Compound Library & Informatics for HTS Screening / LEADID_HTS
Scripps Research Institute Molecular Screening Center

Louis Scampavia
Associate Director and Associate Prof. of Molecular Medicine
Director of HTS Chemistry and Technologies
scampl@scripps.edu

http://hts.florida.scripps.edu/

January 2017
Provide an overview and understanding:

1. Compound management practices for HTS
2. Informatics used to curate chemical libraries
3. Cheminformatics tutorial and tools
4. Compound selection rules
5. Screening Libraries available
   - Diversity sets
   - Focus target sets
6. Alternatives to traditional HTS campaigns
   - Fragment-based Lead Development [FBDD or FBLD]
   - Fragment-based assisted HTS
7. Resources for:
   - Assay development
   - Cheminformatic tools
   - Compound sources
Our goal
The Lead Identification Department, in collaboration with other scientists at Scripps and around the world, identifies compounds that can help better understand and ultimately cure diseases such as cancer, diabetes, Alzheimer’s disease and hepatitis.

How do we do it?
By combining robotics, biology, chemistry and informatics, the Lead Identification Department is able to perform High-Throughput Screening, a robotic process that tests hundreds of thousands of drug-like compounds for biological activity both rapidly and economically.

Why use robotics?
In order to screen large collections of compounds for drug activity economically, biological tests (also known as “assays”) are miniaturized. Miniaturized assays also generate less waste since they use small amounts of materials. Since miniaturized assays are difficult to perform manually, they require the use of specialized robots and microfluidic devices. Moreover, the use of robotics allows us to rapidly produce consistent, high quality assay data.
PLATE AND COMPOUND ANALYTICS
• Dedicated Kalypsys-GNF robotic platform for library storage & cherry-picking
• Capable of storing >2.9 million samples in 1536-well plates
• Platform capable of performing >2,000 cherry-picks/day

• “Offline” compound management automation in a dedicated lab
• Routinely used for sample preparation, dissolution, retrieval & storage
• Rapid, flexible 96/384/1536 tube & plate replication, reformatting, re-arraying
• All processes tracked by barcode and logged in Scripps’ LIMS
Lead ID LC-MS Platform: Automated Sample QC

Auto-sampler:
- 96-, 384-, 1536-well plate formats
- Sealed or unsealed plates
- Temperature controlled

Liquid Chromatography:
- 5-min analysis time
- Sample size ~1uL
- Full UV spectral analysis (190nm-400nm)

Mass Spectrometer:
- Multimode ES-APCI ionization
- Positive/negative ion

ELSD (Evaporative Light Scattering Detector):
- Mass-based detection

Software Automation (Virscidian):
- Data analysis
- Summary reports
- Database archives

Unambiguous QC of compound ID, structure, molecular mass & sample purity

Rapid: 280 samples/day
Low volume: 1uL/test

Automated analysis provided by manufacturer...
Instrument “disconnected” from LIMS
Lead ID Plate Auditor: Sample QA/QC

- Incorporates machine vision, image analysis and analytical spectroscopy
- Automatically and rapidly (<1 min) identifies and annotates issues specific to compound libraries:
  - Empty
  - Colored
  - Precipitate
  - Partial
  - Crystallization
  - Full

- Measurements are non-contact, non-destructive
- Results db integrated with Scripps LIMS

Scripps routinely uses the “general” and “purity” fields for added comments.

<table>
<thead>
<tr>
<th>General Comments</th>
<th>Purity Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>isomers</td>
<td>multiple peaks</td>
</tr>
<tr>
<td>No peaks detected: MS Inconclusive</td>
<td>Pure cmpd; MS inconclusive</td>
</tr>
<tr>
<td>empty well</td>
<td>Pure cmpd; Mole Mass inconclusive</td>
</tr>
<tr>
<td></td>
<td>Pure cmpd; Mole Mass incorrect</td>
</tr>
<tr>
<td></td>
<td>Pure cmpd; Mole Mass out of MS range</td>
</tr>
</tbody>
</table>

Comment Category Manual/Auto inputs:

- Overrides
- Integration
- Processing
- General
- Purity
- Quant
- Flag
Definitions

**SMILES:** The simplified molecular-input line-entry system (SMILES) is a specification in form of a line notation for describing the structure of chemical species using short ASCII strings. SMILES strings can be imported by most molecule editors for conversion back into two-dimensional drawings or three-dimensional models of the molecules.

**MOL files (MOLE):** An MDL Molfile is a file format for holding information about the atoms, bonds, connectivity and coordinates of a molecule.

**SD-files (SDF):** Structure Data files applies MOL file Format for multiple compounds delimited by four dollar signs $$$

**Important note:** These formats provide the means for database management and in silico analysis of compounds. Software can often be used to change formats (e.g. SMILES to MOL or MOL files to SD-files. Vital when ordering compounds is to receive MOL/SDF from vendors to avoid structural corruptions (e.g. chirality errors).
<table>
<thead>
<tr>
<th>Molecule</th>
<th>Structure</th>
<th>SMILES Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dinitrogen</td>
<td>N≡N</td>
<td>N≡N</td>
</tr>
<tr>
<td>Methyl isocyanate (MIC)</td>
<td>CH₃-N≡C=O</td>
<td>CN≡C=O</td>
</tr>
<tr>
<td>Copper(II) sulfate</td>
<td>Cu²⁺ SO₄²⁻</td>
<td>[Cu²⁺][SO₄²⁻]</td>
</tr>
<tr>
<td>Cantharidin (C₁₇H₂₂O₂)</td>
<td><img src="image1.png" alt="Image" /></td>
<td>CCC(O@@H)OCC(C=O)C=C=CC#C=O=C=O</td>
</tr>
<tr>
<td>Pyrethrin II (C₂₂H₂₆O₃)</td>
<td><img src="image2.png" alt="Image" /></td>
<td>COC(=O)C(C=O)C=CC=C=C(C=O)C=C=H2C(C=O)C(=O)(C)2CC=CC=C</td>
</tr>
<tr>
<td>Aflatoxin B1 (C₁₇H₁₂O₉)</td>
<td><img src="image3.png" alt="Image" /></td>
<td>O1C=O[Ca@H][Ca@H]1O2k3c2cc(OC)c4c3OC(=O)C5=C4CCC(=O)5</td>
</tr>
<tr>
<td>Glucose (glucopyranose) (C₆H₁₂O₆)</td>
<td><img src="image4.png" alt="Image" /></td>
<td>CCC(O@@H)O1[C@H]6[C@H]6[C@H]6[C@H]6[C@H]6[C@H]1</td>
</tr>
</tbody>
</table>
MOL example: Glucose

MOL file as shown through chemical viewer software

Same file when opened with a text editor
**SDF example:** This Enamine order of over 20 cmpds came with vendor provided SD-file. Notice the additional metadata provided that can be easily imported into database for future mining.

<table>
<thead>
<tr>
<th></th>
<th>structure</th>
<th>CatalogID</th>
<th>formula</th>
<th>MW</th>
<th>CSTR</th>
<th>CAGG</th>
<th>LOGEE</th>
<th>LOGG</th>
<th>LOGP</th>
<th>rotating_bonds</th>
<th>PISA</th>
<th>hacceptors</th>
<th>tenders</th>
<th>purity</th>
<th>codes</th>
<th>unt</th>
<th>amount</th>
<th>measure</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><img src="image1.png" alt="Image" /></td>
<td>255601080</td>
<td>C19H18N2O2</td>
<td>321.38</td>
<td>0.0</td>
<td>0.0</td>
<td>2.17</td>
<td>1.025</td>
<td>3.0</td>
<td>55.48</td>
<td>5.0</td>
<td>1.0</td>
<td>91.0</td>
<td>70505-1854</td>
<td>Ti</td>
<td>1</td>
<td>3.00</td>
<td>mg</td>
<td>49.92</td>
</tr>
<tr>
<td>2</td>
<td><img src="image2.png" alt="Image" /></td>
<td>249550376</td>
<td>C17H17N2O2</td>
<td>312.973</td>
<td>0.0</td>
<td>0.0</td>
<td>3.709</td>
<td>0.0</td>
<td>0.0</td>
<td>82.49</td>
<td>5.0</td>
<td>1.0</td>
<td>100.0</td>
<td>70504-7342</td>
<td>Ti</td>
<td>1</td>
<td>3.00</td>
<td>mg</td>
<td>49.92</td>
</tr>
<tr>
<td>3</td>
<td><img src="image3.png" alt="Image" /></td>
<td>255803501</td>
<td>C20H22N4O4</td>
<td>534.423</td>
<td>0.0</td>
<td>0.0</td>
<td>1.31</td>
<td>2.185</td>
<td>3.0</td>
<td>82.49</td>
<td>5.0</td>
<td>1.0</td>
<td>92.0</td>
<td>70505-4075</td>
<td>Ti</td>
<td>1</td>
<td>3.00</td>
<td>mg</td>
<td>49.92</td>
</tr>
<tr>
<td>4</td>
<td><img src="image4.png" alt="Image" /></td>
<td>255804162</td>
<td>C21H20N4O4</td>
<td>549.45</td>
<td>0.0</td>
<td>0.0</td>
<td>1.54</td>
<td>2.715</td>
<td>4.0</td>
<td>52.49</td>
<td>5.0</td>
<td>1.0</td>
<td>92.0</td>
<td>70505-4754</td>
<td>Ti</td>
<td>1</td>
<td>3.00</td>
<td>mg</td>
<td>49.92</td>
</tr>
<tr>
<td>5</td>
<td><img src="image5.png" alt="Image" /></td>
<td>255800406</td>
<td>C22H21N3O2</td>
<td>555.415</td>
<td>0.0</td>
<td>0.0</td>
<td>1.47</td>
<td>5.285</td>
<td>4.0</td>
<td>30.35</td>
<td>3.0</td>
<td>1.0</td>
<td>80.0</td>
<td>70505-9220</td>
<td>Ti</td>
<td>1</td>
<td>3.00</td>
<td>mg</td>
<td>49.92</td>
</tr>
</tbody>
</table>

**Critical to always acquire SDF from compound vendors**

- Prevent errors due to manual entry
- **Speed and time for informatics.** How else do you keep a proper inventory of over one million compounds with just a compound management staff of two people?
- Database informatics allows other researcher to mine information quickly.
SDF file Imported

Note data tabs for linked HTS informatics

First cmpd from previous page.

Scripps assigned ID SR-01000006508

Vendor ID and Cat#
Scripps ID Decoder:

Scripps Corporate-ID: SR-XXXXXXXX 11 digits to follow Scripps Research prefix

First two digits (XX) identify the source:
- 00 FL Synthesis represents in-house synthesis
- 01 FL HTS represents original FL HTS library collection
- 02 Roberts Group
- 03 MLPCN represents NIH collection
- 05 FL Academic represents Scripps Hodder library collection
- 06 Eli Lilly

Scripps Sample-ID: SR-XXXXXXXXxx 1 digit extension for batch number

Example Sample-ID: SR-01000001430-1 Batch-1 of a Scripps Research compound initially from the Florida HTS collection. (First batch of this compound sample to have been acquired)
Software supports HTS, Compound Management & Cheminformatics Activities
A. Structure Searches: Exact, Analogs and Sub-Structures


B. Structure Analysis

- Similarity ranking using Tanimoto scoring
- MCS clustering
- Substructure identification
- Determination of fragment collections
- fSP3 hybridization ratio (i.e. natural product likeness)

C. Chemical/Physical properties calculations

- MW, cLogP, LogD, PSA; rotatable bonds; H-donors/acceptors; ring count; heavy atom count…

D. Drug likeness /Affinity Ranking/ Chemistries

- Rule of Five; Rule of Three; Reactive Molecule filtering; Bioavailability ranking; Customized filters…
Chemoinformatics Toolkit

- **ChemAxon Applications**
  - **Instant JChem**: Structural based determination of physical properties, rule of 5, similarity, rule of 3, SP³ hybridization ratio, structure queries, bioavailability…
  - **LibraryMCS**: Hierarchical clustering based on Maximum Common Substructures
  - **Fragmenter**: Molecular fragments based on cleavage rules to create fragment collections

- **Accelrys Pipeline Pilot**
  - **Similarity profiling** e.g. Tanimoto score
  - **Nonhierarchical clustering**
  - **PAINS Reactive Molecule** filtering
  - Chemical/physical properties calculating

- **Accelrys MDL/ISIS**
  - **Database Queries**: structural analogs; substructure searches
  - **ISIS/Excel**: SD file import/export; properties sorting & mining; customized graphics

- **Web-Based Tools**
  - **ChemNavigator**: Discovery of commercially available structures and analogs
  - **Scifinder**: Search structures and analogs, physical properties
Chemoinformatics: **MSC Clustering**

**Maximum Common Substructure (MCS) clustering**: ChemAxon tool used to distill collections into hierarchical order based on the largest shared substructural component.

**Important use**: Hit cmpds can be organized into clusters to help Med. Chemists find common scaffolds for SAR development.
**Tanimoto Similarity Score:** The most popular similarity measure for comparing chemical structures represented by means of fingerprints is the Tanimoto coefficient. The Tanimoto similarity is only applicable for a binary variable, and the Tanimoto coefficient ranges from 0 to +1 (where +1 is the highest similarity).

The *Tanimoto coefficient* between two points, \( a \) and \( b \), with \( k \) dimensions is calculated as:

\[
\frac{\sum_{j=1}^{k} a_j \times b_j}{(\sum_{j=1}^{k} a_j^2 + \sum_{j=1}^{k} b_j^2 - \sum_{j=1}^{k} a_j \times b_j)}
\]

**Rule of Thumb:** Two structures are usually considered similar if \( T > 0.85 \). However, a similarity of \( T > 0.85 \) does not imply similarity in bioactivities.

Selected high scoring hits from the MCHr1 antagonist
IC50 5.9nM
**Chemoinformatics WorkFlow**

**Step-1 Query MDL/ISIS** to select one of the custom library collections and export as SD-file.

**Step-2 Remove Reactive Molecules:** Use *Pipeline Pilot* and Lead-ID propriety *PAINS* filter to remove unwanted compounds; generate a triaged collection and export as SD-file.

**Step-3 Instant JChem Calculations:** Import SD-file to *Instant JChem* and insert “Rule of Five” ranking into SD-file. Export as an appended SD-file.

**Step-4 LibraryMCS:** Import appended SD file and run MCS clustering. Export as SD-file with embedded MCS structure and hierarchical ranking of the customized library collection.
ACD/Labs Freeware: Chemistry Software
http://www.acdlabs.com/resources/freeware/

ACD/ChemSketch Freeware is a drawing package that allows you to draw chemical structures including organics, organometallics, polymers, and Markush structures. It also includes features such as calculation of molecular properties (e.g., molecular weight, density, molar refractivity etc.), 2D and 3D structure cleaning and viewing, functionality for naming structures (fewer than 50 atoms and 3 rings), and prediction of logP.

ChemAxon Freeware: Chemistry Software
https://www.chemaxon.com/my-chemaxon/my-academic-license/

Marvin Suite: MarvinSketch and MarvinViewer (for SMILES and MOL and SDF viewing and drawing), MarvinSpace, Reactor, Library MCS, Instant JChem

Molecular surfaces with ChemAxon MarvinSpace
CHEMICAL SPACE

Drugs

Drugs

Lead-like

Hit-like

Drug-like
**BCUT DESCRIPTORS:** Designed to encode atomic properties that govern intermolecular interactions. Used in diversity analysis. BCUT matrices encode atomic charge, atomic polarizability, and atomic hydrogen bonding ability and the highest and lowest eigenvalues are extracted for use as descriptors. Principal component analysis (PCA) is implemented for eigenvector-based multivariate analyses to ascertain principal component (PC) which have the largest possible variance accounted. Although multi-dimensional PC space is possible its common to only plot those PC with the largest variances in 3D or 2D plot comparisons.

**Translation:** BCUT analysis is one means to visualize chemical library space and compare collections.

**Example:** BCUT ChemBridge 50K comparison to SDDL
Definitions

Chemical Space: is a concept in cheminformatics referring to the property space spanned by all chemical compounds adhering to a given set of construction principles and boundary condition. In pharmacologically active molecules (Lipinski rules, CHNOS only) estimated to be ~10^63

Hit-Like: Compounds that exhibit high affinity toward target (<1uM), selectivity versus other targets and exhibit acceptable/desirable sigmoidal dose response (i.e. Hillslope, EC_{50}, EC_{max}, EC_{min}).

Lead-Like: Must be Hit-like and exhibit low cytotoxicity, synthetic tractability, patentable, chemically stable, desirable to be non-Pan-Assay Interference Compounds (PAINS).

Drug-Like: Lead-like and exhibit good drug-like properties (Lipinski, Veber rules). Those compounds that have acceptable ADME/TOX properties to survive through the completion of human Phase 1 trials

IND: Investigational New Drugs is a FDA approval to allow drug-like compounds to be released to clinical investigators for trial testing. Drug-like compounds must be tested in animal pharmacology/toxicology studies and animal models if possible.

Drugs: Those INDs that have been tested in clinical trials (phase I-III) and exhibit acceptable pharmacology in man as well as being efficacious against diseases.
**Lipophilicity**: The ability of a chemical compound to partition into fats, oils, lipids, and non-polar solvents. Estimated through partition-coefficient (P) or distribution-coefficient (D) is the ratio of concentrations of a compound in a mixture of two immiscible phases at equilibrium

**Log D “distribution” coefficient**: The ratio of the sum of the concentrations of all forms of the compound (ionized plus un-ionized) in each of the two phases (water and octanol) it depends on the pH of the aqueous phase.

\[
\log D_{\text{oct/wat}} = \log \left( \frac{[\text{solute}]_{\text{ionized,octanol}} + [\text{solute}]_{\text{un-ionized,octanol}}}{[\text{solute}]_{\text{ionized,water}} + [\text{solute}]_{\text{un-ionized,water}}} \right)
\]

**Log P “partition” coefficient**: Ratio of the concentrations of a solute between the two phases (water and octanol) specifically for un-ionized solutes. Log P value is a measure of lipophilicity.

\[
\log P_{\text{oct/wat}} = \log \left( \frac{[\text{solute}]_{\text{un-ionized,octanol}}}{[\text{solute}]_{\text{un-ionized,water}}} \right)
\]

log D = log P  only when compounds are non-ionizable at any pH
Not all LogP’s are calculated equal:

** ALOGP:** Atom based formulated on summative atomic contribution. Other atom based algorithms include XLogP, MLogP

** CLogP:** is a fragment-based method using group contribution methods. Formulated on summative atom contribution, atomic hybridization states, fragment and molecular properties contributions (proprietary).

**Simple fragment molecules:** CLOGP method is better for very small molecules in the range of 1–20 atoms.

**Standard small molecules:** The two methods are almost comparable in the range of 21–45 atoms

**Complex molecules:** ALOGP method has better accuracy for molecules with more than 45 atoms; but experimental determination preferred.
Landmark publication in Dr. J. Clinical Pharmacology (1964-1985):

~39% of drugs (NCE and INDs) failed during development phase due to poor biopharmaceutical properties

Poor pharmacokinetics (~39% failures) and lack of clinical efficacy (29% failures)

- Early Drug Design Protocols focused on the isolation of active compounds
- Issues such as pK, toxicity & solubility were addressed much later in the development phase

**Paradigm shift**: It is necessary to anticipate these requirements during drug discovery & promote exclusively those molecules that have the highest chances of success to the development phase.

*Fail Fast and Early in Discovery Phase*
If the focus of a drug design study is solely activity based then this may yield compounds that are effective ligands for the target site but have inadequate properties that would make them a successful drug.

In HTS, library formulation must make this a consideration.
Lipinski's rule of five also known as the Pfizer's rule of five is a rule of thumb to evaluate drug-likeness and estimate if a chemical compound with a certain pharmacological or biological activity has properties that would make it a likely orally active.

- ≤ 5 hydrogen bond donors (nitrogen-hydrogen; oxygen-hydrogen bond donors)
- ≤ 10 hydrogen bond acceptors (nitrogen and oxygen atom acceptors)
- A molecular mass < 500 daltons (rule extension: 180 to 500 dalton range)
- Partition coefficient LogP < 5 (rule extension: -0.5 to 5.6 range)
- Polar surface area < 140 Å² (rule extension: Veber Rule)
- Rotatable Bonds < 12 (rule extension: Veber Rule)

Why called rule of five? All numbers are multiples of five

How were these rules derived?

2,200 INDs were examined for properties based on survival of clinical phase 1 trials (acceptable toxicity and pharmacokinetic profiles)

**Molecular Weight:** Increase size impedes passive diffusion and water solubility. Impedes lipid bilayer membrane penetration.

**Hydrogen bonds:** Increases water solubility but must be broken if compound is to permeate the lipid bilayer membrane. Increase H-bonds reduces partitioning between water and lipid phases.

**LogP:** Increase decreases aqueous solubility which reduces absorption.

**Veber Rule Extension**

**Polar surface area:** As surface size increases, a larger cavity must form in water to solubilize the compound. Crossing a lumen requires that molecules be non-polar. Large polar surface as part of the surface makes the interaction and uptake over a lipid bilayer difficult.

**Rotatable Bonds:** Veber’s experiments concluded that MW was not the critical issue with lumen uptake; but rather the number of rotatable bonds, which comes as an entropic cost.

SDDL collection is 87% compliant with the Pfizer rule of five
In early drug discovery, lipophilicity and molecular weight are often increased to improve the affinity and selectivity of the drug candidate. This practice limits Medicinal Chemistry efforts in optimizing a structure and maintaining drug-likeness. Screening libraries are biased toward lower mass/lipophilicity to enhance MedChem development post-HTS. Candidate drugs that conform to the RO5 tend to have lower attrition rates during clinical trials and increased chance of reaching the market.

Warning: Screening hits selected can be artifacts and not true activity profile between molecule to protein drug-like interactions

PAINS are defined by their ability to show activity across a range of assay platforms and against a range of proteins. High promiscuity!

The most common causes of PAINS activity:

- Metal chelation
- Chemical aggregation
- Redox activity
- Compound fluorescence
- Cysteine oxidation
- Promiscuous binding

<table>
<thead>
<tr>
<th>Substructure</th>
<th>Structure (example)</th>
<th>Class name</th>
<th>MOA comment</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Substructure" /></td>
<td><img src="image2.png" alt="Structure" /></td>
<td>Allylsidine barbutanes and furobarbutanes</td>
<td>Reactive</td>
</tr>
<tr>
<td><img src="image3.png" alt="Substructure" /></td>
<td><img src="image4.png" alt="Structure" /></td>
<td>Allylsidine rhodamines</td>
<td>Reactive, photo-reactive, chelation</td>
</tr>
<tr>
<td><img src="image5.png" alt="Substructure" /></td>
<td><img src="image6.png" alt="Structure" /></td>
<td>Beta-amino ketones</td>
<td>Elimination to form reactive species</td>
</tr>
<tr>
<td><img src="image7.png" alt="Substructure" /></td>
<td><img src="image8.png" alt="Structure" /></td>
<td>Catechols</td>
<td>Chelator, redox-active, oxidizes to quinones</td>
</tr>
<tr>
<td><img src="image9.png" alt="Substructure" /></td>
<td><img src="image10.png" alt="Structure" /></td>
<td>Cyclopropane-fused tetrahydroquinolines</td>
<td>Reaction catalyst unanticipated and/or formation of an electrophilic by-product in vitro</td>
</tr>
<tr>
<td><img src="image11.png" alt="Substructure" /></td>
<td><img src="image12.png" alt="Structure" /></td>
<td>Hydroxyphenylhydrozones</td>
<td>Chelation</td>
</tr>
</tbody>
</table>
Caution is needed:
• Computational PAINS filters are far from comprehensive and vendors still include many PAINS-type structures in their catalogue. There is no universal consensus on what are PAINS compounds. Consequently different filters will yield different results.
  Rules are Evolving

• Many FDA approved drugs (~7%) would be qualified as PAINS compounds!
  • ~5% Scripps’ FDA approved drug collection (~3,250 cmpds) would be flagged as PAINS

• Not all offenders are equally bad. Medicinal chemists can modify promising leads to limit promiscuity.

EARLY HIT DISCOVERY SHOULD PAINS?
• Be strictly enforced?
• Ignored?
• Somewhere in between?
SRIMSC BEST PRACTICES:

- **HITS are analyzes** through the Baell/Holloway algorithm and classified into PAINS A, B, C classes
  - Filter A: 16 substructural elements e.g. phenols; quinones; reactive azo; mannich bases
  - Filter B: 55 substructural elements e.g. cyano-imines; tetrazines; dyes; imidazoles; catechols
  - Filter C: 409 substructural elements e.g. thio_urea; thiophene_amino; cyano_imine; sulfonamide

- **PAINS Hits are flagged but not removed.** Data is presented to medicinal chemistry to triage bad actors

- SRIMSC also applies a “**Promiscuity Index**”. A simple ratio of how many times a compound was used in a HTS campaign (across all target classes) to how many times it was found to be a hit.

FINAL POINTS OF CONSIDERATION:

- **SDDL library** of ~645K compounds is estimated to contain ~6% PAINS cmpds
- **MLPCN Library** of ~360K compounds is estimated to contain ~4% PAINS cmpds
- **SDDL overlaps MLPCN library collection** by 14.8% (identical cmpds found)
- **FDA-Drug Approved library** (Scripps) contains ~5% PAINS cmpds
- **SDDL PAINS-FREE Sub-library** is a collection of 20,559 cmpds curated by Scripps in order to furnish hits with greater target selectivity and lower promiscuity.

• Natural ligands are predominantly three dimensional in their interactions, providing strong and more selective affinities.

• However, traditional HTS libraries are heavily composed of flat aromatic compounds that poorly emulate their natural ligand counterparts.

• FSP$^3$ is the fractional ratio of the number of sp$^3$ hybridized carbons to the total carbon count.

• It has been demonstrated that this hybridization ratio correlates with the success of compound transition from discovery, through clinical testing.
Traditional HTS libraries are heavily composed of flat aromatic compounds that poorly emulate their natural ligand counterparts.

Mean Fsp³ ratio

<table>
<thead>
<tr>
<th>Category</th>
<th>Fsp³</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discovery</td>
<td>0.36</td>
</tr>
<tr>
<td>Phase I</td>
<td>0.38</td>
</tr>
<tr>
<td>Phase II</td>
<td>0.43</td>
</tr>
<tr>
<td>Phase III</td>
<td>0.45</td>
</tr>
<tr>
<td>Drugs</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Lovering, F.; Bikker J.; Humblet C.; (2009); Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success J. Med. Chem. 52, 6752-6756
• Toxicity plays a major role in attrition in the clinic and promiscuity has been linked to toxicity.

• Increasing complexity reduces promiscuity and CYP450 inhibition
Lovering, F.; Bikker J.; Humblet C.; (2009); Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success J. Med. Chem. 52, 6752-6756
Compared to drugs that act in the periphery, brain-penetrant drugs tend to be more lipophilic and rigid, having fewer hydrogen bonds, fewer formal charges, and a lower polar surface area.

A linear relationship between brain penetration and dynamic polar surface area of a drug was found by Kelder et al. (1).

Mahar Doan et al. (2) reported that in their analysis of 18 physicochemical properties, the CNS drug set had fewer hydrogen bond donors, fewer positive charges, greater lipophilicity, lower polar surface area, and reduced flexibility compared with the non-CNS drug set.

Optimal molecular properties for brain penetration have been proposed by Van de Waterbeemmd et al.

References:

2. Doan, KMM; Humphreys, JE; Webster, LO; et al. Passive permeability and P-glycoprotein-mediated efflux differentiate central nervous system (CNS) and non-CNS marketed drugs (2002) JOURNAL OF PHARMACOLOGY AND EXPERIMENTAL THERAPEUTICS Vol: 303 Issue: 3 p1029-1037

Substances cross the blood-brain barrier (BBB) by a variety of mechanisms. These include transmembrane diffusion, saturable transporters, adsorptive endocytosis, and the extracellular pathways.*

<table>
<thead>
<tr>
<th>PROPERTY</th>
<th>CNS Rules</th>
<th>Lipinski Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogP lipophilicity</td>
<td>1.5 to 2.7</td>
<td>-0.5 to 5.6</td>
</tr>
<tr>
<td>Molecular Mass</td>
<td>&lt;400 daltons</td>
<td>180- 500 daltons</td>
</tr>
<tr>
<td>Polar Surface Area</td>
<td>&lt; 90Å²</td>
<td>&lt; 140Å²</td>
</tr>
<tr>
<td>Hydrogen Donors</td>
<td>2.12 ave</td>
<td>≤ 5</td>
</tr>
<tr>
<td>Hydrogen Acceptors</td>
<td>1.5 ave</td>
<td>≤ 10</td>
</tr>
<tr>
<td>Rotatable Bonds</td>
<td>≤ 5</td>
<td>≤ 12</td>
</tr>
<tr>
<td>Hetero-atoms (O + N)</td>
<td>&lt;5 (4.32 ave)</td>
<td>na</td>
</tr>
</tbody>
</table>

* Hassan Pajouhesh and George R. Lenz; Medicinal Chemical Properties of Successful Central Nervous System Drugs NeuroRx. 2005 Oct; 2(4): 541-553. PMCID: PMC1201314
SDDL Collection and CNS/BBB Compliant Compounds:
Within our SDDL collection there are 301,518 compounds that meet the above criteria as CNS-favorable compounds. Selective library plates can be screened for “hits” that have desirable blood-brain barrier (BBB) properties.

Important Note: Small size and PSA provides bandwidth for MedChem modifications

**Dark chemical matter (DCM) compounds**

*Definition:* Compounds that have never shown biological activity, even after being screened repeatedly in many different drug assays.

Studies on DCMs over 650 assays found that:

- 36% of the MLPCN collection are DCMs
- DCMs have higher solubility; less hydrophobic
- Have lower MW
- Fewer aromatic rings than bioactive cmpds
- Concluded DCM cmpds are not dramatically different in structure from cmpds commonly identified as hits
- Almost all of the substructural features in “dark” cmpds can be found in active cmpds

* Wassermann, Lounkine, Glick et.al. Dark chemical matter as a promising starting point for drug lead discovery; Nat. Chem. Biol. 11, 958-966 (2015) DOI: 10.1038/nchembio.1936*
A NEW HOPE FOR DCM!

DCMs can still be of value:

- Structures are non-PAINS
- Structures with no promiscuity
- Potential for high selectivity
- A compound that has not yet been active in a biological assay doesn’t mean that will be the case for all future assay
- Flagging compounds as DCM may prove useful in high-throughput screening to highlight potential opportunities.

Case-point: Novartis testing DCMs from multiple assays had identified four DCM compounds with antifungal activity.

* Wassermann, Lounkine, Glick et.al. Dark chemical matter as a promising starting point for drug lead discovery; Nat. Chem. Biol. 11, 958-966 (2015) DOI: 10.1038/nchembio.1936
SRIMSC Best practices with non-Hits

1st: Post-HTS campaign efforts involve working with Medicinal Chemist to identify ~3-5 tractable chemical series for SAR development.

2nd: Perform in silico analysis for all structural analogs with high similarity to the select tractable series that were also screened.

3rd: Sort analogs by primary screen potency to provide early SAR clues

Examine results for:
* To avoid unnecessary MedChem synthesis efforts
* Provide early clues with respect to SAR vs. potency
* For weaker actives in a given series, does therapeutic window improve (greater selectivity)
Pilot Screening Sets:

- Small library sets ~1K to ~10K cmpds used to generate preliminary data
- Aimed at demonstrating HTS readiness
  - Z-score; S:B ratio; DMSO tolerance; reagent stability; good controls etc.
- Provide an estimate of HIT rate
  - Target ~1% HIT rate. Low rates may indicate non-druggable target; High rates may require new assay design or counterscreens. PAINS fishing?
- Provide early lead for active compounds of interest
  - Novel active compounds can serve as better controls (HTS campaign)
  - Novel actives provide insight on library selection
- Provide critical data in support of HTS grants
Pilot Screening Set Samples used SRIMSC:

**LOPAC Collection:** The *Library of Pharmacologically Active Compounds* (LOPAC) is a collection of annotated small molecules with pharmacology against a broad range of targets. This collection is well suited for preliminary exploration/assay validation in high throughput screening (HTS), high content screening (HCS) and chemical biology. (~1280 cmpds)

**Prestwick Collection:** Annotated library containing off-patent small molecules, with 90% being marketed drugs and 10% being bioactive alkaloids or related substances. The set is selected for structural diversity, broad spectrum activity covering several therapeutic areas (e.g. neuropsychiatry to cardiology, immunology, anti-inflammatory, analgesia) and for safety and bioavailability profiles in humans. (~1200 cmpds)

**TOCRIS Collection:** This annotated small molecule collection is designed for exploratory discovery in high throughput screening (HTS), high content screening (HCS) and chemical biology applications. Tocriscreen represents a diverse and unique collection of compounds with proven bioactivity on a broad range of targets including GPCRs, kinases, ion channels, nuclear receptors and transporters. (~1200 cmpds)

**Clinically Relevant Collection:** Curated by Scripps, this clinically relevant library consists of commercial bioactive compounds identified from the MDL® Comprehensive Medicinal Chemistry database (over 7,500 bioactive compounds used or studied as medicinal agents in humans) or DrugBank database (detailed drug data for nearly 4,800 bioactives tested in humans). (~500 cmpds)
Repurpose Collections: Repurposing has the objective of targeting existing and abandoned drugs to new disease areas including those targeting rare and neglected diseases.

SRIMSC FDA-Approved Collection: is composed of drugs that have reached clinical trial stages in the USA or that are marketed in Europe and/or Asia. Compound have been assigned USAN, USP INN, BAN and/or JAN designations and are included in the USP Dictionary (U.S. Pharmacopeia), the authorized list of established names for drugs in the USA and/or are listed in the Index Nominum, the International Drug Directory. All of these compounds have known and well-characterized bioactivities, safety and bioavailability properties, which could dramatically accelerate drug development and optimization. (~3250 cmpds)

NCI Oncology Drug Set: A set of anticancer drugs to enable oncology research that contains the most current FDA-approved anticancer drugs. The current set (AODV: Approved Oncology Drug) consists of 120 agents and is intended to enable cancer research, drug discovery and combination drug studies. All proprietary agents in this set were obtained by NCI/NIH Developmental Therapeutics Program through commercial sources.

The Pathogen Box: Contains ~400 diverse, drug-like molecules active against neglected diseases of interest. Composition includes drug targeted to Tuberculosis (116); Malaria (125); Kinetoplastids (70); Helminths (32); Cryptosporidiosis (11); Toxoplasmosis; Dengue (5) and reference compounds (26).

Calibr ReFRAME IND Collection: ReFRAME library has restricted use for only rare and neglected diseases. Represents ~10,000 IND status compounds that have been used in clinical trials including those that failed, abandoned or have become drugs. A copy can be obtained by Scripps researchers through inquiry at Calibr.
Focus Libraries

Diversity Collections:

**Cayman Bio-Active Lipids:** This library collection is ideal for prostanoid or other G protein-coupled receptor screening, target validation, secondary screening, validating new assays and for routine pharmacological applications. Include prostaglandins, thromboxanes, cannabinoids, D-my-ino-sitol-phosphates, phosphatidylinositol-phosphates, sphingolipids, inhibitors, receptor agonists and antagonists, ceramide derivatives, and several other complex polyunsaturated fatty acids. (~1000 cmpds)

**Natural Products:** Natural products (NP) historically have provided the most successful source of leads for the development of new drugs, but can be problematic or difficult to implement in an HTS environment. Two NP collections exist including Prof. Ben Shen’s (Scripps-FL) actinomycetes origin compounds. (2,030 cmpds)

**Click-Chemistry Collection:** has been developed by Nobel-laureate Barry Sharpless of TSRI, and provides a powerful means of easily derivatizing hit compounds from screening efforts. This synthetic approach allows scaffolds to be modular and easily modified into stereo-specific analogs under benign and often bio-friendly conditions. The Click Chemistry Collection mimics nature in its organic synthesis approach leading to novel discovery of new pharmaceuticals and relative ease of generating large number of analog structures. (445 cmpds)

**Rule of Three (RO3) library:** Small molecule fragment library compliant with the "Rule-of-Three" guidelines, namely: MW \(\leq 300\); H-bond donors/acceptors \(\leq 3\); cLogP \(\leq 3\); Rotatable Bond Count \(\leq 3\); and Polar Surface Area \(\leq 60\). Although the RO3 compounds can have lower binding affinities, it can provide a better chance for finding leads for development that have drug-likeness parameters. (15,255 cmpds)
“Going small” with fragment-based drug discovery (FBDD/FBLD) uses low molecular weight compounds to probe a therapeutic target. This also includes using smaller tailored libraries and lower screening throughput. This is a consequence of being reliant on biophysical technologies, as compared to classical high-throughput screening (HTS) approaches. FBDD at its core a target-based drug discovery

**Features:**

- Small fragments libraries of just a few thousands of compounds vs HTS (+100K)
- Small fragments (rule of three): MW<300; cLogP<3; rotatable bonds <3; Hydrogen acceptors and donors each <3
- Fragments have low affinities and must be screened at higher concentrations 100uM to 1,000uM range.
- In HTS screening Hits with sub micromolar affinity are sought. FBDD screening seeks Hits with sub millimolar affinity.
- HTS and FBDD are two very different screening paradigms. Require different infrastructures (instruments) and expertise.
FBDD: Fragment Based Drug Discovery

HTS hits may bind by virtue of numerous suboptimal interactions. By contrast, fragment hits are more ligand efficient and involve fewer but more optimized interactions.

FBDD biophysical toolkit
- Fluorescence-based thermal shift (TS)
- 1D/2D NMR
- X-ray crystallography
- Mass spectrometry (MS)

Drug candidate
- Surface plasmon resonance (SPR)
- Isothermal titration calorimetry (ITC)

Rational synthesis
Bioaffinity measurement
Structural binding mode
Fragment elaboration cycle
Fragment screening
Target selection
FBDD pipeline
Library design
Common Instruments needed for FBDD

- Fluorescence-based thermal shift (TS)
- Mass spectrometry (MS)
- NMR (WaterLOGSY)
- Surface plasmon resonance (SPR)
- X-ray crystallography
- In silico (docking)

FBDD Advantages

- Hits are more hydrophilic. Easier to increase affinity by adding hydrophobic groups
- Higher ligand efficiency allowing medicinal chemistry to provide a RO5 ligand.
- Multiple fragments can in theory be found and combined for optimized ligand
- Fragments have fewer steric blocking groups
- Adaptable chemical-space allows for the discoveries on intractable targets.

FBDD Disadvantages

- Infrastructure can be expensive and often requires greater expertise in staff
- Protracted timelines. Hit leads are a product of medicinal chemistry efforts and not screening.
- May be a more costly program requiring significant efforts from medicinal chemists and analytic support.
Can FBDD and HTS be Combined to Leverage Advantages of both?

- Cheminformatic analysis of the NIH MLPCN library indicates that it was serendipitously composed of over 8,000 fragment-like compounds that are compliant with the “Rule of Three”; ideal fragments for FBDD.

- Further analysis has shown that many of these fragments to be representative scaffolds for hierarchical related compounds also found within the MLPCN HTS library.

- Data mining HTS screening results of various campaigns has revealed that the fragment sub-library portion produced similar hit rates to the entire library deck of over 300K compounds; but their hierarchical related compounds had proportionately enhanced hit rates (up to~50X); serving as a hit predictor and guide for compound selection.
**FB assisted HTS Advantages**

- No special infrastructural changes needed. Std. HTS instrumentation
- Only ~5% of the full SDDL needs to be screened.
- Lower costs
- Ideal for high risk targets: orphan targets; potentially undruggable proteins
- Acceptable Timeline ~HTS like
- Added SAR informatics
- Provides two paths forward: FBLD or HTS lead development.

**FB assisted HTS Screening Paradigm**

1. **Fragment HTS**
   - Primary Screen
   - ~1% Hit rate yield ~200 compounds
   - Estimated to be around ~6,500 cmpds
   - Cherry-pick MCS superstructures
   - ~5-10% Hit rate yield ~300-600 compounds

2. **Enrich Cmpd HTS**
   - Primary Screen
   - ~5-10% Hit rate yield ~300-600 compounds

**FBDD Disadvantages**

- Not possible to cover all chemical-space e.g. singletons; Natural Products
- May be limited to PPI assays. Hydrophilicity likely to be problematic with cell-based assays
- Ultimately drug discovery is a numbers game. Larger libraries means greater likely- hood of finding quality hits.
- Unproven technology. Still in the development stages.
HTS Manuals and Guidance:

The collection of chapters in this eBook is written to provide guidance to investigators who are interested in developing assays useful for the evaluation of collections of molecules to identify probes that modulate the activity of biological targets, pathways, and cellular phenotypes.

This manual has been adapted to provide guidelines for scientists in academic, non-profit, government and industrial research laboratories to develop potential assay formats compatible with High Throughput Screening (HTS) and Structure Activity Relationship (SAR) measurements of new and known molecular entities. Topics addressed in this manual include:

- Development of optimal assay reagents.
- Optimization of assay protocols with respect to sensitivity, dynamic range, signal intensity and stability.
- Adopting screening assays from bench scale assays to automation and scale up in microtiter plate formats.
- Statistical concepts and tools for validation of assay performance parameters.
- Secondary follow up assay development for probe validation and SAR.
- Data standards to be followed in reporting screening and SAR assay results.

https://www.ncbi.nlm.nih.gov/books/NBK53196/
The goal of the Academic Drug Discovery Consortium (ADDC) is to build a collaborative network among the growing number of university-led drug discovery centers and programs. With this interactive website, we aim to allow scientists to exchange technical expertise on drug discovery and development strategies as well as form partnerships with each other, biopharma companies, and drug discovery-focused contract service organizations and consultants. The website will also serve as a repository for drug discovery events, educational material, job postings, and partnership opportunities.

http://addconsortium.org
Free Sources of Chemical libraries:

**DTP/NCI vialized and plated compounds:** DTP maintains a repository of synthetic compounds and pure natural products that are available to investigators for non-clinical research purposes. The Repository collection is a uniquely diverse set of more than 200,000 compounds that have been either submitted to DTP for biological evaluation. [https://dtp.cancer.gov/organization/dscb/obtaining/default.htm](https://dtp.cancer.gov/organization/dscb/obtaining/default.htm)

**Plated NCI oncology sets:**
- The NCI Diversity Set V of 1593 compounds is available on 96-well PP U-bottom plates.
- Approved Oncology Drugs Set: contains 129 agents that are most current FDA-approved anticancer drugs
- Mechanistic Set III: 813 compounds derived from test results across 60 NCI tumors lines (96-well PP plate)
- Natural Products Set IV: 419 compounds focused on a variety of scaffold structures having multiple functional groups.

**The Pathogen Box:** Contains ~400 diverse, drug-like molecules active against neglected diseases of interest. [http://www.pathogenbox.org/](http://www.pathogenbox.org/)

**Calibr ReFRAME IND Collection:** ReFRAME library has restricted use for only rare and neglected diseases. Represents ~10,000 IND status compounds. A copy can be obtained through inquiry at Calibr.