



Published in final edited form as:

Methods Mol Biol. 2015 ; 1263: 225–242. doi:10.1007/978-1-4939-2269-7_18.

Principal Component Analysis as a Tool for Library Design: A Case Study Investigating Natural Products, Brand-Name Drugs, Natural Product-Like Libraries, and Drug-Like Libraries

Todd A. Wenderski, Christopher F. Stratton, Renato A. Bauer, Felix Kopp, and Derek S. Tan

Abstract

Principal component analysis (PCA) is a useful tool in the design and planning of chemical libraries. PCA can be used to reveal differences in structural and physicochemical parameters between various classes of compounds by displaying them in a convenient graphical format. Herein, we demonstrate the use of PCA to gain insight into structural features that differentiate natural products, synthetic drugs, natural product-like libraries, and drug-like libraries, and show how the results can be used to guide library design.

Keywords

Principal component analysis (PCA); Medium rings; Macrocycles; Ring expansion; Natural products; Drugs; Libraries; Diversity-oriented synthesis

1 Introduction

Principal component analysis (PCA) is a mathematical method for dimensionality reduction that allows for multidimensional datasets to be visualized using two- or three-dimensional plots with minimal loss of information [1, 2]. When applied in the context of diversity-oriented synthesis, PCA is primarily used to visualize similarities and differences within collections of compounds based on structural and physicochemical parameters, and can be leveraged in library design [3]. Molecular weight, stereocenters, rotatable bonds, hydrophobicity, and aqueous solubility are a few examples of parameters commonly included in such analyses. Herein, we selected 20 structural and physicochemical parameters for analysis based on previously identified correlations of these parameters with oral bioavailability, cell permeability, solubility, and binding selectivity, as well as their ability to distinguish synthetic drugs from natural products (*vide infra*). Each compound in our analysis is represented as a 20-dimensional vector defined by the structural and physicochemical parameters. PCA rotates these vectors onto a new set of orthogonal axes called principal components, in which the variance retained from the original data is maximized on each successive principal component. As such, the three-dimensional plot we show in this example retains 75 % of the variance from the full 20-dimensional dataset.

PCA can also be used to guide the design of chemical libraries. This is important in drug discovery because current drugs are limited in both structure and function. For example, current small-molecule drugs address only about 1 % of the protein targets encoded in the human genome [4], and half of those target only four protein classes: rhodopsin-like G-protein receptors, nuclear receptors, and voltage- and ligand-gated ion channels. In contrast, natural products are known to target a broader range of protein classes and have led to the majority of antibacterials (65 %) and anticancer drugs (75 %) [5]. Therefore, novel libraries of compounds that share the structural features of natural products are attractive for the discovery of lead compounds to evaluate new therapeutic targets.

Along these lines, many macrocycle and medium-ring-containing natural products have compelling biological activities. This key cyclic framework presents functional groups to biological targets in appropriate pharmacophoric conformations [6–8]. Compared to their corresponding linear congeners, macrocycles can provide increased binding affinity [9], improved bioavailability [10], and, in some cases, enhanced cell permeability [11], which are desirable pharmacological properties in the development of new drugs.

However, despite these attractive features of macrocycles and medium rings, they remain severely underexploited in current drug and probe discovery efforts [12, 13], due to challenges associated with their synthesis. To address the underrepresentation of these compounds, we have sought to circumvent the inherent limitations of classical cyclization-based strategies for macrocycle and medium-ring synthesis by developing alternative ring-expansion approaches that are tolerant of a broad range of substitution patterns and functional groups. We recently developed two such methods (Fig. 1) [12, 13], both of which can be employed on gram scale, provide products bearing handles for further diversification, and are transferable to parallel synthesis platforms.

Herein, we describe the use of PCA to assess how libraries of compounds produced using these synthetic routes compare to natural products and what structural and physicochemical parameters distinguish them from synthetic drugs and drug-like libraries. The information harnessed from PCA can also direct downstream modifications of a scaffold to obtain molecules that are more characteristic of a targeted class, such as natural products.

In this example, we show that compounds appearing in the proximity of the drug-like region of the PCA plot can be modified to have greater natural product-like properties by addressing several influential structural and physicochemical parameters. The relative contributions of structural and physicochemical parameters to each principal component (PC) axis are obtained from the loading data and loading plots produced from PCA. In the analysis presented herein, the number of oxygen atoms, hydrogen bond donors, and hydrogen bond acceptors are among the most influential parameters for PC1. Stereochemical density (the number of stereocenters normalized to molecular weight) and the fraction of sp^3 -hybridized carbons are large contributors to PC2. We further demonstrate that these structural and physicochemical parameters can be addressed by chemical modifications of our library members to increase their natural product-like character. Subsequent analysis of these modified compounds in PCA demonstrates their increased penetration into natural product-like regions of the plot. This work illustrates how insights

gleaned from PCA can be used in the planning of chemical libraries to probe targeted areas of chemical space.

2 Materials

This analysis requires the use of several software packages that are either commonly available in chemistry labs or freely available for download. The following software and versions were used for this protocol:

1. Mac OS 10.5.8 or Windows 7 (procedure described for Mac OS with specific changes for Windows 7 users indicated).
2. CS ChemBioDraw Ultra 12.0.3 (CambridgeSoft).
3. Microsoft Office 2008.
4. Instant JChem 5.3.8 (ChemAxon, free Academic License available).
5. Virtual Computational Chemistry (VCC) Laboratory: <http://www.vcclab.org/lab/alogps/start.html> (requires a JAVA-enabled browser).
6. R 2.9.2 (open-source R Project for Statistical Computing, available from <http://www.r-project.org>).

3 Methods

3.1 Calculation of Physicochemical Parameters

1. Obtain SMILES codes for all of the compounds to be included in the PCA (*see* Notes 1 and ²). SMILES codes for known compounds can be obtained from PubChem (<http://pubchem.ncbi.nlm.nih.gov/>) or other online resources. For new compounds, SMILES codes can be generated using ChemBioDraw (*see* Note 3).
2. Create a new MS Excel file containing one column for compound names (Column A) and one column for SMILES codes (Column B) (*see* Note 4). Do not include a header row. Group the compounds by compound class (such as Drugs, Natural Products). Save the MS Excel file as a Text (tab delimited) (.txt) file that will be used in Subheading 3.1, **step 4**. Delete the compound names column and save an additional .txt file that contains only SMILES codes, which will be used later will be used later for batch processing (Subheading 3.1, **step 7**).

¹In this analysis, we used 40 top-selling drugs from [20], 60 diverse natural products, 20 drug-like library members, 23 macrocycle natural products, 32 synthetic macrocycles, 20 medium-ring natural products, 38 synthetic medium rings, and 25 cyclohexadienone precursors to those medium rings. In the analysis described in Subheading 3.3, an additional eight synthetic medium-ring diols were included.

²Much of the raw data used in this analysis is available from the Supplementary Information for refs. [12, 13].

³To obtain the SMILES codes in ChemBioDraw, select the chemical structures and choose Edit > Copy As > SMILES. Paste the SMILES codes into an MS Word document. The compounds in the string are separated by a period (“.”), and can be converted to a table format in MS Excel by saving the MS Word document as a text file (.txt) and importing the data in MS Excel using Data > Get External Data. Select the text file, choose the “Delimited” option in Step 1 of the Text Import Wizard, and in Step 2 of the Wizard specify the delimiters as a period (“.”) in the “Other” field. The imported SMILES codes can be transposed (flipped from row to column format) by copying them, then selecting Edit > Paste Special, and clicking on the “Transpose” option.

⁴Some software does not handle spaces and punctuation in compound names, but underscores can be used instead. For Windows 7 users, spaces are allowed.

3. Using Instant JChem, open a new project (File > New Project), and click on “Next” (*see* Note 5). Enter a project name and then click on “Finish”.
4. From Instant JChem, import the MS Excel file containing the compound names and SMILES codes by selecting File > Import File, and then click on “Next” (*see* Note 6). Click on the folder icon next to the “File to import” field, and then navigate to and select the .txt file containing compound names and SMILES codes. Under “File Format” choose “Delineated text files (*.csv, *.tab, *.txt)”, and then click on “Open”. After Instant JChem has finished scanning the file and indicated the number of fields found, click on “Next”. The “Field details” panel gives a summary of the fields to be imported from the text file. The structure, molecular weight, molecular formula, and compound names of each entry are displayed by default (*see* Note 7). The “Monitor import” window will give a summary of the imported data (*see* Note 8). Once fully processed, click on “Finish”.
5. Use Instant JChem to determine values for physicochemical properties by selecting Data > New Chemical Term Field. Use the “Expression” drop-down menu to choose preset chemical terms (*see* Notes 9 and ¹⁰). Enter an appropriate Name for the column and then click on “Finish” (Fig. 2). For this example, we selected the following 16 terms (Instant JChem input syntax follows each description) (*see* Notes 11 and ¹²):

Molecular weight (MW): mass()

N atom count (N): atomCount(“7”)

O atom count (O): atomCount(“8”)

H-bond donor count (HBD): donorCount()

H-bond acceptor (HBA): acceptorCount()

Rotatable bond count (RotB): rotatableBondCount()

Stereocenter count (nStereo): chiralCenterCount()

⁵Make sure that “IJC Project (with local database)” is highlighted in the “Projects” panel.

⁶Make sure that “localdb [as admin]” is highlighted in the “Projects and schemas” panel.

⁷Remove any undesired fields using the “< Remove” button. Click on “Next” when finished.

⁸Any error messages are saved to a file that can be opened with a text editor for review.

⁹Each chemical term must be added individually.

¹⁰User-defined chemical terms can be saved to the “Expression” menu by clicking on the yellow star to the right of the menu and naming the new term for convenient use in the future.

¹¹The order of the columns can be changed by clicking and dragging on the header row of a column.

¹²We selected the 20 physicochemical parameters used in this example based on several criteria. First, Lipinski parameters (MW 500, logP 5, HBA 10, HBD 5) [14], and Veber parameters (RotB 10, tPSA 140 Å²) [10], have been correlated with oral bioavailability. These parameters partially correlate to cell permeability [11], which is relevant to the utility of new chemical probes discovered from library screening. Second, Tetko’s calculated logS solubility (ALOGpS) [15] was included because compound solubility is critical in screening and is often problematic for commercial drug-like libraries. ALOGPs was included as an alternative method for estimating solubility, and logD was included to approximate aqueous solubility at a physiological relevant pH (7.4). Third, several stereochemical parameters (nStereo, nStMW, Fsp³) were included as approximations of three-dimensional complexity, which have been shown to be a distinguishing factor between synthetic drugs and natural products [16, 17], and also impact protein binding selectivity and frequency [18]. Fourth, relative polar surface area (relPSA) was included because it has been shown to be a distinguishing factor between Gram-positive and Gram-negative antibiotics [19]. Finally, additional physicochemical parameters (N, O, Rings, RngAr, RngSys, RngLg, RRSys) were included because they have been found to differentiate synthetic drugs from natural products [16].

Topological polar surface area (tPSA): PSA()

Number of rings (Rings): ringCount()

Aromatic ring count (RngAr): aromaticRingCount()

Ring system count (RngSys): ringSystemCount()

Size of largest ring (RngLg): largestRingSize()

Fraction of sp³-hybridized carbons (Fsp3): count(filter('atno()==6 && connections()==4'))/atomCount("6")

n-Octanol/water partition coefficient at pH = 7.4 (LogD): logd('7.4')

van der Waals surface area (VWSA): vanDerWaalsSurfaceArea()

Relative polar surface area (relPSA): PSA()/vanDerWaals SurfaceArea()

- Export the table of physicochemical parameters calculated in Instant JChem to an MS Excel file (.xls) by selecting File > Export to File. In the "Specify details" window, click on the purple folder to the right of the "File" field. Name the file and define the file format as "Microsoft Office Excel Workbook (*.xls)". The following window gives the user an option to remove or rearrange columns in the exported file (*see* Note 13). The "Monitor progress" window summarizes the export process. When complete, click on "Finish". Open the .xls file containing these physicochemical parameters in MS Excel; additional physicochemical parameters will be added later (Subheading 3.1, **steps 8 and 9**).

- The following two physicochemical values are calculated using the VCC Lab Website (<http://www.vcclab.org/lab/alogps/start.html>) (*see* Notes 12 and ¹⁴):

n-Octanol/water partition coefficient alt (ALOGPs)

Tetko's logS aqueous solubility (ALOGpS)

To calculate ALOGPs and ALOGpS from the website's window, choose "Upload file" and select "Smiles—SMILES file—default" from the drop-down menu and click on "Proceed with file uploading" (Fig. 3). Click on "Choose file" and select the .txt file consisting of SMILES codes only that was created in Subheading 3.1, **step 2**. Click on "Upload file" and a new pop-up window should be displayed that states "Your file "yourfile.txt" was uploaded successfully." Close the pop-up window displaying this message and a new window will open that is entitled "results.txt". Copy the text from this results window and paste it into a new MS Excel file.

- In the MS Excel sheet/tab that contains the physicochemical parameters calculated by Instant JChem (Subheading 3.1, **step 6**), add two new columns labeled "ALOGPs" and "ALOGpS." Copy the "logP" and "logS" data columns calculated by the VCC Lab website, and paste them into the "ALOGPs" and "ALOGpS"

¹³The Structure column may be removed for faster processing, consistency of row height, and a cleaner appearance in MS Excel.

¹⁴The browser's pop-up blocker must be turned off.

columns, respectively. Close the MS Excel file that contains only the VCC Lab data.

9. The remaining two physicochemical parameters are calculated in MS Excel (*see* Note 12):

$$nStereo \div MW, \text{ stereochemical density (nStMW)}$$

$$Rings \div RingSys, \text{ ring complexity (RRSys)}$$

Create two new columns in the MS Excel file that contains the other physicochemical values. For the nStMW and RRSys columns, set the column's formula according to its respective equation (Fig. 4):

$$nStMW = [\text{column for } nStereo]/[\text{column for } MW]$$

$$RRSys = \text{IF}([\text{column for } RingsSys] = 0, 0, [\text{column for } Rings]/[\text{column for } RingSys]) \text{ (see Note 15).}$$

All of the physicochemical parameters needed for this PCA are now in the MS Excel file.

3.2 Principal Component Analysis

1. In the MS Excel file containing the compound names and physicochemical parameters that was created in Subheading 3.1, verify that the compounds are grouped by compound class (such as Drugs, Natural Products), and insert a new row below each group. Each new row will represent the average compound for a given class. Accordingly, name each new row based on the category it will represent, for example "AVG Drug". For these new rows, fill the cells associated with structural and physicochemical parameter values using the "AVERAGE" function to the left of the cell formula field and select the appropriate cells. Similarly, add two new rows below the last compound and calculate the mean ("AVERAGE") of each physicochemical parameter for the entire dataset (all compounds), as well as the standard deviation of each parameter using the "STDEV" function. Name the MS Excel sheet/tab "Raw".
2. Create a new sheet/tab within the same MS Excel file, naming it "Norm" to contain mean normalized values of the physicochemical parameters. Copy the compound names from the row header and physicochemical descriptors from the column header of the "Raw" sheet/tab and paste them into this "Norm" sheet/tab. Fill each standardized physicochemical parameter cell with the Normval value calculated using the equation

$$\text{Normval} = ([\text{val}] - [\text{column mean}])/[\text{column standard deviation}]$$

where "val" is the value from the corresponding cell in the "Raw" sheet/tab. The number format (Format > Cells...) should be set to four decimal places. Save the MS Excel file. Now, save only the "Norm" sheet/tab as "Data.txt" (Text-Tab

¹⁵This formula avoids errors caused by a zero in the denominator.

Delimited) on the Desktop (Mac). Close the MS Excel file and discard the changes to the file format.

3. Launch the “R” open-source computing package and run the following PCA commands (at the command prompt “R>”) (*see* Note 16):

```
R > read.table("~/Desktop/Data.txt") -> a
```

```
R > prcomp(a) -> b
```

```
R > summary(b)
```

This command gives a table showing the distribution of variance from the full dataset on each principal component (Fig. 5). PCA generates as many principal components as there are parameters, but importantly, the majority of variance is represented in the first few components (*see* refs. 1, 2 for further discussion on PCA and variance). In this example, the first three principal components (PC1–PC3) retain 75 % of the variance from the 20-dimensional dataset. As such, PC1–PC3 can be used to construct a set of two-dimensional plots that will allow the visualization of the data in a more intuitive manner while still retaining the majority of the information from the full dataset. We will therefore focus our remaining analysis on PC1–PC3.

4. To obtain loading information for each structural and physicochemical parameter, continue PCA with the following command in “R”:

```
R > b
```

The resulting table can be copied into MS Excel for convenient access (*see* Note 17) (Fig. 6). The loading data will become useful for directing future library design and will be discussed in Subheading 3.3.

5. Obtain loading plots in “R” using the following command:

```
R > biplot(b, choices = c(1, 2), col = c("gray", "red"))
```

¹⁶For Windows users, the read.table command is slightly different:

```
R > read.table("file path \\Data.txt", header = T, sep="\t", row.names = 1) -> a
```

where *file path* is the entire file path beginning with the drive (usually C:\\). Users can obtain the file path by dragging and dropping the text file directly into R, which returns an error message but reports the file location including the drive.

¹⁷To transfer the “R” output to MS Excel, copy the first section of the table without the column headers (for example the PC1–PC8 data before the first section break) and paste it into an MS Word document. Change the font to Courier and the size to 5 pt such that the text in the document resembles a table. Save the file as a Text-only (.txt) file. From MS Excel, import the data using Data > Get External Data. Select the file, then choose the “Fixed width” option in Step 1 of the Text Import Wizard, click on “Next”, verify column divisions, and then click on “Finish”.

The plot produced illustrates loading data of structural and physicochemical parameters for PC1 vs. PC2 (*see* Notes 18 and ¹⁹) (Fig. 7). Save the loading plot and name it appropriately (*see* Note 20).

6. Generate loading plots of PC1 vs. PC3 and PC3 vs. PC2, using the following commands (and saving each loading plot):

```
R > biplot(b, choices = c(1, 3), col = c("gray", "red"))
```

```
R > biplot(b, choices = c(3, 2), col = c("gray", "red"))
```

These loading plots give insight into the structural and physicochemical parameters that most distinguish a set of compounds by visually displaying each physicochemical parameter's influence on where a compound will appear in the final PCA plots. This information is also valuable to the planning of future libraries and is discussed in Subheading 3.3.

7. In "R", use the following commands to obtain the rotated PCA data (scores) and save the output as a text file (*see* Note 21):

```
R > b$x -> c
```

```
R > write.table(c, "~/Desktop/scores.txt", sep = "\t")
```

8. Open the MS Excel file that contains the "Raw" and "Norm" sheets/tabs (Subheading 3.2, **step 2**), and create a new sheet/tab within that MS Excel file, naming it "PCA". To transfer the scores data obtained from "R" to MS Excel, first open the scores.txt file in a text editor. At the beginning of the document, add a column header such as "Compound" followed by a tab (Fig. 8). Next, copy all of the text in the file and paste it into the MS Excel sheet/tab named "PCA". Change the number format (Format > Cells...) of the PCA cells to three decimal places.
9. In MS Excel, plot PC1 vs. PC2 from the "PCA" sheet/tab by selecting the columns and clicking on the "Chart Wizard" icon in the Standard Toolbar (View > Toolbars > Standard). Under "Standard Types" choose "Scatter XY" and click on "Next". Enter series information (e.g., Drugs, Natural Products) under the "Series" tab and fill the X- and Y-values data fields with corresponding range for each series (for example PC1 data for X-values and PC2 data for Y-values). When done entering the series information, click on "Next". Enter a title for the plot and labels for the axes and then click on "Next." Select "As object in PCA" and click on "Finish" (*see* Note 22).

¹⁸Several PC axes were inverted in this example to maintain resemblance to our previous PCA plots [12, 13] by adding the "ylim" and "xlim" axis limit options to the "biplot" command in "R":

```
R > biplot(b, choices = c(1, 2), col = c("gray", "red"), ylim = c(0.12, -0.12), xlim = c(-0.12, 0.12))
```

This does not impact data interpretation because the signs of all PC axes are arbitrary.

¹⁹If desired, loading plots can also be produced where the scores (compound names) are hidden. To do this, replace "gray" with "white" in the biplot command.

²⁰The plot must be saved before an additional plot command is entered.

²¹For Windows users, the file path information in the write.table command will be different and needs to include the drive (such as C:\\) (cf. Note 16).

²²If desired, change the appearance of the plot by right-clicking on the object you wish to modify, and then select the Format option.

10. Follow a procedure similar to that described in Subheading 3.2, **step 9**, to generate plots for PC1 vs. PC3 and PC3 vs. PC2 using the data from the appropriate columns.
11. Copy the plots produced in MS Excel and use MS PowerPoint or graphical editing software to add colored ovals that encompass the majority of data points for a given class of compounds (Fig. 9).

3.3 Using PCA to Guide Library Design

1. Examine the PCA plots together with the loading plots to identify structural and physicochemical parameters that determine where a particular compound or a collection of compounds appears on the PCA plot. For example, many natural products appear to the left (negative PC1) of the library members in the PC1 vs. PC2 plot (Fig. 9). The corresponding loading plot (Fig. 7) indicates that HBA, tPSA, and O are all major components of PC1. Recalling that each PC axis is a linear combination of structural and physicochemical parameters, Note that the coefficients for each parameter used for a given PC axis were also obtained in Subheading 3.2, **step 4** (Fig. 6). This table provides more quantitative information regarding the parameters that have the greatest impact on the location of a compound with respect to each PC axis.
2. Consider the structural and physicochemical parameters that are the most important in differentiating a collection of compounds from the targeted region of chemical space. In this example, O, HBD, HBA, tPSA, nStereo, aqueous solubility, and Fsp³ are among the most influential in distinguishing our library compounds from natural products. The introduction of additional oxygen atoms and additional stereocenters to our library compounds would likely lead to more natural product-like compounds. A dihydroxylation of the olefins contained in our medium-ring library members would address all of the parameters mentioned above and should result in compounds that are shifted towards natural products in our PCA plots. Leveraging this information, we proceed with the dihydroxylation of multiple medium-ring compounds to produce a collection of PCA-directed derivatives of our initial medium-ring library.
3. Include the collection of modified compounds in a new analysis to evaluate the effectiveness of the PCA-directed modifications in targeting the desired region of the plot (Fig. 10). In this example, our diol products have structural and physicochemical parameters that are more consistent with natural products compared to their parent olefins, and as a result, the compounds are shifted towards natural products in all of the PCA plots. Reiterate this process as necessary to provide a library with the desired structural and physicochemical properties.

Acknowledgments

We thank Tony D. Davis (MSKCC) for suggesting inclusion of the logD, van der Waals surface area, and relative polar surface area parameters, and for providing modifications of this protocol for Windows users. Instant JChem was generously provided by ChemAxon. Financial support from the NIH (P41 GM076267 to D.S.T., P41 GM076267-03S1 to R.A.B., T32 CA062948-Gudas to T.A.W.), Starr Foundation, Tri-Institutional Stem Cell

Initiative, Alfred P. Sloan Foundation (Research Fellowship to D.S.T.), Deutscher Akademischer Austauschdienst (DAAD, postdoctoral fellowship to F.K.), William H. Goodwin and Alice Goodwin and the Commonwealth Foundation for Cancer Research, and the MSKCC Experimental Therapeutics Center is gratefully acknowledged.

References

1. Jolliffe, IT. Principal component analysis. Springer; New York, NY: 2002.
2. Jackson, JE. A user's guide to principal components. Wiley; Hoboken, NJ: 2003.
3. Akella LB, DeCaprio D. Cheminformatics approaches to analyze diversity in compound screening libraries. *Curr Opin Chem Biol.* 2010; 14:325–330. [PubMed: 20457001]
4. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? *Nat Rev Drug Discov.* 2006; 5:993–996. [PubMed: 17139284]
5. Newman DJ, Cragg GM. Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J Nat Prod.* 2012; 75:311–335. [PubMed: 22316239]
6. Sánchez-Pedregal VM, et al. The tubulin- bound conformation of discoder-molide derived by NMR studies in solution supports a common pharmacophore model for epothilone and discodermolide. *Angew Chem Int Ed.* 2006; 45:7388–7394.
7. Canales A, et al. The bound conformation of microtubule-stabilizing agents: NMR insights into the bioactive 3D structure of discodermolide and dictyostatin. *Chem Eur J.* 2008; 14:7557–7569. [PubMed: 18449868]
8. Knust J, Hoffmann RW. Synthesis and conformational analysis of macrocyclic dilactones mimicking the pharmacophore of aplysiatoxin. *Helv Chim Acta.* 2003; 86:1871–1893.
9. Khan AR, et al. Lowering the entropic barrier for binding conformationally flexible inhibitors to enzymes. *Biochemistry.* 1998; 37:16839–16845. [PubMed: 9836576]
10. Veber DF, et al. Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem.* 2002; 45:2615–2623. [PubMed: 12036371]
11. Rezaei T, et al. Testing the conformational hypothesis of passive membrane permeability using synthetic cyclic peptide diastereomers. *J Am Chem Soc.* 2007; 128:2510–2511. [PubMed: 16492015]
12. Kopp F, et al. A diversity-oriented synthesis approach to macrocycles via oxidative ring expansion. *Nat Chem Biol.* 2012; 8:358–365. [PubMed: 22406518]
13. Bauer RA, Wenderski TA, Tan DS. Biomimetic diversity-oriented synthesis of benzannulated medium rings via ring expansion. *Nat Chem Biol.* 2013; 9:21–29. [PubMed: 23160003]
14. Lipinski CA, et al. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev.* 1997; 23:3–25.
15. Tetko IV, et al. Estimation of aqueous solubility of chemical compounds using E-state indices. *J Chem Inf Comput Sci.* 2001; 41:1488–1493. [PubMed: 11749573]
16. 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. [PubMed: 12546556]
17. Lovering F, Bikker J, Humblet C. Escaping from flatland: increasing saturations as an approach to improving clinical success. *J Med Chem.* 2009; 52:6752–6756. [PubMed: 19827778]
18. Clemons PA, et al. Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc Natl Acad Sci U S A.* 2010; 107:18787–18792. [PubMed: 20956335]
19. O'Shea R, Moser HM. Physicochemical properties of antibacterial compounds: implications for drug discovery. *J Med Chem.* 2008; 51:2871–2878. [PubMed: 18260614]
20. McGrath NA, Brichacek M, Njardarson JT. A graphical journey of innovative organic architectures that have improved our lives. *J Chem Educ.* 2010; 87:1348–1349. Also see also: Njardarson Group—Top top-selling drugs Pharmaceuticals poster; <http://cbc.arizona.edu/njardarson/group/top-pharmaceuticals-poster>.

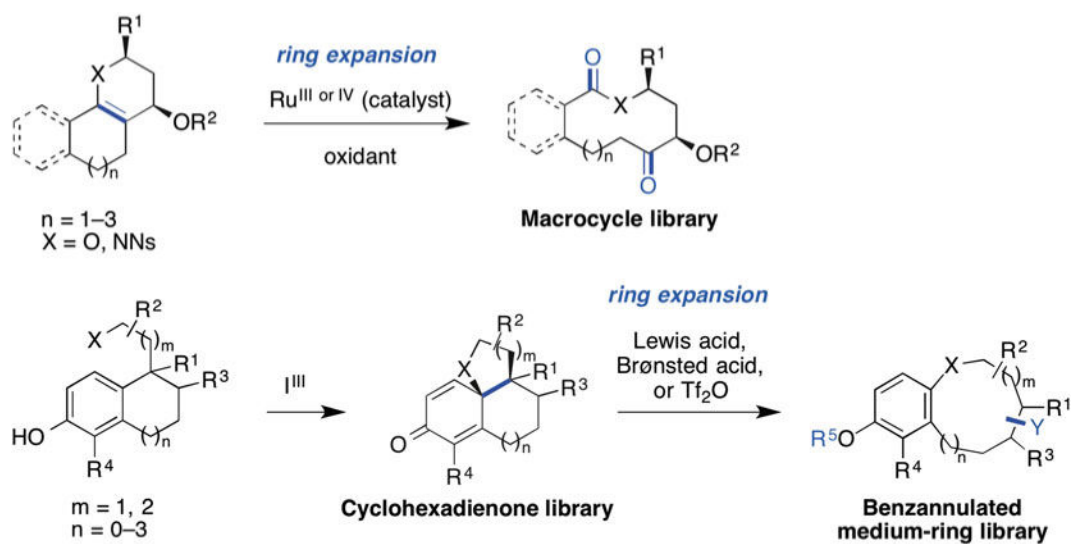
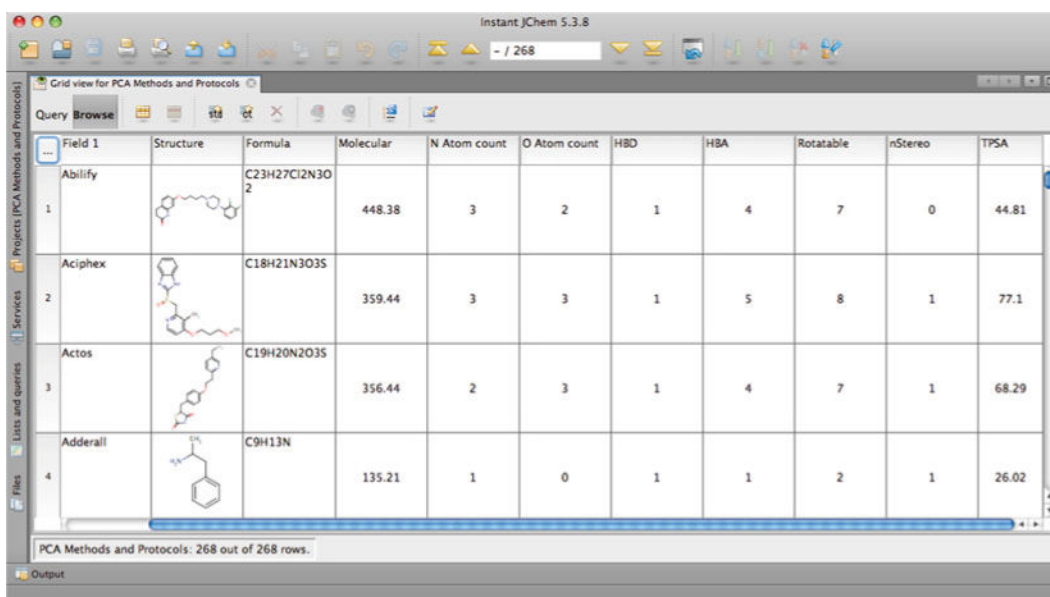


Fig. 1. Recently developed routes to natural product-like macrocycle and medium-ring libraries using ring expansion strategies







Field 1	Structure	Formula	Molecular	N Atom count	O Atom count	HBD	HBA	Rotatable	nStereo	TPSA
1		C ₂₃ H ₂₇ Cl ₂ N ₃ O ₂	448.38	3	2	1	4	7	0	44.81
2		C ₁₈ H ₂₁ N ₃ O ₃ S	359.44	3	3	1	5	8	1	77.1
3		C ₁₉ H ₂₀ N ₂ O ₃ S	356.44	2	3	1	4	7	1	68.29
4		C ₉ H ₁₃ N	135.21	1	0	1	1	2	1	26.02

Fig. 2. Instant JChem table showing selected chemical terms (physicochemical parameters) used for PCA

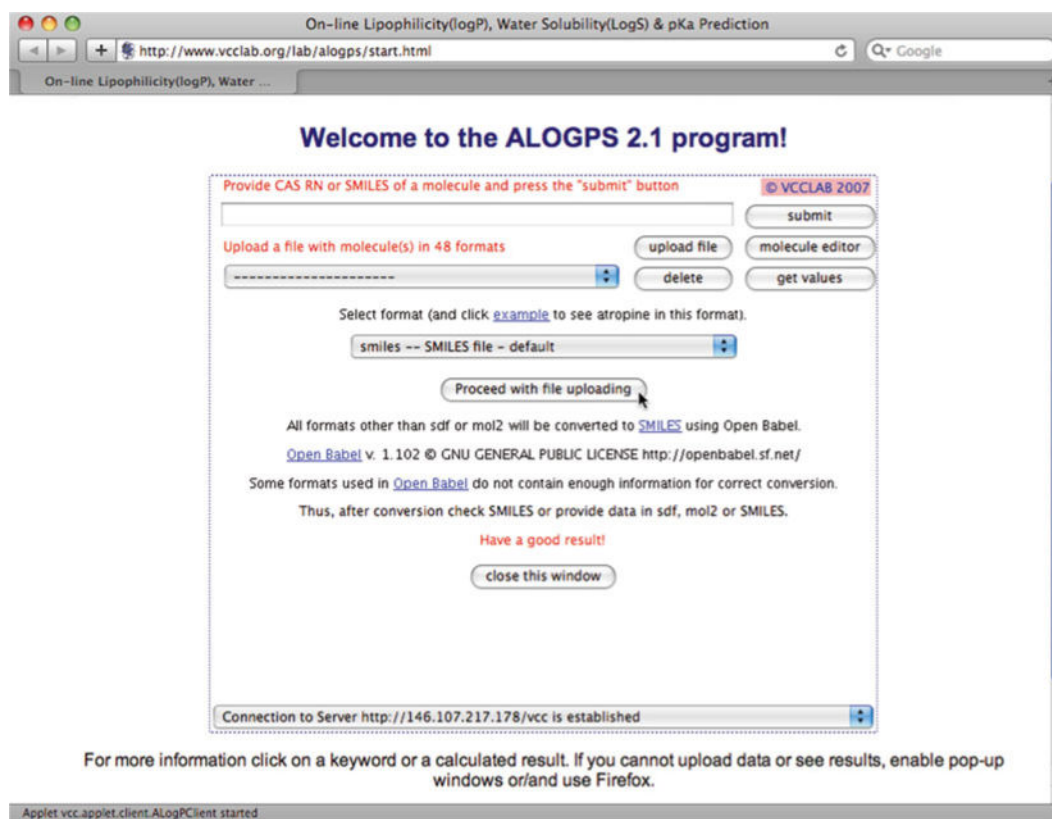
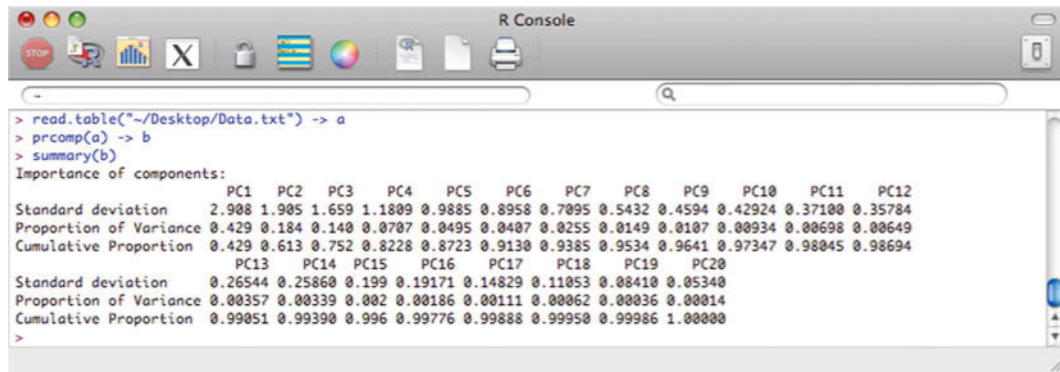


Fig. 3. Uploading SMILES codes to calculate ALogpS and ALogPs at the VCC Lab Website

	M	N	O	P	Q	R	S	T
1 Field 1	Rings	RngAr	RngSys	RngLg	ALOGPs	ALOGpS	nStMW	RRSys
2 Abilify	4	2	3	6	5.21	-4.76	0.0000000	=IF(O2=0,0,M2/O2)
3 Aciphex	3	3	2	6	2.04	-3.03	0.0027821	1.50
4 Actos	3	2	3	6	3.17	-4.91	0.0028055	1.00
5 Adderall	1	1	1	6	1.85	-1.89	0.0073961	1.00
6 Ambien	3	3	2	6	3.15	-3.99	0.0000000	1.50
7 Avandia	3	2	3	6	2.95	-3.97	0.0027978	1.00
8 Benazepril	3	2	2	7	1.14	-4.61	0.0047115	1.50
9 Celebrex	3	3	3	6	3.99	-4.88	0.0000000	1.00
10 Concerta	2	1	2	6	1.47	-3.11	0.0085724	1.00
11 Coreg	4	4	2	6	3.05	-4.96	0.0024602	2.00
12 Crestor	2	2	2	6	1.47	-3.74	0.0041534	1.00
13 Cymbalta	3	3	2	6	4.72	-5	0.0033623	1.50
14 Diovan	3	3	3	6	3.68	-4.27	0.0022961	1.00
15 Effexor	2	1	2	6	2.69	-3.08	0.0036049	1.00
16 Flonase	4	0	1	6	3.69	-4.64	0.0179795	4.00
17 Fosamax	0	0	0	0	-1.34	-1.17	0.0000000	0.00
18 Imitrex	2	2	1	6	1.17	-3.37	0.0000000	2.00
19 Lamictal	2	2	2	6	1.87	-2.72	0.0000000	1.00
20 Levaquin	4	2	2	6	-0.02	-2.4	0.0027673	2.00
21 Levsin	2	2	2	6	2.58	4.74	0.0020027	1.50

Fig. 4.
Final physicochemical parameters calculated in MS Excel, with the RRSys calculation in cell T2 and the referenced cells M2 and O2 highlighted as an example



```
> read.table("~/Desktop/Data.txt") -> a
> prcomp(a) -> b
> summary(b)
Importance of components:
      PC1  PC2  PC3  PC4  PC5  PC6  PC7  PC8  PC9  PC10  PC11  PC12
Standard deviation  2.908 1.905 1.659 1.1809 0.9885 0.8958 0.7095 0.5432 0.4594 0.42924 0.37100 0.35784
Proportion of Variance 0.429 0.184 0.140 0.0707 0.0495 0.0407 0.0255 0.0149 0.0107 0.00934 0.00698 0.00649
Cumulative Proportion 0.429 0.613 0.752 0.8228 0.8723 0.9130 0.9385 0.9534 0.9641 0.97347 0.98045 0.98694
      PC13  PC14  PC15  PC16  PC17  PC18  PC19  PC20
Standard deviation  0.26544 0.25860 0.199 0.19171 0.14829 0.11053 0.08410 0.05340
Proportion of Variance 0.00357 0.00339 0.002 0.00186 0.00111 0.00062 0.00036 0.00014
Cumulative Proportion 0.99051 0.99390 0.996 0.99776 0.99888 0.99950 0.99986 1.00000
>
```

Fig. 5.
“R” output showing the variance retained in each successive principal component of the PCA

	PC1	PC2	PC3
MW	-0.3207	-0.1482	0.1008
N	-0.2080	-0.1055	-0.3331
O	-0.3180	0.0179	0.1159
HBD	-0.3087	0.0677	-0.1180
HBA	-0.3346	-0.0057	-0.0256
RotB	-0.2587	-0.1209	-0.0540
tPSA	-0.3336	0.0109	-0.0818
nStereo	-0.2789	0.0936	0.2882
nStMW	-0.1612	0.2880	0.2964
Rings	-0.1669	-0.2384	0.0698
RngAr	-0.0475	-0.3660	-0.3227
RngSys	-0.1761	-0.2315	-0.1506
RngLg	-0.1459	-0.0128	0.2521
RRSys	-0.0015	-0.0362	0.1920
ALOGPs	0.0859	-0.4127	0.2677
ALOGpS	0.0378	0.4319	-0.2060
Fsp3	-0.0910	0.3464	0.3224
LogD	0.1850	-0.2569	0.3114
relPSA	-0.1797	0.2262	-0.3112
VWSA	-0.3122	-0.1253	0.1626

Fig. 6. Coefficients of each parameter used for PC1–PC3, highlighting those of the greatest magnitude

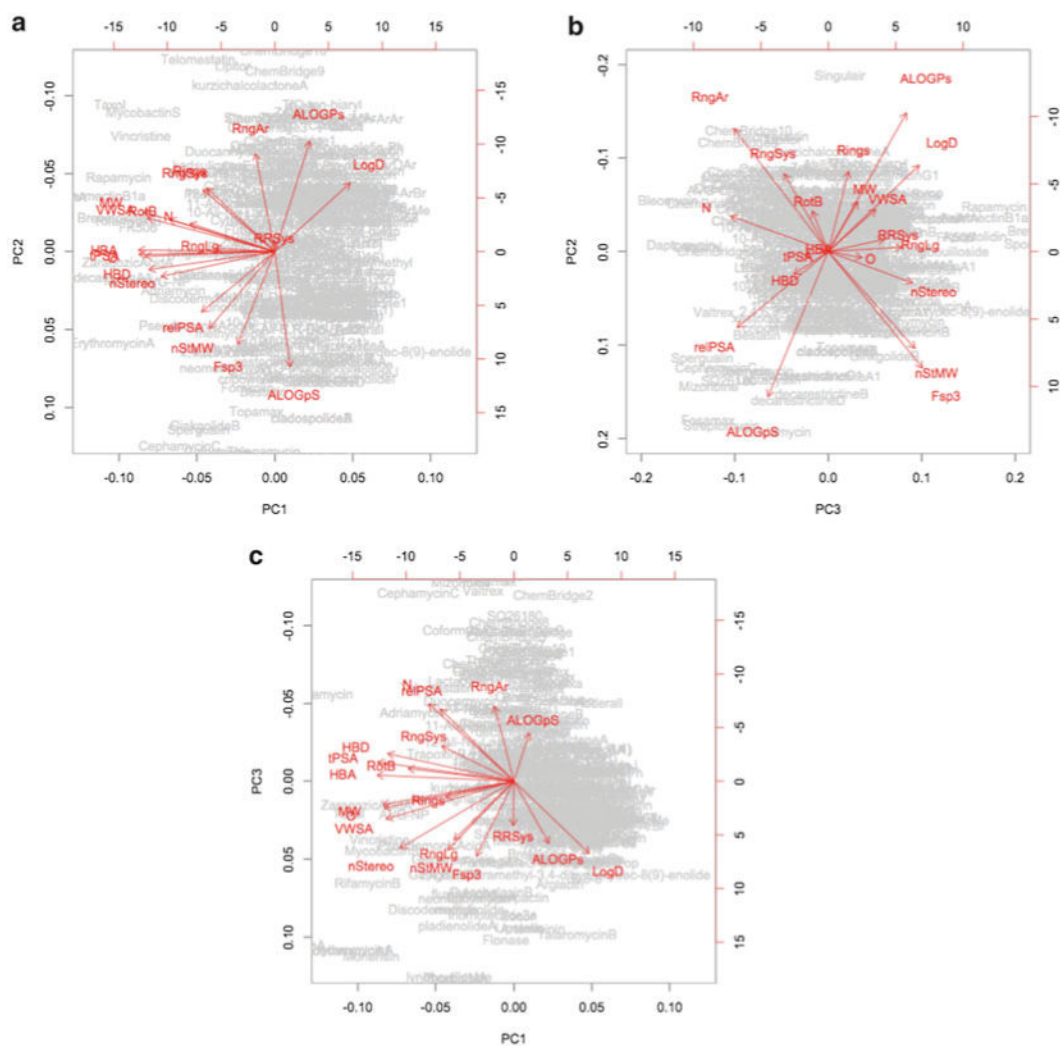
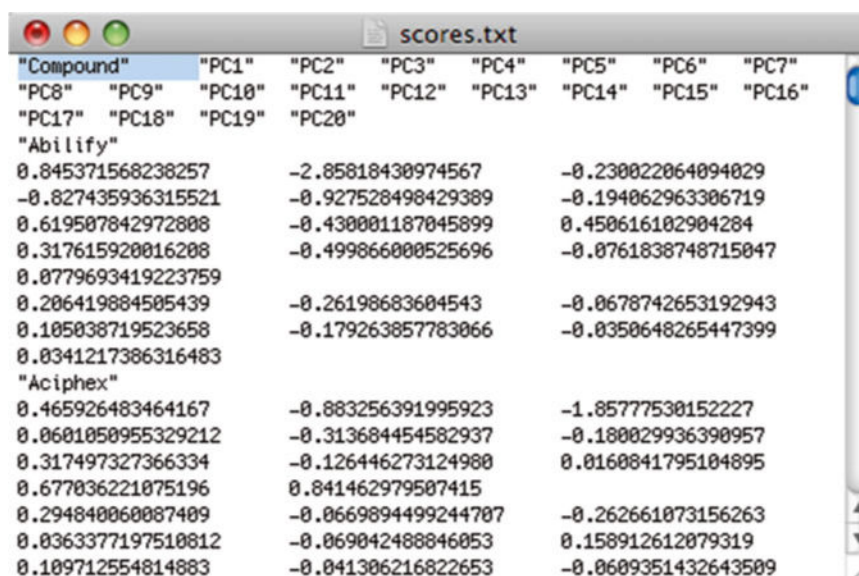


Fig. 7. Loading plots from PCA: **(a)** PC1 vs. PC2, **(b)** PC2 vs. PC3, **(c)** PC1 vs. PC 3



"Compound"	"PC1"	"PC2"	"PC3"	"PC4"	"PC5"	"PC6"	"PC7"	"PC8"	"PC9"	"PC10"	"PC11"	"PC12"	"PC13"	"PC14"	"PC15"	"PC16"	"PC17"	"PC18"	"PC19"	"PC20"		
"Abilify"	0.845371568238257	-2.85818430974567			-0.230022064094029			-0.827435936315521		-0.927528498429389			-0.194062963306719				0.619507842972808		-0.430001187045899		0.450616102904284	
	0.317615920016208	-0.499866000525696			-0.0761838748715047			0.0779693419223759										0.206419884505439		-0.26198683604543		-0.0678742653192943
	0.105038719523658				-0.0350648265447399			0.0341217386316483		-0.179263857783066												
"Aciphex"	0.465926483464167	-0.883256391995923			-1.85777530152227			0.0601050955329212		-0.313684454582937			-0.180029936390957				0.317497327366334		-0.126446273124900		0.0160841795104895	
	0.677036221075196	0.841462979507415						0.294840060087409		-0.0669894499244707			-0.262661073156263									
	0.0363377197510812	-0.069042488846053			0.158912612079319			0.109712554814883		-0.041306216822653			-0.0609351432643509									

Fig. 8.
Modified PCA scores.txt file including a header for compound names

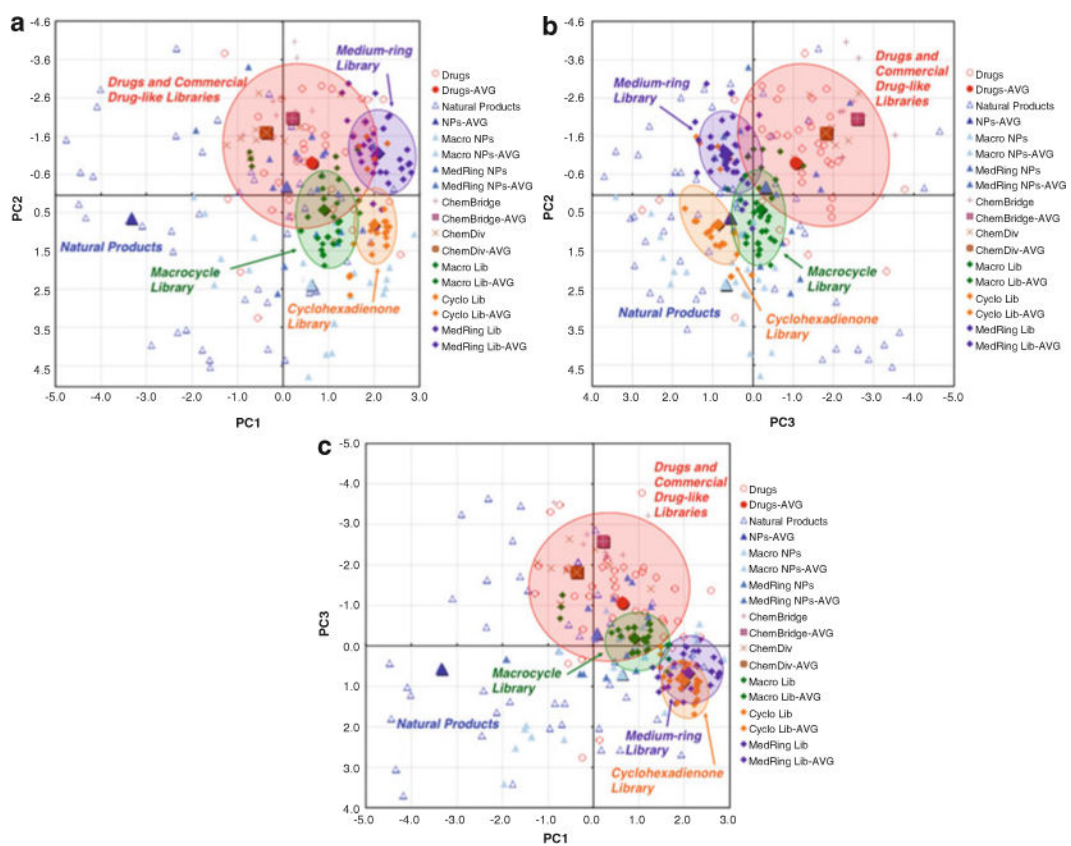


Fig. 9.
Completed PCA plots: (a) PC1 vs. PC2, (b) PC2 vs. PC3, (c) PC1 vs. PC 3

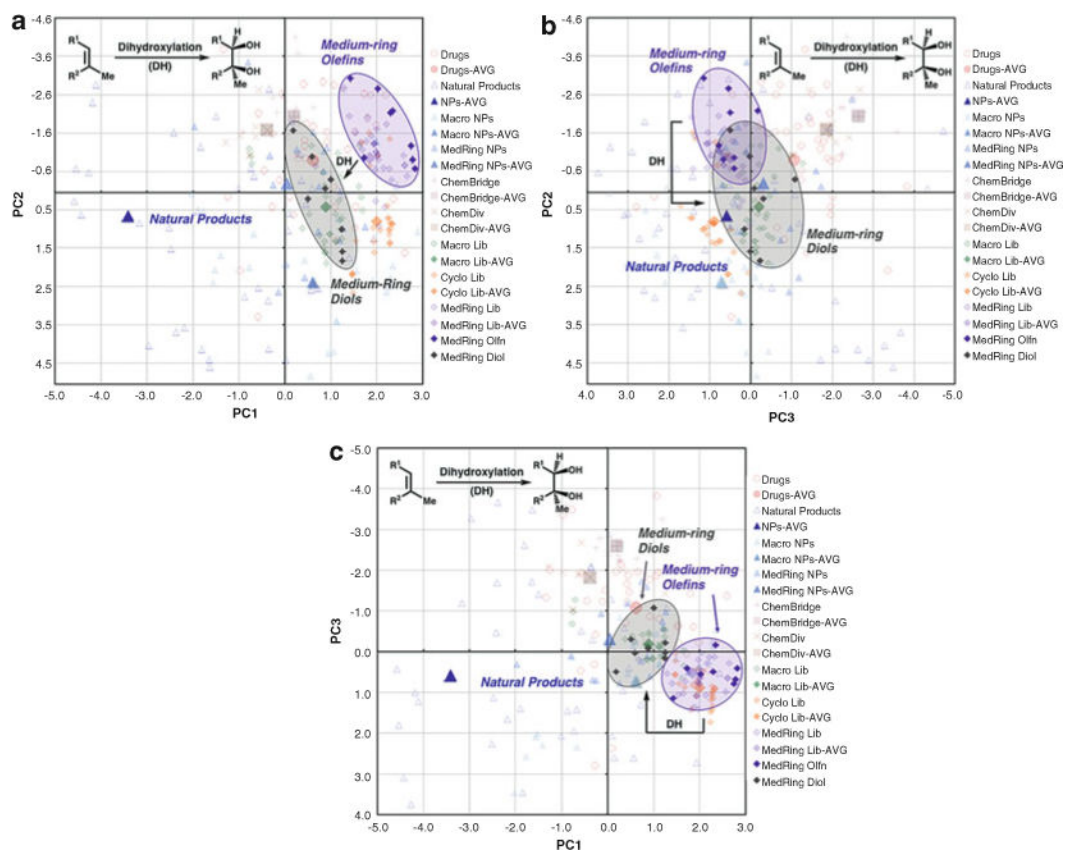


Fig. 10. Completed PCA plots: (a) PC1 vs. PC2, (b) PC2 vs. PC3, (c) PC1 vs. PC 3. DH = dihydroxylation