Integration of Image-Based Cytological Phenotypes with Computational Ligand-Target Prediction to Identify Mechanisms of Action

Daniel W. Young\textsuperscript{1,3}, Andreas Bender\textsuperscript{2}, Jonathan Hoyt\textsuperscript{1}, Elizabeth Mcwhinnie\textsuperscript{1}, Gung-wei Chirn\textsuperscript{1}, Charles Tao\textsuperscript{1}, John A. Tallarico\textsuperscript{4}, Mark Labow\textsuperscript{1}, Jeremy L. Jenkins\textsuperscript{2}, Timothy J. Mitchison\textsuperscript{3}, and Yan Feng\textsuperscript{1}

1. Developmental and Molecular Pathways, Novartis Institutes for BioMedical Research, Cambridge, MA 02139, USA.
2. Lead Discovery Informatics, Novartis Institutes for BioMedical Research, Cambridge, MA 02139, USA.
3. Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA.
4. Global Discovery Chemistry, Novartis Institutes for BioMedical Research, Cambridge, MA 02139, USA.

Main Text Word Count: 4804
Main Text Character Count: 28659
References: 53

#Correspondence: yan.feng@novartis.com
Summary:

High-content screening (HCS) is transforming drug discovery by enabling simultaneous measurement of multiple features of cellular phenotype that are relevant to therapeutic and toxic activities of compounds. HCS studies typically generate immense datasets of image-based phenotypic information, and how best to mine relevant phenotypic data is an unsolved challenge. Here, we introduce factor analysis as a data-driven tool for defining cell phenotypes, and profiling compound activities. This method allows a large data reduction while retaining relevant information, and the data-derived factors used to quantify phenotype have discernable biological meaning. We use factor analysis of cells stained with fluorescent markers of cell cycle state to profile a compound library, and cluster the hits into seven phenotypic categories. We then compare phenotypic profiles, chemical similarity and predicted protein binding activity of active compounds. By integrating these different descriptors of measured and potential biological activity, we can effectively draw mechanism of action inferences.

Introduction:

Drug discovery requires integration of chemical and biological knowledge about many compounds in an efficient manner\(^1\). Profiling compounds by chemical structure has become increasingly sophisticated, but profiling by biological activity has lagged due to the difficulty of collecting and integrating different types of biological information, and also the large expense of data-rich methods such as mRNA expression profiling. High content screening (HCS) combines automated microscopy with image analysis to enable
phenotypic profiling of compounds based on activities on cells visualized by fluorescence cytology. This rapidly developing technology is increasingly used to facilitate both target and lead characterization. The instrumentation and image quantification aspects of HCS, while under constant improvement, are already well advanced. Methods for downstream data processing, and mining of biological data are, by comparison, significantly less refined. Most users score for pre-defined phenotypes of interest, such as nuclear translocation of a transcription factor, largely ignoring the wealth of phenotypic information present in most HCS datasets. Thus, the huge potential of HCS to inform on biological effects relevant to therapeutics and toxicity is largely untapped.

Two problems have limited the use of HCS to report broadly on phenotypic effects of compounds: the large size of the datasets, and the fact that the biological meaning of most of the measurements is unclear. A typical HCS experiment might generate terabytes of image data from which gigabytes of numbers are extracted describing the amount and location of biomolecules on a cell-to-cell basis. Most of these numbers have no obvious biological meaning; for example, while the amount of DNA per nucleus has obvious significance, that of other nuclear measures, such as DNA texture, or nuclear ellipticity, are much less clear. This leads biologists to ignore the non-obvious measurements, even though they may report usefully on compound activities. Here, we introduce Factor Analysis to mine HCS datasets. This method was developed more than a century ago, remains standard in other fields for analyzing large, multidimensional datasets, and was implemented here using standard, commercially available statistics software. It allows a large data-reduction, and quantifies phenotype using data-derived factors that are biologically interpretable in many cases.
The basic supposition underlying Factor Analysis is that groups of variables within a multivariate data set that are highly correlated with each other, but poorly correlated with other variables in the data set, are likely to be measuring a common underlying trait, or “Factor”\(^\text{16}\). In HCS, this translates to the reasonable supposition that groups of image-based cell features that exhibit highly correlated changes between individual cells, following different compound treatments, are likely reporting on a common phenotypic property. If this supposition is true, we should often be able to interpret the biological meaning of the factors, even though they were generated directly from the data without biological assumptions. Here, we use cytological markers of cell cycle, HCS and Factor Analysis to profile the biological effects of a compound library. We find that six factors are sufficient to describe the biological responses, that several of them have interpretable biological meaning, and that they group the active compounds into seven major categories by phenotypic effects. We then compare how phenotypic profiles of active compounds compare with chemical structure and predicted target profiles. The resulting structure-activity relationships are more information-rich than would be possible with a single data type, and allow us to infer mechanisms of action for some compounds.

**Materials and Methods**

*Factor Analysis*

For High-Content applications, data are contained in an \(n \times m\) matrix, \(X\) consisting of a set of \(n\) image-based features measured on \(m\) cells. From a screening stand-point, one is typically not interested in the features contained within \(X\), per se, but rather with the underlying cellular processes that control these features. For this
philosophical reason Factor Analysis is highly appropriate to high-content imaging, as it seeks to identify these underlying processes. In mathematical terms the so-called, Common Factor Model, posits that a set of measured random variables, $X$ is a linear function of common Factors, $F$ and unique Factors, $\varepsilon$:

$$X = LF + \varepsilon$$

In HCS the common factors in $F$ reflect the set of major phenotypic attributes measured in the assay. The loading matrix, $L$ relates the measured variables in $X$ to $F$. Whereas, $\varepsilon$ is a matrix of unique Factors and is comprised of the reliable effects and the random error that is specific to a given variable. Rooted in this model is the concept that the total variance of $X$ is partitioned into common and specific components. Specifically, it can be shown that the following covariance structure exists for $X$,

$$\Sigma = LL^T + \Psi$$

where $^T$ is the transpose operator and $\Psi$ is the covariance of $\varepsilon$, a diagonal matrix whose, $n$ non-zero components are specific variances for the $n$ random variables (cell features). The common portion of the co-variance is the squared Factor loading matrix, $LL'$. 

Fitting the Factor model requires estimating the loading matrix, $L$ and the specific covariance matrix, $\Psi$. With some underlying restrictions placed on the structure of $\Sigma$, the model fit can be accomplished quite easily\textsuperscript{16}. The Factor model fit was performed here using the so-called Principal Factor method\textsuperscript{16,17}; was carried out using the Factor procedure in statistical analysis software, SAS (SAS Institute Inc., Cary NC); and involves the following steps: 1. standardize the data matrix, $X$ to zero-mean and unit-variance column-wise. 2. compute the sample correlation matrix, $R$. 3. Generate the adjusted correlation matrix $R^*$ by setting the diagonal elements (i.e., communalities) of, $R$ to the squared multiple correlations of each variable in $X$ with all other variables. 4.
Perform an eigenanalysis on, \( R^* \) to determine the appropriate number of Factors, \( k \) according to the Kaiser criterion; i.e., where the number of Factors is equal to the number of eigenvalues greater than one. 5. Using an \( k \) Factor Model, estimate the loading matrix, \( L \) through spectral decomposition. Such that,

\[
L = \begin{bmatrix} \sqrt{\lambda_1} e_1 & \sqrt{\lambda_2} e_2 & \ldots & \sqrt{\lambda_k} e_k \end{bmatrix}
\]

Where \( \lambda_i \) is the eigenvalue associated with the eigenvector \( e_i \), derived from the adjusted correlation matrix, \( R^* \) with \( k \) Factors.

The loading matrix, \( L \) relates the inputs variables, \( X \) to the underlying common factors, \( F \). To facilitate understanding of the common factors the loading matrix is rotated for ease of factor interpretability. The justification for factor rotation derives from the fact that there are an infinite number of loading matrices that can be specified with the same statistical properties and that reproduce the same covariance matrix, \( \Sigma \).

An \( n \times n \) orthogonal rotation matrix, \( T \) can be specified such that:

\[
\Sigma = LL^T + \Psi = LTT^TL^T + \Psi = \Lambda\Lambda^T + \Psi
\]

And,

\[
\Lambda = LT = Rotated \ Loading \ Matrix
\]

There are several methods for defining the rotation matrix, \( T \). These approaches are broadly classified based on whether they preserve the independence between factors (i.e., orthogonal rotations) or they permit correlation between factors (i.e., oblique rotations). Here we employ the orthogonal Varimax method, an orthogonal rotation strategy that maximizes the variance in factor loadings. This approach results in a simple structure with factors that have a small number of high loading variables and a large number of zero loading variables, and yields factors that can be readily interpreted based on the set of variables with high loading.
The factor model was fit using the steps described above on a data set comprising two replicate screens on 6547 compounds and ~600 control treatments (no compound), for which 36 cytological features (Supplementary Table 1) were measured on approximately 500-600 cells. A 1% random sample of the entire data set (~0.3 x 10^9 data points) was generated and used to fit the factor model. In practice we have determined that this method of sampling is more than sufficient to produce a stable factor model. Stability is assessed by examining the factor structure (e.g., factor loadings, see Supplementary Data 1) for multiple random samples. Here we computed three random samples and observed essentially identical factor structures in each.

The Factor model provides insight on which cytological traits are prominent in the high-content assay. However, for phenotypic profiling purposes it is of interest to understand how individuals cells score on each Factor. Therefore, after fitting the Factor model and performing the rotations, we estimate a score, $s_F$, on each of the $k$ factors for each observation (i.e., cell) using a regression equation derived from the Factor model; this is accomplished using the Score procedure in SAS®. As a summary statistic for each treatment condition (i.e., well) we compute averages on each of the $k$ factor scores. Each average is determined by computing the mean of a factor score over all cells within a well. After hit selection averages are computed between corresponding replicate wells for profiling.

**Distance Metric**

We considered a phenotypic vector as the set of six well-averaged Factor score estimates, $F_s$. The Euclidean distance between each treatment phenotypic vector, $F_s'$ and
the control (untreated) vector, $\mathbf{F}_s^u$ defines our phenotypic response metric, $\mathbf{P}$ for each treatment:

$$\mathbf{P} = \sqrt{(\mathbf{F}'_s - \mathbf{F}^u_s)^T (\mathbf{F}'_s - \mathbf{F}^u_s)}$$

Where, $T$ is the transpose operator. This metric projects the multidimensional phenotype onto a single response dimension, enables a standard comparison between compounds with various bioactivities, and facilitates hit identification independent of the specific phenotype.

**Compound Transfer**

HELA cells (ATCC, Manassas, Va) were plated in 384 well, black clear bottomed plates (Greiner, Monroe, NC) at a density of 2000 cells per well in 25ul of growth medium (DMEM, 10%FBS, P&S, Invitrogen, Carlsbad, CA) for overnight incubation. Compounds were diluted in DMEM and 5ul of diluted compound was transferred to the 384 well culture plates at a final concentration of 10uM per well using the BioMek FX (Beckman Coulter, Fullerton, CA). Plates were transferred to 37°C and incubated for 20 hours.

**Deoxyuridine Label Preparation**

Rhodamine azide (a gift from Adrian Salic, Harvard Medical School) was added to a solution of 100mM Tris pH 8.5 and 100mM CuSO$_4$ to a final concentration of 1μM. Ascorbic acid was added last to a final concentration of 100mM. The solution was mixed briefly to complete the reaction.

**Cell Staining**
After 20 hours of incubation with compound, cells were pulsed with 500nM 5-ethynl-2’-deoxyuridine (Berry & Associates Inc., Dexter, MI) using a MultiDrop (Thermo Lab Systems, Waltham, MA) and incubated for 40 minutes at 37°C. Cells were fixed in 3.7% paraformaldehyde for 30 minutes at 25°C. Cells were washed once with PBST (phosphate buffered saline (Invitrogen), 0.5% triton X-100 (Sigma Aldrich, St. Louis, MO) using a Biotek Plate washer (Biotek Instruments, Winooski, Vt.) and then stained with rhodamine-azide for 30 minutes at 25°C. Plates were washed again with PBST and then incubated with primary antibodies. Rabbit anti-phospho-histone H3 Ser10 (Upstate, Billerica, MA) and mouse anti-α-tubulin (Sigma Aldrich) were added and plates were incubated at 25°C for 3 hours. Cells were washed once with PBST. Secondary antibodies (donkey anti-mouse Alexa-488, Invitrogen) and goat anti-rabbit-Cy5 (Amersham, Piscataway, NJ) were added for 2 hours at 25°C. Cells were washed once with PBST and stained with Hoescht 33342 (Invitrogen) for 30 minutes at 25°C. Wells were washed once with PBST, filled with PBS and sealed for imaging.

**Imaging**

Plates were imaged with a Cellomics Arrayscan (Cellomics, Pittsburg, PA). Images were collected using the XF93 filter set and 10X PlanFluor objective with camera binning set at 2x2. Individual cell segmentation was done using the Cellomics Morphology Explorer algorithm. Measurements for each cell were made on DNA intensity, nuclear area, deoxyuridine incorporation and phospho-H3 staining.

**Target Prediction Model**

Target prediction was performed using statistical models of substructural features, based on an annotated chemogenomics database which pairs ligand molecular structures
and the biological targets they act on. The underlying assumption made is the “molecular similarity principle” which assumes that similar molecules are likely to show similar properties. We used the WOMBAT database in version 2006.1 as a knowledge base for training, which associates 154,236 ligands with 1,336 protein targets in 256,039 data entries. ECFP_4 fingerprints were calculated for washed and normalized structures and multiple category Naïve Bayes models with Laplacian Correction were trained on all data points, as implemented in PipelinePilot 5.1. The five targets with highest Bayes scores were considered for further analysis. For further details on the target prediction used see the original publication as well as a recent review which gives an overview of currently available methods that is also highlighting some recent applications. The method employed in this work is based on ECFP_4 descriptors which are circular fingerprints, encoding molecules as a set of radial patches which in their completeness again characterize the whole molecule. Circular fingerprints in general have been found to contain significant information regarding bioactivity, but it was recently shown that 3D descriptors show better generalization performance in case no bioactive structures similar to the one under consideration are known. While overall quite high prediction performance of the correct target for >70% of the structures could be achieved in a validation study, the dependence of the method on the available knowledge base (training set) needs to be kept in mind. This is particularly true for novel chemotypes.

**Compound Library**

We screened and profiled a library of 6547 compounds derived from a diversity library (21%) and a library of known bioactive compounds (21%), and a natural products library (58%). The bioactive set comprises those Novartis compounds that were
recommended for promotion into development as drug candidates. This library has been compiled from multiple internal proprietary sources, and includes entries irrespective of whether the compounds succeeded in preclinical or clinical Development, or were introduced into the market. The natural products library consists of ~3800 compounds purified from plant extracts and other natural sources. In all cases compounds are stored lyophilized and have been determined by LC/MS to be at least 90% pure. Lyophilized compounds are resuspended in DMSO for a stock concentration of 10mM. Immediately before treatment samples of compound stock solution are diluted in DMEM to a 6x working concentration of 60µM. We provide a table outlining the available PubChem IDs for our 211 hit set (Supplementary Data 2).

Results

Factor Analysis of High-Content Image Data

We designed a HCS assay to identify compounds that affect cell proliferation, and to profile their cell cycle phenotype, using fluorescent probes for DNA (Ch1), DNA replication (Ch2) and mitosis (Ch3)(Figure 1). Probes for Ch1 (Hoechst 33342 dye) and Ch3 (anti-phosphoH3) were standard. To label sites of DNA replication in Ch2 we pulsed cells briefly with Ethynyl-deoxyuridine (EdU) prior to fixation. Classic Bromo-deoxyuridine (BrdU) staining is not ideal for HCS because many steps are required to visualize the probe, including a DNA denaturation step that perturbs nuclear morphology. EdU is incorporated into DNA during replication like BrdU, but visualization requires only a single reaction using “Click chemistry” to conjugate a Rhodamine-Azide dye to
the Ethynyl group. Images were acquired automatically using 10x objective and widefield imaging. For primary image analysis, the DNA stain was segmented to find nuclei. A nuclear mask was then used to generate 36 cytological features (all nuclear) from the three fluorescent channels (Supplementary Table 1). At least 500 cells were scored per treatment in two replicate experiments. We used the common factor model to map these 36-cytological features into a reduced dimensional space defined by a set of 6 orthogonal factors that reflect the major underlying phenotypic attributes measured in the assay (Figure 2A, and refer to materials and methods section for details). The set of features that load significantly on a given Factor was used to infer the underlying phenotypic attributes associated with that Factor. Figure 2C shows a representative polar plot of loadings versus cytological features for Factor 1. Complete factor structure and underlying phenotypic traits are outlined in Figure 2D and representative images in Supplementary Figure 1. Note the order of numbering the factors is based on the extent to which a given factor accounts for the common variance in the whole dataset.

Factor 1, which accounts for the majority of the common variance, loads highly on 12 features all of which describe the size of the nucleus. Examples of these features include: Area-Ch1, TotalIntensity-Ch1, Length-Ch1, and Width-Ch1. Based on this loading pattern we conclude that this is a Nuclear Size Factor. Thus, the most information rich phenotypic characteristic given our labeling and imaging strategy is the size of the nucleus and the quantity of DNA. Factor 2 loads primarily with four features that describe the extent of EdU probe incorporation. Hence, Factor 2 is a DNA Replication, or S-phase Factor. Factor 3 loads primarily with features that describe DNA concentration, and thus condensation (e.g., AvgIntensity-Ch1), and phosphoH3 intensity (e.g., AvgIntensity-Ch2), and is thus a mitosis and chromosome condensation factor.
Factor 4 is loaded significantly by four features that refer to the shape contour of the nuclear perimeter and is thus a nuclear morphology Factor. Factor 5 loads with four features that describe Ch2 texture, i.e., the morphology of EdU incorporation. It is statistically distinct from factor 2, and must report on some particular aspect of DNA replication, such as early vs. late S-phase. Factor 6 reports mainly on nuclear shape. Taken together, we reduced a dataset of 36 measured cytological features from \( \sim 10^6 \) cells (~7 GB) to six common underlying Factors scored for \( \sim 10^4 \) wells (~3 MB). Moreover, these common underlying Factors reflect a set of orthogonal phenotypic attributes that account for almost all of the co-variance relationships exhibited in the image-based cytological features measured on each cell in our assay.

**Factor-Based Phenotypic Compound Profiling**

We used our high content image assay to screen and profile a library of 6547 compounds derived from a diversity library (21%), a natural products library (58%), and a library of known bioactive compounds (21%); and all compounds were assayed in duplicate at a single dose of 10\( \mu \)M for 20 hours (Figure 3A). Dose response studies using a panel of known cytotoxic compounds with diverse mechanisms of action indicated the appropriateness of these dose and time conditions for phenotypic profiling (Supplementary Figures 2 and 3). Based our the six-Factor model we used regression to estimate scores for each Factor (i.e., Nuclear Size, Replication, Mitosis, Nuclear Morphology, EdU Texture, and Nuclear Ellipticity) on a cell-by-cell basis for each treatment. We summarized each compound treatment effect as the mean score on each of the six factors (i.e., a well average).
Our compound library is expected to contain multiple bioactive compounds with various distinct targets and mechanisms of action, and consequently to generate unique phenotypic read-outs on the six orthogonal Factors. To score the strength of phenotypic perturbation independent of precise phenotype, we computed the Euclidean distance between each compound and the average control (untreated) phenotype for a composite vector consisting of all factor scores for that compound. This Euclidean distance metric projects the multidimensional phenotype onto a single phenotypic response dimension, and allows us to call “hits” independent of their exact phenotype.

We defined hits as compounds whose phenotypic response (i.e., distance) was in the top 5% in both replicate experiments, resulting in 211 compounds or ~3% of the total screening set. Our hit set was enriched in compounds derived from the library of bioactive compounds (Figure 3B). This enrichment was most pronounced when we examined the strongest bioactive compounds in the top 1% distance group. In this set, 48% of the compounds were derived from the bioactive library, compared with 21% in the entire screening set. This indicates our strategy is effective at identifying compounds with substantial biological activity. We observed a generally good correspondence between the two replicate experiments (Figure 3C).

We next profiled the biological activity of the hit compounds, using unsupervised hierarchical clustering of the factor scores. This revealed seven primary clusters, that we will term “phenotypes”, that are shown in Figure 3D along with a bar denoting whether they are in the strongest (green) or second strongest (blue) category in terms of overall strength of the phenotype. We can begin to interpret these phenotypes by looking at how the factors change, and also where compounds with known biological activity are positioned (discussed below). For example, phenotypes 1 and 2, with high chromosome
condensation, correspond in large part to mitotic arrest, phenotype 4 with generally high chromosome condensation but also decreased nuclear size, corresponds in large part to apoptosis, phenotype 5 with increased DNA replication, and phenotypes 6 and 7, with increased nuclear area, decreased DNA replication, and decrease chromosome condensation, probably correspond to cell cycle exit in G1, that is generally understood to increase nuclear cross-sectional area. The strongest hits in our screen (green bars) mostly affect mitotic progression and cell survival, while weaker hits (blue bars) appear to block cell cycle progression via a G1 arrest. This difference in phenotypic strength presumably reflects the more dramatic cytological changes associated with mitosis and death, rather than differences in compound potency.

Comparison to metrics of chemical similarity

Compounds with similar structure have similar function\(^9\), and quantitative structure-activity relationships (SARs) are at the heart of drug discovery. As a step towards phenotype based SARs, we asked whether our phenotypic clustering grouped together structurally similar compounds. For each compound we defined a circular molecular fingerprint using ECFP\(_4\) descriptors that define molecular structure using radial atom neighborhoods (see materials and methods). We computed a similarity matrix based on Tanimoto similarities that describes the relationship between each of the 211 compounds in our hit set. Analogously, we generated a cosine distance based phenotypic similarity matrix using our Factor-Based phenotypic profiles. These matrices, displayed as heat maps, are shown side by side in figure 4. The compounds are ordered by phenotypic similarity using unsupervised clustering, so the 7 primary phenotypes appear as blue boxes on the diagonal in the biological space panel.
In the chemical space side, we observed multiple blocks of structurally similar compounds, that correspond to phenotypes 1, 2, 6 and 7. The blocks of chemical similarity were smaller than the phenotype blocks, because only a subset of compounds causing a given phenotype are similar, and in some cases, multiple blocks of chemical similarity were observed for a given phenotype, especially phenotype 6. These clusters evidently reflect regions where biological effects are dominated by distinct structural compounds classes. The relationship between phenotype space and chemical space we observed in figure 4 is perhaps expected, but has not been visualized before in such quantitative detail.

To quantify the extent to which phenotypic clustering of the active compounds groups structurally related compounds, and to determine if this structure-function concurrence is beyond which would be expected by random chance, we determined the Spearman correlation coefficient for rank ordered phenotypic similarities and the corresponding compound similarities using the matrices from Figure 4. We found an overall modest positive correlation (correlation = 0.0746), presumably reflecting strong correlation within small clusters, and lack of correlation elsewhere. We then generated 1000 random compound similarity matrices by randomized sorting, computed the Spearman correlation coefficient with the phenotypic ordering, and used this to ask if the observed correlation was statistically significant (Supplementary Figure 4). This analysis indicates that the observed correlation is significant (p<0.001) and approximately two-fold above that maximum chance observation. Thus, whereas this analysis comprises both structurally similar and structurally dissimilar compounds, the significance of the association between compound structure and function suggests that the molecular similarity principle holds for our phenotypic compound profiling.
In light of emerging evidence that the molecular similarity principle might not always hold true\(^{30}\) we sought to understand the extent to which small changes in structure are associated with large changes in function, e.g., activity cliffs. To address this concept we compared Tanimoto similarities with phenotypic distance between each compound pair in our screening set. Due to the large number of comparisons we focused our analysis only those comparisons in which at least one compound in a pair was active, and examined phenotypic distance for those compound pairs that exhibit high structural similarity (Tanimoto score \(\geq 0.3\)). Our analysis reveals that approximately 96% of the compounds with significantly similar structure exhibit significantly similar phenotypic readouts (Figure 4b, green box). Alternatively, of the structurally similar compounds active in our assays, only 4% exhibit significant phenotypic divergence (Figure 4b, red box). To understand this divergence further we examined a pair of Scoulerine-related compounds more closely (Figure 4b and Supplementary Figure 6). These two compounds have high molecular similarity (top 0.1% based on similarity) and differ essentially by the presence methoxy or hydroxyl groups (Supplementary Figure 6). Interestingly the compound pair exhibits significantly different phenotypes (top 1% based on phenotypic distance), and this functional divergence is consistent with recent structure-activity studies on the two compounds\(^{31,32}\). Taken together we conclude that activity cliffs do emerge in our phenotypic screen. But, they represent the minority of cases. We are therefore more likely to observe phenotype concordance for structurally similar compounds.

Examples of phenotypic structure-activity relationships (SARs)
We chose several local SARs to examine in more detail, indicated by black bars adjacent to the compound structure similarity matrix in figure 4. Figure 5A shows a sub-cluster that falls within phenotype 4. This sub-cluster is characterized by decreased Nuclear Size, Replication, and EdU texture scores; and increased nuclear morphology score. Unlike the majority of compounds in cluster 4 this sub-cluster does not exhibit a substantial increase in chromosome condensation. Thus, these compounds, though apparently cytotoxic, generate a phenotypic cytotoxicity signature distinct from classic apoptosis. This sub-cluster is enriched in antibiotic compounds that have known cytotoxic effects in mammalian cells. A small structural cluster contained three cyclic hexadepsipeptides, including Aurantimycin and Diperamycin, derived from strains of *Streptomyces*[^33]^[34]. These showed strong phenotypic similarity to a structurally divergent Lysolipin derivative. The second region of local structural convergence within this phenotypic cluster contains several cyclic non-peptide compounds. This includes Kendomycin, an antibiotic with a C-glycosidic core and previously reported mammalian cytotoxicity and endothelin receptor antagonistic activity[^35]; as well as other cyclic compounds that include two antibiotic of the Concanaamycin class with cytotoxic activity and which are potent inhibitors of vacuolar ATPases[^36]. We also note that this phenotypic sub-cluster contains a region with structurally distinct cytotoxic / antibiotic compounds, including Heptelidic Acid Chlorohydrin[^37]. The phenotypic similarity of all these compounds presumably reflects a common target at the level of protein or pathway, that may be vacuolar ATPases or other proteins that function in related areas of vesicular trafficking[^36].

Figure 5B shows sub-cluster within phenotype 6 with high structural convergence. This subcluster is characterized by increased nuclear area and ellipticity,
but decreased DNA Replication, Chromosome Condensation, Nuclear Morphology and EdU texture. It contains eleven corticosteroid compounds with significant structural similarity including Clobetasol-17-Propionate, Dexamethasome, and Triamcinolone. Corticosteroids are known to cause a cell cycle arrest during G1, validating our interpretation of the parent Cluster 6 as a G1 arrest phenotype. However, the local grouping of highly structurally similar compounds within this subcluster indicates a corticosteroid-specific G1-arrest phenotype. Such discrimination is surprising given our choice of cell types and fluorescent probes, and indicate the power of relatively subtle morphology descriptors, such as nuclear shape metrics, to report on biological activity.

Figure 5C shows a larger phenotypic subcluster within phenotype 7, which has various effects on Nuclear Size, and a persistent decrease in DNA Replication and EdU Texture, consistent with a cell cycle arrest. Within this subcluster we find two groups of structurally similar compounds separated by a region of structurally distinct compounds. The two groups display a significant degree of intergroup similarity, presumably because they share a steroid, or steroid-like, structure. The first group contains three cardiac glycosides, Ouabain, Digitalis, and Digoxin. These are well known inhibitors of Na/K pumps, and have be shown to inhibit Topoisomerase I in mammalians cells at nanomolar concentrations. At high doses, cardiac glycosides cause a large drop in intracellular potassium levels leading to an inhibition of protein synthesis. The protein translation inhibitor Emetine is also present within this cardiac glycoside subcluster, suggesting that the protein translation inhibition mechanism of action of these compounds may dominate their phenotypic effect in our assay. Supporting this interpretation, the non-structurally-related translation inhibitor cycloheximide shares this phenotype. The second group of related compounds contains
a set of steroid hormones including progesterone and Danatrol. Progesterone signaling is known to result in growth arrest in G0/G1\textsuperscript{43,44}.

**Integration with Ligand-Target Knowledge Space**

Our observation that multiple distinct structural classes of compounds can produce similar phenotypes, even at our highest phenotypic resolution, could be explained by compounds perturbing common targets via the same, or different binding sites, or by compounds perturbing different components of common pathways. We investigated this possibility by implementing a structure-based target prediction method that has recently been reported\textsuperscript{22}. Statistical models of substructural features were combined with an annotated chemogenomics database (WOMBAT) that associates ligand molecular structures with their cognate biological targets. We used these “known” ligand-target associations to train a Naïve Bayes model that we subsequently employed to predict the targets of our 211 active compounds. Using the top five most probable targets for each compound, we examined the extent to which phenotypic clustering of all the active compounds groups their cognate predicted targets. Notably, we found an increased positive correlation (correlation = 0.136, \(p<0.001\), Supplementary Figure 5) between phenotypes and targets. This is twice the strength in correlation compared with the phenotype to structure comparison, and indicates that the observed divergence in structure-activity relationships can, in part, be accounted for by structurally different compounds having common targets.

Although our results above point to the effectiveness of the target prediction method, in fact, predicting ligand-target association is an imperfect art. Thus, comparisons with the more robust phenotype and chemical similarity measures must be
treated with caution. To provide a sense of its potential utility in pointing to a particular
target, we illustrate results from a subcluster from mitotic arrest phenotype 1, that is
primarily characterized by high chromosome condensation. Within this cluster we
observed four distinct groups of structurally related compounds. The first, second, third,
and fourth groups are characterized by a Colchicine derivative, a set of novel kinase
inhibitors, a Quinoline derivative, and a PseudoLarix Acid B derivative. Our
substructure-based method predicted multiple targets for each compound. We focused
only on the top five targets, and for visualization purposes plot only those targets that are
predicted at least twice within the phenotypic subcluster (Figure 6A). We find that a
majority of all the compounds are predicted to target tubulin (7 out of 13), and as a
consequence should affect mitotic spindle integrity. Additionally, the distinct group of
novel kinase inhibitor compounds is predicted to hit both CDK1 and CDK2. Colchicine
is a well known inhibitor of microtubule dynamics, binding a distinct pocket within
tubulin and causing depolymerization\textsuperscript{45}, it is predicted that this derivative we found
would have similar effects in cells. Several Quinoline derivatives, including this one\textsuperscript{46},
have been shown to also depolymerize microtubules via tubulin interactions\textsuperscript{47}, and
Pseudolarix B has been recently shown to affect tubulin polymerization through a binding
site distinct from the Colchicine pocket\textsuperscript{48}.

To gain mechanistic insight, we examined cytoskeletal morphology and cell cycle
profiles for the set of putative tubulin targeting compounds. We used
immunofluorescence microscopy to detect $\alpha$-tubulin in cells treated with each compound
at the screening dose. As predicted we observe depolymerization of microtubules and
mitotic arrest in cells treated with each of the Colchicine, Quinoline, and PseudoLarix
Acid B derivatives (Figure 6B). Thus integration of compound structure with knowledge
based ligand-target predictions reveals that similar phenotypes produced by different compounds can, in part, be accounted for by targeting different components of common pathways, and by compounds hitting common targets via different binding sites. Moreover, our results indicate that phenotype and predicted targets constitute a useful SAR pair that can overcome the limitations of chemical similarities.

Discussion

In this paper we introduce Factor Analysis as a method to mine HCS data for quantitative phenotypic profiles. Factor analysis was developed more than century ago in the field of psychometrics and it continues to be applied across many diverse fields of science\textsuperscript{10-15}. Compared to other recent efforts to develop phenotypic profiles from HCS data\textsuperscript{49-51}, Factor Analysis had two main benefits. It drastically reduced the size of the dataset early in the data mining process, and it reported phenotypes in terms of six factors with interpretable biological meaning. These benefits were achieved while retaining most of the information in the primary data, as evidenced by the statistical criteria that were used to determine that six factors were sufficient to effectively account for the common variance in the cytological data (Figure 2B). It is possible that Factor Analysis might neglect some subtle effects that could be revealed by more exhaustive methods, but because it is robust and easy to implement with commercial statistics software, it is well suited for routine use in drug discovery.

Other dimensional reduction strategies can be used to analyze HCS data, notably principal component analysis\textsuperscript{50}. Principal component analysis and Factor analysis are similar in their goal of mining interpretable information from high-dimensional data. Yet philosophically and operationally they are different\textsuperscript{52}. Principal component analysis
seeks to reduce the dimensionality of a multivariate data set into a small number of
dimensions that maximally accounts for the total variance. Factor analysis seeks to
account only for the common variance, which is regarded as that variance shared among
variables, and excludes the specific and error variances. In principal component
analysis, the components are modeled as linear combinations of the measured variables.
In factor analysis the measured variables are modeled as linear combinations of the latent
underlying Factors. We have chosen Factor analysis as it emphasizes identifying
interpretable dimensions, or metrics, in phenotype space. Profiling is possible without
using interpretable phenotypic dimensions, but in this case compounds can only be
classified by comparison to each other. Profiling using interpretable phenotypic
dimensions, such as our factors 1-6, enable hypothesis generation based on biological
effects as well as compound classification (see results section, figs 4-5).

One limitation of our study was the use of a single compound concentration and a
single time point. Following phenotype across a range of concentrations and times would
certainly produce more mechanistic information and could perhaps facilitate more precise
mechanism of action inferences in certain cases, but at the cost or requiring a lot more
data collection. Factor analysis could be readily extended to such higher dimensionality
datasets, for example by implementing a titration-invariant similarity score\textsuperscript{49} for data
reduction of concentration-dependent effects.

The phenotypic profile we generated using Factor Analysis can be compared to
other data-rich methods, such as mRNA expression profiles of drug treated cells\textsuperscript{53}, or
proteomic methods. Profiles based on HCS cytology are, perhaps, less rich in detailed
information than some \textquotedblright-omic\textquotedblright methods, but much cheaper to acquire; so profiling
thousands of compounds is feasible. Expression profiling shares with HCS the challenge
of analyzing very large datasets. Recently, a Factor analysis of genome-wide expression
data was shown to have both statistical and computational benefits compared with
existing classification schemes for the prediction of gene function. Profiling methods
that generate profiles by combining multiple cell-based pathway readouts in image-based
protein complementation assays are comparable to standard high-content screening in
content and expense, and are likely amenable to Factor analysis. Different phenotypic
profiling technologies can provide orthogonal information, and it will be useful to
combine them to profile compounds early in the drug discovery pipeline.

The central goal of our study was to investigate structure-activity relationships by
integrating phenotypic information from HCS with chemical knowledge from profiles of
chemical similarity and predicted targets. Such integration would be a powerful tool in
drug discovery. This is not a novel concept, but it has been difficult to achieve at a
practical level, in part because we lack conceptual frameworks for integrating high-
dimensional biological and chemical data, and in part because high dimensional datasets
of biological activity (e.g., microarray data) are typically too expensive to acquire across
a large number of compounds. Figures 4-6 represent considerable progress on the
integrated structure-activity problem, using easy-to-adopt methods. The two chemical
knowledge profiles we use, structural similarity (figures 4-5) and target predictions
(figure 6) differ considerably in their rigor and degree of development, with the former a
well established science, and the latter more of a ongoing challenge of computation
chemists than a practical reality. Thus, our goals in comparing them to phenotypic
profiles were somewhat different in the two cases. In the case of structural similarity, we
knew that clusters of compounds that were related by phenotype and chemistry should
exist in our library, and we used the comparison with phenotypes to find them, and to
examine them in detail to uncover new mechanistic information (figure 5). In the case of target prediction, we used the phenotype data to test how well the prediction algorithm was working, and also to point to one particular target (figure 6). Our analysis revealed that phenotypes correlate better with predicted compound targets than with the compound structures themselves (Supplementary Figure 5). This result provided support for both the effectiveness of the target prediction model and for the idea that different ligand-target interactions account, in part, for divergence in compound structure activity relationships.

Concordance between phenotypic and structural similarity profiles revealed the capability of HCS combined with Factor analysis to make subtle phenotypic distinctions. For example, we readily discriminated the effects of corticosteroid-like and progesterone-like steroids, even though both cause cells to stop proliferating in G0/G1 (figure 5B,C). The subclustering of cytotoxic compounds in figure 5A illustrates even finer phenotypic resolution. Obtaining this degree of distinction of therapeutically relevant mechanisms using a generic cancer cell line and cell cycle probes is remarkable, and attests to the large amount of information that can be derived from microscope images when appropriate mining methods are implemented. Use of primary cells and more disease-relevant probes should further increase the resolution in areas relevant to drug discovery.

Lack of concordance between phenotypic and chemical similarity profiles is illustrated in the cytotoxicity cluster 4. One can envision cell death as a phenotypic end-point for multiple stress pathways that can be invoked by a variety of pharmacologic perturbations. In this regard we observe multiple distinct compound classes appearing within the cytotoxicity cluster, and consequently minimal correlation between structure and phenotype when examined a low phenotypic resolution, i.e., the cluster as whole.
However, when examined at higher phenotypic resolution we can discriminate multiple small groups of structurally related compounds within which we observed highly similar cytotoxicity signatures, for example the cyclic hexidepsipeptides versus the cyclic non-peptide antibiotic compounds (Figure 5A). This indicates that even at the end-point phenotype of cell death observed at a saturating dose we can still generate meaningful structure function relationships.

Computational ligand-target prediction enabled us to demonstrate that by mapping compound structures to targets we improve our ability to discover meaningful structure-activity relationships based on cytological phenotype (supplementary figure 3). Furthermore, our data provide quantitative support to a, perhaps, logical explanation for divergence in structure versus phenotype concordance. To test the effectiveness of the target prediction method at higher phenotypic resolution we looked closely at the predicted targets for four groups of phenotypically similar, yet structurally distinct compounds. Our computational prediction pointed to tubulin as a common target for three of these groups, and our phenotypic data and follow-up experimental work supported this prediction (Figure 6). Ligand-target prediction also revealed multiple highly probable targets that appear within each of four structural groups. Thus, parallel activity on these additional targets could account for subtle phenotypic differences between groups. Taken together, our results show that the combination of cytological phenotypes can improve confidence levels in target prediction both globally, as in our active compound set, (supplementary Figure 3) and with respect to specific targets (Figure 6). Thus quantitative cytological phenotypes, such as those derived here, may represent a new set of compound descriptor data that could be included directly into computational models to bolster compound-target prediction efficiency.
Despite progress on analysis of HCS data, reported here and elsewhere \textsuperscript{51,56}, the use of cytology to profile phenotype in a broad and quantitative manner is still its infancy. We believe the potential is enormous. For example, new markers could be implemented that enable predictive toxicology of active lead compounds. Combined with chemical structure knowledge and ligand-target prediction, as shown here, such approaches could provide detailed mechanistic insight to help guide medicinal chemists early in the lead optimization process. Dealing with complexities of predictive toxicology will require breakthroughs in cytological image analysis, target prediction schemes, and data mining. Our integration here of image-based cytological phenotypes with chemical structure and computational ligand-target prediction represents a step forward in solving this and other difficult drug discovery problems.

**Acknowledgements**

We thank Leah Martell, Mathis Thoma, James Nettles, Brian Dwyer, and Michelle Pflumm for insightful comments and discussions, Craig Mickanin and ShanChuan Zhao for automation support, and Quan Yang for database support. DWY and AB are both Novartis Presidential Postdoctoral Fellows. Work in the TJM lab supported by NIH Grant CA78048.

**Competing interests statement:** The authors declare that they have no competing financial interests.
Reference List

   Ref Type: Journal

   Ref Type: Journal

   Ref Type: Journal

   Ref Type: Journal

   Ref Type: Journal

   Ref Type: Journal

   Ref Type: Journal

   Ref Type: Journal

   Ref Type: Journal

    Ref Type: Journal

    Ref Type: Journal

    Ref Type: Journal

Ref Type: Journal

Ref Type: Journal


Ref Type: Journal

Ref Type: Journal

20 "WOrld of Molecular BioAcTivity (WOMBAT), Available From Sunset Molecular Discovery LLC," in 2007).

21 "PipelinePilot 5.1, Available From Scitegic," in 2007).

Ref Type: Journal

Ref Type: Journal

Ref Type: Journal

Ref Type: Journal


Ref Type: Journal

Ref Type: Journal

Ref Type: Journal

Ref Type: Journal

Ref Type: Journal

Ref Type: Journal

Ref Type: Journal

Ref Type: Journal

Ref Type: Journal

Ref Type: Journal

Ref Type: Journal
Ref Type: Journal

Ref Type: Journal

Ref Type: Journal

Ref Type: Journal

Ref Type: Journal

Ref Type: Journal

Ref Type: Journal

Ref Type: Journal

Ref Type: Journal
Figure 1: High Content Screen

Hela cells were grown in 384 well optical plates for 24 hr prior to compound treatment. Compounds were delivered in an automated manner for a final concentration of 10uM and incubated for approximately 20hrs. Cells were then pulsed for 40 minutes with 500nM 5-ethynl-2’-deoxyuridine (EdU) to label sites of nascent DNA replication (Yellow), followed by fixation in formaldehyde. Rhodamine-azide was conjugated to EdU by click chemistry. Cells were immunolabeled with rabbit anti-phospho-histone H3 Ser10 (pH3) and a Cy5 conjugated goat anti-rabbit secondary antibody (Red). DNA was labeled with Hoechst Dye (Blue). Automated fluorescence microscopy was carried out using a Cellomics Arrayscan, and images were collected with a 10X PlanFluor objective. Individual cell segmentation based on the DNA stain (cytological mask) and quantification was performed using the Cellomics Morphology Explorer algorithm, and 30 cytological features (Supplementary Table 1) were determined for each cell on DNA (Ch1), pH3 (Ch2), and EdU (Ch3) channels. Features were collected for at least 500 cells per well (treatment).

Figure 2: Common Factor Model Defines a Multidimensional Biological Activity Space

A. High content data are contained in an n x m matrix, X consisting of a set of n image-based cytological features measured on m cells. The common factor model maps the n-cytological features to a reduced k-dimensional space described by a set of factors, F that reflect the major underlying phenotypic attributes measured in the assay. The loading matrix L defines the relationship between the measurements in X to the underlying common factors. The diagonal matrix, ε is a matrix of specific variances (see materials
and methods section for details). **B.** The dimensionality of the factor space is determined by an eigenanalysis of the correlation matrix of the data matrix, \( \mathbf{X} \). Prior communality estimates were established as the square multiple correlations of each of the 36 cytological features with all other features, and final communalities were determined from the estimated loading matrix. This dimension \( k \) is determined by Kaiser criterion to be equal to the number of factors with variance greater than unity. Using this criterion we determine that there are 6 significant factors. **C.** After fitting the \( k \)-model, the loading matrix, \( \mathbf{L} \) was rotated to maximize the variance in factor loadings. Supplementary Data File 1 outlines the loadings of each feature on the six common Factors for the unrotated and varimax rotated loading matrices. The loadings, \( \mathbf{L} \) reflect the correlations between cytological features and the common underlying factors. We used polar plots to visualize these loading patterns, and interpret the biological meaning of the underlying factor. All polar plots are included in the supplementary materials. Shown here is the loading pattern for Factor 1 as an example. The first Factor is highly correlated with 12 features all of which describe the size of the nucleus, examples of these features include: AreaCh1, TotalIntensityCh1, LengthCh1, and WidthCh1. Based on this loading pattern we conclude that this is a Nuclear Size Factor. See Supplementary Data File 1 for complete factor model fit details. **D.** The complete factor structure is shown in this schematic. Each of the six factors are drawn with lines connected to the cytological features with which they are most significantly correlated. Our interpretation of the phenotypic attributes characterized by each Factor is shown on the right.

**Figure 3: Screen Layout and Phenotypic Compound Profiling**
A. We screened a 6547 compounds from three libraries that include both natural and synthetic compounds from a diversity set, a natural products library, and a set of known bioactive compounds. Our screen was performed in two replicate experiments. We established a Factor-based phenotypic response metric that reflects the distance in factor space from a treatment to the untreated control population. This metric projects the multidimensional phenotype onto a single response dimension, enables a standard comparison between compounds with various bioactivities, and facilitates hit identification independent of the specific phenotype. Hits were defined as compounds in the top 5% based on phenotypic response in both replicate experiments. This filter criteria results in 211 bioactive compounds or ~3%. B. Pie charts indicating the fraction of each library in our screening set and hits set. We observe an enrichment of known bioactive compounds in our hit collection. C. A scatter plot comparing the factor-based phenotypic response from both replicate experiments. Compounds in the top 2-5% are colored blue, the top 1% are green, and non-hits are gray. D. We performed hierarchical clustering of mean factor scores for each of the 211 hits compounds. Clustering is based on Ward’s linkage criteria and the half Euclidean distance metric. The position of compounds within the top 1% and 2-5% based on phenotypic response is shown. (-1=blue, +1.5=red)

Figure 4: Similarity in Biological Activity is Correlated with Similarity in Compound Structure

A. We examined the relationship between phenotypic profiles (biological activity space) and compound structure (chemical space). We generated a phenotypic similarity matrix for that each compound with each of the remaining 210 compounds. Similarity between
two compounds is determined by the cosine distance between their respective phenotypic vectors. Phenotypic vectors consist of the six mean factor scores for each compound. Analogously, we determined the compound structure similarity matrix comparing each and every compound based on the Tanimoto similarities in compound structure vectors. Compound structure vectors are ECFP_4 fingerprints (see materials and methods). The similarities are organized in 211x211 symmetric matrices that are ordered based on the hierarchical clustering in Figure 3D, and the corresponding dendrogram from Figure 3D is shown. A heatmap is applied to both the phenotypic (black-to-blue) and the compound structure (black-to-yellow) similarity matrices. The colorbars are shown for each, and the scale was selected such that similarities at or below the 75\textsuperscript{th}-percentile are black and are maximally colored (blue or yellow) at the 99\textsuperscript{th}-percentile. Percentiles were established based on distributions consisting only of off-diagonal similarity values. Four black bars adjacent to the compound structure similarity matrix reflect the positioning of the subclusters displayed in Figures 5 and 6. From the top the first bar corresponds to Figure 6, while the second, third and fourth bars correspond to Figures 5A,B, and C, respectively.

B. We assessed the extent of structure activity concordance and discordance. We computed Tanimoto similarities (as above) between each pair of compounds in our screening set, and computed phenotypic distance between each pair of compounds using the Euclidean distance metric. We then compared the Tanimoto similarities with phenotypic distance in a scatter plot. Due to the large number of comparisons we focused our analysis only those comparisons in which at least one compound in a pair was active in our assay, and for illustration purposes plot a 10\% random sample of the entire similarity/distance data set. Compound pairs that exhibit high structural similarity (Tanimoto score ≥ 0.3) and low phenotypic distance (Euclidean
distance \( < 1 \) are considered to exhibit structure-phenotype concordance. Our analysis reveals that approximately 96% of the compound pairs with high structural similarity fall in this class (green box). In contrast, compound pairs with exhibit high structural similarity (Tanimoto score \( \geq 0.3 \)) and high phenotypic distance (Euclidean distance \( \geq 1 \)) are considered to exhibit structure-phenotype discordance (e.g. activity cliffs). Approximately, only 4% of the compound pairs with high structural similarity exhibit fall in this class (red box). The red data point identifies the location of the scoulerine-related compound pair shown in Supplementary Figure 6.

**Figure 5. Factor-Based Phenotypic Profiling Elucidates Structure-Activity Relationships in Biological Activity Space.**

We examined the relationships between clusters containing similar phenotypic profiles and their corresponding structural similarities and member compounds. Three examples are shown here. Factor-based phenotypic profiles and subcluster dendrograms from figure 3D are shown. The heat-map is maximally blue at or below a standardized factor score -1.5 and maximally red at or above a standardized factor score 1.5. The corresponding submatrices from the compound structure similarity matrix are also shown with the identical color map from figure 4. Maximally yellow indicating high similarity and black indicating low similarity. Structures are shown for several member compounds, and the position with the clusters is indicated by number. **A.** Subcluster of compounds that result a cell death phenotype that is generally characterized by low factor 1 (nuclear size) and increased factor 3 (chromosome condensation). The subcluster contains several antibiotic compounds in particular two cyclic depsipeptides with known cytotoxic activity, Aurantimycin A (#78) and Diperamycin (#79). We also observed the cytotoxic antibiotics Heptelidic Acid (#81) and Kendomycin (#83). **B.** Subcluster of
compounds that result in a G1-arrest characterized in general by large nuclear size (factor 1) and low DNA replication and mitosis (Factor 2 and 3). This subcluster consists mainly of corticosteroids, for example Clobetasol-17-propionate (#141), Dexamethasone (#143), and Triamcinolone (#152). C. A larger subcluster phenotypically dominated by low DNA replication, mitosis, EdU texture and average to high nuclear size. The top portion of the cluster contains several cardiac glycosides known to affect Na/K pumps, including Digitoxigenin (#154), Ouabain Octa-Hydrate (#157), and Digoxin (#158). The subcluster also contains two known protein translation inhibitors Emetine (#155) and Cycloheximide (#161). The lower portion contains several steroid hormones including Progesterone (#165) and Danatrol (#169).

Figure 6 – Factor-Based Phenotypic Profiling Provide Biological Support to Structure-Based Target Predictions

A. We show a mitotic subcluster. Factor-based phenotypic profiles and subcluster dendrograms from Figure 3D are shown (-1.5=blue, +1.5=red). The corresponding compound structure similarity submatrix is also shown with the identical color map from Figure 4. (Black=Low Similarity, Yellow=High Similarity). Structures are shown for several member compounds, and the position with the clusters is indicated by number. We predicted targets for each compound as described in the materials and methods. Blue boxes identify related predicted target with corresponding compound. Only genes encoding proteins that are targeted by two or more compounds within the cluster are shown. B. The partial structures of three representative compounds are shown a Colchicine derivative (#39), a Quinoline derivative (#44), and a PseudoLarix Acid B derivative. We show images of cells treated with compound for 20 hours and stained for
DNA by Hoescht Dye and the predicted target α-tubulin. Cell Cycle profiles determined from HCS images using a decision-tree based classification scheme described elsewhere are shown for each compound. Images and profiles of control cells with normal phenotype are shown.

**Supplementary Figure 1**

To establish the biological relevance of the six Factors we examined images of cells scoring both at the extreme high and extreme low ends of each Factor. In this analysis we observed that Factor 1 is proportional to nuclear size and DNA content. High scoring cells on this Factor have large nuclei, and typically classify as late S-phase, G2, and prophase. Some extreme outliers were in fact two nuclei in juxtaposition that had not been segmented (data not shown). Cells scoring low on Factor 1, had smaller nuclei or appear to be apoptotic bodies. Examination of Factor 2 reveals that the EdU texture parameter is a good indicator of S-phase entry and S-phase exit. High values are associated with no EdU incorporation whereas, extreme low values are associated with low levels of EdU staining. Intermediate values were associated with higher replication labeling. Factor 3 is a strong indicator of S-Phase, where as, Factor 4 is a strong indicator of mitosis. As anticipated Factor 5 characterizes nuclear morphology, with high scoring cells having abnormally shaped nuclei, and low score cells having classic round nuclei. Cells scoring high on Factor 6 exhibit an oblong elliptical cross-section relative to the image plane and low scoring cells have more circular nuclear shape.

**Supplementary Figure 2**
We performed a phenotypic dose-response analysis of classic cytotoxic compounds. Factor-based dose-response relationships for cytotoxic compounds across serial dilutions ranging from 10µM to 0.70pM for a 20hrs treatment period. The response is a phenotypic response metric (see methods section). Logistic regression was used to fit sigmoidal dose-response profiles for each compound and were plotted in a heatmap format using MATLAB. As an example, the Microtubule poison, Nocodazole and the protein translation inhibitor, Emetine exhibit similar factor-based EC50 values (~120nM) (left panel). Nocodazole and Emetine dose-response profiles for each of the six orthogonal Factors are shown (Right Panel). Emetine treated cells exhibit a dose dependent increase in nuclear size, with concomitant decreases in EdU texture, DNA replication, chromosome condensation, and nuclear morphology scores. Nocodazole results in a prominent dose-dependent increase in chromosome condensation with decreases in all other factors (Blue=Low, Red=High).

**Supplementary Figure 3**

In order to validate our method we examined the extent to which cytotoxic compounds with similar biological activites exhibit similar factor-based phenotypic profiles. We performed hierarchical cluster analysis on Factor scores using data from the maximum dose (10µm) for each one of our panel of compounds. The dendrogram reflects the emergent hierarchical structure (left) and for illustration purposes the panels reflecting dose responses for constituent factors are shown for each compound. Two main clusters emerge that can be broadly classified based on compounds that result in G2 and mitotic arrest and those that result in a G1-S arrest (light gray vertical bars). Importantly, we find that compounds that cluster together target similar biochemical
processes (dark gray vertical bars). Notably, we find clusters containing: Emetine and MG132, which are both well known inhibitors of protein metabolism; Camptothecin and Etoposide, which both effect topoisomerases; and Nocodazole and Taxol, which both affect microtubule dynamics. This analysis reveals that compounds that effect similar cellular process exhibit similar Factor-based phenotypic profiles and that these phenotypic similarities can be elucidated, quantitatively, at a single saturating dose.

**Supplementary Figure 4**

We determined if the observed visual correlation between the phenotypic similarity matrix and the compound structure similarity matrix was statistically significantly. We determined the Spearman correlation coefficient for rank ordered phenotypic similarities, and compound similarities using the original matrices from Figure 3 (correlation = 0.0746). We then generated 1000 random compound similarity matrices, by randomizing the positions of off-diagonal similarities. For each random similarity matrix we compound the spearman correlation coefficient. This scatterplot shows both the original correlation and the correlations between the phenotypic similarity matrix and the 1000 random generated compound similarity matrices. **B.** The original compound structure similarity matrix and example random similarity matrix are displayed as a heatmaps. Colorbar reports the degree of similarities, values at or below the 75%-percentile in off-diagonal similarities are black. Values are increasingly yellow up to the 99%-percentile in off-diagonal similarities.

**Supplementary Figure 5**
A. We determined the correlation between the phenotypic similarity matrix and the compound target similarity matrix was statistically significantly. Top five compound targets based on Bayes Score (see materials and methods) were used to construct a similarity matrix based on the Tanimoto similarity score. We determined the Spearman correlation coefficient for rank ordered phenotypic similarities, and compound target similarities using the original matrices (correlation = 0.136). We then generated 1000 random compound target similarity matrices, as in supplementary figure 2. For each random similarity matrix we compound the spearman correlation coefficient. This scatterplot shows both the original correlation and the correlations between the phenotypic similarity matrix and the 1000 random generated compound target similarity matrices. For comparison, the correlation results for compound structure similarities (correlation = 0.0746) and corresponding random matrices are shown. B. The original compound target similarity matrix and example random similarity matrix are displayed as a heatmaps. Colorbar reports the degree of similarities, values at or below the 75%-percentile in off-diagonal similarities are black. Values are increasingly green up to the 99%-percentile in off-diagonal similarities.

Supplementary Figure 6

Structures are shown for two Scoulerine-related compounds identified as an activity-cliff pair in the analysis described in Figure 4B (red data point). A. The compound is a known antagonist of the D2 receptor (QSAR 18, 4, 1999, p. 354). B. The drug Scoulerine, which is derived from poppy seeds, is used as a sedative and binds to a series of GPCR receptors, such as the alpha andrenergic receptors, GABA, 5HT receptors, and the D1 a D2 dopamine receptors [REF].
Figure 1 Young DW et. al,

Automated Compound Treatment
24 Hour Incubation Time

Cell Culture - 384 well Optical plates

Automated Microscopy, Image Processing and Quantification

<table>
<thead>
<tr>
<th>CYTOLOGICAL FEATURES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area-Ch1</td>
</tr>
<tr>
<td>Perimeter-Ch1</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Total Intensity-Ch3</td>
</tr>
</tbody>
</table>
Figure 2 - Young DW et. al,

A. COMMON FACTOR MODEL

\[
X_{nm} = \sum_{i=1}^{k} L_{im} F_{im} + \epsilon_{nm}
\]

B. EIGENANALYSIS OF THE ADJUSTED CORRELATION MATRIX

C. FACTOR INTERPRETATION

D. BIOLOGICAL ACTIVITY SPACE

<table>
<thead>
<tr>
<th>CYTOLOGICAL FEATURES (m=30)</th>
<th>PHENOTYPIC ATTRIBUTES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>COMMON FACTOR SPACE (k=6)</td>
</tr>
<tr>
<td>EntropyInterCh1</td>
<td>FACTOR 1</td>
</tr>
<tr>
<td>TotalInterCh1</td>
<td>Integrated Intensity</td>
</tr>
<tr>
<td>AreaCh1</td>
<td>DNA REPLICATION</td>
</tr>
<tr>
<td>LengthCh1</td>
<td>FACTOR 2</td>
</tr>
<tr>
<td>WidthCh1</td>
<td>FACTOR 3</td>
</tr>
<tr>
<td>FiberLengthCh1</td>
<td>FACTOR 4</td>
</tr>
<tr>
<td>EqChOrCh1</td>
<td>FACTOR 5</td>
</tr>
<tr>
<td>EqSphereAreaCh1</td>
<td>FACTOR 6</td>
</tr>
<tr>
<td>EqSphereVolCh1</td>
<td></td>
</tr>
<tr>
<td>EqEllipsoidObVolCh1</td>
<td></td>
</tr>
<tr>
<td>EqEllipsoidPVarCh1</td>
<td></td>
</tr>
<tr>
<td>PerimCh1</td>
<td></td>
</tr>
</tbody>
</table>
Figure 3 - Young DW et. al,

A. Natural and Synthetic Compounds (n=6547)
   High-Content Screen (Replicates = 2)
   Identification of Active Compounds Using a Factor-Based Phenotypic Response Metric (Hits = 211)

B. Screening Set
   - Hits Top 2-5%
     - Diversity Set: 21%
     - Natural Products: 18%
     - Known Bioactives: 58%
   - Hits Top 1%
     - Diversity Set: 21%
     - Natural Products: 32%
     - Known Bioactives: 48%

C. HIGH-CONTENT SCREENING
   Natural and Synthetic Compound Library

D. PHENOTYPIC HIT PROFILING
   TOP-5%
   - Nuclear Size
   - DNA Replication
   - Chr. Condensation
   - Nuclear Morphology
   - EdU Texture
   - Nuclear Elongation
   - Top 1%
   - Top 2-5%
Figure 4a - YoungDW et. al,
Figure 4b – Young DW et. al,

Structure-Activity Discordance ~4%
"Activity Cliffs"

Structure-Activity Concordance ~96%

(Phenotypic Vector – Euclidean Distance)
Figure 5 - YoungDW et al.,

A. PHENOTYPE STRUCTURE

#78 Aurantimycin A

#81 Hexenidic Acid Chlorhydridin

#79 Diperamycin

#83 Kendomycin.

B. PHENOTYPE STRUCTURE

#141: Clobetasol-17-Propionate (DERMOVATE®)

#143: Dexamethasone

#145: Entolox

#152 Triamcinolone

C. PHENOTYPE STRUCTURE

#154 Digitoxigenin

#157 Ouabain Octahydrate

#158 Digoxin

#155 Emetine

#161 Cycloheximide

#165 Progesterone

#169 Danatro
Figure 6 - YoungDW et. al,

A.

B.
Supplementary figure 1 – Young DW et. al,

Factor 1: Nuclear Size

Factor 2: DNA Replication

Factor 3: Chr. Condensation

Factor 4: Nuc. Morphology

Factor 5: Edu Texture

Factor 6: Nuc. Ellipticity
Supplementary Figure 2 - Young DW et. al,

Phenotypic Response Metric
_Dose Response_

Constituent Factors
_Dose Response_

[Diagram showing responses to various compounds and factors]
Supplemental Figure 4 - YoungDW et. al.

A. Phenotype vs. Structure Spearman's Rank Correlation

Original Compound Structure Similarity Matrix
Random Compound Structure Similarity Matrices (n=1000)

B. Structure Similarity Matrix
Example Random Structure Similarity Matrix
Supplementary Figure 5: Young DW et al.

A. Phenotype vs. Compound Similarity Spearman's Rank Correlation

- Original Compound Target Similarity Matrix
- Randomly Permutated Target Similarity Matrix (n=1000)
- Original Compound Structure Similarity Matrix
- Randomly Permutated Structure Similarity Matrix (n=1000)

B. Original Compound Target Similarity Matrix

Example Random Compound Target Similarity Matrix
Supplementary Figure 6 - Young DW et. al,

A.

B.