

# Understanding Cytotoxicity and Cytostaticity in a High-Throughput Screening Collection

Lewis H. Mervin,<sup>†</sup> Qing Cao,<sup>‡</sup> Ian P. Barrett,<sup>§</sup> Mike A. Firth,<sup>§</sup> David Murray,<sup>||</sup> Lisa McWilliams,<sup>||</sup> Malcolm Haddrick,<sup>||</sup> Mark Wigglesworth,<sup>||</sup> Ola Engkvist,<sup>⊥</sup> and Andreas Bender<sup>\*,†</sup>

<sup>†</sup>Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Cambridge, United Kingdom

<sup>‡</sup>Discovery Sciences, AstraZeneca R&D, Waltham, United States

<sup>§</sup>Discovery Sciences, AstraZeneca R&D, Cambridge Science Park, Cambridge, United Kingdom

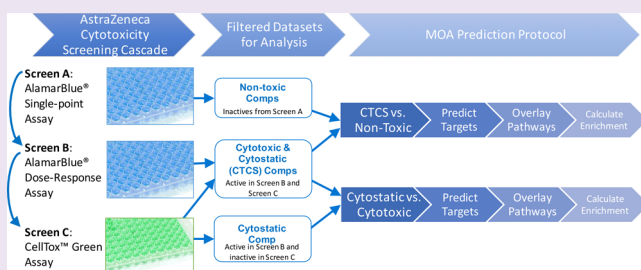
<sup>||</sup>Discovery Sciences, AstraZeneca R&D, Alderley Park, Macclesfield, United Kingdom

<sup>⊥</sup>Discovery Sciences, AstraZeneca R&D, Mölndal, Sweden

## Supporting Information

**ABSTRACT:** While mechanisms of cytotoxicity and cytostaticity have been studied extensively from the biological side, relatively little is currently understood regarding areas of chemical space leading to cytotoxicity and cytostasis in large compound collections. Predicting and rationalizing potential adverse mechanism-of-actions (MoAs) of small molecules is however crucial for screening library design, given the link of even low level cytotoxicity and adverse events observed in man. In this study, we analyzed results from a cell-based cytotoxicity screening cascade, comprising 296 970 nontoxic,

5784 cytotoxic and cytostatic, and 2327 cytostatic-only compounds evaluated on the THP-1 cell-line. We employed an *in silico* MoA analysis protocol, utilizing 9.5 million active and 602 million inactive bioactivity points to generate target predictions, annotate predicted targets with pathways, and calculate enrichment metrics to highlight targets and pathways. Predictions identify known mechanisms for the top ranking targets and pathways for both phenotypes after review and indicate that while processes involved in cytotoxicity versus cytostaticity seem to overlap, differences between both phenotypes seem to exist to some extent. Cytotoxic predictions highlight many kinases, including the potentially novel cytotoxicity-related target STK32C, while cytostatic predictions outline targets linked with response to DNA damage, metabolism, and cytoskeletal machinery. Fragment analysis was also employed to generate a library of toxicophores to improve general understanding of the chemical features driving toxicity. We highlight substructures with potential kinase-dependent and kinase-independent mechanisms of toxicity. We also trained a cytotoxic classification model on proprietary and public compound readouts, and prospectively validated these on 988 novel compounds comprising difficult and trivial testing instances, to establish the applicability domain of models. The proprietary model performed with precision and recall scores of 77.9% and 83.8%, respectively. The MoA results and top ranking substructures with accompanying MoA predictions are available as a platform to assess screening collections.



Profiling compound libraries through phenotypic and high throughput screening (HTS) cascades is a well-established process, originating commonly from a single-concentration assessment of a compound collection, with subsequent hits from this primary screen profiled through a series of follow-up potency, selectivity, and specificity assays.<sup>1</sup> The size, content, and quality of a screening library are intrinsic to the value of the screen output, influencing the future direction of projects and the likelihood of candidate success or attrition.<sup>2</sup> Poor decisions with respect to the identification of good starting points from compound collections can disadvantage the progression of a compound through the drug discovery process and/or lead to the best compounds not being progressed.<sup>3</sup>

Pharmaceutical companies have considered novel ways to improve the chemical equity of compound collections throughout the years, expanding coverage of chemical space

in order to tackle emerging targets.<sup>4</sup> Many have invested in altering the chemical composition of screening libraries in order to improve physicochemical properties, such as reduced average molecular weight and *c* log-P, alleviating many problem hits driven by size and lipophilicity.<sup>5</sup> Current collections are often considered to be of “high-quality.”<sup>6</sup>

Despite the concentrated investment into library optimization, the pharmaceutical industry has however continued to experience high attrition rates due to poor efficacy and the discovery of adverse effects during preclinical animal safety profiling and clinical trials.<sup>7</sup> Safety-related attrition represents a leading cause of project failures at AstraZeneca, where

Received: June 20, 2016

Accepted: August 29, 2016

**Table 1. Literature-Reviewed Important Targets and Pathways Implicated in Small-Molecule Induced Apoptosis and Regulated Necrosis**<sup>25,28,49,a</sup>

	apoptosis		regulated necrosis				
morphology	cytoplasmic shrinkage	translucent cytoplasm					
	chromatin condensation	swelling of organelles					
	nuclear fragmentation	increased volume					
	blebbing of plasma membrane	permeabilization of plasma membrane					
	shedding of apoptotic bodies	mild chromatin condensation, nuclei intact					
mechanism of death	intrinsic apoptosis	necroptosis					
	extrinsic apoptosis	ferroptosis					
		MPT-mediated regulated necrosis					
broad pathways	death receptor pathway	mitotic catastrophe					
	mitochondrial apoptotic pathway	fidelity of DNA regulation/repair					
	TOR pathway inhibition	ER stress					
	cytochrome c/apaf-1/caspase-9 apoptosome complex	FAS pathway					
	TRAIL receptor activation	TNF pathway					
molecular targets	caspases/ effectors	apoptotic factors	DNA regulation/repair/ expression	phosphorylating proteins	ion channels	cytoskeletal proteins	receptors
	•caspases	•APAF-1	•HDACS	•EGFR	•chloride channel 3	•tubulin, complex associated protein 6	•toll-like receptor 2
	•interleukins	•TNFR-1	•TOP1	•RAS		•M-phase phosphoprotein 1	•integrin, B3
	•interferon	•BCL-2	•TOP2	•RAF			
	•procaspase-9		•polymerase (RNA/ DNA)	•MEK			
	•cytochrome P450		•tyrosyl-tRNA synthetase 2	•MAK			
				•CHEK1			
				•JAK			
				•SRC			

<sup>a</sup>Given that some biological processes are more easily modulated by small molecules than others, we could identify in the current study which of those processes are of relevance for the design of small molecule screening libraries. Molecular targets are examples of well studied proteins and are not exhaustive.

unacceptable toxicity profiles were the single most important reason for failure between 2005 and 2010 and accounted for 82% of preclinical project closures.<sup>8</sup> Attrition continues to dominate phase 1 and phase 2 studies, accounting for 62% and 30% of closures, respectively. One reason for high attrition is due to a lack of sufficient consideration for toxicity during compound selection.<sup>9</sup> Indeed, even low-level cytotoxicity can be linked to late-stage adverse effects in subsequent clinical trials, underlining the requirement of considering the toxic tendencies of compounds earlier, even when assembling an HTS library.

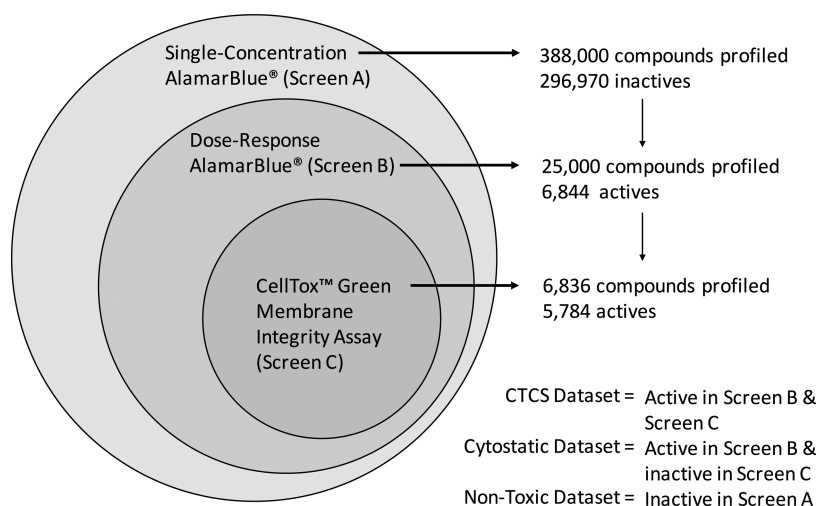
Cytotoxicity has been extensively studied from a biological perspective and is broadly recognized *via* the mutually exclusive categories of “accidental cell death” (ACD) and “regulated cell death” (RCD).<sup>10</sup> ACD can be elicited *via* the detergent properties of small molecules causing cell membrane damage and necrotic lysis, a process lacking specific modulation of cellular targets, rendering it virtually insensitive to biological intervention. In comparison, RCD comprises the biological mechanisms of apoptosis, a genetically encoded process of programmed cell death; autophagy, a survival mechanism involving self-cannibalization of organelles; and *regulated* necrosis, a phenotype exhibiting neither apoptotic nor autophagic characteristics, wherein cells lose membrane integrity and die rapidly.<sup>11</sup>

While RCD processes have conventionally been split into these three mutually exclusive cellular states, how these processes converge to elicit cytotoxicity is less clear.<sup>12</sup> It

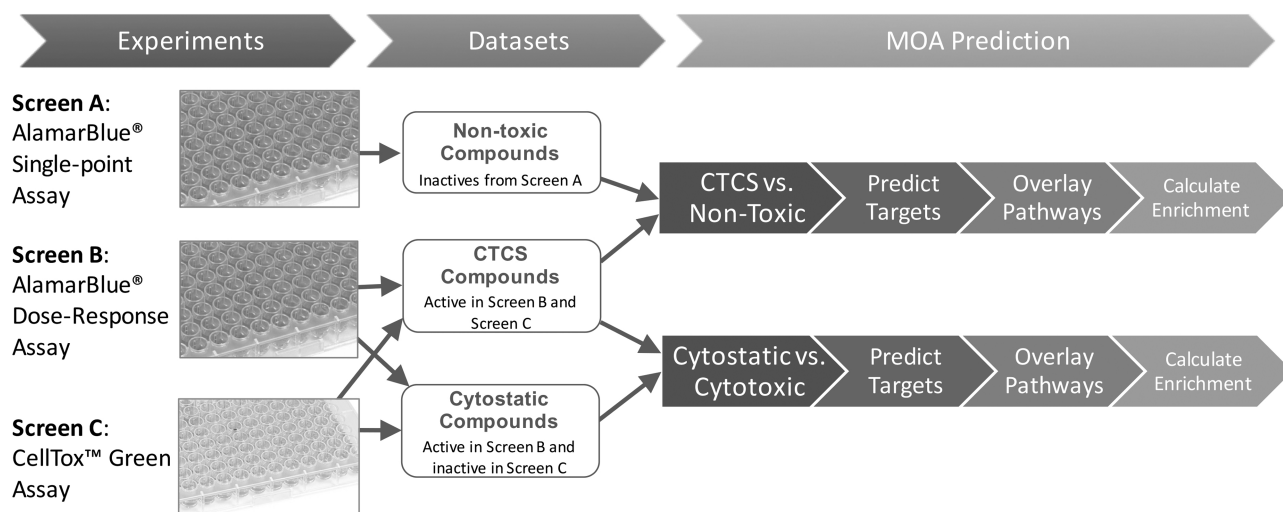
seems that a balanced interplay between apoptosis, autophagy, and regulated necrosis dictates the cellular end point in a specific situation, comprising several central death effector molecules functioning as pro-survival or pro-death mediators.<sup>13</sup> Nevertheless, there are specific targets and pathways more often associated for each of the three conventional processes, although it is nontrivial to review all mechanisms implicated in the phenotype here, due to the complex and interconnected relationships influencing the response.<sup>10–15</sup> Broadly, targets can be split into various general categories (Table 1).

The situation becomes even more complex when considering the cytostatic properties of compounds. Cytostatic compounds do not kill cells but instead invoke the inhibition of cell proliferation and growth. Identified mechanisms comprise DNA damage, DNA polymerase inhibition, increased oncogenic signaling, oxidative stress, and cytoskeletal inhibition.<sup>16</sup>

Various investigations have attempted to discern the cytotoxic and cytostatic effects in cells, where it is argued that either property is dependent on the dose used, time-point measured, phase of the cell cycle upon compound administration, and cellular context.<sup>16</sup> For example, cytostaticity can be evoked in both the S and G<sub>2</sub>-M cell cycle phases, while lethality is shown in the S phase, given that DNA synthesis machinery is often targeted in cytotoxicity.<sup>17</sup> Furthermore, small molecules that are considered cytotoxic are frequently characterized as being cytostatic due to the close overlap of processes. For example, microtubule-targeting agents are inherently cytostatic, since the interference of microtubule



**Figure 1.** Phenotypic screening cascade of the compound collection. Screen A comprises single-concentration AlamarBlue assessment of 388 000 compounds. Subsequent profiling of 25 000 compounds is conducted in screen B *via* a follow-up dose–response AlamarBlue screen. A final CellTox Green membrane integrity screen C is used to profile hits from screen B. A total of 296 970 nontoxic compounds with activity values less than 4.949 are extracted from the primary screen A. A total of 6844 toxic compounds were extracted from dose–response screen B after filtering activity values over 4.949 as a cutoff to define cytotoxicity. Compounds with activity values over 4.949 from screen C are applied as an additional filter to improve the confidence of cytotoxic molecules, producing a library of 5784 CytoToxic and CytoStatic (CTCS) compounds. CTCS compounds are further split into a cytostatic library by filtering for actives from screen B and inactives from screen C, producing a library of 2327 cytostatic compounds.



**Figure 2.** Flow of data through the MoA protocol. Nontoxic, cytotoxic, and cytostatic data sets were extracted from the results of the cytotoxicity experiments. The nontoxic and cytotoxic data sets were passed to the MoA prediction protocol to obtain enriched target and pathway enrichments for a cytotoxic phenotype. The cytostatic and cytotoxic data set were employed for a second MoA prediction step, to obtain enriched cytostatic targets and pathways when compared to cytotoxicity bioactivity profiles.

dynamics leads to stasis in mitosis.<sup>16</sup> However, mitotic arrest is a condition that is poorly tolerated by cells, which is often resolved by cell death (Table 1). Thus, cytotoxic activity is regularly observed after long time points from primarily cytostatic agents.

Although well studied from a biological viewpoint, the links between biological (cytotoxic and cytostatic) space and chemical space are poorly understood.<sup>18</sup> The discontinuous and complex nature of cytotoxicity, due to the multitude of biological processes involved, is often characterized by regions of structure–activity cliffs, with areas of high structural similarity frequently exhibiting low activity similarity.<sup>19</sup> Various studies have utilized machine learning techniques in an attempt to predict the toxic tendencies of small molecules,<sup>18,20</sup> where activity cliff regions are known to be discarded as outliers, cause

overfitting, or increase the prediction error while generating models.<sup>21</sup> Cell-line specific cytotoxicity represents an additional parameter of consideration, with small molecules frequently expressing differential toxicity between cell lines.<sup>22</sup> Protocols are often incapable of maintaining their predictive power outside cell-line training data, which narrows the applicability domain (AD) of models.

The consistency of publicly reported cytotoxicity data has also been shown to be unreliable in many cases, where the misannotation of compounds is known to confound the performance of algorithms.<sup>23</sup> AstraZeneca data obtained through single well controlled and large scale experimentation could be considered of better quality, since toxicity and the associated biological target data to be analyzed are obtained under standardized conditions, from the same compound stocks,

Table 2. Top Enriched Targets for Cytotoxic Compounds versus Non-Toxic Compounds<sup>a</sup>

EGID	name	classification	CTCS hit rate (%)	prediction ratio	Fisher exact test p value	implication in cytotoxicity	associated cell Fate	ref
SNRK	SNF related kinase	kinase	2.44	0.01	$2.58 \times 10^{-176}$	up-regulated in the nucleus during low potassium induced apoptosis; regulates proliferation and $\beta$ -catenin signaling	apoptosis	50
CDK13	cyclin-dependent Kinase 13	kinase	1.43	0.02	$6.13 \times 10^{-95}$	expression is associated with pancreatic cancer	necrosis	51
DSTYK	dual serine/threonine and tyrosine protein kinase	kinase	6.29	0.02	$0.00 \times 10^{00}$	overexpression induces cell death with characteristic apoptotic morphology	apoptosis	52
MAK	male germ cell-associated kinase	kinase	6.14	0.02	$0.00 \times 10^{00}$	role in mitotic defects such as centrosome amplification and lagging chromosomes	necrosis	53
MAP3K6	mitogen-activated protein kinase kinase 6	kinase	7.12	0.02	$0.00 \times 10^{00}$	involved in processes such as stress induced cell cycle arrest, transcription activation and apoptosis	apoptosis	54
STK32A	serine/threonine kinase 32A	kinase	1.69	0.02	$3.99 \times 10^{-109}$	implicated in mitophagy in response to mitochondrial depolarization	autophagy	55
CDKL3	cyclin-dependent kinase-like 3	kinase	4.91	0.02	4.48e-314	RNAi shows a role in cell-cycle control, regulating cell survival and apoptosis	apoptosis	56
IFNG	interferon, gamma	cytokine	1.95	0.02	$3.53 \times 10^{-124}$	induces apoptosis in multiple cell lines	apoptosis	57
NIM1K	NIM1 serine/threonine protein kinase	kinase	9.49	0.03	$0.00 \times 10^{00}$	Nim1/cdr1 is as a positive regulator of mitosis	necrosis	58
WEE1	WEE1 G2 checkpoint kinase	kinase	6.74	0.03	$0.00 \times 10^{00}$	Wee1 markedly accelerates apoptosis	apoptosis	59
CASK	calcium/calmodulin-dependent serine protein kinase	kinase	5.65	0.03	$0.00 \times 10^{00}$	overexpression results in reduced rate of cell viability, while inhibition increases cell growth rate	necrosis	60
STK32C	serine/threonine kinase 32C	kinase	5.69	0.03	$0.00 \times 10^{00}$	unknown function, but is highly expressed in the brain	?	29
IGF1	insulin-like growth factor 1 (somatomedin C)	growth factor	1.09	0.03	$1.41 \times 10^{-65}$	expression has an antiapoptotic effect with increased chance of survival in the presence of genetic damage	apoptosis	61
CDKL5	cyclin-dependent kinase-like 5	kinase	8.54	0.03	$0.00 \times 10^{00}$	associated with strong inhibition of cell proliferation with no increase in apoptotic cell death	necrosis	62
EIF2AK1	eukaryotic translation initiation factor 2-alpha kinase 1	kinase	3.20	0.03	$1.27 \times 10^{-188}$	phosphorylated in response to stress, leading to induction of apoptosis	apoptosis	63

<sup>a</sup>The top 15 enriched targets/pathways are ranked using the *Prediction Ratio* and accompanying Fisher's test p values. Results illustrate a high frequency of hits does not necessarily correspond to a high enrichment score, demonstrating the importance of normalizing predictions to reduce biases in the chemical and bioactivity space (e.g. target model sizes, promiscuity, and sampling bias). The top enriched targets show links to cell death in the literature; show a mix of apoptotic, autophagic, and necrotic agents; and find overlap with targets in Table 1. Fisher's test p values of  $<0.00 \times 10^{00}$  indicate scores that are less than the smallest numerical value allowed in Python.

Table 3. Top Enriched Pathways for the Cytotoxic Compounds versus Non-Toxic Compounds<sup>a</sup>

WikiPathways ID	name	biological link	CTCS hit rate (%)	prediction ratio	Fisher exact test p value
WP3390	uptake and function of anthrax toxins	kinase activity	0.054	0.11	$0.00 \times 10^{00}$
WP1861	mRNA Capping	gene expression	0.058	0.13	$0.00 \times 10^{00}$
WP405	eukaryotic transcription initiation	gene expression	0.059	0.14	$0.00 \times 10^{00}$
WP453	inflammatory response pathway	immune response	0.061	0.15	$0.00 \times 10^{00}$
WP2760	signaling by BMP	cytokine activity	0.083	0.16	$0.00 \times 10^{00}$
WP2752	MyD88-independent TLR3	cytokine activity	0.150	0.16	$0.00 \times 10^{00}$
WP3351	RHO GTPases activate PAKs	cell cycle	0.136	0.16	$0.00 \times 10^{00}$
WP2732	interleukin-2 signaling	cytokine activity	0.125	0.17	$0.00 \times 10^{00}$
WP1845	MAPK targets	kinase activity	0.063	0.17	$1.77 \times 10^{-266}$
WP3305	NoRC negatively regulates rRNA expression	gene expression	0.072	0.17	$9.52 \times 10^{-303}$
WP1808	DSCAM interactions	cell recognition	0.137	0.17	$0.00 \times 10^{00}$
WP1906	RNA polymerase II transcription	gene expression	0.103	0.17	$0.00 \times 10^{00}$
WP1799	co-stimulation by the CD28 family	cytokine activity	0.302	0.17	$0.00 \times 10^{00}$
WP2654	mitotic prophase	cell cycle	0.139	0.18	$0.00 \times 10^{00}$
WP2768	nonhomologous end joining	DNA repair	0.318	0.19	$0.00 \times 10^{00}$

<sup>a</sup>The highest enriched pathways have biological links to the phenotype at different prediction rates. Pathways can be classified into processes implicated in cell cycle, DNA repair, gene expression, and kinase mediated events. WP3390 and WP1808 are examples of tangential signals that should be interpreted with caution. Prediction percentage is calculated from the number of pathway hits normalized by the sum of pathway predictions. Fisher's test p values of " $0.00 \times 1000$ " indicate scores that are less than the smallest numerical value allowed in Python.

within the same laboratories. AstraZeneca data are therefore expected to provide better performance than alternative publicly available data sets assimilated for cytotoxicity analyses.<sup>24</sup>

In order to improve the links between chemical and biological spaces, we analyzed the results of a phenotypic screening cascade from an AstraZeneca diversity-based screening library evaluated on the THP-1 cell line, shown in Figure 1. AstraZeneca work has shown that a cell health assay using THP-1 can identify compounds that later go on to cause specific organ toxicity. Organ specific cell lines such as HepG2 do not predict specific organ toxicity,<sup>22</sup> so the THP-1 cell line is employed in-house as a first-pass assay to identify compounds with the potential to cause cytotoxicity.

The screening initiative presented here comprises a single concentration cytotoxicity assessment of a 388 000 compound collection in an AlamarBlue cell viability assay (screen A), assessing the intracellular reducing potential of living THP-1 cells. The 25 000 most potent hits from this primary screen were subsequently profiled through a follow-up dose–response AlamarBlue (Screen B). A total of 6844 hits from screen B were then assessed in the CellTox Green membrane integrity assay (Screen C) to identify cytotoxic compounds by measuring DNA staining in cells with compromised cell membranes.

Data sets of CytoToxic and CytoStatic (CTCS) compounds and cytostatic-only compounds were extracted from screening data via filtering and analyzed *via in silico* mechanism-of-action (MoA) analysis (Figure 2). Computational MoA protocols have previously been applied for the rationalization of toxicity libraries, where the extension of enriched targets with pathway information was found to underline the pathways implicated in cell death and improve the explanatory power of predictions.<sup>25,26</sup> Furthermore, we have in this work performed fragment analysis, enabling us to generate a library of structural alerts for further filtering of the undesirable liabilities of compounds. Finally, a CTCS classification model has been developed to support better understanding of the MoA of cytotoxic compounds, and to provide guidance into which compounds to add, remove, or flag in a HTS library. The MoA,

cytotoxic substructures and toxicity classification protocols will be employed in AstraZeneca for future HTS triage processes.

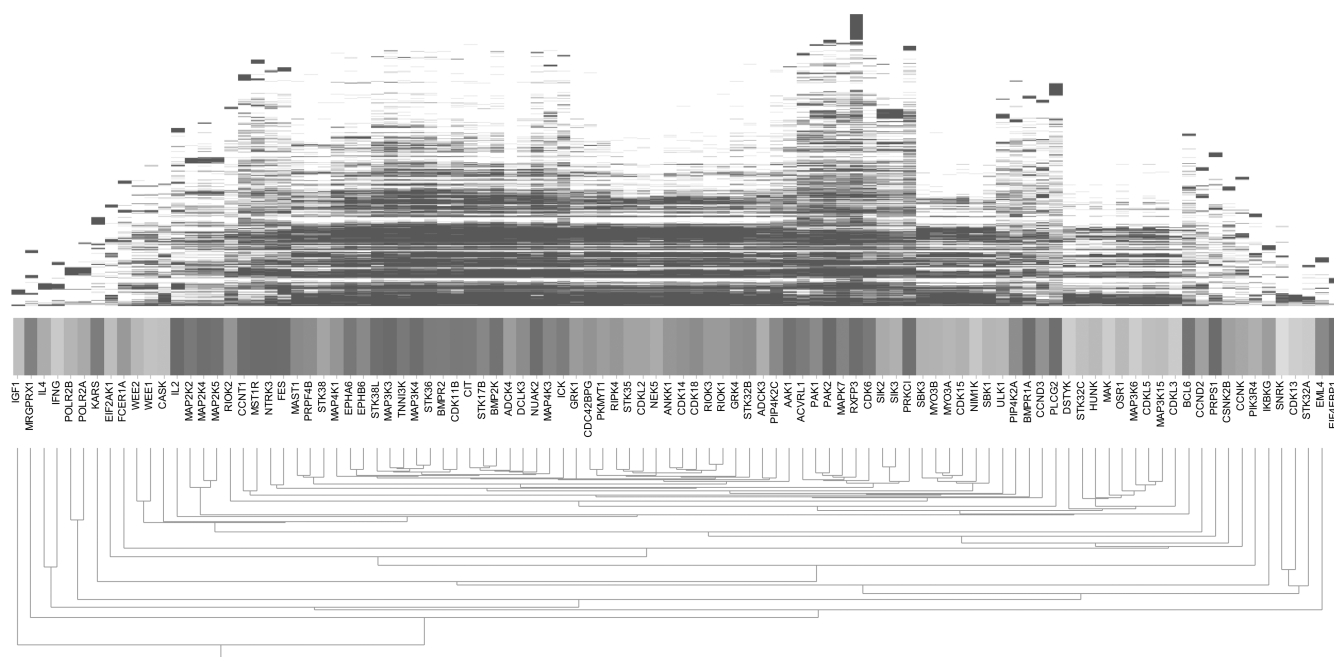
## RESULTS AND DISCUSSION

**Enriched Targets and Pathways for the Cytotoxic and Cytostatic Phenotypes.** The 5784 CTCS compounds identified from the THP-1 screening cascade were subjected to bioactivity prediction, annotated with pathways using WikiPathways,<sup>27</sup> and the subsequent calculation of enrichment metrics when compared to noncytotoxic compound predictions. The predictions from this analysis are shown in Tables 2 and 3, although these are not exhaustive since targets remain significantly enriched (*Prediction Ratio* less than 0.1 and Fisher's exact test p values below a significance level of 5%) outside the top 15 (a comprehensive list of enriched targets is available in the Supporting Information Table 1).

We observe that the most enriched targets and pathways have biological links to the phenotype at different absolute frequencies, a trend that has been previously observed when using *in silico* pathway enrichment, which highlights the need to normalize predictions using an enrichment metric.<sup>25</sup> For example, "Cyclin-Dependent Kinase 13 (CDK13)" (ranked second) comprises a cytotoxic hit rate of 1.43%, while "Dual Serine/Threonine and Tyrosine Protein Kinase (DSTYK)" (ranked third) has a prediction rate of 6.29%.

Table rankings highlight a mixture of apoptotic (8), necrotic (5), and autophagic (1) targets, where 14 of the top ranking targets can be attributed to cell-death via literature evidence. These results find little overlap compared with previous attempts to rationalize cytotoxicity, overlapping only for the cytotoxic effector Wee1.<sup>25,28</sup> This may be due to the dissimilar chemical space encompassed by public screening libraries and industry compound collections, or the number of targets and bioactivity data points available for modeling in every case.

Despite the differences for specific targets, the results find overlap to previous findings when considering target classifications, since the highest enriched targets are dominated by many kinases (13). The two nonkinase targets, "Interferon, Gamma (IFNG)" and "Insulin-Like Growth Factor 1 (Somatomedin C) (IGF1)," are expected to originate from



**Figure 3.** Euclidean biclustering of top 100 enriched target prediction profiles with accompanying *Prediction Ratio* scores. Compound hits are shown in the matrix; accompanying *Prediction Ratio* scores are shown above target labels, scaled from dark gray (increased enrichment) to light gray (decreased enrichment). UPGMA Euclidean distance was calculated for the targets (*x* axis) and compounds (*y* axis). Dendrogram structure highlights that similar protein families cluster together due to similarities in prediction profiles. Clusters also exhibit correlation with *Prediction Ratio* enrichment; hence, similar prediction profiles often produce similar enrichment profiles.

models comprising functional assay data. For example, bioactive training compounds for IFNG may be annotated active due to IFNG production measured in a functional assay, rather than actual biochemical affinity for the isolated protein. The real targets involved in these circumstances are likely to be proteins that influence either their production or associated signaling.

Examples of apoptotic targets among the top rankings include “SNF Related Kinase (SNRK)” and “Dual Serine/Threonine and Tyrosine Protein Kinase (DSTYK),” while examples of necrotic targets include “Calcium/Calmodulin-Dependent Serine Protein Kinase (CASK)” and “Male Germ Cell-Associated Kinase (MAK),” which is reviewed in [Table 1](#). The autophagic target identified in the table is “Serine/Threonine Kinase 32A (STK32A),” which is implicated in mitophagy (degradation of mitochondria by autophagy) in response to mitochondrial depolarization. “Serine/Threonine Kinase 32C (STK32C)” is a highlighted kinase related to STK32A, which does not appear to comprise known links to cytotoxicity. This potentially novel cytotoxicity-related target has unknown function but is highly expressed in the brain and has links to depression.<sup>29</sup> This finding illustrates the potential for *in silico* methods to highlight lesser studied targets.

The main groups of enriched pathways can be split into broad processes including cytokine activity (4), gene expression (4), cell cycle (2), and kinase activity (2), with links to lethality through modulation of these cellular processes.<sup>11</sup> Pathways range from generic processes such as “Mitotic Prophase (WP2654)” to ones with higher granularity, including “MyD88-independent TLR3 (WP2752).” Processes show overlap to the pathways reviewed in [Table 1](#), for example, the pathways “NoRC negatively regulates rRNA expression (WP3305)” and “mRNA Capping (WP1861)” highlight the importance of the gene regulation processes, whereas the links to “Signaling by BMP (WP2760),” “MyD88-independent

TLR3 (WP2752),” and “Interleukin-2 signaling (WP2732)” suggests the importance of cytokines. “Nonhomologous end joining (WP2768)” is the only DNA-repair process implicated in the results from this analysis. This pathway is essential for genomic integrity since it is the principle double-strand break repair pathway in mammalian cells, and perturbation of this pathway is central to cytotoxic action of various anticancer drugs.<sup>30</sup>

Although pathway analysis can provide a broader context for further interrogation of other markers associated with the cytotoxic phenotype, there is still a need to rationalize the associations driving links between targets and pathways and their relation to cell death. For example, the highest-ranking process, “Uptake and Function of Anthrax Toxins (WP3 390),” was highlighted due to links with the hydrolysis of mitogen-activated protein kinases through the action of the anthrax lethal factor endopeptidase,<sup>31</sup> while “DSCAM interactions (WP1808)” was highlighted due to the high enrichment of mitogen-activated protein kinases and PAK1 present in the pathway. Both of these pathways could be considered tangential signals, since the first is quite specific and the latter lacks a direct implication to cytotoxicity.

Aside from the potentially novel cytotoxicity-related target STK32C, MoA analysis fails to highlight many novel targets or pathways lacking unforeseen implications in cytotoxicity within the highest ranked positions, which, due to the amount of work that has been performed around cytotoxicity, is perhaps unsurprising. This indicates one should look further down the rankings when searching for more novel associations with unanticipated cytotoxic targets. For example, when consulting the entire list ([Supporting Information Table 1](#)), “EPH Receptor B6 (EPHB6; *Prediction Ratio* = 0.063, ranked 64)” and “NLR Family, Pyrin Domain Containing 3 (NLRP3; *Prediction Ratio* = 0.106, ranked 113)” represent two examples

of enriched targets within the top ranking 10% that lack literature links to cytotoxicity, which would be suitable candidates for future experimental studies.

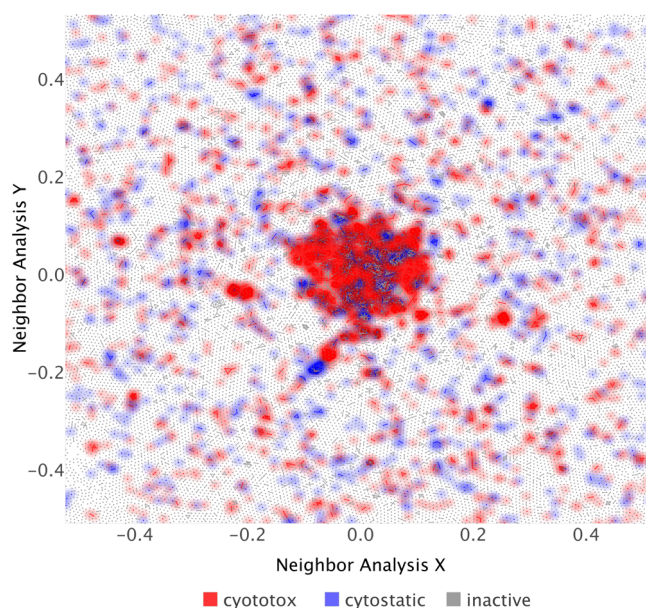
We next performed hierarchical biclustering of the enriched CTCS bioactivity profiles to analyze the polypharmacological trends of predictions (Figure 3). Clustering shows target families, such as the “MAP-Kinases” (“Mitogen-Activated Protein Kinase Kinase 2 (MAP2K2),” “Mitogen-Activated Protein Kinase Kinase 4 (MAP2K4)” and “Mitogen-Activated Protein Kinase Kinase 5 (MAP2K5)”), are frequently positioned together in bioactivity space due to similarity of predictions between these targets. The Rio kinases represent one enriched target family which has not clustered together, with “Rio Kinase 2 (RIOK2)” separating into a distinct branch from the highly associated “Rio Kinase 1 (RIOK1)” and “Rio Kinase 3 (RIOK3)” subcluster. The dispersed topology for such family classes may result from biases in bioactivity space, or selectivity due to protein structure diversity within families.

Clustering also shows correlation with areas of enrichment (*Prediction Ratio* indicated in the row above the dendrogram), where highly enriched clusters such as “Relaxin/Insulin-Like Family Peptide Receptor 3 (RXFP3),” “Mitogen-Activated Protein Kinase 7 (MAPK7),” “p21 Protein (Cdc42/Rac)-Activated Kinase 2 (PAK2),” and “p21 Protein (Cdc42/Rac)-Activated Kinase 1 (PAK1)” can be contrasted with comparatively lower enriched clusters, for example, “SH3 Domain Binding Kinase Family, Member 3 (SBK3),” “Cyclin-Dependent Kinase 15 (CDK15),” “Myosin IIIB (MYO3B),” and “Myosin IIIA (MYO3A)”. Thus, we frequently observe that similar prediction profiles can often comprise similar *Prediction Ratio* scores.

The potentially novel cytotoxicity-related target STK32C is positioned within a subcluster of targets comprising “Hormonally Up-Regulated Neu-Associated Kinase (HUNK),” “Male Germ Cell-Associated Kinase (MAK),” “Protein odd-skipped-related 1 (OSR1),” “Mitogen-Activated Protein Kinase Kinase 6 (MAP3K6),” “Cyclin-dependent kinase-like 5 (CDKL5),” and “Mitogen-activated Protein Kinase Kinase 15 (MAP3K15).” The STK32C node is the most distinct within this highly enriched cluster and illustrates the somewhat unique bioactivity profile generated for this target.

We next visualized the distribution of the cytotoxic, cytostatic, and noncytotoxic compounds in chemical space, *via* the 2D-RBS plot as shown in Figure 4. Visualization shows that the areas' chemical space occupied by compound sets is complex, with small islands of cytotoxic and cytostatic compounds clustering together toward the center of the plot. This 2D representation indicates that the cytotoxic and cytostatic compounds are more frequently similar to each other than to inactive compounds but suggests that areas of cytotoxicity and cytostaticity can be separated to some extent.

In order to better discern the cytostatic properties of compounds, we next conducted additional MoA analysis using a library of cytostatic compounds. These compounds were subjected to target and pathway enrichment versus cytotoxic compounds, with the enriched target and pathway results shown in Tables 4 and 5. Fewer numbers of targets and pathways are shown in this analysis, since smaller numbers comprise low *Prediction Ratio* and Fisher's exact test p-values, which can therefore be considered significant. The reason for the comparatively low enrichment is due to high similarity in chemistry between the two sets, and therefore a high degree of overlap between static and toxic bioactivity profiles. This



**Figure 4.** Visualization of cytotoxic, cytostatic, and non-toxic compounds. Visualization shows that the relationship between the areas' chemical space occupied by compound sets is complex, with small cytotoxic and cytostatic islands of compounds clustering together within the distribution of nontoxic compounds. Thus, cytotoxic and cytostatic are more frequently similar to each other than to inactive compounds. Molecular descriptor used is SkelSpheres, and nearest-neighbor threshold is set to 0.9.

overlap may not be surprising, since it is the entry points into these shared processes that are likely to dictate the fate of a cell.<sup>16</sup>

Enrichment highlights a mixture of enzymes (2), protease (1), lyase (1), ion channel (1), oxidoreductase (1), and imidazoline receptor (1) targets, which can be attributed to cytostasis *via* literature evidence. In comparison to the cytotoxic enrichment results, enriched cytostatic targets do not comprise kinases, which is likely due to the high degree of overlap for predictions, or frequent lethality of this target class. “Bone Morphogenetic Protein 1 (BMP1)” is identified as the highest ranking cytostatic target, which has previously been identified as responsible for mediating a cytostatic response to rapamycin.<sup>32</sup> “Tyrosinase (TYR)” is the second ranking target highlighted for the analysis, where it has been demonstrated that functional mutations in either tyrosinase or tyrosinase-related protein 1 (TYRP1) are less sensitive to the cytostatic effects of deoxyArbutin and its derivatives.<sup>33</sup> Another cytostatic target highlighted in the list is “Nischarin (NISCH),” known to be implicated with cell migration and cytoskeleton reorganization by binding to alpha-5-beta-1 integrin,<sup>34</sup> which can be linked to mitotic arrest and stasis through the interference of microtubule dynamics.<sup>16</sup>

In comparison to the CTCS pathways, the results from cytostatic pathway enrichment, shown in Table 5, illustrate that *Prediction Ratio* enrichment does not necessarily correlate with Fisher's test p values. For example, “Alanine and aspartate metabolism (WP106)” comprises a *Prediction Ratio* of 0.34 and a Fisher's p value of  $1.99 \times 10^{-01}$ . Hence, this pathway is considered moderately significant when filtering using *Prediction Ratio* but is not significant when filtering for a p value at a 5% confidence level and, hence, is the only pathway that is not significant using Fisher's test p value in this table.

Table 4. Top Enriched Targets Comparing Cytostatic Compounds versus Cytotoxic Compounds<sup>a</sup>

EGID	name	classification	cytostatic hit rate %	cytotoxic hit rate %	prediction ratio	Fisher's exact test p value	implication in cytostasis	ref
BMP1	bone morphogenetic protein 1	protease	1.07	0.32	0.30	$5.06 \times 10^{-04}$	activation of BMP signaling in prostate carcinoma cells is cytostatic	32
TYR	tyrosinase	oxidoreductase	1.25	0.43	0.35	$9.54 \times 10^{-04}$	inhibition of tyrosinase activity reduces melanocyte cell number due to inhibition of proliferation rather than initiation of apoptosis	33
NISCH	nischarin	imidazole-1 receptor	0.99	0.35	0.35	$2.91 \times 10^{-03}$	overexpression of this protein profoundly affects cell migration and cytoskeleton reorganization	34
NQO1	NAD(P)H dehydrogenase, quinone 1	enzyme	1.12	0.49	0.44	$7.99 \times 10^{-03}$	selective induction of NQO1 has potent antiproliferative activity in two human cancer cell lines	64
NOD1	nucleotide-binding oligomerization domain containing 1	enzyme	0.95	0.46	0.49	$3.10 \times 10^{-02}$	NOD1 expression is significantly correlated with tumor differentiation	65
CAL3	carbonic anhydrase XIII	lyase	0.99	0.49	0.50	$3.45 \times 10^{-02}$	cytostatic agent (selenocystine) is found to up-regulate CAL3 in keloid fibroblasts using cDNA microarray analysis	66
ATP1A1	ATPase, Na <sup>+</sup> /K <sup>+</sup> transporting, alpha 1 polypeptide	ion channel	0.99	0.49	0.50	$3.45 \times 10^{-02}$	inhibition found to enhance the antiproliferative effect of hyperosmotic conditions through increased G2/M block	67

<sup>a</sup>The top enriched targets have links to the cytostatic phenotype via literature. Comparatively fewer targets are highlighted in this analysis due to lower enrichment as exemplified by higher Prediction Ratio and Fisher's exact test p values. Ranked targets do not contain any kinases in comparison to the cytotoxic results. Results are filtered for a Prediction Ratio less than or equal to 0.5.

The top ranking pathways have links to the cytostatic phenotype. "Collagen biosynthesis and modifying enzymes (WP2725)" are highlighted due to the frequent prediction of cytoskeletal enzymes (e.g., Nischarin) in the analysis, which are essential for mitotic progression during cellular division. "Melanin biosynthesis (WP3377)" has also been highlighted in this analysis due to the high target enrichment of tyrosinase and its importance in melanization.<sup>33</sup> Despite an insignificant Fisher's test p value, "Alanine and aspartate metabolism (WP106)" may have links to cytotoxicity via endoplasmic reticulum (ER) stress. Interference of this metabolic process is known to stimulate ER stress-induced cytostasis after failed attempts to rectify unfolded proteins in the ER,<sup>35</sup> although this link should be extrapolated with caution.

Overall, while proteins and pathways involved in cytotoxicity versus cytostaticity seem to be overlapping, the differences between both types of biological processes indeed seem to exist, and the chemical space modulating both functions is also to some extent distinct, according to the analysis presented here.

**Compound Fragments Enriched for CTCS Compounds.** We next conducted fragment analysis of the CTCS (actives from screen B and screen C) and noncytotoxic compounds from screen A to identify substructures associated with the cytotoxic end point. Substructures were filtered for significant cytotoxicity count (binomial p values greater than  $1.0 \times 10^{-10}$ ) and 10-fold enrichment in the cytotoxic set (*Fragment Ratio* less than 0.1), producing 978 fragments. Target predictions were then generated for the fragment cores, with predictions mapped to their respective target classification class using the ChEMBL classification system. This step enables an approximation for the fraction of fragments that may be attributed to potential kinase-dependent versus kinase-independent mechanisms of cytotoxicity.

The majority of enriched frameworks (612 of the 978 fragments) have target prediction hits comprising 10 or more kinases, which indicates a targeting of kinases in about two-thirds of the fragments involved with cytotoxicity. This high frequency of kinase inhibitor-like fragments likely results from the ubiquitous role kinases play in modulation of signaling pathways and further reflects the kinase target focus of the compound library through the decades. Ten representative substructures with a kinase prediction rate greater than 36 are shown in Table 6. Whether or not to filter screening libraries for these types of substructures would depend on the purpose of the library, as well as the target classes one intends to develop ligands for.

A total of 366 fragments comprise an average kinase prediction rate less than 10, where toxicity may be elicited by nonkinase mediated events, and hence may be cause for concern since this may be instigated *via* unknown mechanisms. Table 7 features 10 representative substructures from this list, and highlights their comparatively lower cytotoxic count when compared to their kinase-like counterparts. Indeed, this analysis fails to identify significant quantities of cytotoxic frameworks that are not attributed to kinase inhibition. This finding is most likely due to the careful selection and previous application of published in-house filters placed on screening collections to remove compounds with undesirable toxicity profiles.<sup>36</sup> The top ranking fragments made available in this study are useful as off-target flags that can be employed to filter for predicted kinase-(in)dependent mechanisms of cytotoxicity.



Table 5. Top Enriched Pathways Comparing Cytostatic Compounds versus Cytotoxic Compounds<sup>a</sup>

WikiPathways ID	name	biological link	cytostatic hit rate (%)	prediction ratio	Fisher exact test p-value
WP2725	collagen biosynthesis and modifying enzymes	cytoskeleton	0.17	0.17	$2.58 \times 10^{-02}$
WP106	alanine and aspartate metabolism	metabolism	0.09	0.34	$1.99 \times 10^{-01}$
WP3377	melanin biosynthesis	cell cycle	1.25	0.35	$3.11 \times 10^{-07}$
WP1872	neurotransmitter uptake and metabolism in glial cells	metabolism	0.26	0.45	$3.78 \times 10^{-02}$
WP1804	DNA damage reversal	DNA damage	0.69	0.46	$9.64 \times 10^{-04}$

<sup>a</sup>The top enriched pathways have biological links to the cytostatic phenotype. Fewer pathways are shown compared to the CTCS vs. non-toxic analysis, due to the lower enrichment values when comparing cytostatic vs. cytotoxic compounds. Enriched pathways are filtered for a Prediction Ratio less than or equal to 0.5. Results show that a low Prediction Ratio does not necessarily correspond with low Fisher's exact test values. For example, a Fisher's test does not find statistical significance for "alanine and aspartate metabolism (WP106)" at a significance level of 5%, although this pathway has a moderately enriched Prediction Ratio of 0.34.

**Cytotoxicity End Point Prediction.** We next trained a random forest (RF) model using ECFP<sub>4</sub> fingerprints on both the public and proprietary parts of the CTCS data set, in order to generate a predictive cytotoxicity model for both the entire existing compound collection, as well as future library enhancement initiatives. These models were validated via 5-fold cross-validation (performance details included in [Supporting Information Table 2](#)) and prospectively validated using 988 novel compounds selected at different cytotoxicity probability intervals and similarity to the training set to prospectively explore the AD *in vitro*.

The 988 compounds selected for cell-based cytotoxicity assessment via the AlamarBlue cell viability assay are visualized in the 2D-RBS plot in [Figure 5](#). Visualization shows testing comprises both compounds predicted as cytotoxic with toxic near-neighbors and compounds predicted to be nontoxic with nontoxic near-neighbors. Testing also includes circumstances when nearest neighbors conflict with the prediction of cytotoxicity, for example a cytotoxic near-neighbor predicted to be nontoxic, or a toxic near-neighbor compound predicted to be nontoxic. 2D-RBS shows highly connected clusters of compounds toward in the center of the plot, comprising small molecules with both toxic nearest neighbors and nontoxic nearest neighbors intermingled throughout chemical space. Relatively few compounds have a toxic nearest neighbor and low RF cytotoxicity probabilities.

Overall, it was found that the proprietary model performs with an averaged precision and recall of 77.9% and 83.8%, while the public version performs with an overall precision and recall of 77.0% and 83.6% (confusion matrix shown in [Table 8](#) and binned precision-recall curves shown in [Supporting Information Figure 1](#)). The confusion matrix ([Table 8](#)) demonstrates the tendency for models to overpredict compounds as toxic (increased recall at the cost of precision). This is due to the specification of sample weight parameters in Scikit-learn,<sup>37</sup> which correct for imbalanced training data *via* penalizing the misclassification of the CTCS minority class. This model behavior is a desirable trait when filtering for cytotoxicity, since short-listed false-negative compounds that may show toxicity further into project timelines are potentially more detrimental than the excessive filtering of compound libraries early in discovery stages.

The AD of the proprietary model was investigated in [Figure 6a](#) for compounds confirmed to be toxic (top-left) and nontoxic (top-right). The binned boxplot of train-to-test Tanimoto coefficient (Tc) similarity (*x*-axis) and RF probability distributions (*y*-axis) exhibits improvement with increasing similarity between a Tc of 0.3 and 0.8, correlating with an overall improvement of median cytotoxicity proba-

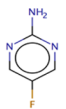
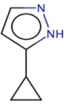
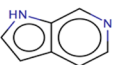
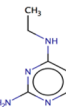
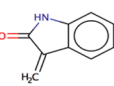
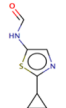
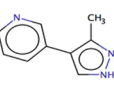
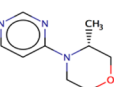
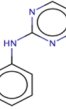
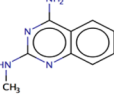
bilities (although there is a drop for the final bin). Inactive compounds show a similar trend with increasing similarity, where cytotoxicity predictions exhibit overall improvement between 0.3 and 1.0 Tc bins, decreasing from 0.22 to 0.02, respectively.

There is evidence of cytotoxicity activity-cliff results given the distribution of predictions. One example is the unexpected drop in the distribution for cytotoxicity probabilities and increasing Tc similarity between the 0.8 and 0.9 similarity bins from 0.865 to 0.345, respectively. To further explore this trend, a scatter plot of the underlying predictions is shown in [Figure 6b](#), separated by shape based on cytotoxic near-neighbor activity in screen B ("x" marker) and nontoxic near-neighbor activity ("o" marker) in screen B. Upon further analysis of the 0.9 Tc similarity, we observe that the two false-negative predictions, assigned cytotoxicity probability scores of 0.15 and 0.17, share nontoxic near-neighbor activity and hence are located in areas of chemical space considered activity cliffs. Conversely, activity-cliff behavior that is correctly predicted by the model can be observed within the 0.8 similarity bin, where the true-positive prediction of a compound with a nontoxic near-neighbor is correctly assigned a moderately high cytotoxic probability of 0.71.

**Conclusion.** MoA analysis of the 5784 CTCS and cytostatic compounds was in this work able to highlight the targets and pathways implicated in cytotoxicity and cytostaticity for a large high-throughput screening collection. The bioactivity profiles generated here can be used as a guideline for off-target activity when considering which compounds to select for screening collections. Review into the known MoAs for cytotoxicity highlighted many kinases, including the potentially novel cytotoxicity-related target STK32C, which illustrates the potential for the MoA protocol to highlight even lesser studied targets. Pathway analysis has highlighted processes implicated in cytokine activity, gene expression, and cell cycle progression. Analysis of the 2327 cytostatic-only compounds has furthermore enabled us to discern the cytostatic properties from the cytotoxic phenotype. Results from this analysis highlighted fewer targets that can be statistically associated with the cytostatic effects of compounds due to proximity to (and often overlap with) cytotoxic effects; however, some differences could still be identified. In particular, enriched pathways for the cytostatic compounds show links to the phenotype *via* DNA damage reversal, metabolism, and cytoskeletal machinery.

Although the pathway enrichment approach presented in this study gives an indication for the types of biology that targets are implicated in, care should be taken when extrapolating enrichments. First, the choice of enrichment metric has been shown here to influence the signal of enriched targets, since a

Table 6. The Top 10 Enriched Cytotoxic Fragments with High Kinase Prediction Rates<sup>a</sup>

Fragment	CTCS Count	CTCS %	Non-Toxic Count	Non-Toxic %	Frag. Ratio	Binomial P-Value	Predictions for Fragment				
							1st Predicted Class	2nd Predicted Class	3rd Predicted Class	Kinase	Non-Kinase
	94	1.63	146	0.05	0.03	3.86E-95	Kinase : 49	Other : 32	Oxidoreductases : 21	49	147
	74	1.28	301	0.10	0.08	6.26E-51	Kinase : 63	Ion_Channel : 31	Other : 21	63	158
	55	0.95	172	0.06	0.06	2.14E-43	Kinase : 130	Other : 38	Oxidoreductases : 25	130	253
	230	3.98	1129	0.38	0.10	1.00E-101	Kinase : 74	Other : 38	GPCR : 14	74	161
	38	0.66	134	0.04	0.07	6.69E-29	Kinase : 38	Other : 24	Ion_Channel : 19	38	137
	11	0.19	1	0.00	0.00	1.41E-18	Kinase : 54	Other : 23	Ion_Channel : 13	54	117
	9	0.16	29	0.01	0.06	3.08E-08	Kinase : 48	Other : 39	Ion_Channel : 26	48	158
	21	0.36	14	0.00	0.01	1.33E-27	Kinase : 62	Ion_Channel : 13	Other : 12	62	116
	498	8.61	883	0.30	0.03	1.00E-101	Kinase : 177	Other : 67	Transporter : 22	177	355
	24	0.41	67	0.02	0.05	8.56E-21	Other : 52	Kinase : 36	Oxidoreductases : 23	36	200

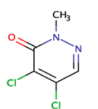
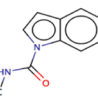
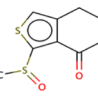
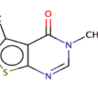
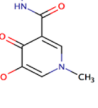
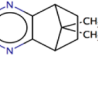
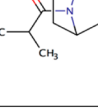
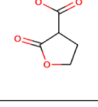
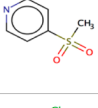
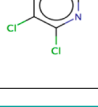
<sup>a</sup>Fragments highlighted in this table would not necessarily require removal from a screening library since they have been historically selected to identify a therapeutic window between efficacy and toxicity. Results show the need to normalize fragments to the number of fragments generated by the toxic or non-toxic data sets. Toxic % and Non-Toxic % were calculated via Toxic Count and Non-toxic Count of fragments and divided by the total number of fragments for that set (5784 CTCS fragments and 296 970 non-toxic fragments).

low *Prediction Ratio* does not necessarily correlate with a low Fisher's test *p* value. Second, there are tangential biological annotations within the top ranking pathways which are indirectly related to the phenotype, and this area of ambiguity is where other approaches could prioritize additional steps to follow up on in the future.

One current limitation of the MoA protocol presented here is that the statistical analysis of the target enrichments assumes that target enrichments are equally meaningful across the spread of models, which is not true in reality. For example, there are some targets that are neglected in our analysis due to little (or a complete lack of) training data. Additionally, the

coverage of chemical space, as well as the number of data points, is different between the target models. This leads to virtually every model behaving differently with respect to sensitivity and specificity. Additionally, the input chemical space has a certain distribution, in relation to the training set, which influences the prediction of each target. An additional limitation is the lack of information regarding the underlying modulation of targets (for example, activating or inactivating mechanism upon binding of compounds, which is lacking from the localized MoA analysis). We aim to pursue this shortcoming *via* the prediction of antagonism or agonism activity of compounds using annotation of functional assays. Future work will

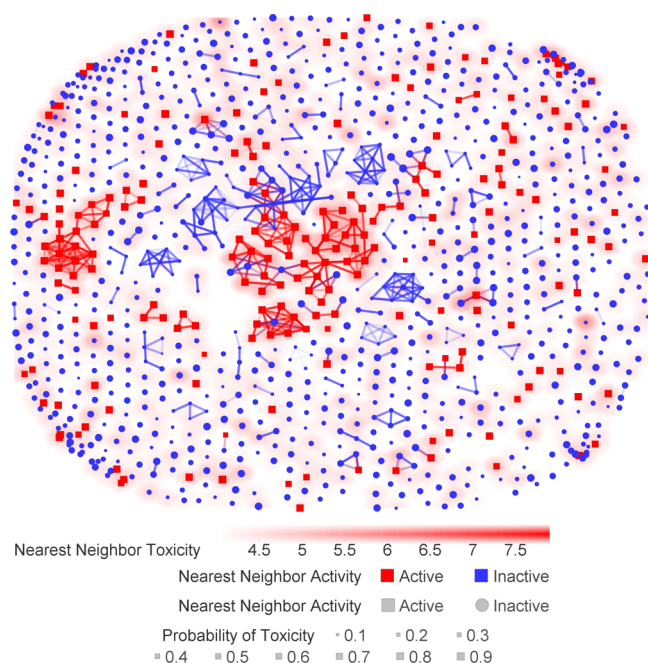
Table 7. Top 10 Enriched Cytotoxic Fragments with Low Kinase Prediction Rates

Fragment	CTCS Count	CTCS %	Non-Toxic Count	Non-Toxic %	Frag. Ratio	Binomial P-Value	Predictions for Fragment				
							1st Predicted Class	2nd Predicted Class	3rd Predicted Class	Kinase	Non-Kinase
	16	0.28	1	0.00	0.00	5.00E-27	Other : 15	Ion_Channel : 8	Kinase : 4	4	41
	7	0.12	1	0.00	0.00	7.14E-12	Other : 5	Oxidoreductases : 4	Ion_Channel : 4	1	18
	6	0.10	1	0.00	0.00	3.28E-10	Ion_Channel : 8	Other : 7	Protease : 6	5	46
	10	0.17	2	0.00	0.00	3.99E-16	Ion_Channel : 13	Other : 11	Oxidoreductases : 5	0	42
	10	0.17	3	0.00	0.01	1.70E-15	Lyases : 10	Other : 7	Oxidoreductases : 7	2	36
	6	0.10	3	0.00	0.01	3.79E-09	Other : 17	Oxidoreductases : 10	Ion_Channel : 10	5	62
	13	0.22	7	0.00	0.01	2.94E-18	Ion_Channel : 6	GPCR : 5	Protease : 3	2	27
	5	0.09	5	0.00	0.02	5.74E-07	Kinase : 2	Other : 2	Isomerases : 1	2	9
	47	0.81	0	0.00	0.00	1.42E-81	Oxidoreductases : 13	Other : 10	Kinase : 5	5	41
	5	0.09	0	0.00	0.00	2.51E-09	Other : 29	Ion_Channel : 22	Oxidoreductases : 10	4	99

concentrate on the prediction of compound MoA to better understand underlying modulation of targets and pathways in relation to complex end points like cytotoxicity.

Fragment analysis has furthered the understanding of the chemical space linked to cytotoxicity induced by small molecules. We have shown that cytotoxic compounds in screening collections often (in the data analyzed here in about two-thirds of the cases) involve kinase inhibitors, which is corroborated by the kinase-dominated CTCS target prediction

profiles. The choice of whether to filter compounds containing these fragments would depend on the use of the library. They have been retained within the compound collection at AstraZeneca since there is a possibility to identify a therapeutic window between efficacy and toxicity. There are low numbers of fragments with predicted kinase-independent activity, which may be cause for concern when selecting which compounds to add to a screening library, since these frameworks may elicit toxicity through unidentified mechanisms. The relatively low



**Figure 5.** Similarity map of compounds for testing. Selected molecules span areas with different toxicity probabilities (from 0.1 to 1.0), inhabit varied chemical space often separated by islands, have diverse near-neighbor toxicity (pIC50; ranging from activity values of 4.5 to 8.1 in screen B), and comprise a mixture of compounds with near-neighbors that are predicted as both toxic and nontoxic compounds (indicated *via* marker shape and color). Molecular descriptor used is SkelSpheres, and near-neighbor threshold is set to 0.90.

**Table 8. Confusion Matrix of Cytotoxicity Predictions<sup>a</sup>**

		proprietary model predicted		public model predicted	
		cytotoxic	nontoxic	cytotoxic	nontoxic
actual class	cytotoxic	<b>264</b>	181	<b>253</b>	192
	nontoxic	19	<b>524</b>	16	<b>527</b>

<sup>a</sup>Proprietary models exhibit superior prediction trends compared to public models, with higher numbers of correct classifications (diagonal boxes in bold). Proprietary models predict with an overall averaged precision and averaged recall score of 77.9% and 83.8% (more specifically, the cytotoxic class of compounds performs with a precision and recall of 59.3% and 93.3%, respectively, whilst the non-toxic class is predicted with a precision and recall of 96.5% and 74.3%). Public models have comparatively higher numbers of false-negatives (toxic compounds predicted as non-toxic, top right) due to the restricted coverage of training data.

frequency of such fragments may result from the structural filters previously employed to select compounds with more desirable chemistry on an empirical basis.

A proprietary random forest (RF) cytotoxicity classification protocol was also developed using the CTCS and nontoxic compounds, which was able to perform with a precision and recall score of 77.9% and 83.8% on 988 untested compounds. It was found that models have a tendency to overpredict cytotoxicity, which is in practice preferred when compared to the potential cost in attrition (higher false negatives) for the under-prediction of toxicity. Predictions give indication for the applicability domain of the model and occurrences of activity cliffs.

While generation of further screening data will aid in the refinement of these predictive models, it also remains to be seen if these data set can be used to flag on-target liability with novel targets as they progress through the early drug discovery portfolio. Additionally, future improvements in the use of more complex cell models may allow generation of data sets that are better representative of a complex human system, and we believe the data presented, along with *in silico* prediction tools, will provide useful annotation to the drug discovery community.

Overall, a decision whether to add a compound to a screening library can be based on the target prediction profiles, highlighted pathways, predictions for the likelihood of cytotoxicity, or presence of substructure flags. Models will aid library design for future large collection screening, by affording better selection of which compounds to insert or remove from compound collections, and as a method to shortlist candidates in further studies. The top ranking targets, pathways, and fragments are available for download and can be deployed as bioactivity and structural flags during the selection of new compounds.

## EXPERIMENTAL SECTION

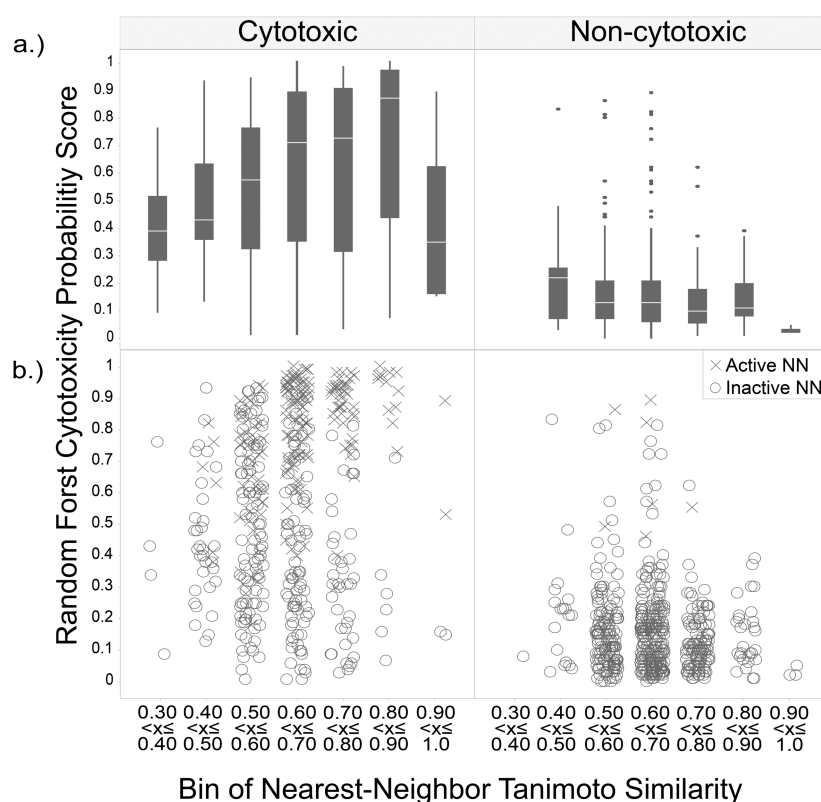
A depiction of the entire screening cascade of the compound collection is outlined in Figure 1.

### AlamarBlue Cytotoxicity Assay Protocol (Screen A and B).

Primary assay (A) test compounds were prepared in 100% v/v dimethyl sulfoxide (DMSO) at 10 mM and screened at a final assay concentration of 50  $\mu$ M. Assay ready compound plates (ARPs) were generated *via* acoustic dispensing using a Labcyte Echo 555 instrument. For the primary screen (A), 20 nL of compound was transferred to each well of a 1536-well black plate (#781076, Greiner). Ranges of volumes were dispensed for the concentration–response screen (B), creating 10-point concentration–response curves with a final compound concentration range between 100  $\mu$ M and 5 nM in 384-well assay ready plates. Wells were backfilled with the appropriate volume of DMSO to ensure a final screening concentration of 1% (v/v) and a total volume of 400 nL. Maximum (0% inhibition of assay response) and minimum (100% inhibition of assay response) compounds controls were added to define relative activity and the reproducibility of the data generated. The maximum control signal was determined using 1% DMSO, with minimum signal controls defined with 50  $\mu$ M Puromycin for both screening stages.

The human monocytic THP-1 cell line sourced from American Type Culture Collection (ATCC) was routinely cultured in suspension in 1700 cm<sup>2</sup> roller bottles (#UY-0183-05, Corning). Growth media consisted of RPMI-1640 (#R7509, sigma), supplemented with 10% fetal calf serum (FCS; #A15-011, PAA) and 200 mM L-glutamine (#35050038, Invitrogen). Cell cultures were maintained at 37 °C in a 95% humidified atmosphere of 5% (v/v) CO<sub>2</sub>/95%(v/v) air on a rotating rack set to 140 rpm. Cells were passaged every 2–3 days depending on cell density.

Cells were collected by centrifugation at 300g and washed in PBS on the day of assay. The cell pellet was resuspended in growth media with the addition of 100 units/mL of penicillin streptomycin (#P4333, Sigma). The number of viable cells was counted using the Beckman ViCell coulter counter to determine cell number per milliliter. Cells were suspended at the required density of 250 000 cells/mL, and 4  $\mu$ L was transferred into the assay ready 1536-well black plates (#781076, Greiner) or 40  $\mu$ L for the 384 assay plates and incubated at 37 °C with 5% (v/v) CO<sub>2</sub> for 48 h at 95% humidity. CellTiter-Blue Viability reagent (#G3580, Promega) was diluted 1:6 in RPMI 1640 growth media excluding penicillin streptomycin and 2  $\mu$ L/well added to all wells of the 1536 plates and 8  $\mu$ L/well added to all wells of the 384 plates using a Multidrop Combi. The assay plate was incubated for a further 2 h under the same conditions as the initial 48-h compound incubation. Plates were then centrifuged at 300g prior to fluorescence



**Figure 6.** Applicability domain analysis of the proprietary cytotoxicity model. (a) Distribution of RF cytotoxicity probabilities for tested compounds. Predictions for cytotoxic compounds are shown on the top-left; predictions for nontoxic predictions are shown on the top-right. Distributions of Tanimoto similarities ( $T_c$ ) are split into seven bins, between  $T_c$  values of 0.3 and 1.0, at intervals of 0.1. The median predictions for cytotoxic prediction increase with increasing similarity between the 0.3 and 0.9 bins, correlating increasing scores from 0.455 to 0.685. Nontoxic compounds show a decrease in median cytotoxicity predictions showing overall improvement between 0.3 and 0.9  $T_c$  bins, decreasing from 0.06 to 0.03. (b) Underlying RF Predictions for the boxplot colored by near-neighbor activity and the identification of activity-cliff behavior. Individual compound predictions are shown below, where the near-neighbor compound activity in the primary screen is indicated *via* marker shape (cytotoxic and nontoxic near-neighbor activity indicated by “x” and “o,” respectively). This highlights instances of activity-cliff behavior and situations when the model correctly and incorrectly classifies compounds. Many false-negative compounds below have nearest-neighbors to the nontoxic data set and are considered cytotoxic activity cliffs.

reading on a fluorescence plate reader at  $\lambda_{\text{ex}} = 540 \text{ nm} \pm 20$  and  $\lambda_{\text{em}} = 590 \text{ nm} \pm 20$ .

**CellTox Green Express Cytotoxicity Assay Protocol (Screen C).** The 384-well assay ready plates were prepared with 200 nL of test compound/DMSO to give a range of concentrations between 100  $\mu\text{M}$  and 5 nM. A total of 20  $\mu\text{L}$  of cell suspension at 625 000 cell/mL was added to each well of the lidded assay plates and incubated for 15 min at RT to equilibrate the plate temperature before transferring to 37  $^\circ\text{C}$ , in a 5% v/v  $\text{CO}_2$ /95% v/v air humidified incubator for 48 h. After 48 h, a 2 $\times$  diluted stock of CellTox Green reagent was prepared in assay dilution buffer, and 20  $\mu\text{L}$  of CellTox Green reagent was dispensed into each well using a Multidrop Combi with a standard volume Multidrop cassette on fast speed. Assay plates were shaken for 15 min on an orbital shaker shielded from light for optimal staining of cells. Assay plate(s) were read on a fluorescence plate reader at  $\lambda_{\text{ex}} = 485 \text{ nm} \pm 20 \text{ nm}$  and  $\lambda_{\text{em}} = 535 \text{ nm} \pm 20 \text{ nm}$ .

## METHODS

**Mining AstraZeneca Collections for Active Bioactivity Training Data.** Bioactivity data sources available at AstraZeneca,<sup>38</sup> comprising both in-house data and public repositories such as ChEMBL, were mined for activity values ( $\text{IC}_{50}/\text{EC}_{50}/K_i/K_d$ ) less than or equal to 10  $\mu\text{M}$  from “binding” or “functional” human protein assays. The 10  $\mu\text{M}$  cutoff for activity specified here is in accordance with previously validated target elucidation methods,<sup>39,40</sup> and assigns both marginally and highly active compounds to targets. HomoloGene<sup>41</sup> was used for the extrapolation of bioactivity data to nonhuman orthologous targets to improve the coverage of chemical

space for 647 models.<sup>42</sup> Compounds were subjected to preprocessing and filtered for targets with more than 10 activities to ensure proteins encompassing sufficient chemical space are retained for training. The resulting data set includes 3 381 388 distinct compounds for 9 565 534 bioactivities spanning 2882 targets. The targets modeled comprise a variety of target classifications; the top three include 488 kinases, 281 transporters, and 233 GPCRs (full list of target classifications and number modeled shown in Supporting Information Table 3). Some degree of care should be employed when extrapolating compound–target associations, since the annotation of bioactivity data may reflect a functional assay versus actual binding affinity. For example, an interleukin annotation may actually reflect assay metadata measuring interleukin production, rather than affinity measurement of the isolated protein.

**Mining Negative Bioactivity Training Data. AstraZeneca Collections.** The HTS bioactivity data from 420 AstraZeneca target-based screens, spanning 400 different targets, was employed as a resource of inactive bioactivity data. Inactive data have coverage for a wide variety of targets, including 88 different GPCRs, 77 kinases, and 31 proteases (full table of target classifications covered by HTS screens employed for inactive data shown in Supporting Information Table 4). These screens were mined for activity values ( $K_i/K_d$ ) greater than 10  $\mu\text{M}$ . In cases where a compound has been measured more than once for a target, it was defined as inactive if it was measured at least twice as many times as inactive versus as active. The resulting compound–target pairs were preprocessed, resulting in a data set of 189 965 064 inactive data points, comprising 2 827 651 distinct compounds for the 400 targets. A breakdown of the numbers of inactive data points added

for each target classification from AZ collections is shown in Supporting Information Table 5.

**PubChem BioAssay.** The NCBI BioAssay<sup>41</sup> database was mined for additional experimentally confirmed inactive data points in a similar procedure to that of Mervin *et al.*,<sup>39</sup> via the EUtils and PubChem PUG REST APIs. This process involved “ESearch” and “ELink” EUtils procedures to obtain a comprehensive list of all Entrez Gene IDs (GIDs) and Protein IDs (PIDs) associated with a given GID. These GIDs and PIDs were used to “ELink” to binding and functional assays held in the NCBI BioAssay database. An additional “ELink” step was used to link from these assays to Compound IDs (CIDs) with a compound–target “activity\_outcome” annotation that has been manually declared as “inactive” upon upload of the screen. Finally, inactive CIDs were mapped to SMILES using the PubChem Power User Gateway (PUG) REST service. CIDs were subjected to preprocessing, producing 419 121 152 inactive data points for 768 014 distinct compounds spanning 2,116 targets.

The AstraZeneca and PubChem inactive data sets were combined, yielding 598 923 798 inactive data points spanning 2161 targets. The active set of target–compound pairs was retained when conflicting inactive bioactivities arose, since these data are calculated from dose–response curves.

**Sphere Exclusion and Undersampling of Negative Bioactivity Data.** A sphere exclusion algorithm was applied to 1500 targets with insufficient numbers of inactive data points, for both public and proprietary data. In this procedure, compounds were randomly sampled from PubChem with a Tanimoto coefficient ( $T_c$ ) similarity to actives lower than 0.4, ensuring that training data comprise at least 10 000 inactive data points, or a ratio of 3:1 inactives to actives per target. A total of 16 188 048 additional putative inactives were sampled in this manner. Conversely, 1003 target models required random undersampling to achieve a 50:1 maximum ratio of inactive to active molecules. The putative inactives were combined with the inactive data set, producing a final data set of 602 887 162 inactive compounds.

**Compound Preprocessing.** Compound structures were standardized using an in-house script<sup>43</sup> set to remove salts, normalize charges, and tautomerize compounds. To ensure only drug-like molecules are retained for training purposes, structures were filtered for duplicates, heavy metals, a molecular weight between 100 and 1000 Da, and the presence of at least one carbon atom.

**Target and Pathway Deconvolution.** RDKit<sup>44</sup> was used to generate 2048 bit Morgan fingerprints<sup>45</sup> with a radius of 2 bonds (known as ECFP<sub>4</sub> fingerprints). This algorithm generates a binary bit-string representation for the presence or absence (represented via a “1” or “0,” respectively) of atom environment features in a molecule, which can be interpreted by the Bernoulli Naïve Bayes (NB) algorithm. This procedure was conducted for each of the compounds in the bioactivity training set. ECFP<sub>4</sub> fingerprints were selected since they have been previously shown to be successful in capturing relevant molecular information for *in silico* bioactivity prediction.<sup>46</sup> The Bernoulli Naïve Bayes (NB) classifier, implemented by Scikit-learn,<sup>37</sup> was trained using the binary matrix of the active and inactive compound–target fingerprints on a per target basis. In this procedure, a NB model is trained for a single target using the active and inactive compounds annotated for that target. Such models generate a posterior likelihood of activity by an input compound for that target. The 2882 individual target models are deployed together when performing target deconvolution, in order to generate the predicted bioactivity spectra for the complete range of targets, which to our knowledge is the largest *in silico* target deconvolution protocol currently published.

The average precision and recall for target prediction models is calculated over a stratified 5-fold split (Supporting Information Figure 2). Overall, target prediction models with positive hits performed with an average precision and recall of 67.9% and 72.4%, respectively, with the variation in target model size and structural diversity of compounds describing each target class known to affect performance of models.<sup>40</sup> Pathway annotations were extracted from the WikiPath-

ways database<sup>47</sup> and aggregated to describe the total pathways predicted for a set of compounds.

**Definition of Non-Toxic, Cytotoxic, and Cytostatic Testing Sets.** Results from a screening cascade were used to generate libraries of cytotoxic and nontoxic compound sets (Figure 1), comprising a single-concentration AlamarBlue cell viability assessment of 388 000 compounds (screen A), subsequent profiling of 25 000 top hits via a follow-up dose–response AlamarBlue procedure (screen B), and CellTox Green membrane integrity assessment of hits from the dose–response screen (screen C).

A total of 296 970 compounds with activity values between –30% and 30% were extracted from the primary screen A for use as a nontoxic compound library (filtering between these values removes false-positive fluorescent compounds). This extensive inactive chemical space affords target prediction and pathway annotation of targets and processes that are not correlated with toxicity and corrects for model promiscuity and biases in bioactivity space. It is possible that toxic compounds may test negative in this set, since these are unable to penetrate into the cell to elicit cytotoxicity. Activity values ( $pIC_{50}$ ) greater than 4.949 in screen B and screen C were used as a cutoff to define cytotoxicity, producing a library of 5784 CytoToxic and CytoStatic (CTCS) compounds. Compounds with activity values ( $pIC_{50}$ ) greater than 4.949 in screen B and activity values ( $pIC_{50}$ ) less than 4.5 in screen C were extracted as a library of 2327 cytostatic-only compounds.

**Enrichment Calculation of Targets and Pathways.** An enrichment calculation was performed to avoid the over- or under-annotation of target predictions or pathway annotation, elucidating the relationships that are more likely to be responsible for the phenotypic response. The *Prediction Ratio* for a given selection of targets/annotations is calculated by comparing the frequency of compounds that are predicted as active for a given target in the CTCS or cytostatic set ( $F_t$ ) versus the frequency of compounds predicted as active for the target in the nontoxic or cytostatic set ( $F_b$ ) when normalizing for the total number of compound–target predictions in each set ( $N$ ). If  $F_b = 0$ , then the *Prediction Ratio* is assigned a score of 0, which indicates that a target is perfectly enriched in the test set.

$$\text{Prediction Ratio} = \frac{F_t/N_t}{F_b/N_b} \quad (1)$$

The *Prediction Ratio* is used to score the enrichment of the cytotoxic targets and pathways with nontoxic compounds ( $t = \text{CTCS}$ ,  $b = \text{nontoxic}$ ) or to compare cytostatic targets and pathways with cytotoxic predictions ( $t = \text{cytostatic}$ ,  $b = \text{cytotoxic}$ ). The Fisher’s exact test was calculated to indicate the probability of obtaining a prediction distribution at least as extreme as the one observed. This statistical measure was selected since it has been shown to be used for calculating enrichment for the deconvolution of phenotypic screens.<sup>48</sup>

**Cytotoxicity End point Modeling and 5-Fold Validation.** A random forest (RF) of 100 trees, with the number of features set to “auto” and class weight set to “balanced,” was implemented in Scikit-learn. The model was trained using ECFP<sub>4</sub> fingerprints generated in RDKit (in the same manner as employed when training the target prediction models) for the 5784 proprietary toxic compounds and 17 352 randomly under-sampled nontoxic compounds, while supplying the *fit* method with the sample weights of the CTCS to nontoxic compound ratio. Cytotoxic probabilities are defined as the majority predicted class probabilities of the trees in the forest, where the class probability for a tree is computed as the fraction of the same number of class samples in a leaf.<sup>37</sup> A second RF was trained on 3720 publicly available cytotoxic compounds and 11 160 randomly under-sampled nontoxic compounds. The proprietary model performs with average precision and recall values of 88.0% and 93.1% during 5-fold stratified cross-validation (Supporting Information Table 2).

**Fragment Analysis of the CTCS Compounds.** The CTCSs were fragmented at single nonring carbon–carbon bonds, filtered between 4 and 20 heavy atoms, producing 12 346 unique fragments. The *Fragment Ratio* for the resulting fragments was calculated by comparing the frequency of R-group fragment occurrences in the

cytotoxic set ( $F_{\text{tox}}$ ) versus the frequency of fragments in the nontoxic set ( $F_{\text{nontox}}$ ) when normalizing for the total number of compounds in each set ( $N$ ). If  $F_{\text{nontox}} = 0$ , then the *Fragment Ratio* is assigned a score of 0, which indicates that a fragment is perfectly enriched in the cytotoxic set.

$$\text{Fragment Ratio} = \frac{F_{\text{tox}}/N_{\text{tox}}}{F_{\text{nontox}}/N_{\text{nontox}}} \quad (2)$$

Fragments are filtered for duplicate molecular frameworks, retaining the first ratio ranking instance. Binomial probability was calculated to correct for the probability that a fragment can occur by chance. Target profiles are generated for fragment cores and mapped to the ChEMBL protein classification system. Predictions are then averaged over all fragment cores. The “other” category comprises targets that are not currently assigned classification using internal methods. Representative fragments highlighted in the tables are further filtered for an atom count of less than or equal to 16.

## ■ ASSOCIATED CONTENT

### ● Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: [10.1021/acschembio.6b00538](https://doi.org/10.1021/acschembio.6b00538).

Precision-Recall Curve of the Cytotoxicity Classification Model Shows Performances for Similarity Bins of Validated Compounds (Figure 1) and Precision and Recall Values for the 5-fold Cross Validation of Target Prediction Models (Figure 2) (PDF)

Comprehensive List of Enriched Targets and Pathways (XLSX)

5-fold Cross Validation Performance of the Cytotoxicity Endpoint Prediction Protocol (XLSX)

Target Classification Covered using AstraZeneca Collections of Active Bioactivity Data (XLSX)

Target Classifications Covered using Inactive Training Data from AstraZeneca HTS Screens (XLSX)

Number of Inactive Data Added using Inactive AstraZeneca HTS collections (XLSX)

## ■ AUTHOR INFORMATION

### Corresponding Author

\*E-mail: [ab454@cam.ac.uk](mailto:ab454@cam.ac.uk).

### Author Contributions

L.H.M. created, implemented, and applied the MoA protocol and cytotoxicity classification protocols presented in this work and wrote this manuscript. Q.C. performed the fragment analysis. I.P.B. and M.A.F. helped with data curation and analysis of results. M.W., D.M., and L.M. performed the experimental screenings. A.B., O.E., D.M., and M.W. conceived the main theme on which the work was performed and ensured that scientific aspect of the study was rationally valid. All authors contributed to revising the final draft of the manuscript.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

L.H.M. would like to thank BBSRC and AstraZeneca and for their funding, L. Scheidt and Z. You for proofreading the manuscript, I. Feierberg for advice on HTS data collection, and S. Ashenden for helpful discussions.

## ■ REFERENCES

- (1) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; and Sittampalam, G. S. (2011) Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discovery* 10, 188–195.
- (2) Hay, M.; Thomas, D. W.; Craighead, J. L.; Economides, C.; and Rosenthal, J. (2014) Clinical development success rates for investigational drugs. *Nat. Biotechnol.* 32, 40–51.
- (3) Wigglesworth, M. J.; Murray, D. C.; Blackett, C. J.; Kossenjans, M.; and Nissink, J. W. (2015) Increasing the delivery of next generation therapeutics from high throughput screening libraries. *Curr. Opin. Chem. Biol.* 26, 104–110.
- (4) Rees, S.; Gribbon, P.; Birmingham, K.; Janzen, W. P.; and Pairaudeau, G. (2016) Towards a hit for every target. *Nat. Rev. Drug Discovery* 15, 1–2.
- (5) Waring, M. J.; Arrowsmith, J.; Leach, A. R.; Leeson, P. D.; Mandrell, S.; Owen, R. M.; Pairaudeau, G.; Pennie, W. D.; Pickett, S. D.; Wang, J.; Wallace, O.; and Weir, A. (2015) An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat. Rev. Drug Discovery* 14, 475–486.
- (6) Mignani, S.; Huber, S.; Tomás, H.; Rodrigues, J.; and Majoral, J. P. (2016) Compound high-quality criteria: a new vision to guide the development of drugs, current situation. *Drug Discovery Today* 21, 573.
- (7) Hoelder, S.; Clarke, P. A.; and Workman, P. (2012) Discovery of small molecule cancer drugs: successes, challenges and opportunities. *Mol. Oncol.* 6, 155–176.
- (8) Cook, D.; Brown, D.; Alexander, R.; March, R.; Morgan, P.; Satterthwaite, G.; and Pangalos, M. N. (2014) Lessons learned from the fate of AstraZeneca’s drug pipeline: a five-dimensional framework. *Nat. Rev. Drug Discovery* 13, 419–431.
- (9) Roberts, R. A.; Kavanagh, S. L.; Mellor, H. R.; Pollard, C. E.; Robinson, S.; and Platz, S. J. (2014) Reducing attrition in drug development: smart loading preclinical safety assessment. *Drug Discovery Today* 19, 341–347.
- (10) Galluzzi, L.; Bravo-San Pedro, J. M.; Vitale, I.; Aaronson, S. A.; Abrams, J. M.; Adam, D.; Alnemri, E. S.; Altucci, L.; Andrews, D.; Annicchiarico-Petruzzelli, M.; Baehrecke, E. H.; Bazan, N. G.; Bertrand, M. J.; Bianchi, K.; Blagosklonny, M. V.; Blomgren, K.; Borner, C.; Bredesen, D. E.; Brenner, C.; Campanella, M.; Candi, E.; Cecconi, F.; Chan, F. K.; Chandel, N. S.; Cheng, E. H.; Chipuk, J. E.; Cidlowski, J. A.; Ciechanover, A.; Dawson, T. M.; Dawson, V. L.; De Laurenzi, V.; De Maria, R.; Debatin, K. M.; Di Daniele, N.; Dixit, V. M.; Dynlacht, B. D.; El-Deiry, W. S.; Fimia, G. M.; Flavell, R. A.; Fulda, S.; Garrido, C.; Gougeon, M. L.; Green, D. R.; Gronemeyer, H.; Hajnoczky, G.; Hardwick, J. M.; Hengartner, M. O.; Ichijo, H.; Joseph, B.; Jost, P. J.; Kaufmann, T.; Kepp, O.; Klionsky, D. J.; Knight, R. A.; Kumar, S.; Lemasters, J. J.; Levine, B.; Linkermann, A.; Lipton, S. A.; Lockshin, R. A.; López-Otín, C.; Lugli, E.; Madeo, F.; Malorni, W.; Marine, J. C.; Martin, S. J.; Martinou, J. C.; Medema, J. P.; Meier, P.; Melino, S.; Mizushima, N.; Moll, U.; Muñoz-Pinedo, C.; Nuñez, G.; Oberst, A.; Panaretakis, T.; Penninger, J. M.; Peter, M. E.; Piacentini, M.; Pinton, P.; Prehn, J. H.; Puthalath, H.; Rabinovich, G. A.; Ravichandran, K. S.; Rizzuto, R.; Rodrigues, C. M.; Rubinsztein, D. C.; Rudel, T.; Shi, Y.; Simon, H. U.; Stockwell, B. R.; Szabadkai, G.; Tait, S. W.; Tang, H. L.; Tavernarakis, N.; Tsujimoto, Y.; Vanden Berghe, T.; Vandenberghe, P.; Villunger, A.; Wagner, E. F.; Walczak, H.; White, E.; Wood, W. G.; Yuan, J.; Zakeri, Z.; Zhivotovskiy, B.; Melino, G.; and Kroemer, G. (2015) Essential versus accessory aspects of cell death: recommendations of the NCCD 2015. *Cell Death Differ.* 22, 58–73.
- (11) Chaabane, W.; User, S. D.; El-Gazzah, M.; Jaksik, R.; Sajjadi, E.; Rzeszowska-Wolny, J.; and Los, M. J. (2013) Autophagy, apoptosis, mitoptosis and necrosis: interdependence between those pathways and effects on cancer. *Arch. Immunol. Ther. Exp.* 61, 43–58.
- (12) Galluzzi, L.; Vitale, I.; Abrams, J. M.; Alnemri, E. S.; Baehrecke, E. H.; Blagosklonny, M. V.; Dawson, T. M.; Dawson, V. L.; El-Deiry, W. S.; Fulda, S.; Gottlieb, E.; Green, D. R.; Hengartner, M. O.; Kepp, O.; Knight, R. A.; Kumar, S.; Lipton, S. A.; Lu, X.; Madeo, F.; Malorni, W.; Mehlen, P.; Nuñez, G.; Peter, M. E.; Piacentini, M.; Rubinsztein, D.

- C., Shi, Y., Simon, H. U., Vandenabeele, P., White, E., Yuan, J., Zhivotovskiy, B., Melino, G., and Kroemer, G. (2012) Molecular definitions of cell death subroutines: recommendations of the Nomenclature Committee on Cell Death 2012. *Cell Death Differ.* 19, 107–120.
- (13) Nikolettou, V., Markaki, M., Palikaras, K., and Tavernarakis, N. (2013) Crosstalk between apoptosis, necrosis and autophagy. *Biochim. Biophys. Acta, Mol. Cell Res.* 1833, 3448–3459.
- (14) Mariño, G., Niso-Santano, M., Baehrecke, E. H., and Kroemer, G. (2014) Self-consumption: the interplay of autophagy and apoptosis. *Nat. Rev. Mol. Cell Biol.* 15, 81–94.
- (15) Hotchkiss, R. S., Strasser, A., McDunn, J. E., and Swanson, P. E. (2009) Cell death. *N. Engl. J. Med.* 361, 1570–1583.
- (16) Rixe, O., and Fojo, T. (2007) Is cell death a critical end point for anticancer therapies or is cytostasis sufficient? *Clin. Cancer Res.* 13, 7280–7287.
- (17) Lupi, M., Cappella, P., Matera, G., Natoli, C., and Ubezio, P. (2006) Interpreting cell cycle effects of drugs: the case of melphalan. *Cancer Chemother. Pharmacol.* 57, 443–457.
- (18) Raies, A. B., and Bajic, V. B. (2016) In silico toxicology: computational methods for the prediction of chemical toxicity. *WIREs Comput. Mol. Sci.* 6, 147.
- (19) Cruz-Monteagudo, M., Medina-Franco, J. L., Pérez-Castillo, Y., Nicolotti, O., Cordeiro, M. N., and Borges, F. (2014) Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discovery Today* 19, 1069–1080.
- (20) Patlewicz, G., and Fitzpatrick, J. M. (2016) Current and Future Perspectives on the Development, Evaluation, and Application of In Silico Approaches for Predicting Toxicity. *Chem. Res. Toxicol.* 29, 438–451.
- (21) Stumpfe, D., and Bajorath, J. (2012) Exploring activity cliffs in medicinal chemistry. *J. Med. Chem.* 55, 2932–2942.
- (22) Lin, Z., and Will, Y. (2012) Evaluation of drugs with specific organ toxicities in organ-specific cell lines. *Toxicol. Sci.* 126, 114–127.
- (23) Cortés-Ciriano, I., and Bender, A. (2016) How Consistent are Publicly Reported Cytotoxicity Data? Large-Scale Statistical Analysis of the Concordance of Public Independent Cytotoxicity Measurements. *ChemMedChem* 11, 57–71.
- (24) Pohjala, L., Tammela, P., Samanta, S. K., Yli-Kauhaluoma, J., and Vuorela, P. (2007) Assessing the data quality in predictive toxicology using a panel of cell lines and cytotoxicity assays. *Anal. Biochem.* 362, 221–228.
- (25) Liggi, S., Drakakis, G., Koutsoukas, A., Cortes-Ciriano, I., Martínez-Alonso, P., Malliavin, T. E., Velazquez-Campoy, A., Brewerton, S. C., Bodkin, M. J., Evans, D. A., Glen, R. C., Carrodegua, J. A., and Bender, A. (2014) Extending in silico mechanism-of-action analysis by annotating targets with pathways: application to cellular cytotoxicity readouts. *Future Med. Chem.* 6, 2029–2056.
- (26) Liggi, S., Drakakis, G., Hendry, A. E., Hanson, K. M., Brewerton, S. C., Wheeler, G. N., Bodkin, M. J., Evans, D. A., and Bender, A. (2013) Extensions to in silico bioactivity predictions using pathway annotations and differential pharmacology analysis: application to xenopus laevis phenotypic readouts. *Mol. Inf.* 32, 1009–1024.
- (27) Kelder, T., van Iersel, M. P., Hanspers, K., Kutmon, M., Conklin, B. R., Evelo, C. T., and Pico, A. R. (2012) WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.* 40, D1301–1307.
- (28) Flachner, B., Lörincz, Z., Carotti, A., Nicolotti, O., Kuchipudi, P., Remez, N., Sanz, F., Tóvári, J., Szabó, M. J., Bertók, B., Cseh, S., Mestres, J., and Dormán, G. (2012) A chemocentric approach to the identification of cancer targets. *PLoS One* 7, e35582.
- (29) Dempster, E. L., Wong, C. C., Lester, K. J., Burrage, J., Gregory, A. M., Mill, J., and Eley, T. C. (2014) Genome-wide methylomic analysis of monozygotic twins discordant for adolescent depression. *Biol. Psychiatry* 76, 977–983.
- (30) Heo, J., Li, J., Summerlin, M., Hays, A., Katyal, S., McKinnon, P. J., Nitiss, K. C., Nitiss, J. L., and Hanakahi, L. A. (2015) TDP1 promotes assembly of non-homologous end joining protein complexes on DNA. *DNA Repair* 30, 28–37.
- (31) Turk, B. E., Wong, T. Y., Schwarzenbacher, R., Jarrell, E. T., Leppla, S. H., Collier, R. J., Liddington, R. C., and Cantley, L. C. (2004) The structural basis for substrate and inhibitor selectivity of the anthrax lethal factor. *Nat. Struct. Mol. Biol.* 11, 60–66.
- (32) Wahdan-Alaswad, R. S., Bane, K. L., Song, K., Shola, D. T., Garcia, J. A., and Danielpour, D. (2012) Inhibition of mTORC1 kinase activates Smads 1 and 5 but not Smad8 in human prostate cancer cells, mediating cytostatic response to rapamycin. *Mol. Cancer Res.* 10, 821–833.
- (33) Chawla, S., Kvalnes, K., deLong, M. A., Wickett, R., Manga, P., and Boissy, R. E. (2012) DeoxyArbutin and its derivatives inhibit tyrosinase activity and melanin synthesis without inducing reactive oxygen species or apoptosis. *J. Drugs Dermatol* 11, e28–34.
- (34) Alahari, S. K., Reddig, P. J., and Juliano, R. L. (2004) The integrin-binding protein Nischarin regulates cell migration by inhibiting PAK. *EMBO J.* 23, 2777–2788.
- (35) Sun, S., Han, J., Ralph, W. M., Chandrasekaran, A., Liu, K., Auburn, K. J., and Carter, T. H. (2004) Endoplasmic reticulum stress as a correlate of cytotoxicity in human tumor cells exposed to diindolylmethane in vitro. *Cell Stress Chaperones* 9, 76–87.
- (36) Cumming, J. G., Davis, A. M., Muresan, S., Haerberlein, M., and Chen, H. (2013) Chemical predictive modelling to improve compound quality. *Nat. Rev. Drug Discovery* 12, 948–962.
- (37) Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.
- (38) Muresan, S., Petrov, P., Southan, C., Kjellberg, M. J., Kogej, T., Tyrchan, C., Varkonyi, P., and Xie, P. H. (2011) Making every SAR point count: the development of Chemistry Connect for the large-scale integration of structure and bioactivity data. *Drug Discovery Today* 16, 1019–1030.
- (39) Mervin, L. H., Afzal, A. M., Drakakis, G., Lewis, R., Engkvist, O., and Bender, A. (2015) Target prediction utilising negative bioactivity data covering large chemical space. *J. Cheminf.* 7, 51.
- (40) Koutsoukas, A., Lowe, R., Kalantarmotamedi, Y., Mussa, H. Y., Klafke, W., Mitchell, J. B., Glen, R. C., and Bender, A. (2013) In silico target predictions: defining a benchmarking data set and comparison of performance of the multiclass Naïve Bayes and Parzen-Rosenblatt window. *J. Chem. Inf. Model.* 53, 1957–1966.
- (41) Coordinators, N. R. (2013) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 41, D8–D20.
- (42) Gfeller, D., and Zoete, V. (2015) Protein homology reveals new targets for bioactive small molecules. *Bioinformatics* 31, 2721–2727.
- (43) OEChem T, version 2.0, OpenEye Scientific Software, Santa Fe, NM.
- (44) Landrum, G. (2012) RDKit: Open-source cheminformatics. <http://www.rdkit.org/>.
- (45) Morgan, H. L. (1965) The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* 5, 107–113.
- (46) Wale, N., and Karypis, G. (2009) Target fishing for chemical compounds using target-ligand activity data and ranking based methods. *J. Chem. Inf. Model.* 49, 2190–2201.
- (47) Pico, A. R., Kelder, T., van Iersel, M. P., Hanspers, K., Conklin, B. R., and Evelo, C. (2008) WikiPathways: pathway editing for the people. *PLoS Biol.* 6, e184.
- (48) Bornot, A., Blackett, C., Engkvist, O., Murray, C., and Bendtsen, C. (2014) The Role of Historical Bioactivity Data in the Deconvolution of Phenotypic Screens. *J. Biomol. Screening* 19, 696–706.
- (49) Thaker, N. G., Zhang, F., McDonald, P. R., Shun, T. Y., Lewen, M. D., Pollack, I. F., and Lazo, J. S. (2009) Identification of survival genes in human glioblastoma cells by small interfering RNA screening. *Mol. Pharmacol.* 76, 1246–1255.
- (50) Rines, A. K., Burke, M. A., Fernandez, R. P., Volpert, O. V., and Ardehali, H. (2012) Snf1-related kinase inhibits colon cancer cell



proliferation through calcyclin-binding protein-dependent reduction of  $\beta$ -catenin. *FASEB J.* 26, 4685–4695.

(51) Ansari, D., Andersson, R., Bauden, M. P., Andersson, B., Connolly, J. B., Welinder, C., Sasor, A., and Marko-Varga, G. (2015) Protein deep sequencing applied to biobank samples from patients with pancreatic cancer. *J. Cancer Res. Clin. Oncol.* 141, 369–380.

(52) Zha, J., Zhou, Q., Xu, L. G., Chen, D., Li, L., Zhai, Z., and Shu, H. B. (2004) RIP5 is a RIP-homologous inducer of cell death. *Biochem. Biophys. Res. Commun.* 319, 298–303.

(53) Wang, L. Y., and Kung, H. J. (2012) Male germ cell-associated kinase is overexpressed in prostate cancer cells and causes mitotic defects via deregulation of APC/CCDH1. *Oncogene* 31, 2907–2918.

(54) Parray, A. A., Baba, R. A., Bhat, H. F., Wani, L., Mokhdomi, T. A., Mushtaq, U., Bhat, S. S., Kirmani, D., Kuchay, S., Wani, M. M., and Khanday, F. A. (2014) MKK6 is upregulated in human esophageal, stomach, and colon cancers. *Cancer Invest.* 32, 416–422.

(55) Orvedahl, A., Sumpster, R., Jr., Xiao, G., Ng, A., Zou, Z., Tang, Y., Narimatsu, M., Gilpin, C., Sun, Q., Roth, M., Forst, C. V., Wrana, J. L., Zhang, Y. E., Luby-Phelps, K., Xavier, R. J., Xie, Y., and Levine, B. (2011) Image-based genome-wide siRNA screen identifies selective autophagy factors. *Nature* 480, 113–117.

(56) MacKeigan, J. P., Murphy, L. O., and Blenis, J. (2005) Sensitized RNAi screen of human kinases and phosphatases identifies new regulators of apoptosis and chemoresistance. *Nat. Cell Biol.* 7, 591–600.

(57) Chawla-Sarkar, M., Lindner, D. J., Liu, Y. F., Williams, B. R., Sen, G. C., Silverman, R. H., and Borden, E. C. (2003) Apoptosis and interferons: role of interferon-stimulated genes as mediators of apoptosis. *Apoptosis* 8, 237–249.

(58) Parker, L. L., Walter, S. A., Young, P. G., and Piwnica-Worms, H. (1993) Phosphorylation and inactivation of the mitotic inhibitor Wee1 by the nim1/cdr1 kinase. *Nature* 363, 736–738.

(59) Wroble, B. N., Finkielstein, C. V., and Sible, J. C. (2007) Wee1 kinase alters cyclin E/Cdk2 and promotes apoptosis during the early embryonic development of *Xenopus laevis*. *BMC Dev. Biol.* 7, 119.

(60) Qi, J., Su, Y., Sun, R., Zhang, F., Luo, X., Yang, Z., and Luo, X. (2005) CASK inhibits ECV304 cell growth and interacts with Id1. *Biochem. Biophys. Res. Commun.* 328, 517–521.

(61) Pollak, M. N., Schernhammer, E. S., and Hankinson, S. E. (2004) Insulin-like growth factors and neoplasia. *Nat. Rev. Cancer* 4, 505–518.

(62) Valli, E., Trazzi, S., Fuchs, C., Erriquez, D., Bartesaghi, R., Perini, G., and Ciani, E. (2012) CDKL5, a novel MYCN-repressed gene, blocks cell cycle and promotes differentiation of neuronal cells. *Biochim. Biophys. Acta, Gene Regul. Mech.* 1819, 1173–1185.

(63) Clemens, M. J. (2001) Initiation factor eIF2 alpha phosphorylation in stress responses and apoptosis. *Prog. Mol. Subcell. Biol.* 27, 57–89.

(64) Zhang, W., and Go, M. L. (2009) Functionalized 3-benzylidene-indolin-2-ones: inducers of NAD(P)H-quinone oxidoreductase 1 (NQO1) with antiproliferative activity. *Bioorg. Med. Chem.* 17, 2077–2090.

(65) Wang, X., Jiang, W., Duan, N., Qian, Y., Zhou, Q., Ye, P., Jiang, H., Bai, Y., Zhang, W., and Wang, W. (2014) NOD1, RIP2 and Caspase12 are potentially novel biomarkers for oral squamous cell carcinoma development and progression. *Int. J. Clin. Exp. Pathol.* 7, 1677–1686.

(66) De Felice, B., Garbi, C., Wilson, R. R., Santoriello, M., and Nacca, M. (2011) Effect of selenocystine on gene expression profiles in human keloid fibroblasts. *Genomics* 97, 265–276.

(67) Mavrogomatou, E., Papadimitriou, K., Urban, J. P., Papadopoulos, V., and Kletsas, D. (2015) Deficiency in the  $\alpha 1$  subunit of Na<sup>+</sup>/K<sup>+</sup>-ATPase enhances the anti-proliferative effect of high osmolality in nucleus pulposus intervertebral disc cells. *J. Cell. Physiol.* 230, 3037–3048.