Expanding the Range of 'Druggable' Targets with Natural Product-based Libraries: An Academic Perspective

Renato A. Bauer,[1] Jacqueline M. Wurst,[1] and Derek S. Tan[1,2,*]

[1]Tri-Institutional Training Program in Chemical Biology,
[2]Molecular Pharmacology & Chemistry Program and Tri-Institutional Research Program,
Memorial Sloan–Kettering Cancer Center, 1275 York Ave, Box 422, New York, NY 10065, USA

Supplementary Information

A. Principal component analysis protocol and results

A total of 127 compounds (**Table S1**) were selected for analysis, comprising the top 40 brand-name small molecule drugs by revenue in 2006 [1], 60 natural products with diverse structures, biosythetic origins and biological activities, including the 24 identified by Ganesan as having led directly to an approved drug since 1970 [2], 10 drug-like pyrrazolecarboxamides in the MLSMR from ChemBridge, 10 drug-like dihydrotriazolopyrimidines in the MLSMR from Chem Div, and the 7 natural products and library-derived probes discussed in the manuscript. This relatively small dataset allows for identification of individual compounds in the resulting chemical space plot while retaining robustness – removal of pladienolide from the dataset resulted in a rmsd change of 0.04% on the PC1 axis and 0.33% on the PC2 axis for the positions of the remaining 126 compounds. Notably, analyses with an additional 10 different scaffolds each from the ChemBridge and Chem Div drug-like libraries gave similar results (not shown).

A set of 20 physicochemical properties (**Table S2**) for all 127 compounds were then obtained from PubChem and/or calculated using free online cheminformatics tools (Molinspiration [3], VCCLab [4,5]), ChemDraw, and manual inspection: molecular weight (MW) [6], nitrogens (N), oxygens (O), calculated 1-octanol/water partition coefficient (XLogP) [5,6], hydrogen bond donors (HBD) [6], hydrogen bond acceptors (HBA) [6], rotatable bonds (RotB) [7], topological polar surface area (tPSA) [7,8], an alternative calculated logP (ALOGPs) [9], calculated aqueous solubility (ALOGpS) [8], stereogenic centers (nStereo), *R* stereogenic centers (R), *S* stereogenic centers (S), nStereo/MW (nStMW), R – S (RSdelta), Rings, aromatic rings (RngAr), ring systems (RngSys), largest ring outline (RngLg), and Rings/RngSys (RRSys). These parameters were selected based on several criteria. First, Lipinski parameters [6] (MW ≤ 500, logP ≤ 5, HBA ≤ 10, HBD ≤ 5) and Veber parameters [7] (RotB ≤ 10, tPSA ≤ 140 $Å^2$) have been correlated with oral bioavailability and are frequently used to filter drug-like libraries. Second, Tetko's calculated aqueous solubility (ALOGpS) [8] was included as compound solubility is critical in screening. Third, since we did not have a convenient means to assess three-dimensional descriptors, several stereochemical parameters (nStereo, R, S, RSdelta) were included as a

first-order approximation of three-dimensional complexity – the last three are pertinant only to libraries containing enantiomeric compounds as an arbitrary and imperfect indication of stereochemical diversity, but were retained from other analyses we have performed for consistency. A molecular weight-normalized value (nStMW) was also included as a measure of 'stereochemical density'. Fourth, several additional parameters found by Feher and Schmidt to differentiate synthetic drugs and natural products were included [9]. Synthetic drugs tend to have more nitrogens, while natural products tend to have more oxygens (N,O). Natural products also tend to have fewer aromatic rings and more complex, fused ring systems (Rings, RngAr, RngSys, RngLg, RRSys). These data were assembled conveniently in a Microsoft Excel spreadsheet, and average values calculated for each compound series (**Table S3**).

Table S1. Compounds used in principal component analysis.

Series	**Compounds**			
Best-Selling	Lipitor	Lexapro	Topamax	Coreg
Brand Name	Nexium	Seroquel	Toprol	Valtrex
Drugs	Prevacid	Protonix	Zetia	Adderall
(40 entries)	Flonase	Ambien	Fosamax	Aciphex
	Serevent	Actos	Abilify	Cymbalta
	Singulair	Zoloft	Levaquin	Crestor
	Effexor	Wellbutrin	Lamictal	Diovan
	Plavix	Avandia	Celebrex	Tricor
	Zocor	Risperdal	Benazepril	Concerta
	Norvasc	Zyprexa	Zyrtec	Imitrex
ChemBridge	*PubChem*	5771429	5309975	5308431
Library	*Compound CIDs:*	5771374	5309772	5309246
(10 entries)	5771496	5771371	5309762	5309020
ChemDiv	*PubChem*	2529482	2474145	2490046
Library	*Compound CIDs:*	2474174	1340935	2490068
(10 entries)	2529498	2471337	2490059	1342784
Nat Prods	CephamycinC	Mizoribine	Coformycin	Compactin
(compliant)	Spergualin	SQ26180	Arglabin	Artemisinin
(12 entries)	Forskolin	Thienamycin	Bestatin	Plaunotol
Nat Prods	Daptomycin	Validamycin	MidecamycinA1	Rapamycin
(non-comp)	EchinocandinB	Avermectin B1a	Taxol	FK506
(12 entries)	Calicheamicin γ1	Cyclosporin A	Pseudomonic Acid A	Lipstatin
Natural	Geldanamycin	Trapoxin B	Talaromycin B	Bleomycin
Products	Actinonin	Vincristine	Spongistatin 1	Brefeldin A
(other)	Discodermolide	Colchicine	Radicicol	Cytochalasin B
(36 entries)	Monensin	Trichostatin	Salicylihalamide A	Epothilone A
	Calyculin A	Fumagillin	Brevetoxin B	Apoptolidin
	Amphotericin B	Staurosporine	Rifamycin B	Lactacystin
	Adriamycin	Erythromycin A	Quinine	Duocarmycin A
	Ginkgolide B	Streptomycin	Mycobactin S	Zaragozic Acid A
	Phorbol MA	Penicillin G	Telomestatin	Vancomycin
COCB	FR901464	Lactam Carboxamide	Pladienolide B	Gemmacin
(7 entries)	Robotnikinin	Abyssomicin C	Avrainvillamide	

Table S2. Structural and physicochemical parameters used in PCA.

Parameter	Description	Method of Determination
MW	molecular weight	ChemDraw Analysis Window
N	number of nitrogens	ChemDraw Analysis Window
O	number of oxygens	ChemDraw Analysis Window
XlogP	calc *n*-octanol/water partition coefficient	http://www.vcclab.org
HBD	number of hydrogen bond donors	http://www.molinspiration.com
HBA	number of hydrogen bond acceptors	http://www.molinspiration.com
RotB	number of rotatable bonds	http://www.molinspiration.com
tPSA	topological polar surface area	http://www.molinspiration.com
ALOGPs	calc *n*-octanol/water partition coeff (alt)	http://www.vcclab.org
ALOGpS	calculated aqueous solubility	http://www.vcclab.org
nStereo	number of stereocenters	http://www.molinspiration.com
R	number of R stereocenters	ChemDraw Show Stereochemistry
S	number of S stereocenters	ChemDraw Show Stereochemistry
RSdelta	*R – S*	Microsoft Excel
nStMW	*nStereo ÷ MW* (stereochemical density)	Microsoft Excel
Rings	number of rings	Manual inspection
RngAr	number of aromatic rings	Manual inspection
RngSys	number of ring systems	Manual inspection
RngLg	number of atoms in largest ring outline	Manual inspection
RRSys	*Rings ÷ RngSys* (ring complexity)	Microsoft Excel

Table S3. Average structural and physicochemical parameters by compound series. † = nStMW × 1000 for clarity

	Drug	NP	ChBr	ChDv	COCB
MW	361	629	381	446	488
N	2.2	2.6	4.3	4.7	2.0
O	2.9	9.7	3.1	3.4	5.6
XlogP	2.7	1.5	2.9	1.8	3.3
HBD	1.5	4.9	1.1	1.9	2.1
HBA	5.4	10.8	5.9	7.7	6.7
RotB	6.3	9.7	5.3	6.1	6.1
tPSA	69	183	103	94	106
ALOGPs	2.8	2.1	3.3	2.7	3.4
ALOGpS	-3.9	-3.8	-4.0	-3.8	-4.7
nStereo	1.4	9.1	0.0	1.0	5.3
R	0.6	4.1	0.0	0.5	2.7
S	0.8	5.0	0.0	0.5	2.6
nStMW†	3.7	13.9	0.0	1.3	11.2
RSdelta	-0.2	-0.9	0.0	0.0	0.1
Rings	2.9	3.8	3.2	4.2	4.7
RngAr	2.1	1.0	2.9	2.9	1.1
RngSys	2.1	2.0	3.1	3.1	2.4
RngLg	8.4	15.8	6.3	9.4	12.6
RRSys	1.4	2.3	1.0	1.4	2.7

To provide a visual representation of the position of each compound in chemical space, we then carried out principal component analysis with the "R" open source statistical computing package[10] to reduce the 20-dimensional vector corresponding to each compound to a 2-dimensional vector, with minimal loss of information. The detailed protocol is as follows:

1) In MS Excel, a "Raw" worksheet was created with compounds in rows and physicochemical descriptors in columns. Note that compound names must not have spaces or other punctuation.
2) Average values were calculated for individual compound categories (*e.g.*, for "Drugs", "Natural Products", etc.) as well as MAX and MIN values for each column.
3) A "Norm" worksheet was created and normalized values were generated by normalizing each column to a range of 0–1 using the equation:

$$normval = (val - MIN) / (MAX - MIN)$$

4) With the upper left cell blank (R requires this to recognize a header row), the Number format was designated for all data columns to 4 decimal places.
5) The Excel workbook was saved.
6) The "Norm" worksheet was saved as "Data.txt" (Text–Tab Delimited) on the Desktop (Mac).
7) The Excel workbook was closed and the changes discarded.
8) The "R" open source computing package was opened and the following commands were entered:

```
9)  R> read.table("~/Desktop/Data.txt") -> a     { read data into dataframe a
10) R> t(a) -> b                                 { transposed dataframe a to b
11) R> prcomp(b) -> c                            { PCA of dataframe b results to c
12) R> summary(c)                                { summary of %contributions
                                                 { copied to a text file for reference
13) R> biplot(c,ylabs = NULL)                    { plot of data and eigenvectors
                                                 { saved as a screenshot for reference
14) R> c
```

15) This final command gave a line listing of the results. The first section of the data was selected and copied (PC1–PC6, without top headers).
16) These results were pasted into a MS Word text file and the font changed to Courier 8 pt.
17) This MS Word file was Saved as... "Results.txt" (Text Only with Line Breaks)
18) Excel was opened again and the results were imported by selecting Get External Data... in the Data menu, then Import from Text file.
19) The "Fixed" width button was left checked and dividers were adjusted, making sure to include minus signs in the second column (PC1) rather than the first (compound names).
20) This data file was imported into a new Excel worksheet "Results".
21) The first three columns (compound names, PC1, PC2) where copied into a new worksheet "PCA", and the Number format was designated to 3 decimal places.
22) Each group of compounds was then sorted in order of ascending PC1 to facilitate its location on the PCA plot.
23) With the PC1 and PC2 columns selected, the Scatter XY plot was selected in the Chart Wizard.
24) Series information for each set of compounds, *e.g.* Drugs, AVG Drug, etc., was entered and the chart formatted as desired.
25) Subsequent revisions to the dataset could then be made easily by using Paste Special... Values from new Results worksheets onto the existing PCA worksheet.

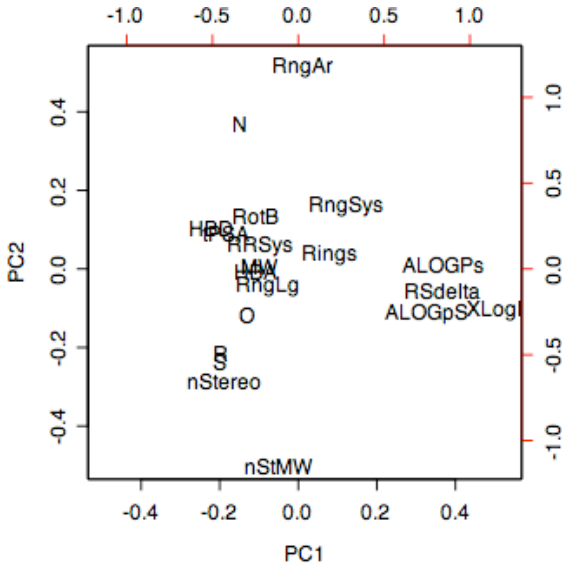
Compounds and compound family averages were plotted on the two newly generated axes (principal components), which were unitless, orthogonal, and are based on linear combinations of the original 20 variables. The Summary information from R above indicated that the first principle component (PC1), or eigenvector, represented 56.1% of the variance in the original

dataset, and PC2 represented an additional 16.6%, for a total of 72.7% (**Table S4**). The Biplot information from R above illustrated the component loadings, with MW and tPSA shifting compounds to the left, XlogP and AlogPs shifting compounds to the right, RngAr shifting compounds to the top, and nStMW shifting compounds to the bottom (**Figure S1**).

Table S4. Summary results from R indicating standard deviation and percent contribution for each principal component.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	1.963	1.068	0.774	0.693	0.478	0.427	0.374	0.263	0.218	0.196
Proportion of Variance	0.561	0.166	0.087	0.070	0.033	0.027	0.020	0.010	0.007	0.006
Cumulative Proportion	0.561	0.727	0.814	0.884	0.918	0.944	0.964	0.975	0.981	0.987

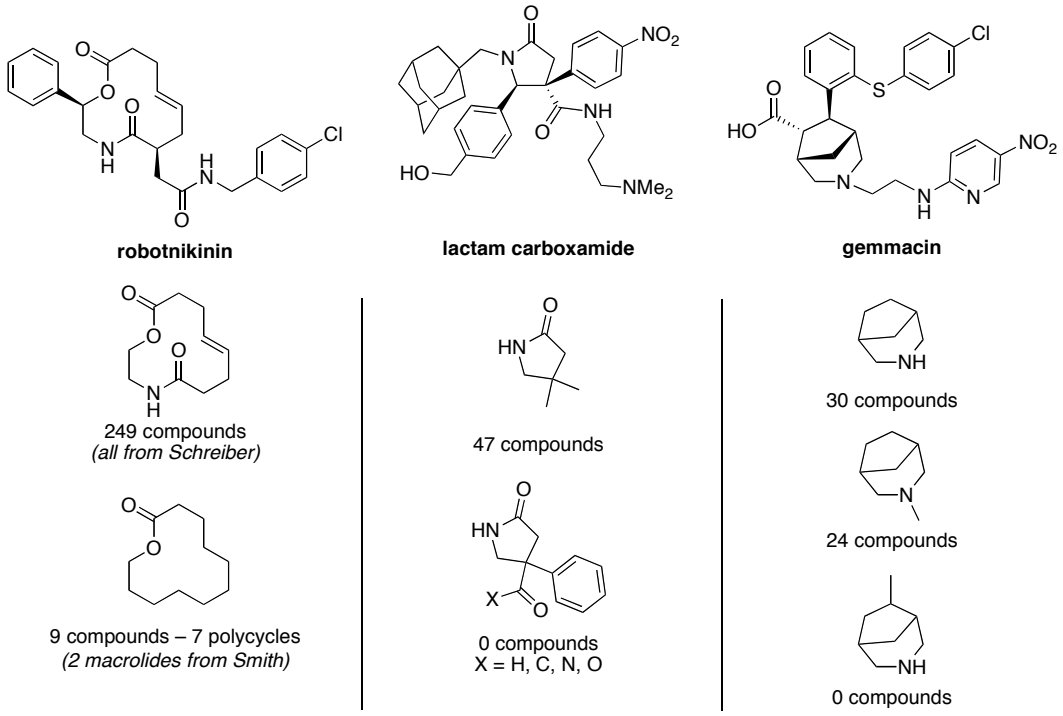
Figure S1. Plot of component loadings from R PCA.



B. PubChem substructure search protocol and results

Substructure searches for scaffolds found in the PubChem database were carried out using the substructure search function at http://pubchem.ncbi.nlm.nih.gov/. (Select "Chemical structure search", followed by the "Substructure/Superstructure" tab). SMILES codes were entered for each scaffold to be searched and the results restricted to compounds in the MLSMR (*Filters: Data source: From = MLSMR*).

Figure S2. Results of PubChem substructure searches for core scaffolds of library-derived probes.



C. Supplementary references

1. Top 200 Brand-Name Drugs by Retail Dollars in 2006; http://drugtopics.modernmedicine.com/

2. Ganesan: **The impact of natural products upon modern drug discovery**. *Curr Opin Chem Biol* 2008, **12**:306–317.

3. MolInspiration – free on-line cheminformatics tool; http://www.molinspiration.com/cgi-bin/properties

4. Tetko, I. V., Virtual Computational Chemistry Laboratory; http://www.vcclab.org/lab/alogps/

5. Tetko IV, Tanchuk VY, Kasheva TN, Villa AEP: **Internet software for the calculation of the lipophilicity and aqueous solubility of chemical compounds**. *J Chem Inf Comput Sci* 2001, **41**:246–252.

6. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ: **Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings**. *Adv Drug Deliv Rev* 1997, **23**:3–25.

7. Veber DF, Johnson SR, Cheng H-Y, Smith BR, Ward KW, Kopple KD: **Molecular properties that influence the oral bioavailability of drug candidates**. *J Med Chem* 2002, **45**:2615–2623.

8. Tetko IV, Tanchuk VY, Kasheva TN, Villa AEP: **Estimation of aqueous solubility of chemical compounds using E-state indices**. *J Chem Inf Comput Sci* 2001, **41**:1488–1493.

9. Feher M, Schmidt JM: **Property distributions: Differences between drugs, natural products, and molecules from combinatorial chemistry**. *J Chem Inf Comput Sci* 2003, **43**:218–227.

10. The R Project for Statistical Computing; http://www.r-project.org/