}$ and $\mathcal{G}$ as:

$$d(\mathcal{T}, \mathcal{G}) = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \underset{t \in \mathcal{T}}{min}\ spd(t, g)$$

To compute a Z-score, this distance is compared to a reference distribution which measures the shortest path distance with 1000 randomly permuted sets of proteins $\mathcal{T}'$ and $\mathcal{G}'$ that match the size and degrees of the original drug combination targets $\mathcal{T}$ and disease genes $\mathcal{D}$ in the network.

Functional overlap between drug-drug pair targets and disease genes. Let $\mathcal{U}$ and $\mathcal{V}$ be the multisets of Gene Ontology (GO) terms associated with the union of targets of drugs in a combination and cell line mutated proteins respectively. Note that in the multisets $\mathcal{U}$ and $\mathcal{V}$, $\mathcal{U}_i$ and $\mathcal{V}_i$ correspond to the number of occurrences of the ith element in the multiset. We measure the functional overlap as the Jaccard Similarity between $\mathcal{U}$ and $\mathcal{V}$:

$$\frac{\sum_i min(\mathcal{U}_i, \mathcal{V}_i)}{\sum_i max(\mathcal{U}_i, \mathcal{V}_i)}$$

To compute a Z-score, this overlap is compared to a reference distribution which measures the multiset overlap when $\mathcal{U}$ and $\mathcal{V}$ are obtained with 1000 randomly permuted sets of proteins that match the size and degrees of the original drug combination targets and disease genes in the network.

Predicting food molecule-drug interactions. We calibrate a graph neural model to predict phenotype-specific drug-drug ZIP synergy scores. We evaluate performance and interpretability of our model, and experimentally test selected new drug-drug interactions

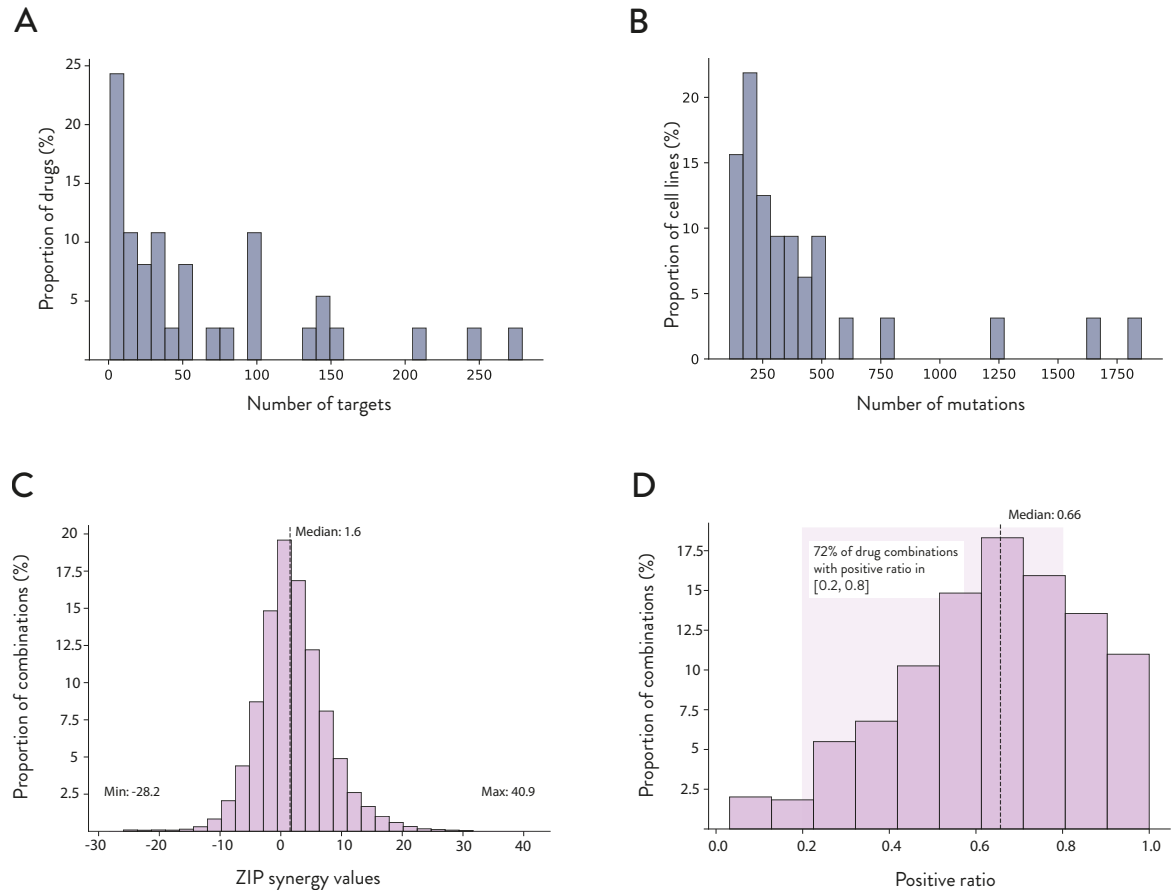


Figure 4.2: Merck and Co's dataset[180], an unbiased oncology screen. **(A)** Distribution of drug targets for the 37 drugs in Merck and Co's (see Supplementary Table 8.3). Drug targets were extracted from STITCH[90]. Only high confidence drug-target associations were kept. **(B)** Distribution of mutations for the 32 cell lines in Merck and Co's dataset (see Supplementary Table 8.3). Cell line mutations were extracted from COSMIC CLP[91]. Mutations were filtered to keep only those recorded as "pathogenic". **(C)** Distribution of ZIP synergy scores. The ZIP synergy score represents the deviation, in percentage, of a drug combination effect compared to the expected effect under the hypothesis that the drugs do not interact. Synergy scores of samples in Merck and Co's dataset distribute evenly around 1.6. **(D)** Distribution of ZIP synergy positive ratio values per combination across cell lines. For a given drug combination, the positive ratio represents the number of cell lines in which the combination had a positive synergy value divided by the total number of cell lines the given combination was tested in. Positive ratios of 1 and 0 represent combinations with positive and negative synergy value, respectively, in all cell lines in which the combination was tested. 72% of combinations have positive ratios within [0.2, 0.8], ensuring that a high proportion of combinations have balanced positive/negative synergy scores across cell lines.

predicted. We then use the trained and validated model to predict phenotype-specific food molecule-drug ZIP synergy scores (Figure 4.3A). We formulate the synergy prediction problem as a regression task, in which, given a drug-drug pair and a cancer-derived cell line, we predict a synergy score. The graph deep learning model propagates or diffuses the information from drug protein targets and cell line mutated genes across the multiscale interactome, in a learnable fashion, to learn the effects of drugs and phenotype on proteins and biological functions. The learned drug and phenotype effects, which we call *profiles*, are used to predict synergy score of the drug-drug pair. Processed data and code is available at bitbucket.org/iAnalytica/foodcomb/.

Regression model. We cast the problem of predicting phenotype-specific drug-drug synergy scores as a regression task, in which, given a pair of drugs, and a cell line, we predict a score representing the drug-dug synergy in that specific cell line (Figure 4.3B-C). Drug protein targets and cell line mutations are used as input data to the model. More precisely, for each drug-drug combination (i, j, c) of drugs i and j tested in cell line c, we use input data:

- Protein targets of each drug on the PPI network, each one represented as a node feature vector $\mathbf{x}_i \in \{0,1\}^{|\mathcal{V}|}$ and $\mathbf{x}_j \in \{0,1\}^{|\mathcal{V}|}$, i.e. one binary feature per node: 1 if the protein is a target of drug i/j, and 0 otherwise.
- Mutations in cell line c, represented by a binary node feature vector $\mathbf{x}_c \in \{0,1\}^{|\mathcal{V}|}$, i.e. one binary feature per node: 1 if the corresponding gene is mutated in cell line c, 0 otherwise.

For a given drug-drug combination (i, j, c) of drugs i and j tested in cell line c, the proposed graph neural model takes protein targets on the PPI of drugs i and j, and mutations set of cell line c, and outputs a real-valued score representing the synergistic effect of drugs i and j in cell line c (Figure 4.3B-C). We propagate the effects of each drug and cancer mutations across the multiscale interactome using GNNs. The learned diffusion profile represents the degree of affection of proteins and biological functions as a result of drugs and cancer mutations effects. Each drug and cell line diffusion profile is computed using GNNs which diffuse information in a learnable fashion through several *diffusion channels*, which are later aggregated for each protein and biological function.

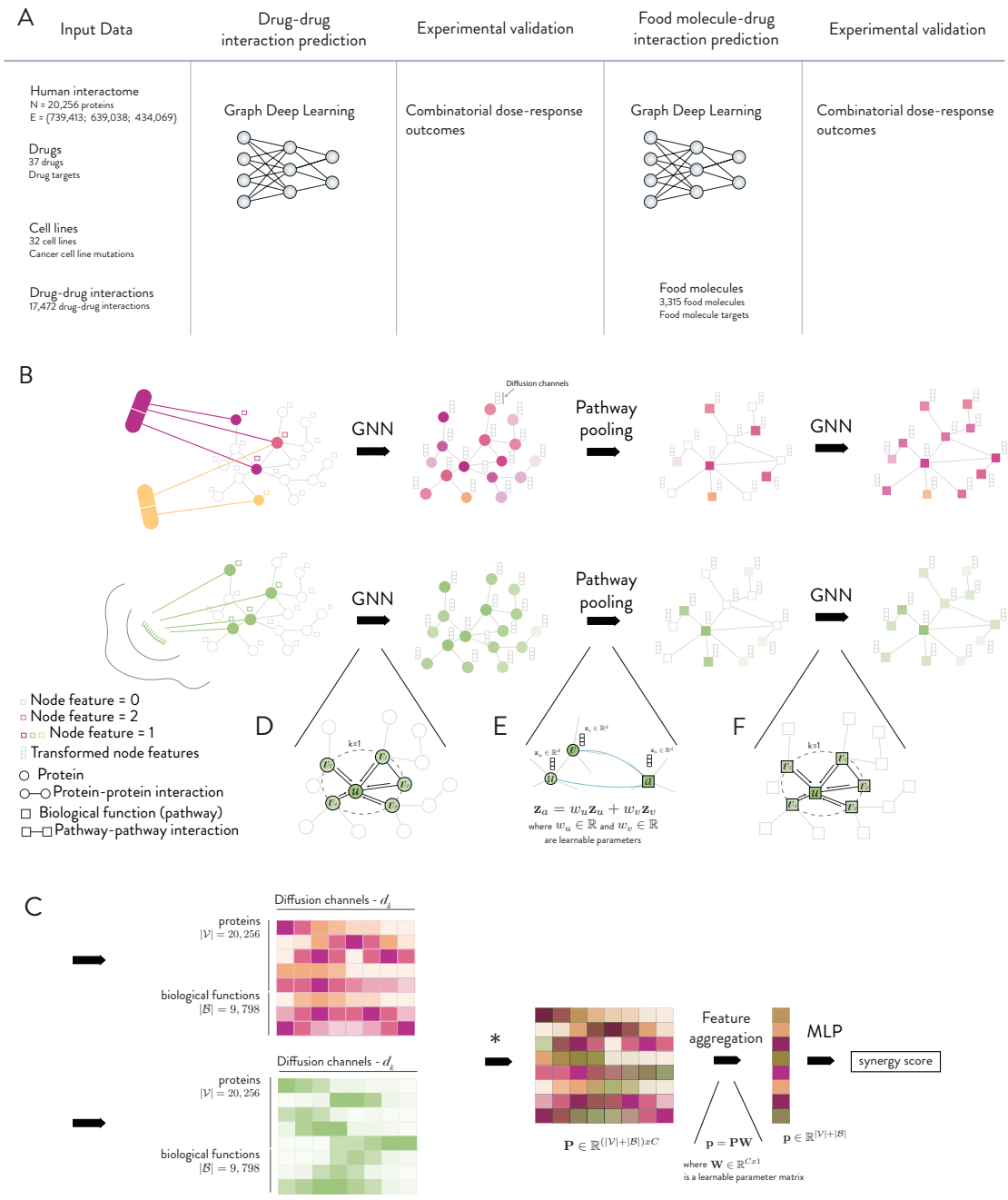


Figure 4.3: Study design for food molecule-drug interaction prediction. (A) A graph deep learning model is trained with annotated drug-drug interactions. After experimental validation, the trained model is used to predict food molecule-drug interactions. **(B -C)** We introduce a graph neural network (GNN)-based model for the prediction of synergy scores. Drug targets and cancer cell line mutations are encoded onto the PPI as node (protein) features. A first GNN diffuses information across the PPI, generating transformed node features capturing the initial node features and PPI topology. Node features are *pooled* or *propagated* to node's biological functions (pathways), using a learnable node feature transformation process. A second GNN diffuses information across the pathway ontology network generating transformed node embeddings capturing the initial pathway features (diffused from the PPI) and the pathway ontology network topology. Drug and cell line protein and biological function embeddings are concatenated and multiplied. A final feature aggregation reduces the protein and biological function embeddings dimensionality to 1. This embedding is fed to a Multi-Layer Perceptron (MLP) to predict a synergy score. **(D)** Shown is the generic message-passing mechanism followed by GNNs to propagate information across the graph. For a given protein node u, in each layer, information from the 1-hop neighborhood $\mathcal{N}(u)$ of u is aggregated with the current node information to generate a transformed representation of node u. This process is repeated for each node in the graph. **(E)** Shown is the pooling or *propagation* process from proteins to their biological functions. **(F)** Shown is the message-passing mechanism followed by GNNs for a single node. For a given pathway node u, in each layer, information from the 1-hop neighborhood $\mathcal{N}(u)$ of u is aggregated with the current node information to generate a transformed representation of node u.

Learning drug and cell line diffusion profiles. For a given pairwise drug combination (i, j, c) of drugs i and j tested in cell line c, graph convolutional networks propagate the effects of drug targets $\mathbf{x}_i$ and $\mathbf{x}_j$ and cell line mutations $\mathbf{x}_c$ across the multiscale interactome (Figure 4.3B-F) in a learnable fashion. Graph neural networks are able to diffuse information across the multiscale interactome along a number of *diffusion channels* d_k in each layer k, adding flexibility to the diffusion process, compared to traditional network propagation methods, as the graph neural layer is able to learn different diffusion processes in parallel. A graph neural network first propagates effects of drug protein targets and cell line mutated proteins across the PPI, in parallel, generating transformed node (protein) embeddings capturing the initial node features and the PPI topology (Figure 4.3D). Transformed node (protein) representations are *pooled* or *propagated* to node's biological functions (pathways), using a learnable node feature transformation process (Figure 4.3E). A second graph neural network propagates information across the biological function ontology network, generating transformed node (pathway) embeddings capturing the initial pathway features and the pathway ontology network topology (Figure 4.3F). For each, drug combination and cell line, their protein and biological function embeddings are concatenated, and then multiplied (Figure 4.3C). A final feature aggregation step reduces the protein and biological function embeddings dimensionality to 1. Intuitively, final node representations capture the degree of affection of each protein and biological function as a result of the drug and cell mutations effects, oriented by initial node (protein) feature and topological similarity between nodes in the PPI and the pathway ontology network.

For a given node, each layer of a graph neural network propagates or diffuses information from its first-order neighborhood and applies the same propagation across all locations in the graph (Figure 4.3D,F). Stacking K convolutional layers in the neural network then translates into information propagation across the K-th order neighborhood (i.e. a node's information is propagated to all of its neighbors at up to K-hop distance). Graph neural layers are typically followed by a non-linear function to allow for higher representation power. We use two variants of GNNs: Graph Convolutional Network (GCN)[118] and Chebyshev Network (ChebNet)[120].

To generate representations of drugs i, j and cell line c, we apply K convolutional layers to the initial representations $\mathbf{x}_{ij} = \mathbf{x}_i + \mathbf{x}_j$ and $\mathbf{x}_c$ on the PPI and K convolutional layers

on the pathway ontology network. Concatenated protein and pathway embeddings $\mathbf{Z}_{ij} \in \mathbb{R}^{|\mathcal{V}|+|\mathcal{B}|\ x\ d_K}$, and $\mathbf{Z}_c \in \mathbb{R}^{|\mathcal{V}|+|\mathcal{B}|\ x\ d_K}$ are multiplied element-wise $\mathbf{Z}_{ijc} = \mathbf{Z}_{ij} \odot \mathbf{Z}_c$ and then fed to a fully-connected layer to reduce the dimensionality of representations by aggregating node features to a single dimension:

$$\mathbf{z}_{ijc} = \mathbf{Z}_{ijc}.\mathbf{W}_{fc}$$

Where $\mathbf{z}_{ijc} \in R^{|\mathcal{V}|+|\mathcal{B}|}$ represents the learned combined effects of drugs i and j and cell line c on the multiscale interactome, and $\mathbf{W}_{fc} \in \mathbb{R}^{d_K x 1}$ is a vector of weights aggregating node features, learned during training.

Computing synergy value from drug and cell line diffusion profiles. Learned representation $\mathbf{z}_{ijc}$ of the combination (i, j, c) is fed to a Multi-Layer Perceptron (MLP) for the prediction of synergy scores (Figure 4.3C).

$$s_{ijc} = \ MLP(\ \mathbf{z}_{ijc}\ )$$

Each layer of this neural network takes the form:

$$\mathbf{h}_{ijc}^{k+1} \ = \ \sigma\ (\ \mathbf{W}\ \mathbf{h}_{ijc}^{k}\ )$$

Where $\mathbf{h}_{ijc}^{k+1} \in \ \mathbb{R}^{d_{k+1}}$ is the representation of the combination (i, j, c) in step $k+1$; $\mathbf{h}_{ijc}^{k} \in \mathbb{R}^{d_k}$ is the representation of the combination (i, j, c) in step k, with $\mathbf{h}_{ijc}^{0} = \mathbf{z}_{ijc} \in \mathbb{R}^{|\mathcal{V}|+|\mathcal{B}|}$; and σ is a non-linear element-wise activation function.

Attributing predictions to proteins and biological functions for interpretability. We seek to assess which features in the diffusion profiles are more relevant for the synergy prediction task. We can obtain a score/ranking of features importance for the prediction using attribution methods. Typically, attribution methods assign scores to each input feature that reflect the contribution of that feature to the model prediction. A recently introduced set of methods use the notion of *conductance* to extend the notion of attribution to the understanding the importance of hidden units in a neural network[204].

Given a trained model, an input, and a hidden unit, an attribution method assigns a score to the given hidden unit that reflects the contribution of that hidden unit to the model

prediction. Formally, suppose we have a function $F\colon \mathbb{R}^n \to \mathbb{R}$ that represents a deep neural network. We let $\mathbf{x} \in \mathbb{R}$ be the input at hand, and $\mathbf{x}' \in \mathbb{R}$ be a baseline input. For image networks, the baseline could be the black image. To compute the conductance or importance of each neuron with respect to each input feature, we use the Neuron Conductance Method[204]. This method provides neuron conductance values contributing to the prediction relative to a baseline input. The gradient conductance through neuron y for the attribution to an input variable i is:

$$Cond_i^y(\mathbf{x}) ::= (x_i - x_i') \int_{\alpha=0}^{1} \frac{\delta F(\mathbf{x}' + \alpha \times (\mathbf{x} - \mathbf{x}'))}{\delta y} \frac{\delta y}{\delta x_i} \delta\alpha$$

The total conductance or importance of the hidden neuron y can be computed by summing over the input variables:

$$Cond^y(\mathbf{x}) ::= \sum_i (x_i - x_i') \int_{\alpha=0}^{1} \frac{\delta F(\mathbf{x}' + \alpha \times (\mathbf{x} - \mathbf{x}'))}{\delta y} \frac{\delta y}{\delta x_i} \delta\alpha$$

In our case, given a drug combination (i, j, c), and an input $\mathbf{x} = [\mathbf{x}_i | \, \mathbf{x}_j | \, \mathbf{x}_c]$, $F(x)$ denotes the real-valued synergy score. We compute total conductance of neurons in the first layer of the prediction MLP. This reveals what neurons, in this case representing proteins and biological functions, were most relevant to the model's decision.

Neuron scores provided by the Neuron Conductance method sum up to the difference between the model prediction for sample $\mathbf{x}$ and baseline $\mathbf{x}'$ (i.e. $\sum_{y_j} Cond^{y_j}(\mathbf{x}) = F(\mathbf{x}) - F(\mathbf{x}')$ for all neurons y_j in layer j - first layer of the MLP). We choose a baseline with all initial feature values set to 0, to represent lack of information, which results in a predicted synergy score for baseline $\mathbf{x}'$ close to 0. Then, the most important features for the prediction task will be the most positively (negatively) scored features for combinations with positive (negative) synergy values. In practice, we use the efficient method proposed by[205] to compute neuron conductance.

Model training. During model training, we optimize model parameters using Mean Squared Error (MSE) loss:

$$L(i, j, c) = (s_{ijc} - \hat{s}_{ijc})^2$$

Where s_{ijc} and $\hat{s}_{ijc}$ are the real and predicted synergy score of (i, j, c), respectively. We

train the model for a maximum of 50 epochs (training iterations) with the Adam optimizer and early stopping with a window size of 20: The training stops if the validation loss does not decrease at least $1e-4$ for 20 consecutive epochs. We implement our model using PyTorch[154] and the Torch Geometric Library[155].

Evaluating predictions of phenotype-specific drug-drug synergy. We evaluate how effectively a model predicts phenotype-specific synergy scores by using MSE, and Pearson correlation. For each triplet (i, j, c) representing combination of drugs i and j tested in cell line c, a model produces a real-valued synergy score. For each model, we obtain the MSE and Pearson correlation values across all triplets. We create hold-out test and train datasets without shared combinations. To ensure robust model optimization, we perform five-fold cross-validation on the training dataset. We compute the evaluation metrics on the hold-out test set for the optimized models in each of the 5 splits. The reported metric correspond to the performance metrics on the held-out set averaged across folds.

Experimental setup. We view the problem of predicting drug-drug synergy scores as a regression task. Here, each combination (i, j, c) of drugs i, and j tested in cell line c is associated with a synergy score s_{ijc}. We split drug pairs into training and test sets, keeping 10% of the combinations in the test set, and ensuring no overlap of combinations between the sets. To optimize model hyperparameters, we use Bayesian hyperparameter optimization with 5-fold cross-validation on the training set. Bayesian hyperparameter optimization learns to select optimal hyperparameters based on performance using prior hyperparameter settings, allowing for rapid identification of the best set of hyperparameters for any model. Using 5-fold cross-validation, each of the hyperparameter combinations is tested 5 times, and the average performance across the 5 splits is the final performance of a given hyperparameter combination. We use 20 steps of Bayesian hyperparameter optimization, with the final selected hyperparameter combination being the one with highest performance. Hyperparameter search space is: number of layers in MLP between 1-8; number of layers in GNN between 1-3; dimension of representations in MLP between 8-64; dimensions of representations in GNNs between 4-24; PPI threshold in [700, 800, 900]; number of hops for ChebNet in [2, 3, 4, 6]; L2 regularization in MLP between 0-0.005; dropout in MLP between 0-0.4; and learning rate in [1e-5, 1e-4, 1e-3, 1e-2, 1e-1].

We evaluate the performance of the proposed model against a set of proposed baselines using input data in the form $\mathbf{x}_{ijc}$ where $\mathbf{x}_{ijc} = (\mathbf{x}_i + \mathbf{x}_j) \cdot \mathbf{x}_c$ following the approach we used in our proposed model.

- **MLP:** We feed input data to an MLP which applies multiple non-linear transformations to predict a synergy score. Hyperparameter settings are optimized using Bayesian hyperparameter optimization and the same 5-fold cross-validation strategy used to optimize the proposed model.
- **Linear models with Lasso/Elastic-Net regularization:** We feed input data to a linear regression model which applies linear transformations to predict a synergy score. Hyperparameter settings are optimized using 5-fold cross-validation. We standardized input data and extracted features using Principal Component Analysis (explained variance ratio = 99%).
- **Random Forest:** We feed input data to a Random Forest mode, an ensemble learning method of decision trees. We standardized input data and extracted features using Principal Component Analysis (explained variance ratio = 99%).

Performance is measured using the test set MSE and Pearson correlation value averaged across the 5-fold cross-validation splits. Lower values of MSE and higher values of Pearson correlation indicate better performance.

Ensembling to predict novel drug-drug and food molecule-drug synergies. After we evaluated performance of our model in the drug-drug synergy prediction task, we conduct predictions for novel drug-drug combinations and food molecule-drug combinations. To do this, we use the entire dataset to train our model given that we are no longer interested in measuring performance on the test set. We train new models using 20 random data splits, each with 90% training data and 10% validation data to perform early stopping. For each of the random splits, we use 20 steps of Bayesian hyperparameter optimization, with the final selected hyperparameter combination for each split being the one with the highest performance. After models training, for each new combination, we average synergy score predictions from all models. We predict synergy in a lung cancer cell line (A549) and a colorectal cancer cell line (HCT116) for new drug-drug combinations comprised by all

possible combinations of anticancer drugs, as recorded in DrugCentral[89](n=189), excluding those in the training dataset. We predict synergy of food molecule-drug combinations in a lung cancer cell line (A549) for all possible combinations between food molecules from FooDB[149] (n=3,315), and any drug in the training set (n=37).

The selection of A549 and HCT116 cell lines in our study, initially based on availability, can be justified given their relevance in cancer research. A549, a lung carcinoma cell line, and HCT116, a colorectal carcinoma cell line, are extensively utilized due to their well-characterized genetics and their representativeness of lung and colorectal cancer biology, respectively. This choice offers a broad spectrum for assessing drug-drug synergies, allowing for both validation within a known context (HCT116 in the training set) and exploration of the model's generalizability in an unfamiliar setting (A549, not in the training set). Furthermore, testing food molecule-drug combinations in A549 aids in exploring potential therapies in lung cancer, aligning with the study's objective of examining the efficacy and specificity of predicted drug combinations across diverse cancer types. This approach enhances the robustness and relevance of our study, despite the initial pragmatic decision based on cell line availability.

Experimental validation of drug-drug and food molecule-drug combinations

Source of molecules. CRC cell line HCT116 (CCL-247) and LC cell line A549 were purchased from the American Type Culture Collection (ATCC). Cells were resuscitated from liquid nitrogen storage. They were cultured in Corning T-75 flasks with growth medium and subculture once when 70-80% confluence was reached. Cell culture and treatment media for cell culture was prepared using Roswell Park Memorial Institute (RPMI)-1640 (R5886 Sigma Aldrich), fetal calf serum (FCS; 02-00-850 First Link ltd., 10%), Penicillin-Streptomycin-Glutamine (PSG; G1146 Sigma Aldrich, 1%).

A549 and HCT116 cells were plated (6000 cells/well) to 96 well plates in $100\mu L$ growth medium excluding outer wells to avoid edge effect. Molecules source are as follows: Temozolomide (T2577 Sigma-Aldrich), Rucaparib camsylate (PZ0036 Sigma-Aldrich), Sorafenib (SML2653 Sigma-Aldrich), Mitomycin (M0440 Sigma-Aldrich; 51854, Cell Signalling Technology), Sunitinib malate (PZ0012, Sigma Aldrich), Bosutinib (PZ0192, Sigma Aldrich), Midostaurin hydrate (M1323, Sigma Aldrich), Nintedanib (S1010, Selleckchem), Brassinolide (S3882, Selleckchem), Dactolisib (S1009, Selleckchem), Linoelaidic acid (90160, Cayman

Chemical), and Galangin (92342, Sigma Aldrich).

Molecules preparation. Molecules were dissolved in dimethyl sulfoxide (DMSO; D8418, Sigma Aldrich) to prepare different concentration stocks for each molecule. Dilutions were prepared using IC50 values obtained experimentally in the laboratory for each compound. For combinatorial experiments, each molecule was added with $50\mu L$ on top of the $100\mu L$ of media, thereby leading to a total assay volume of $200\mu L$ to achieve the desired drug concentrations with the treatment time for 48 hours.

Combinations tested. For combinations, cells were treated with a 5 by 5 matrix of drug concentrations. See Table 4.1 for details on combinations tested and replicates.

Crystal violet assay. Concentrations were applied using a full matrix approach and cells were incubated with the combination treatment for 72 Hr. In all drug treatments final DMSO concentration was kept at a maximum of 0.5% v/v. DMSO was added to all wells, including controls. For both monotherapy and combination therapy, cell death was quantified through crystal violet assay, staining proteins and DNA of cells retaining adherence and therefore still viable. We followed the protocol in Feoktistova et al[206]. After removing media containing detached dead cells, crystal violet was added to the wells (0.5%m/v, MeOH 25% v/v) and incubated for 20 min at room temperature on a bench rocker with a frequency of 20 oscillations per minute. We then washed the plates four times in a stream of tap water and let them the plates air-dry for at least 2 hours at room temperature. We then added $200\mu L$ of methanol to each well (10% v/v) incubated the plates for 20 minutes to solubilize crystal violet. Absorbance was read at 570 nm using a Tecan Sunrise plate reader.

Calculation of synergy. In our experiments, technical replicates were employed to select the replicate with highest precision. The data from these technical replicates was manually assessed and the replicate with the least visible failures was selected, that is, the replicates with the least saturated or blank wells. Each selected biological replicate was analyzed separately using SynergyFinder 3.0[207]. We fitted a 4-parameter logistic regression curve to the dose-response data. SynergyFinder merges biological replicates data and computes average ZIP synergy values. We include all synergy results in Appendix 8.2.

In the comparison between predicted and experimentally measured synergy values, a direct statistical comparison to establish significance is not feasible due to the uneven distribution of replicates available for study—20 for predicted synergy and merely 2 or 3 for

experimentally measured synergy. This significant disparity in the number of replicates hinders the application of standard statistical tests for determining the significance of differences between these values. Consequently, we adopt a trend-based approach rather than focusing on exact numerical values. We consider synergy values below -10 as antagonistic, those between -10 and 10 as additive, and values above 10 as synergistic. This classification provides a practical framework for interpreting the data, emphasizing the broader trends observed in the interactions under study, while circumventing the limitations posed by the small sample size in certain measurements. Thus, our comparisons prioritize the identification of general patterns over the precision of individual numerical comparisons, aligning with the practical constraints and the overarching goals of the research.

Table 4.1: Predicted combinations tested in A549 and HCT116

Cell line	Molecule 1	Molecule 2	Technical replicates	Biological replicates
A549	Temozolomide	Rucaparib	3	2
	Sunitinib	Bosutinib	3	2
	Midostaurin	Sunitinib	3	3
	Mitomycin	Sorafenib	3	3
	Sunitinib	Rucaparib	3	2
	Sunitinib	Brassinolide	3	3
	Sunitinib	Linoelaidic acid	3	3
	Dactolisib	Galangin	3	3
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

4.8 Results

Using learnable information propagation in the multiscale interactome accurately predicts drug-drug interactions in cell lines. Our proposed model predicts phenotype-specific PD drug-drug interactions 55% more effectively than traditional neural network models by using diffusion across the multiscale interactome (ChebNet-P vs MLP: MSE 19.11 vs 28.23; Pearson correlation 0.68 vs 0.44; Figure 4.4A). The proposed model's improvement

is especially pronounced relative to standard machine learning models, where it surpasses the Elastic-Net, Lasso, and Random Forest models by 74% (Pearson correlation). We note that performance is almost identical with the molecular-scale and multiscale interactomes (ChebNet-P vs ChebNet: MSE 19.11 vs 18.41; Pearson correlation 0.68 vs 0.70; Figure 4.4A). Model performance is relatively consistent across cell lines with only a handful of cell lines showing poor convergence (Figure 4.4B).

Our model discovers proteins and biological functions responsible for the drug-drug interactions predicted. Learnable information propagation across the multiscale interactome models effects that drug combinations and phenotypes exert through both proteins and biological functions, modeling the range of PD interactions that rely on both. We seek to determine whether the model has learned the drug-drug interacting logic behind PD interactions leading to drug-drug synergy/antagonism. Given our proposed model ChebNet-P, we investigate attributions to proteins and biological functions across all samples in the hold-out test set. In Figure 4.4D, top attributed proteins and biological functions are shown. Individual node size corresponds to the relative discriminating capacity of a given protein/biological function. A high proportion of discriminating genes have been documented as cancer-driving genes by the Cancer Gene Census initiative (genes highlighted in blue in Figure 4.4D). These genes have documented activity relevant to cancer, along with evidence of mutations in cancer which change the activity of the gene product in a way that promotes oncogenic transformation. Top predictive biological functions participate in processes related to cell replication and apoptosis. We include in Supplementary files the attribution scores for all proteins and biological functions in our dataset.

Our model offers insights into mechanisms responsible for drug-drug interactions. Diffusion profiles offer further insight into the potential mechanisms behind drug-drug interactions predicted. Consider the combination of Sorafenib and MK-2206 with a measured synergy of 12.51 in cell line A2058. We show in Figure 4.4E a network with top-scored nodes in diffusion profiles computed by our model ChebNet-P. A2058 is a cell line isolated from a lymph node metastasis from a 43 year old male patient with malignant melanoma. A2058 is characterized by BRAF and MAPK over-activation[208,209] as well as over-activation of the PI3K-AKT pathway[210]. Sorafenib is a kinase inhibitor used in the treatment of unre-

A

Model	PPI	
	MSE	Pearson's r
ChebNet-P	19.11±0.88	0.68±0.02
GCN-P	21.78±0.54	0.61±0.01
ChebNet	18.41±0.49	0.70±0.00
GCN	22.03±0.85	0.61±0.02
MLP	28.23±0.12	0.44±0.00
Linear–Elastic-Net	29.53±0.06	0.39±0.00
Linear–Lasso	29.54±0.07	0.39±0.00
Random Forest	29.97±0.08	0.39±0.00

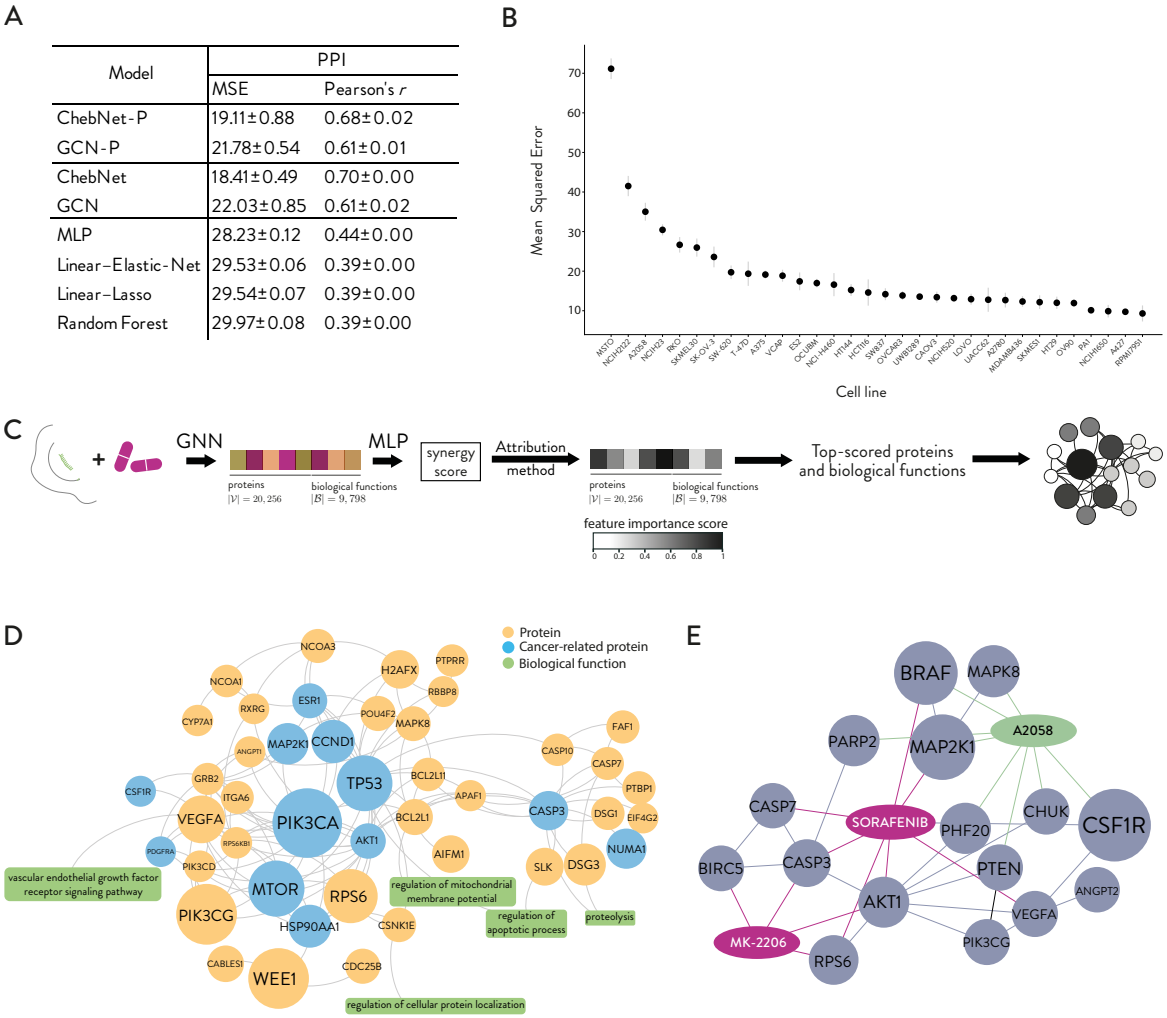


Figure 4.4: Our model predicts drug-drug synergy accurately and is interpretable. **(A)** Mean Squared Error (MSE), and Pearson correlation on the drug-drug ZIP synergy prediction task. Average performance values on held-out test set for the 5-fold cross-validation splits used for model hyperparameter tuning. **(B)** Test Mean Squared Error across 5 folds for each cell line showing that model performance remains consistent across cell lines with only a handful of cell lines showing poor convergence. **(C)** For all samples in the test set, we obtain predictive importance scores of proteins and biological functions across 5 folds and average them (see *Methods*). Then, we draw a network with the proteins and biological functions most relevant for model predictions. **(D)** Shown are the top predictive proteins and biological functions in our proposed ChebNet-P model. Proteins and biological functions that most contribute to the predictive power of our proposed model are related to cancer and cell growth. Node size corresponds to the relative discriminating capacity of a given protein and biological function. Node color illustrates the node type. **(E)** Sorafenib and MK-2206 were predicted with a synergy score of 12.5. Their synergistic interaction could be explained by Sorafenib's targeting of multiple intracellular and cell surfaces kinases, which reduces blood flow to the tumor, and by MK-2206's inhibition of AKT, which helps sensitize tumors to Sorafenib.

sectable liver carcinoma and advanced renal carcinoma[88]. MK-2206 is an experimental drug being investigated for the treatment of several types of cancers. Both drugs are currently in clinical trials, alone or associated with other drugs, for the treatment of melanoma[211,212]. Sorafenib's activity against A2058 could be explained by its interaction with multiple intracellular and cell surface kinases (including BRAF and VEGFA) which reduces blood flow to the tumor[88]. However, some tumors enhance their PI3K/AKT activity in order to reduce the efficacy of Sorafenib against cancer cell survival[213]. Considering this, MK-2206, an AKT inhibitor[214,215], can help resensitizing Sorafenib-resistant tumors, potentially explaining the synergistic relationship found between Sorafenib and MK-2206 in A2058.

Experimental validation of drug-drug interactions in A549. We used our proposed model ChebNet-P to predict synergy values for 17,766 new drug-drug combinations against A549, a cell line model to study non-small cell lung cancer (NSCLC; see subsection "Ensembling to predict novel drug-drug and food molecule-drug synergies" in the *Methods* section for details). 17,759 of combinations were predicted with additive interaction (synergy value within [-10,10]), 6 with synergistic (synergy value $>$ 10) and 1 with antagonistic (synergy value $<$ -10) interaction (Figure 4.5A). We also note that the average standard deviation is 2.45, with 95% of predictions having a standard deviation ≤ 4. Interestingly, out of the 171 combinations with standard deviation ≥ 7 (including the distinct cluster in Figure 4.5A), 169 of those combinations include the drug Histamine. To identify why predictions for combinations with this drug might have unusually high standard deviations, we investigate the number of drug targets. We find that, on average, drugs in the training set have 63 targets, with a maximum of 279 targets whereas Histamine has 443 targets. This large difference in the number of targets can explain the large standard deviation, resulting from Histamine being out of distribution compared to the training dataset. We filtered combinations based on synergy value and availability and assessed their synergy interaction score. We tested 4 additive and 1 antagonistic combination, achieving consistency with predicted synergy values in 4 out of 5 combinations (Figure 4.5B). We highlight 2 use cases and explore potential mechanisms of action.

Consider the combination of Rucaparib and Sunitinib, with a predicted synergy value of 3.85 ± 3.42 against A549 cells and an experimentally measured synergy value of 4.84 ± 1.59 (Figure 4.5C). Sunitinib is a promising FDA-approved tyrosine kinase inhibitor that has

been evaluated in patients with metastatic NSCLC[216]. In the context of its interaction with Rucaparib, it acts as the primary agent against A549 cells, whereas Rucaparib acts as an adjuvant (Figure 4.5D). In this context, Sunitinib's activity against A549 cells is driven by its effects on STK11/LKB1, a kinase that phosphorylates AMPK and triggers essential adaptations to subvert metabolic stress and promote tumor cell survival[217,218]. One of these metabolic adaptations is glycolysis enhancement, a response previously observed in both NSCLC cell lines and human tissue samples, where the circular RNA MYLK (circMYLK) enhances glucose consumption and lactate production[219]. The common MYLK inhibition by both drugs[220,221] synergizes with Rucaparib's immunomodulatory effects against PARP1 in the NSCLC cell line A549[222]. Rucaparib, as PARP1 inhibitor, promotes glycolysis-induced epithelial-mesenchymal transition (EMT)[223,224]. The EMT and metabolic disruption induced by Rucaparib and Sunitinib leads to a synergistic apoptotic induction, where Sunitinib activates the general apoptosis inducer caspase-3[225] whose active zymogen can cleave to the Rucaparib target PARP1 contributing to mitochondrial instability, increasing the apoptosis-inducing factor (AIF) release[226] which promotes the apoptotic response even further.

Now consider the combination of Sunitinib and Bosutinib, with a predicted synergy value of 6.61 ± 4.11 and an experimentally measured synergy value of 8.62 ± 1.47 (Figure 4.5E). The additive interaction between Sunitinib and Bosutinib in A549 could be explained as a result of the inhibition of cancer cell proliferation by the suppression of up- and downstream regulators of growth factor signaling, and apoptosis induction (Figure 4.5F). This cancer cell survival/death imbalance induction can be explained by the antagonistic effects of both agents against the Src-family kinase Fgr, a recognized proto-oncogene involved in cell proliferation, differentiation, migration, tumor microenvironment interactions, and inflammatory response regulation[227,228]. This interaction is complemented with the inhibition of EGFR-related pro-survival pathways by Bosutinib, and with the inhibition of PTK2B/Pyk2, a downstream regulator of EGFR response linked to lung cancer survival and proliferation[229], by Sunitinib and Bosutinib. This synergistic inhibition of EGFR signaling could eventually lead to cancer resistance. This phenomenon could be promoted by the non-receptor tyrosine kinase FRK, a regulatory protein that can induce EGFR intracellular phosphorylation and enhance a phenotypical plasticity in NSCLC[230,231]. Fortunately, these processes could be interrupted by Bosutinib[220,221], suppressing the FRK role in EMT and metabolic adaptation

to growth factor suppression. MAP3K serine/threonine kinases are common downstream effectors of the aforementioned EGFR-related pathways. Bosutinib can inhibit MAP3K2[232], a protein associated with cell proliferation, migration, and invasion in A549, triggering apoptosis[233,234]. In addition to this proapoptotic environment triggered by Fgr, FRK, EGFR, Pyk2 and MAP3K2 inhibition, Sunitinib and Bosutinib also inhibit MAP3K7 function, promoting proteasomal degradation, and apoptosis induction[235]. Finally, apoptotic mediators are also targeted by this synergic association through the induction of the executor caspase-3 activity by Sunitinib[225].

Experimental validation of drug-drug interactions in HCT116. We used our proposed model ChebNet-P to predict synergy values for 17,668 new drug-drug combinations against HCT116, a cell line model to study colorectal cancer (CRC; see subsection "Ensembling to predict novel drug-drug and food molecule-drug synergies" in the *Methods* section for details). 17,506 of combinations were predicted with additive interaction (synergy value within [-10,10]), and 162 with synergistic (synergy value > 10) interaction (Figure 4.8A). We also note that the average standard deviation is 2.35, with 94% of predictions having a standard deviation ≤ 4. Interestingly, out of the 192 combinations with standard deviation ≥ 7 (including the distinct cluster in Figure 4.8A), 186 of those combinations include the drug Histamine. To identify why predictions for combinations with this drug might have unusually high standard deviations, we investigate the number of drug targets. We find that, on average, drugs in the training set have 63 targets, with a maximum of 279 targets whereas Histamine has 443 targets. This large difference in the number of targets can explain the large standard deviation, resulting from Histamine being out of distribution compared to the training dataset. We filtered combinations based on synergy value and availability and assessed their synergy interaction score. We tested 1 additive and 5 synergistic combinations, achieving consistency with predicted synergy values in 4 out of 6 combinations (Figure 4.8B). We highlight 2 use cases and explore potential mechanisms of action.

Consider the combination of Sunitinib and Nintedanib, with a predicted synergy value of 13.85 ± 4.52 and an experimentally measured synergy value of 13.99 ± 4.08 (Figure 4.8C). Sunitinib, an approved RTK inhibitor for metastatic renal carcinoma[236], is also used to treat gastrointestinal stromal tumor and pancreatic neuroendocrine tumors. Due to the poor effectiveness in colorectal cancer monotherapy, Sunitinib is not used in patient treat-

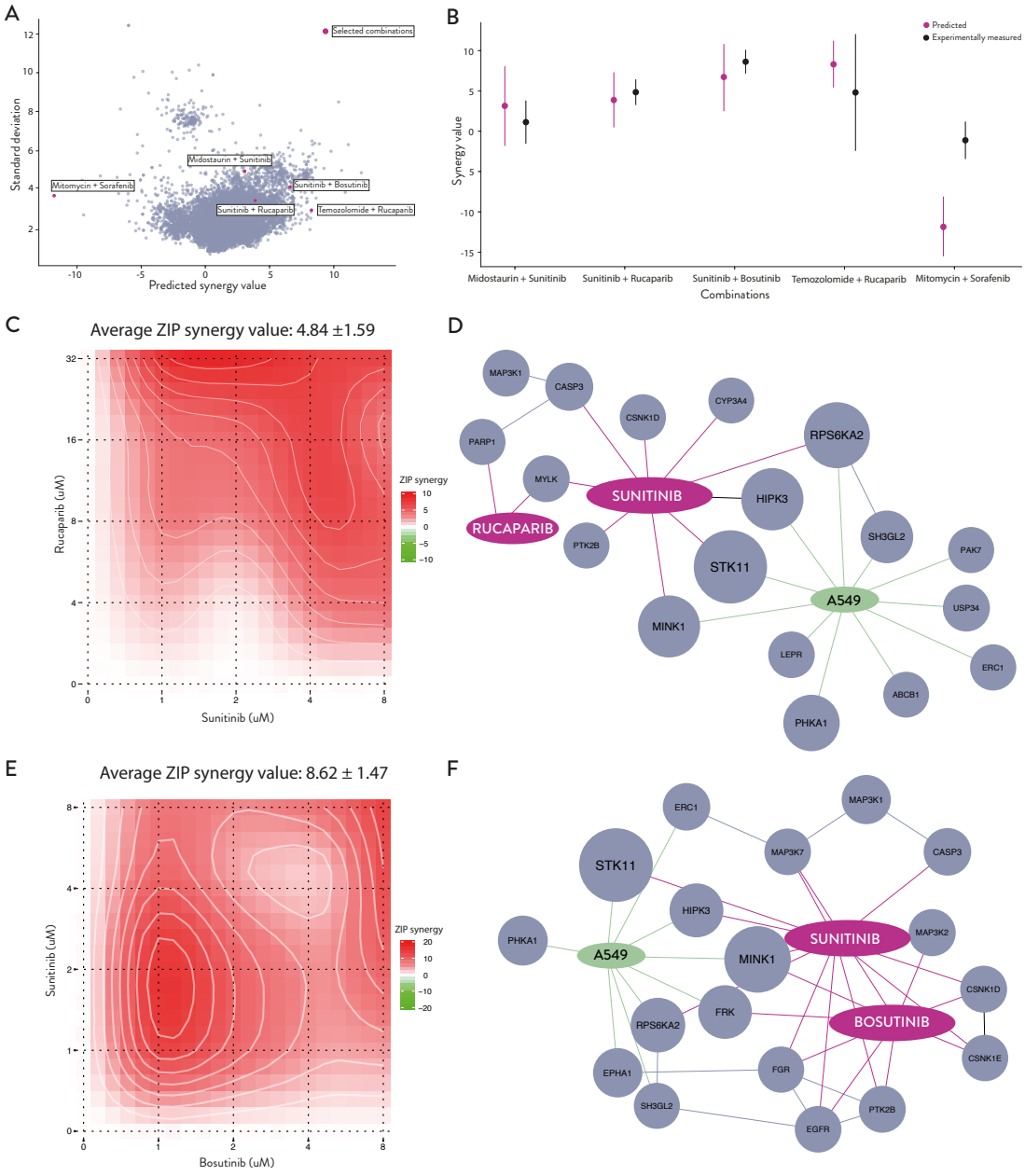


Figure 4.5: Experimental validation of drug-drug interactions in A549. (A) We predicted synergy values for 17,766 new drug-drug combinations against A549 (see *Methods*). Shown are the combinations we subsequently tested experimentally. **(B)** Shown are predicted and experimentally measured synergy values for five combinations of which four were consistent with predicted results. **(C)** Shown is the synergy map of Rucaparib and Sunitinib computed using SynergyFinder. Rucaparib and Sunitinib were predicted with a synergy score of 3.85 ± 3.42. Experimental testing showed a synergy of 4.84 ± 1.59. **(D)** Network representation of the top-scored proteins in diffusion profiles computed by our model CheNet-P for the prediction of synergy score between Rucaparib and Sunitinib. Sunitinib's activity against A549 cells is driven by its effects on STK11/LKB1, a kinase that phosphorylates AMPK and triggers essential adaptations to subvert metabolic stress and promote tumor cell survival[217,218]; whereas Rucaparib acts as an adjuvant. Their combined effects lead to epithelial-mesenchymal transition (EMT) and metabolic disruption resulting in a synergistic apoptotic induction. We include in Supplementary files the embedding scores for all proteins and biological functions in our dataset. **(E)** Shown is the synergy map of Sunitinib and Bosutinib computed using SynergyFinder. Sunitinib and Bosutinib were predicted with a synergy score of 6.61 ± 4.11. Experimental testing showed a synergy of 8.62 ± 1.47. **(F)** Network representation of the top-scored proteins in diffusion profiles computed by our model CheNet-P for the prediction of synergy score between Sunitinib and Bosutinib. The additive interaction between Sunitinib and Bosutinib in A549 could be explained as a result of the the proapoptotic environment triggered through the inhibition of Fgr, FRK, EGFR, Pyk2 and MAP3K2. Sunitinib and Bosutinib also inhibit MAP3K7 function, promoting proteasomal degradation, and apoptosis induction. We include in Supplementary files the embedding scores for all proteins and biological functions in our dataset.

ment, nevertheless it has an acceptable safety profile[237,238]. Nintedanib, another approved RTK inhibitor for idiopathic pulmonary fibrosis and non-small-cell lung carcinoma[239], was shown to be a potent drug in *in vitro* and *in vivo* studies[240]. Furthermore, Nintedanib was shown to have similar safety profile with tolerable adverse effects in a phase II study for advanced renal cell cancer patients[241]. However, Nintedanib also did not meet a satisfactory primary end points in a Phase III study of chemo-refractory metastatic CRC, despite having a tolerable safety profile[242]. The synergistic interaction between Nintedanib and Sunitinib against HCT116 cells could be explained through their targeting of the FMS-like tyrosine kinases isoforms 3 and 4, FLT3 and FLT4/VEGFR-3, respectively (Figure 4.8D). These receptor protein kinases, often found mutated in CRC[243–245], are involved in cellular responses relevant in cancer, including cell proliferation, apoptosis suppression, and cell migration and invasion. The antagonistic effects of Nintedanib and Sunitinib against FLT3 and FLT4[246–248] lead to suppressive effects against other targets, including the oncogene and serine/threonine kinase MELK, and the fibroblast growth factor receptor 2 (FGFR2). These combined effects of Nintedanib and Sunitinib suppress relevant functions in CRC related with cell survival, tumor migration/invasion, and phenotypical plasticity[249–252]. Analyzing downstream mediators involved in the aforementioned cellular processes, the inhibitory action of Nintedanib and Sunitinib against TNIK emerges as another target that synergizes the effects on CRC stemness. TNIK blockage impairs EMT progression in both in vivo and in vitro CRC models[253]. These synergistic effects on phenotypical plasticity lead to a decrease of migration and invasion potential of cancer cells, fundamental processes in cancer progression and metastasis. Regarding cancer migration and invasion, these drugs also impair the function of the proto-oncogene MERTK, a relevant mediator triggering mutated-Ras/IL-33-induced tumor cell migration[254,255]. Nintedanib also inhibits the discoidin domain receptor 2 (DDR2), a receptor tyrosine kinase associated with the fibrillary collagen disposition in mesenchymal tissues and micro-metastasis formation. Altogether, these inhibition patterns generate a suppression link between the migration/invasion potential and EMT progression in CRC[256,257]. Lastly, the inhibitory effect of Nintedanib and Sunitinib against the ABL1/c-Abl non-receptor tyrosine kinase and the effect of Sunitinib against HIPK2 lead to an inhibition of cell proliferation and cell cycle progression in human colorectal cancer cell lines and patient-derived biopsies[258–260]. These effects, together with the downstream

inhibition of MAP3K1 and MAP3K2, highlight the synergistic tumor suppressive effect of Nintedanib and Sunitinib against CRC[261].

Now consider the combination of Sorafenib and Bosutinib, with a predicted synergy value of 11.67 ± 3.03 and an experimentally measured synergy value of 5.23 ± 2.54 (Figure 4.8E). We observe the discrepancy between the predicted and experimentally measured synergy score. We also observe that the overlap of cancer-promoting kinases and receptors which Sorafenib and Bosutinib target in this cell line (Figure 4.8F) is smaller compared to the previous drug pair (Sunitinib and Nintedanib). Nonetheless, this drug pair is still borderline synergistic in the inhibition of HCT116. The synergistic interaction between Sorafenib and Bosutinib against HCT116 cells could be attributed to a general multi-pronged inhibition of cell survival pathways in CRC. These agents can act as antagonists of FLT3, a receptor tyrosine kinase overexpressed in 5.5% of CRC patients[262]. Sorafenib enhances these effects by inhibiting FLT4/VEGFR-3[263], both essential receptors for cell proliferation induction in CRC progression. These pharmacological effects synergize with the antagonistic effect of Sorafenib against DDR2, a biomarker of cell proliferation, CRC staging and prognosis[264,265], and the inhibitory effects of Bosutinib against MERTK, a receptor associated with cell proliferation, migration, and EMT in different types of cancer[254,255]. These inhibitory effects on different cancer-related receptors are linked to the antagonistic activity of Sorafenib and Bosutinib against the intracellular kinases ABL1 and ABL2, proteins strongly associated with cell survival, cell adhesion promotion, and metastatic progression in different types of cancer[266,267]. Furthermore, these effects against CRC survival are also promoted by Bosutinib's inhibition of downstream mediators like MAP3K2 and MELK, whose overexpression in CRC is associated with cell proliferation, tumor growth, and cell cycle and metastatic progression[251,268–270]. This cascade of survival response inhibition is further affected by Sorafenib's promotion of the activity of executor caspases 3/7, essential for apoptosis progression[271], and Sorafenib's inhibition of BIRC2/cIAP1, a suppressor of TNF -induced pro-apoptotic effects that is frequently overexpressed in CRC[272,273].

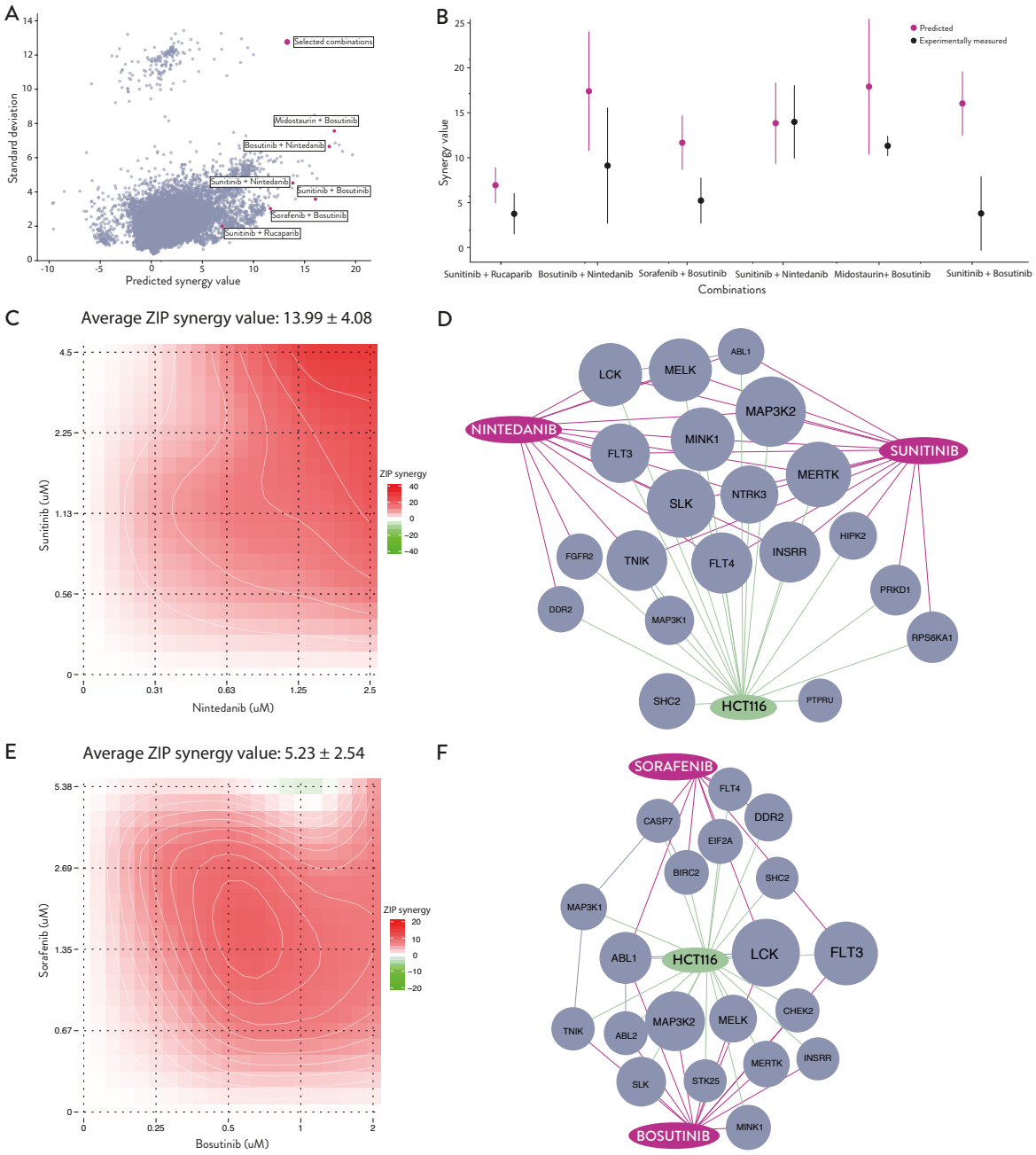


Figure 4.6: Experimental validation of drug-drug interactions in HCT116. **(A)** We predicted synergy scores for 17,668 new drug-drug combinations against HCT116 (see *Methods*). Shown are the combinations we subsequently tested experimentally. **(B)** Shown are predicted and experimentally measured synergy values for six combinations, of which four were consistent with predicted results. **(C)** Shown is the synergy map of Sunitinib and Nintedanib computed using SynergyFinder. Sunitinib and Nintedanib were predicted with a synergy value of 13.85 ± 4.52. Experimental testing showed a synergy of 13.99 ± 4.08. **(D)** Network representation of the top-scored proteins in diffusion profiles computed by our model CheNet-P for the prediction of synergy score between Sunitinib and Nintedanib. Their synergistic interaction could be explained through their targeting of multiple genes involved in cancer-related processes including FLT3, FLT4, MELK, FGFR2, TNIK, MERTK, DDR2, and HIPK2. This leads to a reduction in cell proliferation, cell survival, cell cycle progression, and cell migration and invasion. We include in Supplementary files the embedding scores for all proteins and biological functions in our dataset. **(E)** Shown is the synergy map of Sorafenib and Bosutinib computed using SynergyFinder. Sorafenib and Bosutinib were predicted with a synergy value of 11.67 ± 3.03. Experimental testing showed a synergy of 5.23 ± 2.54. **(F)** Network representation of the top-scored proteins in diffusion profiles computed by our model CheNet-P for the prediction of synergy score between Sorafenib and Bosutinib. The additive interaction between Sorafenib and Bosutinib against HCT116 cells could be attributed to a general multi-pronged inhibition of cell survival pathways in CRC through their antagonistic effects on FLT3, FLT4/VEGFR-3, DDR2, MERTK, ABL1, ABL2, MAP3K2, and MELK. We include in Supplementary files the embedding scores for all proteins and biological functions in our dataset.

Prediction and validation of phenotype-specific food molecule-drug interactions in A549. We used our proposed model ChebNet-P to predict synergy values for 122,655 new food molecule-drug combinations against A549, a cell line model to study non-small cell lung cancer (NSCLC; see subsection "Ensembling to predict novel drug-drug and food molecule-drug synergies" in the *Methods* section for details). 120,843 combinations were predicted with additive interaction (synergy value within [-10,10]), 1,811 with synergistic (synergy value > 10) and 1 with antagonistic (synergy value $<$ -10) interaction (Figure 4.7A). We also note that the average standard deviation is 2.48, with 89% of predictions having a standard deviation ≤ 4. We filtered combinations based on synergy value and availability and assessed their synergy interaction score. We tested 1 additive and 2 synergistic combination, achieving consistency with predicted synergy values in all combinations tested (Figure 4.7B). We highlight one use case and explore potential mechanisms of action.

Consider the combination of Sunitinib and Brassinolide, with a predicted synergy value of 4.40 ± 3.53 and an experimentally measured synergy value of 3.86 ± 3.34 (Figure 4.7C). The additive interaction between Sunitinib and Brassinolide in A549 cells could be explained by their combined inhibition of different cell processes such as metabolic adaptation, cell growth, cell cycle progression, and cell survival (Figure 4.7D). Regarding metabolic adaptation and cell growth promotion, Sunitinib inhibits STK11/LKB1, reducing the glucose, lipid and protein biosynthesis induced by AMPK phosphorylation and mTOR activation[217]. These effects on the AMPK/mTOR axis are enhanced by Sunitinib's effect against HIPK3, which reduces the anabolic/pro-autophagic profile that this pathway promotes in NSCLC[274,275]. Regarding cell cycle progression and tumor cell survival, Sunitinib inhibits MINK1, a kinase involved in the downregulation of the canonical tumor suppressor gene p53 via AKT/MDM-2 induction, leading to cell cycle arrest and apoptosis induction[276]. This later response is promoted also by Sunitinib's inhibition of relevant downstream pro-survival regulators including MAP3K1 and RPS6KA2/RSK3, leading to a general impairment of the Ras/Raf/MEK/ERK pathway and its subsequent effectors[235,277,278]. These changes drive cancer cell fate balance towards proapoptotic conditions, where the synergistic mechanism between Sunitinib and Brassinolide takes place. This apoptotic induction is reinforced by Sunitinib via CSNK1D/CK1 and CSNK1E/CK1 inhibition, which reduces Bid phosphorylation and leads to the subsequent promotion of the Fas-mediated extrinsic pathway of apoptosis[279]. This extracellular activa-

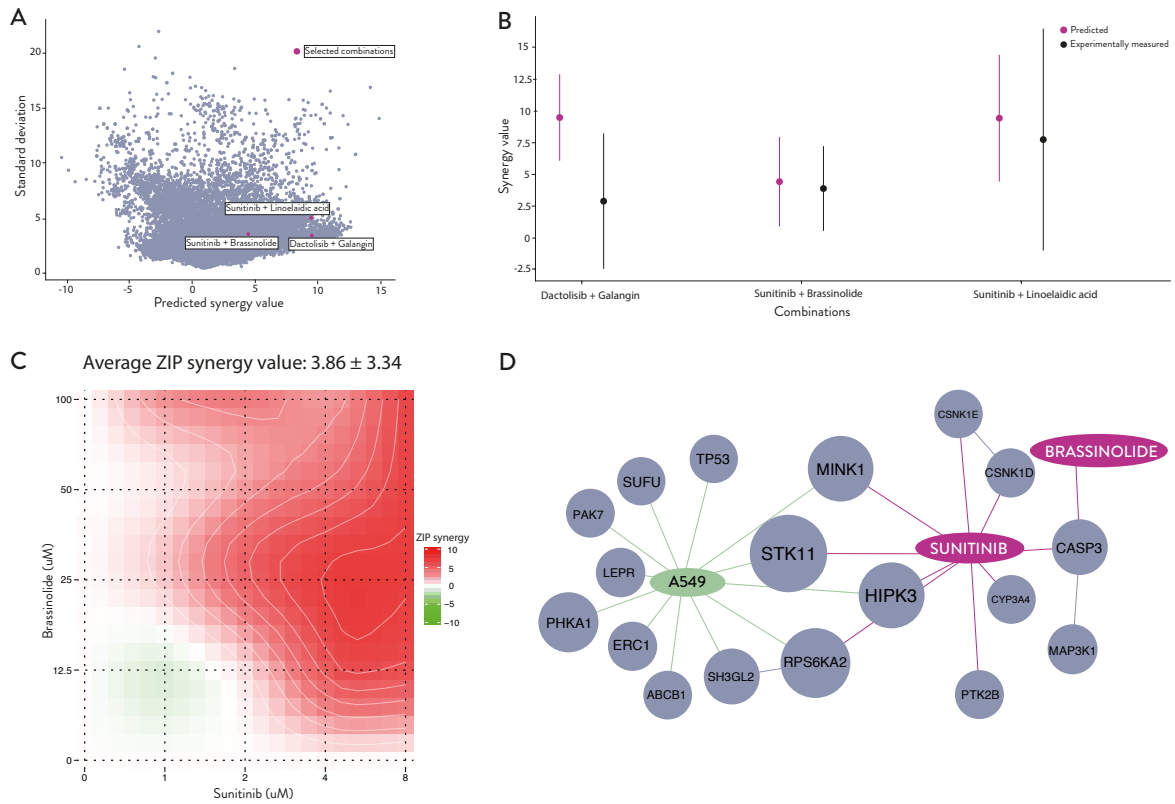


Figure 4.7: Experimental validation of food molecule-drug interactions in A549. (A) We predicted synergy scores for 122,655 new food molecule-drug combinations against lung cancer (cell line A549; see *Methods*). Shown are the combinations we subsequently tested experimentally. **(B)** Shown are predicted and experimentally measured synergy values for 3 combinations chosen based on availability and synergy scores. All combinations tested were consistent with predicted results. **(C)** Shown is the synergy map of Sunitinib and Brassinolide computed using SynergyFinder (see *Methods*). Sunitinib and Brassinolide were predicted with a synergy value of 4.40 ± 3.53. Experimental testing showed a synergy of 3.86 ± 3.34. **(D)** Network representation of the top-scored proteins in diffusion profiles computed by our model CheNet-P for the prediction of synergy score between Sunitinib and Brassinolide. Their additive interaction against A549 cells could be explained by their combined inhibition of different cell processes such as metabolic adaptation, cell growth, cell cycle progression, and cell survival. Sunitinib induces extrinsic apoptotic pathways through its inhibition of CSNK1D/CK1 and CSNK1E/CK1. Brassinolide induces reactive oxygen species (ROS) production which leads to apoptosis induction via intrinsic pathway induction.This activation of both extrinsic and intrinsic apoptotic signaling pathways converges in the activation of executor caspases 3/7. We include in Supplementary files the embedding scores for all proteins and biological functions in our dataset.

tion of apoptosis is synergized by the effects reported for brassinosteroids like Brassinolide on NSCLC cell models like A549 cells, where these compounds induce reactive oxygen species (ROS) production. This leads to apoptosis induction via intrinsic pathway induction by triggering a proapoptotic imbalance of Bcl-2-family proteins[280,281]. This activation of both extrinsic and intrinsic apoptotic signaling pathways converges in the activation of executor caspases 3/7, where Sunitinib and Brassinolide show another synergistic effect by promoting the activity of these later-stage zymogens[282,283].

4.9 Summary

In this Chapter, we introduce an approach to systematically investigate and predict phenotype-specific pharmacodynamic (PD) interactions between food molecules and drugs. PD interactions can positively or negatively influence pharmacological effects of drugs offering an opportunity for improved drug efficacy upon lower concentrations and hence lower toxicity. Whereas PD drug-drug interactions have been extensively studied, knowledge of PD food molecule-drug interactions is limited mainly due to the lack of in-vitro studies. Here, we propose a graph neural network (GNN)-based model for the task of PD food molecule-drug interaction prediction. We first calibrate this model using a publicly available dataset of in-vitro PD drug-drug interactions, and then transfer this model for the task of predicting PD food molecule-drug interactions.

Our model leverages learnable diffusion across a network representing proteins and biological functions to learn effects of molecules and disease (cancer) in the form of diffusion profiles. We use diffusion profiles to predict PD drug-drug interactions, achieving up to 55% higher performance over traditional deep learning models (Figure 4.4A).

Importantly, model predictions become transparent using attribution methods. We showed that the proposed model identifies proteins and biological functions related to cancer and cell growth as the most relevant for the prediction of PD interactions (Figure 4.4D). For individual use cases, diffusion profiles computed by our model facilitates obtaining insights about the intracellular mechanisms that could lead to the predicted PD interactions (Figure 4.4E; Figure 4.5D,F; Figure 4.8D,F, Figure 4.7D). Experimental validation of PD drug-drug interactions showed consistency with predicted results in 8 out of 11 combinations tested, further validating our approach (Figure 4.5B, Figure 4.8B). We then used our proposed model to predict PD food molecule-drug interactions. Selected food molecule-drug interacting pairs were validated experimentally, showing consistency in all 3 combinations tested. However, it is worth noting the high standard deviation in experimental results in 2 out of the 3 combinations tested which might point to the need of additional confirmatory experiments.

The work in this Chapter offers the first systematic approach to predict PD food molecule-drug interactions in a supervised setting to prioritize experimental testing of combinations. The somewhat high variance in model predictions for the set of new drug-drug and food

molecule-drug combinations highlights that exact prediction scores should be considered with caution and that experimental tests are indeed needed to validate predictions. Future research directions could consider adding an active learning component to further calibrate the model based on experimental results. Extensive validation of interacting combinations and the creation of an interaction map between all molecules within foods and drugs could further orient diet approaches accompanying therapeutics in practice.

Our work adopts a well-established approach for DL-based PD interaction prediction, i.e., first obtaining representations of molecules and disease and then passing them through a neural network for the prediction of a phenotype-specific synergy score. Existing works focus on PD drug-drug interaction prediction, where extensive heterogeneous datasets exist including protein targets, drug chemical fingerprints, drug physico-chemical properties, and gene expression upon drug treatment[187–195]. Given that our goal is the prediction of PD food molecule-drug interactions, we could only leverage protein targets to build our model, as this is the only widely available data type for food molecules. Hence, our challenge was to exploit protein target data and disease genes, a significantly sparse data type, to predict PD interactions. Our initial analyses revealed that drugs might interact in a given context (disease) when their protein targets are closer than expected by chance to disease genes in a network of protein-protein interactions (Figure 4.1A). Our analyses also showed that 97% of interacting drugs affect the same biological functions as mutated genes of cell lines where they interact more often than is expected by chance (Figure 4.1B). Considering these findings, we leveraged GNNs to propagate the information represented by drug targets and disease genes across a heterogeneous network, generating diffusion profiles that represent effects of drug and disease on proteins and biological functions controlled by network proximity of each protein and biological function to proteins targeted by drugs and genes mutated in disease. However, we note that our proposed model achieved almost identical performance when using a multiscale and molecular-scale interactome (ChebNet-P vs ChebNet: MSE 19.11 vs 18.41; Pearson correlation 0.68 vs 0.70; Figure 4.4A). This is possibly be due to the fact that 99% of interacting drug-drug combinations target proteins that are closer than expected by chance to disease-associated genes, which can be captured in molecular-scale diffusion profiles. Using diffusion profiles computed by our GNN model, we achieved considerably higher performance than traditional DL algorithms (Figure 4.4A), highlighting the utility

of learnable network propagation to amplify sparse biological signal for the prediction of phenotype-specific PD interactions.

This capability is relevant not only in the search for synergistic molecules with existing therapeutics but also in the context of clinical trials. The potential for PD interactions between food molecules and drugs to influence drug efficacy and toxicity raises the possibility that certain drug trials may have encountered varied outcomes due to uncontrolled dietary factors. Compounds present in food can modulate drug metabolism, absorption, and biological response, which may synergize with or counteract the therapeutic effects of drugs[284]. This unaccounted variability could result in drug trials failing to meet their endpoints, not due to the inefficacy of the drug itself, but because of the concurrent consumption of incompatible foods. Hence, the integration of dietary control and monitoring into clinical trial protocols is imperative to ensure accurate assessment of a drug's performance, underscoring the necessity of personalized dietary guidelines in conjunction with pharmacological treatment to optimize therapeutic outcomes.

It is important to note the limitations of our study as well, the first related to the experimental evaluation of combinations. In the course of our experiments, while our goal was to maintain a standard of three biological replicates (n=3) for robust statistical analysis, we encountered practical challenges that led to a variation in the number of replicates. Some biological replicates failed, and due to time constraints, it was not feasible to repeat all experiments to consistently achieve n=3 across all test conditions. As a result, certain experimental conditions were analyzed with only two biological replicates (n=2). In the analysis of our results, we did not employ specific statistical methods to adjust for the varying number of biological replicates. While this is a deviation from the ideal experimental setup, it is a reflection of the real-world constraints encountered during the research process. We recognize that this limitation can impact the statistical power and confidence levels of our findings, particularly in cases where only two biological replicates were used. Future iterations of this research would benefit from repeating experiments to ensure a uniform number of biological replicates across all conditions, enhancing the reliability and validity of the findings.

A second limitation of our study is the lack of established synergistic pairs as positive controls. This omission was due in part to the exploratory nature of our model's predictions

and a preliminary oversight in not conducting an exhaustive literature search for known synergistic food molecule-drug interactions. Upon subsequent reflection, it is recognized that the integration of positive controls would have been advantageous, providing a robust validation framework for our model's predictions.

The literature pertaining to food molecule-drug synergies is still emerging, which may contribute to the challenge of identifying suitable positive controls. This scarcity highlights a gap in the current research landscape, emphasizing the novelty of our research. Nonetheless, it is acknowledged that a more thorough literature review could have potentially revealed documented interactions to be used as positive controls. Moving forward, this can be addressed by diligently identifying any such combinations that have been previously verified in similar cell lines. This would entail a comprehensive and systematic literature review and may involve the development of a curated database of known interactions, enhancing the experimental rigor of future studies.

5 A principled method for lead discovery

5.1 Abstract

Phenotype-driven drug discovery approaches have gained attention as an alternative to target-driven strategies in the search for new therapeutics. These approaches aim to identify genetic interventions or chemical agents, also called perturbagens, that reverse the overall effects of diseases by leveraging phenotype signatures. The goal is, given a signature characterizing the effects of disease, to find a perturbagen reversing those effects. However, current methods are limited in their ability to predict arbitrary perturbagens because they rely on predefined compound libraries, restricting the search space. In this work, we propose a novel problem formulation for phenotype-driven lead discovery where the goal is to predict an arbitrary perturbagen, represented as a set of targets, to reverse the effects of disease. We borrow concepts from the causal inference field to formulate the task as an optimal perturbagen discovery problem where the desired perturbagen is that which maximizes the likelihood of shifting the system to a desired outcome. To solve this task, we introduce PDGrapher, a two-module approach comprising a perturbagen discovery module and a response prediction module, both based on graph neural networks. We evaluate PDGrapher in a random and leave-cell-out setting across four datasets of genetic and chemical interventions. PDGrapher is able to predict perturbagens that shift cell line gene expression phenotypes from an initial to a desired state for up to 12% more samples in the test set compared to the second-best-performing baseline. Additionally, ground-truth perturbagens are up to 35% higher in the ranking of perturbagens predicted by PDGrapher compared to the second best-performing baseline. Our results reveal the potential of reframing phenotype-driven lead discovery to expand the lead search space.

5.2 Background

The identification of new potential therapeutics against disease is a key task in drug discovery and development[285]. Target-driven approaches, which focus on designing maximally selective compounds against a disease-associated target, have dominated the drug discovery landscape for the past three decades[286]. In recent years, there has been a revival of alter-

native approaches that fall under the umbrella of phenotype-driven drug discovery. These approaches go beyond the paradigm of one drug, one gene, one disease that prevails in target-driven approaches by looking for compounds that reverse the overall effects of diseases as measured by various HTS technologies[287]. The goal is, given a *signature* characterizing the effects of disease, to find a *perturbagen* or *intervention* reversing those effects[285]. The process of identifying compounds with potential activity against disease is commonly known as *lead discovery* and serves as the initial stage in the extensive drug discovery and development pipeline[288–290]. Lead discovery provides the starting point: potential candidates to be tested and optimized against disease.

Phenotype-driven lead discovery emerged as a result of the compilation of large amounts of data describing biological systems generated by HTS technologies, in particular by the creation of the connectivity map (CMap[291]) and the Library of Integrated Network-Based Cellular Signatures (LINCS[292]). CMap and LINCS are large databases that contain gene expression profiles of dozens of cell lines treated with thousands of chemical compounds. CMap researchers hypothesized that new leads for a given disease could be discovered by finding compounds with gene expression signatures similar to those of known drugs that treat the disease. To realize this, they introduced the concept of the connectivity score, a metric that captures the similarity between the gene expression signatures of two given chemical compounds[293]. This connectivity metric was available through the CMap online portal, allowing researchers to query the CMap database with a given compound signature (not necessarily included within CMap) to find (dis)similar chemical compounds[285]. This concept gained widespread popularity, and researchers quickly proposed improvements to the connectivity score[294].

However, the applicability of these methods is limited to diseases with known therapeutics[295]. As a result, this concept was expanded to incorporate disease signatures. Typically, a disease expression signature is constructed by comparing the disease with healthy expression samples. Then, disease expression signatures are compared to compound-induced gene expression signatures to recover small molecules that reverse the disease signature, that is, compounds that increase the expression of down-regulated and decrease the expression of up-regulated genes in disease[296]. Studies have shown that the reversal of gene expression positively correlates with drug efficacy in vitro[296], and subsequent efforts have resulted in

the compilation of large repositories and web tools for in vitro compound screening[291–297]. This strategy has resulted in various successful therapeutic discoveries against new and understudied diseases[298–300].

Systematic evaluations have shown that gene expression signatures adequately represent chemical compounds, so similarities obtained using connectivity scores represent similarities according to the anatomical therapeutic chemical (ATC) classification of drugs[301–303]. However, further studies have shown that the power of chemical compound and disease signatures to predict existing drug-disease associations (Food and Drug Administration-approved drug indications) is limited[304]. Some argue that this could be solved by developing more powerful connectivity scores[304]. However, several other reasons could explain why the similarity between drug and disease signatures does not correlate well with existing drug indications, including the low quality of disease signatures and incomplete knowledge of drug-disease associations[304]. More important is the fact that drug signatures are computed using responses on an array of cancer cell lines rather than healthy tissue or the disease tissue itself, a factor that potentially explains the better performance of the connectivity score in neoplastic disease signatures[304].

In recent years, representation learning approaches to predict responses to untested perturbagens or perturbagen combinations have emerged[305–308]. These methods have facilitated lead discovery by predicting response to untested perturbagens, which increases the search space of connectivity score-based methods. However, existing methods cannot predict responses to an arbitrary, unseen perturbagen, represented simply as a set of targets. If our goal is to find new therapeutic leads, we argue that the current phenotype-driven approach, even when enhanced by representation learning, is limiting because it only allows for the prediction of compounds that exist within a library, reducing the potential search space. We propose a new problem formulation for phenotype-driven lead discovery in which the goal is, given a *signature* characterizing the effects of disease, to find what genes a new *perturbagen* should target to reverse those effects.

The problem of finding which elements of a system should be perturbed is typically referred to as *optimal intervention discovery* and has been addressed within the causal discovery literature in the context of *experimental design*. Experimental design refers to the problem of finding the minimum number of interventions necessary to recover the ground

truth graph that encodes causal relationships between elements of a system[309,311]. Other approaches to intervention design aim to find interventions that produce a desired change in an associated outcome[312–315] or that change a data distribution from an initial state to a target state[316,317].

Here, we build on the concept of optimal intervention design to introduce PDGrapher, an approach to predict arbitrary sets of genes to be targeted to shift cell gene expression from an initial to a desired state in genetic and chemical perturbagen datasets. Our approach is composed of two modules. First, a perturbagen discovery module takes the initial and desired target cell states and outputs a candidate perturbagen (set of genes). Then, a response prediction module takes the initial state and predicted perturbagen and predicts the cell response to the predicted perturbagen. Our response prediction and perturbagen discovery modules are graph neural network (GNN)-based models operating on an input graph where edge mutilations represent the effects of interventions on our noisy causal graph of gene-gene interactions. We consider a Markovianity setting (without unobserved confounders), a strong constraint that could potentially be addressed in future works.

5.3 Related work

Learning optimal interventions. The problem of learning interventions to achieve a desired state has gained interest in recent years. A few recent works have formulated this problem as finding optimal interventions to optimize an associated outcome[312–315]. These works offer varied approaches. For example, Mueller et al.[312] aimed to learn an intervention policy defined by a covariate transformation that produces the largest post-intervention improvement with high uncertainty. Pacchiano et al.[313] formalize the task as a bandit optimization problem in which each bandit arm corresponds to a covariate to intervene, and the goal is to recover an almost optimal arm in the least number of arm pulls possible. Mueller et al.[314] and Hie et al.[315] approached the problem of sequence-based data where each sequence is associated with an outcome and the goal is to find mutations in the input sequence that increase the desired outcome. Other recent works have formulated this problem as finding optimal interventions to shift the system to a desired state. Zhang et al.[316,317] aimed to find an intervention that applied to a distribution helps match a desired distribution. Specifically,

given a distribution P over $\mathbf{X}$, and a desired distribution Q over $\mathbf{X}$, the goal is to find an optimal matching intervention I such that P^I best matches Q under some metric. They address the special case of soft interventions (shift interventions) and use the expectation of distributions as the distance metric.

Neural networks and structural causal models (SCMs). Causal representation learning has been a growing trend in recent years[318]. It aims to combine the strength of traditional causal learning methods with the strong capabilities of deep learning in the face of large and noisy data. Bottlenecks of traditional causal learning methods include unstructured high-dimensional variables, combinatorial optimization problems, unknown intervention, unobserved confounders, selection bias, and estimation bias[318]. There are three areas in which deep learning helps to overcome these bottlenecks[318]: first, in learning causal variables from high-dimensional unstructured data; second, in learning the causal structure between causal variables, called *causal discovery* within the causal inference literature; and third, in facilitating the inference of interventional and counterfactual queries. Within the last branch, a promising approach aims to join structural causal models (SCMs) and neural models to facilitate interventional and counterfactual querying. Parafita et al. put forward the requirements that any deep learning (DL) model should fulfill to approximate causal queries and introduced normalizing causal flows as a specific instantiation[319]. Pawlowski et al. followed a similar approach to introduce a model capable of computing counterfactual queries[320]. Xia et al. approached the problem from a different perspective, introducing a neural causal model (NCM), which is a type of SCM that has neural networks as structural equations[321]. Together with the NCM, the authors introduced an algorithm that provably performs identification and inference of interventional queries[321]. A follow-up work extended the NCM framework for identification and inference of counterfactual queries[322]. Our work is inspired by the concept of NCMs by considering the graph in which we operate as a noisy version of a causal graph and our model operating on the graph as a proxy for the structural equations.

Interventions in graph neural networks (GNNs). Graph Neural Networks (GNNs) are a type of neural model that fall under the umbrella term of geometric deep learning[26,323]. These models operate on graph-structured data to compute transformed representations that are useful for downstream predictive tasks. Given their ability to operate over graphs by

design, they appear especially relevant in connection to NCMs. A recent work by Zecevic et al. explored this connection and introduced interventional GNNs in which interventions are represented through mutilations in the input graph and interventional inference as GNN computations on the mutilated graph[324]. We borrow this concept in our work and additionally extend the representational capabilities of GNNs by assigning learnable embeddings to input nodes.

5.4 Contributions and impact

Contributions

- We propose a new formulation for the task of phenotype-drive therapeutic lead discovery to predict arbitrary leads not limited to existing curated libraries of compounds.
- We propose PDGrapher, a causally-inspired GNN approach for the task of perturbagen discovery as defined previously. Our model is able to predict perturbagens to move cell line phenotype signatures from an initial to a desired state across a set of genetic and chemical interventions datasets.

Impact. We introduce a novel formulation for the task of phenotype-driven lead discovery that increases the flexibility of current approaches. We propose PDGrapher as an initial approach to address the modified problem formulation. Our results show the promise of reframing phenotype-driven lead discovery to allow for the prediction of arbitrary leads, which can open the search space and point to promising directions.

This Chapter is based on the following publication:

1. **Gonzalez, G.**, Herath, I., Veselkov, K., Bronstein, M., and Zitnik, M. (in preparation). Combinatorial discovery of therapeutic targets using a causally inspired neural network.

5.5 Problem definition

Preliminaries. A calligraphic letter $\mathcal{X}$ indicates a set, an italic uppercase letter X denotes a graph; an uppercase $\mathbf{X}$ denotes a matrix; lowercase $\mathbf{x}$ denotes a vector, and a monospaced letter $\mathtt{X}$ indicates a tuple. An uppercase letter X indicates a random variable, and lowercase

letter $\mathtt{x}$ indicates its corresponding value; a bold uppercase $\mathbf{X}$ denotes a set of random variables, and a lowercase letter $\mathbf{x}$ indicates its corresponding values. We denote $P(\mathbf{X})$ as a probability distribution over a set of random variables $\mathbf{X}$ and $P(\mathbf{X} = \mathbf{x})$ as the probability of $\mathbf{X}$ being equal to the value of $\mathbf{x}$ under the distribution $P(\mathbf{X})$. For simplicity, $P(\mathbf{X} = \mathbf{x})$ is abbreviated as $P(\mathbf{x})$.

Problem formulation: learning optimal inverse perturbagens. Intuitively, given a diseased cell line sample, we would like to predict which genes need to be targeted to reverse the effects of disease, that is, which genes need to be perturbed to shift the cell gene expression state as close as possible to the healthy state. Next, we formalize our problem. Let $\mathtt{M} =< \mathbf{E}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) >$ be an SCM associated with causal graph G. Let $\mathcal{T} = \{\mathtt{T}_1, ..., \mathtt{T}_m\}$ be a dataset of paired healthy and diseased samples, where each element is a three-tuple $\mathtt{T} =< \mathbf{v}^h, \mathbf{U}, \mathbf{v}^d >$ with $\mathbf{v}^h \in [0,1]^N$ being gene expression values of a healthy cell line (variable states before perturbation), $\mathbf{V_U}$ being the disease-causing perturbed variable (gene) set in $\mathbf{V}$, and $\mathbf{v}^d \in [0,1]^N$ being gene expression values of a diseased cell line (variable states after perturbation), with $\mathbf{V_U} = \mathbf{v'_U} = \mathbf{v}^d_\mathbf{U}$. Our goal is to find, for each sample $\mathtt{T} =< \mathbf{v}^h, \mathbf{U}, \mathbf{v}^d >$, the perturbation set $\mathbf{U}'$ with the highest likelihood of reversing variable states from diseased $\mathbf{v}^d$ to healthy $\mathbf{v}^h$ state. For clarity and to increase generality, we refer to the desired variable states as *treated* ($\mathbf{v}^t$). Our goal can then be expressed as

$$argmax\ _{\mathbf{U}'} P^{G^{\mathbf{U}}}(\mathbf{V} = \mathbf{v}^t \mid do(\mathbf{U}')) \tag{2}$$

where $P^{G^{\mathbf{U}}}$ represents the probability over graph G mutilated upon perturbations in variables in $\mathbf{U}$. Under markovianity assumptions, the above interventional probability can be expressed as a conditional probability on the mutilated graph $G^{\mathbf{U}'}$:

$$argmax\ _{\mathbf{U}'} P^{G^{\mathbf{U}'}}(\mathbf{V} = \mathbf{v}^t \mid \mathbf{U}') \tag{3}$$

which under the causal Markov condition, is:

$$argmax\ _{\mathbf{U}'} \prod_i P(\mathsf{V}_i = \mathsf{v}_i^t \mid \mathbf{Pa}_{\mathsf{V}_i}) \tag{4}$$

where $\mathbf{Pa}_{\mathsf{V}_i}$ represents the parents of variable V_i according to graph $G^{\mathbf{U}'}$ (that is, the

mutilated graph upon intervening on variables in $\mathbf{U}'$). Here, state of a variable $\mathsf{V}_j \in \mathbf{Pa}_{\mathsf{V}_i}$ will be equal to an arbitrary value v'_j if $\mathsf{V}_j \in \mathbf{U}'$. Therefore, intervening on the variable set $\mathbf{U}'$ modifies the graph used to obtain conditional probabilities in addition to determining the state of variables in $\mathbf{U}'$.

5.6 Datasets

We compiled and processed a total of six primary data sources and two additional repositories of biological information. Data sources included PPIs, healthy and diseased cell line gene expression, diseased cell line gene expression upon chemical or genetic perturbagens, phenotype-associated genes, drug targets, and drug indications. The following is a description of the data sources and preprocessing steps.

Human protein-protein interaction network. We built a PPI network by aggregating proteins and connections from BIOGRID[325] (accessed in March 2022), HuRI[326], and Menche et al.[327] In this graph, nodes represent human proteins, and edges exist between nodes if there is physical interaction between the proteins. We downloaded a gene ID mapping file from www.genenames.org/download/custom/. Using this file, we mapped proteins in BIOGRID and Menche et al.[327] from Entrez gene ID[328] to HUGO gene nomenclature committee ID[329], and proteins in HuRI from Ensembl gene ID[330] to HUGO gene nomenclature committee ID[329]. Our final PPI is made up of the union of nodes and edges, resulting in a graph with 15,742 nodes and 222,498 edges.

Gene expression data. We downloaded the LINCS[331] library level-3 gene expression data from https://clue.io/releases/data-dashboard (accessed in February 2022). Level-3 data consist of quantile-normalized samples across each plate and is the lowest level in the LINCS library that can be compared across plates. LINCS contains gene expression measurements for 12,327 genes upon genetic and chemical perturbations. There are 387,317 samples treated with CRISPR genetic interventions (treated samples), with 5,156 unique knocked-out genes across 27 unique cell lines. There are an average of 17.18 replicates per cell line-knocked-out gene pair. The number of unique genes knocked-out in each cell line varies from one to 5,114, with an average of 2,042.14 unique genes knocked out per cell line.

Control data for CRISPR interventions, that is diseased samples, are genetic perturbagens

that either do not contain a gene-specific sequence or whose gene-specific sequence targets a gene not expressed in the human genome. There are 47,781 diseased samples across 50 cell lines. The number of diseased samples for each cell line varies from one to 6,890, with an average of 955.62 diseased samples per cell line.

There are 1,313,292 samples treated with chemical interventions (treated samples), with 31,234 unique compounds across 229 unique cell lines. There are an average of 7.96 replicates per cell line-compound pair. The number of compounds tested in each cell line varies from one to 19,509, with an average of 719.69 unique compounds tested per cell line. Drugs are administered at different doses and measured at different timepoints after treatment. On average, there are 2.73 different doses per compound-cell line pair, with a minimum of one and a maximum of 26 different doses. On average, gene expression is measured at 1.25 timepoints per compound-cell line pair, with a minimum of one and a maximum of 13 different timepoints. The control data for chemical interventions, that is diseased samples, is treatment with vehicle (dimethyl sulfoxide). There are a total of 76,795 diseased samples across 226 cell lines. The number of diseased samples for each cell line varies from one to 7,336, with an average of 339.80 diseased samples per cell line. On average, gene expression of diseased samples is measured at 1.4 timepoints, with a minimum of one and a maximum of 5 different timepoints.

We filtered cell lines to keep those treated with at least 4,000 unique genetic or chemical perturbagens, resulting in 10 selected cell lines for each, genetic and chemical datasets. To find healthy cell line counterparts, we extracted all cell lines with "unknown" tumor phase in the downloaded LINCS dataset (N = 145). The rationale behind filtering cell lines using "unknown" tumor phase is that these are potentially healthy cell lines as there is no tumour phase attached to them, making then a good starting point for literature search. Then, we filtered the cell lines by tissue type. To find the exact match to diseased cell lines and confirm their experimental use as healthy counterparts, we performed a manual literature search. We extracted healthy counterparts for three out of the 10 diseased cell lines: cell line NL20 as the healthy counterpart for A549, cell line MCF10A as the healthy counterpart for MCF7, and cell line RWPE1 as the healthy counterpart for PC3.

Genetic interventions correspond to gene experiment knockouts in which the gene expression of the knocked-out gene after the intervention is zero. Chemical interventions correspond

Dataset type	Cancer type	Samples type	N samples	Category	N perturbagens
Genetic	Lung cancer	healthy	50	vehicle	-
		diseased	4,327	vector	-
		treated	24,255	CRISPR	3,711
	Breast cancer	healthy	113	untreated	-
		diseased	4,852	vector	-
		treated	18,774	CRISPR	3,090
	Prostate cancer	healthy	185	vector	-
		diseased	6,890	vector	-
		treated	21,229	CRISPR	3,710
Chemical	Lung cancer	healthy	50	vehicle	-
		diseased	5,261	vehicle	-
		treated	23,100	compound	1,041
	Breast cancer	healthy	2,675	untreated	-
		diseased	7,336	vehicle	-
		treated	35,421	compound	1,154
	Prostate cancer	healthy	185	vector	-
		diseased	7,202	vehicle	-
		treated	32,555	compound	1,182

Table 5.1: Table with number of healthy, diseased, and treated samples for lung cancer (A549), breat cancer (MCF7), and prostate cancer (PC3), across genetic and chemical perturbagens.

to small-molecule treatments where each molecule targets one or more proteins. Chemical interventions were performed at different dose levels and measured at different timepoints. We included replicates measured at all time points and doses. For each cell line and condition (healthy, diseased, and treated), we log-normalized level 3 gene expression data, and applied a min-max normalization to transform gene expression values into the range $[0, 1]$ following established practices in the field. We matched genes in LINCS to proteins in our PPI using the HUGO gene nomenclature committee ID[329], resulting in 10,716 overlapping genes and 149,270 edges. Furthermore, we excluded treated samples from our datasets whose targeted genes were not included in the PPI.

For each cell line, there are healthy, diseased, and treated gene expression samples treated with a number of genetic or chemical perturbagens (Table 5.1). For healthy counterparts, samples with the corresponding treatment ("vector" for genetic perturbagens, and "vehicle" for chemical perturbagens) were not available; therefore, we used the closest possible one (see "Sample category" in Table 5.1).

Disease-associated genes. We extracted disease-associated genes from COSMIC[332] (accessed in October 2022) in addition to expert-curated genes available at https://cancer.sanger.ac.uk/cosmic/curation. Genes were represented using the HUGO gene nomenclature committee ID. For each cell line in our dataset, we extracted cancer-causing mutations as the list of genes with "verified" *mutation verification status* present in the list of genes curated by experts. Mapping the resulting genes to our list of genes in the PPI resulted in eight disease-associated genes for lung cancer cell line A549, eight disease-associated genes for breast cancer cell line MCF7, and one disease-associated gene for prostate cancer cell line PC3. Therefore, we filtered out cell line PC3 and proceeded with only MCF7 and A549.

Drug targets. We downloaded drug-related data from DrugBank[88] (accessed in November 2022). We extracted drug names and synonyms, chemical identifiers, drug targets, and all available synonyms for each target. We mapped drugs in DrugBank with chemical perturbagens in LINCS by InChI key[333], resulting in 1,522 out of 31,234 unique LINCS compounds mapped to DrugBank with information on at least one target. We mapped drug targets to our PPI network using the HUGO gene nomenclature committee ID, excluding any drug targets that did not map. Chemical interventions target multiple genes, with a minimum of one, a maximum of 300, and an average of 2.44 targets per compound.

List of cancer drugs for baseline models using cancer targets. We extracted a list of cancer drugs by cancer type from the National Cancer Institute (https://www.cancer.gov/about-cancer/treatment/types/targeted-therapies/approved-drug-listtargeted-therapy-approved-for-breast-cancer; accessed in November 2022). We mapped drug names to DrugBank to obtain cancer drug targets. In total, there are 24 drugs associated with breast cancer (cell line MCF7) and 30 drugs associated with lung cancer (cell line A549).

Uncertainty in databases. It is crucial to address the inherent uncertainties and limitations associated with the databases we used: BIOGRID, HuRI, LINCS, COSMIC, and DrugBank resources. While they are invaluable for providing extensive data on protein-protein interactions, gene expression, cancer-associated genes, and drug targets, they are subject to several forms of uncertainty.

Firstly, the data in these databases is often generated from a variety of experimental

methods and conditions, which can lead to variability in the results. This heterogeneity can affect the reliability and reproducibility of the data. For instance, differences in experimental techniques, cell lines used, or conditions of the experiment can lead to discrepancies in protein-protein interactions or gene expression profiles.

Secondly, these databases are continuously updated as new research and discoveries emerge. While this ensures the inclusion of the latest information, it also implies that the data is dynamic and may change over time, potentially impacting longitudinal studies or comparisons made at different time points. Furthermore, the accuracy of database entries is reliant on the original experimental data and the curation process. Misinterpretations or errors in the primary literature can propagate into the database, and while curators endeavor to maintain high-quality data, complete accuracy might not be guaranteed.

Lastly, there is also the consideration of completeness. No database can be entirely comprehensive, and there may be gaps in the data, especially regarding newly discovered or less well-studied proteins, genes, or interactions. Here, we do not take special measures to account for the data uncertainty in our model, hence, we acknowledge that the uncertainty in the databases can translate into uncertainties in our results. However, we posit that the representation power of deep learning models can compensate to some extent for this uncertainty by capturing true signal within the noisy data.

Disease intervention data. Disease intervention datasets consist of gene expression measurements of healthy cell lines, disease-associated genes, and gene expression measurements of diseased cell lines. Gene expression samples of healthy and diseased cell lines were retrieved from LINCS[331], and disease-associated genes were retrieved from COSMIC[332]. Each dataset $\mathcal{T} = \{\mathtt{T}_1, \dots, \mathtt{T}_M\}$ is a collection of paired healthy-diseased cell lines where in each sample $\mathtt{T} =< \mathbf{x}^h, \mathcal{U}, \mathbf{x}^d >$, $\mathbf{x}^h$ corresponds to gene expression values of healthy cell line; set $\mathcal{U}$ is comprised by a randomized subset of disease-associated genes; and $\mathbf{x}^d$ corresponds to gene expression values of diseased cell lines (that is, upon mutations on genes in $\mathcal{U}$). To select the randomized set of disease-associated genes, we first select at random a proportion $p \in \{0.25, 0.50, 0.75, 1\}$, and then select N disease-associated genes at random where N is the proportion multiplied by the total number of disease-associated genes. Given that there are more diseased samples available than healthy samples (see Table 5.1), when building the triplets, we select a random sample from the set of healthy samples and, therefore, have

non-unique healthy samples during training. In total, we built two datasets of disease interventions: one comprised by gene expression of the healthy cell line MCF10A, breast cancer mutations, and gene expression of the breast cancer cell line MCF7; the second comprised by gene expression of the healthy cell line NL20, lung cancer mutations, and gene expression of the lung cancer cell line A549. More details on data compilation and processing can be found in the *Methods* section.

Treatment intervention data (genetic). Genetic treatment intervention datasets consist of single-gene knockout experiments using CRISPR/Cas9-mediated gene knockout. Genetic treatment intervention data is comprised of gene expression measurements of diseased cell lines, single knocked-out genes, and gene expression measurements of treated cell lines. Gene expression samples of diseased and treated cell lines, and knocked-out genes were retrieved from LINCS[331]. Each dataset $\mathcal{T} = \{\mathtt{T}_1, \dots, \mathtt{T}_M\}$ is a collection of paired diseased-treated cell lines where in each sample $\mathtt{T} =< \mathbf{x}^d, \mathcal{U}', \mathbf{x}^t >$, $\mathbf{x}^d$ corresponds to gene expression values of the diseased cell line; set $\mathcal{U}'$ is comprised by the knocked-out gene; and $\mathbf{x}^t$ corresponds to gene expression values of treated cell lines (that is, upon knocking-out the gene in $\mathcal{U}'$). Given that there are more treated samples available than diseased samples (see Table 5.1), when building the triplets, we select a random sample from the set of diseased samples and therefore have non-unique diseased samples during training. In total, we built two datasets of treatment interventions: one comprised of gene expression of the diseased cell line MCF7, knocked-out genes, and gene expression of the treated cell line MCF7, and the second comprised by gene expression of the diseased cell line A549, knocked-out genes, and gene expression of the treated cell line A549. More details on data compilation and processing can be found in the *Methods* section.

Treatment intervention data (chemical). Chemical treatment intervention datasets consist of chemical compound treatment experiments. The data is comprised by gene expression measurements of diseased cell lines, chemical compound targets, and gene expression measurements of treated cell lines. Gene expression samples of diseased and treated cell lines were retrieved from LINCS[331], and chemical compound targets were retrieved from DrugBank. Each dataset $\mathcal{T} = \{\mathtt{T}_1, \dots, \mathtt{T}_M\}$ is a collection of paired diseased-treated cell lines where in each sample $\mathtt{T} =< \mathbf{x}^d, \mathcal{U}', \mathbf{x}^t >$, $\mathbf{x}^d$ corresponds to gene expression values of the

diseased cell line; set $\mathcal{U}'$ is comprised by the chemical compound targets; and $\mathbf{x}^t$ correspond to gene expression values of the treated cell lines (that is, treated with the chemical perturbagen targeting genes in $\mathcal{U}'$). Given that there are more treated samples available than diseased samples (see Table 5.1), when building the triplets, we select a random sample from the set of diseased samples, and therefore have non-unique diseased samples during training. In total, we built two datasets of treatment interventions: one comprised by gene expression of the diseased cell line MCF7, chemical compound target genes, and gene expression of the treated cell line MCF7, and the second comprised by gene expression of the diseased cell line A549, chemical compound target genes, and gene expression of the treated cell line A549. More details on data compilation and processing can be found in the *Methods* section.

5.7 Methods

Problem formulation - representation-learning-based learning of optimal inverse perturbagens. In the previous section, we drew on the SCM framework to introduce a generic formulation for the task of learning optimal inverse perturbagens. Instead of approaching the problem from a purely causal inference perspective, we draw upon representation learning to approximate the queries of interest and address the limiting real-world setting of a noisy and incomplete causal graph. Formulating our problem using the SCM framework allows for explicit modeling of interventions and the formulation of interventional queries. Inspired by this principled problem formulation, we next introduce the problem formulation using a representation learning paradigm.

We let $G = (\mathcal{V}, \mathcal{E})$ denote a graph with $|\mathcal{V}| = n$ nodes, and $|\mathcal{E}|$ edges, which contains partial information on causal relationships between nodes in $\mathcal{V}$ and some noisy relationships. We refer to this graph as *noisy causal graph*. Let $\mathcal{T} = \{\mathtt{T}_1, ..., \mathtt{T}_\mathsf{M}\}$ be a dataset with an individual sample being a three-tuple $\mathtt{T} =< \mathbf{x}^h, \mathcal{U}, \mathbf{x}^d >$ with $\mathbf{x}^h \in [0,1]^n$ being the node states (attributes) of healthy cell sample (before perturbation), $\mathcal{U}$ being the set of disease-causing perturbed nodes in $\mathcal{V}$, and $\mathbf{x}^d \in [0,1]^n$ being the node states (attributes) of diseased cell sample (after perturbation), where $\mathbf{x}^d_\mathcal{U} = \mathbf{x}'_\mathcal{U}$ (that is, $\mathbf{x}^d_\mathcal{U}$ is set to arbitrary values). We denote by $G^\mathcal{U} = (\mathcal{V}, \mathcal{E}^\mathcal{U})$ the graph resulting from the mutilation of edges in G as a result of perturbing nodes in $\mathcal{U}$ (one graph per perturbagen; we avoid using superindices for

simplicity). Here again, we refer to the desired variable states as *treated* ($\mathbf{x}^t$). Our goal is then to learn a function:

$$f : G^{\mathcal{U}'},\ \mathbf{x}^d, \mathbf{x}^h \rightarrow argmax\ _{\mathcal{U}'} P^{G^{\mathcal{U}'}}(\mathbf{x} = \mathbf{x}^t | \mathbf{x}^d, \mathcal{U}') \tag{5}$$

that given the graph $G^{\mathcal{U}'}$ and diseased node states $\mathbf{x}^d$, predicts what perturbagen (node set $\mathcal{U}'$) would have the highest chance of shifting the node states to the treated state $\mathbf{x}^t$. We note here that $P^{G^{\mathcal{U}'}}$ represents probabilities over graph $G^{\mathcal{U}}$ mutilated upon perturbations in nodes in $\mathcal{U}'$. Under causal Markov condition, we can factorize $P^{G^{\mathcal{U}'}}$ over graph $G^{\mathcal{U}'}$:

$$f : G^{\mathcal{U}'},\ \mathbf{x}^d, \mathbf{x}^h \rightarrow argmax\ _{\mathcal{U}'} \prod_i P(\mathbf{x}_i = \mathbf{x}_i^t | \mathbf{x}_{\mathcal{PA}_i}) \tag{6}$$

that is, the probability of each node depends only on its parents $\mathcal{PA}_i$ in graph $G^{\mathcal{U}'}$.

We assume (i) real-valued node states; (ii) G being fixed and given; and (iii) atomic and non-atomic perturbagens (intervening on individual nodes or sets of nodes). Given that the value of each node should depend only on its parents on the graph $G^{\mathcal{U}'}$, a message passing framework appears especially suited to compute the factorized probabilities P.

Comment on the problem formulation. In the SCM framework, the conditional probabilities in equation 4 are computed recursively on the graph, each being an expectation over exogenous variables $\mathbf{E}$. Therefore, node states from the previous time point are not necessary. To translate this query into the realm of representation learning, we discard the existence of noise variables and directly try to learn a function encoding the transition from an initial state to a desired state.

Intuitively, an exhaustive approach to solving equation 6 would be to search the space of all potential perturbagens $\mathcal{U}'$ and score how effective each is in achieving the desired treated state. This is, indeed, how many cell response prediction approaches can be used for perturbagen discovery[307,308,334]. However, with moderately sized graphs, this is highly computationally expensive, if not intractable. Instead, we propose to search for potential perturbagens in a smart way with a perturbagen discovery module (f_p), and a way to score each potential perturbagen with a response prediction module (f_r).

Relationship to conventional graph prediction tasks. The fact that the prediction

for each variable is dependent only on its parents in a graph, GNNs appear especially suited for this problem. We can formulate the query of interest under a graph representation learning paradigm as follows: Given a graph $G = (\mathcal{V}, \mathcal{E})$, with paired sets of node attributes $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m\}$ and node labels $\mathcal{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_m\}$ where each $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$, with $\mathbf{y}_i \in [0, 1]$, we aim at training a neural message passing architecture that given node attributes $\mathbf{X}_i$ predicts the corresponding node labels $\mathbf{Y}_i$. There are, however, some major differences between our problem formulation and the conventional graph prediction tasks, namely, graph and node classification (summarized in Table 5.2).

In node classification, a single graph G is paired with node attributes $\mathbf{X}$ and the task is to predict the node labels $\mathbf{Y}$. Our formulation differs in that we have m paired sets of node attributes $\mathcal{X}$ and labels $\mathcal{Y}$ instead of a single set, yet they are similar in that there is a single graph in which GNNs operate. In graph classification, a set of graphs $\mathcal{G} = \{G_1, \dots, G_m\}$ is paired with a set of node attributes $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m\}$ and the task is to predict a label for each graph $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_m\}$. Here, graphs have a varying structure, and both the topological information and node attributes are predictive of graph labels. In our formulation, a single graph is used in combination with each of the node attributes $\mathbf{X}_i$, and the goal is to predict a label for each node, not for the whole graph.

Task	Number of graphs	Number of node attribute sets	Label dimensions
Graph Classification	m	m	m x 1 (one for each graph)
Node Classification	1	1	1 x n (one for each node)
Ours	1	m	m x n (one for each node of each graph)

Table 5.2: Our problem formulation is similar to conventional node and graph classification tasks albeit some major differences exist.

Proposed approach. We propose an approach to learn optimal inverse perturbagens composed of two modules. First, a perturbagen discovery module f_p searches the space of potential perturbagens to predict a suitable candidate $\mathcal{U}'$. Next, a response prediction module f_r checks the goodness of the predicted set $\mathcal{U}'$, that is how effective intervening on variables in $\mathcal{U}'$ is in shifting node states to the desired treated state $\mathbf{x}^t$.

$$(1) \quad \mathbf{x}^d, \mathbf{x}^t \xrightarrow{f_p} \hat{\mathcal{U}}'$$

$$(2) \quad \mathbf{x}^d, \hat{\mathcal{U}}' \xrightarrow{f_r} \hat{\mathbf{x}}^t$$

Model optimization. We optimize our response prediction module f_r using cross-entropy loss on known triplets $< \mathbf{x}^h, \mathcal{U}, \mathbf{x}^d >$ and $< \mathbf{x}^d, \mathcal{U}', \mathbf{x}^t >$:

$$\mathcal{L}_{f_r} = CE\left(\mathbf{x}^d,\ f_r(\mathbf{x}^h,\ \mathcal{U})\right) + CE\left(\mathbf{x}^t,\ f_r(\mathbf{x}^d,\ \mathcal{U}')\right)$$

We optimize our intervention discovery module f_p using a cycle loss such that the response upon a predicted $\mathcal{U}'$ is as close to the desired treated state as possible. In addition, we provide a supervisory signal for the prediction of $\mathcal{U}'$ in the form of cross-entropy loss.

$$\mathcal{L}_{f_p} = CE\left(\mathbf{x}^t,\ f_r(\mathbf{x}^d,\ f_p(\mathbf{x}^d,\ \mathbf{x}^t))\right) + CE\left(\mathcal{U}', f_p(\mathbf{x}^d,\ \mathbf{x}^t)\right) \quad (with\ f_r\ frozen)$$

We train f_p and f_r in parallel and implement early stopping separately (see *Experimental setup* for more details). Trained modules f_p and f_r are then used to predict, for each diseased cell sample, which nodes should be perturbed ($\mathcal{U}'$) to achieve a desired treated state (Figure 5.1C).

Response prediction module. Our response prediction module f_r should learn to map pre-perturbagen node values to post-perturbagen node values through learning relationships between connected nodes (equivalent to learning structural equations in SCMs) and propagating the effects of perturbations downstream in the graph (analogous to the recursive nature of query computations in SCMs).

Given a triplet $< \mathbf{x}^h, \mathcal{U}, \mathbf{x}^d >$, we propose a neural model operating on a mutilated graph, $G^{\mathcal{U}}$ where the node attributes are the concatenation of $\mathbf{x}^h$ and $\mathbf{x}'_{\mathcal{U}}$, predicting diseased node values $\mathbf{x}^d$. Each node i has a two-dimensional attribute vector $\mathbf{d}_i = [\mathbf{x}^h_i \,||\, \mathbf{x}'_{\mathcal{U}}]$, where the first element is its gene expression value $\mathbf{x}^h_i$, and the second is a perturbation flag: a binary label indicating whether a perturbation occurs at node i. In practice, we embed each node feature into a high-dimensional continuous space by assigning learnable embeddings to each node based on the value of each input feature dimension. Specifically, for each node, we use the binary perturbation flag to assign a d-dimensional learnable embedding, which is different between nodes but shared across samples for each node. To embed the gene expression value $\mathbf{x}^h_i \in [0, 1]$, we first calculate thresholds using quantiles to assign the gene expression value to one of B bins, then we use the bin index to assign a d-dimensional learnable embedding, which

is different between nodes but shared across samples for each node. To increase our model's representation power, we additionally concatenate a d-dimensional positional embedding (d-dimensional vector initialized at random following a normal distribution). Concatenating these three embeddings results in an input node representation of dimensionality $3.d$

For each node $i \in \mathcal{V}$, an embedding $\mathbf{z}_i$ is computed using a GNN operating on the node's neighbors' attributes. The most general formulation of a GNN layer is:

$$\mathbf{h}'_i = \phi\left(\mathbf{h}_i, \bigoplus_{j \in \mathcal{N}^i} \psi(\mathbf{h}_i, \mathbf{h}_j)\right)$$

where $\mathbf{h}'_i$ represents the updated information of node i, and $\mathbf{h}_i$ represents the information of node i in the previous layer, with embedded $\mathbf{d}_i$ being the input to the first layer. ψ is a *MESSAGE* function, $\bigoplus$ an *AGGREGATE* function (permutation-invariant, typically summation), and ϕ is an *UPDATE* function. We obtain an embedding $\mathbf{z}_i$ for node i by stacking K GNN layers. Node embedding $\mathbf{z}_i \in \mathbb{R}$ is then passed to a multilayer feed-forward neural network to obtain an estimate of the post-perturbation node values $\mathbf{x}^d$.

Perturbation discovery module. Our perturbagen prediction module f_p should learn which nodes in the graph should be perturbed to shift node states (attributes) from diseased $\mathbf{x}^d$ to a desired treated state $\mathbf{x}^t$.

Given a triplet $< \mathbf{x}^d, \mathcal{U}', \mathbf{x}^t >$, we propose a neural model operating on graph $G^{\mathcal{U}'}$ with node features $\mathbf{x}^d$ and $\mathbf{x}^t$ that predicts a ranking for each node, where the top P ranked nodes should be predicted as the nodes in $\mathcal{U}'$. Each node i has a two-dimensional attribute vector: $\mathbf{d}_i = [\mathbf{x}_i^d \mid\mid \mathbf{x}_i^t]$. In practice, we represent these binary features in a continuous space using the same approach as described for our response prediction module f_r.

For each node $i \in \mathcal{V}$, an embedding $\mathbf{z}_i$ is computed using a GNN operating on the node's neighbors' attributes. The most general formulation of a GNN layer is:

$$\mathbf{h}'_i = \phi\left(\mathbf{h}_i, \bigoplus_{j \in \mathcal{N}^i} \psi(\mathbf{h}_i, \mathbf{h}_j)\right)$$

where $\mathbf{h}'_i$ represents the updated information of node i, and $\mathbf{h}_i$ represents the information of node i in the previous layer, with embedded $\mathbf{d}_i$ being the input to the first layer. ψ is a *MESSAGE* function, $\bigoplus$ an *AGGREGATE* function (permutation-invariant, typically

summation), and ϕ is an *UPDATE* function. We obtain an embedding $\mathbf{z}_i$ for node i by stacking K GNN layers. Node embedding $\mathbf{z}_i \in \mathbb{R}$ is then passed to a multilayer feed-forward neural network to predict a real-valued number for node i.

Network proximity between predicted and ground truth perturbagens. Let $\mathcal{P}$ be the set of predicted perturbagen targets, $\mathcal{R}$ be the set of ground truth targets, and $spd(p, r)$ be the shortest-path distance between nodes in P and R. We measure the closest distance between P and R as:

$$d(P, R) = \frac{1}{|R||P|} \sum_{r \in R} \sum_{p \in P} spd(p, r)$$

Model implementation and training. We implemented PDGrapher using PyTorch[154] and the Torch Geometric Library[155]. The implemented architecture yields a neural network with the following hyperparameters: number of GNN layers, and number of prediction layers. We set the number of prediction layers to two, and perform a grid search over the number of GNN layers (1-3 layers). We train our model using a 5-fold cross-validation strategy and report PDGrapher's performance resulting from the best-performing hyperparameter setting.

5.8 Experimental setup

Baselines. We evaluated the performance of our intervention discovery model against a set of baselines:

- **Random baseline:** Given a sample $\mathtt{T} =< \mathbf{x}^d, \mathcal{U}', \mathbf{x}^t >$, the random baseline returns N random genes as the prediction of genes in $\mathcal{U}'$, where N is the number of genes in $\mathcal{U}'$.
- **Cancer genes:** Given a sample $\mathtt{T} =< \mathbf{x}^d, \mathcal{U}', \mathbf{x}^t >$, the cancer genes baseline returns the top N genes from an ordered list where the first M genes are disease-associated genes (cancer-causing genes), the remaining $N - M$ genes are ranked randomly, and N is the number of genes in $\mathcal{U}'$.
- **Cancer drug targets:** Given a sample $\mathtt{T} =< \mathbf{x}^d, \mathcal{U}', \mathbf{x}^t >$, the cancer genes baseline returns the top N genes from an ordered list where the first M genes are cancer drug targets and the remaining $N - M$ genes are ranked randomly, where N is the number

of genes in $\mathcal{U}'$.

- **scGen[307]:** scGen is a widely-used gold-standard latent variable model for response prediction[335–338]. Given a set of observed cell type in control and perturbed state, scGen predicts the response of a new cell type to the perturbagen seen in training. To utilize scGen as a baseline, for each dataset type, we first fit it to our LINCS gene expression data to predict response to perturbagens, training one model per perturbagen (chemical or genetic). Then, given a sample of paired diseased-treated cell line states, $\mathtt{T} =< \mathbf{x}^d, \mathcal{U}', \mathbf{x}^t >$, we compute the response of cell line with gene expression $\mathbf{x}^d$ to all perturbagens. The predicted perturbagen is that whose predicted response is closest to $\mathbf{x}^t$ in R^2 score.

Dataset splits. We trained our model on two different settings. **Random splits:** Our dataset was split randomly into training and test sets to measure our model performance in an IID setting. **Leave-cell-line-out splits:** To measure model performance on unseen cell lines, we trained our model with random splits on one cell line, and tested on a new cell line.

Model training. For all dataset split settings, we trained our model using five-fold cross-validation and reported performance metrics as the average on the test sets. Within each fold, we further split the training set into training and validation sets (8:2) to perform early stopping: we train the model on the training set until the validation loss has not decreased at least 10^{-5} for 15 continuous epochs.

Evaluation metrics. We report the average sample-wise R^2 score, and the average perturbagen-wise R^2 score to measure performance in the prediction of $\mathbf{x}^t$. The sample-wise R^2 score is computed as the square of the Pearson correlation between the predicted sample $\hat{\mathbf{x}}^t \in \mathbb{R}^N$ and the real sample $\mathbf{x}^t \in \mathbb{R}^N$. The perturbagen-wise R^2 score is adopted from scGen, a state-of-the-art method for response prediction[307]. It is computed as the square of the Pearson correlation of a linear least-squares regression between a set of predicted treated samples $\hat{\mathbf{X}}^t \in \mathbb{R}^{N \times S}$ and a set of real treated samples $\mathbf{X}^t \in \mathbb{R}^{N \times S}$ for the same perturbagen. Higher values indicate better performance in predicting the treated sample $\mathbf{x}^t$ given the diseased sample $\mathbf{x}^d$ and predicted perturbagen.

We also report the average ranking of real interventions (genes) in the predicted ordered

list of interventions to measure the ability of our model to rank interventions correctly. We normalize the ranking to the range $[0, 1]$ as $1 - ranking/N$ where higher values indicate better performance, that is, the model ranks ground truth perturbed genes closer to the top of the predicted list. In addition, we report the accuracy of perturbagen discovery (PD accuracy) as the Jaccard score between the predicted set of genes $\mathcal{U}'$ and the ground truth set of genes $\mathcal{U}$. We report the PD accuracy as the average Jaccard score only for the predicted perturbagens that have at least one overlapping gene with the ground-truth perturbagen. We additionally report the proportion of test samples in which the overlap is non-zero which allows us to more clearly see the trade-off between accuracy and partial recall of ground-truth perturbagens.

5.9 Results

Overview of datasets and PDGrapher. We have four datasets across two treatment types (genetic and chemical interventions), two cell lines (lung cancer cell line A549 and breast cancer cell line MCF7), and one noisy causal graph: Genetic-PPI-Lung, Genetic-PPI-Breast, Chemical-PPI-Lung, and Chemical-PPI-Breast. Genetic interventions are single-gene knockout experiments mediated by CRISPR/Cas9-mediated gene knockout, while chemical interventions are multiple-gene treatment experiments with chemical compounds. We utilize a PPI network as our noisy causal graph, which has 10,716 nodes and 149,270 edges. Each data set consists of disease and treatment intervention data. Disease intervention data contains paired healthy and diseased gene expression samples, together with disease-associated genes. Treatment intervention data contains paired diseased and treated gene expression samples, together with a genetic or chemical perturbagen. Figures 5.1A and 5.1B shows a summary of the number of samples in each cell line and dataset type.

Given a diseased cell state (gene expression), the goal of PDGrapher is to predict the interventions that, if performed on the diseased cell, would shift the cell to a desired treated state (Figure 5.1C). PDGrapher predicts perturbagens to shift cell states, under the assumption that an optimal perturbagen will be that which shifts the cell gene expression as closely as possible to the desired gene expression. Our approach consists of two modules (Figure 5.1D). First, a perturbagen discovery module f_p takes the initial and desired cell states and

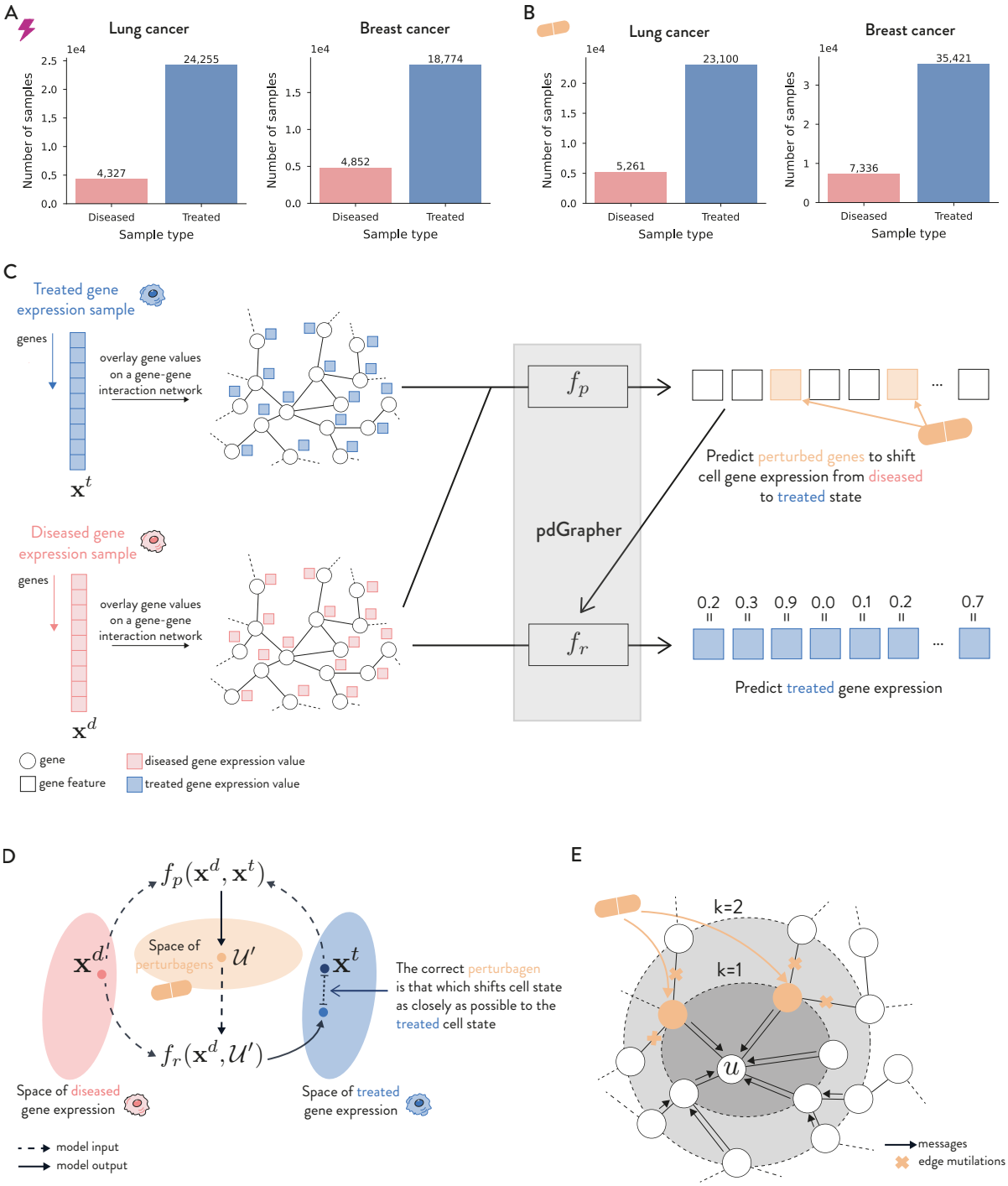


Figure 5.1: Overview of datasets and PDGrapher. **(A,B)** We train PDGrapher across 2 genetic (A) and 2 chemical (B) datasets, where each dataset is comprised by healthy, diseased, and treated gene expression samples for one cell line, and one noisy causal graph (PPI). **(C)** Given a paired diseased and treated gene expression samples, PDGrapher first predicts a candidate perturbagen to shift cell gene expression from diseased to treated state through a perturbagen discovery module f_p. Then PDGrapher takes the diseased gene expression sample and the predicted perturbagen and predicts the treated gene expression through a response prediction module f_r. **(D)** Our perturbagen discovery module f_p is optimized using 2 losses: a cross-entropy cycle loss to predicted a perturbagen $\mathcal{U}'$ which would shift diseased cell state to a state as close as possible to the treated state: $CE\left(\mathbf{x}^t,\ f_r(\mathbf{x}^d,\ f_p(\mathbf{x}^d,\ \mathbf{x}^t))\right)$ $(with\ f_r\ frozen)$, and a cross-entropy supervision loss that directly supervises the prediction of $\mathcal{U}'$: $CE\left(\mathcal{U}', f_p(\mathbf{x}^d,\ \mathbf{x}^t)\right)$. See *Methods* for more details. **(E)** PDGrapher has two modules, f_r and f_p, both of which are GNN-based models. Depicted is f_r, which takes as input a diseased gene expression and a perturbagen and represents perturbagens effects in the noisy causal graph as edge mutilations. Both f_r and f_p follow the standard message-passing framework, where node representations are updated by aggregating the information from neighbors in the graph.

outputs a candidate perturbagen $\mathcal{U}'$. Then a response prediction module f_r takes the initial state and predicted perturbagen $\mathcal{U}'$ and predicts the cell response upon perturbing genes in $\mathcal{U}'$. Our response prediction and perturbagen discovery modules are GNN-based models operating on an input graph, where edge mutilations represent the effects of interventions on our noisy causal graphs (Figure 5.1E).

We have two objective functions, one for each of our modules f_r and f_p. We optimize our response prediction module f_r with all available data on cell state transitions (that is, disease and treatment intervention data). We train f_r so that cell states upon interventions are as close as possible to the known perturbed cell states. We optimize our perturbagen discovery module with treatment intervention data, that is, given a diseased cell state, we predict the intervention set $\mathcal{U}'$ that caused the corresponding treated cell state. Our objective function for the perturbagen discovery module consists of two parts: (1) a cycle loss that optimizes the parameters of f_p such that the response upon intervening on the predicted perturbagen $\mathcal{U}'$, as measured by f_r, is as close as possible to the real treated state, and (2) a supervision loss on the intervention set $\mathcal{U}'$ that directly pushes our model to predict the correct interventions. We train both models simultaneously and perform early stopping separately so that each model finishes training when they have reached convergence.

We evaluate PDGrapher's performance in two separate settings (Figure 5.2A): (1) a random setting where we split samples randomly between training and test sets; and (2) a leave-cell-out setting where we train PDGrapher in one cell line and measure its performance in the remaining cell line.

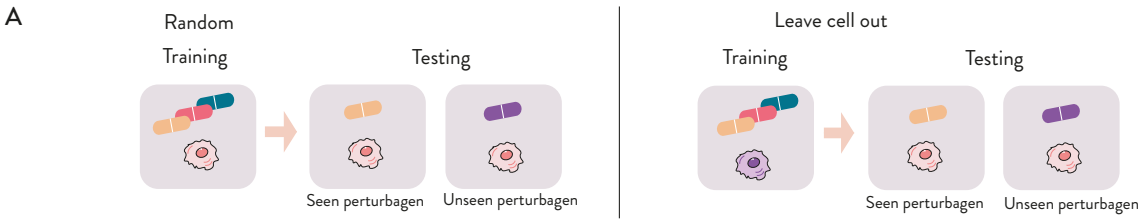
PDGrapher predicts genetic perturbagens to achieve a desired cell state. We assess the ability of PDGrapher for the task of perturbagen discovery across our genetic datasets. Specifically, we measure whether, given paired diseased-treated gene expression samples, PDGrapher can predict the gene that was targeted (knocked-out) in the diseased sample to generate the treated sample (PD accuracy in Figures 5.2B and 5.2C). As an additional indirect metric of PDGrapher's ability to predict perturbagens, we measure whether, given a diseased gene expression sample and a predicted perturbagen, PDGrapher can reconstruct the treated sample given the diseased sample and the predicted perturbagen (R^2 in Figures 5.2B and 5.2C). Intuitively, PDGrapher will predict the treated sample with higher accuracy if the predicted perturbagen is close to the ground-truth perturbagen. For each

perturbagen (single knocked-out gene), we have multiple replicates, therefore, our model might predict different perturbagens for each replicate. To account for this, we report two variants of the R^2 score, one where we compute the average across samples of the R^2 score for each pair of predicted and real treated samples (sample-wise R^2), and one where we compute the average across samples of the R^2 score between the average of all predicted samples and average of all real treated samples for a given perturbagen (perturbagen-wise R^2). We split our datasets into training, validation, and test sets, and report performance metrics, averaged across five-cross-validation splits, in Figures 5.2B and 5.2C.

We compare PDGrapher across our genetic datasets (Genetic-PPI-Lung in Figure 5.2B, Genetic-PPI-Breast in Figure 5.2C) to four methods that can predict interventions: *Random*, *Cancer genes*, *Cancer targets*, and an approach leveraging a response prediction model, *scGen*[307], to search for potential perturbagens (see *Experimental setup* section for more details).

All models except PDGrapher incorrectly predicted the genetic perturbagen (knocked-out gene) in all samples in the test sets (proportion non-zeros = 0.00 ± 0.00). PDGrapher predicts the ground-truth perturbagen in 2% and 1% of the samples in the test set for Genetic-PPI-Lung (PD accuracy = 1.00 ± 0.00; proportion of non-zeros = 0.02 ± 0.00) and Genetic-PPI-Breast (PD accuracy = 1.00 ± 0.00; proportion of non-zeros = 0.01 ± 0.00), respectively. Perhaps more telling than the PD accuracy is the ranking of predicted genes where we see that PDGrapher predicted ground-truth genes 14% and 16% more often than the second-best-performing baseline for Genetic-PPI-Lung (Ranking = 0.65 ± 0.06 vs 0.51 ± 0.00) and Genetic-PPI-Breast (Ranking = 0.67 ± 0.06 vs 0.51 ± 0.01), respectively. In terms of treated sample reconstruction, we observe that scGen has almost identical R^2 scores to PDGrapher in Genetic-PPI-Lung, whereas PDGrapher outperforms scGen by 37% in sample-wise R^2 and by 39% in perturbagen-wise R^2 in Genetic-PPI-Breast. We also note that training time for PDGrapher is approximately 30 times shorter compared to scGen.

We observe that ground-truth genes are consistently at the top of ranking of genes predicted by PDGrapher as compared to the uniform distribution observed in the predicted ranking by the *Random* baseline for both Genetic-PPI-Lung and Genetic-PPI-Breast (Figures 5.2D and 5.2E). Additionally, the predicted perturbagens by PDGrapher are closer in the network to the ground-truth perturbagens than those predicted by the *Random* baseline



B

Model	Sample-wise R^2	Perturbagen-wise R^2	Ranking	PD accuracy	Proportion of non-zeros	Training time (mins)
Random	-	-	0.50 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0
Cancer genes	-	-	0.50 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0
Cancer targets	-	-	0.51 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0
scGen	**0.87 ± 0.00**	**0.90 ± 0.00**	-	0.00 ± 0.00	0.00 ± 0.00	3,644.20
PDGrapher	**0.87 ± 0.03**	0.89 ± 0.03	**0.65 ± 0.06**	**1.00 ± 0.00**	**0.02 ± 0.00**	119.09

C

Model	Sample-wise R^2	Perturbagen-wise R^2	Ranking	PD accuracy	Proportion of non-zeros	Training time (mins)
Random	-	-	0.50 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0
Cancer genes	-	-	0.50 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0
Cancer targets	-	-	0.51 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0
scGen	0.33 ± 0.35	0.34 ± 0.36	-	0.00 ± 0.00	0.00 ± 0.00	1,287.89
PDGrapher	**0.71 ± 0.18**	**0.74 ± 0.17**	**0.67 ± 0.11**	**1.00 ± 0.00**	**0.01 ± 0.00**	135.07

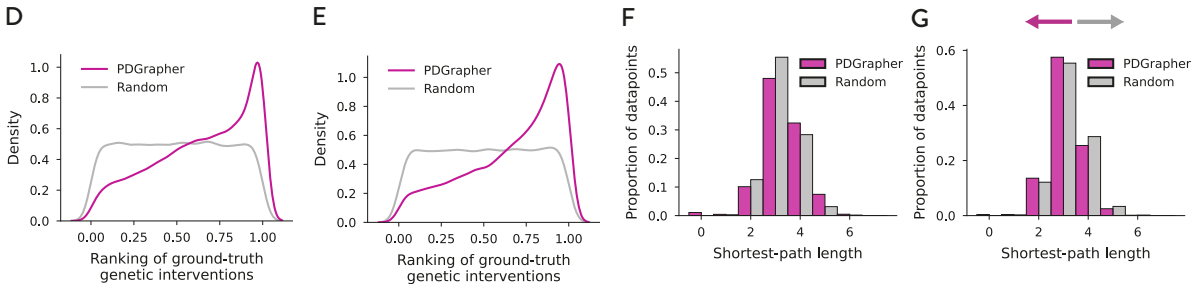


Figure 5.2: PDGrapher predicts genetic perturbagens to achieve a desired cell state. (A) We evaluate PDGrapher's performance across two settings: randomly splitting samples between the training and test sets, and the leave-cell-out setting where we train in one cell line and evaluate PDGrapher's performance in the remaining cell line. **(B,C)** Performance metrics in the perturbagen discovery task for Genetic-PPI-Lung (B) and Genetic-PPI-Breast (C). PD accuracy = perturbagen discovery accuracy; Proportion non-zeros = proportion of samples from the test set in which the predicted perturbagen has an overlap of 0 with the ground truth perturbagen. **(D,E)** Ground-truth interventions (knocked-out genes) are at the top of predicted rankings by PDGrapher more often than would be expected with a *Random* predictor for both Chemical-PPI-Lung (D) and Genetic-PPI-Breast (E). **(F,G)** Interventions predicted by PDGrapher are closer in the network to ground-truth interventions compared to the *Random* baseline for Genetic-PPI-Breast (G), whereas the opposite is observed for Chemical-PPI-Lung (F). *There are dashes in performance tables where it was not possible to obtain a given metric given a model's output.

for Genetic-PPI-Breast (Figures 5.2F and 5.2G).

PDGrapher predicts chemical perturbagens to achieve a desired cell state. We assess the ability of PDGrapher for the task of perturbagen discovery across our chemical datasets. Specifically, we measure whether given paired diseased-treated gene expression samples, PDGrapher can predict the genes that were targeted (by the chemical compound) in the diseased sample to generate the treated sample (PD accuracy in Figures 5.3A and 5.3B). As an additional indirect metric of PDGrapher's ability to predict perturbagens, we measure whether given a diseased gene expression sample and a predicted perturbagen, PDGrapher can reconstruct the treated sample (R^2 in Figures 5.3A and 5.3B). Intuitively, PDGrapher will predict the treated sample with higher accuracy if the predicted perturbagen is close to the ground-truth perturbagen. For each perturbagen (chemical compound), we have multiple replicates, therefore, our model might predict different perturbagens for each replicate. To account for this, we report two variants of the R^2 score, one where we compute the average across samples of the R^2 score for each pair of predicted and real treated samples (sample-wise R^2), and one where we compute the average across samples of the R^2 score between the average of all predicted samples and average of all real treated samples for a given perturbagen (perturbagen-wise R^2). We split our datasets into training, validation, and test sets, and report performance metrics, averaged across 5 cross-validation splits, in Figures 5.3A and 5.3B.

We compare PDGrapher across our chemical datasets (Chemical-PPI-Lung in Figure 5.3A; Chemical-PPI-Breast in Figure 5.3B) to four methods that can predict interventions: *Random*, *Cancer genes*, *Cancer targets*, and an approach leveraging a response prediction model, *scGen*[307], to search for potential perturbagens (see *Experimental setup* section for more details)

scGen shows the highest perturbagen discovery accuracy excluding predicted perturbagens with zero overlap with the ground truth perturbagen (PD accuracy: 0.33 ± 0.07 vs 0.14 ± 0.01 for Chemical-PPI-Lung and 0.35 ± 0.04 vs 0.17 ± 0.01 for Chemical-PPI-Breast). However, scGen predicts perturbagens with an overlap of at least one gene in only 4% and 3% of test samples, compared to 13% and 14% of test samples for PDGrapher in Chemical-PPI-Lung and Chemical-PPI-Breast, respectively. This indicates that even though the overlap is higher for scGen compared to PDGrapher for the predicted perturbagens in which at

least one gene overlaps with the ground-truth perturbagen genes, PDGrapher is able to predict a perturbagen that partially overlaps with the ground truth perturbagen for a higher proportion of the test set.

A

Model	Sample-wise R^2	Perturbagen-wise R^2	Ranking	PD accuracy	Proportion of non-zeros	Training time (mins)
Random	-	-	0.50 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0
Cancer genes	-	-	0.50 ± 0.00	0.10 ± 0.01	0.02 ± 0.00	0
Cancer targets	-	-	0.56 ± 0.00	0.13 ± 0.00	0.05 ± 0.00	0
scGen	0.81 ± 0.00	**0.88 ± 0.00**	-	**0.33 ± 0.07**	0.03 ± 0.01	2,018.23
PDGrapher	**0.82 ± 0.08**	0.85 ± 0.07	**0.84 ± 0.08**	0.14 ± 0.01	**0.14 ± 0.01**	78.20

B

Model	Sample-wise R^2	Perturbagen-wise R^2	Ranking	PD accuracy	Proportion of non-zeros	Training time (mins)
Random	-	-	0.50 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0
Cancer genes	-	-	0.50 ± 0.00	0.13 ± 0.00	0.02 ± 0.00	0
Cancer targets	-	-	0.55 ± 0.00	0.15 ± 0.01	0.05 ± 0.00	0
scGen	0.78 ± 0.00	**0.88 ± 0.00**	-	**0.35 ± 0.04**	0.03 ± 0.00	2,782.65
PDGrapher	**0.81 ± 0.09**	0.85 ± 0.08	**0.90 ± 0.04**	0.17 ± 0.01	**0.15 ± 0.01**	118.72

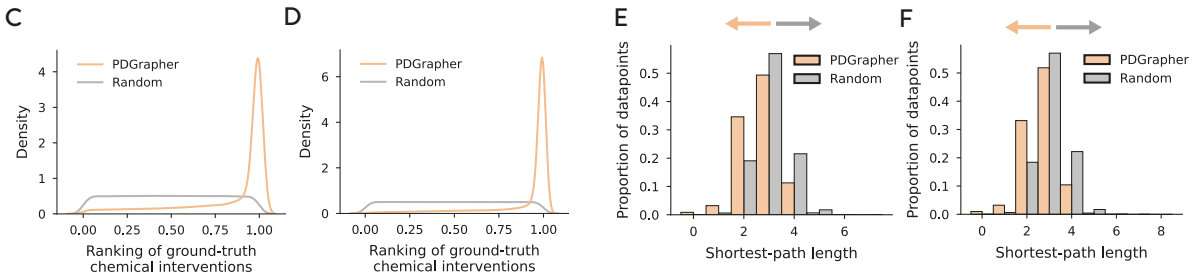


Figure 5.3: PDGrapher predicts chemical perturbagens to achieve a desired cell state. (A, B) Performance metrics in the perturbagen discovery task for Chemical-PPI-Lung (A) and Chemical-PPI-Breast (B). PD accuracy = perturbagen discovery accuracy; Proportion non-zeros = proportion of samples from the test set in which the predicted perturbagen has an overlap of 0 with the ground truth perturbagen. **(C,D)** Ground-truth interventions (targeted genes) are at the top of predicted rankings by PDGrapher more often than would be expected by chance, for both Chemical-PPI-Lung (C) and Chemical-PPI-Breast (D). **(E,F)** Interventions predicted by PDGrapher are closer in the network to ground-truth interventions compared to random, for both Chemical-PPI-Lung (E) and Chemical-PPI-Breast (F). *There are dashes in performance tables where it was not possible to obtain a given metric given a model's output.

We observe that the ground-truth genes are very close to the top of the predicted gene list by PDGrapher, with ranking scores of 0.84 ± 0.08 and 0.90 ± 0.04 for Chemical-PPI-Lung and Chemical-PPI-Breast, respectively. In terms of treated sample reconstruction, we observe that scGen and PDGrapher have almost identical performance. We also note that training time for PDGrapher is approximately 23 times smaller compared to scGen.

We observe that ground-truth interventions are consistently at the top of ranking of genes predicted by PDGrapher as compared to the uniform distribution observed in the predicted ranking by the *Random* baseline for both Genetic-PPI-Lung and Genetic-PPI-Breast (Figures 5.3C and 5.3D). Additionally, the predicted perturbagens by PDGrapher are closer in the network to the ground-truth perturbagens than those predicted by the

Random baseline for Genetic-PPI-Lung and Genetic-PPI-Breast (Figures 5.3E and 5.3F).

PDGrapher can learn optimal perturbagens to achieve a desired cell state for unseen cell lines. We assess PDGrapher's ability in the task of perturbagen discovery across our genetic and chemical datasets in unseen cell lines. In this evaluation setting (leave cell out in Figure 5.2A), we optimize our model in one cell line and measure predictive performance in the remaining cell line. For each perturbagen type (genetic/chemical), we use each of the optimized models in one cell line in the random setting to predict perturbagens for another cell line. We report performance metrics, averaged across 5 splits, in Figures 5.4A and 5.4B and Figures 5.5A and 5.5B. Overall, we observe similar trends across our genetic and chemical datasets compared to the random setting. PDGrapher shows the highest ranking performance across all genetic and chemical datasets.

A

Model	Sample-wise R^2	Perturbagen-wise R^2	Ranking	PD accuracy	Proportion of non-zeros
Random	-	-	0.50 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Cancer genes	-	-	0.50 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Cancer targets	-	-	0.50 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
scGen	0.22 ± 0.24	0.24 ± 0.27	-	0.00 ± 0.00	0.00 ± 0.00
PDGrapher	**0.46 ± 0.04**	**0.53 ± 0.04**	**0.64 ± 0.09**	**1.0 ± 0.0**	**0.01 ± 0.00**

B

Model	Sample-wise R^2	Perturbagen-wise R^2	Ranking	PD accuracy	Proportion of non-zeros
Random	-	-	0.50 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Cancer genes	-	-	0.50 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Cancer targets	-	-	0.50 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
scGen	**0.58 ± 0.00**	0.62 ± 0.00	-	0.00 ± 0.00	0.00 ± 0.00
PDGrapher	**0.58 ± 0.09**	**0.63 ± 0.09**	**0.65 ± 0.07**	**1.0 ± 0.0**	**0.01 ± 0.00**

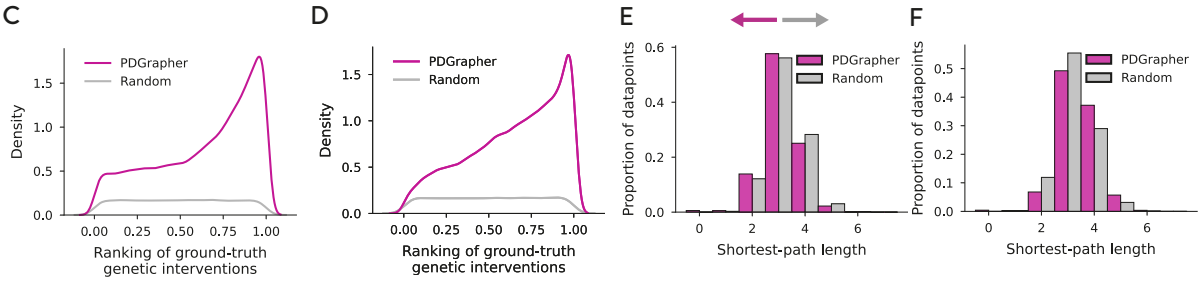


Figure 5.4: PDGrapher predicts genetic perturbagens to achieve a desired cell state for new cell lines. (A,B) Performance metrics in the perturbagen discovery task in the leave-cell-out setting for Genetic-PPI-Lung (A) and Genetic-PPI-Breast (B). PD accuracy = perturbagen discovery accuracy; Proportion non-zeros = proportion of samples from the test set in which the predicted perturbagen has an overlap of 0 with the ground truth perturbagen. **(C,D)** Ground-truth interventions (knocked-out genes) are at the top of predicted rankings by PDGrapher more often than would be expected by chance, for both Genetic-PPI-Lung (C) and Genetic-PPI-Breast (D) **(E,F)** Interventions predicted by PDGrapher are closer in the network to ground-truth interventions compared to the *Random* baseline, for Genetic-PPI-Lung (D), whereas the opposite is observed for Genetic-PPI-Breast (F). *There are dashes in performance tables where it was not possible to obtain a given metric given a model's output.

PDGrapher also shows the highest treated sample reconstruction performance and perturbagen discovery accuracy for both Genetic-PPI-Lung and Genetic-PPI-Breast. We observe the same trade-off between perturbagen discovery accuracy for predictions with at least one overlapping gene with the ground-truth perturbagen and the proportion of predicted perturbagens with at least one gene overlapping with the ground-truth perturbagen. For the predicted perturbagens with at least one overlapping gene with the ground-truth perturbagen scGen shows a higher accuracy compared to PDGrapher; however, PDGrapher predicts a higher proportion of perturbagens with at least one overlapping gene with ground-truth perturbagens.

We observe that ground-truth interventions are consistently at the top of the ranking of genes predicted by PDGrapher as compared to the uniform distribution observed in the predicted ranking by the *Random* baseline across all datasets (Figures 5.3C and 5.3D; Figures 5.4C and 5.4D). Additionally, the predicted perturbagens by PDGrapher are closer in the network to the ground-truth perturbagens than those predicted by the *Random* baseline for Genetic-PPI-Lung, Chemical-PPI-Lung, and Chemical-PPI-Breast (Figures 5.3E and 5.3F; Figures 5.4E and 5.4F).

PDGrapher's performance across different training settings. We benchmarked three variants of PDGrapher in the random training setting: PDGrapher-SuperCycle trained with the cycle and supervision loss, PDGrapher-Cycle trained only with the cycle loss, and PDGrapher-Super trained only with the supervision loss (Figure 5.6). Across all datasets, PDGrapher-Super achieved the highest perturbagen discovery accuracy, with the highest proportion of predicted perturbagens having at least one overlapping gene with the ground-truth perturbagen. Additionally, PDGrapher-Super also exhibits the best predicted gene ranking. In terms of treated sample reconstruction, PDGrapher-Cycle shows the highest R^2 performance in genetic datasets, whereas PDGrapher-SuperCycle shows the highest R^2 performance in chemical datasets.

A

Model	Sample-wise R^2	Perturbagen-wise R^2	Ranking	PD accuracy	Proportion of non-zeros
Random	-	-	0.50 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Cancer genes	-	-	0.50 ± 0.00	0.11 ± 0.00	0.02 ± 0.00
Cancer targets	-	-	0.55 ± 0.00	0.13 ± 0.00	0.04 ± 0.00
scGen	**0.56 ± 0.00**	**0.58 ± 0.00**	-	**0.38 ± 0.07**	0.04 ± 0.01
PDGrapher	0.43 ± 0.14	0.49 ± 0.14	**0.90 ± 0.04**	0.16 ± 0.01	**0.13 ± 0.01**

B

Model	Sample-wise R^2	Perturbagen-wise R^2	Ranking	PD accuracy	Proportion of non-zeros
Random	-	-	0.50 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Cancer genes	-	-	0.50 ± 0.00	0.11 ± 0.01	0.03 ± 0.00
Cancer targets	-	-	0.50 ± 0.00	0.17 ± 0.00	0.05 ± 0.00
scGen	**0.52 ± 0.00**	**0.57 ± 0.00**	-	**0.35 ± 0.02**	0.03 ± 0.00
PDGrapher	0.49 ± 0.15	0.55 ± 0.14	**0.65 ± 0.00**	0.14 ± 0.00	**0.14 ± 0.01**

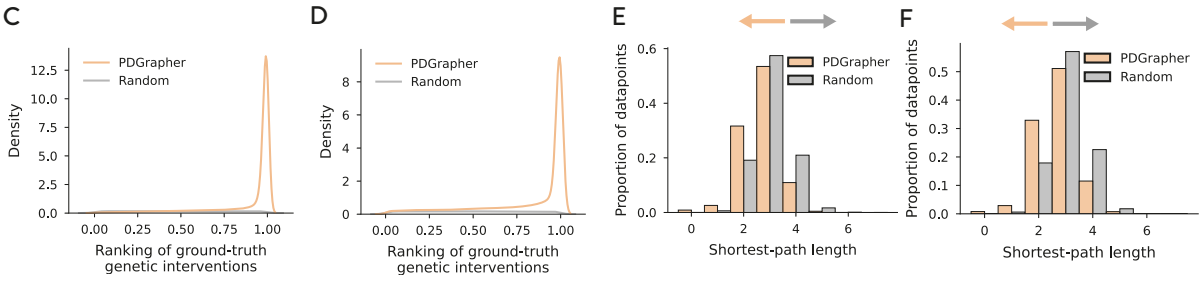


Figure 5.5: PDGrapher predicts chemical perturbagens to achieve a desired cell state for new cell lines. (A,B) Performance metrics in the perturbagen discovery task in the leave-cell-out setting for Chemical-PPI-Lung (A) and Chemical-PPI-Breast (B). PD accuracy = perturbagen discovery accuracy; Proportion non-zeros = proportion of samples from the test set in which the predicted perturbagen has an overlap of 0 with the ground truth perturbagen. **(C,D)** Ground-truth interventions (targeted genes) are at the top of predicted rankings by PDGrapher more often than would be expected by chance, for both Chemical-PPI-Lung (C) and Chemical-PPI-Breast (D). **(E,F)** Interventions predicted by PDGrapher are closer in the network to ground-truth interventions compared to the *Random* baseline, for both Chemical-PPI-Lung (E) and Chemical-PPI-Breast (F). *There are dashes in performance tables where it was not possible to obtain a given metric given a model's output.

A

Model	Sample-wise R^2	Perturbagen-wise R^2	Ranking	PD accuracy	Proportion of non-zeros
PDGrapher - Cycle	**0.89 ± 0.04**	**0.91 ± 0.04**	0.51 ± 0.03	0.00 ± 0.00	0.00 ± 0.00
PDGrapher - SuperCycle	0.87 ± 0.03	0.89 ± 0.03	0.65 ± 0.06	1.00 ± 0.00	0.02 ± 0.00
PDGrapher - Super	0.10 ± 0.11	0.10 ± 0.12	**0.84 ± 0.00**	1.00 ± 0.00	0.02 ± 0.00

B

Model	Sample-wise R^2	Perturbagen-wise R^2	Ranking	PD accuracy	Proportion of non-zeros
PDGrapher - Cycle	**0.87 ± 0.07**	**0.89 ± 0.06**	0.52 ± 0.03	0.00 ± 0.00	0.00 ± 0.00
PDGrapher - SuperCycle	0.71 ± 0.18	0.74 ± 0.17	0.67 ± 0.11	1.00 ± 0.00	0.01 ± 0.00
PDGrapher - Super	0.27 ± 0.15	0.28 ± 0.15	**0.86 ± 0.00**	1.00 ± 0.00	0.01 ± 0.00

C

Model	Sample-wise R^2	Perturbagen-wise R^2	Ranking	PD accuracy	Proportion of non-zeros
PDGrapher - Cycle	0.52 ± 0.23	0.58 ± 0.23	0.51 ± 0.05	0.07 ± 0.03	0.01 ± 0.00
PDGrapher - SuperCycle	**0.82 ± 0.08**	**0.85 ± 0.07**	0.84 ± 0.08	0.14 ± 0.01	0.14 ± 0.01
PDGrapher - Super	0.01 ± 0.01	0.01 ± 0.02	**0.99 ± 0.00**	**0.16 ± 0.01**	**0.15 ± 0.00**

D

Model	Sample-wise R^2	Perturbagen-wise R^2	Ranking	PD accuracy	Proportion of non-zeros
PDGrapher - Cycle	0.58 ± 0.33	0.61 ± 0.32	0.50 ± 0.02	0.08 ± 0.03	0.01 ± 0.00
PDGrapher - SuperCycle	**0.81 ± 0.09**	**0.85 ± 0.08**	0.90 ± 0.04	0.17 ± 0.01	0.15 ± 0.01
PDGrapher - Super	0.00 ± 0.00	0.01 ± 0.01	**0.99 ± 0.00**	**0.20 ± 0.00**	0.15 ± 0.00

Figure 5.6: Ablation of PDGrapher's loss components. **(A)** Performance metrics of PDGrapher in Genetic-PPI-Lung. **(B)** Performance metrics of PDGrapher in Genetic-PPI-Breast. **(C)** Performance metrics of PDGrapher in Chemical-PPI-Lung. **(D)** Performance metrics of PDGrapher in Genetic-PPI-Lung.

5.10 Summary

In this work, we address the problem of therapeutic lead discovery. While target-driven approaches have dominated the lead discovery landscape for the past three decades, interest in phenotype-driven approaches has resurfaced in recent years as a way to go beyond the one drug, one gene, one disease paradigm predominant in target-driven approaches. In phenotype-driven approaches, the goal is, given a *signature* characterizing the effects of disease, to find a *perturbagen* or *intervention* reversing those effects. Existing target-driven approaches compare disease and compound signatures to propose maximally anti-correlated compounds as potential leads. This, however, limits the search space to existing curated libraries of compounds. We introduce a novel problem formulation for phenotype-driven lead discovery in which the goal is, given a *signature* characterizing the effects of disease, to find what genes a *new perturbagen* should target to reverse those effects.

We introduce PDGrapher, a causally inspired model for perturbagen design. Given an

initial cell state represented as a gene expression signature, PDGrapher predicts what genes should be targeted to shift the cell to a desired target state. PDGrapher has two modules, a perturbagen discovery module, which proposes an set of therapeutic targets (genes) given the initial and desired cell gene expression states, and a response discovery module, which measures the response of applying the predicted perturbagen on the initial cell state. Both modules are GNN-based models operating on gene-gene networks representing a noisy causal graph where interventions are represented using mutilations in the graph.

We evaluate PDGrapher's performance in two settings: (i) randomly splitting samples into training and test sets, and (ii) training with one cell line and predicting in a different cell line. PDGrapher successfully predicts perturbagens to achieve a desired gene expression cell state across four genetic and chemical datasets in the random setting (Figures 5.2B, and 5.2C; Figures 5.3A, and 5.3B). PDGrapher's performance is lower in the leave-cell-out setting, however, it still still higher than existing baselines (Figures 5.4A, and 5.4B; Figures 5.5A, and 5.5B).

Our results show the promise of developing methods for therapeutic lead discovery to expand the search space and guide lead design. PDGrapher was tested to shift cell line gene expression from an initial to a desired state using large datasets of genetic and chemical interventions. By providing the set of genes to be intervened-on, rather than a perturbagen itself, we can increase the flexibility and power of phenotype-driven lead discovery.

Certain assumptions and limitations of our work could be addressed in the future. Throughout, we assumed a Markovianity setting (implying no unobserved confounders) which is a strong constraint and often difficult or impossible to test. Future work could relax this assumption during the problem formulation phase, leading to an updated problem formulation and approach. Another limitation of our work is the use of PPIs as an approximation of the underlying causal graph of genes. This graph is undoubtedly noisy and incomplete. Here, we hypothesized that a representation learning approach would circumvent the approximations in the causal graph. Theoretically determining the effects of using a causal graph approximation in the predicted likelihoods appears to be a natural future direction.

It is important to address whether our study led to the discovery of any novel lead therapeutics. The primary objective of this study was to develop and validate PDGrapher, a novel model for phenotype-driven therapeutic lead discovery. While the focus was on the

methodology and its potential to innovate the lead discovery process, the direct discovery of novel lead therapeutics was not the central aim of this particular study. However, PDGrapher has demonstrated significant promise in accurately predicting therapeutic targets for existing therapeutics, laying a robust foundation for future experimental validations and drug development processes. The model's capability to propose therapeutic targets based on gene expression signatures marks an important step forward in identifying novel leads. Future research, building upon this groundwork, could leverage PDGrapher to discover and validate new therapeutic agents, fulfilling the broader aim of advancing phenotype-driven approaches in drug discovery. This study, thus, serves as a crucial stepping stone towards the practical application of our model in identifying novel, effective therapeutic leads.

6 Conclusions

6.1 Summary of contributions

In this dissertation, we developed techniques for computational lead identification within the drug discovery pipeline. Given the renewed interested in natural compounds as potential therapeutic leads, we first address our research question Q1 (can we build graph deep learning algorithms that leverage natural compounds for drug discovery?) in Chapters 3 and 4. To answer research question Q1, we developed methods to predict plant-based molecules with therapeutic properties alone (Chapter 3) or in combination (Chapter 4) with existing drugs. Then, we turned out focus to the task of lead identification in research question Q2 (given the limiations of the current lead identification problem formulation, can we propose a novel formulation for network-based lead identification together with an algorithm to tackle it?). To answer this question, we outlined the limitations of the current lead identification problem formulation and proposed a novel problem formulation to overcome said limitations, together with a causally grounded algorithm to address it (Chapter 5).

- In Chapter 3, we addressed our research question **Q1.1: can graph deep learning algorithms surpass traditional network-based algorithms in the prediction of plant-based compounds with therapeutic properties? In addition, are these methods able to capture inductive biases that exist in biological data?.** To address this question, we introduced a graph deep learning model to predict plant-based molecules with anticancer properties. We proposed an end-to-end graph neural model that operates directly on graphs to (i) learn how to propagate information across the PPI network to capture effects of drugs and food molecules on proteins and (ii) identify anticancer molecules from the learned diffusion profiles. Our proposed GNN model outperformed the baseline approach in the anticancer therapeutic prediction task by up to 67%. Using attribution methods, we also showed that our proposed model captures knowledge of biological pathways to predict anticancer molecules based on the molecules' effects on cancer-related protein communities. This project, along with our previous related work[121], are part of the large-scale DreamLab project funded by the Vodafone Foundation which has resulted in considerable public engagement with

the publication of a Hyperfoods cookbook with recipes rich in our predicted anticancer compounds. The Hyperfoods cookbook can be found here.

- In Chapter 4, we addressed our research question **Q1.2: can we build a graph deep learning algorithm to predict plant-based compounds that synergize with existing drugs?.** To address this question, we introduced a GNN-based model to predict plant-based molecules that synergize with existing FDA-approved anticancer drugs. Our model takes as input a plant-based molecule, a drug, and a phenotype (cancer cell line) and outputs the predicted synergy score between the plant-based molecule and drug against the phenotype. We validated this model computationally by training and testing the model for the prediction of drug-drug synergy, and experimentally by testing synergy in the laboratory. In our experimental validations, 8/11 drug-drug combinations tested were consistent with predictions, and 3/3 plant-based molecule-drug combinations tested were consistent with predictions. This project is part of the large-scale DreamLab project funded by the Vodafone Foundation. Our results serve as a proof of concept for the large-scale prediction of interactions between plant-based molecules and drugs. This work is expected to be extended to other disease areas and developed further to propose dietary recommendations for patients.

- In Chapter 5, we addressed our research question **Q2: given the limitations of the current lead identification problem formulation, can we propose a novel formulation for network-based lead identification together with an algorithm to tackle it?.** To address this question, we first outlined the limitations of the current network-based lead discovery problem formulation. Then, we introduced a novel problem formulation for the lead discovery task and a causally inspired GNN-based model to predict which genes a lead perturbagen (genetic or chemical agent) should target to reverse the effects of disease. Our model was able to successfully predict targeted genes to shift cell line gene expression from an initial to a desired target state across various genetic and chemical perturbation datasets. We believe this work opens the door for the development of new, more flexible, methods that can leverage high-throughput screening data for the task of lead discovery where new potential leads are not limited to compounds from a curated library.

6.2 Future directions

During the course of the final project presented in this dissertation (Chapter 5), the importance of perturbational data in biology has become increasingly clear. Perturbations are ubiquitous in biology, manifesting in various forms such as diseases, drugs, knockout experiments, and mutations. These perturbations can affect genes, proteins, and molecules, leading to alterations in gene expression, cell function, and cell phenotype. At the molecular level, changes in aminoacid sequences or molecular atoms can impact the conformation and function of proteins and molecules, respectively. Perturbations introduce changes in a system, which can have downstream effects that may be beneficial or detrimental. In the context of drug discovery, the response of a system to perturbations is crucial to evaluating and optimizing potential therapies, for instance, determining whether the druggability of a molecule is enhanced by modifying one of its atoms. Moreover, identifying the elements of a system that would need to be perturbed or intervened on to achieve a desired outcome is essential to discovering new therapies.

Exploiting interventional data in biology is the most promising research direction I hope to contribute meaningfully in the years following the submission of this dissertation. To do this, I believe that developing methods that combine the rigor of causal inference with the flexibility and power of deep learning is key. In particular, I would like to explore method development and applications within the recent trend of causal representation learning[318]. Some specific examples of future directions I would like to work on are:

- **Exploring the connection between SCMs and GNNs.** Both SCMs and GNNs are powerful frameworks for modeling complex systems and making predictions about their behavior. SCMs provide a formal language for specifying the causal relationships between variables in a system, and they enable the identification of causal effects and interventions. GNNs, on the other hand, are a class of neural networks that operate on graph-structured data, enabling the modeling of relationships between entities in a system. The connection between SCMs and GNNs lies in their shared focus on modeling complex systems as graphs, and their ability to leverage the graph structure to make predictions and inferences. Recent work has shown the possibility of modeling structural equations in SCMs using neural networks, enabling fast computa-

tion of causal effects and counterfactual inference[318,321,322]. Given that GNNs operate on graph-structured data, it is reasonable to think that an integration of SCMs with GNNs would be smoother and provide more computationally tractable models. This integration could provide a powerful framework for modeling and analyzing complex systems, by combining the causal reasoning of SCMs with the flexible and expressive modeling capabilities of GNNs. This could benefit various fields where data is represented as complex graph-based systems and the rigor of causal inference is needed such as in biology, economics, and social science.

- **Leveraging perturbational data for protein design.** Perturbational data involves the systematic introduction of mutations in aminoacids of a protein which translates into changes in sequence, and potentially structure, followed by measurements of the resulting changes in protein properties, such as activity, binding affinity, or thermodynamic stability. I am interested in causal inference in this scenario and in investigating how perturbational data can be leveraged for protein design by providing insights into the effects of sequence and structural changes on protein function and stability.

7 Appendix to Chapter 3

7.1 Neratinib and Calcitriol over-represented pathways

Table 7.1: We compute embeddings of the top correctly (neratinib) and incorrectly (calcitriol) classified drugs as anticancer for our best-performing model (ChebNet) and obtain the top 25 over-represented pathways using GSEA[152].

Molecule	Pathways
Neratinib	KEGG ERBB SIGNALING PATHWAY
	KEGG MAPK SIGNALING PATHWAY
	KEGG GLIOMA
	KEGG PROSTATE CANCER
	KEGG PANCREATIC CANCER
	KEGG NON SMALL CELL LUNG CANCER
	KEGG ENDOMETRIAL CANCER
	KEGG NEUROTROPHIN SIGNALING PATHWAY
	KEGG ACUTE MYELOID LEUKEMIA
	KEGG FC EPSILON RI SIGNALING PATHWAY
	KEGG T CELL RECEPTOR SIGNALING PATHWAY
	KEGG RENAL CELL CARCINOMA
	KEGG CHRONIC MYELOID LEUKEMIA
	KEGG B CELL RECEPTOR SIGNALING PATHWAY
	KEGG INSULIN SIGNALING PATHWAY
	KEGG FOCAL ADHESION
	KEGG MELANOMA
	KEGG BLADDER CANCER

	KEGG GNRH SIGNALING PATHWAY
	KEGG VEGF SIGNALING PATHWAY
	KEGG FC GAMMA R MEDIATED PHAGOCYTOSIS
	KEGG PATHWAYS IN CANCER
	KEGG COLORECTAL CANCER
	KEGG MTOR SIGNALING PATHWAY
	KEGG TOLL LIKE RECEPTOR SIGNALING PATHWAY
Calcitriol	KEGG RIBOSOME
	KEGG SPLICEOSOME
	KEGG PROTEASOME
	KEGG PARKINSONS DISEASE
	KEGG SNARE INTERACTIONS IN VESICULAR TRANSPORT
	KEGG HUNTINGTONS DISEASE
	KEGG OXIDATIVE PHOSPHORYLATION
	KEGG PROPANOATE METABOLISM
	KEGG BASAL TRANSCRIPTION FACTORS
	KEGG ALANINE ASPARTATE AND GLUTAMATE METABOLISM
	KEGG TRYPTOPHAN METABOLISM
	KEGG GLYCINE SERINE AND THREONINE METABOLISM
	KEGG PYRUVATE METABOLISM
	KEGG RNA POLYMERASE
	KEGG AMINOACYL TRNA BIOSYNTHESIS
	KEGG BETA ALANINE METABOLISM
	KEGG ARGININE AND PROLINE METABOLISM
	KEGG GLYOXYLATE AND DICARBOXYLATE METABOLISM

	KEGG VALINE LEUCINE AND ISOLEUCINE DEGRADATION
	KEGG ALZHEIMERS DISEASE
	KEGG BUTANOATE METABOLISM
	KEGG ETHER LIPID METABOLISM
	KEGG NOTCH SIGNALING PATHWAY
	KEGG NUCLEOTIDE EXCISION REPAIR
	KEGG RNA DEGRADATION

7.2 Over-represented pathways for top predicted anticancer molecules

Table 7.2: Top over-represented pathways in KEGG[103] using GSEA[152] computed from embeddings of food molecules with highest anticancer likeness

Name	**Top 10 KEGG pathways**
Pancreatin	SMALL CELL LUNG CANCER, BASAL TRANSCRIPTION FACTORS, CHRONIC MYELOID LEUKEMIA, ARRHYTHMOGENIC RIGHT VENTRICULAR CARDIOMYOPATHY ARVC, ECM RECEPTOR INTERACTION, GLIOMA, INSULIN SIGNALING PATHWAY, CELL CYCLE, ADHERENS JUNCTION, FOCAL ADHESION
Anthracene	BLADDER CANCER, MELANOMA, NON SMALL CELL LUNG CANCER, ENDOMETRIAL CANCER, GLIOMA, SNARE INTERACTIONS IN VESICULAR TRANSPORT, ERBB SIGNALING PATHWAY, PROSTATE CANCER, DORSO VENTRAL AXIS FORMATION, ACUTE MYELOID LEUKEMIA
2, 2'-Bis(4-hydroxyphenyl) propane	STEROID HORMONE BIOSYNTHESIS, RETINOL METABOLISM, DRUG METABOLISM CYTOCHROME P450, DRUG METABOLISM OTHER ENZYMES, PROSTATE CANCER, STARCH AND SUCROSE METABOLISM, GLYCINE SERINE AND THREONINE METABOLISM, ASCORBATE AND ALDARATE METABOLISM, PENTOSE AND GLUCURONATE INTERCONVERSIONS, ALDOSTERONE REGULATED SODIUM REABSORPTION

Coumestrol	CELL CYCLE, MISMATCH REPAIR, DNA REPLICATION, T CELL RECEPTOR SIGNALING PATHWAY, STEROID HORMONE BIOSYNTHESIS, RIBOSOME, NUCLEOTIDE EXCISION REPAIR, COMPLEMENT AND COAGULATION CASCADES, PENTOSE AND GLUCURONATE INTERCONVERSIONS, B CELL RECEPTOR SIGNALING PATHWAY
Genistein	PROSTATE CANCER, NON SMALL CELL LUNG CANCER, PATHWAYS IN CANCER, GLIOMA, STEROID HORMONE BIOSYNTHESIS, ABC TRANSPORTERS, PENTOSE AND GLUCURONATE INTERCONVERSIONS, MELANOGENESIS, ENDOMETRIAL CANCER, ASCORBATE AND ALDARATE METABOLISM
Pterostilbene	COLORECTAL CANCER, PROSTATE CANCER, ENDOMETRIAL CANCER, DORSO VENTRAL AXIS FORMATION, NON SMALL CELL LUNG CANCER, CHRONIC MYELOID LEUKEMIA, APOPTOSIS, GLIOMA, PANCREATIC CANCER, MTOR SIGNALING PATHWAY
Mercenene	RIBOSOME, ERBB SIGNALING PATHWAY, CHRONIC MYELOID LEUKEMIA, UBIQUITIN MEDIATED PROTEOLYSIS, RNA POLYMERASE, REGULATION OF ACTIN CYTOSKELETON, BIOSYNTHESIS OF UNSATURATED FATTY ACIDS, VEGF SIGNALING PATHWAY, NOD LIKE RECEPTOR SIGNALING PATHWAY, FOCAL ADHESION
Myristicin	HOMOLOGOUS RECOMBINATION, DNA REPLICATION, NUCLEOTIDE EXCISION REPAIR, TERPENOID BACKBONE BIOSYNTHESIS, NITROGEN METABOLISM, MISMATCH REPAIR, CELL CYCLE, BASE EXCISION REPAIR, RIBOSOME, PPAR SIGNALING PATHWAY
2, 5-Dihydro-4, 5-dimethyl-2-(1-methylpropyl)thiazole	ABC TRANSPORTERS, HYPERTROPHIC CARDIOMYOPATHY HCM, PANCREATIC CANCER, GLYOXYLATE AND DICARBOXYLATE METABOLISM, PRION DISEASES, NON SMALL CELL LUNG CANCER, RNA DEGRADATION, DORSO VENTRAL AXIS FORMATION, RETINOL METABOLISM, GLUTATHIONE METABOLISM
1-Methyl-6-phenyl-1H-imidazo[4, 5-b]pyridin-2-amine	BLADDER CANCER, RETINOL METABOLISM, PRION DISEASES, GLIOMA, MATURITY ONSET DIABETES OF THE YOUNG, PROSTATE CANCER, NON SMALL CELL LUNG CANCER, THYROID CANCER, CELL CYCLE, P53 SIGNALING PATHWAY

5a-Androstane-3a, 17a-diol	STEROID HORMONE BIOSYNTHESIS, PENTOSE AND GLUCURONATE INTERCONVERSIONS, ASCORBATE AND ALDARATE METABOLISM, PORPHYRIN AND CHLOROPHYLL METABOLISM, RETINOL METABOLISM, METABOLISM OF XENOBIOTICS BY CYTOCHROME P450, DRUG METABOLISM OTHER ENZYMES, STARCH AND SUCROSE METABOLISM, VIRAL MYOCARDITIS, DRUG METABOLISM CYTOCHROME P450
Isoursodeoxycholic acid	TOLL LIKE RECEPTOR SIGNALING PATHWAY, TGF BETA SIGNALING PATHWAY, THYROID CANCER, MTOR SIGNALING PATHWAY, ADHERENS JUNCTION, CITRATE CYCLE TCA CYCLE, PROTEIN EXPORT, OOCYTE MEIOSIS, ANTIGEN PROCESSING AND PRESENTATION, TYPE I DIABETES MELLITUS
11alpha-Hydroxyprogesterone	ADHERENS JUNCTION, THYROID CANCER, WNT SIGNALING PATHWAY, PATHWAYS IN CANCER, PRION DISEASES, LYSOSOME, BASAL CELL CARCINOMA, TGF BETA SIGNALING PATHWAY, STEROID HORMONE BIOSYNTHESIS, DNA REPLICATION
Dihydrotestosterone	STEROID HORMONE BIOSYNTHESIS, ADIPOCYTOKINE SIGNALING PATHWAY, VASOPRESSIN REGULATED WATER REABSORPTION, ALANINE ASPARTATE AND GLUTAMATE METABOLISM, LEISHMANIA INFECTION, TGF BETA SIGNALING PATHWAY, PPAR SIGNALING PATHWAY, THYROID CANCER, HISTIDINE METABOLISM, PROSTATE CANCER
Testosterone enanthate	PROTEASOME, PYRUVATE METABOLISM, CITRATE CYCLE TCA CYCLE, PROGESTERONE MEDIATED OOCYTE MATURATION, CYTOSOLIC DNA SENSING PATHWAY, GAP JUNCTION, GLYCOLYSIS GLUCONEOGENESIS, CALCIUM SIGNALING PATHWAY, TASTE TRANSDUCTION, ACUTE MYELOID LEUKEMIA
5b-Dihydrotestosterone	THYROID CANCER, PPAR SIGNALING PATHWAY, PATHOGENIC ESCHERICHIA COLI INFECTION, PRION DISEASES, ARRHYTHMOGENIC RIGHT VENTRICULAR CARDIOMYOPATHY ARVC, UBIQUITIN MEDIATED PROTEOLYSIS, BUTANOATE METABOLISM, PROPANOATE METABOLISM, RNA DEGRADATION, HYPERTROPHIC CARDIOMYOPATHY HCM
Methylarsonite	PANCREATIC CANCER, DORSO VENTRAL AXIS FORMATION, GLIOMA, PROSTATE CANCER, CHRONIC MYELOID LEUKEMIA, NON SMALL CELL LUNG CANCER, FOCAL ADHESION, ALDOSTERONE REGULATED SODIUM REABSORPTION, P53 SIGNALING PATHWAY, SMALL CELL LUNG CANCER

Platinum	COMPLEMENT AND COAGULATION CASCADES, PATHWAYS IN CANCER, P53 SIGNALING PATHWAY, GLIOMA, PROSTATE CANCER, MELANOMA, NON SMALL CELL LUNG CANCER, HEMATOPOIETIC CELL LINEAGE, FOCAL ADHESION, SMALL CELL LUNG CANCER
17- - hydroxypregnenolone	RNA POLYMERASE, ADHERENS JUNCTION, PYRIMIDINE METABOLISM, BLADDER CANCER, MATURITY ONSET DIABETES OF THE YOUNG, PURINE METABOLISM, ENDOMETRIAL CANCER, NON SMALL CELL LUNG CANCER, CYTOKINE CYTOKINE RECEPTOR INTERACTION, GLYCOSYLPHOSPHATIDYLINOSITOL GPI ANCHOR BIOSYNTHESIS
androst-4-en-3, 17-dione	STEROID HORMONE BIOSYNTHESIS, RETINOL METABOLISM, METABOLISM OF XENOBIOTICS BY CYTOCHROME P450, PARKINSONS DISEASE, TOLL LIKE RECEPTOR SIGNALING PATHWAY, ABC TRANSPORTERS, PPAR SIGNALING PATHWAY, TGF BETA SIGNALING PATHWAY, PROXIMAL TUBULE BICARBONATE RECLAMATION, NEUROACTIVE LIGAND RECEPTOR INTERACTION

8 Appendix to Chapter 4

8.1 Drugs and cell lines used in training

Table 8.3: We extracted drug-drug interaction data of the Merck and Co's study[180], and filtered samples to include only those with available drug targets and cell line mutations. The final set contains 546 unique pairwise drug interactions between 37 drugs in 32 cell lines. This table contains such drugs and cell lines.

Drugs	**Cell lines**
ZOLINZA	LOVO
ETOPOSIDE	MSTO
TEMOZOLOMIDE	HT144
METHOTREXATE	A427
MRK-003	UACC62
MK-2206	UWB1289
AZD1775	A2780
MK-8669	NCIH23
VINBLASTINE	NCI-H460
SN-38	HCT116
PACLITAXEL	SW-620
SUNITINIB	A375
GELDANAMYCIN	MDAMB436
PD325901	NCIH2122
ERLOTINIB	T-47D
LAPATINIB	VCAP
METFORMIN	A2058
MK-5108	PA1
VINORELBINE	NCIH1650
BORTEZOMIB	SW837
MITOMYCINE	RKO
GEMCITABINE	SK-OV-3

L778123	OV90
CARBOPLATIN	HT29
DASATINIB	CAOV3
DOXORUBICIN	SKMEL30
MK-4827	SKMES1
5-FU	OVCAR3
DEXAMETHASONE	RPMI7951
CYCLOPHOSPHAMIDE	OCUBM
BEZ-235	ES2
OXALIPLATIN	NCIH520
ABT-888	
TOPOTECAN	
SORAFENIB	
DINACICLIB	
MK-8776	

8.2 Synergy experiments results

The following are the synergy matrices as outputted by SynergyFinder.

8.2.1 Drug-drug combinations tested in A549

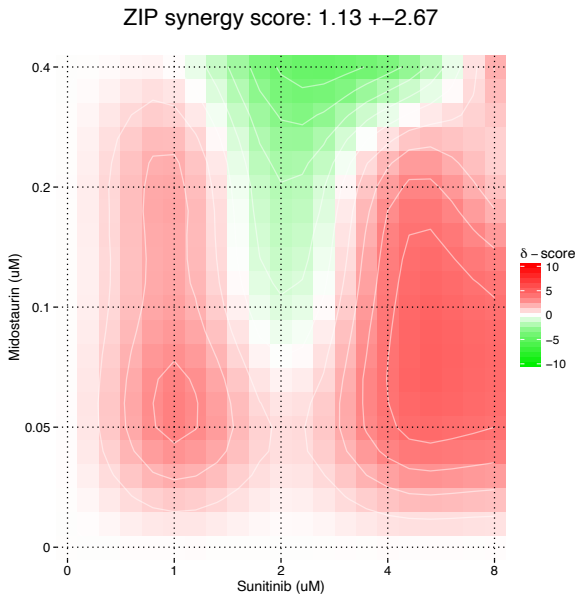


Figure 8.1: Synergy map for Midostaurin + Sunitinib

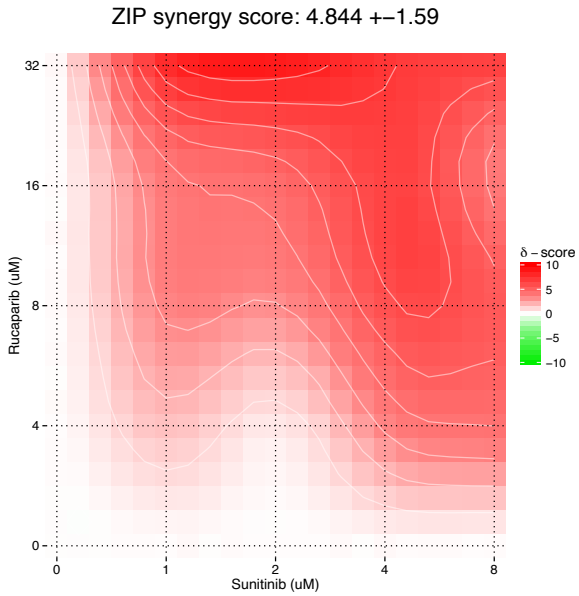


Figure 8.2: Synergy map for Sunitinib + Rucaparib

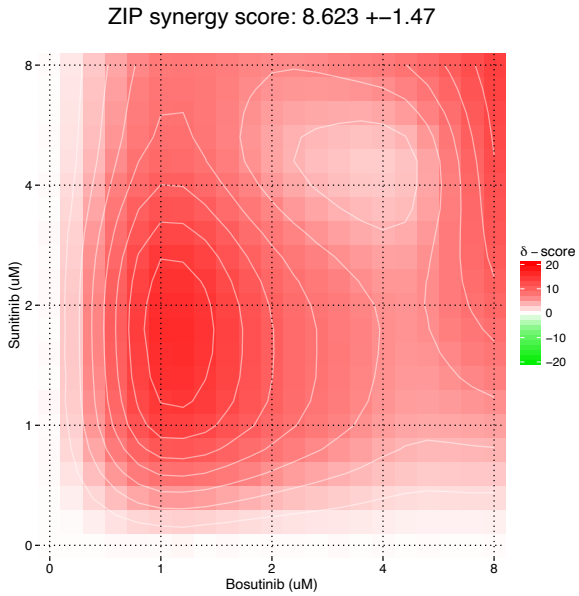


Figure 8.3: Synergy map for Sunitinib + Bosutinib

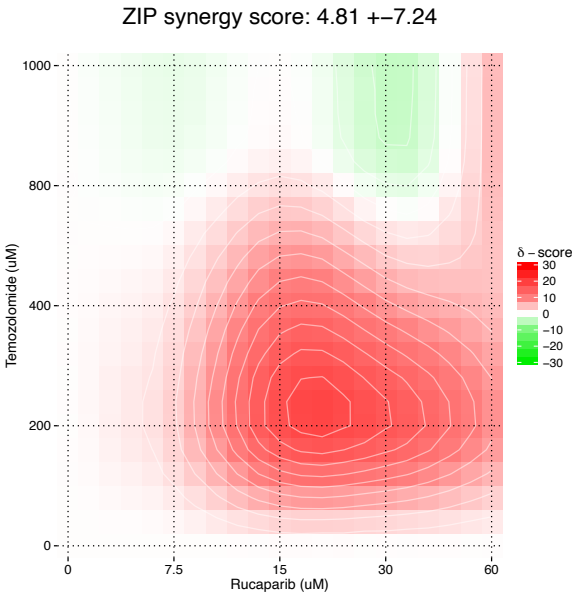


Figure 8.4: Synergy map for Temozolomide + Rucaparib

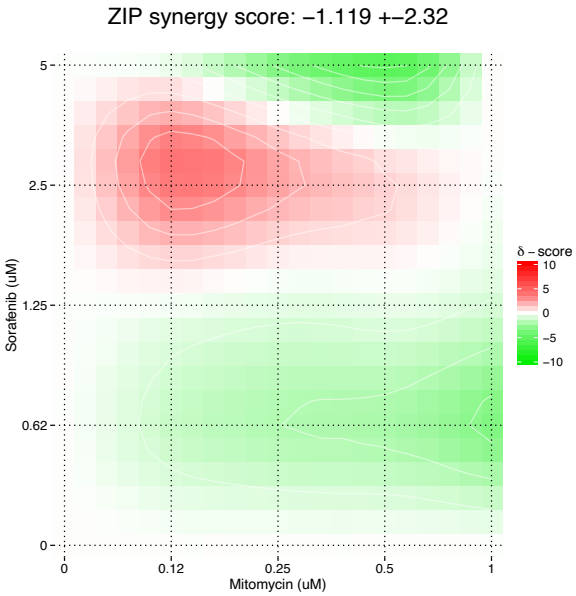


Figure 8.5: Synergy map for Mitomycin + Sorafenib

8.2.2 Drug-drug combinations tested in HCT116

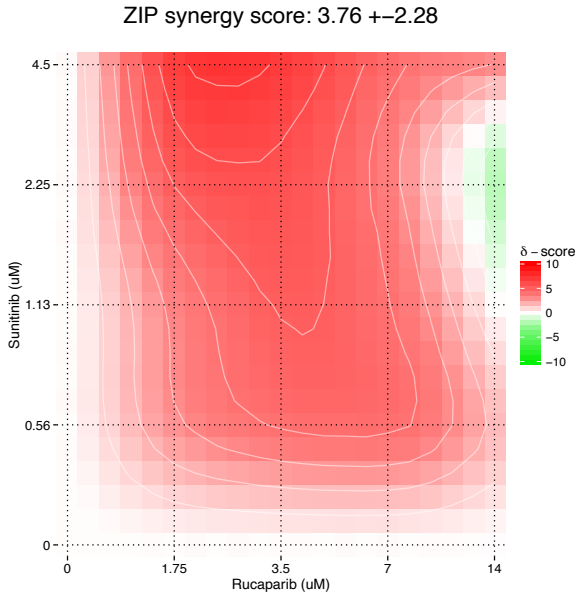


Figure 8.6: Synergy map for Sunitinib + Rucaparib

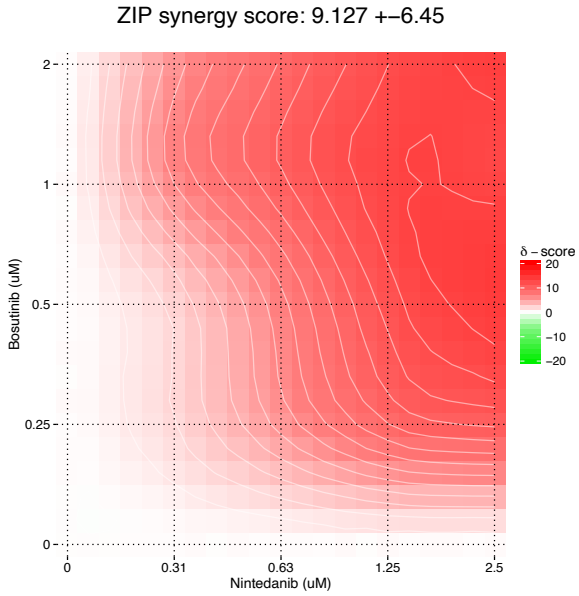


Figure 8.7: Synergy map for Bosutinib + Nintedanib

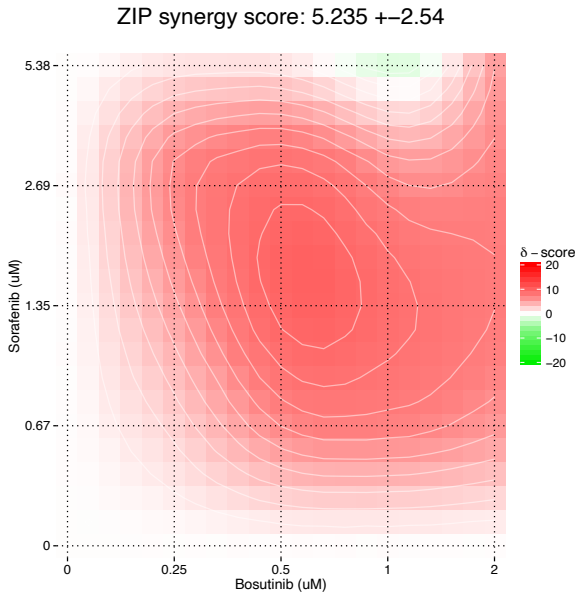


Figure 8.8: Synergy map for Sorafenib + Bosutinib

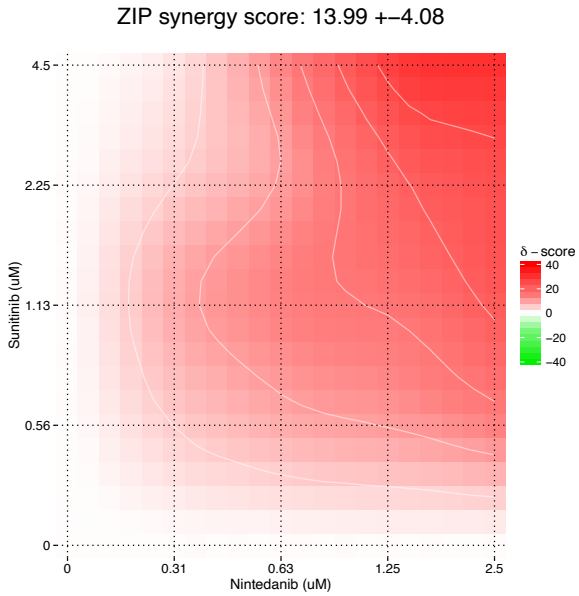


Figure 8.9: Synergy map for Sunitinib + Nintedanib

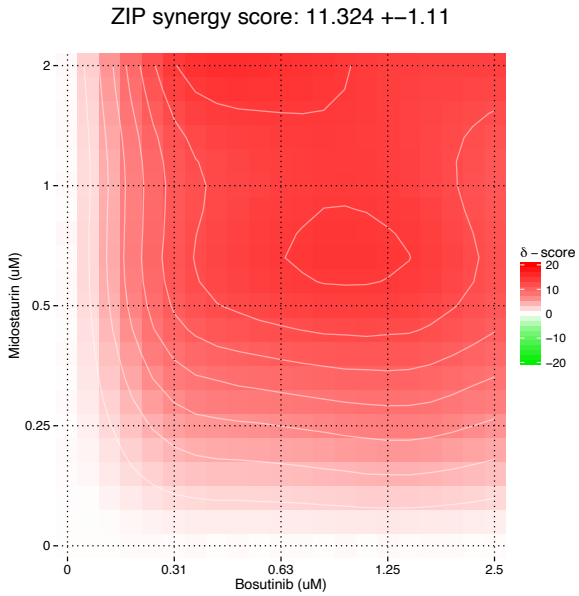


Figure 8.10: Synergy map for Midostaurin + Bosutinib

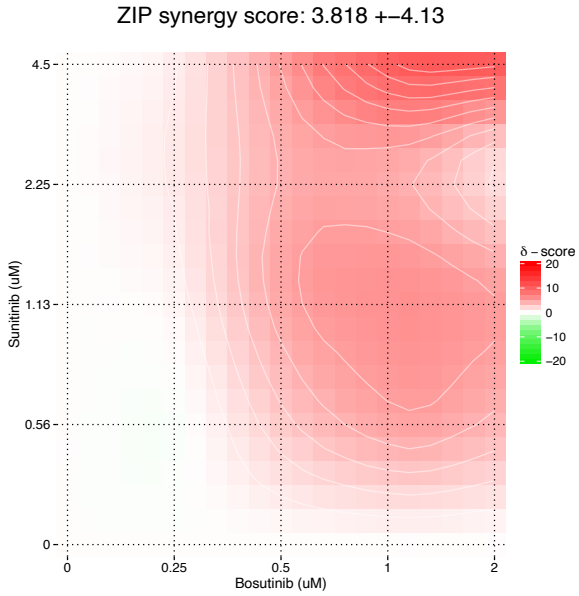


Figure 8.11: Synergy map for Sunitinib + Bosutinib

8.2.3 Food molecule-drug combinations tested in A549

Here, it is important to note that the calibration of the dose-response curve for Galangin appeared to be more complex than initially anticipated. This complexity necessitated a series of recalibrations following each combination replicate assessment, suggesting that the response to Galangin might not strictly adhere to a linear dose-response relationship. Therefore, we show the individual synergy maps for each replicate of Galangin + Dactosilib. It seemed possible that we were observing phenomena akin to a bell-shaped dose-response curve, potentially indicating receptor desensitization at higher doses or the activation of different cellular pathways at various concentrations. While these interpretations are speculative, they highlight the potential intricacies in Galangin's interaction with the biological system.

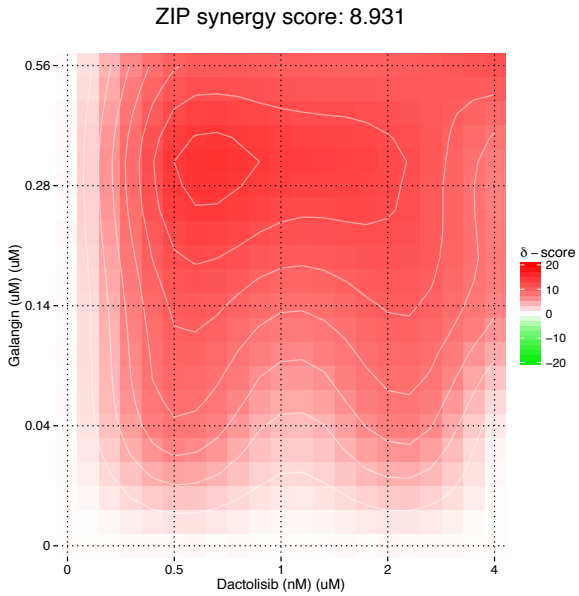


Figure 8.12: Synergy map for Dactosilib + Galangin. Replicate 1.

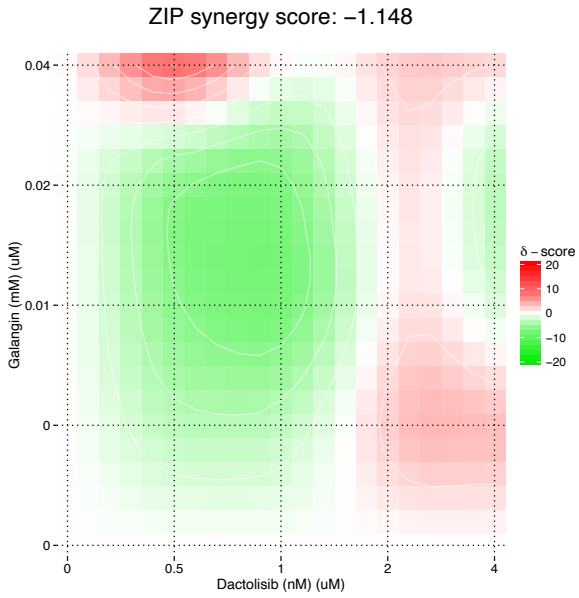


Figure 8.13: Synergy map for Dactosilib + Galangin. Replicate 2.

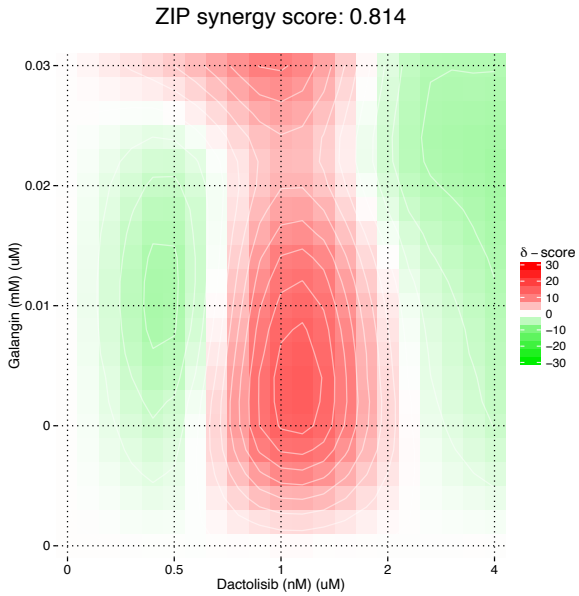


Figure 8.14: Synergy map for Dactosilib + Galangin. Replicate 3.

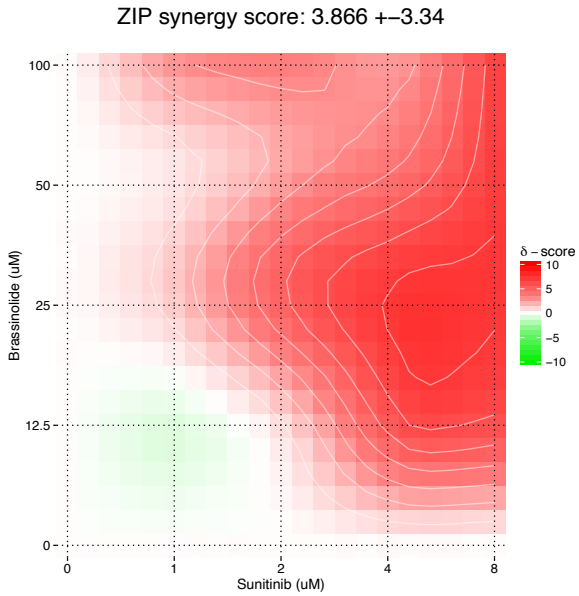


Figure 8.15: Synergy map for Sunitinib + Brassinolide

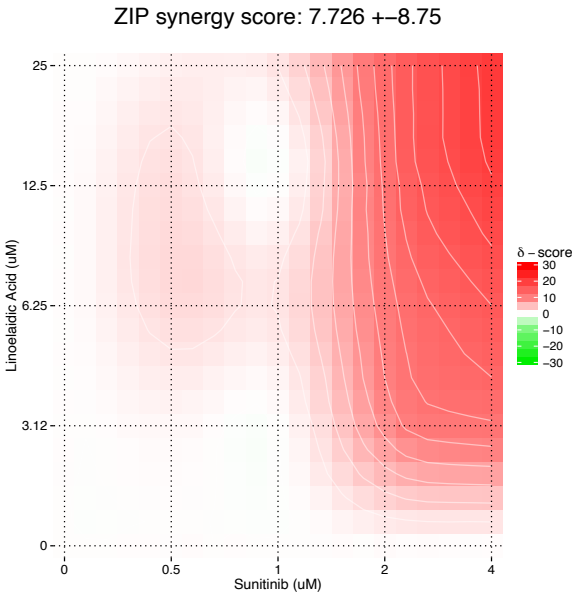


Figure 8.16: Synergy map for Sunitinib + Linoelaidic acid

References

1. Petrovska, B. B. Historical review of medicinal plants' usage. *Pharmacognosy Reviews* **6,** 1–5. ISSN: 09737847 (2012).
2. Dias, D. A., Urban, S. & Roessner, U. A Historical overview of natural products in drug discovery. *Metabolites* **2,** 303–336. ISSN: 22181989 (2012).
3. Yuan, H., Ma, Q., Ye, L. & Piao, G. The traditional medicine and modern medicine from natural products. *Molecules* **21.** ISSN: 14203049 (2016).
4. Drews, J. Drug discovery: A historical perspective. *Science* **287,** 1960–1964. ISSN: 00368075 (2000).
5. Bernardini, S., Tiezzi, A., Laghezza Masci, V. & Ovidi, E. Natural products for human health: an historical overview of the drug discovery approaches. *Natural Product Research* **32,** 1926–1950. ISSN: 14786427 (2018).
6. Thomas, M. *et al.* in *Artificial Intelligence in Drug Design* 1–59 (2021). ISBN: 9781071617861.
7. Shen, B. A New Golden Age of Natural Products Drug Discovery. *Cell* **163,** 1297–1300. ISSN: 10974172 (2015).
8. Newman, D. J. & Cragg, G. M. Natural Products As Sources of New Drugs over the 30 Years from 1981 to 2010.
9. Hopkins, A. L. Network pharmacology: The next paradigm in drug discovery. *Nature Chemical Biology* **4,** 682–690. ISSN: 15524469 (2008).
10. Boezio, B., Audouze, K., Ducrot, P. & Taboureau, O. Network-based Approaches in Pharmacology. *Molecular Informatics* **36,** 1–10. ISSN: 18681751 (2017).
11. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nature Reviews Drug Discovery* **1,** 727–730. ISSN: 14741776 (2002).
12. Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there ? *Nature reviews. Drug discovery* **5,** 10 (2006).
13. Imming, P., Sinning, C. & Meyer, A. Drugs, their targets and the nature and number of drug targets. *Nature Reviews Drug Discovery* **5,** 821–835 (2006).
14. Casas, A. I. *et al.* From single drug targets to synergistic network pharmacology in ischemic stroke. *Proceedings of the National Academy of Sciences of the United States of America* **116,** 7129–7136. ISSN: 10916490 (2019).
15. Goh, K. I. *et al.* The human disease network. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 8685–8690. ISSN: 00278424 (2007).

16. Iskar, M., Zeller, G., Zhao, X. M., van Noort, V. & Bork, P. Drug discovery in the age of systems biology: The rise of computational approaches for data integration. *Current Opinion in Biotechnology* **23,** 609–616. ISSN: 09581669 (2012).

17. Barabási, A. L., Gulbahce, N. & Loscalzo, J. Network medicine: A network-based approach to human disease. *Nature Reviews Genetics* **12,** 56–68. ISSN: 14710056 (2011).

18. Galindez, G., Sadegh, S., Baumbach, J., Kacprowski, T. & List, M. Network-based approaches for modeling disease regulation and progression. *Computational and Structural Biotechnology Journal* **21,** 780–795. ISSN: 2001-0370 (Jan. 2023).

19. Ding, H., Takigawa, I., Mamitsuka, H. & Zhu, S. Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Briefings in Bioinformatics* **15,** 734–747. ISSN: 1467-5463 (Sept. 2014).

20. Lotfi Shahreza, M., Ghadiri, N., Mousavi, S. R., Varshosaz, J. & Green, J. R. A review of network-based approaches to drug repositioning. *Briefings in Bioinformatics* **19,** 878–892. ISSN: 1467-5463 (Sept. 2018).

21. Wu, Z., Li, W., Liu, G. & Tang, Y. Network-based methods for prediction of drug-target interactions. *Frontiers in Pharmacology* **9,** 1134. ISSN: 16639812 (Oct. 2018).

22. Zhao, Q., Yu, H., Ji, M., Zhao, Y. & Chen, X. Computational Model Development of Drug-Target Interaction Prediction: A Review. *Current Protein & Peptide Science* **20,** 492–494. ISSN: 13892037 (Jan. 2019).

23. Wu, Z., Wang, Y. & Chen, L. Network-based drug repositioning. *Molecular BioSystems* **9,** 1268–1281. ISSN: 1742-2051 (May 2013).

24. Song, T. *et al.* Network-Based Approaches for Drug Repositioning. *Molecular Informatics* **41,** 1–12. ISSN: 18681751 (2022).

25. Sachdev, K. & Gupta, M. K. A comprehensive review of computational techniques for the prediction of drug side effects. *Drug Development Research* **81,** 650–670. ISSN: 1098-2299 (Sept. 2020).

26. Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A. & Vandergheynst, P. Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine* **34,** 18–42. ISSN: 1053-5888 (July 2017).

27. Li, M. M., Huang, K. & Zitnik, M. Graph representation learning in biomedicine and healthcare. *Nature Biomedical Engineering.* ISSN: 2157846X (2022).

28. Zhang, Z. *et al.* Graph neural network approaches for drug-target interactions. *Current Opinion in Structural Biology* **73,** 102327. ISSN: 1879033X (2022).

29. Gaudelet, T. *et al.* Utilizing graph machine learning within drug discovery and development. *Briefings in Bioinformatics* **22.** ISSN: 14774054 (2021).

30. Lee, C. Y. & Chen, Y. P. P. Prediction of drug adverse events using deep learning in pharmaceutical discovery. *Briefings in Bioinformatics* **22,** 1884–1901. ISSN: 14774054 (Mar. 2021).

31. Callaway, E. & Cyranoski, D. Anti-parasite drugs sweep Nobel prize in medicine 2015. *Nature* **526,** 174–175 (2015).

32. Maddams, J., Utley, M. & Møller, H. Projections of cancer prevalence in the United Kingdom, 2010-2040. *British Journal of Cancer* **107,** 1195–1202. ISSN: 00070920 (2012).

33. Fitzmaurice, C. *et al.* The Global Burden of Cancer 2013. *JAMA Oncology* **1,** 505–527. ISSN: 23742445 (2015).

34. Cao, W., Chen, H. D., Yu, Y. W., Li, N. & Chen, W. Q. Changing profiles of cancer burden worldwide and in China: A secondary analysis of the global cancer statistics 2020. *Chinese Medical Journal* **134,** 783–791. ISSN: 25425641 (2021).

35. Schiller, J. H. Current standards of care in small-cell and non-small-cell lung cancer. *Oncology* **61,** 3–13. ISSN: 00302414 (2001).

36. Maughan, K. L., Lutterbie, M. A. & Ham, P. S. Treatment of breast cancer. *American Family Physician* **81,** 1339–1346. ISSN: 0002838X (2010).

37. Stintzing, S. Management of colorectal cancer. *F1000Prime Reports* **6.** ISSN: 20517599 (2014).

38. Debela, D. T. *et al.* New approaches and procedures for cancer treatment: Current perspectives. *SAGE Open Medicine* **9.** ISSN: 20503121 (2021).

39. Duma, N., Santana-Davila, R. & Molina, J. R. Non-Small Cell Lung Cancer: Epidemiology, Screening, Diagnosis, and Treatment. *Mayo Clinic Proceedings* **94,** 1623–1640 (2019).

40. Herbst, R. S., Morgensztern, D. & Boshoff, C. The biology and management of non-small cell lung cancer. *Nature* **553,** 446–454 (2018).

41. Planchard, D. *et al.* Metastatic non-small cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology* **29,** iv192–iv237 (2018).

42. Rudin, C. M., Brambilla, E., Faivre-Finn, C. & Sage, J. Small-Cell Lung Cancer. *Nature Reviews Disease Primers* **7,** 3 (2021).

43. Gazdar, A. F., Bunn, P. A. & Minna, J. D. Small-cell lung cancer: what we know, what we need to know and the path forward. *Nature Reviews Cancer* **17,** 725–737 (2017).

44. West, H. *et al.* Treatment Patterns and Outcomes in Resected Early-stage Non-small Cell Lung Cancer: An Analysis of the SEER-Medicare Data. *Clinical Lung Cancer* **24,** 260–268. ISSN: 1525-7304 (2023).

45. Soares, M. *et al.* Treatment and outcomes for early non-small-cell lung cancer: A retrospective analysis of a Portuguese hospital database. *Lung Cancer Management* **10,** LMT46 (2021).

46. Benson, A. B. I. *et al.* NCCN Guidelines Insights: Colon Cancer, Version 2.2018. *Journal of the National Comprehensive Cancer Network* **16,** 359–369 (2018).

47. Grothey, A. *et al.* Duration of Adjuvant Chemotherapy for Stage III Colon Cancer. *New England Journal of Medicine* **378,** 1177–1188 (2018).

48. Van Cutsem, E. *et al.* ESMO consensus guidelines for the management of patients with metastatic colorectal cancer. *Annals of Oncology* **27,** 1386–1422 (2016).

49. André, T. *et al.* Pembrolizumab in Microsatellite-Instability–High Advanced Colorectal Cancer. *New England Journal of Medicine* **383,** 2207–2218 (2020).

50. Overman, M. J. *et al.* Nivolumab in patients with metastatic DNA mismatch repair-deficient or microsatellite instability-high colorectal cancer (CheckMate 142): an open-label, multicentre, phase 2 study. *The Lancet Oncology* **18,** 1182–1191 (2017).

51. Le, D. T. *et al.* Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* **357,** 409–413 (2015).

52. Overman, M. J. *et al.* Nivolumab in patients with metastatic DNA mismatch repair-deficient or microsatellite instability-high colorectal cancer (CheckMate 142): an open-label, multicentre, phase 2 study. *The Lancet Oncology* **19,** 1182–1191 (2018).

53. Lee, S. & Meyerhardt, J. Impact of Diet and Exercise on Colorectal Cancer. *Hematology/Oncology Clinics of North America* **36,** 471–489 (2022).

54. Kumar, A., Chinnathambi, S., Kumar, M. & Pandian, G. Food Intake and Colorectal Cancer. *Nutrition and Cancer* **75,** 1710–1742 (2023).

55. Vallis, J. & Wang, P. in *Gastrointestinal Cancers* (ed Morgado-Diaz, J.) chap. 2 (Exon Publications, Brisbane (AU), Sept. 2022).

56. Borgas, P., Gonzalez, G., Veselkov, K. & Mirnezami, R. Phytochemically rich dietary components and the risk of colorectal cancer: A systematic review and meta-analysis of observational studies. *World Journal of Clinical Oncology* **12,** 482–499 (2021).

57. Waks, A. G. & Winer, E. P. Breast Cancer Treatment: A Review. *JAMA* **321,** 288–300 (2019).

58. Harbeck, N. & Gnant, M. Breast cancer. *The Lancet* **393,** 1134–1150 (2019).

59. Finn, R. S. *et al.* Palbociclib and Letrozole in Advanced Breast Cancer. *New England Journal of Medicine* **375,** 1925–1936 (2016).

60. Slamon, D. *et al.* Adjuvant trastuzumab in HER2-positive breast cancer. *New England Journal of Medicine* **365,** 1273–1283 (2011).

61. Lee, A. & Djamgoz, M. B. A. Triple-negative Breast Cancer: Emerging Therapeutic Modalities and Novel Combination Therapies. *Cancer Treatment Reviews* **62,** 110–122 (2017).

62. Robson, M. *et al.* Olaparib for Metastatic Breast Cancer in Patients with a Germline BRCA Mutation. *New England Journal of Medicine* **377,** 523–533 (2017).

63. World Cancer Research Fund/American Institute for Cancer Research. *Diet, Nutrition, Physical Activity and Cancer: a Global Perspective. A Summary of the Third Expert Report* 2018.

64. Farvid, M. S. *et al.* Consumption of red and processed meat and breast cancer incidence: A systematic review and meta-analysis of prospective studies. *International Journal of Cancer* **143,** 2787–2799 (2018).

65. Song, M., Garrett, W. S. & Chan, A. T. Nutrients and foods for the primary prevention of colorectal cancer and colorectal adenoma: a systematic review and meta-analysis. *Nutrients* **7,** 9252–9271 (2015).

66. DeClercq, V., Nearing, J. & Sweeney, E. Plant-Based Diets and Cancer Risk: What is the Evidence? *Current Nutrition Reports* **11.** Epub 2022 Mar 25, 354–369 (2022).

67. Pandey, K. B. & Rizvi, S. I. Plant polyphenols as dietary antioxidants in human health and disease. *Oxidative Medicine and Cellular Longevity* **2,** 270–278 (2009).

68. Reuter, S., Gupta, S., Chaturvedi, M. & Aggarwal, B. Oxidative stress, inflammation, and cancer: how are they linked? *Free Radical Biology and Medicine* **49,** 1603–1616 (2010).

69. Kampa, M., Nifli, A., Notas, G. & Castanas, E. in *Reviews of Physiology, Biochemistry and Pharmacology* (eds Amara, S. *et al.*) (Springer, Berlin, Heidelberg, 2007).

70. Niedzwiecki, A., Roomi, M., Kalinovsky, T. & Rath, M. Anticancer Efficacy of Polyphenols and Their Combinations. *Nutrients* **8,** 552 (2016).

71. Briguglio, G. *et al.* Polyphenols in cancer prevention: New insights (Review). *International Journal of Functional Nutrition* **1,** 9 (2020).

72. Li, Y., Zhang, T. & Korkaya, H. Sulforaphane, a dietary component of broccoli/broccoli sprouts, inhibits breast cancer stem cells. *Clinical Cancer Research* **16,** 2580–2590 (2010).

73. Navarro, S. L., Li, F. & Lampe, J. W. Mechanisms of action of isothiocyanates in cancer chemoprevention: an update. *Food Function* **2,** 579–587 (2011).

74. Ahmad, A., Sakr, W. A. & Rahman, K. W. Anticancer Properties of Indole Compounds: Mechanism of Apoptosis Induction and Role in Chemotherapy. *Current Drug Targets* **11,** 652–666 (2010).

75. Juge, N., Mithen, R. & Traka, M. Molecular basis for chemoprevention by sulforaphane: a comprehensive review. *Cellular and Molecular Life Sciences* **64,** 1105. https://doi.org/10.1007/s00018-007-6484-5 (2007).

76. Lee, C. *et al.* Fasting cycles retard growth of tumors and sensitize a range of cancer cell types to chemotherapy. *Science Translational Medicine* **4,** 124ra27 (2012).

77. O'Flanagan, C. H., Smith, L. A., McDonell, S. B. & Hursting, S. D. When less may be more: calorie restriction and response to cancer therapy. *BMC Medicine* **15,** 106 (2017).

78. Allen, B. G. *et al.* Ketogenic diets as an adjuvant cancer therapy: History and potential mechanism. *Redox Biology* **2,** 963–970. ISSN: 2213-2317 (2014).

79. Arends, J. *et al.* ESPEN guidelines on nutrition in cancer patients. *Clinical nutrition* **36,** 11–48 (2017).

80. Rock, C. L. *et al.* Nutrition and physical activity guidelines for cancer survivors. *CA: A Cancer Journal for Clinicians* **62,** 243–274 (2012).

81. Zhang, F. F. *et al.* Preventable cancer burden associated with poor diet in the United States. *JNCI Cancer Spectrum* **3,** pkz034 (2019).

82. Kushi, L. H. *et al.* American Cancer Society Guidelines on nutrition and physical activity for cancer prevention: reducing the risk of cancer with healthy food choices and physical activity. *CA: a cancer journal for clinicians* **62,** 30–67 (2012).

83. Eseberri, I. *et al.* Variability in the Beneficial Effects of Phenolic Compounds: A Review. *Nutrients* **14,** 1925 (2022).

84. Holst, B. & Williamson, G. Nutrients and phytochemicals: from bioavailability to bioefficacy beyond antioxidants. *Current Opinion in Biotechnology* **19,** 73–82. ISSN: 0958-1669 (2008).

85. Rein, M. J. *et al.* Bioavailability of bioactive food compounds: a challenging journey to bioefficacy. *British Journal of Clinical Pharmacology* **75,** 588–602 (2013).

86. Briguglio, M. *et al.* Food Bioactive Compounds and Their Interference in Drug Pharmacokinetic/Pharmacodynamic Profiles. *Pharmaceutics* **10,** 277 (2018).

87. Huang, S.-M. & Lesko, L. Drug-Drug, Drug—Dietary Supplement, and Drug—Citrus Fruit and Other Food Interactions: What Have We Learned? *The Journal of Clinical Pharmacology* **44,** 559–569 (2004).

88. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **46,** D1074–D1082. ISSN: 0305-1048 (Jan. 2018).

89. Ursu, O. *et al.* DrugCentral: online drug compendium. *Nucleic acids research* **45,** D932–D939. ISSN: 1362-4962 (Jan. 2017).

90. Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J. & Bork, P. STITCH: interaction networks of chemicals and proteins. *Nucleic acids research* **36,** 684–8. ISSN: 1362-4962 (Jan. 2008).

91. Bamford, S. *et al.* The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British Journal of Cancer* **91,** 355–358. ISSN: 00070920 (2004).

92. Garg, N. *et al.* FlavorDB: A database of flavor molecules. *Nucleic Acids Research* **46,** D1210–D1216. ISSN: 13624962 (2018).

93. Ni, Y., Jensen, K., Kouskoumvekaki, I. & Panagiotou, G. NutriChem 2.0: exploring the effect of plant-based foods on human health and drug efficacy. *Database* **2017,** 1–6. ISSN: 1758-0463 (Jan. 2017).

94. Jensen, K., Ni, Y., Panagiotou, G. & Kouskoumvekaki, I. Developing a Molecular Roadmap of Drug-Food Interactions. *PLOS Computational Biology* **11** (ed Tan, K.) e1004048. ISSN: 1553-7358 (Feb. 2015).

95. Zheng, T., Ni, Y., Li, J., Chow, B. K. & Panagiotou, G. Designing dietary recommendations using system level interactomics analysis and network-based inference. *Frontiers in Physiology* **8,** 1–16. ISSN: 1664042X (2017).

96. Pavlopoulos, G. A. *et al.* Using graph theory to analyze biological networks. *BioData mining* **4,** 10. ISSN: 1756-0381 (Apr. 2011).

97. Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2015 update. *Nucleic acids research* **43,** 470–8. ISSN: 1362-4962 (Jan. 2015).

98. Huttlin, E. L. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* **545,** 505–509. ISSN: 0028-0836 (May 2017).

99. Keshava Prasad, T. S. *et al.* Human Protein Reference Database–2009 update. *Nucleic Acids Research* **37,** D767–D772. ISSN: 0305-1048 (Jan. 2009).

100. Von Mering, C. *et al.* STRING: a database of predicted functional associations between proteins. *Nucleic acids research* **31,** 258–61. ISSN: 1362-4962 (Jan. 2003).

101. Bader, G. D., Cary, M. P. & Sander, C. Pathguide: a pathway resource list. *Nucleic acids research* **34,** 504–506. ISSN: 13624962 (2006).

102. Romero, P. *et al.* Computational prediction of human metabolic pathways from the complete human genome. *Genome biology* **6,** 1–17. ISSN: 14656914 (2005).

103. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28,** 27–30. ISSN: 13624962 (Jan. 2000).

104. Schaefer, C. F. *et al.* PID: The pathway interaction database. *Nucleic Acids Research* **37,** 674–679. ISSN: 03051048 (2009).

105. Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research* **41,** 377–386. ISSN: 03051048 (2013).

106. Croft, D. *et al.* Reactome: A database of reactions, pathways and biological processes. *Nucleic Acids Research* **39,** 691–697. ISSN: 03051048 (2011).

107. Luo, Y. *et al.* A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature Communications* **8,** 573. ISSN: 2041-1723 (Dec. 2017).

108. Yamanishi, Y., Kotera, M., Kanehisa, M. & Goto, S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* **26,** i246–i254. ISSN: 1367-4803 (June 2010).

109. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nature Methods* **10,** 1108–1118. ISSN: 15487091 (2013).

110. Zhang, W., Ma, J. & Ideker, T. Classifying tumors by supervised network propagation. *Bioinformatics* **34,** i484–i493. ISSN: 1367-4803 (July 2018).

111. Liu, H., Song, Y., Guan, J., Luo, L. & Zhuang, Z. Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks. *BMC bioinformatics* **17,** 539. ISSN: 1471-2105 (Dec. 2016).

112. Wang, W., Yang, S., Zhang, X. & Li, J. Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics* **30,** 2923–2930. ISSN: 1460-2059 (Oct. 2014).

113. Li, J. & Lu, Z. Pathway-based drug repositioning using causal inference. *BMC Bioinformatics* **14,** S3. ISSN: 1471-2105 (Oct. 2013).

114. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: A universal amplifier of genetic associations. *Nature Reviews Genetics* **18,** 551–562. ISSN: 14710064 (2017).

115. Hamilton, W. L., Ying, R. & Leskovec, J. Representation Learning on Graphs: Methods and Applications. *IEEE Data Eng. Bull.* (2017).

116. Perozzi, B., Al-Rfou, R. & Skiena, S. *DeepWalk: Online Learning of Social Representations* New York, New York, USA, 2014.

117. Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. *KDD 16,* 855–864 (2016).

118. Kipf, T. N. & Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *International Conference on Learning Representations (ICLR)* (Sept. 2017).

119. Hamilton, W. L., Ying, R. & Leskovec, J. Inductive representation learning on large graphs. *31st Conference on Neural Information Processing Systems (NIPS 2017)* (June 2017).

120. Defferrard, M., Bresson, X. & Vandergheynst, P. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. *Advances in neural information processing systems (NIPS)* (June 2016).

121. Veselkov, K. *et al.* HyperFoods: Machine intelligent mapping of cancer-beating molecules in foods. *Scientific Reports* **9,** 9237 (Dec. 2019).

122. Prince, M. J. *et al.* The burden of disease in older people and implications for health policy and practice. *The Lancet* **385,** 549–562. ISSN: 0140-6736 (Feb. 2015).

123. Shi, Y. & Hu, F. B. The global implications of diabetes and cancer. *The Lancet* **383,** 1947–1948. ISSN: 1474547X (June 2014).

124. Smittenaar, C. R., Petersen, K. A., Stewart, K. & Moitt, N. Cancer incidence and mortality projections in the UK until 2035. *British Journal of Cancer* **115,** 1147–1155. ISSN: 15321827 (2016).

125. WCRF/AICR. Food, Nutrition, and the Prevention of Cancer: A Global Perspective. *World Cancer Research Fund / American Institute for Cancer Research* (1997).

126. Donaldson, M. S. Nutrition and cancer: A review of the evidence for an anti-cancer diet. *Nutrition Journal* **3,** 19. ISSN: 1475-2891 (Dec. 2004).

127. Schwingshackl, L. *et al.* Food groups and risk of colorectal cancer. *International Journal of Cancer* **142,** 1748–1758. ISSN: 10970215 (May 2018).

128. Wu, Q. J. *et al.* Cruciferous vegetables intake and the risk of colorectal cancer: A meta-analysis of observational studies. *Annals of Oncology* **24,** 1079–1087. ISSN: 09237534 (Apr. 2013).

129. Borgas, P., Gonzalez, G., Veselkov, K. & Mirnezami, R. Phytochemically rich dietary components and the risk of colorectal cancer: A systematic review and meta-analysis of observational studies. *World Journal of Clinical Oncology* **12,** 482–499. ISSN: 2218-4333 (June 2021).

130. Huang, M. Y., Zhang, L. L., Ding, J. & Lu, J. J. Anticancer drug discovery from Chinese medicinal herbs. *Chinese Medicine (United Kingdom)* **13,** 1–9. ISSN: 17498546 (2018).

131. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *Journal of Natural Products* **83,** 770–803. ISSN: 15206025 (2020).

132. Huang, M., Lu, J. J. & Ding, J. Natural Products in Cancer Therapy: Past, Present and Future. *Natural Products and Bioprospecting* **11,** 5–13. ISSN: 21922209 (2021).

133. Menche, J. *et al.* Integrating personalized gene expression profiles into predictive disease-associated gene pools. *npj Systems Biology and Applications* **3,** 10. ISSN: 2056-7189 (2017).

134. Creixell, P. *et al.* Pathway and network analysis of cancer genomes. *Nature Methods* **12,** 615–621. ISSN: 15487105 (June 2015).

135. Krogan, N. J., Lippman, S., Agard, D. A., Ashworth, A. & Ideker, T. *The Cancer Cell Map Initiative: Defining the Hallmark Networks of Cancer* May 2015.

136. Veselkov, K. *et al.* HyperFoods: Machine intelligent mapping of cancer-beating molecules in foods. *Scientific Reports* **9,** 9237. ISSN: 2045-2322 (Dec. 2019).

137. Zitnik, M., Agrawal, M. & Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* **34,** i457–i466 (2018).

138. Rhee, S., Seo, S. & Kim, S. Hybrid Approach of Relation Network and Localized Graph Convolutional Filtering for Breast Cancer Subtype Classification. *Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI),* 3527–3534 (2018).

139. Xie, Y., Peng, J. & Zhou, Y. Integrating Protein-Protein Interaction Information into Drug Response Prediction by Graph Neural Encoding. *Submitted to BMC bioinformatics* (2020).

140. Pan, X. *et al.* Deep learning for drug repurposing: Methods, databases, and applications. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **12,** 1–21. ISSN: 17590884 (2022).

141. Yu, J. L., Dai, Q. Q. & Li, G. B. Deep learning in target prediction and drug repositioning: Recent advances and challenges. *Drug Discovery Today* **27,** 1796–1814. ISSN: 18785832 (2022).

142. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28,** 31–6 (1988).

143. David, L., Thakkar, A., Mercado, R. & Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *Journal of Cheminformatics* **12,** 1–22. ISSN: 17582946 (2020).

144. Jaeger, S., Fulle, S. & Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *Journal of Chemical Information and Modeling* **58,** 27–35. ISSN: 1549960X (2018).

145. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings,* 1–12 (2013).

146. Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* **180,** 688–702. ISSN: 1097-4172 (Feb. 2020).

147. Liu, H. *et al.* Drug Repositioning for SARS-CoV-2 Based on Graph Neural Network. *Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020,* 319–322 (2020).

148. Ioannidis, V. N., Zheng, D. & Karypis, G. Few-shot link prediction via graph neural networks for Covid-19 drug-repurposing (2020).

149. Wishart Research Group. *FooDB* 2017. http://foodb.ca.

150. Brown, A. S. & Patel, C. J. A standard database for drug repositioning. *Scientific Data* **4,** 170029. ISSN: 2052-4463 (Mar. 2017).

151. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D. & Pletnev, I. InChI - The worldwide chemical structure identifier standard. *Journal of Cheminformatics* **5,** 1. ISSN: 17582946 (2013).

152. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 15545–50. ISSN: 0027-8424 (Oct. 2005).

153. Xu, K. *et al.* Representation Learning on Graphs with Jumping Knowledge Networks (2018).

154. Paszke, A. *et al.* Automatic differentiation in PyTorch. *31st Conference on Neural Information Processing Systems (NIPS)* (2017).

155. Fey, M. & Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. *ICLR Workshop on Representation Learning on Graphs and Manifolds* (Mar. 2019).

156. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. *Proceedings of the 34th International Conference on Machine Learning (ICLR)* (2017).

157. Rauf, A. *et al.* Resveratrol as an anti-cancer agent: A review. *Critical Reviews in Food Science and Nutrition* **58,** 1428–1447. ISSN: 1040-8398 (June 2018).

158. Chen, R.-J. *et al.* Apoptotic and Nonapoptotic Activities of Pterostilbene against Cancer. *International Journal of Molecular Sciences* **19,** 287. ISSN: 1422-0067 (Jan. 2018).

159. Li, Y.-R., Li, S. & Lin, C.-C. Effect of resveratrol and pterostilbene on aging and longevity. *BioFactors* **44,** 69–82. ISSN: 09516433 (Jan. 2018).

160. Schmeer, M. R., Horton, D. & Tanimura, A. Mercenene, a tumor inhibitor from Mercenaria: Purification and characterization studies. *Life Sciences* **5,** 1169–1178. ISSN: 00243205 (1966).

161. Brubaker, D., Hooper, F. B. & Piscitelli, J. Effects of Mercenene on the Growth of MBA Hela Cell Line and Strain L Cells. *Proceedings of the Pennsylvania Academy of Science,* **45,** 16–19 (1971).

162. Rall, D. P., Ben, M. & McCarthy, D. M. 1,3-Bis(2-chloroethyl)-l-nitrosourea. *Proc. Am. A. Cancer Res.* **4** (1963).

163. DeVita, V. T., Denham, C., Davidson, J. D. & Oliverio, V. T. The physiological disposition of the carcinostatic 1,3-bis(2-chloroethyl)-1-nitrosourea (BCNU) in man and animals. *Clinical Pharmacology andTherapeutics* **8,** 566–577. ISSN: 15326535 (1967).

164. Genkinger, J. M. & Koushik, A. Meat consumption and cancer risk. *PLoS Medicine* **4,** e345. https://doi.org/10.1371/journal.pmed.0040345 (2007).

165. Farvid, M. S. *et al.* Consumption of red meat and processed meat and cancer incidence: a systematic review and meta-analysis of prospective studies. *European Journal of Epidemiology* **36,** 937–951. eprint: 2021Aug29 https://doi.org/10.1007/s10654-021-00741-9 (Sept. 2021).

166. Fleisher, D., Li, C., Zhou, Y., Pao, L. & Karim, A. Drug, Meal and Formulation Interactions Influencing Drug Absorption After Oral Administration. *Clinical Pharmacokinetics* **36,** 233–254. ISSN: 0312-5963 (Nov. 1999).

167. Schmidt, L. E. & Dalhoff, K. Food-Drug Interactions. *Drugs* **62,** 1481–1502. ISSN: 0012-6667 (Sept. 2002).

168. Koziolek, M. *et al.* The mechanisms of pharmacokinetic food-drug interactions – A perspective from the UNGAP group. *European Journal of Pharmaceutical Sciences* **134,** 31–59. ISSN: 0928-0987 (June 2019).

169. Dahan, A. & Altman, H. Food–drug interaction: grapefruit juice augments drug bioavailability—mechanism, extent and relevance. *European Journal of Clinical Nutrition* **58,** 1–9. ISSN: 0954-3007 (Jan. 2004).

170. O'Shea, J. P., Holm, R., O'Driscoll, C. M. & Griffin, B. T. Food for thought: formulating away the food effect – a PEARRL review. *Journal of Pharmacy and Pharmacology* **71,** 510–535. ISSN: 2042-7158 (Mar. 2019).

171. Brown, C., Taniguchi, G. & Yip, K. The Monoamine Oxidase Inhibitor-Tyramine Interaction. *The Journal of Clinical Pharmacology* **29,** 529–532. ISSN: 00912700 (June 1989).

172. Sathyanarayana Rao, T. S. & Yeragani, V. K. Hypertensive crisis and cheese. *Indian Journal of Psychiatry* **51,** 65–66. ISSN: 00195545 (2009).

173. López-Muñoz, F., D'Ocón, P., Romero, A., Guerra, J. A. & Álamo, C. Role of serendipity in the discovery of classical antidepressant drugs: Applying operational criteria and patterns of discovery. *World Journal of Psychiatry* **12,** 588–602. ISSN: 2220-3206 (2022).

174. Blackwell, B. & Marley, E. Interactions of Cheese and of Its Constituents With Monoamine Oxidase Inhibitors. *British Journal of Pharmacology and Chemotherapy* **26,** 120–141. ISSN: 14765381 (1966).

175. Briguglio, M. *et al.* Food Bioactive Compounds and Their Interference in Drug Pharmacokinetic/Pharmacodynamic Profiles. *Pharmaceutics* **10.** ISSN: 1999-4923 (Dec. 2018).

176. Niu, J., Straubinger, R. M. & Mager, D. E. Pharmacodynamic Drug-Drug Interactions. *Clinical pharmacology and therapeutics* **105,** 1395. ISSN: 15326535 (June 2019).

177. Laponogov, I. *et al.* Network machine learning maps phytochemically rich "Hyperfoods" to fight COVID-19. *Human Genomics* **15,** 1 (Dec. 2021).

178. Gonzalez, G., Gong, S., Laponogov, I., Bronstein, M. & Veselkov, K. Predicting anticancer hyperfoods with graph convolutional networks. *Human Genomics* **15,** 33 (Dec. 2021).

179. Ruiz, C., Zitnik, M. & Leskovec, J. Discovery of disease treatment mechanisms through the multiscale interactome.

180. O'Neil, J. *et al.* An Unbiased Oncology Compound Screen to Identify Novel Combination Strategies. *Molecular Cancer Therapeutics* **15,** 1155–1162. ISSN: 1535-7163 (June 2016).

181. Yadav, B., Wennerberg, K., Aittokallio, T. & Tang, J. Searching for Drug Synergy in Complex Dose–Response Landscapes Using an Interaction Potency Model. *Computational and Structural Biotechnology Journal* **13,** 504–513. ISSN: 2001-0370 (Jan. 2015).

182. Holbeck, S. L. *et al.* The National Cancer Institute ALMANAC: A comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. *Cancer Research* **77,** 3564–3576. ISSN: 15387445 (2017).

183. Menden, M. P. *et al.* Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nature Communications* **10,** 1–17. ISSN: 20411723 (2019).

184. *AstraZeneca-Sanger drug combination prediction DREAM challenge.* https://www.synapse.org/#!Synapse:syn4231880/%20wiki/

185. Liu, H. *et al.* DrugCombDB: A comprehensive database of drug combinations toward the discovery of combinatorial therapy. *Nucleic Acids Research* **48,** D871–D881. ISSN: 13624962 (2020).

186. Zagidullin, B. *et al.* DrugComb: an integrative cancer drug combination data portal. *Nucleic Acids Research* **47,** W43–W51. ISSN: 0305-1048 (July 2019).

187. Preuer, K. *et al.* DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics* **34** (ed Wren, J.) 1538–1546. ISSN: 1367-4803 (May 2018).

188. Zhang, H., Feng, J., Zeng, A., Payne, P. & Li, F. Predicting Tumor Cell Response to Synergistic Drug Combinations Using a Novel Simplified Deep Learning Model. *AMIA ... Annual Symposium proceedings. AMIA Symposium* **2020,** 1364–1372. ISSN: 1942597X (2020).

189. Zhang, T., Zhang, L., Payne, P. & Li, F. in *Translational Bioinformatics for Therapeutic Development. Methods in Molecular Biology* (ed Markowitz, J.) 143–175 (Humana Press, 2021). ISBN: 978-1-0716-0848-7.

190. Liu, Q. & Xie, L. TranSynergy: Mechanism-driven interpretable deep neural network for the synergistic prediction and pathway deconvolution of drug combinations. *PLoS Computational Biology* **17,** 1–22. ISSN: 15537358 (2021).

191. Jiang, P. *et al.* Deep graph embedding for prioritizing synergistic anticancer drug combinations. *Computational and Structural Biotechnology Journal* **18,** 427–438. ISSN: 2001-0370 (Jan. 2020).

192. Wang, J., Liu, X., Shen, S., Deng, L. & Liu, H. DeepDDS: Deep graph neural network with attention mechanism to predict synergistic drug combinations. *Briefings in Bioinformatics* **23,** 1–11. ISSN: 14774054 (2022).

193. Zhang, P., Tu, S., Zhang, W. & Xu, L. Predicting cell line-specific synergistic drug combinations through a relational graph convolutional network with attention mechanism. *Briefings in Bioinformatics,* 1–13. ISSN: 1467-5463 (2022).

194. Liu, X. *et al.* Multi-way relation-enhanced hypergraph representation learning for anti-cancer drug synergy prediction. *Bioinformatics* **38,** 4782–4789. ISSN: 1367-4803 (2022).

195. Wang, X. *et al.* PRODeepSyn: Predicting anticancer synergistic drug combinations by embedding cell lines with protein-protein interaction network. *Briefings in Bioinformatics* **23,** 1–14. ISSN: 14774054 (2022).

196. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170,** 564–576. ISSN: 10974172 (2017).

197. Kuru, H. I., Tastan, O. & Cicek, A. E. MatchMaker: A Deep Learning Framework for Drug Synergy Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **19,** 2334–2344. ISSN: 15579964 (2022).

198. Feng, Y., You, H., Zhang, Z., Ji, R. & Gao, Y. Hypergraph Neural Networks.

199. Ianevski, A., He, L., Aittokallio, T. & Tang, J. SynergyFinder: A web application for analyzing drug combination dose-response matrix data. *Bioinformatics* **33,** 2413–2415. ISSN: 14602059 (Aug. 2017).

200. Brown, G. R. *et al.* Gene: a gene-centered information resource at NCBI. *Nucleic Acids Research* **43,** D36–D42. ISSN: 1362-4962 (Jan. 2015).

201. Bateman, A. *et al.* UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **45,** D158–D169. ISSN: 0305-1048 (Jan. 2017).

202. Carbon, S. *et al.* The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research* **47,** D330–D338. ISSN: 13624962 (Jan. 2019).

203. Ashburner, M. *et al. Gene ontology: Tool for the unification of biology* May 2000.

204. Dhamdhere, K., Ai, G., Google, S. & Yan, Q. *How Important Is a Neuron?* tech. rep. ().

205. Shrikumar, A., Su, J. & Kundaje, A. *Computationally Efficient Measures of Internal Neuron Importance* tech. rep. ().

206. Feoktistova, M., Geserick, P. & Leverkus, M. Crystal violet assay for determining viability of cultured cells. *Cold Spring Harbor Protocols* **2016,** 343–346. ISSN: 15596095 (2016).

207. Ianevski, A., Giri, A. K. & Aittokallio, T. SynergyFinder 3.0: an interactive analysis and consensus interpretation of multi-drug synergies across multiple samples. *Nucleic Acids Research* **50,** W739–W743. ISSN: 13624962 (2022).

208. Xiao, J., Egger, M. E., McMasters, K. M. & Hao, H. Differential expression of ABCB5 in BRAF inhibitor-resistant melanoma cell lines. *BMC Cancer* **18,** 1–10. ISSN: 14712407 (2018).

209. Panka, D. J., Wang, W., Atkins, M. B. & Mier, J. W. The raf inhibitor BAY 43-9006 (sorafenib) induces caspase-independent apoptosis in melanoma cells. *Cancer Research* **66,** 1611–1619. ISSN: 00085472 (2006).

210. Davies, M. A. The role of the PI3K-AKT pathway in melanoma. *Cancer Journal* **18,** 142–147. ISSN: 15289117 (2012).

211. ClinicalTrials.gov. *Search of: MK2206 | Melanoma* https://clinicaltrials.gov/ct2/results?cond=Melanoma&term=MK2206&cntry=&state=&city=&dist=.

212. ClinicalTrials.gov. *Search of: Sorafenib | Melanoma* https://clinicaltrials.gov/ct2/results?term=sorafenib&cond=Melanoma&Search=Apply&recrs=d&recrs=e&age_v=&gndr=&type=Intr&rslt=.

213. Yan, D. *et al.* Activation of AKT/AP1/FoxM1 signaling confers sorafenib resistance to liver cancer cells. *Oncology Reports* **42,** 785–796. ISSN: 17912431 (2019).

214. National Cancer Institute, U. *Akt inhibitor MK2206 - NCI Drug Dictionary* https://www.cancer.gov/publications/dictionaries/cancer-drug/def/akt-inhibitor-mk2206.

215. Zhai, B. *et al.* MK2206 overcomes the resistance of human liver cancer stem cells to sorafenib by inhibition of pAkt and upregulation of pERK. *Tumor Biology* **37,** 8047–8055. ISSN: 14230380 (2016).

216. Novello, S. *et al.* Phase II study of sunitinib in patients with non-small cell lung cancer and irradiated brain metastases. *Journal of Thoracic Oncology* **6,** 1260–1266. ISSN: 15561380 (2011).

217. Sumbly, V. & Landry, I. Unraveling the Role of STK11/LKB1 in Non-small Cell Lung Cancer. *Cureus* **2,** 1–7 (2022).

218. Momcilovic, M. & Shackelford, D. B. Targeting LKB1 in cancer-exposing and exploiting vulnerabilities. *British Journal of Cancer* **113,** 574–584. ISSN: 15321827 (2015).

219. Xiong, S. *et al.* Circular RNA MYLK promotes glycolysis and proliferation of non-small cell lung cancer cells by sponging MiR-195-5p and increasing glucose transporter member 3 expression. *Cancer Management and Research* **12,** 5469–5478. ISSN: 11791322 (2020).

220. *Home | IUPHAR/BPS Guide to PHARMACOLOGY* https://www.guidetopharmacology.org/.

221. *HMS LINCS Database - HMS LINCS Project* https://lincs.hms.harvard.edu/db/.

222. Chabanon, R. M. *et al.* PARP inhibition enhances tumor cell–intrinsic immunity in ERCC1-deficient non–small cell lung cancer. *Journal of Clinical Investigation* **129,** 1211–1228. ISSN: 15588238 (2019).

223. Kumar, M. *et al.* PARP-1 induces EMT in non-small cell lung carcinoma cells via modulating the transcription factors Smad4, p65 and ZEB1. *Life Sciences* **269,** 118994. ISSN: 18790631 (2021).

224. Marcucci, F. & Rumio, C. Tumor Cell Glycolysis—At the Crossroad of Epithelial–Mesenchymal Transition and Autophagy. *Cells* **11.** ISSN: 20734409 (2022).

225. Korashy, H. M. *et al.* Sunitinib inhibits breast cancer cell proliferation by inducing apoptosis, cell-cycle arrest and DNA repair while inhibiting NF-κB signaling pathways. *Anticancer Research* **37,** 4899–4909. ISSN: 17917530 (2017).

226. Mashimo, M. *et al.* The 89-kDa PARP1 cleavage fragment serves as a cytoplasmic PAR carrier to induce AIF-mediated apoptosis. *Journal of Biological Chemistry* **296,** 100046. ISSN: 1083351X (2021).

227. Gridling, M. *et al.* Identification of Kinase Inhibitor Targets in the Lung Cancer Microenvironment by Chemical and Phosphoproteomics. *Molecular Cancer Therapeutics* **13,** 2751–2762. ISSN: 15378276 (2014).

228. Carneiro, P. J. *et al.* Bosutinib therapy ameliorates lung inflammation and fibrosis in experimental silicosis. *Frontiers in Physiology* **8,** 1–15. ISSN: 1664042X (2017).

229. Shen, T. & Guo, Q. EGFR signaling pathway occupies an important position in cancer-related downstream signaling pathways of Pyk2. *Cell Biology International* **44,** 2–13. ISSN: 10958355 (2020).

230. Zhang, L. *et al.* FRK plays an oncogenic role in non-small cell lung cancer by enhancing the stemness phenotype via induction of metabolic reprogramming. *International Journal of Cancer* **146,** 208–222. ISSN: 10970215 (2020).

231. Goel, R. K. & Lukong, K. E. Understanding the cellular roles of Fyn-related kinase (FRK): implications in cancer biology. *Cancer and Metastasis Reviews* **35,** 179–199. ISSN: 15737233 (2016).

232. Ahmad, S., Hughes, M. A., Johnson, G. L. & Scott, J. E. Development and validation of a high-throughput intrinsic ATPase activity assay for the discovery of MEKK2 inhibitors. *Journal of Biomolecular Screening* **18,** 388–399. ISSN: 10870571 (2013).

233. Huang, T. *et al.* MicroRNA-186 suppresses cell proliferation and metastasis through targeting MAP3K2 in non-small cell lung cancer. *International Journal of Oncology* **49,** 1437–1444. ISSN: 17912423 (2016).

234. Yuan, M. *et al.* SRC and MEK co-inhibition synergistically enhances the anti-tumor effect in both Non-Small Cell Lung Cancer (NSCLC) and erlotinib-resistant NSCLC. *Frontiers in Oncology* **9,** 1–11. ISSN: 2234943X (2019).

235. Suddason, T. & Gallagher, E. A RING to rule them all? Insights into the Map3k1 PHD motif provide a new mechanistic understanding into the diverse roles of Map3k1. *Cell Death and Differentiation* **22,** 540–548. ISSN: 14765403 (2015).

236. Moran, M. *et al.* Sunitinib for Metastatic Renal Cell Carcinoma: A Systematic Review and Meta-Analysis of Real-World and Clinical Trials Data. *Targeted Oncology* **14,** 405–416. ISSN: 1776260X (Aug. 2019).

237. Saltz, L. B. *et al.* Phase II trial of sunitinib in patients with metastatic colorectal cancer after failure of standard therapy. *Journal of Clinical Oncology* **25,** 4793–4799. ISSN: 0732183X (Oct. 2007).

238. Al Baghdadi, T. *et al.* Sunitinib in Patients with Metastatic Colorectal Cancer (mCRC) with FLT-3 Amplification: Results from the Targeted Agent and Profiling Utilization Registry (TAPUR) Study. *Targeted Oncology* **15,** 743–750. ISSN: 1776260X (Dec. 2020).

239. Lamb, Y. N. Correction to: Nintedanib: A Review in Fibrotic Interstitial Lung Diseases (Drugs, (2021), 81, 5, (575-586), 10.1007/s40265-021-01487-0). *Drugs* **81,** 1133–1133. ISSN: 11791950 (May 2021).

240. Hilberg, F. *et al.* BIBF 1120: Triple Angiokinase Inhibitor with Sustained Receptor Blockade and Good Antitumor Efficacy. *Cancer Research* **68,** 4774–4782. ISSN: 0008-5472 (June 2008).

241. Eisen, T. *et al.* A randomised, phase II study of nintedanib or sunitinib in previously untreated patients with advanced renal cell cancer: 3-year results. *British Journal of Cancer* **113,** 1140. ISSN: 15321827 (Oct. 2015).

242. Van Cutsem, E. *et al.* Nintedanib for the treatment of patients with refractory metastatic colorectal cancer (LUME-Colon 1): A phase III, international, randomized, placebo-controlled study. *Annals of Oncology* **29,** 1955–1963. ISSN: 15698041 (Sept. 2018).

243. Hasegawa, H. *et al.* FMS-like tyrosine kinase 3 (FLT3) amplification in patients with metastatic colorectal cancer. *Cancer Science* **112,** 314–322. ISSN: 13497006 (2021).

244. Xiao, X. *et al.* Genistein suppresses FLT4 and inhibits human colorectal cancer metastasis. *Oncotarget* **6,** 3225–3239. ISSN: 19492553 (2015).

245. Mendes Oliveira, D. *et al.* Next-generation sequencing analysis of receptor-type tyrosine kinase genes in surgically resected colon cancer: Identification of gain-of-function mutations in the RET proto-oncogene. *Journal of Experimental and Clinical Cancer Research* **37,** 1–12. ISSN: 17569966 (2018).

246. Kennedy, V. E. & Smith, C. C. FLT3 Mutations in Acute Myeloid Leukemia: Key Concepts and Emerging Controversies. *Frontiers in Oncology* **10,** 1–20. ISSN: 2234943X (2020).

247. Landi, C. *et al.* Common molecular pathways targeted by nintedanib in cancer and IPF: A bioinformatic study. *Pulmonary Pharmacology and Therapeutics* **64,** 101941. ISSN: 15229629 (2020).

248. Aparicio-Gallego, G. *et al.* New insights into molecular mechanisms of sunitinib-associated side effects. *Molecular Cancer Therapeutics* **10,** 2215–2223. ISSN: 15357163 (2011).

249. Hilberg, F. *et al.* Triple angiokinase inhibitor nintedanib directly inhibits tumor cell growth and induces tumor shrinkage via blocking oncogenic receptor tyrosine kinases. *Journal of Pharmacology and Experimental Therapeutics* **364,** 494–503. ISSN: 15210103 (2018).

250. Tang, B. *et al.* xCT contributes to colorectal cancer tumorigenesis through upregulation of the MELK oncogene and activation of the AKT/mTOR cascade. *Cell Death and Disease* **13.** ISSN: 20414889 (2022).

251. Liu, G. *et al.* MELK Accelerates the Progression of Colorectal Cancer via Activating the FAK/Src Pathway. *Biochemical Genetics* **58,** 771–782. ISSN: 15734927 (2020).

252. Li, P. *et al.* FGFR2 Promotes Expression of PD-L1 in Colorectal Cancer via the JAK/STAT3 Signaling Pathway. *The Journal of Immunology* **202,** 3065–3075. ISSN: 0022-1767 (2019).

253. Masuda, M. *et al.* TNIK inhibition abrogates colorectal cancer stemness. *Nature Communications* **7,** 1–7. ISSN: 20411723 (2016).

254. Chen, C. J. & Liu, Y. P. Mertk inhibition: Potential as a treatment strategy in egfr tyrosine kinase inhibitor-resistant non-small cell lung cancer. *Pharmaceuticals* **14,** 1–25. ISSN: 14248247 (2021).

255. Ohta, S. *et al.* The role of MerTK in promoting cell migration is enhanced by the oncogenic Ras/IL-33 signaling axis. *FEBS Journal* **289,** 1950–1967. ISSN: 17424658 (2022).

256. Lafitte, M., Sirvent, A. & Roche, S. Collagen Kinase Receptors as Potential Therapeutic Targets in Metastatic Colon Cancer. *Frontiers in Oncology* **10,** 1–8. ISSN: 2234943X (2020).

257. Jia, S. *et al.* Discoidin domain receptor 2 signaling regulates fibroblast apoptosis through PDK1/akt. *American Journal of Respiratory Cell and Molecular Biology* **59,** 295–305. ISSN: 15354989 (2018).

258. Liu, Y. *et al.* C1222C Deletion in Exon 8 of ABL1 Is Involved in Carcinogenesis and Cell Cycle Control of Colorectal Cancer Through IRS1/PI3K/Akt Pathway. *Frontiers in Oncology* **10,** 1–12. ISSN: 2234943X (2020).

259. Rausch, J. L. *et al.* Opposing roles of KIT and ABL1 in the therapeutic response of gastrointestinal stromal tumor (GIST) cells to imatinib mesylate. *Oncotarget* **8,** 4471–4483. ISSN: 19492553 (2017).

260. Di Segni, M. *et al.* HIPK2 Cooperates with KRAS Signaling and Associates with Colorectal Cancer Progression. *Molecular Cancer Research* **20,** 686–698. ISSN: 15573125 (2022).

261. Li, Y. *et al.* MiR-372-3p Functions as a Tumor Suppressor in Colon Cancer by Targeting MAP3K2. *Frontiers in Genetics* **13,** 1–12. ISSN: 16648021 (2022).

262. Lim, S. H. *et al.* The implication of FLT3 amplification for FLT targeted therapeutics in solid tumors. *Oncotarget* **8,** 3237–3245. ISSN: 19492553 (2017).

263. Hsu, M. C., Pan, M. R. & Hung, W. C. Two birds, one stone: Double hits on tumor growth and lymphangiogenesis by targeting vascular endothelial growth factor receptor 3. *Cells* **8.** ISSN: 20734409 (2019).

264. Gao, Y., Zhou, J. & Li, J. Discoidin domain receptors orchestrate cancer progression: A focus on cancer therapies. *Cancer Science* **112,** 962–969. ISSN: 13497006 (2021).

265. Sasaki, S. *et al.* DDR2 expression is associated with a high frequency of peritoneal dissemination and poor prognosis in colorectal cancer. *Anticancer Research* **37,** 2587–2591. ISSN: 17917530 (2017).

266. Luttman, J. H., Colemon, A., Mayro, B. & Pendergast, A. M. Role of the ABL tyrosine kinases in the epithelial–mesenchymal transition and the metastatic cascade. *Cell Communication and Signaling* **19,** 1–16. ISSN: 1478811X (2021).

267. Wang, J. & Pendergast, A. M. The Emerging Role of ABL Kinases in Solid Tumors. *Trends in Cancer* **1,** 110–123. ISSN: 24058033 (Oct. 2015).

268. Slattery, M. L. *et al.* The MAPK-signaling pathway in colorectal cancer: Dysregulated genes and their association with micrornas. *Cancer Informatics* **17.** ISSN: 11769351 (2018).

269. Thangaraj, K., Ponnusamy, L., Natarajan, S. R. & Manoharan, R. MELK/MPK38 in cancer: from mechanistic aspects to therapeutic strategies. *Drug Discovery Today* **25,** 2161–2173. ISSN: 18785832 (2020).

270. Tang, Q. *et al.* MELK is an oncogenic kinase essential for metastasis, mitotic progression, and programmed death in lung carcinoma. *Signal Transduction and Targeted Therapy* **5.** ISSN: 20593635 (2020).

271. Nass, N. *et al.* Validation of VX2 as a hepatocellular carcinoma model: Comparison of the molecular reaction of VX2 and HepG2 tumor cells to sorafenib in vitro. *Anticancer Research* **37,** 87–93. ISSN: 17917530 (2017).

272. Samanta, D. *et al.* BIRC2 Expression Impairs Anti-Cancer Immunity and Immunotherapy Efficacy. *Cell Reports* **32,** 108073. ISSN: 22111247 (2020).

273. Dagenais, M. *et al.* A critical role for cellular inhibitor of protein 2 (cIAP2) in colitis-associated colorectal cancer and intestinal homeostasis mediated by the inflammasome and survival pathways. *Mucosal Immunology* **9,** 146–158. ISSN: 19353456 (2016).

274. Zeng, K. *et al.* CircHIPK3 promotes colorectal cancer growth and metastasis by sponging miR-7 article. *Cell Death and Disease* **9.** ISSN: 20414889 (2018).

275. SHAO, Q., HUANG, Y., ZHANG, C., GAO, X. & GAO, S. Emerging landscape of circHIPK3 and its role in cancer and other diseases (Review). *Molecular Medicine Reports* **23.** ISSN: 17913004 (2021).

276. Mohanty, S. *et al.* CRISPR-based kinome-screening revealed MINK1 as a druggable player to rewire 5FU-resistance in OSCC through AKT/MDM2/p53 axis. *Oncogene* **41,** 4929–4940. ISSN: 14765594 (2022).

277. Casalvieri, K. A., Matheson, C. J., Backos, D. S. & Reigan, P. Selective Targeting of RSK Isoforms in Cancer. *Trends in Cancer* **3,** 302–312. ISSN: 24058033 (2017).

278. Poomakkoth, N., Issa, A., Abdulrahman, N., Abdelaziz, S. G. & Mraiche, F. p90 ribosomal S6 kinase: A potential therapeutic target in lung cancer. *Journal of Translational Medicine* **14,** 4–9. ISSN: 14795876 (2016).

279. Knippschild, U. *et al.* The CK1 family: Contribution to cellular stress response and its role in carcinogenesis. *Frontiers in Oncology* **4 MAY,** 1–32. ISSN: 2234943X (2014).

280. Ma, F. *et al.* Effects of brassinosteroids on cancer cells: A review. *Journal of Biochemical and Molecular Toxicology* **36,** 1–9. ISSN: 10990461 (2022).

281. Kohli, S. K. *et al.* Therapeutic potential of brassinosteroids in biomedical and clinical research. *Biomolecules* **10.** ISSN: 2218273X (2020).

282. Xie, Y., Li, C., Huang, Y., Jia, Z. & Cao, J. A novel multikinase inhibitor R8 exhibits potent inhibition on cancer cells through both apoptosis and autophagic cell death. *Oncotarget* **8,** 87209–87220. ISSN: 19492553 (2017).

283. Lorca, M. *et al.* Cancer and brassinosteroids: Mechanisms of action, SAR and future perspectives. *Steroids* **190.** ISSN: 18785867 (2023).

284. Bushra, R., Aslam, N. & Khan, A. Y. Food-drug interactions. *Oman Medical Journal* **26,** 77–83. ISSN: 20705204 (2011).

285. Musa, A. *et al.* A review of connectivity map and computational approaches in pharmacogenomics. *Briefings in Bioinformatics* **19,** 506–523. ISSN: 14774054 (May 2018).

286. Moffat, J. G., Vincent, F., Lee, J. A., Eder, J. & Prunotto, M. Opportunities and challenges in phenotypic drug discovery: An industry perspective. *Nature Reviews Drug Discovery* **16,** 531–543. ISSN: 14741784 (2017).

287. Vincent, F. *et al. Phenotypic drug discovery: recent successes, lessons learned and new directions* Dec. 2022.

288. Hughes, J. P., Rees, S. S., Kalindjian, S. B. & Philpott, K. L. Principles of early drug discovery. *British Journal of Pharmacology* **162,** 1239–1249. ISSN: 00071188 (2011).

289. Reichert, J. M. Trends in development and approval times for new therapeutics in the United States. *Nature Reviews Drug Discovery* **2,** 695–702. ISSN: 14741776 (2003).

290. Mohs, R. C. & Greig, N. H. Drug discovery and development: Role of basic biological research. *Alzheimer's and Dementia: Translational Research and Clinical Interventions* **3,** 651–657. ISSN: 23528737 (2017).

291. Keenan, A. B. *et al.* Connectivity Mapping: Methods and Applications (2019).

292. Keenan, A. B. *et al.* The Library of Integrated Network-based Cellular Signatures (LINCS) NIH Program: System-level Cataloging of Human Cells Response to Perturbations. *Cell System* **6,** 13–24. ISSN: 1472-3263 (2018).

293. Lamb, J. *et al.* The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *science* **313,** 1929–1935 (2006).

294. Samart, K., Tuyishime, P., Krishnan, A. & Ravi, J. Reconciling multiple connectivity scores for drug repurposing. *Briefings in Bioinformatics* **22.** ISSN: 14774054 (Nov. 2021).

295. Guney, E. Reproducible drug repurposing: When similarity does not suffice. *Pacific Symposium on Biocomputing* **0,** 132–143. ISSN: 23356936 (2017).

296. Chen, B. *et al.* Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. *Nature Communications* **8.** ISSN: 20411723 (July 2017).

297. Zeng, B. *et al.* OCTAD: an open workspace for virtually screening therapeutics targeting precise cancer patient groups using gene expression features. *Nature Protocols 2020 16:2* **16,** 728–753 (Dec. 2020).

298. Chen, B. *et al.* Computational Discovery of Niclosamide Ethanolamine, a Repurposed Drug Candidate That Reduces Growth of Hepatocellular Carcinoma Cells In Vitro and in Mice by Inhibiting Cell Division Cycle 37 Signaling. *Gastroenterology* **152,** 2022–2036. ISSN: 15280012 (June 2017).

299. Pessetto, Z. Y. *et al.* In silico and in vitro drug screening identifies new therapeutic approaches for Ewing sarcoma. *Oncotarget* **8,** 4079–4095. ISSN: 1949-2553 (Nov. 2016).

300. Morselli Gysi, D. *et al.* Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proceedings of the National Academy of Sciences of the United States of America* **118.** ISSN: 1091-6490 (May 2021).

301. Iskar, M. *et al.* Drug-Induced Regulation of Target Expression. *PLOS Computational Biology* **6,** e1000925. ISSN: 1553-7358 (2010).

302. Cheng, J. *et al.* Evaluation of analytical methods for connectivity map data. *Pacific Symposium on Biocomputing,* 5–16. ISSN: 23356936 (2013).

303. Duran-Frigola, M. *et al.* Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker. *Nature Biotechnology* **38,** 1087–1096. ISSN: 15461696 (2020).

304. Jie Cheng, Lun Yang, Vinod Kumar & Pankaj Agarwal. Systematic evaluation of connectivity map for disease indications. *Genome Medicine* **6.** ISSN: 1756994X (Oct. 2014).

305. Zhu, J. *et al.* Prediction of drug efficacy from transcriptional profiles with deep learning. *Nature Biotechnology 2021 39:11* **39,** 1444–1452. ISSN: 1546-1696 (June 2021).

306. Pham, T. H., Qiu, Y., Zeng, J., Xie, L. & Zhang, P. A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing. *Nature Machine Intelligence 2021 3:3* **3,** 247–257. ISSN: 2522-5839 (Feb. 2021).

307. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nature Methods 2019 16:8* **16,** 715–721. ISSN: 1548-7105 (July 2019).

308. Hetzel, L. *et al.* Predicting Cellular Responses to Novel Drug Perturbations at a Single-Cell Resolution.

309. Hauser, A. & Bühlmann, P. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning* **55,** 926–939. ISSN: 0888613X (2014).

310. Ghassami, A. E., Salehkaleybar, S., Kiyavash, N. & Bareinboim, E. Budgeted experiment design for causal structure learning. *35th International Conference on Machine Learning, ICML 2018* **4,** 2788–2801 (2018).

311. Agrawal, R., Squires, C., Yang, K., Shanmugam, K. & Uhler, C. ABCD-strategy: Budgeted experimental design for targeted causal structure discovery. *AISTATS 2019 - 22nd International Conference on Artificial Intelligence and Statistics* **89** (2020).

312. Mueller, J., Reshef, D. N., Du, G. & Jaakkola, T. Learning Optimal Interventions (2016).

313. Pacchiano, A. & Barton, R. A. Neural Design for Genetic Perturbation Experiments, 1–37.

314. Mueller, J., Gifford, D. & Jaakkola, T. Sequence to Better Sequence : Continuous Revision of Combinatorial Structures (2017).

315. Hie, B., Bryson, B. D., Zhong, E. D. & Berger, B. Learning Mutational Semantics, 1–13 (2020).

316. Zhang, J., Squires, C. & Uhler, C. Matching a Desired Causal State via Shift Interventions (2021).

317. Zhang, J., Cammarata, L., Squires, C., Sapsis, T. P. & Uhler, C. Active Learning for Optimal Intervention Design in Causal Models (2022).

318. Deng, Z., Zheng, X., Tian, H. U. & Zeng, D. D. Deep Causal Learning: Representation, Discovery and Inference.

319. Parafita, Á. & Vitrià, J. Causal Inference with Deep Causal Graphs (2020).

320. Pawlowski, N., Castro, D. C. & Glocker, B. Deep structural causal models for tractable counterfactual inference. *Advances in Neural Information Processing Systems* **2020-Decem.** ISSN: 10495258 (2020).

321. Xia, K., Lee, K.-Z., Bengio, Y. & Bareinboim, E. The Causal-Neural Connection: Expressiveness, Learnability, and Inference.

322. Xia, K., Pan, Y. & Bareinboim, E. Neural Causal Models for Counterfactual Identification and Estimation. **2,** 1–57 (2022).

323. Bronstein, M. M., Bruna, J., Cohen, T. & Veličković, P. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges (2021).

324. Zečević, M., Dhami, D. S., Veličković, P. & Kersting, K. Relating Graph Neural Networks to Structural Causal Models (2021).

325. Oughtred, R. *et al.* The BioGRID interaction database: 2019 update. *Nucleic Acids Research* **47,** D529–D541. ISSN: 0305-1048 (Jan. 2019).

326. Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature 2020 580:7803* **580,** 402–408. ISSN: 1476-4687 (Apr. 2020).

327. Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* **347,** 841. ISSN: 10959203 (Feb. 2015).

328. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research* **35,** D26. ISSN: 03051048 (Jan. 2007).

329. Tweedie, S. *et al.* Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Research* **49,** D939–D946. ISSN: 0305-1048 (Jan. 2021).

330. Cunningham, F. *et al.* Ensembl 2022. *Database issue Nucleic Acids Research* **50,** 989 (2022).

331. Stathias, V. *et al.* LINCS Data Portal 2.0: next generation access point for perturbation-response signatures. *Nucleic Acids Research* **48,** D431–D439. ISSN: 0305-1048 (Jan. 2020).

332. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research* **47,** D941–D947. ISSN: 0305-1048 (Jan. 2019).

333. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics* **7,** 1–34. ISSN: 17582946 (May 2015).

334. Lotfollahi, M. *et al.* Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular systems biology,* e11517. ISSN: 1744-4292 (2023).

335. Heumos, L. *et al.* Best practices for single-cell analysis across modalities. *Nature Reviews Genetics 2023,* 1–23. ISSN: 1471-0064 (Mar. 2023).

336. Gayoso, A. *et al.* A Python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology 2022 40:2* **40,** 163–166. ISSN: 1546-1696 (Feb. 2022).

337. Jovic, D. *et al.* Single-cell RNA sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine* **12,** e694. ISSN: 2001-1326 (Mar. 2022).

338. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods 2021 19:1* **19,** 41–50. ISSN: 1548-7105 (Dec. 2021).