JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

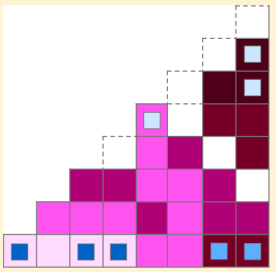
pubs.acs.org/jcim

Navigating High-Dimensional Activity Landscapes: Design and Application of the Ligand-Target Differentiation Map

Preeti Iyer,[†] Dilyana Dimova,[†] Martin Vogt, and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany

ABSTRACT: The transformation of high-dimensional bioactivity spaces into activity landscape representations is as of yet an unsolved problem in computational medicinal chemistry. High-dimensional activity spaces result from the experimental evaluation of compound sets on large numbers of targets. We introduce a first concept to represent and navigate high-dimensional activity landscapes that is based on a data structure termed ligand-target differentiation (LTD) map. This approach is designed to reduce the complexity of high-dimensional bioactivity spaces and enable the identification and further analysis of compound subsets with interesting activity and structural relationships. Its utility has been demonstrated using a set of more than 1400 inhibitors with exact activity measurements for varying numbers of 172 kinases.



1. INTRODUCTION

The experimental evaluation of compounds on arrays of biological targets, often referred to as compound profiling, has become an important source of activity data for pharmaceutical research and chemical biology.[1] Compound profiling is often carried out for major therapeutic target families such as G protein coupled receptors[2] or protein kinases.[3] In profiling campaigns, structurally diverse or, alternatively, focused compound collections are screened against varying numbers of targets. This is often (but not always) done for targets providing a representative subset of a given family. The resulting profiling data constitute high-dimensional bioactivity spaces, which are generally difficult to represent and navigate.[4] However, in such activity spaces, ligand-binding profiles of targets, compound activity patterns, and ligand-target relationships can be explored. Furthermore, it might be attempted to identify chemical probes that differentiate between related targets[4] or prioritize compounds for further chemical exploration and discovery efforts.[3]

On the basis of the activity landscape concept,[5,6] representations of activity spaces are often generated by integrating structure and activity relationships between sets of compounds.[6] Activity landscapes provide an intuitive access to structure−activity relationship (SAR) information but are usually focused on a specific biological activity, i.e., a single target. However, the activity landscape concept has also been extended to pairs of targets[7] or more than two targets[8] in order to explore the target selectivity of active compounds or the formation of multi-target activity cliffs.[9] Only recently, multi-target activity landscape representations have been introduced, including annotated molecular networks, the original multi-target landscape design,[8] structural similarity and activity similarity difference maps that utilize plots of compound activity versus structural similarity,[10] and a landscape layout based on self-organizing maps to group structurally similar compounds together and encode their activity relationships.[11] However, in these representations, activities against only a few targets (e.g., three or four) can be captured in a meaningful and interpretable way. The representation of high-dimensional activity spaces (e.g., involving 50, 100, or more targets) in an activity landscape format has thus far not been reported.

The design of high-dimensional activity landscapes has been hampered by the limited availability of compound profiling data in the public domain. Although a number of pharmaceutical companies have already generated large bodies of profiling data for popular therapeutic targets, most of this data is kept proprietary, for understandable reasons. A notable exception has been a recent study by a group from Abbott Laboratories.[12] For the generation of kinase interaction networks and the exploration of polypharmacology patterns, a total of 3858 compounds were tested against varying numbers of 172 kinases representing a diverse sample of the kinome.[12] As a part of this investigation, structures and activity data for a subset of 1496 of these compounds were made publicly available, hence providing a significant source of profiling data for further studies. Using this data set, we have developed and applied a first concept for the design and analysis of high-dimensional activity landscapes that is reported herein.

2. ACTIVITY DATA

From the publicly released Abbott data set, all compounds with unique 2D molecular representations[13] were extracted for which a K_i value for at least one kinase was available, leading to the selection of 1473 compounds. These compounds were annotated with pK_i values for one to 122 kinases. The activity annotations included all 172 kinases investigated in the Abbott study. A maximum of 101 kinases were shared between individual compounds. In our analysis, only absolute equilibrium constants were considered as activity measurements and threshold measurements were ignored. For activity landscape

Received: May 9, 2012
Published: July 14, 2012

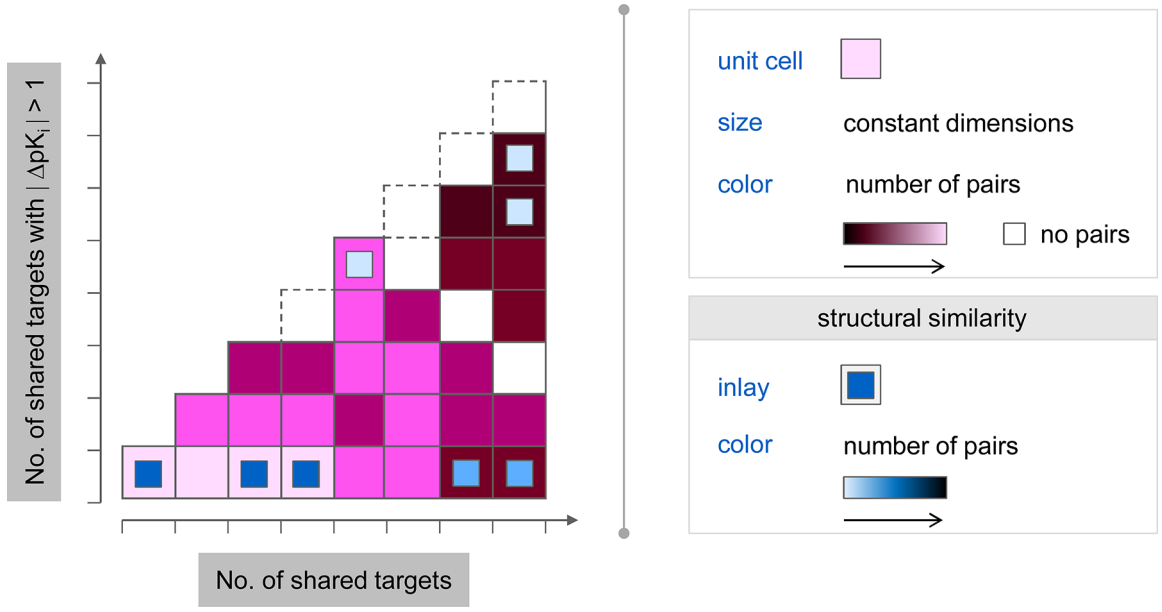


Figure 1. Design principles of the ligand-target differentiation map. The schematic illustration summarizes the basic design elements of the ligand-target differentiation map. Unit cells have constant dimensions and delineate a well-defined range of shared targets (*x*-axis) and of shared targets with qualifying potency differences (*y*-axis). Compound pairs are assigned to a unit cell if the underlying target relationship defined by the cell is met. Color coding accounts for the number of qualifying compound pairs from light pink (many pairs) over magenta to black (a single pair). Inlay squares indicate structural relationships and are "inversely" color-coded according to the frequency of structurally related pairs in a unit cell from black (many pairs) over dark blue to light blue (a single pair).

design, this data set presented a challenging test case because the underlying profiling matrix was high-dimensional yet incomplete, i.e., compounds were assayed against varying numbers of targets and activity profiles only partly overlapped in many instances. The activity profile of a compound consists of all of its target annotations.

3. LIGAND-TARGET DIFFERENTIATION MAP

On the basis of our evaluation, annotated molecular network representations, which were previously utilized for the generation of single- and multi-target activity landscapes, were not suitable for capturing and representing high-dimensional activity spaces. The design of high-dimensional activity landscapes presents challenges that go beyond the analysis of single-target SARs[6] or multi-target SAR discontinuity patterns.[6,8] In particular, ligand-target relationships need to be systematically explored and compared in light of structural features of active compounds. Therefore, it was required to investigate new representation concepts. In the following, the basic design principles of the Ligand-Target Differentiation (LTD) map, our central data structure for high-dimensional activity landscape analysis, and its elements are discussed. In addition, the extraction of compound and activity information from the map is illustrated.

3.1. Design Concept. As an activity landscape representation, the LTD map must systematically account for compound potency and similarity relationships in high-dimensional data sets. A major goal of such representations is to provide complete coverage of experimental data. For the analysis of compound profiling data, it must be considered that high-dimensional matrices are often incomplete. Hence, the graphical data structure must be flexible and capable of capturing profiling matrices of different composition.

Figure 1 shows a schematic representation of an LTD map to illustrate its design principles. The LTD map is based on four general principles:

(1) The basic unit of the data structure is a compound pair.
(2) Compounds are differentiated according to the number of targets they share.
(3) Compounds are further differentiated according to their activity differences against these targets.
(4) Structural relationships between all molecules are monitored.

Accordingly, all pairwise target, activity profile, and structural relationships between active compounds are initially determined. The LTD map then relates the number of targets shared by any pair of compounds to the number of targets against which these two compounds display significant differences in potency using a "unit cell" as its basic data element. In our analysis, the threshold for the potency difference between compounds in a pair against a common target is set to 1 order of magnitude (a flexible criterion, depending on data set characteristics). The relative frequencies of detected target and structural relationships are then captured through color coding and map annotation. Figure 1 summarizes the basic elements of the LTD map, and Figure 2 shows the LTD map representation of the entire kinase inhibitor data set, as further discussed in the following.

3.2. Elements of Graphical Representation. In the map, a unit cell is represented as a square. The constant dimensions of a unit cell in Figure 2 are five by five. The numbers of shared targets and shared targets with more than an order of magnitude potency difference are reported along the *x*-axis and the *y*-axis, respectively. The LTD map thus consists of an array of squares that account for all pairwise target and potency difference relationships in a data set and span the entire ranges. The number of compound pairs falling into each cell is

dx.doi.org/10.1021/ci3002765 | J. Chem. Inf. Model. 2012, 52, 1962−1969

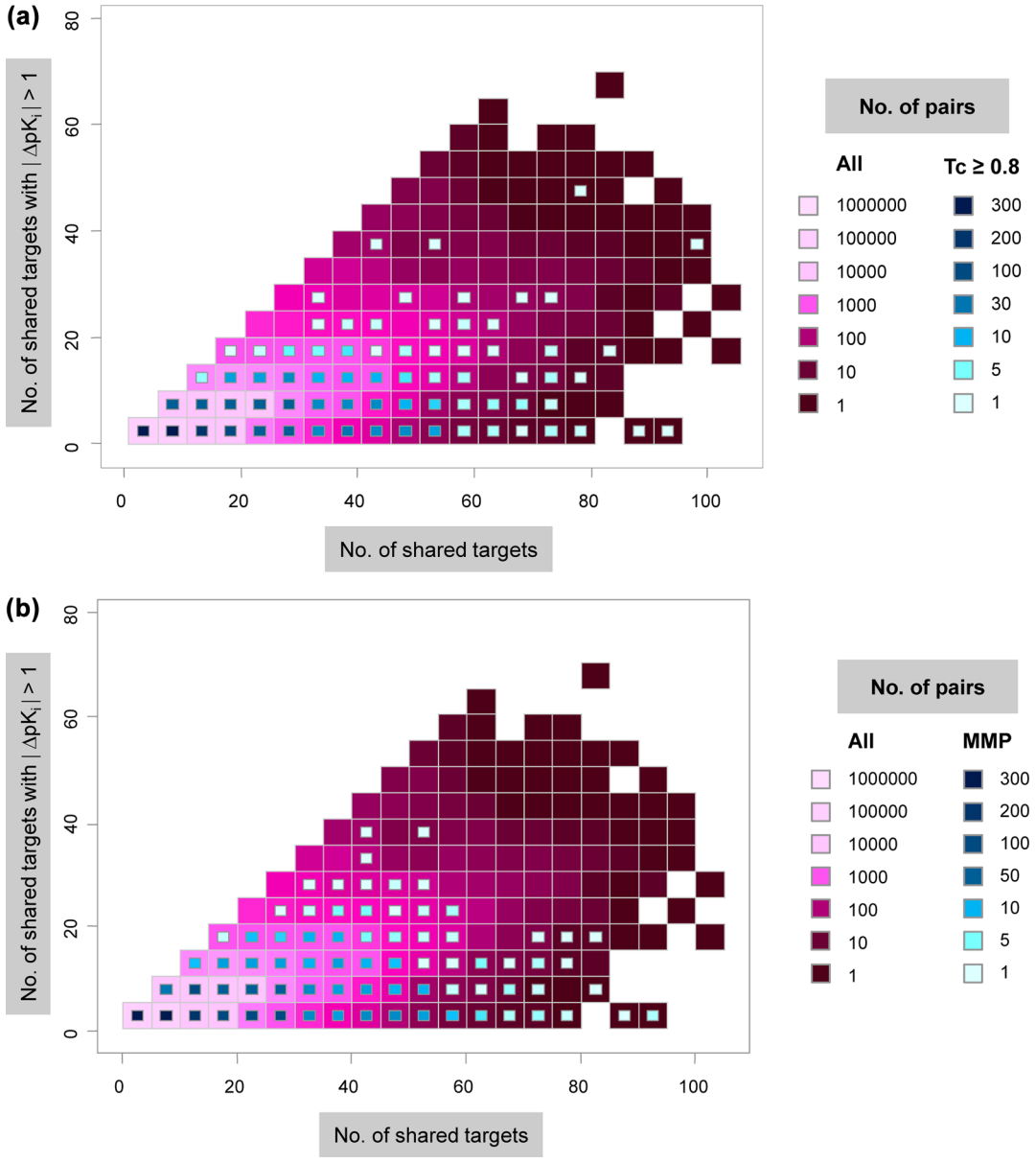


Figure 2. LTD map of kinase inhibitor data. Two versions of the LTD map of the kinase inhibitor data set are shown that only differ in the way structural relationships between active compounds are determined. In panel (a), MACCS Tanimoto similarity relationships are shown, and in panel (b) MMP-based substructure equivalences are shown (i.e., cells with inlays contain pairs of compounds with a common core structure).

determined, and a continuous color code (from black over magenta to light pink) is used to monitor the frequency of relationships within the cells (i.e., a cell colored in black contains a single pair, and a cell in light(est) pink contains the maximally observed number of pairs). An "empty" cell within the map indicates that no compound pairs are falling into the respective data intervals.

Structural relationship information is also incorporated into the LTD map. For our analysis, compound similarity was assessed in two complementary ways. Pairwise whole molecule Tanimoto similarity[14] was calculated using MACCS structural keys.[15] As a similarity threshold for selected compound relationships, a Tanimoto coefficient of 0.8 was applied. Furthermore, matched molecular pair (MMP) analysis[16,17] was carried out to identify substructure relationships between compounds. For this purpose, all compound pairs were identified that formed an MMP, i.e., that shared a given key fragment (core structure), as described previously.[17] It should be noted that increasing the potency and similarity thresholds decreases the number of qualifying compound pairs and hence the information content of the analysis. Data noise and information content must be balanced.

All unit cells that contain compound pairs with structural relationships are then marked through the addition of "square inlays", as schematically illustrated in Figure 1. Furthermore, Figure 2a captures structural relationships between the kinase inhibitors on the basis of pairwise Tanimoto similarity calculations and Figure 2b on the basis of common (MMP-based) core structures. An inversely shaded color code (from light blue over dark blue to black) is applied to the inlays in order to account for the frequency of pairwise structural relationships per cell (i.e., a cell with a light(est) blue square inlay contains a single relationship). Hence, annotation of cells

dx.doi.org/10.1021/ci3002765 | J. Chem. Inf. Model. 2012, 52, 1962−1969

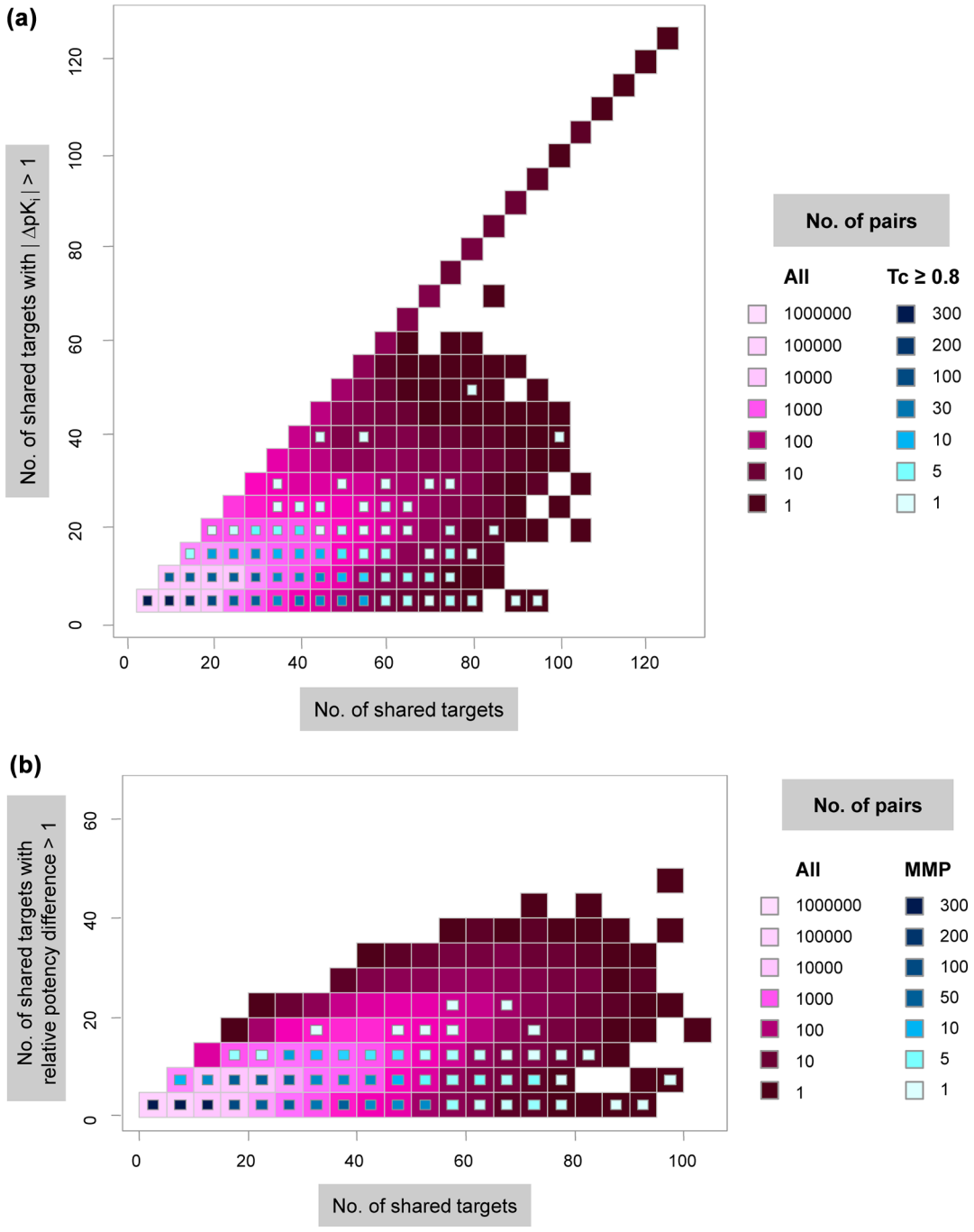


Figure 3. Data set and map modifications. In panel (a), the LTD map of the kinase inhibitor data set was calculated after addition of a hypothetical compound with 1 mM potency against all 172 kinases. In panel (b), an alternative version of the LTD map is shown for the original data set where potency deviations from mean compound potency were used instead of absolute potency differences, as rationalized in the text.

monitors the distribution of structural relationships in high-dimensional activity space.

A consistent numbering scheme is applied to cells in LTD maps, adhering to "top down" (vertical) followed by "from left to right" (horizontal) reading directions. Thus, if multiple cells have the same number of targets with significant potency differences, they are numbered in the order of increasing numbers of shared targets. LTD maps were drawn using routines implemented in the R environment.[18]

3.3. Interpretation. The LTD map provides an immediate view of the data distribution in high-dimensional activity space, as illustrated in Figures 1 and 2. For the kinase inhibitor data set containing activities against a total of 172 different kinases, individual compound pairs share up to 101 kinases and up to 69 kinases with potency differences of more than 1 order of magnitude. As clearly delineated by the cell color code, the bulk of the activity data falls into the map section spanned by zero to ~20 shared targets and zero to ~10 targets with significant potency differences. In addition, the section of highly populated cells extends to ~60 shared targets and ~30 targets with potency differences. For further increasing numbers of shared targets and targets with qualifying potency differences, the number of compound pairs rapidly declines. The inlay view of structural relationships between compound pairs adds further

dx.doi.org/10.1021/ci3002765 | J. Chem. Inf. Model. 2012, 52, 1962−1969

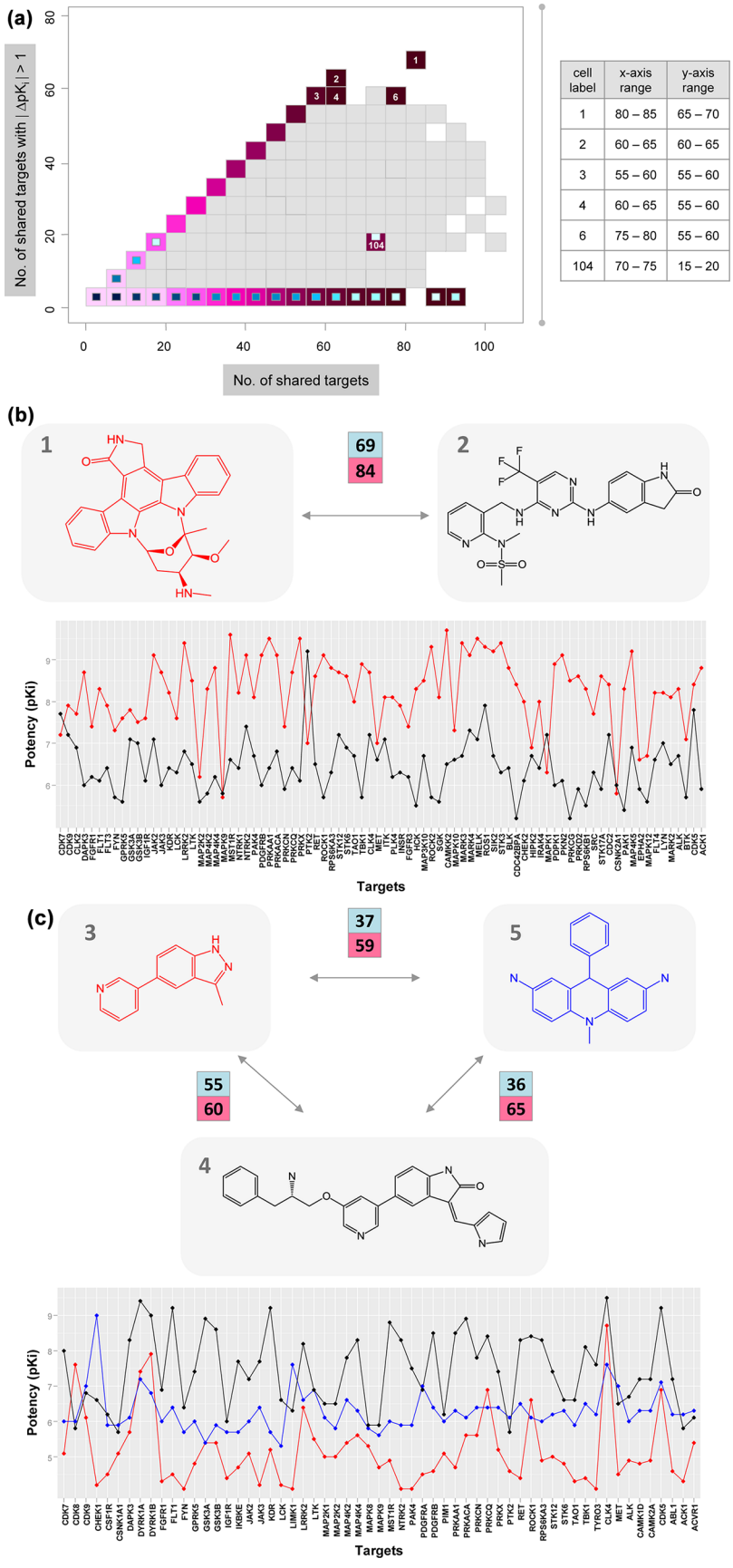


Figure 4. continued

dx.doi.org/10.1021/ci3002765 | J. Chem. Inf. Model. 2012, 52, 1962−1969

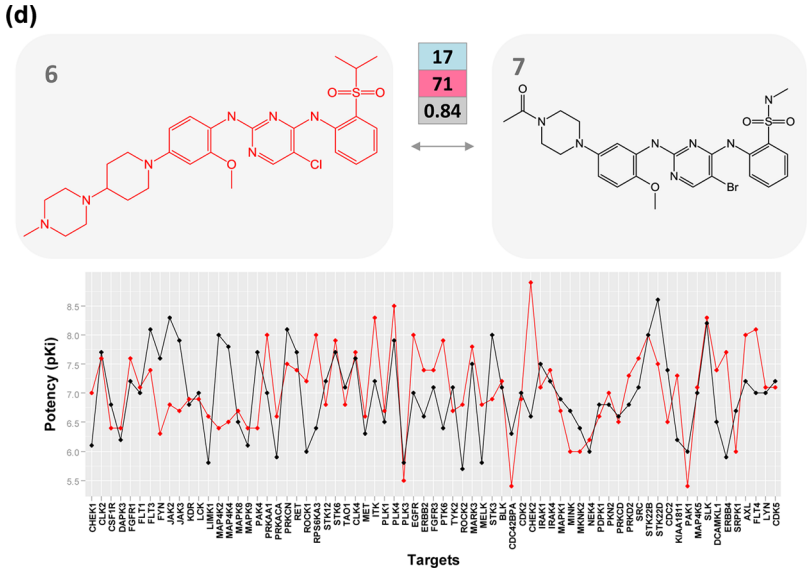


Figure 4. Compound information. The figure illustrates how compound information is extracted from the LTD map. In panel (a), the LTD map in Figure 1a is displayed in simplified form. Map boundaries referred to in the text and exemplary cells are color-coded while other regions are displayed in light gray. In addition, selected cells are numbered (following the numbering scheme described in the text). The table insert reports the *x*-axis and *y*-axis range for each labeled cell. In panels (b)−(d), compounds taken from selected cells are shown together with their activity profiles. Cells from which the compounds originate are specified. For pairwise comparisons, the total number of shared kinases and the subset of kinases with significant potency differences are reported on a pink and light blue background, respectively. In addition, for compound pairs with MACCS Tanimoto similarity above the threshold value, the Tanimoto coefficient is given on a gray background. In the activity profiles, pK_i values for all shared targets are reported. Compounds and corresponding profiles are color-coded. Kinase abbreviations are used according to ref 12.

information to the map. Cells are highlighted that contain compound pairs with structural relationships. In Figure 2a and b, Tanimoto similarity relationships and (MMP-based) substructure equivalences are displayed, respectively. These figures show how structural similarity relationships are distributed in high-dimensional activity space. In the kinase inhibitor data set, the distributions of Tanimoto similarity and substructure relationships are similar, as revealed in Figure 2. Most similarity relationships are detected between compounds that share only a few targets. Frequent similarity relationships still occur between compounds sharing up to ∼50 targets and ∼10 targets with significant potency differences (dark to medium blue inlay area in Figure 2). For increasing numbers of shared targets and targets with potency differences, only few structural relationships between inhibitors are detected. The bottom row of cells in the map contains compound pairs that have similar activity against all kinases against which they have been tested. This region contains many structurally similar compounds (as one might expect). By contrast, compound pairs forming the leftmost (pseudo-diagonal) cell layer display potency differences against all, or nearly all, of their shared targets.

On the basis of the information provided by the LTD map, activity data can be further analyzed by selecting compounds from map segments or individual cells of interest. In the following, representative examples are discussed.

3.4. Modifications of LTD Maps. Compounds forming pairs on the diagonal of the LTD maps include those that display differential activity against kinases. In addition, cells on the diagonal might also include compounds having consistently higher or lower potency against shared targets. These compounds are less interesting for further analysis. As an extreme case, a data set might contain one or more compounds with artificially high or low potency against many targets, which would result in the formation of many artificial compound pairs and complicate the analysis of the LTD map. Such compounds were not present in the kinase inhibitor set. However, if such compounds exist in a data set, they will exclusively form cells on the diagonal, which alters the map appearance in a characteristic manner. This is demonstrated in Figure 3a that shows the LTD map of the kinase inhibitor data set recalculated after addition of a hypothetical compound with 1 mM potency (pK_i value of 3) against all 172 kinases. Because such compounds only occur in cells on the diagonal, they (and the pairs they form) can be easily identified and removed from further analysis.

However, to principally omit this potential complication, the LTD map can be modified by considering potency differences with respect to the average potency of compounds for the set of shared targets instead of absolute potency differences. Thus, for each compound in a pair, the average potency for its shared targets is calculated and subtracted from each individual potency value, which yields relative potency values. These values reflect whether a compound shows above or below average potency values for its shared targets. The differences between these relative potency values are then used for map construction. The modified LTD map calculated for the potency difference threshold of 1 order of magnitude as before is shown in Figure 3b. The diagonal cells are less densely populated, which is due to the fact that the median of relative potency differences is only 0.54 compared to 0.80 for absolute differences. Compounds with artificially high or low potencies against many targets no longer form pairs with large relative potency differences and do not populate diagonal cells in this map. Only compounds with differentiated potency profiles (i.e., compounds with selectivity) can induce signals and form pairs that populate prominent cells. Accordingly, the modified LTD map in Figure 3b remained essentially constant when it was recalculated after addition of the hypothetical compound.

dx.doi.org/10.1021/ci3002765 | J. Chem. Inf. Model. 2012, 52, 1962−1969

3.5. Alternative Potency Difference Threshold Values.

It should also be mentioned that potency difference threshold values can not only be prespecified but can also be determined in a meaningful manner on the basis of the potency distributions within a given data set, as demonstrated in the following for the kinase inhibitor set. In order to evaluate the significance of potency differences for a target observed for pairs of compounds, the distribution of potency differences for all pairs of compounds and all shared targets was analyzed. Thus, for all 786,776 compounds pairs, the potency differences for one to 101 targets shared between them were pooled. Observed potency differences ranged from zero to a maximum of 7.1. The mean value was 1.0, with a standard deviation of 0.8. These values reflected the asymmetric nature of the distribution, with values extending from about one standard deviation below the average to more than seven standard deviations above the average. This was also indicated by the median of 0.8 and the interquartile range with a first quartile of 0.4 and a third quartile of 1.5. On the basis of these values, a potency difference threshold of 1 order of magnitude applied in our analysis was a reasonable choice for this data set because it directed the analysis toward compound pairs with a high number of above average potency differences.

On the basis of these considerations, a statistical analysis of the significance of compound pairs with a certain number of targets with above average potency difference with respect to the total number of commonly annotated targets might be performed. For example, by taking the median potency difference as a threshold value, half the potency differences between a pair of compounds would be expected to have values beyond the threshold. This gives rise to a binomial distribution with $p = 0.5$. In this case, significance at the 0.01 level would correspond, for instance, to compound pairs with 15 large potency differences given a total of 20 shared targets. These values can guide the analysis of the LTD map in order to identify interesting compound pairs especially considering the incompleteness of data set annotations.

4. COMPOUND DATA ANALYSIS

Figure 4a points at regions in the LTD map in Figure 2 that contain interesting compounds for further exploration. Pairs can be automatically extracted from cells. The consistent numbering scheme of cells in LTD maps is also illustrated in Figure 4a. Cell 1 contains a pair of structurally distinct inhibitors that share 84 targets and yield significant potency differences for 69 of them. This represents the largest number of shared targets with qualifying potency differences detected within the data set. The structures of these compounds and their activity profiles are shown in Figure 4b. Compound 1 has mostly higher potency than compound 2, which explains the overall large number of potency differences (a situation observed in a number of instances in this region of the map). However, both compounds display different activity profiles and significant differentiation potential against many kinases, with in part large differences in potency, especially in the case of compound 1. Furthermore, the adjacent cells 2−4 contain compound pairs active against ∼60 kinases with significant potency differences against many of them. Figure 4c shows three exemplary compounds taken from these cells. Compounds 3 and 4 form a pair and have different potencies against 55 of the 60 targets they share. As an additional example, compound 5 is included in the comparison. Compound 3 has overall lower potency than compound 4 but the traces of their activity profiles show notable similarities. Compound 4 has high kinase differentiation potential, often with relative differences in potency between kinases of 3 orders of magnitude or more. By contrast, compound 5 has mostly intermediate potency and shows rather limited ability to differentiate between kinases. Cell 104 in Figure 4a maps to another interesting region in the LTD map. In this region, compounds have similar activity against many kinases and yield only a limited number of significant potency differences. Figure 4d shows a pair of structurally similar compounds taken from cell 104. Their activity profiles are overall also similar but reveal a number of notable potency differences against individual kinases. Thus, the comparison of compounds with many shared targets that include only a limited number of targets with significant potency differences might identify potential selectivity probes. Taken together, these examples illustrate how compound information can be extracted from the LTD map and how compound subsets with desired properties can be selected for further studies.

5. CONCLUDING REMARKS

We have introduced the ligand-target differentiation map that is designed to navigate high-dimensional bioactivity spaces taking structural relationships between active compounds into account. As such, the LTD map represents a high-dimensional activity landscape. A hallmark of the activity landscape concept is the graphical integration of compound structure and activity relationships. One of the difficulties involved in exploring the design of high-dimensional landscapes has been the limited availability of relevant compound profiling data, at least in the public domain. A notable exception is provided by the publicly available kinase inhibitor data set from Abbott that contains activity data for more than 1400 inhibitors and 172 different kinases and probably represents the largest high-dimensional compound profiling set currently available in the public domain. Our exploratory efforts have been much supported by the availability of this data set, leading to the design of the LTD map structure. A key feature of the newly introduced approach is the reduction of the inherent complexity of variable high-dimensional activity spaces. This is accomplished by systematically accounting for pairwise differences between multi-target activity profiles of test compounds. Such differences are graphically represented by monitoring the number of common targets with significant potency differences as a function of the total number of targets that are shared by compound pairs. This representation greatly simplifies the navigation of high-dimensional activity spaces and makes it possible to quickly focus on regions in data sets that are most interesting for further analysis. Another key feature of the LTD map is its basic data element, a constantly sized cell that contains compound pairs with well-defined relationships. Through color coding and cell annotation, both activity and structural relationship information is provided. From individual cells or groups of cells, compound pairs and subsets with well-defined relationships can be selected, as demonstrated herein. The generation of the LTD map is straightforward, and the representation is also applicable to smaller data sets of lower dimensionality. Thus, LTD maps should be helpful for many applications in multi-target compound data analysis. In addition, it is hoped that the approach introduced herein might also catalyze the development of alternative high-dimensional activity landscape views.

AUTHOR INFORMATION

Corresponding Author

*Tel: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

Author Contributions

†The contributions of these two authors should be considered equal.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Rix, U.; Superti-Furga, G. Target profiling of small molecules by chemical proteomics. *Nature Chem. Biol.* **2009**, *5*, 616−624.
(2) Allen, J. A.; Roth, B. L. Strategies to discover unexpected targets for drugs active at G protein-coupled receptors. *Annu. Rev. Pharmacol. Toxicol.* **2011**, *51*, 117−144.
(3) Goldstein, D. M.; Gray, N. S.; Zarrinkar, P. P. High-throughput kinase profiling as a platform for drug discovery. *Nature Rev. Drug. Discov* **2008**, *6*, 391−397.
(4) Bajorath, J. Computational analysis of ligand relationships within target families. *Curr. Opin. Chem. Biol.* **2008**, *12*, 352−358.
(5) Bajorath, J.; Maggiora, G.; Lajiness, M., organizers. The Emerging Concepts of Activity Landscapes and Activity Cliffs and Their Role in Drug Research; 240th National Meeting of the American Chemical Society, Divisions of Chemical Information and Computers in Chemistry, Boston, MA, August 22−26, 2010.
(6) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity landscape representations for structure-activity relationship analysis. *J. Med. Chem.* **2010**, *53*, 8209−8223.
(7) Peltason, L.; Hu, Y.; Bajorath, J. From structure-activity to structure-selectivity relationships: Quantitative assessment, selectivity cliffs, and key compounds. *ChemMedChem* **2009**, *4*, 1864−1873.
(8) Dimova, D.; Wawer, M.; Wassermann, A. M.; Bajorath, J. Design of multi-target activity landscapes that capture hierarchical activity cliff distributions. *J. Chem. Inf. Model.* **2011**, *51*, 256−288.
(9) Stumpfe, D.; Bajorath, J. Exploring activity cliffs in medicinal chemistry. *J. Med. Chem.* **2012**, *55*, 2932−2942.
(10) Medina-Franco, J. L.; Yongye, A. B.; Perez-Villanueva, J.; Houghten, R. A.; Martinez-Mayorga, K. Multi-target structure-activity relationships characterized by activity-difference maps and consensus similarity measures. *J. Chem. Inf. Model.* **2011**, *51*, 2427−2439.
(11) Iyer, P.; Bajorath, J. Representation of multi-target activity landscapes through target pair-based compound encoding in self-organizing maps. *Chem. Biol. Drug Des.* **2011**, *78*, 778−786.
(12) Metz, J. T.; Johnson, E. F.; Soni, N. B.; Merta, P. J.; Kifle, L.; Hajduk, P. J. Navigating the kinome. *Nature Chem. Biol.* **2011**, *7*, 200−202.
(13) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(14) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(15) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, USA.
(16) Kenny, P. W.; Sadowski, J. Structure modification in chemical databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2005; pp 271−285.
(17) Wassermann, A. M.; Bajorath, J. Large-scale exploration of bioisosteric replacements on the basis of matched molecular pairs. *Future Med. Chem.* **2011**, *3*, 425−436.
(18) *R: A Language and Environment for Statistical Computing*; R Development Core Team, R Foundation for Statistical Computing: Vienna, Austria, 2008.