Bioorganic & Medicinal Chemistry 20 (2012) 5416–5427

Contents lists available at SciVerse ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Chemotography for multi-target SAR analysis in the context of biological pathways

Eugen Lounkine *, Peter Kutchukian, Paula Petrone, John W. Davies, Meir Glick

Lead Discovery Informatics, Novartis Institutes for Biomedical Research, 250 Massachusetts Ave., Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Available online 20 February 2012

Keywords:
Chemical space
Color
Visualization
Molecular fingerprints
Chemogenomics
Systems pharmacology
Molecular pathways

ABSTRACT

The increasing amount of chemogenomics data, that is, activity measurements of many compounds across a variety of biological targets, allows for better understanding of pharmacology in a broad biological context. Rather than assessing activity at individual biological targets, today understanding of compound interaction with complex biological systems and molecular pathways is often sought in phenotypic screens. This perspective poses novel challenges to structure–activity relationship (SAR) assessment. Today, the bottleneck of drug discovery lies in the understanding of SAR of rich datasets that go beyond single targets in the context of biological pathways, potential off-targets, and complex selectivity profiles. To aid in the understanding and interpretation of such complex SAR, we introduce Chemotography (chemotype chromatography), which encodes chemical space using a color spectrum by combining clustering and multidimensional scaling. Rich biological data in our approach were visualized using spatial dimensions traditionally reserved for chemical space. This allowed us to analyze SAR in the context of target hierarchies and phylogenetic trees, two-target activity scatter plots, and biological pathways. Chemotography, in combination with the Kyoto Encyclopedia of Genes and Genomes (KEGG), also allowed us to extract pathway-relevant SAR from the ChEMBL database. We identified chemotypes showing polypharmacology and selectivity-conferring scaffolds, even in cases where individual compounds have not been tested against all relevant targets. In addition, we analyzed SAR in ChEMBL across the entire Kinome, going beyond individual compounds. Our method combines the strengths of chemical space visualization for SAR analysis and graphical representation of complex biological data. Chemotography is a new paradigm for chemogenomic data visualization and its versatile applications presented here may allow for improved assessment of SAR in biological context, such as phenotypic assay hit lists.

1. Introduction

Large amounts of chemical and biological data, such as annotated hit lists from phenotypic screening campaigns, require computational methods to aid in the discovery of useful structure–activity relationships (SARs), as well as interpretation of chemogenomics data in relevant biological context, for example molecular pathways and multi-target activity profiles. A common way to analyze large numbers of biologically active compounds, such as phenotypic screening hit lists,[1] relies on grouping compounds based on their molecular structures with subsequent analysis of their biologically relevant properties.[2,3] Historically, optimizing potency at a single primary target has been the paramount goal of hit list assessment and SAR analysis. This single target-minded philosophy required only one degree of freedom for biological data to be visualized, that is, potency at the primary target. Consequently, techniques to visualize chemical space have used spatial degrees of freedom for chemical descriptors, and used marker size, shape, or color to encode for desirable biological properties.[4–10] Statistical dimension-reduction approaches, such as principal concept analysis, multi-dimensional scaling, generative topographic mapping, etc., aim at reproducing distances and/or local neighborhood relationships from high-dimensional chemical space to few, typically two, dimensions, which serve as the basis for scatter plots.[2,7,11] Network and graph-based approaches[5,12,13] utilize spatial degrees of freedom to layout substructure hierarchies or pair-wise similarity relationships between compounds. In all representations spatial proximity of markers representing chemical structures implies some degree of chemical similarity, explicitly in statistical dimension-reduction approaches and implicitly via layout in network-based representations. The ensuing spatial distribution of markers greatly depends on the chemical descriptors or molecular fingerprints chosen,[8,14] making it hard to

Abbreviations: GPCR, G-protein coupled receptor; LGIC, ligand-gated ion channel; SAR, structure–activity relationship.

* Corresponding author. Tel.: +1 617 871 4953.
E-mail address: eugen.lounkine@novartis.com (E. Lounkine).

0968-0896/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2012.02.034

compare different chemical representations that focus on distinct chemical characteristics.[12,15]

SAR assessment in the chemogenomic era is departing from potency at individual targets towards a holistic understanding of how compounds influence complex biological phenotypes.[1,16,17] Compound activity at multiple targets is analyzed in order to avoid off-target activity[18,19] or because a well-defined multi-target activity profile is desired, for example for kinase inhibitors with well-defined multi-kinase polypharmacology.[20–22] These approaches are more concerned with biological dimensions than with the comparison of chemical structures. A successful visualization technique for kinase profiling has been the projection of a single compound's potency onto the phylogenetic tree of the Kinome.[20,22] In these visualizations, two spatial dimensions are used to display phylogenetic target relationships and compound potency is visualized as markers of different size located at relevant kinase targets. Despite its popularity, this approach is not easily extendable to multiple compounds. The common practice of encoding chemical similarity using spatial proximity cannot be applied, because space is already used by the phylogenetic tree. Thus, a rich representation of biological data comes at the cost of missing chemical information.

Here, we propose a new visualization paradigm to reconcile chemical space visualization for SAR analysis and rich biological data. In our approach, we reserve spatial degrees of freedom for biological data, such as target hierarchies and molecular pathways, and encode chemical space using a one-dimensional color spectrum that combines molecular clustering and dimensionality reduction.

2. Materials and methods

2.1. Pathway and Kinome maps

We selected four pathway maps from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Table 1).[23] From each pathway, we extracted the Entrez gene ids of all targets and selected compounds with activity against the targets from the ChEMBL database (version 11).[24] Only high-confidence annotations (ChEMBL confidence score 9) were extracted that had an IC_{50} value with a qualifier of '=' or '<'. The compounds were processed using Accelrys Pipeline Pilot (version 8) and represented using InChIKeys.[25] The largest fragment of each molecule was retained. If multiple measurements were present for a compound against one and the same target, we calculated the geometric mean of all relevant IC_{50} values. On average, pathways contained 107 targets (unique Entrez gene ids), with 37% of targets having reported compound activity in ChEMBL and 3937 compounds (Table 1). Targets were organized using the ChEMBL taxonomy, which consists of eight levels ranging from a broad classification into enzymes, ion channels, GPCRs, transporters, and transcription factors to individual genes. Each pathway was analyzed separately.

In addition to KEGG pathways, we extracted 9804 kinase inhibitors from ChEMBL using the same criteria as for pathway-relevant compounds (Table 1) and mapped them to a phylogenetic tree of the kinome.[22]

2.2. Clustering

We explored three distinct molecular representations. First, in a topology-centric approach, the molecular graphs were abstracted to carbon skeletons (CSK): side chains were removed, and all non-hydrogen atoms were replaced with carbons and all bonds were made single.[26] Each unique CSK represented one cluster. As the cluster center, we selected a random compound from each cluster. Second, we used extended connectivity fingerprints (Pipeline Pilot ECFP_4), in combination with the Tanimoto coefficient (Tc) and k-medoids clustering, as implemented in Pipeline Pilot. Diverse cluster centers were picked such that each cluster contained 20 compounds on average. Centers were then twice recalculated to minimize the average distance to any of the cluster members. Third, we utilized 2D pharmacophore fingerprints (Pipeline Pilot PHFP_3) in combination with Tc and k-medoids clustering (Table 2).

2.3. Multidimensional scaling

We calculated compound similarity within clusters and also between cluster centers. For CSK and ECFP_4, we used ECFP_4 fingerprints to calculate the similarity; for PHFP_3, we used PHFP_3 itself. We then applied multidimensional scaling (MDS) using the R statistical package (version 2.12, as packaged with Pipeline Pilot) to each cluster (Fig. 1) in order to project distances from the high-dimensional fingerprint space to one dimension. MDS assigned a real number to each compound such that the relative distances (1-Tc) in high-dimensional space were maximally preserved along the one-dimensional chemical coordinate.[7,27] For each cluster, we rescaled the chemical coordinate to the interval [0, 0.99] (Fig. 1; compounds **1a** and **1b** are more similar to each other than to compounds **1c**–**e**). We also applied MDS to the cluster centers, and ordered clusters based on their chemical coordinate value. We then assigned integer coordinates to clusters based on this ordering. Adding cluster coordinates to the real-valued intra-cluster coordinates yielded a one-dimensional chemical coordinate for each compound (Fig. 1; compounds **1f**–**j** are cluster centers with increasing chemical coordinate), which combined the coarse–grained cluster assignment and the detailed intra-cluster chemical distances of compounds.

2.4. Encoding chemical space using color

We visualized chemogenomics data using TIBCO Spotfire (version 3.3.1). We reserved spatial dimensions for bioactivity data and used a color spectrum ranging from red to violet to encode the chemical variable. In each visualization, compounds with the

Table 1
KEGG pathways and ChEMBL data sets

KEGG Id	Description	Map targets[a]	ChEMBL targets[b]	Compounds[c]
hsa04012	ErbB signaling pathway	87	36	4254
hsa04020	Calcium signaling pathway	177	81	6539
hsa05220	Chronic myeloid leukemia	73	33	2756
hsa05323	Rheumatoid arthritis	91	14	2199
N/A	Kinome map	395	152	9804

[a] Number of targets in KEGG map/Kinome.
[b] Number of targets from map for which active compounds in ChEMBL could be found.
[c] Number of unique ChEMBL compounds active at one or more map targets.

Table 2
Clustering methods

Method	Molecular representation	Chemical focus	Clustering
CSK	Carbon skeleton	Scaffold topology	Unique CSK
ECFP4	Extended connectivity fingerprints (ECFP_4)	Atom environments	Divisive k-medoids clustering
PHFP3	2D Pharmacophore fingerprints (PHFP_3)	Functional group graph distances	Divisive k-medoids clustering

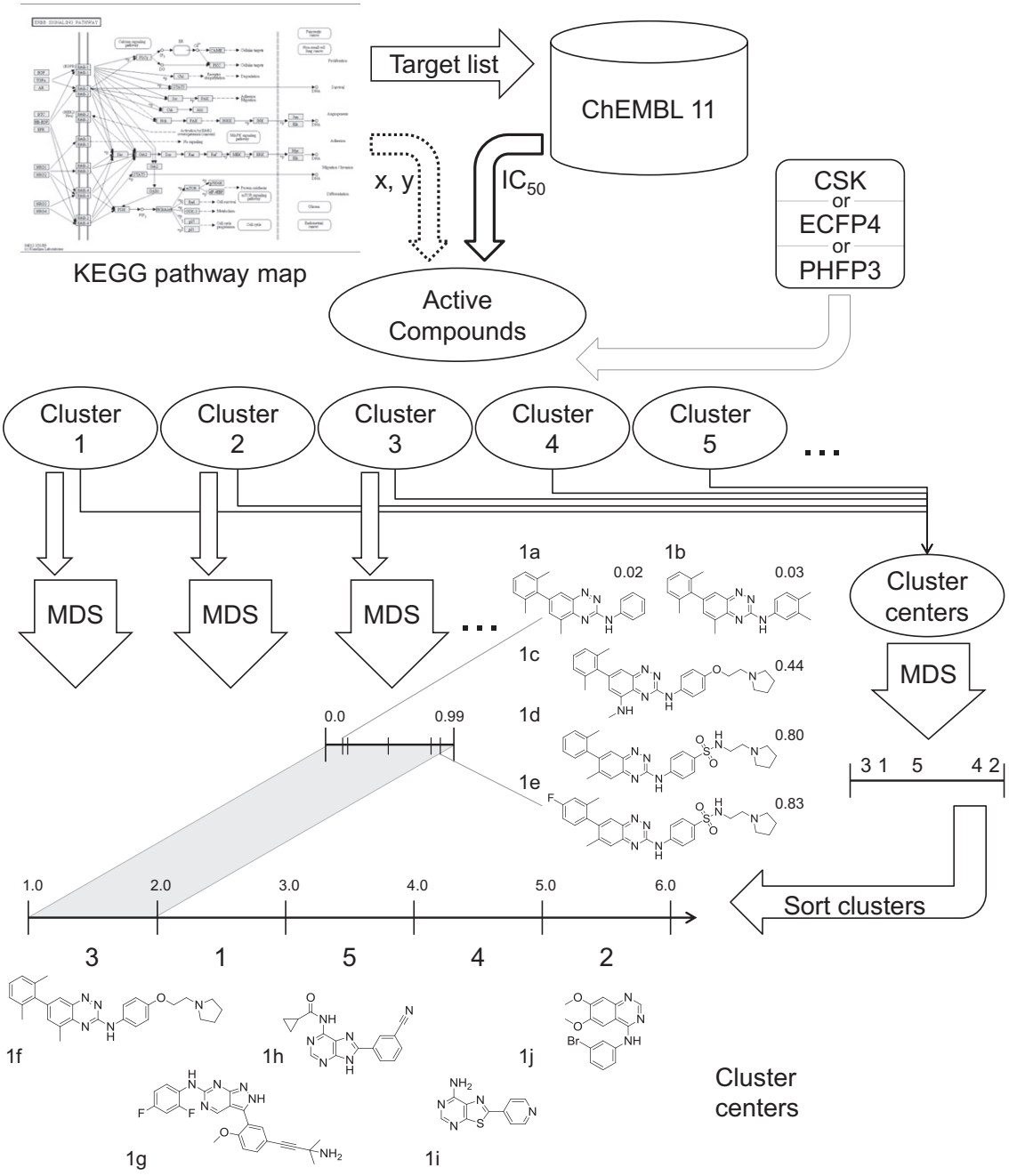

Figure 1. Clustering and multidimensional scaling. Starting from a biologically relevant KEGG pathway map or the kinome phylogenetic tree, active compounds were extracted from the ChEMBL database and encoded using one of three molecular representations (CSK, ECFP4, PHFP3, cf. Table 2). The maps defined coordinates of individual targets, and consequently active compounds. Compounds were clustered based on chemical structure and multidimensional scaling was carried out for each cluster and the pooled cluster centers. The one-dimensional coordinate reflected similarity relationships within each cluster (compounds **1a**–**e**), as well as between clusters (cluster centers **1f**–**j**). The numbers next to the compounds are within-cluster chemical coordinates calculated for the Chronic Myeloid Leukemia data set.

lowest chemical coordinate were represented by red markers and compounds with the maximal coordinate were violet. Compounds with chemotypes in between were assigned orange, yellow, green, and blue, respectively (Fig. 2a).

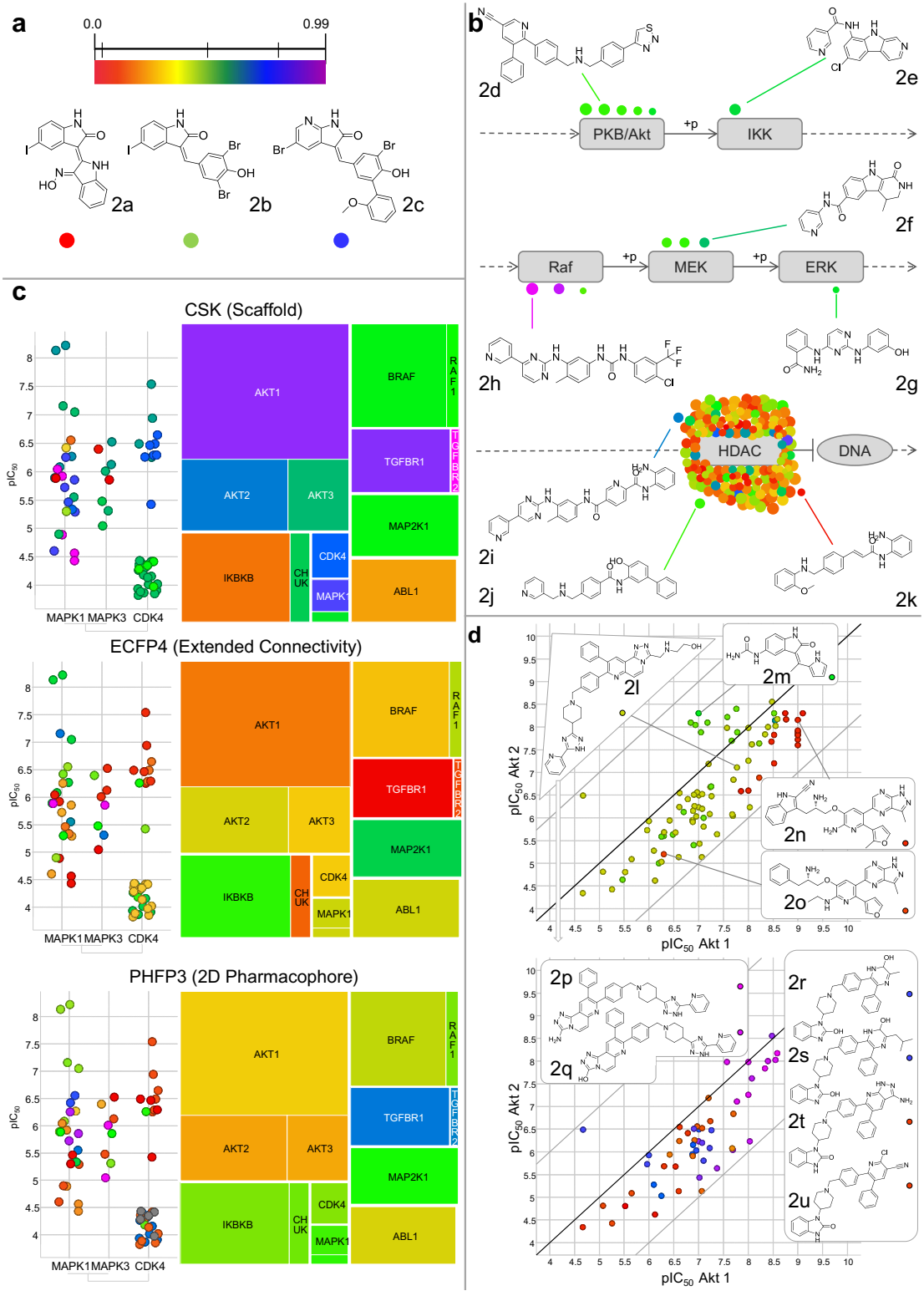

Figure 2. Chemotype color-coding. (a) For individual visualizations, part of the one-dimensional chemical coordinate was mapped to a color spectrum ranging from red (minimum) to violet (maximum). Compounds were represented by markers colored according to their chemical coordinate. (b) In pathway maps, markers from individual clusters were projected onto the targets with a random offset ('jitter') to avoid marker overlap. Marker size codes for potency (large markers correspond to potent compounds). Exemplary structures are shown together with their colored markers. Most of the compounds from the particular cluster shown here were structurally diverse HDAC inhibitors. (c) In activity scatter plots (left), the *x* coordinate was used for the target hierarchy (categorical), and the *y* coordinate for pIC_{50} values (continuous). Markers are placed with a random x-offset. Tree maps (right) summarize the number of active compounds for each target and are color-coded based on the average chemical coordinate. Targets were thus compared in both target hierarchy and chemical space. Three distinct chemical representations are shown, with different similarity patterns emerging dependent on the chemical space used. (d) Two scatter plots of Akt1 versus Ak2 selectivity. The upper plot utilizes the color spectrum to encode the entire chemical space (ECFP4), whereas the bottom plot focuses on the yellow cluster (compound **2l**). Color-coding the smaller part of the chemical space allowed us to further distinguish chemotypes within that cluster, separating highly potent (violet, **2p**, **2q**) from less potent compounds (red, **2t**, **2u**).

3. Results and discussion

3.1. Encoding chemical space using color

We devised a hierarchical, one-dimensional encoding of chemical space using clustering and multidimensional scaling (MDS) (Fig. 1). All compounds relevant to a pathway or the kinome were projected onto one chemical coordinate. SAR analysis often relies on clustering of structurally similar compounds and analyzing each cluster separately, as well as the comparison of chemotypes represented by each cluster. We incorporated this intra- and inter-cluster view into our one-dimensional representation of chemical space. Each cluster was mapped to an integer along the chemical coordinate; clusters were sorted based on the similarity of cluster centers (Fig. 1). Compounds within each cluster were distributed within a unit-sized section of the chemical coordinate. This formalism allowed us to compare different chemical space representations, including two distinct molecular fingerprints and a compound classification based on molecular scaffold topology (Table 2). Furthermore, application of MDS to each cluster and the cluster centers separately substantially reduced the computational cost. For example, MDS of the kinome data set would require nearly 50 million (9804∗9803/2) compound–compound comparisons. However, clustering the data set yielded 491 clusters, amounting to a total number of only 360,891 comparisons for MDS. Although additional computational cost is added dependent on the clustering procedure, the k-medoids clustering and topological classification chosen here were not as expensive as full MDS.

Although chemical space often is visualized using two spatial dimensions, we deliberately chose to use only one dimension. Low-dimensional projection of the high-dimensional chemical space inevitably masks some of the similarity relationships between the compounds and this limitation applies equally to our one-dimensional approach as well as to two- or three-dimensional scatter plot representations.[2,7,28] Similarly, in network graphs the spatial dimensions have no explicit chemical or biological meaning, but are used by layout algorithms to maximize visibility of relevant compound pairs (i.e., edges in network graphs).[2,5,13] In all cases, the relative position of markers representing compounds encodes chemical similarity and groups compounds, whereas the absolute coordinates are often meaningless. An exception is very simple chemical space definitions consisting of only two descriptors, such as molecular weight and lipophilicity. However, irrespective of the number of dimensions (one, two, or three), interactive visualizations of chemical space demand manual confirmation of representative structures. Then, markers close to such selected structures can be assumed to have similar chemical properties, dependent on the chemical space chosen. These neighbor relationships are often of higher importance in SAR analysis than correct representation of all distances including highly dissimilar compound pairs.[7] Hence, we facilitated the identification of neighbors by integrating clustering and MDS in one dimension, rather than relying on two or more dimensions.

One-dimensional neighbor relationships can be encoded using color, and this is exploited in color-coding potency, gene expression magnitude, for example with a palette ranging from red to green. Markers of the same color are thereby perceived as having a similar quantity. In our approach we combined the two observations:

1. 'For SAR analysis, only relative proximity in chemical space needs to be visualized.'
2. 'Color encodes proximity.'

We encoded chemical space using a color spectrum ranging from red to violet (compounds **1a**–**c** in Fig. 2a). A caveat that pertains to all methods that utilize color is differences in color perception and color vision deficiency. In these cases, the palette could be adjusted (e.g., ranging from yellow to blue for red–green blindness). Despite this obvious limitation, Chemotography provided two major advances over spatially encoded chemical space. First, two spatial dimensions could be used for biologically relevant data, such as potency at two or more targets, or target membership in biological pathways (Fig. 2a–d). At the same time, structural similarity between compounds was visualized using color: two red markers represented more similar compounds than a red and a blue marker. Second, different molecular fingerprints focusing on distinct chemical properties (e.g., ring topology vs pharmacophore) could be easily compared, because marker positions were defined by biological compound activity, rather than its chemical structure. Conversely, the traditional spatial approach can only reveal general trends when comparing different chemical space representations, but not distinctive patterns, because markers change their position. Furthermore, our hierarchical design of the chemical coordinate allowed us to map either the entire chemical space to the color spectrum in low-resolution, as well as portions of the chemical space, for example individual clusters or few closely related clusters. We found that the color code was sufficient to group similar clusters and distinguish distinct clusters. To further emphasize cluster membership, markers of different shapes might be used in addition to color and size.

We explored four distinct visualization applications of Chemotography including pathway maps, multi-target activity profiles, target hierarchy tree maps, and selectivity scatter plots. In the following, we describe the various visualizations (Fig. 2a–d) using the KEGG map for Chronic Myeloid Leukemia.

3.1.1. Pathway maps

We visualized distribution of active compounds and their structural similarity using Chemotography in biological context using KEGG pathway maps (Fig. 2b). Spatial dimensions were used to lay out biologically relevant networks of proteins. Compounds were represented by colored markers and placed on top of their targets in pathway maps with a small random offset ('jitter') to reduce marker overlapping. Marker size was used to encode potency (larger markers meaning higher potency). In addition to individual compounds that bind different targets in a pathway or multiple related pathways, Chemotography allowed selection of very similar compounds binding distinct targets in a pathway (e.g., compounds **2d**–**g** and **2j**, and compounds **2h** and **2i**). Since not all compounds have been tested against all targets, identification of similar compounds binding different targets can elucidate chemotypes that modulate multiple pathways of interest. For example, in Chronic Myeloid Leukemia-related pathways, compounds belonging to one ECFP4 cluster inhibiting kinases PKB/Akt were chemically similar to some compounds inhibiting the kinase MEK, but also to some histone deacetylase inhibitors (HDAC, compound **2j**). HDAC inhibitors spanned a wide range of related (same ECFP4 cluster), yet distinct (ranging from red to blue) chemotypes. Furthermore, some HDAC inhibitors (compound **2i**) also shared many chemical features with a Raf inhibitor (**2h**). Conversely, no single compound in our dataset has been tested against both targets. Thus, Chemotography led to the identification of similar compounds active against distinct targets in related, disease-relevant biological pathways.

3.1.2. Activity profiles

In order to provide a high resolution of potency against all targets in a pathway, we also explored utilization of one spatial dimension for the target (class) and the second spatial dimension for potency (left panels in Fig. 2c). This allowed identification of

chemotypes characteristic of potent compounds for individual targets, as well as target families. Furthermore, chemotypes characteristic of high- and low-potency compounds were distinguished from chemical classes that spanned a wide range of potency values. Different target hierarchy levels allowed for direct comparison of SAR both within and across target families (Fig. 2c, Supplementary Fig. S1). For example, Chemotography made apparent that compounds with very distinct scaffolds have been tested at MAPK1, while only a subset of compounds with similar scaffolds have been tested at MAPK3 (Fig. 2c, top panel). For CDK4, there was a separation of two related low-activity scaffold clusters (green markers) and more potent compounds with scaffolds similar to those of MAPK1 and MAPK3 inhibitors (turquoise and blue markers). In this case, visualization of more than one CSK cluster allowed us to distinguish different scaffolds using color. However, as explained above, marker shape could be used to further emphasize cluster membership, and therefore different scaffolds in the CSK chemical space. Chemotography activity profiles also allowed us to detect similar compounds with comparable activity profiles, even when individual compounds had not been tested against all targets. This was revealed by connecting markers representing one and the same compound (Supplementary Fig. S1), which also elucidated chemotypes that have been tested across multiple target families.

3.1.3. Chemical space comparison

Chemical space can be defined in various ways, emphasizing on different molecular features.[15] We have evaluated three distinct chemical space representations including a topological carbon skeleton classification (CSK), extended connectivity fingerprints (ECFP4), which focus on local atom environments[29] and 2D pharmacophore fingerprints (PHFP3), which emphasize graph distances between functional groups. We chose these three representations because they are complementary to each other. In particular, the CSK formalism focused on ring topology only. Other scaffold definitions[5,26] [BM, ST] incorporate more chemical information, but are also more sensitive to features captured by ECFP4 and PHFP3. Hence, we chose CSK to increase the diversity of chemical space representations.

When encoding chemical distances using spatial dimensions, chemical space representations generally result in very distinct arrangements of individual compounds.[8,14] In a one-target setting, this rearrangement can be desired, as it identifies chemical features that group together compounds with a desired property, such as high potency, which can be encoded using color or marker size. However, it is often hard to encode multiple target activities, and follow subsets of compounds from one representation to the other, because compound markers can dramatically change their relative position.

By contrast, in our approach the spatial arrangement of markers identifying compounds was given by their biological activity, and hence did not change when a distinct representation of chemical space was used. Instead, we changed the color of the markers. This allowed us to focus on subsets of compounds that were similar in one chemical space, and directly compare it to a different chemical space (Fig. 2c). For example, the two very similar low-potency CDK4 scaffold clusters (green in Fig. 2c) were further distinguished by ECFP4 (green and orange), and yet further discriminated by PHFP3 (ranging from red to blue, with gray markers identifying compounds for which no PHFP3 could be calculated). A caveat of this color-based chemical space comparison is that colors generally do not match across different representations—a 'red' ECFP4 compound will be generally not 'red' if represented using CSK or PHFP3. However, the position of compounds was invariantly defined by their biological activity, allowing for their identification across distinct chemical representations.

3.1.4. Target tree maps

We explored tree maps as an additional way to visualize the chemotype distribution across different targets. Whereas tree maps have been used to reflect chemical hierarchical clustering,[2] we have applied them to visualization of global target coverage by different chemotypes (Fig. 2c). In tree maps, each rectangle represented a target, or target class if higher hierarchy levels were considered. The rectangles were nested according to the target hierarchy, such that closely related targets were close to each other. The size of the rectangles reflected the number of compounds active at each target. We mapped the average chemical coordinate of active compounds to rectangle color. This allowed for a global overview of chemotype distribution across all targets relevant to a pathway. Again, we were able to compare different chemical representations. For example, scaffolds of AKT kinase inhibitors overall differed substantially from those of BRAF and ABL1 inhibitors, whereas from an atom environment (ECFP4) and 2D pharmacophore (PHFP3) perspective they were more similar. Conversely, TGF beta receptor inhibitor scaffolds were overall similar to AKT1 inhibitors, but the compounds were distinguished using other chemical representations (Fig. 2c).

Two caveats of tree maps were apparent. First, as mentioned in Section 3.1.3, colors were generally not transferrable across different chemical representations, but could only be used to compare targets within the considered chemical space. Second, mapping of the average value of the chemical coordinate did not distinguish between a set of 'green' chemotypes from a mixture of 'red' and 'blue' chemotypes. Therefore, green color could be misleading, and needed to be corroborated by other visualizations, such as scatter plots. Another way to increase the information content of tree maps was the direct inclusion of compounds at the lowest target hierarchy level and mapping rectangle size to compound potency. This representation showed the distribution of all chemotypes, and their potency for all targets (Supplementary Fig. S2). Thus, adjusting the trade-off between chemical and biological resolution via the hierarchy level allowed for an overview of the entire relevant biology and chemical space in one single image.

3.1.5. Selectivity determinants

In cases where compounds have been tested against multiple targets, we used scatter plots to analyze chemical determinants of both compound activity and selectivity (Fig. 2d). Color-coding using the entire chemical coordinate (upper plot in Fig. 2d) allowed us to identify clusters of compounds that were preferentially active and/or selective ('red' cluster in Fig. 2d, upper panel), as well as compound groups that spanned a large range of different potency and selectivity values ('yellow' cluster), signifying a heterogeneous SAR. Color-coding this particular cluster only (bottom plot in Fig. 2d) then provided information about changes to a common core that yielded highly potent, nonselective (compounds **2p** and **2q**, purple) compounds, medium-potency compounds with selectivity potential (**2r** and **2s**, blue) and molecules with low potency against AKT1 and AKT2 (**2t** and **2u**, red).

3.2. Application to pathway-relevant compound sets

Using the visualization techniques described in Section 3.1, we analyzed compounds acting at targets that were part of three additional signaling and/or disease-associated pathways. In this section, we provide representative examples of Chemotography in combination with distinct visualization techniques.

3.2.1. ErbB signaling pathway—SAR in pathway context

The ErbB signaling pathway map contained predominantly kinases. Kinase inhibitors in general can have diverse selectivity profiles.[20] Often it is possible to identify compounds with specific

activity profiles by looking at each compound in isolation. However, the utility of a particular activity distribution of a group of compounds is only apparent in biological context, that is interactions between different pathway members.

Using pathway Chemotography (Fig. 2b), we identified compound **3a** that inhibited multiple downstream ErbB pathways (Fig. 3, compound **3a**, encoded green). Because this visualization put the compound activity into pathway context, it was immediately apparent that compound **3a** inhibited GSK3beta as well as Akt/PKB, which in turn inhibited GSK3beta. The effects of inhibiting both targets may cancel each other out. Therefore, albeit hitting many targets in this map, for the Akt–GSK3beta pathway this compound might not be a good starting point for lead optimization. Conversely, the structurally related (same cluster), yet distinct (encoded violet) compound **3b** was a micromolar inhibitor of both Akt and Ribosomal protein S6 kinase (Rsk). These pathway members are connected by positive interactions, and thus simultaneous inhibition would make sense for protein synthesis modulation. From the activity scatter plot (cf. Fig. 2c) we inferred that the violet chemotype was generally less potent at Akt (pIC_{50} range 6.2–7.2) than the green chemotype (pIC_{50} range 6.8–8.5). However, a chemotype encoded red (representative compound **3c**) spanned a broad potency range (pIC_{50} range 4.5–7.4) of GSK3beta inhibitors. Chemotography allowed us to detect at-a-glance these multitarget SARs, which are laborious to detect using conventional methods. Furthermore, guided by Chemotography, we identified compound **3d** (violet), which was structurally similar to compound **3b**, but had no activity reported for Rsk in ChEMBL. We looked up activity annotations in a commercial database (GVKBio) and found that it was indeed a potent (reported activity 1 nM) Rsk inhibitor.[30] Thus, Chemotography may be applied prospectively to identify compounds with biologically useful polypharmacology. Supplementary Figure 3 depicts the original KEGG pathway with projected compound markers.

3.2.2. Calcium signaling pathway—scaffold analysis

Calcium is one of the key second messengers involved in multiple cellular processes. Consequently, the calcium signaling pathway map in KEGG included receptors of various classes such as GPCRs, Kinases, and LGICs. We sought to identify scaffolds that represented compounds with reported activity at members of different families. Whereas the carbon skeleton representation captured the topology of the scaffold, on its own it was insufficient to distinguish different pharmacophores arising from functional groups added to the common core. Chemotography allowed us to combine the CSK clustering scheme with pharmacophore elucidation.

We used the tree-map to inspect individual CSK clusters, but color-coded them using PHFP3. This view combined the information on scaffold topology and compounds with similar pharmacophores.

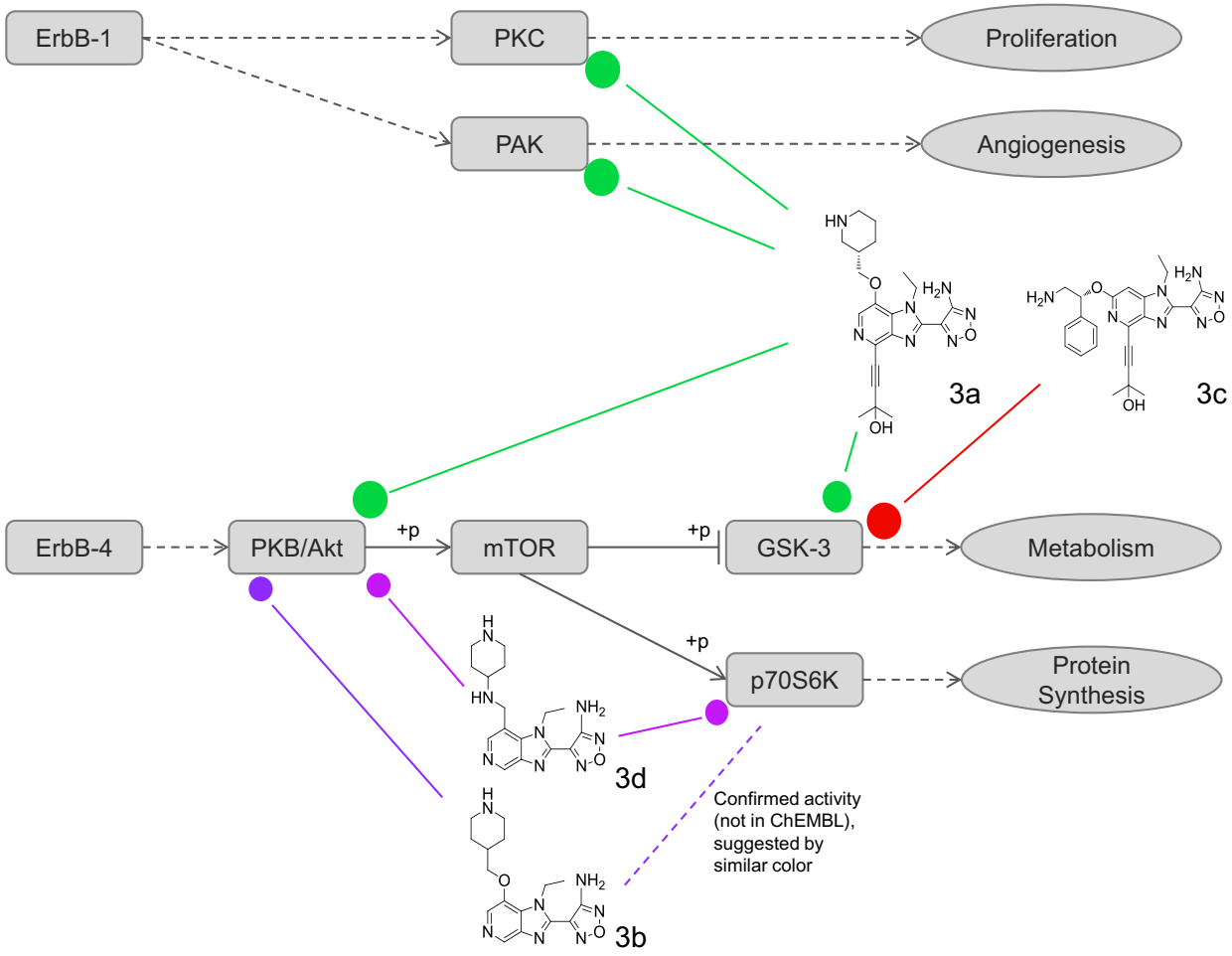

Figure 3. SAR in context of ErbB pathway. Part of the KEGG ErbB pathway map is shown with selected polypharmacologic compounds. Compound **3a** was color-coded green and showed extensive polypharmacology across multiple targets of ErbB pathways. However, using the pathway information it was evident that it inhibited both GSK-3 and its inhibitor, Akt. Compounds **3b** and **3d** were color-coded violet. The dashed line represents confirmed p70S6K activity of compound **3d** that was not reported in ChEMBL. Compound **3c** was color-coded red and also inhibited GSK-3.

Furthermore, the number of compounds screened at each individual target was readily apparent *via* the rectangle size. We identified four scaffolds with surprising activity profiles spanning multiple target families (Fig. 4). Scaffold **4s1** was small and many compounds with distinct pharmacophores matched its topology. We identified similar compounds acting at two unrelated target classes: serotonin receptors, which release calcium via Gq, and Fak2 kinases, which are downstream of the released calcium (compounds **4s1a** and **4s1b** in Fig. 4). These compounds were color-coded orange, indicating the similar pharmacophore. Conversely, compound **4s1c** (inhibitor of nitric oxide synthetase, which is downstream of calcium release) was color-coded violet and had a distinct pharmacophore. Scaffold **4s2** represented NOS inhibitors (**4s2a**), as well as a variety of metabotropic glutamate receptor antagonists (GRM5, **4s2b** and **4s2c**), some of which had considerable pharmacophore overlap with NOS (**4s2b**, color-coded green). In addition, an ErbB2 inhibitor with a related pharmacophore could be identified (**4s2f**). Serotonin receptor antagonists **4s2d** and **4s2e** had the same scaffold and similar pharmacophores (both color-coded violet), which they shared with **4s2c**. Scaffold **4s3** was characteristic of voltage-gated calcium channel blockers (representative compound **4s3a**), as well as serotonin (HTR7, **4s2b**) and leukotriene receptor antagonists

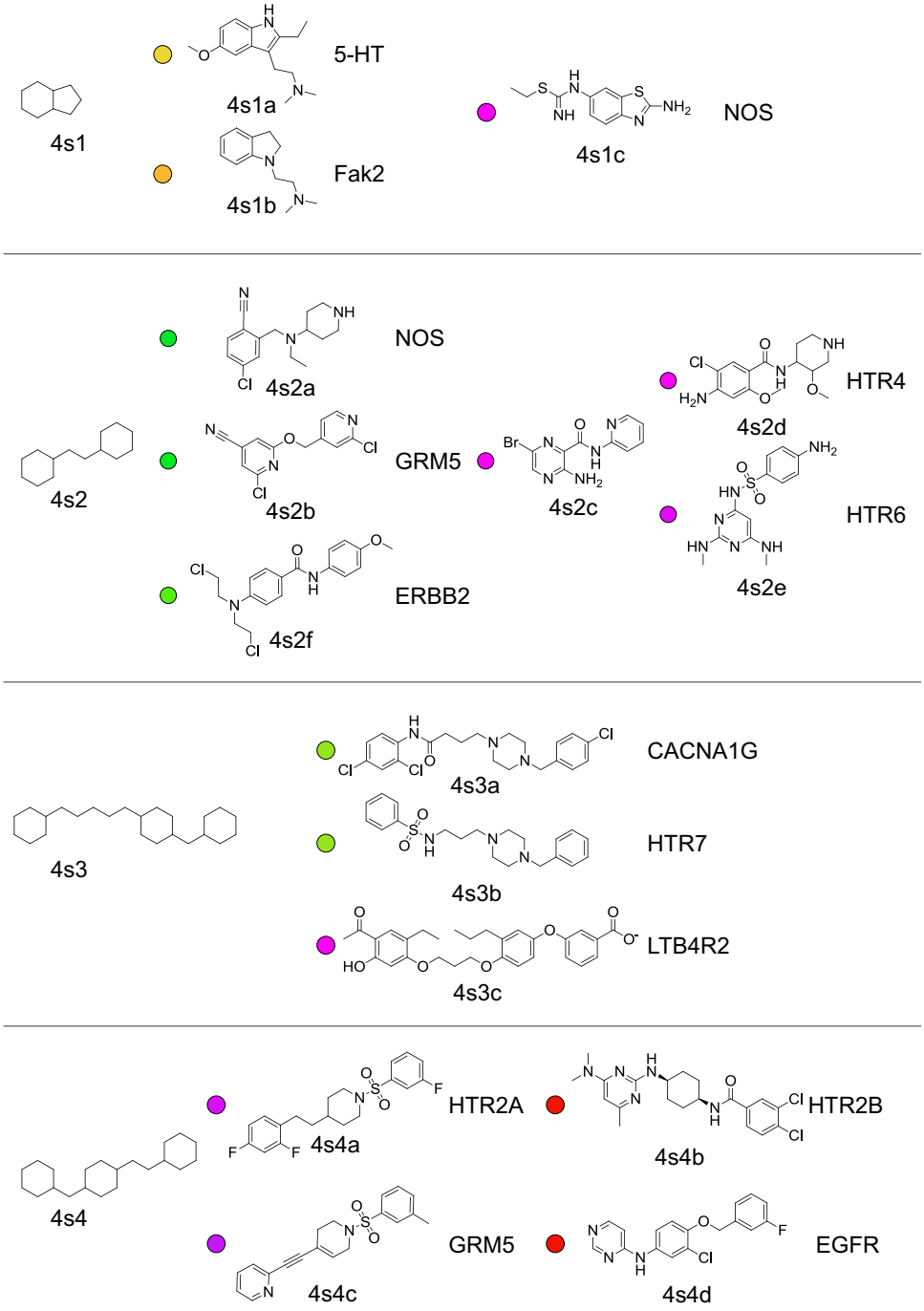

Figure 4. Scaffold polypharmacology in context of calcium signaling pathways Four carbon skeletons representing distinct scaffold topologies are shown (**4s1**–**4s4**) together with corresponding compounds. Targets are indicated next to each compound. Markers next to compounds are color-coded based on PHFP3.

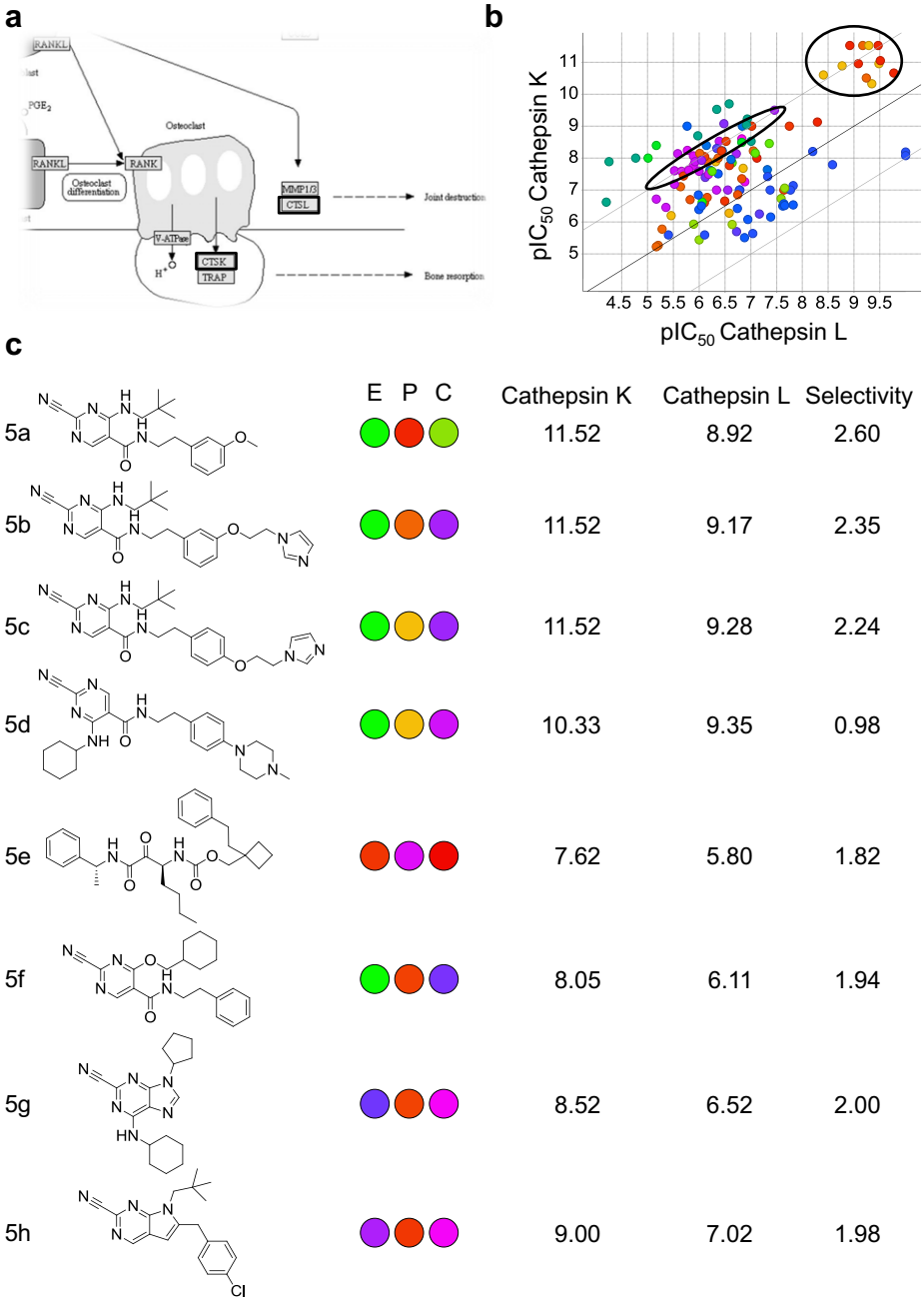

Figure 5. Cathepsin selectivity in context of rheumatoid arthritis. (a) Part of KEGG rheumatoid arthritis map; boxed: cathepsins K and L. (b) Selectivity scatter plot for Cathepsin L versus K colored using PHFP3. (c) Exemplary compounds active at both cathepsins are shown with pIC_{50} values and log-selectivity. Marker colors are shown for each compound and chemical representation: E: ECFP4; P: PHFP3; C: CSK.

(LTB4R2, compound **4s3c**). Although HTR7 and LTB4R2 are GPCRs, the pharmacophore of **4s3a** was more similar to **4s3b** (both color-coded green) than to **4s3c** (color-coded violet). Target tree map Chemotography (cf. Fig. 2c) revealed these surprising discrepancies between target family and chemical similarity via distinct adjacent colors. Some scaffolds represented compounds with distinct pharmacophores, which nevertheless bound to closely related targets. For example, scaffold **4s4** represented compounds with distinct pharmacophores (**4s4a**, color-coded red and **4s4b**, color-coded violet) that bound to serotonin HTR2A and HTR2B receptors. By contrast, these two compounds shared the pharmacophore with GRM5 and EGFR inhibitors (compounds **4s4c** and **4s4d**), respectively. Because we used the spatial dimensions for the target hierarchy in the tree map and activity scatter plot (cf. Fig. 2c), such surprising multi-target SARs were easy to spot.

3.2.3. Rheumatoid arthritis—selectivity

In the KEGG map for rheumatoid arthritis cathepsins L and K both played a role as downstream targets (Fig. 5a). We found that compounds with distinct chemotypes inhibited chatepsin L and/or cathepsin K (Fig. 5b). We compared the three different chemical representations to identify patterns that distinguished active and selective compounds. The atom environments of compounds **5a**–**d** were very similar (same ECFP4 cluster, color-coded green). Among these, potent nitrile compounds with decreasing selectivity for cathepsin K were distinguished by their pharmacophore (Fig. 5c, pharmacophore coordinate increased from compounds **5a** to **5d**). Chemotype color-coding allowed for visual detection of SAR patterns for both targets at once. Thus, we were able to select compounds (**5e**–**5h**) that all were approximately 100-fold more potent for cathepsin K, but differed in their activity (cathepsin K

pIC_{50} range 7.62–9.00). In contrast to highly potent compounds **5a**–**d**, which had very similar atom environments, the ECFP4 coordinate differed substantially (from **5e**: orange to **5g**: blue). Three of these compounds (**5f**–**h**), however, had a very similar

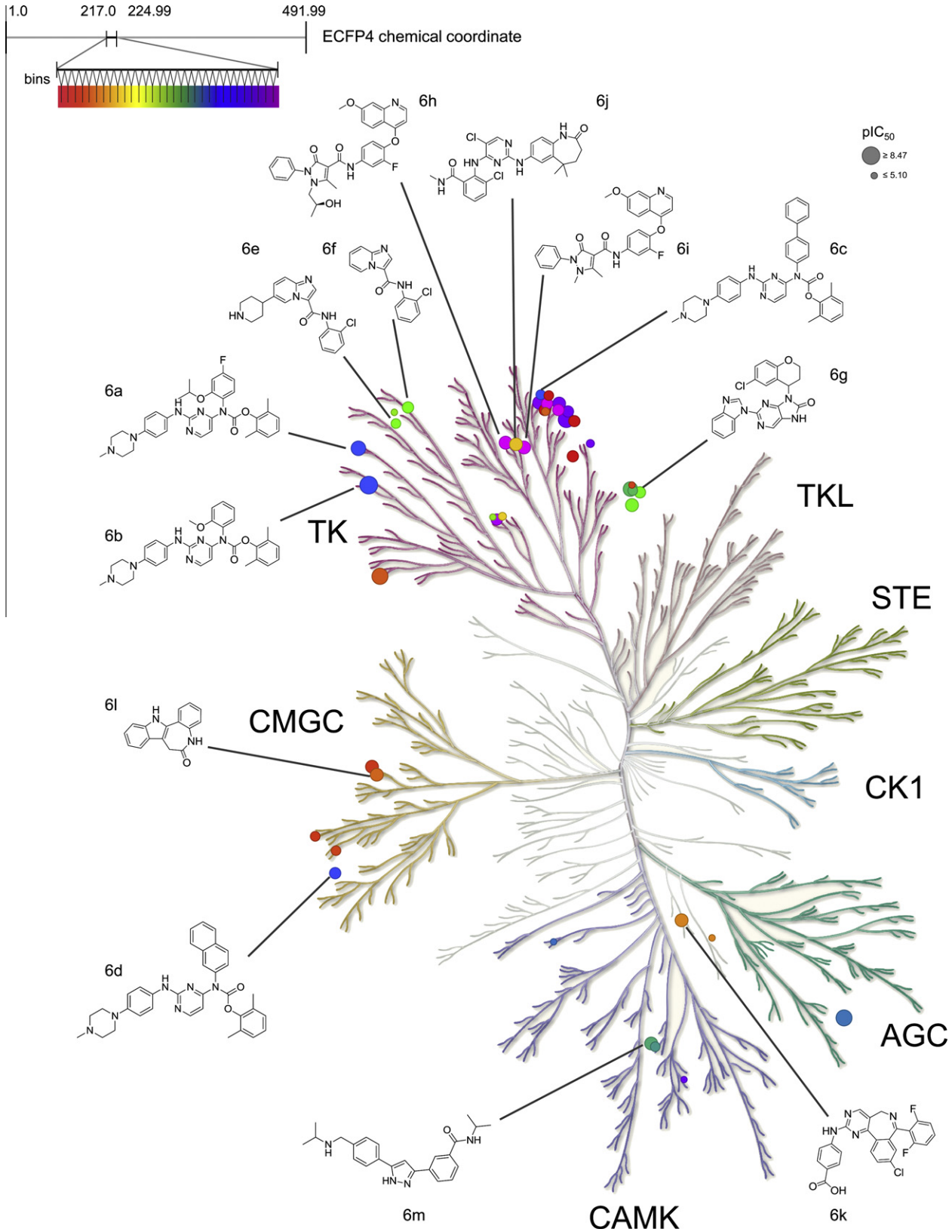

Figure 6. Kinome SAR Part of the ECFP4 chemical coordinate of kinase inhibitors spanning eight clusters was binned and average coordinate values of compounds from each bin were mapped to colors. Markers represent series of similar kinase inhibitor (same chemical coordinate bin) active at individual kinases (positioned according to target position in tree, with small random offset); marker size corresponds to average pIC_{50} values.

pharmacophore and carbon skeleton. Thus, Chemotography enabled us to identify the pharmacophore and carbon skeleton representations as most instructive in distinguishing selective and active from non-selective cathepsin inhibitors in the rheumatoid arthritis map. Although more sophisticated machine-learning methods might be able to select optimal descriptors to distinguish these compounds, our method allowed for intuitive visualization and comparison of three different chemical spaces in the context of cathepsin selectivity. Groups of selective and/or active compounds sharing a similar scaffold or pharmacophore could be easily identified.

3.3. Kinome SAR

In addition to pathway-centric applications, we used Chemotography in combination with a phylogenetic tree of the kinome[22] for Kinome-wide SAR analysis (Fig. 6). The potency of individual compounds has been previously projected onto the kinome tree,[20] and this representation has become a valuable visualization tool in kinome profiling for the distinction of selectivity profiles of kinase inhibitors. We wanted to compare not only the potency of individual compounds, but the activity distribution of different chemotypes that were defined by structurally similar compounds. Chemotography enabled comparison of a large number of compounds, visualizing the distribution of tested and potent chemotypes across the whole kinome (Fig. 6).

Similar to pathway SAR analysis, we used the spatial degrees of freedom to position compounds according to their activity at different kinases. Marker size represented pIC_{50} values. We mapped 9804 unique compounds to one or more of 152 kinase targets that we found in the ChEMBL database. Because we wanted to visualize polypharmacology of multiple compounds, each marker represented a particular compound/kinase pair, analogous to one-compound kinome activity visualizations.[20] Thus, each compound was represented by multiple markers of exactly the same color. Going beyond a particular compound, our method also shows structurally similar compounds via similar colors. Even though not all compounds have been tested against all kinases, Chemotography allows for identification of chemotypes with related polypharmacology profiles. We expected that mapping such a vast number of compounds with multiple target annotations to a phylogenetic tree would make it hard to distinguish activity at closely related targets, as well as minor potency and structural differences. We found two ways to decrease the number of markers and increase the information content of the kinome SAR visualization. First, we limited color-coding to a portion of chemical space (Fig. 6). This made minute changes in color more likely to represent very similar compounds with minor structural alterations. We found that spanning 5–10 clusters (i.e., integers on our chemical coordinate, cf. Fig. 1) yielded most intuitive results. Second, we binned the chemical coordinate and showed one marker for each bin. We mapped the average value of the chemical coordinate to the color of each marker, and the average pIC_{50} value to the marker size. This way, each marker represented a series of similar compounds. Distinct from color averaging in target tree maps, chemotype bins always represented neighbors in chemical space. We were able to distinguish overall potent series of compounds from less potent ones and identify similar compounds that have been found active at kinases belonging to different families. For example, compounds **6a**–**d** were color-coded blue; this chemotype was overall most potent at the TK family kinase LCK (average $pIC_{50} = 8.47$) and has been tested at kinases of both TK and CMGC families (Fig. 6). Conversely, the 'green' chemotype (compounds **6e**–**g**) has only been found to be active at TK kinases. Compounds **6j**–**l** were color-coded bright to dark orange and have been tested across several families. By contrast, none of these individual compounds has been tested across multiple families. Compound **6m** showed a chemotype distinct from the other compounds and has only been tested at PKD1. Thus, Chemotography allowed us to distinguish the different polypharmacology profiles of distinct chemical series, each encoded with a distinct color spectrum, such as 'green' imidazopyridines and 'blue' *N*-methylpiperazines.

With the increasing availability of interactive visualization tools such as TIBCO Spotfire, which we have used here, it becomes possible to incorporate many activity annotations and other biologically relevant data, such as biological network data (Fig. 2b). This tendency shifts the focus from one-target SAR to a holistic understanding of chemotypes and their role in complex biological systems. Taken together, our results demonstrated that Chemotography substantially extends the visualization toolbox currently available for chemogenomic data analysis. Given sufficient activity annotation, it may be useful in hypothesis generation for phenotypic screening campaigns, where it might hint towards phenotype-relevant biological pathways and structural selectivity determinants of compounds that yield a well-defined phenotypic response.

4. Conclusion

Traditionally, for SAR analysis, similarity relationships between compounds in high-dimensional chemical space have been projected onto few (two or three) spatial dimensions. Color was then often used to encode biologically relevant properties. This practice limited SAR analysis to one potency at a time, and different chemical space representations were not easily comparable. Chemotography introduced here extended SAR analysis to multiple targets in their specific biological context, including activity in biological pathways and selectivity across the kinome. Spatial dimensions in our approach were used for biology, and color was used to indicate similar chemotypes. Our method allowed for a comprehensive biological interpretation of SAR, and easy comparison of different chemical space representations, including atom environments, 2D pharmacophores, and scaffold topology. Chemotography is versatile and allows visualization and easy comparison of multiple active compounds in a broad biological context. This method will be therefore particularly useful for assessment of hit lists derived from phenotypic screens, as it reconciles visualization of chemical space and rich biological data.

Acknowledgements

E.L., P.K., and P.P. are Postdoctoral fellows supported by the Education Office of the Novartis Institutes for Biomedical Research.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2012.02.034.

References and notes

1. Young, D. W.; Bender, A.; Hoyt, J.; McWhinnie, E.; Chirn, G.-W.; Tao, C. Y.; Tallarico, J. A.; Labow, M.; Jenkins, J. L.; Mitchison, T. J.; Feng, Y. *Nat. Chem. Biol.* **2008**, *4*, 59.
2. Wawer, M.; Lounkine, E.; Wassermann, A. M.; Bajorath, J. *Drug Discovery Today* **2010**, *15*, 630.
3. Glick, M.; Jacoby, E. *Curr. Opin. Chem. Biol.* **2011**, *15*, 540.
4. Oprea, T. I.; Gottfries, J.; Sherbukhin, V.; Svensson, P.; Kühler, T. C. *J. Mol. Graphics Modell.* **2000**, *18*(512–524), 541.
5. Wetzel, S.; Klein, K.; Renner, S.; Rauh, D.; Oprea, T. I.; Mutzel, P.; Waldmann, H. *Nat. Chem. Biol.* **2009**, *5*, 581.
6. Garcia-Serna, R.; Ursu, O.; Oprea, T. I.; Mestres, J. *Bioinformatics* **2010**, *26*, 985.
7. Owen, J. R.; Nabney, I. T.; Medina-Franco, J. L.; López-Vallejo, F. *J. Chem. Inf. Model.* **2011**, *51*, 1552.
8. Iyer, P.; Hu, Y.; Bajorath, J. *J. Chem. Inf. Model.* **2011**, *51*, 3026.

9. Iyer, P.; Bajorath, J. *Chem. Biol. Drug Des.* **2011**, *78*, 778.
10. Maniyar, D. M.; Nabney, I. T.; Williams, B. S.; Sewing, A. *J. Chem. Inf. Model.* **2006**, *46*, 1806.
11. Reutlinger, M.; Guba, W.; Martin, R. E.; Alanine, A. I.; Hoffmann, T.; Klenner, A.; Hiss, J. A.; Schneider, P.; Schneider, G. *Angew. Chem. Int. Ed.* **2011**, *50*, 11633.
12. Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. *Drug Discovery Today* **2009**, *14*, 698.
13. Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. *J. Med. Chem.* **2008**, *51*, 6075.
14. Medina-Franco, J. L.; Yongye, A. B.; Pérez-Villanueva, J.; Houghten, R. A.; Martínez-Mayorga, K. *J. Chem. Inf. Model.* **2011**, *51*, 2427.
15. Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. *J. Chem. Inf. Model.* **2009**, *49*, 108.
16. Lamb, J.; Crawford, E. D.; Peck, D., et al *Science* **2006**, *313*, 1929.
17. Feng, Y.; Mitchison, T. J.; Bender, A.; Young, D. W.; Tallarico, J. A. *Nat. Rev. Drug Disc.* **2009**, *8*, 567.
18. Nigsch, F.; Lounkine, E.; McCarren, P.; Cornett, B.; Glick, M.; Azzaoui, K.; Urban, L.; Marc, P.; Müller, A.; Hahne, F.; Heard, D. J.; Jenkins, J. L. *Expert Opin. Drug Metab. Toxicol.* **2011**, *7*, 1497.
19. Giacomini, K. M.; Krauss, R. M.; Roden, D. M.; Eichelbaum, M.; Hayden, M. R.; Nakamura, Y. *Nature* **2007**, *446*, 975.
20. Anastassiadis, T.; Deacon, S. W.; Devarajan, K.; Ma, H.; Peterson, J. R. *Nat. Biotechnol.* **2011**, *29*, 1039.
21. Gomase, V. S.; Tagore, S. *Curr. Drug Metab.* **2008**, *9*, 255.
22. Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. *Science* **2002**, *298*, 1912.
23. Kanehisa, M.; Goto, S.; Furumichi, M.; Tanabe, M.; Hirakawa, M. *Nucleic Acids Res.* **2010**, *38*, D355.
24. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. *Nucleic Acids Res.* **2011**.
25. Stein, S.; Heller, S.; Tchekhovski, D. *Nimes Int. Chem. Inf. Conf. Proc.* **2003**, 131.
26. Bemis, G. W.; Murcko, M. A. *J. Med. Chem* **1996**, *39*, 2887.
27. Bronstein, A. M.; Bronstein, M. M.; Kimmel, R. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 1168.
28. Rosén, J.; Lövgren, A.; Kogej, T.; Muresan, S.; Gottfries, J.; Backlund, A. *J. Comput. Aided Mol. Des.* **2009**, *23*, 253.
29. Rogers, D.; Hahn, M. *J. Chem. Inf. Model.* **2010**, *50*, 742.
30. Heerding, D. A.; Rhodes, N.; Leber, J. D., et al *J. Med. Chem.* **2008**, *51*, 5663.