HiTSEE: A Visualization Tool for Hit Selection and Analysis in High-Throughput Screening Experiments

Enrico Bertini Hendrik Strobelt Joachim Braun Oliver Deussen Ulrich Groth
Thomas U. Mayer Dorit Merhof*

University of Konstanz

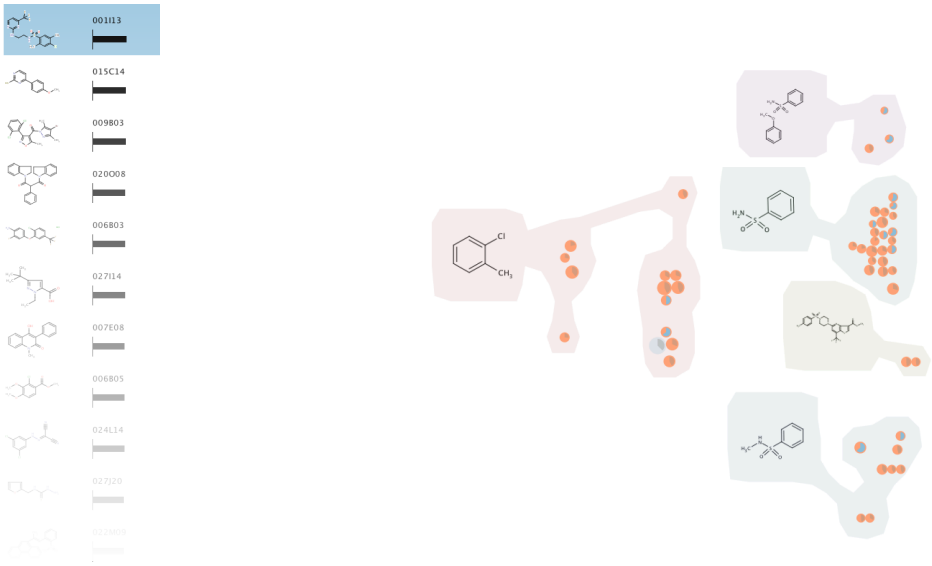


Figure 1: HiTSEE's main view elements

Abstract

We present HiTSEE (High-Throughput Screening Exploration Environment) a visualization tool for the analysis of large chemical screens for the analysis of biochemical processes. The tool supports the analysis of structure-activity relationships (SAR analysis) and, through a flexible interaction mechanism, the navigation of large chemical spaces. Our approach based on the projection of one or few molecules of interest and the expansion around their neighborhood allows for the exploration of large chemical libraries without the need to create an all encompassing overview of the whole library. We describe the requirements we collected during our collaboration with biologists and chemists, the design rationale behind the tool, and two case studies.

1 Introduction

Genetics has been widely used in the past to study complex biological processes within a cellular system and to elucidate the function of proteins. As genes encode proteins, gene function can be modulated through a mutation, which in turn perturbs the function of the protein of interest and either affects its activity or entirely suppress its expression ('knockout'). As a result, the physiological effect observed in the phenotype allows to identify protein function.

*emails: [surname.name]@uni-konstanz.de

Although genetic approaches have proven to be extremely powerful in elucidating the principles of a wide range of biological processes, there are a number of substantial limitations to this approach, most importantly the lack of temporal control required to study dynamic processes, since a protein cannot be turned on or off on demand. A more recent approach to study protein function which overcomes this limitation is chemical genetics. In chemical genetics, biological systems are studied using cell-permeable small molecules (compounds), which inhibit the protein under investigation (chemical knock-out). This approach makes it possible to perturb protein function rapidly, reversibly and conditionally with temporal and quantitative control, both in cultured cells or whole organisms [10].

The foundation of chemical screens are commercially available compound libraries comprising hundreds of thousands of small molecules which cover a high degree of structural diversity. In order to switch a protein off, a compound needs to be identified which inhibits the protein under investigation and hence allows to study its function. For this purpose, high-throughput screening is performed, which is a major technological breakthrough in biology experimentation [6].

Although experimentation capabilities have increased significantly over the last years, resulting in vast amounts of data generated in high-throughput screenings, analysis methods that are able to handle and process large amounts of data lag behind and don't scale equally fast. For this reason, many sites where high-throughput screenings are performed use sub-optimal solutions which are either too slow or suffer from a limited scope of anal-

ysis.

The development of HiTSEE stems from the analysis of HTS data analysis practices performed by several researchers at the School of Chemical Biology at the University of Konstanz and by the analysis of existing HTS tools. We discovered that electronic spreadsheets is the main data analysis tool employed by the researchers and that their data exploration capabilities are, as a consequence, extremely limited. These practices not only leave room to several kinds of mistakes, but they also hinder the possibility to effectively explore the chemical space and relate activity levels to structural features.

At the same time, all the tools we have analyzed did not completely fit the needs of our researchers. While the whole field of Chemoinformatics has developed numerous and impressive computational tools for drug discovery, there is a lack of flexible visualization tools that allow for lower-scale smooth exploration of chemical spaces. During our analysis we reviewed a number of visualization tools for structure-activity relationships (we provide a full description and comparison in Section 7) but none of them seemed to fit the needs we encountered. We believe this is due to three main factors: (1) the tools tend to focus either on gaining an overview of a chemical space or on the exploration of the neighborhood of a single compound; (2) the tools tend to focus either on the comparison of entire molecules or on their fragments; (3) many tools offer limited navigation and interaction capabilities.

HiTSEE addresses these issues by providing a multi-view interactive system in which it is possible to project one or more compounds of interest and explore a neighborhood. The tool features flexible navigation capabilities that allow the user to easily jump from one chemical context to another.

The main contributions of this paper are: the in depth analysis of the HTS problem with a group of involved researchers in biochemistry, the design rationale and development of a flexible visual HTS analysis tool, and its interaction paradigm.

The validity of HiTSEE is demonstrated by two case studies performed by experts from biochemistry. The presented approach is of major interest for biologists involved in high-throughput experiments and visualization designers that want to learn from a real design study.

The paper is organized as follows: the following section describes HTS in more details to provide the right context to readers not familiar with the process, Section 3 describes the data processing steps needed before the data could enter into the system and the tasks collected during our collaboration, Section 4 describes HiTSEE and its design, Section 5 illustrates the case studies, Section 7 provides a critical review of similar visualization systems, Section 6 offers some reflections and lessons learn from the process, and Section 8 provides the conclusions and outlines our plans for future work.

2 High-Throughput Screening (HTS)

The main motivation behind HTS is the need to quickly test a large number of chemical compounds against a biological target (e.g., protein or cell) to restrict the scope of the analysis to a manageable number of potentially relevant molecules called *hits*.

The researcher uses predefined chemical libraries containing thousands of chemical compounds to be used in testing. A robot handles the preparation and execution of the experiment. The chemical compounds are arranged in *assay plates* and mixed with the biological target. The plate is a plastic rectangular container featuring a grid of *wells* in which the compounds and the target are placed for reaction. Figure 2 shows an example of *assay plate*.

After the time has passed for the biological target to react (or not react), the machine can measure different external physical features like reflectivity, fluorescence or absorbance to determine an activity level (e.g., the level of binding between a protein and the chemical compound). Together with this main information the machine can provide additional parameters for each well that help in the interpretation of the experiment and in the assessment of its quality. In addition to containing target samples, the assay normally contains control elements to baseline the measurement results.

Figure 2: Picture of an assay plate.

Typically, the generated data is organized in a format resembling the physical structure of the well plates. Figure 3 provides an example coming from one of the spreadsheets used by one of the biologists in our group.

<>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
A	117	124	101	127	117	86	117	108	103	114	136	109	104	111	122	109	109	100	109	105	18	20
B	116	112	105	116	109	115	115	115	115	121	118	120	270	105	112	116	126	104	70	116	20	20
C	120	123	100	116	118	136	36	112	111	113	114	112	119	122	127	103	114	117	108	116	22	21
D	123	105	115	111	120	111	115	125	117	140	110	135	207	125	116	235	131	110	101	132	21	21
E	116	116	113	107	110	105	30	112	132	117	115	114	105	126	113	113	114	112	103	112	21	26
F	130	103	114	114	131	111	110	106	181	123	118	138	113	113	114	266	130	105	100	119	21	26
G	125	100	108	107	108	110	107	122	108	107	113	116	112	118	116	108	111	127	104	116	13	20
H	110	110	138	106	14	106	120	110	114	106	107	116	113	103	113	122	106	111	107	116	21	20
I	116	107	115	115	110	34	110	78	104	110	112	113	126	110	116	115	102	116	110	113	100	25
J	36	108	133	111	108	114	151	260	111	134	114	36	124	114	116	131	117	114	115	107	100	16
K	115	111	106	115	121	110	107	114	107	113	111	115	112	114	120	115	103	103	114	112	102	13
L	132	109	290	118	108	110	113	126	104	152	87	114	134	113	120	152	119	112	113	113	101	24
M	116	36	108	105	112	111	111	109	110	111	133	114	110	122	103	118	118	108	113	110	103	23
N	115	109	119	106	106	103	107	123	111	124	102	107	175	107	107	122	112	118	110	109	107	14
O	107	58	111	103	119	103	113	105	109	109	115	101	233	107	105	111	111	119	119	119	114	25
P	142	140	114	109	131	103	102	101	101	111	107	110	97	105	107	56	-12	109	33	105	100	20

Figure 3: Example of how data generated by an HTS system is formatted in a spreadsheet used by one of the biologists in our group.

Once the data has been collected, the researcher goes through the following stages.

Data Processing & Quality Control. The researcher normalizes the data against the control values and analyzes the result to check for abnormal behaviors. A number of biases and outliers can exist in a HTS experiment. In our environment we typically check for large variances in the readouts between one experiment and another, assay plates with a low Z-factor (a measure of assay quality based on the controls). This phase is also supported by a visual tool we have developed, which is not the main focus of this paper, to visually explore the values in the plates. The researchers can directly filter out compounds with not reliable readouts or simply mark them for future analysis.

Hit Selection. This phase's goal is to identify the compounds that reacted in the experiment. Typically, the researcher organizes the results in a list sorted according to activity level and chooses a threshold value above which the compounds are considered active. These compounds are called *hits*.

Hit Confirmation. After hit identification, the researcher re-tests the selected hits in a new and more focused screening (e.g., testing different concentrations of the compounds and calculating the *half maximal inhibitory concentration* (IC_{50})) in which confirmation of activity is sought. This phase is normally done manually given the small number of molecules to test.

Hit Exploration. After hit confirmation the researcher explores the chemical space around the hits. Typically, he or she is in search of relationships between molecular structure and activity to isolate molecular fragments that induce activity (a process called *structure-activity relationship (SAR) analysis*).

Hit Expansion. Similar to hit exploration, hit expansion focuses on exploring the space around the hits but the focus is on the identification of alternative molecules that retain the desired properties and meet additional requirements (e.g., solubility).

Each one of these steps can involve data and computational resources. While in our environment we provide support for all these stages, HiTSEE provides support only for a subset of them, namely: *selection*, *exploration*, and *expansion*.

3 Data Processing and Tasks

In the following we provide additional details about the data and describe how it is processed before entering into the description of HiTSEE. Then, we describe how we gathered the requirements for the design of the tool and the main motivations behind our focus on a subset of the HTS tasks.

3.1 Data Processing

As mentioned in the previous section, the basic data format returned by the screening machine is a flat file organized around the shape of a well plate (Figure 3). HiTSEE leverages on a whole data analysis environment based on the KNIME platform [1]. KNIME is a well-known data mining framework based on a work-flow paradigm where data is processed by connecting data processing *nodes* one to another (Figure 4). It features an extensive and extensible library of nodes with a variety of purposes, e.g., querying, data mining algorithms, biochemical libraries. Within KINIME we have implemented a number of additional nodes to process the HTS data. More specifically, we process the data through the following nodes:

1. *Data Normalization* - The system permits to apply several kinds of normalization and takes into account different plate formats. Typically at this stage the system normalizes the data taking into account the values found in the control cells using the average value of the positive and negative controls.

2. *Quality Control* - At this stage the researcher can use a variety of tools we have developed to asses the quality of the experiment and to filter out or mark values with unusual behaviors. Many of the functions we have implemented leverage on a plate view through which the user can observe the distribution of the activity levels across the plates.

3. *Fingerprints Generation* - Before entering HiTSEE, the molecular structure of each compound (described by the SMILE format in the database) has to be translated into a format that allows structural similarity comparisons between the molecules. Thus, we transform the molecular descriptions into binary vectors called *fingerprints*.

Since this last fingerprint generation step is critical for the way HiTSEE arranges the molecules in its main view we provide additional details about it.

3.1.1 Fingerprints Generation

Chemoinformatics applications use fingerprints (FPs) as a way to allow similarity search and comparison between molecules. The basic idea behind FPs is to describe each molecule with a numeric vector that captures relevant properties of molecules. While FPs can capture a variety of molecular features, structural fingerprints are above all the most popular. Structural fingerprints are based on the concept of molecular fragments, that is, subsets of atoms and bonds found in the original sets, and describe each molecule in terms of the presence or absence of a molecular fragment. A fingerprint is thus a (normally very long) binary vector where each entry represents a fragment, and the value is set to one if the fragment is contained in the molecule and to zero otherwise. Through such a binary representation it is possible to compare the structural similarity of molecules: two molecules with similar vectors contain similar molecular fragments.

Static and dynamic fingerprints exist: the former employ a predefined set of common molecular fragments, independently from the specific chemical library under observation; the latter generates a customized set of fragments extracted from the library itself.

For more information on fingerprints and related techniques in chemoinformatics Leach and Gillet [8] provide an excellent introduction to the aforementioned concepts.

We tested several fingerprint libraries and eventually decided to build our own fingerprint generator based on a existing fragment mining function within KNIME. The main motivation behind our choice is the additional flexibility and transparency gained by being in control of all the steps required to generate the FPs. One major problem we had with existing FPs libraries is their extremely limited transparency: virtually all the libraries we have tested work in a black-box fashion and it is thus impossible to control and understand the fingerprint generation process. In particular, given the complex hash-based compression mechanisms they employ it is not possible to trace back the connection between the bits in the vectors and the molecular fragments they refer to.

Our fingerprint generation function developed in KNIME leverages on an existing function called MoSS (Molecular Substructure Search), which implements the fragment mining method developed by Borgelt and Berthold in [2]. The user can run the fragment mining function on the chemical library under observation and generate a number of relevant fragments. Our developed node takes these fragments as an input and builds, for each compound in the library, a fingerprint based on them. This approach has two main advantages: (1) the researcher can control the features of the generated fragments by setting up the controls of the MoSS node (e.g., size of the fragments, presence of aromatic rings, etc.), (2) the system keeps track of the connection between the bits in the vector and the fragments they refer to. This last feature is useful in the comparison of interesting molecules since it is possible to directly visualize and highlight which fragments they share (as we do in the Fingerprint View described in Section 4.3).

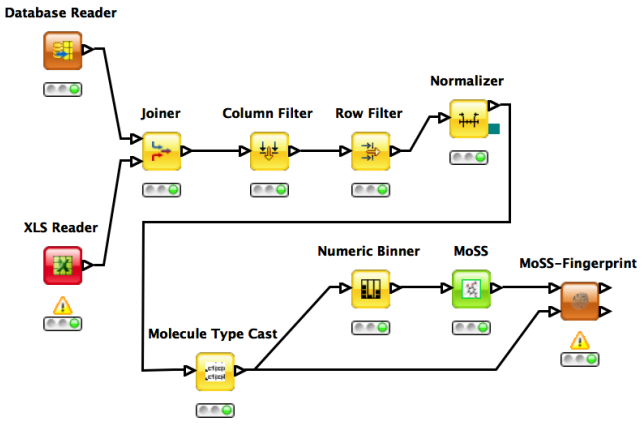


Figure 4: The KNIME workflow for fingerprint generation

3.2 Tasks

The requirements we have gathered to design HiTSEE are the result of a long-term collaboration between the department of Information Science and the Konstanz Research School in Chemical Biology

[1] www.knime.org

at the University of Konstanz. We organized regular meetings between the involved groups to get acquainted with the biochemical problems and to gather information about current practices and data analysis needs. HiTSEE is the last in a number of developed prototypes designed over one year and a half of collaboration. We used the prototypes as a way to probe the design space, to better understand the domain problems, and eventually to isolate the tasks that needed a real support in terms of visual analytics tools.

While we originally developed prototypes for a diverse number of tasks all over the range of the HTS steps described in Section 2, e.g., data processing, quality control, and chemical libraries overviews, HiTSEE has been designed specifically to support hit selection, exploration and expansion. More precisely we provide support for two main visual analytics tasks:

1. *Setting a threshold in hit selection.* One of the challenges we have encountered early on in the process is the definition of an activity threshold value in the hit selection process. From our observations and discussions with the domain experts we realized that the hits are normally selected through a fuzzy process. The researcher sorts the molecules according to their activity value and chooses a threshold going by eye, searching for a trade-off between the number of hits (to be kept low for later, more in depth, testing) and the risk of missing important molecules. One need voiced by our collaborators was the possibility to gain, already at this stage, a better view on the selected hits in order to make the hit selection process more informed.

2. *Exploring the neighborhood of confirmed hits.* A second major need we spotted during our collaboration consists in the exploration of the neighborhood of one or more confirmed hits in the hit expansion phase. This stage starts from one or more molecules resulted to be active in a secondary screening. At this point, the researcher wants to explore the neighborhood to: (1) understand how little structural changes influence the chemical behavior with the selected target; (2) find a trade-off between the activity level expressed by the compounds and other chemical features of interest. In our specific case, for instance, the solubility of the compounds (measured in LogP values) is a critical element to isolate molecules of interest.

HiTSEE supports these two tasks in an integrated environment in which the user can project elements of interest in a scatter plot view, expand the projected items to include their neighbors, and perform several interactive operations that support flexible navigation and details on demand. In the following we describe HiTSEE in detail and explain how it supports the aforementioned tasks.

4 HiTSEE

HiTSEE's interface is organized around four main views: list+projection view (Fig. 5 (left)), molecules detail (Fig. 5 (right)) and substructure search view (Fig. 8), fingerprint view (Fig. 6), that support exploration, in-detail investigation and structural queries. The *list+projection view* permits to select molecules of interest and to project them in a scatter plot visualization to form clusters of (structurally) similar compounds. The view supports the investigation of relationships between activity levels, structural features, and other chemical properties. The *molecules detail and substructure search view* shows the molecular structure of compounds selected in the projection view and permits to trigger substructure searches. The *fingerprint view* shows the distribution of the fingerprint's fragments in the chemical library and supports the selection of meaningful thresholds in the hit selection phase.

In the following we describe each component in detail together with the interaction capabilities offered by each one.

4.1 List+Projection View

The List+Projection view consists of two interactive elements: a compounds list and a linked scatter plot view. (Fig. 5 (left)) The compound list organizes the full set of compounds in the library in a list format sorted by activity level. Each item is represented by its molecular structure and by a bar with length proportional to its activity level.

The user can select one or more items in the list, project them in the scatter plot view, and expand the selection to a user-defined number of neighbors. The neighbors are the compounds that are structurally most similar to the current selection. The structural similarity is calculated from the fingerprint bit-vectors (see Section 3.1.1).

The compounds are represented by circles and positioned in the view through a multidimensional scaling (MDS) projection in a way that compounds with similar structures occupy similar positions. Size represents the activity level and color is used to distinguish between compounds included in the initial selection and those added by the expansion mechanism. Each circle contains also a small pie chart representing additional chemical properties of interest (in our case the LogP value). The pie chart is designed in a way to turn its fill color into a more prominent one (darker blue) when the value of interest goes beyond a predefined threshold.

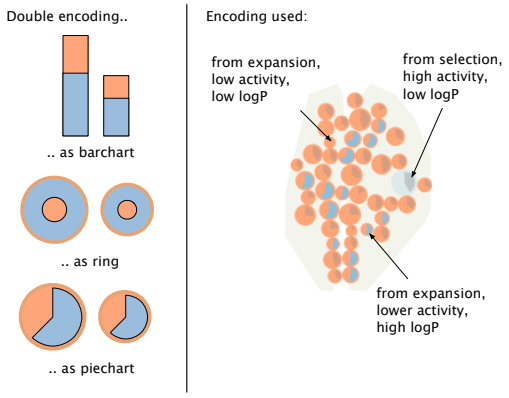
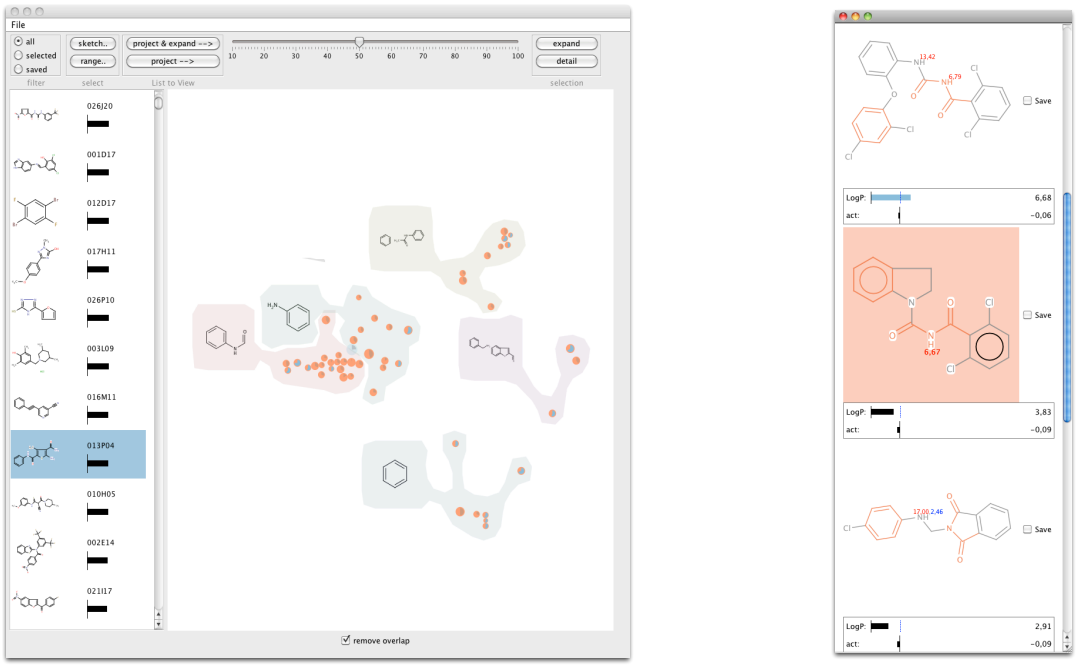


Figure 7: *Left:* Different alternatives for visually encoding activity (length/radius) and logP (proportion of shape). *Right:* HiTSEE's mapping of origin (direct selection or expansion), activity level, and logP value to the visual features: color (orange, blue), size (circle radius), and angle (pie chart).

The MDS projection takes as input a distance matrix of metric distance values. For each pair of compounds, we calculate the Tanimoto distance between their fingerprint bit-vectors. Two problems emerge from MDS-based projections: overlapping items and fuzzy boundaries between the groupings. To cope with these two issues we implemented two additional features. First, we used an overlap removal mechanism based on a 2D rigid body physics simulation (see Section 4.4) that permits to displace the items from their original position in areas where they overlap. Second, in order to facilitate the grouping of the items, we cluster the items and draw a "bubble" around them to reinforce the perception of grouping. The k-means clustering algorithm takes as an input the screen-space positions of the items and clusters them into bubble sets [5]. For each cluster, we determine the common substructure of all containing compounds and position it left to the cluster.

We designed the projection view trying to optimize its visual effectiveness towards reading patterns with biological interest. In the following we provide a summary of the rationale behind our main design choices.

Since position is the strongest visual variable, we use it to convey



List+Projection View

Detail View

Figure 5: HiTSEE interface with *list+projection view* (left) and *detail view* (right).

molecular similarity (through the proximity data given by MDS), which is the most important piece of information in the data. Activity level is mapped to circle size (with a square root mapping to take into account the area effect) to allow for an easy discrimination among the molecules. While visual variables like bar length allow for a more accurate comparison of values [4], we decided to use circles and their size because: (1) they cluster more naturally then shapes with other aspect ratios, (2) they are more robust to the overlapping removal mechanism, (3) they allow for easy discrimination between high vs. low activity molecules while keeping the visualization compact, (4) reading the activity values accurately is not the main purpose of the visualization (as long as major differences can be spotted). We encode a third parameter (LogP) to angle by using a pie chart embedded in the circles. While a number of alternatives exist, as for instance stacked bars and nested circles (see Figure 7), we decided to use a pie chart because it matches well with the circular shape we adopted and scales better than nested rings to items of different size.

4.2 Molecules Detail View and Substructure Search View

From the projection view the user can select a group of interesting compounds to investigate them in detail. Figure 5 (right) shows the detail view with its core features.

The selected set of compounds is visualized as an ordered list of high resolution molecule renderings. The common substructure of all molecules in the selected set is highlighted in each molecule (red). We map the chemical features *activity* and *logP* into small bar charts at the bottom, the *pKa* values are rendered directly into the molecule.

During the investigation of the molecules two operations are available. First, we permit to mark molecules as "saved" for later reinvestigation. Second, we permit the user to start a search on a particular pattern by selecting a molecular fragment and issuing a query for retrieving all the compounds containing the selected fragment. We support this function by providing the substructure search

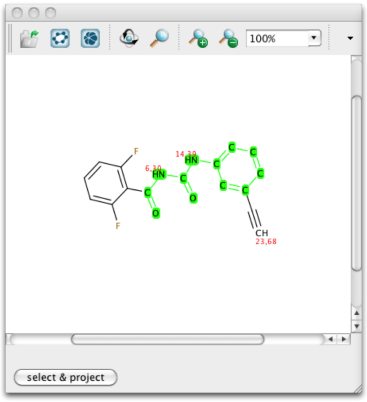


Figure 8: The HiTSEE *HiTSEE substructure search view*

view, which opens up when the user clicks on a molecule in the detail view.

The substructure search view (Figure 8) is based on the JChem Marvin Sketch applet that provides a common interactive method for selecting substructures. The user starts a search on the selected substructure, the search results are highlighted as selections in the List+Projection view, and the user can project them in the projection view for investigation (see Section 4.1).

4.3 Fingerprint View

Another method to make intelligent selections in the List+Projection view is provided by the fingerprint view (Figure 6). This supports the user in making a range selection of activity levels. We show a subset of all (sorted) activity levels on top of the view and a matrix of all molecular fragments that form the fingerprint in the lower part. The user can select a certain range of activity

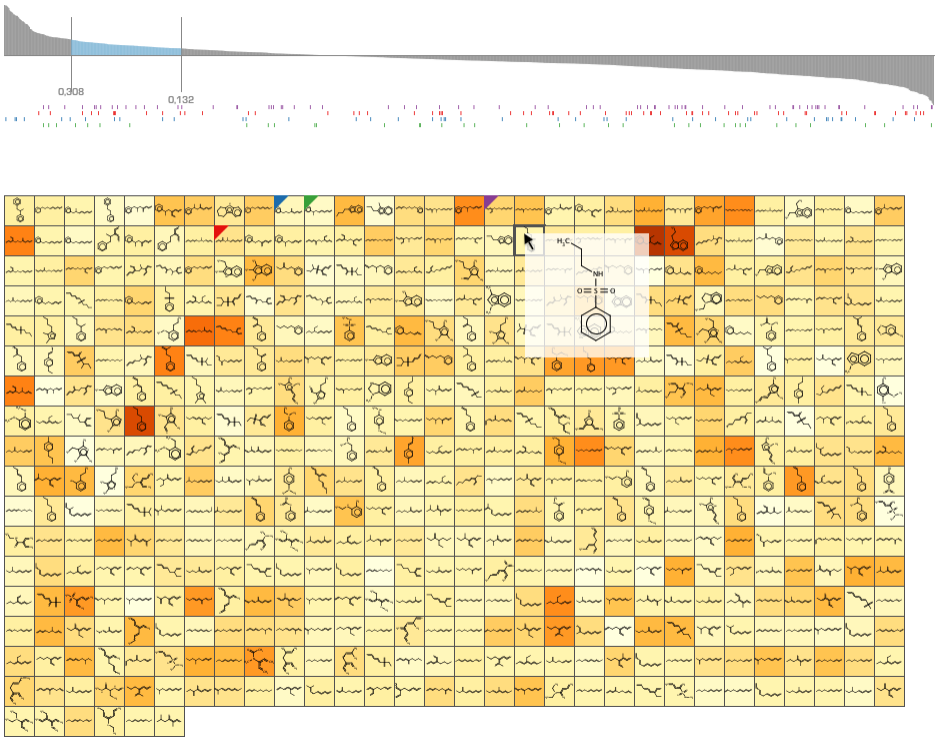


Figure 6: HiTSEE *fingerprint view*

levels and modify the upper and lower thresholds interactively. The coloring of the fingerprint cells shows how frequent a fragment is in the current selection (red - very frequent, yellow - less frequent, gray - not occurring). Conversely, fragments can be selected to highlight compounds which contain this fragment (colored ticks underneath the activity levels). This provides a view on the fragment distribution in larger scale and supports the selection of a range of compounds in which the chosen fragments are (not) contained.

4.4 Implementation Details

The prototype is programmed in Java. For rendering molecules, finding common substructures, and making the interactive selection we use the JChem library version 5.4, 2011, ChemAxon (`http://www.chemaxon.com`). The MDS projection is done by the Java Library for Multidimensional Scaling v0.2 (`http://www.inf.uni-konstanz.de/algo/software/mdsj/`), and the overlap removal is based on JBox2D v2.01 (`http://www.jbox2d.org/`). The cluster shapes are retrieved with BubbleSets (`https://github.com/JosuaKrause/Bubble-Sets`)

5 Case Studies

We used HTS-data generated by a screen looking for a specific inhibitor of Kif18A [3] to proof the effectiveness and usability of HiTSEE.

Kif18A belongs to the family of mitotic kinesins. Kinesins are ATP dependent motor proteins, which utilize the energy derived from ATP hydrolysis to produce mechanical force. Kif18A belongs to the kinesin-8 family whose members are known to be required for the correct segregation of chromosomes in mitosis. Besides its key function in mitosis, Kif18A is characterized by its unique enzymatic properties since it integrates both motility and microtubule depolymerization activity. Due to its central function in mitosis and intriguing enzymatic properties we performed a small molecule screen to identify small molecules that inhibit the ATPase activity of Kif18A. The published results of the small molecule screen can be applied as a proof-of-concept principle to validate HiTSEE.

5.1 Hit Selection

The first step after a high throughput screen is to decide which positive results are counted as a hit and therefore the compounds are sorted according to their activity level. With HiTSEE the compounds are directly sorted according to their activity level and the corresponding structures are represented in a list. After choosing the 30 compounds with the highest activity and projecting them, we have a first set for hit selection. In the projection we could see the common structure of the clusters (Fig. 9a). The only common motif in this case was the phenyl moiety, which is not really significant. Also blue filling of the dots indicates a LogP value above 5, which could cause solubility problems in aqueous media. Nevertheless, by removing the overlap we are able to see the structure of the active compounds and get first hints for structures relevant for activity. In the detailed view of all these selected compounds we can easily compare their structures by eye and spot new common or interesting structures like the diphenyl sulfide moiety. With the project and expand option we were able to see structures related to the hits with lower activity levels (indicated by orange dots, Fig. 9b). Clusters of highly active and less active compounds made us feel more confident to select the highly active compound as a hit, because the structural motif is spread over an activity range.

5.2 Hit Expansion

After one hit has been selected we can choose structural elements and search the entire library for that motif. In our case we started from projecting in the view the 30 compounds with the highest activity. After selecting all the hits and going through the detailed view we decided to look for compounds containing a diphenyl sul-

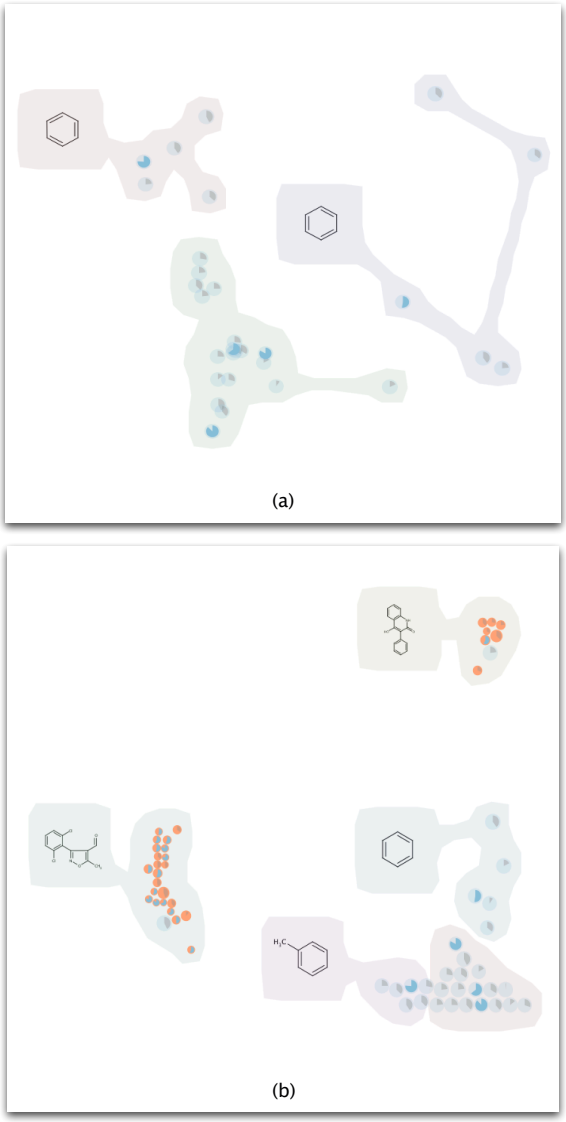


Figure 9: The *projection view* for case study *Hit Selection*

fide moiety which was present in several hits (Fig. 10). The result was that we received nice clusters of compounds. This clustering gives us more confidence to choosing the selected hit for further testing because now we know that there are other compounds containing the same structural moiety with different activity level. This finding gives us a hint that the selected hit is not a false positive because the substructure is present in different compounds. Using the detailed view of a cluster, a list of compounds with common substructures and diverse activity levels is obtained. The fact that the common motif is not only present in the highly active compounds but also in less active ones enabled us to establish first SARs and to feel confident in choosing this compound for further investigations. The search did not give BTB-1 as a result but it gave a whole set of BTB-1 analogues which are also very active. Going on with these results we would choose the diphenyl sulfone moiety as the lead structure. After investigating compounds with this structural motif, establishing SARs we would finally end at BTB-1. HiTSEE allows to confirm hits by exploring the chemical space around them and revealing less active compounds bearing same structural moieties.

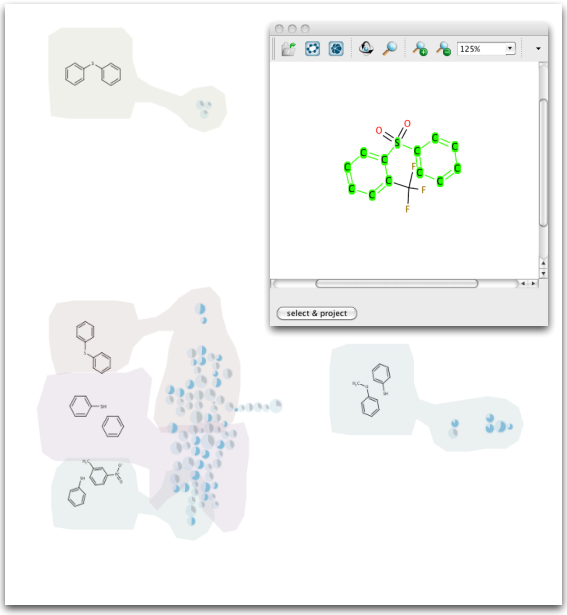


Figure 10: The *projection view* and *substructure search view* for case study *Hit Expansion*

6 Lessons Learned

In this section we highlight two main lessons we have learned during the design of HiTSEE. The first lesson concerns the collaboration process between computer science experts and biochemistry experts, the second concerns the design strategy of HiTSEE.

During our collaboration we noticed that not only it is necessary to make sure that all the parties get acquainted with their respective research background and goals (e.g., the computer scientists have to understand the domain problem, the biochemists have to picture the capabilities that computer systems can offer) but also that their collaboration and influence lay on a steady exchange of ideas. In our experience it is not sufficient to meet at scheduled times in more or less formal meetings to report on the state of the work. A tighter integration is needed. Our project experienced big leaps every time we had computer scientists and biochemists sitting together one next to another and walking through the data analysis steps together for extended periods of time.

From a visualization design point of view we learned that trying to achieve an overview of the whole data at all cost is not always the best strategy. Before developing HiTSEE we provided the users with a number of prototypes based on the idea of visualizing the entire chemical library under observation, or large portions of it. All our attempts in this direction failed because we misinterpreted the needs of our collaborators. The bottleneck in their analysis was not in spotting elements of interest as a way to kick-start the process, but rather to effectively and efficiently explore a fairly large number of compounds similar to few selected ones.

We believe that visualization researchers and designers should take this advice in serious consideration and always ask if creating an overview is the best strategy to cope with the current problem. Especially when dealing with large data sets, trying to obtain full overviews might end up being not only impractical, but also not useful (or sub-optimal).

7 Related Work

While there are a number of free and commercial tools that support one or more phases of HTS (e.g., Spotfire), in the following we focus our review on visualization tools that specifically address

hit selection, exploration and expansion and more specifically the understanding of structure-activity relationships.

SARANEA [9] is a visualization tool to support structure-activity relationship and selectivity analysis. It is based on a network graph visualization. The graph is built by connecting molecules with an edge if their similarity value is higher than a predefined threshold and projected using the classical Fruchterman-Reingold algorithm. The main feature of the tool is the calculation and visualization (through color) of a "cliff index", which describes whether the compound has a big shift in potency compared to its neighborhood. HiTSEE can also help in the detection of activity cliffs by spotting big changes in size within a given cluster. While SARANEA is more targeted towards the exploration of a full set of compounds, HiTSEE leverages on the idea of having a small set of initial compounds (often a single one) and to explore their direct neighborhood.

Scaffold Hunter [11] is another visualization tool that can be used to find relations between structural features and activity level. A tree structure is built starting from one compound and building molecular scaffolds by removing, through a series of chemistry and medicinal chemistry rules, rings in the periphery. The visual representation is a radial tree depicting the hierarchical scaffold structure and the activity level of the scaffolds. The tool can be used to investigate the potential of the scaffold to be at the origin of activation of biological processes. Scaffold Hunter shares with HiTSEE the idea of starting the analysis from one (or more in HiTSEE) compounds of interest and explore their neighborhood. HiTSEE however is focused more on structural similarities between the compounds rather than the scaffolds.

DrugViz [12] is a newly developed plug-in in the Cytoscape environment in which the analysis is centered on a network representation of interactions between biological targets. The system permits to select similar targets and find common compounds or, in alternative, select similar compounds and see how they distribute in the target network. While the system permits to investigate relationships between biological targets and molecular structures the visualization is not targeted towards the visualization of structure-activity relationships.

The SAR Map [1] (and its extension Enhanced SAR Map) permits to focus on one single molecule of interest and explore all its substituents through R-group analysis. R-group analysis takes as an input a list of compounds with a common scaffold and generates all the possible variations. The SAR map is a heat map where the columns and the rows represent substituents of two selected variation sites. Each cell represents one specific compound (formed by attaching the substituents) and a color map, or more complex visualizations, provide rich information about each compound. HiTSEE also permits to visualize the variations around a subset of structurally similar compounds (in the molecule details view) but this functionality takes place in the larger context of similarity analysis among entire molecules.

ChemGPS-NP [7], similarly to HiTSEE, projects molecules in a low-dimensional space using a PCA projection. The visualization is designed in a way to reflect properties that are relevant for bioactivity. The visualization however does not address directly the correlation between activity and structural similarity.

In summary, HiTSEE has the unique advantage of allowing a flexible and smooth navigation in the chemical space by conjugating two contrasting needs: the need to create visual summaries of chemical libraries and the need to explore the neighborhood of selected compounds.

8 Conclusion and Future Work

We presented HiTSEE a visualization tool for the analysis of high-throughput screening data for biochemistry experiments. HiTSEE proposes a smooth interface and interaction paradigm that permits to explore the chemical space and find relationships between activity values and molecular structures. The paper presented a series of requirements, their impact on the design of the tool, and its effectiveness through a case study.

There are a number of issues we plan to address in the future. The projection view changes abruptly when it is modified by some external events, making it difficult to preserve the mental map of the projected items. We plan to develop methods to reduce the changes from one view to another and to implement smooth animations that help relating the new projection to the old one. As the analysis gets more complex and the user goes through several steps, it becomes difficult to remember previous steps and return to interesting states previously visited. We plan to implement a history and save mechanism that support this specific need. We also plan to perform additional tests on our fingerprints and run a thorough comparison with other solutions.

Further development of HiTSEE and a freely available version for KNIME will be available at: `http://hitsee.hs8.de`

Acknowledgements

The authors wish to thank Michael Höhn and Roland Jungnickel for their great help. This work was partially supported by the DFG Research Training Group GK-1042 "Explorative Analysis and Visualization of Large Information Spaces", the Konstanz Research School Chemical Biology (KoRS-CB), and the "Interdisciplinary Center for interactive Data Analysis, Modeling and Visual Exploration" (INCIDE).

References

[1] D. K. Agrafiotis, M. Shemanarev, P. J. Connolly, M. Farnum, and V. S. Lobanov. Sar Maps: A new SAR visualization technique for medicinal chemists. *Journal of Medicinal Chemistry*, 50(24):5926–5937, 2007.

[2] C. Borgelt and M. R. Berthold. Mining molecular fragments: Finding relevant substructures of molecules. In *Proc. of IEEE International Conference on Data Mining*, ICDM '02, 2002.

[3] M. Catarinella, T. Grüner, T. Strittmatter, A. Marx, and T. Mayer. Btb-1: A small molecule inhibitor of the mitotic motor protein kif18a. *Angewandte Chemie International Edition*, 48(48):9072–9076, 2009.

[4] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.

[5] C. Collins, G. Penn, and S. Carpendale. Bubble sets: Revealing set relations with isocontours over existing visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 15:1009–1016, November 2009.

[6] R. Hertzberg and A. Pope. High-throughput screening: new technology for the 21st century. *Current Opinion in Chemical Biology*, 4(4):445451, 2000.

[7] J. Larsson, J. Gottfries, S. Muresan, and A. Backlund. ChemGPS-NP: Tuned for navigation in biologically relevant chemical space. *Journal of Natural Products*, 70(5):789–794, 2007.

[8] A. Leach and V. Gillet. *An Introduction to Chemoinformatics*. Springer, 2007.

[9] E. Lounkine, M. Wawer, A. M. Wassermann, and J. Bajorath. Saranea: A freely available program to mine structureactivity and structureselectivity relationship information in compound data sets. *Journal of Chemical Information and Modeling*, 50(1):68–78, 2010.

[10] T. Mayer. Chemical genetics: tailoring tools for cell biology. *Trends in Cell Biology*, 13(5):270–277, 2003.

[11] S. Wetzel, K. Klein, S. Renner, D. Rauh, T. I. Oprea, P. Mutzel, and H. Waldmann. Interactive exploration of chemical space with scaffold hunter. *Nat Chem Biol*, 5(8):581–583, 2009.

[12] B. Xiong, K. Liu, J. Wu, D. L. Burk, H. Jiang, and J. Shen. DrugViz: a Cytoscape plugin for visualizing and analyzing small molecule drugs in biological networks. *Bioinformatics*, 24(18):2117–2118, 2008.