scientific reports

OPEN

Genomic-driven nutritional interventions for radiotherapy-resistant rectal cancer patient

Joshua Southern[1,8], Guadalupe Gonzalez[1,2,8], Pia Borgas[3], Liam Poynter[4], Ivan Laponogov[4], Yoyo Zhong[4], Reza Mirnezami[5], Dennis Veselkov[1], Michael Bronstein[6] & Kirill Veselkov[2,7]✉

Radiotherapy response of rectal cancer patients is dependent on a myriad of molecular mechanisms including response to stress, cell death, and cell metabolism. Modulation of lipid metabolism emerges as a unique strategy to improve radiotherapy outcomes due to its accessibility by bioactive molecules within foods. Even though a few radioresponse modulators have been identified using experimental techniques, trying to experimentally identify all potential modulators is intractable. Here we introduce a machine learning (ML) approach to interrogate the space of bioactive molecules within food for potential modulators of radiotherapy response and provide phytochemically-enriched recipes that encapsulate the benefits of discovered radiotherapy modulators. Potential radioresponse modulators were identified using a genomic-driven network ML approach, metric learning and domain knowledge. Then, recipes from the Recipe1M database were optimized to provide ingredient substitutions maximizing the number of predicted modulators whilst preserving the recipe's culinary attributes. This work provides a pipeline for the design of genomic-driven nutritional interventions to improve outcomes of rectal cancer patients undergoing radiotherapy.

Mesorectal excision is the surgical standard of care in rectal cancer (RC)[1]. The additive benefit of radiotherapy (RT) in reducing local recurrence in advanced RC has been extensively documented[2–5]. However, there is considerable variability in radioresponse across patients, with patients showing either (1) complete tumor destruction, (2) moderate tumor regression, or (3) negligible tumor shrinkage. For patients in the last category, the delay in proceeding to tumor excision while completing RT may increase the likelihood of distant metastases, therefore, modulation of radioresponse to improve RT outcomes is a critical need.

RT response is governed by various molecular mechanisms including response to stress, cell death, and cell metabolism[6]. Current strategies to improve radioresponse focus on the modulation of cell death and response to stress using chemotherapies, such as fluorouracil (5-FU), capecitabine, gemcitabine, and cisplatin, to enhance tumor sensitivity to RT[7]. However, combined therapy often produces mixed results and can increase toxicity in normal tissues[7]. In contrast, bioactive molecules within foods appear as a promising alternative to modulate radioresponse, through lipid metabolism modulation[8,9]. Proteins corresponding to up-regulated genes in RT-resistant RC patients participate in lipid biosynthetic and metabolic pathways with various roles (Fig. 1A). The up-regulation of most of these genes translates into increased lipid availability, which leads to a myriad of downstream tumor-promoting effects. For example, CDS1- and CDS2-encoded proteins regulate growth and maturation of lipid droplets which serve as storage, providing nutrients necessary for cell growth, and can serve as additional nutrients for the uncontrolled growth of cancerous cells[10]. Moreover, over-expression of PLA2G5, the ELOVL family of genes, FASN and the PLP family of genes translates into increased lipid availability leading to downstream activation of inflammation and stress pathways. Proteins encoded by these genes increase lipid availability through different mechanisms: PLA2G5-encoded protein through the generation of lysophospholipids and free fatty acids, including arachidonic acid[10,11]; encoded proteins by the ELOVL family of genes through

[1]Department of Computing, Imperial College London, London SW7 2BX, UK. [2]Prescient Design, Genentech, Basel 4052, Switzerland. [3]North Middlesex University Hospital, London N18 1QX, UK. [4]Department of Surgery and Cancer, Imperial College London, London SW7 2BX, UK. [5]Royal Free Hospital, London NW3 2QG, UK. [6]Department of Computer Science, University of Oxford, Oxford OX1 3QD, UK. [7]Department of Environmental Health Sciences, Yale University, New Haven, CT 06510, USA. [8]These authors contributed equally: Joshua Southern and Guadalupe Gonzalez. ✉email: kirill.veselkov04@imperial.ac.uk

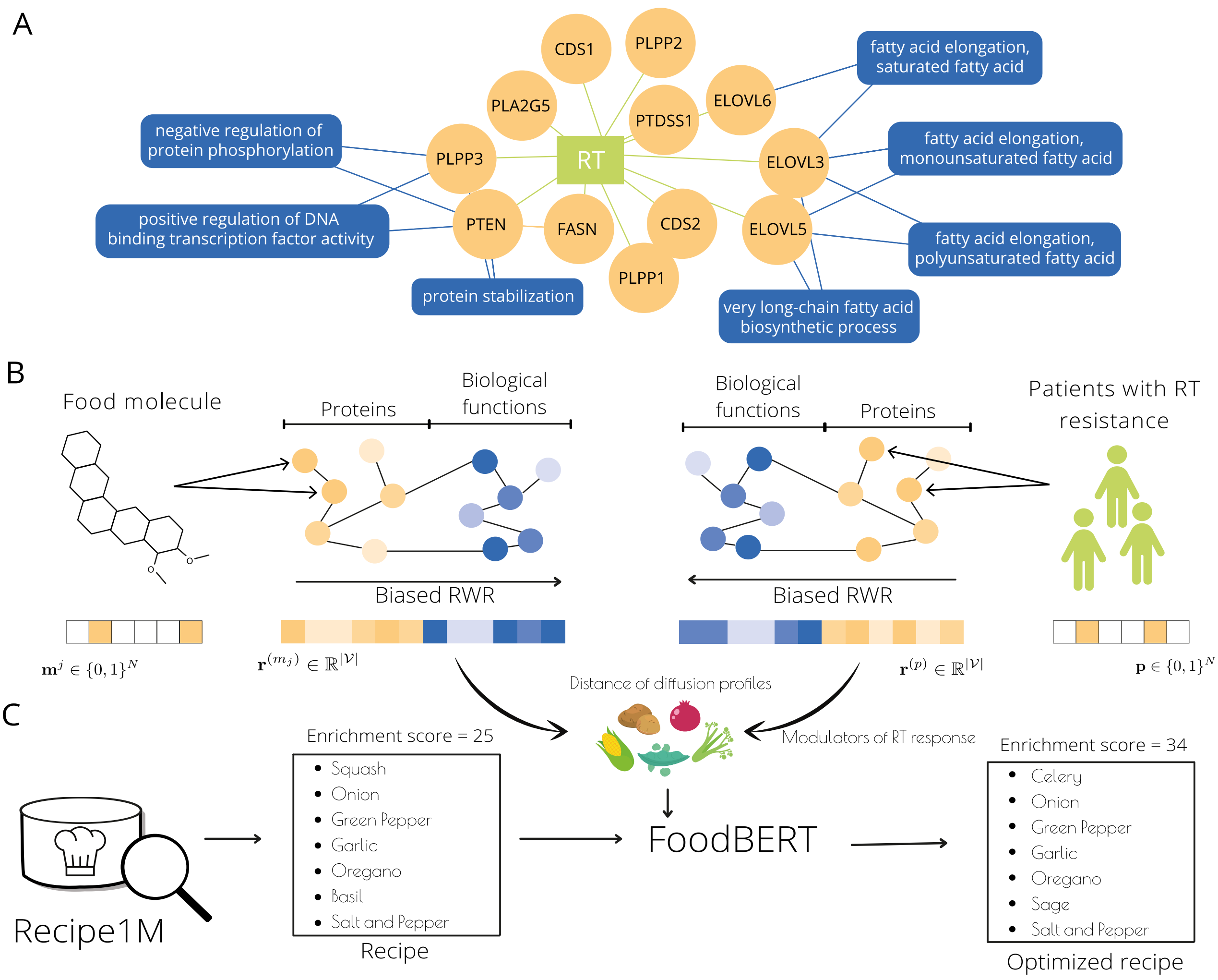
A
CDS1
PLPP2
PLA2G5
PTDSS1
ELOVL6
fatty acid elongation, saturated fatty acid
negative regulation of protein phosphorylation
PLPP3
RT
ELOVL3
fatty acid elongation, monounsaturated fatty acid
positive regulation of DNA binding transcription factor activity
PTEN
FASN
CDS2
ELOVL5
fatty acid elongation, polyunsaturated fatty acid
PLPP1
protein stabilization
very long-chain fatty acid biosynthetic process
B
Food molecule
Proteins
Biological functions
Biological functions
Proteins
Patients with RT resistance
Biased RWR
Biased RWR
$\mathbf{m}^j \in \{0,1\}^N$
$\mathbf{r}^{(m_j)} \in \mathbb{R}^{|\mathcal{V}|}$
$\mathbf{r}^{(p)} \in \mathbb{R}^{|\mathcal{V}|}$
$\mathbf{p} \in \{0,1\}^N$
Distance of diffusion profiles
C
Enrichment score = 25
Squash
Onion
Green Pepper
Garlic
Oregano
Basil
Salt and Pepper
Recipe
Modulators of RT response
FoodBERT
Recipe1M
Enrichment score = 34
Celery
Onion
Green Pepper
Garlic
Oregano
Sage
Salt and Pepper
Optimized recipe

Figure 1. Overview of approach. (**A**) Network representation of over-expressed proteins (yellow) and biological functions (blue) in RT-resistant RC patients. Over-expressed proteins were experimentally identified by one of the authors of this work. Their corresponding biological functions were extracted from the Gene Ontology's Biological Processes[32]. Over-expression of all genes but PTEN leads to increased lipid availability resulting in cancer-promoting effects including increased nutrient storage (CDS1, CDS2), activation of stress and response signaling pathways (PLA2G5, ELOVL family, FASN and PLP family), and increased immunosuppressive properties (PTDSS1). (**B**) Radioresponse modulators identification module. Food protein targets and RT-resistant-associated proteins are mapped onto a multiscale interactome of proteins and biological functions. A biased random walk with restarts (RWR) propagates the effects of food molecules and phenotype, revealing the most affected proteins and biological functions. Top food molecules with the most similar propagated profiles to the phenotype are used to create a list of food ingredients with potentially beneficial RT response modulation activity. (**C**) Recipe generation module. Using FoodBERT, Recipe1M recipes are optimized to increase the number of ingredients with beneficial radioresponse modulation properties.

the elongation of long chain fatty acids to provide precursors for synthesis of sphingolipids and ceramides[10,12]; FASN-encoded protein through the synthesis of long-chain fatty acids[10]; and encoded proteins by the PLP family of genes through the hydrolysis and uptake of lipids from extracellular space[10,13]. Increased lipid availability in cancer cells can also lead to increased immunosuppressive properties, as is the case with PTDSS1 over-expression, whose encoded protein catalyzes the formation of phosphatidylserine which, exposed on the surface of tumor cells, increases their immunosuppressive properties and facilitates tumor growth and metastasis[10,14]. On the other hand, PTEN has documented tumor-suppressing properties[10]. Loss of PTEN leads to elevated de novo lipogenesis through induction of SREBP and FASN expression[15]. Therefore, over-expression of PTEN in this context might be a compensatory mechanism to inhibit FASN in an attempt to decrease lipogenesis.

Bioactive molecules in food can modulate lipid metabolism, have a promising safety profile in toxicity studies, and have documented chemopreventive and chemotherapeutic effects[16–18]. This means dietary interventions could be a promising strategy to increase treatment efficacy, prevent resistance acquisition and reduce side effects[19]. However, experimental large-scale testing of chemotherapeutic or chemopreventive properties of bioactive molecules within food is not generally feasible due to a large number of food-based bioactive molecules. As a result, a unique wave of research has leveraged network machine learning (ML) and genomic data to carry out a large-scale screening of anticancer molecules within food[20–22]. Building on these works, we propose a computational genomic-driven approach to mine the space of bioactive molecules within food for potential radioresponse modulators and propose phytochemically-enriched recipes to improve radioresponse of RC patients.

The proposed pipeline, shown in Fig. 1, comprises (1) Identifying over-expressed proteins in RT-resistant RC patients (Fig. 1A) (2) a radioresponse modulators identification module (Fig. 1B) and (3) a recipe generation module (Fig. 1C). In order to identify radioresponse modulators, we map food molecule protein-coding gene targets and RT resistance dysregulated genes onto a heterogeneous network representing proteins and biological functions. Using a network propagation algorithm combined with metric learning, we learn effects of food molecules and the phenotype across the heterogeneous network, and find food molecules with similar effects to those observed in the phenotype. The third stage involves recipe optimization to maximise the number of ingredients with these molecules. Dietary recommendations can then be proposed for RT-resistant RC patients using these recipes and taking into account other user-specific requirements such as taste preferences and allergies.

Results

Random walks and metric learning to predict drug-phenotype associations.

We compute propagated profiles of drugs and diseases on the multiscale-interactome using random-walk with restarts and then use metric learning to minimise the distance between a disease and drugs that treat this disease and maximise the distance between a disease and drugs with no known benefit. To show the gain of combining metric learning with the random walk algorithm, we evaluated the improvement on the multiscale-based drug-disease prediction task proposed in[23]. We show that the addition of metric learning improves the random walk diffusion profiles resulting in a 20% increase in performance ($AUROC = 0.714$ *vs* 0.876). Additionally, the choice of the restart probability only has a small effect on the results in the initial implementation and no influence when combined with metric learning. These results, shown in Fig. 2, confirm the benefit of fixing the restart probability and instead of optimising the weights of the walker, optimizing an MLP by directly back-propagating information from the prediction task using a triplet loss function.

The model identifies molecules with therapeutic potential to reverse RT resistance.

Using propagated profiles, we find the top 100 food molecules closest to the phenotype. These molecules affect similar proteins and biological functions as those responsible for radioresistance, however, diffusion profiles do not provide information whether the modulation is positive or negative. Experimental evidence indicates that the phenotype-associated genes are over-expressed in patients exhibiting RT resistance leading to a positive modulation of lipid metabolism (Fig. 1A). Therefore, we use domain knowledge and literature search to filter out identified molecules with positive regulatory effects on lipid metabolism, leaving 33 modulators to retrieve the list of ingredients (Appendix A). Modulators belong to a myriad of compound classes including flavonoids, isoflavonoids, and bezenoids, in alignment with the current knowledge on chemotherapeutic bioactive molecules within foods[11]. Overall, predicted modulators are involved in cell signaling, cell growth and lipid metabolism. For example, Mangiferol and Dihydrosphingosine modulate downstream effects linked to fatty acid biosynthetic and elongation pathways, down-regulating stress and inflammation processes (Figure 3). Additionally, we have compiled a list of ingredients with the highest number of modulators (Appendix B).

Highest scoring foods modulating RT response.

In order to validate the recipes, we explored the mechanisms by which the substituted ingredients could modulate radiotherapy response. The tables in appendices A

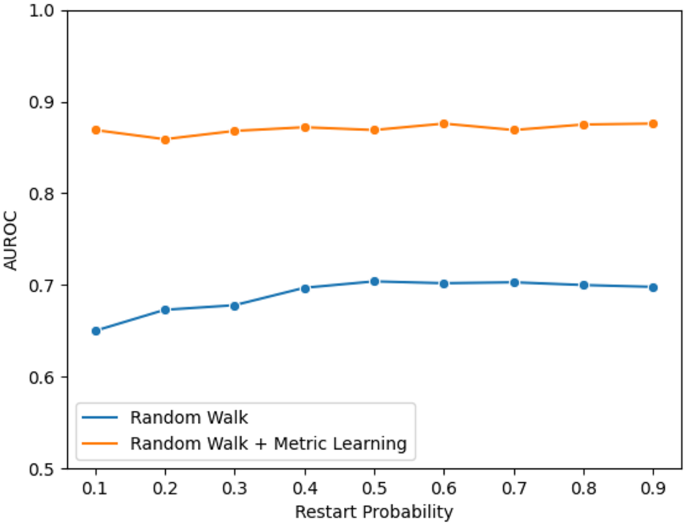

Figure 2. The presented model outperforms the baseline approach across all values of restart probability. AUROC of purely random walk based approach[23] and the proposed approach combining the method with metric learning as a function of restart probability on the drug-disease prediction task.

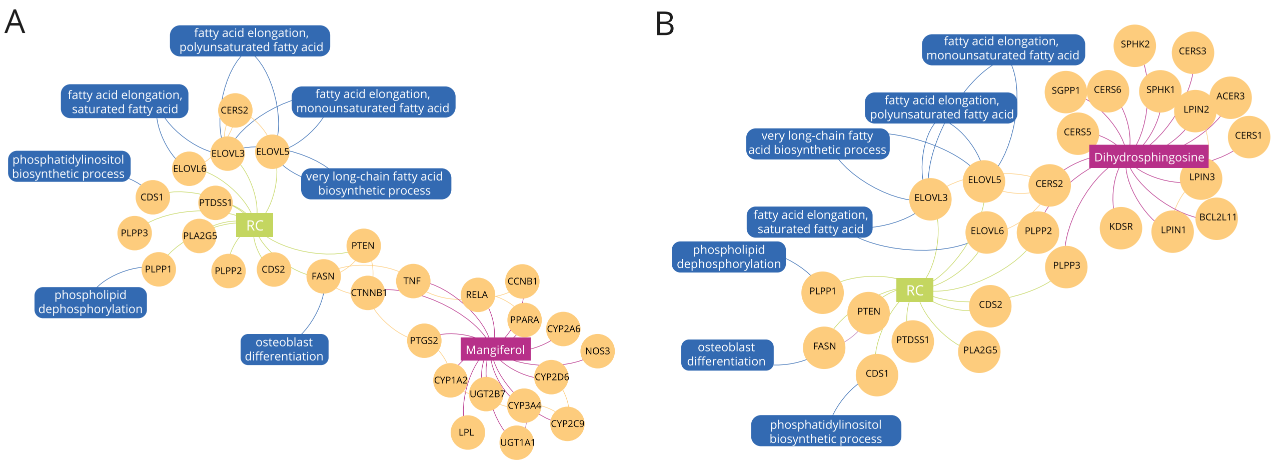
A
fatty acid elongation, polyunsaturated fatty acid
fatty acid elongation, saturated fatty acid
fatty acid elongation, monounsaturated fatty acid
phosphatidylinositol biosynthetic process
very long-chain fatty acid biosynthetic process
phospholipid dephosphorylation
osteoblast differentiation
CERS2
ELOVL3
ELOVL5
ELOVL6
CDS1
PTDSS1
PLPP3
PLA2G5
RC
PLPP1
PLPP2
CDS2
PTEN
FASN
CTNNB1
TNF
CCNB1
RELA
PPARA
CYP2A6
PTGS2
Mangiferol
NOS3
CYP1A2
UGT2B7
CYP2D6
CYP3A4
LPL
UGT1A1
CYP2C9
B
fatty acid elongation, monounsaturated fatty acid
fatty acid elongation, polyunsaturated fatty acid
very long-chain fatty acid biosynthetic process
fatty acid elongation, saturated fatty acid
phospholipid dephosphorylation
osteoblast differentiation
phosphatidylinositol biosynthetic process
SPHK2
CERS3
SGPP1
CERS6
SPHK1
ACER3
LPIN2
CERS5
CERS1
Dihydrosphingosine
LPIN3
ELOVL5
CERS2
ELOVL3
ELOVL6
PLPP2
KDSR
LPIN1
BCL2L11
PLPP3
PLPP1
RC
PTEN
CDS2
PTDSS1
FASN
PLA2G5
CDS1

Figure 3. The multiscale interactome identifies proteins and biological functions related to RT modulation. (**A**) The multiscale interactome involving Mangiferol and RT modulation. Mangiferol modulates RT response by targeting CTNNB1 and TNF which has downstream effects in osteoblast differentiation by being linked to FASN. (**B**) The multiscale interactome involving Dihydrosphingosine and RT modulation. Dihydrosphingosine modulates RT response primarily by targeting PLPP2, PLPP3 and CERS2, which are linked to the ELOVL family of genes, inhibiting fatty acid biosynthetic and elongation processes.

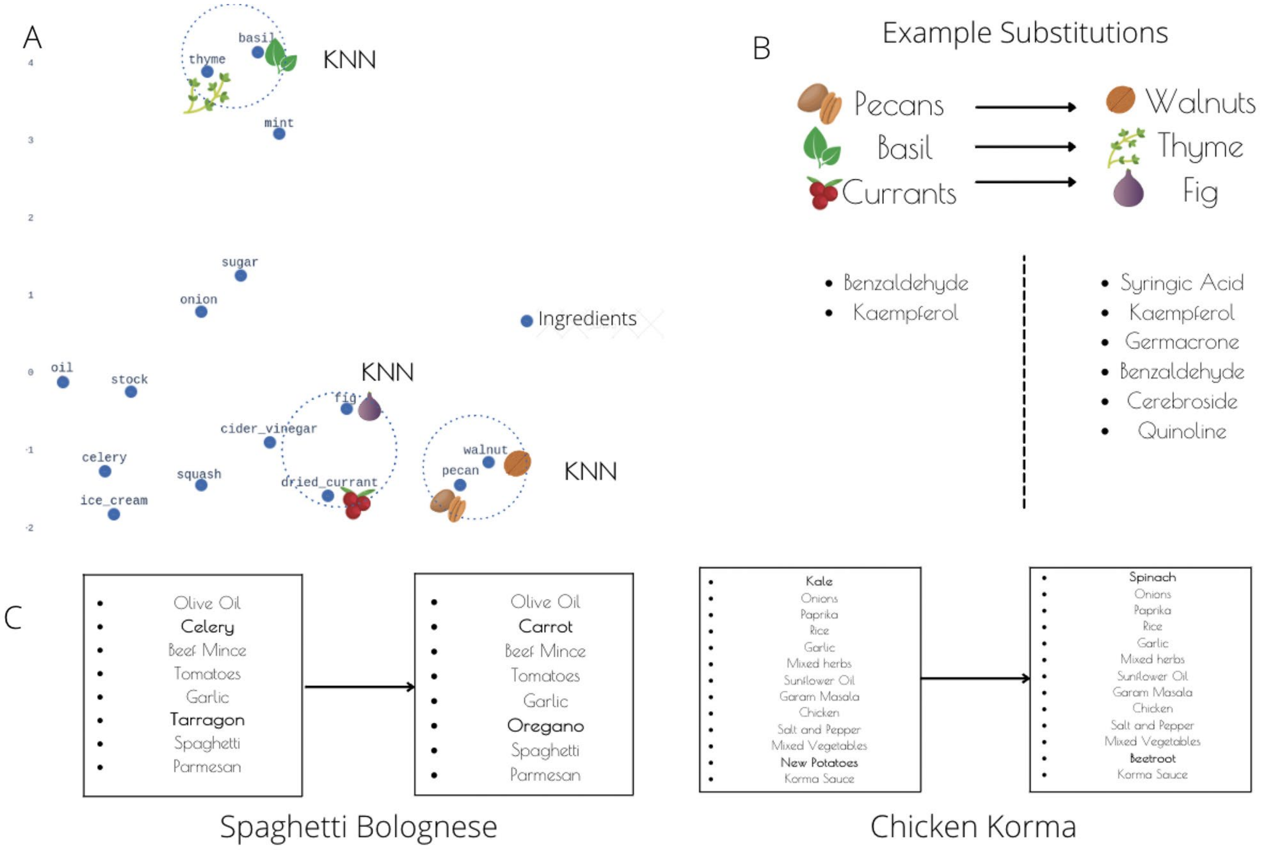

Figure 4. Ingredient substitutions using FoodBert embeddings and food-chemical information. (**A**) Visualisation of FoodBert embeddings. 2D representations found using PCA of the 768 dimensional FoodBert embeddings for some ingredients. Ingredients close in this space appear in similar contexts. (**B**) Example ingredient substitutions. Some example substitutions found by the K-nearest neighbor algorithm in the embedding space and additional filtering to increase the number of beneficial bioactive molecules. (**C**) Some example ingredient substitutions within popular recipes.

and B give a more extensive treatment of the RT response modulators within food and their potential mechanism for modulation. In Fig. 4, we show how a chicken korma recipe is mutated by substituting kale for spinach and new potatoes for beetroot. Whilst it is difficult to evaluate these substitutions from a culinary perspective, the substitutions do increase the number of potential radioresponse modulators. New potatoes contain none of the found potential modulators, whereas beetroot contains both kaempferol and syringic acid. It has been shown that syringic acid-treated cells developed anti-cancer activities by losing MMP, cell viability, and enhancing intracellular ROS and kaempferol has been shown to be a potential chemo-therapeutic agent to be used alone or in combination with 5-FU to overcome colon cancer drug resistance[2425]. Additionally, spinach also contains kaempferol as well as alpha-lipoic acid. Alpha-lipoic acid can effectively induce apoptosis in human colon cancer cells by a mechanism that is initiated by an increased uptake of oxidizable substrates into mitochondria[26]. The addition of these molecules in the recipe, which have been found using our drug-disease association model, and have demonstrated chemotherapeutic effects could be beneficial to radiotherapy-resistant rectal cancer patient as an added measure alongside their standard treatment.

Discussion

In 2017, dietary risk factors were attributed to approximately 11 million deaths globally, equivalent to about 1 in 5 deaths[27]. This stark statistic emphasises the global need for dietary improvements. Furthermore, evidence has mounted on the potential benefits of drug-like molecules in foods against diseases such as cancer[28,29], Covid-19[30] and other health conditions[31]. The prospect of dietary recommendations both for the general population and patients with specific diseases becomes increasingly important. We delved into understanding the role of bioactive food molecules as potential modulators of radiotherapy response. This was achieved by expanding a drug-disease prediction model based on RWR with metric learning, pinpointing radioresponse modulators and showcasing enhanced results on a benchmark dataset. The integration of these analytical methodologies is pivotal; it not only facilitates a comprehensive understanding of intricate interactions but also combines the strengths of prediction and metric learning, ensuring a system-wide appraisal of the potential therapeutic influence of bioactive food molecules on radiotherapy efficacy. By utilising propagated profiles from our model, we identified radioresponse modulators in food, subsequently integrating this with experimental evidence from literature reviews to determine the modulation direction - either positive or negative. It is important to acknowledge, however, that while our findings are encouraging, the model's transfer-ability may necessitate further validations. This arises from discrepancies, albeit reasonable, in data distribution between the dataset for optimisation of propagation weights and the datasets for food molecules and phenotypes (Appendix C). In this study, we adopted an assumption of direct correspondence between effects on genes and proteins, neglecting potential post-translational modifications. These modifications could be profoundly influenced by dietary intake and merit further exploration in subsequent studies. In terms of advancements, future iterations of the recipe recommendation module could contemplate the de novo creation of recipes using text or cooking graph representations, surpassing current NLP-based models. Furthermore, the optimal timing for dietary interventions, aimed at maximising radiotherapy outcomes, was beyond our current scope but warrants attention, potentially encompassing clinical trials assessing the interplay between dietary intervention timing and therapy outcomes. The current work, in general, provides a framework for the discussion of methodological approaches for the task of modulating radioresponse using bioactive molecules within foods. We consider this work as a first milestone approach in the design of genome-guided phytochemically-enriched recipes to improve RT outcomes in RC patients and envision its use as a baseline for future work. Our approach, centred on lipid metabolism modulation, offers a novel avenue to augment radiotherapy outcomes. Nevertheless, individual biological and health variations signify that it might not universally benefit all patients. Aspects like obesity and BMI, which intrinsically modify lipid metabolism and various physiological processes, could dictate the intervention's effectiveness. In such instances, personalised strategies, ranging from dietary modifications to manage weight to pharmacological measures addressing obesity-related comorbidities, might be indispensable. By employing machine learning, our study enables recipe adjustments in line with identified potential radiotherapy modulators. This presents an opportunity for bespoke recipe alterations aligning with individual patient requirements, considering elements like obesity and BMI. Such a comprehensive, personalised treatment paradigm accentuates the essence of optimising radiotherapy outcomes and overall patient health. The flexibility of our approach also encompasses patient-specific data such as allergies, cost considerations, food preferences, and concurrent treatments, ensuring dietary compatibility and synergy.

Contrasting with gut microbiota modification strategies, our method prioritises direct dietary alterations aimed at cellular mechanisms, including lipid metabolism, rather than reshaping the gut microbiome. Nevertheless, dietary effects on gut microbiota composition and functionality are undeniable and can sway health outcomes, including therapy responses. Since the gut microbiota orchestrates the bioavailability of bioactive food molecules, these two strategies might be synergistically combined for a comprehensive therapeutic approach encompassing both cellular mechanisms and microbial interactions. Our proposed pipeline possesses the adaptability to address any disease given the knowledge of target genes, offering a holistic framework for recipe recommendations that complement prevailing treatment standards. We foresee evaluating these findings via clinical trials, providing participants with enriched recipes and evaluating dietary intervention impacts through outcomes such as progression-free survival (PFS) or disease-free survival (DFS). Moreover, the approach, although primarily centred on radiotherapy for rectal cancer, hints at the broader applicability, extending possibly to other therapeutic modalities or diseases.

Conclusion

We introduce a network machine learning pipeline for predicting radioresponse modulators within foods and generating recipes to enhance RT response in RC patients. For the identification of radioresponse modulators within foods, we adopted a genomic-driven approach, hypothesising that these modulators should exhibit similar effects on protein networks as those observed in RT-resistant RC patients. To model the genomic effects of food molecules and phenotype, we integrated metric learning with biased RWR, mapping the influence of food molecules and the phenotype across a multiscale interactome. This process illuminated the proteins and biological functions most impacted. Overall, this study establishes a foundation for discussing methodological strategies aimed at modulating radioresponse through bioactive molecules in foods. We view this as a pioneering step in creating genome-guided, phytochemically-enriched recipes to enhance RT outcomes in RC patients and see its potential as a reference point for subsequent research.

Methods

Identifying radioresponse modulators.

We propose the approach outlined in Fig. 1A for the identification of radioresponse modulators within foods. The core of our model is a graph $G = (\mathcal{V}, E)$ representing the multiscale interactome described by[23], where nodes are proteins and biological functions, and edges represent protein-protein, protein-biological functions, and biological function-biological function interactions. Protein-protein interactions describe physical interactions between proteins. Protein-biological function interactions connect proteins to the biological functions they affect and biological function-biological function interactions represent the hierarchy of biological functions using the Gene Ontology's Biological Processes[32]. For more details on the construction of the multiscale interactome, we refer the reader to[23]. Specifically, our graph G has $|\mathcal{V}| = N + M = 27,458$ nodes of which $N = 17,660$ are proteins and $M = 9798$ are biological functions. The phenotype, i.e., the over-expressed proteins in patients exhibiting RT resistance, is modeled as an N-dimensional vector $\mathbf{p} \in \{0, 1\}^N$ where $p_i = 1$ if gene i is over-expressed and 0 otherwise. Similarly, protein targets of food molecules are represented as N-dimensional vectors $\mathbf{m}^j \in \{0, 1\}^N$ where $m_i^j = 1$ if protein i is targeted by food molecule j and 0 otherwise. Information of 2100 food molecules and their targets are obtained from FoodDB[33] and STITCH[34] datasets. Using the multiscale interactome allows us to explain identified molecules, even when they seem unrelated to the phenotype. It additionally allows us to identify which biological functions are being modulated in cases where a short protein-protein path exists between food molecule targeted proteins and RC-resistant over-expressed genes, adding a level of interpretability.

Network propagation algorithm and metric learning. We combine a network propagation algorithm based on biased random walks with restarts with deep metric learning. The network propagation algorithm starts from initial nodes encoded in binary vectors encoding food molecules and the phenotype. At every step, the walker can restart its walk or jump to an adjacent node. The outputted diffusion profile measures how often each node in the multiscale interactome is visited by the RWR, encoding the effect of food molecules and the phenotype on every protein and biological function. In[23], they optimise the edge weights of the algorithm for a multiscale-based drug-disease prediction task, in which an AUROC = 0.705 was achieved. The task involves predicting whether a drug treats a disease based on known drug-disease pairs taken from the Drug Repurposing Database, the Drug Repurposing Hub and the Drug Indication Database with only FDA-approved treatment relationships. Given that optimising the edge weights of the random walk algorithm has a very small effect on the prediction task (a fixed random walk probability of $\alpha = 0.64$ and edge-weights all being 1 gives 0.702 AUROC), we propose to fix the edge weights and optimise the weights of a multilayer perceptron (MLP) instead, using deep metric learning in order to minimise the distance between known drug-disease pair embeddings and maximise the distance to unknown drug-disease pairs. We set the propagation value of the RWR to 10 times the mean maximum propagated value over all drugs after propagating with $\alpha = 0$, giving a value of $\alpha = 0.64$. For each disease, we randomly sample both a positive drug (a drug which is known to be beneficial against the disease) and a negative drug (a drug which has no known benefit). This triple (disease, positive drug, negative drug) is passed to a MLP in order to get an embedding for the disease and the two drugs. A triplet loss is then used in order to minimise the distance between the disease and positive drug and maximise the distance to the negative drug. We use 5-fold cross-validation to optimize the model in the set of drugs and diseases (N = 1651), and use the trained model to give a ranking of food molecules based on distance to the phenotype embedding. In each split, we train the model for a maximum of 100 epochs using the Adam optimizer. Final propagation profiles reflect protein and biological functions affected. However, the model alone is not sufficient to filter out toxic molecules or metals from the food molecule database. Additionally, it is difficult for the model to learn whether the molecules affect the biological functions disrupted by the phenotype rather than directly targeting disease proteins or their regulators.

Filtering predictions. Using propagated profiles or the entity embeddings, we find the top 100 food molecules closest to the phenotype. These molecules affect similar proteins and biological functions as those responsible for radioresistance. Experimental evidence indicates that the phenotype-associated genes are over-expressed in patients exhibiting RT resistance leading to a positive modulation of lipid metabolism. Therefore, we use domain knowledge and literature search to filter out identified molecules with positive regulatory effects on lipid metabolism, leaving 33 modulators to retrieve the list of ingredients (Appendix A). Modulators belong to a myriad of compound classes including flavonoids, isoflavonoids, and bezenoids, in alignment with the current knowledge on chemotherapeutic bioactive molecules within foods[11]. Overall, predicted modulators are involved in cell signaling, cell growth and lipid metabolism. For example, Genistein works by inhibiting the Arachidonic Acid pathway, making it a suitable natural agent for cancer prevention and therapy[11].

Recipe optimisation module. Having found radioresponse modulators in the previous step, we propose to provide patients with recipes that maximize the number of ingredients with these molecules (Fig. 1C). Associations between foods and the molecules they contain are taken from FoodDB[35], and a baseline set of recipes from the Recipe1M dataset[36]. Ingredients from these two datasets were preprocessed (turned to lowercase, spaces and plurals removed) and matched if they shared the first or last two words, or if they had the same word in the first or in the last position. This meant that ingredients such as king oyster mushroom and dried porcini mushroom were treated as being the same ingredient.

After combining these datasets, an enrichment score is calculated for each recipe based on the number of radioresponse modulators that they contain. Additionally, ingredient context embeddings from the BERT model[37] are used to optimize the recipes and provide recommended ingredient substitutions to patients. These substitutions are done to increase the amount of anti-RT-resistance molecules whilst also preserving the recipe's culinary attributes. Ingredient substitutions for the Recipe1M dataset were then found using the same method outlined in[38]. Starting with the bert-base-cased model in the Hugging Face library[39], the BERT vocabulary was extended to include all the ingredients in the dataset. The BERT model, with a hidden representation of dimension 768, was then trained on the cooking instructions for each recipe in the dataset. Given that BERT gives different embeddings for the same ingredient in different contexts, there ends up being approximately 285,000 embeddings for all ingredients. For all the embeddings of a single ingredient, the 200 nearest neighbors were found using KNN and a substitute score given to other ingredients based on how often it appeared in the 200 nearest neighbors for all the embeddings. Suggested substitutes were then found for an ingredient by finding ingredients which had a score of over 100 and which were greater than 1/10 of the highest score for that ingredient.

To visualize the embedding space, we averaged all the embeddings for the same ingredient in order get a single embedding of dimension 768 for each ingredient. A 2D projection of this space using Principal Component Analysis is shown for a few of the ingredients in Fig. 4A. The suggested ingredient substitutes for a particular ingredient were then filtered to only include ingredients that had a higher number of molecules with potential for RT modulation than the initial ingredient. Some examples of these substitutions are shown in Fig. 4B. The number of beneficial molecules for each ingredient was found using the FoodDB database and is shown in Appendix A. Recipes in the Recipe1M dataset were optimized by looping through the ingredients and randomly selecting a substitute within the filtered list of substitutes. Additionally, it was constrained such that the same substitute can not be made for different ingredients within the recipe and a substitute suggestion which is already in the recipe is not allowed. Some examples of a mutated recipe are shown in Fig. 4C.

Dietary recommendations. When recommending recipes to a patient, it is also important to take into account other factors such as allergies, food preferences and general nutritional guidance. The flexibility of our approach and scoring function makes this possible. We showcase this by further optimising our recipes to take into account allergies and food preferences. Additional input is given to the model in the form of a list of user allergies and a dictionary of user food preferences. The allergy list contains which of the 14 main food allergens the user has and the food preference dictionary has keys corresponding to ingredients and values being a score of 1–5 indicating the patient's like of the food (1 indicating a strong dislike and 5 a strong like). In order to take into account this information, we create a database containing all the unique ingredients and whether they satisfy each of the 14 allergies. Given an allergy list input, we loop through all recipes and make an ingredient substitution for all ingredients where the patient is allergic. If there doesn't exist a substitution or all substituted ingredients also cause allergies then the recipe is removed. We then optimise these new recipes as before to take into account both the number of radioresponse modulators and also the patient's food preferences. This is done by making ingredient substitutions in a recipe if either the patient prefers the new ingredient or if there is an increase in the number of radioresponse modulators whilst also enforcing that there is not a reduction in the other.

Data availability

All data used in the paper is publicly available. Genome data can be collected from STRING[40] (https://string-db.org), UniProt[41] (https://www.uniprot.org), COSMIC[42] (https://cancer.sanger.ac.uk/cosmic), and NCBI Gene[43] (https://www.ncbi.nlm.nih.gov/gene/). Drug data can be extracted from DrugBank[44] (https://www.drugbank.ca), DrugCentral[45] (http://drugcentral.org), and STITCH[46] (http://stitch.embl.de). Food data can be extracted from FooDB[47] (https://foodb.ca) and STITCH[46] (http://stitch.embl.de). The recipes can be obtained from Recipe1M[36] (http://pic2recipe.csail.mit.edu/) and the Multiscale Interactome data and analysis from (github.com/snap-stanford/multiscale-interactome)[23].

Received: 7 December 2022; Accepted: 31 August 2023
Published online: 08 September 2023

References

1. Heald, R., Husband, E. & Ryall, R. The mesorectum in rectal cancer surgery-the clue to pelvic recurrence?. *Br. J. Surg.* **69**, 613–616. https://doi.org/10.1002/BJS.1800691019 (1982).
2. Kreis, M. E. *et al.* Use of preoperative magnetic resonance imaging to select patients with rectal cancer for neoadjuvant chemoradiation-interim analysis of the German OCUM Trial (NCT01325649). *J. Gastrointest. Surg.* **20**, 25–33. https://doi.org/10.1007/S11605-015-3011-0 (2015).
3. Sebag-Montefiore, D. *et al.* Preoperative radiotherapy versus selective postoperative chemoradiotherapy in patients with rectal cancer (MRC CR07 and NCIC-CTG C016): A multicentre, randomised trial. *The Lancet* **373**, 811–820. https://doi.org/10.1016/S0140-6736(09)60484-0 (2009).

4. Erlandsson, J. *et al.* Optimal fractionation of preoperative radiotherapy and timing to surgery for rectal cancer (Stockholm III): A multicentre, randomised, non-blinded, phase 3, non-inferiority trial. *Lancet Oncol.* **18**, 336–346. https://doi.org/10.1016/S1470-2045(17)30086-4 (2017).
5. Gijn, W. V. *et al.* Preoperative radiotherapy combined with total mesorectal excision for resectable rectal cancer: 12-year follow-up of the multicentre, randomised controlled TME trial. *Lancet Oncol.* **12**, 575–582. https://doi.org/10.1016/S1470-2045(11)70097-3 (2011).
6. Poynter, L. *et al.* Network mapping of molecular biomarkers influencing radiation response in rectal cancer. *Clin. Colorectal Cancer* **18**, e210–e222. https://doi.org/10.1016/J.CLCC.2019.01.004 (2019).
7. Buckley, A. M., Lynam-Lennon, N., O'Neill, H. & O'Sullivan, J. Targeting hallmarks of cancer to enhance radiosensitivity in gastrointestinal cancers. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 298–313. https://doi.org/10.1038/s41575-019-0247-2 (2020).
8. Gavrilas, L. I. *et al.* Plant-derived bioactive compounds in colorectal cancer: Insights from combined regimens with conventional chemotherapy to overcome drug-resistance. *Biomedicines* **10**, 85 (2022).
9. Mahmod, A. I., Haif, S. K., Kamal, A., Al-Ataby, I. A. & Talib, W. H. Chemoprevention effect of the Mediterranean diet on colorectal cancer: Current studies and future prospects. *Front. Nutr.* **9**, 924192 (2022).
10. Stelzer, G. *et al.* The GeneCards suite: From gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinform.* **54**, 1–1. https://doi.org/10.1002/CPBI.5 (2016).
11. Yarla, N. S. *et al.* Targeting arachidonic acid pathway by natural products for cancer prevention and therapy. *Semin. Cancer Biol.* **40–41**, 48–81. https://doi.org/10.1016/J.SEMCANCER.2016.02.001 (2016).
12. Hama, K. *et al.* Very long-chain fatty acids are accumulated in triacylglycerol and nonesterified forms in colorectal cancer tissues. *Sci. Rep.* **11**, 1–10. https://doi.org/10.1038/s41598-021-85603-w (2021).
13. Tang, X. & Brindley, D. N. Lipid phosphate phosphatases and cancer. *Biomolecules* **10**, 1–24. https://doi.org/10.3390/BIOM10091263 (2020).
14. Chang, W., Fa, H., Xiao, D. & Wang, J. Targeting phosphatidylserine for cancer therapy: Prospects and challenges. *Theranostics* **10**, 9214. https://doi.org/10.7150/THNO.45125 (2020).
15. Chen, C.-Y., Chen, J., He, L. & Stiles, B. L. PTEN: Tumor suppressor and metabolic regulator. *Front. Endocrinol.* **0**, 338. https://doi.org/10.3389/FENDO.2018.00338 (2018).
16. Kim, Y. S., Young, M. R., Bobe, G., Colburn, N. H. & Milner, J. A. Bioactive food components, inflammatory targets, and cancer prevention. *Cancer Prev. Res.* **2**, 200–208. https://doi.org/10.1158/1940-6207.CAPR-08-0141 (2009).
17. Pan, M.-H., Lai, C.-S., Dushenkov, S. & Ho, C.-T. Modulation of inflammatory genes by natural dietary bioactive compounds. *J. Agric. Food Chem.* **57**, 4467–4477. https://doi.org/10.1021/JF900612N (2009).
18. Samadi, A. K. *et al.* A multi-targeted approach to suppress tumor-promoting inflammation. *Semin. Cancer Biol.* **35**, S151–S184. https://doi.org/10.1016/J.SEMCANCER.2015.03.006 (2015).
19. Nencioni, A., Caffa, I., Cortellino, S. & Longo, V. D. Fasting and cancer: Molecular mechanisms and clinical application. *Nat. Rev. Cancer* **18**, 707–719 (2018).
20. Veselkov, K. *et al.* HyperFoods: Machine intelligent mapping of cancer-beating molecules in foods. *Sci. Rep.* **9**, 9237. https://doi.org/10.1038/s41598-019-45349-y (2019).
21. Gonzalez, G., Gong, S., Laponogov, I., Bronstein, M. & Veselkov, K. Predicting anticancer hyperfoods with graph convolutional networks. *Hum. Genom.* **15**, 33. https://doi.org/10.1186/s40246-021-00333-4 (2021).
22. Laponogov, I. *et al.* Network machine learning maps phytochemically rich Hyperfoods to fight COVID-19. *Hum. Genom.* **15**, 1. https://doi.org/10.1186/s40246-020-00297-x (2021).
23. Ruiz, C., Zitnik, M. & Leskovec, J. Identification of disease treatment mechanisms through the multiscale interactome. *Nat. Commun.* **12**, 1–15. https://doi.org/10.1038/s41467-021-21770-8 (2021).
24. Pei, J., Velu, P., Zareian, M., Feng, Z. & Vijayalakshmi, A. Effects of syringic acid on apoptosis, inflammation, and akt/mtor signaling pathway in gastric cancer cells. *Front. Nutr.* **8**, 1109. https://doi.org/10.3389/fnut.2021.788929 (2021).
25. Riahi-Chebbi, I. *et al.* The Phenolic compound Kaempferol overcomes 5-fluorouracil resistance in human resistant LS174 colon cancer cells. *Sci. Rep.* **9**, 195. https://doi.org/10.1038/s41598-018-36808-z (2019).
26. Wenzel, U., Nickel, A. & Daniel, H. alpha-lipoic acid induces apoptosis in human colon cancer cells by increasing mitochondrial respiration with a concomitant o2-*-generation. *Apoptosis Int. J. Program. Cell Death* **10**, 359–68. https://doi.org/10.1007/s10495-005-0810-x (2005).
27. Afshin, A. *et al.* Health effects of dietary risks in 195 countries, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *The Lancet* **393**, 1958–1972. https://doi.org/10.1016/S0140-6736(19)30041-8 (2019).
28. Gonzalez, G., Gong, S., Laponogov, I., Bronstein, M. & Veselkov, K. Predicting anticancer hyperfoods with graph convolutional networks. *Hum. Genom.* **15**, 741. https://doi.org/10.1186/s40246-021-00333-4 (2021).
29. Mittelman, S. D. The role of diet in cancer prevention and chemotherapy efficacy. *Annu. Rev. Nutr.* **40**, 273–297 (2020).
30. Laponogov, I. *et al.* Network machine learning maps phytochemically rich hyperfoods to fight COVID-19. *Hum. Genom.* **15**, 741. https://doi.org/10.1186/s40246-020-00297-x (2021).
31. Cory, H., Passarelli, S., Szeto, J., Tamez, M. & Mattei, J. The role of polyphenols in human health and food systems: A mini-review. *Front. Nutr.* **5**, 753. https://doi.org/10.3389/fnut.2018.00087 (2018).
32. The Gene Ontology Consortium. The gene Ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.***47**, D330–D338. https://doi.org/10.1093/NAR/GKY1055 (2019).
33. Wishart, D. S. *et al.* DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.***46**, D1074–D1082. https://doi.org/10.1093/nar/gkx1037 (2018).
34. Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J. & Bork, P. STITCH: Interaction networks of chemicals and proteins. *Nucleic Acids Res.* **36**, 684–8. https://doi.org/10.1093/nar/gkm795 (2008).
35. Harrington, R. A., Adhikari, V., Rayner, M. & Scarborough, P. Nutrient composition databases in the age of big data: FoodDB, a comprehensive, real-time database infrastructure. *BMJ Open* **9**, 1–10. https://doi.org/10.1136/bmjopen-2018-026652 (2019).
36. Marin, J. *et al.* Recipe1M+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 187–203. https://doi.org/10.1109/TPAMI.2019.2927476 (2021).
37. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019– 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies—Proceedings of the Conference, vol. 1* 4171–4186. https://doi.org/10.18653/v1/N19-1423 (2019).
38. Pellegrini, C., Özsoy, E., Wintergerst, M. & Groh, G. Exploiting food embeddings for ingredient substitution. In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies*. https://doi.org/10.5220/0010202000670077 (Science and Technology Publications, 2021).
39. Wolf, T. *et al.* HuggingFace's Transformers: State-of-the-art Natural Language Processing. *CoRR***abs/1910.0**, https://doi.org/10.18653/v1/2020.emnlp-demos.6 (2019).
40. Szklarczyk, D. *et al.* The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612. https://doi.org/10.1093/nar/gkab835 (2021).
41. The UniProt Consortium. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.***45**, D158–D169. https://doi.org/10.1093/nar/gkw1099 (2016).

42. Tate, J. G. *et al.* COSMIC: The catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**, D941–D947. https://doi.org/10.1093/nar/gky1015 (2019).
43. Brown, G. R. *et al.* Gene: A gene-centered information resource at NCBI. *Nucleic Acids Res.* **43**, D36–D42. https://doi.org/10.1093/nar/gku1055 (2014).
44. Wishart, D. S. *et al.* DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082. https://doi.org/10.1093/nar/gkx1037 (2017).
45. Ursu, O. *et al.* DrugCentral: Online drug compendium. *Nucleic Acids Res.* **45**, D932–D939. https://doi.org/10.1093/nar/gkw993 (2016).
46. Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J. & Bork, P. STITCH: Interaction networks of chemicals and proteins. *Nucleic Acids Res.* **36**, D684–D688. https://doi.org/10.1093/nar/gkm795 (2007).
47. Wishart Research Group. FooDB. http://foodb.ca (2022).

Acknowledgements

J.S. is supported by the UKRI CDT in AI for Healthcare http://ai4health.io (Grant No. P/S023283/1) and by the Vodafone Foundation through the DreamLab/DRUGS and CORONA-AI projects. K.V, D.V, I.L., and G.G were supported by the Vodafone Foundation as part of the ongoing DreamLab/DRUGS and CORONA-AI projects and the ERC Proof of Concept Grant No. 899932 (Hyperfoods). M.B, D.V., and G.G were supported by the ERC-Consolidator Grant No. 724228 (LEMAN). The work was additionally funded by the UK Research and Innovation (Grant No. 10058099), and the European Union (Grant No. 101095359), as part of the AIDA project.

Author contributions

K.V. and M.B. designed the concept and supervised the study. J.S and G.G. developed the methodology, implemented the computational workflow. J.S, G.G., R.M., P.B., L.P., I.L, Y.Z. aggregated the data sets. L.P. did the gene expression studies and provided the list of DE genes; K.V., M.B., I.L., R.M., D.V. designed research and helped with idea creation, Y.Z. benchmarked the models, P.B. helped with filtering radioresponse modulators. All authors contributed to writing the manuscript and results interpretation. The authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-41833-8.

Correspondence and requests for materials should be addressed to K.V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.