AWARD NUMBER: W81XWH-15-1-0423

TITLE: Novel High-Fidelity Screening of Environmental Chemicals and Carcinogens and Mechanisms in Colorectal Cancer

PRINCIPAL INVESTIGATOR: Dr. Sivanesan Dakshanamurthy

CONTRACTING ORGANIZATION: Georgetown University
Washington, DC 20057

REPORT DATE: September 2016

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*



| **1. REPORT DATE** | **2. REPORT TYPE** | **3. DATES COVERED** |
|---|---|---|
| September 2016 | Annual Report (First Year) | 1 Sep 2015 - 31 Aug 2016 |

**4. TITLE AND SUBTITLE**

Novel High-Fidelity Screening of Environmental Chemicals and Carcinogens and Mechanisms in Colorectal Cancer.

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**
W81XWH-15-1-0423

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**
**Sivanesan Dakshanamurthy**, Henri Wathieu, Stephen Byers, Abiola Ojo.

**E-Mail**: sd233@georgetown.edu

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
GEORGETOWN UNIVERSITY (THE)37TH & O STS NWWASHINGTON DC 20057-0001

**8. PERFORMING ORGANIZATION REPORT**

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

**10. SPONSOR/MONITOR'S ACRONYM(S)**

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION / AVAILABILITY STATEMENT**

Approved for Public Release; Distribution Unlimited

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**

Environmental chemicals (ECs) include chemical warfare agents and carcinogens confer potential cancer risks to Military personnel and their families during service. The relative contribution of EC exposure and genetic susceptibility in the etiology of cancer is poorly understood making the translation of existing data into meaningful prevention and/or therapeutic strategies for Military personnel difficult. To that end, we develop a platform for the predictive biological assessment of ECs through two foci: (1) predicting empirical EC-protein target associations by a proteochemometric method called Tox-TMFS and incorporating systems biology analysis to model the cancer-linked cellular activity of the EC using Net-TMFS, now called as DrugGenEx-Net, and (2) a novel method entitled the "chemo-phenotypic based toxicity measurement" (CPTM) that integrates EC-target biological effects with ADME toxicokinetic parameters and intrinsic chemical reactivity properties into a quantifiable "toxicity score" (Zts). We hypothesize that our novel *in-silico* screening method will identify mechanisms through finding targets/pathways with unprecedented accuracy and identify which ECs have the potential to influence the development of cancer. As biological effects are driven by EC interactions with biological entities such as proteins, we completed a computational systems biology model called Tox-TMFS that predicts EC-protein target signatures and relates them to higher-order effects that include protein-protein interactions, signaling pathways, and molecular functions using DrugGenEx-Net. We have carried out and validated Tox-TMFS and DrugGenEx-Net procedures by querying predicted EC-target and EC-molecular function signatures against external databases. We further investigated cancers that have been causally linked to perturbations of particular gene products/pathways, such as VDR signaling in Colorectal Cancer (CRC), and proposed 20 biological testing candidate ECs associated to CRC and its associated protein targets and pathways. This biological testing is currently underway. Lastly, we have applied the CPTM method on ECs to quantify the "toxicity potential" and validate the method by correlating $Z_{ts}$ scores with documented toxicity effects from external databases. Biological assessments of EC toxicity for 20 toxicants are in the preliminary phases. This work is the first platform of its kind in the toxicological sciences rooted in chemistry and systems toxicology.

**15. SUBJECT TERMS**

Toxicity, Environmental Chemicals, TOX-TMFS, CPTM, Cancer Cellular Network Model, Chemical Reactivity, Chemical Promiscuity, Pharmacokinetics, Colorectal Cancer, N,N'-disalicylidene-1,2-diaminopropane, Pyraclostrobin, Paclobutrazol, Vitamin D Receptor, Wnt/Beta Catenin Signaling, Transforming growth factor beta.

| **16. SECURITY CLASSIFICATION OF:** U | | | **17. LIMITATION OF ABSTRACT** | **18. NUMBER OF PAGES** | **19a. NAME OF RESPONSIBLE PERSON** USAMRMC |
|---|---|---|---|---|---|
| **a. REPORT** | **b. ABSTRACT** | **c. THIS PAGE** | Unclassified | 119 | **19b. TELEPHONE NUMBER** *(include area code)* |
| Unclassified | Unclassified | Unclassified | | | |

## Table of Contents

# 1. Introduction……..

Environmental chemical (EC) exposure due to military occupation, travel, industrial, social and general living situations is ubiquitous. Industrial and technological progress has significantly increased the amount of ECs and human interactions with them. Biological characterization of these chemicals is challenging and inefficient, even with available high-throughput technologies. We are developing an in-silico method for characterizing relative toxicity called the Chemo-Phenotypic Based Toxicity Measurement (CPTM). This method provides the first comprehensive integration of EC biological effects, chemical reactivity features, and pharmacokinetic properties. As biological effects are driven by EC interactions with biological entities such as proteins, we completed a computational systems biology model called Tox-TMFS that predicts EC-protein target signatures and relates them to higher-order effects that include protein-protein interactions, signaling pathways, and molecular functions. For proof-of-concept, we are in the process of simulating ECs using CPTM. Validations of EC-biological effect signatures are currently being performed at all levels. Focused assessments are in progress for EC associations with vitamin D receptor signaling pathway and colorectal cancer. CPTM was used to quantify EC "toxicity score" (Zts), which serves as a holistic metric of potential toxicity by which ECs may be ranked relative to each other. Assessment of general toxicity for a panel of 20 ECs is currently underway. CPTM with integrated TOX-TMFS and DRUGGENEX-NET is, to our knowledge, the first comprehensive systems biology- and chemistry-based platform for efficient high-throughput study of EC toxicity. As the EC space grows exponentially with new commercial materials, environmental waste products, and pharmaceuticals, CPTM platform is positioned to streamline the comprehensive assessment of these chemicals for focused subsequent assays and increased efficiency in toxicology.

## 2. Keywords……

Toxicity, Environmental Chemicals, TOX-TMFS, CPTM, Cancer Cellular Network Model, Chemical Reactivity, Chemical Promiscuity, Pharmacokinetics, Colorectal Cancer, N,N'-disalicylidene-1,2-diaminopropane, Pyraclostrobin, Paclobutrazol, Vitamin D Receptor, Wnt/Beta Catenin Signaling, Transforming growth factor beta.

# 3. Accomplishments…….

## A. Summary of Accomplishments

**What were the major goals of the project?**

**Major Task 1. Screening of Potential EC-Protein Interactions Using the Tox-TMFS Method**.

*Subtask 1 included the in-silico screening of ECs against human protein targets using the Tox-TMFS method, and this task is 75% complete. Subtask 2 entails the computation of intrinsic chemical properties for those chemicals being assessed for protein binding, and the incorporation of these properties in calculating the Tox-TMFS-yielded protein binding Z-scores, and was completed in January 2016. In Subtask 3, each EC would have putative protein targets ranked 1-40 based on that Z-score, with an assessment of the promiscuity of each EC, completed in January 2016. The milestone for this Major Task was to obtain top 40 protein targets for each of ECs, and this milestone is 80% completed*.

**Major Task 2. Development of a Cancer Cellular Network Model Using the DrugGenEx-Net Method**.

*Subtask 1, under this major task involved the linking of EC-protein interactions to Protein-protein interactions, molecular functions, and pathways in the cell, constructing EC perturbation networks assessed by way of the OMIM Morbid Map database to create Chemical-Disease interaction networks, a subtask completed in March 2016. Subtask 2 entailed a relationship analysis between the cellular networks for ECs as elucidated in subtask 1, completed in March 2016. Subtask 3 was a CRC-specific analysis of EC Cancer Cellular Network Models, 75% complete. Milestone 1 under this Major task involved the development of biological perturbation networks for each EC in the context of CRC, which was completed in June 2016. Milestone 2 pinpointed key ECs to be assessed biologically for perturbation activity in CRC-associated pathways VDR, TGF-beta and Wnt/Beta-catenin signaling pathways, completed in June 2016*.

**Major Task 3. Prediction of EC Toxicity Using the CPTM Model**.

*Subtask 1 incorporates the Intrinsic properties of environmental chemicals computed in Major Task 1 and requires assessment of additional kinetic properties for the ECs of this work, 60% complete. Subtask 2 requires development of the CPTM model and computation of general toxicity using kinetic and intrinsic EC parameters, 60% complete. The milestone for this major task entails the computation of toxicity scores for all ECs assessed, and identification of key ECs for biological testing*.

**Major Task 4. Biological testing of candidate Environmental Chemicals**.

*Subtask 1 for biological testing entails CRC-associated target binding studies for a small number of ECs, 25% completed. Subtask 2 is to perform Wnt/Beta-catenin activation studies using reporter assays and mammalian two hybrid assays for VDR/beta-catenin interactions for the previously identified CRC-linked ECs of interest, which is 15% complete. Subtask 3 was to be performed cell viability and apoptosis measurements for CPTM model validation using cells from relevant tissue sources, 15% completed. The milestone for this major task is the complete biological testing of 40 ECs, which is 15% complete*.

**What was accomplished under these goals?**

**Major Task 1. Screening of Potential EC-Protein Interactions Using the Tox-TMFS Method**.

*We have screened 420 ECs following the development of Tox-TMFS, against 2,335 protein targets, exceeding the 254 structures anticipated. These 420 ECs were assessed for physiochemical descriptors and the Tox-TMFS procedure was completed, yielding top 40 predicted protein interactions for each of the assessed ECs.*

**Major Task 2. Development of a Cancer Cellular Network Model Using the DrugGenEx-Net Method**.

*By way of the* DrugGenEx-Net *method, protein binders for the characterized ECs were combined with known protein interactions to create protein signatures for each EC, and each protein was annotated to its Protein-protein interactions, pathways, and molecular functions in the cell. These cellular network models were further enriched with oncologic disease OMIM profiles to create cancer-specific networks. The ECs N,N'-disalicylidene-1,2-diaminopropane, paclobutrazol, and pyraclostrobin were selected for biological testing of CRC-linked pathway perturbation based on their multiscale activity profiles*.

**Major Task 3. Prediction of EC Toxicity Using the CPTM Model**.

*Integrating biological interactions, physicochemical descriptors, and pharmacokinetic properties, a toxicity risk quantification score (Zts) was calculated for each of the TMFS-assessed ECs under the CPTM model. These predicted toxicity potentials were validated using known EC toxic assay results*.

**Major Task 4. Biological testing of candidate Environmental Chemicals**.

*Reporter assays for VDR binding studies are under way for N,N'-disalicylidene-1,2-diaminopropane and in the preparatory phases for other ECs predicted to bind to VDR. Wnt/beta-catenin pathways reporter assay activation studies and VDR/beta-catenin distruption by mammalian two hybrid assays have been initiated for the selected ECs N,N'-disalicylidene-1,2-diaminopropane, paclobutrazol, and pyraclostrobin. Overall toxicity assessments will be further performed for 20 selected ECs of various predicted toxic potential*.

**What opportunities for training and professional development has the project provided?**

Development of the CPTM procedure provided valuable training opportunities for a fourth year pharmacology student seeking to acquire proficiency in the computational assessment of biopharmacological and toxicological properties. A bioinformatics research assistant, and a prospective medical student has been trained to perform systems toxicology and other in-silico assessments under the computation of intrinsic and kinetic properties of ECs by Tox-TMFS, and

to manage datasets in building cancer cellular networks for ECs by the DrugGenEx-Net technique. Lastly, MSD-MAP, a metabolomic cancer-association assessment platform tangential to this toxicological project, was developed in conjunction with a student completing a Masters in biomedical sciences seeking exposure to computational biology.

**How were the results disseminated to communities of interest?**

Tangential projects developed using the principles central to Tox-TMFS and CPTM, as described in the Statement of Work, included a metabolite-disease association platform entitled MSD-MAP, a network-based drug repurporsing platform entitled DrugGenEx-Net, and a drug repurposing-oriented protein interaction prediction model called RepurposeVS. The procedural techniques we developed were utilized in increasing the public understanding of systems biology based chemical-phenotypic association assessments by way of a polypharmacology-centered review article called "Harnessing Polypharmacology with Computer-Aided Drug Design and Systems Biology."

The cancer cellular network models built in this work will be disseminated in our online database and visualization tool called the Chemical Interactome Cellular Network Interface (CICNI), a platform that can be used to build systems-based chemical phenotypic prediction models similar to CPTM, with applications in the prevention, causative mechanistic understanding, and treatment of disease.

**What do you plan to do during the next reporting period to accomplish the goals?**

The next reporting period will entail a completion of Major Task 1 subtask 1, wherein ECs for which we have information on chemical structure will be processed for computation of intrinsic and kinetic characteristics, with assessment of protein binding by the Tox-TMFS procedure. This will equally accomplish the milestone for Major Task 1, and all characterized chemicals will have full EC-protein association networks. Upon this characterization, cancer cellular network models will be annotated for these ECs, and incorporated into the assessments for biological testing for mechanistic perturbation of

VDR, TGF beta, and Wnt/Beta-catenin signaling pathways. Completion of these chemical parameters will lastly allow for the conclusion of our CPTM procedure, and we will have obtained ranked potential toxicity prediction scores for the ECs. In this process, ECs of interest will emerge for which we will assess cell viability and apoptosis measurements according to subtask 3 in Major Task 4. As was described in our accomplishments, biological testing for candidate ECs at the levels of protein binding studies using reporter and surface plasmon resonance assays, wnt/beta-catenin pathway activation studies, VDR/beta-catenin interaction disruption studies, and CPTM-linked toxicity assays are in the preliminary stages, and this full panel of toxicology testing will be carried out in the upcoming months.

In accordance with Major Task 1, we will continue the physicochemical assessment and building protein binding signatures for ECs using Tox-TMFS. Using the CPTM model, the general toxicity for these ECs will equally be assessed from kinetic and intrinsic properties of these compounds. The primary task from this point is to continue our biological testing at the levels of protein binding, cancer-associated pathway perturbation assessments, and in vitro assays for measures of general toxicity. These biological assessments will be utilized to validate our phenotypic predictions, which we arrive to solely from the starting point of chemical structure, and will also provide novel findings of biological perturbation and toxicity activity of key ECs, having implications for the ways in which EC exposure toxicity is measured and preemptively utilized to make policy and other decisions.

## B. Detailed Accomplishments

### Specific Aim 1. Toxicity Screens

**Major Task 1**. Screening of Potential EC-Protein Interactions Using the Tox-TMFS Method

Subtask 1. In-silico screen of ECs against human protein targets using the Tox-TMFS method.

Subtask 1 was listed to be completed in 1-8 months. This task is 75% complete.

Subtask 2. Computation of TMFS method descriptors as described previously. Generation of Tox-TMFS Z-score for each ECs.

Subtask 2 was listed to be completed in 2-8 months. This task was completed in 4 months, in January 2016.

Subtask 3. For each ECs rank order the protein target hits as top 1-40 based on the Tox-TMFS Z-score. Predict potential EC-protein (ECP) interactions called "Tox" signatures. Analyze the predicted top 40 ECP associations in terms of chemical promiscuity for each target i.e. chemicals that interact with multiple targets with potentially bad effects.

Subtask 3 was listed to be completed in 3-8 months. This task was completed in 4 months, in January 2016.

Milestone(s) Achieved. Using these computational biology, and data analysis methods, potential EC-cancer proteins associations (top 1- 40) will be achieved.

Major Task 1. Milestone Achieved was listed to be completed in 1-8 months. This milestone is 90% complete.

### Accomplishments

**Major Task 1**. Screening of Potential EC-Protein Interactions Using the Tox-TMFS Method.

In following the stated subtask 1, we applied our TMFS method **[1]** for the prediction of binding signatures for the environmental chemicals (ECs) against 2,335 protein targets **(Figure 1)**. In short, ECs were docked into target pockets identified by their reference ligand positions using 20 angstrom grids in GLIDE **[2]**. ECs retained after docking were then subjected to Schrodinger's QikProp **[2]** module to generate the following physicochemical descriptors as required by subtask 2: (1) # H-bond acceptors, (2) # H-bond donors, (3) dipole moment, (4) electron affinity, (5) globularity, (6) ionization potential, (7) molecular weight, (8) # rotatable bonds, (9) solvent-accessible surface area (SASA), and (10) volume. These physicochemical properties were compared to those of the reference ligand using a Tanimoto similarity score. EC and shapes were computed using a spherical harmonics expansion approach and compared to that of the reference ligand using a Euclidean distance metric. EC shapes were also compared to that of the protein target pocket. The process of computing and comparing shapes previously reported **[3]**. Docking scores and shape metrics were normalized to a score range between 0 and 1 to be implemented with Tanimoto similarity scores into a comprehensive "Z-score" that represents the quantitative likelihood of binding for an EC as described in **[1]**:

$$Z = w_k Y(s_p, s_l) + \sum_{i=1}^{1} [w_i f(s_p, s_l) + w_{i+1} f(s_p, s_l)] + \sum_{n=1}^{j} X_n(s_c, s_l) + CS(OLIC) \quad \textbf{[Equation 1]}$$

In accordance with subtask 3, EC-protein (ECP) interactions called "Tox" signatures were ranked by descending Z-score, and the top 40 protein targets for each EC were further narrowed to Z-scores greater than the value of 13.5 out of 18 (75%) considered likely to bind to the respective protein target. Daidzein, an EC which we investigate as biologically associated with CRC in our analyses of cancer cellular network models (**Major Task 2**) and promiscuity (**Major Task 3**), was found to have 25 novel protein interactions after implementation of the Z-score cutoff of 13.5 **(Table 1)**.

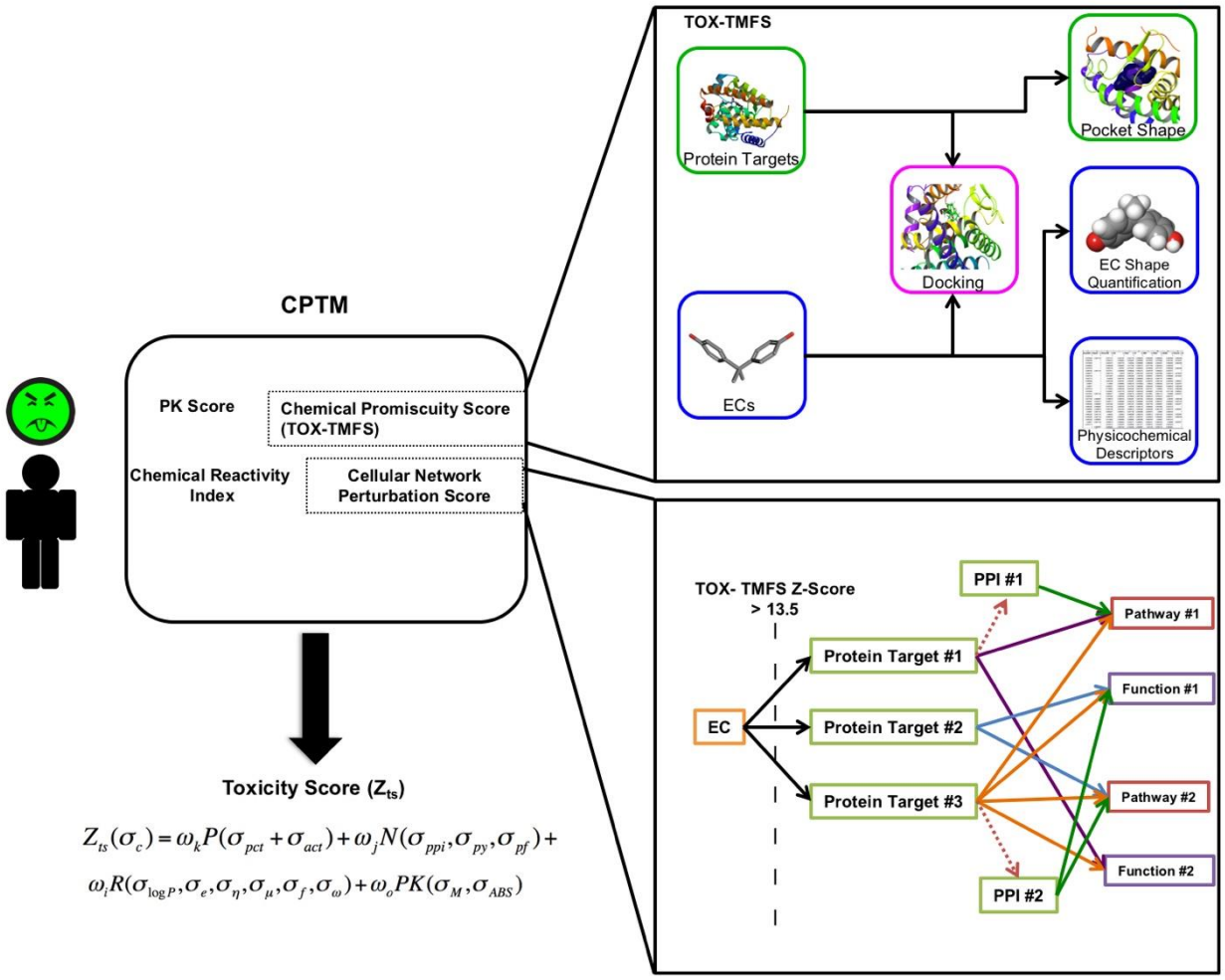



**Figure 1**. Schematic of the Chemo-Phenotypic Toxicity Measurement (CPTM) for the generation of toxicity scores (Zts), highlighting Tox-TMFS-derived protein interactions and resulting cellular networks. An environmental chemical is given a toxicity score based on the combination of biological, chemical, and pharmacokinetic terms. Tox-TMFS is used to predict novel EC-protein target associations in addition to experimentally determined signatures. The number of targets, which include signatures and their associated protein-protein interactions, pathways, and functions, contribute to an EC's chemical promiscuity score and cellular network perturbation score. The pharmacokinetic (PK) score is determined by the total number of potential reactions that an EC can undergo into metabolites as well as the predicted percent human oral absorption. The chemical reactivity index is composed of physicochemical properties associated with general mechanisms of toxicity. The individual score terms

are combined to give a comprehensive toxicity score (Zts) to rank ECs, where higher scores denote greater potential toxicity.

| PDB ID | Protein Name | TMFS Z-Score |
|---|---|---|
| 1X76 | Estrogen Receptor Beta Steroid Receptor Coactivator-1 | 16.6982 |
| 2QGW | Estrogen Receptor Alpha Ligand Binding Domain Complexed With A Chloro-Indazole Compound | 16.4182 |
| 2QAB | Estrogen Receptor Alpha Ligand Binding Domain Mutant 537S Complexed With An Ethyl Indazole Compound | 16.4003 |
| 3L5R | Macrophage Migration Inhibitory Factor | 16.2179 |
| 1YY4 | Estrogen Receptor Beta Steroid Receptor Coactivator-1 | 16.1505 |
| 3MB6 | Casein Kinase Ii Subunit Alpha | 15.8822 |
| 2A5U | Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit, Gamma Isoform | 15.8815 |
| 2UZE | Cell Division Protein Kinase 2 Cyclin A2 | 15.4576 |
| 3K99 | Heat Shock Protein Hsp 90-Alpha | 15.3548 |
| 2W1H | Cell Division Protein Kinase 2 | 15.21 |
| 3ENE | Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Gamma Isoform | 15.1988 |
| 3MAX | Histone Deacetylase 2 | 15.1063 |
| 3CN4 | Transthyretin | 14.9438 |
| 3INL | Aldehyde Dehydrogenase | 14.9217 |
| 2C3K | Serine/Threonine-Protein Kinase Chk1 | 14.768 |
| 2R2W | Plasminogen Activator, Urokinase | 14.6887 |
| 1OWE | Urokinase-Type Plasminogen Activator | 14.6524 |
| 1G4K | Stromelysin-1 | 14.6277 |
| 2QR9 | Estrogen Receptor Alpha Ligand Binding Domain Complexed With An Oxabicyclic Derivative Compound | 14.5379 |
| 1YC4 | Heat Shock Protein Hsp 90-Alpha | 14.3872 |
| 3LKA | Macrophage Metalloelastase | 14.1458 |
| 3BHV | Cell Division Protein Kinase 2 Cyclin-A2 | 13.9828 |
| 2OK1 | Mitogen-Activated Protein Kinase 10 | 13.6429 |
| 2WQO | Serine/Threonine-Protein Kinase Nek2 | 13.5631 |
| 2Z2W | Wee1-Like Protein Kinase | 13.5532 |

**Table 1**. Tox signature of Daidzein, predicted by Tox-TMFS.

Hundreds of ECs have been screened following the development of Tox-TMFS, but Subtask 1 has not been fully screened but is nearing completion. At 2,335, the number of protein targets tested however has far exceeded the intended 254 structures anticipated. The EC-target signature space predicted by Tox-TMFS, which in its methodology and entirety accomplishes the milestone for **Major Task 1**, is large and highly interconnected **(Figure 2)**. A two-tiered validation of EC-target as well as EC-function signatures was performed using the Comparative Toxicogenomics Database (CTD) **[4]**, a comprehensive resource for experimental EC data. 898 (~14%) EC-target interactions and 14,461 EC-function signatures were validated. **Table 2** lists the validations for EC-function signatures derived for methyltestosterone. While the absolute quantity of validations is less than the total prediction space, this is expected as many of these EC associations have yet to be tested experimentally. These results suggest that predicted EC-target interactions using Tox-TMFS are accurate and can be used to infer higher-order biological associations.

| PBD ID | Protein Name | GO Term | GO Function |
|---|---|---|---|
| 1E3K | Progesterone receptor | GO:0003707 | steroid hormone receptor activity |
| 1FDS | Estradiol 17-beta-dehydrogenase 1 | GO:0006629 | lipid metabolic process |
| 1FDS | Estradiol 17-beta-dehydrogenase 1 | GO:0008152 | metabolic process resulting in cell growth |
| 1FDS | Estradiol 17-beta-dehydrogenase 1 | GO:0008202 | steroid metabolic process |
| 1GS4 | Androgen receptor | GO:0003707 | steroid hormone receptor activity |
| 1J96 | Aldo-keto reductase family 1 member C2 | GO:0006629 | lipid metabolic process |
| 1J96 | Aldo-keto reductase family 1 member C2 | GO:0008202 | steroid metabolic process |
| 1J99 | Bile salt sulfotransferase | GO:0005515 | protein binding |
| 1J99 | Bile salt sulfotransferase | GO:0006629 | lipid metabolic process |
| 1J99 | Bile salt sulfotransferase | GO:0008202 | steroid metabolic process |
| 1JTV | Estradiol 17-beta-dehydrogenase 1 | GO:0006629 | lipid metabolic process |
| 1JTV | Estradiol 17-beta-dehydrogenase 1 | GO:0008152 | metabolic process resulting in cell growth |
| 1JTV | Estradiol 17-beta-dehydrogenase 1 | GO:0008202 | steroid metabolic process |
| 1M2Z | Nuclear receptor coactivator 2 | GO:0003707 | steroid hormone receptor activity |
| 1PQ9 | Oxysterols receptor LXR-beta | GO:0003707 | steroid hormone receptor activity |
| 1QYX | Estradiol 17-beta-dehydrogenase 1 | GO:0006629 | lipid metabolic process |
| 1QYX | Estradiol 17-beta-dehydrogenase 1 | GO:0008152 | metabolic process resulting in cell growth |
| 1QYX | Estradiol 17-beta-dehydrogenase 1 | GO:0008202 | steroid metabolic process |
| 1SQN | Progesterone receptor | GO:0003707 | steroid hormone receptor activity |
| 1SR7 | Progesterone receptor | GO:0003707 | steroid hormone receptor activity |
| 1T65 | Androgen receptor | GO:0003707 | steroid hormone receptor activity |
| 1UPV | Oxysterols receptor LXR-beta | GO:0003707 | steroid hormone receptor activity |
| 1X7E | Estrogen receptor | GO:0003707 | steroid hormone receptor activity |
| 1XF0 | Aldo-keto reductase family 1 member C3 | GO:0008202 | steroid metabolic process |
| 1XF0 | Aldo-keto reductase family 1 member C3 | GO:0008584 | male gonad development |
| 1XF0 | Aldo-keto reductase family 1 member C3 | GO:0010942 | positive regulation of cell death |
| 1ZAF | Nuclear receptor coactivator 1 | GO:0003707 | steroid hormone receptor activity |
| 1ZKY | Nuclear receptor coactivator 2 | GO:0003707 | steroid hormone receptor activity |
| 2AA6 | Mineralocorticoid receptor | GO:0003707 | steroid hormone receptor activity |
| 2AAX | Mineralocorticoid receptor | GO:0003707 | steroid hormone receptor activity |
| 2AB2 | Mineralocorticoid receptor | GO:0003707 | steroid hormone receptor activity |
| 2HZQ | Apolipoprotein D | GO:0005515 | protein binding |
| 2HZQ | Apolipoprotein D | GO:0006629 | lipid metabolic process |
| 2PNU | Androgen receptor | GO:0003707 | steroid hormone receptor activity |
| 2POG | Estrogen receptor | GO:0003707 | steroid hormone receptor activity |
| 2Q7I | Androgen receptor | GO:0003707 | steroid hormone receptor activity |
| 2Q7K | Androgen receptor | GO:0003707 | steroid hormone receptor activity |
| 2QAB | Estrogen receptor | GO:0003707 | steroid hormone receptor activity |
| 2RBE | Corticosteroid 11-beta-dehydrogenase isozyme 1 | GO:0006629 | lipid metabolic process |
| 2RBE | Corticosteroid 11-beta-dehydrogenase isozyme 1 | GO:0008152 | metabolic process resulting in cell growth |
| 2RBE | Corticosteroid 11-beta-dehydrogenase isozyme 1 | GO:0008202 | steroid metabolic process |
| 2VCT | Glutathione S-transferase A2 | GO:0008152 | metabolic process resulting in cell growth |
| 2W8Y | Progesterone receptor | GO:0003707 | steroid hormone receptor activity |
| 3BUR | 3-oxo-5-beta-steroid 4-dehydrogenase | GO:0005496 | steroid binding |
| 3BUR | 3-oxo-5-beta-steroid 4-dehydrogenase | GO:0006629 | lipid metabolic process |
| 3BUR | 3-oxo-5-beta-steroid 4-dehydrogenase | GO:0008202 | steroid metabolic process |
| 3BUR | 3-oxo-5-beta-steroid 4-dehydrogenase | GO:0008209 | androgen metabolic process |
| 3BYZ | Corticosteroid 11-beta-dehydrogenase isozyme 1 | GO:0006629 | lipid metabolic process |
| 3BYZ | Corticosteroid 11-beta-dehydrogenase isozyme 1 | GO:0008152 | metabolic process resulting in cell growth |
| 3BYZ | Corticosteroid 11-beta-dehydrogenase isozyme 1 | GO:0008202 | steroid metabolic process |
| 3BZU | Corticosteroid 11-beta-dehydrogenase isozyme 1 | GO:0006629 | lipid metabolic process |
| 3BZU | Corticosteroid 11-beta-dehydrogenase isozyme 1 | GO:0008152 | metabolic process resulting in cell growth |
| 3BZU | Corticosteroid 11-beta-dehydrogenase isozyme 1 | GO:0008202 | steroid metabolic process |
| 3CAS | 3-oxo-5-beta-steroid 4-dehydrogenase | GO:0005496 | steroid binding |
| 3CAS | 3-oxo-5-beta-steroid 4-dehydrogenase | GO:0006629 | lipid metabolic process |
| 3CAS | 3-oxo-5-beta-steroid 4-dehydrogenase | GO:0008202 | steroid metabolic process |
| 3CAS | 3-oxo-5-beta-steroid 4-dehydrogenase | GO:0008209 | androgen metabolic process |
| 3CAV | 3-oxo-5-beta-steroid 4-dehydrogenase | GO:0005496 | steroid binding |
| 3CAV | 3-oxo-5-beta-steroid 4-dehydrogenase | GO:0006629 | lipid metabolic process |
| 3CAV | 3-oxo-5-beta-steroid 4-dehydrogenase | GO:0008202 | steroid metabolic process |
| 3CAV | 3-oxo-5-beta-steroid 4-dehydrogenase | GO:0008209 | androgen metabolic process |
| 3DHE | Estradiol 17-beta-dehydrogenase 1 | GO:0006629 | lipid metabolic process |
| 3DHE | Estradiol 17-beta-dehydrogenase 1 | GO:0008152 | metabolic process resulting in cell growth |
| 3DHE | Estradiol 17-beta-dehydrogenase 1 | GO:0008202 | steroid metabolic process |
| 3KLM | Estradiol 17-beta-dehydrogenase 1 | GO:0006629 | lipid metabolic process |
| 3KLM | Estradiol 17-beta-dehydrogenase 1 | GO:0008152 | metabolic process resulting in cell growth |
| 3KLM | Estradiol 17-beta-dehydrogenase 1 | GO:0008202 | steroid metabolic process |
| 3L3X | Androgen receptor | GO:0003707 | steroid hormone receptor activity |
| 3L3Z | Androgen receptor | GO:0003707 | steroid hormone receptor activity |

**Table 2**. EC-function signature predicted via protein interactions for the EC methyltestosterone.
**Major Task 2**. Development of a Cancer Cellular Network Model Using the DrugGenEx-Net Method.

Subtask 1. Using DrugGenEx-Net, a novel molecular profiling method, cancer cellular networks such as protein-protein interaction (PPI), function, and pathways will be created by linking the predicted chemical-protein interaction signatures obtained from major task 1. DrugGenEx-Net method is used to construct a disease networks (here cancer) from empirically predicted chemical-target (CP) interactions to explore their relationships in human diseases, and mechanistic insights. CP signatures were predicted using the Tox-TMFS method. Chemicals were associated with diseases via their predicted targets using the OMIM Morbid Map database to create CP-bipartite networks. It is also possible that a chemical is connected to a disease because the chemical targets multiple proteins or because a protein is associated with multiple diseases, both cases being indistinguishable from the network.

Subtask 1. This task was listed to be completed in 3-8 months. This task was completed in 6 months, in March 2016.

Subtask 2. Relationship analyses among top 40 ECP interactions (selected based on Tox-TMFS Z-score) and PPI/pathway/function.
Subtask 2 was listed to be completed in 4-10 months. This task was completed in 6 months, in March 2016.

Subtask 3. Generate EC-CRC network through predicted ECP-protein target associations for CRC-related targets and further incorporation of signaling pathways and molecular functions. Subtask 3 was listed to be completed in 5-16 months. This task is 75% complete.

Milestone(s) Achieved 1: Potential mechanistic biological network perturbations of cancer (types listed in the RFA) and colorectal cancer (CRC) network for each ECs will be identified.
Milestone Achieved 1 was listed to be completed in 8-16 months. This milestone was completed in 9 months, in June 2016.

Milestone(s) Achieved 2. Small number of ECs that mechanistically contribute to the perturbations of VDR, TGFβ and Wnt/β-catenin, signaling pathways will be selected for biological testing described in aim 2.

Milestone Achieved 2 was listed to be completed in 8-16 months. This milestone was completed in 9 months, in June 2016.

**Accomplishments**

**Major Task 2**. Development of a Cancer Cellular Network Model Using the DrugGenEx-Net Method.

Confident in the accuracy of protein-level predictions, we employed these signatures for assessing the biological effects of the EC set. Under subtasks 1 and 2, networks for these biological effects have so far been modeled against 12 oncologic diseases. Colorectal cancer (CRC) is a test case for which the biological network perturbations for each EC was assessed and for which a small number of prioritized ECs were selected for biological testing.

**Subtask 1** was accomplished when ECs were related to higher-order biological effects through their predicted and annotated targets by way of our DrugGenEx-Net method **(Figure 2)**. The Comparative Toxicogenomics Database (CTD) **[4]** and The Toxin and Toxin Target Database (T3DB) **[5]** were used to obtain experimentally annotated targets for ECs. Protein-protein interactions (PPIs) were obtained for the entire protein dataset using the ExPASy STRING database **[6]**. A confidence score cutoff of 0.95 was used to extract associations likely to be true positives. Annotation of functions from Gene Ontology **[7,8]** was performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) Functional Annotation Tool **[9,10]**, while annotation for pathways from KEGG **[11]**, REACTOME **[12]**, PharmGKB **[13]**, NetPath **[14]**, BioCarta **[15]**, WikiPathway **[16]**, and Pathway Interaction Database **[17]** was performed using ConsensusPathDB **[18]**. Enriched pathways and functions for individual ECs were selected using P-value < 0.05 and FDR < 0.25. Cytoscape v2.8.3 was used to create EC-effect networks **[19]**.

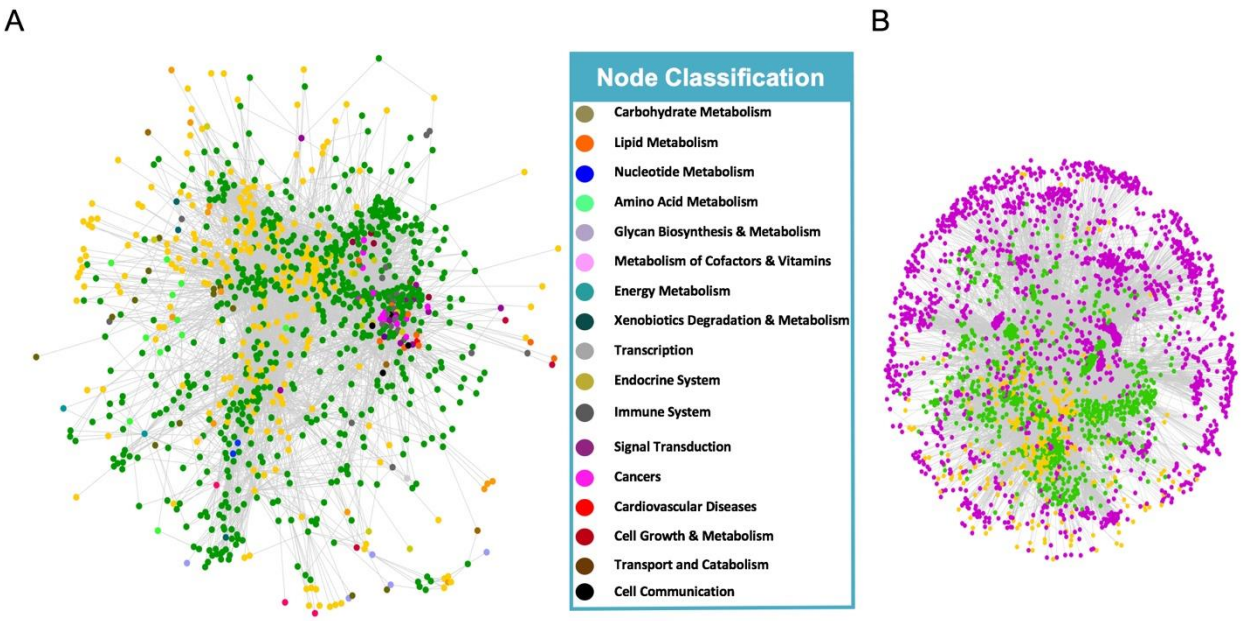



**Figure 2**. Tox-TMFS and DrugGenEx-Net generated EC network biological space from predicted EC-protein target interactions. Tox-TMFS was applied to a dataset of 420 environmental chemical (EC) structures and 2,335 human protein target crystal structures. (A) EC-protein-pathway tripartite network; pathways are further sub-classified into multiple categories per the KEGG classification. (B) EC (yellow nodes)-protein target (green nodes)-function (purple nodes) tripartite network.

Known EC-cancer associations were recapitulated by way of developing Cancer EC-biological effect networks through cancer-related targets linked to the OMIM database **[20]**. **Table 3** catalogs the ECs correctly predicted to be associated with Lung, Prostate, and Colorectal cancers by way of their multi-level targets.

| Lung Cancer - OMIM #211980 | | |
|---|---|---|
| CASRN | Environmental Chemical | Reference |
| 101-14-4 | 4,4'-Methylenebis(2-chloroaniline) | Butler *et al*, Cancer Res, 1989 |
| 71-43-2 | Benzene | Yin *et al*, Environ Health Perspect, 1996 |
| 2921-88-2 | Chlorpyrifos | Alavanja *et al*, Am J Epidemiol, 2004 |
| 7440-48-4 | Cobalt | Suh *et al*, Regul Toxicol Pharmacol, 2016 |
| 60-57-1 | Dieldrin | Bonner *et al*, Environ Health Perspect, 2016 |
| 64-17-5 | Ethanol | Bagnardi *et al*, Am J Epidemiol, 2010 |
| 58-89-9 | Gamma-Hexachlorocyclohexane | Rafnsson, Scand J Work Environ Health, 2006 |
| 51218-45-2 | Metolachlor | Alavanja *et al*, Am J Epidemiol, 2004 |
| 7440-02-0 | Nickel | Chiou *et al*, Toxicol Lett, 2015 |
| 40487-42-1 | Pendimethalin | Alavanja *et al*, Am J Epidemiol, 2004 |
| 51-71-8 | Phenelzine | Toth, Cancer Res, 1976 |
| Prostate Cancer - OMIM #176807 | | |
| CASRN | Environmental Chemical | Reference |
| 309-00-2 | Aldrin | Koutros *et al*, Am J Epidemiol, 2013 |
| 319-85-7 | Beta-Hexachlorocyclohexane | Kumar *et al*, Chemosphere, 2010 |
| 80-05-7 | Bisphenol A | Prins *et al*, Endocrinology, 2014 |
| 143-50-0 | Chlordecone | Multigner *et al*, J Clin Oncol, 2010 |
| 50-29-3 | Clofenotane | Settimi *et al*, Int J Cancer, 2003 |
| 56-72-4 | Coumaphos | Koutros *et al*, Am J Epidemiol, 2013 |
| 115-32-2 | Dicofol | Settimi *et al*, Int J Cancer, 2003 |
| 521-18-6 | Dihydrotestosterone | Gupta *et al*, BMC Urol, 2014 |
| 57-63-6 | Ethinyl Estradiol | Shirai *et al*, Cancer Res, 1986 |
| 58-89-9 | Gamma-Hexachlorocyclohexane | Kalantzi *et al*, Carcinogenesis, 2004 |
| 70-18-8 | Glutathione | Keshari *et al*, J Nucl Med, 2013 |
| 76-44-8 | Heptachlor | Mills *et al*, J Occup Environ Med. 2003 |
| 72-55-9 | P,P'-DDE | Kumar *et al*, Chemosphere, 2010 |
| 52645-53-1 | Permethrin | Koutros *et al*, Am J Epidemiol, 2013 |
| Colorectal Cancer - OMIM #114500 | | |
| CASRN | Environmental Chemical | Reference |
| 52645-53-1 | Permethrin | Andreotti *et al*, Cancer Causes Control, 2010 |
| 2921-88-2 | Chlorpyrifos | Lee *et al*, Int J Cancer, 2007 |
| 116-06-3 | Aldicarb | Lee *et al*, Int J Cancer, 2007 |
| 789-02-6 | O.P'-Ddt | Andreotti *et al*, Cancer Causes Control, 2010 |
| 72-55-9 | P.P'-Dde | Song *et al*, PLOS ONE, 2014 |
| 51218-45-2 | Metolachlor | Andreotti *et al*, Cancer Causes Control, 2010 |
| 34256-82-1 | Acetochlor | Lerro *et al*, Int J Cancer, 2015 |
| 60-57-1 | Dieldrin | Lee *et al*, Int J Cancer, 2010 |
| 8001-35-2 | Toxaphene | Purdue *et al*, Int J Cancer, 2007 |
| 1582-09-8 | Trifluralin | Andreotti *et al*, Cancer Causes Control, 2010 |
| 15972-60-8 | Alachlor | Leet *et al*, Am J Ind Med, 1996 |
| 21725-46-2 | Cyanazine | Andreotti *et al*, Cancer Causes Control, 2010 |
| 309-00-2 | Aldrin | Swaen *et al*, Toxicol Ind Health, 2002 |
| 319-85-7 | Beta-Hexachlorocyclohexane | Soliman *et al*, Arch Environ Health, 1997 |
| 319-84-6 | Alpha-Hexachlorocyclohexane | Howsam *et al*, Environ Health Perspect, 2004 |
| 1912-24-9 | Atrazine | Lerro *et al*, Int J Cancer, 2015 |
| 76-44-8 | Heptachlor | Purdue *et al*, Int J Cancer, 2007 |
| 118-74-1 | Hexachlorobenzene | Howsam *et al*, Environ Health Perspect, 2004 |
| 333-41-5 | Diazinon | Weichenthal *et al*, Environ Health Perspect, 2010 |
| 121-75-5 | Malathion | Andreotti *et al*, Cancer Causes Control, 2010 |

**Table 3**. Selected validations of ECs perturbing Lung and Prostate Cancer by building Cancer Cellular Network Models using DrugGenEx-Net.

CRC is a solid malignancy typically diagnosed in late adulthood (>50 years). EC exposures over many years may be attributable to a subset of sporadic CRCs. A CRC EC-biological effect network was generated through CRC-related targets (**Figure 3a**). ECs were associated with CRC if their targets were annotated with CRC in the OMIM database. Centrally clustered ECs include butyl benzyl phthalate, oxazepam, bifenazate, sulfaquinoxaline and genistein. The uses of these compounds vary and are found in distinct environment spaces (i.e. plasticizers, pesticides, etc.). However, their chemical structures are similar in that they contain two substituted benzyl groups that are consistently spaced apart (**Figure 3b**). This structure is similar to polychlorinated biphenyls, which are pesticides that correlate with CRC. As molecules with structural similarity tend to exhibit similar biological properties, these molecules may be associated with CRC through similar mechanisms. Alternatively, molecules such as oleic acid and progesterone are less clustered and found peripherally in the network. Such molecules are of distinct chemical structure classes (fatty acid and steroidal hormone, respectively; **Figure 3c**) that tend to bind to more specific targets. Collectively, Tox-TMFS predicted ECs of various chemical structures to be associated with CRC.

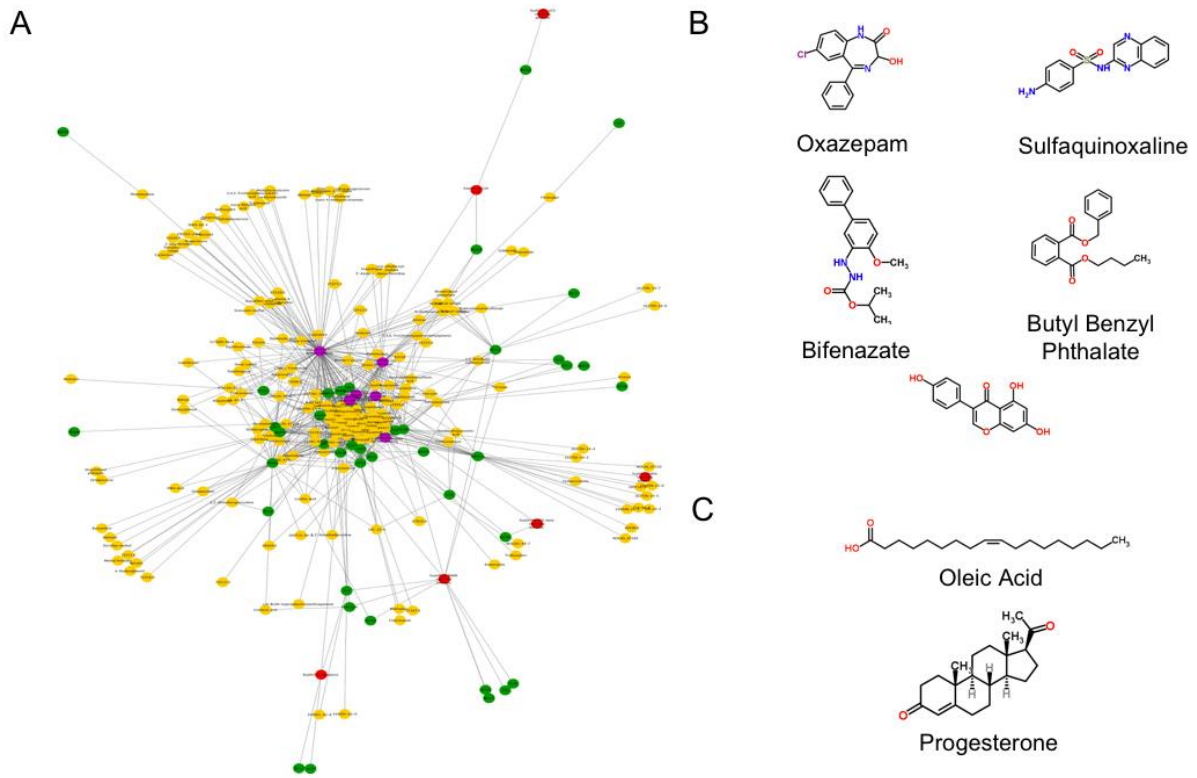



**Figure 3**. Environmental chemicals (ECs) predicted to be involved in colorectal cancer (CRC). (A) ECs (yellow nodes) predicted by Tox-TMFS to associate with CRC through disease-related protein targets (green nodes). Relevant targets were chosen if they were found to be related to CRC in OMIM, the medical literature, through CRC-enriched GO molecular functions (purple nodes), or through relevant sub-pathways (red nodes) found under the main KEGG pathway hsa05210:colorectal cancer. (B & C) Subset of chemical structures for predicted CRC-associated ECs that clustered centrally (B) or are found peripherally (C) in the CRC network.

Of the predicted CRC-associated ECs, literature validation revealed some to be beneficial and others to be harmful (note that ECs are defined as chemicals encountered in the environment that are not necessarily deleterious). Beneficial associations include phytoestrogens (genistein and daidzein) **[21]**, flavonoids (quercetin) **[22]** and non-steroidal anti-inflammatory drugs (diclofenac) **[23]**. These molecules are commonly known for their anti-inflammatory properties and ability to inhibit multiple kinases, which are important for inhibiting cancer growth. Warfarin, an anticoagulant, was also found to be beneficial **[24]**. Alternatively, a CRC-promoting association was found for the laxative phenolphthalein **[25]**. However, there is no consensus yet on this association as epidemiological data is limited **[26]**.

The vitamin D receptor (VDR) is a type II nuclear receptor implicated in various downstream cellular processes **(Figure 4)**. Binding of the active form of vitamin D3 (calcitriol) activates VDR and exhibits pleiotropic beneficial health effects **[27]**, such as anti-inflammation and cancer prevention. Low serum vitamin D levels, inhibition or downregulation of VDR have been conversely associated with autoimmune diseases such as Crohn's **[28]**, rheumatoid arthritis **[29]** and psoriasis **[30]**. Exposure to certain ECs may pathologically dysregulate VDR signaling.

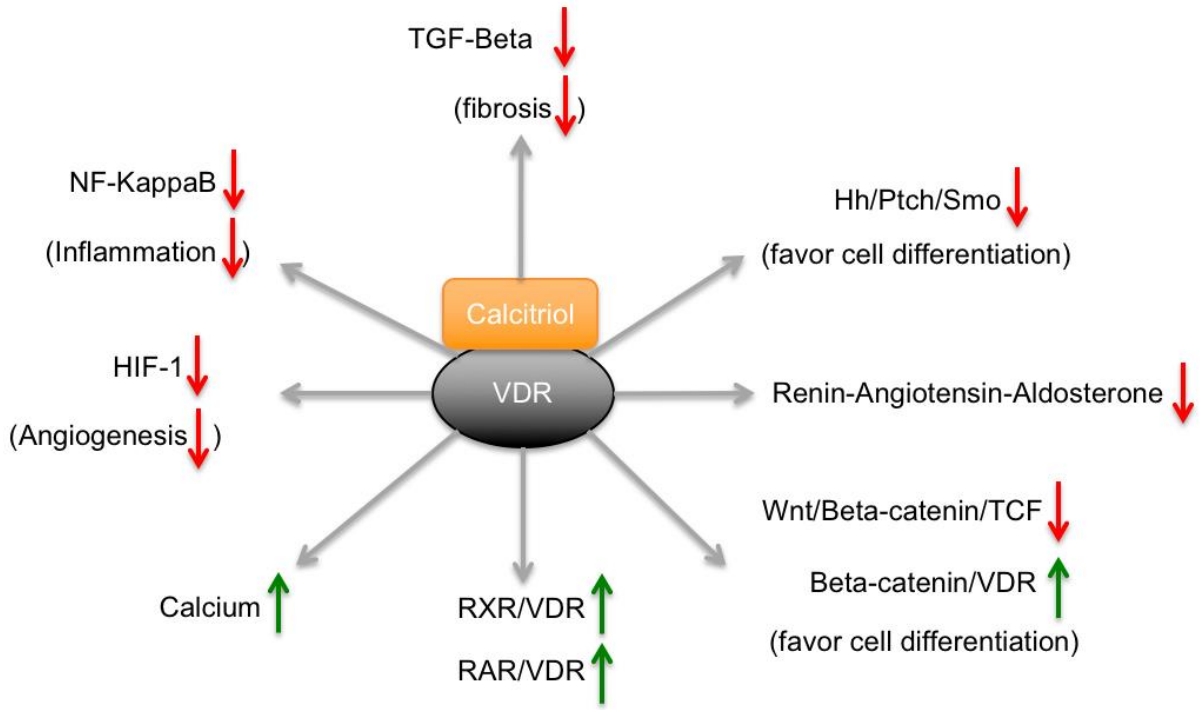



**Figure 4**. Downstream effects of VDR activation by its cognate ligand calcitriol (activated Vitamin D3). Arrows denote directionality of biological response.

Tox-TMFS predicted nordihydroguaiaretic acid (NDGA) and N,N'-disalicylidene-1,2-diaminopropane to bind VDR. NDGA is a phenolic compound derived from the Creosote bush used as a natural supplement for its anti-oxidant activity. However, medicinal use of NDGA is controversial as NDGA also exhibits pro-oxidant activity resulting in hepato- and renal toxicity **[31]**. NDGA has also recently been identified as a novel VDR antagonist **[32]**. We hypothesize that NDGA-induced VDR antagonism induces toxicity.

N,N'-disalicylidene-1,2-diaminopropane is a salen-type ligand used as a fuel additive in motor oils to deactivate metals. Motor oil exposure has been linked to autoimmune diseases such as rheumatoid

arthritis **[33]**, pro-inflammatory skin conditions **[34]**, skin cancers **[35]**, and other diseases. Deployed military personnel, aerospace workers and automobile mechanics are occupation groups with significantly increased exposure to motor oils. Exposure usually occurs via direct contact or through inhalation, such as when flying in poorly ventilated aircrafts for extended periods or being in close proximity to fire burn pits in active warzones like Afghanistan and Iraq. The increase in the prevalence of certain pathologies- lung cancer and malignant melanoma in aerospace workers **[36]** and rectal cancer in automobile manufacturing workers **[37]** - may be attributed to multiple effects from motor oil-induced dysregulation of VDR signaling **(Figure 4)**. To date, there is no established link between motor oil exposure and CRC prevalence. Our results suggest a long-term epidemiological study to define this association given the extended time course for CRC development.

VDR signaling is highly complex with multifaceted health implications **(Figure 4)**. Perturbation of the VDR network through antagonism or decreased VDR protein expression is pathological. Combining Tox-TMFS predictions with literature findings highlights the potential of EC exposure to cause a myriad of diseases through VDR signaling disruption. This informed our prioritization of three environmental chemicals for biological testing as part of **subtask 3**, emphasizing important CRC pathways including VDR, TGFß, and Wnt/ß-catenin signaling pathways.

In selecting candidates for biological testing, we examined all cancer cellular network models for each EC characterized by way of Tox-TMFS predicted and previously known protein interactions, and, as described above and accomplishing **Milestone 1**, extrapolated cellular actions associated with those interactions at the levels of PPI, pathways, and functions using DrugGenEx-Net. N,N'-disalicylidene-1,2-diaminopropane, as discussed above, was predicted to bind to VDR, one of the components of the EC mechanistic biological network that was annotated as associated with CRC. **Figure 5** illustrates the full biological action network of N,N'-disalicylidene-1,2-diaminopropane, with activity associated with VDR highlighted in pink. VDR perturbation constitutes a significant share of the predicted interactome of this EC, but it is not the sole subnetwork associated with predicted targets using the OMIM database. We additionally predicted by way of Tox-TMFS that this EC would bind to CRC targets MAOB (Monoamine oxidase B) and MMP12 (Matrix Metallopeptidase 12), and perturb a set of CRC-linked

pathways, among them P53 and tyrosine metabolism, two pathways which have been strongly implicated in CRC disease progression **[38]** (**Table 4**). Importantly, as shown in **Table 4**, N,N'-disalicylidene-1,2-diaminopropane was additionally linked to TGF-beta, Wnt/Beta-catenin, and the Renin-Angiotensin System, through the three CRC-linked protein targets and others which were not directly associated with CRC. These pathways are important downstream components associated with VDR **(Figure 4)**, and further inform our hypothesis that the motor oil additive N,N'-disalicylidene-1,2-diaminopropane induces CRC toxicity by way of interaction with VDR and its offshoot cellular mechanisms. In accomplishing milestone 2, N,N'-disalicylidene-1,2-diaminopropane was therefore selected for biological testing for VDR binding, Wnt/Beta-Catenin pathway activation studies, and other tests as discussed under **Major Task 4**.

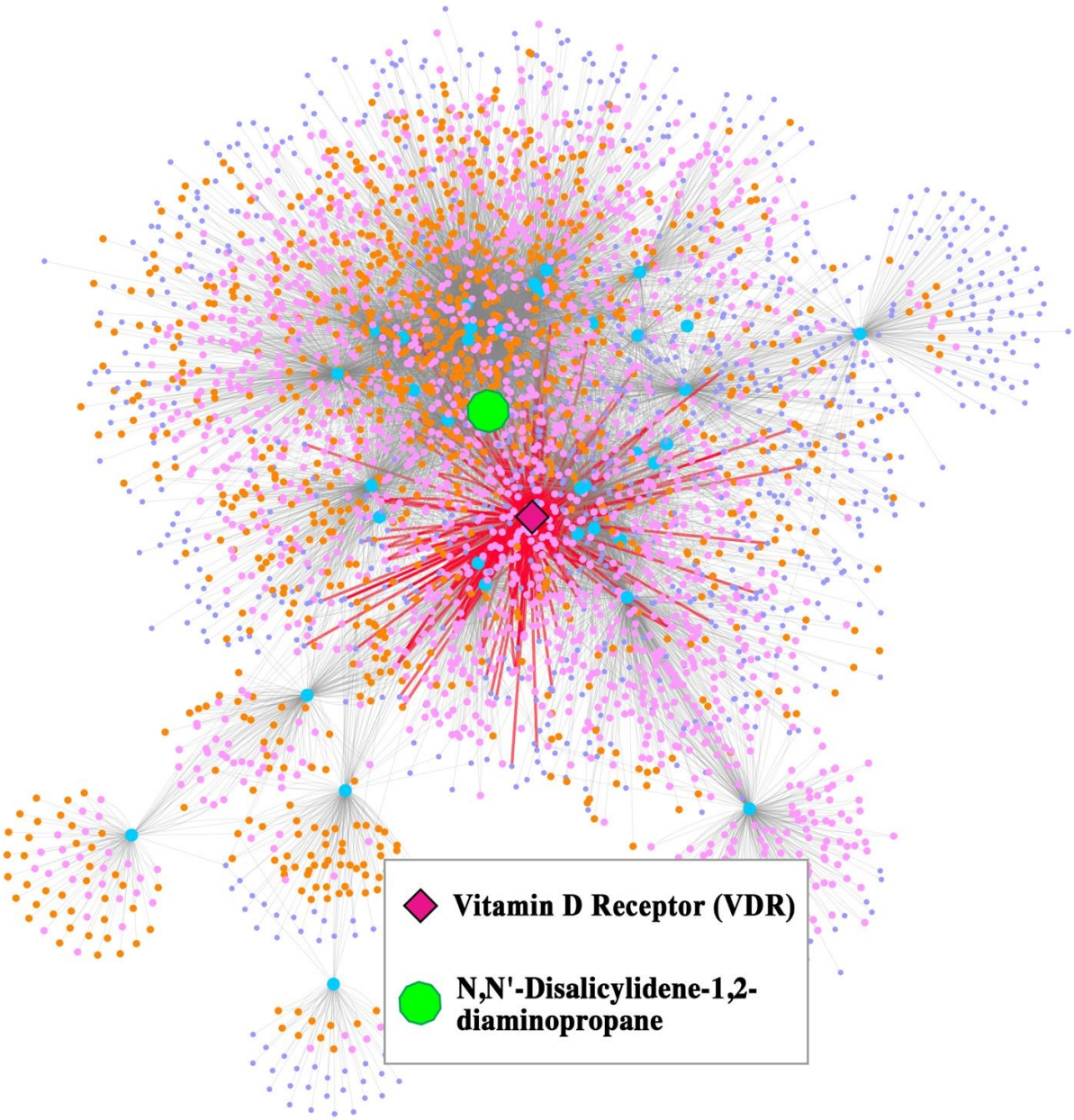
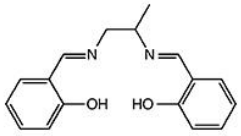



**Figure 5**. Entire Cellular Network Model for N,N'-disalicylidene-1,2-diaminopropane with highlighted Vitamin D Receptor (VDR) subnetwork. Green node is the EC, blue nodes are direct protein interactions, light pink nodes are PPIs, orange nodes are pathways, and purple nodes are molecular functions. Dark pink node is the direct interacting protein Vitamin D Receptor (VDR), with protruding edges in red to signify direct perturbation network of VDR. This network excludes disease associations from the OMIM database, instead describing all possible disease-associated biological activity perturbed from an EC-centric viewpoint.

N,N'-Disalicylidene-1,2-diaminopropane



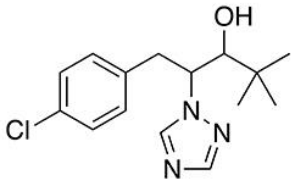
| UniprotID | Protein Name | Gene Symbol | Tox-TMFS Predicted? |
|---|---|---|---|
| P11473 | Vitamin D3 receptor | VDR | Yes |
| P27338 | Amine oxidase [flavin-containing] B | MAOB | Yes |
| P39900 | Macrophage metalloelastase | MMP12 | Yes |

| PathwayID | Pathway Name | Source |
|---|---|---|
| hsa04110 | Cell cycle | KEGG |
| hsa05200 | Pathways in cancer | KEGG |
| REACT_604 | Hemostasis | Reactome |
| hsa04914 | Progesterone-mediated oocyte maturation | KEGG |
| hsa05222 | Small cell lung cancer | KEGG |
| hsa04114 | Oocyte meiosis | KEGG |
| Pathway_TGF_beta_Receptor | TGF_beta_Receptor | NetPath |
| betacatenin_nuc_pathway | Regulation of nuclear beta catenin signaling and target gene transcription | PID |
| hsa04115 | p53 signaling pathway | KEGG |
| REACT_115566 | Cell Cycle | Reactome |
| REACT_383 | DNA Replication | Reactome |
| hsa04060 | Cytokine-cytokine receptor interaction | KEGG |
| REACT_118779 | Extracellular matrix organization | Reactome |
| hsa04350 | TGF-beta signaling pathway | KEGG |
| hsa04916 | Melanogenesis | KEGG |
| hsa04310 | Wnt signaling pathway | KEGG |
| REACT_15518 | Transmembrane transport of small molecules | Reactome |
| REACT_160300 | Binding and Uptake of Ligands by Scavenger Receptors | Reactome |
| hsa05219 | Bladder cancer | KEGG |
| hsa04610 | Complement and coagulation cascades | KEGG |
| hsa04512 | ECM-receptor interaction | KEGG |
| tgfbpathway | tgf beta signaling pathway | BioCarta |
| hsa04360 | Axon guidance | KEGG |
| REACT_111045 | Developmental Biology | Reactome |
| hsa04062 | Chemokine signaling pathway | KEGG |
| hsa04920 | Adipocytokine signaling pathway | KEGG |
| REACT_111217 | Metabolism | Reactome |
| hsa00590 | Arachidonic acid metabolism | KEGG |
| hsa00982 | Drug metabolism - cytochrome P450 | KEGG |
| hsa04270 | Vascular smooth muscle contraction | KEGG |
| hsa04724 | Glutamatergic synapse | KEGG |
| REACT_13685 | Neuronal System | Reactome |
| hsa00350 | Tyrosine metabolism | KEGG |
| hsa00360 | Phenylalanine metabolism | KEGG |
| hsa04960 | Aldosterone-regulated sodium reabsorption | KEGG |
| PA165110622 | Agents Acting on the Renin-Angiotensin System Pathway, Pharmacodynamics | PharmGKB |
| hsa04961 | Endocrine and other factor-regulated calcium reabsorption | KEGG |
| hsa05144 | Malaria | KEGG |
| hsa04978 | Mineral absorption | KEGG |
| hsa04514 | Cell adhesion molecules (CAMs) | KEGG |
| hsa05414 | Dilated cardiomyopathy | KEGG |
| hsa04672 | Intestinal immune network for IgA production | KEGG |
| hsa00040 | Pentose and glucuronate interconversions | KEGG |

**Table 4**. CRC-linked direct proteins and pathways associated with biological testing candidate EC N,N'-disalicylidene-1,2-diaminopropane, with components of interest labeled red.

It is the case for many of our assessed ECs that several or all of the chemical-protein target interactions were previously established in the scientific literature. Cancer-linked biological activity annotated to those interactions, however, are largely uncharacterized at the present time. This leaves opportunities for testing of predicted cancer-linked pathway activity and cancer progression extrapolated by way of our cancer cellular network models using the DrugGenEx-Net procedure. As delineated in **Table 5**, Paclobutrazol (PBZ), a plant growth retardant and triazole fungicide, interacts with four CRC-associated proteins, including MAOB (Monoamine oxidase B), ABCB1 (ATP-binding cassette sub-family B member 1), CXCL10 (C-X-C Motif Chemokine Ligand 10), and MMP1 (Interstitial Collagenase). Associated with these and non-CRC target proteins were a large set of CRC pathways, including TGF-Beta, VDR, and Beta-Catenin signaling pathways. Outlined in **Figure 6** are the direct protein targets predicted to interact with Paclobutrazol and associated with these key CRC pathways. Omitted are the PPIs also linked to these pathways and interacting with the direct protein highlighted in Figure 6 as well

as other direct proteins. We hypothesize that PBZ, due to its predicted perturbation of the VDR pathway and associated downstream activity, as well as predicted associations with CRC-linked cellular components such as P53 and PPAR signaling pathways, will exhibit pathological perturbation on CRC disease models. We chose PBZ as a candidate for disruption of VDR/Beta-Catenin interactions. This assessment is further discussed under **Major Task 4**.




Paclobutrazol

| UniprotID | Protein Name | Gene Symbol | Tox-TMFS Predicted? | PathwayID | Pathway Name | Source |
|---|---|---|---|---|---|---|
| P27338 | Amine oxidase [flavin-containing] B | MAOB | Yes | hsa04060 | Cytokine-cytokine receptor interaction | KEGG |
| P08183 | Multidrug resistance protein 1 | ABCB1 | No | hsa05323 | Rheumatoid arthritis | KEGG |
| P02778 | C-X-C motif chemokine 10 | CXCL10 | No | hsa04110 | Cell cycle | KEGG |
| P03956 | Interstitial collagenase | MMP1 | No | hsa04350 | TGF-beta signaling pathway | KEGG |
| | | | | hsa05200 | Pathways in cancer | KEGG |
| | | | | Pathway_TGF_beta_Receptor | TGF_beta_Receptor | NetPath |
| | | | | REACT_118779 | Extracellular matrix organization | Reactome |
| | | | | REACT_604 | Hemostasis | Reactome |
| | | | | WP2877 | Vitamin D Receptor Pathway | Wikipathway |
| | | | | REACT_15518 | Transmembrane transport of small molecules | Reactome |
| | | | | tgfbpathway | tgf beta signaling pathway | BioCarta |
| | | | | hsa05219 | Bladder cancer | KEGG |
| | | | | hsa04610 | Complement and coagulation cascades | KEGG |
| | | | | hsa04115 | p53 signaling pathway | KEGG |
| | | | | betacatenin_nuc_pathway | Regulation of nuclear beta catenin signaling and target gene transcription | PID |
| | | | | REACT_111217 | Metabolism | Reactome |
| | | | | hsa00140 | Steroid hormone biosynthesis | KEGG |
| | | | | hsa00830 | Retinol metabolism | KEGG |
| | | | | hsa00980 | Metabolism of xenobiotics by cytochrome P450 | KEGG |
| | | | | hsa00590 | Arachidonic acid metabolism | KEGG |
| | | | | hsa00982 | Drug metabolism - cytochrome P450 | KEGG |
| | | | | hsa00983 | Drug metabolism - other enzymes | KEGG |
| | | | | hsa04976 | Bile secretion | KEGG |
| | | | | hsa04060 | Cytokine-cytokine receptor interaction | KEGG |
| | | | | hsa04640 | Hematopoietic cell lineage | KEGG |
| | | | | hsa04672 | Intestinal immune network for IgA production | KEGG |
| | | | | hsa05200 | Pathways in cancer | KEGG |
| | | | | hsa05414 | Dilated cardiomyopathy | KEGG |
| | | | | REACT_118779 | Extracellular matrix organization | Reactome |
| | | | | REACT_604 | Hemostasis | Reactome |
| | | | | hsa00350 | Tyrosine metabolism | KEGG |
| | | | | hsa00360 | Phenylalanine metabolism | KEGG |
| | | | | WP2877 | Vitamin D Receptor Pathway | Wikipathway |
| | | | | hsa02010 | ABC transporters | KEGG |
| | | | | REACT_15518 | Transmembrane transport of small molecules | Reactome |
| | | | | hsa05144 | Malaria | KEGG |
| | | | | PA165110622 | Agents Acting on the Renin-Angiotensin System Pathway, Pharmacodynamics | PharmGKB |
| | | | | hsa04062 | Chemokine signaling pathway | KEGG |
| | | | | hsa04961 | Endocrine and other factor-regulated calcium reabsorption | KEGG |
| | | | | hsa03320 | PPAR signaling pathway | KEGG |
| | | | | hsa04610 | Complement and coagulation cascades | KEGG |
| | | | | hsa04514 | Cell adhesion molecules (CAMs) | KEGG |
| | | | | hsa00040 | Pentose and glucuronate interconversions | KEGG |
| | | | | hsa04960 | Aldosterone-regulated sodium reabsorption | KEGG |
| | | | | REACT_111045 | Developmental Biology | Reactome |

**Table 5**. CRC-linked direct proteins and pathways associated with biological testing candidate EC Paclobutrazol, with components of interest labeled red.

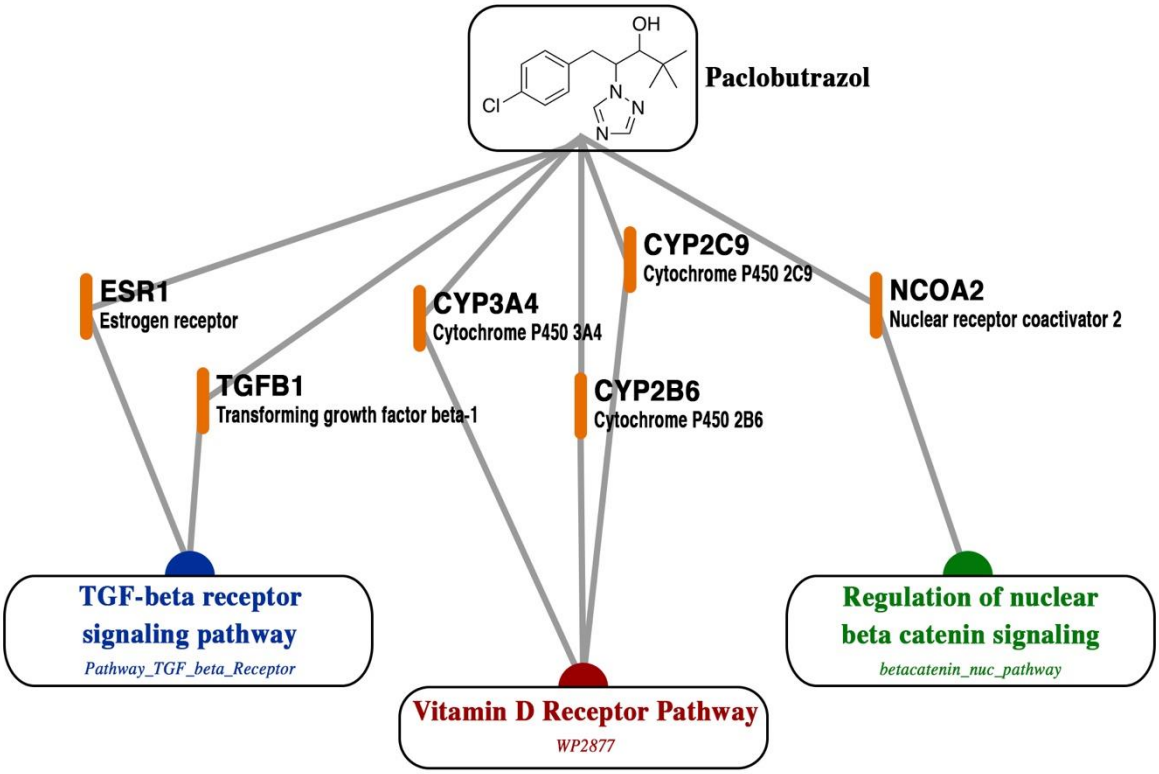



**Figure 6**. Predicted direct protein targets of Paclobutrazol involved in TGF-beta Receptor Signaling Pathway, and Vitamin D Receptor Pathway.

Pyraclostrobin is an agricultural fungicide which we predicted to have the largest CRC cellular perturbation network. **Table 6** lists protein targets for this EC which are CRC-associated, as well as the CRC-linked pathways represented in its perturbation network. The activity predicted is similar in nature to that of PBZ, including predicted interaction with TGF-Beta, VDR, and Beta-Catenin signaling pathways, and also including the Wnt signaling pathway. As outlined under **Major task 4**, pyraclostrobin was prioritized for biological testing in Wnt/Beta-Catenin pathway activation studies using reporter assays, and for VDR/Bet-catenin interactions which we will test by mammalian two hybrid assay.

| Pyraclostrobin | | | | | | |
|---|---|---|---|---|---|---|
| UniprotID | Protein Name | Gene Symbol | Tox-TMFS Predicted? | PathwayID | Pathway Name | Source |
| P08183 | Multidrug resistance protein 1 | ABCB1 | No | hsa04060 | Cytokine-cytokine receptor interaction | KEGG |
| O95342 | Bile salt export pump | ABCB11 | No | hsa04350 | TGF-beta signaling pathway | KEGG |
| P02778 | C-X-C motif chemokine 10 | CXCL10 | No | hsa03440 | Homologous recombination | KEGG |
| P10145 | Interleukin-8 | CXCL8 | No | hsa04512 | ECM-receptor interaction | KEGG |
| P03956 | Interstitial collagenase | MMP1 | No | hsa04974 | Protein digestion and absorption | KEGG |
| | | | | REACT_118779 | Extracellular matrix organization | Reactome |
| | | | | hsa05323 | Rheumatoid arthritis | KEGG |
| | | | | hsa04110 | Cell cycle | KEGG |
| | | | | hsa04115 | p53 signaling pathway | KEGG |
| | | | | hsa04310 | Wnt signaling pathway | KEGG |
| | | | | hsa05200 | Pathways in cancer | KEGG |
| | | | | hsa05217 | Basal cell carcinoma | KEGG |
| | | | | hsa05222 | Small cell lung cancer | KEGG |
| | | | | Pathway_TGF_beta_Receptor | TGF_beta_Receptor | NetPath |
| | | | | REACT_115566 | Cell Cycle | Reactome |
| | | | | REACT_115566 | Cell Cycle | Reactome |
| | | | | REACT_604 | Hemostasis | Reactome |
| | | | | REACT_160300 | Binding and Uptake of Ligands by Scavenger Receptors | Reactome |
| | | | | WP1531 | Vitamin D Metabolism | Wikipathway |
| | | | | WP2877 | Vitamin D Receptor Pathway | Wikipathway |
| | | | | hsa05219 | Bladder cancer | KEGG |
| | | | | betacatenin_nuc_pathway | Regulation of nuclear beta catenin signaling and target gene transcription | PID |
| | | | | REACT_15518 | Transmembrane transport of small molecules | Reactome |
| | | | | hsa04916 | Melanogenesis | KEGG |
| | | | | hsa04610 | Complement and coagulation cascades | KEGG |
| | | | | hsa00040 | Pentose and glucuronate interconversions | KEGG |
| | | | | hsa04060 | Cytokine-cytokine receptor interaction | KEGG |
| | | | | hsa04080 | Neuroactive ligand-receptor interaction | KEGG |
| | | | | hsa04020 | Calcium signaling pathway | KEGG |
| | | | | hsa04640 | Hematopoietic cell lineage | KEGG |
| | | | | hsa04970 | Salivary secretion | KEGG |
| | | | | hsa04972 | Pancreatic secretion | KEGG |
| | | | | hsa04514 | Cell adhesion molecules (CAMs) | KEGG |
| | | | | hsa04672 | Intestinal immune network for IgA production | KEGG |
| | | | | hsa04974 | Protein digestion and absorption | KEGG |
| | | | | REACT_118779 | Extracellular matrix organization | Reactome |
| | | | | hsa00140 | Steroid hormone biosynthesis | KEGG |
| | | | | hsa00830 | Retinol metabolism | KEGG |
| | | | | hsa00980 | Metabolism of xenobiotics by cytochrome P450 | KEGG |
| | | | | REACT_111217 | Metabolism | Reactome |
| | | | | hsa00982 | Drug metabolism - cytochrome P450 | KEGG |
| | | | | hsa00590 | Arachidonic acid metabolism | KEGG |
| | | | | hsa00983 | Drug metabolism - other enzymes | KEGG |
| | | | | hsa04976 | Bile secretion | KEGG |
| | | | | hsa03320 | PPAR signaling pathway | KEGG |
| | | | | hsa04920 | Adipocytokine signaling pathway | KEGG |
| | | | | hsa05200 | Pathways in cancer | KEGG |
| | | | | REACT_604 | Hemostasis | Reactome |
| | | | | hsa02010 | ABC transporters | KEGG |
| | | | | REACT_160300 | Binding and Uptake of Ligands by Scavenger Receptors | Reactome |
| | | | | WP2877 | Vitamin D Receptor Pathway | Wikipathway |
| | | | | hsa04062 | Chemokine signaling pathway | KEGG |
| | | | | hsa05144 | Malaria | KEGG |
| | | | | REACT_15518 | Transmembrane transport of small molecules | Reactome |
| | | | | hsa04610 | Complement and coagulation cascades | KEGG |
| | | | | hsa04961 | Endocrine and other factor-regulated calcium reabsorption | KEGG |
| | | | | hsa00053 | Ascorbate and aldarate metabolism | KEGG |
| | | | | hsa00860 | Porphyrin and chlorophyll metabolism | KEGG |
| | | | | REACT_111045 | Developmental Biology | Reactome |
| | | | | hsa04950 | Maturity onset diabetes of the young | KEGG |

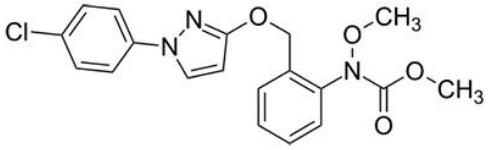



**Table 6**. CRC-linked direct proteins and pathways associated with biological testing candidate EC Pyraclostrobin, with components of interest labeled red.

In the course of our assessments of ECs, we accomplished the task of building cancer cellular network models using the DrugGenEx-Net method and making cellular and human level outcome predictions in applying these principles to both drugs and metabolites. In our published platform entitled DrugGenEx-Net, we assessed hundreds of pharmaceutical species, both FDA-approved and experimental drugs, by a proteochemometric method analogous to Tox-TMFS, to protein binding signature for each drug. The drug action space was then developed by annotation to protein-linked pathways, functions, and PPIs,

after which a gene expression-derived disease perturbation space was assessed for likeness to the drug space in order to determine therapeutic potential against neurodegenerative disorders and autoimmune diseases **[39]**. Furthermore, we assessed the disease-causing, therapeutic, and biomarker applications of metabolites in a platform entitled MSD-MAP (Multi Scale Disease-Metabolite Association) using the network principles of the work discussed here. MSD-MAP produces cancer cellular network models for endogenous and exogenous metabolites that may be central to the development of a given cancer, by extrapolating cellular activity linked to those metabolites from known and TMFS-predicted protein binding signatures. The resulting networks are assessed for coincidence with cancer-associated biological perturbation networks. This project has been accepted. We also published a review article discussing the current possibilities by which polypharmacology of a chemical compound can be modeled by computational means, and how this polypharmacology is employed to predict phenotypic outcomes such as adverse reactions, diseases, and therapeutic effects by combining principles of systems biology and network analysis **[40]**. In our development of Tox-TMFS, we developed a drug-repurposing variant on this proteochemometric model called RepurposeVS **[41]**. These works were important in building and applying the principles utilized in this work for predicting the mechanistic and phenotypic outcomes resulting from exposure to a chemical compound, particularly those utilized in **Major Tasks 1 and 2**.

**Major Task 3**. Prediction of EC Toxicity Using the CPTM Model

Subtask 1. Computation of the Intrinsic and Kinetic Properties of Environmental Chemicals.
Subtask 1 was listed to be completed in 8-12 months. This task is 60% complete.

Subtask 2. Our novel computational model CPTM will compute a potential toxicity score using the data generated from sub aims 1a-1c. Again this process is iterative as described above in Major Task 2.
Subtask 2 was listed to be completed in 8-16 months. This task is 60% complete.

Milestone(s) Achieved. The CPTM method simulates the interaction of chemicals, proteins and cells in physiological processes, and measure ECs toxicity in terms of a “toxicity score” (Zts) for colorectal

cancer in general and its pathways. During month 10 we expect to have first set of ECs to do biological testing.

Milestone Achieved was listed to be completed in 8-16 months. This task is 90% complete.

**Accomplishments**

**Major Task 3**. Prediction of EC Toxicity Using the CPTM Model

The Chemo-Phenotypic based Toxicity Measurement (CPTM) method (**Figure 1**) is a comprehensive integration of biological interactions, physicochemical descriptors, and pharmacokinetic (PK) properties, calculated in accordance with subtask 1 and all utilized to quantify an EC's toxicity risk through a "toxicity score" ($Z_{ts}$) under subtask 2. ECs with higher $Z_{ts}$ values are predicted to have higher overall human toxicity relative to ECs with lower $Z_{ts}$ values. An EC's toxicity score is calculated using the following equation **(2)**:

$$Z_{ts}(s_c) = w_k P(s_{pct} + s_{act}) + w_j N(s_{ppi}, s_{py}, s_{pf}) + w_i R(s_{\log P}, s_e, s_I, s_A, s_h, s_m, s_f, s_w) + w_o PK(s_M, s_{ABS})$$

where the "P" term represents the normalized chemical promiscuity score of a given EC ($s_c$), along with its designated weight ($\omega_k$ = 2). The promiscuity score is the sum of two factors: (1)$\sigma_{pct}$ represents predicted chemical-protein interactions using TOX-TMFS, and (2) $\sigma_{act}$ represents annotated chemical-protein interactions, which are obtained from the CTD. The "N" term represents the normalized cellular network perturbation score of an EC, which is derived from the total number of higher-order EC-biological effect associations (**Figure 1**) at the level of protein-protein interactions ($s_{ppi}$), pathways ($s_{py}$), and functions ($s_{pf}$), along with its designated weight ($\omega_j$ = 1). The "R" term represents the normalized score of the EC's chemical reactive index with ($w_i$=1). This index is composed of physicochemical properties associated with general mechanisms of toxicity according to the toxicology literature **[42]**. The properties included in the chemical reactive index are: log P ($s_{\log P}$), HOMO-LUMO energy band gap ($s_e$), ionization potential ($I$), electron affinity ($A$), chemical hardness ($s_h$), electronic chemical potential ($s_m$), electrophilic ($f^-$) Fukui index ($s_f$), and electrophilicity index ($s_w$). EC electron affinity ($A$) and ionization potentials ($I$) were obtained using QikProp and used to calculate the electrophilicity index ($w$), chemical hardness ($\eta$), and chemical potential ($\mu$):

$$w = \frac{\mu^2}{2\eta} \qquad \textbf{(3)}$$

where $m = -\frac{1}{2}(I + A)$ and $m = \frac{1}{2}(I - A)$. Jaguar [Schrodinger Inc. Computational modeling package**]** was used to calculate EC electrophilic Fukui indices and HOMO-LUMO gap energies.

The "PK" term represents the normalized pharmacokinetic score containing the number of potential metabolic reactions ($s_M$) the EC may undergo as well as its percent human oral absorption ($s_{ABS}$). Both properties are calculated using QikProp. We make the assumption that a toxicant's bioavailability is directly correlated with its toxicity, and that toxicants likely undergo metabolism to more reactive intermediates that give rise to toxicity. By way of this pharmacokinetic term, we seek to address the fact that many environmental toxicants will arrive at the colonic epithelium or other disease-relevant tissue in an altered state, i.e. as a metabolite. The metabolic reaction term, as detailed in Figure 7, is prospective in predicting the total number of reactions based on chemical structure and circumvents the issues arising from animal model genetic backgrounds (i.e. differential cytochrome P450 expression levels). While this may lead to some overestimated risks, we believe it is prudent to have this overestimation for a few toxicants than to underestimate the risk of the majority of toxicants.

The rigorous mathematical derivation of each CPTM input parameter is laid out in **Figure 7**, wherein colored headings correspond to major components of the CPTM calculation, such as the electrophilicity index column corresponding to **Equation 3**. Daidzein and Genistein, which have high toxicity scores, have intrinsic and kinetic qualities that match their diverse bioactivity and resulting toxicity, while fenaminosulf, which we predicted to have the lowest general toxicity for an EC, exhibits correspondingly mild calculated intrinsic and kinetic qualities.

| CASRN | EC Name | genes | NORM_GENE | ppi | NORM_PPI | pathways | NORM_PATH | functions | NORM_FUNCTION | BIOSCORE | NORM_BIOSCORE | f_minus_min | f_minus_max | f_plus_min | f_plus_max | HOMO | LUMO | GAPhartree | GAPev | NORM_HL_GAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 486-66-8 | Daidzein | 107 | 1 | 3016 | 1 | 1812 | 1 | 1466 | 1 | 4 | 1 | 0.0001 | 0.1263 | 0.0001 | 0.1526 | -0.176 | -0.065 | 0.11148 | 3.0334 | 0.48747184 |
| 446-72-0 | Genistein | 82 | 0.76635514 | 2582 | 0.8561008 | 1477 | 0.81512141 | 1180 | 0.804911323 | 3.2424887 | 0.810622168 | 0 | 0.1722 | 0.0001 | 0.1295 | -0.174 | -0.063 | 0.11131 | 3.0288 | 0.488253414 |
| 140-56-7 | Fenaminosulf | 6 | 0.056074766 | 163 | 0.0540451 | 201 | 0.11092715 | 47 | 0.032060027 | 0.253107 | 0.06327676 | 0 | 0.3828 | 0 | 0.255 | -0.164 | -0.129 | 0.03476 | 0.9458 | 0.840191256 |

| CASRN | EC Name | #NandO | #acid | #amide | #amidine | #amine | #in34 | #in56 | #metab | NORM_METAB_RXNS | PercentHumanOralAbsorption | NORM_ABSORPTION | #nonHatm | #noncon | #ringatoms | #rotor | #rtvFG | #stars | CNS | HumanOralAbsorption |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 486-66-8 | Daidzein | 5 | 0 | 0 | 0 | 0 | 0 | 16 | 3 | 0.375 | 76.685 | 0.76685 | 20 | 0 | 16 | 4 | 0 | 0 | -2 | 3 |
| 446-72-0 | Genistein | 4 | 0 | 0 | 0 | 0 | 0 | 16 | 2 | 0.25 | 83.786 | 0.83786 | 19 | 0 | 16 | 3 | 0 | 0 | -1 | 3 |
| 140-56-7 | Fenaminosulf | 6 | 1 | 0 | 0 | 0 | 0 | 6 | 2 | 0.25 | 60.258 | 0.60258 | 15 | 0 | 6 | 4 | 1 | 0 | -2 | 3 |

| CASRN | EC Name | RuleOfFive | RuleOfThree | Ionization penalty | Ionization penalty charging | Ionization penalty neutral | Gas Phase Energy | Potential Energy-MMFF94s | RMS Derivative-MMFF94s | Relative Potential Energy-MMFF94s | ACxDN^.5/SA | ClQPlogS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 486-66-8 | Daidzein | 0 | 0 | 0.0018 | 0 | 0.0018 | -953.7521465 | 238.5509644 | 0.000954362 | 0 | 0.0110738 | -4.101 |
| 446-72-0 | Genistein | 0 | 0 | 0.0012 | 0 | 0.0012 | -878.518446 | 232.3827667 | 0.043901496 | 0 | 0.0120425 | -3.71 |
| 140-56-7 | Fenaminosulf | 0 | 0 | 0.0074 | 0 | 0.0074 | -1099.444893 | 156.5025635 | 0.029301824 | 0 | 0.0162735 | -1.419 |

| CASRN | EC Name | EA(eV) (more negative, more toxicity) | NEW_EV_Range | NORM_EV | FISA | FOSA | IP(eV) (more negative, more toxicity) | NEW_IP_Range | NORM_IP | HARDNESS (0.5*(IP-EA)) | CHEMICAL POTENTIAL (0.5*(-IP+EA)) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 486-66-8 | Daidzein | 0.652 | 1.791 | 0.407416 | 185.7 | 0 | 9.034 | 3.413 | 0.686858523 | 4.191 | -4.843 |
| 446-72-0 | Genistein | 0.594 | 1.849 | 0.42061 | 147.2 | 0 | 8.957 | 3.49 | 0.702354599 | 4.1815 | -4.7755 |
| 140-56-7 | Fenaminosulf | 1.144 | 1.299 | 0.295496 | 174.6 | 155.03 | 8.417 | 4.03 | 0.811028376 | 3.6365 | -4.7805 |

| CASRN | EC Name | Electrophilicity Index (CP^2/2H) | Jm | PISA | PSA | PercentHumanOralAbsorption | QPPCaco | QPPMDCK | QPlogBB | QPlogHERG | QPlogKhsa | QPlogKp | QPlogPC16 | QPlogPoct | NEW_LogP_Range | NORM_LOGP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 486-66-8 | Daidzein | 2.798216297 | 0.084 | 293.25 | 98.21 | 76.685 | 171.921 | 73.762 | -1.302 | -4.987 | -0.111 | -3.524 | 9.508 | 1.662 | 5.299 | 0.49366499 |
| 446-72-0 | Genistein | 2.726940123 | 0.458 | 322.54 | 79.06 | 83.786 | 398.101 | 182.806 | -0.896 | -5.078 | -0.149 | -2.808 | 9.27 | 1.76 | 5.201 | 0.484535122 |
| 140-56-7 | Fenaminosulf | 3.142194452 | 0.724 | 129.77 | 92.87 | 60.258 | 55.408 | 28.098 | -1.252 | -2.628 | -1.049 | -3.897 | 7.319 | 0.36 | 6.601 | 0.614961804 |

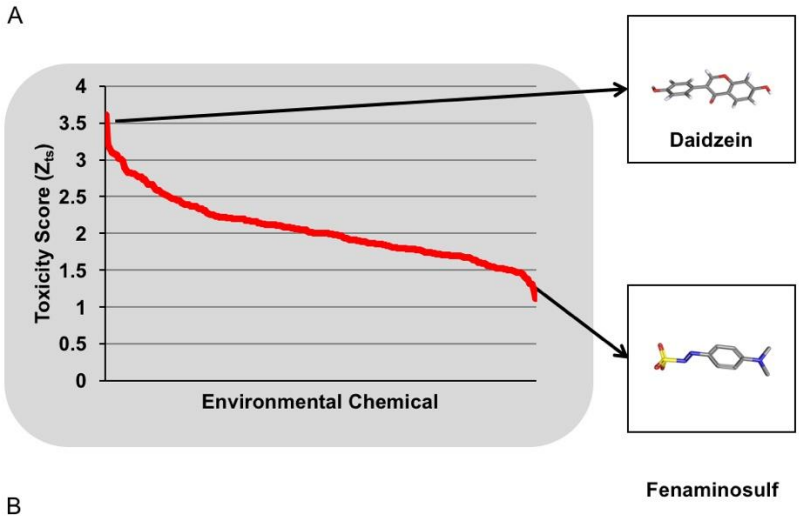
| CASRN | EC Name | QPlogPoct | QPlogPw | QPlogS | QPpolrz | SASA | SAamideO | SAfluorine | WPSA | acceptHB | dip^2/V | dipole | donorHB | glob | mol MW | volume |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 486-66-8 | Daidzein | 13.919 | 9.87 | -2.982 | 26.242 | 478.91 | 0 | 0 | 0 | 3.75 | 0.0139 | 3.347 | 2 | 0.873 | 270.241 | 803.7 |
| 446-72-0 | Genistein | 14.115 | 10.252 | -2.937 | 26.406 | 469.74 | 0 | 0 | 0 | 4 | 0.0219 | 4.145 | 2 | 0.876 | 254.242 | 784.75 |
| 140-56-7 | Fenaminosulf | 14.126 | 10.869 | -1.604 | 22.148 | 460.87 | 0 | 0 | 1.448 | 7.5 | 0.1182 | 9.356 | 1 | 0.859 | 229.253 | 740.4 |

**Figure 7**. Detailed sample computation of intrinsic and kinetic properties for ECs Daidzein, Genistein, and Fenaminosulf. Colored headings correspond to parameters directly incorporated into the CPTM general toxicity calculation.

We developed the Chemo-Phenotypic Based Toxicity Measurement (CPTM) to integrate with Tox-TMFS as a quantitative predictive tool for applied toxicology. CPTM is a metric defining potential clinical toxicity for a given EC molecule derived from predicted EC-target signatures via Tox-TMFS. It is a quantitation, reflected in terms of a "toxicity score" ($Z_{ts}$), of an EC's promiscuity of biological effects, chemical reactivity, potential bioavailability and potential metabolism (**Equation 2**). ECs exhibiting the highest $Z_{ts}$ are predicted to be the most toxic.

The molecules with the top two $Z_{ts}$ are daidzein and genistein **(Figure 8a; Table 7)**. Daidzein and genistein are isoflavones typified as endocrine disruptors and found in soy-based products. They have been shown to regulate the activity of many proteins and are thought to elicit multiple effects. The molecules with the lowest two $Z_{ts}$ are fenaminosulf, a fungicide, and benoxacor, a pesticide safetener, implying that they are relatively safe for human use. To validate CPTM, we queried the Hazardous Substance Data Bank (HSDB) for a subset of ECs to obtain relevant toxicological data (see Methods). A direct correlation between an EC's $Z_{ts}$ and the total number of observed toxic effects was found **(Figure 8b; Table 8)**. $Z_{ts}$ is therefore a reliable quantitative metric for predicting the extent of EC toxic effects.

**Table 7** reveals the top 30 ECs by CPTM-calculated $Z_{ts}$ (overall toxicity score), which is derived from the 958 ECs assessed by CPTM to accomplish the milestone for **Major Task 3**. From this assessment as well as the network analyses carried out under **Major Task 2**, we have determined candidates and commenced biological testing for the first set of ECs to be tested. This is further discussed in **Major Task 4**.



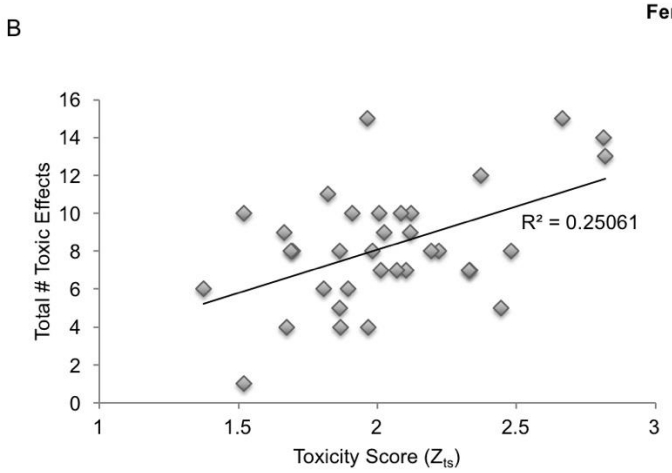




**Figure 8**. CPTM method arrangement of EC toxicity scores ($Z_{ts}$) and validation. (A) Waterfall plot of toxicity scores (Zts) for each environmental chemicals using the CPTM method. (B) Scatter plot of EC toxicity score against total number of toxic effects noted in the Hazardous Substance Data Bank (HSDB).

| CASRN | Environmental Chemical Name | $Z_{ts}$-Score | Promiscuity | F Minus Max | HOMO-LUMO Gap Normalized | EV Normalized | Ionization Potentials Normalized | Hardness | Chemical Potential | Electrophilicity Index | Log P Normalized | Metabolic Rxns Normalized | Absorption Normalized |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 486-66-8 | Daidzein | **3.62** | 1.00 | 0.13 | 0.49 | 0.41 | 0.69 | 4.19 | -4.84 | 2.80 | 0.49 | 0.38 | 0.77 |
| 446-72-0 | Genistein | **3.21** | 0.81 | 0.17 | 0.49 | 0.42 | 0.70 | 4.18 | -4.78 | 2.73 | 0.48 | 0.25 | 0.84 |
| 162706-14-1 | 5-[2-methyl-3-(pyridin-3-yl)-1H-indol… | **3.16** | 0.74 | 0.67 | 1.13 | 0.52 | 0.85 | 4.03 | -4.21 | 2.20 | 0.25 | 0.50 | 0.90 |
| 491-80-5 | Biochanin A | **3.10** | 0.74 | 0.13 | 0.49 | 0.41 | 0.70 | 4.17 | -4.80 | 2.76 | 0.42 | 0.38 | 0.91 |
| 50-65-7 | Niclosamide | **3.09** | 0.81 | 0.18 | 0.61 | 0.25 | 0.60 | 4.07 | -5.42 | 3.62 | 0.36 | 0.25 | 0.87 |
| 520-36-5 | Apigenin | **3.07** | 0.73 | 0.28 | 0.48 | 0.37 | 0.64 | 4.22 | -5.03 | 3.00 | 0.50 | 0.38 | 0.74 |
| 4065-45-6 | Sulisobenzone | **3.07** | 0.74 | 0.33 | 0.87 | 0.45 | 0.58 | 4.56 | -5.02 | 2.76 | 0.50 | 0.25 | 0.63 |
| 117-39-5 | Quercetin | **3.01** | 0.63 | 0.20 | 0.50 | 0.39 | 0.68 | 4.18 | -4.91 | 2.89 | 0.61 | 0.63 | 0.52 |
| 50892-23-4 | Pirinixic acid | **3.01** | 0.64 | 0.51 | 0.98 | 0.42 | 0.71 | 4.16 | -4.76 | 2.72 | 0.32 | 0.63 | 0.84 |
| 2528-16-7 | Mono-benzyl phthalate | **2.98** | 0.73 | 0.51 | 0.92 | 0.42 | 0.55 | 4.56 | -5.14 | 2.90 | 0.36 | 0.13 | 0.85 |
| 4291-63-8 | Cladribine | **2.89** | 0.57 | 0.22 | 0.36 | 0.48 | 0.78 | 4.11 | -4.46 | 2.41 | 0.70 | 0.38 | 0.60 |
| 69-74-9 | Cytarabine hydrochloride | **2.86** | 0.46 | 0.62 | 0.38 | 0.55 | 0.68 | 4.52 | -4.53 | 2.27 | 0.84 | 0.50 | 0.42 |
| 149877-41-8 | Bifenazate | **2.82** | 0.68 | 0.33 | 0.49 | 0.52 | 0.88 | 3.95 | -4.10 | 2.13 | 0.30 | 0.13 | 1.00 |
| 349495-42-7 | 1-ethyl-5-(4-hydroxyph… | **2.82** | 0.59 | 0.14 | 0.51 | 0.39 | 0.71 | 4.10 | -4.83 | 2.85 | 0.46 | 0.38 | 0.87 |
| 654055-01-3 | Morin hydrate | **2.82** | 0.54 | 0.11 | 0.52 | 0.34 | 0.65 | 4.12 | -5.08 | 3.13 | 0.61 | 0.63 | 0.54 |
| 27314-13-2 | Norflurazon | **2.81** | 0.59 | 0.27 | 0.50 | 0.34 | 0.71 | 3.97 | -4.93 | 3.06 | 0.32 | 0.50 | 1.00 |
| 60168-88-9 | Fenarimol | **2.80** | 0.58 | 0.18 | 0.40 | 0.38 | 0.60 | 4.34 | -5.11 | 3.00 | 0.28 | 0.63 | 1.00 |
| 493-52-7 | Methyl red | **2.78** | 0.56 | 0.56 | 0.82 | 0.48 | 0.91 | 3.80 | -4.14 | 2.25 | 0.37 | 0.38 | 0.87 |
| 194098-25-4 | N-{4-[(methylamino)methyl]phenyl… | **2.77** | 0.49 | 0.70 | 0.72 | 0.57 | 0.81 | 4.24 | -4.20 | 2.08 | 0.44 | 0.50 | 0.80 |
| 22839-47-0 | Aspartame | **2.77** | 0.38 | 0.80 | 0.72 | 0.55 | 0.62 | 4.68 | -4.70 | 2.36 | 0.74 | 0.75 | 0.36 |
| 78473-71-9 | Enterolactone | **2.74** | 0.51 | 0.14 | 0.28 | 0.55 | 0.63 | 4.66 | -4.67 | 2.34 | 0.42 | 0.63 | 0.85 |
| 10161-33-8 | 17beta-Trenbolone | **2.73** | 0.42 | 0.72 | 0.59 | 0.37 | 0.72 | 4.03 | -4.84 | 2.90 | 0.38 | 0.75 | 1.00 |
| 81335-37-7 | Imazaquin | **2.70** | 0.61 | 0.63 | 0.88 | 0.21 | 0.64 | 3.88 | -5.41 | 3.77 | 0.44 | 0.00 | 0.76 |
| 474-86-2 | Equilin | **2.67** | 0.45 | 0.20 | 0.40 | 0.59 | 0.70 | 4.58 | -4.41 | 2.13 | 0.35 | 0.63 | 1.00 |
| 59-40-5 | Sulfaquinoxaline | **2.66** | 0.51 | 0.36 | 0.63 | 0.24 | 0.73 | 3.71 | -5.09 | 3.49 | 0.57 | 0.25 | 0.72 |
| 188425-85-6 | Boscalid | **2.66** | 0.64 | 0.21 | 0.44 | 0.38 | 0.70 | 4.09 | -4.87 | 2.89 | 0.20 | 0.25 | 1.00 |
| 139149-55-6 | 1-[(3S)-6-(benzyloxy)-2,3-dihydro… | **2.63** | 0.44 | 0.22 | 0.39 | 0.59 | 0.71 | 4.54 | -4.38 | 2.11 | 0.51 | 0.50 | 0.80 |
| 94-91-7 | N,N'-Disalicylidene-1,2-diaminopropane | **2.59** | 0.53 | 0.09 | 0.44 | 0.50 | 0.68 | 4.42 | -4.68 | 2.48 | 0.34 | 0.25 | 1.00 |
| 604-75-1 | Oxazepam | **2.59** | 0.53 | 0.30 | 0.50 | 0.38 | 0.67 | 4.17 | -4.94 | 2.93 | 0.44 | 0.13 | 0.90 |
| 13684-63-4 | Phenmedipham | **2.58** | 0.49 | 0.23 | 0.29 | 0.54 | 0.72 | 4.42 | -4.47 | 2.26 | 0.37 | 0.38 | 0.96 |

**Table 7**. CPTM-derived Top 30 ECs by $Z_{ts}$ toxicity score.

As the $Z_{ts}$ of genistein, a frequently encountered molecule, is among the highest, this EC was analyzed in greater detail. Considering all predicted genistein-target associations, enriched pathways and functions include those involved in cancers and cardiovascular illnesses. Pathways include PPAR signaling (hsa03320), T cell receptor signaling pathway (hsa04660) and pathways in cancer (hsa05200). Functions include negative regulation of cholesterol storage (GO:0010887), lipid storage (GO:0010888) and foam cell differentiation (GO:0010745), all of which promote cardiovascular health, as well as regulation of apoptosis (GO:0042981) and cell proliferation (GO:0042127), which are cancer-mediating functions.

These predicted associations are also supported by experimental evidence **[43]**. Similar analysis can be performed for any other ECs in the dataset.

| Environmental Chemical | Toxicity Score (Zts) | Total Toxic Effects | Subchronic | Chronic | Carcinogenicity | Developmental/Reproductive | Genotoxicity |
|---|---|---|---|---|---|---|---|
| Bifenazate | 2.81869 | 13 | 5 | 5 | 0 | 3 | 0 |
| Norflurazon | 2.81408 | 14 | 4 | 4 | 3 | 3 | 0 |
| Boscalid | 2.66387 | 15 | 6 | 3 | 1 | 5 | 0 |
| Imazapic | 2.48170 | 8 | 0 | 8 | 0 | 0 | 0 |
| Zoxamide | 2.44380 | 5 | 3 | 1 | 0 | 1 | 0 |
| Cycloheximide | 2.37212 | 12 | 10 | 0 | 0 | 2 | 0 |
| Fluroxypyr | 2.33090 | 7 | 5 | 2 | 0 | 0 | 0 |
| Auramine | 2.32929 | 7 | 0 | 0 | 0 | 3 | 4 |
| Phenolphthalein | 2.22190 | 8 | 1 | 0 | 5 | 0 | 2 |
| Prohexadione | 2.19326 | 8 | 3 | 3 | 1 | 1 | 0 |
| Triticonazole | 2.12015 | 10 | 6 | 3 | 1 | 0 | 0 |
| Butralin | 2.11813 | 9 | 6 | 0 | 1 | 1 | 1 |
| Fenoxycarb | 2.10192 | 7 | 3 | 1 | 3 | 0 | 0 |
| Pendimethalin | 2.08564 | 10 | 4 | 3 | 1 | 0 | 2 |
| Myclobutanil | 2.06824 | 7 | 4 | 1 | 0 | 2 | 0 |
| Triadimefon | 2.02297 | 9 | 4 | 0 | 2 | 2 | 1 |
| Quinoxyfen | 2.01315 | 7 | 4 | 3 | 0 | 0 | 0 |
| Isoxaben | 2.00681 | 10 | 0 | 4 | 3 | 2 | 1 |
| 2,4,5-Trichlorophenoxyacetic acid | 1.98101 | 8 | 6 | 2 | 0 | 0 | 0 |
| Triclopyr | 1.96640 | 4 | 1 | 3 | 0 | 0 | 0 |
| Amitraz | 1.96346 | 15 | 7 | 1 | 1 | 6 | 0 |
| MGK-264 | 1.90823 | 10 | 4 | 1 | 3 | 2 | 0 |
| Dichlorophen | 1.89372 | 6 | 5 | 0 | 0 | 0 | 1 |
| Fenhexamid | 1.86599 | 4 | 2 | 2 | 0 | 0 | 0 |
| Clodinafop-propargyl | 1.86494 | 8 | 3 | 0 | 3 | 2 | 0 |
| Sethoxydim | 1.86358 | 5 | 2 | 1 | 0 | 2 | 0 |
| Oleic acid | 1.82216 | 11 | 5 | 0 | 4 | 1 | 1 |
| Pyriproxyfen | 1.80500 | 6 | 2 | 3 | 0 | 1 | 0 |
| Triglycidyl isocyanurate | 1.69468 | 8 | 5 | 0 | 0 | 0 | 3 |
| Octadecanoic acid | 1.68960 | 8 | 3 | 1 | 4 | 0 | 0 |
| Bisphenol AF | 1.67427 | 4 | 2 | 0 | 0 | 2 | 0 |
| Imidacloprid | 1.66371 | 9 | 5 | 2 | 1 | 1 | 0 |
| Nitrofen | 1.52059 | 10 | 1 | 0 | 2 | 5 | 2 |
| Laurocapram | 1.52022 | 1 | 1 | 0 | 0 | 0 | 0 |
| Halofenozide | 1.37465 | 6 | 3 | 0 | 0 | 2 | 1 |

**Table 8**. Comparison of CPTM toxicity scores (Zts) with established toxic effects in literature. Known EC effects were tabulated from the Hazardous Substance Data Bank (HSDB). The total number of unique toxicity effects were compared to the EC's toxicity score (Zts). Color circles represent relative toxicity category, with red being most toxic and green being least toxic.

**Specific Aim 2. VDR and β-catenin Pathways in CRC**

**Major Task 4**. Biological testing of candidate Environmental Chemicals.
Subtask 1. Perform vitamin D receptor binding studies on the small number of ECs (selected from Aim 1) using reporter and surface plasmon resonance assays.
Subtask 1 was listed to be completed in 3-24 months. This task is 15% complete.

Subtask 2. Perform Wnt/Beta-Catenin pathway activation studies using the reporter assays, and for some ECs which may disrupt VDR/beta-catenin interactions which we could test by mammalian two hybrid assay.
Subtask 2 was listed to be completed in 4-24 months. This task is 15% complete.

Subtask 3. We will perform cell viability and apoptosis measurements on selected ECs. We expect to test around 20 of the highest toxicity ECs predicted by the CPTM model in addition to top hits came from earlier experiments. Cells: We will use Keratinocytes, Fibroblasts, and Airway epithelial cells.

Subtask 3 was listed to be completed in 4-24 months. This task is 15% complete.

Milestone(s) Achieved: Milestone Achieved was listed to be completed in 4-24 months. This task is 15% complete.

**Accomplishments**

**Major Task 4**. Biological testing of candidate Environmental Chemicals.

Under **subtask 1**, reporter assays for VDR binding studies are under way for the selected EC N,N'-disalicylidene-1,2-diaminopropane. 19 other ECs predicted to bind to VDR/pathway components or other CRC target pathway components have been purchased and handled for preparation of in-vitro assays. Since this process is iterative, **subtask 1** may be completed by testing for binding against other proteins for which Tox-TMFS predicted interactions for key ECs, with preference to those relevant to CRC.

In accordance with **Subtask 2**, Wnt/Beta-Catenin pathway activation studies using reporter assays as well as VDR/Beta-catenin disruption by mammalian two hybrid assay are at the preliminary stages for the three chemicals, N,N'-disalicylidene-1,2-diaminopropane, Pyraclostrobin, and Paclobutrazol, described in **Major Task 2**. 20 ECs scoring highly in terms of overall toxicity (Zts) have been chosen for **subtask 3**, but cell viability and apoptosis has not commenced.

**What do you plan to do during the next reporting period to accomplish the goals?**

The next reporting period will entail a completion of **Major Task 1 subtask 1**, wherein ECs for which we have information on chemical structure will be processed for computation of intrinsic and kinetic characteristics, with assessment of protein binding by the Tox-TMFS procedure. This will equally accomplish the milestone for **Major Task 1**, and all characterized chemicals will have full EC-protein association networks. Upon this characterization, cancer cellular network models will be annotated for these ECs, and incorporated into the assessments for biological testing for mechanistic perturbation of VDR, TGF beta, and Wnt/Beta-catenin signaling pathways. Completion of these chemical parameters will lastly allow for the conclusion of our CPTM procedure, and we will have obtained ranked potential toxicity prediction scores for the ECs. In this process, ECs of interest will emerge for which we will assess cell viability and apoptosis measurements according to **subtask 3 in Major Task 4**. As was

described in our accomplishments, biological testing for candidate ECs at the levels of protein binding studies using reporter and surface plasmon resonance assays, wnt/beta-catenin pathway activation studies, VDR/beta-catenin interaction disruption studies using mammalian two hybrid assays, and CPTM-linked toxicity assays are in the preliminary stages, and this full panel of toxicology testing will be carried out in the upcoming months.

In accordance with **Major Task 1**, we will continue the physicochemical assessment and building protein biding signatures for ECs using Tox-TMFS. Using the CPTM model, the general toxicity for these ECs will equally be assessed from kinetic and intrinsic properties of these compounds. The primary task from this point is to continue our biological testing at the levels of protein binding, cancer-associated pathway perturbation assessments, and in vitro assays for measures of general toxicity. These biological assessments will be utilized to validate our phenotypic predictions, which we arrive to solely from the starting point of chemical structure, and will also provide novel findings of biological perturbation and toxicity activity of key ECs, having implications for the ways in which EC exposure toxicity is measured and preemptively utilized to make policy and other decisions.

**References**

[1] Dakshanamurthy, S., Issa, N. T., Assefnia, S., Seshasayee, A., Peters, O. J., Madhavan, S., ... & Byers, S. W. (2012). Predicting new indications for approved drugs using a proteochemometric method. *Journal of medicinal chemistry*, *55*(15), 6832-6848.
[2] Small-Molecule Drug Discovery Suite 2013-3: Glide, version 6.1 and QikProp, version 3.8, Schrödinger, LLC, New York, NY, 2013.
[3] Kahraman, A.; Morris, R.; Laskowski, R.; Thornton, J. Shape Variation in Protein Binding Pockets and Their Ligands. *J. Mol. Biol.* **2007**, *368*, 283–301.
[4] Davis AP, Murphy CG, Johnson R, Lay JM, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL, Rosenstein MC, Wiegers TC, Mattingly CJ. The Comparative Toxicogenomics Database: update 2013. Nucleic Acids Res. 2013 Jan 1;41(D1):D1104-14.
[5] Wishart, D., *et al*. T3DB: the toxic exposome database. Nucleic Acids Res. 2015 Jan;43(Database issue):D928-34.
[6] Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, et al. (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res D1: D808-D815.
[7] The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. Nat Genet 25: 25-29.
[8] Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, et al. (2009) AmiGO: online access to ontology and annotation data. Bioinformatics 25: 288-289.
[9] Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc 4: 44-57.
[10] Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1-13.
[11] Kanehisa M, Goto S (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28: 27-30.
[12] Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D428-32.
[13] Hewett *et al*. PharmGKB: the Pharmacogenetics Knowledge Base. Nucleic Acids Res. 2002 Jan 1;30(1):163-5.
[14] Kandasamy K *et al*. NetPath: a public resource of curated signal transduction pathways. Genome Biol. 2010 Jan 12;11(1):R3.
[15] Nishimura D. BioCarta. Biotech Software & Internet Report. July 2004, 2(3): 117-120.
[16] Pico *et al*, WikiPathways: pathway editing for the people. PLoS Biol. 2008 Jul 22;6(7):e184.
[17] Schaefer *et al*. PID: the Pathway Interaction Database. Nucleic Acids Res. 2009 Jan;37(Database issue):D674-9.
[18] Kamburov *et al*. ConsensusPathDB: toward a more complete picture of cell biology. Nucleic Acids Res. 2011 Jan;39(Database issue):D712-7.
[19] Shannon *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003 Nov;13(11):2498-504.
[20] Hamosh *et al*. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D514-7.

[21] Lepri, Sandra Regina, Leonardo Campos Zanelatto, Patrícia Benites Gonçalves da Silva, Daniele Sartori, Lucia Regina Ribeiro, and Mario Sergio Mantovani. "The effects of genistein and daidzein on cell proliferation kinetics in HT29 colon cancer cells: the expression of CTNNBIP1 (β-catenin), APC (adenomatous polyposis coli) and BIRC5 (survivin)." Human cell 27, no. 2 (2014): 78-84.
[22] McClellan, Jamie L., Jennifer L. Steiner, Reilly T. Enos, and E. Angela Murphy. "Effects of quercetin in a mouse model of colitis associated colon cancer." The FASEB Journal 27 (2013): 235-3.
[23] Gupta, Rajnish A., and Raymond N. DuBois. "Colorectal cancer prevention and treatment by inhibition of cyclooxygenase-2." Nature Reviews Cancer 1, no. 1 (2001): 11-21.
[24] Bobek, Vladimir, and Josef Kovařík. "Antitumor and antimetastatic effect of warfarin and heparins." Biomedicine & Pharmacotherapy 58, no. 4 (2004): 213-219.
[25] Coogan, Patricia F., Lynn Rosenberg, Julie R. Palmer, Brian L. Strom, Ann G. Zauber, Paul D. Stolley, and Samuel Shapiro. "Phenolphthalein laxatives and risk of cancer." Journal of the National Cancer Institute 92, no. 23 (2000): 1943-1944.
[26] Report on Carcinogens, Twelfth Edition (2011).
[27] Lai, Yu-Hsien, and Te-Chao Fang. "The Pleiotropic Effect of Vitamin D." ISRN nephrology 2013 (2013).
[28] Froicu M, Cantorna MT. Vitamin D and the vitamin D receptor are critical for control of the innate immune response to colonic injury. BMC Immunol. 2007; 8:5.
[29] Laragione, Teresina, Anish Shah, and Pércio S. Gulko. "The vitamin D receptor regulates rheumatoid arthritis synovial fibroblast invasion and morphology."Molecular Medicine 18, no. 1 (2012): 194.
[30] Trémezaygues, Lea, and Jörg Reichrath. "Vitamin D analogs in the treatment of psoriasis: Where are we standing and where will we be going?." Dermato-endocrinology 3, no. 3 (2011): 180.
[31] Lambert, Joshua, Robert Dorr, and Barbara Timmermann. "Nordihydroguaiaretic acid: a review of its numerous and varied biological activities." Pharmaceutical biology 42, no. 2 (2004): 149-158.
[32] Sahu, Saura C., Dennis I. Ruggles, and Michael W. O'Donnell. "Prooxidant activity and toxicity of nordihydroguaiaretic acid in clone-9 rat hepatocyte cultures." Food and chemical toxicology 44, no. 10 (2006): 1751-1757.
[33] Teske, Kelly, Premchendar Nandhikonda, Jonathan W. Bogart, Belaynesh Feleke, Preetpal Sidhu, Nina Y. Yuan, Joshua Preston et al. "Identification of VDR Antagonists among Nuclear Receptor Ligands Using Virtual Screening." (2014).
[34] Sverdrup, Berit, Henrik Källberg, Camilla Bengtsson, Ingvar Lundberg, Leonid Padyukov, Lars Alfredsson, and Lars Klareskog. "Association between occupational exposure to mineral oil and rheumatoid arthritis: results from the Swedish EIRA case–control study." Arthritis research & therapy 7, no. 6 (2005): R1296.
[35] Wolf, Ronni, Moshe Movshowitz, and Sarah Brenner. "Supplemental tests in the evaluation of occupational hand dermatitis in soldiers." International journal of dermatology 35, no. 3 (1996): 173-176.
[36] Mehlman, MA. "Dangerous and cancer-causing properties of products and chemicals in the oil refining and petrochemical industry. VIII. Health effects of motor fuels: carcinogenicity of gasoline--scientific update." Environ Res 59, no. 1 (1992), 238-249.
[37] Zhao, Yingxu, Anusha Krishnadasan, Nola Kennedy, Hal Morgenstern, and Beate Ritz. "Estimated effects of solvents and mineral oils on cancer incidence and mortality in a cohort of aerospace workers." American journal of industrial medicine 48, no. 4 (2005): 249-258.

[38] Malloy, Elizabeth J., Katie L. Miller, and Ellen A. Eisen. "Rectal cancer and exposure to metalworking fluids in the automobile manufacturing industry."Occupational and environmental medicine 64, no. 4 (2007): 244-249.
[39] Issa *et al*. DrugGenEx-Net: a novel computational platform for systems pharmacology and gene expression-based drug repurposing. BMC Bioinformatics. 2016 May 5;17(1):202.
[40] Wathieu *et al*. Harnessing Polypharmacology with Computer-Aided Drug Design and Systems Biology. Curr Pharm Des. 2016;22(21):3097-108.
[41] Issa NT, Peters OJ, Byers SW, Dakshanamurthy S. RepurposeVS: A Drug Repurposing-Focused Computational Method for Accurate Drug-Target Signature Predictions. Comb Chem High Throughput Screen. 2015;18(8):784-94.
[42] Tsakovska, I., Lessigiarska, I., Netzeva, T., & Worth, A. P. (2008). A mini review of mammalian toxicity (Q) SAR models. QSAR & Combinatorial Science, 27(1), 41-48.
[43] Schrödinger Release 2013-3: Jaguar, version 8.0, Schrödinger, LLC, New York, NY, 2013.

## 4. Impact……..

**What was the impact on the development of the principal discipline(s) of the project?**

ECs include toxins in air, water, food, and soil. Continuous exposure from any of these sources could put military personnel at increased risk of cancer. For example “agent orange,” a dioxin derivative (plant herbicide) used in Vietnam, causes cancer specifically in the kidney, colon and blood. In this project, by way of the Tox-TMFS and DrugGenEx-Net procedures, we reveal molecular targets and pathways relevant to these and other cancer types. The CPTM model, by contrast, will reveal uncharacterized connections between ECs and cancer types described in this FOA, thus predicting which chemicals may be risk factors. The resulting outcomes can be used as a foundation for future research to understand the mechanisms of initiation, and progression of cancer resulting from toxin exposure.

Our work offers contributions and has significant implications for the way in which toxicity is assessed on a high throughput and automated scale, for chemical species of all classes. While the validated EC-target space thus far is relatively small, many of the predicted ECs have not yet been studied for CRC. This study will lay the ground work to reveal potential risks and mechanisms whereby such chemicals cause cancer. Moreover, CRC is a late-onset adult cancer with an important time-course consideration for human epidemiological evaluation of ECs. Tox-TMFS provides biologically plausible hypotheses for small-molecule ECs to focus toxicology studies for CRC as well as investigate environmental molecules such as phytoestrogens to prevent or treat CRC. Our CPTM model offers a new approach for multi-dimensional toxicological characterization stemming only from the chemical structure of a given EC. In particular, it helps to justify the prediction of phenotypic consequences of toxicants, and other chemicals using network characteristics and systems biological understanding in an efficient manner, ultimately decreasing the need for human capital and financial resources.

Tox-TMFS with the CPTM method is the first platform of its kind in the toxicological sciences rooted in chemistry and systems biopharmacology. As the EC space grows exponentially with new commercial materials, environmental waste products, and pharmaceuticals, our platform is positioned to streamline the comprehensive assessment of these chemicals for focused subsequent assays and increased efficiency

in toxicology. Thus, using synergistic effects of chemical and cellular components to bring about toxicity and cancer risk is an innovative approach in this field, and will provide insights into mechanisms of cancers initiation and contribute to the treatment and prevention paradigm. Our method also has the potential to be utilized for diseases apart from oncologic diseases, such as neurodegenerative diseases, autoimmune diseases, and others.

**What was the impact on other disciplines?**

Having a robust computational platform that allows for accurate and rapid toxicity assessments of chemicals commonly encountered in the environment benefits public health by increasing the efficiency of toxicity studies and guiding healthcare professionals in noticing the gamut of clinical presentations arising from exposure. Such a platform would not only help reassess already-known ECs given the integration of the currently rich corpus of biomedical data, but would also benefit future industrial endeavors as preconceived knowledge of potential toxicity would guide the production of new chemicals with safety at the forefront. Tox-TMFS combined with CPTM is a step in this direction. This project has additional in the realm of toxicology assessments for therapeutics and mechanistic understanding of diseases in that context. Perturbations in the biological signal network either by up regulation or down regulation of target proteins will act as indicators of future cancer. Pharmacological interventions with existing drugs such as drug repurposing on the characterized perturbation signals could serve to both prevent and treat cancer.

**What was the impact on society beyond science and technology?**

Outside of pure scientific and technological advancements, CPTM a quantitative tool, determining the target signatures of ECs by using many reference chemicals with known biology as chemical probes. It provides the framework to better understand the potential for cancer risk of ECs with incomplete toxicity information, and may ultimately serve as a tool to predict unwanted biological activity, and provide a potential prevention as well as chemical design guidelines prior to exposure. The ability to assess the impact of militarily relevant environmental carcinogens, for which little toxicity risk data are available, for cancer or to foresee such effects during the early stages of chemical development and use, before

potential exposure occurs, is an urgent need. The proposed CPTM (chemo-phenotypic based toxicity measurement) technology would serve as a toxicity guideline for exposure of service members and their families to relevant chemicals.

Our work therefore has the potential for profound impact in precisely gearing regulatory policies regarding the limitation of military and civilian exposure to environmental chemicals as a waste or as a resource for industrial, chemical, or other application, and on the use of such chemicals as biological weapons. In addition, having a more complete framework for how to anticipate and treat the pathological conditions resulting from specific chemical exposures can inform decision making at the level of clinical care and cultural behaviors exhibited by populations which may influence the level of exposure to a given chemical.

## 5. Changes/Problems……..

**Changes in approach and reasons for change**

Nothing to report**.**

**Actual or anticipated problems or delays and actions or plans to resolve them**

**Problems**

Some of the computational tasks, and biological testing were delayed because of significant delay in recruiting suitable computational biologist who has expertise in computational biology modeling, and biology. There is an additional delay in acquiring chemicals and reagents from commercial vendors because of availability issues, even some chemicals are not available commercially.

**Actions and Plans to Resolve**

Because of post doc. recruitment delay, therefore in this reporting period, additional PI effort were used, a research assistant from other project, and research intern were involved in this project. The personnel who has appropriate expertise is underway actively to avoid any delay. We are currently coordinating with suppliers to ensure that correct chemicals and reagents are prepared for use in our biological testing of candidates. The chemicals which are not available will be skipped, and to avoid further delay next highest ranked commercially available chemicals are being ordered.

## 6. Products…….

**Journal Publications**

1. Issa NT, Peters OJ, Byers SW, Dakshanamurthy S. RepurposeVS: A Drug Repurposing-Focused Computational Method for Accurate Drug-Target Signature Predictions. Comb Chem High Throughput Screen, 2015, 18(8):784-94. [Published]. *Acknowledgement of federal support (yes).*

2. Wathieu H, Issa NT, Byers SW, Dakshanamurthy S. Harnessing Polypharmacology with Computer-Aided Drug Design and Systems Biology. Curr Pharm Des, 2016 Feb 24. 22(21): 3097-108. [Published]. *Acknowledgement of federal support (yes).*

3. Issa NT, Kruger J, Wathieu H, Raja R, Byers SW, and Dakshanamurthy S. DrugGenEx-Net: A Novel Computational Platform for Systems Pharmacology and Gene Expression-Based Drug Repurposing. BMC Bioinformatics, 2016 May 5, 17:202. [Published]. *Acknowledgement of federal support (yes).*

4. Wathieu H, Issa NT, Mohandoss M, Byers SW, Dakshanamurthy S. MSD-MAP: A Network-based Systems Biology Platform for Predicting Disease-Metabolite Links. Comb Chem High Throughput Screen, 2016. [Accepted]. *Acknowledgement of federal support (yes).*

**Website(s) or other Internet site(s)**

Nothing to report**.**

**Technologies or Techniques**.

In the course of developing our core methodologies for EC biological assessment, called Tox-TMFS, RepurposeVS, MSD-MAP, and DrugGenEx-Net, we produced and published two technologies consisting of goal-specific implementations of these methodologies. DrugGenEx-Net predicts protein binding signatures for FDA-approved drugs and produces extrapolated drug action networks by way of

multi scale annotation of those interacting proteins. Similar to our cancer cellular perturbation network strategy in this work, DrugGenEx-Net identifies drugs which coincide most significantly with biological factors perturbed by neurodegenerative and autoimmune pathologies, thereby presenting opportunities for drug repurposing. It has been published along with a tutorial that describes its stepwise implementation. MSD-MAP, the Multi Scale Disease-Metabolite Association Platform, produces similar cellular action networks for metabolites rather than drugs or ECs, and infers disease causage, biomarker potential, or therapeutic potential for those metabolites in the context of three cancers. MSD-MAP utilizes public data and therefore reproduction of the technology can be facilitated by consulting with the methods described in the publication. In addition, the CPTM method, as described previously, is the first computational technique to produce reliable general toxicity predictions based on the intrinsic and kinetic properties of a chemical, including those yielded by Tox-TMFS and DrugGenEx-Net. The resulting data, and technique procedure will be shared through publications in the journal.

**Other Products**

Biological components and relationship data generated by our CPTM, cellular perturbation networks will be deposited to the Chemical Interactome Cellular Network Interface (CICNI), an online database and visualization tool for systems biological networks developed in our laboratory. This is a user-friendly platform is valuable as a research tool for precision pharmacology and toxicology. It promotes the shared goal of reducing costs for achieving a mechanistic understanding and producing hypotheses in terms of prevention, diagnosis, and treatment of diseases and other clinically relevant phenotypes with an *in-silico* precision approach.

Utilizing our CPTM method in conjunction with protein interactome prediction and cancer cellular network models, we are identifying and testing various ECs as having potential general toxicity and disease-specific toxicity. We predicted a common motor oil additive called N,N'-disalicylidene-1,2-diaminopropane, for example, to have cellular perturbation characteristics strongly coinciding with Colorectal Cancer and its known causative mechanisms. The CPTM tool has implications for the prevention and mechanistic understanding of key pathologies and how ECs are complicit in them, and thus contributes to the public good.

## 7. PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS……

**What individuals have worked on the project?**

1.

Name: Sivanesan Dakshanamurthy

Project Role: PI

Nearest person month worked: 5

Contribution to Project: Dr. Dakshanamurthy has performed computational work in the development, execution of CPTM model, simulations, data analysis, and database integration. He also overseeing the overall project goals and plan, design, execute Aim 1 and Aim 2 of the project.

Funding Support: In addition to the funding support from this DoD award, effort from Georgetown University-Lombardi Cancer Center Institutional support funds has also been used.

2.

Name: Stephen W Byers

Project Role: Co-PI

Nearest person month worked: 0.72

Contribution to Project: Dr. Byers has involving in the plan, and design, and execute of VDR-pathway components reporter assays, Aim 2 of the project.

Funding Support: This DoD Award

3.

Name: Henri Wathieu

Project Role: Research Assistant

Nearest person month worked: 3.4

Contribution to Project: Mr. Wathieu performed, data simulations, data analysis, data integration involving CPTM model, Aim 1of the project.

Funding Support: Effort from Georgetown University-Lombardi Cancer Center Institutional support funds has been used.

4.

Name: Abiola Ojo

Project Role: Research Intern

Nearest person month worked: 1.4

Contribution to Project: Mr. Abiola has performed data analysis, data curation data integration involving CPTM model, Aim 1 of the project.

Funding Support: Effort from SOAR-MHHD Research Internship Program.

**Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?**

*Nothing to Report.*

**What other organizations were involved as partners?**

*Nothing to Report.*

## 8. Special Reporting Requirements……..

*Nothing to Report.*

# 9. Appendices.....

**APPENDIX A**

Issa NT, Peters OJ, Byers SW, Dakshanamurthy S. RepurposeVS: A Drug Repurposing-Focused Computational Method for Accurate Drug-Target Signature Predictions. Comb Chem High Throughput Screen, 2015, 18(8):784-94. [Published]. *Acknowledgement of federal support (yes).*

**APPENDIX B**

Wathieu H, Issa NT, Byers SW, Dakshanamurthy S. Harnessing Polypharmacology with Computer-Aided Drug Design and Systems Biology. Curr Pharm Des, 2016 Feb 24. 22(21): 3097-108. [Published]. *Acknowledgement of federal support (yes)*

**APPENDIX C**

Issa NT, Kruger J, Wathieu H, Raja R, Byers SW, and Dakshanamurthy S. DrugGenEx-Net: A Novel Computational Platform for Systems Pharmacology and Gene Expression-Based Drug Repurposing. BMC Bioinformatics, 2016 May 5, 17:202. [Published]. *Acknowledgement of federal support (yes)*

**APPENDIX D**

Wathieu H, Issa NT, Mohandoss M, Byers SW, Dakshanamurthy S. MSD-MAP: A Network-based Systems Biology Platform for Predicting Disease-Metabolite Links. Comb Chem High Throughput Screen, 2016. [Accepted]. *Acknowledgement of federal support (yes)*

# RepurposeVS: A Drug Repurposing-Focused Computational Method for Accurate Drug-Target Signature Predictions

Naiem T. Issa[1], Oakland J. Peters[1], Stephen W. Byers[1,2] and Sivanesan Dakshanamurthy[*,1,2]

[1]*Department of Oncology, Georgetown Lombardi Cancer Center, Washington, D.C. 20057, USA*

[2]*Department of Biochemistry, Molecular and Cellular Biology, Georgetown University Medical Center, Washington, D.C. 20057, USA*

**S. Dakshanamurthy**

**Abstract:** We describe here RepurposeVS for the reliable prediction of drug-target signatures using X-ray protein crystal structures. RepurposeVS is a virtual screening method that incorporates docking, drug-centric and protein-centric 2D/3D fingerprints with a rigorous mathematical normalization procedure to account for the variability in units and provide high-resolution contextual information for drug-target binding. Validity was confirmed by the following: (1) providing the greatest enrichment of known drug binders for multiple protein targets in virtual screening experiments, (2) determining that similarly shaped protein target pockets are predicted to bind drugs of similar 3D shapes when RepurposeVS is applied to 2,335 human protein targets, and (3) determining true biological associations *in vitro* for mebendazole (MBZ) across many predicted kinase targets for potential cancer repurposing. Since RepurposeVS is a drug repurposing-focused method, benchmarking was conducted on a set of 3,671 FDA approved and experimental drugs rather than the Database of Useful Decoys (DUD-E) so as to streamline downstream repurposing experiments. We further apply RepurposeVS to explore the overall potential drug repurposing space for currently approved drugs. RepurposeVS is not computationally intensive and increases performance accuracy, thus serving as an efficient and powerful *in silico* tool to predict drug-target associations in drug repurposing.

**Keywords:** Cancer, drug, interaction, mebendazole, repositioning, repurposing, virtual screening.

## 1. INTRODUCTION

Drug repurposing- the process of utilizing drugs approved for one indication for another- is an efficient method for bolstering the pharmaceutical pipeline [1]. Given that approved drugs have known well-tolerated toxicity profiles, they can, therefore, be streamlined back into the development pipeline directly at phase II. Despite some successes, drug repurposing remains a challenge for two main reasons: (1) validating druggable therapeutic target(s) associated with the disease, and (2) confidently establishing the repertoire of protein target interactions for the FDA approved drug set. This manuscript will focus on the latter aspect.

A variety of methods for establishing drug-target interactions are employed in both academia and industry. High-throughput screening (HTS) strategies are used for establishing interactions for large drug libraries against protein targets of interest [2]. These approaches, however, have multiple obstacles. These include the financial cost per assay run, development of appropriate screening assays, maintaining biochemical relevance of the target given the assay (i.e. target immobilization in 96-well plates may alter binding site properties), among others [3]. The amount of potential druggable disease-related targets is also exponentially increasing [4] along with the number of synthesizable drugs [5]. Creating the vast possible drug-target space of true interactions and further narrowing it to that of physiologic- and disease-relevance remains a great challenge.

Computer-aided methods allow for a substantial increase in efficiency in establishing drug-target interactions and are constantly becoming more accurate as the biophysical mechanisms behind molecular recognition become better understood [6]. Such methods are typically used in virtual screenings against a protein target of interest, where large drug libraries (>1,000,000 structures) are subjected to an algorithm that quantifies the drugs' "fit" into the binding site. The first few hundred or thousand drugs are then validated experimentally, and the potential drug-target space has been drastically reduced to that with the greatest biological plausibility.

Many efforts for computationally predicting drug-target interactions exist, spanning both chemo-centric [7, 8] and target-based methodologies [9, 10]. Chemo-centric approaches utilize physical and chemical information obtained from ligands. Some approaches relate receptors based on the chemical similarity [7] as well as shape similarity [8] between ligands. Large public databases that aid in extracting ligand-based data for informatics also exist [9]. Target-based approaches, on the other hand, rely on docking [10-13] or binding site similarity [14]. Docking has driven some successful drug repurposing attempts [15-19], but scoring functions are generally considered inaccurate in calculating free energies of binding due to difficulty in predicting bioactive poses and variable contributions of weak interactions [20]. Alternatively, binding site comparison

*Address correspondence to this author at the Department of Oncology, Georgetown University, Washington D.C. 20057, USA;
Tel: ++1-202-687-2347; E-mail: sd233@georgetown.edu.

1875-5402/15 $58.00+.00 

methods [21, 22] are implemented under the premise that similar binding sites should bind similar molecules. The use of binding site similarities has been successful in identifying novel targets for known drugs [23, 24] under the assumption that drugs interact with proteins containing similar binding sites [25, 26].

While chemo-centric and target-based methods have their own strengths and limitations, few computational methods attempt to combine ligand- and protein-based approaches [27, 28]. In this work, we present RepurposeVS, a comprehensive method for predicting FDA approved and experimental drug-protein target interactions through computationally efficient virtual screenings. RepurposeVS combines high-throughput docking with quantified shape, atom pair, and other descriptor similarity information of query drugs to reference experimentally derived crystal structure complexes. Furthermore, the utilized normalization procedure provides biological context of binding and allows for cross-protein comparison of drug binding signatures instead of protein-specific rank-ordering of drugs. This enables a standardized prioritization of predicted drug-target signatures for the entire proteome cohort in a study and the future incorporation of new signatures when novel protein target structures become available.

To assess accuracy, RepurposeVS was compared to the GLIDE docking algorithm in virtual screening experiments for prioritization of known drug binders for multiple pharmaceutically relevant protein targets. As RepurposeVS is a drug repurposing-driven method, the drug set chosen for benchmarking includes 3,671 FDA approved and experimental drugs. This drug set is composed of diverse chemical structures and chemotypes, as well as streamlines the generation of drug repurposing hypotheses for later experimental testing. Although benchmarks for virtual screening methods typically utilize the Database of Decoys (DUD-E) [29], we are focused on drug repurposing and therefore the ability of RepurposeVS to enrich for actives from an approved drug set rather than a chemical set of closely related analogues that may or may not be clinically relevant. RepurposeVS provided the greatest enrichment for known active drugs and was then scaled up to predict drug-target signatures across 2,335 human protein targets. Cursory global validation across the entire protein target set was then achieved by recapitulating the phenomenon of similarly shaped protein pockets binding drugs of similar shape [30, 31]. Biological validation was further obtained for the anti-hookworm drug mebendazole *via* kinase binding assays, thus providing further evidence to its anti-cancer efficacy for repurposing. Finally, RepurposeVS was used to explore the entire potential drug repurposing space by devising a "repurposing potential score". With its high accuracy and ease of implementation, RepurposeVS is an efficient computational method for the accurate prediction of drug-protein target signatures to drive drug repurposing efforts forward.

## 2. MATERIALS AND METHODS

### 2.1. Drug and Protein Target Dataset

Drugs were obtained from the DrugBank [32], FDA [33] and BindingDB [34]. LigPrep [35] was used to prepare and minimize drug structures at neutral pH of 7.0. Human protein target crystal structures containing a reference drug in the binding pocket with X-ray resolution <2.5 angstrom were chosen from RCSB (www.rcsb.org). After processing, the dataset included 3,671 drugs and 2,335 protein target crystal structures. Known active drugs for the benchmark protein targets HSP90A (PDB: 4O05), CA4 (PDB: 3FW3), ALDR1 (PDB: 3RX3), ACE (PDB: 1O86), PPARG (PDB: 3VSO), ADRB2 (PDB: 3NYA), VEGFR2 (PDB: 2P2H), ESR1 (PDB: 3ERD), AR (PDB: 3L3Z), BACE1 (PDB: 3VF3), GR (PDB: 4P6X), and HMGCR (PDB: 1HWK) were obtained *via* DrugBank annotations.

### 2.2. RepurposeVS Procedure

The workflow for RepurposeVS, modeled after the "Train Match, Fit, Streamline" (TMFS) protocol [36], is outlined in Fig. (**1**). A 3D comprehensive conformer library was generated using ConfGen [37] for each drug. From this library, the conformer whose 3D shape was most similar to that of the reference ligand bioactive pose was chosen for all subsequent steps. GLIDE [38] docking was performed to obtain free energies of binding, QikProp [39] was used for generating ligand-based 2D descriptors, and 3D shape descriptors for drug and protein binding sites were generated using spherical harmonics expansion coefficients Java software package provided to us by the Thornton group [40]. Reference-occupied protein target pocket shapes were determined using protomol information from sc-PDB [41]. Atom Pair (AP) similarity normalized scores were calculated directly using Strike [42].

The RepurposeVS Z-score ranking equation for a query drug $q$ against protein target $p$ with reference drug $r$ is as follows:

$$Z(q,p,r) = \omega_j Y(p,q) + \omega_k P(r,q) + \sum_{m=1}^{1}[\omega_m f_m(p,q) + \omega'_m f'_m(r,q)] + \sum_{n=1}^{N} X_n(r,q) + CS(OLIC) \quad (1)$$

$Y$ represents the rigorously normalized docking score based on the method outlined in **Section 2.2.1** below with weight $\omega_j = 4$. $P$ represents the normalized AP similarity tanimoto coefficient ($T_c$) of a query drug $q$ to the reference drug $r$ along with its designated weight ($\omega_k = 4$). The first summation corresponds to the shape similarity metric composed of two functions: (1) $\omega_m f_m(p,q)\ \omega_m f_m(\sigma_p, \sigma_l)$, where $f_m$ is the shape function corresponding to a similarity quantification between pocket shape of the protein target $p$ and the query drug $q$ with weighting factor $\omega_m = 2$, and (2) $\omega'_m f'_m(r,q)$, where $f'_m$ is a shape function corresponding to a similarity quantification between reference drug shape $r$ and query drug shape $q$ with weighting factor $\omega'_m = 2$. Shape similarities are represented as Euclidean distances between the spherical harmonics expansion coefficients, as described in [43]. The second summation term corresponds to the combined similarity of $N = 10$ query drug-based descriptors terms ($X_n$) to reference drug $r$. Normalized $T_c$ scores were calculated for the following descriptors: (1) number of H-bond acceptors, (2) number of H-bond donors, (3) dipole, (4) electron affinity, (5) globularity, (6) molecular

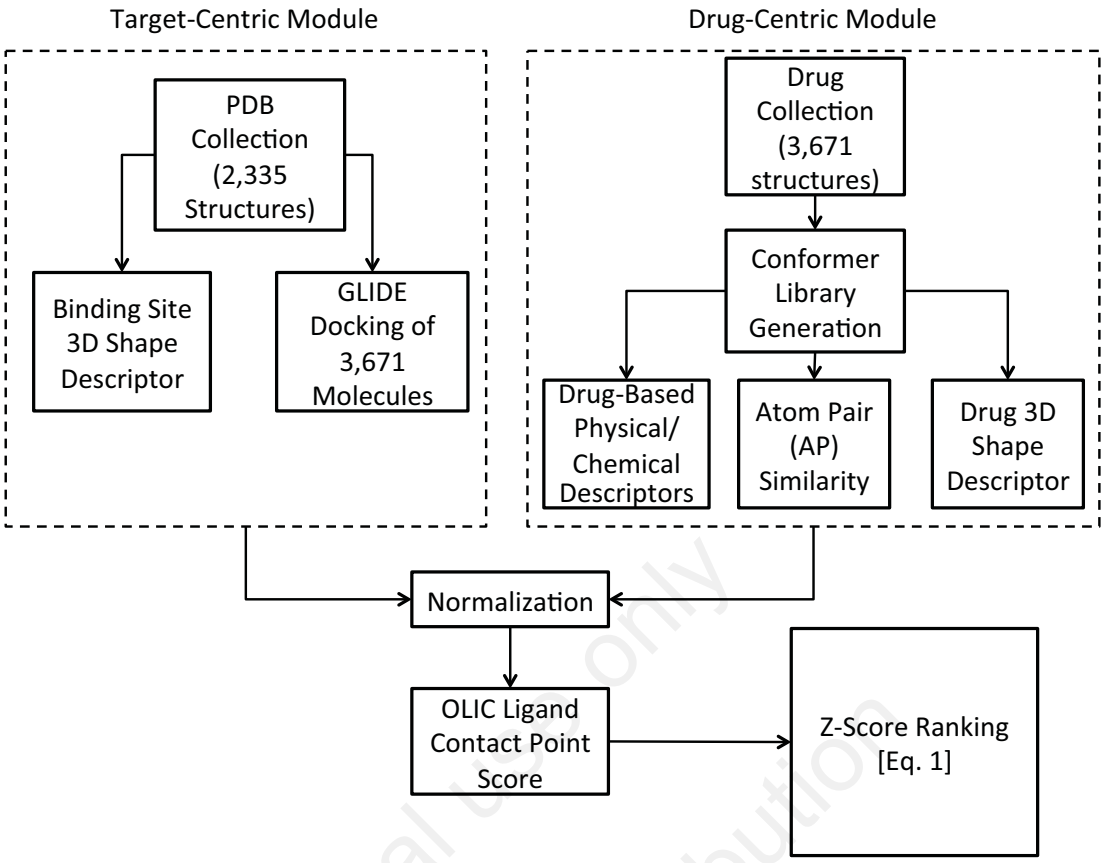



**Fig. (1).** Workflow of RepurposeVS algorithm.

weight, (7) ClogP, (8) number of rotatable bonds, (9) solvent-accessible surface area, and (10) volume.

The $CS(OLIC)$ term is a correction term called "optimal ligand interaction correction" (OLIC), an algorithm that obtains a better estimate of drug-protein interactions on the reference binding site by assuming that drugs will have similar experimental activity if their interaction involves similar binding site residues and makes similar interaction patterns to the reference drug. The following equation was used to determine binding site energies for reference drug (2) and query drugs (3):

$$S(OLIC-r)_p = \sum_{n=1}^{NR} \omega_n E_{n,p} \quad (2)$$

$$S(OLIC-q)_{q,p} = \sum_{n=1}^{NQ} \omega_n E_{n,q,p} \quad (3)$$

The sums are over the number of contact points NR or NQ between protein $p$ and its reference drug or query drugs, respectively. Contact points for drugs are described as those that overlap with established reference drug-protein contacts. Drugs that match or cover most of the interactions as that of the reference scored higher. Their corresponding energies are evaluated and compared with the energy of the reference drug. Energy of the test ligands scored higher if it is close or higher than the energy of the reference drug. The $E_{n,p}$ corresponds to energy for the nth contact point for the reference drug-protein complex $(p,r)$. The $E_{n,q,p}$ term corresponds to energy for the nth contact point for the $q$th query drug and protein $p$. Weighting factors specific to each contact point used and are dependent on the particular drug-target complex.

The correction term *CS(OLIC)* has been determined as a difference between the two sums:

$$CS(OLIC) = S(OLIC-r)_p - S(OLIC-q)_{q,p} \quad (4)$$

The additive combination of the aforementioned normalized terms with their respective weights results in the final RepurposeVS comprehensive Z-score (1) to rank drugs for a given target.

### *2.2.1. Rigorous Normalization Procedure of RepurposeVS Terms*

RepurposeVS contains distinct parameters in (1) that are represented in different units, which correspondingly contain very different raw numeric ranges. For example, docking scores are expressed in kJ/mol where small changes in number correspond to large changes in the free energies of binding. Shape similarity terms are quantified by Euclidean distances and, therefore, function on an independent range of values that are incompatible with other terms in the equation. Consequently, to better allow RepurposeVS parameters to be compared and weighted intelligently, raw values for the docking score and shape similarity terms $Y$, $f_m$, and $f'_m$

were normalized onto the $N(x): R \rightarrow (0,1)$ unit range using a sigmoid function to preserve order and provide symmetry. The normalization function is defined as follows:

$$N_\alpha(x) = 1 - |1 - S_\alpha(x)| \quad (5)$$

where x is the raw parameter, *S(x)* is a sigmoid function, and $\alpha$ is a tunable scalar coefficient chosen to maximize the information-preserving variance in the image of $N(x)$ (5). Since the range varied significantly between parameters, the coefficient $\alpha$ varied as well.

For the sigmoid function, the hyperbolic tangent function is chosen because it is well-behaved and computationally tractable, yielding (6). Since some RepurposeVS parameters required subtly different normalization properties. Hence, (6) was re-expressed for easier modification in special normalization cases. Expressing (6) in terms of the simpler logistic function *L*, shown in (7), yields the equivalent function in (8):

$$N_\alpha(x) = 1 - |\tanh(\frac{\alpha x}{2})| \quad (6)$$

$$L_\alpha(x) = \frac{1}{1 + e^{-\alpha x}} \quad (7)$$

$$N_\alpha(x) = 1 - 2 * |\frac{1}{2} - L_\alpha(x)| \quad (8)$$

To check the information preserving quality of this normalization, we formed histograms of both the un-normalized, or raw (Fig. **2A**), and normalized population distributions (Fig. **2B**) for the shape similarity parameter using α=0.1. A scatter plot of the un-normalized shape parameter *versus* the normalized shape parameter was also formed (Fig. **2C**). Fig. (**2A**) shows that in this case normalization results in a good fit for a symmetric and centered (at 0.5) Gaussian distribution implying that the normalized data will be statistically well behaved. By comparing the un-normalized to normalized distributions, we can see that the input distribution was not significantly

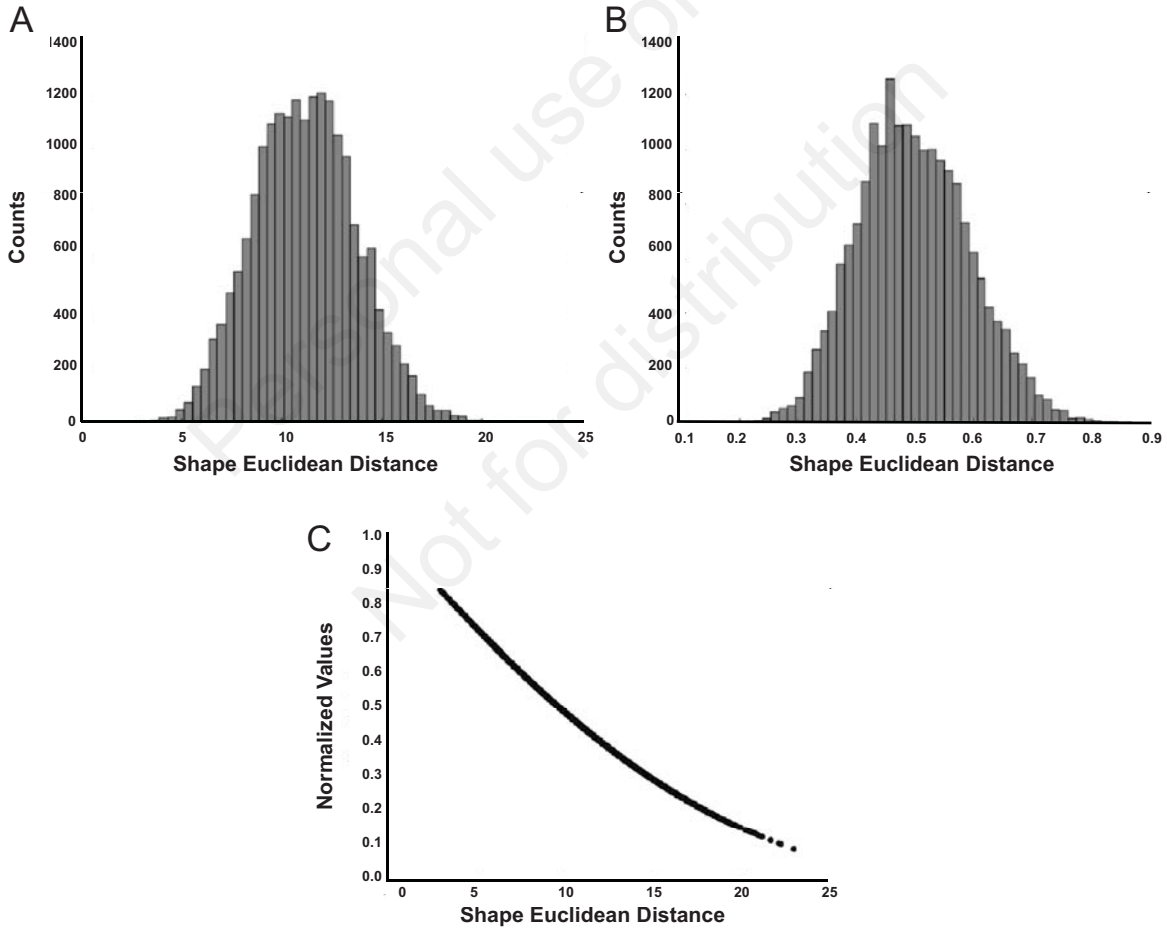





**Fig. (2).** Normalization of RepurposeVS parameters. (**A**) Histogram of raw (non-normalized) scores for post-docking shape similarity Euclidean distance calculations of 2,207 unique drug-protein target pairs. The Y-axis shows counts of data points *versus* X-axis Euclidean distances using a bin-width of 0.5. (**B**) Histogram of normalized scores for the same 2,207 shape similarity calculations shown in (**A**). Normalization equation is shown in Eq. (14). The normalization preserves the Gaussian shape of the distribution, and centers the new distribution on the 0.5 mid-point of the 0-1 unit range. (**C**) Scatterplot showing the relationship between non-normalized shape similarity Euclidean distances (X-axis) and the resultant normalized values (Y-axis), for the data points shown in (**A**) and (**B**). The approximately linear relationship shown implies that the normalization does little to distort the population, although some bending is visible at the high-end (shape Euclidean distance values above 15).

distorted by our normalization function $N(x)$. Fig. (**2C**) shows an approximately linear relationship between the majority of raw and normalized data point pairs, implying that the coefficient was a good choice for capturing the dynamic range of most of the dataset. The procedure was repeated for docking scores $Y$ using α=0.25 (data not shown). This normalization procedure allows for RepurposeVS to better predict viable drug-protein signatures in an absolute manner, where relativistic knowledge of other drugs in an experimental cohort is not necessary to quantify and establish binding signatures. Thus, resulting Z-scores can be pooled across all protein target systems for global objective prioritization of drug-target predictions.

### 2.3. Drug Shape Deviation Score

To determine shape similarity for drugs shared between unique protein target pairs, the "Drug Shape Deviation Score" ($\overline{F}$) metric was created. For analysis, target pairs must have at least three drugs predicted in common (i.e. within top 40 ranking for each protein target *via* Z-score). A "permutation of differences" (9)-(12) approach was applied to arrive at a score within the 0-1 unit range that reflects the average shape deviation of the predicted common drugs for a protein target pair. The process is as follows:

$$V = \{v_1, v_2, ..., v_n\} \tag{9}$$

where, for a given protein target pair, $V$ is the set of common drugs,

$$F = \{f_k = \ |a_{k_1} - a_{k_2}| \ \big| \ \ a_k = [v_i, v_j] \in C_2(V)\} \tag{10}$$

$$|F| \ = \binom{n}{2} = \frac{(n-1)n}{n} \tag{11}$$

$$\overline{F} = \sum_{k=1}^{|F|} \frac{f_k}{|k|} \tag{12}$$

$a_k$ is the Euclidean distance between a pair of common drugs, $C_2(V)$ is the set of all combinations of two elements from $V$ without replacement to generate the number of difference values, $F$ is the set of differences between the Euclidean distances *via* all possible permutations, $|F|$ is the number of elements within set $F$, and $\overline{F}$ is the average across all Euclidean distance values.

### 2.4. Kinase Binding Assay

Kinase assays were performed using Kinomescan, by Discoverx, CA, USA and Caliper LabChip 3000 by Caliper Life sciences, USA as described previously [36]. The determination of MBZ thermodynamic binding affinities ($K_d$) to kinase targets predicted by RepurposeVS was performed by using active site-directed competition binding [44]. Kinase-tagged T7 phage strains were grown in parallel in 24-well blocks in an E. coli host derived from the BL21 strain. E. coli bacteria were grown to log-phase and infected with T7 phage from a frozen stock (multiplicity of infection = 0.4) and incubated with shaking at 32°C until lysis (90-150 minutes). The lysates were centrifuged (6,000 x g) and filtered (0.2 μm sieves) to remove cell debris. The remaining kinases were produced in HEK-293 cells and subsequently tagged with DNA for qPCR detection. Streptavidin-coated magnetic beads were treated with control (biotinylated) for 30 minutes at room temperature to generate affinity resins for kinase assays. The liganded beads were blocked with excess biotin and washed with blocking buffer (SeaBlock (Pierce), 1% BSA, 0.05 % Tween 20, 1 mM DTT) to remove unbound ligand and to reduce non-specific phage binding. Binding reactions were assembled by combining kinases, control liganded affinity beads, and mebendazole in 1x binding buffer (20 % SeaBlock, 0.17x PBS, 0.05 % Tween 20, 6 mM DTT). Mebendazole was prepared as 40x stocks in 100% DMSO and directly diluted into the assay. All reactions were performed in polypropylene 384-well plates in a final volume of 0.04 ml. The assay plates were incubated at room temperature with shaking for 1 hour and the affinity beads were washed with wash buffer (1x PBS, 0.05 % Tween 20). The beads were then re-suspended in elution buffer (1x PBS, 0.05 % Tween 20, 0.5 μM non-biotinylated affinity ligand) and incubated at room temperature with shaking for 30 minutes. The kinase concentration in the eluates was measured by qPCR. Drugs that bind the kinase active site and directly prevent kinase binding to the immobilized ligand will reduce the amount of kinase captured, whereas drugs that do not bind the kinase have no effect on the amount of kinase captured. The amount of kinase captured in test *versus* control samples were measured by using a quantitative, precise and ultra-sensitive qPCR method that detects the associated DNA label. Using (13), the primary screen binding interactions are reported as '% Ctrl' (Percent kinase remaining activity), where lower numbers indicate stronger hits.

$$\text{Percent Control (\%Ctrl)} = \frac{\text{Mebendazole signal - Positive control signal}}{\text{DMSO Negative control signal - Positive control signal}} \text{ x } 100 \tag{13}$$

In a similar manner, binding constants ($K_d$) for mebendazole-kinase interactions are calculated by measuring the amount of kinase captured as a function of the mebendazole concentration in a dose response manner. An 11-point 3-fold serial dilution of each test compound was prepared in 100% DMSO at 100x final test concentration and subsequently diluted to 1x in the assay (final DMSO concentration = 1%). Most $K_d$s were determined using a starting concentration = 30,000 nM. If the initial $K_d$ determined was < 0.5 nM (the lowest concentration tested), the measurement was repeated with a serial dilution starting at a lower starting concentration. Binding constants ($K_d$) were calculated with a standard dose-response curve (drug dose (x-axis) - qPCR signal (y-axis)) using the Hill equation in (14) with the Hill Slope set to -1.

$$\text{Response (Y)} = \text{Background} + \frac{\text{Signal(max) - Background}}{1 + \left(\frac{k_d}{\text{Drug Dose (X)}}\right) \text{Hill Slope}} \tag{14}$$

### 2.5. Repurposing Potential

Original drug class indications, obtained from DrugBank [32], were given a "Repurposing Potential Score" ($T$) based

on the number of drugs studied for a given approved indication class and the number of unique RepurposeVS-predicted disease classes for that indication class. (15) represents the repurposing potential score (*T*) for a given disease class *i* :

$$T = \frac{d_i}{d_{neo}} + \frac{k_i}{k_{neo}} \quad (15)$$

where $d_i$ and $d_{neo}$ correspond the number of drugs approved for disease classes "$i$" and neoplasms, respectively, and $k_i$ and $k_{neo}$ correspond to the number of predicted new disease classes excluding the original for disease classes "$i$" and neoplasms. All disease classes are normalized to the neoplastic disease class since it contained both the greatest number of drugs with unique indications and predicted new diseases classes. The Online Mendelian Inheritance in Man (OMIM) [45] and the Comparative Toxicogenomics Database (CTD) [46] were used to annotate disease classes for predicted drug-protein target interactions.

## 3. REPURPOSEVS PERFORMS SUPERIORLY TO GLIDE DOCKING IN PRIORITIZING KNOWN BINDERS FOR PROTEIN TARGETS IN VIRTUAL SCREENING

Virtual screenings were performed on a set of 12 pharmaceutically relevant protein targets to assess the accuracy of RepurposeVS. RepurposeVS performed superiorly to GLIDE, a docking algorithm found to be accurate in high-throughput virtual screenings [47], in enriching for known drug binders to a protein target over a set of 3,671 drugs (Fig. **3A**, **B**). Using a paired, one-tailed student's t-test, RepurposeVS performed statistically significantly better than GLIDE ($P<0.05$). Receiver operating curves demonstrate that RepurposeVS increased accuracy the most for solvent-exposed binding pockets, such as VEGFR2 kinase domain and $\beta_2$-adrenergic G protein-coupled receptor, whereas minimal increase occurred for buried pockets such as the estrogen and androgen nuclear receptors (Fig. **S1**). This differential may be attributed to greater flexibility in binding pose in exposed sites, which are specifically reflected by the docking score and pocket shape terms. Altering the weights $\omega_k$ and $\omega_m$ for docking score and pocket shape, respectively, had no appreciable effect on performance (Fig. **3A**). This suggests that the other parameters in compensate for the imprecision derived from the nature of exposed pockets and that RepurposeVS is a robust method applicable to diverse protein targets.

## 4. GLOBAL VALIDATION OF REPURPOSEVS USING SHAPE SIMILARITY

RepurposeVS was applied to a set of 2,335 human protein target crystal structures and globally validated using the concept of similarly shaped drugs binding to protein

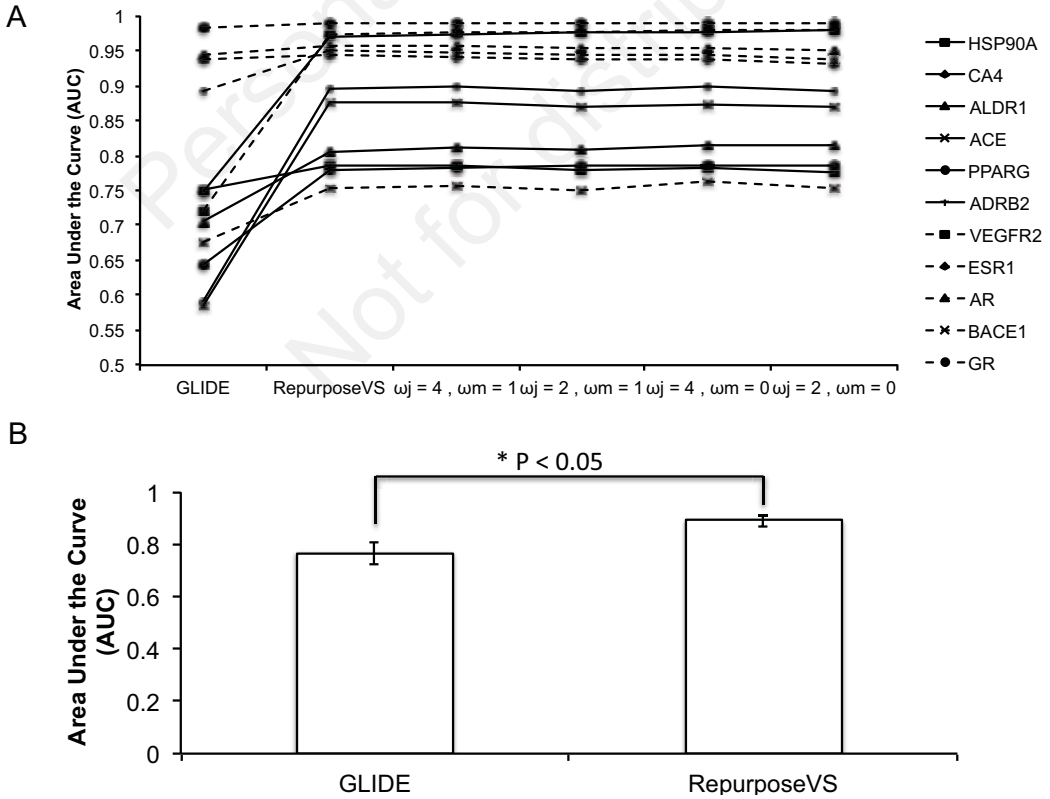



**Fig. (3).** Areas under the curve (AUCs) for virtual screening of approved active drugs across 12 protein targets. (A) Outcomes of GLIDE docking and RepurposeVS in virtual screening experiments enriching for true active drugs for the noted protein targets. The remaining conditions reflect adjusted weights for the docking parameter (ωj) and protein shape parameter (ωm) in RepurposeVS (Eq. 1). (B) Average AUC across all 12 targets for each method.

target sites of similar shape. Shape complementarity is a critical aspect of biomolecular recognition, though it may not explain all possible binding modes. Nonetheless, it has generally been noted that drugs that interact with protein binding sites of similar shapes tend to exhibit shape similarity to each other. We first determined that the notion of similarly shaped drugs bind to similar protein pockets is upheld using the reference co-crystallized molecules for the protein target set (Fig. **4**). Similarity between protein target pockets was quantified using two metrics: (1) Euclidean distance of the space-filling protomol structure (Fig. **4A**), and (2) root-mean-square deviation (RMSD) of binding site residues 6Å from the geometric center of the bound molecule (Fig. **4B**). The former metric characterizes the binding site occupancy volume whereas the latter metric is a topological term reflective of the binding site Cα backbone. RMSD values were calculated using Maestro [48]. There exists a direct correlation between drug-drug shape Euclidean distances and protein-protein binding site shape Euclidean distances (Fig. **4A**) and backbone RMSDs (Fig. **4B**). This implies that for true biochemical associations, determined *via* crystal structures, similarly shaped molecules bind protein pockets of similar shape and topology. Using the "Drug Shape Deviation Score", $\overline{F}$ (12), a similar trend was observed for drugs predicted by RepurposeVS (top 40 by Z-score) to bind the same protein targets (Fig. **4**). Thus, RepurposeVS is a valid method for determining drug-target associations across a large and diverse protein target set *via* the pharmacological metric of similarly shaped drugs binding similarly shaped protein pockets.

## 5. *IN VITRO* BIOLOGICAL VALIDATION OF REPURPOSEVS USING MEBENDAZOLE FOR CANCER DRUG REPURPOSING

To biologically confirm RepurposeVS predictions *in vitro*, we tested the binding of protein kinase target hits to mebendazole (MBZ) for cancer drug repurposing. MBZ was originally approved for its potent nanomolar inhibition of hookworm tubulin. It is thought that its cross-over effect on mammalian tubulin, though with 1000x less potency, is responsible for its anti-cancer efficacy *in vitro* [49]. Using kinase binding assays, nano- and micromolar inhibition of several predicted kinase targets of MBZ was confirmed (Table **1**). MBZ appears to inhibit kinases found within two main branches of the kinome phylogenetic tree, with nanomolar potency clustering on one branch and micromolar potency on the other (Fig. **5**). However, intra-branch variability in potency is also obvserved. It is likely that the semi-promiscuous nature of MBZ (Fig. **5**) towards kinases is a result of a small fragment that allows it to interact with the benzimidazole moiety acting as head group anchor connecting loop residues between the c-lobe and n-lobe.



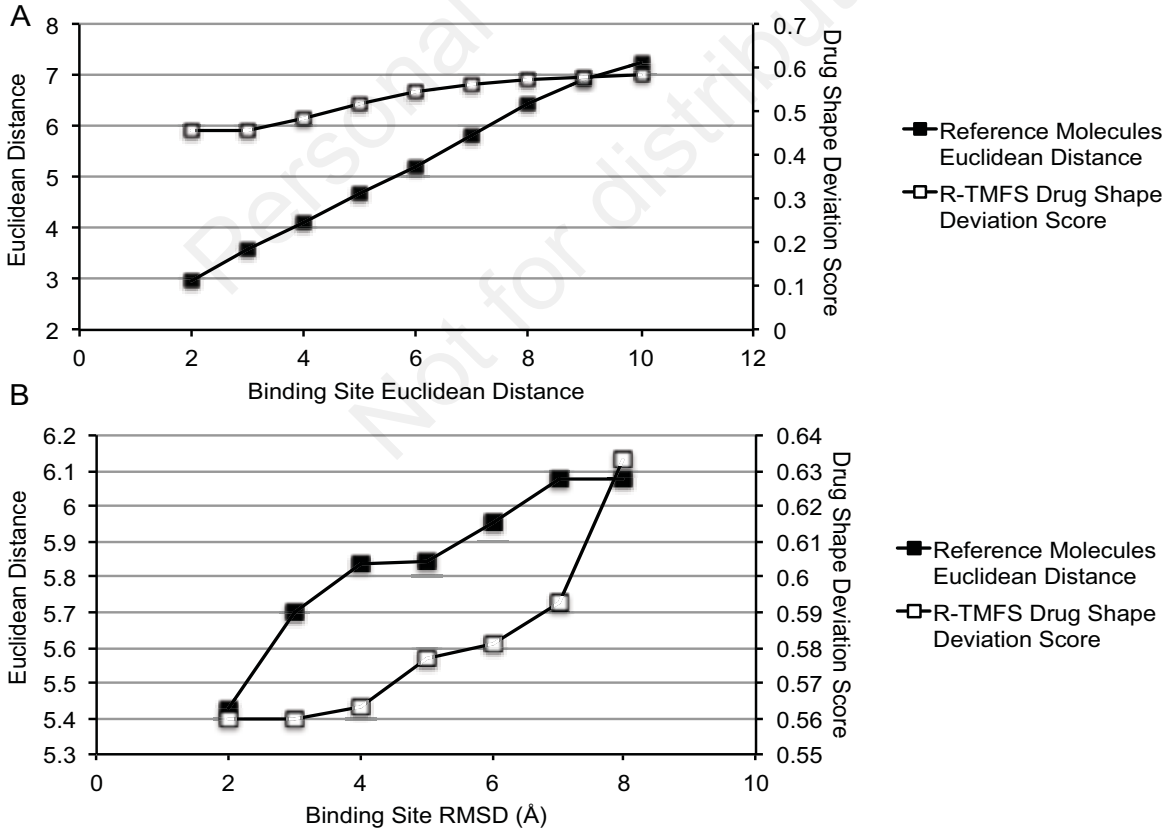



**Fig. (4).** Trends in drug shape as a function of binding site shape and structural differences between protein target pairs. Line plots depicting shape Euclidean distances between co-crystallized reference molecules and normalized "Drug Shape Deviation Scores" ($\overline{F}$) against (A) binding pocket shape differences quantified by Euclidean distances and (B) backbone root-mean-squared deviation (RMSD) in angstroms. The data was binned into 1-unit groups with their means represented in the plot. Smaller Euclidean distances or RMSDs imply greater similarity.

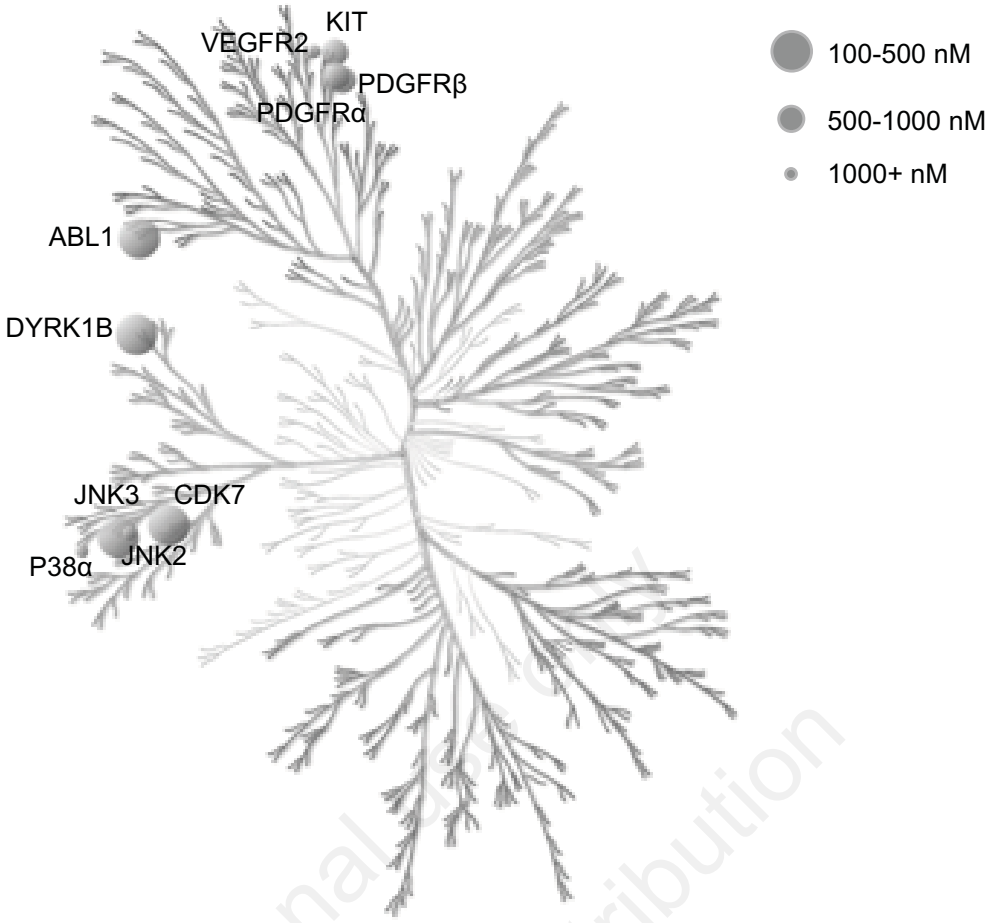



**Fig. (5).** Validated mebendazole (MBZ) kinase targets predicted from RepurposeVS. Kinases for which binding affinities were determined are shown on the kinome phylogenetic tree.

MBZ is also dually able to form water-mediated contacts or directly interact with ATP site cavity-forming residues in the absence of water molecules. Our predicted kinase hits and the activity data of MBZ indicate that its anti-cancer properties may be due to a synergistic inhibition of tubulin as well as kinase activity. Interestingly, for lung cancer, combined inhibition of microtubules and DYRK1B, a MBZ target (Table **1**), is a more potent therapeutic strategy than microtubule inhibitors alone [50]. In this instance, a single drug such as MBZ, which has both properties, would be advantageous. RepurposeVS, thus, is able to reliably predict targets for MBZ that contribute to its repurposing for cancers.

## 6. DRUG REPURPOSING POTENTIAL

RepurposeVS was used to provide a cursory assessment the potential repurposing space for FDA approved drugs based on their drug classes (Fig. **6**). We devised a repurposing potential score (*T*) in (15) for this purpose. Anti-neoplastic agents are shown to have the greatest repurposing potential with regards to the number of drugs and the diversity of newly predicted disease categories, with a total of 47 drugs and 8 disease categories. The nutritional-metabolic and neoplasm disease classes are also predicted to have the greatest number of drugs with the greatest number of unique original indications repurposed to them with 143 drugs/29 indications and 123 drugs/22 indications, respectively.

The overrepresentation of anti-neoplastic drugs is expected as tumor development is due to perturbations in a variety of cell processes that are likely shared with other diseases. Dysregulated kinase signaling, for example, is a ubiquitous pathogenic disease mechanism given the role of kinases in signal transduction. Thus, kinase inhibitors would be expected to potentially be useful in other diseases. In addition, some cancer drugs exhibit polypharmacology that simultaneously alter multiple cell processes. Alternatively, anti-infection agents exhibit relatively low repurposing potential (Fig. **6**). This emphasizes the selectivity of these agents towards non-human targets for efficacy and desired therapeutic indices [51]. Some of these drugs, however, exhibit modest repurposing potential. These include anti-bacterial agents, possibly attributed to structural similarity between bacterial motifs and human proteins [52]. Antipsychotic agents and other psychiatry-approved drugs also are predicted to have modest repurposing potential. These drugs typically exhibit polypharmacology through GPCR-mediated interactions [53], and some are being repurposed for cancer therapy [54]. The outcomes of the potential drug repurposing space are in pharmacological and clinical agreement with the known properties of the mentioned drugs, further confirming the ability of

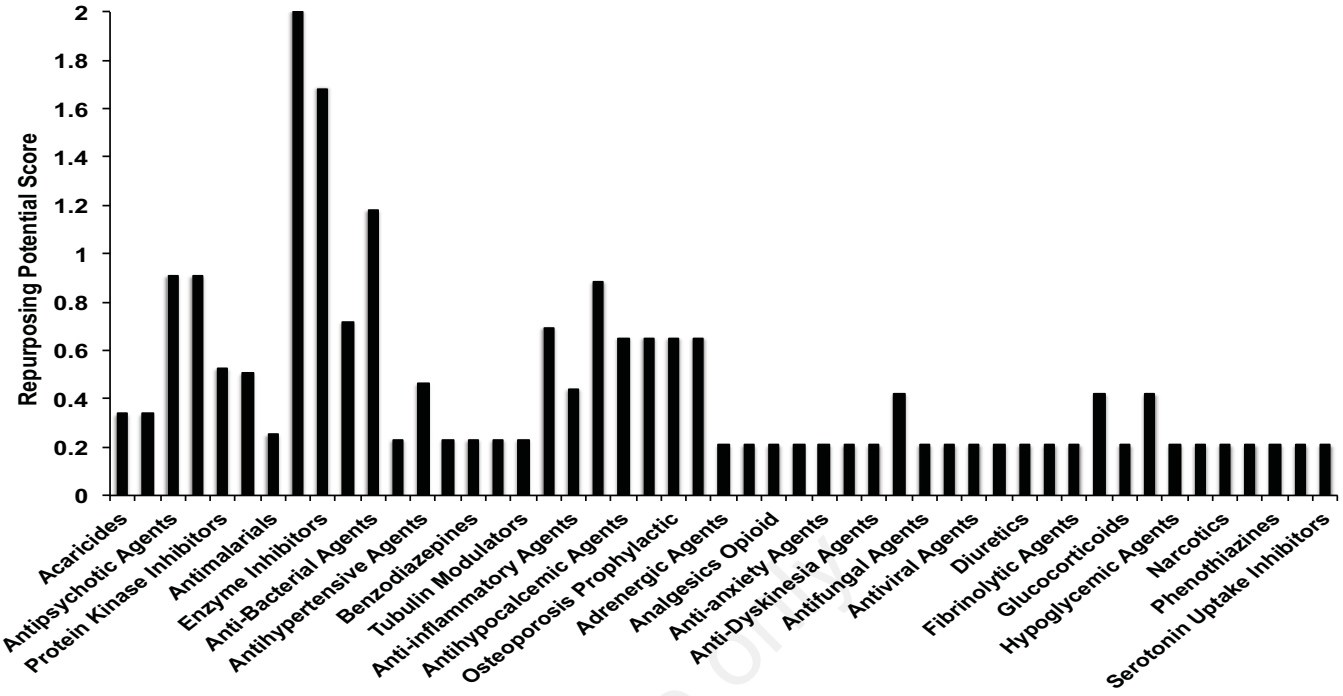



**Fig. (6).** Histogram of predicted repurposing potential of approved drugs/indications to new disease classes. The "Repurposing Potential Score" is calculated using (15) (see Materials and Methods).

RepurposeVS to empirically predict drug-target signatures for higher-order pharmacologic assessment.

**Table 1. Binding affinities of MBZ for predicted kinase hits.**

| Kinase Target | Percent Control at 10μM | Binding Affinity ($K_d$) in nM |
|---|---|---|
| ABL1(E255K)-phosphorylated | 2.2 | N/D |
| ABL1(T315I)-phosphorylated | 3.2 | N/D |
| ABL1-nonphosphorylated | 2 | N/D |
| ABL1-phosphorylated | 0.9 | 120 |
| CDK7 | 11 | 390 |
| CSNK1D | 36 | N/D |
| DYRK1A | 34 | N/D |
| DYRK1B | 5.6 | 340 |
| GSK3B | 35 | N/D |
| JAK3 | 29 | N/D |
| JNK1 | 14 | N/D |
| JNK2 | 9.6 | 1090 |
| JNK3 | 3 | 410 |
| KIT (D816V) | 7.4 (33) | 750 |
| MET | 32 | N/D |
| P38-alpha | 17 | 1660 |
| PDGFR-A | 7.8 | 820 |
| PDGFR-B | 3.2 | 660 |
| PIK3CG | 18 | N/D |
| SRC | 34 | N/D |
| ULK2 | 30 | N/D |
| VEGFR-2 | 30 | 3600 |

## CONCLUSION

RepurposeVS is a combined drug-centric and protein-centric computational method for formulating drug-target signature predictions in drug repurposing. Validity was confirmed through benchmark virtual screenings using 12 protein targets of pharmaceutical interest to better enrich for their respective known approved drugs over GLIDE docking. RepurposeVS was also validated by recapitulating that drugs of similar shapes were predicted to bind similarly shaped protein pockets when defining pocket shapes through drug occupancy, and also by confirming predicted kinase hits of mebendazole *via* kinase binding assays. Finally, RepurposeVS was used to quantify "repurposing potential scores" for drugs categorized by disease indication and showed that anti-infection compounds had the least repurposing potential whereas anti-neoplastic drugs had the greatest. One limitation, however, is that diverse binding modes and protein flexibility are not accounted for. However, RepurposeVS aims only to reestablish the experimental binding states obtained from crystallography so as to decrease false positive and false negative outcomes in virtual screenings. Overall, we believe RepurposeVS to be an efficient computational method to aid drug repurposing endeavors.

## ABBREVIATIONS

AP = Atom-pair

FDA = Food and Drug Administration

HTS = High-throughput screening

MBZ = Mebendazole

Tc = Tanimoto coefficient

VEGFR2 = Vascular endothelial growth factor receptor

## CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

## ACKNOWLEDGEMENTS

The authors wish to acknowledge DOD grants BC062416, BC096277 and CA140882 (SB, SD), R01 CA170653 (SB, SD), CCSG grant NIH-P30 CA51008 and Georgetown Lombardi Cancer Center. We acknowledge the DiscoverX, CA, USA and Caliper Life Sciences, USA for the assay. This project has been funded in whole or in part with Federal funds (Grant #UL1TR000101) from the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), through the Clinical and Translational Science Awards Program (CTSA).

## SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's web site along with the published article.

## REFERENCES

[1] Ashburn, T.T.; Thor, K.B. Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.,* **2004**, *3*, 673-683.

[2] Fox, S.; Farr-Jones, S.; Yund, M.A. High-throughput screening for drug discovery: continually transitioning into new technologies. *J. Biomol. Screen.,* **1999**, *4*, 183-186.

[3] Beibette, J. Gaining confidence in high-throughput screening. *Proc. Natl. Acad. Sci. U S A.,* **2012**, *109*, 649-650.

[4] Griffith, M.; Griffith, O.; Coffman, A.C.; Weible, J.V.; McMichael, J.F.; Spies, N.C.; Koval, J.; Das, I.; Callaway, M.B.; Eldred, J.M.; Miller, C.A.; Subramanian, J.; Govindan, R.; Kumar, R.D.; Bose, R.; Ding, L.; Walker, J.R.; Larson, D.E.; Dooling, D.J.; Smith, S.M.; Ley, T.J.; Mardis, E.R.; Wilson, R.K. DGIdb: mining the druggable genome. *Nat. Methods,* **2013**, *10*, 1209-1210.

[5] Irwin, J.J.; Sterling, T.; Mysinger, M.M.; Bolstad, E.S.; Coleman, R.G. ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.*, **2012**, *52*, 1757-1768.

[6] Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.*, **2002**, *1*, 882-894.

[7] Keiser, M.J.; Roth, B.L.; Armbruster, B.N.; Ernsberger, P.; Irwin, J.J., Shoichet, B.K. Relating protein Pharmacology by ligand chemistry. *Nat. Biotechnol., 25*, 197-206.

[8] Warner, W.A.; Sanchez, R.; Dawoodian, A.; Li, E.; Momand, J. Identification of FDA-approved drugs that computationally bind to MDM2. *Chem. Biol. Drug Des*., **2012**, *80*, 631-637.

[9] Bolton, E.E.; Chen, J.; Kim, S.; Han, L.; He, S.; Shi, W.; Simonyan, V.; Sun, Y.; Thiessen, P.A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S.H. PubChem3D: a new resource for scientists. *J. Cheminform*., **2011**, *3*, 32.

[10] Chen, Y.Z.; Zhi, D.G. Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecules. *Proteins*, **2001**, *43*, 217-226.

[11] Li, H,.; Gao, Z.; Kang, L.; Zhang, H.; Yang, K.; Yu, K.; Luo, X.; Zhu, W.; Chen, K.; Shen, J.; Wang, X.; Jiang, H. TarFisDock: a web server for identifying drug targets with docking approach. *Nucleic Acids Res.*, **2006**, *34*, W219-W224.

[12] Paul, N.; Kellenberger, E.; Bret, G.; Muller, P.; Rognan, D. Recovering the true targets of specific ligands by virtual screening of the protein data bank. *Proteins*, **2004**, *54*, 671-680.

[13] Kellenberger, E.; Foata, N.; Rognan, D. Ranking targets in structure-based virtual screening of three-dimensional protein libraries: methods and problems. J. Chem. Inf. Model., **2008**, *48*, 1014-1025.

[14] Kellenberger, E.; Schalon, C.; Rognan, D. How to measure the similarity between protein ligand-binding sites? *Curr. Comput. Aided Drug Des.,* **2008**, *4*, 209-220

[15] Yang, L.; Chen, J.; He, L. Harvesting candidate genes responsible for serious adverse drug reactions from a chemical-protein interactome. *PLoS Comput. Biol.,* **2009**, *5*, e1000441.

[16] Zahler, S.; Tietze, S.; Totzke, F.; Kubbutat, M.; Meijer, L.; Vollmar, A.M.; Apostolakis, J. Inverse *in silico* screening for identification of kinase inhibitor targets. *Chem. Biol.,* **2007**, *14*, 1207-1214.

[17] Tang, L.; Li, M.H.; Cao, P.; Wang, F.; Chang, W.R.; Bach, S.; Reinhardt, J.; Ferandin, Y.; Galons, H.; Wan, Y.; Gray, N.; Meijer, L.; Jiang, T.; Liang, D.C. Crystal structure of pyridoxal kinase in complex with roscovitine and derivatives. *J. Biol. Chem.,* **2005**, *280*, 31220-31229.

[18] Do, Q.T.; Renimel, I.; Andre, P.; Lugnier, C.; Muller, C.D.; Bernanrd, P. Reverse pharmacognosy: application of selnergy, a new tool for lead discovery. The example of epsilon-viniferin. *Curr. Drug. Discov. Technol.*, **2005**, *2*, 161-167.

[19] Cai, J.; Han, C.; Hu, T.; Zhang, J.; Wu, D.; Wang, F.; Liu, Y.; Ding, J.; Chen, K.; Yue, J.; Shen, X.; Jiang, H. Peptide deformylase is a potential target for anti-Helicobacter pylori drugs: reverse docking, enzymatic assay, and X-ray crystallography validation. *Protein Sci*., **2006**, *15*, 2071-2081.

[20] Bohari, M.H.; Sastry, G.N. FDA approved drugs complexed to their targets: evaluating pose prediction accuracy for docking protocols. *J. Mol. Model*., **2012**, *18*, 4263-4274.

[21] Meslamani, J.; Rognan, D.; Kellenberger, E. sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins. *Bioinformatics*, **2011**, *27*, 1324-1326.

[22] Haupt, V.J.; Schroeder, M. Old friends in new guise: repositioning of known drugs with structural bioinformatics. *Brief. Bioinform*., **2011**, *12,* 312-326.

[23] Kinnings, S.L.; Liu, N.; Buchmeier, N.; Tonge, P.J.; Xie, L.; Bourne, P.E. Drug discovery using chemical systems biology: repositioning the safe medicine comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol*., **2009**, *5*, e1000423.

[24] Defranchi, E.; Schalon, C.; Messa, M.; Onofri, F.; Benfenati, F.; Rognan D. Binding of protein kinase inhibitors to synapsin I inferred from pair-wise binding site similarity measurements. *PLoS One*, **2010**, *5*, e12214.

[25] Das, S.; Kokardekar, A.; Breneman, C.M. Rapid comparison of protein binding site surfaces with property encoded shape distributions. *J. Chem. Inf. Model*., **2009**, *49*, 2863-2872.

[26] Kahraman, A.; Morris, R.J.; Laskowski, R.A.; Favia, A.D.; Thornton, J.M. On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins. *Proteins*, **2010**, *78*, 1120-1136.

[27] Muegge; I. Synergies of Virtual Screening Approaches. *Mini Rev. Med. Chem.*, **2008**, *8*, 927-933.

[28] Broccatelli, F.; Brown, N. Best of both worlds: on the complementarity of ligand-based and structure-based virtual screening. *J. Chem. Inf. Model*., **2014**, *6*, 1634-1641.

[29] Huang, N.; Shoichet, B.K.; Irwin, J.J.; Benchmarking sets for molecular docking. *J. Med. Chem.*, **2006**, *49*, 6789-6801.

[30] Das, S.; Kokardekar, A.; Breneman, C.M. Rapid comparison of protein binding site surfaces with property encoded shape distributions. *J. Chem. Inf. Model*., **2009**, *49*, 2863-2872

[31] Haupt, V.J.; Daminelli, S.; Schroeder, M. Drug promiscuity in PDB: protein binding site similarity is key. *PLoS One*, **2013**, *8*, e65894.

[32] Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A.C.; Wishart, D.S. DrugBank 3.0: A comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.,* **2011**, *39*, D1035-D1041.

[33] U.S. Food and Drug Administration. www.FDA.gov (Accessed **2012**).

[34] Liu, T.; Lin, Y.; Wen, X.; Jorrisen, R.N.; Gilson, M.K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res*., **2007**, *35*, D198-D201.

[35] Schrödinger Release 2013-3: LigPrep, version 2.8, Schrödinger, LLC, New York, NY, **2013**.

[36] Dakshanamurthy, S.; Issa, N.T.; Assefnia, S.; Seshasayee, A.; Peters, O.J.; Madhvan, S.; Uren, A.; Brown, M.L.; Byers, S.W. Predicting new indications for approved drugs using a proteochemometric method. *J. Med. Chem*., **2012**, *55*, 6832-6848.

[37] Small-Molecule Drug Discovery Suite 2013-3: Confgen, version 2.6, Schrödinger, LLC, New York, NY, **2013**.
[38] Small-Molecule Drug Discovery Suite 2013-3: Glide, version 6.1, Schrödinger, LLC, New York, NY, **2013**.
[39] Small-Molecule Drug Discovery Suite 2013-3: QikProp, version 3.8, Schrödinger, LLC, New York, NY, **2013**.
[40] Kahraman, A.; Morris, R.; Laskowski, R.; Thornton, J. Shape variation in protein binding pockets and their ligands. *J. Mol. Biol.,* **2009**, *368*, 283-301.
[41] sc-PDB Home Page. http://bioinfo-pharma.u-strasbg.fr/scPDB/
[42] Small-Molecule Drug Discovery Suite 2013-3: Strike, version 2.4, Schrödinger, LLC, New York, NY, **2013**.
[43] Morris, R.J.; Najmanovich, R.J.; Kahraman, A.; Thornton, J.M. Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics*, **2005**, *21*, 2347-2355.
[44] Fabian, M.A.; Biggs, W.H. 3rd; Treiber, D.K.; Atteridge, C.E.; Azimioara, M.D.; Benedetti, M.G.; Carter, T.A.; Ciceri, P.; Edeen, P.T.; Floyd, M.; Ford, J.M.; Galvin, M.; Gerlach, J.L.; Grotzfield, R.M.; Herrgard, S.; Insko, D.E.; Insko, M.A.; Lai, A.G.; Lelias, J.M.; Mehta, S.A.; Milanov, Z.V.; Velasco, A.M.; Wodicka, L.M.; Patel, H.K.; Zarrinkar, P.P; Lockhart, D.J. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol*., **2005**, *23*, 329-336.
[45] Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, (Baltimore, MD), {2014}. World Wide Web URL: http://omim.org/.
[46] Davis, A.P.; Murphy, C.G.; Johnson, R.; Lay, J.M.; Lennon-Hopkins, K.; Saraceni-Richards, C.; Sciaky, D.; King, B.L.; Rosenstein, M.C.; Wiegers, T.C.; Mattingly, C.J. Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Res*., **2013**, *41*, D1: D1104-D1114.
[47] Warren, G.L.; Andrews, C.W.; Capelli, A-M.; Clarke, B.; LaLonde, J.; Lambert, M.H.; Lindvall, M.; Nevins, N.; Semus, S.F.; Senger, S.; Tedesco, G.; Wall, I.D.; Woolven, J.M.; Peishoff, C.E.; Head, M.S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.*, **2006**, *49*, 5912-5931.
[48] Small-Molecule Drug Discovery Suite 2013-3: Maestro, version 9.6, Schrödinger, LLC, New York, NY, **2013**.
[49] Sasaki, J.; Ramesh, R.; Chada, S.; Gomyo, Y.; Roth, J.A.; Mukhopadhyay, T. The anthelmintic drug mebendazole induces mitotic arrest and apoptosis by depolymerizing tubulin in non-small cell lung cancer cells. *Mol. Cancer Ther*., **2002**, *1*, 1201-1209.
[50] Li, L.; Liu, Y.; Zhang, Q.; Zhou, H.; Zhang, Y.; Yan, B. Comparison of cancer cell survival triggered by microtubule damage after turning Dyrk1B kinase on and off. *ACS Chem. Biol*., **2014**, *9*, 731-742.
[51] Pereira, M.P.; Kelley, S.O. Maximizing the therapeutic window of an antimicrobial drug by imparting mitochondrial sequestration in human cells. *J. Am. Chem. Soc*., **2011**, *133*, 3260-3263.
[52] Trost, B.; Lucchese, G.; Stufano, A.; Bickis, M.; Kusalik, A.; Kanduc, D. No human protein is exempt from bacterial motifs, not even one. *Self Nonself*, **2010**, *4*, 328-334.
[53] Roth, B.L.; Sheffler, D.J.l; Kroeze, W.K. Magic shotguns *versus* magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discov.*, **2004**, *4*, 353-359.
[54] Yeh, C.T.; Wu, A.T.; Chang, P.M.; Chen, K.Y.; Yang, C.N.; Yang, S.C.; Ho, C.C.; Chen, C.C.; Kuo, Y.L.; Lee, P.Y.; Liu, Y.W.; Yen, C.C.; Hsiao, M.; Lu, P.J.; Lai, J.M.; Wang, L.S.; Wu, C.H.; Chiou, J.F.; Yang, P.C.; Huang, C.Y. Trifluoperazine, an antipsychotic agent, inhibits cancer stem cell growth and overcomes drug resistance of lung cancer. *Am. J. Respir. Crit. Care Med.*, **2012**, *186*, 1180-1188.

**Received: January 20, 2015** **Revised: May 18, 2015** **Accepted: June 17, 2015**

*Send Orders for Reprints to reprints@benthamscience.ae*

*Current Pharmaceutical Design*, 2016, *22*, 000-000 

# Harnessing Polypharmacology with Computer-Aided Drug Design and Systems Biology

Henri Wathieu[1], Naiem T. Issa[1], Stephen W Byers[1,2] and Sivanesan Dakshanamurthy[1,2,*]

[1]*Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington DC, 20057 USA;* [2]*Department of Biochemistry & Molecular Biology, Georgetown University, Washington DC, 20057, USA*

Sivanesan Dakshanamurthy

**Abstract:** The ascent of polypharmacology in drug development has many implications for disease therapy, most notably in the efforts of drug discovery, drug repositioning, precision medicine and combination therapy. The single-target approach to drug development has encountered difficulties in predicting drugs that are both clinically efficacious and avoid toxicity. By contrast, polypharmacology offers the possibility of a controlled distribution of effects on a biological system. This review addresses possibilities and bottlenecks in the efficient computational application of polypharmacology. The two major areas we address are the discovery and prediction of multiple protein targets using the tools of computer-aided drug design, and the use of these protein targets in predicting therapeutic potential in the context of biological networks. The successful application of polypharmacology to systems biology and pharmacology has the potential to markedly accelerate the pace of development of novel therapies for multiple diseases, and has implications for the intellectual property landscape, likely requiring targeted changes in patent law.

**Keywords:** ??????????????????????????????????.

## 1. INTRODUCTION

Many approved drugs, though originally intended for specific mechanisms of action, often exhibit multi-target profiles that likely contribute to their efficacies [1]. Describing the extent of this polypharmacology is desirable for the high-resolution mechanistic knowledge of how an active compound might perturb specific diseases [2]. The promise of polypharmacology is to achieve an expansive understanding of the molecular mechanisms and response phenotypes for a given drug candidate, in order to select disease-modulated biological factors such as proteins (on-targets) while avoiding those that cause toxicity (off-targets). Such a paradigm of simultaneous breadth and precision necessitates the large-scale consolidation of data relating drug effects and disease effects on complex physiological networks. Moreover, it is of paramount importance to develop computational platforms that utilize these networks to prioritize drug candidates exhibiting polypharmacology with the most therapeutic and least toxic potential.

It is generally recognized that a "reductionist" approach, which seeks to develop drugs affecting a single disease-related molecular entity, neglects the multi-genic and multi-pathway nature of many disease mechanisms. Single-target drug development has yielded some key successes, but many "specific" drugs are now known to be considerably less selective than previously thought [3]. The apparent inevitability of polypharmacology may be the result of protein modification in an evolutionary past marked by high chemical diversity, wherein it offered an adaptive survival advantage [4]. Reductionist approaches coincided with the concern that more promiscuous drugs would inevitably cause unpredictable side effects [5]. The recent emergence of powerful new technologies and hubs of data with which to characterize drug and disease signatures alike allows for increased understanding and cataloging of multi-target activities [6]. It is apparent that despite remaining concerns associated with promiscuity, embracing polypharmacology is becoming an important part of contemporary drug discovery [7].

*Address correspondence to this author at the Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington DC, 20057 USA; Tel: 202-687-2347;
E-mail: sd233@georgetown.edu

Many qualities of therapeutic agents exhibiting polypharmacology make them rather attractive compared to drugs bind to single protein. It has been posited through network models that multi-targeted agents, which often bind at low affinities, can be more efficient in partially inhibiting a small number of proteins, causing the distributed attenuation or amplification of a biological network [8, 9]. Paradoxically, this low-affinity multi-protein activity can cause fewer side effects compared to those resulting from a fully inhibited single protein and its affected downstream components [8]. Another important advantage of multi-targeted approach is the increased propensity to delay or prevent drug resistance, especially in the case of malignancies in which mutator phenotypes result in rapid adaptation and the emergence of drug resistant clones [7]. The rational design of a polypharmacological agent may seek to target multiple essential functions to overcome known or predicted compensatory signaling pathways employed by a disease, lessening the probability that disease mechanisms will circumvent drug actions [10]. Polypharmacology may therefore allow for a level of control previously sought in reductionist endeavors, which suffered from a narrow toolkit by comparison.

The conceptual framework of polypharmacology is considered in some emerging drug therapy realms but not others. Here, we comment on those of drug repositioning and precision medicine. Drug repositioning, or the identification and use of current drugs for new medical indications, has gained a great deal of attention as a way to circumvent the limitations posed by de novo drug design [11, 12]. Successful computational prediction of alternative drug targets requires an understanding of the most essential disease mechanisms, and measures of how existing drugs may be appropriated to those mechanisms. Describing and exploiting the polypharmacology of approved drugs will play an important role in discovering opportunities for repositioning. In addition, modeling the overlap of biological system, disease perturbation, and drug action networks is a key element of systems biology-based *in silico* drug development. While disease perturbation and drug perturbation spaces are varying parameters de facto, biological system and disease perturbation spaces are too often considered as static frameworks. Progressing from this static model to a more nuanced understanding of patient and disease heterogeneity is the goal of preci-

1381-6128/16 $58.00+.00 

sion medicine [13], and computational methods in the context of polypharmacology will likely pave the way for such an exquisitely tuned approach to therapy.

In this review, we provide an overview of recent efforts that have been made to navigate two major bottlenecks in the productive employment of polypharmacology. The first bottleneck is the formulation of ligand-protein interactions, for which we consider some currently available computer-aided drug design methods. These broadly include structure-based approaches, chemoinformatics and protein-based methods, QSAR and proteo-chemometric approaches, and text mining strategies such as natural language processing. The second bottleneck has been the extension of ligand-protein interactions to a biological network-based understanding of drug action, and the application of that understanding by computational means to predict drug response phenotypes and therapeutic efficacy in treating diseases. We also consider the ways in which the principles of polypharmacology are being used to model the synergistic effects of combination therapies against a given disease. Lastly, we reflect on the current limitations and future prospects for the computational harnessing of polypharmacology and the implications for the intellectual property landscape.

## 2. POLYPHARMACOLOGY AND COMPUTER-AIDED DRUG DESIGN

Annotating the "one drug - many target" space with high confidence is a challenging task for the drug development community. While high-throughput biological screening has been helpful in identifying drug-disease interactions [14], costs are prohibitive for many investigative groups, and assays may not necessarily reflect true biophysical processes or provide enough structural and/or mechanistic insight to move rapidly to Investigational New Drug (IND) Applications. Computer-aided drug design (CADD) has had a profound impact on efficiently assessing hundreds of drug-protein possibilities and engaging a broad research community. In this review, we first discuss CADD methods (Fig. **1**) that have been of great utility for establishing drug-protein interactions, and some notable efforts that have exploited such methods in a polypharmacology framework.

### 2.1. Structure-based Approaches

Over the past decade, structure-based approaches have resulted in the successful prediction of a wide variety of ligand-protein associations. These approaches rely on the availability of three-dimensional (3D) structural data of proteins. Protein structures may be experimentally derived and deposited in the Protein Data Bank [15], or predicted through homology modeling [16]. Millions of ligand structures can be found through publicly available databases such as Zinc [17], DrugBank [18], and PubChem [19]. In structure-based approaches, the goal is to leverage information contained within 3D structures for formulating ligand-protein associations.

Docking is a structure-based method to virtually screen large compound libraries and obtain a rank-ordered list of molecules based on their binding potential to a protein of interest [20]. Many algorithms differ in how a molecule is fit within the binding pocket or secondary binding site [21], but the common underlying premise is that molecules that are able to achieve more negative free energies of binding are assumed more likely to bind. Docking is a popular choice due to its computational efficiency in screening large molecule libraries. As the number of potentially synthesizable molecules is growing exponentially, there has been increased interest in massively parallelizing docking [22] or an accelerating process using hardware such as GPUs [23].

The use of docking has led to several notable successes in drug discovery and repurposing. For instance, using docking, the antipsychotic haloperidol has been repurposed as a highly selective HIV-1 and HIV-2 protease inhibitor [24], and the anti-leukemia

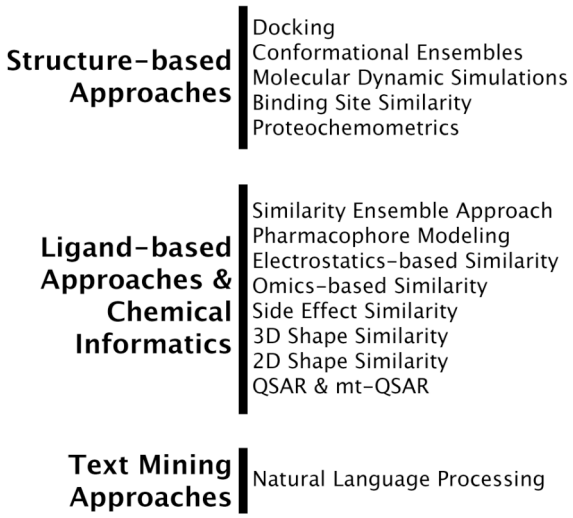



**Fig. (1). Computer-Aided Drug Design Strategies.** Computer-Aided Drug Design (CADD) tools described in this review can be functionally categorized into structure-based approaches, ligand-based approaches & chemical informatics, and text mining approaches. Structure-based approaches utilize protein-based information, sometimes in conjunction with ligand information. Ligand-based approaches are predicated on common ligand characteristics that predict putative proteins, and informatics are often required to link those characteristics for drug-proteins interaction predictions. Text mining approaches seek to extract known or predicted drug-proteins information from the text.

ABL Proto-Oncogene 1, Non-Receptor Tyrosine Kinase (ABL1) inhibitor nilotinib was found to inhibit Mitogen-Activated Protein Kinase 14 (MAPK14) [25]. Mendez-Lucio *et al.* recently characterized the anti-inflammatory drug olsalazine as a DNA methyltransferase inhibitor, indicating its possible repurposing for epigenetic modifications [26]. In the realm of oncologic diseases, the antipsychotic drug fluspirilene was found to disrupt the interaction of MDM2 Proto-Oncogene, E3 Ubiquitin Protein Ligase (MDM2) with Tumor Protein P53 (TP53) [27]. Chen *et al.* also utilized docking to find that the active ingredient of the Chinese traditional medicine Danshen, Transhinone IIA, bind to retinoid acid receptor alpha (RARA) in acute promyelocytic leukemia [28]. Docking has also begun to emerge in a polypharmacology framework, and was used to identify the tricyclic anti-depressant protriptyline as potent multi-target inhibitor of the Alzheimer's disease-associated proteins acetylcholinesterase (ACHE), beta-secretase (BACE1), and amyloid-beta aggregation [29]. More recently, Banerjee *et al.* screened newly designed ligands, to identify dual inhibitors against FabG4 and HtdX that could constitute promising agents against drug resistant Mtb strains or latent stage tuberculosis [30].

While conventional docking has aided in the identification of multiple protein targets for many approved drugs, it has drawbacks [31-33]. These mainly relate to a lack of protein structure flexibility and modeling of solvation effects. As proteins can adopt many conformational states, the use of a single conformation in a docking study may not be sufficient to account for all possible ligand chemotypes and binding states *in vivo*. One remedy is to employ conformational ensembles in docking, allowing for enriched virtual screening outcomes [34]. An example of this is in the investigation of undecaprenyl diphosphate synthase enzyme (UPPS), a bacterial protein essential for cell wall biosynthesis [35]. Experimental crystal structures of UPPS show that it is able to adopt three distinct conformations depending on substrate or ligand characteristics. Docking was performed for a set of 112 known active and inactive compounds on each conformation, and two of the three conforma-

tions (open and ajar states) were found to provide the best performance with respect to ranking active compounds over inactive compounds. In this case the use of these conformations for subsequent screening of large compound databases resulted in the discovery of UPPS inhibitors with promising antibiotic activity [36]. In such cases where crystal structure diversity is available for a protein of interest, use of a conformational ensemble leads to better virtual screening outcomes in docking. However, for proteins with low crystal structure diversity or no identified structure, computational molecular simulations can be used to explore the conformational space of proteins using an initial crystal structure or homology model.

As noted above, multi-targeted drugs often have lower affinity to multiple proteins [8], and such wide-reaching but weaker interactions may be instrumental to clinical efficacy. Inadequate simulation of physiological conditions in structure-based methods leads to false negatives, particularly for weaker ligand-protein interactions [8]. These deficiencies are particularly important for polpharmacology. To address this concern, molecular dynamic (MD) simulations allow for studying protein flexibility in implicit or explicit aqueous (i.e. cytoplasmic) and lipid (i.e. plasma membrane) environments [37]. Long time-scale simulations allow for the identification of diverse conformations that a protein may adopt, even rare conformational states that are transient. High- and low-populated conformations obtained from MD simulations can be used to represent an ensemble for virtual screenings. Sinko *et al.* identified a rare conformational state of UPPS through long MD simulations that allowed for the identification of a class of inhibitors different from other those for other conformations [38].

MD simulations have also been extensively used for studying protein allostery. This is exemplified by G protein-coupled receptors (GPCRs), where allosteric modulators are able to bias signaling toward G protein or beta-arrestin signaling pathways [39]. Pharmaceutical-induced signaling bias of GPCRs has shown clinical potential, as in the case of angiotensin II type 1 receptor (AGTR1) in cardiovascular disease where ligand TRV120027 competitively antagonizes G protein signaling but stimulates the recruitment of beta-arrestin to increase cardiomyocyte contractility and reduce mean arterial pressure [40]. Dror and colleagues employed MD simulations to mechanistically explore the association of allosteric modulating ligands to the M2 muscarinic acetylcholine receptor (CHRM2) [41]. This knowledge will pave the way for future structure-based virtual screening of GPCR allosteric modulators and potentially identify drug repurposing opportunities.

Structure-based approaches encompassing docking and molecular dynamics are critical in establishing ligand-protein associations in polypharmacology. With the development of more accurate atomistic models and high-performance computing, as well as the crystal structure or homologous template characterization of large numbers of proteins, these methods will achieve greater accuracy in their predictions and can be implemented on a large scale. Consideration of allosteric mechanisms also will unlock many new drug-repurposing avenues.

### 2.2. Chemoinformatics and Protein-based Methods

Chemical informatics approaches have been popular for large-scale computationally efficient screening to identify new ligand-protein associations. These methods focus on compound-specific properties, which may be structural, topological, chemical, physical or biological attributes (i.e. induced gene expression or proteomic changes). Many software packages, both open source and proprietary, that calculate structural and physiochemical properties such as substructure fragments, atomic connectivity indices, electronegativity, solvent-accessible surface area, electron affinity, solubility, among hundreds of others are available. These include Dragon [44], QikProp [45], and others. It is typically assumed that molecules with similar properties tend to interact with similar proteins [46]. In addition, methods based on ligand to ligand and ligand-binding pocket shape similarity are also useful given the importance of shape complementarity in biological associations [47].

Use of chemical informatics and 3D shape similarity metrics have been successful in identifying alternative targets for existing drugs. The cardiovascular drug S-bepridil was found to have potential anti-cancer properties through interaction with MDM2, a negative regulator of TP53, using a similarity method that combines both three-dimensional shape and chemical properties [48]. Vasudevan *et al.* utilized shape screening to find chlorprothixene and promazine as antagonists of the H1 histamine receptor (HRH1) [49]. The similarity ensemble approach (SEA) [50] is another chemical similarity-based method that has been successful in predicting drug polypharmacology. SEA identified the opioid drug methadone as an antagonist of the muscarinic M3 receptor (CHRM3), ementine as an adrenergic alpha2 antagonist, and loperamide as an NK2 (TACR2) antagonist. SEA has also been used to identify drug-protein interactions that lead to adverse drug reactions (ADRs) [51], as discussed later, as well as phosphodiesterase-4 (PDE4) to be a novel target of the approved angiotensin converting enzyme (ACE) inhibitor moexipril [52].

Drug-induced biological effects have also been leveraged to discover new interactions. Protein targets are identified using a "guilt by association" approach [54] where drugs that exhibit a biological profile similar to that of a drug or active compound with known targets are inferred to also interact with those targets. Gene expression analysis and transcriptomics have been particularly fruitful here for finding similar targets for structurally diverse ligands [55]. The Connectivity Map (cMap) [56] is a publicly accessible database of drug-induced gene expression changes across multiple cell lines from which similarity analyses are conducted. Li *et al.* used cMap to identify 148 targets for 20 polyphenols, showing polypharmacological mechanisms that extend beyond anti-oxidant activity [57]. Babcock *et al.* discovered novel hERG (KCNH2) binding properties for fendiline, cloperastine, ethopropazine, and sulconazole through cMap [58]. Other investigations of gene expression profiles have found potential repurposing of quinacrine, an anti-protozoal, for myeloid leukemia by modulating RNA polymerase I to affect ribosomal biogenesis [59]. Other biological manifestations, such as side effect similarity [60] or biological network perturbation characteristics [61-63], have also been employed as parameters for deducing drug-protein associations. In one study, chemical, side effect, and therapeutic similarities were combined to derive drug-protein interaction predictions, and as a result seven alternative targets for seven approved drugs were found [53].

Pharmacophore modeling is another informatics method for identifying ligand-protein interaction signatures from chemical features. A pharmacophore is an abstract representation of molecular features important for drug binding in three-dimensional space. These features include hydrogen bond acceptor and donor groups, hydrophobic areas, aromatic centers, and others, and their optimal positioning for bio-molecular recognition. Pharmacophore models can be derived in multiple ways: (1) from crystalized protein complexes, (2) a set of known active molecules, and (3) apoprotein structures. When using ligand-protein complex structures, pharmacophores can be obtained by assessing which molecular features are most important for that particular binding mode. They can for instance be determined through energy optimization, as is done with the e-Pharmacophore module in the Schrodinger software suite [64] or GALAHAD provided by Tripos [65]. In the absence of a protein structure, pharmacophore models can be generated from an active molecule based on shared chemical features and their alignments e.g. using GASP [66]. When ligand information is absent but a protein structure is available, pharmacophores can be generated by inferring complementary functional interactions from important residues within a proposed binding pocket. This is considered a receptor- or structure-based pharmacophore and software created

for this purpose include FLAP [67], and SNOOKER [68]. Pharmacophore models are then used in virtual screening to identify active compounds that are able to recapitulate important functional contact points. Additional use of exclusion spheres that mark the atomic radii of binding site residues can further filter out molecules that are too big or that may form steric interactions with a protein, thus acting as a sort of molecular sieve. Pharmacophore screening has resulted in many novel drug-protein associations for potential repurposing. Ai *et al.* discovered the negative modulating activity of nitazoxanide, an anti-protozoal agent, on both mGluR1 (GRM1) and mGluR5 (GRM5) receptors [69]. Levit *et al.* identified the FDA approved drugs glimepiride and salsalate, a second-generation sulfonylurea used for the treatment of type 2 diabetes mellitus, as an agonist of the human bitter taste receptor type 2 member 14 (TAS2R14) [70]. Krautscheid *et al.* used pharmacophore modeling to establish the unexpected associations of haloperidol, eprazinone, and fenbrutazate with neurokinin receptors [71]. Rolofylline, an adenosine A1 antagonist, was also found to be a micromolar inhibitor of cGMP-specific phosphodiesterase type 5 (PDE5) [72].

Like pharmacophore screening, electrostatics-based virtual screening methods have also revealed novel biological associations. As the molecular electrostatic potential (MESP) distribution is critical in binding and high-resolution molecular recognition [73, 74], it is assumed that protein bind molecules exhibiting similar ESP distributions or atomic partial charges [75]. Armstrong *et al.* developed a similarity-based ligand screening method called ElectroShape that combines partial charge information at atom-occupied coordinates with three-dimensional shape [76]. Shape and electrostatic similarity was also used by Muchmore *et al.* to identify a novel antagonist of melanin-concentrating hormone receptor 1 [77]. The commercial software EON (www.eyesopen.com/eon) is also available for coupled shape- and electrostatics-based molecular similarity screening and has been successfully used to find many ligand-protein- such as those of 5-(4-piperidyl)-3-isoxzaolol as a potent fibrinolysis inhibitor [78].

Protein-based methods are also used to identify ligand-protein associations, where ligands are assumed to bind protein of similar properties [79]. In particular, binding site shape has been exploited in drug repurposing. De Franchi *et al.* used the SiteAlign [80] algorithm to discover staurosporine as a potent inhibitor of synapsin I, which regulates neurotransmitter release, due to its pocket shape similarity to Pim-1 kinase [81]. Kinnings *et al.* employed the Sequence-Order Independent Profile-Profile Alignment (SOIPPA) algorithm [82] and found entacapone, an anti-Parkinson drug that bind to catechol-O-methyltransferase (COMT), to inhibit enoyl-acyl carrier protein reductase (ENR), a therapeutic target for Mycobacterium tuberculosis [83]. SOIPPA was also utilized by Durrant *et al.* to predict targets for an inhibitor of Trypanosoma brucei RNA editing ligase 1 based on binding site shape even though the proteins did not show global similarity in sequence or structure [84]. In addition, protein-based electrostatic properties can be used for successful predictions. Voet *et al.* exploited the concept of electrostatic complementarity between ligand and protein [85, 86]. They developed an algorithm called Elekit for identifying small molecule inhibitors of protein-protein interactions based on the electrostatic potential distribution at the protein-protein interface [87].

In mapping the polypharmacology of a given ligand, it is crucial to note that binding site similarities may be more important than global structural or sequence similarities of potential protein [42]. Candidate ligands are often tested against a large panel of related proteins, to determine the relative selectivity of two active compounds that act as inhibitors of a given protein subclass. This strategy considers the mechanistic description of compound promiscuity as an opportunity to fine-tune the prediction of clinical efficacy. Möller-Acuña *et al.* recently investigated the observed polypharmacological action of the drug SB-206553 on the structurally and functionally dissimilar 5-HT2 and α7 nACh receptors [43]. Using docking and molecular dynamics methodologies in concert, they found that each receptor had a binding site with hydrophobic pockets of chemically and structurally similar residues that could explain common affinity by SB-206553 [43]. Interestingly, SB-206553 acts as an inverse agonist of 5-HT2Rs and as a positive allosteric modulator of α7 nAChR, and this dual activity may account for its anxiolytic and anti-addictive properties [43].

The chemoinformatics and protein-based approaches described here are commonly used with other modalities to enrich virtual screening as each approach has its own strengths. For example, a combination of pharmacophore models, homology modeling and docking led to the identification of novel inhibitors of Topoisomerase II-alpha (TOP2A) from the NCI2000 drug database [88]. Dobi *et al.* performed a cascade screening where a 2D similarity search was performed after pharmacophore matching to find potent antagonists of the 5-$HT_6$ receptor (HTR6) implicated in neurologic disorders [89]. Markt and co-workers utilized a virtual screening workflow that integrated pharmacophore modeling with 3D shape and electrostatic similarity screening to establish novel PPAR ligands for cardiovascular diseases [90].

### 2.3. QSAR and Proteo-chemometric Approaches

Quantitative Structure-Activity Relationship (QSAR) methods employ regression modeling and machine learning to classify compound activity against single protein classes and identify chemical and structural parameters that are associated with binding [91]. These molecular descriptors may be different from one model to another, depending on the test set used. Models derived from a bioactivity data series with respect to a protein are then used to predict binding affinities of new compounds to that protein. Many chemical-centric QSAR models have been developed over the last decade and have been successful in identifying novel drug candidates for well-established proteins [92]. Examples include indole aryl sulfones against the HIV-1 reverse transcriptase non-nucleoside binding site [93] and capsazepine as an anti-inflammatory through blocking tumor necrosis factor alpha (TNF) [94].

Although success has been achieved using QSARs, the applicability domain (AD) is generally limited to that single protein or chemical congener series [95]. For instance, a QSAR model created for a Class I GPCR may not be applicable to a different GPCR class or even a GPCR from within the same class. This is primarily due to the training set used, binding mode, and experimental conditions used to obtain binding affinity values. Furthermore, QSAR models may not be accurate for the same protein containing different single nucleotide polymorphisms (SNPs) that alter protein structure and binding properties. To address these problems, proteochemometric and multi-target QSAR approaches have been developed.

Proteochemometric (PCM) models integrate protein -based descriptors, such as amino acid sequence, with chemical-centric descriptors [96]. In contrast to classical QSAR approaches, which are oriented to one protein, PCM models are derived from a set of many proteins and ligands simultaneously and attempt to consider the entire ligand-protein interaction space to help overcome "activity cliffs" [97, 98], situations where similar drugs do not result in similar activities [99]. This is particularly useful for discovering drugs that interact with proteins harboring a particular mutational status [81]. PCM modeling has been successfully used for G protein-coupled receptors (GPCRs) [101], HIV-1 protease [102], kinases [103], penicillin-binding proteins of infectious agents [104], along with many others. Our group established novel proteochemometric methods called TMFS and RepurposeVS [105]. TMFS and RepurposeVS differ from other PCM models in that they leverage information from single drug-protein crystal structure complexes to find drugs that are able to recapitulate those properties. This is especially useful for exploiting differential binding modes within a single protein, such as agonist and antagonist conformations. For example, the estrogen nuclear receptor (ESR1) has

been co-crystallized with agonists such as diethylstilbestrol as well as antagonists such as 4-OH tamoxifen. Leveraging those individual structures permits TMFS to find approved drugs that could either activate or inhibit estrogen receptor signaling. TMFS has allowed us to identify clinically important drug repurposing opportunities. One is the repurposing of mebendazole, an anti-hookworm tubulin inhibitor, as an anti-angiogenic agent through inhibition of vascular endothelial growth factor receptor 2 (VEGFR2) kinase domain. Another is the repurposing of the anti-inflammatory cyclooxygenase-2 (COX-2) inhibitor celecoxib and its analog dimethylcelecoxib (DMC) as inhibitors of cadherin-11 (CDH11), a cell adhesion molecule implicated in poor-prognosis malignancies and rheumatoid arthritis.

In addition to traditional QSAR and PCM methods, multi-target QSAR (mt-QSAR) models have been developed to calculate the probability of compound activity on multiple protein targets, such as kinases [106], GPCRs [107], bacterial strains [108], or even a combination of different target types [109]. Machine learning methods such as artificial neural networks (ANN), Markov models, support vector machines machines (SVMs), linear discriminant analysis, and others are integrated to discover important biological and chemical properties and can be used to model a particular set of proteins. Liu *et al.* utilized mt-QSAR to identify co-inhibitors of HIV and HCV viruses [110]. Garcia *et al.* discovered glycogen synthase kinase 3 beta (GSK3B) inhibitors that could treat Alzheimer's disease as well as parasitic infections [111]. Speck-Planche *et al.* were able to use mt-QSAR for the simultaneous prediction of anti-infective compounds along with their toxicological profiles [112]. Mt-QSAR is not limited to proteins, but can also be applied to larger systems such as cell lines and bacterial species. For example, Speck-Planche *et al.* built a mt-QSAR model for ten colon cancer cell lines to identify anti-colon cancer agents [108]. Similarly, Prado-Prado *et al.* built the first unified mt-QSAR model of using artificial neural networks that predicts compound activity against different parasite infections using 500 drugs tested against 16 parasite species found in the literature [113].

### 2.4. Natural Language Processing

The computational methods described above far provide efficient screening for predicted biological targets. Analyses to support CADD, however, often rely on access to "Big Data" that are already available.

In practice, the accessibility of existing information is a major limitation to data analysis in bioinformatics. Databases such as those described in this review have been developed to facilitate this hurdle, but much of the desired data are accessible only as a reference, and not as a repository. Empirically determined biological activities, in particular, are strewn over a growing mass of scientific literature. Such information, presented in "natural language" form, is bound by free text and is unstructured, carrying semantic variation and ambiguity. Despite the immediacy of such a valuable resource, the volume of research publications alone is such that manual human curation of information from these sources, in a manner that is comprehensive and up-to-date, is not feasible [114].

Natural Language Processing (NLP) is a method of information extraction (IE) within text mining that addresses the problem of accessibility by extracting structured and meaningful information from natural human language by computational means. In the context of CADD and polypharmacology, the central challenge of NLP is to accurately identify biologically or chemically relevant components from text and determine their semantic associations.

To extract ligand-protein associations from a text of heterogeneous syntax and semantic features, NLP typically divides the text into word boundaries, called tokens [115]. After identifying these tokens and their parts of speech, parsing tasks use syntactic components to validate token sequences and thus extract relationships between entities of a desired type [116]. Rule-based and statistical parsing are the two overarching approaches to building an NLP system [116]. A rule-based approach implements hand crafted rules of grammar, whereas a statistical approach utilizes probabilities to train machine-learning algorithms based on previously annotated scientific literature (corpora), ultimately to determine the most likely parse of a sentence or phrase [116]. A statistical system can also be built conjointly with a rule-based system.

NLP technologies vary in complexity and approach. Many of the existing efforts to improve NLP modeling consist of addressing a single subtask of NLP. Named Entity Recognition (NER), for instance, is required for any IE endeavor. In biology, however, the inconsistency of nomenclature means that identifying relevant entities in a standardized way is an especially arduous task [117]. The Critical Assessment of Information Extraction system in Biology (BioCreAtIvE) community-wide competition is a recent and ongoing effort that has brought about both annotated training corpora and NER tools for literature mentions and relationships of genes, drugs, and other entities [118]. One such outcome of BioCreAtIvE is CHEMDNER, a new corpus that can be used for training chemical entity taggers in NLP models, was created to support the building of statistical models that can identify and classify entity mentions [119]. The corpus entails 3, 000 manually annotated ("Gold Standard") PubMed abstracts, 17, 000 automatically annotated ("Silver Standard") abstracts, and more [119].

Building upon early attempts [120], some current databases already employ NLP to link components relevant to systems biology, usually along with other manual or computational extraction methods. STITCH is a repository of protein-chemical and chemical-chemical interactions that derives its information from experimental and manually curated data sources, in addition to extraction from the literature using NLP and other text mining strategies [121]. STRING is a related database focusing on direct and indirect protein-protein interactions (PPIs), and incorporates statistical entity co-occurrence analyses on large quantities of full text articles, as well as rule-based NLP tasks such as part-of-speech tagging, semantic tagging, and formula-based grammar [122, 123]. SIDER contains adverse drug reactions (ADRs) associated with a given drug based on FDA package inserts, along with frequency information for each ADR [124]. An NER strategy was developed to extract mentions of relevant components from the package inserts, using STITCH and STRING databases for drugs and proteins, and the UMLS Metathesaurus and MedDRA for ADRs and diseases [124]. The Stanford Dependencies [125] NLP tool served the development of SIDER by parsing sentences that denote a disease indication for a given drug [124]. In this context, NLP can be particularly useful in, for example, separating ADRs from drug effects that are contingent on pre-existing conditions, and other tasks that require a high degree of context-dependent semantic parsing.

NLP may also be used directly in a drug development effort predicated on polypharmacology. Yu and colleagues recently revealed the polypharmacology of mifepristone (RU486), a synthetic steroid that has clinically established anticancer properties and could be effective as a cancer metastasis chemopreventive [126]. In addition to running multiple assays relating, among other parameters, the promising cell adhesion and migration effects of mifepristone, the authors identified 513 genes affected by the drug using NLP, and then carried out functional interpretation of these proteins using pathway and GO analyses [126]. This NLP endeavor required gene mention tagging of full text articles using the open source biological entity NER tool ABNER [127], followed by extraction of multiple genes that were combined into a single term, and gene name normalization [126]. Finally, a hypergeometric distribution was applied to narrow down genes that co-occurred with mifepristone with sufficient statistical significance [126]. Zeng *et al.* provide a helpful survey of NLP techniques utilized in bioinformatics [127]. Presently NLP is mainly employed in the curation of databases. Later in this review, we consider the practical applications of

such a growing catalog of biological associations in the context of polypharmacology and drug repositioning.

## 3. POLYPHARMACOLOGY AND NETWORKS

Attempts to identify and exploit biological targets have often failed to consider the fluctuating nature and downstream effects of targeted components at multiple stages of biological action, from molecular mechanisms to observed phenotypic response [6]. Lacking a fuller understanding of action of drugs, in many cases, the unanticipated side effects, drug resistance, lack of therapeutic efficacy, and other mishaps that are all-too-often encountered in drug development [7].

Networks carry the potential to support a topological rendition of all atomic/molecular, macromolecular, cellular, tissue, organ, and organismal biological features, as well as the many dynamic relationships that exist between them [6]. Upon bridging native biological networks, one can investigate how such networks are perturbed by diseases and modulated by pharmacological agents, and thus the discussion of polypharmacology necessarily arises (Fig. **2**). In this higher-resolution context, polypharmacology can be extended to signify multiple immediate targets and their downstream biological components. This level of precision also requires a sense of the qualitative and, ideally, the quantitative relationships between these components. To develop the most powerful tools for predicting therapeutic efficacy of an and its greater response phenotype based on polypharmacology, a network approach therefore seems essential.

### 3.1. Predicting Drug Response Phenotypes with Polypharmacology

As previously addressed, the inherent promiscuity surrounding drugs that exhibit polypharmacology is of some concern to the drug development community. Specific proteins are known to be therapeutically useful, but just how drugs cause toxicity or side effects has been somewhat unclear in the past, along with the extent to which it is helpful to invoke a simplistic causal connection between single target and single observed effects [9]. Nevertheless, distinguishing between on- and off-targets, and identifying and characterizing the off-target effects, has been a major undertaking in polypharmacology. Much of it has been accomplished by statistically linking off-target drug actions and drug response phenotypes.

Torcetrapid, a Cholesteryl Ester Transfer Protein (CETP) inhibitor, had promising characteristics for the treatment of cardiovascular diseases (CVD), but resulted in deadly off-target hypertensive effects and was withdrawn in phase III clinical trials [128]. Xie *et al.*, without any ADR-related data, predicted off-targets of CETP inhibitors using structure-based methods. The off-targets were found to be members of lipid metabolism and signaling pathway networks that modulate processes linked to the known adverse effects [129]. Chang *et al.* also attempted to explain the adverse effects of torcetrapid, instead by building an *in silico* reduced kidney metabolic model, manifesting itself as a network from which predictions about gene activity in the kidney could be derived [130]. Upon predicting putative torcetrapid off-targets, they found that many of them perturbed the renal function model and had formerly been shown to impact renal function in patients with corresponding gene deficiencies [130]. Such studies signaled the potential role that a systems biology and polypharmacology approach could have in drawing a link between the modulation of specific genes, or sets of genes, and the development of certain phenotype responses. This direction could more easily address drug toxicity a priori, and therefore streamline the process of drug development.

Various systems biology approaches offer the power to predict drug effect profiles, dependent only on knowledge of protein -level mechanisms. Some key efforts have taken extensive network approaches to identify ADRs that were disproportionally found in drugs predicted and known to bind to a given protein, resulting in drug-protein-ADR networks [131, 132]. Kuhn *et al.* noted that conducting an overrepresentation analysis in this way could easily result in false positives, because drugs tend to bind to sets of pharmacologically similar proteins, while only one of those may be the driving force behind a side effect [133]. Focused on narrowing

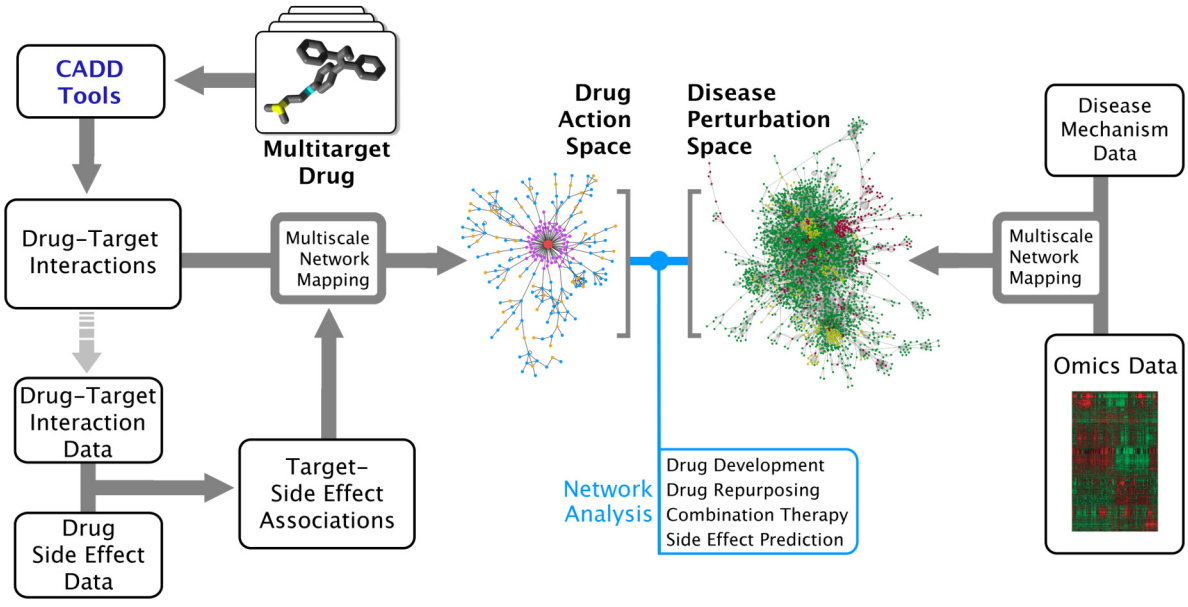



**Fig. (2). Using Polypharmacology in Network Analysis.** CADD Tools allow for the polypharmacological characterization of multi-target drugs, resulting in a panel of drug-protein interactions for each drug. These interactions may be mapped to associated or enriched multi-scale biological factors to a network of drug action space. Curated interactions may also be integrated with drug side effect data to make predictions about target-side effect associations, which may in turn contribute to the drug action space network. Disease-related omics data or known mechanistic data can be analyzed to derive putative therapeutic biological factors, and multi-scale network mapping allows for the creation of a disease perturbation space. Using network analysis, the coincidence of the disease perturbation and drug action spaces may reveal potential de novo or repositioned indications, combination therapies, and side effect predictions.

down these causal relationships, they confirmed in a mouse model that a selective drug could often counteract the side effect induced by a previously administered drug, if it had an opposite effect on the primary protein [133]. Furthermore, they successfully mapped known gene-phenotype pairs from knockout mice studies to their own predicted target-side effect pairs [133]. They thus developed a drug-protein-ADR network, and predicted that 732 of 1428 studied side effects were largely caused by, individual proteins [133].

A few studies have attempted to integrate quantitative characteristics of ligand-protein interactions to predict phenotypic responses, corresponding to a more comprehensive network topology that affords value to edges connecting the nodal biological components. Simon *et al.* used docking to predict targets for 1226 FDA-approved drugs among 149 protein candidates, and correlated by canonical correlation analysis (CCA) the calculated binding free energies for each protein with the presence or absence of 177 known drug effect categories for a given drug [134]. The primary advantage of using *in silico* binding fingerprints rather than experimentally verified derived binary binding annotations is that it allows for the incorporation of potentially important weak interactions, as well as nonspecific binding of the compound to a receptor, and therefore provides a more comprehensive topological network space from which to mechanistically resolve drug promiscuity and derive phenotype associations.

Duran-Frigola *et al.* did not limit their side effect network characterization to mapping proteins [135]. Rather, they examined chemical and network biological characteristics and statistically deduced, for each drug, which factors best revealed the molecular bases for a side effect [135]. 6% of examined drug-ADR associations were best explained by the chemistry of the compounds, suggestive of nonspecific drug actions [135]. While this method deviates from a strictly polypharmacology-based interpretation of drug action, it does afford value to integrate quantitative chemical parameters accounting for nonspecific binding events that may not yet be modeled *in silico* using biological networks, but nevertheless influence drug response. Using such integrated strategies may increase predictive power, especially considering the success in linking specific chemical fragments to drug side effects [136]. Recently, Pérez-Nueno integrated their physiochemical property-based tool for the prediction of polypharmacology [137] with known side effect data and carried out a correlation and discrimination procedure [138]. This integrative tool, called GESSE, ultimately predicts side effects from the 3D structure and chemical properties of a drug, and elucidates the polypharmacology of that drug at multiple levels of biological action, such as pathways and cellular functions [138]. Such integrative methods will be essential to achieving maximal mitigation of shortcomings in the process of drug development.

### 3.2. Matching Disease and Drug Spaces for Therapeutic Potential

Upon acquiring a full picture of drug action, being able to implement it computationally to address a therapeutic need is the second critical bottleneck of polypharmacology. This challenge has been addressed by multiple integrative strategies, all of which have sought to determine the space of disease-regulated cellular components through "omics" technologies and subsequent analyses, or other approaches. Following this, disease-related components may somehow be prioritized for drug identification, often based on prior knowledge or statistically predicted importance. Herein lies a key obstacle; an ideal polypharmacological agent should be directed at only those proteins that need to be perturbed for clinical efficacy, and should not modify other factors. Furthermore, such an agent would target the combination of necessary molecules that results in the least toxicity. After all, on- and off-targets both carry the capacity for potential undesired effects when perturbed, but only on-targets are deemed to have therapeutic benefit. A great challenge in this undertaking, however, has been to distinguish between simply disease-modulated proteins and those that, when affected by a drug candidate, translate to a superior patient outcome.

Occasionally, groups seeking to apply polypharmacology to address a medical indication have prioritized the disease perturbation space based on previous work that determined biological network subsets to be critically important to the course of a disease [10, 139, 140]. Zhao *et al.* used networks to elucidate the polypharmacology of the medicinal herb derivative Astragaloside IV (AGS-IV) in an effort to explain its known therapeutic effects on cardiovascular diseases [141]. Interestingly, AGS-IV was found to exhibit far weaker action against key proteins compared to other CVD drugs, but had comparable effectiveness at the cellular level [141]. This indicated multiple weak interactions [141]. AGS-IV was first tested by *in silico* inverse docking within the signaling pathways known to be implicated in the actions of other drugs approved for CVDs, and 39 potential proteins were established, with three validated *in vitro* [141]. Notably, 69% of 39 putative AGS-IV had been previously associated with other CVD drugs in the literature [141]. Upon constructing PPI (Protein-Protein Interactions), drug-protein, and protein-pathway association networks for those proteins, they found that AGS-IV likely acts by modulating proteins involved in crosstalk between calcium, MAPK, and VEGF signaling, among others [141]. This approach demonstrates the usefulness of identifying proteins by testing protein candidates that are members of pathways modulated by other drugs for the same indication. This has pragmatic implications in providing both an efficient means of distinguishing between the actions of drugs for the same disease indication, and for repurposing a drug by testing its alignment to the most important pathways for the new indication.

There have been various other large-scale efforts to build clinically relevant disease perturbation networks based on the polypharmacology of drugs known to be effective against the disease in question. Seeking to derive and apply drug-protein interactions, Cheng *et al.* found that network-based inference, which uses complex network theory and topological similarity of drug-protein-disease networks, out-performed strategies hinging on protein and drug similarity parameters [62]. They discovered new polypharmacological features of some dipeptidyl peptidase-4 (DPP4) inhibitors and ER ligands, defining on- and off-targets by presence or lack, respectively, of known association with diseases in the network [62]. Arooj *et al.* recently developed a computational method to identify the off-targets of human chymase inhibitors with molecular docking [142]. Then, they employed structural and functional similarity to elucidate the roles of each off-target in biological pathways as well as their disease and phenotypic associations [142].

Xie *et al.* built a drug-abuse chemogenomics knowledgebase (DA-KB) to compile the known molecular interaction networks involved in drug abuse, particularly those encompassing GPCRs [143]. This network approach extended to mapping signaling pathways corresponding to DA-related proteins, as well as the distribution of GPCRs in human tissues and linking this information to side effects caused by abused drugs [143]. They further sought to characterize within this network the polypharmacology of DA-related protein ligands and illicit substances, and other medicines that target the central nervous system [143]. Using this polypharmacological network in concert with pharmacophore modeling, it was demonstrated that some cannabinoid ligands can interact simultaneously with cannabinoid receptor type 1 (CNR1), mu-opioid (OPRM1), and dopamine receptor D1 (DRD1) targets, allowing for the possibility of drug repurposing to mitigate cocaine craving [143]. A related study introduced a domain-specific chemogenomics knowledgebase called AlzPlatform, geared for Alzheimer's Disease [144]. These integrative efforts combine the curation of extensive disease-related biological networks from the literature, the linking of branches of this disease space to phenotypic outcomes, and a multi-faceted effort to predict new disease-related biological molecules

for both novel and FDA-approved drugs [144]. In a recent study, Jansson *et al.*. explored the potential for polypharmacology to address the highly adaptable nature of cancer, positing that addressing multiple cancer-modulated mechanisms may thwart resistance to conventional targeted therapies that often results from clonal selection [145]. They identified and characterized multiple drug-protein interactions of di-2-pyridylketone thiosemicarbazones, which they determined effectively confront the "triad of death" of cancer, which encompasses tumor growth, drug resistance, and metastasis [145]. Thus, by molecularly characterizing a few key factors of disease dysregulation that are the most clinically problematic, researchers may begin to prioritize drugs by their polypharmacology that addresses these key factors simultaneously. Applying a multi-scale network analysis to this framework may prove fruitful.

While building a disease perturbation network around biological components modulated by drugs that exhibit therapeutic action can effectively prioritize network subsets that are therapeutically relevant, it falls short of maximizing the potential of a network approach on two counts. First, this approach is not possible in diseases for which effective drug therapies do not exist, such as specific subtypes of cancer. Second, it fails to consider the likelihood that other protein targets exist which might, result in greater therapeutic effect and less toxicity. While multi-scale network analysis may help to predict alternative targets based on factors such as connectivity relative to established targets and literature-derived essential disease components as described above [143-145], the scope of this approach at the genomic level is not maximized. Fortuitously, in recent years the bioinformatics field has benefited from omics technologies that greatly facilitate the quantification of cellular variables on a large scale, allowing researchers to better build networks, and link changes in the levels or states of these variables to clinical endpoints or disease conditions using patient data [6, 9]. Analysis of omics data, in itself and paired with experimental validation, has allowed for the derivation of many important single therapeutic targets for various diseases. Omics data stands as a useful measure of the relative activities of known molecules, and linking this activity to clinical parameters, or comparing the expression of different tissues such as by differential gene expression on a large scale, may elucidate a panel of proteins that should be prioritized in parallel with an *in silico* polypharmacology-based drug development strategy.

BioProfiling.de is a web portal that supports several analytical tools for high-throughput cell biology [146]. Among these are DRUGSURV [147] and PPISURV [148], tools that apply network-based statistical analyses to deduce the effect of a drug or its direct or indirect proteins on overall patient survival in cancers, thus allowing for drug-repurposing opportunities. DRUGSURV calculates those genes from various clinical microarray expression datasets whose up- or down-regulation is significantly associated with patient survival outcomes in cancer [147]. For experimental and approved drugs, a "drug signature" network was developed that consisted of both direct and indirect protein [147]. By calculating the significance of coincidence between the protein and survival-linked gene spaces, DRUGSURV attempts to address both anticancer potential and drug efficiency in clinical trials [147]. PPISURV is a related tool, centered on the observation that the expression of a well-known cancer-linked gene is often not correlated with survival outcome [148]. The functions of TP53, for example, are mostly controlled at a post-translational level, and its gene expression does not correlate with survival in many survival types [148]. However, a statistically significant portion of its interaction partners, are positively correlated with survival in a broad spectrum of cancers [148]. This kind of phenomenon reveals the need for multi-scale biological networks in future polypharmacology-based endeavors seeking to relate drug and disease signatures.

There remains much progress to be made in both the prioritization of on-targets and the extent of multi-scale network mapping of both on- and off-targets. In a fashion similar to the correlation-based linkage of off-targets to drug response phenotypes, it is possible that linking on-targets to clinical outcomes holds promise as omics technologies progress. These measures are crucial to maximizing the power of a network approach. Meanwhile, the extension of network topology for both disease perturbation and drug action on biological networks must be developed, in conjunction with more advanced connectivity metrics.

### 3.3. Polypharmacology and Combination Therapies

Multi-target drug development shares the objective of combination therapy to simultaneously target multiple, sometimes redundant, mechanisms of disease action for an effective and durable drug response. Anighoro and colleagues, in a recent review, outline the primary advantages of using multi-target drugs over combination or standalone therapies [5]. They note that any combination therapy should ideally be comprised of two or more drugs with nonoverlapping mechanisms of therapeutic action, resistance, and toxicity [5]. Combination therapies have two potential key advantages over monotherapy, synergistic action and typically lower individual drug dosages, qualities that they have led to some key successes [149]. Presently there is a lack of *in silico* solutions for predicting synergistic drug action. The Dialogue for Reverse Engineering Assessments and Methods (DREAM) consortium recently initiated an open challenge to develop in-silico solutions [150]. Subsequently, of the 32 methods assessed for efficacy relative to established experimental combination screening data, four performed better than chance [150]. While this effort certainly illuminated many of the key characteristics of drug synergy that should be prioritized in building an *in silico* prediction model of synergistic effect, there remains much to be explored.

Vitali *et al.*. developed a tool that ranks drug pairs by a multicomponent synergistic score for combination therapies using topological features of PPI networks, and evaluated the efficacy of this tool with a gene expression-based disease network for Type 2 Diabetes Mellitus [151]. Tang *et al.*. presented anticancer combination metrics that incorporate treatment efficacy screening data and predicted drug-protein binding affinities [152].

In a recent integrative approach, Sun *et al.* combined targeting networks and transcriptomic profiles to rank chemotherapeutic agent pairs by potential synergy against three types of cancer [153]. To build a comprehensive targeting network for each drug pair, they collected known drug-protein pairs and mapped them to biological factors at the levels of PPIs, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, and Gene Ontology (GO) biological processes, defining designated cancer networks (CNs) as the disease space [153]. From 14 potential features of the molecular mechanisms of existing synergistic combinations, many of which were newly designed, only the features that were significantly different between known synergistic drug pairs and unlabeled combinations were utilized to predict synergistic potential [153]. Calculated features indicative of synergistic potential made use of connectivity characteristics of the pathway networks, as well as some molecular and pharmacological drug parameters [153]. Drug combinations were ranked according to these features, in addition to information from correlations between the gene expression profiles of cell lines perturbed by the drugs in question [153]. Such an exhaustive systems-based approach yielded significant improvements in predicting known combinational activity compared to previous platforms, including those proposed for the DREAM consortium [153]. Building strong combinatorial metrics will require network-based approaches that combine synergistic prediction, prioritization of on-targets to maximize therapeutic potential, and the target-based systematic prediction of drug response phenotypes to minimize toxicity.

## 4. CURRENT LIMITATIONS

The general principles of polypharmacology are far from understood, and the computational methods utilized to apply these principles to predict drug response phenotypes and therapeutic efficacy are in their infancy. Nevertheless, computer-aided methods for establishing drug-protein interactions have significantly improved over the past decade. Cumulatively, structure-based methods, chemical informatics, machine learning, and natural language processing have resulted in the identification of alternate targets for many known drugs. We are currently undergoing the development of increasingly precise and accurate tools from which drug-protein interactions can be derived and multi-scale networks can be built. Together with a rapid expansion of databases storing and providing this information.

Nevertheless, a significant challenge remains to fully catalogue polypharmacology, because of the lack of complete protein structural and of putative and experimental data. The availability of such data will facilitate the development of computational models attempting to make clinical outcome predictions. These may relate to drug response phenotypes by way of off-target drug action, or the calibrated effects of on-targets and avoidance of off-targets by polypharmacological agents. Beyond data availability, there is a sizable gap between the power of computational prediction and experimental validation, one that should be addressed by greater multi-scale network mapping of biological processes with quantitative modeling of all pertinent interactions. Improvements are unquestionably needed in the methods used to select proteins and drugs most therapeutically and clinically useful for a given disease. Ultimately, this would require an integration of both pharmacokinetic and pharmacodynamic facets of drug action to adequately prioritize drugs for a given indication. The lack of factors such as time-dependent ADME parameters, for instance, is a drawback to network analyses as they are currently being built.

## 5. FUTURE PROSPECTS

The rise of cloud-based computing and the resulting ability of individuals, small research groups and startups to carry out "high performance computing" and big data analytics at manageable costs is set to level the playing field and dramatically stimulate innovation in many fields [154]. In the drug discovery, polypharmacology and personalized medicine arena we are on the brink of an era in which high fidelity molecular profiling can be linked to individualized drug treatment regimens. To facilitate this approach, the advances in computation, drug screening and drug repurposing need to be better linked to electronic health records in a manner that protects patient privacy. Discussions among third party payers (Insurance Industry), drug producers (Pharmaceutical Industry) and intellectual property experts should be aligned to focus on modifying patent law to more clearly reflect the new reality that most innovation in these areas is a result of input from many partners **[155]**. We need to move away from the "prisoners dilemma" approach to invention and recognize that cooperation and sharing benefits the group as a whole.

## CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

## ACKNOWLEDGEMENTS

The authors wish to acknowledge DOD grants BC062416 and BC096277 (SB, SD), R01 CA170653 (SB, SD), CCSG grant NIH-P30 CA51008 and Georgetown Lombardi Cancer Center. SB and SD were supported by NIH-R01 CA170653 and DOD grants CA1408852, CA140882 and PC140268.

## REFERENCES

[1] Peters JU. Polypharmacology: Foe or Friend? J Med Chem 2013; 56: 8955-71.
[2] Pujol A, Mosca R, Farrés J, Aloy P. Unveiling the role of network and systems biology in drug discovery. Trends Pharmacol Sci 2010; 31: 115-23.
[3] Frantz S. Drug discovery: playing dirty. Nature 2005; 437: 942-3.
[4] Jalencas X, Mestres J. On the origins of drug polypharmacology. Med Chem Commun 2013; 4: 80-7.
[5] Anighoro A, Bajorath J, Rastelli G. Polypharmacology: challenges and opportunities in drug discovery. J Med Chem 2014; 57: 7874-87.
[6] Zhao S, Iyengar R Systems pharmacology: network analysis to identify multiscale mechanisms of drug action. Annu Rev Pharmacol Toxicol 2012; 52: 505-21.
[7] Hopkins AL Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 2008; 4: 682-90.
[8] Csermely P, Agoston V, Pongor S. The efficiency of multi-target drugs: the network approach might help drug design. Trends Pharmacol Sci 2005; 26: 178-82.
[9] Xie L, Xie L, Kinnings SL, Bourne PE. Novel computational approaches to polypharmacology as a means to define responses to individual drugs. Annu Rev Pharmacol Toxicol 2012; 52: 361-79.
[10] Apsel B, Blair JA, Gonzalez B, *et al*. Targeted polypharmacology: discovery of dual inhibitors of tyrosine and phosphoinositide kinases. Nat Chem Biol 2008; 4: 691-9.
[11] Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov 2004; 3: 673-83.
[12] Chong CR, Sullivan DJ Jr. New uses for old drugs. Nature 2007; 448: 645-6.
[13] Mirnezami R, Nicholson J, Darzi A. Preparing for precision medicine. N Engl J Med 2012; 366: 489-91.
[14] Bajorath J. Integration of virtual and high-throughput screening. Nat Rev Drug Discovery 2002; 1: 882-94.
[15] Berman HM, Westbrook J, Feng Z, *et al*. The Protein Data Bank. Nucleic Acids Res 2000; 28: 235-42.
[16] Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T. Protein structure homology modeling using SWISS-MODEL workspace. Nat Protoc 2008; 4: 1-13.
[17] Irwin JJ, Shoichet BK. ZINC-a free database of commercially available compounds for virtual screening. J Chem Inf Model 2005; 45: 177-82.
[18] Law V, Knox C, Djoumbou Y, *et al*. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 2014; 42: D1091-7.
[19] Wang Y, Xiao J, Suzek TO, *et al*. PubChem's BioAssay Database. Nucleic Acids Res 2012; 40: D400-12.
[20] Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discovery 2004; 3: 935-49.
[21] Kontoyianni M, McClellan LM, Sokol GS. Evaluation of docking performance: comparative data on docking algorithms. J Med Chem 2004; 47: 558-65.
[22] Ellingson SR, Dakshanamurthy S, Brown M, Smith JC, Baudry J. Accelerating virtual high-throughput ligand docking: current technology and case study on a petascale computer. Concurrency Computat Pract Exper 2014; 26: 1268-77.
[23] Ritchie DW, Venkatraman V. Ultra-fast FFT protein docking on graphics processors. Bioinformatics 2010; 26: 2398-405.
[24] DesJarlais RL, Seibel GL, Kuntz ID, *et al*. Structure-based design of nonpeptide inhibitors specific for the human immunodeficiency virus 1 protease. Proc Natl Acad Sci USA 1990; 87: 6644-8.
[25] Li YY, An J, Jones SJM. A computational approach to finding novel targets for existing drugs. PLoS Comput Biol 2011; 7: 1-13.
[26] Méndez-Lucio O, Tran J, Medina-Franco JL, Meurice N, Muller M. Toward Drug Repurposing in Epigenetics: Olsalazine as a Hypomethylating Compound Active in a Cellular Context. ChemMedChem 2014; 9: 560-5.
[27] Patil SP, Pacitti MF, Gilroy KS, *et al*. Identification of antipsychotic drug fluspirilene as a potential p53-MDM2 inhibitor: a combined computational and experimental study. J Comput Aided Mol Des 2015; 29: 155-63.
[28] Chen SJ. A Potential Target of Tanshinone IIA for Acute Promyelocytic Leukemia Revealed by Inverse Docking and Drug Repurposing. Asian Pac J Cancer Prev 2013; 15: 4301-5.

[29] Bansode SB, Jana AK, Batkulwar KB, *et al*. Molecular Investigations of Protriptyline as a Multi-Target Directed Ligand in Alzheimer's Disease. PLoS One. 2014 Aug 20; [cited 2015 June 16]; 9: e105196. Available from: (http: //journals.plos.org/plosone/article?id=10.1371/journal.pone.0105196)

[30] Banerjee DR, Biswas R, Das AK, Basak A. Design, synthesis and characterization of dual inhibitors against new targets FabG4 and HtdX of Mycobacterium tuberculosis. Eur J Med Chem 2015; 100: 223-34.

[31] Bohari MH, Sastry GN. FDA approved drugs complexed to their targets: evaluating pose prediction accuracy of docking protocols. J Mol Model 2012; 18: 4263-74.

[32] Warren GL, Andrews CW, Capelli AM, *et al*. A critical assessment of docking programs and scoring functions. J Med Chem 2006; 49: 5912-31.

[33] Chen YC. Beware of docking!. Trends Pharmacol Sci 2015; 36: 78-95.

[34] Amaro RE, Li WW. Emerging methods for ensemble-based virtual screening. Curr Top Med Chem 2010; 10: 3-13.

[35] Oldfield E. Targeting Isoprenoid Biosynthesis for Drug Discovery: Bench to Bedside. Acc Chem Res 2010; 43: 1216–26.

[36] Zhu W, Zhang YH, Sinko W, *et al*. Antibacterial drug leads targeting isoprenoid biosynthesis. Proc Natl Acad Sci USA 2013; 110: 123-8.

[37] Durrant J, McCammon JA. Molecular dynamics simulations and drug discovery. BMC Biol 2011; 9: 71.

[38] Sinko W, de Oliveira CAF, Williams S, *et al*. Applying molecular dynamics simulations to identify rarely sampled ligand-bound conformational states of undecaprenyl pyrophosphate synthase, an antibacterial target. Chem Biol Drug Des 2011; 77: 412-20.

[39] Whalen EJ, Rajagopal S, Lefkowitz RJ. Therapeutic potential of beta-arrestin- and G protein-biased agonists. Trends Mol Med 2011; 17: 126-39.

[40] Violin JD, DeWire SM, Yamashita D, *et al*. Selectively engaging beta-arrestins at the angiotensin II type 1 receptor reduces blood pressure and increases cardiac performance. J Pharmacol Exp Ther 2010; 335: 572-9.

[41] Dror RO, Green HF, Valant C, *et al*. Structural basis for modulation of a G-protein-coupled receptor by allosteric drugs. Nature 2013; 503: 295-9.

[42] Salentin S, Haupt VJ, Daminelli S, Schroeder M. Polypharmacology rescored: protein-ligand interaction profiles for remote binding site similarity assessment. Prog Biophys Mol Biol 2014; 116: 174-86.

[43] Möller-Acuña P, Contreras-Riquelme JS, Rojas-Fuentes C, *et al*. Similarities between the Binding Sites of SB-206553 at Serotonin Type 2 and Alpha7 Acetylcholine Nicotinic Receptors: Rationale for Its Polypharmacological Profile. PLoS One. 2015 August 5; [cited 2015 October 28]; 10: e0134444. Available from: (http: //dx.plos.org/10.1371/journal.pone.0134444)

[44] Tetko IV, Gasteiger J, Todeschini R, *et al*. Virtual computational chemistry laboratory - design and description. J Comput Aid Mol Des 2005; 19: 453-63.

[45] Small-Molecule Drug Discovery Suite 2015-2: QikProp, version 4.4, Schrödinger, LLC, New York, NY, 2015.

[46] Haupt VJ, Daminelli S, Schroeder M. Drug promiscuity in PDB: protein binding site similarity is key. PLoS One 2013; 8: e65894.

[47] Li Y, Zhang X, Cao D. The Role of Shape Complementarity in the Protein-Protein Interactions. Sci Rep 2013; 3: 1-7.

[48] Warner WA, Sanchez R, Dawoodian A, Li E, Momand J. Identification of FDA-approved Drugs that Computationally Bind to MDM2. Chem Biol Drug Des 2012; 80: 631-7.

[49] Vasudevan SR, Moore JB, Schymura Y, Churchill GC. Shape-Based Reprofiling of FDA-Approved Drugs for the H1 Histamine Receptor. J Med Chem 2012; 55: 7054-60.

[50] Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. Nat Biotechnol 2007; 25: 197-206.

[51] Lounkine E, Keiser MJ, Whitebread S, *et al*. Large-scale prediction and testing of drug activity on side-effect targets. Nature 2012; 486: 361-7.

[52] Cameron RT, Coleman RG, Day JP, *et al*. Chemical informatics uncovers a new role for moexipril as a novel inhibitor of cAMP phosphodiesterase-4 (PDE4).Biochem Pharmacol 2013; 85: 1297-305.

[53] Cheng F, Li W, Wu Z, *et al*. Prediction of polypharmacological profiles of drugs by the integration of chemical, side effect, and therapeutic space. J Chem Inf Model 2013; 53: 753-62.

[54] Chiang AP, Butte AJ. Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. Clin Pharmacol Ther 2009; 86: 507-10.

[55] Wang K, Sun J, Zhou S, *et al*. Prediction of drug-target interactions for drug repositioning only based on genomic expression similarity. PLoS Comput Biol 2013; 9: e1003315.

[56] Lamb J, Crawford ED, Peck D, *et al*. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 2006; 313: 1929-35.

[57] Li B, Xiong M, Zhang HY. Elucidating polypharmacological mechanisms of polyphenols by gene module profile analysis. Int J Mol Sci 2014; 15: 11245-54.

[58] Babcock JJ, Du F, Xu K, Wheelan SJ, Li M. Integrated analysis of drug-induced gene expression profiles predicts novel hERG inhibitors. PloS one 2013; 8: e69513.

[59] Eriksson A, Österroos A, Hassan S, *et al*. Drug screen in patient cells suggests quinacrine to be repositioned for treatment of acute myeloid leukemia. Blood Cancer J 2015; 5: e307.

[60] Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science 2008; 321: 263-6.

[61] Alaimo S, Pulvirenti A, Giugno R, Ferro A. Drug-target interaction prediction through domain-tuned network-based inference. Bioinformatics 2013; 29: 2004-8.

[62] Cheng F, Liu C, Jiang J, *et al*. Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput Biol 2012; 8: e1002503.

[63] Yadav B, Gopalacharyulu P, Pemovska T, *et al*. From drug response profiling to target addiction scoring in cancer cell models. Dis Model Mech 2015; 8: 1255-64.

[64] Dixon SL, Smondyrev AM, Knoll EH, *et al*. PHASE: A New Engine for Pharmacophore Perception, 3D QSAR Model Development, and 3D Database Screening. 1. Methodology and Preliminary Results. J Comput Aided Mol Des 2006; 20: 647-71.

[65] Richmond N, Abrams C, Wolohan P, Abrahamian E, Willett P, Clark R. GALAHAD: 1. Pharmacophore identication by hypermolecular alignment of ligands in 3D. J Comput Aided Mol Des 2006; 20: 567-87.

[66] Jones G, Willett P, Glen RC. GASP: genetic algorithm superposition program. In: Güner OF Ed. Pharmacophore Perception, Development, and Use in Drug Design. La Jolla, CA: International University Line 2000; pp. 85-106.

[67] Baroni M, Cruciani G, Sciabola S, Perrucio F, Mason JS. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application. J Chem Inf Model 2007; 47: 279-94.

[68] Sanders MP, Verhoeven S, de Graaf C, *et al*. Snooker: a structure-based pharmacophore generation tool applied to class A GPCRs. J Chem Inf Model 2011; 51: 2277-92.

[69] Ai N, Wood RD, Welsh WJ. Identification of Nitazoxanide as a Group I Metabotropic Glutamate Receptor Negative Modulator for the Treatment of Neuropathic Pain: An In Silico Drug Repositioning Study. Pharm Res 2015; 32: 2798-807.

[70] Levit A, Nowak S, Peters M, *et al*. The bitter pill: clinical drugs that activate the human bitter taste receptor TAS2R14. FASEB J 2014; 28: 1181-97.

[71] Krautscheid Y, Senning CJÅ, Sartori SB, Singewald N, Schuster D, Stuppner H. Pharmacophore modeling, virtual screening, and *in vitro* testing reveal haloperidol, eprazinone, and fenbutrazate as neurokinin receptors ligands. J Chem Inf Mod 2014; 54: 1747-57.

[72] Meslamani J, Bhajun R, Martz F, Rognan D. Computational profiling of bioactive compounds using a target-dependent composite workflow. J Chem Inf Mod 2013; 53: 2322-33.

[73] Dakshanamurthy D, Basu G, Go N. The role of electrostatics in discrimination of Adenine and Guanine by Proteins. Genome Inf 2002; 13: 316-7.

[74] Basu G, Dakshanamurthy S, Kawabata T, Go N. Electrostatic potential of nucleotide-free protein is sufficient for discrimination between adenine and guanine-specific binding sites. J Mol Biol 2004; 342: 1053-66.

[75] Jennings A, Tennant M. Selection of molecules based on shape and electrostatic similarity: proof of concept of "electroforms". J Chem Inf Mod 2007; 47: 1829-38.

[76] Armstrong MS, Morris GM, Finn PW, *et al*. ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. J Comput Aided Mol Des 2010; 24: 789-801.

[77] Muchmore SW, Souers AJ, Akritopoulou-Zanze I. The Use of Three-Dimensional Shape and Electrostatic Similarity Searching in the Identification of a Melanin-Concentrating Hormone Receptor 1 Antagonist. Chem Biol Drug Des 2006; 67: 174-6.

[78] Boström J, Grant JA, Fjellström O, Thelin A, Gustafsson D. Potent fibrinolysis inhibitor discovered by shape and electrostatic complementarity to the drug tranexamic acid. J Med Chem 2013; 56: 3273-80.

[79] Haupt VJ, Schroeder M. Old friends in new guise: repositioning of known drugs with structural bioinformatics. Brief Bioinform 2011; 12: 312-26.

[80] Schalon C, Surgand JS, Kellenberger E, Rognan D. A simple and fuzzy method to align and compare druggable ligand-binding sites. Proteins 2008; 71: 1755-78.

[81] Defranchi E, Schalon C, Messa M, Onofri F, Benfenati F, Rognan D. Binding of protein kinase inhibitors to synapsin I inferred from pair-wise binding site similarity measurements. PLoS One 2010; 5: e12214.

[82] Xie L, Bourne PE. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile–profile alignments. Proc Natl Acad Sci USA 2008; 105: 5441-6.

[83] Kinnings SL, Liu N, Buchmeier N, Tonge PJ, Xie L, Bourne PE. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. PLoS Comput Biol 2009; 5: e1000423.

[84] Durrant JD, Amaro RE, Xie L, *et al*. A multidimensional strategy to detect polypharmacological targets in the absence of structural and sequence homology. PLoS Comput Biol 2010; 6: e1000648.

[85] Nray-Szab G. Analysis of molecular recognition: steric electrostatic and hydrophobic complementarity. J Mol Recognit 1996; 6: 205-10.

[86] Chau PL, Dean PM. Electrostatic complementarity between proteins and ligands. 1. Charge disposition, dielectric and interface effects. J Comput-Aided Mol Des 1994; 8: 51325.

[87] Voet A, Berenger F, Zhang KY. Electrostatic similarities between protein and small molecule ligands facilitate the design of protein-protein interaction inhibitors. PloS one 2013; 8: 75762.

[88] Drwal MN, Marinello J, Manzo SG, Wakelin LP, Capranico G, Griffith R. Novel DNA Topoisomerase IIα Inhibitors from Combined Ligand-and Structure-Based Virtual Screening. PloS one 2014; 9: e114904.

[89] Dobi K, Flachner B, Pukáncsik M, *et al*. Combination of Pharmacophore Matching, 2D Similarity Search, and *In Vitro* Biological Assays in the Selection of Potential 5-HT6 Antagonists from Large Commercial Repositories. Chem Biol Drug Des. 2015; 86: 864-80.

[90] Markt P, Petersen RK, Flindt EN, *et al*. Discovery of novel PPAR ligands by a virtual screening approach based on pharmacophore modeling, 3D shape, and electrostatic similarity screening. J Med Chem 2008; 51: 6303-17.

[91] Winkler DA. The role of quantitative structure-activity relationships (QSAR) in biomolecular discovery. Brief Bioinform 2002; 3: 73-86.

[92] Singh A, Singh R. QSAR and its Role in Target-Ligand Interaction. The Open Bioinformatics Journal 2013; 7: 63-7.

[93] Ragno R, Artico M, De Martino G, *et al*. Docking and 3-D QSAR studies on indolyl aryl sulfones. Binding mode exploration at the HIV-1 reverse transcriptase non-nucleoside binding site and design of highly active N-(2-hydroxyethyl) carboxamide and N-(2-hydroxyethyl) carbohydrazide derivatives. J Med Chem 2005; 48: 213-23.

[94] Shukla A, Sharma P, Prakash O, *et al*. QSAR and Docking Studies on Capsazepine Derivatives for Immunomodulatory and Anti-Inflammatory Activity. PLoS One. 2014 Jul 8; [cited 2015 Jul 14]; 9: e100797. Available from: (http: //dx.plos.org/10.1371/journal.pone.0100797)

[95] Roy K, Kar S. Importance of Applicability Domain of QSAR Models. In: Roy K, Kar S, Eds. Quantitative Structure–Activity Relationships in Drug Design, Predictive Toxicology, and Risk Assessment. Hershey, PA: IGI Global 2015; pp. 180-211.

[96] Lapinsh M, Prusis P, Mutule I, Mutulis F, Wikberg JE. QSAR and proteo- chemometric analysis of the interaction of a series of organic compounds with melanocortin receptor subtypes. J Med Chem 2003; 46: 2572-9.

[97] Guha R, Van Drie JH. Structure-activity landscape index: identifying and quantifying activity cliffs. J Chem Inf Model. 2008; 48: 646-58.

[98] Wawer M, Peltason L, Bajorath J. Elucidation of structure-activity relationship pathways in biological screening data. J Med Chem 2009; 52: 1075-80.

[99] Gedeck P, Rohde B, Bartels C. QSAR-how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. J Chem Inf Model 2006; 46: 1924-36.

[100] Van Westen GJ, Wegner JK, Geluykens P, *et al*. Which compound to select in lead optimization? Prospectively validated proteochemometric models guide preclinical development. PloS One 2011; 6: e27518.

[101] Gao J, Huang Q, Wu D, *et al*. Study on human GPCR–inhibitor interactions by proteochemometric modeling. Gene 2013; 518: 124-31.

[102] Lapins M, Eklund M, Spjuth O, Prusis P, Wikberg JE. Proteochemometric modeling of HIV protease susceptibility. BMC Bioinformatics 2008; 9: 181.

[103] Lapins M, Wikberg JE. Kinome-wide interaction modelling using alignment- based and alignment-independent approaches for kinase description and linear and non-linear data analysis techniques. BMC Bioinformatics 2010; 11: 339.

[104] Nabu S, Nantasenamat C, Owasirikul W, *et al*. Proteochemometric model for predicting the inhibition of penicillin-binding proteins. J Comput Aided Mol Des 2015; 29: 127-41.

[105] (a) Dakshanamurthy S, Issa NT, Assefnia S, *et al*. Predicting new indications for approved drugs using a proteochemometric method. J Med Chem 2012; 55: 6832-48. (**b**) Issa NT, Peters OJ, Byers SW, Dakshanamurthy S. RepurposeVS: A Drug Repurposing-Focused Computational Method for Accurate Drug-Target Signature Predictions. Comb Chem High Throughput Screen 2015; 18: 784-94.

[106] Munteanu CR, Magalhaes AL, Uriarte E, Gonzalez-Diaz H. Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices. J Theor Biol 2009; 257: 303-11.

[107] Rolland C, Gozalbes R, Nicolaï E, *et al*. G-protein-coupled receptor affinity prediction based on the use of a profiling dataset: QSAR design, synthesis, and experimental validation. J Med Chem 2005; 48: 6563-74.

[108] Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. Rational drug design for anti-cancer chemotherapy: Multi-target QSAR models for the in silico discovery of anti-colorectal cancer agents. Bioorg Med Chem 2012; 20: 4848-55.

[109] Cheng F, Zhou Y, Li J, Li W, Liu G, Tang Y. Prediction of chemical–protein interactions: multitarget-QSAR versus computational chemogenomic methods. Mol BioSyst 2012; 8: 2373-84.

[110] Liu Q, Zhou H, Liu L, Chen X, Zhu R, Cao Z. Multi-target QSAR modelling in the analysis and design of HIV-HCV co-inhibitors: an in-silico study. BMC Bioinformatics 2011; 12: 294-314.

[111] Garcia I, Fall Y, Gomez G, Gonzalez-Diaz H. First computational chemistry multi-target model for anti-Alzheimer, anti-parasitic, anti-fungi, and anti-bacterial activity of GSK-3 inhibitors *in vitro*, *in vivo*, and in different cellular lines. Mol Divers 2011; 15: 561-7.

[112] Speck-Planche A, Kleandrova VV, Cordeiro MN. New insights toward the discovery of antibacterial agents: Multi-tasking QSBER model for the simultaneous prediction of anti-tuberculosis activity and toxicological profiles of drugs. Eur J Pharm Sci 2013; 48: 812-8.

[113] Prado-Prado FJ, Garcia-Mera X, Gonzalez-Diaz H. Multi-target spectral moment QSAR versus ANN for antiparasitic drugs against different parasite species. Bioorg Med Chem 2010; 18: 2225-31.

[114] Baumgartner WA Jr, Cohen KB, Fox LM, Acquaah-Mensah G, Hunter L. Manual curation is not sufficient for annotation of genomic databases. Bioinformatics 2007; 23: i41-8.

[115] Andronis C, Sharma A, Virvilis V, Deftereos S, Persidis A. Literature mining, ontologies and information visualization for drug repurposing. Brief Bioinform 2011; 12: 357-68.

[116] Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc 2011; 18: 544-51.

[117] Chen L, Liu H, Friedman C. Gene name ambiguity of eukaryotic nomenclatures. Bioinformatics 2005; 21: 248-56.

[118] Smith L, Tanabe LK, Johnson nee Ando R, *et al*. Overview of BioCreative II gene mention recognition. Genome Biol 2009; 9: S2.

[119] Krallinger M, Rabal O, Leitner F, *et al*. The CHEMDNER corpus of chemicals and drugs and its annotation principles. J Cheminform 2015; 7: S2.

[120] Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pac Symp Biocomput 2000; 517-28.

[121] Kuhn M, Szklarczyk D, Franceschini A, von Mering C, Jensen LJ, Bork P. STITCH 3: zooming in on protein-chemical interactions. Nucleic Acids Res 2012; 40: D876-80.

[122] Franceschini A, Szklarczyk D, Frankild S, *et al*. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 2013; 41: D808-15.

[123] Saric J, Jensen LJ, Ouzounova R, Rojas I, Bork P. Extraction of regulatory gene/protein networks from Medline. Bioinformatics 2006; 22: 645-50.

[124] Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2015 Oct 19; [cited 2015 November 2]. (http: //nar.oxfordjournals.org/content/early/2015/10/19/nar.gkv1075.full)

[125] de Marneffe M-C, Manning CD. The Stanford typed dependencies representation. CrossParser '08 2008. doi: 10.3115/1608858.1608859.

[126] Yu S, Yang X, Zhu Y, *et al*. Systems pharmacology of mifepristone (RU486) reveals its 47 hub targets and network: comprehensive analysis and pharmacological focus on FAK-Src-Paxillin complex. Sci Rep 2015; 5: 7830.

[127] Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics 2005; 21: 3191-2.

[128] Cutler DM. The demise of the blockbuster? N Engl J Med 2007; 356: 1292-3.

[129] Xie L, Li J, Xie L, Bourne PE. Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. PLoS Comput Biol 2009; 5: e1000387.

[130] Chang RL, Xie L, Xie L, Bourne PE, Palsson BØ. Drug off-target effects predicted using structural analysis in the context of a metabolic network model. PLoS Comput Biol 2010; 6: e1000938.

[131] Lounkine E, Keiser MJ, Whitebread S, *et al*. Large-scale prediction and testing of drug activity on side-effect targets. Nature 2012; 486: 361-7.

[132] Mizutani S, Pauwels E, Stoven V, Goto S, Yamanishi Y. Relating drug-protein interaction network with drug side effects. Bioinformatics 2012; 28: i522-8.

[133] Kuhn M, Al Banchaabouchi M, Campillos M, *et al*. Systematic identification of proteins that elicit drug side effects. Mol Syst Biol 2013; 9: 663.

[134] Simon Z, Peragovics A, Vigh-Smeller M, *et al*. Drug effect prediction by polypharmacology-based interaction profiling. J Chem Inf Model 2012; 52: 134-45.

[135] Duran-Frigola M, Aloy P. Analysis of chemical and biological features yields mechanistic insights into drug side effects. Chem Biol 2013; 20: 594-603.

[136] Pauwels E, Stoven V, Yamanishi Y. Predicting drug side-effect profiles: a chemical fragment-based approach. BMC Bioinformatics 2011; 12: 169.

[137] Pérez-Nueno VI, Venkatraman V, Mavridis L, Ritchie DW. Detecting drug promiscuity using Gaussian ensemble screening. J Chem Inf Model 2012; 52(8): 1948-61.

[138] Pérez-Nueno VI, Souchet M, Karaboga AS, Ritchie DW. GESSE: Predicting Drug Side Effects from Drug–Target Relationships. J Chem Inf Model 2015; 55(9): 1804-23.

[139] Xie L, Evangelidis T, Xie L, Bourne PE. Drug discovery using chemical systems biology: weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir. PLoS Comput Biol 2011; 7: e1002037.

[140] Knight ZA, Lin H, Shokat KM. Targeting the cancer kinome through polypharmacology. Nat Rev Cancer 2010; 10: 130-7.

[141] Zhao J, Yang P, Li F, *et al*. Therapeutic effects of astragaloside IV on myocardial injuries: multi-target identification and network analysis. PLoS One 2012; 7: e44938.

[142] Arooj M, Sakkiah S, Cao GP, Kim S, Arulalapperumal V, Lee KW. Finding off-targets, biological pathways, and target diseases for chymase inhibitors via structure-based systems biology approach. Proteins 2015; 83: 1209-24.

[143] Xie XQ, Wang L, Liu H, Ouyang Q, Fang C, Su W. Chemogenomics knowledgebased polypharmacology analyses of drug abuse related G-protein coupled receptors and their ligands. Front Pharmacol 2014; 5: 3.

[144] Liu H, Wang L, Lv M, *et al*. AlzPlatform: An Alzheimer's Disease Domain-Specific Chemogenomics Knowledgebase for Polypharmacology and Target Identification Research. J Chem Inf Model 2014; 54: 1050–60.

[145] Jansson PJ, Kalinowski DS, Lane DJ, *et al*. The renaissance of polypharmacology in the development of anti-cancer therapeutics: Inhibition of the "Triad of Death" in cancer by Di-2-pyridylketone thiosemicarbazones. Pharmacol Res 2015; 100: 255-60.

[146] Antonov AV. BioProfiling.de: analytical web portal for high-throughput cell biology. Nucleic Acids Res 2011; 39: W323-7.

[147] Amelio I, Gostev M, Knight RA, Willis AE, Melino G, Antonov AV. DRUGSURV: a resource for repositioning of approved and experimental drugs in oncology based on patient survival information. Cell Death Dis 2014; 5: e1051.

[148] Antonov AV, Krestyaninova M, Knight RA, Rodchenkov I, Melino G, Barlev NA. PPISURV: a novel bioinformatics tool for uncovering the hidden role of specific genes in cancer survival outcome. Oncogene 2014; 33: 1621-8.

[149] Lehár J, Krueger AS, Avery W, *et al*. Synergistic drug combinations tend to improve therapeutically relevant selectivity. Nat Biotechnol 2009; 27: 659-66.

[150] Bansal M, Yang J, Karan C, *et al*. A community computational challenge to predict the activity of pairs of compounds. Nat Biotechnol 2014; 32: 1213-22.

[151] Vitali F, Mulas F, Marini P, Bellazzi R. Network-based target ranking for polypharmacological therapies. J Biomed Inform 2013; 46: 876-81.

[152] Tang J, Karhinen L, Xu T. Target inhibition networks: predicting selective combinations of druggable targets to block cancer survival pathways. PLoS Comput Biol 2013; 9: e1003226.

[153] Sun Y, Sheng Z, Ma C, *et al*. Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer. Nat Commun 2015; 6: 8481.

[154] Byers SW, Dakshanamurthy S. Diviner Intervention - Computational drug discovery and drug repurposing technology. International Innovation 2015; 174: 10-2.

[155] Roin B. The Case for Tailoring Patent Awards Based on Time-to-Market. UCLA L Rev 2014; 61: 672.

**Received: January 6, 2016** **Accepted: February 23, 2016**

Issa et al. BMC Bioinformatics (2016) 17:202
DOI 10.1186/s12859-016-1065-y

BMC Bioinformatics

RESEARCH ARTICLE Open Access

# DrugGenEx-Net: a novel computational platform for systems pharmacology and gene expression-based drug repurposing

Naiem T. Issa[1], Jordan Kruger[2†], Henri Wathieu[3†], Rajarajan Raja[4], Stephen W. Byers[1,2] and Sivanesan Dakshanamurthy[1,2*]

## Abstract

**Background:** The targeting of disease-related proteins is important for drug discovery, and yet target-based discovery has not been fruitful. Contextualizing overall biological processes is critical to formulating successful drug-disease hypotheses. Network pharmacology helps to overcome target-based bottlenecks through systems biology analytics, such as protein-protein interaction (PPI) networks and pathway regulation.

**Results:** We present a systems polypharmacology platform entitled DrugGenEx-Net (DGE-NET). DGE-NET predicts empirical drug-target (DT) interactions, integrates interaction pairs into a multi-tiered network analysis, and ultimately predicts disease-specific drug polypharmacology through systems-based gene expression analysis. Incorporation of established biological network annotations for protein target-disease, −signaling pathway, −molecular function, and protein-protein interactions enhances predicted DT effects on disease pathophysiology. Over 50 drug-disease and 100 drug-pathway predictions are validated. For example, the predicted systems pharmacology of the cholesterol-lowering agent ezetimibe corroborates its potential carcinogenicity.
When disease-specific gene expression analysis is integrated, DGE-NET prioritizes known therapeutics/experimental drugs as well as their contra-indications. Proof-of-concept is established for immune-related rheumatoid arthritis and inflammatory bowel disease, as well as neuro-degenerative Alzheimer's and Parkinson's diseases.

**Conclusions:** DGE-NET is a novel computational method that predicting drug therapeutic and counter-therapeutic indications by uniquely integrating systems pharmacology with gene expression analysis. DGE-NET correctly predicts various drug-disease indications by linking the biological activity of drugs and diseases at multiple tiers of biological action, and is therefore a useful approach to identifying drug candidates for re-purposing.

**Keywords:** DrugGenEx-NET, TMFS, Polypharmacology, Gene expression analysis, Rheumatoid arthritis, Inflammatory bowel disease, Parkinson's disease, Alzheimer's disease

## Background

Modern drug discovery endeavors are only rarely translated into acceptable clinical success rates [1]. Pre-clinical drug discovery initiatives have been gene-centric with a focus on finding drugs for targets of interest with high binding affinity and selectivity [2]. It is increasingly accepted, however, that disease states exhibit biological complexity, and that the gene-centric view neglects physiologic context by isolating the target in an artificial environment [3]. Furthermore, drugs arising from de novo design are likely to have many unknown targets given the limited scope of biochemical assays, thus leading to both clinical toxicity and unanticipated novel disease indications [4]. Systems pharmacology, the integration of systems biology with network pharmacology, is a mechanism-centric solution that considers the global physiological environment of disease states and allows for the discovery of drugs or combinations of

* Correspondence: sd233@georgetown.edu
†Equal contributors
[1]Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington DC 20057, USA
[2]Department of Biochemistry & Molecular Biology, Georgetown University, Washington DC 20057, USA
Full list of author information is available at the end of the article

drugs that may simultaneously target multiple nodes of the disease-associated network [5]. Initiatives utilizing network analysis have led to successful drug discovery efforts [6–11].

As most FDA-approved drugs are considered safe and simultaneously exhibit multi-target effects, drug repurposing is an optimal strategy for harnessing the strength of polypharmacology [12]. Current methods do not utilize high-throughput approaches to empirically determine drug-target associations and subsequently contextualize them using systems biology. Here, we have created a novel computational systems pharmacology platform, entitled DGE-NET, that: (1) accurately predicts drug-protein target interactions, (2) assesses drug effects through systems analysis of cumulative predicted targets for each drug, and (3) formulates drug-disease associations through gene expression analysis and polypharmacology.

DGE-NET was first applied to a set of 3,671 FDA approved and experimental drugs across 2,335 human protein target crystal structures for potential drug repurposing. Drugs were then associated with biological effects, which include molecular functions, signaling pathways, protein-protein interactions (PPIs) and diseases, through association with their predicted targets. Drug-biological effect predictions were validated at multiple tiers using findings in the literature and experimentally determined associations from annotated databases. Over 50 drug-disease and 100 drug-pathway associations were validated. DGE-NET also provided further evidence for unexpected toxicities, such as the potential carcinogenic properties of the cholesterol absorption blocker ezetimibe. Drug-target and drug-biological effect signatures were also statistically associated with clinical disease-relevant protein targets, PPIs, pathways, and functions obtained from differential gene expression analysis. DGE-NET incorporated a novel drug prioritization scheme that ranks drugs matched to a disease based on its polypharmacology at each tier of biological action.

For proof-of-concept, DGE-NET was applied to human-derived gene expression datasets obtained for rheumatoid arthritis (RA), inflammatory bowel disease (IBD), Alzheimer's disease (AD), and Parkinson's disease (PD). DGE-NET was validated by prioritizing approved drugs and biologics as well as those currently being examined repurposing, and also revealed drugs contraindicated in those conditions, such as tetracyclines in IBD. DGE-NET is first computational platform we know of that predicts novel protein binding signatures of FDA-approved drugs and subsequently matches drug action at multiple levels of biological activity to gene expression-based characterization of disease perturbation. It stands as an effort to address the pressing need for models that account for the complexity of multi-tiered interactions for better simulations of disease states and predictive therapeutics. In summary, DGE-NET is a novel computational method for gene expression- and systems polypharmacology-driven drug repurposing.

## Methods

### Collection of FDA-approved drugs, experimental molecules, and protein target curation

Spatial Data Files (SDF) of drugs and experimental molecules containing spatial atom connectivity information were obtained from DrugBank [13], the NCGC Pharmaceutical Collection [14], FDA (www.FDA.gov), and BindingDB [15]. Energy-minimized 3D structures were prepared using Schrodinger's LigPrep [16] algorithm at pH 7.0. Human protein crystal structures were obtained from RCSB (www.rcsb.org). Only X-ray structures with <2.5 angstrom resolution and a reference co-crystallized ligand were chosen. Protein structures were further processed to remove non-biologically relevant chains (i.e. those that do interact with the ligand), metal ions, and all heteroatoms (i.e. non-cofactors, solvent molecules). Structures were then prepared using ProteinPrep in Schrodinger to relax the structures and optimize hydrogen bonds at pH 7.0. After processing, the dataset included 3,671 drugs and 2,335 protein target crystal structures.

### Predicting Drug-Target (DT) signatures

DGE-NET utilizes a modified version of our "Train, Match, Fit and Streamline" (TMFS) method [17] for generating reliable binding signature predictions. Briefly, TMFS is a proteochemometric method that predicts the binding potential of a protein-ligand complex by integrating docking, three-dimensional shape, and ligand physicochemical descriptors (Fig. 1). GLIDE [18] was used to dock molecules into protein pockets identified by the reference ligand, and QikProp [19] was used to generate the following ligand-specific physicochemical descriptors: (1) solvent-accessible surface area, (2) volume, (3) dipole, (4) # H-bond acceptors, (5) # H-bond donors, (6) globularity, (7) ionization potential, and (8) electron affinity. Strike [20] was used to generate Tanimoto similarity coefficients to quantify the similarity of ligand physicochemical descriptors to that of the bioactive reference molecules found in the protein complex crystal structures. Ligand and pocket 3D shapes were quantified using a spherical harmonics expansion approach [21] and ligand-reference molecule/ligand-protein pocket shape similarities were quantified using a Euclidean distance metric. After docking scores, shape similarity, Euclidian distance scores, and ligand-based descriptor similarity scores were derived by the tools described above, a common scheme was used to normalize these scores, wherein each is transformed into a 0–1 range, 1 being the most favorable score present. These metrics were combined into a comprehensive Z-score

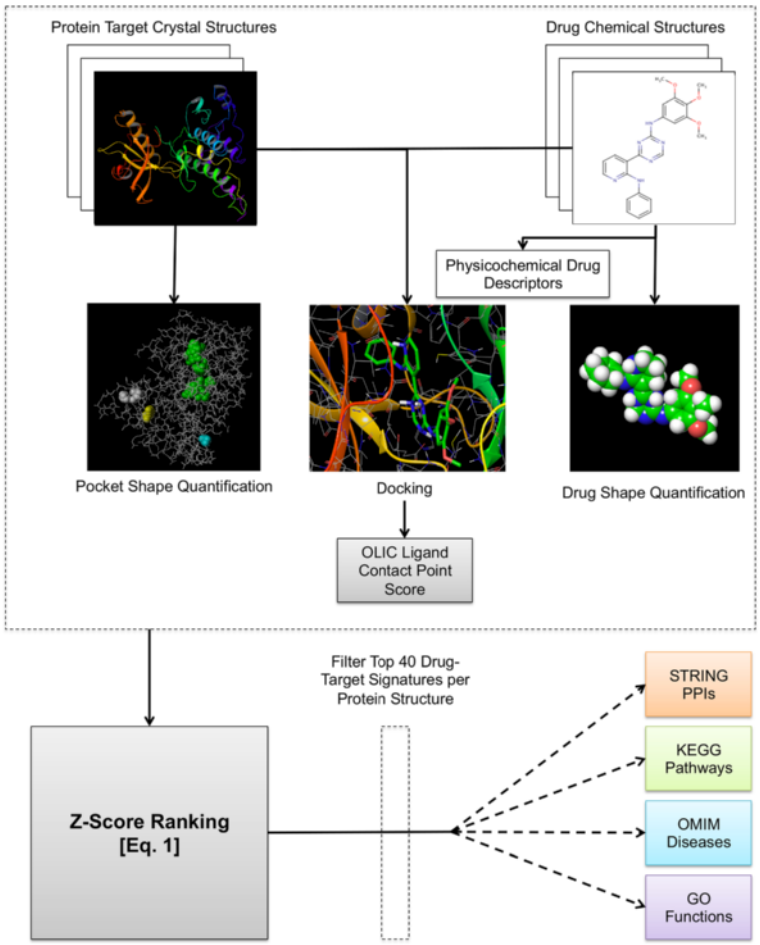



**Fig. 1** Workflow for predicting drug-target signatures and relating network pharmacology

that was used to rank ligands such that the top-ranking molecules are considered most likely to bind. The Z-score for a unique ligand ($l$) –protein ($p$) co-crystallized with reference ligand $r$ is as follows:

$$Z(l,r,p) = w_k Y(l,p) + \sum_{m=1}^{M}\left[w_m f_m(l,p) + w'_m f'_m(l,r)\right] + \sum_{n=1}^{N}[X_n(l,r) + CS(OLIC)] \quad (1)$$

$Y$ is the normalized docking score with weight $w_{k=4}$. The first summation term is the normalized shape similarity score for ligand-to-protein pocket $f_m(l,p)$ and ligand-to-reference $f'_m(l,r)$ with weights $w_{m=1}$ and $w'_{m=2}$, and the second summation term corresponds to the sum of the Tanimoto similarity coefficients between the ligand and reference for physicochemical descriptors. Aforementioned weights for docking, protein shape similarity, and ligand shape similarity, respectively, were found to maximize the accuracy of TMFS in predicting top protein targets from publically available experimental data. Lastly, $CS(OLIC)$ is a correction term based on the similarity of contact points created between the ligand and reference to the protein target. It was assumed that drugs have similar experimental activity if their interaction involves similar binding site residues and interaction patterns to that of the reference. The top 40-scoring drugs were considered as "hits" for a given protein target for subsequent network analysis. The top 40 drugs were chosen as they represent the top 1 % of all the drugs in our dataset, a fraction that is typically employed in virtual screening protocols [17].

### Relating drug-target predictions to diseases, pathways, functions, and protein-protein interactions

Predicted drug-target associations were associated with diseases, signaling pathways and molecular functions for network analysis (Fig. 1). Protein targets were cross-referenced using the unique PDB entry with UniProt

[22]. Because many crystal structures may correspond to the same protein, collapsing them using UniProt reduces the total number of protein target nodes. A list of genes associated with the protein were obtained from each UniProt entry and mapped to Online Mendelian Inheritance in Man (OMIM) Morbidity Map [23] gene-disease associations, a procedure modeled after Yildirim et al. [24]. Drugs are connected to a disease via mapping of their target genes to their associated disease. Thus, a drug is connected to a disease if its predicted targets have disease genes associated with the disorder. In the DT-disease network, all disorders associated with a predicted protein target will be associated with the drug.

Disease-associated targets were also annotated with KEGG pathway [25, 26] and Gene Ontology (GO) molecular function [27, 28] information using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) Functional Annotation Tool (FAT) [29, 30]. FAT was also used to annotate pathways and functions for a given drug via its predicted direct and indirect targets through protein-protein interactions using FDR <0.25. Protein-protein interactions (PPIs) were extracted from the ExPASy STRING database [31] using a confidence score cutoff of 0.95. Any PPI pairs where one of the partners did not exist in our protein target dataset were excluded. A gene list comprised of a drug's predict direct targets as well as those targets' interacting partners was subjected to DAVID annotation. For example, if Drug A was predicted to interact with Target A and Target B, and Target A also interacted with Protein C while Target B interacted with Protein D and Protein E, then the gene list for Drug A would consist of the following: Target A, Target B, Protein C, Protein D, and Protein E.

### Annotating disease and pathway categories

The disease categories from Medical Subject Headings (MeSH) were used for annotation of disease names corresponding to OMIM disorder entries. Approximately 93 % of the diseases were mapped to a disease category. The Comparative Toxicogenomics Database (CTD) [32] was used to map 75 % of the diseases; the remaining diseases were manually curated, with 71 % of these providing a partial or close match. Diseases that mapped to multiple disease categories were manually evaluated to determine a primary disease category. This was done by determining what the primary clinically treated category is for a disease. For example, the disease systemic lupus erythematosus is primarily an autoimmune disorder but can be considered as "skin and connective tissue" if the disease process involves the facial malar rash. Diseases in which a primary category could not be determined were categorized as multiple. Pathways were manually organized into categories based on metabolic/cellular processes and diseases as annotated by KEGG.

### Incorporation of disease gene expression data with systems pharmacology

A schematic of DGE-NET is illustrated in Fig. 2. Differential gene expression analysis on Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) microarray data was performed for RA (GSE55235 and GSE55457), IBD (GSE52746 and GSE11223), AD (GSE29378), and PD (GSE7621). Differentially expressed genes between normal and diseased patient biopsies with adjusted $P$ values < 0.05 (using GEO2R [33]) were obtained. GEO2R is a R-based publicly accessible web tool for analyzing GEO-deposited gene expression data (http://www.ncbi.nlm.nih.gov/geo/

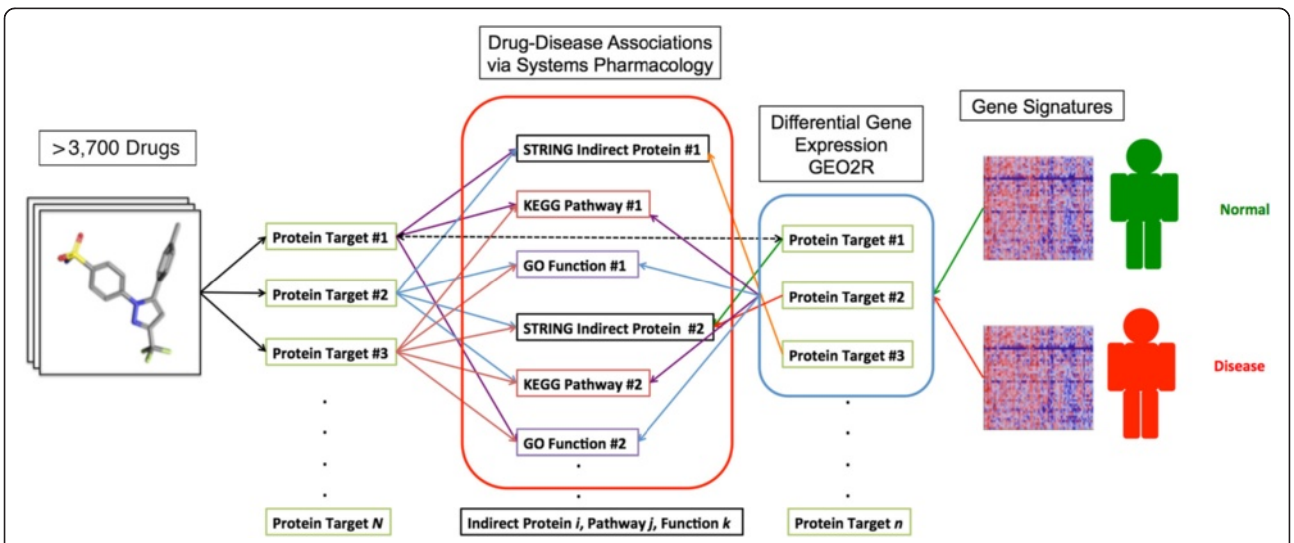



**Fig. 2** Schematic of DGE-NET used to associate drugs with diseases. Differential gene expression analysis of diseased versus non-diseased states is used to establish a disease-related gene set. DAVID and STRING analysis of this gene set provides disease-related pathways, functions, and protein-protein-interactions

geo2r/). The differential gene list was subjected to functional systems biology annotation as noted above. For disease sets, multiple testing correction yielded few genes having significantly differential expression. Nominal $P$-values < 0.05 were therefore used to allow for robust overrepresentation analyses. For the IBD set (normal colonic tissue control versus active IBD without anti-TNF therapy), the top 1,500 up-regulated and top 1,500 down-regulated genes were taken to create a list of 3,000 genes – the maximum number that DAVID accepts. All other datasets resulted in differential gene lists of fewer than 3,000 genes.

Using drug-target signatures from TMFS and the DGIdb [34], a comprehensive resource of experimentally determined drug-target associations curated from multiple large publically available databases, drugs were associated with diseases using the hypergeometric test (Fig. 3a) in R [35] at each of the following biological levels: direct protein targets, cell signaling pathways, molecular functions and PPIs. Drugs with $P < 0.05$ had their $P$-values log-transformed and normalized to the value of the most significantly-associated drug, resulting in values on the 0–1 unit range as illustrated in Fig. 3b. All non-significant $P$-values were automatically normalized to a value of 0. Normalization minimizes discrepancies found in the $P$-value ranges between different biological effect categories.

For each drug $i$, normalized values corresponding to each biological effect tier were used to calculate a drug-disease association Z-score used for ranking:

$$Z_i = aA + bB + cC + dD \quad (2)$$

where $A$, $B$, $C$, and $D$ correspond to the normalized values for drug-direct target, –pathway, –function, and –PPI associations, respectively. In illustrative Fig. 3c, A, B, C, and D correspond to zgene, zpathway, zfunction, and zPPI, respectively. Associated weights a, b, c, and d were set to the values of 2, 1, 0.5, and 0.25, respectively, as to prioritize direct binding of disease-regulated gene products with each subsequent level of activity receiving lower weights (Fig. 3c). This configuration was determined to best prioritize experimentally validated drugs for the given indication, and allowed for drugs highly associated with disease mechanisms at pathway, function, and indirect proteins levels to be recognized as candidates even when gene-level significance of association was poor. PPIs were given the least weight as many interactions tend to occur simultaneously within the diseased cell and prioritizing relevant interactions is difficult due to the simultaneous expression of thousands of proteins. Drugs are ranked in descending order by Z-score (Fig. 3d). High Z-scores indicate a drug's potential to most significantly and simultaneously target the greatest amount of direct proteins, pathways, functions and PPIs associated with the disease.

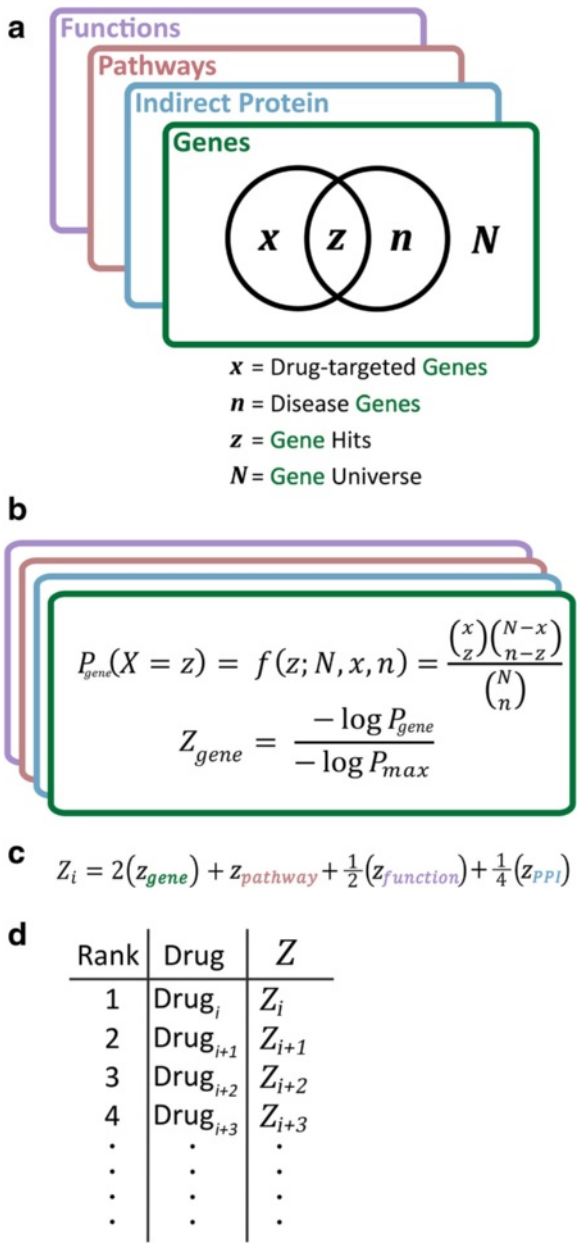



**Fig. 3** Hypergeometric test schematic for drug-disease association at each level of biological activity. Each drug is associated with a given disease at each level of biological action by the hypergeometric test. **a** Given a gene, pathway, function, or indirect protein 'universe', the hypergeometric test allows one to determine the probability that coincident drawings between two samples drawn from that universe is due to random chance. Therefore, the statistical significance of having hits (common items) between drug-associated biological factors and disease-associated factors is derived. **b** Computation of hypergeometric $p$-values and subsequent normalization for integration into cumulative score. **c** Computation of drug-disease association Z-score. **d** Ranking scheme by drug-disease association Z-score in descending order. That is, Zi exhibits the highest system-wide statistical association (highest-magnitude Z-score), followed by Zi + 1, Zi + 2, Zi + 3, and so forth

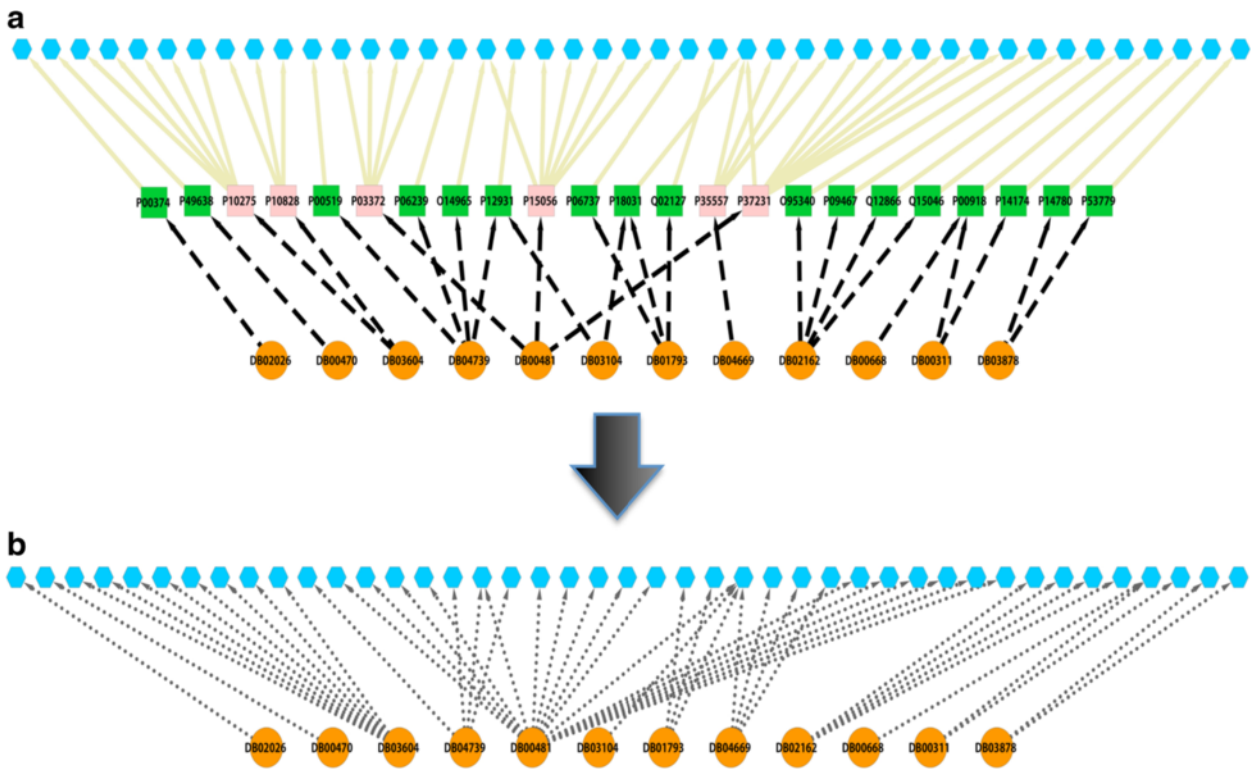



**Fig. 4** Formation of drug-target (DT) disease networks. A random sample of drugs with predicted protein targets known to be associated with a disease in OMIM were selected to illustrate the process of associating drugs with diseases. **a** Drugs (orange circle nodes) are connected using a charcoal dashed edge to predicted protein targets (square nodes); the protein targets are connected using a solid tan edge to a disease if the protein has disease genes associated with the disease. Pink nodes represent proteins associated with multiple diseases, while green nodes represent proteins associated with a single. These interactions were used to form a drug-target disease network. **b** The drugs (orange circle nodes) are connected to a disease if a predicted drug-target has disease genes associated with the disease

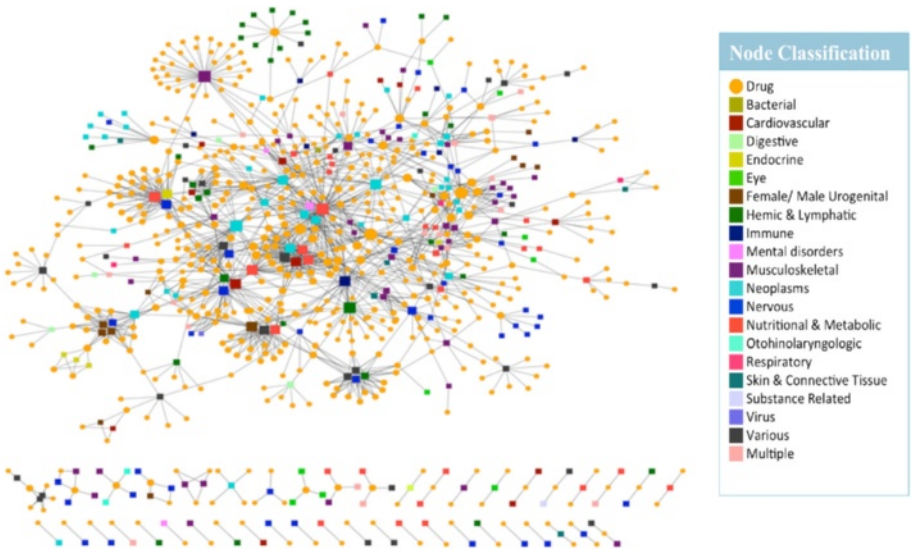



**Fig. 5** Predicted drug-target (DT) disease network. The DT disease bipartite network is generated using the top 1-ranked DT predictions and disorder-disease gene associations from OMIM. Drug nodes (circles) are connected to disease nodes (squares) if a drug is predicted to target a protein that has disease genes associated with the disease. Disease nodes are colored according to their MeSH disease category; color classification given in legend. The size of node is proportional to the number of degrees (connections)

Thus, drugs with the highest Z-scores are prioritized for repurposing due to their systems-wide effects.

## Results & discussions

### Prediction of empirical drug-disease associations

DGE-NET predicted drug associations to diseases with known etiologies by way of direct gene aberrations, as annotated in OMIM (Fig. 4). The DT-disease network contains 562 drugs (only those appearing as the top 1-ranked for their respective protein target) and 296 diseases, with the largest component containing 498 drugs (Fig. 5; Additional file 1: Table S1). The neoplasm and "nutritional and metabolic" disease classes are found centrally, reflecting the large number of drugs already approved for them and a notable potential for repurposing. Given their topology in the network, associated drugs have potential polypharmacology to other disease classes. More specialized diseases tend to occupy peripheral areas of the DT-disease network, exhibiting a smaller degree of node connectivity and suggesting increasingly unique pathogenic factors. Such diseases include digestive, urogenital, "hemic and lymphatic", and respiratory disorders. By contrast, the DT-cancer network exhibits high connectivity, with the average degree of drug nodes being 1.7 and 57 of 159 having a degree higher than 1 (Fig. 6). 26 drugs are predicted to target colorectal cancer, several of which are also predicted to target breast cancer. This is reflected in clinical practice, where several drugs are utilized across multiple cancers. The biologically sensible topology of the network provides further validation: biologically-related cancers are

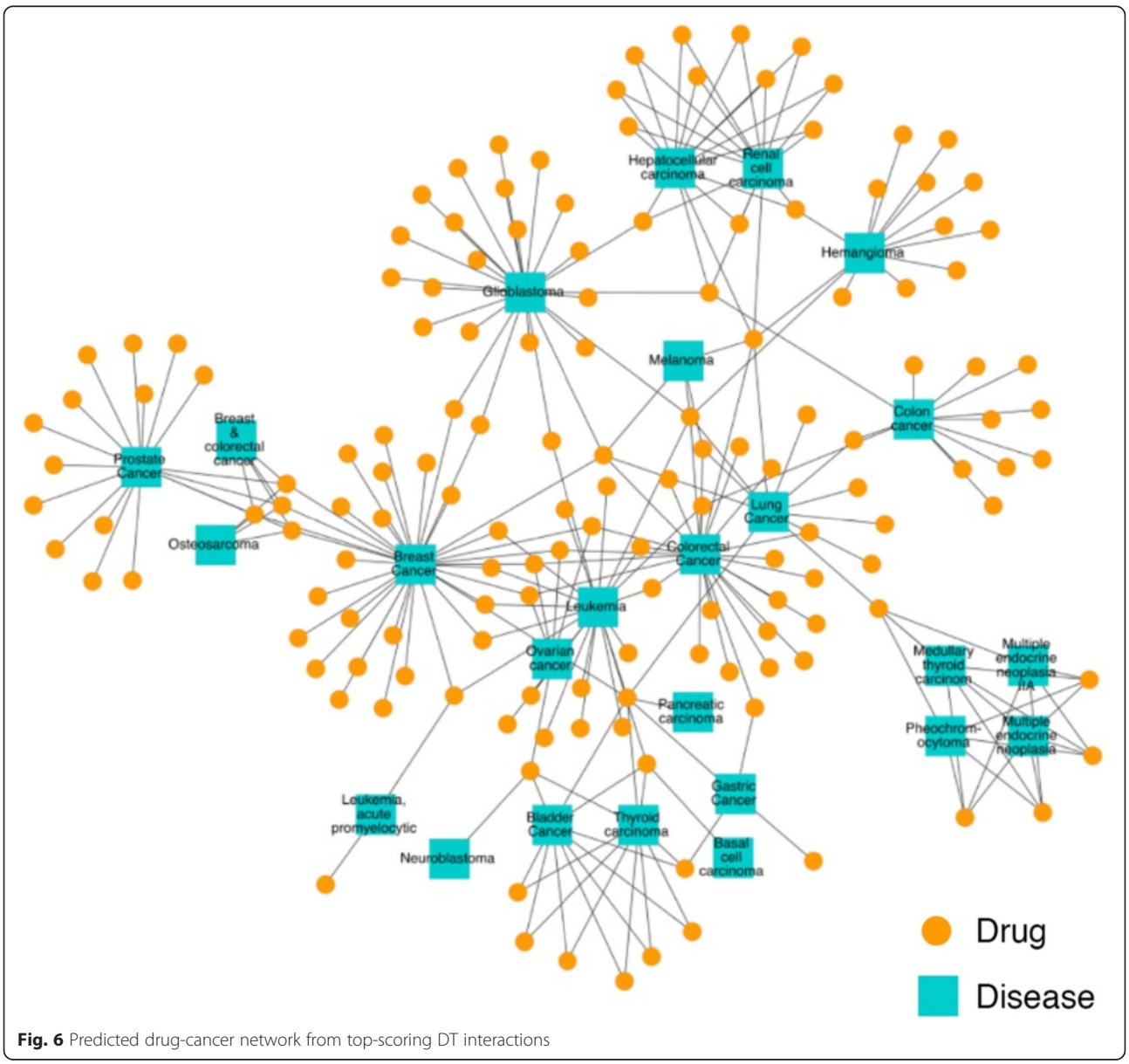



**Fig. 6** Predicted drug-cancer network from top-scoring DT interactions

clustered together through their predicted drugs. For instance, the bottom right cluster contains the endocrine gland tumors medullary thyroid carcinoma, multiple endocrine neoplasia (MEN), and pheochromocytoma, whereas the unique endothelial-originating hemangioma is found isolated in the top right.

Drug-disease predictions were validated via data found in the primary literature (Additional file 2: Table S2). Out of 526 predicted drug-disease associations, 51 were validated. Full coverage is not attainable, as many drug-disease associations have not yet been examined. Nonetheless, some predicted drug-disease combinations have been well studied, such as lisinopril for diabetes-associated microvascular complications [36]. Other associations include the anti-hookworm mebendazole for hepatocellular carcinoma and the antibiotic ceftriaxone for bladder cancer. Thus, for diseases with strong single-gene known associations, DGE-NET is able to reliably predict clinically relevant drug-disease associations by forming accurate drug-target associations. These data collectively demonstrate the ability of DGE-NET to establish known and novel drug-disease associations.

### Expansion of the drug-target prediction space to systems pharmacology

Many diseases exhibit complexity in implicating multiple perturbations rather than single deciding gene associations, and this necessitates a complex systems pharmacology perspective for clinical treatment. Drugs were therefore associated with pathways using KEGG annotations of their predicted targets. Mazindol (DB00368) and sulfadiazine (DB00359) had the least number of predicted pathways (Fig. 7). Mazindol is a tricyclic anorexigenic known to affect the noradrenergic, dopaminergic and serotonergic pathways (KEGG Drug D00367). Sulfadiazine is a sulfonamide used to treat bacterial infections by specifically inhibiting the folate biosynthesis pathway (KEGG Drug D00587). DGE-NET was able to recapitulate their specificity for those pathways. Alternatively, kinase inhibitors and nucleoside analogs such as nelarabine (DB01280) disrupt multiple pathways (Fig. 7). The KEGG Drug corpus was also used to validate 103 drug-pathway associations across 59 drugs (Table 1). Thus, DGE-NET is able to reliably associate drugs with biological pathways important in disease processes.

DGE-NET also related predicted DT signatures to molecular functions (Fig. 7). Deferasirox, an iron chelator, was predicted to affect the greatest number of molecular functions. According to the Institute for Safe Practices, deferasirox was the second most suspected drug in reported patient deaths [37]. This may be due to its potential to disrupt many molecular functions as predicted by DGE-NET. Anti-neoplastic drugs were also predicted to alter a large number of functions (Fig. 8). This reflects

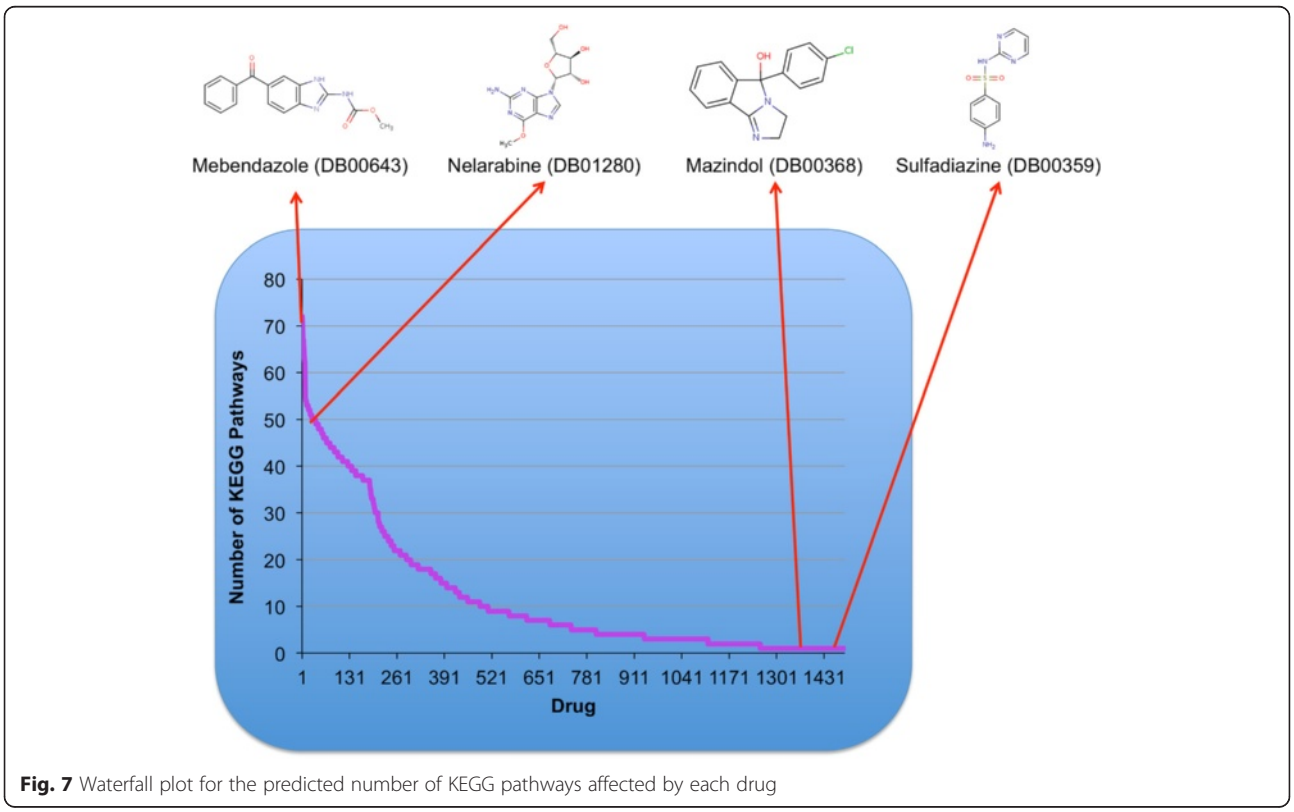



**Fig. 7** Waterfall plot for the predicted number of KEGG pathways affected by each drug

**Table 1** Validations of predicted drug-pathway associations via the KEGG Drug database

| Drug | KEGG Drug ID | KEGG Pathway |
|---|---|---|
| Acetohexamide | D00219 | Type II diabetes mellitus |
| Aripiprazole | D01164 | Gap junction |
| Bezafibrate | D01366 | Adipocytokine signaling pathway |
| Bicalutamide | D00961 | Pathways in cancer, Prostate cancer |
| Candesartan | D00626 | Vascular smooth muscle contraction |
| Carvedilol | D00255 | Vascular smooth muscle contraction |
| Celecoxib | D00567 | VEGF signaling pathway |
| Cilostazol | D01896 | Insulin signaling pathway |
| Clozapine | D00283 | Gap junction |
| Conivaptan | D01236 | Vascular smooth muscle contraction |
| Danazol | D00289 | Oocyet meiosis, Progesterone-mediated oocyte maturation, Pathways in cancer |
| Dasatinib | D03658 | MAPK signaling pathway, ErbB signaling pathway, Cytokine-cytokine receptor interaction, VEGF signaling pathway, Pathways in cancer, Chronic myeloid leukemia |
| Diflunisal | D00130 | VEGF signaling pathway |
| Domperidone | D01745 | Gap junction |
| Droperidol | D00308 | Gap junction |
| Drospirenone | D03917 | Aldosterone-regulated sodium transport |
| Dydrogesterone | D01217 | Oocyte meiosis, Progesterone-mediated oocyte meiosis |
| Eltrombopag | D03978 | Cytokine-cytokine receptor interaction, Jak-STAT signaling pathway |
| Epoprostenol | D00106 | Vascular smooth muscle contraction |
| Eprosartan | D04040 | Vascular smooth muscle contraction |
| Erlotinib | D07907 | MAPK signaling pathway, ErbB signaling pathway, Cytokine-cytokine receptor interaction, Pathways in cancer, Pancreatic cancer, Non-small cell lung cancer |
| Fenofibrate | D00565 | Adipocytokine signaling pathway |
| Floxuridine | D04197 | Pyrimidine metabolism |
| Flupenthixol | D01044 | Gap |
| Flurbiprofen | D00330 | VEGF signaling pathway |
| Flutamide | D00586 | Pathways in cancer, Prostate cancer |
| Gemcitabine | D02368 | Purine metabolism, Pyrimidine metabolism |
| Gliclazide | D01599 | Type II diabetes mellitus |
| Glipizide | D00335 | Type II diabetes mellitus |
| Haloperidol | D00136 | Gap junction |
| Imatinib | D01441 | MAPK signaling pathway, Cytokine-cytokine receptor interaction, Hematopoietic cell lineage, Pathways in cancer, Chronic myeloid leukemia |
| Indacaterol | D09318 | Endocytosis |
| Indomethacin | D00141 | VEGF signaling pathway |
| Ketoprofen | D00132 | VEGF signaling pathway |
| Lapatinib | D04024 | MAPK signaling pathway, ErbB signaling pathway, Cytokine-cytokine receptor pathway, Pathways in cancer |
| Levonorgestrel | D00950 | Oocyte meiosis, Progesterone-mediated oocyte maturation |
| Losartan | D08146 | Vascular smooth muscle contraction |
| Methysergide | D02357 | Gap junction |
| Milrinone | D00417 | Progesterone-mediated oocyte maturation |
| Mitiglinide | D01854 | Type II diabetes mellitus |
| Naproxen | D00118 | VEGF signaling pathway |
| Nilutamide | D00965 | Pathways in cancer, Prostate cancer |
| Norethindrone | D00182 | Oocyte meiosis, Progesterone-mediated oocyte maturation |
| Olmesartan | D01204 | Vascular smooth muscle contraction |
| Oxaprozin | D00463 | VEGF signaling pathway |
| Piroxicam | D00127 | VEGF signaling pathway |
| Progesterone | D00066 | Oocyte meiosis, Progesterone-mediated oocyte maturation |
| Propericiazine | D01485 | Gap junction |
| Regadenoson | D05711 | Vascular smooth muscle contraction |
| Risperidone | D00426 | Vascular smooth muscle contraction, Gap junction |
| Salsalate | D00428 | VEGF signaling pathway |
| Silodosin | D01965 | Vascular smooth muscle contraction |
| Sorafenib | D08524 | MAPK signaling pathway, ErbB signaling pathway, Cytokine-cytokine receptor interaction, Chemokine signaling pathway, mTOR signaling pathway, VEGF signaling pathway, Natural killer cell mediated cytotoxicity, Pathways in cancer, Renal cell carcinoma |
| Sulindac | D00120 | VEGF signaling pathway |
| Sunitinib | D06402 | MAPK signaling pathway, Cytokine-cytokine receptor interaction, VEGF signaling pathway, Pathways in cancer |
| Telmisartan | D00627 | Vascular smooth muscle contraction |
| Testosterone | D00075 | Pathways in cancer, Prostate cancer |
| Vandetanib | D06407 | MAPK signaling pathway, ErbB signaling pathway, Cytokine-cytokine receptor interaction, VEGF signaling pathway, Pathways in cancer |

their polypharmacology as a class of drugs, as they are designed to affect cell signaling and growth through multiple mechanisms. As a result, these drugs also exhibit high toxicity. Such analysis of molecular function can have the advantage of identifying broad- or specific-acting drugs for enriched clinical efficacy or minimized toxicity.

The incorporation of protein-protein interactions (PPIs) further increased the robustness of DGE-NET,

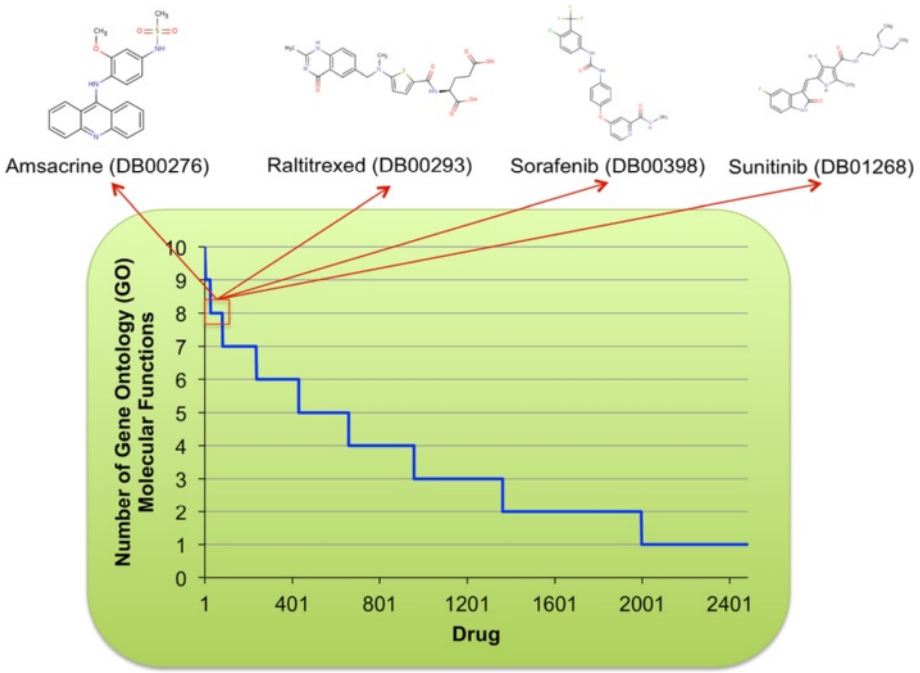



**Fig. 8** Waterfall plot for the predicted number of GO molecular functions affected by each drug. Inset highlights four anti-neoplastic drugs predicted to disrupt the greatest number of functions from the anti-neoplastic drug class

providing insight into unexpected biological similarities among drugs. For example, fluoxymesterone (DB01185) and amscarine (DB00276) are chemically and structurally unrelated. However, our method predicted that they would bind androgen receptor and B-Raf, respectively, both of which interact with MAPK1. It is through the PPI with MAPK1 that these drugs link to pathways in cancer (KEGG hsa:05200). Other drug-PPI validations are listed in Table 2 [38–48]. To highlight the importance of PPIs in attaining a mechanistic understanding of drug effects, we specifically assessed the predicted effects of ezetimibe (Fig. 9; Additional file 3: Table S3). Ezetimibe (DB00983) is a cholesterol-lowering drug used for improving cardiovascular health and has also been associated with increased incidence of cancer [49, 50]. PPIs derived from predicted targets for ezetimibe are highly clustered, indicating that the affected biological space is tightly coordinated through those targets and greatly perturbed by the actions of ezetimibe (Fig. 9). These clustered interacting targets are mainly involved in cell growth, differentiation and signal transduction. Functional annotation using both direct and indirect ezetimibe targets implicates pathways and functions involved in carcinogenesis (Additional file 3: Table S3). Thus, the present DGE-NET prediction of ezetimibe's pro-tumorigenic effects warrants further investigation.

**Table 2** Validations of predicted drug-PPI interactions

| Drug Name | Protein #1 (direct binding partner) | Protein #2 (PPI) | Reference |
|---|---|---|---|
| Bicalutamide | ABL1 | CASP9 | Danquah et al. Pharm Res. 26(9):2081–92. (2009) [38] |
| | ABL1 | CCNA2 | Katayama et al. Int J Oncol. 36(3):553–62. (2010) [39] |
| | ABL1 | MAPK11 | Malinowska et al. Endocr Relat Cancer. 16:155–169. (2009) [40] |
| Cladribine | ADA | DCK | Sasvári-Székely et al. Biochem Pharmacol. 56(9):1175–1179. (1998) [41] |
| Chlordiazepoxide | AKT1 | NR3C1 | Curtin et al. Brain Behav, Immun. 23(4): 535–547. (2009) [42] |
| Progeterone | AR | F2 | Oger et al. Arterioscler Thromb Vasc Biol. 23:1671–1676. (2003) [43] |
| Cyproterone | AR | CASP3 | Eckle et al. Toxicol Pathol. 32:9–15. (2004) [44] |
| | AR | NR3C1 | Honer et al. Mol Pharmacol. 63(5):1012–1020. (2003) [45] |
| Telmisasrtan | BCL2 | IL2 | Syrbe et al. Hypertens Res. 30(6):521–527. (2007) [46] |
| Sorafenib | BRAF | PRKCQ | Jane et al. J Pharmacol Exp Ther. 319(3):1070–1080. (2006) [47] |
| Methotrexate | DHFR | CDK2 | Maddika et al. J Cell Sci. 121:979–988. (2008) [48] |

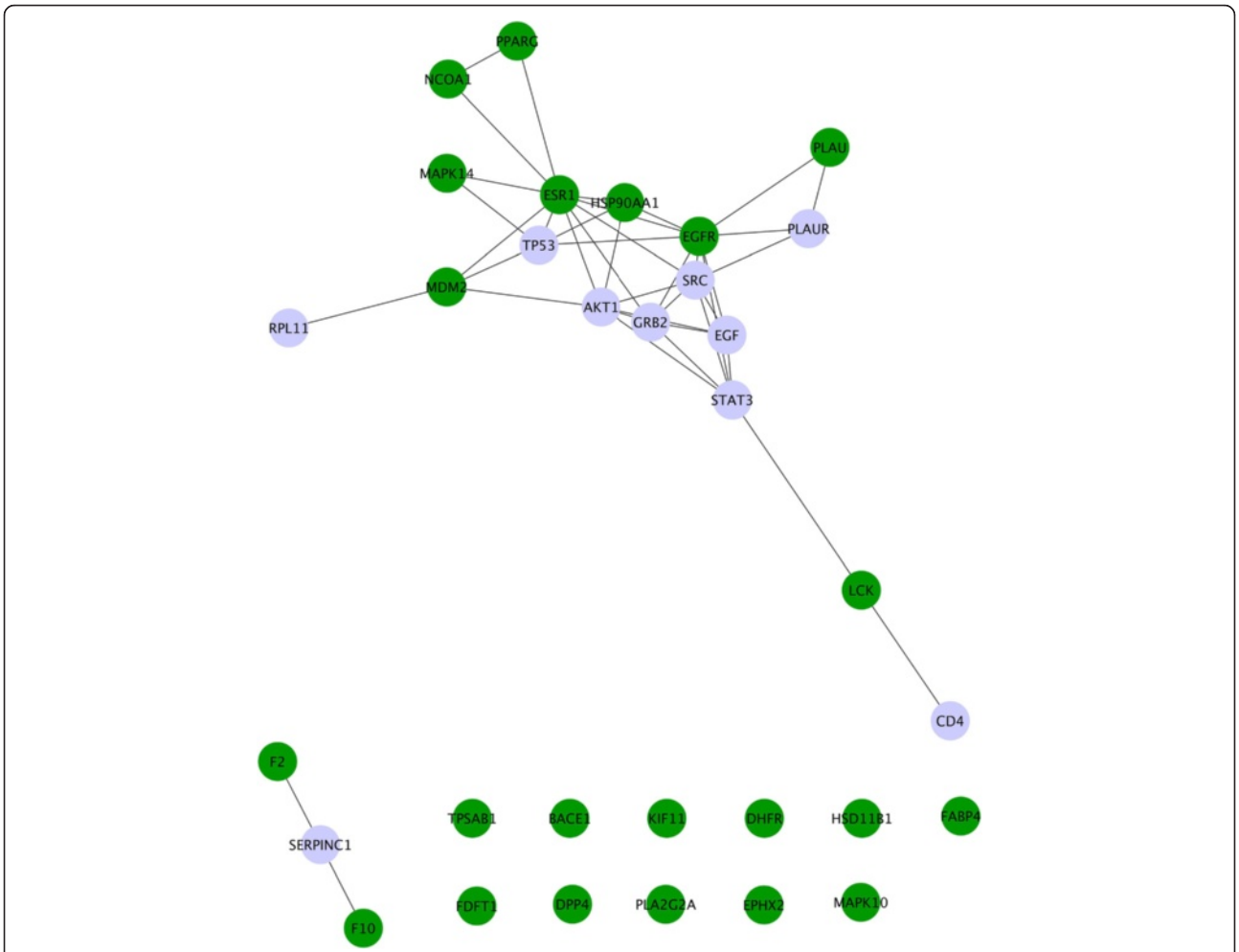



**Fig. 9** Ezetimibe protein-protein interaction (PPI) network. Direct targets (green nodes) predicted for ezetimibe from TMFS were used to establish interactions between direct targets as well as indirect targets (light purple nodes) using the ExPASy STRING database with a confidence score cutoff greater than 0.95

### Incorporation of autoimmune disease-related gene expression data for polypharmacology-driven drug repurposing

Autoimmune diseases are systemic or local pro-inflammatory pathologies with multiple etiologies. Current therapeutics such as corticosteroids, methotrexate and anti-TNF biologics focus on regulating inflammation, and immunosuppression. In addition to acting non-specifically these medications do not address the full extent of effector tissue pathobiology. A treatment approach rooted in polypharmacology may be more efficacious and offers the potential for limiting side effects. For proof-of-concept, we apply DGE-NET as a gene expression-based polypharmacology prediction method (Fig. 2) for rheumatoid arthritis and inflammatory bowel disease.

Rheumatoid arthritis (RA) is a painful multi-joint destructive disease. Joint synovium, usually 1–2 cells thick, becomes inflamed and reaches multicellular thickness due to infiltration of immune effector cells and activation and subsequent proliferation of fibroblast-like synoviocytes (FLS). Cellular molecular cross-talk, infiltration and proliferation lead to pannus formation, which acts analogously to an invasive tumor and causes joint destruction. As FLS cells are critical mediators of RA, we applied our method using differentially expressed genes when comparing activated FLS cells from RA patients and quiescent FLS cells from non-RA patients (GSE55235 and GSE55457). A consensus drug list was constructed by combining the top 100 (~ Top 10 % of total drug database) predicted drugs for each study and extracting those that are present in one or both lists, ranked by mean association Z-score (Additional file 4: Table S4). Shown in Table 3 are those drugs from the consensus drug list that are currently used for RA, or have been found to be potentially useful in the clinic [51–54, 97–103]. Drugs currently used in the clinic were recapitulated in our list, such as anti-TNF biologics adalimumab and etanercept, as well as the NSAID sulindac. Non-approved

**Table 3** Validations of predicted drug indications for RA and IBD from consensus drug lists, ordered by drug list ranking

| Rheumatoid Arthritis (RA) | Reference for Validation | Inflammatory Bowel Disease (IBD) | Reference for Validation |
|---|---|---|---|
| Alvocidib | Sekine et al. J Immunol. 180(3):1954–1961 (2008) [51] | Sulfasalazine | Klotz et al. N Engl J Med. 303(26):1499–1502 (1980) [105] |
| Karenitecin | Liu et al. Med Res Rev. 35(4):753-89 (2015) [53] | Olsalazine | Baumgart et al. Lancet. 369(9573):1641–1657 (2007) [106] |
| Sulindac | Brogden et al. Drugs. 16(2):97–114 (1978) [97] | Tetomilast | Keshavarzian et al. Expert Opin Investig Drugs. 16(9):1489–1506 (2007) [107] |
| Sunitinib | Fuyura et al. Mod Rheumatol. 24(3):487–491 (2013) [52] | Inosine | Mabley et al. Am J Physiol Gastrointest Liver Physiol. 284(1):G138-G144 (2003) [108] |
| INCB28050 | Taylor et al. Ann Rheum Dis. 73:A31 (2014) [99] | Thioproperazine | Lechin et al. J Clin Gastroenterol. 4(5):445–450 (1982) [56] |
| Amodiaquine | Kersley et al. Lancet. 2(7108):886–888 (1959) [54] | Etoricoxib | El Miedany et al. Am J Gastroenterol. 101(2):311–317 (2006) [109] |
| Raltitrexed | van der Heijden et al. Scand J Rheumatol. 43(1):9–16 (2014) [100] | Balsalazide | Carter et al. Gut. 53(Suppl 5):V1-V16 (2004) [110] |
| BIRB 796 | Page et al. Arthritis Rheum. 62(11):3221–3231 (2010) [101] | Thalidomide | Gerich et al. Ailment Pharmacol Ther. 41(5):429–437 (2015) [59] |
| Adalimumab | Weinblatt et al. Arthritis Rheum. 48(1):35–45 (2003) [102] | Rosiglitazone | Ramakers et al. J Clin Immunol. 27(3):275–283 (2007) [57] |
| Etanercept | Moreland et al. Ann Intern Med. 130(6):478–486 (1999) [103] | Irbesartan | Ray et al. Gut. 62(S1):A525-A525 (2013) [60] |
| Minocycline | O'Dell et al. Arthritis Rheum. 40(5):842–848 (1997) [104] | Chloroquine | Nagar et al. Int Immunopharmacol. 21(2):328–335 (2014) [111] |

drugs currently being studied for RA also appeared. These include kinase inhibitors such as alvocidib [51] and sunitinib [52], the topoisomerase inhibitor karenitecin [53], and the chloroquine-related compound amodiaquine [54]. Predicted RA indication for these drugs, which are generally anti-cancer agents, illustrates an important mechanistic underpinning of RA with respect to FLS cells in that activated FLS mimic cancer cell progression [55]. Regardless of the activating stimulus (e.g. TNF-α), our polypharmacological method focuses on downstream gene expression, signaling, and functional effects in activated FLS cells. This highlights the cancer-like mechanisms of pathogenesis and prioritizes those drugs that are able to simultaneously disrupt the greatest number of those mechanisms. In addition, because antibodies have single-target effects, we were surprised by their predicted indications for RA. However, if that target has many pathology-related pleiotropic downstream effects, such as TNF-α, then such drugs would be prioritized due to the pathway and function terms in our equation. Thus, DGE-NET is capable of making important polypharmacological associations beyond immediate gene targets.

DGE-NET also predicted drugs for inflammatory bowel disease (IBD), also a multi-etiological immune-related collection of disorders. Differential gene expression analysis was performed by, comparing normal and inflamed bowel tissues (GSE52746 and GSE11223). Like the RA dataset, a consensus list of the Top 100 drugs obtained from each IBD study was constructed (Additional file 5: Table S5), and therapeutic validations from this list are recapitulated in Table 3 [56–60, 104–110]. Our method predicted the known IBD drug sulfasalazine, serving as an important litmus. Other predicted drugs that are promising in experimental settings and from diverse chemical classes include the antipsychotic thioproperazine [56], the anti-diabetic thiazlidinedione rosiglitazone [57], the leukotriene receptor antagonist tetomilast [58], and thalidomide [59]. Interestingly, DGE-NET predicted the angiotensin receptor blocker (ARB) irbesartan as potential therapy. A recent preliminary study implicates the role of angiotensin receptors in intestinal fibrosis in Crohn's disease [60], a type of IBD, but greater investigation is needed.

In addition to recapitulating known drug associations, we predicted the drugs topotecan and mebendazole for repurposing to rheumatoid arthritis. Topotecan is a DNA topoisomerase 1 (Top1) inhibitor used for NSCLC cancer and has been given both orally and intravenously. Topoisomerases have been implicated in rheumatoid arthritis etiology [61], and the established Top1 inhibitor camptothecin (CPT) has been shown to be effective in a murine collagen-induced RA model [62]. Koo et al. developed a novel nanocarrier for CPT called CPT-SSM-VIP, which denotes micelles to overcome solubility issues and vasoactive intestinal peptide (VIP) for active targeting. As CPT provides evidence for Top1 inhibition

in RA, we also pursued topotecan. Although it can be inferred that topotecan could be an effective anti-arthritic via topoisomerase, many other unreported targets were predicted for topotecan that could mediate potential efficacy. These include multiple tyrosine-protein kinases (BTK, CSK, LCK, TTK, ITK, LYN), non-tyrosine kinases (AURK1, PIK3CG), as well as cyclin A2. Mebendazole is an anti-hookworm tubulin inhibitor with anti-cancer potential through mammalian crossover tubulin [63] and kinase inhibition [64]. We previously predicted many novel protein kinase targets for mebendazole [17]. Kinase inhibition is a sought after therapeutic strategy for rheumatoid arthritis, especially as non-biologic treatment alternatives and for methotrexate-resistant cases [65–67]. Inhibitors of spleen tyrosine kinase (Syk) and Janus kinases (Jaks) have shown short-term efficacy, but other kinases inhibitors with good long-term effect profiles may also exist. Other kinases implicated in RA pathogenesis include aurora kinases [68] and cyclin-dependent kinases (CDKs) [69]. Mebendazole may serve as a good non-biologic disease-modifying antirheumatic drug (DMARD) given its historic use, low toxicity profile, and its effect on multiple kinases.

In another proof-of-concept, we applied DGE-NET to two neurodegenerative disorders, Alzheimer's disease (AD) and Parkinson's disease (PD). Table 4 summarizes those drugs predicted to be in the top 50 for AD and PD by DGE-NET that are currently validated for standard or potential therapeutic use [70–86]. The complete top 50 predicted drugs for these diseases and their validations are found in Additional file 6: Table S6. Others listed are currently undergoing pre-clinical or clinical investigation. Of note is that memantine, an approved drug for AD, appears beyond the top 50 but within the top 500. This drug exhibits less polypharmacology but is still effective given the importance of its direct targets and pathways for AD disease processes (i.e. NMDA receptor antagonism reducing glutamate excitotoxicity of neurons [87]). Thus, it can be hypothesized thatdrugs found higher up in the rank list may be more effective than the current clinical standards of care as those drugs theoretically alter a greater proportion of disease-associated protein targets and biological effects simultaneously.

Sunitinib has been identified as a lead candidate having the potential to mitigate the development of oxidant injury to endothelial cells associated with AD [79]. Sunitinib could affect the vascular activation mechanisms of pathogenesis in AD by reducing the expression of amyloid beta, thrombin, tumor necrosis factor alpha, interleukin-1 beta, interleukin-6, and matrix metalloproteinase 9, and other factors associated with neurodegenerative disorders [79, 88, 89]. This anti-angiogenic property has been

**Table 4** Validations of top 50 predicted drug indications for AD and PD, ordered by ranking

| Alzheimer's Disease (AD) | Reference for Validation | Parkinson's Disease (PD) | Reference for Validation |
|---|---|---|---|
| Rasagiline | Weinreb et al. Neurotherapeutics. (6)1:163–74. (2009) [70] | Dextroamphetamine | Parkes et al. J Neurol Neurosurg Psychiatry. 38(3):232–7 (1975) [83] |
| Interferons | Grimaldi et al. J Neuroinflammation. 11:30 (2014) [71] | Orphenadrine | Bersani et al. Clin Neuropharmacol. 13(6):500–6 (1990) [84] |
| Calcium | Woods et al. Adv Exp Med Biol. 740:1193–217 (2012) [72] | Quinacrine | Tariq et al. Brain Res Bull. 54(1):77–82 (2001) [85] |
| Dovitinib | Li et al. Medical Hypotheses. (80)4:341–44. (2013) [73] | Atomoxetine | Weintraub et al. Neurology. 75(5):448–55 (2010) [86] |
| Somatropin Recombinant | Ling et al. Growth Horm IGF Res. (17)4:336–41 (2007) [74] | | |
| Aripiprazole | De Deyn et al. Expert Opin. Pharmacother. (14)4:459–74 (2013) [75] | | |
| Clozapine | Tariot et al. Clin Geriatr Med. (17)2:359–76 (2001) [76] | | |
| Quercetin | Ansari et al. J Nutr Biochem. 20(4):269–75 (2009) [77] | | |
| Flavopiridol | Pallàs et al. Med Hypotheses. 64(1):120–3 (2005) [78] | | |
| Sunitinib | Grammas et al. J Alzheimers Dis. 40(3):619–30 (2014) [79] | | |
| Risperidone | Katz et al. Int J Geriatr Psychiatry. (60)2:107–15 (2007) [80] | | |
| Genistein | Valles et al. Brain Res. 1312:138–44 (2010) [81] | | |
| Dasatinib | Dhawan et al. J Neuroinflammation. 9:117 (2012) [82] | | |

previously shown to be a major component of the anti-cancer acitivity of sunitinib [90]. Figure 10 illustrates the polypharmacology of sunitinib, at each level of biological activity, predicted by DGE-NET to coincide with significantly AD-associated factors. Single-agent or combination therapies that exploit multiple aspects of disease process are assumed to be efficacious, requiring lower dosages than current therapies and reducing the likelihood of resistance.

In addition to therapeutic drug repurposing candidates, DGE-NET reported drugs that are known to be contra-indicated for their respective diseases. Minocycline and tretinoin, both of which are used to treat acne, may have IBD toxicity. Minocycline is a tetracycline antimicrobial with a potential association with IBD (Additional file 5: Table S5) [91]. Tretinoin is a topical retinoid that is structurally related to isotretinoin, an oral medication used for more severe acne. While tretinoin itself is safe, isotretinoin has been implicated in causing IBD (Additional file 5: Table S5) [92], though this finding is controversial. It could be extrapolated that if tretinoin was given orally and at higher doses that IBD may be a consequence. Others include methysergide, a prophylactic drug that is contra-indicated for RA and other collagen diseases (Additional file 4: Table S4) [93], indomethacin, a non-selective non-steroidal anti-inflammatory drug known to exacerbate IBD (Additional file 5: Table S5) [94, 95], quetiapine, an atypical antipsychotic associated with increased cognitive decline in AD (Additional file 6: Table S6) [96], and methamphetamine, which has been linked with an increased risk of PD (Additional file 6: Table S6), [97]. The appearance of these drugs is likely due to DGE-NET not discriminating between agonistic and antagonistic effects of drugs but rather forming non-directional drug-target-effect associations. Counter-therapeutic drug actions are therefore incorporated, so long as they correspond with disease-associated biological activity.

## Conclusions

DGE-NET is able to predict drug-target interactions and contextualize their biological effects at the levels of protein-protein interactions, biological pathways, and molecular functions. It further integrates gene expression signatures for identification of systems-based disease-relevant targets and prioritization of drugs that exhibit a desired polypharmacology. DGE-NET recapitulated known therapeutic and contraindicated drugs for rheumatoid arthritis and inflammatory bowel disease and led to the identification of mebendazole

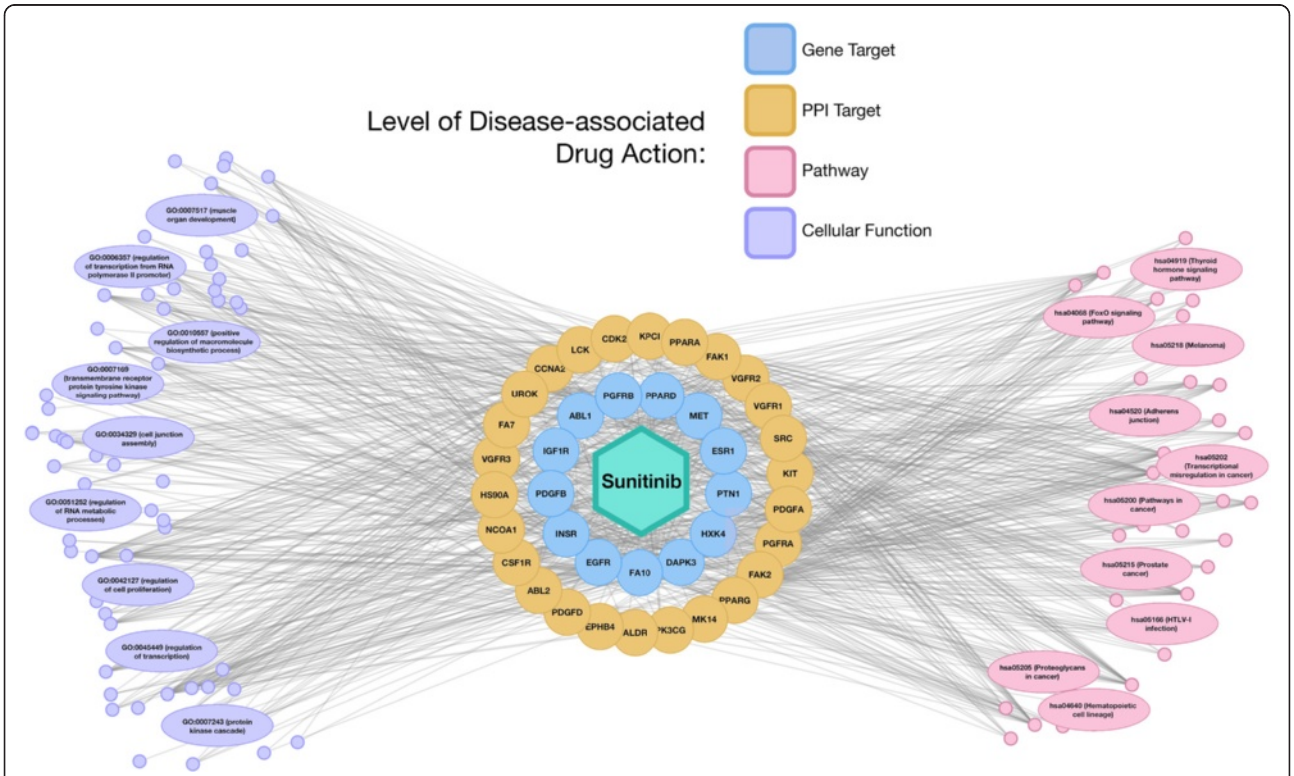



**Fig. 10** Predicted sunitinib drug action network on AD. Direct protein targets predicted by DGE-NET for sunitinib that are also significantly AD-modulated are in large orange and blue circles. Blue circles are genes overexpressed in AD with statistical significance, while orange circles are protein partners of those genes. Pink circles are KEGG pathways, and purple circles are GO cellular functions, enriched at $p$-value < .01 in the up-regulated genes of AD. The top 10 significantly enriched cellular functions and pathways are detailed in large ovals

as drug repurposing candidate for rheumatoid arthritis. Its ability to do so can also be extended to other small molecules with the potential to act as endogenous drugs to alter physiology, such as metabolites. We are currently pursuing the application of DGE-NET to cancer-associated metabolites to potentially explain the mechanisms behind metabolite-disease phenotypic associations. DGE-NET ultimately assists in the formulation of drug-disease hypotheses poised for clinical success.

Differential gene expression analysis is one way of assessing disease pathogenesis to find therapeutic targets. DGE-NET is the first computational tool that associates drugs with diseases through multiple tiers of systems biology obtained via gene expression analysis. This not only aids in finding effective drugs but helps bypass issues that arise from traditional gene sequencing approaches such as un-actionable mutations in single nucleotide polymorphisms, which is currently an important limitation in oncology. Importantly, DGE-NET in its current form does not differentiate agonist or antagonist effects of drugs. The next iteration will include this improvement so that DGE-NET can better discriminate between therapeutic agents and drugs that are contraindicated.

## Availability of data and materials

Because DGE-NET is applied to publicly available data, the authors have provided a tutorial which describes the stepwise implementation of DGE-NET, in Additional file 7.

## Additional files

**Additional file 1: Table S1.** Predicted drug-target-disease associations using OMIM. For each human protein target crystal structure, the top 40-ranked drugs were associated with a disease through their predicted target. (XLSX 464 kb)

**Additional file 2: Table S2.** Validations of predicted drug-disease associations from the literature. (XLS 38 kb)

**Additional file 3: Table S3.** Predicted systems pharmacology of ezetimibe. Targets predicted to directly associate with ezetimibe and their interacting protein partners (confidence score cutoff greater than 0.95) were used to infer pathways and molecular functions (FDR < 0.25) that could be perturbed by ezetimibe. (XLS 42 kb)

**Additional file 4: Table S4.** Consensus drug rank list for Rheumatoid Arthritis. (XLS 47 kb)

**Additional file 5: Table S5.** Consensus drug rank list for Inflammatory Bowel Disease. (XLS 51 kb)

**Additional file 6: Table S6.** Top 50 predicted drugs and validations for Alzheimer's Disease and Parkinson's Disease. (XLS 39 kb)

**Additional file 7:** Tutorial outlining the manual implementation of the DrugGenEx-NET methodology. (DOCX 14 kb)

#### Abbreviations

AD: Alzheimer's disease; CDK: Cyclin-dependent kinase; CPT: Camptothecin; CTD: Comparative toxicogenomics database; DAVID: Database for annotation, visualization, and integrated discovery; DGE-NET: DrugGenEx-Net; DMARD: Disease-modifying antirheumatic drug; DT: Drug-target; FAT: Functional annotation tool; FLS: Fibroblast-like synoviocytes; IBD: Inflammatory bowel disease; MEN: Multiple endocrine neoplasia; MeSH: Medical subject headings; OMIM: Online mendelian inheritance in man; PD: Parkinson's disease; PPI: Protein-protein interactions; RA: Rheumatoid arthritis; SDF: Spatial data file; TMFS: Train, match, fit, streamline; VIP: Vasoactive intestinal peptide.

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

NTW, JK, HW, RR and SD performed the computational studies, NTW, HW, and SD drafted the manuscript. SWB provided advice, and helped to draft the manuscript. All authors read and approved the final manuscript.

#### Acknowledgements

The authors wish to acknowledge DOD grant CA140882 (SD), R01 CA170653 (SD, SB), CCSG grant NIH-P30 CA51008 and Georgetown Lombardi Cancer Center.

#### Author details

[1]Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington DC 20057, USA. [2]Department of Biochemistry & Molecular Biology, Georgetown University, Washington DC 20057, USA. [3]Georgetown University Medical Center, Washington DC 20057, USA. [4]George Mason University, 4400 University Dr, Fairfax, VA 22030, USA.

Received: 29 December 2015 Accepted: 29 April 2016
Published online: 05 May 2016

#### References

1. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov. 2010;9:203–14.
2. Bajorath J. Rational drug discovery revisited: interfacing experimental programs with bio- and chemo-informatics. Drug Discov Today. 2001;6:989–95.
3. Chanda SK, Caldwell JS. Fulfilling the promise: drug discovery in the post-genomic era. Drug Discov Today. 2003;8:168–74.
4. Pujol A et al. Unveiling the role of network and systems biology in drug discovery. Trends Pharmacol Sci. 2010;31(3):115–23.
5. Anighoro A, Bajorath J, Rastelli G. Polypharmacology: challenges and opportunities in drug discovery. J Med Chem. 2014;57:7874–87.
6. Cheng F, Liu C, Jiang J, Lu W, Li W, et al. Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput Biol. 2012;8:e1002503.
7. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science. 2008;321:263–6.
8. Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, et al. Large-scale prediction and testing of drug activity on side-effect targets. Nature. 2012;486:361–7.
9. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313:1929–35.
10. Chen B, Ding Y, Wild DJ. Assessing drug target association using semantic linked data. PLoS Comput Biol. 2012;8:e1002574.
11. Hu G, Agarwal P. Human disease-drug network based on genomic expression profiles. PLoS One. 2009;4:e6536.
12. Medina-Franco JL, Giulianotti MA, Welmaker GS, Houghten RA. Shifting from the single to the multitarget paradigm in drug discovery. Drug Discov Today. 2013;18(9–10):495–501.
13. Knox C, Law V, Jewison T, Liu P, Ly S, et al. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res. 2011;39:D1035–41.
14. Huang R, Southall N, Wang Y, Yasgar A, Shinn P, et al. The NCGC Pharmaceutical Collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics. Sci Transl Med. 2011;3(80):80ps16.
15. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res. 2007;35:D198–201.
16. Schrödinger, LLC. Schrödinger Release 2013–3: LigPrep, version 2.8. New York: Schrödinger, LLC; 2013.

17. Dakshanamurthy S, Issa NT, Assefina S, Seshasayee A, Peters OJ, et al. Predicting new indications for approved drugs using a proteochemometric method. J Med Chem. 2012;55:6832–48.
18. Schrödinger, LLC. Small-Molecule Drug Discovery Suite 2013–3: Glide, version 6.1. New York: Schrödinger, LLC; 2013.
19. Schrödinger, LLC. Small-Molecule Drug Discovery Suite 2013–3: QikProp, version 3.8. New York: Schrödinger, LLC; 2013.
20. Schrödinger, LLC. Small-Molecule Drug Discovery Suite 2013–3: Strike, version 2.4. New York: Schrödinger, LLC; 2013.
21. Kahraman A, Morris R, Laskowski R, Thornton J. Shape variation in protein binding pockets and their ligands. J Mol Biol. 2007;368:283–301.
22. The UniProt Consortium. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. 2014;D1:D191–8.
23. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, et al. The human disease network. Proc Natl Acad Sci U S A. 2007;104:8685–90.
24. Yildirim MA, Goh KI, Cusic ME, Barabasi AL, Vidal M. Drug-target network. Nat Biotechnol. 2007;25:1119–26.
25. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28:27–30.
26. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, et al. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 2014;42:D199–205.
27. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.
28. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, et al. AmiGO: online access to ontology and annotation data. Bioinformatics. 2009;25:288–9.
29. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 2009; 4:44–57.
30. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37:1–13.
31. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2013;D1:D808–15.
32. Davis AP, Murphy CG, Johnson R, Lay JM, Lennon-Hopkins K, et al. The comparative toxicogenomics database: update 2013. Nucleic Acids Res. 2013;D1:D1104–14.
33. Barrett T, Whilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: archive for functional genomics data sets- update. Nucleic Acids Res. 2013; D41:D991–5.
34. Griffith M, Griffith OL, Coffman AC, Weible JV, McMichael JF, et al. DGIdb – Mining the druggable genome for personalized medicine. Nat Methods. 2013;10:1209–10.
35. R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
36. Zuanetti G, Latini R, Maggioni AP, Franzosi M, Santoro L, et al. Effect of the ACE inhibitor lisinopril on mortality in diabetic patients with acute myocardial infarction: data from the GISSI-3 study. Circulation. 1997;96: 4239–45.
37. ISMP (2010). "ISMP QuarterWatch(TM)" 15 (12). ISMP Medication Safety Alert. pp. 1–3
38. Danquah M, Li F, Duke III CB, Miller DD, Mahato RI. Micellar delivery of bicalutamide and embelin for treating prostate cancer. Pharm Res. 2009;26: 2081–92.
39. Katayama H, Murashima T, Saeki Y, Nishizawa Y. The pure anti-androgen bicalutamide inhibits cyclin A expression both in androgen-dependent and-independent cell lines. Int J Oncol. 2010;36:553–62.
40. Malinowska K, Neuwirt H, Cavarretta IT, Bektic J, Steiner H, Dietrich H, Moser PL, Fuchs D, Hobisch A, Culig Z. Interleukin-6 stimulation of growth of prostate cancer in vitro and in vivo through activation of the androgen receptor. Endocr Relat Cancer. 2009;16(1):155–69. doi:10.1677/ERC-08-0174. Epub2008Nov14.
41. Sasvári-Székely M, Spasokoukotskaja T, Szóke M, Csapó Z, Turi Á, Szántó I, et al. Activation of deoxycytidine kinase during inhibition of DNA synthesis by 2- chloro-2′-deoxyadenosine (Cladribine) in human lymphocytes. Biochem Pharmacol. 1998;56:1175–9.
42. Curtin NM, Boyle NT, Mills KH, Connor TJ. Psychological stress suppresses innate IFN-γ production via glucocorticoid receptor activation: Reversal by the anxiolytic chlordiazepoxide. Brain Behav Immun. 2009;23:535–47.
43. Oger E, Alhenc-Gelas M, Lacut K, Blouch MT, Roudaut N, Kerlan V, et al. Differential effects of oral and transdermal estrogen/progesterone regimens on sensitivity to activated protein C among postmenopausal women a randomized trial. Arterioscler Thromb Vasc Biol. 2003;23:1671–6.
44. Eckle VS, Buchmann A, Bursch W, Schulte-Hermann R, Schwarz M. Immunohistochemical detection of activated caspases in apoptotic hepatocytes in rat liver. Toxicol Pathol. 2004;32:9–15.
45. Honer C, Nam K, Fink C, Marshall P, Ksander G, Chatelain RE, et al. Glucocorticoid receptor antagonism by cyproterone acetate and RU486. Mol Pharmacol. 2003;63:1012–20.
46. Syrbe U, Moebes A, Scholze J, Swidsinski A, Dorffel Y. Effects of the angiotension II type 1 receptor antagonist Telmisartan on monocyte adhesion and activation in patients with essential hypertension. Hypertens Res. 2007;30:521.
47. Jane EP, Premkumar DR, Pollack IF. Coadministration of sorafenib with rottlerin potently inhibits cell proliferation and migration in human malignant glioma cells. J Pharmacol Exp Ther. 2006;19:1070–80.
48. Maddika S, Ande SR, Wiechec E, Hansen LL, Wesselborg S, Los M. Akt-mediated phosphorylation of CDK2 regulates its dual role in cell cycle progression and apoptosis. J Cell Sci. 2008;121:979–88.
49. Drazen JM, D'Agostino RB, Ware JH, Morrissey S, Curfman GD. Ezetimibe and cancer – an uncertain association. N Engl J Med. 2008;359:1398–9.
50. Rossebo AB, Pedersen TR, Boman K, et al. Intensive lipid lowering with simvastatin and ezetimibe in aortic stenosis. N Engl J Med. 2008;359:1343–56.
51. Sekine C et al. Successful treatment of animal models of rheumatoid arthritis with small-molecule cyclin-dependent kinase inhibitors. J Immunol. 2008;180(3):1954–61.
52. Furuya K et al. Therapeutic effects of sunitinib, one of the anti-angiogenetic drugs, in a murine arthritis. Mod Rheumatol. 2014;24(3):487–91.
53. Liu YQ, Li WQ, Morris-Natschke SL, et al. Perspectives on biologically active camptothecin derivatives. Med Res Rev. 2015;35(4):753–89.
54. Kersley GD, Palin AG. Amodiaquine and hydroxychloroquine in rheumatoid arthritis. Lancet. 1959;274(7108):886–8.
55. Ainola MM et al. Pannus invasion and cartilage degradation in rheumatoid arthritis: involvement of MMP-3 and interleukin-1b. Clin Exp Rheumatol. 2005;23:644–50.
56. Lechin F et al. Treatment of ulcerative colitis with thioproperazine. J Clin Gastroenterol. 1982;4(5):445–50.
57. Ramakers JD et al. The PPARγ agonist rosiglitazone impairs colonic inflammation in mice with experimental colitis. J Clin Immunol. 2007;27(3): 275–83.
58. Schreiber S, et al. A randomized, placebo-controlled, phase II study of tetomilast in active ulcerative colitis. Gastroenterology. 2007;132(1):76–86.
59. Gerich ME, et al. Long-term outcomes of thalidomide in refractory Crohn's disease. Aliment Pharmacol Ther. 2015;41(5):429–37.
60. Ray S et al. PTH-102 preliminary evidence for a role of the renin angiotensin system in intestinal fibrosis in crohn's disease using angiotensin receptor immunohistochemistry. Gut. 2013;62 Suppl 1:A252–2.
61. Jackson JK et al. Topoisomerase inhibitors as anti-arthritic agents. Inflamm Res. 2008;57(3):126–34.
62. Koo OM, Rubinstein I, Önyüksel H. Actively targeted low-dose camptothecin as a safe, long-acting, disease-modifying nanomedicine for rheumatoid arthritis. Pharm Res. 2011;28(4):776–87.
63. Sasaki J, Ramesh R, Chada S, Gomyo Y, Roth JA, Mukhopadhyay T. The anthelmintic drug mebendazole induces mitotic arrest and apoptosis by depolymerizing tubulin in non-small cell lung cancer cells. Mol Cancer Ther. 2002;1:1201–9.
64. Li L, Liu Y, Zhang Q, Zhou H, Zhang Y, Yan B. Comparison of cancer cell survival triggered by microtubule damage after turning Dyrk1B kinase on and off. ACS Chem Biol. 2014;9:731–42.
65. Opar A. Kinase inhibitors attract attention as oral rheumatoid arthritis drugs. Nat Rev Drug Discov. 2010;9(4):257–8.
66. Gomez-Puerta JA, Mócsai A. Tyrosine kinase inhibitors for the treatment of rheumatoid arthritis. Curr Top Med Chem. 2013;13:760.
67. Cohen S et al. Co-administration of the JAK inhibitor CP-690,550 and methotrexate is well tolerated in patients with rheumatoid arthritis without need for dose adjustment. Br J Clin Pharmacol. 2010;69(2):143–51.
68. Glant TT, et al. Differentially expressed epigenome modifiers, including aurora kinases A and B, in immune cells in rheumatoid arthritis in humans and mouse models. Arth Rheum. 2013;65(7):1725–35.

69. Raychaudhuri S, Remmers EF, Lee AT, et al. Common variants at CD40 and other loci confer risk of rheumatoid arthritis. Nat Genet. 2008;40(10):1216–23.
70. Weinreb O, Mandel S, Bar-Am O, Yogev-Falach M, Avramovich-Tirosh Y, Amit T, Youdim MB. Multifunctional neuroprotective derivatives of rasagiline as anti-Alzheimer's disease drugs. Neurotherapeutics. 2009;6(1):163–74.
71. Grimaldi LM, Zappalà G, Iemolo F, Castellano AE, Ruggieri S, Bruno G, Paolillo A. A pilot study on the use of interferon beta-1a in early Alzheimer's disease subjects. J Neuroinflammation. 2014;11:30.
72. Woods NK, Padmanabhan J. Neuronal calcium signaling and Alzheimer's disease. Adv Exp Med Biol. 2012;740:1193–217.
73. Li JS, Yao ZX. Modulation of FGF receptor signaling as an intervention and potential therapy for myelin breakdown in Alzheimer's disease. Med Hypotheses. 2013;80(4):341–4.
74. Ling FA, Hui DZ, Ji SM. Protective effect of recombinant human somatotropin on amyloid beta-peptide induced learning and memory deficits in mice. Growth Horm IGF Res. 2007;17(4):336–41.
75. De Deyn PP, Drenth AF, Kremer BP, Oude Voshaar RC, Van Dam D. Aripiprazole in the treatment of Alzheimer's disease. Expert Opin Pharmacother. 2013;14(4):459–74.
76. Tariot PN, Ryan JM, Porsteinsson AP, Loy R, Schneider LS. Pharmacologic therapy for behavioral symptoms of Alzheimer's disease. Clin Geriatr Med. 2001;17(2):359–76.
77. Ansari MA, Abdul HM, Joshi G, Opii WO, Butterfield DA. Protective effect of quercetin in primary neurons against Abeta(1–42): relevance to Alzheimer's disease. J Nutr Biochem. 2009;20(4):269–75.
78. Pallàs M, Verdaguer E, Jordà EG, Jiménez A, Canudas AM, Camins A. Flavopiridol: an antitumor drug with potential application in the treatment of neurodegenerative diseases. Med Hypotheses. 2005;64(1):120–3.
79. Grammas P, Martinez J, Sanchez A, Yin X, Riley J, Gay D, Desobry K, Tripathy D, Luo J, Evola M, Young A. A new paradigm for the treatment of Alzheimer's disease: targeting vascular activation. J Alzheimers Dis. 2014;40(3):619–30.
80. Katz IR, Jeste DV, Mintzer JE, Clyde C, Napolitano J, Brecher M. Comparison of risperidone and placebo for psychosis and behavioral disturbances associated with dementia: a randomized, double-blind trial. Risperidone Stud Group J Clin Psychiatry. 1999;60(2):107–15.
81. Valles SL, Dolz-Gaiton P, Gambini J, Borras C, Lloret A, Pallardo FV, Viña J. Estradiol or genistein prevent Alzheimer's disease-associated inflammation correlating with an increase PPAR gamma expression in cultured astrocytes. Brain Res. 2010;1312:138–44.
82. Dhawan G, Combs CK. Inhibition of Src kinase activity attenuates amyloid associated microgliosis in a murine model of Alzheimer's disease. J Neuroinflammation. 2012;2(9):117.
83. Parkes JD, Tarsy D, Marsden CD, Bovill KT, Phipps JA, Rose P, Asselman P. Amphetamines in the treatment of Parkinson's disease. J Neurol Neurosurg Psychiatry. 1975;38(3):232–7.
84. Bersani G, Grispini A, Marini S, Pasini A, Valducci M, Ciani N. 5-HT2 antagonist ritanserin in neuroleptic-induced parkinsonism: a double-blind comparison with orphenadrine and placebo. Clin Neuropharmacol. 1990; 13(6):500–6.
85. Tariq M, Khan HA, Al Moutaery K, Al DS. Protective effect of quinacrine on striatal dopamine levels in 6-OHDA and MPTP models of Parkinsonism in rodents. Brain Res Bull. 2001;54(1):77–82.
86. Weintraub D, Mavandadi S, Mamikonyan E, Siderowf AD, Duda JE, Hurtig HI, Colcher A, Horn SS, Nazem S, Ten Have TR, Stern MB. Atomoxetine for depression and other neuropsychiatric symptoms in Parkinson disease. Neurology. 2010;75(5):448–55.
87. Hashimoto R, Hough C, Nakazawa T, Yamamoto T, Chuang DM. Lithium protection against glutamate excitotoxicity in rat cerebral cortical neurons: involvement of NMDA receptor inhibition possibly by decreasing NR2B tyrosine phosphorylation. J Neurochem. 2002;80(4):589–97.
88. Perez-Gracia JL, Prior C, Guillén-Grima F, et al. Identification of TNF-alpha and MMP-9 as potential baseline predictive serum markers of sunitinib activity in patients with renal cell carcinoma using a human cytokine array. Br J Cancer. 2009;101(11):1876–83. doi:10.1038/sj.bjc.6605409.
89. Wrasidlo W, Crews LA, Tsigelny IF, Stocking E, Kouznetsova VL, Price D, Paulino A, Gonzales T, Overk CR, Patrick C, Rockenstein E, Masliah E. Neuroprotective effects of the anti-cancer drug sunitinib in models of HIV neurotoxicity suggests potential for the treatment of neurodegenerative disorders. Br J Pharmacol. 2014;171(24):5757–73. doi:10.1111/bph.12875.
90. Sanchez A, Tripathy D, Yin X, Luo J, Martinez JM, and Grammas P. Sunitinib enhances neuronal survival in vitro via NF-κB-mediated signaling and expression of cyclooxygenase-2 and inducible nitric oxide synthase. J Neuroinflammation 2013. 10(93). doi: 10.1186/1742-2094-10-93.
91. Margolis DJ et al. Potential association between the oral tetracycline class of antimicrobials used to treat acne and inflammatory bowel disease. Am J Gastroenterol. 2010;105(12):2610–6.
92. On SC, Zeichner J. Isotretinoin updates. Dermatol Ther. 2013;26(5):377–89.
93. US Natl Inst Health; DailyMed. Current Medication Information for sansert (methysergide maleate) tablet, coated (February 2006). Available from, as of March 6, 2012: http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=3fae28ee-700e-4d4f-a040-02ef01a2aeb4
94. Felder JB, Korelitz BI, Rajapakse R, Schwarz S, Horatagis AP, Gleim G. Effects of nonsteroidal antiinflammatory drugs on inflammatory bowel disease: a case–control study. Am J Gastroenterol. 2000;95(8):1949–54.
95. Takeuchi K, Smale S, Premchand P, Maiden L, Sherwood R, Thjodleifsson B, Bjornsson E, Bjarnason I. Prevalence and mechanism of nonsteroidal anti-inflammatory drug-induced clinical relapse in patients with inflammatory bowel disease. Clin Gastroenterol Hepatol. 2006;4(2):196–202.
96. Ballard C, Margallo-Lana M, Juszczak E, Douglas S, Swann A, Thomas A, O'Brien J, Everratt A, Sadler S, Maddison C, Lee L, Bannister C, Elvish R, Jacoby R. Quetiapine and rivastigmine and cognitive decline in Alzheimer's disease: randomised double blind placebo controlled trial. BMJ. 2005; 330(7496):874.
97. Callaghan RC, Cunningham JK, Sykes J, Kish SJ. Increased risk of Parkinson's disease in individuals hospitalized with conditions related to the use of methamphetamine or other amphetamine-type drugs. Drug Alcohol Depend. 2012;120(1–3):35–40.
98. Brogden RN, Heel RC, Speight TM, Avery GS. Sulindac: a review of its pharmacological properties and therapeutic efficacy in rheumatic diseases. Drugs. 1978;16(2):97–114.
99. Taylor P, Genovese M, Keystone E, Schlichting D, Beattie S, Macias W. Baricitinib, an oral janus kinase inhibitor, in the treatment of rheumatoid arthritis: safety and efficacy in an open-label, long-term extension study1. Ann Rheum Dis. 2014;73:A31. doi:10.1136/annrheumdis-2013-205124.71.
100. van der Heijden JW, Assaraf YG, Gerards AH, Oerlemans R, Lems WF, Scheper RJ, Dijkmans BA, Jansen G. Methotrexate analogues display enhanced inhibition of TNF-α production in whole blood from RA patients. Scand J Rheumatol. 2014;43(1):9–16. doi:10.3109/03009742.2013.797490. Epub 30 Aug 2013.
101. Page TH, Brown A, Timms EM, Foxwell BM, Ray KP. Inhibitors of p38 suppress cytokine production in rheumatoid arthritis synovial membranes: does variable inhibition of interleukin-6 production limit effectiveness in vivo? Arthritis Rheum. 2010;62(11):3221–31. doi:10.1002/art.27631.
102. Weinblatt ME, Keystone EC, Furst DE, Moreland LW, Weisman MH, Birbara CA, Teoh LA, Fischkoff SA, Chartash EK. Adalimumab, a fully human anti-tumor necrosis factor alpha monoclonal antibody, for the treatment of rheumatoid arthritis in patients taking concomitant methotrexate: the ARMADA trial. Arthritis Rheum. 2003;48(1):35–45.
103. Moreland LW, Schiff MH, Baumgartner SW, Tindall EA, Fleischmann RM, Bulpitt KJ, Weaver AL, Keystone EC, Furst DE, Mease PJ, Ruderman EM, Horwitz DA,Arkfeld DG, Garrison L, Burge DJ, Blosch CM, Lange ML, McDonnell ND, Weinblatt ME. Etanercept therapy in rheumatoid arthritis. A randomized, controlled trial. Ann Intern Med. 1999;130(6):478–86.
104. O'Dell JR, Haire CE, Palmer W, Drymalski W, Wees S, Blakely K, Churchill M, Eckhoff PJ, Weaver A, Doud D, Erikson N, Dietz F, Olson R, Maloley P, Klassen LW, Moore GF. Treatment of early rheumatoid arthritis with minocycline or placebo: results of a randomized, double-blind, placebo-controlled trial. Arthritis Rheum. 1999;40(5):842–8.
105. Klotz U, Maier K, Fischer C, Heinkel K. Therapeutic efficacy of sulfasalazine and its metabolites in patients with ulcerative colitis and Crohn's disease. N Engl J Med. 1980;303(26):1499–502.
106. Baumgart DC, Sandborn WJ. Inflammatory bowel disease: clinical aspects and established and evolving therapies. Lancet. 2007;369(9573):1641–57.
107. Keshavarzian A, Mutlu E, Guzman JP, Forsyth C, Banan A. Phosphodiesterase 4 inhibitors and inflammatory bowel disease: emerging therapies in inflammatory bowel disease. Expert Opin Investig Drugs. 2007;16(9): 1489–506.
108. Mabley JG, Pacher P, Liaudet L, Soriano FG, Haskó G, Marton A, Szabo C, Salzman AL. Inosine reduces inflammation and improves survival in a murine model of colitis. Am J Physiol Gastrointest Liver Physiol. 2003;284(1): G138–44. Epub 28 Aug 2002.

109. El Miedany Y, Youssef S, Ahmed I, El Gaafary M. The gastrointestinal safety and effect on disease activity of etoricoxib, a selective cox-2 inhibitor in inflammatory bowel diseases. Am J Gastroenterol. 2006;101(2):311–7.
110. Carter MJ, Lobo AJ, Travis SP, IBD Section, British Society of Gastroenterology. Guidelines for the management of inflammatory bowel disease in adults. Gut. 2004;53 Suppl 5:V1–V16.
111. Nagar J, Ranade S, Kamath V, Singh S, Karunanithi P, Subramani S, Venkatesh K, Srivastava R, Dudhgaonkar S, Vikramadithyan RK. Therapeutic potential of chloroquine in a murine model of inflammatory bowel disease. Int Immunopharmacol. 2014;21(2):328–35. doi:10.1016/j.intimp.2014.05.005. Epub 21 May 2014.

**MSD-MAP: A Network-based Systems Biology Platform for Predicting Disease-Metabolite Links**

Henri Wathieu[1], Naiem T Issa[1], Manisha Mohandoss[2], Stephen W Byers[1,2], and Sivanesan Dakshanamurthy[1,2,*]

[1]Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington DC, 20057 USA.

[2]Department of Biochemistry & Molecular Biology, Georgetown University, Washington DC, 20057 USA.

*****Corresponding author.

All correspondence should be addressed to:

Dr. Sivanesan Dakshanamurthy, Ph.D.

e-mail: sd233@georgetown.edu

Phone: 202-687-2347

**Running Title**

Linking Metabolites and Diseases With a Systems Biology Platform

**Keywords**

Biomarker, Colorectal Cancer, Esophageal Cancer, Gene Expression Analysis, Metabolomics, MSD-MAP, Prostate Cancer, Systems Biology

**Abstract**
Metabolites constitute phenotypic end products of gene expression, and are key players in biological networks. For this reason, the field of metabolomics has been useful in predicting, explaining, and affecting the mechanisms of disease phenotypes. MSD-MAP (Multi Scale Disease-Metabolite Association Platform) is a powerful computational tool for hypothesizing new links between diseases and metabolites, and characterizing the functional basis of those links in a systems biology context. Upon integrating both predicted and known metabolite-protein associations, MSD-MAP takes a two-pronged approach to associating metabolites to a disease, relying on network-based characterization of disease perturbation at multiple levels of biological activity as well as statistical matching of metabolite- and disease-associated biological profiles. MSD-MAP successfully recapitulated cross-disease links of cancer-associated metabolites, and predicted key metabolites associated with colorectal, esophageal, and prostate cancers after the integration of patient-based gene expression analysis. For example, the catecholamine dopamine was correctly predicted to be strongly associated with colorectal cancer based on statistical coincidence with its disease perturbation network.

**Introduction**
Metabolic dysregulation is a hallmark of many diseases, especially cancers **[1**]. Alteration of metabolic reaction pathway end-products or intermediates beyond homeostatic levels may be indicative or pathognomonic of disease. These metabolites are contextually viewed as potential biomarkers for diagnosis or even therapeutic targets (i.e. pharmacologically targeting enzymes responsible for the direct increase or decrease of a particular metabolite ascribed to a disease process).

Metabolomic profiling of diseases is a relatively well-established approach, but has been mostly restricted to a framework of canonical metabolic pathways. The small size and ubiquitous nature of metabolites, however, renders them the potential to bind non-canonical targets that fall outside of tradition metabolic regulatory network and elicit pathogenic effects. Therefore, we conduct analyses using both experimentally-verified profiles of metabolites from the Human Metabolome Database (HMDB) **[2]**, as well as profiles derived from our own proteochemometric method RepurposeVS **[3]**.

MSD-MAP (Multi Scale Disease-Metabolite Association Platform) uses predicted and experimentally verified metabolite-protein interactions to place metabolites in a multi scale systems biology disease network in two ways: (1) direct association by way of protein interactions to pathway, function and disease associations, in a procedure modeled after Yildrim et al **[4]**, and (2) statistical calculation of physiological coincidence to multiple cancers, by way of multi scale mapping of biological components from patient-derived gene expression data. The latter method is modeled after our drug repurposing platform, DrugGenEx-NET **[5]**.

Studies attempting to identify metabolic profiles for a given disease typically characterize the metabolomic composition of various biofluids, and in some cases the diseased tissue itself **[6]**. Because MSD-MAP is applied to transcriptomic profiles of primary tumors in cancer, resulting biological networks are thought to be highly reflective of disease perturbation. The platform is not limited to predicting metabolites that may prove useful as disease biomarkers or that elucidate disease perturbation. In addition to incorporating metabolite-disease associations, the platform necessarily produces associations that have the potential to occur but, for various reasons (such as low concentration of metabolite, not achieving binding capacity due to short half-life, or cellular compartmentalization/localization of metabolites and protein targets), may not actually arise in conventional metabolic profiling efforts. MSD-MAP can thus predict metabolites that have potential for therapeutic intervention against a given disease or to serve as a biomarker of disease progression, based on the intersection between disease perturbation and metabolite action in a way that is analogous to pharmacological action. Therapeutic metabolites may encapsulate any metabolites cataloged by HMDB and incorporated into our database.

MSD-MAP was applied to cancer-associated metabolites to provide a novel network analysis of predicted cancer metabolite-target associations and find other diseases that may share those associations. To our knowledge, the prediction of metabolite-target signatures using only protein crystal structures and metabolite chemical structures has not been performed before. Having observed and reported here high fidelity in the ability of RepurposeVS to reproduce known protein interaction profiles of several metabolites, we believe that incorporating this tool allows us to exploit a broader pathophysiological space, which is a key advantage of this approach. MSD-MAP is an entirely computer-based platform, thus allowing it to become a feasible tool for initial exploratory studies. We demonstrate the predictive power of MSD-MAP in establishing metabolite-disease associations using previously established links from the

literature. These validations support MSD-MAP as a tool to identify (1) potential biomarkers pointing to a particular diagnosis or therapy, (2) metabolites that may themselves be useful as therapeutic agents, (3) metabolites that may contribute to disease perturbation and (4) metabolic characterization allowing for the elucidation of biochemical disease mechanisms.

**Methods**

*Collection of Cancer Metabolites and Protein Target Curation to Predict Metabolite-Protein Interactions*

Metabolites associated with cancers were obtained from HMDB as spatial data files (SDFs) containing atom and bond connectivity information **[2]**. Energy-minimized 3D structures were then prepared from those SDF files using Schrodinger's LigPrep[17] algorithm at neutral pH of 7.0. Human protein crystal structures were obtained from RCSB (www.rcsb.org). Only X-ray structures with <2.5 angstrom resolution and a reference co-crystallized ligand were chosen. Protein structures were further processed using an in-house collection of BASH and PERL scripts that directly manipulated the crystal structure PDB files to remove non-biologically relevant chains (i.e. those that do not interact with the ligand), metal ions, and all heteroatoms (i.e. non-cofactors, solvent molecules), as well as to add hydrogens optimally using the Schrodinger ProteinPrep application. After processing, the dataset included 56 metabolites and 2,335 protein target crystal structures.

*Predicting Metabolite-Protein Interactions*

The RepurposeVS approach **[3]**, which is an enhancement of the original TMFS approach developed by our group **[7]**, was used to predict metabolite-protein signatures. In short, RepurposeVS is a proteochemometric method that integrates docking, shape, and ligand physicochemical descriptors for generating reliable binding signature predictions. The comprehensive 'Z-score' represents the quantitative likelihood of binding for ranking purposes (Eq. 1):

$$Z = \omega_k Y(\sigma_p, \sigma_l) + \sum_{i=1}^{1} [\omega_i f(\sigma_p, \sigma_l) + \omega_{i+1} f(\sigma_p, \sigma_l)] + \sum_{n=1}^{j} X_n(\sigma_c, \sigma_l) + CS(OLIC) \quad (1)$$

In determining which metabolite-protein signatures were considered for the subsequent network analyses, a Z-score threshold of 13.5 out of a maximum score of 15.0 (75%) was set. Using a threshold of 75% selected for true-positive metabolites with Z-scores near those of drug-target signatures validated in previous studies **[3,7]**. Furthermore, as metabolites differ substantially from drugs and drug-like molecules in physicochemical properties and may occupy binding pockets differently, setting the Z-score threshold at 75% such also allows for a relaxation of the criteria to retain metabolite-target interactions.

*Construction of Metabolite-Protein-Disease/Pathway/Function Networks*

Protein targets were cross-referenced using the unique PDB entry with UniProt **[8]**. A list of disease genes associated with the protein were obtained from each UniProt entry and mapped to the Online Mendelian Inheritance in Man (OMIM) Morbidity Map **[9]** disease gene-disease associations, a procedure modeled after Yildirim *et al.* **[4]**. Metabolites are connected to a disease by mapping their target genes to their associated disease. Thus, a metabolite is connected to a disease if its predicted target has disease genes associated with the disorder. In the DT-disease network, all disorders associated with a predicted protein target will be associated with the metabolite. **Figure 1a** depicts this approach.

Disease-associated targets were also annotated with pathway and cellular function information using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) Functional Annotation Tool (FAT) **[10]**. FAT and DAVID Functional Clustering were also used to annotate functions from the Gene Ontology (GO) **[11]**. Results were manually curated and functions with a false discovery rate (FDR) <0.25 were selected from each cluster. Metabolites are connected to a pathway or function if they have a predicted target associated with them.

*Curation of Experimentally Determined Metabolite-Protein Interactions*

All known metabolite-target interactions were derived from HMDB **[2]**. This resulted in a database of containing 22,126 metabolites annotated with interacting proteins whose genes were measured in the gene expression data utilized in this study, out of 41,993 total metabolites currently characterized by HMDB. Of the 22,126 metabolites annotated, 20,595 were labeled as endogenous, and 1,531 exogenous.

*Annotating Biological Effects of Metabolites*

Metabolites were associated with biological pathways and cellular functions through their experimentally verified (known) and RepurposeVS-predicted direct protein interactions if such an annotation existed. Experimentally known metabolite-protein target associations were obtained from annotations in the HMDB. Two curated databases, those of known and predicted metabolite-protein interactions, were separately annotated. Annotations were retrieved via DAVID Functional Annotation Tool **[2]** and ConsensusPathDB **[12]**. These collectively allowed for a comprehensive and up-to-date annotation dataset using publicly available tools. Pathways were derived from BioCarta **[13]**, Edinburgh Human Metabolic Network **[14]**, HumanCyc **[15]**, INOH **[16]**, KEGG **[17]**, PharmGKB **[18]**, NCI Pathway Interaction Database **[19]**, Reactome **[20]**, SMPDB **[21]**, and WikiPathways **[22]**. Functions were obtained from the Gene Ontology **[11]**.

*Obtaining Cancer-associated Gene Expression Using Differential Analysis*

Human cancer RNA-seq gene expression data from the The Cancer Genome Atlas (TCGA) **[23]** were obtained from the UCSC Cancer Genome Browser **[24]**. These cancers include Colorectal Cancer (CRC), Esophageal Cancer (EC), and Prostate Cancer (PC). Cancer-associated genes were obtained using differential gene expression analysis (GEA) performed in R **[25]**, where all cancer-derived samples were compared to all normal samples of the same tissue. Genes up-regulated and down-regulated in cancer with t-test Q-value < 0.05 (corresponding to Benjamini-Hochberg corrected P-values), and fold change > 1.0 and < -1.0, respectively, were considered differentially associated. Up-regulated and down-regulated gene lists were subsequently considered separately.

*Annotating Biological Effects of Cancer-Associated Genes*

Protein-protein interactions (PPIs) for differentially expressed genes were obtained from the STRING database using a high confidence score cutoff of >0.7 **[26]**. Similar to the metabolite-protein interaction database, PPIs were filtered into a sub-dataset including only protein partners having probes in the TCGA RNAseq dataset in question. DAVID Functional Annotation and ConsensusPathDB were also used to perform overrepresentation analyses on up- or down-regulated disease-associated gene lists using P-value < 0.05 for discovery purposes.

*Matching Disease Perturbation and Metabolite Action on Multiple Scales*

MSD-MAP prioritizes metabolites associated with a given disease based on statistically significant coincidence across the biological effect tiers of direct protein interactors, PPIs (indirect protein targets), pathways, and functions. **Figure 1b** illustrates this methodology. The hypergeometric test in R was used to statistically associate metabolites in this fashion **(Figure 2)**. With respect to each biological level, metabolites associated with P-value < 0.05 had P-values log-transformed and normalized in relation to the value of the most significantly-associated metabolite, creating transformed values in the 0-1 unit range. All non-significant P-values were automatically assigned a value of 0.

For each metabolite i and for a given disease corresponding to either up-regulated or down-regulated biological perturbation, normalized values for each biological effect tier were integrated to calculate an association Z-score used for ranking (Eq. 2):

$$Z_i = 2(Protein) + PPI + Pathway + Function + PPI$$

where Protein, Pathway, Function, and PPI correspond to the normalized significance score of that metabolite to the respective biological tier. Each of these normalized significance scores quantify the statistical coincidence of associated biological activity and metabolite with respect to the biological tier in question. The Protein variable was given the greatest weight to emphasize the metabolite-protein interaction as the biological tier of primary utility.

In analyses derived from known metabolite-protein associations, metabolites were ranked in descending order by Z-score. High Z-scores indicate a metabolite's statistical potential to be simultaneously associated with the greatest amount of direct proteins, pathways, functions and PPIs associated with the given disease condition. Because novel protein interactions were predicted for only 56 metabolites, these metabolites were not ranked, but instead evaluated relative to a given disease based on stand-alone Z-scores and significance at each level of biological action.

**Results and Discussion**

*Application of MSD-MAP to Cancer-Associated Metabolites: Predicting Protein Associations and Cross-Disease Links*

MSD-MAP was used to predict protein targets for cancer-associated metabolites **(Table S1)**, of which 15 metabolite-target signatures were validated using HMDB MetaboCards **(Table 1)**. The low number of validations relative to the number of predictions can be attributed to the vast experimentally unexplored space of metabolite association with non-metabolism-associated targets. Metabolites were predicted to associate with a diverse set of diseases spanning multiple categories. The metabolite orotidylic acid, for instance, forms a distinct hub and is associated with many

diseases through their predicted targets **(Figure 3)**. This implies that such metabolites are not uniquely associated with cancers, but a variety of other diseases as well. Orotidylic acid (OMP) is a pyrimidine nucleotide intermediate in uridine monophosphate (UMP) biosynthesis. As a nucleotide structural analogue to adenosine monophosphate (AMP) and guanosine monophosphate (GMP), OMP has the potential to bind to similar targets. Indeed, MSD-MAP predicted that OMP would bind many G-protein, kinase and ATPase targets **(Table S1)**.

Given the diverse target set, OMP was also predicted to be involved in a variety of diseases. Buildup of OMP and orotic acid due to enzymatic deficiency, amino acid imbalance or consumption can lead to orotic aciduria and clinical symptoms such as megaloblastic anemia and mental retardation. In particular, a deficiency or mutation in the enzyme UMP Synthase, which catalyzes the conversion of OMP to UMP, has been associated with orotic aciduria **[53]**. This result is reflected in the predicted association between OMP and orotic aciduria by MSD-MAP, as shown in **Figure 3**. Interestingly, OMP was also predicted to be involved in Charcot-Marie-Tooth (CMT) disease and Arts Syndrome. CMT disease is a hereditary motor and sensory neuropathy with multiple clinical categories based on mutation types, and Arts syndrome is a rare and severe neurological and immune system disorder. One associated mutation in both disorders is in the PRPS1 gene causing loss of function of the enzyme phosphoribosylpyrophosphate synthetase-1 (PRPS-1) **[27]**. Recently, orotic aciduria has been described for the first time in patients with PRPS1 mutations **[27]**. Elevated levels of orotic acid could potentially allow it bind many unidentified targets across different cell types to give these poly-symptomatic disorders. This data confirms the predicted associations of OMP with CMT and Arts syndrome by MSD-MAP. Another example is the association of OMP with spastic paraplegia. Arginase deficiency, which increases orotic acid excretion **[28]**, leads to spastic paraplegia in early childhood **[29]**. **Figure 3** reflects these validated OMP-related diseases, all of which have been shown to correspond with increased amounts of orotic acid. MSD-MAP was thus able to predict the phenotypic effects of orotodylic acid and can potentially be used to establish other known and unknown metabolite-target-disease associations.

*Application of MSD-MAP to Prediction of Disease-associated Metabolites using Differential Gene Expression and Systems Biology Network Analysis*

*1. Colorectal Cancer*

Colorectal cancer (CRC) has diverse environmental and genetic risk factors. Its origins and progression have implicated mechanisms involving chromosomal instability, microsatellite instability, microRNA activity, abnormal DNA methylation, inflammation, and others **[30]**. Various studies have used metabolomic profiling to arrive at important physiological or pathophysiological states of CRCs that could pave the way for improved early detection and diagnosis **[31-33]**. MSD-MAP also identifies many metabolites altered in CRC in addition to their known physiological mechanisms. **Table 2** contains selected metabolites predicted to be associated with CRC, and validated by the literature, as well as the statistical significance of coincidence at each level of biological importance.

*1a. CRC-metabolite associations derived from known metabolite-protein interactions*

In normal metabolism, the amino acid tyrosine is a precursor to dopamine, and in CRC conditions, serum dopamine levels are low and tyrosine levels high **[34,35]**, indicating a deviation from the normal metabolic process. Dopamine has been implicated in protection against experimental carcinogenesis in rats, and its agonists exhibit gastroprotective effects **[36]**. Previous findings also suggest that decreased expression of dopamine receptors, when brought about by polymorphisms in corresponding genes, is linked to depletion of dopamine receptors and is associated with increased risk of sporadic CRC **[37]**. MSD-MAP also found dopamine to be strongly associated with factors reduced in CRC including various dopamine receptors and related pathways and functions (**Figure 4)**. Correspondingly, MSD-MAP predicted tyrosine to be associated with upregulated activity.

Cysteamine, the simplest aminothiol found in the body, is a degradation product of the amino acid cysteine. Cysteamine and cysteine were both found at higher levels in CRC compared to normal patient serum **[34].** Taurine, in turn, can be derived from both cysteine and cysteamine, and was also found to be highly concentrated in CRC patient serum compared to normal **[34]**. MSD-MAP predicted significant association of taurine with upregulated CRC modulation at every level of biological activity. The abnormal metabolism of cysteine in CRC also appears to be biochemically linked to the downregulation of tyrosine metabolism discussed above. For instance, taurine was significantly associated with downregulated biological pathways enriched in CRC gene expression, including tyrosine metabolism (KEGG hsa:00350). More notably, cysteamine has the effect of decreasing norepinephrine levels in colon wall tissues **[38]**. These data concur with the finding of MSD-MAP that norepinephrine is significantly associated with downregulated CRC biology, as was the case with dopamine and other catecholamines. Tetrahydrobiopterin, a

cofactor central to synthesis of catecholamines, was also found to be significantly associated at multiple levels with components upregulated in CRC. Thus, MSD-MAP was able to recapitulate the important roles of cysteine-tyrosine metabolic disruption in the pathophysiology of CRC, and supports the status of related metabolites as potential CRC biomarkers.

Historically, a key risk factor for development of CRC is obesity as well as a high fat and low fiber diet **[39]**. There is increasing evidence that the elevated fecal concentration of bile acids associated with such a diet may be responsible for this link. Secondary bile acids in particular may constitute endogenous metabolites that promote the development of malignant tumors **[40]**. Deoxycholic acid (DCA), a secondary bile acid, has been repeatedly identified as a promoter of CRC proliferation, invasion, and migration, possibly by way of activating the Wnt/β-Catenin Signaling Pathway and causing DNA damage **[41-43]**. DCA has also been established as a potential CRC biomarker relative to primary bile acids **[44]**. MSD-MAP successfully predicted DCA to be strongly associated with biological factors upregulated in CRC at every level, implicating abnormal expression of various bile acid transporters interacting with DCA at the protein and indirect protein (PPI) levels. Such transporters are central to the precise control of bile acid levels in the body **[45]**.

In our CRC study, ursodeoxycholic acid (UDCA) was significantly associated with upregulated disease activity at the levels of protein, function, and PPIs. This secondary bile acid is structurally related to DCA and has a similar annotated metabolic profile in MSD-MAP. In contrast to DCA, however, UDCA has been proposed as a chemopreventive agent for CRC and regulates corresponding signaling pathways differently **[46-49]**.

Halofuginone, an exogenous anti-coccidial quinoazolinone derivative, inhibits the synthesis of alpha-1 type I collagen (COL1A1), a component of the extracellular matrix the expression of which MSD-MAP found to be highly upregulated in CRC. Our model predicted halofuginone to be significantly associated with upregulated CRC activity at every scale mapped. Notably, at the pathway and function levels, respectively, halofuginone was predicted to act on CRC-perturbed □3 integrin cell surface interactions (NCI PID: integrin3pathway) and regulation of cell adhesion (GO:0030155), by inhibiting gelatinase A (MMP-2). Given the association between higher MMP-2 levels and tumor invasion in CRCs **[50]**, this activity could explain the observed anti-cancer activity of halofuginone in several CRC cell lines **[51]**.

*1b. CRC-metabolite associations derived from predicted metabolite-protein interactions*

In our analysis of 56 cancer-linked metabolites for which we predicted protein interactions and subsequently annotated metabolite action at multiple scales, MSD-MAP predicted that 11-dehydrocorticosterone (11-DHC), a mineral corticosteroid that is an inactive 11-keto derivative of glucocorticoids, is significantly associated with physiology downregulated in CRC at every level of biological perturbation. The expression of 11β-hydroxysteroid dehydrogenase types 1 and 2, enzymes that regulate the activation of glucocorticoids, is modulated in CRCs such that active glucocorticoids were produced from their inactive 11-keto derivatives **[52]**, corresponding to decreased levels of 11-DHC. These data agree with the predicted link of inactive 11-DHC to downregulated components, and previous work that has shown that immunomodulatory glucocorticoids including the active stress hormone cortisol are produced by both CRCs in both primary tumor and cell line samples **[53]**. Interestingly, MSD-MAP predicted novel interactions of 11-DHC with proteins downregulated in CRC, including androgen receptor (AR), progesterone receptor (PGR), estrogen receptor (ESR1), mineralocorticoid receptor (NR3C2), glucocorticoid receptor (NR3C1), and Aldo-Keto Reductase Family 1 Member C2 (AKR1C2). All are implicated in steroid hormone biosynthesis. Together, these findings substantiate the hypothesis that synthesis of glucocorticoids constitutes a mechanism for tumor immune escape in CRCs, and that depletion of 11-DHC may prove to be a useful biomarker for CRCs.

*2. Esophageal Cancer*

Metabolomic profiling of esophageal cancer (EC) has revealed that glycolysis followed by lactic acid fermentation is the preferred method for energy production in metabolic disease perturbation of this disease, **[54,55]**. Central pathways include changes in amino acid metabolism, biosynthesis and degradation, ketone bodies synthesis and degradation, tricarboxylic acid (TCA) cycle, and fatty acid metabolism **[54]**. Select literature-validated EC-metabolite associations and statistical significance at each level of analysis are listed in **Table 3**.

*EC-metabolite associations derived from known metabolite-protein interactions*

Aberrations in choline metabolism have been observed in esophageal cancer and are perhaps central to its energy demands **[54]**. Our study recapitulates the involvement of creatine and phospho-creatine , both key intermediates in energy metabolism, in downregulated EC biological activity by way of their association with the arginine and proline metabolism pathway (KEGG hsa:00330) wherein these metabolites play crucial roles. Additionally, L-glutamine, an amino acid found in lower-than-normal levels in EC mucosae **[54]**, was significantly linked to EC and had various network associations with alanine, aspartate and glutamate metabolism (KEGG hsa:00250), metabolites that are also linked to proteins downregulated in EC.

Hyaluronic acid (HA), the most abundant glycosaminoglycan in the body, has been associated with poor prognosis in some cancers **[56]**, and observed at elevated levels in the serum of patients with esophageal squamous cell carcinoma **[57]** compared to normal patients. In our model, HA is significantly associated with upregulated EC proteins, pathways, functions, and PPIs. Associated proteins at the protein and PPI levels largely involve HA biosynthesis and related metabolic processes, which are themselves implicated in EC-modulated cellular functions essential to malignancy, such as cell adhesion (GO:0007155), regulation of cell growth (GO:0001558), and angiogenesis (GO:0001525). Thus, MSD-MAP recapitulates the high EC serum levels of hyaluronic acid by association with up-regulated components of disease perturbation.

*3. Prostate Cancer*

Prostate cancer (PC) is unusual in that it exhibits little dependence on glycolysis for energy, instead relying heavily on fatty acid metabolism, probably due to the resulting availability of acetyl-coenzyme A and increased activity of the TCA cycle **[58]**. Refer to **Table 4** for selected literature validations of PC-associated metabolites with significance of association and multiple scales of biological activity.

*3a. PC-metabolite associations derived from known metabolite-protein interactions*

MSD-MAP predicted testosterone to be strongly associated with PC at each biological level. Testosterone has previously been associated with prostate cancer in various ways and is generally found in higher levels in the serum of prostate cancer patients **[59,60]**. Central to this association is the interaction with the heavily upregulated Cytochrome P450 2J2 (CYP2J2), which, among other functions, is important for the metabolism of arachidonic acid. This, in addition to the linoleic acid metabolism (KEGG hsa:00591) and icosanoid metabolic process (GO:0006690), implicate testosterone levels to be reflective of PC perturbation involving eicosanoid biosynthesis and fatty acid metabolism **[61]**. Furthermore, significantly reduced levels of arachidonic acid occur in malignant prostatic tissue **[62]**. Remarkably MSD-MAP also associates arachidonic acid (and other fatty acids linoleic acid, palmitic acid, and others) to PC with significance across all mapped levels.

L-glutamic acid levels are increased in cancerous prostatectomy tissues, and our MSD-MAP analysis found GABA synthesis to be central to the predicted network association between L-glutamic acid and PC. In PC, gamma-aminobutyric acid (GABA) has been previously reported to increase cellular proliferation via the ionotropic GABAa receptor (GABAar) and to promote cellular invasiveness via the metabotropic GABAb receptor **[59]**. GABA itself was found to be significantly associated with upregulated PC physiology at each level of analysis. In a related exogenous association, dihydroergotoxine, which binds the GABAar chloride ion-channel, was also strongly associated with components downregulated in PC by action on PC-downregulated pathway GABA receptor activation (Reactome pathway REACT_25199), which can decrease cellular proliferation in PC **[63]**.

Folate, Vitamin B12 and Homocysteine are known to be involved in the aberrant methylation of DNA and of tumor suppressor genes, and have implications in prostate cancer development **[64]**. S-adenosylhomocystiene (SAH), an amino acid derivative that has been detected at higher than normal levels in PC tissue **[65]**, is involved in the methylation process and gives rise to homocysteine. Our study also predicted SAH to be strongly associated with PC-upregulated biological perturbation at the levels of pathways, functions, and PPI, emphasizing the roles of factors such as DNA methylation (Reactome pathway REACT_268237), and histone modifications (WikiPathway WP2369). Consequently, MSD-MAP predicts the usefulness of SAH as a potential biomarker in PC.

*3b. PC-metabolite associations derived from predicted metabolite-protein interactions*

MSD-MAP was able to independently predict that linoleic acid is associated with PC-downregulated biology by way of novel predicted interactions with four down-regulated proteins: cAMP-specific 3',5'-cyclic phosphodiesterase 4A (PDE4A), cGMP-specific 3',5'-cyclic phosphodiesterase (PDE5A), retinoic acid receptor beta (RARB), and retinoic acid receptor gamma (RARG). Loss of retinoic acid receptors has been associated with tumorigenicity in PCs **[66]**,

and these interactions may provide important insights into the pathophysiology of prostate cancer in relation to its dependence on fatty acid metabolism. Androstenedione, a precursor of testosterone, was also predicted by novel interactions to be strongly associated at each level with PC-downregulated activity **[67]** including the proteins NR3C1 and Aldo-keto reductase family 1 member C2 (AKR1C2), as well as closely related pathways such as arachidonic acid metabolism (KEGG hsa:00590), signaling by retinoic acid (Reactome pathway REACT_267785), and synthesis of prostaglandins (PG) and thromboxanes (TX) (Reactome pathway REACT_150149). Relying only on metabolic action networks extrapolated from proteochemometrically predicted metabolite-protein interactions, MSD-MAP was therefore able to link familiar PC-associated metabolites to PC-derived mechanisms.

*4. Associating 2-methoxyestradiol with all cancers using predicted metabolite-protein interactions*

When analyzing differential gene expression comparing tumor and normal tissue, those genes involved in mechanisms basic to cancer, rather than important to a subtype or other categorical comparison, are the most highly emphasized in the results. The implications of this are evident in our application of MSD-MAP to three different cancers, as a few chief metabolites exhibit significant association with all three cancers. For example, using metabolite-protein interactions predicted by proteochemometric means, which were then mapped to a multi scale metabolite action space as previously described, we found that 2-methoxyestradiol (2-ME), an endogenous metabolite of estradiol, is significantly ($P < .05$) associated with down-regulated biological activity for CRC, EC, and PC, at all levels: proteins, pathways, functions, and PPIs. 2-ME has repeatedly been identified as a drug candidate for multiple cancers **[68]**, exhibiting both antiangiogenic and antiproliferative properties **[69,70]**. In the case of three cancers examined in the present study, the predicted activity of 2-ME on the protein cAMP-specific 3',5'-cyclic phosphodiesterase 4D (PDE4D) and on the cAMP signaling pathway (KEGG hsa:04024) is central to the metabolite-disease association. This is corroborated with findings that 2-ME activates phosphodiesterases **[71]**, which is often enough to inhibit growth or activate apoptosis in many cancer cell types, thus offering a possible mechanism of action for this pan-cancer therapeutic indication **[72]**.

## Conclusions

The study of metabolomics has yielded many insights into diverse disease mechanisms and metabolites that are useful as biological markers for diagnosis or indication of specified disease progression. MSD-MAP revealed novel metabolite-target-disease signatures that mechanistically explain some metabolite-disease phenotypic associations, and successfully matched multi scale physiological networks extrapolated from predicted and known proteomic profiles of metabolites to disease networks derived from differential gene expression. Applying MSD-MAP to colorectal, esophageal, and prostate cancers, we demonstrated that the metabolite-disease links postulated by statistical matching of metabolite and disease networks can identify disease biomarkers suggested by conventional metabolic profiling, implicate new mechanisms in our understanding of disease pathogenesis and pathophysiology, and detect metabolites that have therapeutic potential against these diseases. MSD-MAP validates the role of gene expression-based profiling followed by systems biology analysis as a tool for elucidating the mechanisms behind metabolic dysregulation of diverse pathologies. Our results also indicate that previously undescribed profiles of metabolites that lie outside of recognized biochemical pathway membership can correspond to the transcriptomic behavior of these cancers, and may in themselves be central to the biological perturbation of such diseases. MSD-MAP is applied to characterizations of disease stemming from differential gene expression analysis, and these analyses are not restricted to comparing cancer and normal tissues. Therefore, MSD-MAP could be practically applied to the study of metabolites specific to disease subtypes, metabolic markers of disease progression, or metabolites linked to clinical parameters such as menopausal status. In addition, new insights may be gained by restricting our CRC analysis to bacterial metabolites, which are known to influence the etiology of the disease **[73]**.

## Conflict of Interests

Authors report no conflict of interest.

## Acknowledgements

The authors wish to acknowledge DOD grant CA140882 (SD), R01 CA170653 (SD, SB), CCSG grant NIH-P30 CA51008 and Georgetown Lombardi Cancer Center.

**List of Abbreviations**

MSD-MAP: Multi Scale Disease-Metabolite Association Platform
HMDB: Human Metabolome Database
SDF: Spatial Data File
TMFS: Train, match, fit, streamline
OMIM: Online Mendelian Inheritance in Man
DAVID: Database for Annotation, Visualization, and Integrated Discovery
FAT: Functional Annotation Tool
GO: Gene Ontology
FDR: False Discovery Rate
TCGA: The Cancer Genome Atlas
CRC: Colorectal Cancer
EC: Esophageal Cancer
PC: Prostate Cancer
PPI: Protein-protein Interaction
OMP: Orotidylic Acid
UMP: Uridine Monophosphate
AMP: Adenosine Monophosphate
GMP: Guanosine Monophosphate
CMT: Charcot-Marie-Tooth
PRPS-1: Phosphoribosylpyrophosphate Synthetase-1
DCA: Deoxycholic Acid
USCA: Ursodeoxycholic Acid
COL1A1: Alpha-1 Type 2 Collagen
MMP-2: Gelatinase A
11-DHC: 11-dehydrocorticosterone
AR: Androgen Receptor
PGR: Progesterone Receptor
ESR1: Estrogen Receptor
NR3C2: Mineralocoricoid Receptor
NR3C1: Glucocorticoid Receptor
AKR1C2: Aldo-Keto Reductase Family Member C2
TCA: Tricarboxylic Acid
HA: Hyaluronic Acid
CYP2J2: Cytochrome P450 2J2
GABA: Gamma-aminobutyric Acid
GABAar: Gamma-aminobutyric Receptor
SAH: S-adenosylhomocysteine
PDE4A: cAMP-specific 3',5'-cyclic phosphodiesterase 4A
PDE5A: cGMP-specific 3',5'-cyclic phosphodiesterase
RARB: Retinoic Acid Receptor Beta
RARG: Retinoic Acid Receptor Gamma
AKR1C2: Aldo-keto reductase family 1 member C2
TX: Thromboxane
2-ME: 2-Methoxyestradiol
PDE4D: cAMP-specific 3',5'-cyclic phosphodiesterase 4D

**References**


1. Warburg, O. On respiratory impairment in cancer cells. *Science*, **1956**, 124(3215), 269-270.

2. Wishart, D.S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A.C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M.A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncic, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D.D.; Wagner, J.; Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.; Duggan, G.E.; MacInnis, G.D.; Weljie, A.M.; Dowlatabadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie, M.; Sykes, B.D.; Vogel, H.J.; Querengesser, L. HMDB: the Human Metabolome Database. *Nucleic Acids Res.*, **2007**, 35(Database issue), D521-D526.

3. Issa, N.T.; Peters, O.J.; Byers, S.W.; Dakshanamurthy, S. RepurposeVS: A Drug Repurposing-Focused Computational Method for Accurate Drug-Target Signature Predictions. *Comb. Chem. High Throughput Screen*, **2015**, 18(8), 784-794.

4. Yildirim, M.A.; Goh, K.I.; Cusick, M.E.; Barabási, A.L.; Vidal, M. Drug-target network. *Nat. Biotechnol.*, **2007**, 25(10), 1119-1126.

5. Issa, N.T.; Kruger, J.; Wathieu, H.; Raja, R.; Byers, S.W.; Dakshanamurthy, S. DrugGenEx-Net: A Novel Computational Platform for Systems Pharmacology and Gene Expression-Based Drug Repurposing. *BMC Bioinformatics*, **2016**, 17(1), 202. http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1065-y (Accessed May 27, 2016).

6. Claudino, W.M.; Quattrone, A.; Biganzoli, L.; Pestrin, M.; Bertini, I.; Di Leo, A. Metabolomics: available results, current research projects in breast cancer, and future applications. *J. Clin. Oncol.*, **2007**, 25(19), 2840-2846.

7. Dakshanamurthy, S.; Issa, N.T.; Assefnia, S.; Seshasayee, A.; Peters, O.J.; Madhavan, S.; Uren, A.; Brown, M.L.; Byers, S.W. Predicting new indications for approved drugs using a proteochemometric method. *J. Med. Chem.*, **2012**, 55(15), 6832-6848.

8. UniProt Consortium. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **2009**, 37(Database issue), D169–D174.

9. Hamosh, A.; Scott, A.F.; Amberger, J.S.; Bocchini, C.A.; McKusick, V.A. Online Mendelian Inheritance in Man (OMIM); a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **2005**, 33(Database issue), D514-517.

10. Dennis, G. Jr.; Sherman, B.T.; Hosack, D.A.; Yang, J.; Gao, W.; Lane, H.C.; Lempicki, R.A. DAVID: Database for Annotation; Visualization; and Integrated Discovery. *Genome Biol.*, **2003**; 4(5), P3.

11. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; Harris, M.A.; Hill, D.P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J.C.; Richardson, J.E.; Ringwald, M.; Rubin, G.M.; Sherlock, G. Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **2000**, 25(1), 25–29.

12. Kamburov, A.; Wierling, C.; Lehrach, H.; Herwig, R. ConsensusPathDB--a database for integrating human functional interaction networks. *Nucleic Acids Res.*, **2009**, 37(Database issue), D623-D628.

13. Nishimura, D. Biocarta. *Biotech Softw. Int. Rep.*, **2004**, 2(3), 117-120.

14. Ma, H.; Sorokin, A.; Mazein, A.; Selkov, A.; Selkov, E.; Demin, O.; Goryanin, I. The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol. Syst. Biol.*, **2007**, 3, 135.

15. Caspi, R.; Billington, R.; Ferrer, L.; Foerster, H.; Fulcher, C.A.; Keseler, I.M.; Kothari, A.; Krummenacker, M.; Latendresse, M.; Mueller, L.A.; Ong, Q.; Paley, S.; Subhraveti, P.; Weaver, D.S.; Karp, P.D. The MetaCyc Database

of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **2008**, 36(Database issue), D623-D631.

16. Yamamoto, S.; Sakai, N.; Nakamura, H.; Fukagawa, H.; Fukuda, K.; Takagi, T. INOH: ontology-based highly structured database of signal transduction pathways. *Database (Oxford)*, **2011**, 2011, bar052.

17. Kanehisa, M.; Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **2000**, 28(1), 27-30.

18. Hewett, M.; Oliver, D.E.; Rubin, D.L.; Easton, K.L.; Stuart, J.M.; Altman, R.B.; Klein, T.E. PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res.*, **2002**, 30(1), 163-165.

19. Schaefer, C.F.; Anthony, K.; Krupa, S.; Buchoff, J.; Day, M.; Hannay, T.; Buetow, K.H. PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **2009**, 37(Database issue), D674-D679.

20. Joshi-Tope, G.; Gillespie, M.; Vastrik, I.; D'Eustachio, P.; Schmidt, E.; de Bono, B.; Jassal, B.; Gopinath, G.R.; Wu, G.R.; Matthews, L.; Lewis, S.; Birney, E.; Stein, L. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, **2005**, 33(Database issue), D428-D432.

21. Frolkis, A.; Knox, C.; Lim, E.; Jewison, T.; Law, V.; Hau, D.D.; Liu, P.; Gautam, B.; Ly, S.; Guo, A.C.; Xia, J.; Liang, Y.; Shrivastava, S.; Wishart, D.S. SMPDB: The Small Molecule Pathway Database. *Nucleic Acids Res.*, **2010**, 38(Database issue), D480-D487.

22. Pico, A.R.; Kelder, T.; van Iersel, M.P.; Hanspers, K.; Conklin, B.R.; Evelo, C. WikiPathways: pathway editing for the people. *PLoS Biol.*, **2008**, 6(7), e184.

23. (TCGA) Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **2008**, 455, 1061–1068.

24. Kent, W.J.; Sugnet, C.W.; Furey, T.S.; Roskin, K.M.; Pringle, T.H.; Zahler, A.M.; Haussler, D. The human genome browser at UCSC. *Genome Res.*, **2002**, 12(6), 996-1006.

25. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, **2014**. Vienna, Austria. http://www.R-project.org/.

26. Szklarczyk, D.; Franceschini, A.; Kuhn, M.; Simonovic, M.; Roth, A.; Minguez, P.; Doerks, T.; Stark, M.; Muller, J.; Bork, P.; Jensen, L.J.; von Mering, C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **2011**, 39(Database issue), D561-D568.

27. de Brouwer A.P.; van Bokhoven H.; Nabuurs S.B.; Arts W.F.; Christodoulou J.; Duley J. PRPS1 mutations: four distinct syndromes and potential treatment. *Am. J. Hum. Genet.*, **2010**, 86(4), 506-18.

28. Brosnan, M.E.; Brosnan, J.T. Orotic acid excretion and arginine metabolism. *J. Nutr.*, **2007**, 137(6 Suppl 2), 1656S-1661S.

29. Tsang, J.P.; Poon, W.L.; Luk, H.M.; Fung, C.W.; Ching, C.K.; Mak, C.M.; Lam, C.W.; Siu, T.S.; Tam, S.; Wong, V.C. Arginase deficiency with new phenotype and a novel mutation: contemporary summary. *Pediatr. Neurol.*, **2012**, 4(4), 263-269.

30. Colussi, D.; Brandi, G.; Bazzoli, F.; Ricciardiello, L. Molecular pathways involved in colorectal cancer: implications for disease behavior and prevention. *Int. J. Mol. Sci.*, **2013**, 14(8), 16365-16385.

31. Hirayama, A.; Kami, K.; Sugimoto, M.; Sugawara, M.; Toki, N.; Onozuka, H.; Kinoshita, T.; Saito, N.; Ochiai, A.; Tomita, M.; Esumi, H.; Soga, T. Quantitative metabolome profiling of colon and stomach cancer microenvironment by capillary electrophoresis time-of-flight mass spectrometry. *Cancer Res.*, **2009**, 69(11), 4918-4925.

32. Wang, H.; Tso, V.K.; Slupsky, C.M.; Fedorak, R.N. Metabolomics and detection of colorectal cancer in humans: a systematic review. *Future Oncol.*, **2010**, 6(9), 1395-1406.

33. Nishiumi, S.; Kobayashi, T.; Ikeda, A.; Yoshie, T.; Kibi, M.; Izumi, Y.; Okuno, T.; Hayashi, N.; Kawano, S.; Takenawa, T.; Azuma, T.; Yoshida, M. A novel serum metabolomics-based diagnostic approach for colorectal cancer. PLoS One, **2012**, 7(7), e40459.

34. Markowitz, S.D.; Bertagnolli, M.M. Molecular Origins of Cancer: Molecular Basis of Colorectal Cancer. *N. Engl. J. Med.*, **2009**, 361(25), 2449-2460.

35. Qiu, Y.; Cai, G.; Su, M.; Chen, T.; Zheng, X.; Xu, Y.; Ni, Y.; Zhao, A.; Xu, L.X.; Cai, S.; Jia, W. Serum metabolite profiling of human colorectal cancer using GC-TOFMS and UPLC-QTOFMS. *J. Proteome Res.*, **2009**, 8(10), 4844-4850.

36. Glavin, G.B.; Szabo, S. Dopamine in gastrointestinal disease. *Dig. Dis. Sci.* **1990**, 35(9), 1153-1161.

37. Gemignani, F.; Landi, S.; Moreno, V.; Gioia-Patricola, L.; Chabrier, A.; Guino, E.; Navarro, M.; Cambray, M.; Capellà, G.; Canzian, F. Polymorphisms of the dopamine receptor gene DRD2 and colorectal cancer risk. *Cancer Epidemiol. Biomarkers Prev.*, **2005**, 14(7), 1633-1638.

38. Tatsuta, M.; Iishi, H.; Baba, M.; Taniguchi, H. Tissue norepinephrine depletion as a mechanism for cysteamine inhibition of colon carcinogenesis induced by azoxymethane in Wistar rats. *Int. J. Cancer*, **1989**, 44(6), 1008-1011.

39. Rogers, A.E.; Zeisel, S.H.; Groopman, J. Diet and carcinogenesis. *Carcinogenesis*, **1993**, 14(11), 2205-2217.

40. Ajouz, H.; Mukherji, D.; Shamseddine, A. Secondary bile acids: an underrecognized cause of colon cancer. *World J. Surg. Oncol.*, **2014**, 12, 164.

41. Pai, R.; Tarnawski, A.S.; Tran, T. Deoxycholic Acid Activates β-Catenin Signaling Pathway and Increases Colon Cell Cancer Growth and Invasiveness. *Mol. Biol. Cell*, **2004**, 15(5), 2156-2163.

42. Ochsenkühn, T.; Bayerdörffer, E.; Meining, A.; Schinkel, M.; Thiede, C.; Nüssler, V.; Sackmann, M.; Hatz, R.; Neubauer, A.; Paumgartner, G. Colonic mucosal proliferation is related to serum deoxycholic acid levels. *Cancer*, **1999**, 85(8), 1664-1669.

43. Milovic, V.; Teller, I.C.; Murphy, G.M.; Caspary, W.F.; Stein, J. Deoxycholic acid stimulates migration in colon cancer cells. *Eur. J. Gastroenterol. Hepatol.*, **2001**, 13(8), 945-949.

44. Kamano, T.; Mikami, Y.; Kurasawa, T.; Tsurumaru, M.; Matsumoto, M.; Kano, M.; Motegi, K. Ratio of primary and secondary bile acids in feces: possible marker for colorectal cancer? *Dis. Colon Rectum*, **1999**, 42(5), 668-672.

45. Da Silva, T.C.; Polli, J.E.; Swaan, P.W. The solute carrier family 10 (SLC10): beyond bile acid transport. *Mol. Aspects Med.*, **2013**, 34(2-3), 252-269.

46. Feldman, R.; Martinez, J.D. Growth suppression by ursodeoxycholic acid involves caveolin-1 enhanced degradation of EGFR. *Biochim. Biophys. Acta.*, **2009**, 1793(8), 1387-1394.

47. Im, E.; Martinez, J.D. Ursodeoxycholic acid (UDCA) can inhibit deoxycholic acid (DCA)-induced apoptosis via modulation of EGFR/Raf-1/ERK signaling in human colon cancer cells. *J. Nutr.*, **2004**, 134(2), 483-486.

48. Alberts, D.S.; Martínez, M.E.; Hess, L.M.; Einspahr, J.G.; Green, S.B.; Bhattacharyya, A.K.; Guillen, J.; Krutzsch, M.; Batta, A.K.; Salen, G.; Fales, L.; Koonce, K.; Parish, D.; Clouser, M.; Roe, D.; Lance, P.; Phoenix and Tucson Gastroenterologist Networks. Phase III trial of ursodeoxycholic acid to prevent colorectal adenoma recurrence. *J. Natl. Cancer. Inst.*, **2005**, 97(11), 846-853.

49. Centuori, S.M.; Martinez, J.D. Differential regulation of EGFR-MAPK signaling by deoxycholic acid (DCA) and ursodeoxycholic acid (UDCA) in colon cancer. *Dig. Dis. Sci.*, **2014**, 59(10), 2367-2380.

50. Said, A.H.; Raufman, J.P.; Xie, G. The Role of Matrix Metalloproteinases in Colorectal Cancer. *Cancers*, **2014**, 6(1), 366-375.

51. Chen, G.Q.; Tang, C.F.; Shi, X.K.; Lin, C.Y.; Fatima, S.; Pan, X.H.; Yang, D.J.; Zhang, G.; Lu, A.P.; Lin, S.H.; Bian, Z.X. Halofuginone inhibits colorectal cancer growth through suppression of Akt/mTORC1 signaling and glucose metabolism. *Oncotarget*, **2015**, 6(27), 24148-24162.

52. Zbánková, S.; Bryndová, J.; Kment, M.; Pácha, J. Expression of 11beta-hydroxysteroid dehydrogenase types 1 and 2 in colorectal cancer. *Cancer Lett.*, **2004**, 210(1), 95-100.

53. Sidler, D.; Renzulli, P.; Schnoz, C.; Berger, B.; Schneider-Jakob, S.; Flück, C.; Inderbitzin, D.; Corazza, N.; Candinas, D.; Brunner, T. Colon cancer cells produce immunoregulatory glucocorticoids. *Oncogene*, **2011**, 30(21), 2411-2419.

54. Wang, L.; Chen, J.; Chen, L.; Deng, P.; Bu, Q.; Xiang, P.; Li, M.; Lu, W.; Xu, Y.; Lin, H.; Wu, T.; Wang, H.; Hu, J.; Shao, X.; Cen, X.; Zhao, Y.L. 1H-NMR based metabonomic profiling of human esophageal cancer tissue. *Molecular Cancer*, **2013**, 12, 25.

55. Zhang, J.; Bowers, J.; Liu, L.; Wei, S.; Gowda, GA.; Hammoud, Z.; Raftery, D. Esophageal Cancer Metabolite Biomarkers Detected by LC-MS and NMR Methods. *PLoS ONE*, **2012**, 7(1), e30181.

56. Wang, C.; Tammi, M.; Guo, H.; Tammi, R. Hyaluronan distribution in the normal epithelium of esophagus, stomach, and colon and their cancers. *Am. J. Pathol.*, **1996**, 148(6), 1861-1869.

57. Aghcheli, K.; Parsian, H.; Qujeq, D.; Talebi, M.; Mosapour, A.; Khalilipour, E.; Islami, F.; Semnani, S.; Malekzadeh, R. Serum hyaluronic acid and laminin as potential tumor markers for upper gastrointestinal cancers. *Eur. J. Intern. Med.*, **2012**, 23(1), 58-64.

58. Liu, Y. Fatty acid oxidation is a dominant bioenergetic pathway in prostate cancer. *Prostate Cancer Prostatic Dis.*, **2006**, 9(3), 230-234.

59. Titus, M.A.; Schell, M.J.; Lih, F.B.; Tomer, K.B.; Mohler, J.L. Testosterone and dihydrotestosterone tissue levels in recurrent prostate cancer. *Clin. Cancer Res.*, **2005**, 11(13), 4653-4657.

60. Hyde, Z.; Flicker, L.; McCaul, K.A.; Almeida, O.P.; Hankey, G.J.; Chubb, S.A.; Yeap, B.B. Associations between testosterone levels and incident prostate, lung, and colorectal cancer. A population-based study. *Cancer Epidemiol. Biomarkers Prev.*, **2012**, 21(8), 1319-1329.

61. Yang, Y.J.; Lee, S.H.; Hong, S.J.; Chung, B.C. Comparison of fatty acid profiles in the serum of patients with prostate cancer and benign prostatic hyperplasia. *Clin. Biochem.*, **1999**, 32(6), 405-409.

62. Chaudry, A.A.; Wahle, K.W.; McClinton, S.; Moffat, L.E. Arachidonic acid metabolism in benign and malignant prostatic tissue in vitro: effects of fatty acids and cyclooxygenase inhibitors. *Int. J. Cancer*, **1994**, 57(2), 176-180.

63. Abdul, M.; Mccray, S.D.; Hoosein, N.M. Expression of gamma-aminobutyric acid receptor (subtype A) in prostate cancer. *Acta. Oncol.*, **2008**, 47(8), 1546-1550.

64. Fang, M.; Chen, D.; Yang, C.S. Dietary polyphenols may affect DNA methylation. *J. Nutr.*, **2007**, 137(1 Suppl), 223S-228S.

65. McDunn, J.E.; Li, Z.; Adam, K.P.; Neri, B.P.; Wolfert, R.L.; Milburn, M.V.; Lotan, Y.; Wheeler, T.M. Metabolomic signatures of aggressive prostate cancer. *Prostate*, **2013**, 73(14), 1547-1560.

66. Nakayama, T.; Watanabe, M.; Yamanaka, M.; Hirokawa, Y.; Suzuki, H.; Ito, H.; Yatani, R.; Shiraishi, T. The role of epigenetic modifications in retinoic acid receptor beta2 gene expression in human prostate cancers. *Lab. Invest.*, **2001**, 81(7), 1049-1057.

67. Stanbrough, M.; Bubley, G.J.; Ross, K.; Golub, TR.; Rubin, M.A.; Penning, T.M.; Febbo, P.G.; Balk, S.P. Increased expression of genes converting adrenal androgens to testosterone in androgen-independent prostate cancer. *Cancer Res.*, **2006**, 66(5), 2815-2825.

68. Lakhani, N.J.; Sarkar, M.A.; Venitz, J.; Figg, W.D. 2-Methoxyestradiol, a promising anticancer agent. *Pharmacotherapy*, **2003**, 23(2), 165-172.

69. Pribluda, V.S.; Gubish, E.R. Jr.; Lavallee, T.M.; Treston, A.; Swartz, G.M.; Green, S.J. 2-Methoxyestradiol: an endogenous antiangiogenic and antiproliferative drug candidate. *Cancer Metastasis Rev.*, **2000**, 19(1-2), 173-179.

70. Klauber, N.; Parangi, S.; Flynn, E.; Hamel, E.; D'Amato, R.J. Inhibition of angiogenesis and breast cancer in mice by the microtubule inhibitors 2-methoxyestradiol and taxol. *Cancer Res.*, **1997**, 57(1), 81-86.

71. Tofovic, P.S.; Rosado, B.M.; Dubey, K.R.; Mi, Z.; Jackson, E.K. Effects of estradiol metabolites on cAMP production and degradation. *Prilozi*, **2009**, 30(1), 5-24.

72. Fajardo, A.M.; Piazza, G.A.; Tinsley, H.N. The Role of Cyclic Nucleotide Signaling Pathways in Cancer: Targets for Prevention and Treatment. *Cancers*, **2014**, 6(1), 436-458.

73. Louis, P.; Hold, G.L.; Flint, H.J. The gut microbiota, bacterial metabolites and colorectal cancer. Nat. Rev. Microbiol., **2014**, 12(10), 661-672.

74. Vermeersch, K.A.; Styczynski, MP. Applications of metabolomics in cancer research. *J. Carcinog.*, **2013**, 12, 9.

75. Monleón, D.; Morales, J.M.; Barrasa, A.; López, J.A.; Vázquez, C.; Celda, B. Metabolite profiling of fecal water extracts from human colorectal cancer. *NMR Biomed.*, **2009** 22(3), 342-348.

76. Chan, E.C.; Koh, P.K.; Mal, M.; Cheah, P.Y.; Eu, K.W.; Backshall, A.; Cavill, R.; Nicholson, J.K.; Keun, H.C. Metabolic profiling of human colorectal cancer using high-resolution magic angle spinning nuclear magnetic resonance (HR-MAS NMR) spectroscopy and gas chromatography mass spectrometry (GC/MS). *J Proteome Res.*, **2009**, 8(1), 352-361.

77. Abbassi-Ghadi, N.; Kumar, S.; Huang, J.; Goldin, R.; Takats, Z.; Hanna, G.B. Metabolomic profiling of oesophago-gastric cancer: a systematic review. *Eur. J. Cancer.*, **2013**, 49(17), 3625-3637.

78. Sreekumar, A.; Poisson, L.M.; Rajendiran, T.M.; Khan, A.P.; Cao, Q.; Yu, J.; Laxman, B.; Mehra, R.; Lonigro, R.J.; Li, Y.; Nyati, M.K.; Ahsan, A.; Kalyana-Sundaram, S.; Han, B.; Cao, X.; Byun, J.; Omenn, G.S.; Ghosh, D.; Pennathur, S.; Alexander, D.C.; Berger, A.; Shuster, J.R.; Wei, J.T.; Varambally, S.; Beecher, C.; Chinnaiyan, A.M. Metabolomic Profiles Delineate Potential Role for Sarcosine in Prostate Cancer Progression. *Nature*, **2009**, 457(7231), 910-914.

79. Yang, B.; Sun, H.; Lin, W.; Hou, W.; Li, H.; Zhang, L.; Li, F.; Gu, Y.; Song, Y.; Li, Q.; Zhang, F. Evaluation of global DNA hypomethylation in human prostate cancer and prostatic intraepithelial neoplasm tissues by immunohistochemistry. *Urol Oncol.*, **2013**, 31(5), 628-634.

80. Nyman, D.W.; Suzanne Stratton, M.; Kopplin, M.J.; Dalkin, B.L.; Nagle, R.B.; Jay Gandolfi, A. Selenium and selenomethionine levels in prostate cancer patients. *Cancer Detect Prev.*, **2004**, 28(1), 8-16.

**Figure Legends**

**Figure 1. Schematic of two-pronged method for disease-metabolite association in MSD-MAP.** In part (A), interacting proteins for a given metabolite, as predicted by RepurposeVS, are used to annotate pathways, functions, and diseases using the databases indicated. In part (B), patient gene expression (RNAseq) profiles are used to perform differential gene expression, comparing primary tumors of a given cancer to corresponding normal tissue samples. The final Protein List resulting from this analysis contains either down- or up-regulated genes. Overrepresentation analysis is performed to enrich the Protein List and obtain associated pathways and functions. Proteins interacting with differential proteins by protein-protein interactions are associated with CRC indirectly. Predicted or known interacting proteins for a given metabolite are matched to these biological factors and subsequently tested for statistical significance of coincidence by hypergeometric test. Colored lines represent "hits," while circular nodes represent equivalency.

**Figure 2. Hypergeometric test for significance of disease-metabolite association at each level of biological activity.**

**Figure 3. Predicted metabolite-disease network.** Metabolites (orange nodes) were connected with diseases (blue nodes) through their predicted protein target. Inset shows higher resolution of hub comprised of orotidylic acid and UDP-N-acetylmuraminate. A buildup of orotidylic acid is linked to orotic aciduria [46]. Orotic aciduria, in turn, coincides with PRPS-1 loss-of-function mutations that are characteristic of Charcot-Marie-Tooth Disease and Arts Syndrome [27], and with Arginase deficiency leading to Spastic Paraplegia in early childhood [28, 29].

**Figure 4. Multi scale network association between Dopamine and colorectal cancer biological activity.** Known protein interactions for the neurotransmitter dopamine are considered a "hit" (green lines) when they are significantly downwardly expressed in colorectal cancer (CRC) primary tumors compared to normal colon. These are termed Direct Proteins (leftmost column). Indirect Proteins also interact with dopamine (blue lines), but were not significantly downregulated in CRC. They are linked via protein-protein interactions (dotted lines) with CRC-regulated proteins, including the Direct Proteins and others which do not interact with dopamine. Coinciding between CRC-downregulated pathways and dopamine-linked (red lines) pathways are 138 individual pathways (represented by the red vertical rectangles), seven of which are highlighted to reveal connectivity. 220 CRC-downregulated cellular

functions are also associated with dopamine (orange vertical rectangles connected via orange lines), of which seven are highlighted in the rightmost column.

**Figure 1**

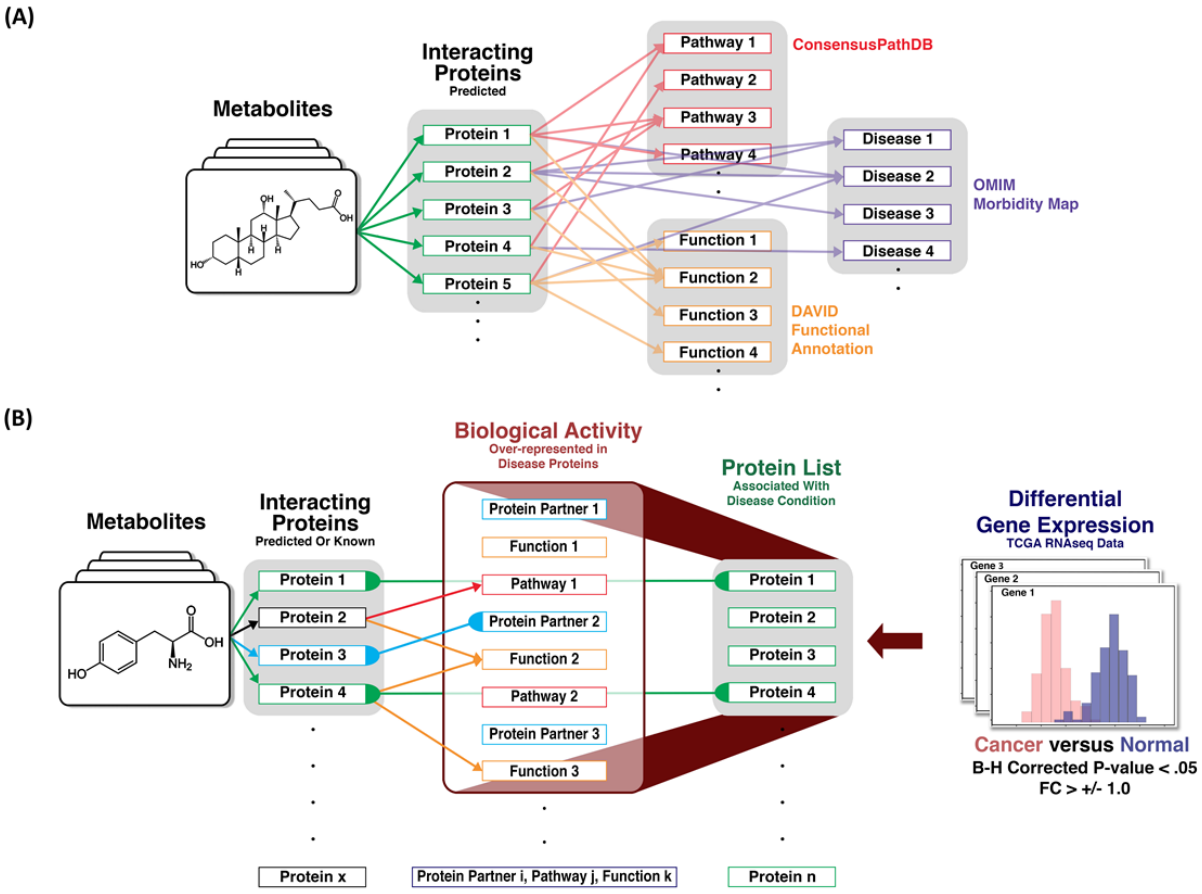

(A)
Metabolites
Interacting Proteins
Predicted
Protein 1
Protein 2
Protein 3
Protein 4
Protein 5
Pathway 1
Pathway 2
Pathway 3
Pathway 4
ConsensusPathDB
Function 1
Function 2
Function 3
Function 4
DAVID Functional Annotation
Disease 1
Disease 2
Disease 3
Disease 4
OMIM Morbidity Map
(B)
Metabolites
Interacting Proteins
Predicted Or Known
Protein 1
Protein 2
Protein 3
Protein 4
Protein x
Biological Activity
Over-represented in Disease Proteins
Protein Partner 1
Function 1
Pathway 1
Protein Partner 2
Function 2
Pathway 2
Protein Partner 3
Function 3
Protein Partner i, Pathway j, Function k
Protein List
Associated With Disease Condition
Protein 1
Protein 2
Protein 3
Protein 4
Protein n
Differential Gene Expression
TCGA RNAseq Data
Gene 3
Gene 2
Gene 1
Cancer versus Normal
B-H Corrected P-value < .05
FC > +/- 1.0

**Figure 2**

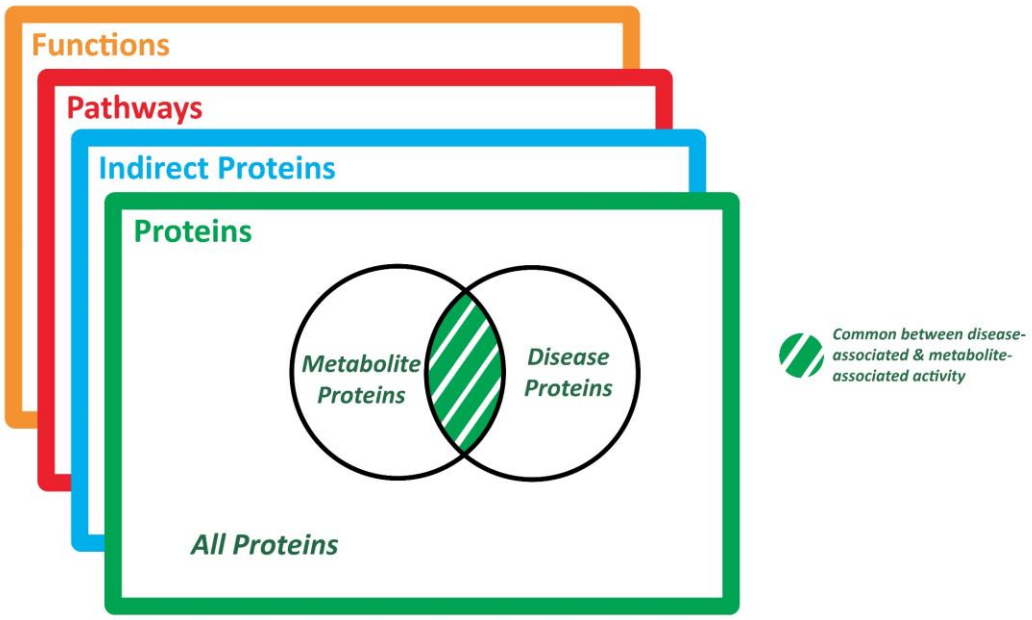

Functions
Pathways
Indirect Proteins
Proteins
Metabolite Proteins
Disease Proteins
All Proteins
Common between disease-associated & metabolite-associated activity

**Figure 3**

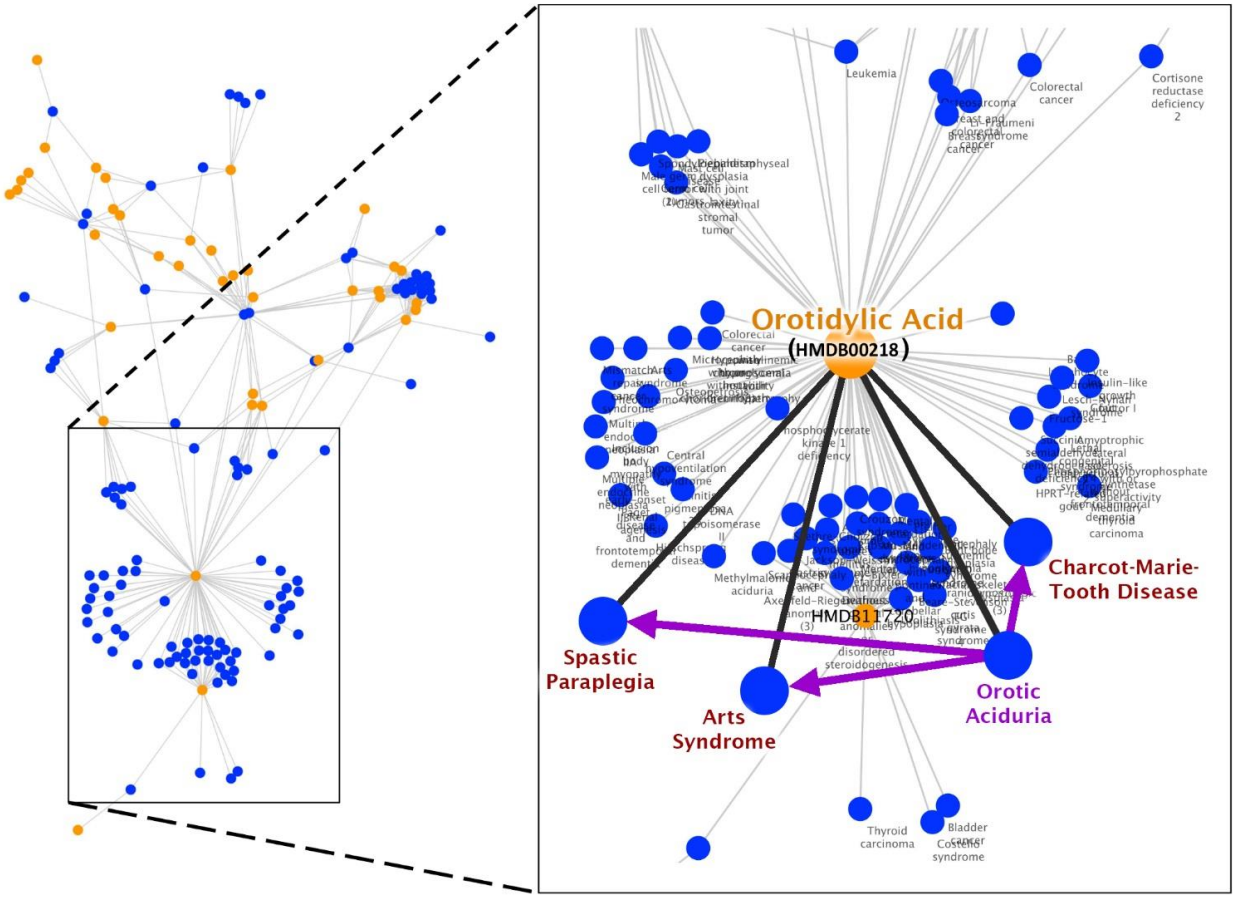
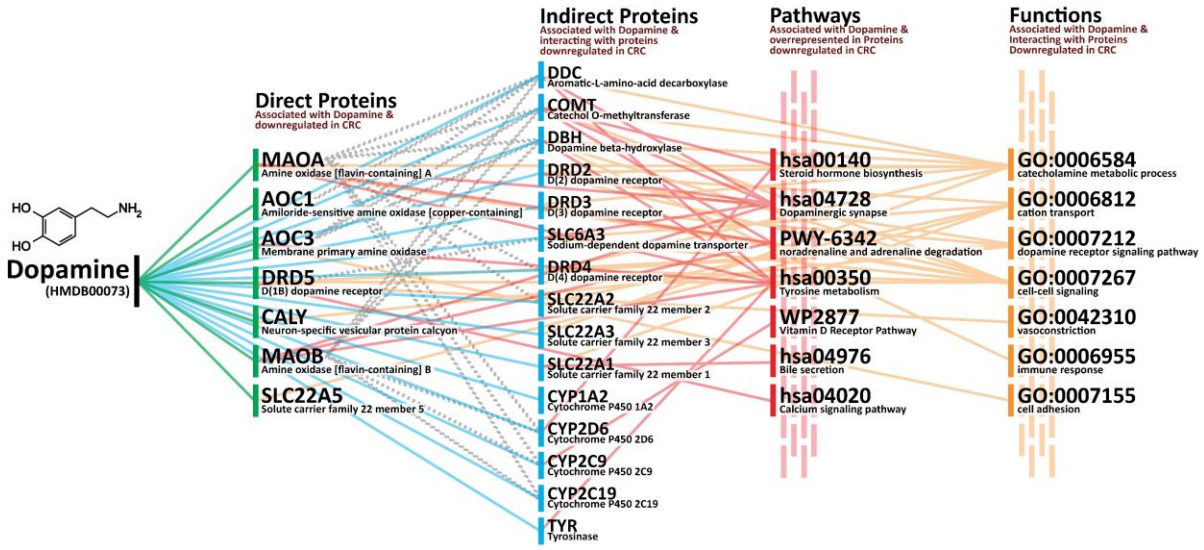

Orotidylic Acid
(HMDB00218)
Leukemia
Colorectal cancer
Cortisone reductase deficiency 2
Charcot-Marie-Tooth Disease
Spastic Paraplegia
Arts Syndrome
Orotic Aciduria
HMDB11720
Thyroid carcinoma
Bladder cancer

**Figure 4**


Dopamine
(HMDB00073)
Direct Proteins
Associated with Dopamine & downregulated in CRC
MAOA
Amine oxidase [flavin-containing] A
AOC1
Amiloride-sensitive amine oxidase [copper-containing]
AOC3
Membrane primary amine oxidase
DRD5
D(1B) dopamine receptor
CALY
Neuron-specific vesicular protein calcyon
MAOB
Amine oxidase [flavin-containing] B
SLC22A5
Solute carrier family 22 member 5
Indirect Proteins
Associated with Dopamine & interacting with proteins downregulated in CRC
DDC
Aromatic-L-amino-acid decarboxylase
COMT
Catechol O-methyltransferase
DBH
Dopamine beta-hydroxylase
DRD2
D(2) dopamine receptor
DRD3
D(3) dopamine receptor
SLC6A3
Sodium-dependent dopamine transporter
DRD4
D(4) dopamine receptor
SLC22A2
Solute carrier family 22 member 2
SLC22A3
Solute carrier family 22 member 3
SLC22A1
Solute carrier family 22 member 1
CYP1A2
Cytochrome P450 1A2
CYP2D6
Cytochrome P450 2D6
CYP2C9
Cytochrome P450 2C9
CYP2C19
Cytochrome P450 2C19
TYR
Tyrosinase
Pathways
Associated with Dopamine & overrepresented in Proteins downregulated in CRC
hsa00140
Steroid hormone biosynthesis
hsa04728
Dopaminergic synapse
PWY-6342
noradrenaline and adrenaline degradation
hsa00350
Tyrosine metabolism
WP2877
Vitamin D Receptor Pathway
hsa04976
Bile secretion
hsa04020
Calcium signaling pathway
Functions
Associated with Dopamine & Interacting with Proteins Downregulated in CRC
GO:0006584
catecholamine metabolic process
GO:0006812
cation transport
GO:0007212
dopamine receptor signaling pathway
GO:0007267
cell-cell signaling
GO:0042310
vasoconstriction
GO:0006955
immune response
GO:0007155
cell adhesion

**Table_S1.xls.** List of MSD_MAP-predicted binding signatures between cancer metabolites and target proteins**.**

**Table 1.** Validations of predicted metabolite protein targets**.**

| **Metabolite Name** | **HMDB ID** | **Protein Uniprot** | **Protein Name** |
|---|---|---|---|
| Androstenedione | HMDB00053 | P10275 | Androgen Receptor |
| Androstenedione | HMDB00053 | P42330 | Aldo-keto reductase family 1 member C3 |
| Androstenedione | HMDB00053 | P10275 | Androgen Receptor |
| Androstenedione | HMDB00053 | P51857 | 3-oxo-5-beta-steroid 4-dehydrogenase |
| Cholesterol | HMDB00067 | P35398 | Nuclear receptor ROR-alpha |
| Citric acid | HMDB00094 | P12931 | Proto-oncogene tyrosine-protein kinase Src |
| Orotidylic acid | HMDB00218 | P11172 | Uridine 5'-monophosphate synthase |
| Tetrahydrodeoxycorticosterone | HMDB00879 | P52895 | Aldo-keto reductase family 1 member C2 |
| Tetrahydrodeoxycorticosterone | HMDB00879 | Q04828 | Aldo-keto reductase family 1 member C1 |
| Tetrahydrodeoxycorticosterone | HMDB00879 | P42330 | Aldo-keto reductase family 1 member C3 |
| Tetrahydrodeoxycorticosterone | HMDB00879 | P52895 | Aldo-keto reductase family 1 member C2 |
| Alpha-Linolenic acid | HMDB01388 | P14555 | Phospholipase A2, membrane associated |
| 11-Dehydrocorticosterone | HMDB04029 | P51857 | 3-oxo-5-beta-steroid 4-dehydrogenase |
| 11-Dehydrocorticosterone | HMDB04029 | P28845 | Corticosteroid 11-beta-dehydrogenase isozyme 1 |
| 11-Dehydrocorticosterone | HMDB04029 | P51857 | 3-oxo-5-beta-steroid 4-dehydrogenase |

**Table 2.** Validated metabolites significantly associated with colorectal cancer (CRC) gene expression-derived multi scale biological activity**.**

| Metabolite Name | HMDB ID | Associated with Up-regulated or Down-regulated CRC Activity? | Hypergeometric P-value for Protein Level | Hypergeometric P-value for Pathway Level | Hypergeometric P-value for Function Level | Hypergeometric P-value for PPI Level | Reference |
|---|---|---|---|---|---|---|---|
| 2-hydroxybutyrate | HMDB00008 | Down | 1.000 | < .001 | < .001 | < .001 | [74] |
| 2-Hydroxybutyric acid | HMDB00008 | Down | 1.000 | < .001 | < .001 | < .001 | [74] |
| Deoxyuridine | HMDB00012 | Down | 1.000 | < .001 | < .001 | 0.002 | [34] |
| Butyric acid | HMDB00039 | Down | 0.110 | 0.039 | < .001 | < .001 | [75] |
| Acetic acid | HMDB00042 | Down | 0.190 | 0.002 | < .001 | < .001 | [75] |
| Beta-Alanine | HMDB00056 | Up | 0.059 | 0.047 | < .001 | < .001 | [34] |
| Cholesterol | HMDB00067 | Down | 0.274 | < .001 | < .001 | < .001 | [75] |
| Dopamine | HMDB00073 | Down | 0.004 | < .001 | < .001 | < .001 | [34] |
| Citric acid | HMDB00094 | Down | 0.202 | 0.001 | < .001 | < .001 | [34] |
| Homovanillic acid | HMDB00118 | Down | 1.000 | < .001 | < .001 | 0.019 | [34] |
| D-Glucose | HMDB00122 | Down | 0.024 | 0.077 | < .001 | < .001 | [76] |
| L-Glycine | HMDB00123 | Down | 0.058 | < .001 | < .001 | < .001 | [76] |
| Glycine | HMDB00123 | Down | 0.058 | < .001 | < .001 | < .001 | [76] |
| D-Glucuronic acid | HMDB00127 | Down | < .001 | < .001 | < .001 | < .001 | [34] |
| Glycerol | HMDB00131 | Down | 0.005 | < .001 | < .001 | < .001 | [34] |
| Glyceric acid | HMDB00139 | Up | 1.000 | 0.017 | < .001 | < .001 | [34] |
| L-Glutamic acid | HMDB00148 | Up | 0.001 | 0.062 | < .001 | < .001 | [34] |
| L-Tyrosine | HMDB00158 | Up | 0.059 | < .001 | < .001 | < .001 | [34] |
| L-Phenylalanine | HMDB00159 | Up | 0.065 | 0.010 | < .001 | 0.004 | [76] |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| D-Mannose | HMDB00169 | Down | 0.174 | < .001 | < .001 | < .001 | [76] |
| L-Histidine | HMDB00177 | Down | 0.380 | < .001 | < .001 | 0.002 | [34] |
| Aspartic acid | HMDB00191 | Up | 0.361 | 0.032 | < .001 | < .001 | [64] |
| L-Aspartic acid | HMDB00191 | Up | 0.361 | 0.032 | < .001 | < .001 | [64] |
| Ornithine | HMDB00214 | Down | 0.324 | 0.012 | < .001 | 0.030 | [34] |
| Palmitic acid | HMDB00220 | Down | 0.047 | < .001 | < .001 | < .001 | [76] |
| Pyruvic acid | HMDB00243 | Down | 0.064 | < .001 | < .001 | < .001 | [34] |
| Taurine | HMDB00251 | Up | 0.019 | 0.004 | 0.041 | 0.001 | [34] |
| Urea | HMDB00294 | Down | 0.366 | 0.014 | < .001 | 0.006 | [35] |
| Uridine | HMDB00296 | Down | 0.281 | 0.009 | < .001 | 0.013 | [71] |
| L-Cysteine | HMDB00574 | Down | 0.283 | 0.029 | < .001 | < .001 | [75] |
| Dodecanoic acid | HMDB00638 | Down | 0.181 | 0.008 | < .001 | 0.010 | [34] |
| L-Glutamine | HMDB00641 | Up | 0.130 | 0.013 | < .001 | < .001 | [34] |
| L-Methionine | HMDB00696 | Down | 0.195 | 0.029 | < .001 | 0.050 | [34] |
| Stearic acid | HMDB00827 | Down | 0.041 | < .001 | < .001 | < .001 | [76] |
| Arachidonic acid | HMDB01043 | Down | 0.042 | < .001 | < .001 | < .001 | [76] |
| Putrescine | HMDB01414 | Down | 0.096 | 0.024 | < .001 | < .001 | [34] |
| Phosphate | HMDB01429 | Down | 0.068 | < .001 | < .001 | < .001 | [76] |
| Paraxanthine | HMDB01860 | Down | 0.041 | < .001 | < .001 | < .001 | [34] |
| Aspirin | HMDB01879 | Down | 0.142 | < .001 | < .001 | 0.013 | [34] |
| Oleamide | HMDB02117 | Down | 0.077 | < .001 | < .001 | 0.006 | [35] |
| Marganic acid | HMDB02259 | Down | 0.024 | < .001 | < .001 | < .001 | [76] |
| Heptadecanoic acid | HMDB02259 | Down | 0.024 | < .001 | < .001 | < .001 | [76] |
| 1-Monooleoylglycerol | HMDB11567 | Down | 0.120 | 0.005 | < .001 | 0.006 | [76] |

**Table 3.** Validated metabolites significantly associated with esophageal cancer (EC) gene expression-derived multi scale biological activity**.**

| **Metabolite Name** | **HMDB ID** | **Associated with Up-regulated or Down-regulated EC Activity?** | **Hypergeometric P-value for Protein Level** | **Hypergeometric P-value for Pathway Level** | **Hypergeometric P-value for Function Level** | **Hypergeometric P-value for PPI Level** | **Reference** |
|---|---|---|---|---|---|---|---|
| 4-Hydroxyphenylpyruvic acid | HMDB00707 | Up | 0.273 | 0.006 | < .001 | 0.010 | [54] |
| Acetic acid | HMDB00042 | Up | 0.058 | 0.001 | < .001 | < .001 | [54] |
| Acetoacetic acid | HMDB00060 | Down | 0.198 | < .001 | < .001 | 0.001 | [54] |
| Adenosine monophosphate | HMDB00045 | Down | 0.010 | 0.037 | < .001 | < .001 | [54] |
| Choline | HMDB00097 | Down | 0.126 | < .001 | < .001 | 0.017 | [54] |
| Creatine | HMDB00064 | Down | 0.002 | 0.001 | < .001 | 0.043 | [54] |
| Ethanol | HMDB00108 | Down | 0.007 | < .001 | < .001 | < .001 | [54] |
| Formic acid | HMDB00142 | Down | 0.007 | 0.006 | < .001 | < .001 | [54] |
| Gamma-Aminobutyric acid | HMDB00112 | Up | 0.216 | < .001 | < .001 | 0.061 | [54] |
| D-Glucose | HMDB00122 | Up | 0.168 | 0.004 | < .001 | 0.003 | [54] |
| L-Glutamic acid | HMDB00148 | Down | 0.102 | < .001 | < .001 | < .001 | [54] |
| 4-Hydroxybutyric acid | HMDB00710 | Down | 0.048 | 0.009 | 0.015 | 1.000 | [54] |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| L-Glutamine | HMDB00641 | Up | 0.024 | 0.002 | < .001 | 0.003 | [54] |
| L-Aspartic acid | HMDB00191 | Down | 0.359 | < .001 | < .001 | < .001 | [54] |
| L-Tyrosine | HMDB00158 | Up | 1.000 | 0.038 | < .001 | 0.003 | [54] |
| NAD | HMDB00902 | Down | 0.045 | < .001 | < .001 | < .001 | [54] |
| L-Phenylalanine | HMDB00159 | Down | 1.000 | < .001 | < .001 | 0.010 | [54] |
| Uracil | HMDB00300 | Up | 0.075 | < .001 | < .001 | 0.001 | [54] |
| Malonic acid | HMDB00691 | Down | 1.000 | 0.004 | < .001 | 0.024 | [77] |
| Fumaric acid | HMDB00134 | Down | 1.000 | 0.025 | < .001 | 0.009 | [77] |
| l-Serine | HMDB00187 | Up | 0.214 | 0.034 | < .001 | 0.015 | [77] |
| L-Aspartate | HMDB00191 | Down | 0.359 | < .001 | < .001 | < .001 | [54] |
| Pyruvic acid | HMDB00243 | Down | < .001 | < .001 | < .001 | < .001 | [77] |
| Inosine | HMDB00195 | Up | 0.151 | < .001 | < .001 | 0.003 | [77] |
| Uridine | HMDB00296 | Up | 0.003 | < .001 | < .001 | < .001 | [77] |
| Cytidine | HMDB00089 | Up | 0.103 | < .001 | < .001 | < .001 | [77] |
| A-glucose | HMDB03345 | Up | 0.339 | < .001 | < .001 | < .001 | [77] |
| B-hydroxybutyrate | HMDB00060 | Down | 0.198 | < .001 | < .001 | 0.001 | [54] |
| L-Margaric acid | HMDB00827 | Down | 0.047 | < .001 | < .001 | < .001 | [77] |
| Myristic acid | HMDB00806 | Down | 0.037 | 0.001 | < .001 | < .001 | [77] |
| Linoleic acid | HMDB00673 | Down | 0.024 | 0.009 | < .001 | < .001 | [77] |

**Table 4.** Validated metabolites significantly associated with prostate cancer (PC) gene expression-derived multi scale biological activity**.**

| **Metabolite Name** | **HMDB ID** | **Associated with Up-regulated or Down-regulated PC** | **Hypergeometric P-value for Protein Level** | **Hypergeometric P-value for Pathway Level** | **Hypergeometric P-value for Function Level** | **Hypergeometric P-value for PPI Level** | **Reference** |
|---|---|---|---|---|---|---|---|

| | | **Activity?** | | | | | |
|---|---|---|---|---|---|---|---|
| Glycine | HMDB00123 | Down | 0.185 | < .001 | < .001 | < .001 | [65] |
| L-Glutamic acid | HMDB00148 | Up | 0.038 | < .001 | < .001 | < .001 | [65] |
| S-Adenosylhomocysteine | HMDB00939 | Down | 0.010 | 0.047 | < .001 | 0.087 | [65] |
| Phosphate | HMDB01429 | Down | 0.076 | < .001 | < .001 | < .001 | [65] |
| Choline | HMDB00097 | Down | 0.063 | < .001 | < .001 | < .001 | [65] |
| Glycerol | HMDB00131 | Down | 0.014 | < .001 | < .001 | < .001 | [65] |
| Adenosine | HMDB00050 | Down | 0.281 | 0.001 | < .001 | < .001 | [65] |
| ADP | HMDB01341 | Down | 0.048 | < .001 | < .001 | < .001 | [65] |
| Citric acid | HMDB00094 | Up | 0.367 | < .001 | 0.001 | < .001 | [65] |
| N-Acetyl-L-alanine | HMDB00766 | Down | 0.299 | 0.003 | < .001 | 0.017 | [78] |
| N-Acetylglutamic acid | HMDB01138 | Down | 0.299 | 0.003 | < .001 | 0.017 | [78] |
| Deoxyuridine triphosphate | HMDB01191 | Down | 1.000 | < .001 | < .001 | < .001 | [78] |
| 5-Methylcytosine | HMDB02894 | Up | 1.000 | < .001 | < .001 | 0.019 | [79] |
| Selenomethionine | HMDB03966 | Down | 0.350 | < .001 | < .001 | 0.014 | [80] |
| Bradykinin | HMDB04246 | Down | 0.050 | < .001 | < .001 | 0.032 | [78] |