

# Chemogenomic Data Analysis: Prediction of Small-Molecule Targets and the Advent of Biological Fingerprints

Andreas Bender<sup>\*,a</sup>, Daniel W. Young<sup>b,§</sup>, Jeremy L. Jenkins<sup>a</sup>, Martin Serrano<sup>c</sup>, Dmitri Mikhailov<sup>a</sup>, Paul A. Clemons<sup>c</sup> and John W. Davies<sup>a</sup>

<sup>a</sup>Lead Discovery Informatics, Novartis Institutes for BioMedical Research Inc., 250 Massachusetts Ave., Cambridge, MA 02139, USA

<sup>b</sup>Developmental and Molecular Pathways, Novartis Institutes for BioMedical Research Inc., Cambridge, MA 02139, USA

<sup>c</sup>Chemical Biology Program, Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA 02142, USA

<sup>§</sup>Current Address: Wolf, Greenfield & Sacks, P.C., 600 Atlantic Avenue, Boston, MA 02210, USA

**Abstract:** Chemogenomics comprises a systematic relationship between targets and ligands that are used as target modulators in living systems such as cells or organisms. In recent years, data on small molecule-bioactivity relationships have become increasingly available, and consequently so have the number of approaches used to translate bioactivity data into knowledge. This review will focus on two aspects of chemogenomics. Firstly, in cases such as cell-based screens, the question of which target(s) a compound is modulating in order to cause the observed phenotype is crucial. *In silico* target prediction tools can suggest likely biological targets of small molecules *via* data mining in target-annotated chemical databases. This review presents some of the current tools available for this task and shows some sample applications relevant to a pharmaceutical industry setting. These applications are the prediction of false-positives in cell-based reporter gene assays, the prediction of targets by linking bioassay data with protein domain annotations, and the direct prediction of adverse reactions. Secondly, in recent years a shift from structure-derived chemical descriptors to biological descriptors has occurred. Here, the effect of a compound on a number of biological endpoints is used to make predictions about other properties, such as putative targets, associated adverse reactions, and pathways modulated by the compound. This review further summarizes these “performance” descriptors and their applications, focusing on gene expression profiles and high-content screening data. The advent of such biological fingerprints suggests that the field of drug discovery is currently at a crossroads, where single target bioassay results are supplanted by multidimensional biological fingerprints that reflect a new awareness of biological networks and polypharmacology.

**Keywords:** Chemogenomics, chemical space, activity spectra, multiple-target drugs, multiple ligands, high-content screening, target prediction, target fishing, gene expression profiles, affinity fingerprints.

## WHAT IS CHEMOGENOMICS?

Chemogenomics is the study of the relationships between targets by relating the structures and activities of their ligands [1-3]. It is related to the concepts of chemical genetics [4, 5] and chemical genomics [6] in that all approaches investigate the perturbation of biological systems with small molecules. While chemogenomics approaches emphasize inherent relationships between targets, chemical genetics and chemical genomics approaches emphasize the influence of the small molecule on the biological system (by analogy to standard genetic techniques such as “knock-out” organisms). In recent years, small molecule-bioactivity information has become available on a large scale and in an electronic form, along with the ability to process this information virtually in real-time on desktop computers. These advances enable the extraction of information from datasets on an unprecedented scale. For example, researchers are no longer limited to considering the characteristic features shared by antagonists of

one particular receptor, but rather can extract characteristic features *for antagonists of all (known) receptors, and how they relate to each other.*

The hierarchical nature of target class similarity [3] has a profound influence on the way novel ligands are discovered. Although high-throughput screening has long been the method of choice for lead discovery in the pharmaceutical industry, knowledge about ligands of one target and the distance between targets in biological space facilitates educated guesses as to which molecules are suitable for novel targets. Thus, we can relate targets by the similarity of ligands to which they bind [2, 7-9] – a “central paradigm of chemogenomics”.

Chemogenomics applications rely heavily on the available information source, which most often means small molecule databases annotated with bioactivity data. For an overview of current chemogenomics databases (as well as target prediction methods) the reader is referred to a recent review [10]. The present review will extend the scope of this previous literature survey on target prediction by presenting applications of the concept. In addition we will focus on a comparatively novel area: employing biological fingerprints

\*Address correspondence to this author at the Lead Discovery Informatics, Novartis Institutes for BioMedical Research Inc., 250 Massachusetts Ave., Cambridge, MA 02139, USA; Tel: +1 (617) 871-3972; E-mail: Andreas.Bender@novartis.com

(instead of structural ones) to describe the behavior of a molecule and to make predictions about other properties.

### THE RATIONALE BEHIND TARGET PREDICTION

Directly following from the lock-and-key analogy for small molecules and their receptors in the human body, the question arises: into *which* locks will a key fit, *i.e.*, with which targets will a compound interact. Early medicines, such as Aspirin and its precursors, have been used for thousands of years without knowing the protein targets, and their relatively safe use has been established empirically. This situation also applies to some current drugs on the market such as the Bcr-Abl/c-KIT inhibitor Gleevec where, interestingly, promiscuity has been increasingly viewed in a positive light because attacks at multiple points in signaling pathways may be more efficient than antagonizing single targets, and further, may limit the risk of developing drug resistance [11]. (Of course, promiscuity also bears the disadvantage of increased likelihood of interfering with undesired pathways [12, 13]). Still, in order to conduct proper clinical trials (both ethically and economically) it is beneficial to know as much as possible about the system one is investigating. Both (desired) on-target effects as well as (undesired) off-target effects can be better understood by knowing the targets modulated by a given drug. This is beneficial on both accounts: for the desired targets, the mode-of-action of an active ingredient can be established, facilitating its rational optimization (e.g. by crystallization of the target protein). In addition, (off-)targets completely unrelated to the actual target can be established, enabling us to predict other, secondary side-effects [10].

Before the advent of molecular biology and protein expression systems, all drug discovery was necessarily “systems based” drug discovery. However, over the past few decades a shift towards target-based screening took place, leading to higher-throughput biochemical assays. This step towards target-based screening was caused by a “selectivity paradigm” which maintained that selective interference with metabolic or signaling networks is an ideal characteristic for medicines, thus promoting drugs to be as selective as possible. More recently it was realized that cellular networks are often able to compensate for single-point modifications [14], resulting in less effective single-target drugs than their potent affinity values might lead one to believe.

Several experimental strategies for elucidating the target of a compound exist such as affinity chromatography, expression cloning and protein microarrays [15]. Still, these approaches usually require a large amount of time and capital expenditure. Thus, this review describes two alternative state-of-art *in silico* concepts used to predict compound mode-of-action. Firstly, statistical models can be generated on the structures of compounds known to show a certain bioactivity, which is undoubtedly one of the major applications of a chemogenomics database. While structure-based chemical descriptors have proven to work quite well for target prediction [16], in recent years a shift from structure-derived descriptors to biological descriptors could be observed. Here, the affinity of a compound to a *panel* of targets was used to generate models of a very diverse nature. For example, “affinity fingerprints” [17] can be used to make predictions about putative targets, about adverse reactions, or

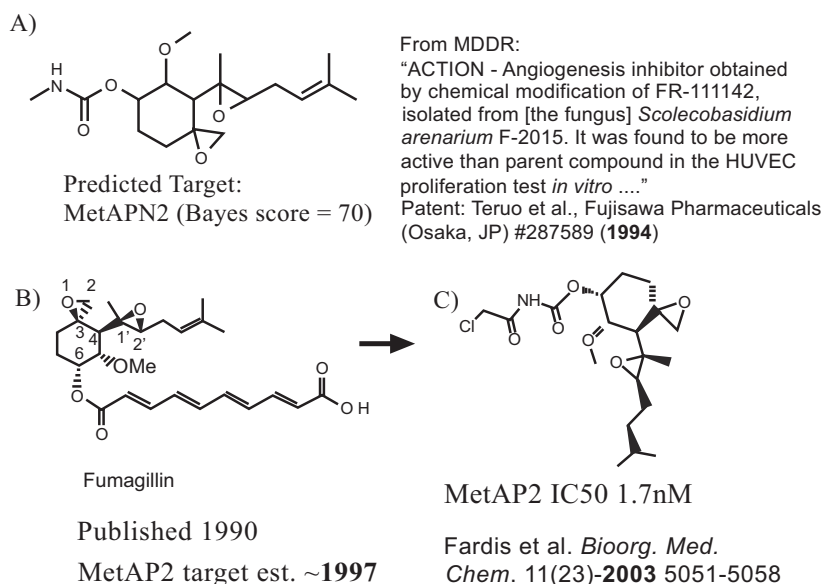
about pathways modulated by the compound. Given the multi-dimensional nature of these descriptors, since their construction comprises activities across multiple potential targets, they can be seen as an inherently “chemogenomics” representation of small molecules. The second half of this article provides a review of such descriptors as well as their applications.

### APPLICATIONS OF *IN SILICO* TARGET PREDICTION BASED ON CHEMICAL DESCRIPTORS

Current approaches to predicting targets of small molecules can be broadly grouped into four classes: chemical similarity searching, data mining/machine learning, panel docking, and the analysis of bioactivity spectra. While the first two approaches are briefly discussed in the following, bioactivity spectra have the entire following chapter dedicated to it and docking-based approaches are omitted here. For a more comprehensive review of target prediction methods based on chemical descriptors see a recent review on the topic [10].

Chemical similarity searching [18] for target prediction compares a compound structure to a database of compounds with known targets. By employing the “molecular similarity principle” the target of similar compounds may also be the target of the query structure. Likewise, it is (within limitations) well-established that similar targets bind to similar compounds [2]. Thus, conventional similarity searching is inverted – while usually *new ligands* for a *known target* are desired, in this case *new targets* for a *known ligand* are proposed. Most commonly, due to their speed, 2D descriptors such as circular fingerprints [19] are employed for target prediction approaches [16]. Similarity searching as a method for ‘target fishing’ (*i.e.*, identification) can also be performed with 3D chemical descriptors [20]; in this case it was found that while 2D descriptors are powerful for similarity searching in annotated databases, 3D descriptors are often more appropriate when the orphan compound has low 2D similarity to all database molecules [20].

Data mining in annotated chemical databases is a second, somewhat more sophisticated method of predicting targets for small molecules. Multidimensional models resulting from data mining differ from similarity searching in that information from multiple ligands can be considered in parallel to make target suggestions. One caveat however is that a *systematic nomenclature* for target information is required for forming distinct activity class sets. As one of the first such applications, the PASS (Prediction of Activity Spectra for Substances) application used circular fingerprints to train a Bayesian-type classifier on the target activities listed in the MDDR database [21]. An approach based on a more comprehensive list of targets has been presented more recently, using circular fingerprints and a Bayesian classifier, but this time trained on a total of 964 targets [16]. Generic activity classes from the MDDR database were successfully assigned to target-specific activities, as examples from antineoplastic, antihypertensive and kinase inhibitor classes showed. One example for an antiangiogenic compound is shown in Fig. 1. From the patent literature and the MDDR abstract it was known that the compound shown was an “... angiogenesis inhibitor obtained by chemical modification of FR-111142, isolated from [the fungus] *Scolecobasidium arenarium* F-



**Fig. (1).** An angiogenesis inhibitor with unknown mode of action was given (A), for which we predicted the most likely target to be MetAPN2, a target known to be involved in angiogenesis. The underlying bases of the prediction were structures like Fumagillin (B), for which both effect and target have been established previously. More recently, a structure similar (C) to the query structure has been published, which shows significant similarity and which is a known ligand of MetAP2. Thus, the prediction is both internally consistent with the phenotype of inhibiting angiogenesis, as well as being consistent with the "molecular similarity principle" that similar molecules have similar biological properties.

2015. It was found to be more active than parent compound in the HUVEC proliferation test *in vitro*....". Despite this annotation, the reasons for its antiangiogenic properties were unknown. The Bayesian model predicted Methionine Aminopeptidase-2 (MetAPN2) as a target for this compound, with a relatively high score of 70 (scores above roughly 30 bear significance in practice; unpublished data). Browsing through the structures used for model generation for MetAPN2, two compounds could be found in the database, also shown in Fig. 1. Fumagillin was established as an inhibitor of Methionine Aminopeptidase-2 around 1997, and the second compound shown is indeed an inhibitor of MetAPN2, at a very low IC<sub>50</sub> of 1.7nM. The only two modifications are a methylether group instead of an epoxide ring and a carbonyl and chloride function instead of a simple methyl function in the other case – two rather minor modifications. Given the overall structure of the compound, the assumption that MetAPN2 is a target of the first structure given is very reasonable.

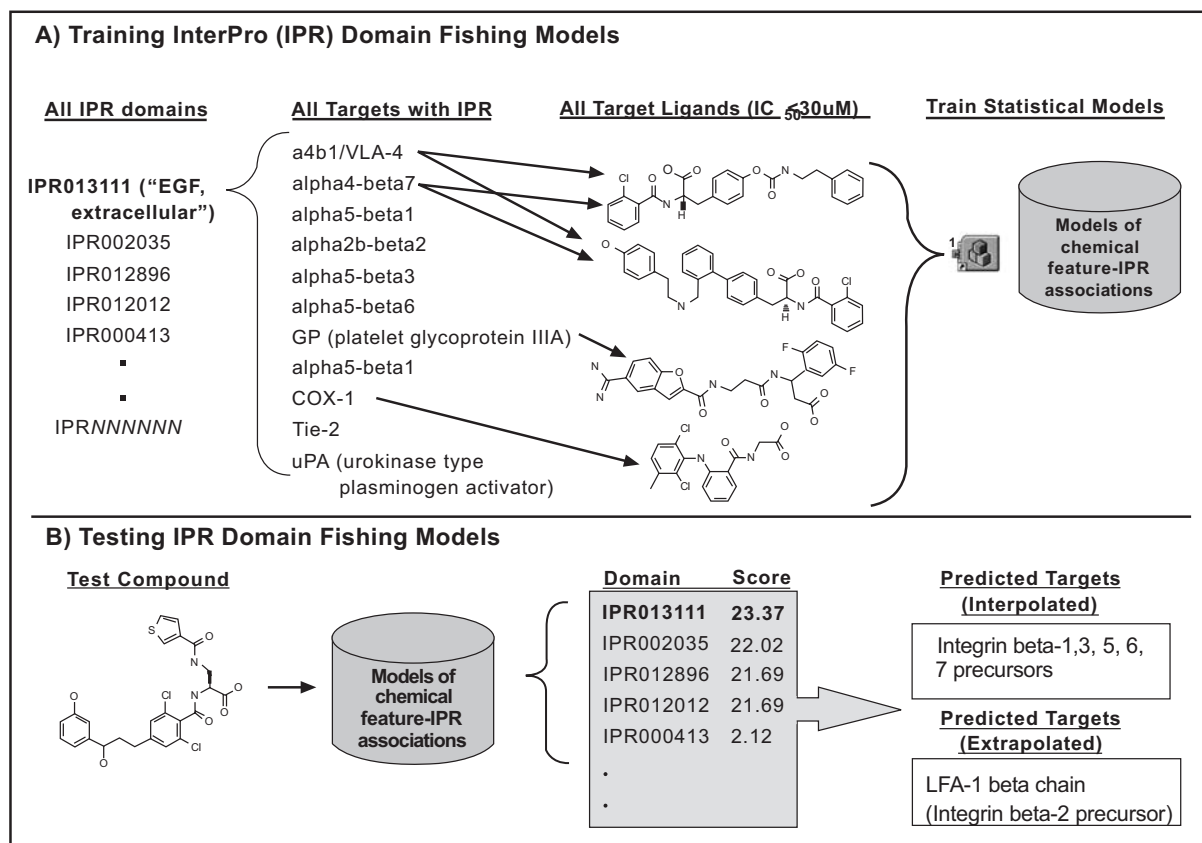
#### APPLICATION OF TARGET PREDICTION TOOLS: TRIAGING FREQUENT HITTERS IN REPORTER GENE ASSAYS

One practical application the authors recently explored in a day-to-day pharmaceutical setting is the triaging of hit lists from high-throughput reporter gene assays (RGAs) [22]. In RGAs, target and final readout are often several signaling steps apart, leading to the possibility of interference of the small molecule with undesired targets. In particular, we observed that some compounds occur as "frequent hitters" across various RGAs irrespective of cell or promoter types. A Bayesian model built on the chemical fingerprints of RGA frequent hitters achieved an average compound hit rate (frac-

tion of active compounds) of 50% on an external validation set in guessing a priori which compounds would be generically active in RGAs (compared to a 2% hit rate in a random selection of compounds), illustrating the predictive power of the developed model. Physicochemical properties on the other hand showed few differences between frequent hitters and selective compounds. Additionally, using *in silico* target prediction tools as described in the previous section, it was found that the predicted targets of frequent hitters are more often associated with undesirable cellular effects such as cytotoxicity and inhibition of the luciferase reporter. The most frequently predicted targets relate to apoptosis and cell differentiation and they include inhibitors of mixed lineage kinases, cyclin-dependent kinases, topoisomerase II, protein phosphatase 1 and protein phosphatase 2, each of which can be rationalized. Thus, *in silico* target prediction enables the concomitant prediction of false positives in reporter gene assays as well as mode-of-action leading to the false positive readout.

#### DOMAIN PREDICTION

While the prediction of targets is able to add knowledge to cell-based screens, one shortcoming is the restriction of the predictions to the training set: targets can only be predicted if they exist already in the database and have known ligands. In order to provide extrapolative abilities to targets outside of the training set, the authors recently extended the same statistical modeling approach beyond *target-based* models to protein *domain-based* models (Fig. 2). The rationale behind this approach was that similar ligands are more likely to bind not only to the same target, but also to the same protein folds or amino acid sequence when they occur in other proteins. For example, if ligands of a protein P1 with



**Fig. (2).** Illustration of the domain prediction workflow. In **A**), ligands are annotated with the Interpro domains they target and models are generated based on ligand features associated with each Interpro ligand set. In **B**) the model is validated, predicting Interpro domains—and thus targets by association—that the ligands are likely to bind. By employing Interpro domains as abstractions of the targets the models are able to extrapolate to targets even if no ligands of it are known (but if it shares the same Interpro domain of targets for which ligands are known). This process is entirely automated in a Pipeline Pilot protocol (Scitegic, Inc.).

fold A are known, and this fold is shared by both protein P1 and protein P2, a target prediction of protein fold A for a test compound would encompass both protein P1 and protein P2, even though no ligands for protein P2 were present in our database. Thus, ligand-based "domain fishing" can widen the net, so to speak, to hypothesize additional targets for an orphan compound.

This probabilistic linkage of (ligand) chemical features to protein folds attempts to create association rules in order to predict receptor-ligand binding affinities and the responsible receptor-ligand interacting motifs [23, 24]. This work is intriguing in that it uses learning to find a priori protein features involved in ligand binding. Small-molecule binding site annotation for protein sequences is still not well documented but an effort has been made to compile known small-molecule binding domains from X-ray crystal structures into a Small Molecule Interaction Domain (SMID) database [25]. This domain prediction method is predicated on annotating the targets from a chemogenomics database (e.g. WOMBAT [26]) with their established InterPro domains. The InterPro database attempts to integrate multiple databases with various protein signature detecting techniques. The source databases are compilations of sequence motifs and clusters signifying protein domains, folds, functional sites, and families,

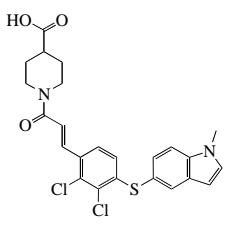
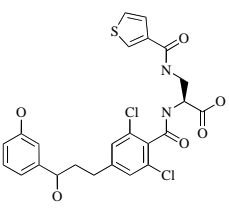
many of which are nested within a parent-child hierarchy [27]. Once ligand chemical structures for targets are linked to the target Interpro domains, machine learning approaches such as Bayesian classification can be used to discover tens of thousands of chemical substructures statistically associated (or not associated) with thousands of protein domains, which can, as the target prediction models themselves, be used to predict *protein folds* for a compound fed into the model. Notably, associations learned by the Bayesian classifier may not reflect an actual small-molecule binding interaction; however, as observed Strombergsson *et al.* [23], even local protein substructures not in contact with ligands may encode certain information during model building that may be useful. The relationship between Interpro domains within a given protein is not random, but rather it is a pre-established functional linkage that is likely to co-occur in other proteins. Thus, association of a given target's ligands to all of its domains, while not physically "accurate", may translate to correct probabilistic target predictions for test molecules, especially given the hierarchical nature of Interpro domains. While simplifications are employed in the domain-based target prediction model, such as unifying domains to a standard nomenclature (weeding out their differences which might remain by assigning identical Interpro

domains) and not accounting for actual domain accessibility by the ligand, in the author's experience the advantage of being able to predict targets outside the training set (as well as protein targets from other organisms than those in the training set!) was outweighing its limitations.

Fig. 2 illustrates how target prediction can be performed *via* "domain fishing". A practical example is provided in Table 1 where the two displayed compounds, labeled compounds 19 and 20 to be consistent with the Arkin and Wells review article [28], are known protein-protein interaction (PPI) inhibitors of the LFA-1/ICAM complex. Each compound binds to leukocyte function associated molecule 1 (LFA-1) at allosteric sites that do not appear to directly interfere with its interaction with ICAM. Compound 19 binds to the I-domain in the  $\alpha$ -chain of LFA-1, whereas compound 20 binds at the interface of the  $\beta$  and  $\alpha$  chains interacting with the I-like domain in the  $\beta$  chain. Importantly, the training set

for the models (WOMBAT) contains inhibitors of the target LFA-1, but *only compounds that bind to the I-domain in the  $\alpha$  chain*, such as Compound 19. Thus, the target binding site of Compound 19 is easily predictable, but the allosteric binding site of Compound 20 is not easily predictable since there are no compounds with the same mode-of-action in the training set. Using the domain fishing method for Compound 19, the six IPR domains predicted in order of statistical significance are: 1) Integrins alpha chain; 2) Integrin alpha chain, C-terminal cytoplasmic region; 3) FG-GAP; 4) Integrin alpha beta-propellor; 5) Integrin alpha-2; and 6) von Willebrand factor, type A. Given these 6 domains, a Cumulative Domain Score for all known proteins is used to rank them as likely targets. The Cumulative Domain Score is the sum of the Bayes scores for every occurrence of the predicted domains found in a given protein; finding the protein with the highest Cumulative Domain Score yields a target prediction *via* domain fishing. As expected for compound 19, the most

Table 1. Target Prediction *via* Probabilistic Association of Chemical Features with InterPro Domains in Proteins

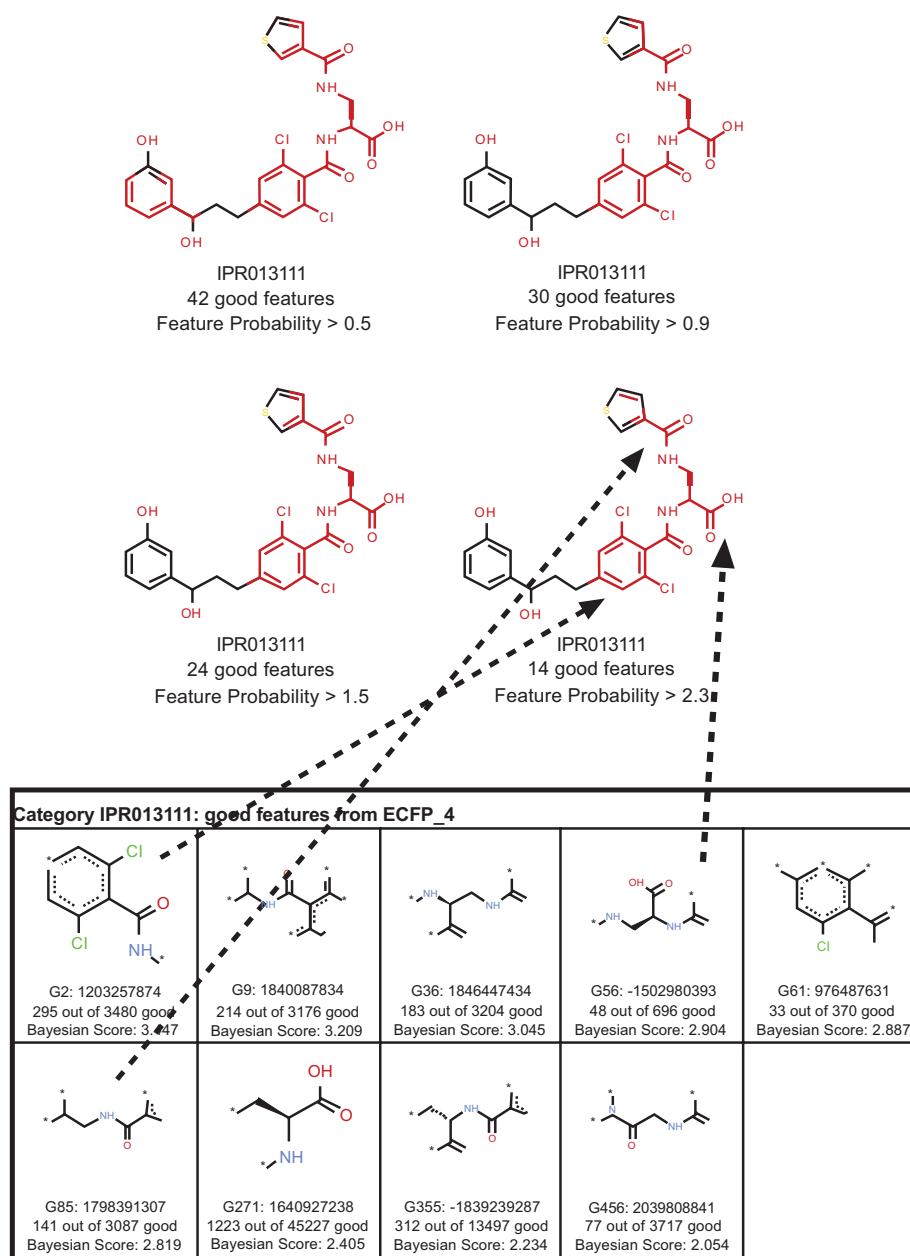
Cpd. Nr.		Known Target Binding Site	Domains Predicted <sup>2</sup>	Bayes Score	Predicted Domain Descriptions	Highest Scoring Target Extrapolated	Extrapolated Target Description	Number Predicted Domains Found in Target	Cumulative Domain Score	Extrapolated Target in Training Set?
19		alpha subunit of LFA-1 in I-domain	1) IPR000413 2) IPR013513 3) IPR013517 4) IPR013519 5) IPR013649 6) IPR002035	1) 43.99 2) 43.99 3) 43.99 4) 43.99 5) 43.99 6) 41.11	1) Integrins alpha chain 2) Integrin alpha chain, C-terminal cytoplasmic region 3) FG-GAP 4) Integrin alpha beta-propellor 5) Integrin alpha-2 6) von Willebrand factor, type A	P20701	Integrin alpha-L precursor (Leukocyte adhesion glycoprotein LFA-1 alpha chain)	6	261.06	YES
20		Between beta and alpha subunits of LFA-1 in I-like domain	1) IPR013111 2) IPR002035 3) IPR012896 4) IPR012012 5) IPR003659 6) IPR002369 7) IPR001169	1) 23.37 2) 22.02 3) 21.69 4) 21.69 5) 21.69 6) 21.69 7) 21.69	1) EGF, extracellular 2) von Willebrand factor, type A 3) Integrin beta tail 4) Integrin, beta subunit 5) Plexin / semaphorin / Integrin 6) Integrin, beta chain N-terminal 7) Integrin beta, C-terminal	P05107	Integrin beta-2 precursor (Cell surface adhesion glycoproteins LFA-1/CR3/p150,95 subunit beta) (Complement receptor C3 subunit beta) (CD18 antigen)	7	153.84	NO

Compound IDs are in line with the original publication. For compound 19, the correct target is predicted *via* the prediction of its interaction domains. Ligands of this target have been known before, so in principle also a model based on ligand-target pairings could have been used. For compound 20, the correct target is predicted *without* having the target in the training set. This extrapolation is achieved by generating models on *protein folds* instead of individual targets

probable target is the  $\alpha$  chain of LFA-1. In contrast, for compound **20** the seven IPR domains predicted in order of statistical significance are: 1) EGF, extracellular; 2) von Willebrand factor, type A; 3) Integrin beta tail; 4) Integrin, beta subunit; 5) Plexin/semaphorin/integrin 6) Integrin beta, C-terminal; 7) Integrin, beta chain N-terminal. On the basis of the Cumulative Domain Score, the most probable target for Compound **20** is: Integrin beta-2 precursor (a.k.a. cell surface adhesion glycoproteins LFA-1/CR3/p150,95 subunit beta). These results demonstrate that the domain fishing method is able to correctly pull out a protein target—or polypeptide chain in this case—that is outside of the training set. How is this possible? As described in Fig. 2, the source

for this leap in predictive ability is that other targets in WOMBAT contain some of the domains found in the LFA-1  $\beta$  chain. For example, chemical features found in ligands from other genetically related target classes (integrins a4b7, a4b1/VLA-4, a5b1, a5b6, a5b5, and a5b3) are contributing to the Bayesian models for IPR domains found in the "correct" protein target, LFA-1  $\beta$  chain.

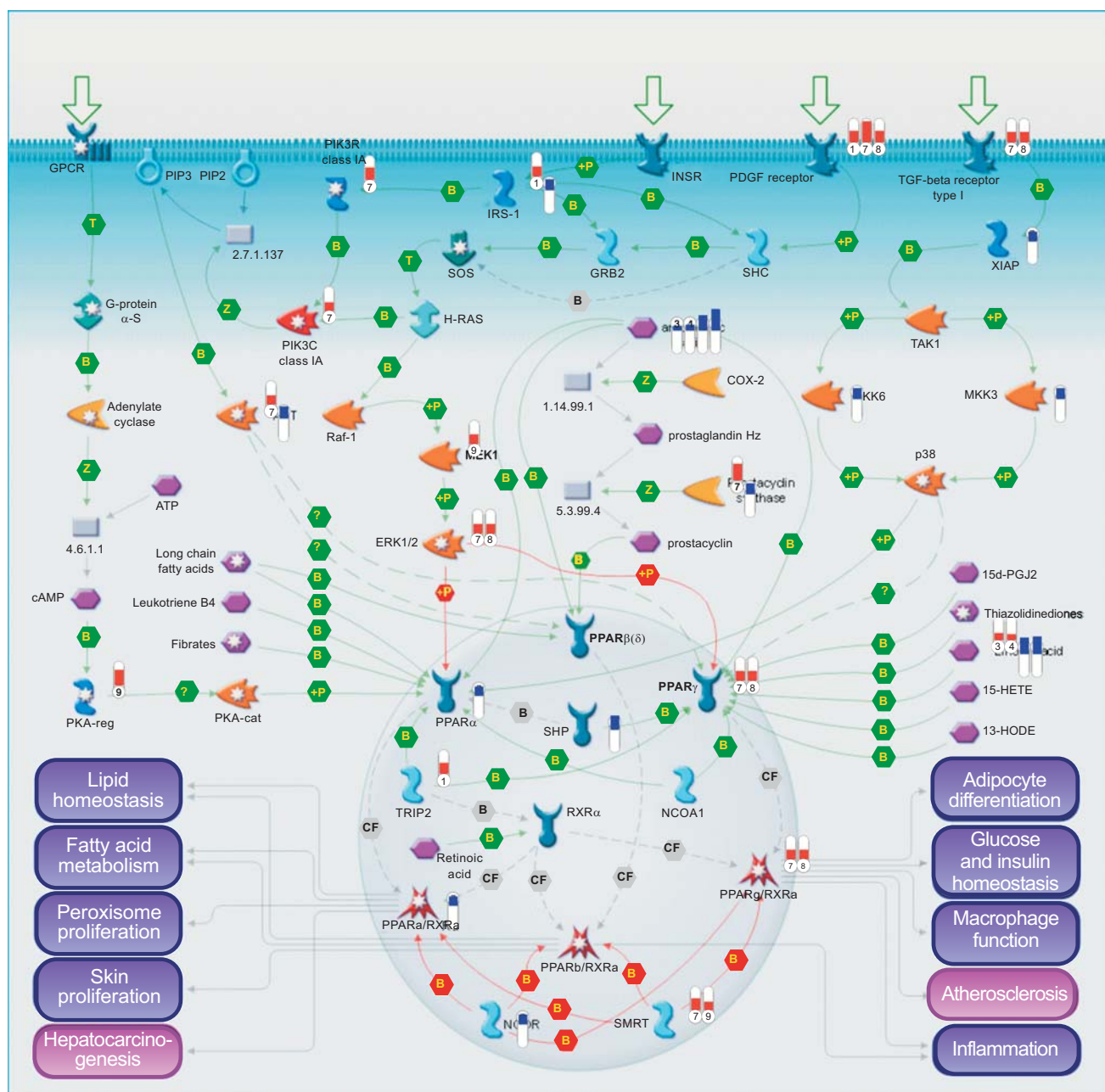
Because the Bayesian models described above consist of substructural features (extended connectivity fingerprints [29]) with computed probabilities, the actual chemical features associated with IPRs can be "back-projected" onto any chemical structure of interest. Fig. 3 highlights in compound **20** the chemical features that are statistically associated with



**Fig. (3).** Back-projection of substructural features associated with binding to Interpro domains. Knowledge about features positively contributing to binding can be used in a variety of settings, one of them, in line with current chemogenomics principles, being the generation of small molecule libraries targeting particular protein folds.

IPR013111, "EGF, extracellular" domain (shown in red). Note that because extended connectivity fingerprints are "circular", there may be sub-features nested inside larger super-features. Compound **20** is shown multiple times, with increasing thresholds of probability for feature correlations. Above a feature probability of 0.5, the Bayesian model for IPR013111 finds more chemical structures to highlight than above a more stringent feature probability threshold of 2.3. Feature probability is calculated as described previously [16]. This approach to chemical image rendering immediately suggests the portions of the molecule that are critical for EGF activity across multiple targets (in red), and conversely, suggests the uncolored moieties may confer selectivity among EGF-containing targets. This approach is applicable for any InterPro domain.

GeneGO (St. Joseph, MI) has taken target prediction one step further in their MetaDrug software with a "systems pharmacology" approach. The software provides a list of predicted targets for an input compound structure based on 2D chemical similarity of the compound or its predicted metabolites to a high-quality bioactivity database. The predicted targets are then linked to curated biological pathways and processes to create a theoretical network map for the compound (Fig. 4). The advantage of this approach is that biological pathways or processes found in common among the list of predicted targets will increase confidence in the target predictions, particularly if the pathways are related to a phenotype known to be induced by the compound.



**Fig. (4).** Illustration of GeneGO pathway maps which can be used to understand potential effects of either predicted or known targets of drugs. Shown here are the different ways in which peroxisome proliferator-activated receptors (PPARs) can be activated as well as known downstream effects of this event. (Image courtesy of Julie Bryant and Ally Perlina, GeneGO, St. Joseph, Michigan).

## ADVERSE DRUG REACTION PREDICTION

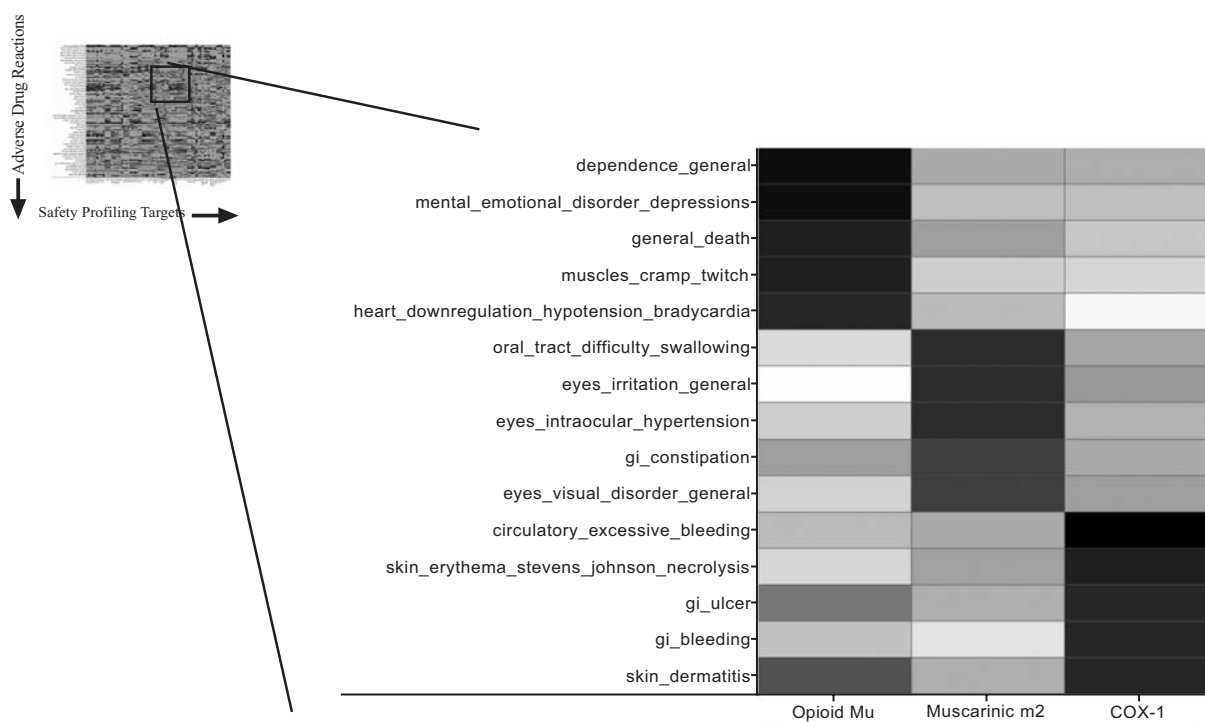
Adverse drug reactions are often detected very late in clinical stages, so early “red flag” approaches can be of tremendous value. Apart from the prediction of targets, models can also be generated for adverse drug reactions (ADRs) based purely on chemical structure [10]. This is possible because, like target-annotated chemical databases, ADR-annotated chemical databases are available, such as the World Drug Index (Elsevier). Bayesian models can thus be trained on all chemicals that cause the same ADR, independent of the exact mechanism causing the effect, and statistically-associated chemical substructures can be extracted. Further, by calculating similarities between the models for a particular target activity and an adverse reaction, likely links between the compound target- and the compound effect worlds can be established. This approach is illustrated in Fig. 5, for the  $\mu$  opioid and muscarinic M2 receptors as well as cyclooxygenase-1. For both activities against targets as well as adverse reactions statistical models are generated on the basis of the chemical structure of compounds. Importantly, the target- and ADR-annotated databases can be different sources altogether; since both models share the common language of “molecular structure” they can be compared by correlation and distance measures between the models – giving indications which targets might be involved in which kind of adverse reaction. On the other hand, target proteins highly correlated with adverse reactions one is interested in can be used to augment target panels used in safety profiling.

The prediction not only of targets, but also of any phenomenological effects, is likely to gain importance and wider use in the future, as are possibilities to infer putative links between them.

## ACTIVITY SPECTRA AND THEIR APPLICATION TO TARGET PREDICTIONS

The activities of a compound across a series of biological readouts such as the affinity to a series of targets, or expression profiles of gene or protein microarrays, can also be seen as molecular descriptors. In many ways, such descriptors are even more relevant than those based purely on chemical structure: Affinities to a series of targets take measured perturbations of biological systems into account, including determinants of affinity that are very difficult to model (as recent literature on docking studies shows [30]). Several names for the description of a molecule by a series of experimental bioactivity data have been coined, such as “affinity fingerprints” [17], “chemical genomic profiles” [31], “chemical-genetic fingerprints” [32], “bioactivity spectra” [33], or just “biospectra” [34]. Each of these compound representations can be used in two directions: to predict targets (or phenotypes) for compounds, as well as to predict compounds active against targets (*e.g.*, for virtual screening purposes where experimental data are used as molecular descriptors).

The history of predicting one molecular property as a function of its other properties is not new, and one of the

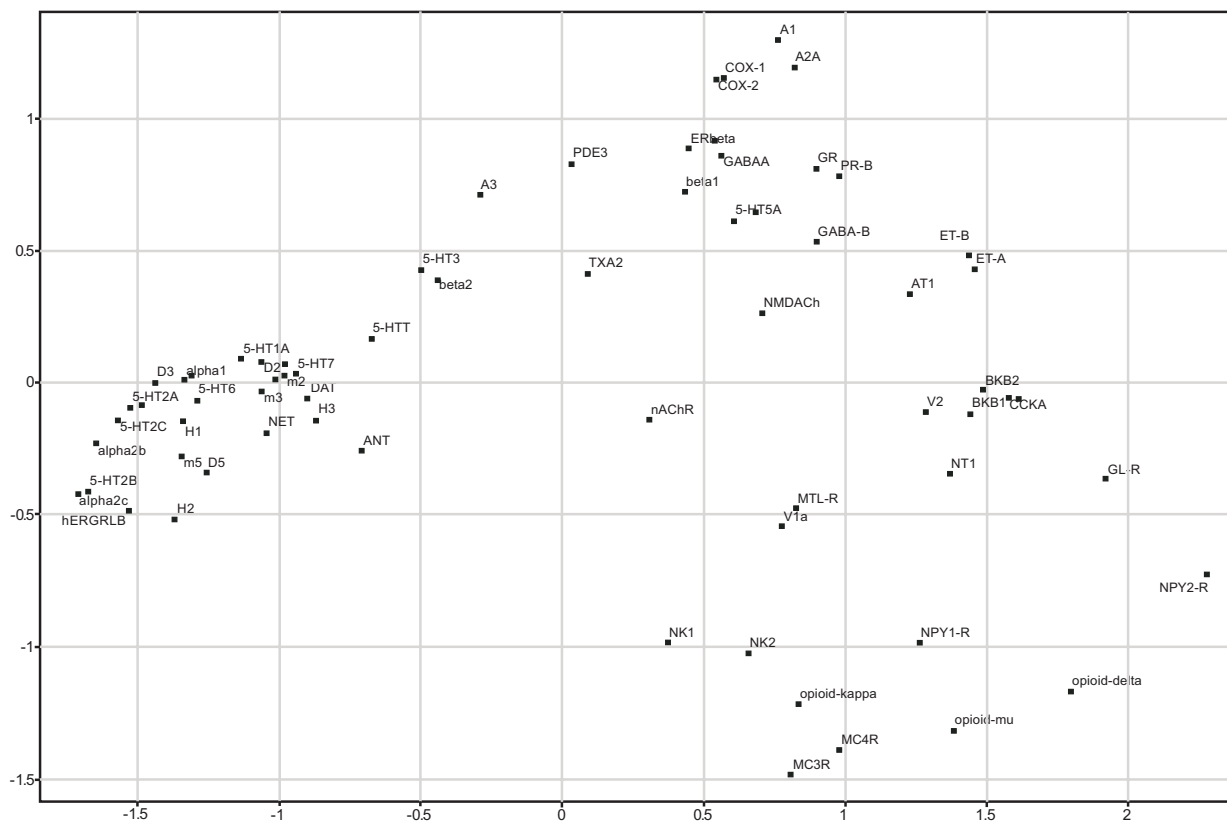


**Fig. (5).** Predictions of targets and predictions of adverse effects can be combined *via* the two statistical models based on molecular features, thereby merging side effect and molecular target space. The side effect space can be replaced by any type of observation or phenotypic space, thus linking empirical “observation space” as well as molecular “mode of action space”. (Bender *et al.*, Analysis of Pharmacology Data and the Prediction of Adverse Drug Reactions and Off-Target Effect From Chemical Structure, ChemMedChem, 2007, 2, 861. Copyright 2007 John Wiley & Sons Ltd.).

seminal works in this area are the “affinity fingerprints” proposed by Kauvar [17]. In this work, the affinity against an information-optimized panel of 18 receptors was used to characterize compounds, and this profile could then later be used to establish the similarity of molecular structures *via* the similarities of their activity profiles. This work was later extended to include “virtual affinity fingerprints” [35], based on interaction energies derived from docking of the ligands to a receptor panel. Acknowledging that docking is a computationally rather demanding, but yet not very reliable way to investigate ligand-target interactions (“no statistically significant relationship existed between docking scores and ligand affinity”, to quote from a recent comparative study involving 10 docking programs and 37 scoring functions on a panel of eight proteins [30]), some of the authors of this article sought a better value for their computational expense, leading to a small molecule analogue of affinity fingerprints termed “Bayes Affinity Fingerprints” [8]. In this case, one small molecule is compared to the features contained in Bayes models for a large number of activity classes ( $10^3$ ); subsequently, the profile of scores from the Bayes models for that compound can be used as a descriptor for traditional activities like database searching, clustering, etc. In addition, correlations between classes can be calculated, thus enabling visualization of activity class similarity in chemical space (Fig. 6). This kind of knowledge can be analyzed in a variety of directions, such as the prediction of secondary targets or

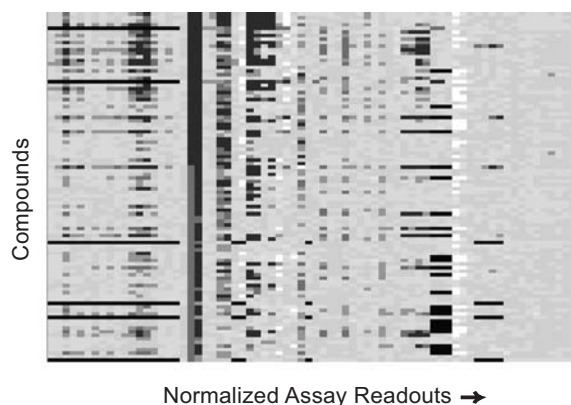
for the design of target-family focused screening libraries.

Similar ideas were pursued by Paolini *et al.* [36] and Keiser *et al.* [9], in the sense that the relationships between targets, based on small molecules interacting with them, were established. Paolini *et al.* [36] compiled a comprehensive database of bioactive molecules, spanning 2,876 targets (1,306 proteins) with a total of 276,122 active ligands and over 600,000 data points linking compounds and targets. Based on shared interactions of ligands with multiple targets, the authors were able to map “global pharmacology space” – and to establish properties such as target promiscuity indices, promiscuity of gene families, and also target- as well as time-resolved physicochemical properties of bioactive compounds. Keiser *et al.* [9] started on a smaller dataset, derived from the MDDR database, and used ligand data from 246 enzymes and receptors to relate all targets with each other. Introducing novel methods to establish the significance of the similarity of two groups of compounds (by applying cutoff scores), these authors also experimentally showed that the similarity of groups of ligands can be used in a prospective manner to elucidate yet unknown bioactivities. Here, methadone, emetine and loperamide were tested for antagonistic properties against the muscarinic M3, the adrenergic  $\alpha 2$  and the neurokinin NK2 receptors, respectively, and the predictions made *in silico* were validated experimentally.



**Fig. (6).** By calculating similarities not only between individual molecules but between the models generated from a set of molecules, the similarity of each part of chemical space occupied can be calculated, predicting properties such as inherent side-effect liability of activity classes. (Reproduced with permission from J. Chem. Inf. Model. 2006, 46, 2445-2456. Copyright 2006 American Chemical Society).

As another example of chemogenomic profiles, databases such as ChemBank [37, 38] have been rapidly growing over the past few years. ChemBank provides a single point of access to public screening data across a wide variety of assays, which can in principle be used to describe compounds by their associated activity profiles. At present, one can make a compound selection in the user interface and visualize and download such assay performance profiles (Fig. 7); additional downloads for these compounds might include compound structures, names, or molecular descriptors. Such data provide a wealth of information to analyze relationships between structure and activity, such as what changes to a molecular structure are more relevant to one class of assay vs another class. In the future, ChemBank will incorporate additional analysis and visualization features to provide cheminformatics functionality directly to users. In the spirit of the work by Kauvar [17], such a future ability to identify similar activity profiles directly within the database will afford a new breadth of biology within phenotypic fingerprints, including binding and enzyme-inhibition assays, cell-based assays (including high-content screens, discussed below), and other new screening technologies [39].



**Fig. (7).** View of the public screening data accessible *via* ChemBank. Compounds are listed along the vertical axis, with normalized assay results given along the horizontal axis (the brighter the area the more active the compound; about 10% of the black area also contains compounds not tested against the particular target). Besides the individual results this representation is of interest to ongoing chemogenomics efforts since the assay readouts can also be used as one form of small molecule affinity fingerprints.

## PROPERTY PREDICTION USING “BIOLOGICAL DESCRIPTORS”

### Gene Expression Data as Descriptors – The “Connectivity Map”

In the simplest scenario, compound activity refers to inhibition or protein binding measurements, but this scenario does not take the intracellular (or intercellular) signaling networks into account. One increasingly popular possibility to take such interdependencies into account is to describe the effect of a compound on a cell, for example by measuring gene expression (mRNAs) and its changes upon compound administration [40]. One of the pioneering works in the area was performed by Covell *et al.* [41] at the National Cancer Institute. For the NCI cancer screening data, Covell corre-

lated the cytotoxic response for certain areas of bioactivity space with gene expression profiles, which were used for the prediction of modes of action for novel chemicals, based on gene expression profiles.

In more recent work, termed the “Connectivity Map” [42], mRNA expression patterns were used as perturbation signatures of biological systems caused by the addition of small molecule perturbants. In this case, 164 compounds representing drugs and other small molecule modulators were chosen, of which some shared molecular targets (such as histone deacetylase, HDAC) while others shared the same clinical indication (such as antidiabetic compounds). Included were both obvious gene-expression modulators (such as nuclear hormone receptor ligands) as well as compounds with functions not obviously related to gene expression. Although hierarchical clustering approaches are often used to make sense of the resulting gene expression data, in this case several factors made this less attractive, one of which being that cell type and batch effects masked weaker perturbations introduced by small-molecule modulation of cellular system. Instead, a rank-based similarity algorithm was used which employs only information whether a gene is upregulated or downregulated between a query signature and the database signatures, giving a real-valued “connectivity score” between +1 and -1 between query signature and database signatures. One aspect of the Connectivity Map linked to the theme of this review is the prediction of modes-of-action of small molecules. Gene-expression signatures for different scaffolds of HDAC inhibitors in different cell types were overall similar, facilitating the prediction of targets for novel compounds. Another example focused on estrogen receptor agonists and antagonists, where agonists and antagonists were found to produce opposite gene expression signatures. The application of phenothiazine antipsychotics was found to correlate negatively with prostaglandin synthesis, which is consistent with the recent discovery of phenothiazines as potent COX/LOX inhibitors. A prospective mode of action analysis was performed using the Connectivity Map on gedunin, a natural product known to abrogate androgen receptor activation in prostate cancer cells, but with an unknown mechanism. *Via* gene expression profiling and querying of the Connectivity Map database, HSP90 was suggested as a target, consistent with the finding that stability of the androgen receptor is dependent on HSP90 activity.

Not only structure-based signatures can be used for querying the gene expression database, but also those derived from disease states. By using an obesity gene expression signature published earlier, it was found that PPAR $\gamma$  agonists received high connectivity scores, and all such compounds are indeed inducers of adipogenesis *in vitro*. For signatures derived from a comparison between Alzheimer disease and normal tissue, querying the Connectivity Map yielded two instances of 4,5-dianilinophthalimide with negatively correlated gene expression behavior. This compound was recently identified in a cell-free screen as reversing the formation of fibrils thought to be associated with the progression of Alzheimer’s disease. Additional applications of this technology are possible for resistance against drugs in cancer chemotherapy: by defining a gene expression signature between children having acute lymphoblastic leukemia which were not resistant to dexamethasone, the most highly correlated expression profile was the one related to the mTOR inhibitor

rapamycin. Indeed, rapamycin was able to increase dexamethasone sensitivity by a factor of more than 50.

### HIGH-CONTENT SCREENING DATA AS DESCRIPTORS

A complementary set of biological descriptors that can be used to measure the effect of compounds on living cells are the readouts of high-content screening (HCS) campaigns [43-45]. High content screening combines automated fluorescence microscopy with state-of-the-art image processing and quantification to generate a biological fingerprint that is based on the quantity, activity, and organization of biomolecules within the spatial context of the cellular milieu (Fig. 8). Employed in a lead discovery and validation pipeline, high content imaging enables screening and profiling compounds based on a rich set of phenotypic information obtained at a single-cell level. Several commercial and open-access high-content packages are now available to enable microscopic image acquisition, processing, and quantification (see [46-48] for recent reviews). Quantitative approaches for defining biological activity space based on high-content image data represent an area of active research that has been driven primarily by the end-user and motivated by the discovery goals of each particular assay.

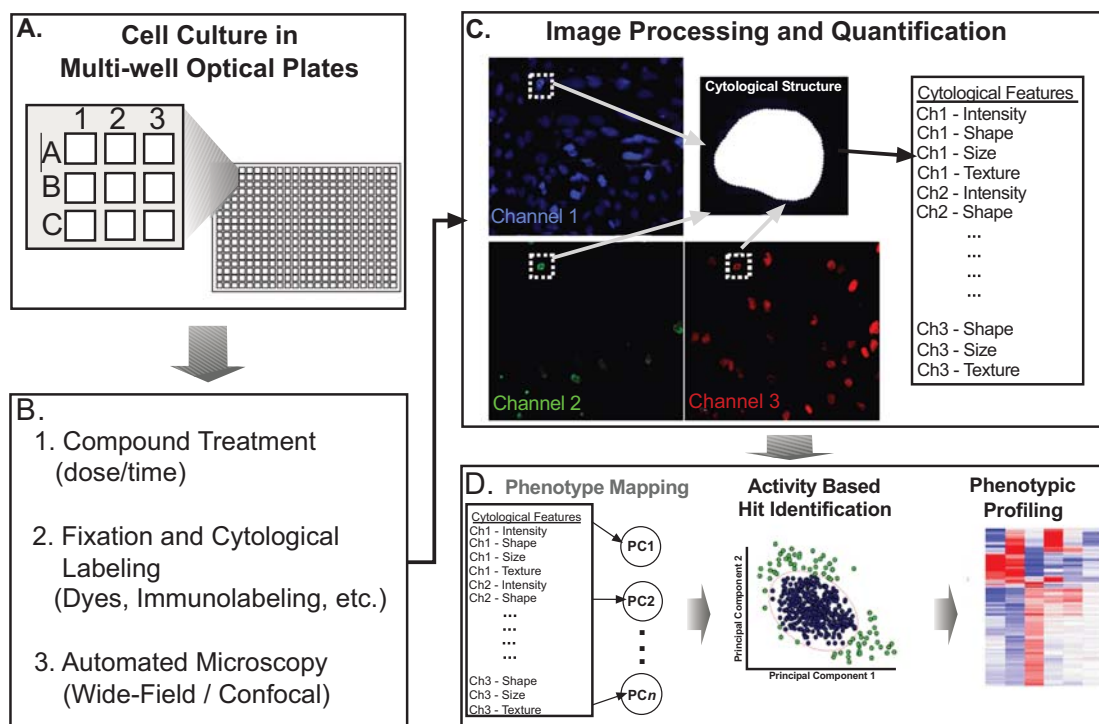
On the other hand, high content assays provide an alternative approach to probing the biological activity of small-molecules in distinct signaling pathways *via* quantification of protein translocation, or receptor internalization [49-51]. Recent studies have taken full advantage of the image quantification capabilities to examine small-molecule activity on more complex phenotypes defined by unbiased, or minimally biased, morphological descriptors of cell structure [52, 53]. Such unbiased approaches can lead not only to the discovery of biologically active compounds with insight on mechanism of action, but also facilitate the discovery of unexpected cellular phenotypes that, for example, could be indicative of cellular toxicity. Mitchison and colleagues recently demonstrated the potential for “hypothesis-free” automated microscopy to generate small-molecule structure activity relationships and mechanism of action inferences [43, 52]. The authors designed a microscopic image-based compound profiling scheme in which multiple distinct cellular components and processes were probed in parallel. Using fluorescent stains and indirect immunofluorescence they quantified more than 90 cytological features in HeLa Cells that described nuclear and cytoskeletal structure, signaling pathway activity, and transcription factor localization on a single cell level. For each cytological feature they used the Kolmogorov-Smirnov (KS) statistic to quantify small-molecule effects, and developed a titration independent similarity score (TISS) to compare KS dose-response profiles among a collection of bioactive drugs with known mechanisms of action. Unsupervised clustering based on TISS values revealed that drugs with common targets grouped together, and enabled the assignment of biological mechanisms of action to blinded drug samples. This work revealed that multidimensional sets of microscopic cytological features can be used to profile compounds and infer mechanisms of action.

In similar work Tanaka *et al.* [53], used an unbiased image-based cell-morphology screen to profile a small panel of 107 kinase inhibitor scaffolds in a dose-response series

across four cancer cell lines and one normal diploid cell line. Their biological fingerprint consisted of morphological descriptors, staining intensities, and spatial properties of nuclei, cytoskeletal proteins, and the Golgi apparatus. Principal components analysis was employed to reduce the dimensionality of the biological fingerprints across the panel of cell lines down to three dimensions. Principal components scores were then used to visualize trajectories of dose-dependent phenotypic change for the panel of compounds. Known compounds provided phenotypic landmarks in the principal component space and, three distinct phenotypes for bioactive compounds could be observed. Tanaka *et al.*, focused on an outlier phenotype that was produced by a single compound of the pyrazolopyrimidine class, Hydroxy-PP. The authors went on to use compound affinity chromatography and mass spectroscopy with Hydroxy-PP to look for a target; combining this with co-crystallization studies they identified and validated CBR1 a dehydrogenase-reductase, as a direct molecular target. The discovery of Hydroxy-PP biological activity was the result of the emergence of an unexpected phenotype that would not have been detectable in a single readout cell-based assay, and thus highlights a key benefit of the image based biological fingerprint.

More recently, MacDonald *et al.* [54], reported on a high-content assay where they monitored signal transduction pathways by detecting protein-protein interactions within the spatial cellular context. Here, the assay employed a previously described protein complementation assay (PCA) [55], where cells are engineered to simultaneously express two proteins – from a common signaling pathway – that are fused to complementary fragments of a fluorescent protein reporter. When the two proteins physically interact they bring together the complementary fragments and generate a fluorescent report. The group constructed 49 such PCAs that probed a diverse set of cellular pathways (for example, cell cycle, ubiquitin proteolysis, and stress response), and monitored the activity and intracellular localization of those pathways independently by automated microscopy in response to 107 different bioactive compounds. Hierarchical clustering of compounds based on the pathway activity biological fingerprints successfully recapitulated known structure-activity relationships. For example, proteasome inhibitors, ALLN and MG132 clustered together. Notably, unexpected phenotypes did emerge for several known drugs, such as the depression medication Zoloft™ (sertraline) that exhibited antiproliferative activity. Follow-up work showed that this antiproliferative phenotype of sertraline was observable with EC50 values in the micromolar range on a panel of five tumor cell lines.

As high-content imaging becomes increasingly more accessible and user friendly new applications are emerging in both pharma and academia. New analysis and phenotypic mapping strategies that facilitate integration of biological fingerprints directly with chemoinformatics structure and target knowledge databases will be useful for the global profiling and annotation of compound libraries in the drug discovery setting. High-content screens using time-lapse imaging in live cells will enable compound dose-response analysis in space and time on a single cell level. Ellenberg and colleagues have already employed such kinetic cytological assays in an siRNA-based genetic screen where morphological phenotypic fingerprints were resolved over time in living cells to profile



**Fig. (8).** (A) In the basic HCS scheme cells are cultured, typically in 96- or 384-well optical plates, under a predetermined set of treatment vs control conditions. (B) Cells in culture are treated with compounds in an automated or semi-automated fashion, and incubated at a dose and time appropriate for the biology under investigation. The biological assay generally incorporates a fluorescent labeling strategy aimed at quantification of protein or nucleic acid concentration, organization, and/or localization using fluorescent dyes, fluorescent proteins, or indirect immunofluorescence. Automated fluorescence microscopy (wide-field or confocal) is performed and multiple images are acquired per fluorescent channel and experimental condition on a well-by-well basis. (C) Image processing and quantification is either carried out subsequently or “on-the-fly”, and involves two main steps: object detection and parameter quantification. Object detection refers to the identification of cellular structures of interest within each image, and is facilitated by a series of image processing and segmentation steps. Examples of such structures include cell bodies, nuclei, vesicles, and cytoskeleton. Parameter quantification refers to the measurement and calculation of features that characterize entities of interest, examples include signal intensity, size, morphology and texture. (D) Cytological features can be used to define a biological activity fingerprint directly, or can be mapped using principal components analysis to a phenotype space of reduced dimensionality. Changes in cytological features, or principal components, can be quantified to identify biologically active compounds. Supervised and unsupervised clustering strategies can be employed to explore structure-activity relationships, and make inference about compound mechanisms of action.

genes that affect cell growth and survival [56]. Importantly, high-content imaging provides an efficient, easy to implement method that is orthogonal to transcript profiling and single readout cell-based assays to generate phenotypic profiles of compounds and provide a quantitative framework for mechanism of action inferences in biological activity space.

Bioactivity spectra have also shown great promise with respect to mining pharmacology data and predicting ADRs as well as primary targets, shown for example by using Cerep’s BioPrint database [57]. In this case, targets as well as ADRs can be predicted for a compound on the basis of its “profile similarity” to other compounds, where the profile is determined in a panel of about 80 ligand-binding assays [57]. “Biological spectra analysis” has recently been found to provide an unbiased means to cluster compounds without introducing a bias by the choice of particular descriptors [58, 59]. Using a dataset of 1,567 compounds measured against 92 targets, clustering by this approach was found to be reasonable, and in some cases similar bioactivities of compounds were grouped (as well as predicted) despite different underlying scaffolds.

More recently, this analysis was extended to an analysis (and, possibly, prediction) of adverse drug reactions [60].

## CONCLUSIONS

This review presented a cross-section of what chemogenomics is able to address today, focused on two areas: target prediction of small molecules with molecular descriptor-derived models, and the increasing application of biological readouts as more complex descriptors of biological systems (which can in turn be used to predict modes of action of a novel compound). Target prediction can elucidate on-target as well as off-target effects, as demonstrated by analysis of hit lists from reporter gene assays in order to identify false-positive readouts. In recent years, a shift from structure-derived chemical descriptors to “biological descriptors” is evident in the literature. Biological descriptors can be more complex in nature, as they can be related more intimately with the networked architecture of cells found in living systems. This review provided a brief overview of these descriptors, focusing on gene expression profiles and high-content screening data as well as their application. These recent advances

suggest that we are currently at a crossroad in the drug discovery process, where single-target bioassay results are supplanted by biological fingerprints from multiple targets to reflect our new awareness of polypharmacology. Chemogenomics concepts and data-analysis methods will play an increasing role in interpreting the increasing quantity and availability of multidimensional activity data.

## ACKNOWLEDGEMENT

AB and DWY thank the Education Office of Novartis for a Postdoctoral Fellowship.

## REFERENCES

- [1] Bredel, M.; Jacoby, E. *Nat. Rev. Genet.*, **2004**, *5*, 262.
- [2] Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 391.
- [3] Schuffenhauer, A.; Zimmermann, J.; Stoop, R.; van der Vyver, J.J.; Lecchini, S.; Jacoby, E. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 947.
- [4] Schreiber, S.L. *Bioorg. Med. Chem.*, **1998**, *6*, 1127.
- [5] Spring, D.R. *Chem. Soc. Rev.*, **2005**, *34*, 472.
- [6] Wagner, B.K.; Haggarty, S.J.; Clemons, P.A. *Am. J. Pharmacogenomics*, **2004**, *4*, 313.
- [7] Mitchell, J.B.O. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1617.
- [8] Bender, A.; Jenkins, J.L.; Glick, M.; Deng, Z.; Nettles, J.H.; Davies, J.W. *J. Chem. Inf. Model.*, **2006**, *46*, 2445.
- [9] Keiser, M.J.; Roth, B.L.; Armbruster, B.N.; Ernsterberger, P.; Irwin, J.J.; Shoichet, B.K. *Nat. Biotechnol.*, **2007**, *25*, 197.
- [10] Jenkins, J.L.; Bender, A.; Davies, J.W. *Drug Discov. Today Technol.*, **2007**, *3*, 413.
- [11] Hampton, T. *JAMA*, **2004**, *292*, 419.
- [12] Bender, A.; Scheiber, J.; Glick, M.; Davies, J.W.; Azzaoui, K.; Hamon, J.; Urban, L.; Whitebread, S.; Jenkins, J.L. *ChemMedChem*, **2007**, *2*, 861.
- [13] Azzaoui, K.; Hamon, J.; Faller, B.; Whitebread, S.; Jacoby, E.; Bender, A.; Jenkins, J.L.; Urban, L. *ChemMedChem*, **2007**, *2*, 874.
- [14] Csermely, P.; Agoston, V.; Pongor, S. *Trends Pharmacol. Sci.*, **2005**, *26*, 178.
- [15] Hart, C.P. *Drug Discov. Today*, **2005**, *10*, 513.
- [16] Nidhi; Glick, M.; Davies, J.W.; Jenkins, J.L. *J. Chem. Inf. Model.*, **2006**, *46*, 1124.
- [17] Kauvar, L.M.; Higgins, D.L.; Villar, H.O.; Sportsman, J.R.; Engqvist-Goldstein, A.; Bukar, R.; Bauer, K.E.; Dilley, H.; Rocke, D.M. *Chem. Biol.*, **1995**, *2*, 107.
- [18] Bender, A.; Glen, R.C. *Org. Biomol. Chem.*, **2004**, *2*, 3204.
- [19] Glen, R.C.; Bender, A.; Amby, C.H.; Carlsson, L.; Boyer, S.; Smith, J. *IDrugs*, **2006**, *9*, 199.
- [20] Nettles, J.H.; Jenkins, J.L.; Bender, A.; Deng, Z.; Davies, J.W.; Glick, M. *J. Med. Chem.*, **2006**, *49*, 6802.
- [21] Poroikov, V.V.; Filimonov, D.A.; Borodina, Y.V.; Lagunin, A.A.; Kos, A. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 1349.
- [22] Crisman, T.J.; Parker, C.N.; Jenkins, J.L.; Scheiber, J.; Thoma, M.; Kang, Z.B.; Kim, R.; Bender, A.; Nettles, J.H.; Davies, J.W.; Glick, M. *J. Chem. Inf. Model.*, **2007**, *47*, 1319-1327.
- [23] Strombergsson, H.; Kryshatovych, A.; Prusis, P.; Fidelis, K.; Wikberg, J.E.; Komorowski, J.; Hvidsten, T.R. *Proteins*, **2006**, *65*, 568.
- [24] Strombergsson, H.; Prusis, P.; Midelfart, H.; Lapinsh, M.; Wikberg, J.E.; Komorowski, J. *Proteins*, **2006**, *63*, 24.
- [25] Snyder, K.A.; Feldman, H.J.; Dumontier, M.; Salama, J.J.; Hogue, C.W.V. *BMC Bioinformatics*, **2006**, *7*.
- [26] World of Biomolecular Activity (WOMBAT), SunSet Molecular Discovery LLC, <http://www.sunsetmolecular.com>.
- [27] Mulder, N.J.; Apweiler, R.; Attwood, T.K.; Bairoch, A.; Bateman, A.; Binns, D.; Bradley, P.; Bork, P.; Bucher, P.; Cerutti, L.; Copley, R.; Courcelle, E.; Das, U.; Durbin, R.; Fleischmann, W.; Gough, J.; Haft, D.; Harte, N.; Hulo, N.; Kahn, D.; Kanapin, A.; Krestyaninova, M.; Lonsdale, D.; Lopez, R.; Letunic, I.; Madera, M.; Maslen, J.; McDowall, J.; Mitchell, A.; Nikolskaya, A.N.; Orchard, S.; Pagni, M.; Ponting, C.P.; Quevillon, E.; Selengut, J.; Sigrist, C.J.; Silventoinen, V.; Stud-
- [28] holme, D.J.; Vaughan, R.; Wu, C.H. *Nucleic Acids Res.*, **2005**, *33*, D201.
- [29] Arkin, M.R.; Wells, J.A. *Nat. Rev. Drug Discov.*, **2004**, *3*, 301.
- [30] SciTegic, Inc., San Diego, CA. - <http://www.scitegic.com>.
- [31] Warren, G.L.; Andrews, C.W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M.H.; Lindvall, M.K.; Nevins, N.; Semus, S.F.; Senger, S.; Tedesco, G.; Wall, I.D.; Woolven, J.M.; Peishoff, C.E.; Head, M.S. *J. Med. Chem.*, **2006**, *49*, 5912.
- [32] Haggarty, S.J.; Clemons, P.A.; Schreiber, S.L. *J. Am. Chem. Soc.*, **2003**, *125*, 10543.
- [33] Kim, Y.K.; Arai, M.A.; Arai, T.; Lamenza, J.O.; Dean, E.F., 3rd; Patterson, N.; Clemons, P.A.; Schreiber, S.L. *J. Am. Chem. Soc.*, **2004**, *126*, 14740.
- [34] Filimonov, D.A.; Poroikov, V.V.; Karaicheva, E.I. *Exp. Clin. Pharmacol. (Rus)*, **1995**, *58*, 56.
- [35] Fliri, A.F.; Loging, W.T.; Thadeio, P.F.; Volkmann, R.A. *J. Med. Chem.*, **2005**, *48*, 6918.
- [36] Briem, H. *Perspect. Drug Discov. Des.*, **2000**, *20*, 231.
- [37] Paolini, G.V.; Shapland, R.H.; van Hoom, W.P.; Mason, J.S.; Hopkins, A.L. *Nat. Biotechnol.*, **2006**, *24*, 805.
- [38] Tolliday, N.; Clemons, P.A.; Ferraiolo, P.; Koehler, A.N.; Lewis, T.A.; Li, X.; Schreiber, S.L.; Gerhard, D.S.; Eliasof, S. *Cancer Res.*, **2006**, *66*, 8935.
- [39] ChemBank. <http://chembank.broad.harvard.edu/>.
- [40] Clemons, P.A. *Curr. Opin. Chem. Biol.*, **2004**, *8*, 334.
- [41] McInnes, C.; Fischer, P.M. *Curr. Pharm. Des.*, **2005**, *11*, 1845.
- [42] Mansfield, M.L.; Covell, D.G.; Jernigan, R.L. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 259.
- [43] Warren, G.L.; Andrews, C.W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M.H.; Lindvall, M.K.; Nevins, N.; Semus, S.F.; Senger, S.; Tedesco, G.; Wall, I.D.; Woolven, J.M.; Peishoff, C.E.; Head, M.S. *J. Med. Chem.*, **2006**, *49*, 5912-5931.
- [44] Mitchison, T.J. *Chembiochem*, **2005**, *6*, 33.
- [45] Taylor, D.L.; Giuliano, K.A. *Drug Discov. Today Technol.*, **2005**, *2*, 149.
- [46] Nichols, A. *Methods Mol. Biol.*, **2007**, *356*, 379.
- [47] Giuliano, K.A.; Haskins, J.R.; Taylor, D.L. *Assay Drug Dev. Technol.*, **2003**, *1*, 565.
- [48] Sun, D.; Chuaqui, C.; Deng, Z.; Bowes, S.; Chin, D.; Singh, J.; Cullen, P.; Hankins, G.; Lee, W.C.; Donnelly, J.; Friedman, J.; Josiah, S. *Chem. Biol. Drug Des.*, **2006**, *67*, 385.
- [49] Carpenter, A.E.; Jones, T.R.; Lamprecht, M.R.; Clarke, C.; Kang, I.H.; Friman, O.; Guertin, D.A.; Chang, J.H.; Lindquist, R.A.; Moffat, J.; Golland, P.; Sabatini, D.M. *Genome Biol.*, **2006**, *7*, R100.
- [50] Raychaudhury, C.; Ray, S.K.; Ghosh, J.J.; Roy, A.B.; Basak, S.C. *J. Comput. Chem.*, **1984**, *5*, 581.
- [51] Granas, C.; Lundholt, B.K.; Heydorn, A.; Linde, V.; Pedersen, H.C.; Krog-Jensen, C.; Rosenkilde, M.M.; Pagliaro, L. *Comb. Chem. High Throughput Screen.*, **2005**, *8*, 301.
- [52] Heydorn, A.; Lundholt, B.K.; Praestegaard, M.; Pagliaro, L. *Methods Enzymol.*, **2006**, *414*, 513.
- [53] Perlman, Z.E.; Slack, M.D.; Feng, Y.; Mitchison, T.J.; Wu, L.F.; Altschuler, S.J. *Science*, **2004**, *306*, 1194.
- [54] Tanaka, H.T.; Ikeda, M.; Chiaki, H. Curvature-based face surface recognition using spherical correlation - Principal directions for curved object recognition. In *Automatic Face and Gesture Recognition - Third IEEE International Conference Proceedings*, I E E E, Computer Soc Press: Los Alamitos, **1998**; pp 372.
- [55] MacDonald, M.L.; Lamerdin, J.; Owens, S.; Keon, B.H.; Bilter, G.K.; Shang, Z.; Huang, Z.; Yu, H.; Dias, J.; Minami, T.; Michnick, S.W.; Westwick, J.K. *Nat. Chem. Biol.*, **2006**, *2*, 329.
- [56] Michnick, S.W.; Remy, I.; Campbell-Valois, F.X.; Vallee-Belisle, A.; Pelletier, J.N. *Methods Enzymol.*, **2000**, *328*, 208.
- [57] Neumann, B.; Held, M.; Liebel, U.; Erfle, H.; Rogers, P.; Pepperkok, R.; Ellenberg, J. *Nat. Methods*, **2006**, *3*, 385.
- [58] Krejcsa, C.M.; Horvath, D.; Rogalski, S.L.; Penzotti, J.E.; Mao, B.; Barbosa, F.; Migeon, J. C. *Curr. Op. Drug Discov. Devel.*, **2003**, *6*, 470.
- [59] Fliri, A.F.; Loging, W.T.; Thadeio, P.F.; Volkmann, R.A. *Proc. Natl. Acad. Sci. USA*, **2005**, *102*, 261.
- [60] Fliri, A.F.; Loging, W.T.; Thadeio, P.F.; Volkmann, R.A. *J. Med. Chem.*, **2005**, *48*, 6918.
- [61] Fliri, A.F.; Loging, W.T.; Thadeio, P.F.; Volkmann, R.A. *Nat. Chem. Biol.*, **2005**, *1*, 389.