Natural Product Target Network in Cancer: Definition and Application

By Steve Chamberlin, ND, Mac

Master's Thesis

Presented to the
Division of Bioinformatics and Computational Biology
Department of Medical Informatics and Clinical Epidemiology
and the
Oregon Health & Science University
School of Medicine
in partial fulfillment of the requirements of

Masters of Biomedical Informatics

April 2018

School of Medicine

Oregon Health & Science University

Certificate of Approval

This is to certify that the Master's Thesis of

Steven R. Chamberlin

"Natural Product Target Network in Cancer: Definition and Application"

Has been approved

_____

Shannon McWeeney
Thesis Advisor
Department of Medical Informatics and Clinical Epidemiology

_____

Guanming Wu
Thesis Committee
Department of Medical Informatics and Clinical Epidemiology

_____

Lynne Shinto
Thesis Committee
Department of Neurology

Table of Contents

Acknowledgments

First, I would like to acknowledge the tremendous support and guidance of my advisor, Shannon McWeeney, on this project. I came in to this program with an idea that I felt was a bit of a long shot. Dr. McWeeney guided me into a project that for which there was no compromise of my dream. I also want to acknowledge the support of the other two members of my committee, Guanming Wu and Lynne Shinto. All of my committee members have brought very complementary support to this project. In addition, I have also received support from Molly Kulesz Martin in dermatology, Aurora Blucher (PhD candidate), and Garbielle Choonoo in the BioDev lab. I could not have completed this without the help from everyone. And the support of the fellows and staff in DMICE has also been critical to my project and to my sanity!

Abstract.

There is a body of research demonstrating in-vitro and in-vivo synergy between Natural Products and anti-neoplastic drugs for some cancers. Natural products (NP) are isolated compounds derived from plants and fungi. The core premise of my work is that the examination of the target space associated with natural products will increase the number of potential therapeutically accessible targets and lead to novel combination therapies for cancer treatment. When considering Reactome pathways only targeted by natural products, at all levels of target evidence, there is an increase in coverage of 61%, relative to pathways covered by FDA approved antineoplastic drugs. The Cancer Targetome is the target network associated with FDA approved antineoplastic drugs. Not only are the number of pathways targeted increased when considering the natural product target space, but target coverage, or interactions, is increased in pathways already targeted by Cancer Targetome drugs. We also examined the distribution of cancer driver genes across pathways and found 24 pathways enriched for cancer drivers that had no cancer Targetome drug interactions (based on binding affinity threshold of < 100 nM) but had at least one target interaction with a natural product at that same binding threshold. Assessment of network context highlighted the fact that natural products show target family groupings both distinct from and in common with cancer drugs, further highlighting the potential for natural products in the cancer therapeutic space. These findings support the potential for discovery of novel combination therapies when considering natural products.

1. Background

The National Cancer Institute lists eight categories of cancer treatments. These include surgery, radiation, chemotherapy, immunotherapy, targeted therapy, stem cell transplant and precision medicine [1]. Historically, surgery, radiation and chemotherapy were the primary forms of treatment. In the late 1990s the FDA started approving targeted therapies for cancer, i.e. a therapy directed towards unique molecular characteristics that drive oncogenesis. Imatinib treatment has shown an 80% decrease in 5-year mortality with chronic myeloid leukemia patients [2]. While some of the early targeted therapies have resulted in dramatic clinical responses, drug resistance often develops after an initial positive response. This adaptation to treatment is known as acquired drug resistance, as opposed to intrinsic resistance, which exists prior to any cancer therapy [3].

Acquired drug resistance is seen with both cytotoxic chemotherapies and targeted therapies, although mechanisms differ. Knowledge of the molecular mechanisms of resistance can inform therapeutic strategies. In cancer, these mechanisms can include compensatory and redundant molecular signaling, target mutations acquired during treatment, increased expression of the targeted protein, inactivation of pro-apoptotic pathways, inhibition of DNA repair mechanisms, epithelial-mesenchymal transition, activation of pro-survival signaling, and upregulation of tumor cell efflux transporters [3, 4] . For drug resistance caused by mutations in drug targets or redundant cell pathways, ‘rational combinatorial targeted therapy’ is a possible solution [5]. This ‘rational’ approach is done within the framework of network pharmacology, which brings together systems biology, network analysis, redundancy, and consideration of all

drug target effects, beyond therapeutic intention, for designing therapies [6]. Knowledge of molecular signaling pathways can be used to design multi-target strategies to block redundant pathways or newly mutated targets. Simultaneous targeting of multiple cancer hallmarks is another approach [6]. Along with reduction in drug resistance, this approach can also lead to decreased adverse effects and increased efficacy [7] . For these combinations, multiple therapeutic agents can be used andthese methods can also take advantage of poly-pharmacological characteristics of each single agent [8]. Drugs can also work together through pharmacokinetic mechanisms, coalistic mechanisms [9], and through independent actions when used in combination [10]. A coalisitic interaction is when two compounds interact in a biological context to form a new third compound.

Computational methods, including approaches in graph theory [11], assessment of differential gene expression [12], and modeling via ordinary differential equations have been successfully developed to predict synergy between compounds in-silico. Predicting these effective drug combinations often require up-to-date comprehensive knowledge of the target space associated with a set of compounds, such as FDA approved cancer drugs. Thus, the core premise of my thesis project is that the examination of the target space associated with natural products (NP) will increase the number of potential therapeutically accessible targets and lead to novel combination therapies for cancer treatment.

Natural products can be broadly defined as any compound derived from a living source (animal, plant, microbial, fungi). This absolute definition would include ‘natural’ cosmetics, ‘natural’ foods, wood, silk, bioplastics and even coal. A more specific definition was based on the following from the National Center for Complementary and Integrative Health (NCCIH) [13]:

> ‘Natural products include a large and diverse group of substances from a variety of sources. They are produced by marine organisms, bacteria, fungi, and plants. The term encompasses complex extracts from these producers, but also the isolated compounds derived from those extracts. It also includes vitamins, minerals and probiotics.’

We have further refined this definition with the following inclusion and exclusion criteria. Only NPs from plants and fungi were included, and only isolated compounds. In addition, only plants and fungi that have a history of traditional medicinal use were included. These compounds already have some historic use as therapeutic agents, and some are already classified as safe for human consumption by the FDA’s Dietary Supplement Health and Education Act of 1994.

Natural product compounds show greater structural diversity, bioactivity and complexity than compounds in synthetic drug libraries, have the ability to inhibit some targets considered ‘undruggable’, such as protein-protein interactions, and inherently target biologically relevant space since they are mostly secondary metabolites, or signaling molecules [14]. There is also limited overlap between the molecular space targeted by natural products and targeted by synthetic drug libraries [14]. These characteristics not only indicate the potential for new targets for therapy, but also can help reduce the cost of the development of new treatments, since these molecules already exist in nature, and offer additional options for combination therapies. As a matter of fact, natural products, or natural product derivatives, are the source of 33% of cancer drugs developed between 1981-2014 [15]. There is also a body of literature demonstrating in-vitro and in-vivo natural product synergies with cancer drugs [16] [17], overcoming drug resistance with the addition of natural products [18], and paradoxical synergy in cancer cells with

simultaneous antagonism in healthy tissue [17]. Therefore, the exploration of the natural product target space could offer the potential to improve existing drug therapy outcomes and reduce side effects.

There is a need to not only compile a comprehensive set of natural product targets from public sources, but also to characterize these compound-target relationships, within a variety of relevant contexts (**Fig 1)**, so that this knowledge can be applied to the prediction of therapeutic combinations with approved cancer drugs or other natural products. To the author's knowledge, a characterization such as this has not been done for natural products and compared to the companion space for cancer drugs. Such a characterization was recently completed for all FDA-approved antineoplastic drugs from multiple public resources by Blucher, *et al*, the Cancer Targetome (CT) [19]. The methods and evidence framework from that project were applied here to develop and characterize a natural product target network (natural product compounds linked to associated targets) and to compare it to the Cancer Targetome. The characterization of this network includes target/pathway coverage, compound promiscuity assessment, inclusion of established cancer driver genes, and the potential to inhibit molecular compensatory signaling mechanisms. The characterization of this natural product target space will not only identify potential new targets, but also evaluate the importance of these targets in a molecular context for cancer. In particular, the functional changes created with the addition of natural product targets to the Cancer Targetome will be evaluated for potential of synergy between these compounds.

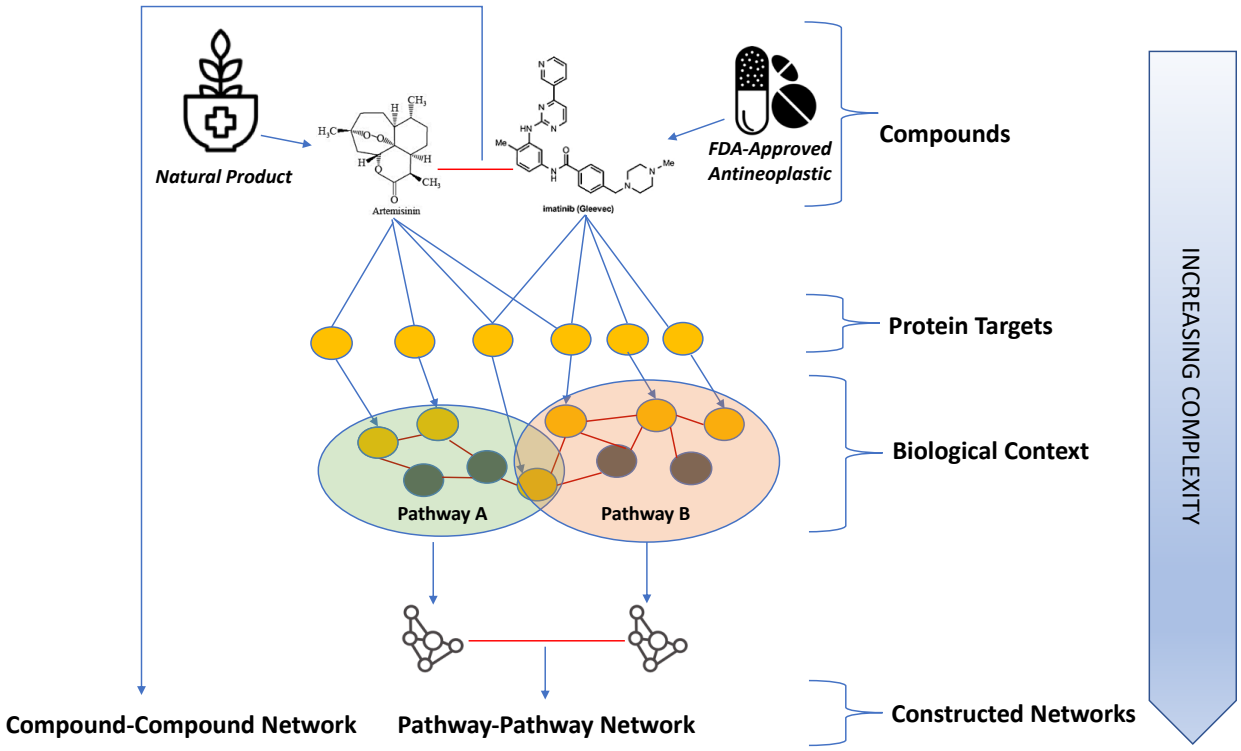

Figure 1. Natural Product Target Network Evaluation. The targets associated with both natural products and anti-neoplastic drugs were evaluated in different contexts of increasing complexity. Complementary and distinct coverage of protein targets and pathways by the two compound classes were assessed. Target importance and relationships were evaluated in biological contexts, which include protein-protein interaction networks and molecular pathways. Pathway relationships and shared target space were assessed through the construction of a pathway-pathway network and a compound-compound network. Red lines indicate the existence of an edge between nodes in these networks. Two compounds have an edge if they share at least one target, and two pathways have an edge if they share at least one protein. Graphics from Creative Mahira, Becris, Takao Umehara [20].

2. Methods

2.1 Data Collection

A total of seven public databases were used to construct the natural product target network used in this project, as sources of both compounds and targets. Two of the seven databases were the source of natural product chemical compounds, TarNet (2016) [21] and The Traditional Chinese Medicine Integrated Database for herb molecular mechanism analysis (TCMID, version 2.01) [22]. These databases were chosen because they contain only compounds from plants used from medicinal traditions.

Four of the target source databases used were not specific to natural product compounds, and were also used in the creation of the Cancer Targetome. These included DrugBank, version 5.0.7 [23], Therapeutic Targets Database, version 4.3.02 [24], the International Union of Basic and Clinical Pharmacology (IUPHAR) / British Pharmacological Society (BPS) database, version 2017.4 [25], and BindingDB (7/1/2017). Additional human target interactions were included from three databases that contained only natural product information. These include TarNet, TCMID and the Universal Natural Products Database (UNPD, 2016) [26].

All seven databases were used for the human target interaction information for these compounds.

2.1.1 Base Natural Product Compounds

The TarNet database, the first of the two sources of NP compounds used, contains information about 12,187 compounds derived from 894 medicinal plants. These plants are used in four traditions of botanical medicine: Chinese, Japanese, European and American [21]. This database also contains 10,783 bio-targets associated with the plant compounds. Both the plant-compound and compound-biotarget relationships were derived by text-mining and manual curation. A random sample of 246 literature references, used by TarNet to support the compound-biotarget relationships, were manually checked for accuracy and also for the presence of binding affinity data in the source.

The second source, Traditional Chinese Medicine Integrated Database (TCMID) comprises plants, associated compounds, and bio-targets curated from Traditional Chinese Medicine (TCM). The goal of this database seeks to translate the common factors between modern western medicine and TCM. The TCMID database contains 8,159 plants, 43,413 associated compounds, and 17,521 bio-targets compiled through a combination of other databases and text mining [20][27].

The compounds from these two databases were combined and redundancies were removed. This was done through the use of the multiple keys associated with each compound in each database. TarNet compounds were classified by chemical name, Chemical Abstracts Service (CAS) number, simplified molecular-input line-entry system (SMILES), and International Union of Pure and Applied Chemistry's International Chemical Identifier Key (INCHIKEY). TCMID compounds were classified by PubChem Compound ID, SMILES and chemical name. The Chemical Translation Service (CTS) from the University of California at Davis was used to help resolve some of the missing key data in TarNet [28]. The two compound lists were then combined and scrubbed for possible drugs that are not of NP origin using lists

from DrugBank [23] and the FDA. To check whether these compounds are consistent with the natural product definition given for this project, 135 entries were randomly selected and curated. The five keys found in the source databases: chemical name, SMILES, INCHIKEY, Pubchem Compound ID and CAS number, were then used to retrieve target information from a variety of sources as detailed below.

2.1.2 Biological Target Retrieval

Molecular target interaction information for the NP compounds was retrieved from seven publicly available data sources. Four widely used data sources were chosen based on the rationale used by Blucher, *et al* for the Cancer Targetome [19]. These sources include DrugBank [23], Therapeutic Targets Database [24], the International Union of Basic and Clinical Pharmacology (IUPHAR) / British Pharmacological Society (BPS) Guide to Pharmacology [25], and BindingDB [29]. These four sources contain substantial information about NPs, but are not limited to this class of compounds. Three additional sources of target interaction data, which are limited only to NPs, were also used. These include TarNet [21], TCMID [22], and the Universal Natural Products Database (UNPD) [26]. All of the interaction data used for this project was based on literature and/or experimental data, no computationally predicted interactions were used.

This data was retrieved by systematically merging each of the five NP keys, from the base NP data, individually to the seven target data sources. The data retrieved by the individual key was then associated back to the unique five key combination of values from which it was derived and data redundancies were resolved to create the final NP target network database.

The Evidence Level framework developed by Blucher, *et al* was then applied to this data as follows. Each NP-target interaction could have more than one piece of supporting evidence. Each piece of evidence was assigned one of three levels. Evidence Level I only had an entry in one of the databases for the interaction, without a supporting literature reference or experimental binding value. Evidence Level II has a supporting literature reference in the database, and Evidence Level III would also have an experimental binding value, in addition to the literature reference. For NP-target interactions with multiple evidence entries from different databases, the maximum Evidence Level was assigned. For a single target, the maximum Evidence Level from all NP-target interactions was assigned. Analyses for this project will consider all levels of evidence, when appropriate, and also focus on interactions only with evidence of strong binding affinity, less then 100 nM. Less stringent levels can still be considered biologically relevant, but below 100 nM is considered significant for drug binding [30, 31].

2.2 Target, Pathway, and Tumor Type Coverage

Analyses compared the two target networks above, i.e. the NP target network and the Cancer Targetome (CT). First, the two sets of targets associated with the two networks were mapped onto two sets of molecular pathways. The Reactome Pathway Knowledgebase [32] was the source of the pathway information, which leverages biological entity reactions.

A total of 1,944 hierarchically structured pathways used from this database for the first pathway mapping, which used all of these pathways. For a subsequent pathway mapping a set of pan-cancer aberrant pathways were derived from the full set of Reactome pathways by performing an over representation analysis with likely cancer driver genes. The genes catalogued

for the Cancer Genome Interpreter [33] were used for this analysis. These genes have either experimental, clinical or in-silico evidence showing that their mutations can drive tumorigenesis. There are 837 genes catalogued representing 193 different tumor types. A hypergeometric test was used to identify pathways enriched with these driver genes. The Benjamini Yekuteli method was used to control the false discovery rate for multiple testing with dependencies [34]. If a pathway contained at least one molecular target with evidence of an interaction with either a NP or a CT drug, the entire pathway was considered targeted by one of those two compound categories. Pathways, from both pathway sets, were then classified as either targeted by NP only, CT drugs only, both NP and CT drugs, or neither. This same mapping classification was also applied to at the protein target level, as opposed to the pathway level, from all of the targets associated with both target networks, and to targets only associated with the pan-cancer aberrant pathways. And finally, tumor types were identified for that were associated with cancer driver genes targeted only by NPs. Only high affinity interactions (IC50, EC50, Ki, KD less than 100 nM) were considered for this analysis.

2.3 Molecular Interaction Network Topology

For this analysis, the NP targets, CT drug targets and cancer driver gene products were projected onto biological networks and the topological features of these targets were evaluated and compared. Target oriented topology research often uses a large non-specific protein-protein interaction network (PPI) for the biological context, but it has also been suggested that the base network should be either be more specific to the tissue or disease of interest, or have greater biological relevance [35]. For this reason, two specific interaction networks were used for this

evaluation. For the first biological network, the protein interactions from the Reactome Functional Interaction Network were used. This network integrates uncurated relationships from sources such as PPI databases and others, with curated interaction information derived from pathway data in Reactome and other databases [36]. These functional interactions have a higher likelihood of being functional in a biological context than an interaction from an uncurated PPI database. The second network is a PPI constructed by Wang, *et al* [37] from the integration of four manually curated human cancer signaling networks with protein interactions from BioGRID [38-42].

Network measures that were considered include degree centrality, betweenness centrality, eigenvector centrality, and average shortest distance to cancer driver genes in the network. Degree centrality is the number of connections a node in a network has to neighboring nodes. In a biological network this could be the number of different interaction partners a single protein might have. Betweenness centrality is the number of shortest paths in a network that pass through a specific node. In a biological network this measure can capture the node's ability to control communication [35]. Eigenvector centrality is a measure of how connected the nodes are that are connected to the node of interest. This is a measure of importance of the node's neighbors in the biological network. The average shortest distance is the path between two nodes that contains the minimum number of nodes, or steps. In a biological network this can be related to the number of reaction steps between two proteins [39][43]. Proteins that are closer could have a higher likelihood of impacting each other, if used as a therapeutic target.

To compare the topology measures, the nodes in the network were classified in four ways: targeted only by NPs, targeted only by CT drugs, targeted by both, and not targeted by either. These were compared to each other for similarities and differences to evaluate any

additional therapeutic potential achieved with the addition of NPs. Only targets with strong binding affinities, less than 100 nM, were considered.

2.4 Compensatory Pathways

To assess compound synergy through pathway crosstalk inhibition we constructed a pathway-pathway network. In this network each node is an entire pathway, and there is an edge between pathways if they share at least one gene. This structure was chosen as a good method to evaluate the NP and CT drug target network's crosstalk inhibition potential since it is possible to measure global patterns in this framework and assess the two target sets in their entirety. Pathways in this network were then classified as being associated with NPs only, CT drugs only, both or none. A compound was associated with a pathway only if there was at least one target with a strong binding affinity, less than 100 nM, in a pathway. Average shortest path distance between the two pathway sets was assessed as a prediction of synergy. This metric has been shown to be correlated with synergy [44]. Reactome entity level pathways (non-hierarchical) were chosen to construct the network, of which there are 290, and an edge was created between two pathways if at least one gene was shared between them. In addition, the previously described cancer driver genes were used to do an over representation analysis on this set of 290 pathways, identifying possible aberrant cancer pathways in this network. Orientation to these cancer pathways in this network was then assessed for the pathways associated with NPs only, CT drugs, and both NPs and CT drugs. Distance between NPs and CT drugs was also assessed to estimate the possible synergistic potential between NPs and CT drugs, and also to estimate increased cross talk inhibition by considering NPs.

2.5 Compound-Compound Network

In this network, a node represents a compound, either NP or CT drug, and an edge exists if they share at least one target, both having a binding value of less than 100 nM. The edges were also weighted based on the number of targets shared. This network analysis has also been used for FDA approved drugs [43]. Network communities were identified and assessed for enrichment in either NPs or CT drugs to see if they are clustering around the same targets, or separate sets of targets. The targets of each network community were also classified based on protein families, defined in IUPHAR. Multilevel clustering was used as the community detection method. This method is a greedy algorithm (creates an optimal solution at each step to find a global optimum) that creates communities based on maximizing modularity, and is recommended for networks of this size. This method is also appropriate for unconnected networks and can use edge weights to determine community structure. [45].

2.6 Targets, Pathways, Tumor Types and Cancer Drivers Per Compound

The multi-targeting aspect of these two compound classes (NP and CT drug) is considered the basis for their poly-pharmacological effects [8]. These effects can be undesirable, such as adverse events, or they could be the mechanism of the therapeutic effect. For this reason, it is important to map out and compare this characteristic of the two target networks, NP and CT, and to understand how NPs might differ from the CT drugs.

For most analyses in this project we only considered target interactions with binding affinities less than 100 nM, but for this analysis we considered two binding affinity thresholds: less than 1000nM and less than 100nM. Distributions of interactions per compound were then compared between NPs and CT drugs for four categories: targets per compound, pathways per compound, cancer driver genes per compound, and number of tumor types targeted per compound. A two sample Kolomogorov Smirnov test was used to compare the two distributions, NP and CT drugs.

2.7 In-Silico Synergy Prediction Methods for Drugs

There is a body of research for methods to predict synergistic drug combinations for drugs. Two of these methods were selected to test their ability to predict natural product/drug synergistic combinations. For this, a literature search was done to create a curated truth dataset of known natural product/drug synergistic pairs. From these pairs the individual compounds were reshuffled to create a set of unknown, or unlabeled pairs. These two sets were combined to create a semi-supervised dataset to be used to test the prediction methods.

The first method is the Ranking System of Anti-Cancer Synergy (RACS) by Sun *et al[11].* This is a semi-supervised method trained on 26 known synergistic pairs of drugs from the Drug Combination Database (DCDB). Seven features were selected from a pool of fourteen as the base model. These features are all derived from evaluations of the two target networks associated with the two compounds that are being ranked for synergy. Features include Gene Ontology-based mutual information entropy, average distance in a protein-protein interaction network (PPI), drug combination interference, efficacy for betweenness, degree, and eigenvector centrality, and unrelated mapped pathway pairs. Gene Ontology-based mutual information entropy is a measure of the overlap in biological function associated with the two compounds.

Average distance is the average shortest path between the two target networks in a PPI created from BioGrid [42], MINT [46], HPRD [47] and DIP [48]. Drug combination interference is a measure of the change in network information transmitting efficiency seen when removing the targets associated with the two compounds, either separately or simultaneously. Efficacy is a based on three measures: betweenness, degree and eigenvector centrality. It is an assessment of the difference in these measures between cancer targets and non-cancer targets in the PPI when projecting the two compound target networks into this PPI. And unrelated mapped pathway pairs is the percent of cancer pathways associated with the two compounds, via their target networks, that do not communicate through shared genes or a PPI. Manifold ranking was then used to rank pairs based on similarity to a the known 26 'bait' pairs. This method then does a secondary screen of candidate combinations using two measures based on cell line gene expression data after treatment with compounds of interest. One measure looks at the overlap of upregulated and downregulated targets between the two compounds, and the other measure looks at percentage of disease pathways targeted by the two compounds. Input data includes cell line gene expression data and a drug-target-disease network.

The other method was developed by Liu *et al* [12]. The main point of this paper was that combining structural dissimilarity with gene expression similarity increases predictive power for synergy. The two features evaluated for this model include compound structural similarity as measured by Tanimoto coefficient based on the molecular binary fingerprint downloaded from Pubchem. The other feature is based on differentially expressed gene signatures created for each compound individually. This feature is the number of upregulated and downregulated genes that overlap between the two compounds.

2.8 Analysis Tools

The analysis was done using Base R (v3.3.1). Packages used included Dplyr (v0.7.4) and iGraph (v1.1.2). Work was also done in Cytoscape (v3.4.0).

3. Results

3.1 Natural Product Data Distributions

The final list of NPs, compiled from TarNet and TCMID, contained 50,109 compounds, uniquely identified by the five keys found in the source databases: chemical name, SMILES, INCHIKEY, Pubchem Compound ID and CAS number. Of these compounds, target interaction data was retrieved for 4,991 from at least one of the seven public databases. There are 137 anti-neoplastic drugs in the Cancer Targetome.

Most of the NP-target interactions are classified as Evidence Level I (94%), and most CT interactions are classified as Evidence Level III (95%), although the absolute counts are much larger for NPs (**Fig 2a)**. For stronger affinity interactions (Evidence Level III Exact), the raw numbers of interactions are comparable between natural products and cancer drugs (**Fig 2b**). The large volume of Level I evidence for NPs is primarily from the TarNet database, which was created by text mining. All of these interactions are supported by literature references, which would be Evidence Level II, but the publicly available data from this source does not have a link to these references. Curation of a random sample of these data also shows that approximately 37% of the interactions do have binding data available in the supporting literature, and would therefore be assigned Level III if the link to these references was available.

Targets are reported according to the highest evidence level of all of the associated compound interactions for each target (**Fig 2a**). This perspective allows us to take a target-focused perspective and determine the maximum strength of evidence supporting any compound/drugs interacting with it. So, if we had a set of targets of interest, this would allow us to prioritize according to those targets with stronger evidence for interactions with compounds/drugs. As with the target-interaction distribution, most of the NP associated targets have a maximum Evidence Level I (83%) and most of the CT drug associated targets have a maximum Evidence Level III (85%), but the distributions are more comparable for within Evidence Level III (**Fig 2b)**.

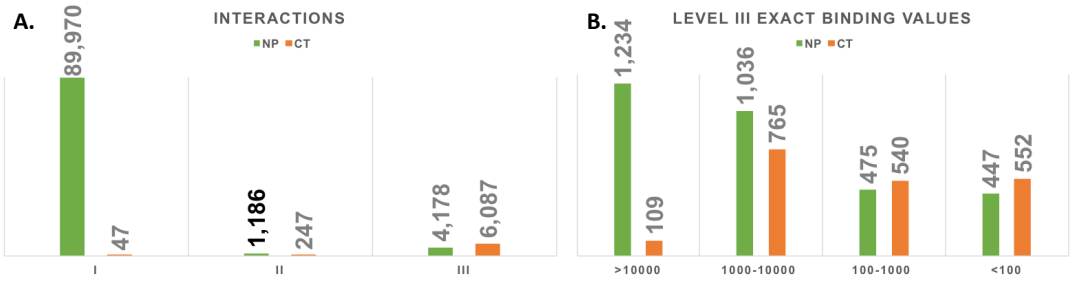

Figure 2. Compound-Target interactions. CT=Cancer Targetome, NP=Natural Product Target Network. **A.** Comparison of interaction distribution by evidence level between NPs and CT drugs. **B.** Comparison of interaction distribution at high affinity only (Evidence Level III exact values). Values are in nM.

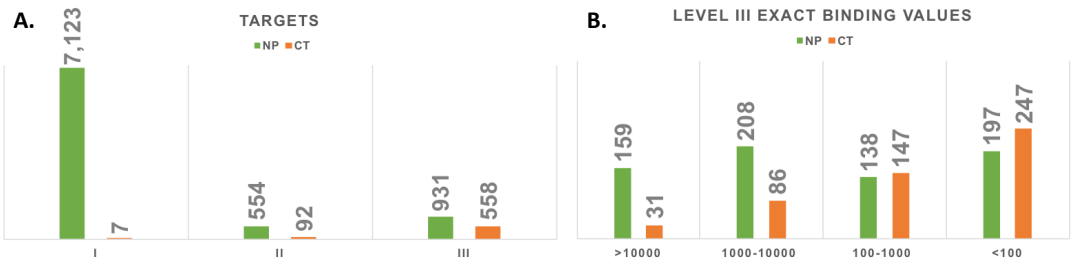

Figure 3. Maximum Compound Interaction Level Per Target. CT=Cancer Targetome, NP=Natural Product Target Network. **A.** Comparison of maximum target evidence level distribution between NPs and CT drugs. **B.** Comparison of maximum target evidence level distribution at high affinity only (Evidence Level III exact values). Values are in nM.

3.2 Natural product space increases coverage of cancer pathways, targets and tumor types

Coverage is considered at all levels of evidence and only for interactions with a binding value of less than 100 nM. Of the 1,944 hierarchically structured pathways contained in the full Reactome database, 533 were considered pan-cancer aberrational, based on over-representation analysis.

For all Reactome pathways, at all levels of evidence, natural products increased coverage by 61%, relative to pathways covered by both natural products and Cancer Targetome drugs, or Cancer Targetome drugs only. Reactome pathway coverage at the 100 nM level was increased by 29%, relatively (**Table 1**). For the aberrational cancer pathways, the NP relative coverage increase for all evidence levels was 12%, and 6% at the 100 nM level (**Table 2**). The percentage of aberrational cancer pathways targeted by both NPs and CT drugs was higher than it is for all

Reactome pathways, which would be expected since the Cancer Targetome is specific to the disease domain associated with this subset of pathways.

When considering the NP contributions within pathways targeted by both NPs and CT drugs, 51% of the individual target interactions in the 1,196 Reactome pathways are with NPs only. And for the 495 Reactome pathways targeted by both at the 100 nM level, 43% of the target interactions are with NPs only. For cancer pathways targeted with affinities less than 100 nM there is a high degree of overlap at the pathway level, but very little overlap at the target level (**Fig 4).** Not only are the number of pathways targeted increased when considering the NP target space, but coverage is increased in pathways already targeted by CT drugs.

When considering all of the targets contained in both target networks, NP and CT, there is a large number of interactions with only NPs (**Table 1).** This is also true for the 7,339 targets associated with the aberrational cancer pathways (**Table 2).** The vast majority of these interactions are Level I evidence from the TarNet database, which are derived from text mining. Manual curation of a random sample of this data suggests that 37% could have valid binding values associated with NPs in the supporting literature. At the higher binding affinity (100 nM), there is a relative increase in target coverage of 65% for all targets from the two networks, and 60% for targets from the aberrational cancer pathways.

Evidence and Binding	# Unique Targets			# Reactome Pathways Targeted			% of Reactome Pathways		
	NP Only	CT Only	Both	NP Only	CT Only	Both	NP Only	CT Only	Both
Evidence Levels I,II,III	7,978	27	630	725	0	1196	37.29%	0.00%	61.52%
Evidence Levels II,III	964	129	521	385	21	1174	19.80%	1.08%	60.39%
Evidence Levels III	453	80	478	275	30	1070	14.15%	1.54%	55.04%
Evidence Level III, Binding LT100	160	210	37	218	266	495	11.21%	13.68%	25.46%

Table 1. Consideration of natural product targets increases both pathway and target coverage in all Reactome pathways. This pathway data includes hierarchically nested Reactome pathways.

Evidence and Binding	# Targets in Cancer Pathways			# Cancer Pathways			% of Cancer Pathways		
	NP Only	CT Only	Both	NP Only	CT Only	Both	NP Only	CT Only	Both
Evidence Levels I,II,III	3,488	16	387	59	0	474	11.07%	0.00%	88.93%
Evidence Levels II,III	422	87	313	24	2	472	4.50%	0.38%	88.56%
Evidence Levels III	250	51	279	35	4	450	6.57%	0.75%	84.43%
Evidence Level III, Binding LT100	102	145	26	24	97	286	4.50%	18.20%	53.66%

Table 2. Consideration of natural product targets increases both pathway and target coverage in cancer pathways. The cancer pathways include 533 pathways and 7,339 associated targets from the Reactome hierarchically nested pathways.

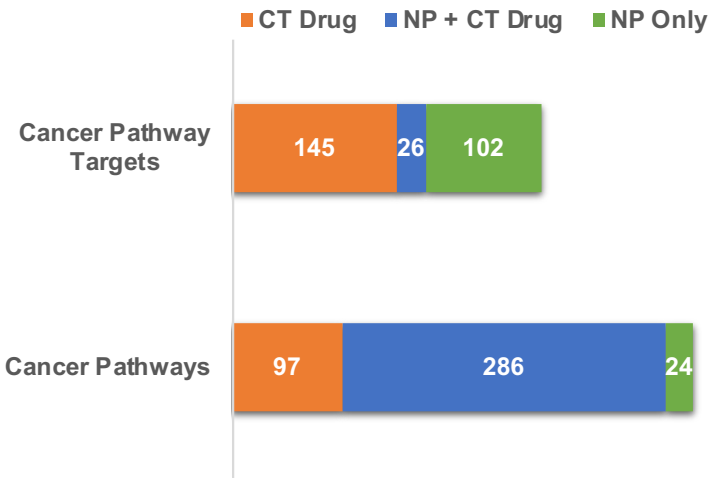

Figure 4. High level of cancer pathway overlap at affinities less than 100 nM, but little overlap at the target level

Target interactions were then assessed for the 837 cancer drivers that were used for the over representation analysis. Only binding values less than 100 nM were considered for this analysis. Cancer drivers with these interactions were then mapped back to their associated tumor types. There were twelve tumor types for which NPs increased driver coverage (**Table 3)**. Five cancer drivers, uniquely targeted by NPs at less than 100 nM were associated with these 12 tumor types. These include Adenylate Cyclase 1 (ADCY1), Matrix Metallopeptidase 2 (MMP2),

Aryl Hydrocarbon Receptor (AHR), Cyclin Dependent Kinase 2 (CDK2), and Mitogen-Activated Protein Kinase 11 (MAP3K11). These cancer drivers are targeted by five NPs. Some of the NPs that target one or more of these drivers include forskolin (from the Indian coleus plant), caffeic acid (found in many foods including coffee), kaempferol (found in many common foods including apples, grapes, tomatoes and green tea), and flavopiridol (a semi-synthetic derivative from the Pithraj tree).

Tumor Type	CT Only	NP Only	Both NP and CT	Total Cancer Drivers
Cutaneous Melanoma	20	2	4	291
Prostate Adenocarcinoma	10	2	3	153
Bladder	10	1	2	195
Esophagous	3	1	3	124
Head and Neck Squamous	9	1	4	188
Hodgkin Lymphoma	0	1	0	12
Large B-cell Lymphoma	1	1	0	3
Lung Adenocarcinoma	18	1	3	209
Neuroblastoma	2	1	0	31
Renal clear cell	6	1	2	116
Small Cell Lung	2	1	0	59
Uterine Corpus Endometroid Carcinon	9	1	4	158

Table 3. Natural Products Improve Coverage of Cancer Drivers Across Cancer Types. Tumor types with drivers targeted only by NPs, at binding values of 100nM or less. The NP Only column lists the number of drivers targeted only by NPs. There are five driver genes associated with these 12 tumor types.

There were 24 pathways in Reactome that were enriched for cancer drivers, had no CT drug interactions at less than 100 nM, and had at least one target interaction with a NP with a binding affinity less than 100 nM. In some of these pathways, such as **'RUNX1 regulates transcription of genes involved in differentiation of myeloid cells'**, the NP interaction was not with a driver gene (**Fig 5).** In this example the interacting NP is a phorbol ester isolated from croton oil. These compounds have been called 'double edged swords', showing both tumor promoting and tumor inhibiting activities, depending on the cancer [49]. This pathway is

involved with the differentiation of myeloid progenitors and also with apoptosis of mature myeloid cells [50]. The three driver genes in this pathway are associated with over 20 tumor types, including acute myeloid leukemia, acute lymphoblastic leukemia, chronic myeloid leukemia, myelodysplastic syndrome, and non-Hodgkin's lymphoma. The close proximity of the NP-target interaction to the drivers in this pathway could indicate a potential to inhibit associated aberrant signals. Further in-vitro testing would be necessary to test the actual effect of this interaction.

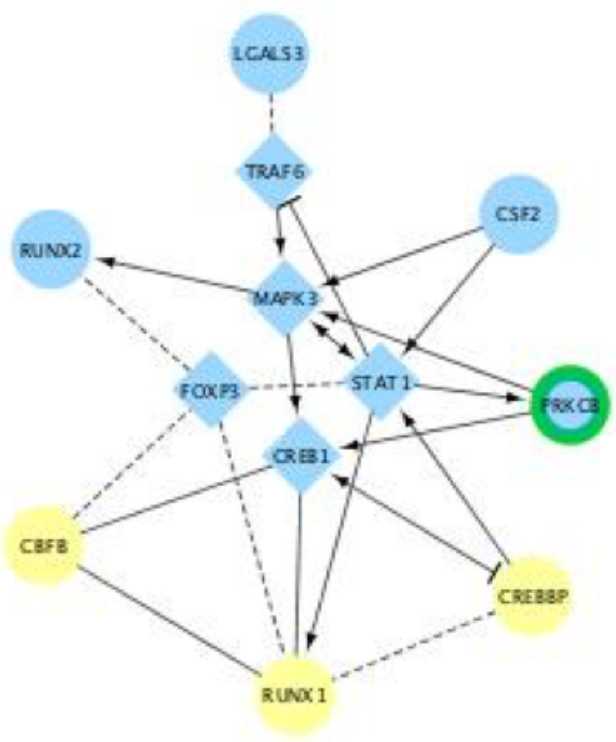

Figure 5. Cancer pathway targeted only by natural products for a non cancer driver target. This Reactome pathway (**RUNX1 regulates transcription of genes involved in differentiation of myeloid cells)** is an example of an aberrant cancer pathway targeted only by NPs at less than 100 nM (green border). No FDA-approved cancer drugs target this pathway with <100nM evidence. Cancer drivers are shown in yellow. The NP target is not a cancer driver. Dashed lines are predicted interactions. The diamond shaped genes are connectors, not part of the formal pathway.

In other pathways, such as **'TP53 Regulates Transcription of Genes Involved in G1 Cell Cycle Arrest'**, the NP interaction was directly with a driver gene (**Fig 6).** The interacting NP for this example is flavopiridol, which is a semisynthetic derivative of a compound extracted from the Pithraj tree. This compound is already in numerous clinical trials for several cancer types [51]. This pathway primarily inhibits the cell cycle transition from the G1 phase to the S phase through multiple mechanisms [50]. Over 40 tumor types are associated with the five cancer drivers in this pathway, including breast, bladder, esophagus, head and neck, lung, prostate, ovary, and hepatocellular.

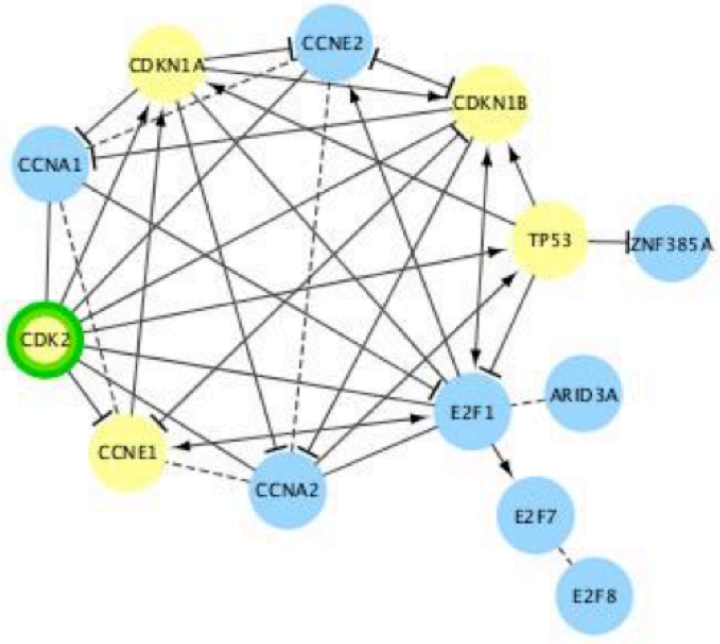

Figure 6. Cancer pathway targeted only by natural products for a cancer driver target. Another aberrant cancer pathway (**TP53 Regulates Transcription of Genes Involved in G1 Cell Cycle Arrest)** targeted only by NPs at less than 100 nM (green border). No FDA-approved cancer drugs target this pathway with <100nM evidence. Cancer drivers are shown in yellow. In this pathway the NP target is also a cancer driver. Dashed lines are predicted interactions.

There were also 286 pathways in Reactome that were enriched for cancer drivers that were targeted by both NPs and CT drugs at binding affinities less than 100 nM. One smaller pathway example of this is **"MAPK3 (ERK1) activation" (Fig 7)**. The two NP compounds targeting the three proteins in this pathway are flavopiridol, mentioned before, and arctigenin, extracted from the burdock plant. This pathway is involved in a wide range of cellular processes, including cytoskeleton remodeling, proliferation, differentiation and regulation of inflammatory responses [50]. Over 20 tumor types are associated with the four cancer drivers found in this pathway, including leukemias, lung cancers, colorectal cancer, melanoma and head and neck squamous cell cancer.

NPs offer a substantial increase in coverage of both pathways and targets compared with FDA-approved antineoplastic drugs. It is now important to assess the functional significance of these targets and pathways in a biological context.

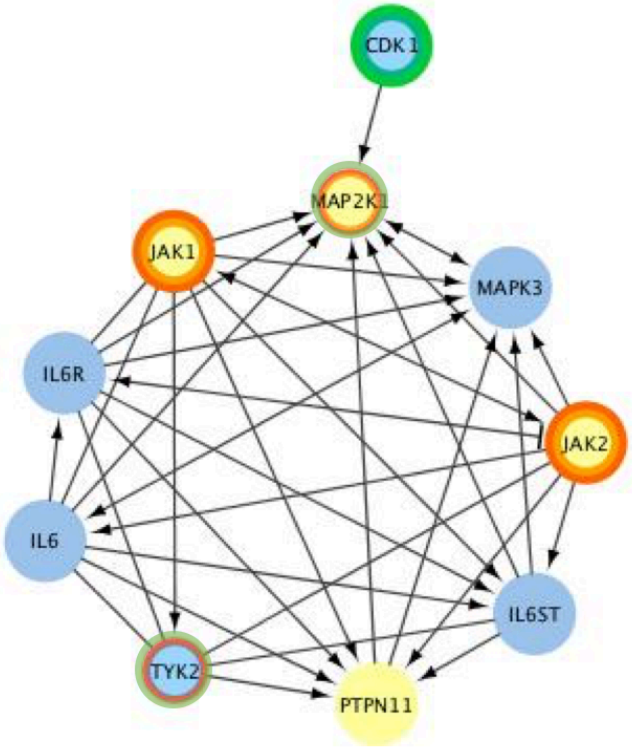

Figure 7. Cancer pathway targeted by both NPs and CT drugs. This Reactome pathway (**MAPK3 (ERK1) activation**) is an example of an aberrant cancer pathway targeted by both NPs (3) and FDA-approved cancer drugs (4) at <100nM evidence. Cancer drivers are shown in yellow. A solid green border is a target for NP only, borders with both green and orange is a target for both a NP and a CT drug, and orange borders are targets for only CT.

3.3 Natural product targets and Cancer Targetome targets show similar topology characteristics in selected biological networks

It is important to consider the target sets associated with NPs and CT drugs in the context of molecular interaction networks, such as protein-protein interactions networks (PPI). The topology of these networks can be related to biological function [52]. Some topology measures commonly considered relevant to biological function include degree centrality, betweenness

centrality, eigenvector centrality, average shortest paths, and clustering coefficients [53, 54]. These measures, and others, have been suggested as a guide to identify potential therapeutic targets for drug development [35, 52, 55, 56]. In general, these biological networks are scale free, meaning that many nodes have low connectivity and are not critical to the function of the network, but some nodes are highly connected and more critical, possibly making good therapeutic targets. Topology measures can guide the identification of these critical nodes. Both degree and betweenness are considered measures of strong importance to the network, but nodes with high values of either can be considered too lethal or toxic to target [35, 52, 55, 56]. It is possible to identify nodes with network influence that have lower essentiality [52] by using measures such as eigenvector centrality, bridging centrality, and others [35, 55, 56]. Eigenvector centrality is a measure of the connectivity of the nodes that are connected to the node of interest. Bridging centrality, developed by Hwang, *et al,* is a measure of how well a node connects separate modular subregions in a network. Other strategies for reducing adverse effects consider targets that influence critical nodes without targeting the critical nodes themselves [35, 57]. For this reason, assessing the shortest path distances from critical nodes, such as cancer drivers, is important. This distance could also be considered an estimate of the number of molecular steps from this node [43].

It has also been shown that cancer genes have topology measures in biological networks that are distinctly different from other nodes. These differences include higher degree and betweenness, shorter paths between them and weaker clustering coefficients [52, 58, 59].

Biological network topology has also been used to predict therapeutic synergy between compounds, an aspect of interest for designing novel combination therapies [60]. Several in-silico methods for prediction of synergy in combination therapy use topology measures [11, 61,

62]. Generally, higher values of measures, such as degree and betweenness, are preferred in these models.

As mentioned in the methods, two biological networks (PPIs) were used. One is the Reactome Functional Interaction network, and the other was created by Wang, *et al*, based on interactions specific to cancer biology. The three target (node) categories include nodes (proteins) targeted by NPs only, targeted by CT drugs only, and targeted by both CT drugs and NPs. In Reactome, these three node categories represented 393 of the 12,227 nodes in the network. In the Wang (cancer) network, these three node categories represented 364 of the 6,306 nodes in the network. Comparisons were done for the topology measures degree, betweenness and eigenvector centrality.

Betweenness Centrality

In the Reactome network, betweenness for all three target node categories was significantly higher than it was for non-targeted nodes in the network. There was no significant difference between these three categories. Consistent with previous research, betweenness was significantly higher for the cancer driver nodes than for non-driver nodes in this network.

In the cancer network created by Wang, *et al* [37], as with the Reactome network, all three target node categories had higher betweenness than the non-target nodes. The nodes targeted only by NPs are not different from the nodes targeted only by CT drugs, but the betweenness for these nodes is less than that for the nodes targeted by both NPs and CT drugs. And, consistent with previous research, the cancer driver nodes were significantly higher for betweenness than the non-cancer driver nodes.

The results for this statistic were mostly consistent between the two networks. Overall nodes targeted by NPs are similar for this characteristic to nodes targeted by CT drugs, and different from non-targeted nodes. Another interesting characteristic is that the average value for the nodes targeted by both NPs and CT drugs was higher for this measure than the average for nodes targeted by either NP only or CT drugs only, although this difference was only significant when compared to CT only drug targets in the cancer network created by Wang, *et al* [37] (p=.0001), and there were only 39 nodes in this category.

Betwenness is the measure of the number of shortest paths in a network that pass through a node. The target with the highest betweenness in the Reactome network, interacting with NPs only at binding less than 100 nM, was nuclear factor kappa beta subunit 1 (NFKB1). The interacting NP compound was rocaglamide, extracted from a variety of *Aglaia* plant species. This gene functions in a dozen large molecular pathways, ranging in size from 72 to 758 genes, most of which are enriched with cancer drivers (**Fig 8**). This protein is a transcription factor found in almost all cell types and is involved in a wide variety of cell processes, including some that are related to cancer, such as inflammation, tumorigenesis, apoptosis and cell growth. This protein is activated by many intracellular and extracellular stimuli, which is consistent with a high degree of betweenness centrality.

Degree Centrality

Degree, in the Reactome network, is not significantly different between all three target node categories, but each of these three categories are all significantly higher than non-targeted

nodes in the network, as was seen with betweenness. And the cancer driver nodes were also higher for this measure than non-driver nodes.

In the Wang, *et al* (cancer) network the nodes targeted by NPs only were not significantly different from non-targeted nodes, and they were significantly lower for this measure than nodes targeted by CT drugs only or nodes targeted by both NPs and CT drugs. The cancer driver nodes were also significantly higher for this measure in this network.

For degree centrality the two networks are not in agreement. The nodes targeted only by NPs are similar to CT drug nodes in the Reactome network, but not in the Wang, *et al* (cancer) network. But nodes targeted by both NPs and CT drugs are again higher on average than either of the other two categories, NP only and CT drugs only, although it is only significant for the comparison to the NP only targeted nodes.

The node with the highest degree targeted only by NPs at a binding affinity less than 100 nM in the Reactome network is the same as the top node for betweenness for this category, nuclear factor kappa beta subunit 1 (NFKB1). In this network there are 515 interactions with other proteins for this protein target. Of these 515 proteins, 90 (17.5%) are cancer drivers.

Eigenvector Centrality

For this measure, in the Reactome network, nodes targeted only by NPs are not different than non-targeted nodes, are significantly higher than nodes targeted by CT drugs only and significantly lower than nodes targeted by both CT drugs and NPs. The nodes targeted only by CT drugs are significantly lower than non-targeted nodes, but not different from nodes targeted by both CT drugs and NPs. And the nodes targeted by both CT drugs and NPs are significantly

higher than non-targeted nodes. Cancer driver nodes are significantly higher for this measure than non-cancer driver nodes in this network.

In the Wang, *et al* (cancer) network, nodes targeted only by NPs are not different from non-targeted nodes, but are significantly less then nodes targeted only by CT drugs, and nodes that are targeted by both CT drugs and NPs. Nodes targeted by CT drugs only and those targeted by both CT drugs and NPs are significantly higher than non-targeted nodes. CT only nodes are not different from nodes targeted by both CT drugs and NPs. And cancer driver nodes are significantly higher than non-cancer driver nodes for this measure.

The two networks are not in agreement for these node categories for this measure. Generally, nodes targeted by NP only are not the same as other targeted nodes in both networks, but the direction of these differences is not the same. This is partly due to the fact that the CT drug only nodes in the Reactome network are generally lower than most other categories, including non-targeted nodes.

The node targeted only by NPs that has the highest eigenvector centrality in the Reactome network is Heterogeneous Nuclear Ribonucleoprotein A1 (HNRNPA1). This means that the other nodes directly connected to HNRNPA1 have a high level of connectivity**.** There are 185 proteins interacting with this protein in this network, with an average degree of 217, which is significantly higher than the network average of 38. While there are only 27 cancer drivers in this list, or 15% of the 185 proteins, one of these proteins, E1A Binding Protein P300 (EP300), has over 1000 interactions listed in Reactome. HNRNPA1 is an abundant core protein found in the cell nucleus and functions in alternative splicing of RNA.

There is a high level of overlap for the top genes ranked by betweenness and degree centrality for the three target categories **(Table 4).** There is less overlap between these lists and

the genes with the highest eigenvector centrality. The three target categories are by definition mutually exclusive, and seem to show some possible complementarity in cancer therapy strategies, based on cancer hallmarks.

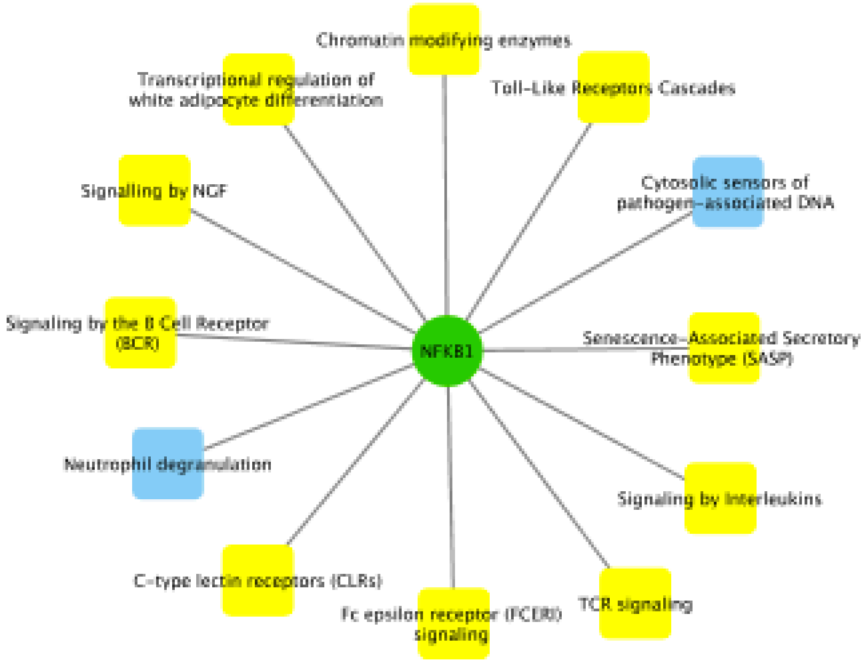

Figure 8. Top natural product only target for betweenness centrality. Nuclear factor kappa beta subunit 1 is shown in the center, illustrating a high betweenness centrality. Squares represent pathways that this gene participates in. If a pathway is yellow, it is enriched for cancer driver genes.

NP Only			CT Only			Both CT and NP		
Gene	Value	Driver ?	Gene	Value	Driver ?	Gene Name	Value	Driver ?
Betweenness Centrality								
Nuclear Factor Kappa B Subunit 1 (NFKB1)	901,715	No	SRC Proto-Oncogene, Non-Receptor Tyrosine Kinase (SRC)	911,785	Yes	Epidermal Growth Factor Receptor (EGFR)	947,674	Yes
RELA Proto-Oncogene, NF-KB Subunit (RELA)	613,037	No	FYN Proto-Oncogene, Src Family Tyrosine Kinase (FYN)	616,175	No	Histone Deacetylase 1 (HDAC1)	669,591	No
Cyclin Dependent Kinase 1 (CDK1)	532,453	No	Janus Kinase 2 (JAK2)	504,217	Yes	Estrogen Receptor 1 (ESR1)	575,569	Yes
MDM2 Proto-Oncogene (MDM2)	395,265	Yes	Retinoid X Receptor Alpha (RXRA)	460,453	No	Histone Deacetylase 2 (HDAC2)	510,638	Yes
Protein Kinase C Beta (PRKCB)	315,021	No	Mitogen-Activated Protein Kinase 14 (MAPK14)	389,578	No	Androgen Receptor (AR)	445,269	Yes
Protein Kinase C Alpha (PRKCA)	251,601	No	Retinoic Acid Receptor Alpha (RARA)	300,939	Yes	Cyclin Dependent Kinase 4 (CDK4)	299,008	Yes
Degree Centrality								
Nuclear Factor Kappa B Subunit 1 (NFKB1)	521	No	SRC Proto-Oncogene, Non-Receptor Tyrosine Kinase (SRC)	569	Yes	Epidermal Growth Factor Receptor (EGFR)	569	Yes
RELA Proto-Oncogene, NF-KB Subunit (RELA)	470	No	FYN Proto-Oncogene, Src Family Tyrosine Kinase (FYN)	497	No	Histone Deacetylase 1 (HDAC1)	497	No
Cyclin Dependent Kinase 1 (CDK1)	468	No	Janus Kinase 2 (JAK2)	398	Yes	Histone Deacetylase 2 (HDAC2)	398	Yes
Protein Kinase C Alpha (PRKCA)	297	No	Mitogen-Activated Protein Kinase 14 (MAPK14)	332	No	Histone Deacetylase 3 (HDAC3)	332	Yes
Protein Kinase C Beta (PRKCB)	296	No	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Delta (PIK3CD)	324	Yes	Cyclin Dependent Kinase 4 (CDK4)	324	Yes
MDM2 Proto-Oncogene (MDM2)	260	Yes	LYN Proto-Oncogene, Src Family Tyrosine Kinase (LYN)	311	No	Estrogen Receptor 1 (ESR1)	311	Yes
Eigenvector Centrality								
Heterogeneous Nuclear Ribonucleoprotein A1 (HNRNPA1)	0.77	No	Retinoic Acid Receptor Alpha (RARA)	0.14	Yes	Histone Deacetylase 3 (HDAC3)	0.13	Yes
Cyclin Dependent Kinase 7 (CDK7)	0.25	No	Aurora Kinase B (AURKB)	0.13	No	Histone Deacetylase 1 (HDAC1)	0.10	No
Cyclin T1 (CCNT1)	0.18	No	Retinoid X Receptor Alpha (RXRA)	0.12	No	Epidermal Growth Factor Receptor (EGFR)	0.09	Yes
Cyclin Dependent Kinase 9 (CDK9)	0.17	No	Nuclear Receptor Corepressor 1 (NCOR1)	0.12	Yes	Histone Deacetylase 2 (HDAC2)	0.08	Yes
Cyclin Dependent Kinase 1 (CDK1)	0.17	No	SRC Proto-Oncogene, Non-Receptor Tyrosine Kinase (SRC)	0.10	Yes	Cyclin Dependent Kinase 4 (CDK4)	0.07	Yes
RELA Proto-Oncogene, NF-KB Subunit (RELA)	0.11	No	Retinoic Acid Receptor Beta (RARB)	0.09	No	Hypoxia Inducible Factor 1 Alpha Subunit (HIF	0.06	No

Table 4. Top genes for the three topology measures (betweenness, degree, eigenvector centrality). The top 6 genes with the highest values of betweenness, degree and eigenvector centrality are shown for each node category: targeted by NP only, targeted by CT drugs only, or targeted by both NP and CT drug. All interactions are less than 100 nM binding affinity.

In both networks, the average shortest path distances from NP only, CT drug only and CT/NP both nodes to cancer driver nodes were significantly shorter than random controls**.** The nodes targeted by both CT drugs and NPs are the closest to the cancer driver nodes, on average. These nodes also have higher betweenness and degree. These nodes represent a subset of both CT drug nodes and NP nodes that appear to be more critical in both networks. These 39 nodes also contain a higher percentage of cancer drivers than the other categories.

3.4 Pathway interactions reveal potential synergistic relationships between natural products and cancer drugs.

It is believed that the use of compensatory and redundant molecular pathways by the cancer cell is one of the mechanisms of acquired drug resistance, and that rational drug combination therapy could inhibit these processes. Several methods have been proposed to select and predict drug combinations that could inhibit this phenomenon [11, 35, 44, 63]. Chen, *et al,* constructed a pathway-pathway interaction network to evaluate drug synergy and as a possible model for the inhibition of pathway crosstalk. In this network, a node represents an entire molecular pathway and an edge represents an interaction between two pathways. Three types of interactions were used to construct three networks. Manually curated interactions from the KEGG database [64], protein-protein interactions, and shared genes were the three edge

definitions. Two sets of pathways were defined by associating the two drugs via the projection of the associated drug target networks onto the pathways. Chen, *et al*, found that drug synergy was correlated with the average shortest distance between the two sets of pathways in this network, most strongly in the network with edges representing shared genes.

For this analysis, the pathway network developed by Chen (*et al)* using shared genes as the criteria for an edge between two pathway nodes was used. This network contained 285 pathways (nodes) with 9,152 edges. These 285 were non-hierarchical Reactome entity level pathways. Of the 285 pathways, 90 were enriched with cancer drivers. The majority of the cancer pathways are targeted by both NPs and CT drugs (73%), and most of the non-cancer pathways are not targeted by either NPs or CT drugs (48%). Only target interactions less than 100 nM were considered.

The average shortest path in this network between pathways targeted by CT drugs, those targeted by NPs and between NP targeted pathways and CT drug targeted pathways were all similar, and closer than random controls. Based on previous research this could correlate with synergy between these classes of compounds, particularly between NPs and CT drugs. Pathways targeted by both CT drugs and NPs were closer on average to cancer pathways than random controls. Pathways targeted by NPs only and those targeted by CT drugs only were not closer to cancer pathways than random controls **(Fig 9).**

There are 559 examples of neighboring pathways (nodes) in this network where one is targeted only by NPs only and the other is a cancer enriched pathway targeted by either CT drugs or both CT drugs and NPs. An example is **DAG and IP3 signaling** and **Signaling by FGFR1**. In the Reactome database, these two pathways share the gene PLCG1 (1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase gamma-1), which creates an edge between them in this network.

Signaling by FGFR1 is enriched with cancer drivers, and has a CT drug target with an interaction less then 100 nM, BRAF (B-Raf Proto-Oncogene, Serine/Threonine Kinase). This pathway initiates intracellular signaling pathways involved with cell proliferation and migration, and other functions. **DAG and IP3 signaling** is not enriched with cancer drivers and does not have any low affinity CT drug targets, but does have 6 high affinity NP targets: PRKCG,E,A,D (Protein Kinase C), PDE1A (Phosphodiesterase 1A) and ADCY1 (Adenylate Cyclase 1). DAG (diacylglycerol) and IP3 (inositol 1,4,5-trisphosphate) are secondary messengers used in intracellular signaling.

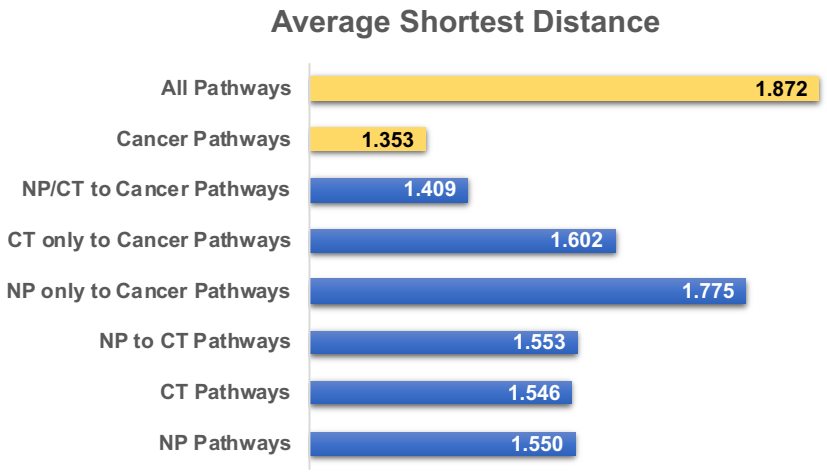

Figure 9. **Pathway-Pathway network analysis.** Average shortest path distances in blue were all shorter than random controls, except for NP only and CT only nodes to cancer nodes (empirical $p<.01$, 1000 permutations).

3.5 Natural products and FDA-approved anti-neoplastic drugs do not have similar distributions for targets and pathways interacting per compound.

We found that there are differences in the distributions of the number of targets and pathways interacting per compound between NPs and CT drugs, and that the binding affinity level impacted whether we saw these differences for tumor types per compound and cancer driver interactions per compound **(Table 5)**. But it is difficult to draw any further conclusions about these differences since the motivations for studying the relationships are not necessarily the same between NPs and FDA approved cancer drugs, and are not captured in this data, although this rationale can be found in the supporting literature references, if a link exists in the data. Previous research has shown an increasing number of target interactions per compound along the drug development pipeline, with approved drugs having the highest level [8],[65]. Also, the total number of targets tested with these compounds, both NPs and CT drugs, is not captured in the public data resources used for this project. And finally, missing data is always a problem when assessing this compound characteristic since all possible target interactions that occur in a biological context cannot be known.

Type	Affinity Level	KS Test	
		p	D
Targets	LT100 All Targets *	0.002222	0.26364
	LT1000 All Targets *	0.0002767	0.27611
	LT100 Cancer Pathway Targets *	0.003176	0.28632
	LT1000 Cancer Pathway Targets *	0.001013	0.27778
Pathways	LT100 All Pathways *	7.37E-12	0.52679
	LT1000 All Pathways *	3.28E-13	0.51282
	LT100 Cancer Pathways *	2.35E-10	0.54144
	LT1000 Cancer Pathways *	9.88E-13	0.54069
Tumors	LT100	0.02306	0.45714
	LT1000 *	0.0008806	0.42083
Drivers	LT100	0.1024	0.37302
	LT1000 *	0.003771	0.37917

Table 5. **Distribution comparisons for Targets, Pathways, Tumors and Cancer Drivers.** Differences between the number of targets, pathways, cancer drivers and tumors interactions per compound for NPs and CT drugs were tested using a two sample Kolmogorov Smirnov test to

detect differences in the distributions. This testing was assessed for binding levels of 100 nM or less, and 1000 nM or less. Tumor associations made via cancer driver interactions.

*indicates significance at p threshold of <.004 (Bonferroni adjustment)

3.5 Natural products show target family groupings both distinct from and in common with cancer drugs

There were a total of 253 compounds and 1,238 edges in the compound network, which included 68 CT drugs and 185 NPs. This network contained 26 unconnected subnetworks, with the majority of nodes (163) in one large connected component. Multilevel clustering created 35 communities, most of which contained 5 or less compounds. Of interest were the three largest communities, each containing over 30 compounds **(Fig 10a).** One community was dominated by CT drugs, but also contained substantial NPs. The other two communities contained primarily NPs.

The first community contained 32 CT drugs and 11 NPs. The community clustered primarily around kinase and other cancer related target families **(Table 6)**. Kinases are extremely well-targeted by current FDA-approved cancer drugs and have been an active area of research following the break through kinase inhibitor imatinib. The next largest community was comprised of 39 NPs and 4 CT drugs. This community contained a variety of cytochrome P450 families and other target families considered therapeutic targets, such as the carbonate dehydratases which are inhibited for the treatment of glaucoma and other conditions **(Table 6)**. The third largest community contained no CT drugs and 30 NPs. The target protein families here were primarily hormone and neurotransmitter receptors **(Table 6).**

The clustering patterns seen in the larger communities in this network seem to indicate NP target interaction research driven by known therapeutic targets as seen in drug therapeutic classifications. Since the only therapeutic classification for the drugs used in this study is cancer, it follows that there is one large cluster containing a majority of the CT drugs and also some NPs, indicating interest in NPs for cancer research. If this analysis was expanded to include drugs from other therapeutic classifications the other communities might also contain the associated drugs, along with the NPs.

Largest Community (32 CT drugs, 11 NPs)	# Targets	Second Largest Community (4 CT drugs, 39 NPs)	# Targets	Third Largest Community (0 CT drugs, 30 NPs)	# Targets
Type XIII RTKs: Ephrin receptor family	12	4.2.1.1 Carbonate dehydratases	3	5-Hydroxytryptamine receptors	11
Src family	11	CYP1 family	3	Adrenoceptors	6
Tec family	5	1.-.-.- Oxidoreductases	1	Ionotropic glutamate receptors	4
Type III RTKs: PDGFR, CSFR, Kit, FLT3 receptor family	5	1.13.11.- Dioxygenases	1	Melatonin receptors	2
Death-associated kinase (DAPK) family	4	ABCC subfamily	1	Acetylcholine receptors (muscarinic)	1
HIPK subfamily	4	Aryl hydrocarbon receptor complex	1	Adenosine receptors	1
Janus kinase (JakA) family	4	Carrier proteins	1	CYP2 family	1
KHS subfamily	4	CFTR	1	Dopamine receptors	1
Myosin Light Chain Kinase (MLCK) family	4	Cyclooxygenase	1	Glucagon receptor family	1
RSK subfamily	4	CYP11, CYP17, CYP19, CYP20 and CYP21 families	1	Glutamate transporter subfamily	1
Type I RTKs: ErbB (epidermal growth factor) receptor family	4	Nucleoside synthesis and metabolism	1	Metabotropic glutamate receptors	1

Table 6. **Natural Products (NP) and Approved Cancer Drugs (CT) Interact with Disjoint and Shared Target Sets.** The top 11 target families in the three largest compound-compound network communities are shown, along with the NP/CT distribution in each. The largest community is dominated by approved cancer drugs and the other two communities are dominated by NPs.

3.6 Some Features from In-Silico Drug Synergy Methods Can Separate NP/Drug Labeled Synergistic Pairs from Unlabeled Pairs

Two unsupervised datasets were created from literature curation. One dataset had 36 labeled synergistic pairs combined with 179 unlabeled pairs for a total of 215 pairs. The second dataset had 11 labeled pairs and 45 unlabeled pairs for a total of 56 pairs. The second dataset was a subset of the first and was restricted by available gene expression data in Connectivity Map, since gene expression profiles were used in both models. All natural products, in both datasets, had to be found in Connectivity Map. Since not all features used in these models require gene

expression data, the larger dataset was used to test the features not dependent on these profiles. These datasets were used to do bivariate tests (Wilcoxon rank sum) on nine features from the two methods.

We found significant differences ($p<.05$) in five of the eight features (**Fig 10)**. The feature derived from gene expression profiles was not tested since there was no overlap in the dataset with 56 pairs. The direction of the difference for mutual information entropy and Tanimoto coefficient was not consistent with that found for drug pairs in the RACS evaluation, but for the efficacy measures it was consistent.

When manifold ranking from RACS was applied to the core model we did not find enrichment of synergistic pairs in the top twenty ranked pairs. This could be due to the weakness of the truth datasets, or that 26 drug pairs were used as the 'bait' in the model. These might not be appropriate for identifying synergy between natural product/drug pairs.

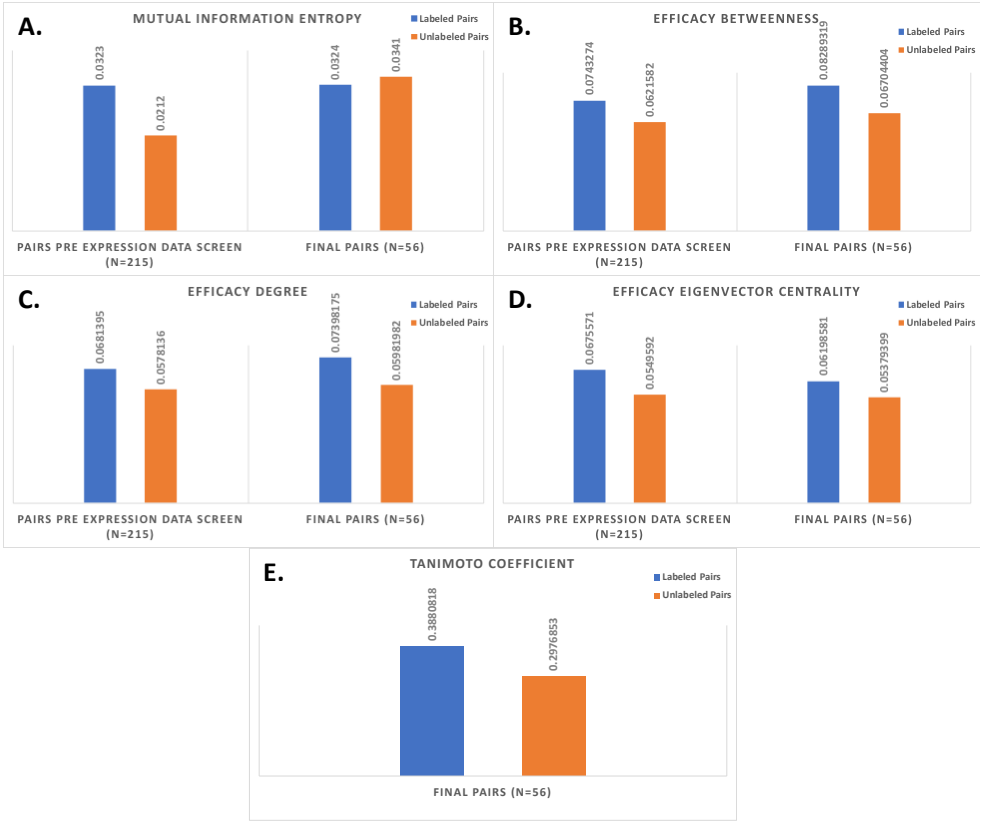

Figure 10. Significantly different features from both methods tested. A. The larger dataset was significant, finding more overlap in biological function for synergistic pairs. **B.** The larger dataset was significantly different with synergistic pairs having a larger value, indicating a greater difference between cancer targets and non-cancer targets for betweenness. **C.** Efficacy for degree had the result as efficacy for betweenness. **D.** Efficacy for eigenvector centrality was significant for the smaller dataset, again with larger values found for synergistic pairs. **E.** Synergistic pairs were found to have greater structural similarity.

4. Discussion.

The intent of this work was not only to assess the number of new potential therapeutic targets available when considering publicly available NP information, but also to assess these targets from a complementary perspective with FDA approved cancer drugs. Most of this analysis only considered high affinity target interactions. At these levels, for the cancer pathways of interest, the increase in number of pathways targeted only by NPs was a small percentage of the total pathways, but these 24 pathways are involved in multiple cancers. Within cancer pathways targeted by both NPs and CT drugs, there is a substantial percentage of targets interacting with only NPs, supporting the potential for new novel combination therapies in cancer. We did see a large increase in both pathway and target coverage when considering all evidence levels, and our random sampling indicates that up to a third of the Level I interactions presented here are accompanied by binding values in the literature. While these binding values can be expected to fall across a wide range, our random sample would indicate that approximately 10% should be less than 100nM. For natural products, we see that the number of targets per compound depends on the evidence level used, and is not currently comparable to what is known for CT drugs. Based on what is not captured in databases but does exist in supporting literature, and also the potential for untested targets, there is a need to focus on both areas to increase knowledge of the NP target space.

It is also important to consider and compare these two target spaces in a biological context. We used two networks to do this, but the Reactome network could be considered more comprehensive than the cancer-specific network. In the Reactome network the new targets identified by considering NPs were similar to drug-targeted nodes, and less similar to non-targeted nodes, when considering betweenness and degree. This similarity was not seen for eigenvector centrality. Degree and betweenness are strong measures of criticality in a network and the fact that the new NP targets are similar to CT drug targets could indicate potential for novel combination therapies. Also of interest is that the nodes targeted by both NPs and CT drugs tended to have higher average values for these measures than either CT drug only or NP only targets. (although this was not statistically significant, is that correct. Again curious about p value is it < 0.20? if so may want to put it in)

It is likely that the research for NP targets that are not targeted by CT drugs is motivated by other drug therapeutic categories, which could also be investigated for cancer therapies. In this study, we have not investigated non-cancer drugs that interact with the targets found to interact only with NPs. But the fact that we saw compound clustering around target families common to NPs and CT drugs, and also clustering around families unique to NPs, when compared to CT drugs, could offer more support for complementary therapeutic relationships between the two compound classifications.

From the perspective of combination therapy and synergy discovery, the focus of this research was on a pharmacodynamic framework. There is a large body of literature for predictive algorithms for combination therapies that is based on network pharmacology, which influenced this work. But NPs can interact with other compounds, either synergistically, additively or antagonistically, through other mechanisms. These can include pharmacokinetic interactions,

interactions with drug efflux transporters and the cell microenvironment. Some of the synergy effects of NPs are not necessarily mediated by direct interaction with a protein target, but also through the up or down regulation of the expression of the protein of interest. Since binding assay results do not necessarily provide functional information about the compound-target interaction, the presence of a strong binding assay result does not necessarily indicate a positive therapeutic effect for cancer. In fact, some of the interactions could promote tumor growth, as is the case in some cancers with the phorbol esters isolated from croton oil. Future efforts should consider both the strength of compound-target binding and the functional effect of the interaction. All of these mechanisms could be considered in future algorithms to predict combination therapies. Briefly, how would you do this in a future study?

A limitation of this study was the quality of publicly available resources for natural product–target interaction data. The two databases chosen are not the only ones available for NPs, but they were appropriate for our NP criteria. The heavy reliance on text mining could explain some of the problems. Some of these data sources appear to contain drugs that do not have a NP source, and also compounds that are apparently used in the extraction of these NP compounds. For this reason, we removed any compounds with a synonym match to drugs using FDA and DrugBank compound lists [Methods]. We also manually curated our high affinity natural product compounds and removed about 11% due to quality concerns. Another issue encountered with these data sources was the inability to download the entire database. There are well curated proprietary natural product databases, such as NapAlert and The Dictionary of Natural Products, for future application of these methods and findings beyond our current analyses including only public resources.

In our analysis, it seemed that some of the natural product targets we reviewed, that were not shared by CT drugs, might have been motivated by other drug therapeutic classifications. We also restricted our natural products to plants and fungi only used in medicinal traditions. There are many more NP compounds in publicly available data sources, and more NP assay data in recently created data sources than we were able to retrieve (NPASS) [66]. In addition, it appears that the large amount of Level I data could be reviewed to retrieve more assay data. For this reason, it would be important to expand this analysis to include all drugs and all known NP compounds available, still within the context of cancer therapies. This would allow for a greater search space for combinations therapies and repurposing opportunities.

I recommend adding a conclusion (short paragraph). Summarize main findings in 2 sentences and then 2-3 sentences of next steps.

References:

1. *National Cancer Institute, Types of Treatment*. Available from: https://www.cancer.gov/about-cancer/treatment/types.
2. Druker, B.J., *Perspectives on the development of imatinib and the future of cancer research.* Nat Med, 2009. **15**(10): p. 1149-52.
3. Holohan, C., et al., *Cancer drug resistance: an evolving paradigm.* Nat Rev Cancer, 2013. **13**(10): p. 714-26.
4. Housman, G., et al., *Drug resistance in cancer: an overview.* Cancers (Basel), 2014. **6**(3): p. 1769-92.
5. Al-Lazikani, B., U. Banerji, and P. Workman, *Combinatorial drug therapy for cancer in the post-genomic era.* Nat Biotechnol, 2012. **30**(7): p. 679-92.
6. Hopkins, A.L., *Network pharmacology: the next paradigm in drug discovery.* Nat Chem Biol, 2008. **4**(11): p. 682-90.
7. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
8. Hu, Y., D. Gupta-Ostermann, and J. Bajorath, *Exploring compound promiscuity patterns and multi-target activity spaces.* Comput Struct Biotechnol J, 2014. **9**: p. e201401003.
9. Jia, J., et al., *Mechanisms of drug combinations: interaction and network perspectives.* Nat Rev Drug Discov, 2009. **8**(2): p. 111-28.

10. Palmer, A.C. and P.K. Sorger, *Combination Cancer Therapy Can Confer Benefit via Patient-to-Patient Variability without Drug Additivity or Synergy.* Cell, 2017. **171**(7): p. 1678-1691.e13.
11. Sun, Y., et al., *Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer.* Nat Commun, 2015. **6**: p. 8481.
12. Liu, Y. and H. Zhao, *Predicting synergistic effects between compounds through their structural similarity and effects on transcriptomes.* Bioinformatics, 2016. **32**(24): p. 3782-3789.
13. Milshteyn, A., J.S. Schneider, and S.F. Brady, *Mining the metabiome: identifying novel natural products from microbial communities.* Chem Biol, 2014. **21**(9): p. 1211-23.
14. Harvey, A.L., R. Edrada-Ebel, and R.J. Quinn, *The re-emergence of natural products for drug discovery in the genomics era.* Nat Rev Drug Discov, 2015. **14**(2): p. 111-29.
15. Newman, D.J. and G.M. Cragg, *Natural Products as Sources of New Drugs from 1981 to 2014.* J Nat Prod, 2016. **79**(3): p. 629-61.
16. Cheng, Y.T., C.C. Yang, and L.F. Shyur, *Phytomedicine-Modulating oxidative stress and the tumor microenvironment for cancer therapy.* Pharmacol Res, 2016. **114**: p. 128-143.
17. Cote, B., et al., *Combinatorial resveratrol and quercetin polymeric micelles mitigate doxorubicin induced cardiotoxicity in vitro and in vivo.* J Control Release, 2015. **213**: p. 128-133.
18. Pearson, H.E., et al., *Overcoming resistance to cetuximab with honokiol, a small-molecule polyphenol.* Mol Cancer Ther, 2017.
19. Blucher, A.S., et al., *Evidence-Based Precision Oncology with the Cancer Targetome.* Trends Pharmacol Sci, 2017. **38**(12): p. 1085-1099.
20. *The Noun Project*. Available from: https://thenounproject.com.
21. Hu, R., et al., *TarNet: An Evidence-Based Database for Natural Medicine Research.* PLoS One, 2016. **11**(6): p. e0157222.
22. Xue, R., et al., *TCMID: Traditional Chinese Medicine integrative database for herb molecular mechanism analysis.* Nucleic Acids Res, 2013. **41**(Database issue): p. D1089-95.
23. Law, V., et al., *DrugBank 4.0: shedding new light on drug metabolism.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1091-7.
24. Qin, C., et al., *Therapeutic target database update 2014: a resource for targeted therapeutics.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1118-23.
25. Harding, S.D., et al., *The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY.* Nucleic Acids Res, 2017.
26. Gu, J., et al., *Use of natural products as chemical library for drug discovery and network pharmacology.* PLoS One, 2013. **8**(4): p. e62839.
27. *Traditional Chinese Medicine Integrated Database*. 2018; Available from: http://www.megabionet.org/tcmid/.
28. Wohlgemuth, G., et al., *The Chemical Translation Service--a web-based tool to improve standardization of metabolomic reports.* Bioinformatics, 2010. **26**(20): p. 2647-8.

29. Gilson, M.K., et al., *BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology.* Nucleic Acids Res, 2016. **44**(D1): p. D1045-53.
30. Paolini, G.V., et al., *Global mapping of pharmacological space.* Nat Biotechnol, 2006. **24**(7): p. 805-15.
31. Wang, Y., et al., *Evidence-Based and Quantitative Prioritization of Tool Compounds in Phenotypic Drug Discovery.* Cell Chem Biol, 2016. **23**(7): p. 862-74.
32. Fabregat, A., et al., *The Reactome Pathway Knowledgebase.* Nucleic Acids Res, 2017.
33. David Tamborero, C.R.-P., Jordi Deu-Pons, Michael P. Schroeder, Ana Vivancos, Ana Rovira, Ignasi Tusquets, Joan Albanell, Jordi Rodon, Josep Tabernero, Carmen de Torres, Rodrigo Dienstmann, Abel Gonzalez-Perez, Nuria Lopez-Bigas, *Cancer Genome Interpreter Annotates The Biological And Clinical Relevance Of Tumor Alterations.* 2017.
34. Benjamini, Y., & Yekutieli, D., *The control of the false discovery rate in multiple testing under dependency. .* Annals of Statistics, 2001. **29**(4): p. 1165-1188.
35. Peng, Q. and N.J. Schork, *Utility of network integrity methods in therapeutic target identification.* Front Genet, 2014. **5**: p. 12.
36. Wu, G., X. Feng, and L. Stein, *A human functional protein interaction network and its application to cancer data analysis.* Genome Biol, 2010. **11**(5): p. R53.
37. Wang, E. *Human Signaling Network*. Available from: http://www.bri.nrc.ca/wang/.
38. Cui, Q., et al., *A map of human cancer signaling.* Mol Syst Biol, 2007. **3**: p. 152.
39. Awan, A., et al., *Regulatory network motifs and hotspots of cancer genes in a mammalian cellular signalling network.* IET Syst Biol, 2007. **1**(5): p. 292-7.
40. Li, L., et al., *The human phosphotyrosine signaling network: evolution and hotspots of hijacking in cancer.* Genome Res, 2012. **22**(7): p. 1222-30.
41. Newman, R.H., et al., *Construction of human activity-based phosphorylation networks.* Mol Syst Biol, 2013. **9**: p. 655.
42. Chatr-Aryamontri, A., et al., *The BioGRID interaction database: 2015 update.* Nucleic Acids Res, 2015. **43**(Database issue): p. D470-8.
43. Yildirim, M.A., et al., *Drug-target network.* Nat Biotechnol, 2007. **25**(10): p. 1119-26.
44. Chen, D., et al., *Synergy evaluation by a pathway-pathway interaction network: a new way to predict drug combination.* Mol Biosyst, 2016. **12**(2): p. 614-23.
45. Yang, Z., R. Algesheimer, and C.J. Tessone, *A Comparative Analysis of Community Detection Algorithms on Artificial Networks.* Sci Rep, 2016. **6**: p. 30750.
46. Ceol, A., et al., *MINT, the molecular interaction database: 2009 update.* Nucleic Acids Res, 2010. **38**(Database issue): p. D532-9.
47. Keshava Prasad, T.S., et al., *Human Protein Reference Database--2009 update.* Nucleic Acids Res, 2009. **37**(Database issue): p. D767-72.
48. Xenarios, I., et al., *DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions.* Nucleic Acids Res, 2002. **30**(1): p. 303-5.
49. Goel, G., et al., *Phorbol esters: structure, biological activity, and toxicity in animals.* Int J Toxicol, 2007. **26**(4): p. 279-88.
50. Ntie-Kang, F., et al., *AfroDb: a select highly potent and diverse natural product library from African medicinal plants.* PLoS One, 2013. **8**(10): p. e78085.

51. Tse, K.Y., et al., *Patient compliance with cervical smear surveillance in a shared-care setting.* Int J Gynaecol Obstet, 2016.
52. Winterbach, W., et al., *Topology of molecular interaction networks.* BMC Syst Biol, 2013. **7**: p. 90.
53. Zhu, X., M. Gerstein, and M. Snyder, *Getting connected: analysis and principles of biological networks.* Genes Dev, 2007. **21**(9): p. 1010-24.
54. Pavlopoulos, G.A., et al., *Using graph theory to analyze biological networks.* BioData Min, 2011. **4**: p. 10.
55. Arrell, D.K. and A. Terzic, *Network systems biology for drug discovery.* Clin Pharmacol Ther, 2010. **88**(1): p. 120-5.
56. Hwang, W.C., A. Zhang, and M. Ramanathan, *Identification of information flow-modulating drug targets: a novel bridging paradigm for drug discovery.* Clin Pharmacol Ther, 2008. **84**(5): p. 563-72.
57. Laderas, T.G., L.M. Heiser, and K. Sonmez, *A Network-Based Model of Oncogenic Collaboration for Prediction of Drug Sensitivity.* Front Genet, 2015. **6**: p. 341.
58. Sun, J. and Z. Zhao, *A comparative study of cancer proteins in the human protein-protein interaction network.* BMC Genomics, 2010. **11 Suppl 3**: p. S5.
59. Xia, J., et al., *Do cancer proteins really interact strongly in the human protein-protein interaction network?* Comput Biol Chem, 2011. **35**(3): p. 121-5.
60. Yin, N., et al., *Synergistic and antagonistic drug combinations depend on network topology.* PLoS One, 2014. **9**(4): p. e93960.
61. Li, S., B. Zhang, and N. Zhang, *Network target for screening synergistic drug combinations with application to traditional Chinese medicine.* BMC Syst Biol, 2011. **5 Suppl 1**: p. S10.
62. Huang, L., et al., *DrugComboRanker: drug combination discovery based on target network analysis.* Bioinformatics, 2014. **30**(12): p. i228-36.
63. Jaeger, S., et al., *Quantification of Pathway Cross-talk Reveals Novel Synergistic Drug Combinations for Breast Cancer.* Cancer Res, 2017. **77**(2): p. 459-469.
64. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.
65. Hu, Y. and J. Bajorath, *High-resolution view of compound promiscuity.* F1000Res, 2013. **2**: p. 144.
66. Zeng, X., et al., *NPASS: natural product activity and species source database for natural product research, discovery and tool development.* Nucleic Acids Res, 2017.
a.