J Comput Aided Mol Des (2013) 27:455–468
DOI 10.1007/s10822-013-9641-y

An informatic pipeline for managing high-throughput screening experiments and analyzing data from stereochemically diverse libraries

Carol A. Mulrooney · David L. Lahr · Michael J. Quintin · Willmen Youngsaye · Dennis Moccia · Jacob K. Asiedu · Evan L. Mulligan · Lakshmi B. Akella · Lisa A. Marcaurelle · Philip Montgomery · Joshua A. Bittker · Paul A. Clemons · Stephen Brudz · Sivaraman Dandapani · Jeremy R. Duvall · Nicola J. Tolliday · Andrea De Souza

Received: 21 January 2013 / Accepted: 28 March 2013 / Published online: 13 April 2013

Abstract Integration of flexible data-analysis tools with cheminformatics methods is a prerequisite for successful identification and validation of "hits" in high-throughput screening (HTS) campaigns. We have designed, developed, and implemented a suite of robust yet flexible cheminformatics tools to support HTS activities at the Broad Institute, three of which are described herein. The "hit-calling" tool allows a researcher to set a hit threshold that can be varied during downstream analysis. The results from the hit-calling exercise are reported to a database for record keeping and further data analysis. The "cherry-picking" tool enables creation of an optimized list of hits for confirmatory and follow-up assays from an HTS hit list. This tool allows filtering by computed chemical property and by substructure. In addition, similarity searches can be performed on hits of interest and sets of related compounds can be selected. The third tool, an "S/SAR viewer," has been designed specifically for the Broad Institute's diversity-oriented synthesis (DOS) collection. The compounds in this collection are rich in chiral centers and the full complement of all possible stereoisomers of a given compound are present in the collection. The S/SAR viewer allows rapid identification of both structure/activity relationships and stereo-structure/activity relationships present in HTS data from the DOS collection. Together, these tools enable the prioritization and analysis of hits from diverse compound collections, and enable informed decisions for follow-up biology and chemistry efforts.

Keywords High-throughput screening · Cheminformatics · Diversity-oriented synthesis · Hit calling · Cherry picking · Structure/activity relationships

C. A. Mulrooney (✉) · D. L. Lahr · M. J. Quintin · W. Youngsaye · D. Moccia · J. K. Asiedu · E. L. Mulligan · L. B. Akella · L. A. Marcaurelle · P. Montgomery · J. A. Bittker · S. Brudz · S. Dandapani · J. R. Duvall · N. J. Tolliday · A. De Souza
Chemical Biology Platform, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
e-mail: carolm@broadinstitute.org
URL: www.broadinstitute.org

P. A. Clemons
Chemical Biology Program, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

Introduction

High-throughput screening (HTS) of small molecules against biological targets or whole cells is a well-established process in drug discovery [1]. Improvements in technologies for miniaturization and automation have enabled the routine testing of hundreds of thousands, or even millions, of compounds. Cheminformatics tools that allow the capture, query, and analysis of vast amounts of data from primary HTS and secondary assays have enabled comparison of the performance of multiple compounds in a single assay or the performance of a single compound in multiple assays [2]. Since HTS has largely remained an industrial activity, pharmaceutical companies have successfully developed informatics frameworks to capture and query huge amounts of disparate types of data. Some examples of these informatics platforms include the ABCD model used by Johnson & Johnson and the OSIRIS system developed by Actelion [3, 4]. However, most of these tools are custom-developed in pharmaceutical companies and are not available for broader use in the open scientific community.

Since many academic and non-profit research institutions now routinely run HTS and many secondary assays,

there exists a need within each of these academic centers for robust and user-friendly cheminformatics tools to store and analyze the vast volume of data from assays. We encountered similar needs in the Chemical Biology Platform at the Broad Institute, a non-profit organization. There are commercial tools that can be used, but customization is required before integration with internal databases can occur. Hence, we decided to build several cheminformatics workflows, configuring tools that will better meet our needs for supporting our probe- and lead-development projects.

The Chemical Biology Platform at the Broad Institute hosts a comprehensive screening and chemistry center in the National Institute of Health (NIH) Molecular Libraries Probe Production Center Network (MLPCN). The MLPCN offers academic researchers access to the large-scale screening capacity and medicinal chemistry necessary to develop new chemical probes [5]. The small molecules used in MLPCN screens are very similar to ones typically used by pharmaceutical companies—rich in heterocycles and dominated by sp^2 centers [6–8]. In addition to this collection, the Broad Institute has invested in the design and synthesis of a large collection of complex small molecules through diversity-oriented synthesis (DOS). The DOS collection is rich in natural product inspired, complex scaffolds that are enriched in sp^3 content and chiral centers [7, 8]. Currently over 100,000 DOS compounds have been synthesized and are part of the Broad Institute's screening collection [9–11].

Here, we outline the development and application of three specific cheminformatics tools—a hit-calling/cherry-pick workflow that helps identify and prioritize hits emerging from HTS (two tools) and a process to quickly identify stereochemical dependencies in HTS data from the DOS collection (one tool). Several examples are provided to demonstrate the utility of these applications.

We typically run HTS with >350,000 compounds, at a single concentration (7–10 μM) in duplicate. The raw data are acquired from a range of detection instruments (plate reader, automated microscope, etc.) and processed for analysis with Genedata Screener assay analyzer module [12] (Fig. 1). All data are normalized (minimally to DMSO neutral controls [2] ideally to both neutral and positive controls), and basic quality-control (QC) evaluation is performed. The results from this evaluation, termed a "QC session", are directly reported to a results database. We then use the "hit-calling" tool to identify hits from HTS based on the QC sessions in the results database. Since many HTS campaigns have hit rates of 0.5–1 %, it is not uncommon to have 1,500–3,500 compounds to consider as possible hits. This list of compounds needs to be reduced to a more manageable number for retesting with a concentration–response curve to confirm activity and gain insight into potency. We use the "cherry-picking" tool to prioritize HTS hits for follow-up studies. When DOS compounds are identified as hits in HTS, we use the "S/SAR viewer" to identify stereochemical dependencies in the HTS data, in addition to the standard workflow, to prioritize compounds for subsequent studies.

Development of a hit-calling and cherry-pick list-creation workflow

Systematic analysis of the corrected HTS data is necessary to maximize the successful identification of suitable compounds for follow-up studies. In the absence of cheminformatics tools, data analysis from an HTS screen is an ad hoc process and may be biased by scientists' pre-conceived notions about compounds or data, including biases that may not reflect their actual preferences [13–15]. Moreover, the use of cheminformatics tools allows the capture and documentation of the decision-making process at each stage, facilitating subsequent reanalysis of the data if desired.

Hit-calling

A hit list is a subset of the compounds screened in an assay that are predicted to be of interest based on the activity results from a primary screen [16–18]. This list provides the basis for a cherry-pick list, which is submitted to the compound management group for re-plating of hits for testing in a dose-response format and subsequent follow-up assays. If feasible, untested or inactive related analogs are included in the selection to explore preliminary structure/activity relationship (SAR) trends. We have developed a series of custom visualizations in TIBCO Spotfire [19] that

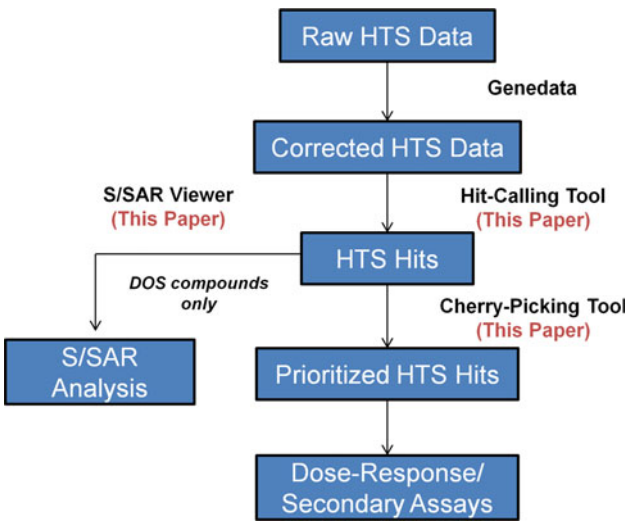


Fig. 1 An overview of activities following HTS highlighting corresponding tools

walk researchers through the hit-calling and cherry-picking processes. The Spotfire software is used along with Pipeline Pilot [20] and the ChemAxon JChem Cartridge for Oracle [21] to streamline data analysis and create a record of the decision-making process.

The first step in the hit-selection process is quality-control assessment of the HTS data (Fig. 2). Using Spotfire, a scatter plot is displayed showing activity within each well of an assay plate. If the primary screen was performed in duplicate, as is customary at the Broad Institute, then both data sets are displayed simultaneously. The user can identify and remove any outlying data points which seem erroneous or invalid (e.g., due to technical artifacts) and that were not detected during the first QC session performed in Genedata. These points can be selected individually, or a threshold can be entered to select all points beyond a certain value. The user can then mask the selected points, so that when the average activity for each tested compound is calculated, invalid data points will not be considered. The user is also presented with the option to override the hit-call result for compounds that have any masked replicates, forcing a final hit-call outcome of 'inactive' or 'inconclusive' regardless of the reported activity. Generally, an inconclusive classification indicates that the accuracy or the reproducibility of measured bioactivity may be questionable.

The next page aggregates all of the non-masked data points to present a scatter plot of activity on a per-sample basis. This is typically done by calculating the mean across replicates of a compound, though the user may choose to

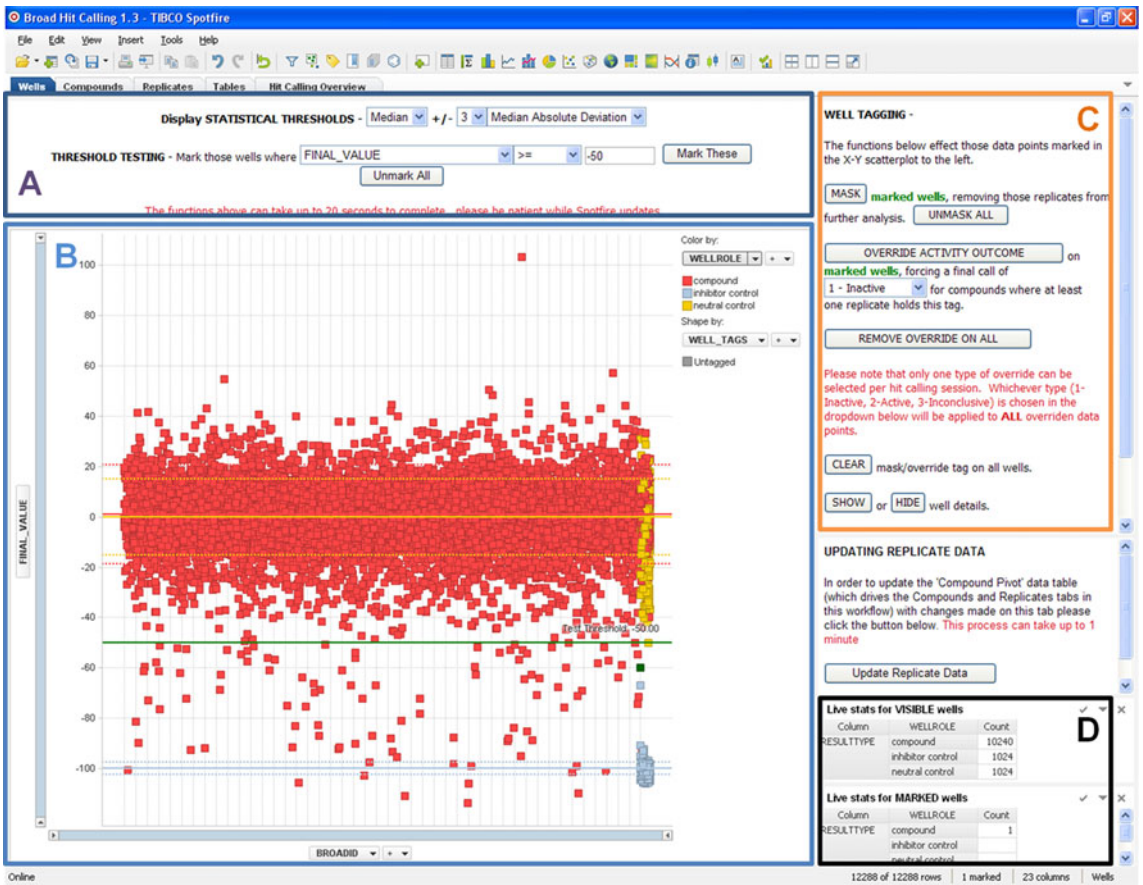


Fig. 2 Hit calling in the Spotfire workflow. The screen is divided into sections **a**–**d**. **a** Drop-down menus permit the user to display various statistical measurements (e.g., mean or median activity) as *solid lines* in the scatter plot (**b**). Deviation thresholds selected by the user will be displayed as *dotted lines* in the scatter plot. Threshold testing allows for quick selection of wells meeting the input criterion to facilitate decision making on large subsets of data. **b** Data points representing individual wells are color-coded by well contents for easier analysis. The *solid lines* represent the testing threshold, statistical measurement, and associated deviations input by the user. The *solid green line* is the activity threshold. The remaining *solid lines* are the statistical median for the various data subsets: *red*: median test compound activity; *orange*: median neutral control compound activity; *blue*: median inhibitor control compound activity. The *colored dotted lines* show the median absolute deviation for the corresponding median values. **c** Individual wells or groups of wells can be masked, discarding these points from further analysis. The user can also override the experimental result and impose a decision (e.g., active, inactive, or inconclusive) on compounds that have a replicate in a masked well. **d** Tabulated summaries of well contents are listed for all unmasked wells and selected wells

view the most- or least-active replicate instead. Data points are color coded to correspond with hit-calling outcomes: active, inactive, or inconclusive. The user is presented with a menu that requires specification of two thresholds: the minimum activity required to be considered active, and the percent of replicates that must pass the activity threshold. A compound that passes only one of these requirements may be marked as active, inconclusive, or inactive at the user's discretion. A table reports the number of compounds receiving each of the three possible outcomes. The user is able to update the two thresholds dynamically in order to fine-tune the number of compounds that will advance to the cherry-pick workflow as active hits.

After the hit-calling session is completed, the hit list can be submitted to a database. We created a tool using the Spotfire application programming interface (API) that permits the application to write data directly to our database. In addition to recording the activity outcome for every sample in the session, we capture each of the decisions made by the user: whether activity values are based on the least or most active replicate, the replicate mean, or the activity and percent-of-active-replicates thresholds.

Cherry-pick list creation

The preliminary list of active compounds must be further limited to a collection of 1,000–1,200 compounds that will then be subjected to re-test in the primary assay using an 8-point dose-response format. The cherry-pick workflow has been configured in Spotfire to permit analysis of active compounds by structure and calculated physicochemical properties. In addition to removing substances possessing reactive functional groups or undesirable c-logP (calculated log-ratio of the octanol:water partition coefficient) values, this workflow also presents the opportunity to eliminate scaffolds deemed intractable from a synthetic chemistry perspective. Optimal structures have functional groups (referred to as "synthetic handles") that can be chemically substituted or rapidly modified to expedite further diversification. Structures lacking such groups, together with structures that cannot be prepared expeditiously, are candidates for elimination, as both groups of compounds will make downstream SAR studies more lengthy and challenging. Similar to the hit-calling workflow, every decision is recorded in the database for future reference. If no compounds successfully retest in dose-response format, it is possible to review the original cherry-pick decisions. With all individual decisions recorded, a new cherry-pick list can be produced that would select a new set of compounds with minimal redundancy. Additionally, compounds previously removed for surmountable liabilities can be given a second opportunity for re-testing.

The cherry-picking workflow comprises multiple sections during which the user is allowed to evaluate the entire primary data set for undesirable functional groups or physical properties, perform Tanimoto similarity searches on fingerprints [22] of promising compounds, and re-examine bioactivity if necessary. Cherry-picking is not a linear process, and the interface is designed to permit the user to move between sections as needed. Decisions performed in one section are captured and reflected throughout the remaining sections. When a hit list is imported into the workflow, the user initially encounters a bioactivity scatter plot similar to what was seen in the hit-calling workflow. By default, compounds declared "active" by the hit-calling session will be selected as cherry picks. Conversely, "inactive" and "inconclusive" substances will be excluded, and these decisions are recorded and displayed in scatter-plot fashion within Spotfire.

The shape and color of the individual data points readily inform the user whether a specific compound was considered "active", "inactive", or "inconclusive" and also show whether that compound was selected or discarded from the cherry-pick list. On the bioactivity page, it is possible to raise or lower the stringency of the assay result cutoff so as to adjust the number of compounds available for cherry-pick selection. Accordingly, decisions performed on this page will automatically specify that these compounds were accepted or rejected on the basis of their bioactivity.

There is also an option to "force pick" specific compounds for addition to the cherry-pick list. This feature can be used, for example, to include inactive analogs in a cherry-pick list to explore preliminary SAR trends. Any compound that has been force-picked will be included in the final cherry-pick list regardless of any liabilities that may be identified on subsequent pages of the workflow.

The second page of the workflow (Fig. 3) presents tabulated results of 52 substructure filters designed to eliminate undesirable functional groups [23–25]. Depending on the screening target, it may be desirable to develop either covalent or non-covalent modulators. Therefore, these substructure filters were implemented to identify chemically reactive moieties such as aziridines, epoxides, and Michael acceptors, so that the user can quickly eliminate these compounds if a non-covalent modulator is sought. In addition to possible alkylating agents, scaffolds posing possible metabolic liabilities, including furans and nitrated compounds, are identified. The number of compounds corresponding to each filter is conveniently displayed in a table that can be used to gauge tentative SAR trends. The information presented in this table reflects the subset of compounds currently being analyzed, meaning the user can limit their structural analysis to only "actives" or even members of an individual cluster. Highlighting any specific filter will display the substance ID and full structures of all compounds bearing that particular functionality (Fig. 3,

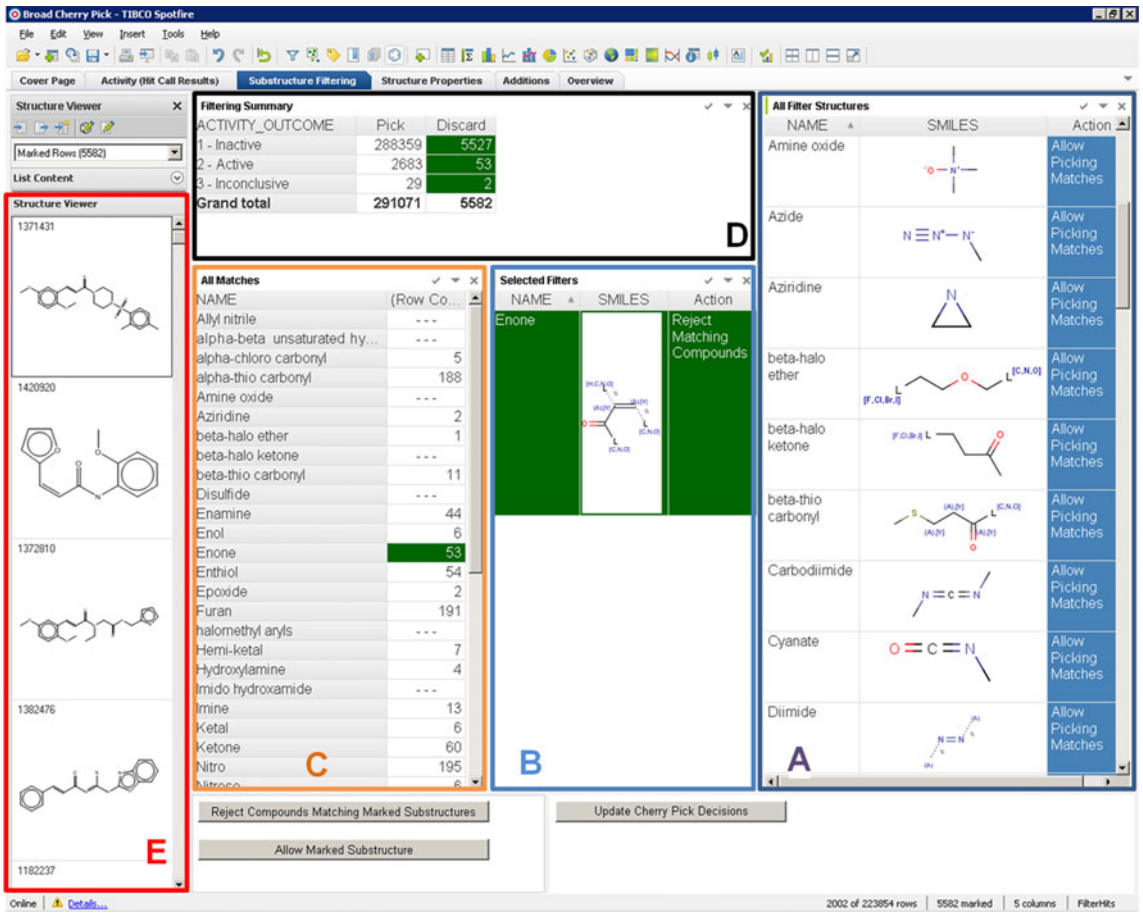


Fig. 3 Substructure filtering facilitates the removal of compounds bearing undesirable functional groups. The screen is divided into sections **a**–**e**. **a** Master list of 52 functional group filters that can be individually applied by the user. **b** Display of currently selected filter(s). **c** Tabular summary of how many unique compounds contain the corresponding functional group. **d** Summary of the decisions made (pick/discard). **e** Structure viewer to display specific compounds associated with the selected functional group filter

section E). With this information a user can look at all active compounds containing "nitro" groups and decide whether to discard the entire class or elect to save select nitrated substances for further analysis.

Physical properties of the screened compounds [26–28] can be analyzed on a different workflow page (Fig. 4). Here, one can investigate possible relationships between bioactivity and molecular weight or c-logP. With data presented in a scatter plot, the user can search for activity dependency on a number of physicochemical properties, including hydrogen-bond donors or acceptors or rotatable bond counts. Given the popularity and efficacy of Lipinski's Rule of 5 [28], as well as ongoing controversy about their specific application [29–31], this section of the workflow conveniently allows the user to apply or relax these guidelines as appropriate.

There is another page of the workflow designed to accommodate the identification and addition of structural analogs. The Broad Institute has access to distinct compound collections for its various screening projects [7, 8]; for the dozens of MLPCN projects that the Broad undertakes annually, the NIH collection is screened in the primary assays. However, in order to leverage the Broad Institute's proprietary compounds, we included a "Manual Additions" page to the cherry-picking workflow that uses Spotfire's inherent database query function in conjunction with the ChemAxon JChem cartridge to enable users to search the entire Broad collection for additional analogs. During this process of "back-filling", an initial Tanimoto similarity search of the primary data is performed to identify both inactive and active analogs. Inactive derivatives are given special consideration, and several are deliberately selected to serve as negative controls for scaffold validation. The search parameters can then be expanded to query the Broad Institute's in-house collection, including the various DOS libraries. Occasionally, there are substances prepared in-house that resemble compounds from the NIH collection, and the search engine provides an avenue to include these Broad compounds for testing. This process facilitates the rapid identification of

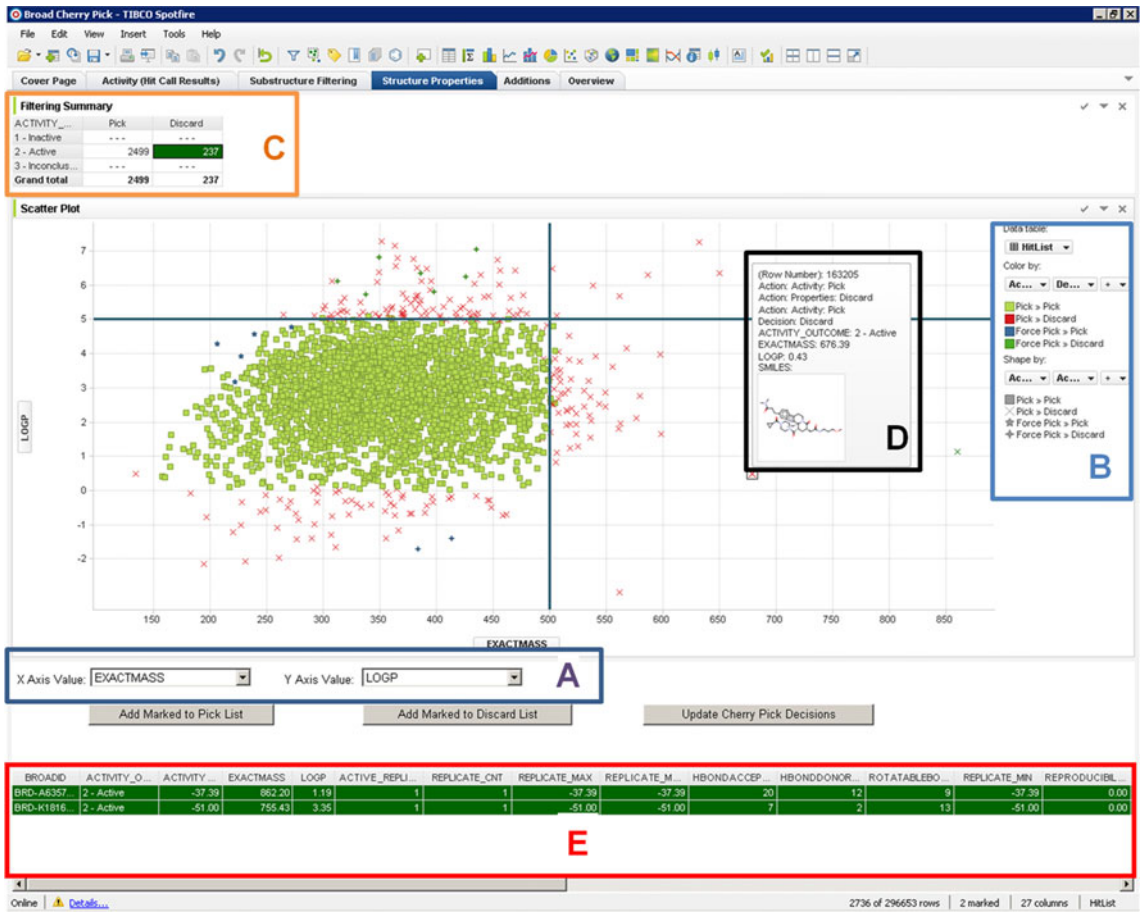


Fig. 4 Physical properties can be readily analyzed and correlated to bioactivity. The screen is divided into sections **a**–**e**: **a** Convenient drop-down lists allow the user to change the horizontal and vertical axes to display bioactivity or any of 7 different physical properties. In this example, c-logP is plotted against exact mass. **b** Data points are displayed in *multiple colors* and *unique shapes*, quickly summarizing cherry-pick decisions made across multiple pages. Here, activity-based decisions (pick or force pick) are displayed. *Green squares* represent compounds with acceptable bioactivity and physical properties. *Red crosses* are compounds with acceptable bioactivity but that fail one or more physical property filters. *Blue stars* are compounds that were force-picked on the bioactivity page with acceptable physical properties. *Green pluses* are force-picked compounds that do not meet the c-logP criterion. **c** A summary of the decisions made (pick/discard) from this page is listed here. **d** Tool tips displaying bioactivity, hit-calling and cherry-pick decisions, selected physical properties and structure can be opened by placing the cursor over individual data points. **e** Detailed compound information is displayed in a table for highlighted data points

valid scaffolds and preliminary SAR trends to guide follow-up medicinal chemistry efforts. In addition to cross-collection searching, the user can manually add compounds by ID number. Such manual additions must be explicitly justified, and the reasons provided by the user are recorded for future reference.

These four workflow pages comprise the primary tools used at the Broad Institute for cherry picking compounds to re-test at different concentrations in screening assays. As mentioned above, this process is a non-linear one and researchers can use each section as they deem appropriate. For larger hit lists, one may want to start with the application of physical property filters to reduce the number of compounds to a more manageable number. Conversely, overly stringent assay cutoffs may lead to a small hit list. In this instance, lowering the bioactivity criterion could increase the number of compounds for consideration.

To provide an example of the cherry-picking process, we examined an MLPCN project conducted recently at our facilities [32]. For this screen, 299,653 compounds from the MLMSR collection were evaluated in a protein-binding fluorescence polarization assay, and substances exhibiting greater than 15 % inhibition at 7.5 μM were classified as "active". Based on this threshold, there were 2,736 actives, 31 inconclusives, and 293,886 inactives. Using our customized Spotfire cherry-picking workflow, this initial hit list of ~2,700 compounds was pared down to 1,058 substances for subsequent re-test in dose–response format (Fig. 5).

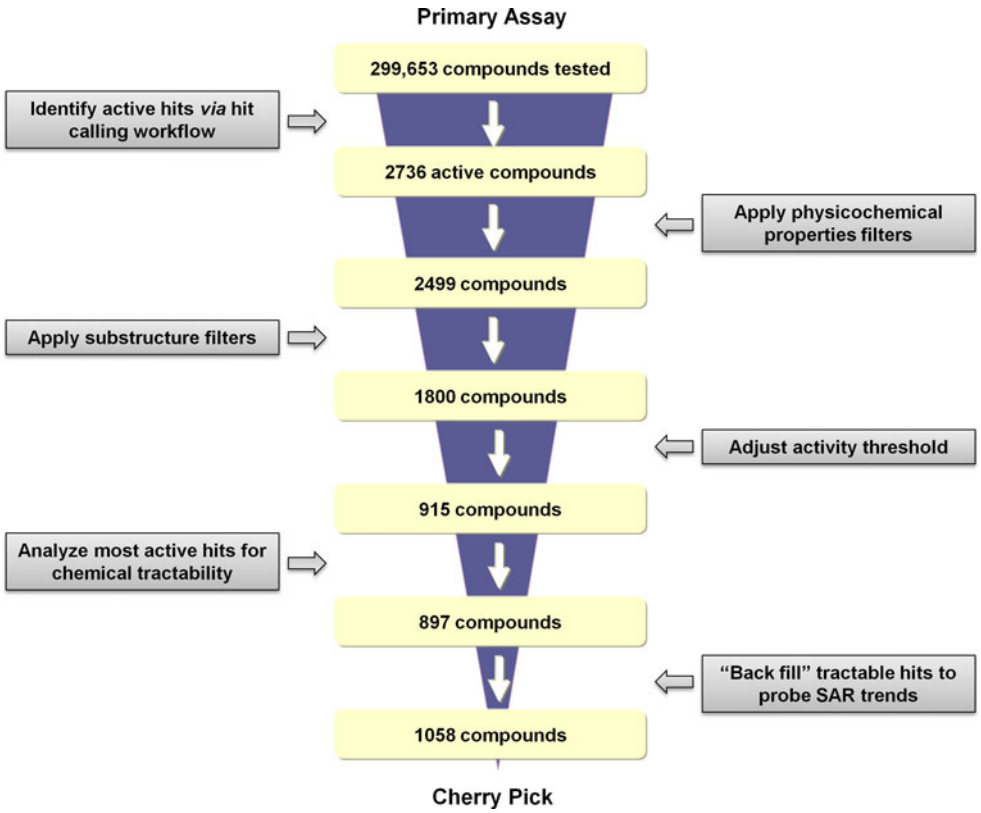


Fig. 5 Hit-calling/cherry-picking workflow

The process began with the evaluation of the 31 inconclusive compounds. This subset of compounds displayed only modest levels of target inhibition (13–29 % at 7.5 μM) and included eight electron-rich furans, two Michael acceptors, and five α-thio amides. Collectively, these substances were not particularly noteworthy, and the entire subset was discarded from the cherry-pick list.

The 2,736 actives were then analyzed by their physical properties. These compounds were generally compliant to Lipinski's guidelines, and there did not appear to be strong correlations between bioactivity and molecular weight, c-logP, hydrogen bond donors, or any other parameter. Consequently, a conservative approach was adopted and only compounds with molecular weights between 150 and 500, while possessing c-logP between 0 and 5, were retained; 237 compounds were removed at this stage, leaving 2,499 actives.

According to the tabulated substructure filter data, 30 of 52 undesirable functionalities were present among the remaining 2,499 actives. The least prevalent features, such as hydroxylamines and pyridinium salts, were immediately discarded without further consideration. Given the number of cherry-pick candidates available, it was considered more efficient to apply structural filters more rigidly. However, the more highly represented functional groups were carefully examined to determine if there were any prominent scaffolds present that would be worth investigating. For example, the thiourea moiety was identified in 51 active compounds. While the bioactivity of these thioureas ranged from 30 to 100 % inhibition, examination of this small collection revealed there was one particular scaffold that appeared repeatedly. The prevalence of this structure among the active compounds suggested this may be a genuine scaffold of interest, and five members of this family were selected for re-testing in order to validate the potential of this scaffold. A similar evaluation of the larger families of compounds was then conducted, and a total of 699 compounds were eliminated for violating one or more of the 30 substructure filters.

To further reduce the number of compounds for retest, the activity threshold was raised so that only compounds demonstrating >45 % inhibition at 7.5 μM were retained. Increasing the stringency of the activity criterion from the original 30 % inhibition cut-off removed another 885 compounds to produce a cherry-pick list containing 915 unique compounds. The 50 most active compounds of the remaining 915 actives were then individually evaluated for synthetic tractability. A number of these candidates were determined to be suboptimal starting points for future SAR investigations, and 18 compounds were subsequently eliminated as being synthetically intractable. The remaining 32 of the 50 were used to back-fill the final list. With the assistance of the Tanimoto search function, a number of analogs were identified for each of the 32 compounds, and

these were added to the cherry-pick list. Several of these analogs were already declared inactive in the primary assay, but were "force picked" as negative controls. Less-active derivatives were also selected in order to enable preliminary SAR analysis. Querying the Broad internal collection revealed that 5 of the 32 hits possessed closely related analogs in our in-house library. These related compounds were also added to the cherry-pick list to provide additional SAR data. Completion of the back-fill increased the final cherry pick list to 1,058 compounds. The entire ~300,000 compound data set and selection criteria were exported to our database for archiving and the finalized cherry-pick list was distributed to compound suppliers at the NIH repository and Broad Institute's compound management group for fulfillment.

Development of a unique tool to visualize stereochemical structure–activity relationships: S/SAR viewer

The DOS collection at the Broad Institute incorporates complex scaffolds, rich in sp^3 content and chiral centers [10]. The collection was also designed such that all possible stereoisomers for a given compound would be present. Thus, the HTS data from our DOS collection enable two levels of analysis. In the first level, traditional SAR can be studied based on the building-block diversity. In the second level, stereo-structure/activity relationships (SSAR) [33–36] can be derived from the sets of stereoisomers that correspond to the same structural isomer. This second level of systematic and comprehensive stereochemical diversity is rarely available in screening collections and, to our knowledge, cheminformatics tools to study such stereochemical dependencies are rare [37, 38] . Therefore, we decided to build the S/SAR-viewer to extract stereochemical dependencies of biological activities residing in HTS data from our DOS collection.

For a scaffold in the DOS collection, SAR visualization would ideally display the biological activity resulting from a combination of both R-group substitutions and the scaffold's stereoisomers. A visualization tool for the DOS collection also needs to be flexible, allowing display for a range of stereoisomeric matrices (2, 3, 4, or more stereocenters), R-group dimensions (2, 3, or more "diversity sites"), and color schemes correlating to the biological activity.

We chose to implement our solution to the SAR and SSAR (or S/SAR for both) visualization challenge using TIBCO's Spotfire software, with a template file containing custom scripts to automate retrieval and display of data. Spotfire has a number of advantages for visualization, including the ability to display high-dimensional data using hierarchical plot axes, plotting colors, and shapes. In a typical S/SAR plot, the dimensions displayed could be structures of the R_1 group, structures of the R_2 group, stereocenter 1 configuration (*S* or *R*), stereocenter 2 configuration, stereocenter 3 configuration, and assay result. It is straightforward to display all of these dimensions simultaneously in Spotfire, and for the user to modify and adjust the display as needed.

The S/SAR visualization process starts by importing the appropriate data, either from the database (guided by menus within Spotfire), or from a manually uploaded file containing compound identifiers and columns of data. Regardless of the source, the system looks up the structure, R-group decomposition, and stereocenter configuration information for each compound in the data set. After this information is loaded and displayed in the S/SAR viewer, the user can select from a drop-down menu containing a list of scaffolds used in the R-group decomposition to visualize different sets of compounds. The basic steps in this process (loading of data, R-group decomposition, display of data, and analysis of the displayed data) are described in more detail in the following sections.

Loading data

One of the S/SAR tool's primary applications is analyzing the results of primary screens of DOS compounds. Therefore, we built functionality to load this specific type of data directly from the database for S/SAR visualization. Within the S/SAR Spotfire template, "information links" have been created for automating the retrieval of data, allowing the user to choose a Genedata Screener assay analyzer QC session to load.

The QC session is not adequately formatted for visualization in the S/SAR viewer as multiple data points can be collected for each compound (typically compounds are tested in duplicate or triplicate). The multiple data points for each compound are stored as separate rows, requiring a "pivot" step during information retrieval so data corresponding to one compound is condensed into a single row. The S/SAR Spotfire template was configured to provide a default aggregation of these data columns by averaging the results, but the user can also create their own custom aggregation method using Spotfire's custom column tools (Table 1).

In addition to database retrieval, the user has the ability to load data from a file, providing greater flexibility in the choice of data to be displayed. For example, users can display data corresponding to secondary screening results, purity of the compounds, or a screening plate's composition.

Table 1 Comparison of data retrieved from database to data pivoted and aggregated in Spotfire in preparation for being displayed

Data from database		Pivoted data			
Compound ID	Intensity	Compound ID	Intensity 1	Intensity 2	Intensity average
1	3.3	1	3.3		3.3
2	1.1	2	1.1	1.3	1.2
2	1.3	3	4		4
3	4				

R-group decomposition

Whether the data to be displayed are retrieved from the database or loaded from a file, the next step is to identify the structural information associated with each of the compounds in the data set—the full structure, the scaffold that matches the structure, the R-groups, and the configuration at each stereocenter of interest (Table 2). This information is stored in the database as a function of compound identifiers assigned to each of the compounds in the DOS collection. An information link in Spotfire is used to retrieve this information for the compounds in the previously loaded data. However, before R-group decomposition lookup can occur, the R-group decomposition must be calculated and stored.

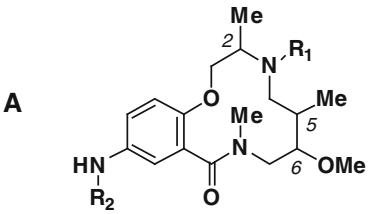
Calculation of the R-group decomposition for the compounds in the DOS collection is carried out using a custom-built application (written using the Grails framework) running on a server, using ChemAxon's JChemBase suite of Java libraries, and engineered to work in parallel architectures (to take advantage of the common multi-core CPUs and distributed systems). A table in the database stores the extended ChemAxon SMILES defining the structure, R-group, and stereocenter label information of the various DOS scaffolds (Fig. 6). The calculation starts by loading these SMILES and comparing the structures against each compound in the DOS collection using the ChemAxon R-group decomposition method. When a match is found, the identity of the core and the corresponding R-groups are associated with the structure. In addition, stereocenter configurations for each stereocenter of interest in the compound are calculated and stored. Finally, the results are uploaded to the database. Throughout this process, numerous validations are performed and warnings can be issued. For example, if a structure is found to match more than one core, a warning is issued so that structural ambiguity can be addressed.

When new compounds are added to the DOS collection, or if an existing structure needs modification ("curation"), then the calculation can be repeated (job scheduling/triggering is managed using Apache Camel). A prototype version of the R-group decomposition calculation can be run in $\sim$10 min for $\sim$96,000 compounds containing $\sim$30 different scaffolds using a laptop with a 2.2 GHz Intel CORE i7 vPro CPU containing 8 cores (8 calculations running in parallel on the same chip). This time measurement illustrates two points. First, it is more efficient to provide this R-group decomposition information via

Table 2 Example of pivoted, aggregated data with R-group decomposition and stereocenter configuration information

Compound ID	Data	Structure		Stereochemical configuration		Collection name	Core
		Rl	R2	C2	C5		
1	3.3	CN	clcccccl	S	R	Nocoll	C1([R1])CCC([R2])CC1
2	1.2	C#N	clcccccl	S	S	Nocoll2	C([R1])CCC([R2])C
3	4	C=C	C1CCC1	R	R	Nocoll Extra	C1([R1])CCCC1[R2]
4	7.9	CN	clcccccl	R	S	Nocoll	C1([R1])CCC([R2])CC1
5	0.15	C#N	clcccccl	R	R	Nocoll2	C([R1])CCC([R2])C
6	0.9	C=C	C1CCC1	S	S	Nocoll Extra	C1([R1])CCCC1[R2]

Columns R_1, R_2, and core contain the SMILES [39] representation of the respective structures. Columns C_2 and C_5 contain the stereocenter configuration as determined by Cahn-Ingold-Prelog (CIP) conventions



B CC(CO)N1CC(C)C(CN(C)[*])OCc2ccccc2-c2ccccc2C1=O |$;;;;;;;;;;;;_R1;;;;;;;;;;;;;;;;;$,$_AV:;C2 Configuration;;;;C5 Configuration;;C6 Configuration;;;;;;;;;;;;;;;;;$|

Fig. 6 **a** Structure with R-group and stereocenters labeled for a core (from the Head-to-Tail Library [6c] within the DOS collection). **b** ChemAxon extended SMILES encoding the structure, R-group, and stereocenter labels

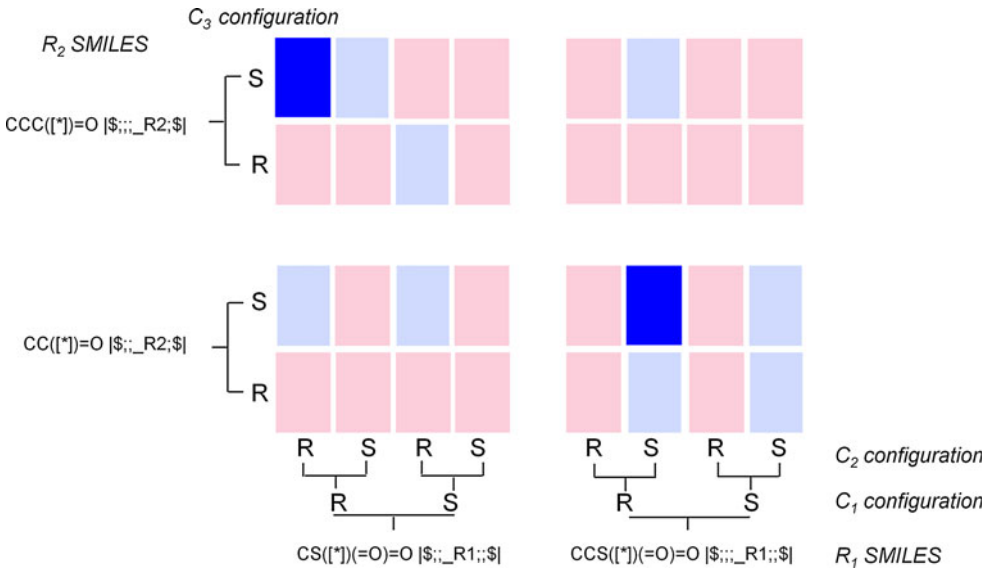


Fig. 7 An example heat map, axes labeled with R-group structures and stereocenter configuration

database lookup (3 min for database lookup for a comparable number of compounds). Second, it is not computationally prohibitive to repeat the entire calculation for even minor changes to the collection or set of core structures.

Data display

After the user has loaded data and R-group decomposition information into the S/SAR Spotfire template, the next step is to create the view where the data will be displayed. The basis for the S/SAR display within Spotfire relies on the ability to setup a hierarchical arrangement on each axis. For SAR visualization, the highest level of the hierarchy is the R-group (R_1, R_2, etc.). The next level is the stereocenter configurations to capture SSAR. For example (Fig. 7), the following hierarchy is used on the horizontal axis: R_1, C_1 Configuration, C_2 Configuration. This hierarchy is represented graphically on the axes of the viewer. For each R-group (represented as SMILES), the possible values of the C_1 stereocenter configuration are present at the next level of the hierarchy. For each of the C_1 stereocenter configuration values, the third level of the hierarchy contains each possible value of the C_2 stereocenter configuration. The empty values are deliberately added during data retrieval, and they provide the vertical and horizontal white space that provides a visual separation between the stereoisomer rectangles of 8 data points.

This hierarchy describes how the heat map (Fig. 7) should be interpreted. Each block of 8 rectangles corresponds to the intersection between R_1 and R_2 on the heat map. The 8 rectangles correspond to data values obtained for each of the 8 stereoisomers that are possible based on the 3 stereocenters in this core (C_1, C_2, C_3). The 2 rows within each block correspond to the *R* and *S* configurations of the C_3 stereocenter in the core, respectively. The left-most column in the block of 8 corresponds to the *R* and *R* configurations for stereocenters C_1 and C_2, respectively. The next column corresponds to the *R* and *S* configurations for these stereocenters, respectively, and so on for the remaining columns. It is important to note that this hierarchical arrangement scales to any number of stereocenters. For example, for a core with a 4th stereocenter, the configuration values for this could be added as a third level of the hierarchy on the vertical axis.

A critical part of the S/SAR heat map is the display of the structure of the R-groups on the axes of the plot. Unfortunately, it is not possible to do this directly for any of the views provided in Spotfire. We have achieved a similar effect by adding additional padding to our data, which consists of entries without compound identifiers or data, but instead containing SMILES for the appropriate R-groups. We then use the labeling mechanism in Spotfire to label just these data points. An option within the labeling further lets us choose to render the value as a chemical structure using external software (an example is shown in Fig. 8). This workaround has limitations concerning Spotfire's zoom and pan functionality, since the structures are part of the graph, not part of the axes. However, we are currently working on an enhanced script using the Spotfire filters that would automatically render the structures in the graph once the zoom or pan has been executed.

The configuration of the S/SAR view has been automated and the user can choose (from a drop-down menu) the scaffold for which they would like to display the data. Once the scaffold is selected, the system then sets the Spotfire filters so that only data for compounds that match the chosen scaffold are presented. The script also reconfigures the axes of the heat map so that the correct set of R-groups and stereocenter configurations are on the axes of the heat map.

Analysis of S/SAR heat map

To illustrate how the S/SAR analysis can be useful in prioritization of hit series, we provide representative S/SAR displays for a set of DOS compounds with a common scaffold (Fig. 8). The scaffold has three stereocenters that lead to a total of 8 possible stereoisomers. The activities are presented as a binary set of (hypothetical) data with the darker color indicating an active compound and the lighter color indicating no activity. Compounds with the highest activities have an R_1 substituent with the structure of a *para*-methoxyphenyl sulfonamide and an R_2 group of a phenylacetamide or a 2-indoleacetamide structure (Fig. 8a). The SAR is seen clearly, but we also can view a clear SSAR trend. By viewing the color-coded blocks, we see the pattern of activity among the stereoisomers is exactly the same for both sets of compounds. Having tractable SAR and SSAR can validate a hit series allowing for prioritization, if necessary.

The chart pictured in Fig. 8b also indicates a set of hits that have the same SAR for the R_1 and R_2 groups as is present in the chart in Fig. 8a. However, there are multiple stereochemical configurations that have strong activity and no definable pattern emerges, suggesting that stereochemistry does not play a key role in this activity. A hit set with this type of activity may be de-prioritized when comparing results among multiple sets of compounds with varying degrees of stereoselectivity due to anticipated off-target activity in down-stream assays.

A recent report on an HTS measuring the suppression of cytokine induced β-cell apoptosis presents an example of the powerful data-analysis capability of the S/SAR viewer [40]. A DOS library was screened in a cell-based assay and a number of hits were obtained, including many with a common scaffold (Fig. 9). The compounds contain an eight-membered ring scaffold with three stereocenters, two of which are contained in the ring system. The red blocks indicate compounds with the highest activity measured as

Fig. 8 **a** An example view of hits with selective SAR and SSAR. **b** A view of selective SAR but non-selective SSAR

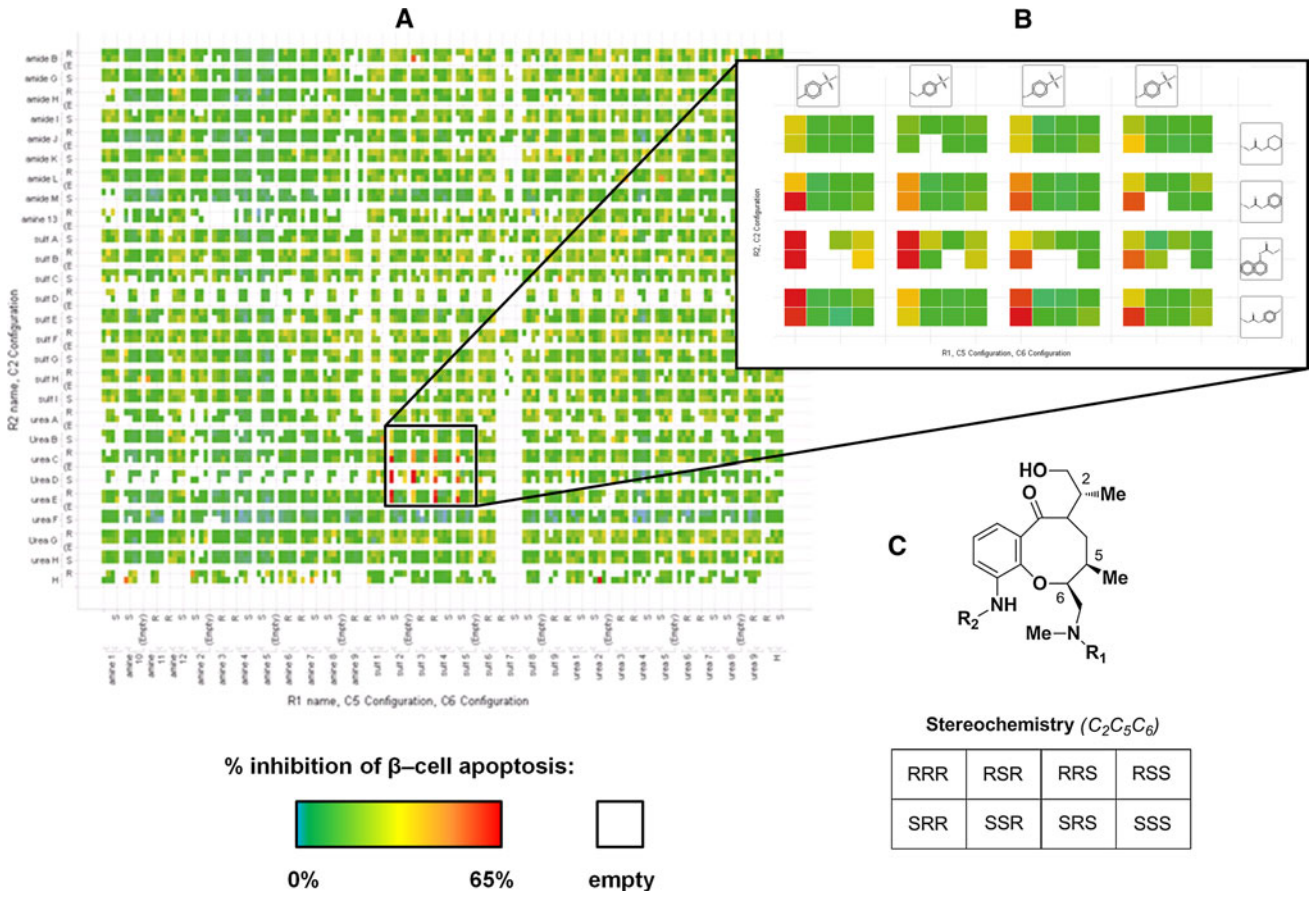


Fig. 9 **a** View of data from the β-cell apoptosis HTS. **b** A magnified view of the hits including R-group structure. **c** Key to core structure and stereochemical assignments

percent inhibition of apoptosis. A cluster of hits that all contain urea substituents at R_2 and sulfonamide groups at R_1 stands out clearly. The pattern of stereochemical configurations is the same for all of these hits; the most active stereoisomer is the *SRR* at C_2:C_5:C_6. To a lesser extent the *RRR* isomer retains some activity, indicating that the exocyclic stereocenter is less critical for activity than the C_5 and C_6 centers. A detailed view is presented showing a number of the hits. At this level, structures become visible, better illustrating the S/SAR for this hit series. The authors used these data to prioritize the hit series and eventually prepared additional active analogs for further probing of this biological system.

Summary

Traditionally, non-profit research institutes do not have access to the extent of resources that belong to large pharmaceutical companies. Therefore, it is of intense interest for an academic screening center to have tools that enable the most informed decision making as early as possible in a project, in order to select candidates most likely to be successful. Cheminformatics plays an important role in the processing and analysis of large amounts of data obtained from HTS campaigns. The hit-calling and cherry-picking workflows that have been developed and implemented for the Chemical Biology Platform at the Broad Institute enable analysis of data from large libraries (>350,000 compounds) screened during HTS campaigns. All decisions are tracked for future reference and the activity thresholds are updated as new data become available. This decision tracking not only improves the ability to easily return to the original steps to change thresholds and filters after viewing results from downstream analyses, but the visibility of these decisions to multiple researchers provides for efficient data sharing and collaboration and promotes use of "best practices". Capture of the decision-making process also provides context for cross-assay analysis of compound activity, such as whether a given compound was considered active in a number of assays [2, 41]. The cherry-pick tool analyzes the hit set for drug-like properties and structural information, and undesirable compounds can be easily removed. The ability to mine the collection for structurally similar compounds and easily

add them to the prioritized list enables built-in SAR to be determined from the confirmatory and follow-up assays.

The DOS compound collection, with its degree of stereochemical complexity, presents additional challenges for the analysis of HTS data. The S/SAR viewer, built using TIBCO Spotfire tools, provides critical information on selectivity among multiple hits from an HTS of this collection. Patterns among data from thousands of compounds and their stereoisomers can be viewed clearly from a single heat map and the presence or absence of stereochemical dependencies can be used as another tool to help prioritize hits for follow up assays and medicinal chemistry.

Together, the tools described here help each research project team make informed decisions to create an optimized HTS hit list that will streamline future assays and medicinal chemistry efforts. To create each tool, we first had to understand and define the process at each step. This was achieved by effective communication between our informatics professionals (application engineers and software developers), our research scientists (chemists and biologists), and our computational scientists. Our experiences in defining the process and workflow for each tool may be applicable to other academic centers, and we are willing to discuss our experiences in more detail than provided by the scope of this article. Ultimately the investment in resources to create these tools will pay dividends in the form of more successful probe and lead development projects.

Acknowledgments The high-throughput screening work was funded by National Institutes of Health-Molecular Libraries Probe Production Centers Network (1 U54 HG005032-1) and the β-cell apoptosis work was in part funded by the National Institute of General Medical Sciences-sponsored Center of Excellence in Chemical Methodology and Library Development (Broad Institute CMLD; P50GM069721) and also the National Institutes of Health Genomics Based Drug Discovery U54 grants RL1CA133834 (administratively linked to NIH Grants RL1HG004671, RL1GM084437, and UL1DE019585). The authors would like to thank Dr. Michael Foley, Dr. Michelle Palmer and Dr. Mary Pat Happ for their help with manuscript preparation. We also thank David DeCaprio for his assistance in the development of a prototype S/SAR viewer.

References

1. Mayr LM, Fuerst P (2008) The future of high-throughput screening. J Biomol Screen 13:443–448
2. Seiler KP, George GA, Happ MP, Bodycombe NE, Carrinski HA, Norton S, Brudz S, Sullivan JP, Muhlich J, Serrano M, Ferraiolo P, Tolliday NJ, Schreiber SL, Clemons PA (2008) ChemBank: a small-molecule screening and cheminformatics resource database. Nucleic Acids Res 36:D351–D359
3. Agrafiotis DK, Alex S, Dai H, Derkinderen A, Farnum M, Gates P, Izrailev S, Jaeger EP, Konstant P, Leung A, Lobanov VS, Marichal P, Martin D, Rassokhin DN, Shemanarev M, Skalkin A, Stong J, Tabruyn T, Vermeiren M, Wan J, Xu XY, Yao X (2007) Advanced biological and chemical discovery (ABCD): centralizing discovery knowledge in an inherently decentralized world. J Chem Inf Model 47:1999–2014
4. Sander T, Freyss J, von Korff M, Reich JR, Rufener C (2009) OSIRIS, an entirely in-house developed drug discovery informatics system. J Chem Inf Model 49:232–246
5. MLPCN website http://mli.nih.gov/mli/mlpcn. Accessed 7 Jan 2013
6. 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
7. Clemons PA, Bodycombe NE, Carrinski HA, Wilson JA, Shamji AF, Wagner BK, Koehler AN, Schreiber SL (2010) Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. Proc Natl Acad Sci 107:18787–18792
8. Clemons PA, Wilson JA, Dančik V, Muller S, Carrinski HA, Wagner BK, Koehler AN, Schreiber SL (2011) Quantifying structure and performance diversity for sets of small molecules comprising small-molecule screening collections. Proc Natl Acad Sci 108:6817–6822
9. Marcaurelle LA, Comer E, Dandapani S, Duvall JR, Gerard B, Kesavan S, Lee MD 4th, Liu H, Lowe JT, Marié J-C, Mulrooney CA, Pandya BA, Rowley A, Ryba TD, Suh B-C, Wei J, Young DW, Akella LB, Ross NT, Zhang Y-L, Fass DM, Reis SA, Zhao W-N, Haggarty SJ, Palmer M, Foley MA (2010) An aldol-based build/couple/pair strategy for the synthesis of medium- and large-sized rings: discovery of macrocyclic histone deacetylase inhibitors. J Am Chem Soc 132:16962–16976
10. Gerard B, Duvall JR, Lowe JT, Murillo T, Wei J, Akella LB, Marcaurelle LA (2011) Synthesis of a stereochemically diverse library of medium-sized lactams and sultams via S_NAr cycloetherification. ACS Comb Sci 13:365–374
11. Fitzgerald ME, Mulrooney CA, Duvall JR, Wei J, Suh B-S, Akella LB, Vrcic A, Marcaurelle LA (2012) Build/couple/pair strategy for the synthesis of stereochemically diverse macrolactams via head-to-tail cyclization. ACS Comb Sci 14:89–96
12. Genedata Screener: http://www.genedata.com/products/screener.html. Accessed 7 Jan 2013
13. Swamidass SJ, Bittker JA, Bodycombe NE, Ryder SP, Clemons PA (2010) An economic framework to prioritize confirmatory tests after a high-throughput screen. J Biomol Screen 15:680–686
14. Swamidass SJ, Calhoun BT, Bittker JA, Bodycombe NE, Clemons PA (2011) Enhancing the rate of scaffold discovery with diversity-oriented prioritization. Bioinformatics 27:2271–2278
15. Swamidass SJ, Calhoun BT, Bittker JA, Bodycombe NE, Clemons PA (2012) Utility-aware screening with clique-oriented prioritization. J Chem Inf Model 52:29–37
16. Wawer M, Bajorath J (2009) Extraction of structure-activity relationship information from high-throughput screening data. Curr Med Chem 16:4049–4057
17. Harper G, Pickett SD (2006) Methods for mining HTS data. Drug Discov Today 11:694–699
18. Gribbon P, Lyons R, Laflin P, Bradley J, Chambers C, Williams BS, Keighley W, Sewing A (2005) Evaluating real-life high-throughput screening data. J Biomol Screen 10:99–107
19. Spotfire. http://spotfire.tibco.com. Accessed 7 Jan 2013
20. Pipeline Pilot. http://accelrys.com/products/pipeline-pilot/. Accessed 7 Jan 2013
21. ChemAxon. http://www.chemaxon.com/jchem/doc/dev/cartridge/index.html. Accessed 7 Jan 2013
22. The Tanimoto search is run through Oracle J Chem cartridge using 2D hashed fingerprints as discussed at http://www.chemaxon.com/jchem/doc/user/fingerprint.html. Accessed 7 Jan 2013

23. Yu B, Reynisson J (2011) Bond stability of the "undesirable" heteroatom-heteroatom molecular moieties for high-throughput screening libraries. Eur J Med Chem 46:5833–5837
24. Meanwell NA (2011) Synopsis of some recent tactical application of bioisosteres in drug design. J Med Chem 54:2529–2591
25. Kalgutkar AS, Gardner I, Obach RS, Shaffer CL, Callegari E, Henne KR, Mutlib AE, Dalvie DK, Lee JS, Nakai Y, O'Donnell JP, Boer J, Harriman SP (2005) A comprehensive listing of bioactivation pathways of organic functional groups. Curr Drug Metab 6:161–225
26. Chuprina A, Lukin O, Demoiseaux R, Buzko A, Shivanyuk A (2010) Drug- and lead-likeness, target class, and molecular diversity analysis of 7.9 million commercially available organic compounds provided by 29 suppliers. J Chem Inf Model 50:470–479
27. Lisurek M, Rupp B, Wichard J, Neuenschwander M, Kries JP, Frank R, Rademann J, Kuehne R (2010) Design of chemical libraries with potentially bioactive molecules applying a maximum common substructure concept. Mol Divers 14:401–408
28. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 23:3–25
29. Wager TT, Hou X, Verhoest PR, Villalobos A (2010) Moving beyond rules: the development of a central nervous system multiparameter optimization (CNS MPO) approach to enable alignment of druglike properties. ACS Chem Neurosci 1:435–449
30. Petit J, Meurice N, Kaiser C, Maggiora G (2011) Softening the Rule of 5-where to draw the line? Bioorg Med Chem 20:5343–5351
31. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL (2012) Quantifying the chemical beauty of drugs. Nat Chem 4:90–98
32. These data will be made available through PubChem this year
33. Zhang QS, Lu HJ, Curran DP (2004) Fluorous mixture synthesis of stereoisomer libraries: total syntheses of (+)-murisolin and fifteen diastereomers. J Am Chem Soc 126:36–37
34. Curran DP, Zhang Q, Richard C, Lu H, Gudipathi V, Wilcox CS (2006) Total synthesis of a 28-member stereoisomer library of murisolins. J Am Chem Soc 128:9561–9573
35. Dandapani S, Jeske M, Curran DP (2005) Synthesis of all 16 stereoisomers of pinesaw fly sex pheromones—tools and tactics for solving problems in fluorous mixture synthesis. J Org Chem 70:9447–9462
36. Wrona IE, Lowe JT, Turbyville TJ, Johnson TR, Beignet J, Beutler JA, Panek JS (2009) Synthesis of a 35-member stereoisomer library of bistramide A: evaluation of effects on actin state, cell cycle and tumor cell growth. J Org Chem 74:1897–1916
37. Tanikawa T, Fridman M, Zhu W, Faulk B, Joseph IC, Kahne D, Wagner BK, Clemons PA (2009) Using biological performance similarity to inform disaccharide library design. J Am Chem Soc 131:5075–5083
38. Kim Y-K, Arai MA, Arai T, Lamenzo JO, Dean EF, Patterson N, Clemons PA, Schreiber SL (2004) Relationship of stereochemical and skeletal diversity of small molecules to cellular measurement space. J Am Chem Soc 126:14740–14745
39. ChemAxon website: http://www.chemaxon.com/marvin/help/formats/smiles-doc.html. Accessed 7 Jan 2013
40. Chou DH-C, Duvall JR, Gerard B, Liu H, Pandya BA, Suh B-C, Forbeck EM, Faloon P, Wagner BK, Marcaurelle LM (2011) Synthesis of a novel suppressor of β-cell apoptosis via diversity-oriented synthesis. ACS Med Chem Lett 2:698–702
41. Petrone PM, Simms B, Nigsch F, Lounkine E, Kutchukian P, Cornett A, Deng Z, Davies JW, Jenkins JL, Glick M (2012) Rethinking molecular similarity: comparing compounds on the basis of biological activity. ACS Chem Biol 7:1399–1409