PREDICTIVE CHEMINFORMATICS ANALYSIS OF DIVERSE CHEMOGENOMICS DATA SOURCES: APPLICATIONS TO DRUG DISCOVERY, ASSAY INTERFERENCE, AND TEXT MINING

Stephen Joseph Capuzzi

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Division of Chemical Biology and Medicinal Chemistry in the Pharmaceutical Sciences Department in the Eshelman School of Pharmacy

Chapel Hill
2018

Approved by:

Alexander Tropsha

Stephen V. Frye

Albert A. Bowers

Nikolay Dokholyan

Dmitri Kireev

© 2018
Stephen Joseph Capuzzi
ALL RIGHTS RESERVED

ABSTRACT

Stephen Joseph Capuzzi: Predictive Cheminformatics Analysis of Diverse Chemogenomics Data Sources: Applications to drug discovery, assay interference, and text mining.
(Under the direction of Alexander Tropsha)

In this dissertation, we describe the cheminformatics analysis of diverse chemogenomics data sources as well as the application of these data to several drug discovery efforts. In Chapter 1, we describe the discovery and characterization of novel Ebola virus inhibitors through QSAR-based virtual screening. In Chapter 2, we report the discovery and analysis of a series of potent and selective doublecortin-like kinase 1 (DCLK1) inhibitors using QSAR modeling, virtual screening, Matched Molecular Pair Analysis (MMPA), and molecular docking. In Chapter 3, we performed a large-scale analysis of publicly available data in PubChem to probe the reliability and applicability of **P**an-**A**ssay **IN**terference compound**S** (PAINS) alerts, a popular computational drug screening tool. In Chapter 4, we explore the PubMed database as a novel source of biomedical data and describe the development of Chemotext, a publicly available web server capable of text-mining the published literature.

To my mother.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Alexander Tropsha. As Markovnikov wrote to Butlerov: “Считаю приличным посвятить небольшой труд свой Вам, многоуважаемый наставник, поскольку проводимые в нём мысли есть дальнейшее развитие установленного Вами….Если в нём и заключается что-нибудь новое, то рождение этого невозможно было бы без исходных положений, заложенных Вами.”

I would also like to thank the various members of the Molecular Modeling Laboratory, especially Dr. Eugene Muratov.

To my committee members, thank you all for your guidance, support and insights.

I would also like to thank my friends and family – too many to name.

Lastly, I would like to acknowledge my funding sources: NIH, NSF, DoD, USP, and EII.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AD	APPLICABILITY DOMAIN
AJAX	ASYNCHRONOUS JAVASCRIPT AND XML
Bla	BETA-LACTAMASE
BSL	BIOSAFETY LEVEL
CADD	COMPUTER-AIDED DRUG DESIGN
CCR	CORRECT CLASSIFICATION RATE
COP	CLINICAL OUTCOME PATHWAYS
DCLK1	DOUBLECORTIN-LIKE KINASE 1
DCM	DARK CHEMICAL MATTER
DTD	DRUG-TARGET-DISEASE
EBOV	EBOLA VIRUS
EV	ENRICHMENT VALUE
FDA	FOOD AND DRUG ADMINISTRATION
FH-NoPAINS	FREQUENT HITTERS – NO PAINS
FH-PAINS	FREQUENT HITTERS - PAINS
FN	FALSE NEGATIVES
FP	FALSE POSITIVES
HCMV	HUMAN CYTOMEGALOVIRUS
HTS	HIGH-THROUGHPUT SCREENING
IH-NoPAINS	INFREQUENT HITTERS- NO PAINS
IH-PAINS	INFREQUENT HITTERS - PAINS
MeSH terms	MEDLINE SUBJECT HEADING
MLT	MACHINE LEARNING TECHNIQUE
MML	MOLECULAR MODELING LABORATORY
MMP	MATCHED MOLECULAR PAIRS

MOA	MECHANISM OF ACTION
MODI	MODELABILITY INDEX
NIH	NATIONAL INSTITUTES OF HEALTH
NIH	NATIONAL INSTITUTES OF HEALTH
NLM	NATIONAL LIBRARY OF MEDICINE
NPC1	NIEMANN PICK C1
NPV	NEGATIVE PREDICTIVE VALUE
PAINS	PAN-ASSAY INTERFERENCE COMPOUNDS
PMID	PUBMED IDENTIFICATION
PPV	POSITIVE PREDICTIVE VALUE
RF	RANDOM FOREST QSAR
RTK	RECEPTOR TYROSINE KINASE
SE	SENSITIVITY
SI	SELECTIVITY INDEX
SI(65)	SELECTIVITY INDEX AT 65%
SI(90)	SELECTIVITY INDEX AT 90%
SLN	SYBYL LINE NOTATION
SP	SPECIFICITY
Tc	TANIMOTO COEFFICIENT
TN	TRUE NEGATIVES
TP	TRUE POSITIVES
TSC	TUMOR STEM CELL
VLP	VIRUS-LIKE PARTICLE
VS	VIRTUAL SCREENING

CHAPTER 1: COMPUTER-AIDED DISCOVERY AND CHARACTERIZATION OF NOVEL EBOLA VIRUS INHIBITORS[1]

1.2 INTRODUCTION

The 2014 Ebola outbreak was the largest and most persistent since the discovery of the Ebola virus (EBOV) in 1976. Alarmingly, a new EBOV outbreak was confirmed in the Democratic Republic of Congo in May 2017.[1] Though advances in the research and development of Ebola therapeutics have been made,[2–4] Ebola drug discovery endeavors are hindered due to the high virulence of the EBOV and its biosafety level 4 (BSL-4) classification.[5] Recently, a biosafety level 2 (BSL-2) Ebola virus-like particle (VLP) entry assay was developed and utilized for a drug repurposing screen of Food and Drug Administration (FDA)-approved drugs.[6–8] The Ebola VLP contains glycoprotein (GP) and the matrix protein VP40 fused to a beta-lactamase reporter for monitoring of VLP entry into cells. Although this BSL-2 Ebola VLP assay enables rapid compound screening, it requires a centrifugation step for assay plates at 1,500 g for 45 minutes at 4 °C that limits its screening throughput. Computational approaches that leverage generated data can be used to design or select small sets of compounds for lead identification in order to reduce the time and costs of high throughput screening. Using the existing data from the Ebola VLP entry assay as well as cytotoxicity data, QSAR models[9] can be built and then employed for virtual screening of large chemical libraries to predict activeompounds against EBOV infection with low

[1] This chapter previously appeared as an article in the Journal of Medicinal Chemistry. The original citation is as follows: Capuzzi, SJ, et. al. "Computer-aided discovery and characterization of novel Ebola virus inhibitors". *Journal of Medicinal Chemistry.* Just Accepted (April, 2018).

expected toxicity. Indeed, QSAR modeling approaches have been previously employed for identification of compounds with efficacy against EBOV.[10,11] Herein, we describe a study that relied on synergistic combination of statistical data modeling and experimental testing for both antiviral inhibitor potency and host cell cytotoxicity (**Figure 1.1**). Our study utilized both BSL-2 and BSL-4 assays to experimentally validate hits identified computationally.

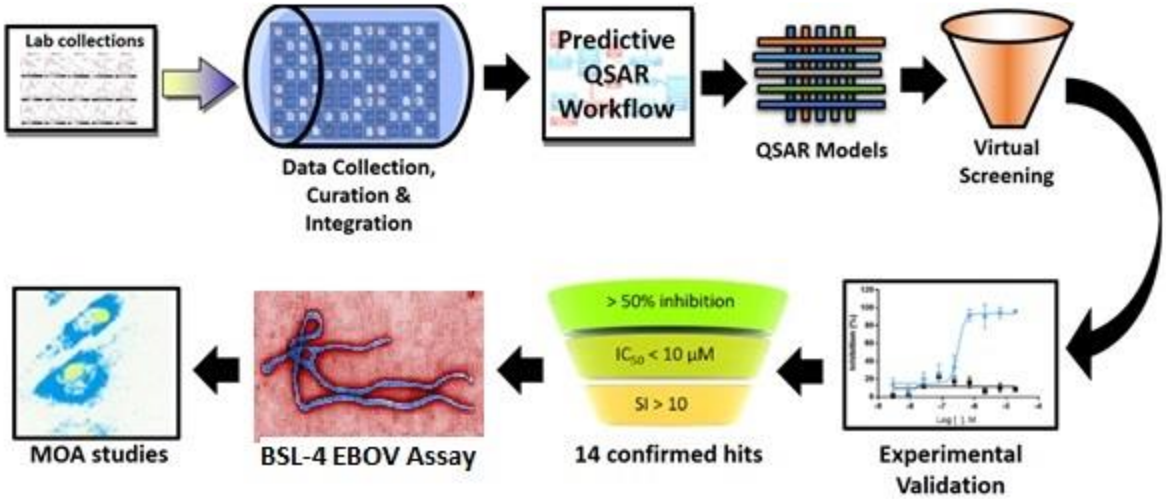


Figure 1.1. Overall study design. The present study synergistically incorporates computational modeling and experimentation.

To identify compounds with anti-EBOV activity and limited host cell cytotoxicity, we designed an integrated QSAR modeling system for virtual compound screening that is combined with experimental testing on a focused set of predicted compounds. In this study, existing antiviral activity and compound cytotoxicity data were collected and carefully curated; respective QSAR models were built and rigorously validated; these models were employed for virtual screening of a large chemical library (~17 million compounds), resulting in 102 hits prioritized for experimental testing; the anti-EBOV activity in the Ebola VLP assay and cytotoxicity in host cells of these hits were determined experimentally in BSL-2 and BSL-4 assays; and the mechanisms of anti-EBOV activity for confirmed hits were identified. Ultimately, 14 potent hits with activity ranging between 0.272 μM and 9.65 μM as well as more than 10-fold selectivity over compound cytotoxicity in host cells were confirmed. Next, five selected hits were shown to inhibit BSL-4 live-EBOV infection in a dose-dependent manner. Two of these hits possessed novel scaffolds, making them candidates for further medicinal chemistry optimization as potential anti-EBOV agents. This study presents the first example of computationally-driven prioritization and experimental discovery of novel potent anti-Ebola compounds with high therapeutic windows in the published literature.

1.3 RESULTS

1.3.1 Model Performance

Prior to the modeling, MODIs of 0.69 and 0.68 were calculated for the P1 and P2 datasets, respectively. For each protocol, three separate software packages (Chembench, HiT QSAR, and GUSAR) employing different descriptors and different machine learning techniques (MLTs) were utilized for model building. In total, six individual models were built and rigorously validated. Results of 5-fold external cross-validation are presented in **Table 1.1**. In order to demonstrate that the predictive power of the models was not due to random correlation between bioactivity and chemical descriptors, 1000 rounds of Y-randomization was performed. No Y-randomized models had a CCR above 0.60.

For P1, models built with HiT QSAR and GUSAR had the highest predictive accuracy, irrespective of the use of different chemical descriptors and MLTs. For P2, HiT QSAR again showed the best performance. Additionally, the CCR of the Chembench model improved by ~7% for P2 over P1. All models were deemed robust and statistically valid (**Table 1.1**).

Table 1.1. Statistical characteristics obtained on 5-fold external CV of all models developed in this study. The results with highest statistical metrics are highlighted in bold. HEK models built with Chembench and HiT QSAR were not used due to poor predictive power. Values below acceptance threshold are underlined.

Model Name	**Descriptors**	**MLT**	**CCR**	**SE**	**SP**	**PPV**	**NPV**
Chembench – P1	Dragon 6.0	RF	0.67	0.69	0.68	0.66	0.68
HiT QSAR – P1	SiRMS	RF	**0.72**	**0.73**	**0.71**	**0.72**	**0.73**
GUSAR – P1	MNA and QNA	SCR-RBF	**0.72**	**0.73**	**0.71**	**0.72**	**0.73**
Chembench – P2	Dragon 6.0	RF	0.72	0.72	0.72	0.72	0.72
HiT QSAR – P2	SiRMS	RF	**0.75**	**0.75**	**0.75**	**0.75**	**0.75**
GUSAR – P2	MNA and QNA	SCR-RBF	0.72	0.69	0.75	0.73	0.71
Chembench – HeLa	Dragon 6.0	RF	0.73	**0.77**	0.68	0.73	**0.72**
HiT QSAR – HeLa	SiRMS	RF	0.64	0.68	0.60	0.66	0.62
GUSAR – HeLa	MNA and QNA	SCR-RBF	**0.75**	0.67	**0.84**	**0.82**	0.69
Chembench – HEK	Dragon 6.0	RF	0.62	0.67	0.57	0.64	0.60
HiT QSAR – HEK	SiRMS	RF	0.53	0.55	0.50	0.56	0.49
GUSAR – HEK	MNA and QNA	SCR-RBF	0.72	0.78	0.71	0.73	0.76

For HeLa and HEK cell lines, MODI of 0.65 and 0.70 were obtained, respectively. For HeLa cytotoxicity, GUSAR yielded the best overall model. Chembench and GUSAR had inverse sensitivity and specificity profiles, indicating that Chembench could better identify toxic compounds, while GUSAR could better identity non-toxic compound. This observation highlights the reciprocal benefit of consensus modeling, *i.e.*, utilizing all the models for VS. No Y-randomized models had a CCR in access of 0.60. All models were deemed robust and statistically valid. For HEK cytotoxicity, GUSAR again proved to be the best overall model. On the other hand, Chembench and HiT QSAR were not statistically validated, as several metrics fell below the 0.60 threshold. Thus, only GUSAR was used for prediction of HEK cytotoxicity. Y-randomized models for GUSAR did not exceeded a CCR of 0.60. A summary of all model performance can be found in **Table 1.1**.

1.3.2 QSAR-Based Virtual Screening

QSAR-based virtual screening (VS) was carried out according to the workflow presented in **Figure 1.2**. Initially, ~17 million compounds (see Methods) were downloaded, prepared, and screened. As previously stated, "hits" were those compounds that were within the AD of the respective model and predicted by all models to have high antiviral activity and limited host cytotoxicity. In total, 102 VS hits were selected for experimental validation in the Ebola-VLP entry assay.

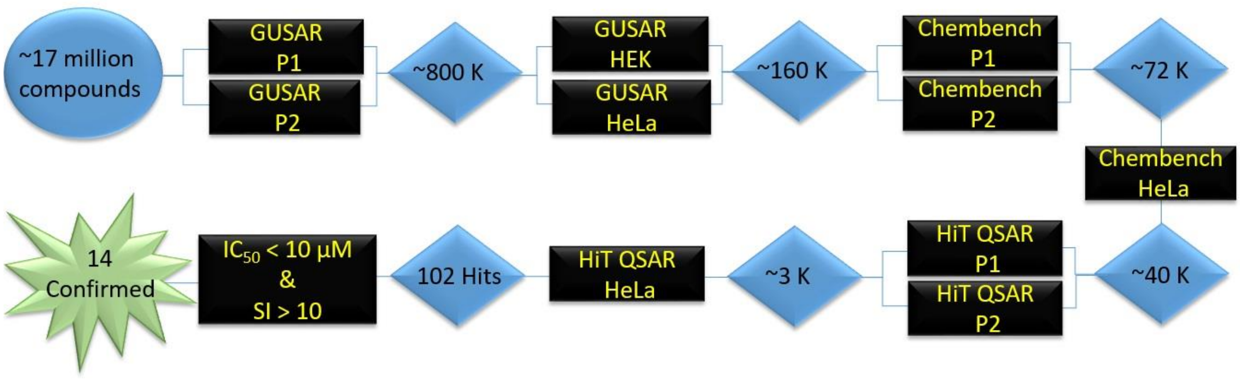


Figure 1.2. **Screening workflow**. A virtual chemical library of ~17 million compounds was screened against a battery of antiviral (P1 and P2) and cytotoxicity (HEK and HeLa) models. Hits selected for experimental validation were predicted to be EBOV inhibitors with limited host cytotoxicity. Then computational hits were experimentally validated, then their activity was evaluated using percent inhibition, IC_{50} values, and selectivity index (SI).

1.3.3 Experimental Evaluation

1.3.3.a Experimental confirmation of Anti-EBOV activity of 14 compounds

Based on the virtual screening results, 102 compounds were purchased and experimentally screened in the Ebola-VLP entry assay in parallel with an ATP content assay to determine compound cytotoxicity in host cells. All compounds were screened at 11 concentration dilutions ranging from 0.001 to 57 µM.[12] Out of 102 compounds tested, 51 showed greater than 50% inhibition, indicating that half of compounds had confirmed anti-EBOV activity. Next, 20 of these 51 compounds exhibited the IC_{50} values under 10 µM. Because the compound cytotoxicity at higher compound concentrations might reduce the Ebola VLP entry in cells, these potential false positive compounds should be deprioritized. After comparing to the compound cytotoxicity data, 14 of these confirmed compounds showed a greater than 10-fold selectivity index (SI) of anti-Ebola VLP entry over compound cytotoxicity. Vindesine and BIX-01294 inhibited the virus in the nanomolar range (**Figure 1.3**).

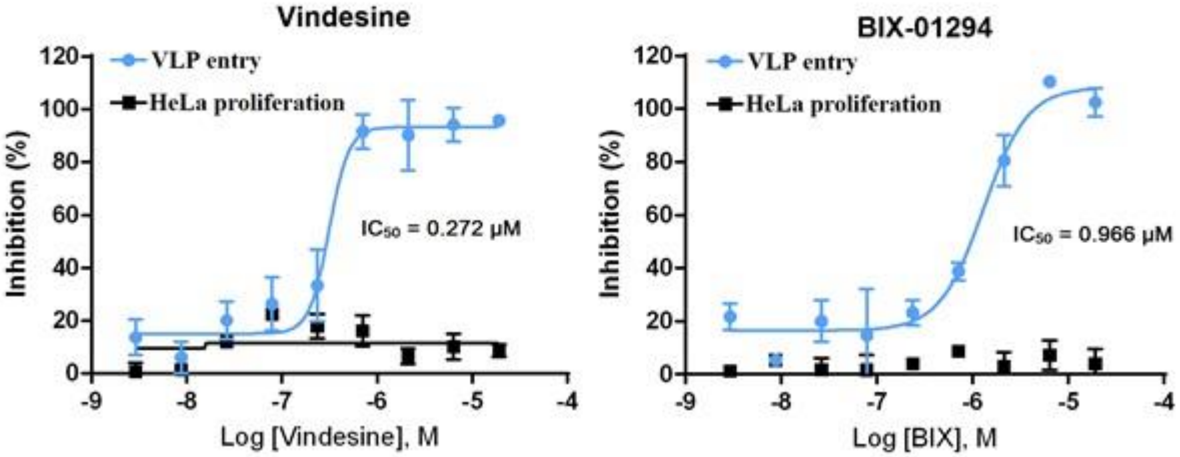


Figure 1.3. Dose response curves for vindesine and BIX-01294. Both antiviral (VLP entry) and host cell cytotoxicity (HeLa) activities are plotted.

Moreover, these 14 confirmed compounds, except for ZINC91973695 and ZINC67869167, have known mechanisms of action (MOAs) and therapeutic indications (**Table 1.2**), including eight anti-cancer, two antihistamines, and two anti-psychotic and anti-inflammatory agents (**Table 1.2**).

Table 1.2. **Experimental results for the top 14 hits.** Most experimentally confirmed hits have known MOAs and therapeutic use indications.

Name	Potency, µM	Selectivity Index	Indication	MOA
Vindesine	0.272	1837	Anticancer	Microtubule Inhibitor
BIX-01294	0.966	45	Anticancer	HMTase Inhibitor
Afimoxifene	1.96	123	Anticancer	Estrogen Receptor Modulator
Tetrandrine	2.16	22	Anti-inflammation	Calcium Channel Blocker
NVP-ADW742	3.05	13	Anticancer	Tyrosine Kinase Inhibitor
Endoxifen	3.05	164	Anticancer	Estrogen Receptor Modulator
ZINC91973695	6.09	82	N/A	N/A
Deptropine	6.58	76	Antihistamine	Anticholinergic
GANT61	6.83	73	Anticancer	Hedgehog Antagonist

ZINC67869167	6.83	73	N/A	N/A
Hh-Ag1.5	7.67	65	Anticancer	Hedgehog Agonist
Cediranib	7.67	65	Anticancer	Tyrosine Kinase Inhibitor
Ebastine	9.56	51	Antihistamine	Histamine H1 Antagonist
Osanetant	9.65	52	Antipsychotic	Neurokinin 3 Receptor Antagonist

The remaining two compounds obtained from the ZINC database (ZINC91973695 and ZINC67869167) have no previously reported bioactivities, anti-Ebola or otherwise. Five hits were further validated in a live EBOV infection assay at bio-safety level-4 (BSL-4). All five hits showed dose-response inhibition against EBOV infection (**Figure 1.4**). Vindesine was the most potent compounds with an IC$_{50}$ of 0.34 µM. The IC$_{50}$ values of NVP-ADW742, BIX-01294, ZINC67869167, and ZINC91973695 were between 1 µM to 10 µM in the live EBOV infection assays.

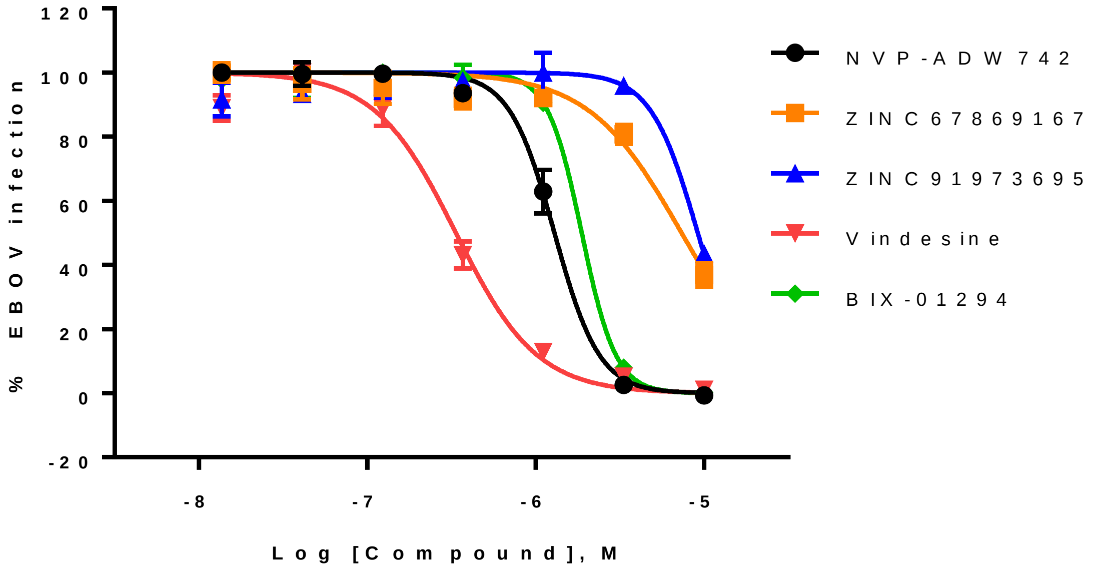


Figure 1.4. Dose-response behavior against BSL-4 live-EBOV infection. Five hits were selected for screening in live-EBOV infection in a BSL-4 assay. Vindesine (red) was the most potent compounds with an IC50 of 0.34 µM. The IC_{50} values of NVP-ADW742 (black), BIX-01294 (green), ZINC67869167 (orange), and ZINC91973695 (blue) were between 1 µM to 10 µM.

1.3.3.b Mechanisms of action against EBOV entry

We probed the chemical biology of these hit compounds in both viral and host systems in order to uncover the mechanisms of anti-EBOV action. We examined the potential sites for drug interaction including Niemann Pick C1 (NPC1) protein, lysosomal function, cathepsin B, and cathepsin L,[13–15] as well as the direct binding of these compounds to the Ebola VLP proteins using thermal shift binding assays.[16] The results of chemical biology studies revealed that these compounds may act via one or more these targets/mechanisms.

The process of EBOV entry into cells involves binding of viral envelop protein(s) to the cell membrane receptor protein/molecule, endocytosis, movement of endocytic vesicles to early/late endosomes and lysosomes, and ejection of viral RNA into the cytosol.[17] Therefore, inhibition of viral protein binding to cell membrane proteins/binding partners can effectively reduce viral entry and subsequent viral replication in cells. Because the cell surface binding protein/molecule for Ebola viral proteins is still unclear, we determined direct binding of these compounds to recombinant Ebola protein. To examine whether these compounds directly interact with the EBOV, their ability of stabilizing Ebola protein was tested in a thermal shift assay using recombinant Ebola VLP. None of the compounds at 50 μM were able to protect Ebola VLP from thermal denaturation. (**Figure 1.5**).

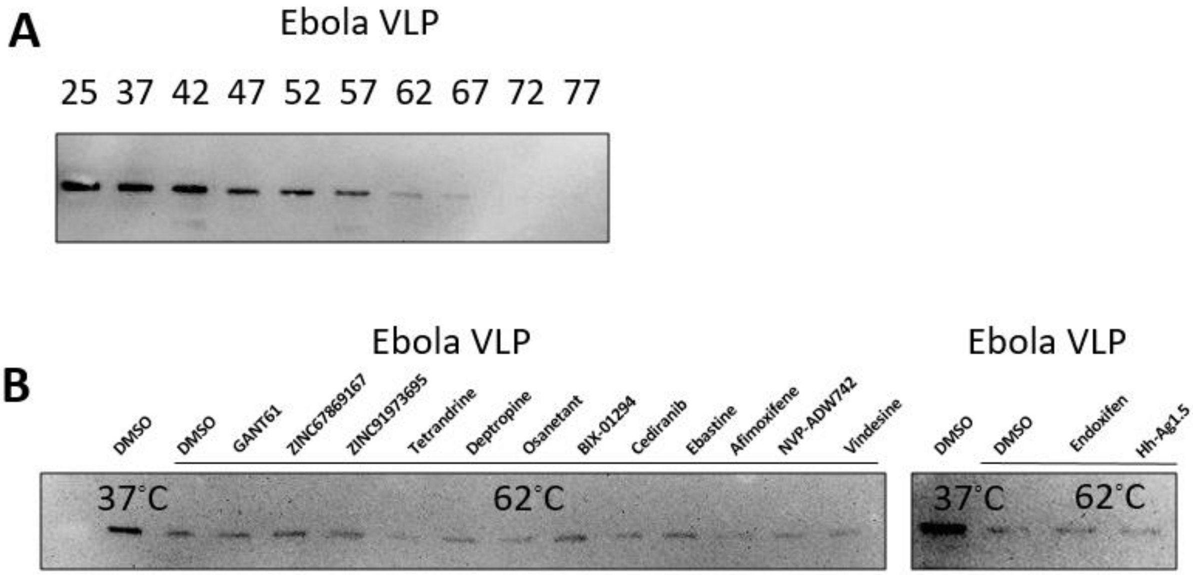


Figure 1.5. Thermal profiling results of Ebola VLP with Ebola entry inhibitors. A, Thermal stability of Ebola VLP at temperatures from 25 °C to 77 °C detected by western blot. **B**, Effects of Ebola entry inhibitors (GANT61, ZINC67869167, ZINC91973695, tetrandrine, deptropine, osanetant, BIX-01294, cediranib, ebastine, afimoxifene, NVP-ADW742, vindesine, endoxifen, Hh-Ag1.5) on thermal stability of Ebola VLP at 62 °C. All experiments were performed in duplicate and data are representative of two independent experiments.

Cathepsin B and L are lysosomal endopeptidases that had been reported to prime EBOV proteins in lysosomes before the viral RNAs are injected into the cytosol for virus replication. Inhibition of cathepsin B or L significantly reduces EBOV infection.[13] GANT61 (an inhibitor of GLI1 and GLI2-induced transcription), deptropine (an antihistamine), and ebastine (an antihistamine) inhibited the enzymatic activity of cathepsin L (**Figure 1.6a and 1.6b**). Only GANT61 inhibited enzymatic activity of cathepsin B (**Figure 1.6c and 1.6d**).

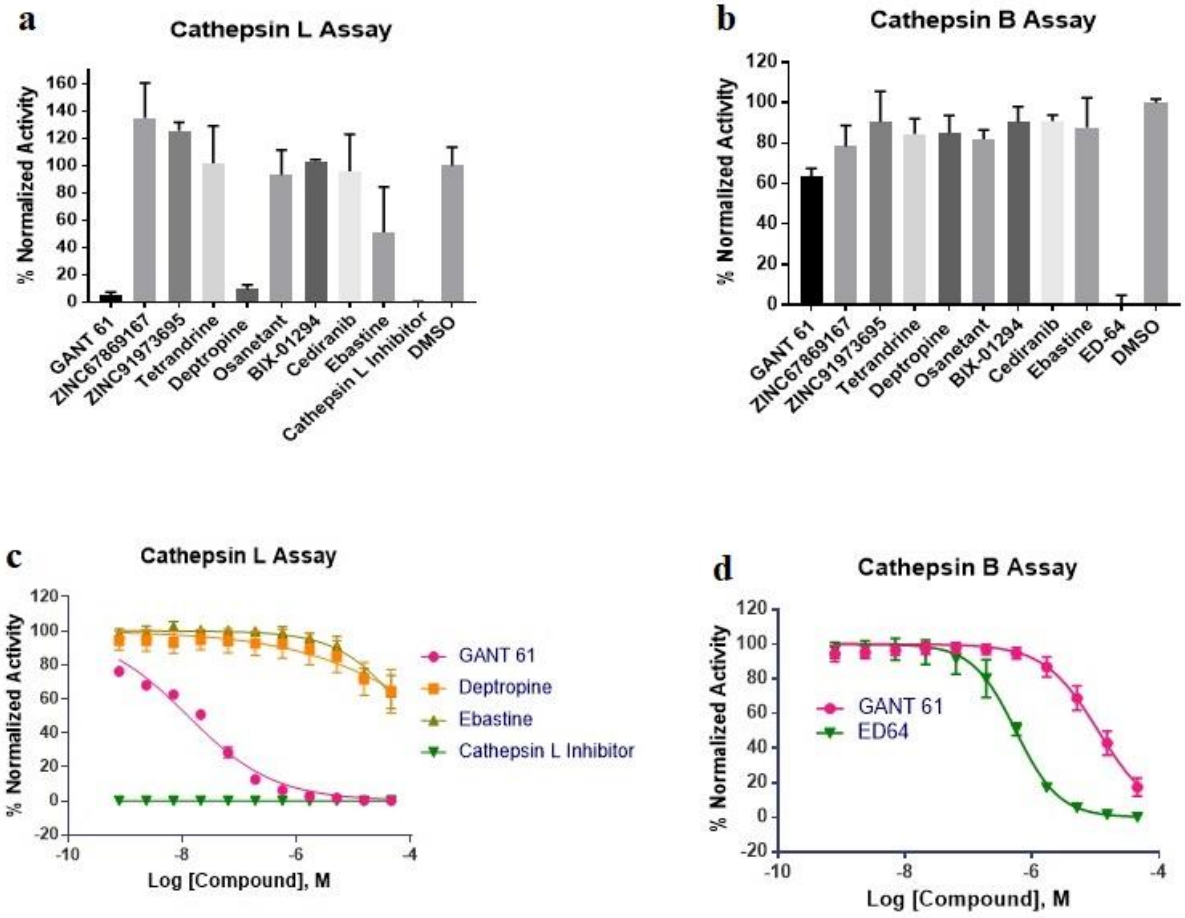


Figure 1.6. Inhibition of protease activities of recombinant cathepsin L and cathepsin B by Ebola entry inhibitors. a and **b**. recombinant cathepsin B or cathepsin L were treated with 10 µM of GANT61, ZINC67869167, ZINC91973695, tetrandrine, deptropine, osanetant, BIX-01294, cediranib, and ebastine. **c**. Dose-response studies of GANT61, deptropine and ebastine in cathepsin

L assay. **d.** Dose-response studies of GANT61 in cathepsin B assay. All experiments were performed in triplicate and data are representative of at least two independent experiments. Data are represented as mean ± SEM.

The NPC1 protein has been reported as an intracellular receptor for EBOV.[14,15,18] Significant reduction of EBOV entry and infection were observed in the NPC1-deficient cells and mouse models.[14,19] Ebastine increased cholesterol accumulation in cells determined by the filipin staining assay, which indicated a functional impairment of NPC1 protein; whereas, the other eight evaluated hits did not impair NPC1 protein function (**Figure 1.7a**).

Lysosomes in cells are enlarged after treatment with certain compounds that damage lysosome functions, resulting in accumulation of lipids and other macromolecules.[20] The enlarged lysosomes are often observed in the patient cells with lysosomal storage diseases caused by mutations in lysosomal proteins and lipid accumulation.[21] EBOV entry is significantly reduced after the lysosomal functions are impaired by compounds. All of the nine evaluated hits increased LysoTracker dye staining in cells, indicating an enlargement in lysosome size (**Figure 1.7b**).

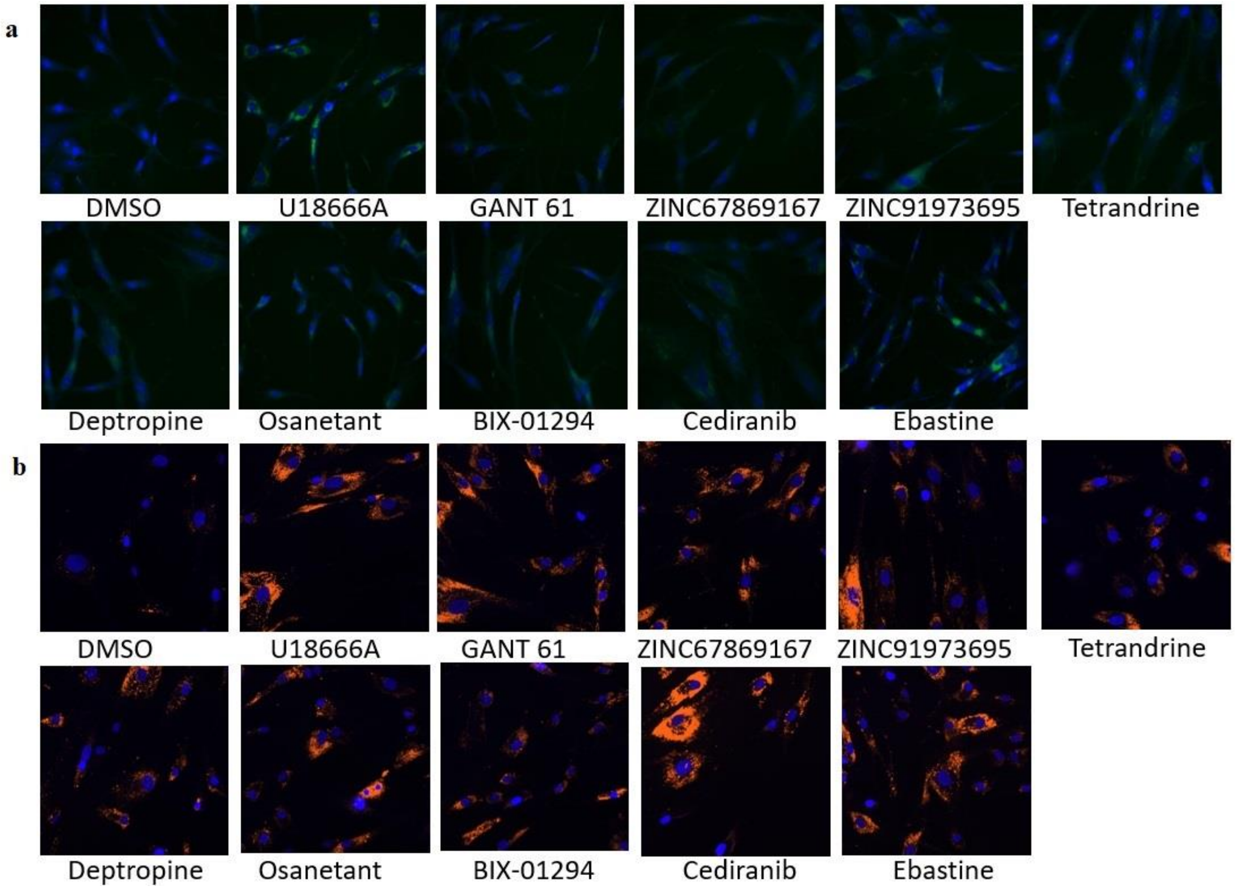


Figure 1.7. Cholesterol accumulation (NPC1 inhibition) and enlargement of lysosome induced by Ebola entry inhibitors. **a.** U18666A and ebastine increased filipin staining in fibroblasts (green: filipin; blue: nuclei). **b**. U1866A, GANT61, ZINC67869167, ZINC91973695, tetrandrine, deptropine, osanetant, BIX-01294, cediranib, and ebastine increased LysoTracker staining in fibroblasts (orange: LysoTracker red; blue: nuclei). All experiments were performed in triplicate and data are representative of at least two independent experiments. Data are represented as mean ± SEM.

1.3.4 Cheminformatics Analysis

1.3.4.a Assay Liabilities

First, using substructural pattern matchers implemented in ZINC15,[22] the 14 experimentally confirmed hits were found to be free of chemical aggregation liabilities[23] and PAINS alerts[24]. Since the assay employed herein relied on a beta-lactamase reporter system, all 14 hits were also checked for potential beta-lactamase inhibition trends using PubChem

Promiscuity.[25] No heightened beta-lactamase assay activity trends were observed, indicating that these hits are not assay artifacts (**Supplementary File 14**).

1.3.4.b Chemical Similarity to Training Set Compounds

Hierarchical clustering analysis revealed that majority of the hits are structurally dissimilar from each other, aside from afimoxifene and endoxifen (**Figure 1.8**).

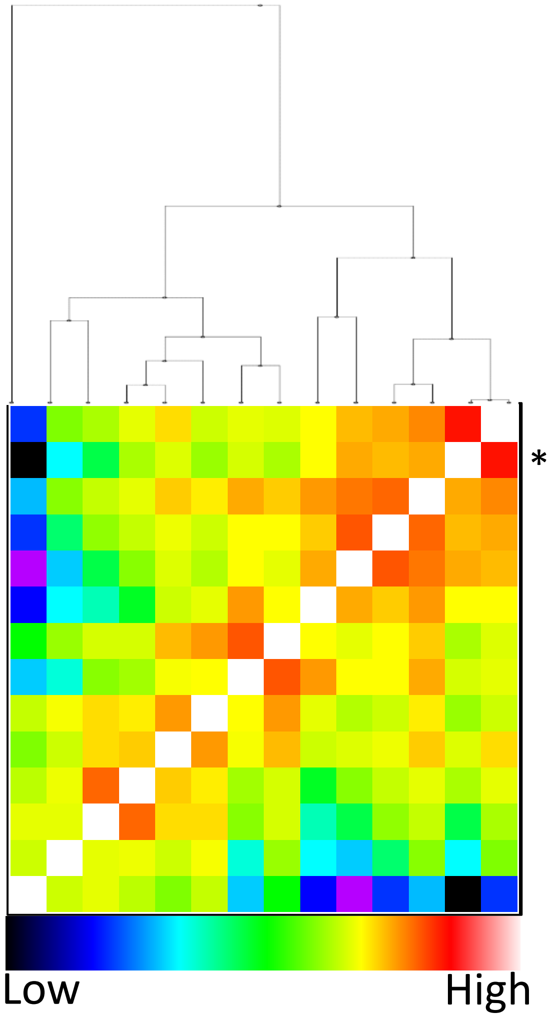


Figure 1.8. Hierarchical clustering of experimental hits. High Tanimoto similarity of afimoxifene and endoxifen is highlighted by an asterisk.

Clustering thus indicates the hits discovered in this study access a wide range of chemical space across several unique chemotypes. The structural similarity based on the Tanimoto coefficient (Tc) of the 14 hits were then compared with compounds in the antiviral training sets (**Table 1.3**) in order to assess the uniqueness of hits. In addition to being highly structurally similar to each other, afimoxifene and endoxifen both have Tc above 0.90 to tamoxifen, which was a previously reported anti-Ebola inhibitor.[26] Likewise, tetrandrine is highly structurally similar (Tc=0.97) to cepharanthine, a training set active compound. The most potent hit in this study, vindesine, had a Tc of 0.96 to vinblastine, which was the most potent hit in the original screen.[26] These hits, while not entirely unique from a chemical perspective, illustrate that the developed QSAR models are robust and that the experimental assays are reproducible. The remaining 10 hits were considerably dissimilar from any training set compounds (Tc = 0.63-0.89), thereby constituting novel anti-Ebola chemotypes as compared to the training set compounds.

1.3.4.c Comparison to Previously Reported EBOV Inhibitors

The potencies and structures of the 14 hits identified in this study were compared to a compiled set of 60 previously published compounds with either *in vitro* or *in vivo* anti-Ebola activity.[2,7,27,28] The full list of previously known published compounds and their potencies can be found in the **Supplementary File 15**.

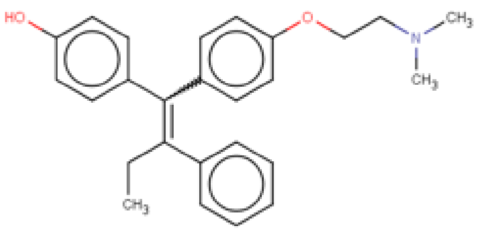
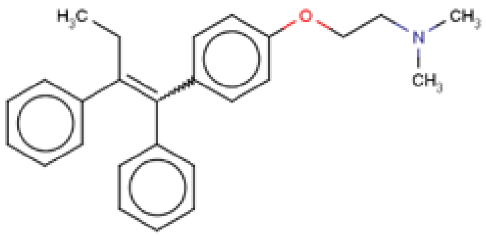
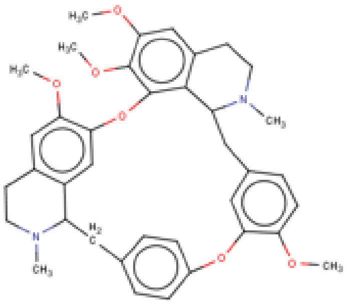
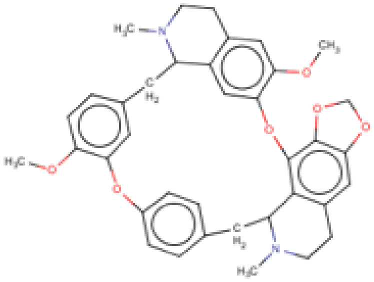
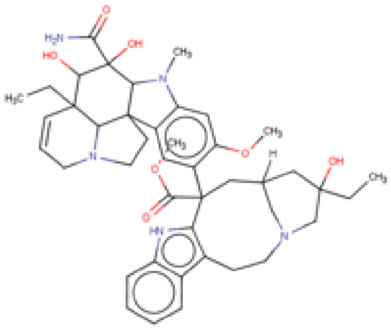
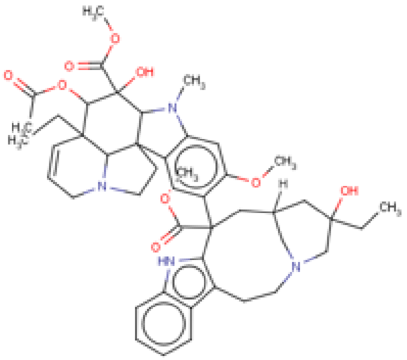
The most potent hit identified from virtual screen was vindesine (0.272 μM), a vinca alkaloid microtubule inhibitor. Previously, other vinca alkaloids were reported as sub-micromolar inhibitors of the EBOV *in vitro*. These vinca alkaloids, vinblastine (0.048 μM), vinorelbine (0.066 μM), and vincristine (0.141 μM),[7] are highly structurally similar to vindesine (**Table 1.4**). Colchicine and nocodazole, microtubule inhibitors that are structurally distinct from the vinca alkaloids, were also previously reported as sub-micromolar inhibitors. The identification of vindesine as one of the most potent hits identified to date highlights the robustness of the developed QSAR models, as well as the efficacy of this class of compounds and compounds with the same associated MOA as viable anti-Ebola compounds.

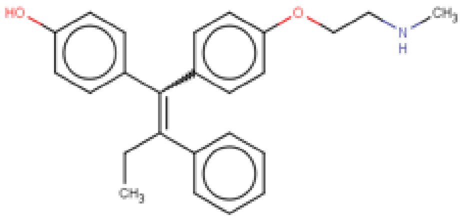
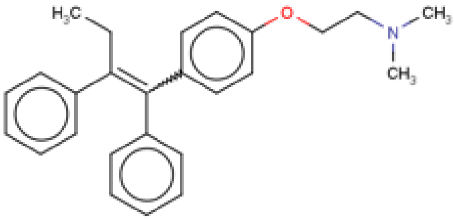
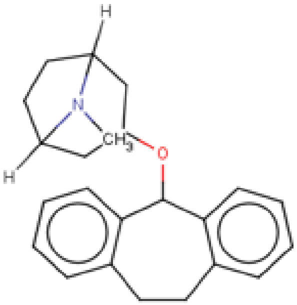
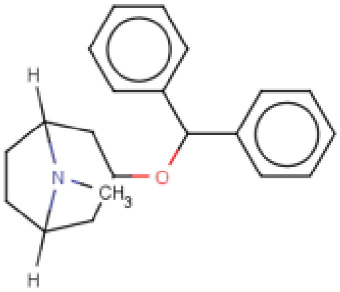
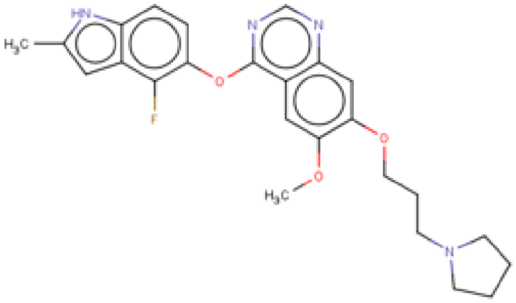
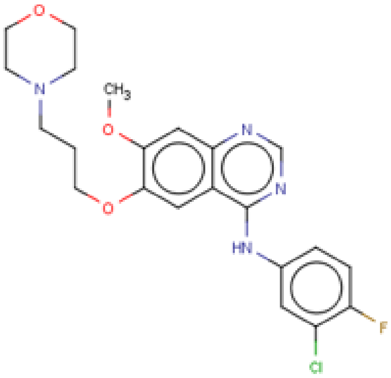
The second most potent hit identified from the virtual screen was BIX-01294 (0.97 μM). This compound is among most potent reported anti-Ebola compounds. Moreover, BIX-01294 is structurally dissimilar from other previously reported compounds (**Table 1.4**) and has a unique primary MOA (G9a histone methyltransferase inhibition).[29] In contrast to vindesine, the identification of BIX-01294 demonstrates the ability to QSAR-based virtual screening to retrieve structurally novel chemotypes.

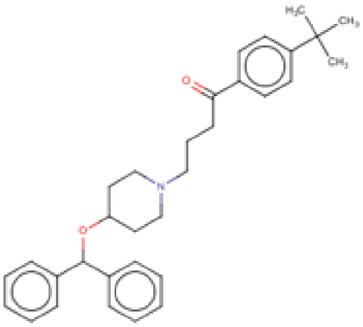
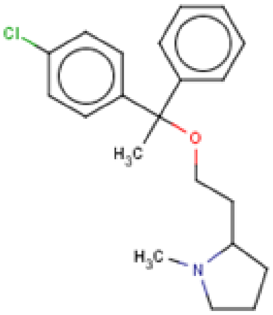
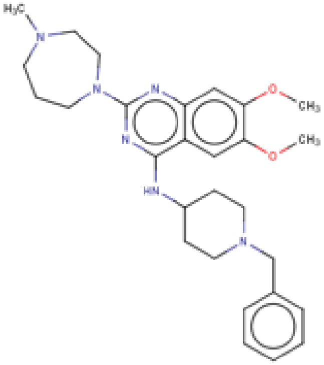
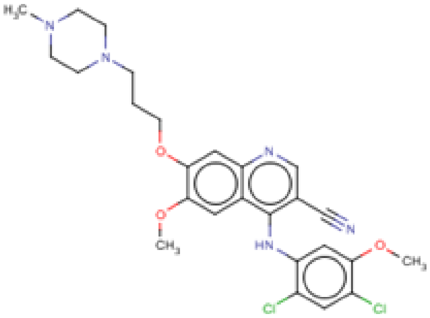
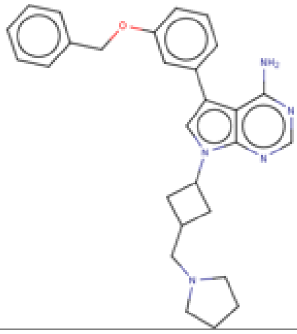
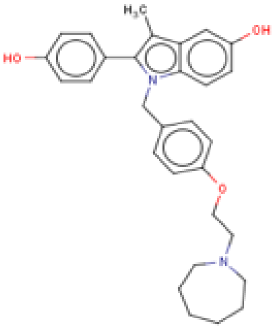
The next most potent series of hits includes afimoxifene (1.36 μM), tetrandrine (2.16 μM), NVP-ADW742 (3.05 μM), and endoxifen (3.05 μM). Afimoxifene and endoxifen are metabolites of tamoxifen (**Table 1.4**), which was previously reported as a sub-micromolar Ebola inhibitor [7]. Likewise, tetrandrine is structurally similar to cepharanthine (**Table 1.4**),[7] as both are isolated from the same plant genus. NVP-ADW742, on the other hand, is structurally dissimilar from any previously reported compound (**Table 1.4**). However, additional tyrosine kinase inhibitors have shown efficacy against the EBOV with a range of potencies *in vitro*, such as sunitinib (1.91 μM) and nilotinib (24.3 μM).[7]

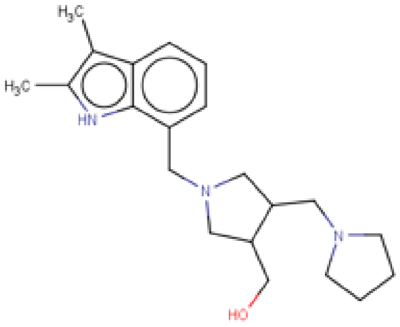
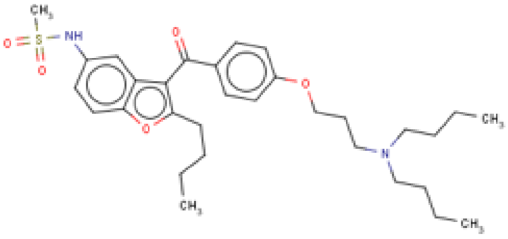
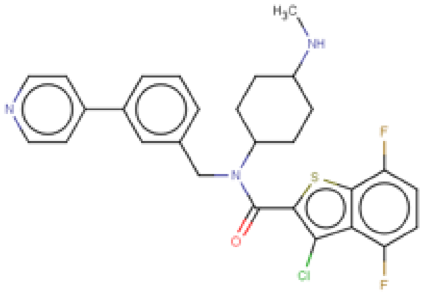
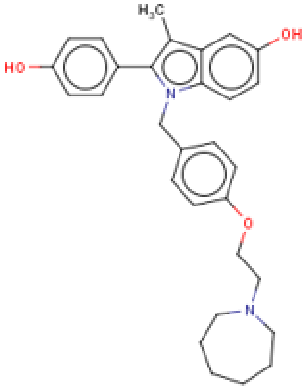
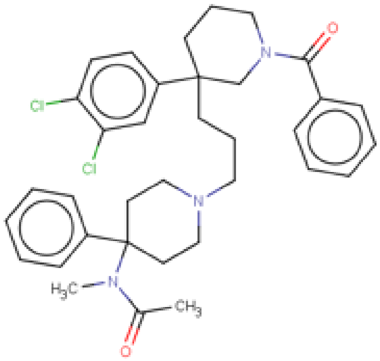
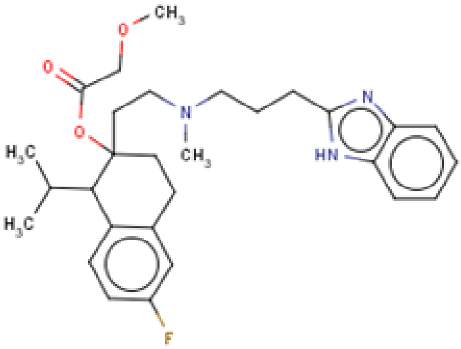
The remaining hits, *i.e.*, ZINC91973695, deptropine, GANT 61, ZINC67869167, Hh-Ag1.5, cediranib, ebastine, osanetant, have potencies ranging from 6.09 μM – 9.65 μM (**Table 1.4**). Each of these hits is structurally unique with respect to previously published compounds (**Table 1.4**). In addition to being structurally novel among EBOV inhibitors, ZINC91973695 and ZINC67869167 have no previously reported bioactivities.

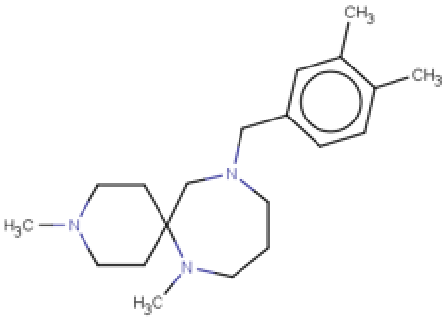
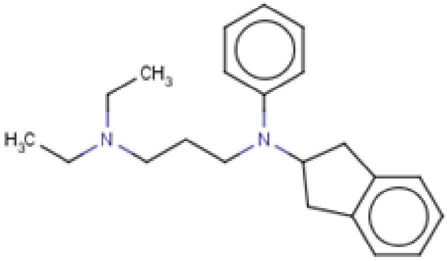
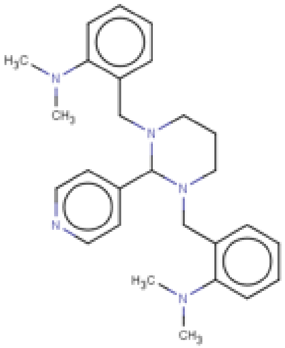
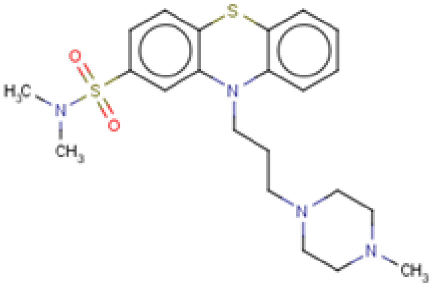
Table 1.3. Structural similarity of top hits to previously published compounds. The Tanimoto coefficient (T_C) between experimentally confirmed hits and compounds in the literature was calculated.

Hit (IC_{50})	T_C	Published Compound (IC_{50})
Afimoxifene (1.96 μM) 	0.99	Tamoxifen (0.73 μM) 
Tetrandrine (2.16 μM) 	0.97	Cepharanthine (1.53 μM) 
Vindesine (0.272 μM) 	0.96	Vinblastine (0.048 μM) 

Hit (IC$_{50}$)	T$_C$	Published Compound (IC$_{50}$)
Endoxifen (3.05 µM) 	0.93	Tamoxifen (0.73 µM) 
Deptropine (6.58 µM) 	0.89	Benztropine (2.64 µM) 
Cediranib (7.67 µM) 	0.79	Gefitinib (9.68 µM) 

Hit (IC_{50})	**T_C**	**Published Compound (IC_{50})**
Ebastine (9.56 µM) 	0.73	Clemastine (1.10 µM) 
BIX-01294 (0.966 µM) 	0.72	Bosutinib (3.85 µM) 
NVP-ADW742 (3.05 µM) 	0.72	Bazedoxifene (3.43 µM) 

Hit (IC_{50})	**T_C**	**Published Compound (IC_{50})**
ZINC91973695 (6.09 µM) 	0.69	Dronedarone (2.20 µM) 
Hh-Ag1.5 (7.67 µM) 	0.66	Bazedoxifene (3.43 µM) 
Osanetant (9.65 µM) 	0.71	Mibefradil (4.32 µM) 

Hit (IC_{50})	**T_C**	**Published Compound (IC_{50})**
ZINC67869167 (6.83 μM) 	0.71	Aprindine (7.69 μM) 
GANT61 (6.83 μM) 	0.47	Thioproperazine (4.32 μM) 

1.4 DISCUSSION

The power of virtual screening is its ability to quickly process millions of compounds and prioritize a small set of highly confident predictions for experimental confirmation. This approach not only saves time and cost as compared to the experimental high throughput screening, but also may lead to the evaluation of additional approved drugs that could be missed in the physical compound screening library. A combination of virtual screening with experimental confirmation is especially useful for challenging assays due to high biosafety requirements, limited patient samples, expensive reagents, or difficult formats (small animal or 3D cell culture). In this study, we prioritized 102 compounds from an *in silico* library of ~17 million compounds for testing in the EBOV entry assay using QSAR modeling and virtual screening. Fourteen of these hits were experimentally confirmed, including 5 selected hits against live-EBOV infections, and their anti-Ebola mechanisms of action were determined using.

The EBOV entry process has been extensively studied. Viral envelope glycoproteins attach to the surface of host cell, and the virus enters through micropinocytosis and endocytosis. Although a cell surface receptor and a few other components are still not clear, several key host factors including cathepsin B/L in the endosome[13] and Niemann Pick C1 protein (NPC1) in the lysosome have been reported as regulators of EBOV entry.[14,15] The chemical biology and anti-Ebola MOAs of the 14 experimentally validated hits were evaluated for interactions with both host and viral targets.

In addition to discovering compounds with unique scaffolds, we also uncovered the anti-Ebola MOAs of these compounds. We have found that the antihistamines ebastine and deptropine inhibited Ebola entry through negatively regulated lysosome function and blocking cathepsin L activity. We also found that osanetant, an anti-psychotic, induces the enlargement of lysosomes and impairs lysosomal function. Additionally, BIX-01294 showed sub-micromolar activity to inhibit EBOV entry. Our LysoTracker dye staining data indicated that BIX-01294 may block EBOV entry through a blockade of lysosome function in host cells. BIX-01294 is a G9a histone methyltransferase (HMTase) inhibitor.[29] HMTases have not been implicated in EBOV entry or replication. Additional chemical biology experiments to test the importance of HMTases in EBOV entry should be performed. Two hedgehog-signaling pathway modulators, Hh-Ag1.5 and GANT61,[30] showed moderate anti-Ebola activity. Our results revealed multiple mechanisms of action involved in the inhibition of EBOV entry by GANT61. GANT61 caused enlargement of lysosomes and inhibited both cathepsin L and cathepsin B, which are known to impair EBOV entry. Two hits from the QSAR-based screen, ZINC91973695 and ZINC67869167, have no previously reported bioactivities. Results of our chemical biology evaluations showed that both compounds induced enlargement of lysosomes, which may implicate the blockage of lysosomal function as a mechanism of action for these two compounds with novel anti-Ebola scaffolds.

A few analogs of previously reported Ebola entry inhibitors or compounds with the same of mechanisms of action were also identified. The most potent hit from our screen (and one of the most potent reported EBOV inhibitors) was vindesine, a vinca alkaloid microtubule inhibitor. Indeed, the vinca alkaloid microtubule inhibitors vinblastine, vincristine, and vinorelbine were also potent hits in the original screen.[26] Likewise, though afimoxifene and endoxifen are novel hits, Selective Estrogen Receptor Modulators (SERMs) have been shown in several studies to have anti-Ebola activity.[26,31,32] The same is true for the receptor tyrosine kinase (RTK) inhibitors cediranib and NVP-ADW742, as sunitinib has been previously reported to have anti-Ebola activity.[26] Last, tetrandrine, a calcium-ion channel blocker, was reported[33] as potent anti-Ebola inhibitor during the course of our study. Thus, these results demonstrate the ability of our QSAR models to reliably retrieve compounds with anti-Ebola activities and confirm the reproducibility of the VLP-assay.

1.5 CONCLUSIONS

Our study is the first case of QSAR-driven experimental discovery of novel anti-Ebola agents with limited host cell toxicity. Robust and predictive QSAR models for both anti-Ebola activity and host cytotoxicity were developed and used for virtual screening of ~17 million compounds in order to identify Ebola inhibitors with high therapeutic windows (selectivity indices). Ultimately, 102 VS hits were tested in both Ebola VLP and ATP content cytotoxicity assays; 14 of these hits had $IC_{50} < 10$ µM and $SI > 10$, which is comparable to the measured potencies of several previously reported compounds. The two most potent hits in the screen were vindesine, a vinca alkaloid microtubule inhibitor, and BIX-01294, an HMTase inhibitor (**Table 1.2**). In a live-EBOV assay, vindesine had an IC_{50} of 0.34 µM. Several of the hits were SERMs and RTKs, which have MOAs known to be related to anti-Ebola activity. We investigated the previously uncharacterized MOAs for anti-Ebola activity of several hits, including both host factors and direct Ebola VLP interactions. Two compounds, ZINC91973695 and ZINC67869167, represent novel chemotypes and can be considered as leads for future anti-Ebola chemical optimization.

In addition to the identification of these compounds, this study demonstrates that FDA-approved drugs, such as vindesine, and compounds that have not yet passed clinical trials for their primary indications, like cediranib, can be repurposed as antivirals. Such compounds are of particular interest, as they may have the potential, pending additional pre-clinical evaluation, to be granted orphan drug status in the United States for EBOV disease. The integrated computational and experimental strategy employed in this study represents an advancement for the rapid discovery of Ebola therapeutics.

1.6 EXPERIMENTAL SECTION

1.6.1 Data Collection, Curation, and Classification

1.6.1.a Antiviral Data

Prior to this work, the Ebola VLP was prepared at the Icahn School of Medicine at Mount Sinai, and VLP-based qHTS screening campaigns were performed at the National Center for Advancing Translational Sciences (NCATS) at the National Institutes of Health (NIH).[26] The results of 4 qHTS screening campaigns (2 primary and 2 confirmatory) were extracted from PubChem (AIDs 1117318, 1117313, 1117312, and 1117308).[34,35] These data are available in **Supplementary Files 1-4**.

Each of the four screens has three readouts, including a blue, green, and ratio (blue/green) channel. The blue channel analyzes the efficacy of the compound at inhibiting VLP entry activity in the host cell. The green channel indicates the healthy and viable cells that loaded with CCF2-AM. The ratio channel screen measures the ratio of blue/green spectra. The beta-lactamase in the VLP hydrolyzes the CCF2-AM dye used in the assay to give a blue fluorescence spectrum. An effective inhibitor will prohibit the beta-lactamase in the VLP from hydrolyzing CCF2-AM, resulting in reduction of the intensity of the blue fluorescence spectrum. A low blue emission spectrum indicates that the compound is inhibitory, while a high green emission spectrum reflects the absence of host cytotoxicity. A simplified schema of the assay is depicted in **Figure 1.9**.

In total, 3121 compounds were tested. These data were then curated according to our well-established protocols.[36–38] Briefly, mixtures, inorganics, and organometallics were removed.

Additionally, replicate compounds were identified and sets were removed if screening results conflicted; if the results were concordant, then one representative compound was selected. After curation, 3104 unique compounds remained.

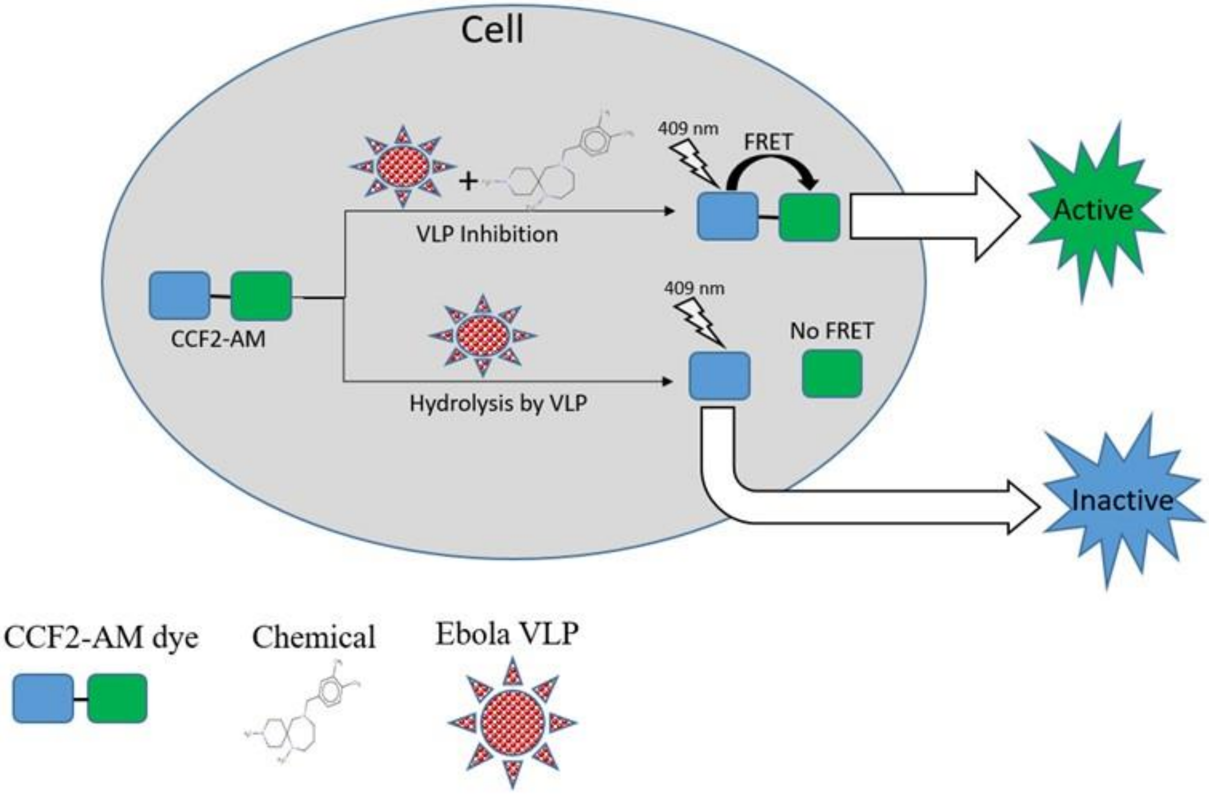


Figure 1.9. Simplified schema of Ebola VLP assay. Ebola VLPs contain Ebola GP and the VP40 protein fused to a beta-lactamase (Bla) reporter. HeLa cells are loaded with the beta-lactamase substrate CCF2-AM. If the VLP enters into the cell, Bla hydrolyzes the substrate CCF2-AM, disrupting the fluorescence resonance energy transfer (FRET) in the substrate, thus causing blue fluorescence. Inhibition of the VLP by a chemical will preserve the substrate FRET, maintaining a green fluorescence. The ratio of blue/green fluorescence intensities represents the VLP activity of inside cells.

1.6.1.b Cytotoxicity Data

Host cell cytotoxicity data for a subset of compounds tested for anti-EBOV activity were obtained from the researchers at the NCATS. Compounds were tested for host cytotoxicity potential in HeLa and HEK cell lines. In total, 171 unique compounds were tested in HeLa cell line, and 156 unique compounds were tested in HEK cell line. All 156 compounds tested in the HEK cell line were also tested in the HeLa cell line. Data curation was performed as above, and one organometallic was removed, leaving 170 and 155 compounds for consideration from the HeLa and HEK cell lines, respectively. These data are available in **Supplementary Files 5-6.**

1.6.1.c Determination of Antiviral Activity

Only compounds with dose-response curve classes[39] of 1.1, 1.2, 2.1, 2.2, and 4 were considered for potential inclusion into the QSAR model training set. In order to comprehensively characterize the results of the screens, two separate protocols were used to classify "active" and "inactive" compounds for subsequent QSAR modeling. In the first protocol (P1), a compound was classified as "active" if and only if the compound had an AC_{50} < 10 µM and Maximum Inhibition ≥ 70% in both a primary and confirmatory screen. Similarly, an "inactive" compound had an AC_{50} ≥ 10 µM and Maximum Inhibition < 70% in both a primary and confirmatory screen.

In the second protocol (P2), an "activity" score was calculated for each compound, *j*, according to the following equation

$$Activity\ score(j) = 50 \times \left[\frac{(\max(\mathrm{AC}_{50}) - \mathrm{AC}_{50}(\mathrm{j}))}{(\max(\mathrm{AC}_{50}) - \min(\mathrm{AC}_{50}))}\right] + 50 \times \left[1 - \frac{(\max(\mathrm{efficacy}) - \mathrm{efficacy}(\mathrm{j}))}{(\max(\mathrm{efficacy}) - \min(\mathrm{efficacy}))}\right]$$

where *activity score(j)* is the relative activity of a specific compound; max(AC_{50}) and min(AC_{50}) are the maximum and minimum AC_{50} in the screen, respectively, and AC_{50}(j) is the AC_{50} of a specific compound; max(efficacy) and min(efficacy) are the maximum and minimum efficacies in the screen, respectively, and efficacy(j) is the efficacy of a specific compound. If a compound had an activity score ≥ 70 in either primary or confirmatory screen, the compound was classified as "active". Similarly, if a compound had an activity score < 70 in either a primary or confirmatory screen, the compound was classified as "inactive".

1.6.1.d Determination of Cytotoxicity

For both the HeLa and HEK cell lines, a compound was considered "toxic" if the associated $pAC_{50} > 4.0$ ($AC_{50} < 100$ μM); whereas, a compound was considered "non-toxic" if the $pAC_{50} \leq 4.0$ ($AC_{50} \geq 100$ μM) or the curve class was 4, indicating no response. Only compounds with dose-response curve classes[39] of 1.1, 1.2, 2.1, 2.2, and 4 were considered.

1.6.1.e Antiviral Training Set Balancing

In both protocols (P1 and P2), the data were imbalanced towards the inactive class. Thus, in order to balance the active and inactive classes in a 1:1 ratio, the inactive class was down-sampled.[40] Fifty percent of the corresponding inactives with the highest Tanimoto similarity[41], *i.e.*, the most similar inactives to the compounds from active class based on MACCS keys fingerprint,[42] were chosen, and the remaining 50% of the corresponding inactives were randomly selected. Important to note that all rationally chosen inactives had different nearest neighbors among actives.[43] For P1 and P2, a total of 166 compounds (83 active and 83 inactive) and 1224 compounds (612 active and 612 inactive) formed the respective training sets. These compounds are available in **Supplementary Files 7-8.**

1.6.1.f Cytotoxicity Training Set Balancing

The "toxic" and "non-toxic" classes were relatively balanced; thus, no down-sampling of the larger class was required. For HeLa and HEK cell lines, a total of 170 compounds (90 toxic and 80 non-toxic) and 155 compounds (83 toxic and 72 non-toxic) formed the respective training sets. These compounds are available in **Supplementary Files 9-10.**

1.6.1.g Modelability Index (MODI)

The MODelability Index (MODI) estimates the likelihood of obtaining predictive QSAR models for a binary data set of compounds.[44] MODI is defined as a weighted ratio of the number of nearest-neighbor pairs of compounds in descriptor space with the same activity class versus the total number of pairs. MODI threshold of 0.65 was previously found to separate the modelable from non-modelable data sets.[44] MODI was calculated for all antiviral and cytotoxicity datasets prior to QSAR modeling in the present study as described earlier.[44]

1.6.2 Computational Methods

1.6.2.a QSAR Model Generation and Validation

Three separate packages, Chembench,[45,46] HiT QSAR,[47] and GUSAR,[48] were employed for consensus classification modeling of both antiviral activity (P1 and P2) and host cytotoxicity (HeLa and HEK). QSAR models built on Chembench used Dragon 6.0 descriptors[49] and the random forest[50] machine-learning algorithm. For models built with HiT QSAR, Simplex Representation of Molecular Structure (SiRMS) descriptors[51] and random forest (RF) were used. GUSAR models utilized a combination of Multilevel Neighborhoods of Atoms (MNA) and Quantitative Neighborhoods of Atoms (QNA) descriptors[52] and a radial-basis function with self-consistent regression (RBF-SCR) as the machine-learning algorithm.[48] We have followed best practices of QSAR modeling developed earlier by our group. All models were rigorously validated using five-fold external cross validation.[9] Y-randomization was performed for all models.[9] Models were statistically evaluated according to, sensitivity (SE), specificity (SP), correct classification rate (CCR), positive predictive value (PPV), and negative predictive value (NPV). These statistical metrics are calculated by the equations 1-5, respectively.

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \qquad (1)$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \qquad (2)$$

$$\mathrm{CCR} = \frac{\mathrm{SE+SP}}{2} \qquad (3)$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \qquad (4)$$

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN+FN}} \qquad (5)$$

Here, TP and TN represent the number of true positives (correct classifications of actives), and true negatives (correct classifications of inactives), respectively; whereas, FP and FN represent the number of false positives (incorrect classifications of actives) and false negatives (incorrect classifications of inactives), respectively.

1.6.2.b Virtual Screening

Two *in silico* libraries, the ZINC drug-like library[22] and previously untested drugs and experimental compounds from the NCATS Chemical Genomics Center Pharmaceutical Collection (NPC), totaling ~17 million compounds after curation (see above), were virtually screened using the developed QSAR models of antiviral activity and host cytotoxicity. A model was deemed acceptable for virtual screening if and only if the CCR, SE, SP, PPV, and NPV were all above 0.60, and no associated Y-randomized model had a CCR above 0.60. An applicability domain (AD) was used for all models. Consensus prediction was utilized, meaning that for a compound to be considered a virtual screening "hit", it must be within the AD of each model and be predicted as "active" and "non-toxic" in *all* developed QSAR models of antiviral activity and host cytotoxicity, respectively (**Figure 1.2**). Once virtual screening "hits" were experimentally validated, the Sequential Agglomerative Hierarchical Nonoverlapping (SAHN) method implemented in the ISIDA/Cluster program[53] was used to probe the uniqueness of hit chemotypes and to identify the most structurally similar compounds in the training set and in the published literature.

1.6.3 Experimental Methods

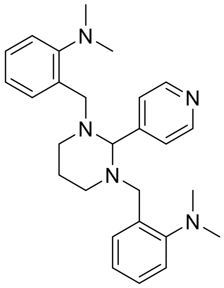
Ebola VLPs containing a beta-lactamase-fused VP40 and GP were prepared in Dr. García-Sastre's lab, as previously described.[6] LiveBLAzer FRET–B/G Loading Kit and CCF2-AM, Dulbecco's modified Eagle's medium (DMEM), and Opti-MEM reduced serum medium were purchased from Life Technologies (Carlsbad, CA, USA). An ATP content cell viability assay kit was purchased from Promega (Madison, WI, USA). 1536-well polystyrene plates were purchased from Greiner Bio-One (Monroe, NC, USA). Compounds were purchased from Sigma-Aldrich (St. Louis, MO, USA), Santa Cruz (Dallas, TX, USA), ChemBridge Corporation (San Diego, CA, USA), Enamine Ltd (Kiev, Ukraine), Maybridge Chemical Company (Altrincham, United Kingdom), Vitas-M Laboratory (Champaign, IL, USA), Ambinter (Orléans, France) and AKos Consulting & Solutions Deutschland GmbH (Steinen-Schlächtenhaus, Germany) at the highest available purity. All of the compounds were dissolved as a 10 mM stock solution in dimethyl sulfoxide (DMSO) and diluted in DMSO at a 1:3 dilution to generate eleven concentrations in 384-well plates, followed by reformatting into one 1536-well compound plate for high throughput screening.

1.6.3.a Materials

All commercially available reagents, compounds, and solvents were purchased and used without further purification. Column chromatography on silica gel was performed on RediSep column using the Teledyne Isco CombiFlash Rf system. Preparative purification was performed on a Waters semi-preparative HPLC. The column used was a Phenomenex Luna C18 (5 micron, 30 × 75 mm) at a flow rate of 45 mL/min. The mobile phase consisted of acetonitrile and water (each containing 0.1% trifluoroacetic acid). A gradient of 10% to 50% acetonitrile over 8 minutes was used during the purification. Fraction collection was triggered by UV detection (220 nm).

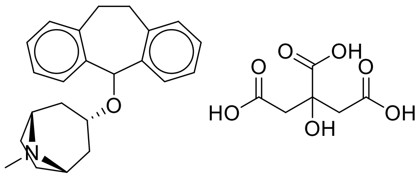
^{1}H spectra were recorded using an INOVA 400 MHz spectrometer (Varian). Samples were analyzed on an Agilent 1200 series LC/MS using a Zorbax Eclipse XDB-C18 reverse phase (5 micron, 4.6 x 150 mm) column and a flow rate of 1.1 mL/min. The mobile phase was a mixture of acetonitrile and H_2O each containing 0.05% trifluoroacetic acid. LC Method A: a gradient of 4% to 100% acetonitrile over 7 minutes was used during analytical analysis. LC Method B: a gradient of 4% to 100% acetonitrile over 3 minutes was used during analytical analysis. High resolution mass spectrometry was recorded on Agilent 6210 Time-of-Flight LC/MS system.

2,2'-((2-(pyridin-4-yl)dihydropyrimidine-1,3(2H,4H)-diyl)bis(methylene))bis(N,N-dimethylaniline) (GANT61)



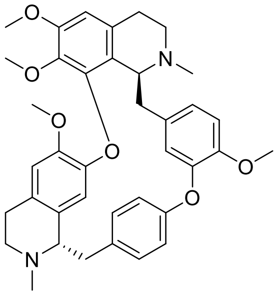
^{1}H NMR (400 MHz, DMSO-d_6) δ 8.59 – 8.52 (m, 2H), 7.70 – 7.64 (m, 2H), 7.47 (dd, J = 7.6, 1.7 Hz, 2H), 7.20 – 7.11 (m, 2H), 7.08 – 6.97 (m, 4H), 4.01 (s, 1H), 3.51 (d, J = 14.1 Hz, 2H), 3.39 (d, J = 14.2 Hz, 2H), 2.81 (dt, J = 11.2, 4.1 Hz, 2H), 2.50 (s, 12H), 2.22 – 2.14 (m, 2H), 1.62 – 1.55 (m, 2H). LC/MS (Method B): (electrospray +ve), m/z 430.3 $(MH)^+$, t_R = 3.826, UV_{254} > 98%.

(1R,3R,5S)-3-((10,11-dihydro-5H-dibenzo[a,d][7]annulen-5-yl)oxy)-8-methyl-8-azabicyclo[3.2.1]octane 2-hydroxypropane-1,2,3-tricarboxylate (Deptropine citrate)



HRMS calculated for : $C_{23}H_{28}NO$ [M + H]$^+$ 334.2165, found 334.2150. LC/MS (Method A): (electrospray +ve), *m/z* 334.1 (MH)$^+$, t_R = 5.000, UV_{254} > 98%.

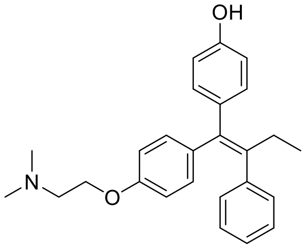
(11S,31S)-16,36,37,54-tetramethoxy-12,32-dimethyl-11,12,13,14,31,32,33,34-octahydro-2,6-dioxa-1(7,1),3(8,1)-diisoquinolina-5(1,3),7(1,4)-dibenzenacyclooctaphane (Tetrandrine)



^{1}H NMR (400 MHz, DMSO-d_6) δ 7.45 (dd, *J* = 8.2, 2.1 Hz, 1H), 7.08 (dd, *J* = 8.2, 2.5 Hz, 1H), 6.91 (d, *J* = 8.2 Hz, 1H), 6.77 (dd, *J* = 8.2, 1.9 Hz, 1H), 6.71 – 6.59 (m, 2H), 6.42 – 6.34 (m, 2H), 6.31 (dd, *J* = 8.3, 2.1 Hz, 1H), 5.92 (s, 1H), 3.89 (dd, *J* = 10.4, 5.9 Hz, 1H), 3.80 (s, 3H), 3.68 (s, 3H), 3.50 (d, *J* = 9.8 Hz, 1H), 3.41 (d, *J* = 4.7 Hz, 1H), 3.32 (s, 3H), 3.29 (s, 4H), 3.17 (dd, *J* = 12.5, 5.9 Hz, 1H), 3.03 (s, 3H), 2.89 – 2.68 (m, 6H), 2.63 (dd, *J* = 13.6, 10.1 Hz, 1H), 2.36 (d, *J*

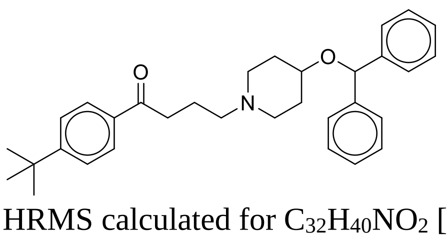
= 16.0 Hz, 1H), 2.28 (d, *J* = 13.5 Hz, 1H), 2.18 (s, 3H). HRMS calculated for $C_{38}H_{44}N_2O_6$ [M + 2H]$^{2+}$ 312.1594, found 312.1602. LC/MS (Method A): (electrospray +ve), *m/z* 623.2 (MH)$^+$, t_R = 3.866, UV_{254} > 98%.

(Z)-4-(1-(4-(2-(dimethylamino)ethoxy)phenyl)-2-phenylbut-1-en-1-yl)phenol (Afimoxifene)



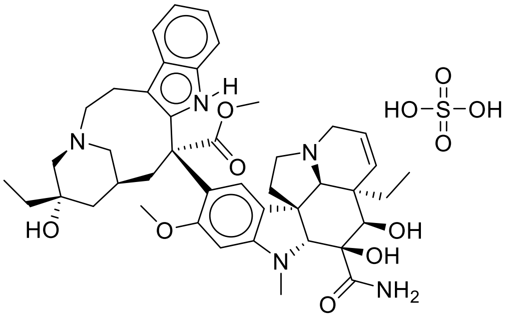
^{1}H NMR (400 MHz, DMSO-d_6) δ 9.40 (s, 1H), 7.22 – 7.13 (m, 2H), 7.09 (ddd, *J* = 6.8, 4.0, 1.3 Hz, 4H), 7.01 – 6.94 (m, 2H), 6.78 – 6.72 (m, 2H), 6.72 – 6.66 (m, 2H), 6.61 – 6.54 (m, 2H), 3.88 (t, *J* = 5.9 Hz, 2H), 2.52 (m, 2H), 2.40 (d, *J* = 7.3 Hz, 2H), 2.15 (s, 5H), 0.84 (t, *J* = 7.4 Hz, 3H). HRMS calculated for $C_{26}H_{30}NO_2$ [M + H]$^+$ 388.2271, found 388.2288. LC/MS (Method A): (electrospray +ve), *m/z* 388.1 (MH)$^+$, t_R = 5.056, UV_{254} > 98%.

4-(4-(benzhydryloxy)piperidin-1-yl)-1-(4-(tert-butyl)phenyl)butan-1-one (Ebastine)



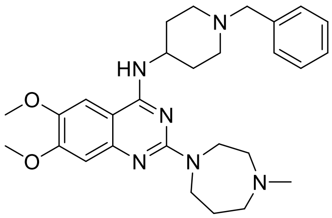
HRMS calculated for $C_{32}H_{40}NO_2$ [M + H]$^+$ 470.3054, found 470.3069. LC/MS (Method A): (electrospray +ve), *m/z* 470.2 (MH)$^+$, t_R = 6.088, UV_{254} > 98%.

methyl (3S,5S,7S,9S)-9-((3aR,3a1R,4R,5S,5aR,10bR)-5-carbamoyl-3a-ethyl-4,5-dihydroxy-8-methoxy-6-methyl-3a,3a1,4,5,5a,6,11,12-octahydro-1H-indolizino[8,1-cd]carbazol-9-yl)-5-ethyl-5-hydroxy-1,4,5,6,7,8,9,10-octahydro-2H-3,7-methano[1]azacycloundecino[5,4-b]indole-9-carboxylate sulfate (Vindesine)



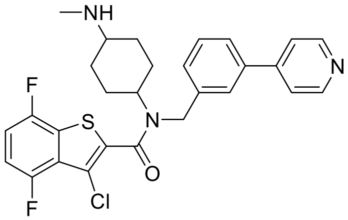
^{1}H NMR (400 MHz, DMSO-d_6) δ 9.67 (s, 1H), 7.52 (d, *J* = 7.9 Hz, 1H), 7.34 (d, *J* = 8.1 Hz, 1H), 7.26 (s, 1H), 7.17 (d, *J* = 3.1 Hz, 1H), 7.08 (t, *J* = 7.5 Hz, 1H), 7.00 (t, *J* = 7.5 Hz, 1H), 6.43 (s, 1H), 6.25 (s, 1H), 5.72 (dd, *J* = 10.7, 4.8 Hz, 1H), 5.61 (d, *J* = 10.6 Hz, 1H), 5.07 (s, 1H), 4.34 (s, 1H), 3.75 (s, 5H), 3.58 (s, 3H), 3.45 (s, 6H), 3.16 (s, 1H), 3.05 (s, 2H), 2.81 (s, 3H), 2.72 (s, 3H), 2.19 (d, *J* = 15.3 Hz, 1H), 1.97 (s, 1H), 1.65 – 1.53 (m, 3H), 1.49 – 1.23 (m, 4H), 0.86 (t, *J* = 7.4 Hz, 3H), 0.73 (t, *J* = 7.3 Hz, 3H). HRMS calculated for $C_{43}H_{57}N_5O_7$ [M + 2H]$^{2+}$ 377.7124, found 377.7140. LC/MS (Method A): (electrospray +ve), *m/z* 754.3 (MH)$^+$, t_R = 3.802, UV_{254} > 98%.

N-(1-benzylpiperidin-4-yl)-6,7-dimethoxy-2-(4-methyl-1,4-diazepan-1-yl)quinazolin-4-amine (BIX-01294)



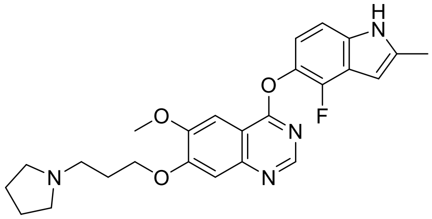
HRMS calculated for $C_{28}H_{39}N_6O_2$ $[M + H]^+$ 491.3129, found 491.3119. LC/MS (Method A): (electrospray +ve), *m/z* 491.3 $(MH)^+$, $t_R = 3.199$, $UV_{254} > 98\%$.

(R)-N-(1-(3-(1-benzoyl-3-(3,4-dichlorophenyl)piperidin-3-yl)propyl)-4-phenylpiperidin-4-yl)-N-methylacetamide (Hh-Ag1.5)



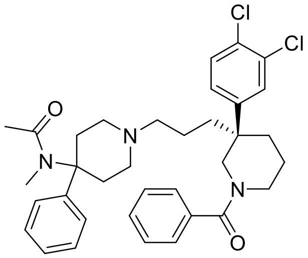
HRMS calculated for $C_{28}H_{27}ClF_2N_3OS$ $[M + H]^+$ 526.1526, found 526.1532. LC/MS (Method A): (electrospray +ve), *m/z* 526.1 $(MH)^+$, $t_R = 4.048$, $UV_{254} > 98\%$.

4-((4-fluoro-2-methyl-1H-indol-5-yl)oxy)-6-methoxy-7-(3-(pyrrolidin-1-yl)propoxy)quinazoline (Cediranib)



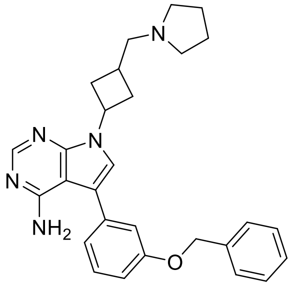
^{1}H NMR (400 MHz, DMSO-d_6) δ 11.34 (d, *J* = 2.4 Hz, 1H), 8.49 (s, 1H), 7.60 (s, 1H), 7.38 (s, 1H), 7.16 (d, *J* = 8.6 Hz, 1H), 6.98 (dd, *J* = 8.6, 7.4 Hz, 1H), 6.24 (d, *J* = 2.0 Hz, 1H), 4.25 (t, *J* = 6.4 Hz, 2H), 3.99 (s, 3H), 2.58 (t, *J* = 7.1 Hz, 2H), 2.46 (m, 5H), 2.41 (s, 3H), 1.99 (dd, *J* = 7.9, 5.7 Hz, 2H), 1.70 (s, 3H). HRMS calculated for $C_{25}H_{28}FN_4O_3$ [M + H]$^+$ 451.2140, found 451.2130. LC/MS (Method A): (electrospray +ve), *m/z* 451.1 (MH)$^+$, t_R = 4.353, UV_{254} > 98%.

(R)-N-(1-(3-(1-benzoyl-3-(3,4-dichlorophenyl)piperidin-3-yl)propyl)-4-phenylpiperidin-4-yl)-N-methylacetamide (Osanetant)



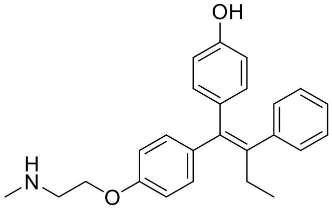
HRMS calculated for $C_{35}H_{42}Cl_2N_3O_2$ [M + H]$^+$ 606.2649, found 606.264. LC/MS (Method A): (electrospray +ve), *m/z* 606.2 (MH)$^+$, t_R = 5.399, UV_{254} > 98%.

5-(3-(benzyloxy)phenyl)-7-(3-(pyrrolidin-1-ylmethyl)cyclobutyl)-7H-pyrrolo[2,3-d]pyrimidin-4-amine (NVP-ADW742)



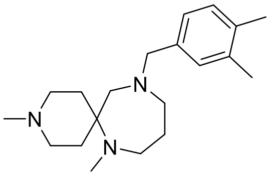
^{1}H NMR (400 MHz, DMSO-d_6) δ 8.12 (s, 1H), 7.65 (s, 1H), 7.51 – 7.44 (m, 2H), 7.45 – 7.29 (m, 4H), 7.15 (t, *J* = 2.0 Hz, 1H), 7.11 – 7.04 (m, 1H), 7.04 – 6.96 (m, 1H), 6.10 (s, 2H), 5.31 (p, *J* = 8.3 Hz, 1H), 5.17 (s, 2H), 2.72 – 2.58 (m, 4H), 2.47-2.44 (m, 5H), 2.25 (ddd, *J* = 12.1, 8.5, 2.9 Hz, 2H), 1.74 – 1.62 (m, 4H). HRMS calculated for C28H32N5O [M + H]$^+$ 454.2601, found 454.2615. LC/MS (Method A): (electrospray +ve), *m/z* 454.2 (MH)$^+$, t$_R$ = 4.106, UV$_{254}$ > 98%.

(E)-4-(1-(4-(2-(methylamino)ethoxy)phenyl)-2-phenylbut-1-en-1-yl)phenol (Endoxifen)



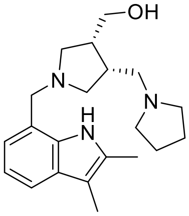
^{1}H NMR (400 MHz, DMSO-d_6, major isomer) δ 9.17 (s, 1H), 7.17 (t, *J* = 7.6 Hz, 2H), 7.08 (d, *J* = 8.9 Hz, 5H), 6.92 (d, *J* = 8.2 Hz, 2H), 6.59 (d, *J* = 8.2 Hz, 2H), 6.39 (d, *J* = 8.1 Hz, 2H), 4.01 (t, *J* = 5.6 Hz, 2H), 2.82 (t, *J* = 5.6 Hz, 2H), 2.39 (q, *J* = 7.3 Hz, 2H), 2.34 (s, 3H), 0.84 (t, *J* = 7.3 Hz, 3H). HRMS calculated for $C_{25}H_{28}NO_2$ [M + H]$^+$ 374.2115, found 374.2105. LC/MS (Method A): (electrospray +ve), *m/z* 374.1 (MH)$^+$, t$_R$ = 5.001, UV$_{254}$ = 60.6% (major isomer).

11-(3,4-dimethylbenzyl)-3,7-dimethyl-3,7,11-triazaspiro[5.6]dodecane (ZINC67869167)



HRMS calculated for $C_{20}H_{34}N_{3}$ $[M + H]^+$ 316.2747, found 316.2757.

((3R,4R)-1-((2,3-dimethyl-1H-indol-7-yl)methyl)-4-(pyrrolidin-1-ylmethyl)pyrrolidin-3-yl)methanol (ZINC91973695)



HRMS calculated for $C_{21}H_{32}N_{3}O$ $[M + H]^+$ 342.254, found 342.2532.

1.6.3.b Cell culture methods

HeLa and HEK293 cells were purchased from the American Type Culture Collection (ATCC, Manassas, VA, USA). The cells were cultured in DMEM supplemented with 10% fetal bovine serum (FBS, GE healthcare, Piscataway, NJ, USA) and 100 U/mL of penicillin and 100 µg/mL of streptomycin (Life Technologies, Carlsbad, CA, USA) at 37 °C in a humidified atmosphere with 5% CO2. Cells were passaged at 90% confluency.

1.6.3.c Ebola VLP beta-lactamase assay for HTS in 1536-well plates

A chemical biology screening campaign was performed. Ebola VLP assay was conducted as previously described.[7] Briefly, HeLa cells were seeded at 750 cells/well in 3 µL of assay medium (DMEM+10% FBS) in 1536-well assay plates. Compounds were prepared in a 1536-well compound plate, and 23 nL of each compound was transferred into 1536-well assay plate using an NX-TR pintool station (WAKO Scientific Solutions, San Diego, CA, USA). After 1 h incubation at 37 °C with 5% CO_2, 1 µL/well of VLP solution was added to the assay plates using a BioRapTR FRD dispenser. The plates were then spinoculated, followed by incubation at 37 °C with 5% CO_2 for 4.5 h. 1 µL CCF2-AM beta-lactamase substrate was added in to each well, and the plates were incubated for 2 h at room temperature. The assay was detected at dual fluorescence intensities (Ex1= 405±20, Em1= 460±20, and Ex2= 405±20, Em2= 530±20 nm) using EnVision plate reader (PerkinElmer, Boston, MA, USA).

1.6.3.d Cell viability assay with the ATP content assay kit

The cell viability assay was performed as previously described.[7] Briefly, HeLa and HEK293 cells were plated at 750 cells/well in 3 µL in 1536-well assay plates, followed by the addition of tested compounds at 23 nL/well. After a 4.5 h incubation at 37 °C and 5% CO_2, cell viability was measured by adding 3 µL of ATP content assay mixture to each well. Luminescence values were obtained using a ViewLux plate reader (PerkinElmer, Boston, MA, USA).

1.6.3.e Ebola live virus assays

Vero E6 cells were plated in the 96-well plate (black with optical bottom). Briefly, serial dilutions of 5 drugs (diluted in DMEM 2% FBS starting at 10 µM) and DMSO as control, were added to the wells, and incubated for 1 h at 37 °C with 5% CO_2. The cells were infected with EBOV/Mayinga-eGFP at a MOI of 0.1 TCID50/cell. The assay was run in triplicate at a biosafety level-4 (BSL-4) facility. The fluorescence was read 72 h after infection using a BioTek Synergy HT.

1.6.3.f Filipin staining and LysoTracker-red staining

The assays were performed as previously described.[54] Fibroblast cells were plated at 1,000 cells/well in 4 µL of assay medium (DMEM + 10% FBS) in 1536-well assay plates and incubated overnight at 37 °C and 5% CO_2. Compounds were added to the assay plate at 23 nL/well. After 24 h incubation at 37 °C and 5% CO_2, 2 µl/well of 50 ng/ml filipin or 0.5 µM LysoTracker Red DND-99 was added to the plate. After 1 hr. incubation at 37 °C and 5% CO_2, the plates were washed twice. The fluorescence intensities were then read with a fluorescence plate reader (GE Healthcare, Chicago, Illinois, USA). U18666A [3-β-(2-[diethylamino]ethoxy)-androst-5-en-17-one, monohydrochloride] was used as the positive control.[55]

1.6.3.g Cathepsin B/L assay

Cathepsin B/L assays were performed as previously described.[8] Briefly, recombinant 5 ng of cathepsin B, or cathepsin L were added into each well in 384-well plate. Indicated drugs were added into the recombinant enzymes, followed by initiation of the reaction by addition of fluorescent substrate. The activity measurements were done using Tecan plate reader (Tecan US, Inc., Morrisville, NC, USA). Cathepsin L inhibitor and ED64 were used as positive controls.[8]

1.6.3.h Thermal shift binding assay with Ebola VLP

The thermal shift binding assay was performed as previously described.[16] Ebola VLPs were pre-incubated with indicated drugs for 10 min at room temperature. The mixture was then subsequently heated at 49 °C for 3 min, followed by centrifugation at 13, 000 x g at 4 °C for 20 min. The supernatant was collected and denatured by heating at 75 °C for 10 min in the presence of SDS loading buffer (Life Technologies, Carlsbad, California, United States). The samples were separated by SDS-PAGE gel electrophoresis and detected by anti-beta-lactamase antibodies (Life Technologies, Carlsbad, California, United States).

1.6.3.i Data analysis and statistics

Half maximal inhibitory concentration (IC_{50}) values of compound activity data were calculated using Prism software (GraphPad Software, Inc. San Diego, CA). All values were expressed as the mean ± SEM ($n \geq 3$).

1.7 ASSOCIATED CONTENT

Supporting Information. Supplementary Files 1-15 are available at https://scapuzzi.web.unc.edu/free-downloads/

Additionally, all the datasets and Chembench models are provided in and on the Chembench Web-Portal (https://chembench.mml.unc.edu/), which provides public access and use of data and models used in this study. The P1, P2, HEK, and HeLa training sets are publicly indexed as "Ebola_SM1" and "Ebola_PCM4", "151105_Ebola_Toxicity_HEK", and "151105_Ebola_Toxicity_HELA", respectively. The Chembench P1, P2, and HeLa models are publicly indexed as "153004_ebola_Strict_Model1_166_DragonH", "151305_ebola_1224_PCM4", and "151105_ebola_tox_HeLa", respectively.

CHAPTER 2: COMPUTATIONAL DISCOVERY AND EXPERIMENTAL VALIDATION OF POTENT INHIBITORS OF THE UNDERSTUDIED KINASE DCLK1

2.2 INTRODUCTION

Doublecortin-like kinase 1 (DCLK1) has been implicated in the development and progression of several cancers.[56,57] Recent studies have shown that DCLK1, which is also referred to as DCAMKL-1, drives tumorigenesis in colon and pancreatic cancer[58,59], is overexpressed in cancers of the liver and esophagus[60,61], and such overexpression is an adverse prognosis factor in bladder and non-small cell lung cancer[62,63]. Notably, in 2013, Nakanishi *et al.* showed that DCLK1 expression uniquely distinguished tumor stem cells (TSCs) in colorectal cancer from healthy stem cells and demonstrated that specific ablation of DCLK1-positive TSCs reduced tumor size without damaging healthy tissue.[64] Given these observations, DCLK1 represents an emergent therapeutic target in oncology, especially, for colorectal cancer.

Despite its growing notoriety in oncology, DCLK1 is still considered as an understudied kinase[65] lacking any potent and moderately selective tool compounds. As per the guidelines for inclusion into the Structural Genomic Consortium's comprehensive kinase chemogenomics set (KGCS), DCLK1 remains a dark, or chemically untargeted, protein kinase.[66] A chemical probe for DCLK1 would be of great scientific and therapeutic value, as it could help unravel the specific biological role of this kinase in various cancers and serve as a potential lead for drug discovery efforts.[67,68]

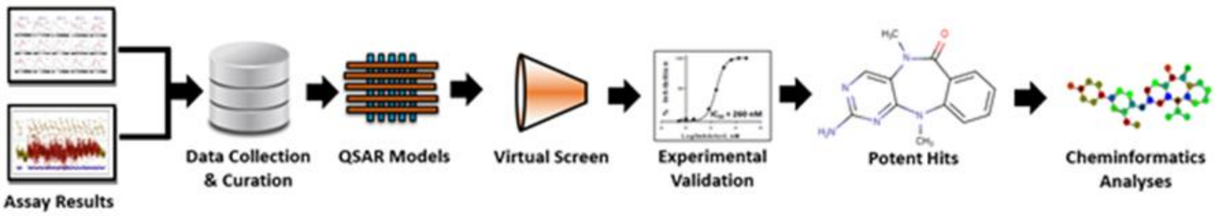
The development of a chemical probe is dependent upon the identification of high quality chemical starting points for potency and selectivity optimization.[66] This process is particularly challenging for dark kinases like DCLK1, where chemogenomics data and SAR studies are limited and often the unintended consequence of screening campaigns for other kinases. Indeed, the handful of compounds in the literature that target DCLK1 have come mainly from a kinome-wide screen of inhibitors bearing pyrimido-diazepine scaffolds[69] or have been reported as an off-target effect during probe development for other kinases such as, ACK1, ERK5, and LRRK2.[69–71] The development of novel DCLK1 inhibitors is critical to progress probe development for this biomedically-relevant, but, so far, dark kinase.

Methods of computer-aided drug design (CADD) are routinely used to leverage prior screening data towards the discovery of novel bio-active compounds while also reducing time and cost. CADD approaches are most effective when large and diverse chemogenomics sets are available. Unfortunately, for DCLK1, experimental screening datasets are small (less than 100 compounds have been tested so far), most compounds are inactive, and active molecules are very limited in chemical diversity.[69] According to best practices previously established by us and others,[9,72] it is not advisable to employ CADD approaches, especially QSAR modeling, for such datasets, as the potential for faulty predictions is high. At the same time, we were confronted by the therapeutic importance of this dark kinase, a lack of tool compounds, and the expressed need to prioritize laborious synthetic efforts presented by the diversity of synthetically feasible compounds. We were thus motivated to apply our expertise in modeling challenging datasets[40,47,73] in an attempt to discover potent DCLK1 inhibitors in close collaboration with our experimental partners.

In order to accomplish this goal, we executed the following steps: (i) the development of QSAR models of DCLK1 inhibition from prior screening data; (ii) virtual screening of focused chemical libraries to identify putative DCLK1 inhibitors; (iii) experimental validation of selected compounds; and (iv) off-target selectivity analyses for experimentally confirmed hits.

Once several high-quality DCLK1 inhibitors were identified, we then (v) derived structural rules and key molecular interactions to guide future design and optimization efforts of these compounds using the cheminformatics techniques of matched molecular pair analysis, QSAR model interpretation, and molecular docking. This joint modeling and experimental effort (**Figure 2.1**) resulted in the discovery of some of the most potent DCLK1 inhibitors to-date. These compounds constitute leads for the development of a chemical probe for this dark kinase.

Figure 2. 1. Overall study design. The workflow combines computational and experimental medicinal chemistry approaches for the discovery for novel potent DCLK1 compounds.



2.3 RESULTS

2.3.1 QSAR Model Development

Modelability Index (MODI),[44] which affords rapid estimation of the feasibility of obtaining predictive QSAR models, was calculated for both training sets. MODI values of 0.79 and 0.89 were obtained for the KINOMEscan and KiNativ training set, respectively. These MODI values were well-above the recommended threshold of 0.65, indicating that despite a high degree of imbalance towards inactive compounds and limited chemical diversity there was meaningful SAR that separates actives from inactives.

Next, we moved to model development and succeeded in developing robust and externally predictive QSAR models. Results of 5-fold external cross-validation are presented in **Table 2.1**. All metrics used to evaluate model performance were above the recommended threshold of 0.60. As such, these metrics demonstrated that active and inactive DCLK1 compounds can be correctly classified through statistically meaningful SAR.

Table 2.33.Statistical characteristics obtained on 5-fold external CV of all models developed in this study.

Model Name	Actives	Inactives	Total	CCR	SE	SP	PPV	NPV
KINOMEscan	8	45	53	0.84	0.75	0.92	0.62	0.95
KiNativ	5	42	47	0.73	0.70	0.75	0.61	0.96

Since externally-validated and predictive QSAR models were developed using all available data, the applicability domain (AD) of the models was maximized, as the imbalanced datasets did not need to be down-sampled.[40] Both the KINOMEscan and KiNativ models were predictive and useful for virtually screening new compounds in so far as these compounds fell within the AD that has been maximized by using all available data. In order to demonstrate that the models were not obtained because of random SAR correlation between bioactivity and chemical descriptors, 1000 rounds of Y-randomization was performed. All Y-randomized models showed a CCR below 0.60.

2.3.2 QSAR-Based Virtual screening

A set of 169 designed compounds possessing the same scaffolds as in **Figure 2.2** was virtually screened with both QSAR models (See Methods). Although all 169 compounds were within the AD of both models, only seven compounds were predicted as active by both models. All seven compounds possessed the 5,11-dihydro-6H-benzo[e]pyrimido[5,4-b][1,4]diazepin-6-one scaffold (**Figure 2.2B**). The remaining compounds were either predicted active only by the KINOMEscan model (29 compounds) or the KiNativ model (two compounds); 131 compounds were predicted inactive by both models. These results are consistent with the distribution between actives and inactives in the training sets, supporting the notion that DCLK1 has highly specific requirements to compound structure to make it active. From the compounds that did not meet the "hit" criteria, four were selected as negative controls for the model validation. Ultimately, 11 compounds, seven putatively active and four predicted inactive, were selected for experimental studies. Virtual screening results for all compounds are provided in the Supporting Information (**Supplementary Table 1**).

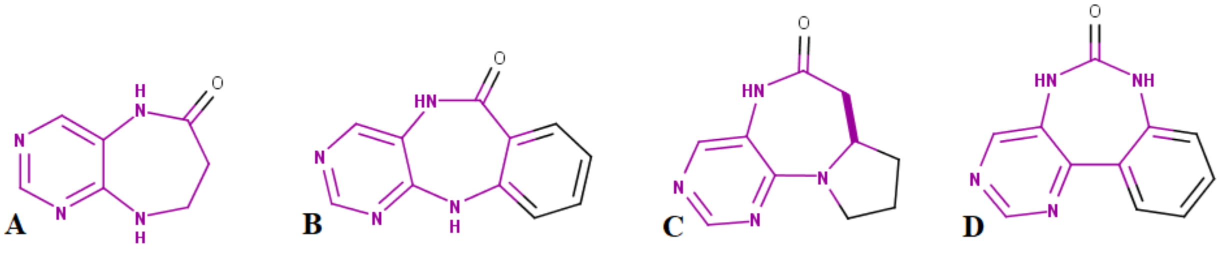


Figure 2.2. Four scaffolds (A-D) based on the pyrimido-diazepine core (purple) possessed by compounds in the modeling datasets. Note A-C are 1,4-diazepines, while D is a 1,3-diazepine.

2.3.3 Experimental Validation

A threshold of activity for the 11 compounds from the virtual screening was set at 10 μM, as both models were developed from compounds screened at this concentration. The IC_{50} values and the structures of these 11 compounds are shown in **Table 2.2**. These results show that QSAR models were ~73% accurate, as they correctly predicted the activity calls for 8 out of 11 compounds. Of the seven putative DCLK1 hits, six had IC_{50} < 10 μM, including four sub-micromolar inhibitors. The top hit, XMD13-44, had an IC_{50} of 52 nM. Two of the four putatively inactive compounds were incorrectly classified, but, ironically, in this case of negative controls, the inaccuracy is a desired outcome. Statistically, however, this observation is expected for the imbalanced dataset modeling.[74] Overall, eight compounds from the virtual screen had IC_{50} < 10 μM for DCLK1. Full dose-response curves are available in **Figure 2.3**.

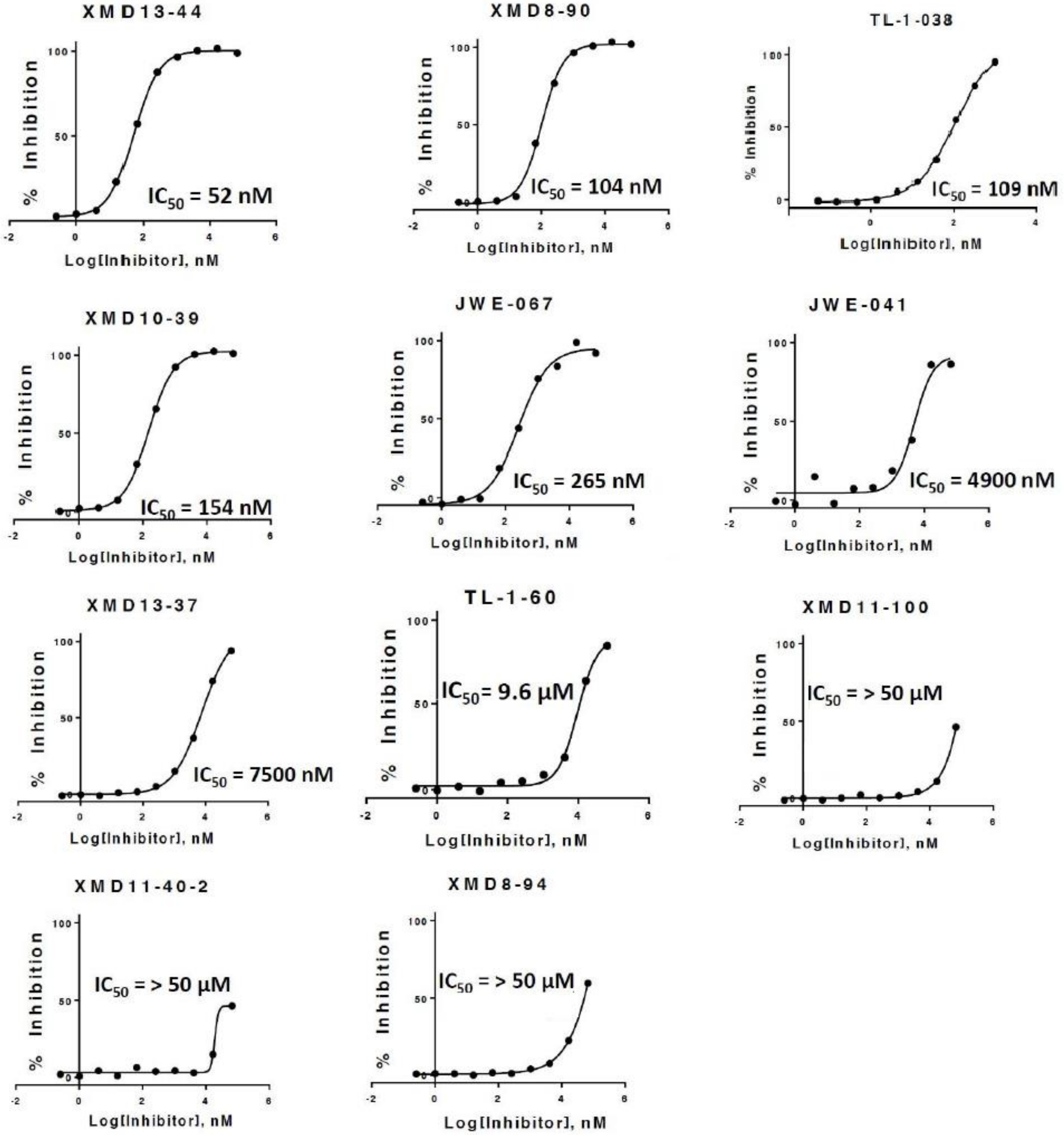


Figure 2. 73. **Dose-response curves for the eleven virtual screening hits.**

The selectivity profiles of the eight compounds with IC_{50} < 10 μM for DCLK1 were determined using a KiNativ screen (**Table 2.3**). The compounds were evaluated both by the selectivity index (SI) at 65% and 90% inhibition at 10 μM, *i.e.*, SI(65) and SI(90), respectively. The Structural Genomics Consortium (SGC) has previously defined SI(65) < 4.0% and SI(90) < 2.0% as acceptable selectivity profiles for a tool compound to be considered for inclusion into their comprehensive kinase chemogenomics set (KCGS).[66] Only XMD13-37 inhibited more than 4% of kinases screened according to SI(65). On the other hand, all eight compounds had acceptable SI(90) profiles. KiNativ screen data are provided for the eight compounds in the Supporting Information (**Supplementary Table 2**).

2.3.4 SAR Analysis and Implications for Future Design

All of the experimentally validated hits possessed the 5,11-dihydro-6H-benzo[e]pyrimido[5,4-b][1,4]diazepin-6-one scaffold, and the four most potent hits shared common structural moieties (**Table 2**). In order to gain insights about structural aspects and key molecular interactions associated with DCLK1 inhibition among these compounds, as well as to guide future design and optimization efforts, SAR analysis was performed using several cheminformatics techniques.

2.3.4.a Matched Molecular Pair Analysis and Model Interpretation

The modeling datasets were investigated for matched molecular pairs (MMPs) bearing the 5,11-dihydro-6H-benzo[e]pyrimido[5,4-b][1,4]diazepin-6-one scaffold. From the KINOMEscan dataset, a series of MMPs was identified with several activity cliffs, *i.e.*, structurally similar compounds from different activity classes.[75,76] The shared scaffold of the MMPs and accompanying structural changes are shown in **Figure 2.4**. Within this series of compounds, only XMD8-85 and XMD8-87 were active; therefore, the remaining associated MMPs constitute activity cliffs. This analysis revealed that the presence of a methoxy substituent at R1 and a co-occurring methyl substituent at R2 correlated with increased DCLK1 inhibition.

On the other hand, the influence of the R3 substituent was unclear, as active compounds XMD8-85 and XMD8-87 are MMPs differing only at this position. Fragment descriptor interpretation (**Figure 2.5**) from the KINOMEscan QSAR model, however, showed that the methyl addition at R3 increased the overall active (inhibitory) character of XMD8-85 relative to XMD8-87. The SAR elucidated by the MMP analysis was also reflected by the model interpretation, as methoxy substituents at R1 and methyl substituents at R2 increased the activity profile in descriptor space.

2.3.4.b Molecular Docking

Training set compounds were docked into the crystal structure of DCLK1 (PDB: 5JZN)[57] in order to evaluate and validate the molecular docking approach. All training set actives, with the exception of just one compound, ranked within the top 15% of the best scored docking poses (**Supplementary Table 3**). This enrichment of training set actives among the best scored docking poses provided validation for the use docking as a method to generate hypotheses related to the protein-ligand interactions for the 11 compounds from the virtual screen. For these 11 compounds, the molecular docking scores correlated well with the experimentally-determined potencies. Indeed, the two most potent compounds, XMD13-44 and XMD8-90, had the best two docking scores (-7.75 and -7.69), respectively. XMD10-100, inactive upon experimental testing, likewise, had the second worst docking score (-4.72). TL-1-060, the compound with the worst docking score (-4.53), was only weakly potent (9.60 μM). All scores and poses for these 11 compounds are provided in the Supporting Information (**Supplementary Table 3**).

2.4 DISCUSSION

Despite small dataset sizes with limited chemical diversity and a small number of active compounds, robust and predictive QSAR models of DCLK1 inhibition were developed from the results of KINOMEscan and KiNativ assays (**Table 2.1**). On further inspection, these training set characteristics were, in fact, crucial for successful QSAR model development. The limited chemical diversity in terms of scaffolds (**Figure 2.2**) among the training sets meant that large changes in DCLK1 bioactivity were caused by slight modifications in a small number of substituents. This observation was supported by the high MODIs of the training sets, which indicated that there was statistically meaningful SAR separating active compounds from inactive compounds. Both the MMP and model interpretation analyses also reflected this observation (**Figures 2.3 and 2.4**), as both showed that the activity profile was modulated by a few substituents at key sites. The successful development of QSAR models for DCLK1 inhibition from small and highly congeneric compounds harkens back to the early days of QSAR modeling and underscores the continuing need to carefully inspect modeling datasets, through methods like MODI, prior to modeling.

After models were built and validated, they were used to virtually screen compounds bearing the same scaffolds as in **Figure 2.2**, resulting in the most potent series of DCLK1 inhibitors to date. Ultimately, six out of seven compounds prioritized by QSAR modeling as DCLK1 inhibitors had IC_{50} values < 10 μM, and four of these were sub-micromolar inhibitors (**Table 2.2**). On the other hand, two compounds predicted to be inactive by the models, JWE-067 and JWE-041, were shown experimentally to inhibit DCLK1 (**Table 2.2**). Since the goal of any drug discovery campaign is to identify compounds with the desired biological profile, in this case DCLK1 inhibition, ironically, this misclassification is not a failure.

Overall, eight compounds from the virtual screen, all of which possessed the 5,11-dihydro-6H-benzo[e]pyrimido[5,4-b][1,4]diazepin-6-one scaffold, had $IC_{50} < 10$ μM, including five sub-micromolar inhibitors (see **Table 2.2**). XMD13-44, a 52 nM inhibitor, was the most potent compound identified through virtual screening, which highlights that these models are capable of not only classifying inhibitors (actives) from non-inhibitors (inactives), but also of identifying highly potent compounds within the same chemical series. By any measure, the hit-rate from QSAR-based virtual screening of compounds (~73%) is enriched in comparison to the ~13% active-calls from the preliminary kinome screens of these same scaffolds (**Table 2.1**).

Chemical probe development also requires selectivity against off-targets.[67,77] The selectivity profiles of the eight compounds with $IC_{50} < 10$ μM for DCLK1 were assessed and quantified according to SI(65) and SI(90) (**Table 2.3**). These selectivity indices are a measure of a compound's kinome promiscuity at certain thresholds of inhibition. For a compound to be considered for possible inclusion into the KCGS developed by the SGC, in addition to sufficient on-target potency, the compound ought to have SI(65) < 4.0% and SI(90) < 2.0%.[66] All eight potent DCLK1 inhibitors were screened against at least 239 additional kinases. Only XMD13-37 did not meet the SI(65) criterion, as it inhibited more than 4% of kinases screened. The most potent DCLK1 inhibitor, XMD13-44, had an acceptable selectivity profile, inhibiting the enzymatic activity of only three off-target kinases by more than 65% (**Supplementary Table 2**). The five sub-micromolar DCLK1 inhibitors, therefore, could be considered useful tool compounds in the KCGS and high-quality starting points for further probe optimization efforts. It should be noted, however, that compounds with this scaffold have been previously reported to competitively inhibit LRRK2,[71] ERK5,[70,78] Aurora A/B,[79] and PI3K-δ/γ kinases[80] and to bind to BRD4 bromodomains[80]. Indeed, the most potent DCLK1 inhibitor, XMD13-44, inhibited ERK5 by more than 90%

(**Supplementary Table 2).** Optimizing DCLK1-selective compounds from this scaffold is, therefore, an on-going effort.[70,80,81]

To this end, SAR analyses were performed to inform optimization efforts through the identification of chemical structures and key protein-ligand interactions that drive DCLK1 inhibition. Cheminformatics approaches, *i.e.*, MMP analysis, QSAR model interpretation, and molecular docking, provided hypotheses about the underlying chemical features and molecular interactions that drive DCLK1 inhibition for lead compounds bearing the 5,11-dihydro-6H-benzo[e]pyrimido[5,4-b][1,4]diazepin-6-one scaffold. These cheminformatics approaches can also be applied to design and optimize follow-up compounds for both potency and off-target selectivity considerations.

In the present study, MMP analysis and model interpretation of fragment descriptors of compounds both in the training set and in the experimentally validated set revealed useful structural insights. For certain compounds possessing the 5,11-dihydro-6H-benzo[e]pyrimido[5,4-b][1,4]diazepin-6-one scaffold (**Figure 2.2**), a methoxy substituent to the phenyl ring (R1) and two co-occurring methyl substituents at R2 and R3 on the diazepine ring were shown to correlate with DCLK1 inhibition. The two most potent hits identified from the virtual screening, XMD13-44 and XMD8-90, align with this observation, as both compounds possess these features at R1, R2, and R3. Similarly, two inactive compounds lack some of these features: XMD11-100 lack methyl substituents at R2 and R3, while XMD11-40-2 lacks the methyl group at R2, though both compounds have a methoxy group at R1. Fragment descriptor analysis indicated that a substituent at the R3 position promoted DCLK1 inhibition (**Figure 2.3**) and may be considered a possible site for future optimization.

While the SAR trends around these three positions correlate with DCLK1 inhibition, not all of them are necessarily required for the desired activity and potency. For instance, the highly potent compound TL-1-038 (109 nM) does have two co-occurring methyl substituents at R2 and R3 on the diazepine ring, but lacks the methoxy group at R1. These derivations from the SAR trends highlight the importance of multivariate features in molecular design, as TL-1-038 possesses a piperidinol group attached to the aniline moiety that is unique among all hits. In fact, potent inhibitors XMD13-44, XMD8-90, and TL-1-038 all differ at this tail region off of the aniline moiety (**Table 2.2**), making it a possible site for further medicinal chemistry efforts.

Molecular docking provided additional insights to SAR trends among MMPs.[82] XMD10-39, a 154 nM inhibitor, differs from training set active compound XMD8-85 and from training set inactive XMD10-78 by a single ethyl substitution at R2. This observation indicates that, for the most part, both methyl and ethyl substitutions at the R2 position are tolerated for DCLK1 inhibition, whereas an isopropyl is not. This slight change in structure that results in a large change in activity is reflected by the molecular docking results (**Figure 2.6**), which shows that the binding pose of inactive XMD10-78 (red) is flipped in the ATP-binding site relative to active compounds XMD8-85 (teal) and XMD10-39 (green). It is worth noting that JWE-067 (0.265 μM) possesses an isopropyl substitution at the R2 position; however, its structure and binding pose are considerably different from the MMPs mentioned above (SI).

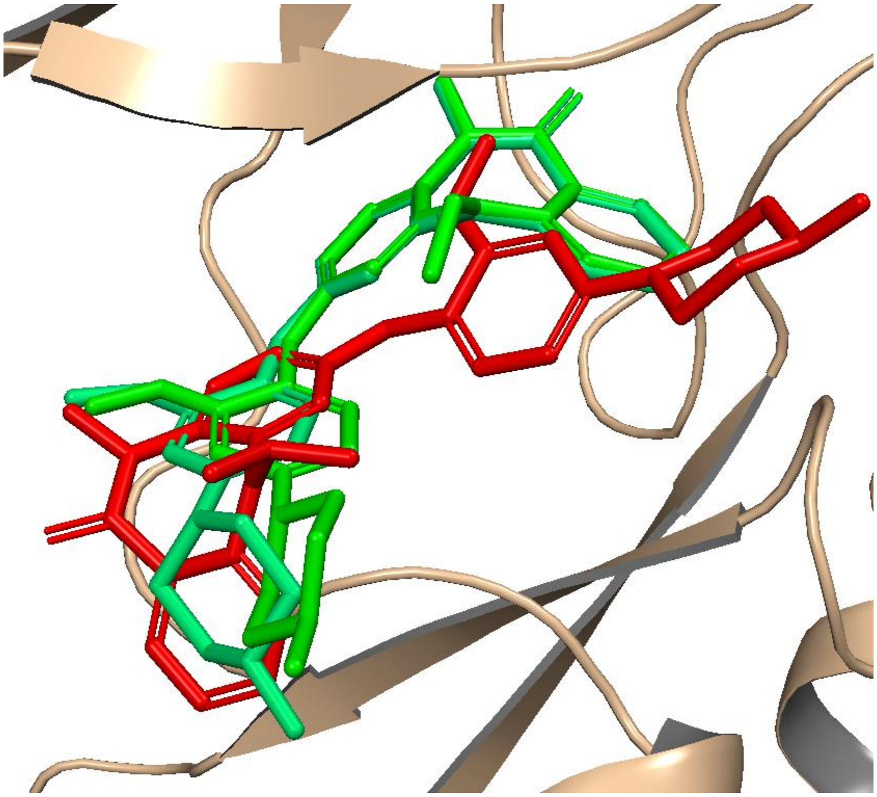


Figure 2.6 Docking of matched molecular pairs. The binding pose of inactive XMD10-78 (red) is flipped in the ATP-binding site relative to active compounds XMD8-85 (teal) and XMD10-39 (green). Note that the 5,11-dihydro-6H-benzo[e]pyrimido[5,4-b][1,4]diazepin-6-one scaffold of the active compounds are perfectly aligned.

Docking provided additional hypotheses about the key protein-ligand interactions for other DCLK1 inhibitors. Using the two most potent compounds as examples, which also have the two best docking scores, key hydrogen-bonds are likely formed with the backbone of the hinge-region valine in the ATP-binding by the 3-position pyrimidine nitrogen and the aniline NH site (**Figure 2.7**).

The recommended modifications to the four most potent lead compounds, *i.e.*, XMD13-44, XMD8-90, XMD10-39, and TL-1-038 (**Table 2.2**), are summarized in **Figure 2.8**. Based on the SAR analysis, a methoxy substitution at R1 paired with either a methyl or ethyl substitution at R2 promotes DCLK1 inhibition (**Figure 2.2**). A methyl substituent at the R3 position is present in all four lead compounds (**Table 2.2**) and was shown to promote DCLK1 inhibition among training set MMPs (**Figure 2.3**). At minimum, a methyl substituent at this position is required. Docking studies also suggest that additional hydrogen bonds may be formed by larger chemical substituents at R3 and residues in the pocket (**Figure 2.7**). This site should be considered for further optimization and SAR studies. Finally, the R4 region off of the aniline moiety is varied among the leads and the training set MMPs, making it another possible site for optimization and combinatorial design.

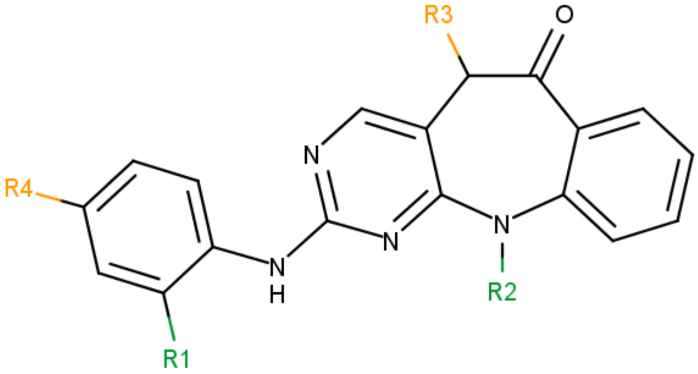


Figure 2.7104. SAR analysis and implications for design. Recommended modifications to the four most potent lead compounds at positions R1-R4 are shown. Sites known to modulate DCLK1 activity are shown in green. Sites proposed for DCLK1 optimization are shown in orange.

2.5 CONCLUSIONS

The integration of QSAR-based virtual screening and experimental medicinal chemistry (**Figure 1.1**) used in this study resulted in the discovery of the most potent and selective series of DCLK1 inhibitors to date. We succeeded to develop robust and predictive QSAR models of DCLK1 inhibition, despite challenges in the available data, which included limited dataset sizes and a very small amount of active compounds. These models were used hen to screen a set of 169 compounds. Ultimately, five sub-micromolar inhibitors with a 5,11-dihydro-6H-benzo[e]pyrimido[5,4-b][1,4]diazepin-6-one scaffold were identified using this approach. The top hit, XMD13-44, had an IC_{50} of 52 nM. Subsequent cheminformatics-based SAR analyses demonstrated that the activity of these compounds depends on certain of structural features. The selectivity profiles of these potent inhibitors against off-target kinases demonstrated their potential utility as DCLK1 tool compounds, though further optimization efforts are currently on-going. Joint modeling and experimental efforts such as those presented here may help accelerate the rate of discovery among dark and understudied kinases. Our study demonstrates that even despite the obvious lack of data for understudied targets, integration of computational and experimental approaches accompanied by close interaction and collaboration between modeling and medicinal chemistry groups can lead to the discovery of potent inhibitors. The significant advantage of this methodology is its easy translation to other kinases and targets, dark or otherwise, for future discovery efforts.

2.6 METHODS

2.6.1 Data Production, Collection, Curation, and Classification

Compounds were synthesized and screened in KINOMEscan[83] and KiNativ[84] assays. Compounds tested in these screens feature four scaffolds based on a common pyrimido-diazepine core (**Figure 2.2 A-D**).[69]

A total of 53 compounds were tested in a kinome-wide screen (>200 kinases) at 10 μM by the KINOMEscan assay. Briefly, KINOMEscan is a competition-based assay in which the kinase of interest is either baited to an immobilized bead by interacting with an active-site directed small molecule or displaced from the bead by binding to the test compound. Compounds with experimental values less than or equal to 1%, *i.e.,* 1% of the kinase not displaced, were considered “active”, while all compounds with values greater than 1% were considered “inactive”. This compound classification resulted in 8 actives and 45 inactives for DCLK1.

A total of 47 compounds were tested in the KiNativ assay in a kinome-wide screen (>200 kinases) at 10 μM. KiNativ is a binding assay in which pretreatment of a kinase with the test compound prevents a covalent interaction between the kinase and a standard probe to take place. Compounds with greater or equal to 50% inhibition were considered "active", while all compounds with less than 50% inhibition were considered "inactive". This compound classification resulted in 5 actives and 42 inactives for DCLK1.

Eleven compounds were tested in both the kinome-wide profiling KINOMEscan and KiNativ assays, with an approximately 82% concordance in classification for activity against DCLK1 (**Supplementary Table 4**). However, due to an insufficient amount of overlap between the two assays, data were not integrated. Instead, two separate models were developed. Prior to modeling, all compounds were curated according to our well-established protocols.[36–38] Training set compounds are available as sdf files in the Supporting Information (Supplementary Files 1 and 2)

2.6.2 Computational Methods

Following the best practices of QSAR modeling advanced earlier by our group,[9] two independent models, *i.e.*, for KINOMEscan and KiNativ data, were developed using the GUSAR modeling package[85,86]. GUSAR is a proprietary software; however, models are available in the Supporting Information (Supplementary Files 3 and 4) and could be used by anyone possessing the software package.

2.6.2.a Molecular descriptors

GUSAR uses a combination of whole-molecule descriptors and 2D fragment descriptors.[85,86] The whole-molecule descriptors generated by GUSAR are topological length and volume; lipophilicity; molecular weight; and numbers of aromatic and halogen atoms, positive and negative charges, and hydrogen bond donors and acceptors. The 2D fragment descriptors are of two types: multilevel neighborhoods of atoms (MNA)[27] descriptors and quantitative neighborhoods of atoms (QNA) [28].

2.6.2.b RBF-SCR algorithm

In this machine-learning algorithm, descriptors are weighted during the calculation of the radial basis function (RBF) by the coefficients obtained from self-consistent regression (SCR).[87] These coefficients reflect the contribution of each particular descriptor to the final equation for the given activity. The absolute value of the coefficient corresponds to its contribution. In the RBF-SCR method, the weights for each descriptor vector used for the calculation of RBF are based on that descriptor's importance for the given activity as determined by SCR.

2.6.2.c Modelability Index (MODI)

The Modelability Index (MODI) estimates the likelihood of obtaining predictive QSAR models for a binary data set of compounds.[44] MODI is defined as a weighted ratio of the number of nearest-neighbor pairs of compounds in descriptor space with the same activity class versus the total number of pairs. MODI threshold of 0.65 was previously found to separate the modelable from non-modelable data sets. MODI was calculated for both training sets to evaluate the feasibility of developing models with high predictive power for the training sets.

2.6.2.d QSAR modeling

Binary classification (active vs. inactive) QSAR models were developed and rigorously validated according to the best practices of QSAR modeling.[9] Models utilized a combination of MNA and QNA descriptors[85,86] and RBF-SCR[87] as the machine-learning algorithm. All models were validated using five-fold external cross validation;[9,73] Y-randomization[88] and applicability domain (AD)[43] were utilized for each QSAR model. Models were statistically evaluated according to sensitivity (SE), specificity (SP), correct classification rate (CCR), positive predictive value (PPV), and negative predictive value (NPV), see the equations 1-5, respectively.

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \qquad (1)$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \qquad (2)$$

$$\mathrm{CCR} = \frac{\mathrm{SE+SP}}{2} \qquad (3)$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \qquad (4)$$

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN+FN}} \qquad (5)$$

Here, TP and TN represent the number of true positives (correct classifications of actives), and true negatives (correct classifications of inactives), respectively; whereas, FP and FN represent the number of false positives (incorrect classifications of actives) and false negatives (incorrect classifications of inactives), respectively. Models were deemed acceptable for virtual screening if and only if the CCR, SE, SP, PPV, and NPV were above 0.60, and no associated Y-randomized model had a CCR above 0.60.

2.6.2.e Virtual screening

A set of 169 compounds with pyrimido-diazepine scaffolds (**Figure 2.2**) considered synthetically feasible and possibly viable as DCLK1 inhibitors were used for prediction and prioritization. As above, the 2D structures were processed following the same curation protocols.[37] Consensus prediction was utilized, meaning that for a compound to be considered a “hit”, it must be within the applicability domain of each model and be predicted as “active” by both.

2.6.2.f Matched Molecular Pair Analysis

The KNIME implementation of the Hussain and Rea algorithm for identifying Match Molecular Pairs (MMPs) was utilized.[89] The number of cuts to acyclic single bonds was set to 1, and the maximum number of heavy atom changes was set to 6 to allow for the addition of six-membered rings.

2.6.2.g Molecular Docking

The lowest energy configurations at pH 7 for the eleven experimentally validated compounds, as well as compounds in the training sets, were generated with the Ligand Preparation application (Schrödinger Release 2016-4: LigPrep, Schrödinger, LLC, New York, NY, 2016). Next, the crystal structure of DCLK1 (PBD: 5JZN)[57] was downloaded from the Protein Databank (http://www.rcsb.org/pdb/home/home.do)[90] and prepared for docking as follows. Briefly, bond orders were assigned, hydrogen atoms were added, selenomethionine residues were converted to methionines, and missing side chains were filled with the Prime application (Schrödinger Release 2016-4: Prime, Schrödinger, LLC, New York, NY, 2016). The protein grid for docking was set by selecting the center of the cocrystal ligand in the PDB structure[57]. The Glide application was used for docking of all compounds into DCLK1 (Schrödinger Release 2016-4: Glide, Schrödinger, LLC, New York, NY, 2016). Flexible ligand sampling was allowed with the XP (extra precision) scoring function.[91]

2.6.3 Experimental Methods

2.6.3a DCLK1 Plasmid Construction

The DNA construct consisting of N-terminally 6-His tagged human DCLK1 residues G351-H689 was obtained from Ana Clara Redondo of the Structural Genomics Consortium at the University of Oxford. The plasmid was co-transformed with lambda phosphatase under chloramphenicol selection into BL21 DE3 *E.coli* cells.

2.6.3b DCLK1 Protein Purification

Protein expression was induced with 0.6 mM IPTG and expression was allowed to continue for ~10 hours at 18 °C. Bacteria were harvested by centrifugation and resuspended in Lysis buffer with protease inhibitors (1 mM Benzamidine and 1 mM PMSF). Lysis was performed by passing 3 times through a homogenizer. Lysate was centrifuged at 20K for 1 hour, and the supernatant was filtered through a 0.2 micron membrane. Protein was captured using Nickle-NTA resin and eluted with imidazole. Eluate was concentrated to 2 mL and passed over a Superdex S200 column. The Lysis buffer was composed of 50mM Hepes (pH 7.8), 350 mM NaCl, 20 mM imidazole, and 5% glycerol. The first wash was the same as the Lysis buffer; the second was the same as Lysis buffer but with 25 mM Imidazole; the elution was the same as Lysis buffer but with 300 mM Imidazole. The S200 gel filtration buffer was composed of 10mM Hepes (pH 7.8), 700 mM NaCl, 1 mM $MgCl_2$, and 5% glycerol.

2.6.3c DCLK1 Mobility Shift Assay

DCLK1 kinase activity was measured *in vitro* using an electrophoretic mobility shift assay. The reaction was assembled in a 384-well plate in a total volume of 20 μl. The reaction comprised 30 nM recombinant DCLK1, one DCLK1 inhibitor or DMSO, 100 μM ATP and 1 μM FAM-labeled peptide substrate (peptide 12: 5-FAM-KKLRRTLSVA-COOH) in a buffer (100 mM HEPES pH 7.5, 0.003% Brij-35, 0.004% Tween-20, 10 mM $MgCl_2$, and 2 mM DTT). DCLK1 inhibitors were dispensed using a Labcyte Echo liquid handler. The reaction was incubated at room temperature for two hours and quenched by addition of 40 μL of termination buffer (100 mM HEPES pH 7.3, 0.015% Brij-35, 0.1% CR-3, 1 x CR-8, and 40 mM EDTA). Substrate and product peptides present in each sample were electrophoretically separated and detected using 12-channel

LabChip3000 microfluidic capillary electrophoresis instrument (Caliper Life Sciences). The change in the relative fluorescence intensities of substrate and product peaks (reflecting enzyme activity) was measured. Capillary electrophoregrams were analyzed using HTS Well Analyzer software (Caliper Life Sciences). The kinase activity in each sample was determined as the product-to-sum ratio (PSR): $P / (S + P)$, where P is the peak height of the product peptide and S is the peak height of the substrate peptide. Negative control samples (DMSO in the absence of inhibitor) and positive control samples (100% inhibition, a tested DCLK1 inhibitor) were assembled in replicates and were used to calculate percent inhibition values for each compound at each concentration. Percent inhibition (%Inhibition) was determined using equation 6:

$$\%\text{Inhibition} = 100 \times \frac{(PSR_{0\%} - PSR_{inh})}{(PSR_{0\%} - PSR_{100\%})} \quad (6)$$

where PSR_{inh} is the product-sum ratio in the presence of inhibitor, $PSR_{0\%}$ is the average product-sum ration in the absence of inhibitor and $PSR_{100\%}$ is the average product-sum ratio in 100%-inhibition control samples. The DCLK1 candidate inhibitors were tested in 8-point dose-response format on each assay plate. The IC_{50} values were determined by fitting the inhibition curves by an eight dose-response model using GraphPad Prism 7 software.

2.6.4 Kinase Selectivity

Kinome-wide profiling in the KiNativ assay was used to assess kinase promiscuity of potent DCLK1 inhibitors. [84] Compounds were screened against at least 239 kinases at 10 μM, and a single percent inhibition value was recorded. PC3 and HeLa cells were used, as well as multiple labeling sites.

For each compound, the selectivity index (SI) was assessed at thresholds of 65% and 90% kinase inhibition, *i.e.*, SI(65) and SI(90). The SI was then computed according to equation 7:

$$SI(90) = 100 * \frac{N_{hits}}{N_{total}} \qquad (7)$$

In this equation, SI(90) is the selectivity index at 90% inhibition, N_{hits} is the number of kinases inhibited by the compound at ≥ 90%, and N_{total} is the total of kinases against which the compound was screened. Likewise, SI(65) is the selectivity index at 65% inhibition.

2.7 ASSOCIATED CONTENT

Supporting Information. Supplementary Files and Tables for Chapter 2 are available at https://scapuzzi.web.unc.edu/free-downloads/

CHAPTER 3: PHANTOM PAINS: A PUBCHEM-WIDE ANALYSIS OF PAN-ASSAY INTERFERENCE COMPOUNDS[2]

3.2 INTRODUCTION

The scientific community is in the grips of the data reproducibility crisis, highlighted by *Nature*'s "Challenges in Irreproducible Research" initiative.[92,93] Oftentimes in drug discovery, compounds active in primary biological screens show no activity in follow-up studies.[94–96] The measured effect of false positives may be due to various compound and assay liabilities including those that interfere with the assay detection technology such as auto-fluorescence, hydrogen peroxide production, metal chelation, chemical aggregation, etc.[24,93–96]

We categorize compound and assay liabilities, as Type 1 and Type 2 behavior, respectively. Type 1 behavior is characterized by *compounds* that affect target activity by an undesirable mechanism of action (MoA).[97] An example of Type 1 behavior is target engagement by colloidal aggregation. The formation of compound aggregates is considered an undesirable MOA since these aggregates, not an individual compound, modulate the target and thus constitutes artefactual activity. Type 2 behavior is characterized by an *assay* result that gives a spurious measurement of compound activity against the target. Examples of Type 2 behavior include singlet oxygen quenching in the AlphaScreen assay platform or inhibition of reporter enzymes by screened compounds. Since the assay itself gives rise to artefacts, Type 2 behavior can be summarized as "assay interference."

[2] This chapter previously appeared as an article in the Journal of Medicinal Chemistry. The original citation is as follows: Capuzzi, S. J., Muratov, E. N., and Tropsha, A. "Phantom PAINS: Problems with the Utility of Alerts for P an-A ssay IN terference Compound S". *J. Chem. Inf. Model.* (February, 2017) *57*, 417–427.

Although assay interference can typically be determined by running orthogonal assays that probe activity with auxiliary readout technologies, *in silico* methods are sought to quickly (and without cost) identify compounds whose "activities" are attributable to interference. The most popular and promulgated computational tool to flag, and often triage, compounds that interfere with the bioactivity detection technology are so-called **P**an-**A**ssay **IN**terference Compound**S** (PAINS) filters.[24]

PAINS filters are composed of 480 "alerts", *i.e.,* substructural features frequently found in PAINS.[24] These alerts are computationally mapped and matched onto full compound structures to flag for potential liabilities.[24] In essence, PAINS alerts are simple binary (yes or no) models that attempt to forecast assay interference propensity by classifying compounds as PAINS (have alerts) or non-PAINS (lack alerts). Like other types of models, PAINS alerts were derived from a training set. This training set consisted of 7900 compounds tested in six AlphaScreen assays.[24] Similarly, like other types of models, PAINS alerts, to be considered reliable, must be validated with external data and be constrained by a defined applicability domain.[98]

During their original development, however, PAINS alerts were not validated using external data, *i.e.*, the predictive power of the alerts to identify compounds that interference with the assay detection technology was not quantified.[24] Moreover, the applicability domain of PAINS alerts has not been appropriately addressed in the published literature. This lack of a defined applicability domain has led to use of alerts beyond the original chemical space of the training set (Type 1 behavior) and beyond the AlphaScreen assay platform (Type 2 behavior).

Despite these concerns, the concept of PAINS alerts (filters) has gained much attention, many supporters, and prompted many follow-up publications.[99–101] Several web-based applications relying on the original work by Baell and Holloway[24] have been developed to flag and filter compounds with PAINS alerts;[102,103] chemical databases, such as ZINC (http://zinc15.docking.org/) and ChEMBL (https://www.ebi.ac.uk/chembl/), also flag compounds containing PAINS alerts. On the scientific blogosphere, publications reporting compounds flagged with PAINS alerts as viable hits have been publicly ridiculed in a practice known as "PAINS-Shaming."[104]

The wide acceptance of the PAINS concept by the scientific community and the availability of PAINS filters have made it common for researchers to triage virtual screening hits flagged with these alerts prior to experimental validation.[105] Similarly, lead compounds resulting from experimental screening campaigns have typically been de-prioritized for follow-up studies if they contained PAINS alerts.[106] Furthermore, scientific journals have begun to recommend that all hit compounds, virtual or otherwise, should be passed through one of the publicly available PAINS filters before the manuscript is considered for publication. For instance, the ACS Journal of Medicinal Chemistry requires that "active compounds from any source must be examined for known classes of assay interference compounds".[107] The authors are asked to "provide firm experimental evidence in at least two different assays that reported compounds with potential PAINS liability are specifically active and their apparent activity is not an artifact".[107] Thus, compounds with potential PAINS liability as those flagged with PAINS alerts.

Amidst the generally wide acceptance of PAINS, there have been a few voices cautioning about the overarching utility of the alerts. Several authors have noted that the application of these alerts could discard viable drug candidates because such alerts have actually been found in approved drugs.[108,109] More substantial criticism of PAINS alerts has emerged as well on internet forums (but not in peer-reviewed publications). Aware of these concerns, in the course of our own recent virtual screening investigations, we re-examined the original study[6] from which the 480 PAINS alerts were derived. We noticed that the study[6] employed a relatively small (93K compounds) and proprietary library (complete chemical structures were not released) tested for one type of activity (protein-protein interaction inhibition) in just six HTS campaigns (three out of six targets were kept confidential) using a single detection technology (AlphaScreen^{TM}).

Though considerable effort was made to divulge as much information as possible, due to the proprietary nature of the original study and unavailability of the chemical library explored therein, the detection of PAINS and the derivation alerts could not be fully and independently reproduced. That being said, upon further inspection of the 92 pages of Supplemental Information[24], we observed that more than half of the PAINS alerts were derived from one or two compounds only (**Figure 3.1**), with 68% (328 out of the 480 alerts) found in four or fewer compounds only, and more than 30% (190 PAINS alerts) found in one compound only showing "pan-assay" activity (**Figure 3.1; Table S1**). This preliminary analysis lead us to hypothesize that the majority of these alerts may have limited extrapolative power due to the constrained applicability domain.

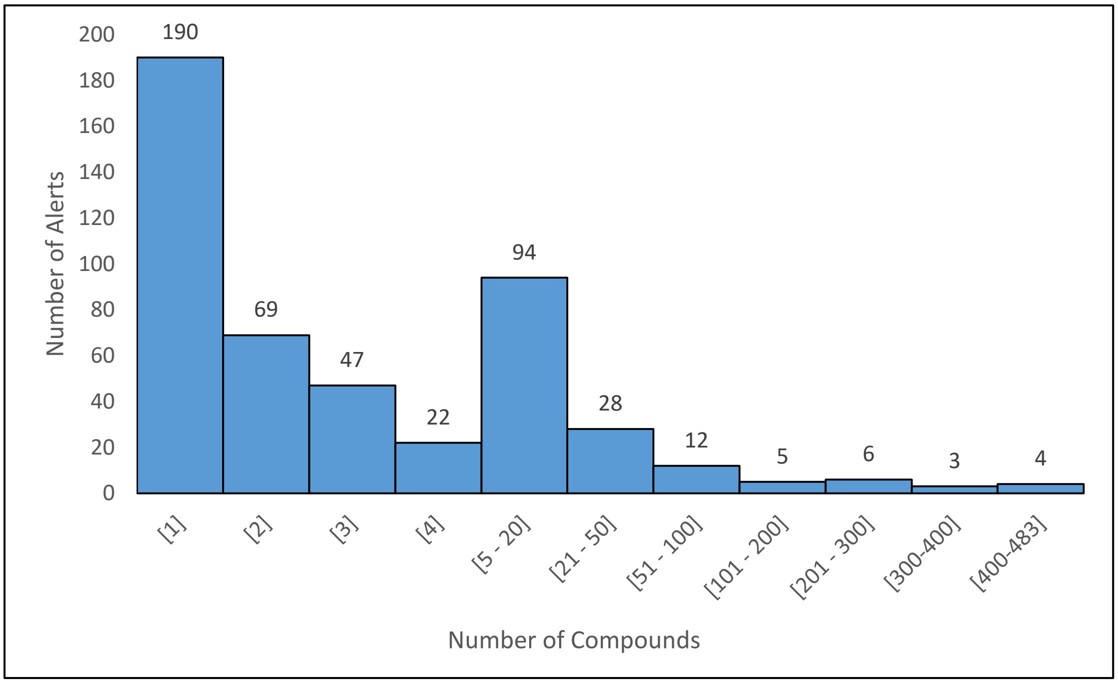


Figure 3.1. **Probing the extrapolative power of PAINS alerts.** A histogram showing the distribution of the number of PAINS alerts (amounting to 480 total) as the function of the number of compounds used to derive each alert. Note that 190 PAINS alerts were derived from one representative compound only whereas only 18 PAINS alerts were derived from samples including more than 100 compounds per alert.

Given the aforementioned limited applicability domain of PAINS alerts, we decided to probe into the "pan-assay" activity of PAINS and the reliability of PAINS alerts by analyzing publicly available data on extensively assayed compounds. To this end, we have (i) assessed the robustness of PAINS alerts at flagging frequent hitters among compounds assayed using the AlphaScreen^{TM} technology as reported in PubChem; (ii) scanned PubChem to investigate the level of "pan-assay" activity of compounds with and without PAINS alerts; (iii) examined the frequency of PAINS alerts in extensively assayed, yet consistently inactive compounds known as "Dark Chemical Matter"[110]; and (iv) profiled the PubChem-wide activity of FDA-approved drugs with and without PAINS alerts. Overall, using publicly available data, this study sought to evaluate the

PAINS concept in general, with an additional focus on specific PAINS alerts established in the original investigation[24], in order to provide both researchers and journal editors with insight into the utility of PAINS alerts as they currently stand.

3.3 RESULTS AND DISCUSSION

3.3.1 Detection of PAINS in chemical libraries tested with the AlphaScreen^{TM} Technology

We have identified six PubChem assays that measure protein-protein interaction (PPI) inhibition using AlphaScreen^{TM}, *i.e.*, the same type of activity and the same technology employed in the original study.[6] The study design is shown in **Figure 3.2.** The six originally studied assays were run at the relatively high compound concentration of 25-50 μM in primary screens, which may account for the high rate of interference, and two of the assays used hexa-his/Ni anchors.[24] We have chosen these six PubChem assays, similar to a study by Schorpp et. al[111], in order to assess the robustness of PAINS alerts to flag frequent hitters across a similar, but not identical, series of assays. It should be noted that the anchorage and screening concentrations reported in PubChem were different from original study. However, current PAINS filters look solely for the presence of specific functional groups in assayed chemicals regardless of the assay conditions; thus, the difference in these conditions does not invalidate the use of PAINS alerts in this investigation.

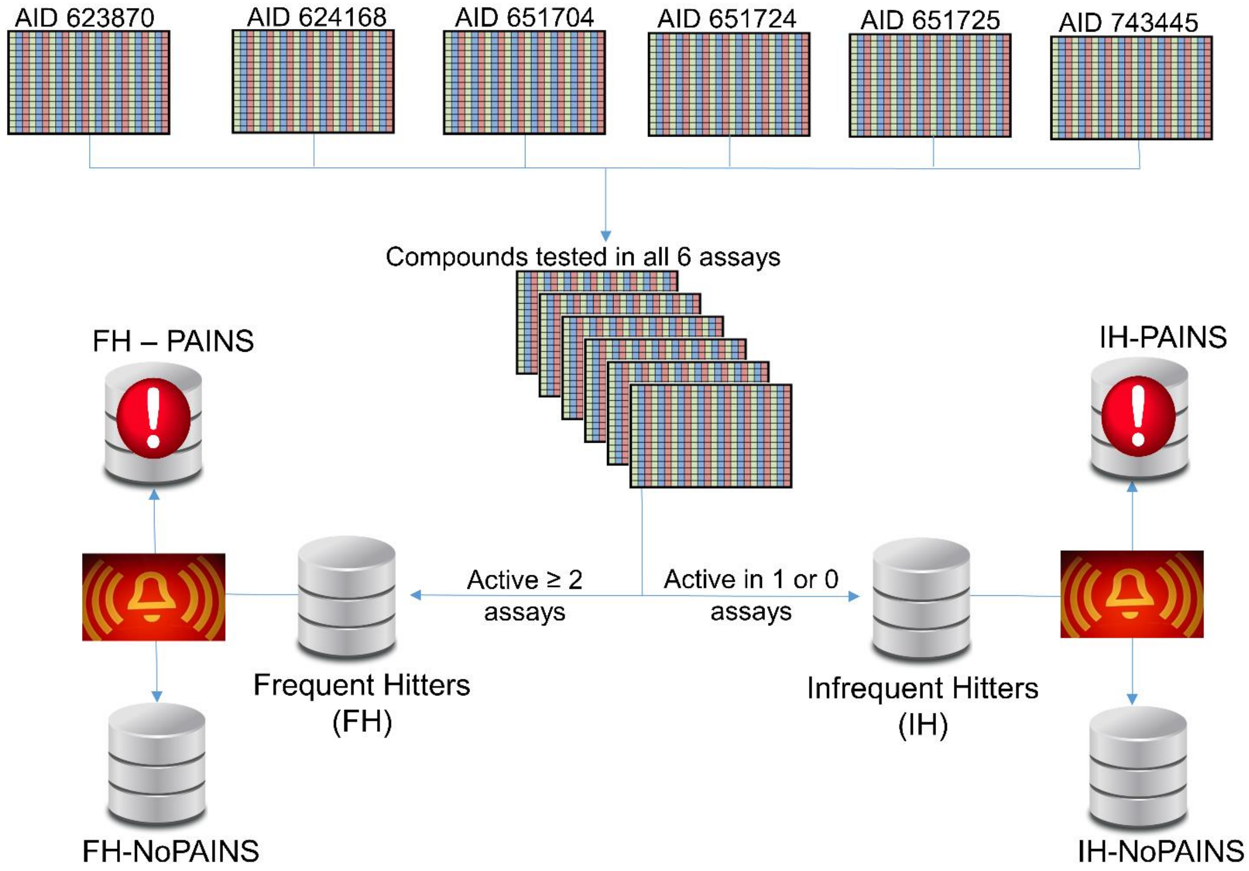


Figure 3. 2. **Study design for examining compounds in PubChem tested with AlphaScreen^{TM} assay technology.** Six assays targeting PPIs and using AlphaScreen^{TM} were identified in PubChem. Only those compounds that were tested in all six assays were considered. Compounds were binned into two categories according to the number of active calls. Compounds in each category were then queried for PAINS alerts. The PubChem-wide bioassay activity of all compounds was then investigated.

As many as 153,339 unique compounds were found in PubChem to have been tested across all six assays (**Table S2**). Activity calls for each compound (Active, Inactive, and Inconclusive) were recorded as defined by the assay depositor. Compounds were then binned into two categories: "Frequent Hitters" (active calls in at least 2 out of 6 assays) and "Infrequent Hitters" (active calls in 1 or 0 assays), which is the same threshold as established in the original study[24]. Both categories were first queried for the presence of PAINS substructural alerts using the SMARTS implementation from PubChem Promiscuity[25], then confirmed using SYBYL Line Notation (SLN) implementation from FAF-Drugs3[102]. Four categories arose: “Frequent Hitters - PAINS” (FH-PAINS), “Frequent Hitters – No PAINS” (FH-NoPAINS), “Infrequent Hitters - PAINS” (IH-PAINS), and “Infrequent Hitters- No PAINS” (IH-NoPAINS). There was a concordance of ~99.9% between the SMARTS and SLN implementations for flagging compounds with PAINS alerts (**Table S3**). The enrichment value (EV), which was previously defined[24] as the percentage of compounds active in at least two of the six assays relative to the number of compounds that displayed no activity across all six assays, was calculated to compare FH-PAINS *vs*. IH-PAINS (**Table S4 and S5**).

The results of our analysis are shown in **Table 3.1**. There were 902 compounds in the “Frequent Hitters” category, and 208 of these only (23%) contained PAINS substructural alerts (FH-PAINS). The remaining 694 “Frequent Hitters” lacked any PAINS alerts (FH-NoPAINS). For the “Infrequent Hitters”, 146,224 (96%) compounds lacked PAINS alerts (IH-NoPAINS), but 6,413 compounds (4%) still contained the alerts (IH-PAINS). Comparing the numbers of IH-PAINS and FH-PAINS leads to the apparent conclusion that, for this series of assays, the majority of compounds containing PAINS alerts (97%) were actually infrequent hitters.

Table 3.1. **Lack of pan-assay activity for compounds with PAINS alerts in PubChem.** The average fraction of activity calls for PAINS and non-PAINS (defined as containing or lacking

PAINS alerts, respectively) across both detection technology-specific assays and all assays in PubChem. The average number of assays in which the compounds were tested are shown in parentheses.

Compound Categories	$N_{compounds}$	Luciferase	β-lactamase	Fluorescence	All Assays
FH-PAINS*	208	12% (93)	4% (9)	7% (312)	10% (546)
FH-NoPAINS*	694	6% (95)	2% (9)	3% (320)	5% (550)
IH-PAINS*	6,413	3% (93)	1% (10)	2% (323)	2% (550)
IH-NoPAINS*	21,500	1.5% (95)	0.5% (9)	1% (326)	1% (555)
Random-PAINS**	14,611	3% (95)	1% (12)	2% (329)	3% (562)
Random-NoPAINS**	58,722	2% (93)	0.6% (13)	0.8% (321)	1% (550)
Drugs-PAINS**	87	9% (71)	7% (40)	6% (223)	24% (602)
Drug-NoPAINS**	1,373	5% (59)	5% (33)	3% (183)	15% (458)

*Defined by the compound profile in PPI assays utilizing AlphaScreen^{TM} ;

**Defined by presence or absence of PAINS alerts.

The enrichment value calculated for PAINS-containing compounds (FH-PAINS and IH-PAINS) was only 3.5% (**Table S4 and S5**). Furthermore, if IH-PAINS that were active in one assay only were taken into consideration, the overall EV fell to 3.2% (**Tables S4 and S5**). The analysis of this series of assays indicates that PAINS alerts are found much more frequently in non-promiscuous compounds.

To probe whether or not this observation is only limited to assays related to PPIs using AlphaScreen^{TM}, we investigated the PubChem-wide bioassay activity of the same compounds. Given that IH-NoPAINS constituted the overwhelming majority of all compounds discussed above (146,224 compounds), the PubChem-wide activity of all IH-NoPAINS was not evaluated due to computational constraints. Instead, a random subset of 21,500 IH-NoPAINS (**Table S7**) were evaluated; this number was selected to preserve approximately the same ratio of frequent to infrequent NoPAINS (1:3.5) as was observed for the PAINS (**cf. Table 3.1**). PubChem Promiscuity[25,112] was used to retrieve the activity calls for all four aforementioned categories of compounds (FH-PAINS, IH-PAINS, FH-NoPAINS, and IH-NoPAINS) tested in luciferase-, beta-lactamase-, and fluorescence-based assays (**See Tables S4-S7**). Lastly, we assessed activity calls across all bioassays in PubChem irrespective of the detection technology (**Table 3.1**).

We found that across all assays, including those that have been reported as particularly susceptible to interference,[96] FH-PAINS were active in more assays than FH-NoPAINS; however, IH-PAINS were active in fewer assays than FH-PAINS (**Table 3.1**). The reduced activity of IH-PAINS in the AlphaScreen^{TM} assays, therefore, is not limited to this detection technology, as it can be observed over all reported assays in PubChem. Also, both categories of frequent hitters in AlphaScreen^{TM} (FH-PAINS and FH-NoPAINS) showed greater PubChem-wide activity than infrequent hitters containing PAINS alerts (IH-PAINS). Therefore, the broader activity spectrum of these frequent hitters is independent of the presence or absence of any PAINS alerts, highlighting the importance of considering molecular entities as a whole rather than chemical fragments when trying to derive any structural rules governing assay promiscuity.

3.3.2 Analysis of PAINS alerts in chemical libraries tested with the AlphaScreen^{TM} Technology

The specific alerts found in the above FH-PAINS and IH-PAINS categories were then investigated on an individual basis. In total, 163 individual PAINS alert types were observed in compounds among the two categories (**Table S8**). It should be noted that multiple PAINS alerts could be present within a single compound (**Figure 3.3**).

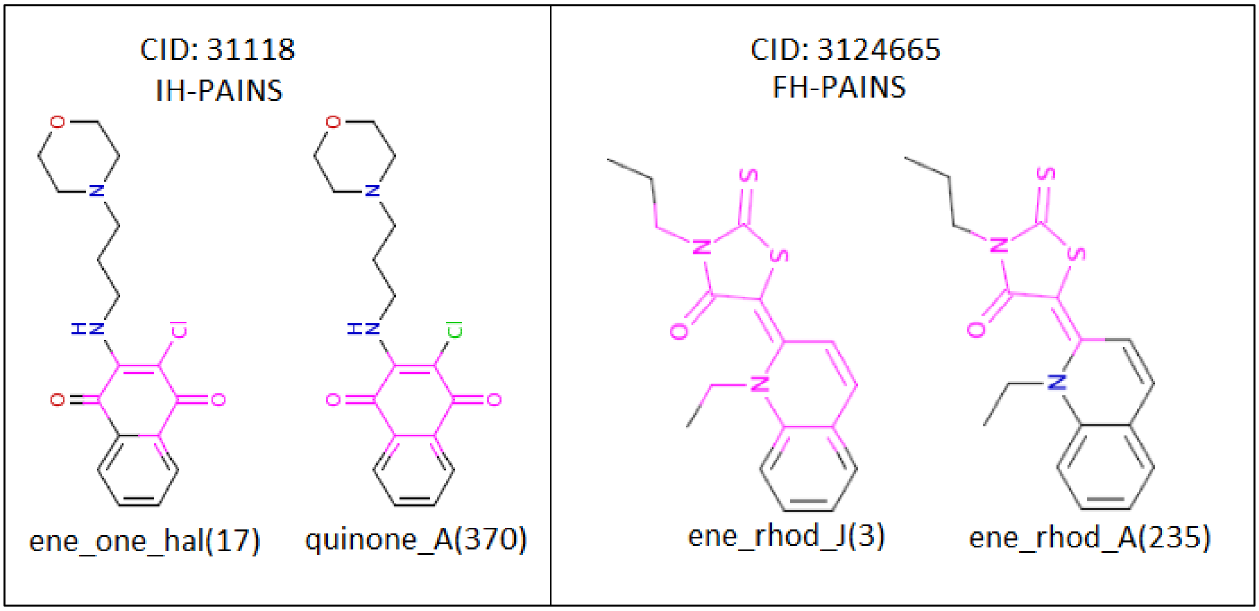


Figure 3. 3. Compounds with multiple PAINS alerts. Two representative compounds from the IH-PAINS and FH-PAINS categories that contain multiple PAINS alerts.

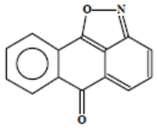
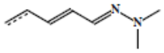
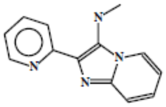
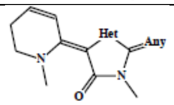
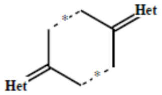
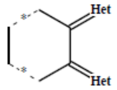
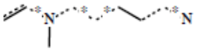
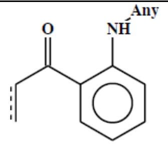
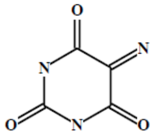
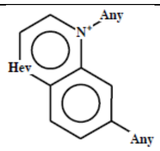
For the 208 FH-PAINS compounds, 41 individual PAINS alerts were detected (**Table S8**). Of these 41 alerts, only 7 alerts, *i.e.*, quinone_A(370), mannich_A(296), ene_six_het_A(483), anil_di_alk_B(251), anil_di_alk_A(478), ene_one_hal(17), and imine_one_A(321), were found in more than 10 FH-PAINS compounds. The remaining 34 PAINS alerts were present in 10 or less FH-PAINS compounds.

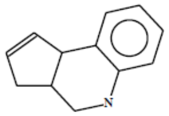
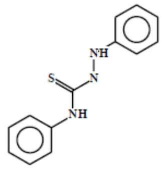
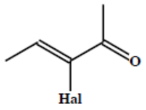
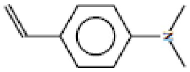
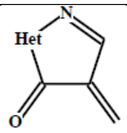
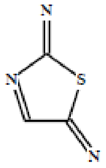
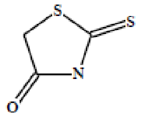
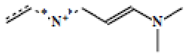
For the 6,314 IH-PAINS compounds, 162 individual PAINS alerts were detected (**Table S8**). Of these 162 alerts, 57 alerts were found in more than 10 IH-PAINS compounds. Moreover, 15 of these alerts were found in more than 100 IH-PAINS compounds. The anil_di_alk_A(478) alert, for example, appeared in 1,083 IH-PAINS compounds.

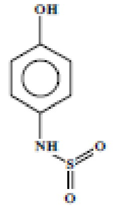
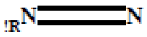
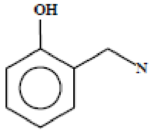
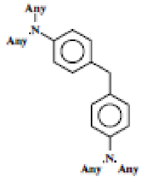
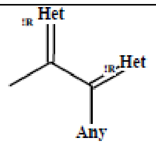
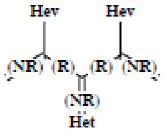
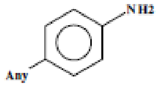
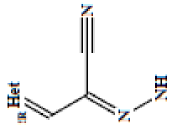
Next, the PAINS alerts that were present in both the FH-PAINS and IH-PAINS were analyzed. Within these two categories, 40 individual PAINS alerts were shared, roughly ~25% of all observed alerts. Only one alert, *i.e.*, anil_no_alk_A(1), was unique to FH-PAINS (**Table S8**); however, only 1 compound possessed this alert, which is consistent with the limited sample size (1 compound) used to derive this alert (**cf. Figure 3.1**). Similarly, 122 alerts were unique to IH-PAINS (**Table S8**). For this series of assays, ~75% PAINS alerts present in the PubChem library analyzed herein (122 out of 163) were found only in the Infrequent Hitters.

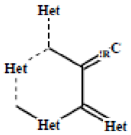
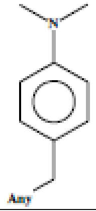
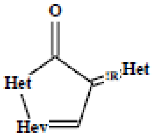
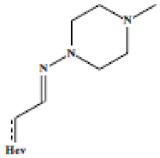
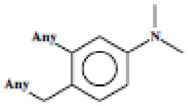
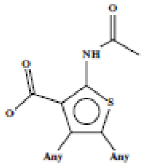
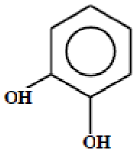
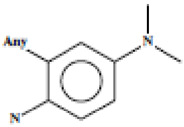
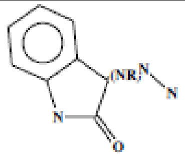
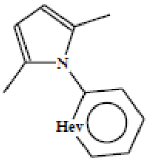
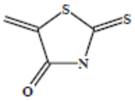
The enrichment value (EV), defined as percentage of compounds active in at least two of the six assays relative to the number of compounds active in 1 or 0 assays, was calculated for each of 40 shared PAINS alerts (**Table 3.2**). Only 6 alerts showed EVs greater or equal to 25%. However, 4 of these 6 alerts had less than 10 representative compounds. Therefore, 2 alerts, *i.e.*, quinone_A(370) and quinone_D(2), were found in 10 or more compounds and had an EV greater or equal to 25%. The remaining 34 shared alerts had EVs less than 25%, and 32 of these 34 alerts had more than 10 representative compounds. Indeed, 6 shared alerts had EVs less than 1.0% despite being present in more than 100 compounds. For this series of assays, the vast majority of PAINS alerts were found among the Infrequent Hitters (IH-PAINS) at much higher frequencies. The full analysis of all 40 shared alerts, including representative compound sizes and EVs, can be found in **Table 3.2**.

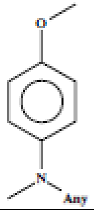
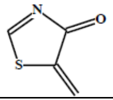
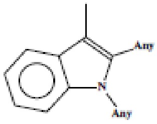
Table 3.2. PAINS enrichment in six PubChem assays employing AlphaScreen^{TM}. Forty alerts were present in both FH-PAINS and IH-PAINS. Two alerts showed EVs greater or equal to 25% and were found in 10 or more total compounds. Six alerts had EVs below 1.0%.

PAINS Alert	Substructure[24]	$N_{FH\text{-}PAINS}$	$N_{IH\text{-}PAINS}$	N_{PAINS}	EV, %
quinone_B(5)		3	1	4	300.0
imine_ene_A(5)		3	2	5	150.0
het_65_Db(5)		4	3	7	133.3
ene_rhod_J(3)		1	3	4	33.3
quinone_A(370)		47	160	207	29.4
quinone_D(2)		9	36	45	25.0
dyes5A(27)		1	5	6	20.0
anthranil_one_A(38)		3	16	19	18.8
imine_one_sixes(27)		3	16	19	18.8
het_pyridiniums_A(39)		5	29	34	17.2

anil_alk_ene(51)		2	12	14	16.7
thio_urea_D(8)		1	6	7	16.7
ene_one_hal(17)		14	89	103	15.7
anil_di_alk_B(251)		16	116	132	13.8
ene_five_het_A(201)		7	51	58	13.7
het_thio_5_imine_A(1)		1	9	10	11.1
rhod_sat_A(33)		4	39	43	10.3
dyes3A(19)		1	11	12	9.1

sulfonamide_B(41)		2	25	27	8.0
azo_A(324)		9	114	123	7.9
mannich_A(296)		36	472	508	7.6
anil_di_alk_F(14)		1	14	15	7.1
imine_one_A(321)		12	215	227	5.6
ene_one_ene_A(57)		2	40	42	5.0
anil_no_alk(40)		2	48	50	4.2
cyano_imine_A(37)		1	26	27	3.9

ene_six_het_A(483)		18	556	574	3.2
anil_di_alk_D(198)		5	155	160	3.2
imine_one_fives(89)		1	38	39	2.6
hzone_pipzn(79)		2	87	89	2.3
anil_di_alk_E(186)		3	148	151	2.0
thiophene_amino_Ab(40)		1	71	72	1.4
catechol_A(92)		1	75	76	1.3
anil_di_alk_A(478)		14	1083	1097	1.3
imine_one_isatin(189)		1	111	112	0.90
pyrrole_A(118)		2	269	271	0.74
ene_rhod_A(235)		4	593	597	0.67

anil_di_alk_C(246)		4	641	645	0.62
ene_five_het_B(90)		1	161	162	0.62
indol_3yl_alk(461)		1	354	355	0.28

3.3.3 Random PAINS in PubChem

We also evaluated the PubChem-wide activity of compounds tested in at least 25 separate bioassays based only on the presence or absence of 480 originally established PAINS alerts, *i.e.*, irrespective of any perceived promiscuity across a selected series of specific assays. Randomly selected compounds that were evaluated in the previous section were excluded. The resultant dataset contained 73,333 individual compounds. The structures of these compounds were searched for PAINS alerts (described above) and binned into two categories: Random-PAINS (14,611 compounds) and Random-NoPAINS (58,722 compounds). We compared these two categories following the same protocol as described in the previous section (**Table 3.1**). The average pan-assay activity of Random-PAINS was just 3%, compared to an average of 1% for Random-NoPAINS (**Table 3.1**), *i.e.*, Random-PAINS were marginally more active than Random-NoPAINS. Additionally, of the 14,611 Random-PAINS only 752 compounds (5% of the total) showed activity in at least 10% of all assays. Of the remaining 13,859 Random-PAINS (95%) that were active in less than 10% of all assays, 1,146 had no activity at all, despite being tested in an average of 443 assays (**Table S9 and S10**). These results indicate that the mere presence of a PAINS substructure does not give rise to any observed pan-assay activity, nor any marked interference trends in luciferase-, beta-lactamase-, or fluorescence-based assays. In fact, only two PAINS alert containing compounds, tanespimycin and dihydrexidine, were active in more than 50% of the assays (**Figure 3.4**). In total 202 PAINS alerts were found among the Random-PAINS category, A PubChem-wide analysis of alerts in random PAINS is described in the Global Analysis of PAINS Alerts Section.

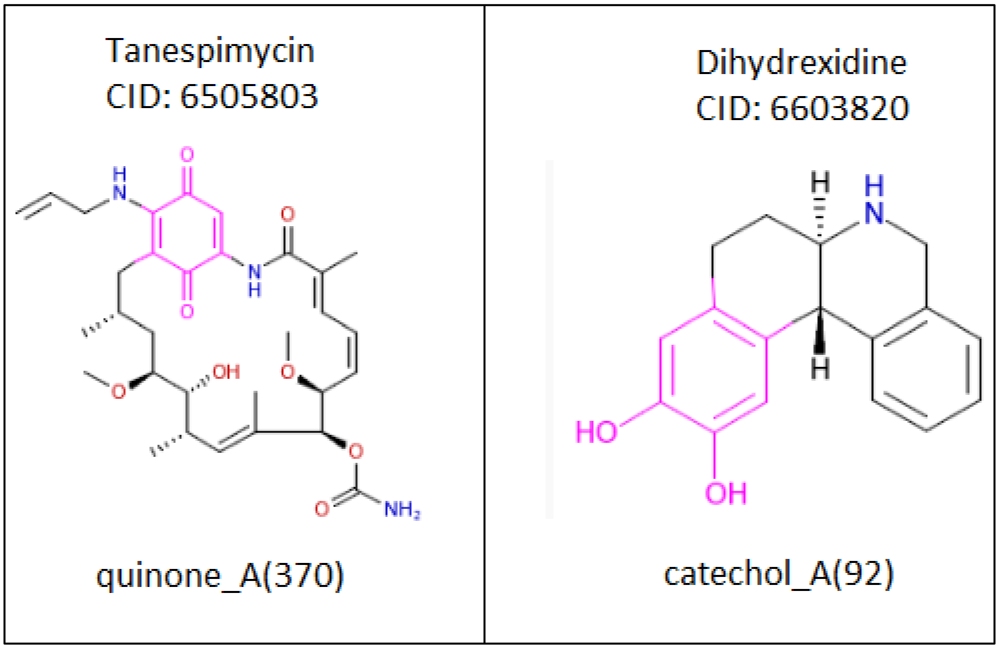


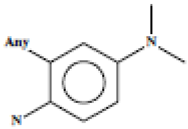
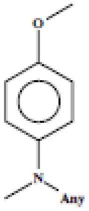
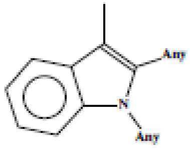
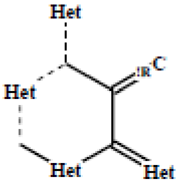
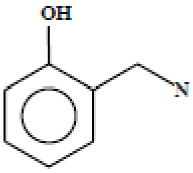
Figure 3. 4. Random-PAINS displaying pan-assay activity. Tanespimycin and dihydrexidine are active in 85% and 50% of all assays in PubChem, respectively.

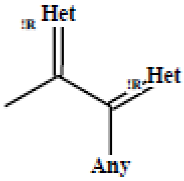
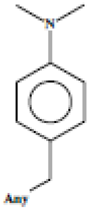
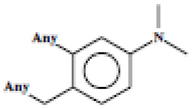
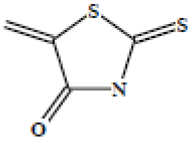
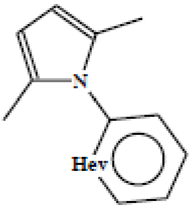
3.3.4 Analysis of PAINS Alerts in Dark Chemical Matter

Following the observation that Random-PAINS can be consistently inactive across a large number of assays, we probed the so-called Dark Chemical Matter (DCM)[110] for the presence of PAINS alerts. DCM was defined by Wasserman et al.[110] as compounds that hav e not yet shown any activity when tested in a minimum of 100 assays. The complete dataset of 139,352 DCM compounds, *i.e.* 128,997 PubChem and 10,355 Novartis DCM compounds, was downloaded from the Supplementary Information of the respective study.[110]

The dataset was examined with FAF3-Drugs[102], and 3,570 DCM compounds containing PAINS substructures were found, encompassing 109 of the 480 original PAINS alerts[24] (**Table S11**). Of these 109 PAINS alerts, 30 alerts were found in more than 10 compounds and 10 alerts were present in 100 or more compounds (**Table 3.3**). This analysis shows that even extensively assayed compounds containing PAINS alerts may be consistently inactive.

Table 3.3. **PAINS alerts enriched in Dark Chemical Matter.** Ten alerts are present in 100 or more in Dark Chemical Matter compounds (N_{DCM}).

PAINS Alert	Substructure[6]	N_{DCM}
anil_di_alk_A(478)		902
anil_di_alk_C(246)		492
indol_3yl_alk(461)		343
ene_six_het_A(483)		256
mannich_A(296)		212

imine_one_A(321)		193
anil_di_alk_D(198)		184
anil_di_alk_E(186)		164
ene_rhod_A(235)		116
pyrrole_A(118)		100

3.3.5 Global Analysis of PAINS Alerts

In this study, a total of 24,802 PAINS compounds were analyzed (208 FH-PAINS, 6413 IH-PAINS, 14611 Random-PAINS, and 3570 DCM-PAINS) covering 220 specific PAINS alert types, which is ~ 46% of the original PAINS alerts.[24] In order to determine if specific PAINS alerts correspond to compounds with elevated assay promiscuity, we performed a global analysis of the PubChem-wide activity of all PAINS compounds investigated herein (**Tables 3.6-3.9**). Since there is no agreed upon threshold of "pan-assay" activity, we selected assay activity of at least 10% as an arbitrary classifier.

Of these 220 PAINS alert types, 32 alerts had greater than 10% assay activity in either all assays, luciferase-, beta-lactamase-, or fluorescence-based assays (**Table 3.4**). However, only 12 of these alerts were present in more than 10 compounds (**Table 3.5**). It should be noted, however, that 6 of these alerts can also be found in DCM.

On the other hand, 176 (~80%) of the total PAINS alerts analyzed (220) were active in less than 10% of all investigated assays and technologies, and 88 alerts were present in DCM (**Tables 3.6 and 3.7**). Eighty-four (84) of 176 alerts were present in more than 10 compounds (**Table 3.6**). Interestingly, 6 of these alerts were found in more than 1,000 compounds (**Table 3.6**). Finally, 12 alerts were found exclusively in DCM-PAINS (**Table 3.8**). Eleven of these alerts were found in less than 10 compounds, while one alert, *i.e.*, hzone_phenol_B(215), was present in exactly 10 compounds.

There are 16 PAINS alerts that were derived from more than 150 compounds (**cf. Fig. 3.1**). Of all 480 alerts, these 16 alerts were created from the most underlying data in the original study[6]. Given the prevalence and heightened promiscuity of compounds possessing these alerts in the original study, we specifically investigated whether any compounds in our collection flagged by these 16 alerts display suspect assay trends (**Table 3.9**). Aside from hzone_phenol_A (479) and hzone_phenol_B(215), which were found exclusively in DCM, 14 of these 16 alerts were frequently assayed and abundantly present in the public collection. All 14 alerts displayed less than 10% activity in all PubChem assays despite being tested on average in more than 500 assays each. Among these specific alerts, the quinone_A(370) alert demonstrated the highest activity in all assays (8.4%).

Our findings using data in the public domain can be corroborated in part by other inquiries into the nature of promiscuous compounds. For instance, while attempting to use PAINS alerts to fill gaps in Eli Lilly's promiscuity filters, Bruns and Watson observed that "PAINS queries matched 286 promiscuous compounds that passed the Lilly rules, compared to 3986 in the non-promiscuous set, for an enrichment factor of 4.0"[113]. Furthermore, they noted that "although 67 PAINS queries matched at least one promiscuous compound, only nine queries matched at least five promiscuous compounds and had an enrichment of at least 5."[113] These findings are consistent with our observations that PAINS alerts in public data frequently flag non-promiscuous compounds or are manifested in only a small number of promiscuous compounds.

On the other hand, another study on frequent hitter behavior by researchers at AstraZeneca showed elevated "Frequent-hitter Incidence %" for 10 out of 15 PAINS alerts.[114] Although the authors state that their "corporate data largely confirm previous observations of the PAINS classes", this study only investigated part of the first tier of the 480 PAINS alerts, *i.e.*, the 15 out of 16 alerts derived from more than 150 compounds (**cf. Figure 3.1**), or ~3% of all alerts.[114] As can be seen, in that study, one-third of the profiled alerts did not show elevated frequent-hitter behavior, which is, in part, aligned with our general observations (**Tables 3.9**).

3.3.6 PAINS Alerts in Drugs

Other groups have noted that many drugs contain PAINS alerts,[24,108,109] and several careful and keen analyses have centered around this phenomenon.[115] It has also been observed that many of these PAINS alerts in drugs (but not all) map to poor ADMETox properties, such as quinone-containing drugs.[24] While this is an interesting observation, we view interference propensity and poor ADMETox properties as separate phenomena. Our group[116] as well as others[19] have also shown that a great majority of toxicity structural alerts, much akin to PAINS alerts, are overly sensitive and not predictive of actual *in vitro* or *in vivo* toxicity. Given that drug repurposing is currently widely used as a boon to traditional drug discovery[117,118], we profiled the PubChem-wide bioassay activity of drugs with and without PAINS alerts (**Table S12**).

A list of 1,460 approved small-molecule drugs was compiled from Drugs@FDA (https://www.accessdata.fda.gov/scripts/cder/drugsatfda/). Structures for these drugs were searched for PAINS alerts.[102,25,112] We identified 87 small-molecule approved drugs possessing 25 individual PAINS alerts (**Table S12**). As observed in the preceding sections, Drugs-PAINS are more active than Drugs-NoPAINS, having activity in 24% and 15% of all bioassays in PubChem, respectively (**Table S12**). According to current filters[25,102], 16 of these drugs possess quinone PAINS alerts. The promiscuity of quinone-containing drugs have been extensively discussed in the PAINS literature[24,115] and is supported by our analysis (**cf. Tables 3.2 and 3.4**). For instance, the chemotherapeutic doxorubicin, which contains the quinone_A(370) alert, has been tested in more than 4,000 assays with active calls ~85% of the time.

At the same time, however, the relationship between polypharmacology and PAINS has not yet been adequately explored. Many drugs show polypharmacological behavior and possibly derive their efficacy from interacting with multiple targets[119]. Indeed, a similar study on promiscuity in extensively assayed compounds found that drugs are more promiscuous than bioactive compounds[119], which is evidenced in our analysis as well (**Table 3.1**). Polypharmacology may well account for the increased activity of both Drugs-PAINS and Drugs-NoPAINS relative to the other categories (**cf. Table 3.1**). While the phenomena of assay interference and polypharmacology have rightfully been contrasted[119], there is very real possibility that a compound may both possess PAINS alerts and display polypharmacological behavior. Given that PAINS-containing drugs are now frequently used in drug repurposing screens, a larger discussion about the utility of PAINS alerts and polypharmacology should take place.

3.3.7 Beyond PAINS Substructures

PAINS concept has been widely accepted by many experienced medicinal chemists both in academia and the pharmaceutical industry. Indeed, the original study from which the PAINS alerts were derived and the impetus behind it are an important step towards reproducibility and the appropriate use of resources in drug discovery. However, our findings based on the analysis of public data suggest that many compounds containing PAINS alerts do not actually show high assay promiscuity, leading to the conclusion that these alerts should not be blindly used, in the absence of orthogonal experimental assays, to deprioritize a compound.

At the same time, it is undeniable that pan-assay interference *compounds* exist and care must be taken to avoid these compounds. Moreover, we recognize that true "PAINS" may be present in the data analyzed herein but have not been classified as such because the current alerts do not cover these compounds[120]. The issue of what constitutes a pan-assay interference compound thus remains unclear. For example, in *How to Triage PAINS-Full Research,* Dahlin and Walters define PAINS as "compounds that are recognized by the substructure filters reported by the Baell and Holloway article."[120] By this definition, all alerts (filters) are treated equally, regardless of the underlying data used to derive the alert or the actual promiscuity of flagged compounds. Yet our analysis indicates that the identification of such compounds should not be restricted to substructures alone. Substructural alerts, PAINS or otherwise, do not take into consideration the whole molecular environment[116], as illustrated by PAINS alerts manifesting in both promiscuous and frequently inactive compounds (DCM). Attempts should then be made to move beyond substructural or fragment-based alerts. For instance, Yang and coworkers in their "BadApple" algorithm have extended the identification of promiscuous compound to larger scaffolds.[121]

In recent publications by Alves et al.[122–124], quantitative structure-activity relationship (QSAR) models were used in conjunction with structural alerts for toxicity to dramatically improve the accuracy of prediction of multiple toxicity endpoints over alerts alone. The authors of the present study advocate the development of a similar approach for PAINS alerts. Such publicly accessible models, if successful, could be employed even for proprietary compounds insofar as chemical descriptors of PAINS-containing compounds could be shared without divulging actual molecular structures (given the proprietary nature of most compounds used to derive and evaluate PAINS so far[24,113,114]). The challenge is to build externally predictive QSAR models capable of classifying PAINS versus non-PAINS compounds. Using such models, predictions of suspect compounds could be made, giving higher confidence in the utility of the alert and the nefarious nature of the compound.

Meanwhile, the concept of PAINS alerts, at the very least, needs a re-defined set of "best practices"[120] that covers the appropriate use of alerts, which may include cross-referencing the promiscuity profiles of structurally similar compounds or alert types in the public domain, annotation of particularly susceptible assays, targets, and conditions, pointers to the appropriate controls, and a generally agreed upon definition of “pan-assay” activity. It would be of great value if a community-wide effort to screen and analyze a large set of commercially available compounds representing the all current PAINS alerts against multiple targets in various assays was performed by several independent groups.

3.4. CONCLUSIONS

It is imperative to establish target selectivity for any compound considered a viable chemical probe or drug candidate through rigorously acquired experimental data and meaningful SAR. Future studies may well establish some generalized approach for detecting frequent hitters engendered by assay interference. However, until such approaches are developed and rigorously validated across a large number of molecules, researchers should be cautioned about using the current PAINS alerts as reliable indicators of non-specific pan assay interference. Though it has been stated elsewhere that compounds flagged with PAINS alerts are not active in all assays or against all targets,[24,120] our analysis provides systematic and data-driven support of this claim across a large series of compounds, alerts, and assays. Our findings do demonstrate, with publicly available data at hand, that majority of the original PAINS alerts are not indicative of pan-assay compound promiscuity, that many compounds without PAINS alerts are as, if not more, promiscuous as those with the alerts, and that many compounds flagged by PAINS alerts show no activity. It is of great importance that reviewers and journal editors request experimental proofs of selectivity, such as orthogonal experimental assays, for hit and lead compounds reported in scientific manuscripts. However, the results of this study strongly suggest that such requests should not be based solely on the results of PAINS filters.[120]

3.5 ASSOCIATED CONTENT

Supporting Information. Supplementary Files and Tables in Chapter 3 are available at https://scapuzzi.web.unc.edu/free-downloads/

CHAPTER 4: CHEMOTEXT: A PUBLICLY-AVAILABLE WEB SERVER FOR MINING DRUG-TARGET-DISEASE RELATIONSHIPS IN PUBMED[3]

4.2 INTRODUCTION

The fundamental goal of small molecule drug discovery is the identification of bioactive compounds for the treatment of disease[125]. Many modern drug discovery projects start with the discovery of novel targets and then progress in the direction of finding ligands of these targets that are expected to affect the disease. Bioactivity data from drug repurposing/discovery campaigns are increasingly available in public databases such as PubChem[35,126] and ChEMBL[127]. At the same time, much information about the biological underpinnings of disease, *i.e.*, effector proteins and pathways, as well as drug targets are stored primarily in the biomedical literature. Thus, biomedically relevant relationships between drugs, biological targets, and diseases, which we call the DTD triangle, can be identified through mining the published biomedical literature.[128,129]

PubMed, the largest repository of published biomedical research, is a freely-accessible search engine maintained by the United States National Library of Medicine (NLM) at the National Institutes of Health (NIH)[130]. PubMed can be used to retrieve scientific articles containing specific search terms that are stored in the Medline bibliographic database. PubMed can also return a list of Medical Subject Headings (MeSH), or so-called MeSH terms[131]. The purpose of these MeSH terms is to index and categorize published studies by the subject matters discussed therein. As

[3] This chapter previously appeared as an article in the Journal of Medicinal Chemistry. The original citation is as follows: Capuzzi, S. J., et. al. “Chemotext: A Publicly Available Web Server for Mining Drug–Target–Disease Relationships in PubMed.” *J. Chem. Inf. Model.* (February, 2018) *58*, 212–218.

most drugs, biological targets, and diseases discussed in biomedical literature are captured by associated MeSH terms, relationships between terms in the DTD triangle (represented by edges of the triangle with vertices representing MESH terms) can be established based on their frequent co-occurrences within articles.

Indeed, such considerations led to the development of the Chemotext approach,[132] which focused on the extraction of MeSH terms describing “chemicals”, “targets”, and “diseases”, *i.e.*, the components of the DTD triangle, that were found to frequently co-occur in abstracts of papers annotated in PubMed. These co-occurrences were regarded as an indication of plausible assertions linking drugs, targets and diseases. Furthermore, Chemotext was conceived as an extension of Swanson’s ABC paradigm[132–134] wherein “A” terms are chemical (drug) - related MeSH terms, “B” terms are so-called “target” MeSH terms, *i.e.,* proteins and pathways, and “C” terms are MeSH terms for diseases (**Figure 4.1**). The underlying hypothesis generation starts with the observation that the name of drug “A” co-occurs in the same articles as the name of target “B” while the name of disease “C” co-occurs in the same or additional articles with the same target “B”. Thus, if drug “A” and disease “C” have not been mentioned together in the same article, an “A-C” connection mediated though target “B” can be inferred. This analysis leads to the identification of a new possible therapeutic use of drug “A”. This reasoning protocol illustrates one of possible uses of Chemotext for drug repurposing, which has emerged in the past decade as a boon to traditional drug discovery.[118,135]

Although efforts have been made to develop tools for text-mining of PubMed, such as “MeSHSim”,[136] “pubmed.mineR”,[137] and IBM-Watson,[138] these current implementations are either available only as R-packages,[136,137] which are not user-friendly, and/or proprietary.[138] To this end, we have developed the publicly-available Chemotext Web server that mines published literature in PubMed in the form of MeSH terms. The goal of Chemotext is to establish text-based drug-target-disease relationships, which, as we show herein, can be used to generate novel drug repurposing hypotheses or elucidate clinical outcomes pathways that mechanistically connect drugs and diseases via intermediary, target-mediated biological effects of drug action. Similar to our Chembench Webportal,[46] the Chemotext Web server is hosted by the Molecular Modeling Laboratory (MML) at the University of North Carolina – Chapel Hill and is freely-available at http://chemotext.mml.unc.edu/.

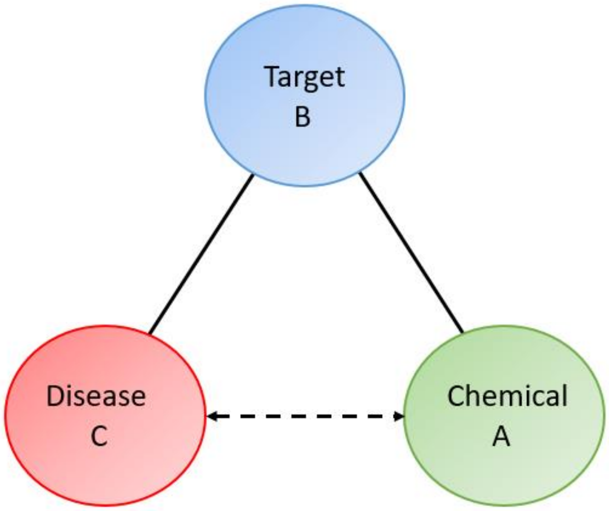


Figure 4. 1. **Swanson's ABC paradigm used in Chemotext**. Chemical A is proposed to have an effect on Disease C since both terms are associated with Target B. Solid lines (edges) indicate an actual text-based relationship, while dashed lines (edges) indicate proposed connections**.**

4.3 METHODS

The Chemotext user interface is written in JavaScript with data retrieval through JQuery's Asynchronous JavaScript and XML (AJAX) functionality.[139] The data are stored in Neo4j, a graph database that uses nodes for articles and drug terminology. The server operates on Red Hat Linux and is hosted by the Longleaf computer cluster at UNC-Chapel Hill. MeSH term data were downloaded directly from the PubMed and Medline repository. Data were input into

Neo4j using Cypher to parse the MeSH XML and to create the article and term nodes and relationships. Cypher queries allow for Neo4j to return sets of MeSH terms and article counts from the input term's article relationships.

Data for calendar year 2016 were retrieved from the MEDLINE/PubMed Baseline Repository (MBR) in June 2017. Data are available at https://mbr.nlm.nih.gov/Downloads.shtml. Currently, the Chemotext database contains 19,282,732 articles and 78,758,882 connections between terms. Chemotext is currently fully functional only with the Google Chrome web browser on both PC and MAC operating systems.

4.4 CHEMOTEXT ENVIRONMENT

Chemotext generates text-based relationships via four modules described below: Find Connected Terms, Find Shared Terms, Path Search, and Find Articles. Within each module, there is a query bar that possesses the full dictionary of MeSH terms with an auto-complete function to facilitate searching. Each module can be executed separately or as part of a larger study design. On its homepage, Chemotext possesses direct link to the Medical Subject Headings search engine in order to facilitate the identification of correct MeSH terms for querying.

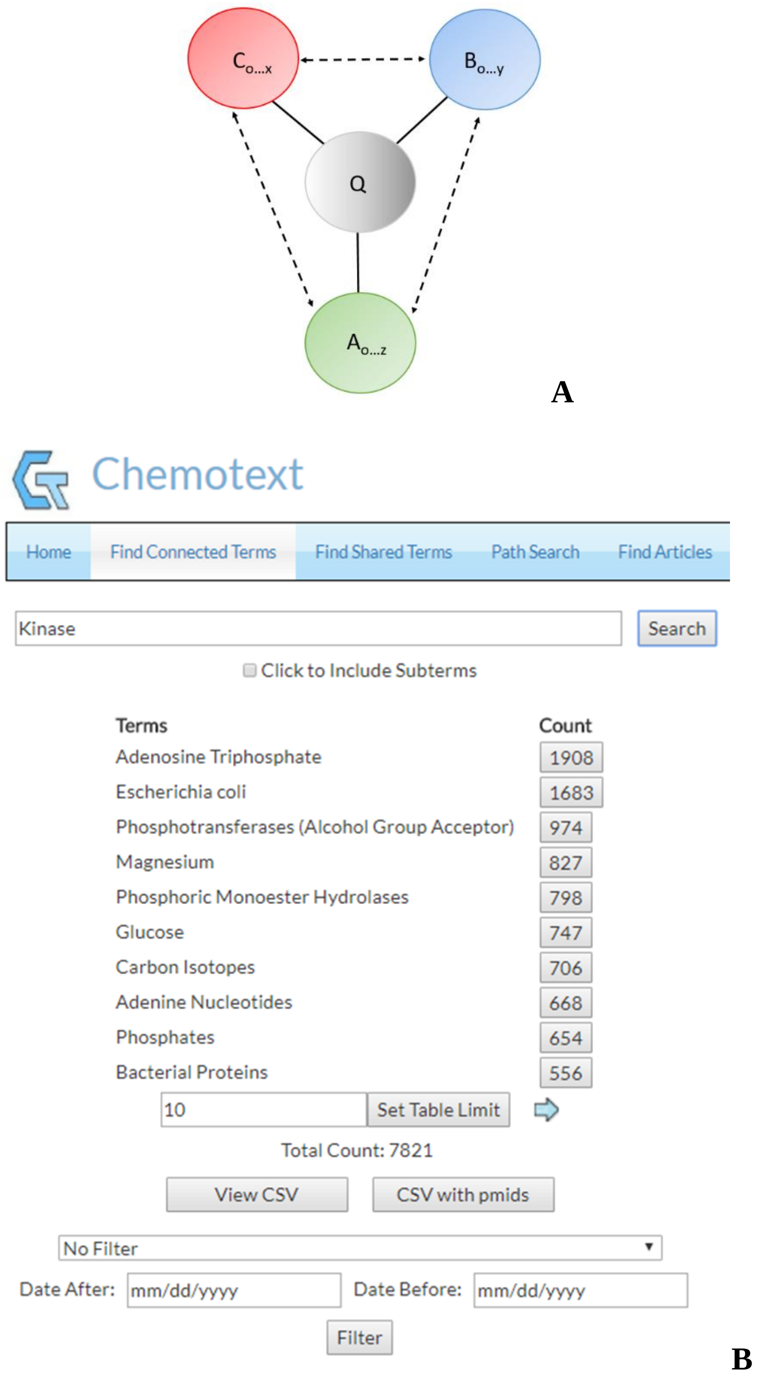
4.4.1 Find Connected Terms

In this module, every MeSH term that occurs in the same article as a query term is returned, and the total number of co-occurring terms and the associated article counts are reported. A schema of this module is presented in **Figure 4.2A**. To illustrate how this module is used, if "Kinase" is queried, 7 821 unique co-occurring MeSH terms are returned (**Figure 4.2B**), such as "Enzyme Inhibitors," an A term with 333 article co-occurrences, and "Neoplasms", a C term with 151 article co-occurrences (**Table S1**).

The resultant terms are rank-ordered by the number of unique articles in which the term co-occurs with the query. Thus, the article count serves as a proxy for the strength of the association between terms in the A-B-C paradigm. For each co-occurring term, the user can click on the article count and view all of the associated article PubMed Identification (PMID) numbers. These PMIDs are linked to PubMed, allowing the user to access and review the article(s) in which the two terms are mentioned together.

The full list of co-occurring terms can be filtered by MeSH term type, *i.e,* by “Chemical” terms, “Proteins-Pathways-Intermediaries-Other”, or by “Disease and Indication”, which correspond to A, B, and C terms, respectively (cf. **Figure 4.1**). Moreover, each MeSH term type (A, B, or C) has additional subtypes that facilitate further refinement of the co-occurring terms. For instance, Chemical (A) terms can be filtered by “Drug” terms, which allows the user to identify which FDA-approved drugs co-occur in the same articles as the query term. The full list of term subtypes for filtering is provided in the Supporting Information (**Table S2**). Aside from type, the co-occurring terms can be filtered by date of publication; thus, all terms appearing in articles published before or after a certain date can be retrieved.

Users are able to download two CSV files. First is a file of the co-occurring terms and the associated article counts, while the second is a file of co-occurring terms, the article counts, and the explicit PMIDs.



Chemotext

Kinase

□ Click to Include Subterms

Terms	Count
Adenosine Triphosphate	1908
Escherichia coli	1683
Phosphotransferases (Alcohol Group Acceptor)	974
Magnesium	827
Phosphoric Monoester Hydrolases	798
Glucose	747
Carbon Isotopes	706
Adenine Nucleotides	668
Phosphates	654
Bacterial Proteins	556

10

Total Count: 7821

No Filter ▾

Date After: mm/dd/yyyy Date Before: mm/dd/yyyy

Figure 4. 2. (A) Schema of the Find Connected Terms Module. A query term (Q) is input and all co-occurring A, B, C terms connections are established and putative connections are proposed. Solid lines indicate actual text-based co-occurrences, while dashed lines indicate proposed connections. It should be noted that Q can be either an A, B, or C term. **(B) Find Connected**

Terms Module Output. All A, B, and C terms (7 821 total) that co-occur in the same articles as the query term "Kinase" are returned with the associated article counts. Resultant terms can be filtered by sub-terms and date, and the results and PMIDs can be downloaded.

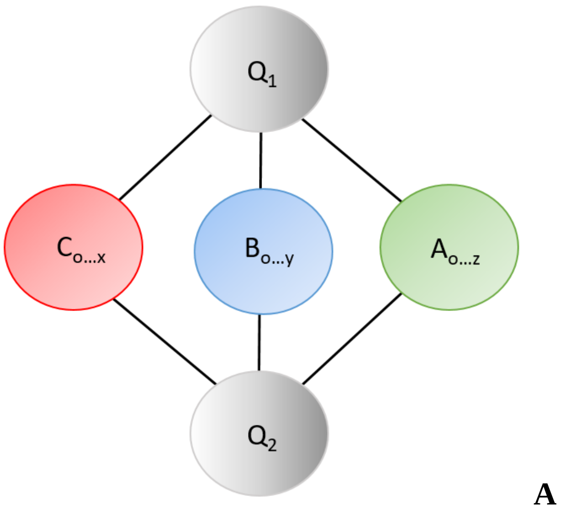
4.4.2 Find Shared Terms

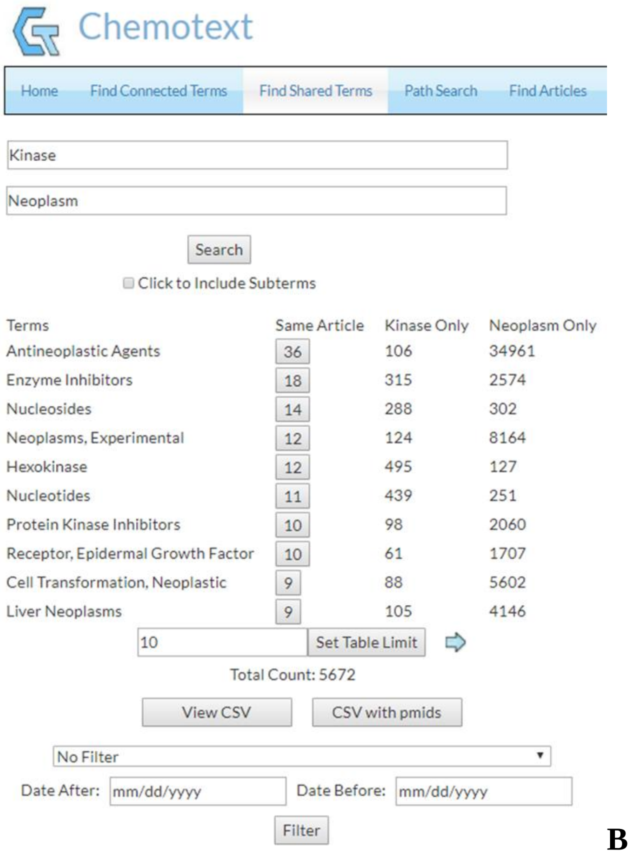
In this module, two query terms are input, and co-occurring terms and the article counts that are shared between the queries are returned. A schema of this module is presented in **Figure 4.3A**.

Thus, this type of search outputs the associated counts of co-occurrence for three instances: (i) when all three terms (query 1, query 2, and co-occurring term) are present in the same article, (ii) when the term co-occurs only in articles with query 1, and (iii) when the term co-occurs only in articles with query 2. For example, when "Kinase" and "Neoplasm" are queried together in this module (**Figure 4.3B**), the term "Antineoplastic Agents" co-occurs in 36 articles with *both* "Kinase" and "Neoplasm", 106 articles with *only* "Kinase", and 34 961 articles with *only* "Neoplasm" (**Table S3**). It should be noted, however, that if a term co-occurs with only *one* of the queries, then this co-occurring term is not returned in this module, as it does not occur with the other query. The term, therefore, is not shared between the two query terms.

The resultant terms are rank-ordered by the number of unique articles in which all three terms co-occur. Since all three terms occur in the same article(s), these associations are considered the strongest.

For each shared co-occurring term, the user can click on the article count and view all of the associated article PMID numbers when all three terms are present in the same article. As stated previously, these PMIDs are linked to PubMed. If for the case where the term co-occurs with query 1 and query 2, but are not necessarily present in the same articles, then the user can obtain these PMIDs and links to articles in the “Find Connected Terms” module. The same previously described filters and downloadable files are available in this module.





Kinase

Neoplasm

□ Click to Include Subterms

Terms	Same Article	Kinase Only	Neoplasm Only
Antineoplastic Agents	36	106	34961
Enzyme Inhibitors	18	315	2574
Nucleosides	14	288	302
Neoplasms, Experimental	12	124	8164
Hexokinase	12	495	127
Nucleotides	11	439	251
Protein Kinase Inhibitors	10	98	2060
Receptor, Epidermal Growth Factor	10	61	1707
Cell Transformation, Neoplastic	9	88	5602
Liver Neoplasms	9	105	4146

10

Total Count: 5672

No Filter

Date After: mm/dd/yyyy Date Before: mm/dd/yyyy

Figure 4. 3. (A) Schema of the Find Shared Terms Module. Two query terms, Q_1 and Q_2, representing any pair of A, B, and C terms, are input, and all co-occurring A, B, and C terms shared between the query terms are established. **(B) Find Connected Terms Module Output.** Two query terms, “Kinase” and “Neoplasm”, are input, and all co-occurring A, B, and C terms shared between the query terms are established (5 672 shared terms). Resultant terms can be filtered by sub-terms and date, and the results and PMIDs can be downloaded.

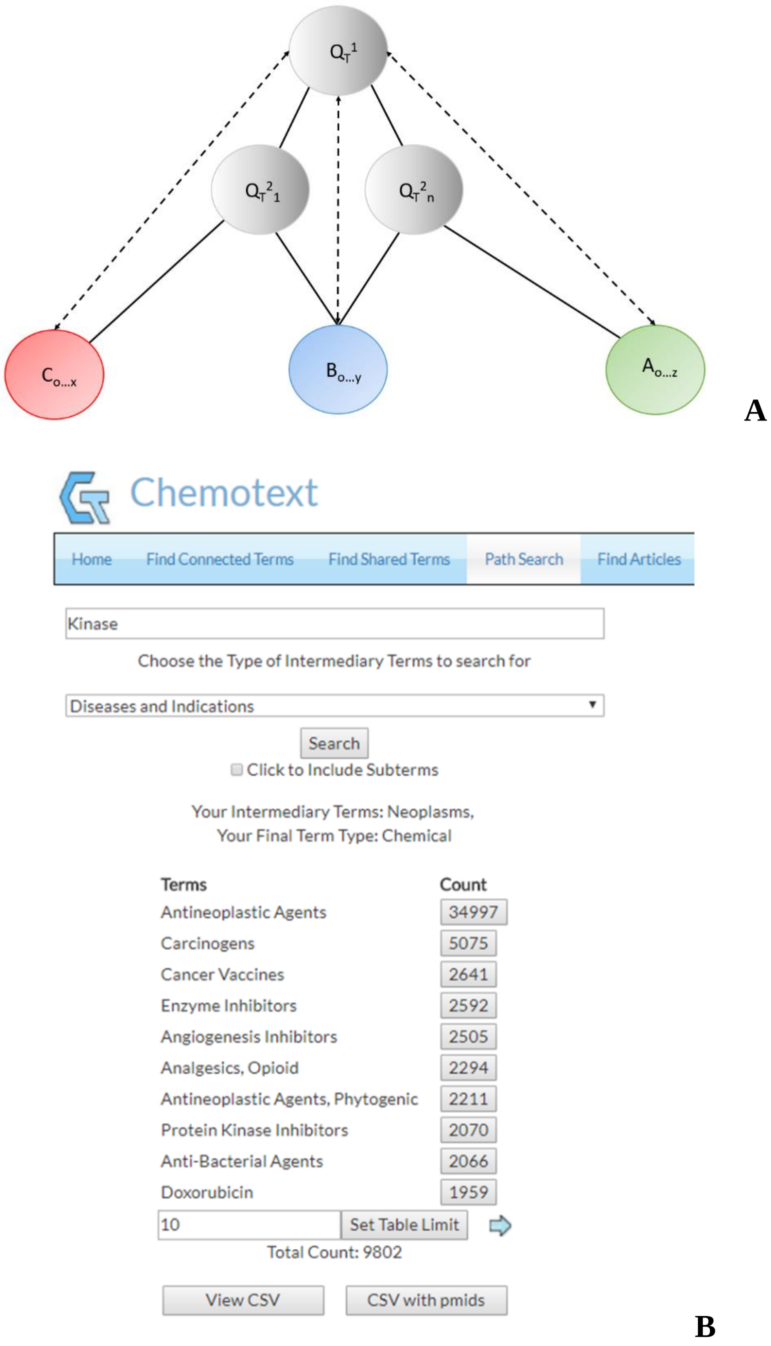
4.4.3 Path Search

In this module, complete text-based A-B-C connections can be made through co-occurring MeSH terms. The name of this module – “Path Search” – indicates that these A-B-C connections can be established through several “paths”, *i.e.*, through multiple intermediary terms or through a single intermediary term. A schema of this module is presented in **Figure 4.4A**.

In the most complex and comprehensive path search, every possible A-B-C connection for a given query term can be established. For instance, if “Kinase” is queried and “Diseases and Indications” are chosen as the intermediary term, 1 242 unique MeSH terms are returned, representing 1 242 unique B-C connections. Examples of these unique B-C connections are as diverse as “Kinase-Neoplasms,” “Kinase-Gout,” and “Kinase-Leprosy.” Next, all 1 242 B-C connections can be queried for associated A-terms, thereby completing every possible A-B-C connection, *i.e.* DTD triangles. In this case, every chemical that can be associated with the B-term “Kinase” as mediated through the 1 242 C-terms is identified.

This path search can be simplified to identify more focused A-B-C connections through a single intermediary term. Using an above example, the single B-C connection of “Kinase-Neoplasms” can be queried for all co-occurring “Chemical” A-terms, resulting in 9802 unique A-B-C connections mediated through the “Kinase-Neoplasms” nodes (**Figure 4.4B**). Of these 9 802 unique A-B-C connection in this path search (**Table S4**), Chemotext retrieves 270 articles that establish the specific A-B-C connection of “Imatinib-Kinase-Neoplasms.” This connection represents a known drug-target-disease relationship, as the tyrosine kinase inhibitor imatinib is used to treat several cancers, including gastrointestinal stromal tumors (GIST) through the blockage of the receptor tyrosine kinase c-kit.[140] In the **Case Study**, we will demonstrate that imatinib can also be repurposed as a treatment for asthma.

In the Path Search module, the intermediary term type can either be the MeSH term type, *i.e.,* “Disease and Indication”, “Proteins-Pathways-Intermediaries-Other”, “Chemical” terms, or the MeSH term subtypes, such as “Viruses”, “Enzymes and Co-Enzymes”, and “Heterocyclic Compounds”. Regardless of the intermediary term type, resultant terms are ranked according to the highest co-occurring article count with the query term. One or more intermediary terms can be selected to complete the path search, and the final connection can either set as the MeSH term type or subtype. The resultant terms are again ranked by highest co-occurring article count with the intermediary terms. Once the path search has been completed, the user can access the articles associated with the final term via the PMID and can download the two previously described CSV files.



Kinase

Choose the Type of Intermediary Terms to search for

Diseases and Indications ▾

□ Click to Include Subterms

Your Intermediary Terms: Neoplasms,
Your Final Term Type: Chemical

Terms	Count
Antineoplastic Agents	34997
Carcinogens	5075
Cancer Vaccines	2641
Enzyme Inhibitors	2592
Angiogenesis Inhibitors	2505
Analgesics, Opioid	2294
Antineoplastic Agents, Phytogenic	2211
Protein Kinase Inhibitors	2070
Anti-Bacterial Agents	2066
Doxorubicin	1959

10 Set Table Limit

Total Count: 9802

View CSV CSV with pmids

Figure 4.4. (A) Schema of the Path Search Module. The first query term, Q_T^1, is the input. Next, a second layer of query terms (Q_T^2) that co-occur with Q_T^1 are selected. The number of terms in the second query layer can range from one ($Q_T^2{}_1$) to all associated terms ($Q_T^2{}_n$). Next, any category of MeSH term that co-occurs with Q_T^2 are returned. Solid lines indicate actual text-based co-occurrences, while dashed lines indicate proposed connections. It should be noted that Q terms can be a combination A, B, or C terms. **(B) Path Search Module Output.** The first query term, “Kinase”, is the input. Co-occurring intermediary C terms, “Disease and Indications”, are returned.

Within this group, “Neoplasms” is selected as the second query layer, and the 9 802 chemical terms that co-occur with that term are returned.

4.4.4 Find Articles

In this module, articles indexed in PubMed can be searched for using specific MeSH terms. Additionally, this module will allow the user to inspect the total number of articles associated with this term. For example, if the term “Neoplasms” is queried, 36 1190 unique hits are returned with direct links to the respective articles.

4.5.1 CASE STUDIES

4.5.1.a Construction of a Clinical Outcome Pathway (COP) for a Drug-Disease Pair

In order to demonstrate the utility of Chemotext, we describe its application for finding the accurate solution of the recent National Center for Advancing Translational Science (NCATS) Biomedical Data Translator Challenge (https://ncats.nih.gov/files/translator_FOA_2017.pdf). The task of this challenge was to construct a clinical outcome pathway (COP) for the drug-disease pair imatinib-asthma. It was stated that a clinical outcome pathway (COP) begins with (i) a molecule physically interacting with (ii) a biological target that affects (iii) a biological pathway relevant to (iv) a particular cell or tissue type that manifest as (v) a clinical phenotype and/or symptom which reflect (vi) a disease or condition. The challenge was to construct a COP for (i) imatinib that successfully reveals its (ii) biological target, (iii) the pathway affected by that target, (iv) the cell or tissue type, and (v) the manifested symptom germane to (vi) asthma in the form of relevant MeSH terms and associated article PMIDs for stages ii-v (**Figure 4.5**).

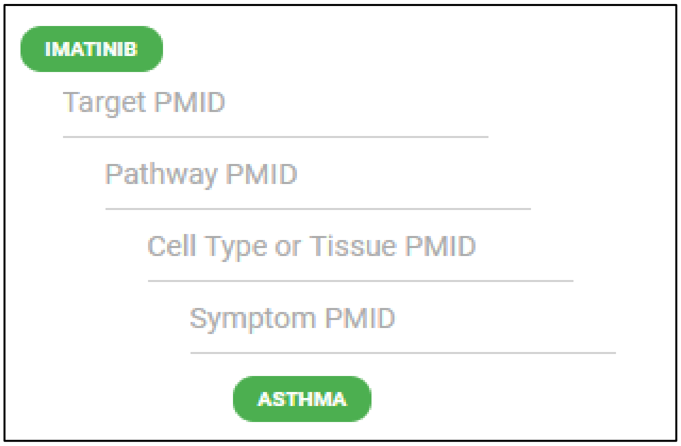


Figure 4.5. NCATS Biomedical Data Translator Challenge #5. The task was to successfully construct a COP connecting imatinib and asthma. Correct MeSH terms and associated article PMIDs had to be identified to solve the challenge.

In the first step of the solution-seeking algorithm, query terms “Imatinib” (i) and “Asthma” (vi) were searched in the Find Shared Terms module. The list of full associations was filtered by “Proteins-Pathways-Intermediaries-Other”. The MeSH term “Proto-Oncogene Proteins c-kit” was the fourth highest ranked shared term (two shared articles) selected as the potential biological target (ii) in the COP. The three more highly ranked terms, *i.e.*, “Allergens”, “Stem Cell Factor”, and “Ovalbumin”, were deemed too broad or generic to be viable solutions. The two articles and their associated PMIDs related to “Proto-Oncogene Proteins c-kit” were then directly accessed through the Chemotext Web server. Both articles, upon visual inspection, confirmed the relevance of this DTD triangle. One article (PMID: 19722748)[141], *i.e.*, “Presence of c-KIT-positive mast cells in obliterative bronchiolitis from diverse causes”, was successfully chosen as the solution to stage (ii) of the COP, as later confirmed by the NCATS Challenge system.

To identify the biological pathway affected (iii) in this COP, query terms “Imatinib” (i) and “Proto-Oncogene Proteins c-kit” (ii) were searched in the Find Shared Terms module in the second step of the solution algorithm. The list of full associations was filtered by “Proteins-Pathways-Intermediaries-Other”. The MeSH terms and associated article counts were downloaded from Chemotext. Next, query terms “Proto-Oncogene Proteins c-kit” (ii) and “Asthma” (vi) were searched in the Find Shared Terms module and the same succeeding steps as above were performed. The intersection of the two lists, *i.e.,* (i-ii) and (ii-vi) was obtained, and MeSH terms were sorted according to article count ranks (**Table S5**). The MeSH term “Phosphatidylinositol 3-Kinases” was one of the most highly ranked shared terms (22nd out of 928 terms). More highly ranked terms, such as “Biomarkers” and “Neoplasm Proteins”, were not selected because they were not relevant to the “Pathway” portion of this COP. Articles and their associated PMIDs related to “Phosphatidylinositol 3-Kinases” were then directly accessed through the Chemotext Web server. One article (PMID: 17546049)[142], *i.e.,* “KIT oncogenic signaling mechanisms in imatinib-resistant gastrointestinal stromal tumor: PI3-kinase/AKT is a crucial survival pathway”, was chosen as the successful solution to stage (iii) of the COP.

In order to identify the cell or tissue type (iv) involved in this COP, “Imatinib” (i) and “Asthma” (vi) were again searched in the Find Shared Terms module. Co-occurring terms were then filtered by “Cells”. This resulted in the correct identification of “Mast Cells” (PMID: 16483568)[143]. Likewise, for the manifested symptom (v), the drug and the disease were queried in the Find Shared Terms module, and resultant connections were filtered by “Diseases and Indications”. The top co-occurring term was “Bronchial Hyperreactivity” (PMID: 24112389)[144]. Both the terms were later confirmed by the NCATS Challenge system as steps in the COP.

The full Imatinib-Asthma COP, as revealed by Chemotext and confirmed by the Challenge system, was: Imatinib (i) → Proto-Oncogene Proteins c-kit (ii) → Phosphatidylinositol 3-Kinases (iii) → Mast Cells (iv) → Bronchial Hyperreactivity (v) → Asthma (iv).

It should be emphasized that expert-based knowledge curation, in conjunction with the results of Chemotext, was key for the successful identification of terms and articles. For instance, in the first step of the solution algorithm, "Proto-Oncogene Proteins c-kit" was the correct target, but there were three more highly ranked terms, *i.e.*, "Allergens", "Stem Cell Factor", and "Ovalbumin". These terms were deemed too broad or generic to be viable solutions to this stage (ii) of the COP. Likewise, in the second step, "Phosphatidylinositol 3-Kinases" ranked 22nd out of 928 terms. More highly ranked terms, such as "Biomarkers" and "Neoplasm Proteins", however, were not biologically relevant (not related to "Pathway") for this COP and thus not investigated. This observation obligates that additional scoring functions - besides of article counts - should be considered to elucidate meaningful relationships.

Last, it should be noted that this COP may have many alternative plausible solutions that have not been investigated herein; we have described a single validated test case merely to illustrate Chemotext's capabilities.

4.5.2.b Drug Repurposing for Human Cytomegalovirus

In April 2017, Arend et. al. published a study that identified kinases upregulated during human cytomegalovirus (HCMV) infection and sought to repurpose kinase inhibitors as novel HCMV antivirals[145]. Chiefly, Arend et al. found that the experimental kinase inhibitors OTSSP167 and dinaciclib were nanomolar inhibitors of HCMV. Since this study was just recently published in 2017, this article has not yet been indexed with MeSH terms and is not yet part of the Chemotext database. This study, thus, can be used to test the ability Chemotext to identify these compounds as HCMV inhibitors solely through text-based relationships, as it represents an A-B-C paradigm.

Using the Path Search module, the relevant MeSH A-term “Cytomegalovirus” was queried, and “Proteins-Pathways-Intermediaries-Other” was selected as the intermediary term type. As a result, 3148 unique A-B connections were returned; however, since the study of interest focused only on kinase targets, the 101 B-terms associated with kinases were selected for further querying. These 101 kinase-associated B-terms range from broad terms, such as “Protein Kinases”, to more specialized terms, such as “ZAP-70 Protein-Tyrosine Kinase”. These B-terms were then queried for “Chemical” terms to complete the A-B-C paradigm.

Ultimately, 16232 unique chemicals, which includes experimental drugs, were associated with the “Cytomegalovirus-Kinase” connection. Among these, the MELK inhibitor OTSSP167 (which is indexed as by its IUPAC name “1-(6-(3,5-dichloro-4-hydroxyphenyl)-4-((4-((dimethylamino)methyl)cyclohexyl)amino)-1,5-naphthyridin-3-yl)ethanone”) and the CDK inhibitor dinaciclib were found to be associated with “Cytomegalovirus” as mediated through kinase B-terms, with 4 and 19 associated articles, respectively. Chemotext was, therefore, successful at identifying drug repurposing candidates for the treatment of HCMV infection.

It is important to note that these experimental drugs had not yet co-occurred in the same articles as “Cytomegalovirus;” thus, this the A-C relationship can only be established via the A-B-C paradigm implemented in Chemotext.

4.6 CONCLUSIONS

We have developed the Chemotext Web server to facilitate the identification of existing drug-target-disease (DTD) relationships and to generate hypotheses about novel relationships by mining of PubMed in the form of MeSH terms via four modules: Connected Terms, Find Shared Terms, Path Search, and Find Articles. In the Connected Terms module (**Figure 4.2A**), the user can query any type of MeSH terms, *i.e.*, an A, B, or C term, and retrieve all MeSH terms that co-occur in the same articles as the query term. This module provides an overview of all text-based associations and makes connections between terms. In the Find Share Terms module (**Figure 4.3A**), two query terms are input, and co-occurring terms that are shared between the queries are returned. For instance, in this module the shared targets between two diseases or between a drug and disease can be identified. In the Path Search module (**Figure 4.4A**), full A-B-C connections can be established through intermediary MeSH terms. We provided an example of using Chemotext to generate drug repurposing candidates. Last, a focused search of PubMed via MeSH term keywords can be performed using the Find Articles module.

The Chemotext Web server was originally conceived of and developed as a text-mining tool for inferring new drug-disease associations[132,146], *i.e*., drug repurposing; however, Chemotext can be used to establish DTD triangles or to mine any type of text-based relationships between biomedical terms or concepts. For example, Chemotext could be used to establish protein-protein interaction networks through co-occurring B-terms or to uncover correlations in disease progression through co-occurring C-terms. The potential number and types of relationships that can be generated with Chemotext are myriad and not limited to the A-B-C paradigm described herein. Indeed, in 2016, Alves et. al.[147] used Chemotext outside of this paradigm to confirm the toxic effects of chemicals predicted as human skin sensitizers in a virtual screening campaign.

In spite of its obvious advantages, Chemotext in its current form has several limitations that must be addressed. First, the deposition of articles into PubMed is ever-growing. As per data availability in MBR, the database of terms that underlies Chemotext, must be updated regularly to capture these articles, and new literature-based connections between terms have to be generated. Additionally, from a functional aspect, relationships derived by Chemotext are limited to MeSH terms indexed in the abstracts of articles. Future implementations will seek to mine full articles, although this form of text-mining is orders of magnitude more difficult. In the same vein, Chemotext currently does not support natural language processing and provides no inference about the nature of the relationship between the terms (agonism vs. antagonism, cause vs. effect, mode of action vs. side effect, etc.). This may lead to a number of false positive hits that are not directly related to the desired effect. From a technical perspective, chemicals can be *queried* by multiple synonyms, *i.e.*, aspirin vs. acetylsalicylic acid vs. dispril. The "Click to Include Subterms" feature of Chemotext ensures that *all* terms associated with a chemical will be investigated. On the other hand, chemicals will be *returned* only by the *main* MeSH term, *i.e.*, aspirin. The user must be aware that the resultant chemical may be indexed by an unfamiliar construction, such as its IUPAC or generic name. Presently, the onus is placed on the user to then identify and investigate the chemical(s) of interest by the corresponding MeSH term. To address these issues and to improve the scope and functionality of Chemotext, regular updates and improvements are underway, such as improving its functionality on additional web browsers like Safari and Firefox and resolving chemical names.

The Chemotext Web server is freely-available at http://Chemotext.mml.unc.edu/index.html (currently fully operational via Google Chrome only). A user-friendly tutorial is also available at the site: http://chemotext.mml.unc.edu/ChemotextAppNote_Tutorial_v3.docx

4.7 ASSOCIATED CONTENT

Supporting Information. Results of Chemotext queries described in the manuscript and other Chemotext related information (**Tables S1-S6**) for Chapter 4 are provided as Supporting Information at https://scapuzzi.web.unc.edu/free-downloads/.

5. REFERENCES

(1) Green, A. Ebola Outbreak in the DR Congo. *The Lancet*. 2017.

(2) Warren, T. K.; Jordan, R.; Lo, M. K.; Ray, A. S.; Mackman, R. L.; Soloveva, V.; Siegel, D.; Perron, M.; Bannister, R.; Hui, H. C.; Larson, N.; Strickley, R.; Wells, J.; Stuthman, K. S.; Van Tongeren, S. A.; Garza, N. L.; Donnelly, G.; Shurtleff, A. C.; Retterer, C. J.; Gharaibeh, D.; Zamani, R.; Kenny, T.; Eaton, B. P.; Grimes, E.; Welch, L. S.; Gomba, L.; Wilhelmsen, C. L.; Nichols, D. K.; Nuss, J. E.; Nagle, E. R.; Kugelman, J. R.; Palacios, G.; Doerffler, E.; Neville, S.; Carra, E.; Clarke, M. O.; Zhang, L.; Lew, W.; Ross, B.; Wang, Q.; Chun, K.; Wolfe, L.; Babusis, D.; Park, Y.; Stray, K. M.; Trancheva, I.; Feng, J. Y.; Barauskas, O.; Xu, Y.; Wong, P.; Braun, M. R.; Flint, M.; McMullan, L. K.; Chen, S.-S.; Fearns, R.; Swaminathan, S.; Mayers, D. L.; Spiropoulou, C. F.; Lee, W. A.; Nichol, S. T.; Cihlar, T.; Bavari, S. Therapeutic Efficacy of the Small Molecule GS-5734 against Ebola Virus in Rhesus Monkeys. *Nature* **2016**, *531* (7594), 381–385.

(3) Smither, S. J.; Eastaugh, L. S.; Steward, J. A.; Nelson, M.; Lenk, R. P.; Lever, M. S. Post-Exposure Efficacy of Oral T-705 (Favipiravir) against Inhalational Ebola Virus Infection in a Mouse Model. *Antiviral Res.* **2014**, *104*, 153–155.

(4) McMullan, L. K.; Flint, M.; Dyall, J.; Albariño, C.; Olinger, G. G.; Foster, S.; Sethna, P.; Hensley, L. E.; Nichol, S. T.; Lanier, E. R.; Spiropoulou, C. F. The Lipid Moiety of Brincidofovir Is Required for in Vitro Antiviral Activity against Ebola Virus. *Antiviral Res.* **2016**, *125*, 71–78.

(5) Burd, E. M. Ebola Virus: A Clear and Present Danger. *J. Clin. Microbiol.* **2015**, *53* (1), 4–8.

(6) Tscherne, D. M.; Manicassamy, B.; García-Sastre, A. An Enzymatic Virus-like Particle Assay for Sensitive Detection of Virus Entry. *J. Virol. Methods* **2010**, *163* (2), 336–343.

(7) Kouznetsova, J.; Sun, W.; Martínez-Romero, C.; Tawa, G.; Shinn, P.; Chen, C. Z.; Schimmer, A.; Sanderson, P.; McKew, J. C.; Zheng, W.; García-Sastre, A. Identification of 53 Compounds That Block Ebola Virus-like Particle Entry via a Repurposing Screen of Approved Drugs. *Emerg. Microbes Infect.* **2014**, *3* (12), e84.

(8) Sun, W.; He, S.; Martínez-Romero, C.; Kouznetsova, J.; Tawa, G.; Xu, M.; Shinn, P.; Fisher, E. G.; Long, Y.; Motabar, O.; Yang, S.; Sanderson, P. E.; Williamson, P. R.; García-Sastre, A.; Qiu, X.; Zheng, W. Synergistic Drug Combination Effectively Blocks Ebola Virus Infection. *Antiviral Res.* **2017**, *137*, 165–172.

(9) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.* **2010**, *29*, 476–488.

(10) Ekins, S.; Freundlich, J. S.; Clark, A. M.; Anantpadma, M.; Davey, R. A.; Madrid, P. Machine Learning Models Identify Molecules Active against the Ebola Virus in Vitro. *F1000Research* **2015**, *4*.

(11) Ekins, S.; Lingerfelt, M. A.; Comer, J. E.; Freiberg, A. N.; Mirsalis, J. C.; O'Loughlin, K.; Harutyunyan, A.; McFarlane, C.; Green, C. E.; Madrid, P. B. Efficacy of Tilorone Dihydrochloride against Ebola Virus Infection. *Antimicrob. Agents Chemother.* **2017**, AAC.01711-17.

(12) Noel T. Southall, Ajit Jadhav, Ruili Huang, Trung Nguyen, and Y. W. Enabling the Large-Scale Analysis of Quantitative High-Throughput. In *Handbook of Drug Screening, Second Edition*; 2009; pp 442–463.

(13) Chandran, K.; Sullivan, N. J.; Felbor, U.; Whelan, S. P.; Cunningham, J. M. Endosomal Proteolysis of the Ebola Virus Glycoprotein Is Necessary for Infection. *Science* **2005**, *308* (5728), 1643–1645.

(14) Côté, M.; Misasi, J.; Ren, T.; Bruchez, A.; Lee, K.; Filone, C. M.; Hensley, L.; Li, Q.; Ory, D.; Chandran, K.; Cunningham, J. Small Molecule Inhibitors Reveal Niemann-Pick C1 Is Essential for Ebola Virus Infection. *Nature* **2011**, *477* (7364), 344–348.

(15) Carette, J. E.; Raaben, M.; Wong, A. C.; Herbert, A. S.; Obernosterer, G.; Mulherkar, N.; Kuehne, A. I.; Kranzusch, P. J.; Griffin, A. M.; Ruthel, G.; Cin, P. D.; Dye, J. M.; Whelan, S. P.; Chandran, K.; Brummelkamp, T. R. Ebola Virus Entry Requires the Cholesterol Transporter Niemann–Pick C1. *Nature* **2011**, *477* (7364), 340–343.

(16) Molina, D. M.; Jafari, R.; Ignatushchenko, M.; Seki, T.; Larsson, E. A.; Dan, C.; Sreekumar, L.; Cao, Y.; Nordlund, P. Monitoring Drug Target Engagement in Cells and Tissues Using the Cellular Thermal Shift Assay. *Science (80-.).* **2013**, *341* (6141).

(17) Jae, L. T.; Brummelkamp, T. R. Emerging Intracellular Receptors for Hemorrhagic Fever Viruses. *Trends Microbiol.* **2015**, *23* (7), 392–400.

(18) Miller, E. H.; Obernosterer, G.; Raaben, M.; Herbert, A. S.; Deffieu, M. S.; Krishnan, A.; Ndungo, E.; Sandesara, R. G.; Carette, J. E.; Kuehne, A. I.; Ruthel, G.; Pfeffer, S. R.; Dye, J. M.; Whelan, S. P.; Brummelkamp, T. R.; Chandran, K. Ebola Virus Entry Requires the Host-Programmed Recognition of an Intracellular Receptor. *EMBO J.* **2012**, *31* (8), 1947–1960.

(19) Herbert, A. S.; Davidson, C.; Kuehne, A. I.; Bakken, R.; Braigen, S. Z.; Gunn, K. E.; Whelan, S. P.; Brummelkamp, T. R.; Twenhafel, N. A.; Chandran, K.; Walkley, S. U.; Dye, J. M. Niemann-Pick C1 Is Essential for Ebolavirus Replication and Pathogenesis in Vivo. *MBio* **2015**, *6* (3), e00565-15.

(20) Jaishy, B.; Abel, E. D. Lipids, Lysosomes, and Autophagy. *J. Lipid Res.* **2016**, *57* (9), 1619–1635.

(21) Vale, G. A. Proceedings: Attractants for Controlling and Surveying Tsetse Populations. *Trans. R. Soc. Trop. Med. Hyg.* **1974**, *68* (1), 11.

(22) Sterling, T.; Irwin, J. J. ZINC 15--Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55* (11), 2324–2337.

(23) Jadhav, A.; Ferreira, R. S.; Klumpp, C.; Mott, B. T.; Austin, C. P.; Inglese, J.; Thomas, C. J.; Maloney, D. J.; Shoichet, B. K.; Simeonov, A. Quantitative Analyses of Aggregation, Autofluorescence, and Reactivity Artifacts in a Screen for Inhibitors of a Thiol Protease. *J. Med. Chem.* **2010**, *53* (1), 37–51.

(24) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719–2740.

(25) Canny, S. A.; Cruz, Y.; Southern, M. R.; Griffin, P. R. PubChem Promiscuity : A Web Resource for Gathering Compound Promiscuity Data from PubChem. *Bioinformatics* **2012**, *28* (1), 140–141.

(26) Kouznetsova, J.; Sun, W.; Martínez-Romero, C.; Tawa, G.; Shinn, P.; Chen, C. Z.; Schimmer, A.; Sanderson, P.; McKew, J. C.; Zheng, W.; García-Sastre, A. Identification of 53 Compounds That Block Ebola Virus-like Particle Entry via a Repurposing Screen of Approved Drugs. *Emerg. Microbes Infect.* **2014**, *3* (12), e84.

(27) Warren, T. K.; Wells, J.; Panchal, R. G.; Stuthman, K. S.; Garza, N. L.; Van Tongeren, S. A.; Dong, L.; Retterer, C. J.; Eaton, B. P.; Pegoraro, G.; Honnold, S.; Bantia, S.; Kotian, P.; Chen, X.; Taubenheim, B. R.; Welch, L. S.; Minning, D. M.; Babu, Y. S.; Sheridan, W. P.; Bavari, S. Protection against Filovirus Diseases by a Novel Broad-Spectrum Nucleoside Analogue BCX4430. *Nature* **2014**, *508* (7496), 402–405.

(28) Yermolina, M. V.; Wang, J.; Caffrey, M.; Rong, L. L.; Wardrop, D. J. Discovery, Synthesis, and Biological Evaluation of a Novel Group of Selective Inhibitors of Filoviral Entry. *J. Med. Chem.* **2011**, *54* (3), 765–781.

(29) Kubicek, S.; O'Sullivan, R. J.; August, E. M.; Hickey, E. R.; Zhang, Q.; Teodoro, M. L.; Rea, S.; Mechtler, K.; Kowalski, J. A.; Homon, C. A.; Kelly, T. A.; Jenuwein, T. Reversal of H3K9me2 by a Small-Molecule Inhibitor for the G9a Histone Methyltransferase. *Mol. Cell* **2007**, *25* (3), 473–481.

(30) Wu, F.; Zhang, Y.; Sun, B.; McMahon, A. P.; Wang, Y. Hedgehog Signaling: From Basic Biology to Cancer Therapy. *Cell Chem. Biol.* **2017**, *24* (3), 252–280.

(31) Fan, H.; Du, X.; Zhang, J.; Zheng, H.; Lu, X.; Wu, Q.; Li, H.; Wang, H.; Shi, Y.; Gao, G.; Zhou, Z.; Tan, D.-X.; Li, X. Selective Inhibition of Ebola Entry with Selective Estrogen Receptor Modulators by Disrupting the Endolysosomal Calcium. *Sci. Rep.* **2017**, *7*, 41226.

(32) Johansen, L. M.; Brannan, J. M.; Delos, S. E.; Shoemaker, C. J.; Stossel, A.; Lear, C.; Hoffstrom, B. G.; DeWald, L. E.; Schornberg, K. L.; Scully, C.; Lehar, J.; Hensley, L. E.; White, J. M.; Olinger, G. G. FDA-Approved Selective Estrogen Receptor Modulators Inhibit Ebola Virus Infection. *Sci. Transl. Med.* **2013**, *5* (190), 190ra79-190ra79.

(33) Sakurai, Y.; Kolokoltsov, A. A.; Chen, C.-C.; Tidwell, M. W.; Bauta, W. E.; Klugbauer, N.; Grimm, C.; Wahl-Schott, C.; Biel, M.; Davey, R. A. Two-Pore Channels Control Ebola Virus Host Cell Entry and Are Drug Targets for Disease Treatment. *Science (80-.).* **2015**, *347* (6225), 995–998.

(34) Wang, Y.; Suzek, T.; Zhang, J.; Wang, J.; He, S.; Cheng, T.; Shoemaker, B. A.; Gindulyte, A.; Bryant, S. H. PubChem BioAssay: 2014 Update. *Nucleic Acids Res.* **2014**, *42* (Database issue), D1075-82.

(35) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2015**, *44* (D1), D1202-13.

(36) Fourches, D.; Muratov, E.; Tropsha, A. Curation of Chemogenomics Data. *Nat. Chem. Biol.* **2015**, *11* (8), 535.

(37) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50* (7), 1189–1204.

(38) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *J. Chem. Inf. Model.* **2016**, *56* (7), 1243–1252.

(39) Wang, Y.; Jadhav, A.; Southal, N.; Huang, R.; Nguyen, D.-T. A Grid Algorithm for High Throughput Fitting of Dose-Response Curve Data. *Curr. Chem. Genomics* **2010**, *4*, 57–66.

(40) Capuzzi, S. J.; Politi, R.; Isayev, O.; Farag, S.; Tropsha, A. QSAR Modeling of Tox21 Challenge Stress Response and Nuclear Receptor Signaling Toxicity Assays. *Front. Environ. Sci.* **2016**, *4*, 3.

(41) Tanimoto, T. IBM Internal Report. In *Armonk: IBM Corp*; 1957.

(42) Accelrys. MACCS structural keys.

(43) Muratov, E. N.; Artemenko, A. G.; Varlamova, E. V; Polischuk, P. G.; Lozitsky, V. P.; Fedchuk, A. S.; Lozitska, R. L.; Gridina, T. L.; Koroleva, L. S.; Sil'nikov, V. N.; Galabov, A. S.; Makarov, V. A.; Riabova, O. B.; Wutzler, P.; Schmidtke, M.; Kuz'min, V. E. Per Aspera Ad Astra: Application of Simplex QSAR Approach in Antiviral Research. *Future Med. Chem.* **2010**, *2* (7), 1205–1226.

(44) Golbraikh, A.; Muratov, E.; Fourches, D.; Tropsha, A. Data Set Modelability by QSAR. *J. Chem. Inf. Model.* **2014**, *54* (1), 1–4.

(45) Walker, T.; Grulke, C. M.; Pozefsky, D.; Tropsha, A. Chembench: A Cheminformatics Workbench. *Bioinformatics* **2010**, *26* (23), 3000–3001.

(46) Capuzzi, S. J.; Kim, I. S.-J.; Lam, W. I.; Thornton, T. E.; Muratov, E. N.; Pozefsky, D.; Tropsha, A. Chembench: A Publicly Accessible, Integrated Cheminformatics Portal. *J. Chem. Inf. Model.* **2017**, *57* (2), 105–108.

(47) Kuz'min, V. E.; Artemenko, a G.; Muratov, E. N. Hierarchical QSAR Technology Based on the Simplex Representation of Molecular Structure. *J. Comput. Aided. Mol. Des.* **2008**, *22* (6–7), 403–421.

(48) Zakharov, A.; Peach, M. L.; Sitzmann, M.; Nicklaus, M. C. A New Approach to Radial

Basis Function Approximation and Its Application to QSAR. *J. Chem. Inf. Model.* **2014**.

(49) Kode. DRAGON 7.0 https://chm.kode-solutions.net/products_dragon.php (accessed May 7, 2016).

(50) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32.

(51) Kuz'min, V. E.; Artemenko, A. G.; Lozytska, R. N.; Fedtchouk, A. S.; Lozitsky, V. P.; Muratov, E. N.; Mescheriakov, A. K. Investigation of Anticancer Activity of Macrocyclic Schiff Bases by Means of 4D-QSAR Based on Simplex Representation of Molecular Structure. *SAR QSAR Environ. Res.* **2005**, *16* (3), 219–230.

(52) Zakharov, A. V; Lagunin, A. A.; Filimonov, D. A.; Poroikov, V. V. Quantitative Prediction of Antitarget Interaction Profiles for Chemical Compounds. *Chem. Res. Toxicol.* **2012**, *25* (11), 2378–2385.

(53) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I.; Marcou, G. ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput. Aided-Drug Des.* **2008**, *4* (3), 191–198.

(54) Xu, M.; Liu, K.; Swaroop, M.; Porter, F. D.; Sidhu, R.; Firnkes, S.; Finkes, S.; Ory, D. S.; Marugan, J. J.; Xiao, J.; Southall, N.; Pavan, W. J.; Davidson, C.; Walkley, S. U.; Remaley, A. T.; Baxa, U.; Sun, W.; McKew, J. C.; Austin, C. P.; Zheng, W. δ-Tocopherol Reduces Lipid Accumulation in Niemann-Pick Type C1 and Wolman Cholesterol Storage Disorders. *J. Biol. Chem.* **2012**, *287* (47), 39349–39360.

(55) Appelqvist, H.; Nilsson, C.; Garner, B.; Brown, A. J.; Kågedal, K.; Ollinger, K. Attenuation of the Lysosomal Death Pathway by Lysosomal Cholesterol Accumulation. *Am. J. Pathol.* **2011**, *178* (2), 629–639.

(56) Liu, Y.-H.; Tsang, J. Y. S.; Ni, Y.-B.; Hlaing, T.; Chan, S.-K.; Chan, K.-F.; Ko, C.-W.; Mujtaba, S. S.; Tse, G. M. Doublecortin-like Kinase 1 Expression Associates with Breast Cancer with Neuroendocrine Differentiation. *Oncotarget* **2016**, *7* (2), 1464–1476.

(57) Patel, O.; Dai, W.; Mentzel, M.; Griffin, M. D. W.; Serindoux, J.; Gay, Y.; Fischer, S.; Sterle, S.; Kropp, A.; Burns, C. J.; Ernst, M.; Buchert, M.; Lucet, I. S. Biochemical and Structural Insights into Doublecortin-like Kinase Domain 1. *Structure* **2016**, *24* (9), 1550–1561.

(58) Westphalen, C. B.; Takemoto, Y.; Tanaka, T.; Macchini, M.; Jiang, Z.; Renz, B. W.; Chen, X.; Ormanns, S.; Nagar, K.; Tailor, Y.; May, R.; Cho, Y.; Asfaha, S.; Worthley, D. L.; Hayakawa, Y.; Urbanska, A. M.; Quante, M.; Reichert, M.; Broyde, J.; Subramaniam, P. S.; Remotti, H.; Su, G. H.; Rustgi, A. K.; Friedman, R. A.; Honig, B.; Califano, A.; Houchen, C. W.; Olive, K. P.; Wang, T. C. Dclk1 Defines Quiescent Pancreatic Progenitors That Promote Injury-Induced Regeneration and Tumorigenesis. *Cell Stem Cell* **2016**, *18* (4), 441–455.

(59) Bailey, J. M.; Alsina, J.; Rasheed, Z. A.; McAllister, F. M.; Fu, Y. Y.; Plentz, R.; Zhang,

H.; Pasricha, P. J.; Bardeesy, N.; Matsui, W.; Maitra, A.; Leach, S. D. DCLK1 Marks a Morphologically Distinct Subpopulation of Cells with Stem Cell Properties in Preinvasive Pancreatic Cancer. *Gastroenterology* **2014**, *146* (1), 245–256.

(60) Ali, N.; Chandrakesan, P.; Nguyen, C. B.; Husain, S.; Gillaspy, A. F.; Huycke, M.; Berry, W. L.; May, R.; Qu, D.; Weygant, N.; Sureban, S. M.; Bronze, M. S.; Dhanasekaran, D. N.; Houchen, C. W. Inflammatory and Oncogenic Roles of a Tumor Stem Cell Marker Doublecortin-like Kinase (DCLK1) in Virus-Induced Chronic Liver Diseases. *Oncotarget* **2015**, *6* (24), 20327–20344.

(61) Whorton, J.; Sureban, S. M.; May, R.; Qu, D.; Lightfoot, S. A.; Madhoun, M.; Johnson, M.; Tierney, W. M.; Maple, J. T.; Vega, K. J.; Houchen, C. W. DCLK1 Is Detectable in Plasma of Patients with Barrett's Esophagus and Esophageal Adenocarcinoma. *Dig. Dis. Sci.* **2015**, *60* (2), 509–513.

(62) Zhang, S.; Zhang, G.; Guo, H. DCAMKL1 Is Associated with the Malignant Status and Poor Outcome in Bladder Cancer. *Tumour Biol.* **2017**, *39* (6), 1010428317703822.

(63) Tao, H.; Tanaka, T.; Okabe, K. Doublecortin and CaM Kinase-like-1 Expression in Pathological Stage I Non-Small Cell Lung Cancer. *J. Cancer Res. Clin. Oncol.* **2017**, *143* (8), 1449–1459.

(64) Nakanishi, Y.; Seno, H.; Fukuoka, A.; Ueo, T.; Yamaga, Y.; Maruno, T.; Nakanishi, N.; Kanda, K.; Komekado, H.; Kawada, M.; Isomura, A.; Kawada, K.; Sakai, Y.; Yanagita, M.; Kageyama, R.; Kawaguchi, Y.; Taketo, M. M.; Yonehara, S.; Chiba, T. Dclk1 Distinguishes between Tumor and Normal Stem Cells in the Intestine. *Nat. Genet.* **2013**, *45* (1), 98–103.

(65) Fedorov, O.; Müller, S.; Knapp, S. The (Un)targeted Cancer Kinome. *Nat. Chem. Biol.* **2010**, *6* (3), 166–169.

(66) Drewry, D. H.; Wells, C. I.; Andrews, D. M.; Angell, R.; Al-Ali, H.; Axtman, A. D.; Capuzzi, S. J.; Elkins, J. M.; Ettmayer, P.; Frederiksen, M.; Gileadi, O.; Gray, N.; Hooper, A.; Knapp, S.; Laufer, S.; Luecking, U.; Michaelides, M.; Mü Ller, S.; Muratov, E.; Aldrin, R.; 17, D.; Saikatendu, K. S.; Treiber, D. K.; Zuercher, W. J.; Willson, T. M. Progress towards a Public Chemogenomic Set for Protein Kinases and a Call for Contributions. *PLoS One* **2017**, *12* (8), e0181585.

(67) Arrowsmith, C. H.; Audia, J. E.; Austin, C.; Baell, J.; Bennett, J.; Blagg, J.; Bountra, C.; Brennan, P. E.; Brown, P. J.; Bunnage, M. E.; Buser-Doepner, C.; Campbell, R. M.; Carter, A. J.; Cohen, P.; Copeland, R. A.; Cravatt, B.; Dahlin, J. L.; Dhanak, D.; Edwards, A. M.; Frederiksen, M.; Frye, S. V; Gray, N.; Grimshaw, C. E.; Hepworth, D.; Howe, T.; Huber, K. V. M.; Jin, J.; Knapp, S.; Kotz, J. D.; Kruger, R. G.; Lowe, D.; Mader, M. M.; Marsden, B.; Mueller-Fahrnow, A.; Müller, S.; O'Hagan, R. C.; Overington, J. P.; Owen, D. R.; Rosenberg, S. H.; Roth, B.; Roth, B.; Ross, R.; Schapira, M.; Schreiber, S. L.; Shoichet, B.; Sundström, M.; Superti-Furga, G.; Taunton, J.; Toledo-Sherman, L.; Walpole, C.; Walters, M. A.; Willson, T. M.; Workman, P.; Young, R. N.; Zuercher, W. J. The Promise and Peril of Chemical Probes. *Nat. Chem. Biol.* **2015**, *11* (8), 536–541.

(68) Elkins, J. M.; Fedele, V.; Szklarz, M.; Abdul Azeez, K. R.; Salah, E.; Mikolajczyk, J.; Romanov, S.; Sepetov, N.; Huang, X.-P.; Roth, B. L.; Al Haj Zen, A.; Fourches, D.; Muratov, E.; Tropsha, A.; Morris, J.; Teicher, B. A.; Kunkel, M.; Polley, E.; Lackey, K. E.; Atkinson, F. L.; Overington, J. P.; Bamborough, P.; Müller, S.; Price, D. J.; Willson, T. M.; Drewry, D. H.; Knapp, S.; Zuercher, W. J. Comprehensive Characterization of the Published Kinase Inhibitor Set. *Nat. Biotechnol.* **2015**, *34* (1), 95–103.

(69) Miduturu, C. V; Deng, X.; Kwiatkowski, N.; Yang, W.; Brault, L.; Filippakopoulos, P.; Chung, E.; Yang, Q.; Schwaller, J.; Knapp, S.; King, R. W.; Lee, J.-D.; Herrgard, S.; Zarrinkar, P.; Gray, N. S. High-Throughput Kinase Profiling: A More Efficient Approach toward the Discovery of New Kinase Inhibitors. *Chem. Biol.* **2011**, *18* (7), 868–879.

(70) Deng, X.; Elkins, J. M.; Zhang, J.; Yang, Q.; Erazo, T.; Gomez, N.; Choi, H. G.; Wang, J.; Dzamko, N.; Lee, J.-D.; Sim, T.; Kim, N.; Alessi, D. R.; Lizcano, J. M.; Knapp, S.; Gray, N. S. Structural Determinants for ERK5 (MAPK7) and Leucine Rich Repeat Kinase 2 Activities of Benzo[e]pyrimido-[5,4-B]diazepine-6(11H)-Ones. *Eur. J. Med. Chem.* **2013**, *70*, 758–767.

(71) Deng, X.; Dzamko, N.; Prescott, A.; Davies, P.; Liu, Q.; Yang, Q.; Lee, J.-D.; Patricelli, M. P.; Nomanbhoy, T. K.; Alessi, D. R.; Gray, N. S. Characterization of a Selective Inhibitor of the Parkinson's Disease Kinase LRRK2. *Nat. Chem. Biol.* **2011**, *7* (4), 203–205.

(72) Dearden, J. C.; Cronin, M. T. D.; Kaiser, K. L. E. How Not to Develop a Quantitative Structure–activity or Structure–property Relationship (QSAR/QSPR). *SAR QSAR Environ. Res.* **2009**, *20* (3–4), 241–266.

(73) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57* (12), 4977–5010.

(74) Zakharov, A. V; Peach, M. L.; Sitzmann, M.; Nicklaus, M. C. QSAR Modeling of Imbalanced High-Throughput Screening Data in PubChem. *J. Chem. Inf. Model.* **2014**, *54* (3), 705–712.

(75) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55* (7), 2932–2942.

(76) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57* (1), 18–28.

(77) Uitdehaag, J. C. M.; Verkaar, F.; Alwan, H.; de Man, J.; Buijsman, R. C.; Zaman, G. J. R. A Guide to Picking the Most Selective Kinase Inhibitor Tool Compounds for Pharmacological Validation of Drug Targets. *Br. J. Pharmacol.* **2012**, *166* (3), 858–876.

(78) Elkins, J. M.; Wang, J.; Deng, X.; Pattison, M. J.; Arthur, J. S. C.; Erazo, T.; Gomez, N.; Lizcano, J. M.; Gray, N. S.; Knapp, S. X-Ray Crystal Structure of ERK5 (MAPK7) in

Complex with a Specific Inhibitor. *J. Med. Chem.* **2013**, *56* (11), 4413–4421.

(79) Kwiatkowski, N.; Deng, X.; Wang, J.; Tan, L.; Villa, F.; Santaguida, S.; Huang, H.-C.; Mitchison, T.; Musacchio, A.; Gray, N. Selective Aurora Kinase Inhibitors Identified Using a Taxol-Induced Checkpoint Sensitivity Screen. *ACS Chem. Biol.* **2012**, *7* (1), 185–196.

(80) Ferguson, F. M.; Ni, J.; Zhang, T.; Tesar, B.; Sim, T.; Kim, N. D.; Deng, X.; Brown, J. R.; Zhao, J. J.; Gray, N. S. Discovery of a Series of 5,11-Dihydro-6H-benzo[e]pyrimido[5,4-b][1,4]diazepin-6-Ones as Selective PI3K-Δ/γ Inhibitors. *ACS Med. Chem. Lett.* **2016**, *7* (10), 908–912.

(81) Deng, X.; Yang, Q.; Kwiatkowski, N.; Sim, T.; McDermott, U.; Settleman, J. E.; Lee, J.-D.; Gray, N. S. Discovery of a Benzo[e]pyrimido-[5,4-b][1,4]diazepin-6(11H)-One as a Potent and Selective Inhibitor of Big MAP Kinase 1. *ACS Med. Chem. Lett.* **2011**, *2* (3), 195–200.

(82) Furtmann, N.; Hu, Y.; Gütschow, M.; Bajorath, J. Identification and Analysis of the Currently Available High-Confidence Three-Dimensional Activity Cliffs. *RSC Adv.* **2015**, *5* (54), 43660–43668.

(83) Fabian, M. A.; Biggs, W. H.; 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.; Grotzfeld, R. M.; Herrgard, S.; Insko, D. E.; Insko, M. A.; Lai, A. G.; Lélias, 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* (3), 329–336.

(84) Patricelli, M. P.; Nomanbhoy, T. K.; Wu, J.; Brown, H.; Zhou, D.; Zhang, J.; Jagannathan, S.; Aban, A.; Okerberg, E.; Herring, C.; Nordin, B.; Weissig, H.; Yang, Q.; Lee, J. D.; Gray, N. S.; Kozarich, J. W. In Situ Kinase Profiling Reveals Functionally Relevant Properties of Native Kinases. *Chem. Biol.* **2011**, *18* (6), 699–710.

(85) Lagunin, A.; Zakharov, A.; Filimonov, D.; Poroikov, V. QSAR Modelling of Rat Acute Toxicity on the Basis of PASS Prediction. *Mol. Inform.* **2011**, *30* (2–3), 241–250.

(86) Filimonov, D. A.; Zakharov, A. V.; Lagunin, A. A.; Poroikov, V. V. QNA-Based “Star Track” QSAR Approach. *SAR QSAR Environ. Res.* **2009**, *20* (7–8), 679–709.

(87) Lagunin, A. A.; Zakharov, A. V.; Filimonov, D. A.; Poroikov, V. V. A New Approach to QSAR Modelling of Acute Toxicity. *SAR QSAR Environ. Res.* **2007**, *18* (3–4), 285–298.

(88) Kuz’min, V. E.; Muratov, E. N.; Artemenko, A. G.; Varlamova, E. V.; Gorb, L.; Wang, J.; Leszczynski, J. Consensus QSAR Modeling of Phosphor-Containing Chiral AChE Inhibitors. *QSAR Comb. Sci.* **2009**, *28* (6–7), 664–677.

(89) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50* (3), 339–348.

(90) Rose, P. W.; Prlić, A.; Bi, C.; Bluhm, W. F.; Christie, C. H.; Dutta, S.; Green, R. K.; Goodsell, D. S.; Westbrook, J. D.; Woo, J.; Young, J.; Zardecki, C.; Berman, H. M.; Bourne, P. E.; Burley, S. K. The RCSB Protein Data Bank: Views of Structural Biology for Basic and Applied Research and Education. *Nucleic Acids Res.* **2015**, *43* (Database issue), D345-56.

(91) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. *J. Med. Chem.* **2006**, *49* (21), 6177–6196.

(92) Collins, F. S.; Tabak, L. A. Policy: NIH Plans to Enhance Reproducibility. *Nature* **2014**, *505* (7485), 612–613.

(93) Frye, S. V; Arkin, M. R.; Arrowsmith, C. H.; Conn, P. J.; Glicksman, M. A.; Hull-Ryde, E. A.; Slusher, B. S. Tackling Reproducibility in Academic Preclinical Drug Discovery. *Nat. Rev. Drug Discov.* **2015**, *14*, 733–734.

(94) Dahlin, J. L.; Baell, J.; Walters, M. A. Assay Interference by Chemical Reactivity. *Assay Guid. Man.* **2015**.

(95) Sassano, M. F.; Doak, A. K.; Roth, B. L.; Shoichet, B. K. Colloidal Aggregation Causes Inhibition of G Protein-Coupled Receptors. *J. Med. Chem.* **2013**, *56* (6), 2406–2414.

(96) Thorne, N.; Auld, D. S.; Inglese, J. Apparent Activity in High-Throughput Screening: Origins of Compound-Dependent Assay Interference. *Curr Opin Chem Biol* **2010**, *14* (3), 315–324.

(97) Kenny, P. W. Comment on The Ecstasy and Agony of Assay Interference Compounds. *J. Chem. Inf. Model.* **2017**, *57* (11), 2640–2645.

(98) Tropsha, a; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation Is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *Qsar Comb. Sci.* **2003**, *22* (3), 69–77.

(99) Dahlin, J. L.; Nissink, J. W. M.; Strasser, J. M.; Francis, S.; Higgins, L.; Zhou, H.; Zhang, Z.; Walters, M. A. PAINS in the Assay: Chemical Mechanisms of Assay Interference and Promiscuous Enzymatic Inhibition Observed during a Sulfhydryl-Scavenging HTS. *J. Med. Chem.* **2015**, *58* (5), 2091–2113.

(100) Erlanson, D. A. Learning from PAINful Lessons. *J. Med. Chem.* **2015**, *58* (5), 2088–2090.

(101) Mok, N. Y.; Maxe, S.; Brenk, R. Locating Sweet Spots for Screening Hits and Evaluating Pan-Assay Interference Filters from the Performance Analysis of Two Lead-like Libraries. *J. Chem. Inf. Model.* **2013**, *53* (3), 534–544.

(102) Lagorce, D.; Sperandio, O.; Baell, J. B.; Miteva, M. A.; Villoutreix, B. O. FAF-Drugs3: A Web Server for Compound Property Calculation and Chemical Library Design. *Nucleic Acids Res.* **2015**, *43* (W1), W200-7.

(103) Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55* (11), 2324–2337.

(104) Baell, J. B. Screening-Based Translation of Public Research Encounters Painful Problems. *ACS Med. Chem. Lett.* **2015**, *6* (3), 229–234.

(105) Neves, B. J.; Dantas, R. F.; Senger, M. R.; Melo-Filho, C. C.; Valente, W. C. G.; de Almeida, A. C. M.; Rezende-Neto, J. M.; Lima, E. F. C.; Paveley, R.; Furnham, N.; Muratov, E.; Kamentsky, L.; Carpenter, A. E.; Braga, R. C.; Silva-Junior, F. P.; Andrade, C. H. Discovery of New Anti-Schistosomal Hits by Integration of QSAR-Based Virtual Screening and High Content Screening. *J. Med. Chem.* **2016**, *59* (15), 7075–7088.

(106) Williamson, A. E.; Ylioja, P. M.; Robertson, M. N.; Antonova-Koch, Y.; Avery, V.; Baell, J. B.; Batchu, H.; Batra, S.; Burrows, J. N.; Bhattacharyya, S.; Calderon, F.; Charman, S. A.; Clark, J.; Crespo, B.; Dean, M.; Debbert, S. L.; Delves, M.; Dennis, A. S. M.; Deroose, F.; Duffy, S.; Fletcher, S.; Giaever, G.; Hallyburton, I.; Gamo, F.-J.; Gebbia, M.; Guy, R. K.; Hungerford, Z.; Kirk, K.; Lafuente-Monasterio, M. J.; Lee, A.; Meister, S.; Nislow, C.; Overington, J. P.; Papadatos, G.; Patiny, L.; Pham, J.; Ralph, S. A.; Ruecker, A.; Ryan, E.; Southan, C.; Srivastava, K.; Swain, C.; Tarnowski, M. J.; Thomson, P.; Turner, P.; Wallace, I. M.; Wells, T. N. C.; White, K.; White, L.; Willis, P.; Winzeler, E. A.; Wittlin, S.; Todd, M. H. Open Source Drug Discovery: Highly Potent Antimalarial Compounds Derived from the Tres Cantos Arylpyrroles. *ACS Cent. Sci.* **2016**, *2* (10), 687–701.

(107) ACS Publications. Guidelines for Authors. *J. Med. Chem.* **2016**, No. January, 5–6.

(108) Chai, C. L.; Mátyus, P. One Size Does Not Fit All: Challenging Some Dogmas and Taboos in Drug Discovery. *Future Med. Chem.* **2016**, *8* (1), 29–38.

(109) Senger, M. R.; Fraga, C. A. M.; Dantas, R. F.; Silva, F. P. Filtering Promiscuous Compounds in Early Drug Discovery: Is It a Good Idea? *Drug Discov. Today* **2016**, *21* (6), 868–872.

(110) Wassermann, A. M.; Lounkine, E.; Hoepfner, D.; Le Goff, G.; King, F. J.; Studer, C.; Peltier, J. M.; Grippo, M. L.; Prindle, V.; Tao, J.; Schuffenhauer, A.; Wallace, I. M.; Chen, S.; Krastel, P.; Cobos-Correa, A.; Parker, C. N.; Davies, J. W.; Glick, M. Dark Chemical Matter as a Promising Starting Point for Drug Lead Discovery. *Nat. Chem. Biol.* **2015**, *11* (12), 958–966.

(111) Schorpp, K.; Rothenaigner, I.; Salmina, E.; Reinshagen, J.; Low, T.; Brenke, J. K.; Gopalakrishnan, J.; Tetko, I. V; Gul, S.; Hadian, K. Identification of Small-Molecule Frequent Hitters from AlphaScreen High-Throughput Screens. *J. Biomol. Screen.* **2014**, *19* (5), 715–726.

(112) Schurer, S.; Vempati, U.; Smith, R.; Southern, M.; Lemmon, V. BioAssay Ontology Annotations Facilitate Cross-Analysis of Diverse High-Throughput Screening Data Sets. *J Biomol Screen.* **2011**, *16* (4), 415–426.

(113) Bruns, R. F.; Watson, I. A. Rules for Identifying Potentially Reactive or Promiscuous

Compounds. *J. Med. Chem.* **2012**, *55* (22), 9763–9772.

(114) M Nissink, J. W.; Blackburn, S. Quantification of Frequent-Hitter Behavior Based on Historical High-Throughput Screening Data. *Future Med. Chem.* **2014**, *6* (10), 1113–1126.

(115) Baell, J. B. Feeling Nature ' S PAINS: Natural Products, Natural Product Drugs, and Pan Assay Interference Compounds (PAINS). *J. Nat. Prod.* **2015**.

(116) Alves, V. M.; Muratov, E. N.; Capuzzi, S. J.; Politi, R.; Low, Y.; Braga, R. C.; Zakharov, A. V.; Sedykh, A.; Mokshyna, E.; Farag, S.; Andrade, C. H.; Kuz'min, V. E.; Fourches, D.; Tropsha, A. Alarms about Structural Alerts. *Green Chem.* **2016**, *18* (16), 4348–4360.

(117) Oprea, T. I.; Mestres, J. Drug Repurposing: Far beyond New Targets for Old Drugs. *AAPS J.* **2012**, *14* (4), 759–763.

(118) Blatt, J.; Farag, S.; Corey, S. J.; Sarrimanolis, Z.; Muratov, E.; Fourches, D.; Tropsha, A.; Janzen, W. P. Expanding the Scopre of Drug Repurposing in Pediatrics: The Children's Pharmacy Collaborative. *Drug Discov. Today* **2014**, *19* (11), 1696–1698.

(119) Jasial, S.; Hu, Y.; Bajorath, J. Determining the Degree of Promiscuity of Extensively Assayed Compounds. *PLoS One* **2016**, *11* (4), e0153873.

(120) Dahlin, J. L.; Walters, M. A. How to Triage PAINS-Full Research. *Assay Drug Dev. Technol.* **2016**, *14* (3), 168–174.

(121) Yang, J. J.; Ursu, O.; Lipinski, C. A.; Sklar, L. A.; Oprea, T. I.; Bologa, C. G. Badapple: Promiscuity Patterns from Noisy Evidence. *J. Cheminform.* **2016**, *8* (1), 29.

(122) Alves, V. M.; Muratov, E. N.; Fourches, D.; Strickland, J.; Kleinstreuer, N.; Andrade, C. H.; Tropsha, A. Predicting Chemically-Induced Skin Reactions. Part I: QSAR Models of Skin Sensitization and Their Application to Identify Potentially Hazardous Compounds. *Toxicol. Appl. Pharmacol.* **2015**, *284* (2), 262–272.

(123) Alves, V. M.; Muratov, E. N.; Fourches, D.; Strickland, J.; Kleinstreuer, N.; Andrade, C. H.; Tropsha, A. Predicting Chemically-Induced Skin Reactions. Part II: QSAR Models of Skin Permeability and the Relationships between Skin Permeability and Skin Sensitization. *Toxicol. Appl. Pharmacol.* **2015**, *284* (2), 273–280.

(124) Alves, V. M.; Capuzzi, S. J.; Muratov, E. N.; Tropsha, A. QSAR Models Can Provide an Alternative to LLNA Testing for Assessing Human Skin Sensitization. *Green Chem.* **2016**.

(125) Frye, S.; Crosby, M.; Edwards, T.; Juliano, R. US Academic Drug Discovery. *Nat. Rev. Drug Discov.* **2011**, *10* (6), 409–410.

(126) Wang, Y.; Suzek, T.; Zhang, J.; Wang, J.; He, S.; Cheng, T.; Shoemaker, B. A.; Gindulyte, A.; Bryant, S. H. PubChem BioAssay: 2014 Update. *Nucleic Acids Res.* **2014**, *42* (Database issue), D1075-82.

(127) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40* (Database issue), D1100-7.

(128) Przybyła, P.; Shardlow, M.; Aubin, S.; Bossy, R.; Eckart de Castilho, R.; Piperidis, S.; McNaught, J.; Ananiadou, S. Text Mining Resources for the Life Sciences. *J. Biol. databases curation* **2016**, *2016*.

(129) Wei, C.-H.; Kao, H.-Y.; Lu, Z. PubTator: A Web-Based Text Mining Tool for Assisting Biocuration. *Nucleic Acids Res.* **2013**, *41* (W1), W518–W522.

(130) Roberts, R. J. PubMed Central: The GenBank of the Published Literature. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98* (2), 381–382.

(131) Lin, J.; DiCuccio, M.; Grigoryan, V.; Wilbur, W. J. Navigating Information Spaces: A Case Study of Related Article Search in PubMed. *Inf. Process. Manag.* **2008**, *44* (5), 1771–1783.

(132) Baker, N. C.; Hemminger, B. M. Mining Connections between Chemicals, Proteins, and Diseases Extracted from Medline Annotations. *J. Biomed. Inform.* **2010**, *43* (4), 510–519.

(133) Swanson, D. R. Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge. *Perspect. Biol. Med.* **1986**, *30* (1), 7–18.

(134) Swanson, D. R. Migraine and Magnesium: Eleven Neglected Connections. *Perspect. Biol. Med.* **1988**, *31* (4), 526–557.

(135) Nosengo, N. Can You Teach Old Drugs New Tricks? *Nature* **2016**, *534* (7607), 314–316.

(136) Zhou, J.; Shui, Y.; Peng, S.; Li, X.; Mamitsuka, H.; Zhu, S. MeSHSim: An R/Bioconductor Package for Measuring Semantic Similarity over MeSH Headings and MEDLINE Documents. *J. Bioinform. Comput. Biol.* **2015**, *13* (6), 1542002.

(137) Rani, J.; Shah, A. B. R.; Ramachandran, S. pubmed.mineR: An R Package with Text-Mining Algorithms to Analyse PubMed Abstracts. *J. Biosci.* **2015**, *40* (4), 671–682.

(138) Spangler, S.; Myers, J. N.; Stanoi, I.; Kato, L.; Lelescu, A.; Labrie, J. J.; Parikh, N.; Lisewski, A. M.; Donehower, L.; Chen, Y.; Lichtarge, O.; Wilkins, A. D.; Bachman, B. J.; Nagarajan, M.; Dayaram, T.; Haas, P.; Regenbogen, S.; Pickering, C. R.; Comer, A. Automated Hypothesis Generation Based on Mining Scientific Literature. *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '14* **2014**, 1877–1886.

(139) McPherson, S. JavaServer Pages: A Developer's Perspective http://www.oracle.com/technetwork/articles/javase/jsp-135132.html (accessed May 7, 2015).

(140) Heinrich, M. C.; Corless, C. L.; Demetri, G. D.; Blanke, C. D.; von Mehren, M.; Joensuu, H.; McGreevey, L. S.; Chen, C.-J.; Van den Abbeele, A. D.; Druker, B. J.; Kiese, B.; Eisenberg, B.; Roberts, P. J.; Singer, S.; Fletcher, C. D. M.; Silberman, S.; Dimitrijevic,

S.; Fletcher, J. A. Kinase Mutations and Imatinib Response in Patients with Metastatic Gastrointestinal Stromal Tumor. *J. Clin. Oncol.* **2003**, *21* (23), 4342–4349.

(141) Fuehrer, N. E.; Marchevsky, A. M.; Jagirdar, J. Presence of c-KIT-Positive Mast Cells in Obliterative Bronchiolitis from Diverse Causes. *Arch. Pathol. Lab. Med.* **2009**, *133* (9), 1420–1425.

(142) Bauer, S.; Duensing, A.; Demetri, G. D.; Fletcher, J. A. KIT Oncogenic Signaling Mechanisms in Imatinib-Resistant Gastrointestinal Stromal Tumor: PI3-kinase/AKT Is a Crucial Survival Pathway. *Oncogene* **2007**, *26* (54), 7560–7568.

(143) Reber, L.; Da Silva, C. A.; Frossard, N. Stem Cell Factor and Its Receptor c-Kit as Targets for Inflammatory Diseases. *Eur. J. Pharmacol.* **2006**, *533* (1–3), 327–340.

(144) Cleary, R. A.; Wang, R.; Wang, T.; Tang, D. D. Role of Abl in Airway Hyperresponsiveness and Airway Remodeling. *Respir. Res.* **2013**, *14* (1), 105.

(145) Arend, K. C.; Lenarcic, E. M.; Vincent, H. A.; Rashid, N.; Lazear, E.; McDonald, I. M.; Gilbert, T. S. K.; East, M. P.; Herring, L. E.; Johnson, G. L.; Graves, L. M.; Moorman, N. J. Kinome Profiling Identifies Druggable Targets for Novel Human Cytomegalovirus (HCMV) Antivirals. *Mol. Cell. Proteomics* **2017**, *16* (4 suppl 1), S263–S276.

(146) Baker, N. C.; Fourches, D.; Tropsha, A. Drug Side Effect Profiles as Molecular Descriptors for Predictive Modeling of Target Bioactivity. *Mol. Inform.* **2015**, *34* (2–3), 160–170.

(147) Alves, V. M.; Capuzzi, S. J.; Muratov, E. N.; Braga, R. C.; Thornton, T. E.; Fourches, D.; Strickland, J.; Kleinstreuer, N.; Andrade, C. H.; Tropsha, A. QSAR Models of Human Data Can Enrich or Replace LLNA Testing for Human Skin Sensitization. *Green Chem.* **2016**, *18* (24), 6501–6515.