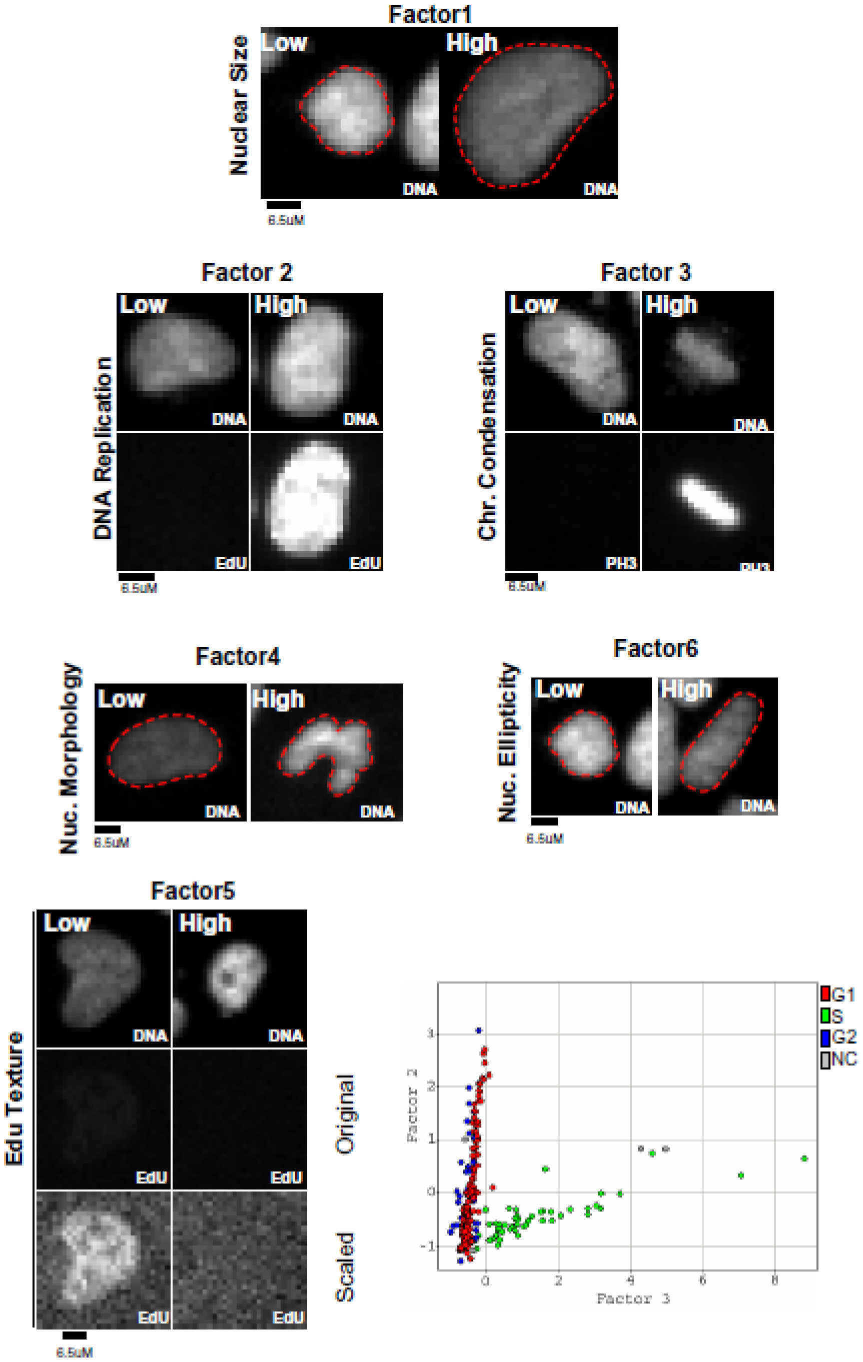
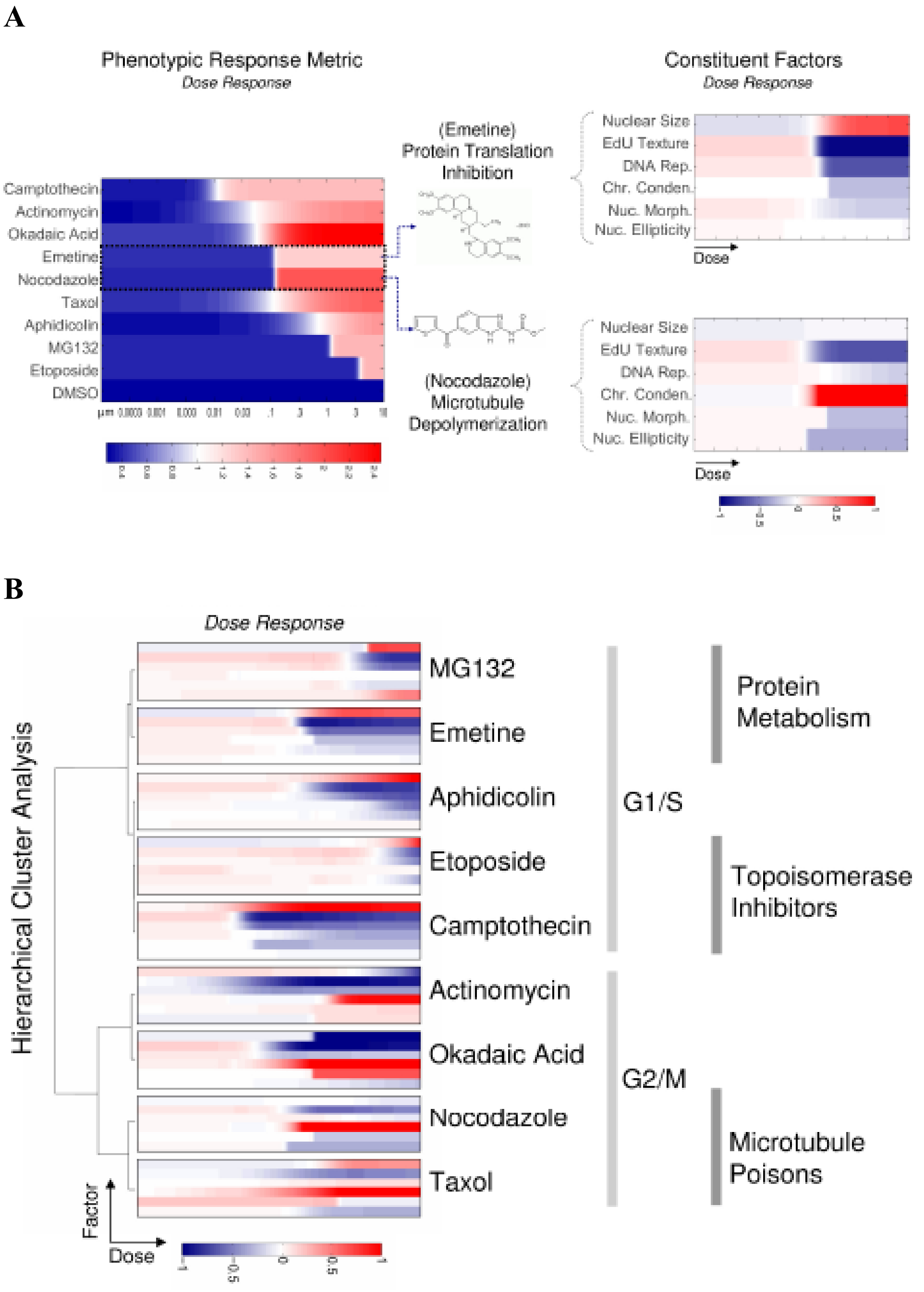
Supplementary Figures

Supplementary Figure 1: Examples of images with high or low values on each factor



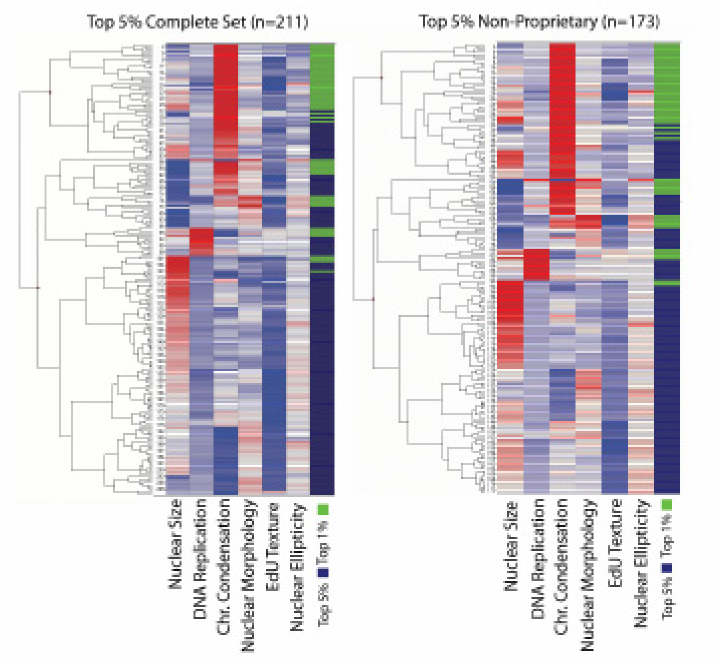
To establish the biological relevance of the six Factors we examined images of cells scoring both at the extreme high and extreme low ends of each Factor. In this analysis we observed that Factor 1 is proportional to nuclear size and DNA content. High scoring cells on this Factor have large nuclei, and 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: Phenotypic dose-response and hierarchical clustering analysis of known cytotoxic compounds



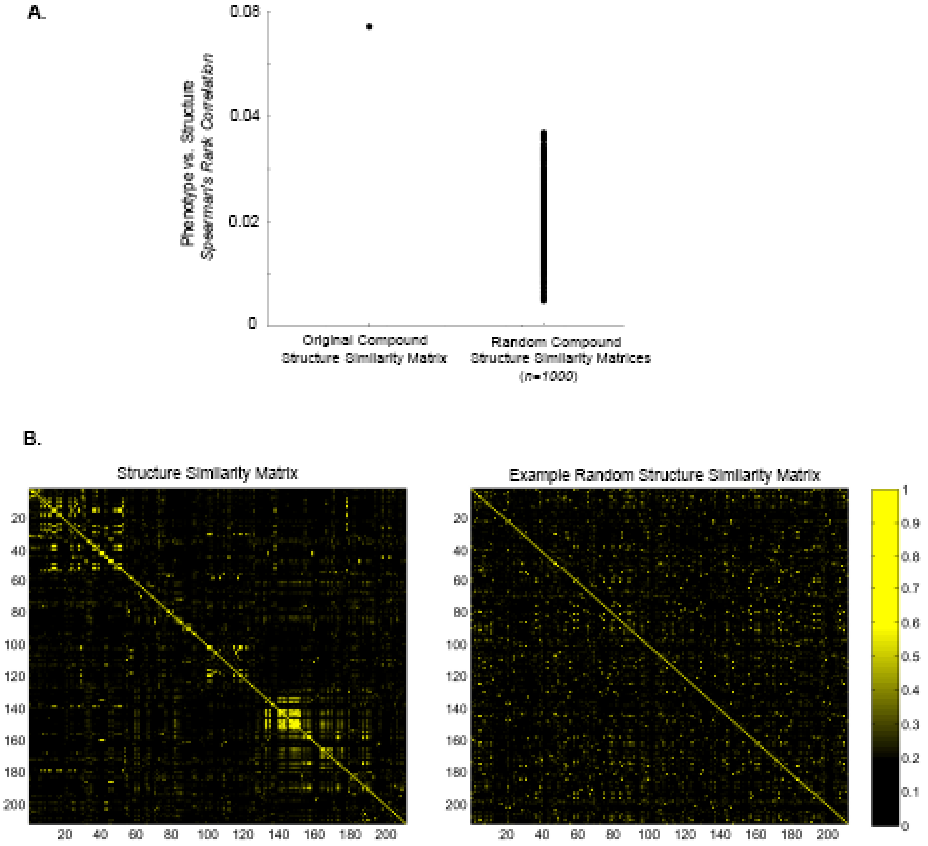
(**A**) We performed a phenotypic dose-response analysis of classic cytotoxic compounds. Factor-based dose-response relationships for cytotoxic compounds across serial dilutions ranging from 10µM to 0.70pM for a 20hrs treatment period. The response is a phenotypic response metric (see methods section). 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). (**B**) 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 3: Hierarchical clustering analysis of non-proprietary hit compounds



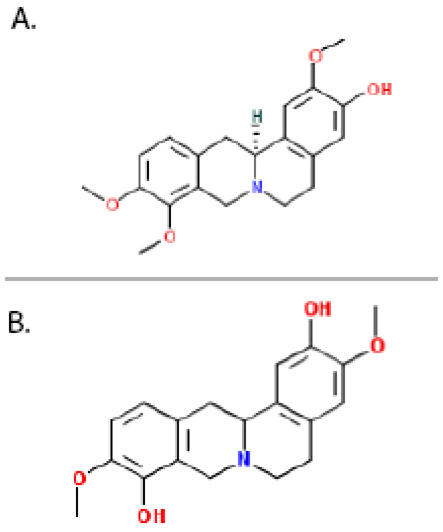
We performed hierarchical clustering of mean factor scores for each of the 173 non-proprietary hit compounds. Clustering is based on Ward's linkage criteria and the half Euclidean distance metric. The position of compounds within the top 1% and 2-5% based on phenotypic response is shown. (-1=blue, +1.5=red). Cluster analysis of a this reduced data set, compared with Figure 3D, retains the overall hierarchical structure and, thereby, indicates robustness of the method and a minimal dependency on individual compounds.

Supplementary Figure 4: Statistical analysis of compound structure and phenotype correlation



We determined if the observed visual correlation between the phenotypic similarity matrix and the compound structure similarity matrix was statistically significantly. (**A**) We determined the Spearman correlation coefficient for rank ordered phenotypic similarities, and compound similarities using the original matrices from Figure 4 (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: Scoulerine-related compound pairs identified as an example of activity-cliff

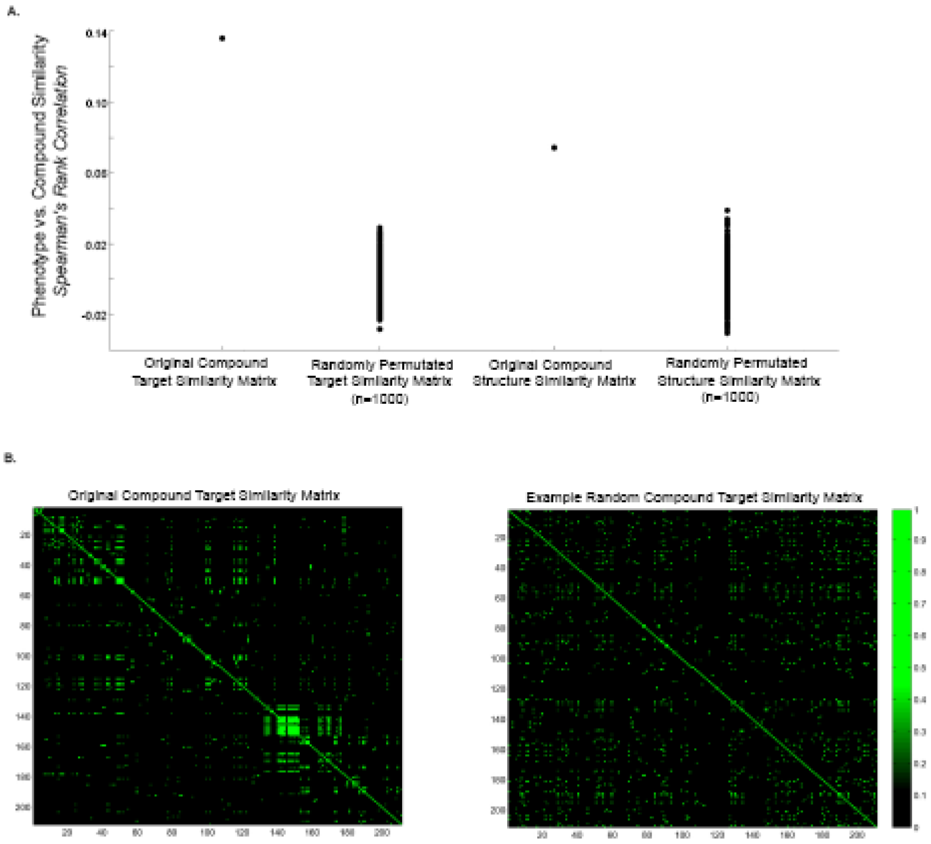


Structures are shown for two Scoulerine-related compounds identified as an activity-cliff pair in the analysis described in Figure 4B (red data point). (**A**) The compound is a known antagonist of the D2 receptor[1]. (**B**) The drug Scoulerine, which is derived from poppy seeds, is used as a sedative and binds to a series of GPCR receptors, such as the alpha andrenergic receptors, GABA, 5HT receptors, and the D1 a D2 dopamine receptors[2].

1. Schaper,K. Free-Wilson-Type Analysis of Non-Additive Substituent Effects on THPB Dopamine Receptor Affinithy Using Artificial Neural Networks. QSAR Struct.-Act.Relat. **19**, 354-360. 1999.
Ref Type: Generic

2. Carrieri,A., Centeno,N.B., Rodrigo,J., Sanz,F., & Carotti,A. Theoretical evidence of a salt bridge disruption as the initiating process for the alpha1d-adrenergic receptor activation: a molecular dynamics and docking study. *Proteins* **43**, 382-394 (2001).

Supplementary Figure 6: Statistical analysis of ligand-target prediction and phenotype correlation



We determined the correlation between the phenotypic similarity matrix and the compound target similarity matrix was statistically significantly. **(A)** Top five compound targets based on Bayes Score (see Methods and Supplementary data 2) 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 5. 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 Table 1 Young DW et al

#	Cytological Feature	Description
1	AreaCh1	Nuclear Area
2	PerimCh1	Nuclear Perimeter
3	ShapeP2ACh1	Nuclear Perimeter to Area Ratio
4	ShapeLWRCh1	Nuclear Bounding Box Length to Width Ratio
5	ShapeBFRCh1	Nuclear Bounding Box Fill Ratio; A Ratio of Nuclear Area to Area of Bounding Box
6	LengthCh1	Bounding Box Length
7	WidthCh1	Bounding Box Width
8	FiberLengthCh1	Nuclear Length
9	ConvexHullAreaRatioCh1	Ratio of the Convex Hull Area (i.e., Area of the smallest convex set of pixels containing the entire nuclear) and the Nuclear Area
10	ConvexHullPerimRatioCh1	Ratio of the Convex Hull perimeter (i.e., perimeter of the smallest convex set of pixels containing the entire nuclear) and the Nuclear perimeter
11	EqCircDiamCh1	Diameter associated with the circle with the equivalent cross sectional area as the nucleus
12	EqSphereVolCh1	Volume of the sphere created by rotating the equivalent circle about its diameter
13	EqSphereAreaCh1	Surface area of the equivalent sphere
14	EqEllipseLWRCh1	Length to width ratio of the ellipse with the equivalent area and aspect ratio
15	EqEllipseProlateVolCh1	Volume of the ellipsoid created by rotating the equivalent ellipse about its major axis
16	EqEllipseOblateVolCh1	Volume of the ellipsoid created by rotating the equivalent ellipse about its minor axis
17	NeighborMinDistCh1	Minimum distance to neighboring nuclei
18	TotalIntenCh1	Integrated pixel intensity within the object
19	AvgIntenCh1	Average pixel intensity within the object
20	VarIntenCh1	Variance in pixel intensity within the object
21	SkewIntenCh1	Skewness in pixel intensity within the object
22	KurtIntenCh1	Kurtosis in pixel intensity within the object
23	EntropyIntenCh1	Texture parameter that quantifies the information content represented in the objects pixel intensities
24	DiffIntenDensityCh1	Texture parameter that accounts for the intensity variation within an object
25	TotalIntenCh2	Integrated pixel intensity within the object on Ch2
26	AvgIntenCh2	Average pixel intensity within the object on Ch2
27	TotalIntenCh3	Integrated pixel intensity within the object on Ch3
28	AvgIntenCh3	Average pixel intensity within the object on Ch3
29	VarIntenCh3	Variance in pixel intensity within the object on Ch3
30	SkewIntenCh3	Skewness in pixel intensity within the object on Ch3
31	EntropyIntenCh3	Texture parameter that quantifies the information content represented an the objects pixel intensities on Ch3
32	DiffIntenDensityCh3	Texture parameter that accounts for the intensity variation within an object on Ch3
33	IntenCoocMaxCh3	These four cooccurance parameters are texture measurements that account for the spatial arrangement of different pixel. They are computed from the co-occurance matrix, which is a matrix of probabilities of intensity co-occurance along a given direction in an image, see Haralick et al
34	IntenCoocContrastCh3	
35	IntenCoocEntropyCh3	
36	IntenCoocASMCh3	

Supplementary Methods

Factor Analysis

For High-Content applications, data are contained in an n x 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, **ε:**

$$\mathbf{X} = \mathbf{LF} + \boldsymbol{\varepsilon}$$

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, **ε** 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**,

$$\boldsymbol{\Sigma} = \mathbf{LL}^{T} + \boldsymbol{\Psi}$$

where ^{T} is the transpose operator and **Ψ** is the covariance of **ε,** 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, **Ψ**. With some underlying restrictions placed on the structure of **Σ,** the model fit can be accomplished quite easily [1]. The Factor model fit was performed here using the so-called Principal Factor method [1, 2]; 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,

$$\mathbf{L} = \left| \sqrt{\lambda_1}\mathbf{e}_1 \quad \sqrt{\lambda_2}\mathbf{e}_2 \quad \cdots \quad \cdots \quad \sqrt{\lambda_k}\mathbf{e}_k \right|$$

Where λ_i is the eigenvalue associated with the eigenvector $\mathbf{e}_{i,}$ 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, $\mathbf{\Sigma}$ **.** An *n* x *n* orthogonal rotation matrix, **T** can be specified such that:

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T + \mathbf{\Psi} = \mathbf{L}\mathbf{T}\mathbf{T}^T\mathbf{L}^T + \mathbf{\Psi} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$$

And,

$$\mathbf{\Lambda} = \mathbf{L}\mathbf{T} = \textit{Rotated Loading Matrix}$$

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 [3]. 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, $\mathbf{F}_s$ 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.

Application of Factor Analysis to HCS data

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 [4-9]. Compared to other recent efforts to develop phenotypic profiles from HCS data[10-12], 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 [11]. 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 [13]. 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 4-6).

One limitation of our screen 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 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. As we demonstrate in Supplementary Figures 2 and 3, factor analysis can be readily extended to such higher dimensionality datasets. It could also be possible to implement a titration-invariant similarity score[10] within the context of factor analysis for data reduction of concentration-dependent effects. Nevertheless, we demonstrate in Supplementary Figure 3, that clustering at a single saturating dose produces meaningful clusters that contain compounds with common or related biochemical and cellular 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[14], or proteomic methods. Profiles based on HCS cytology are, perhaps, less rich in detailed information than some "-omic" 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 [15]. Profiling methods that generate profiles by combining multiple cell-based pathway readouts in image-based protein complementation assays[16] 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.

Distance Metric

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

$$\mathbf{P} = \sqrt{\left|\mathbf{F}_s^t - \mathbf{F}_s^u\right|^T \left|\mathbf{F}_s^t - \mathbf{F}_s^u\right|}$$

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.

Labeling EdU with Rhodamine azide using “Click Chemistry”

Ethnyldeoxyuridine incorporated into cellular DNA was labeled with Rhodamine azide (a gift from Adrian Salic, Harvard Medical School) by a Cu(I)-catalyzed Huisgen 1,3-dipolar cycloaddition reaction (also known as "click reaction")[17]. Rhodamine azide 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 added to cells for 30 minutes at 25°C and then washed with PBS.

Reference List

1. Johnson,R.A. & Wichern,D.W. *Applied Multivariate Statistical Analysis*(Prentice Hall, Inc.,2002).

2. Hatcher,L. *A Step-by-Step Approach to Using SAS for Factor Analysis and Structural Equation Modeling*(SAS Institute, Inc., Cary, NC, USA, 1994).

3. Henry,F.K. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **V23**, 187-200 (1958).

4. Carroll,J.B. & Schweiker,R.F. Factor Analysis in Educational Research. *Review of Educational Research* **21**, 368-388 (1951).

5. Floyd,F.J. & Widaman,K.F. Factor Analysis in the Development and Refinement of Clinical Assessment Instruments. *Psychological Assessment* **7**, 286-299 (1995).

6. Malinowsi,E.R. *Factor Analysis in Chemistry*(John Wiley and Sons, Inc., New York,2002).

7. Spearman,C. "General Intelligence", Objectively Determined and Measured. *American Journal of Psychology* **15**, 201-293 (1904).

8. Stewart,D.W. The Application and Misapplication of Factor Analysis in Marketing Research. *Journal of Marketing Research* **18**, 51-62 (1981).

9. Tinsley,H.E.A. & Tinsley,D.J. Uses of Factor Analysis in Counseling Psychology Research. *Journal of Counseling Psychology* **34**, 414-424 (1987).

10. Perlman,Z.E. *et al.* Multidimensional drug profiling by automated microscopy. *Science* **306**, 1194-1198 (2004).

11. Tanaka,M. *et al.* An unbiased cell morphology-based screen for new, biologically active small molecules. *PLoS Biol.* **3**, e128 (2005).

12. Loo,L.H., Wu,L.F., & Altschuler,S.J. Image-based multivariate profiling of drug responses from single cells. *Nat Meth*(2007).

13. Stewart,D. Difference between Principal Components and Factor Analysis. *Journal of Consumer Psychology* **10**, 75-76 (2001).

14. Butcher,R.A. & Schreiber,S.L. Using genome-wide transcriptional profiling to elucidate small-molecule mechanism. *Current Opinion in Chemical Biology* **9**, 25-30 (2005).

15. Kustra,R., Shioda,R., & Zhu,M. A factor analysis model for functional genomics. *BMC Bioinformatics* **7**, 216 (2006).

16. MacDonald,M.L. *et al.* Identifying off-target effects and hidden phenotypes of drugs in human cells. *Nat Chem Biol* **2**, 329-337 (2006).

17. Kolb, H.C., Finn, M.G., &and Sharpless, K.B. Click Chemistry: Diverse Chemical Function from a Few Good Reactions. *Ang. Chem. Int.* **40**, 2004-2021 (2001).