Chemogenomics Approaches for Receptor Deorphanization and Extensions of the Chemogenomics Concept to Phenotypic Space

Eelke van der Horst[1], Julio E. Peironcely[2,3], Gerard J. P. van Westen[1], Olaf O. van den Hoven[1], Warren R. J. D. Galloway[4], David R. Spring[4], Joerg K. Wegner[5], Herman W. T. van Vlijmen[5], Ad P. IJzerman[1], John P. Overington[6] and Andreas Bender[7,*]

[1]Division of Medicinal Chemistry, Leiden / Amsterdam Center for Drug Research, Einsteinweg 55, 2333 CC Leiden, The Netherlands; [2]Division of Analytical BioSciences, Leiden / Amsterdam Center for Drug Research, Einsteinweg 55, 2333 CC Leiden, The Netherlands; [3]TNO, Quality of Life, Utrechtseweg 48, 3704 HE Zeist, The Netherlands and Netherlands Metabolomics Centre, Einsteinweg 55, 2333CC Leiden, The Netherlands; [4]Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom; [5]Tibotec BVBA, Generaal De Wittelaan L 11B 3, 2800 Mechelen, Belgium; [6]EMBL Outstation – Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom; [7]Unilever Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

Abstract: Chemogenomic approaches, which link ligand chemistry to bioactivity against targets (and, by extension, to phenotypes) are becoming more and more important due to the increasing number of bioactivity data available both in proprietary databases as well as in the public domain. In this article we review chemogenomics approaches applied in four different domains: Firstly, due to the relationship between protein targets from which an approximate relation between their respective bioactive ligands can be inferred, we investigate the extent to which chemogenomics approaches can be applied to receptor deorphanization. In this case it was found that by using knowledge about active compounds of related proteins, in 93% of all cases enrichment better than random could be obtained. Secondly, we analyze different cheminformatics analysis methods with respect to their behavior in chemogenomics studies, such as subgraph mining and Bayesian models. Thirdly, we illustrate how chemogenomics, in its particular flavor of 'proteochemometrics', can be applied to extrapolate bioactivity predictions from given data points to related targets. Finally, we extend the concept of 'chemogenomics' approaches, relating ligand chemistry to bioactivity against related targets, into phenotypic space which then falls into the area of 'chemical genomics' and 'chemical genetics'; given that this is very often the desired endpoint of approaches in not only the pharmaceutical industry, but also in academic probe discovery, this is often the endpoint the experimental scientist is most interested in.

Keywords: Chemogenomics, proteochemometrics, deorphanization, GPCR, virtual screening, G-protein coupled receptors, orphan receptors, target prediction, mode of action analysis.

INTRODUCTION

The term 'chemogenomics'[1], first coined in 2001[2,3] represents the systematic study of ligand chemistry and protein targets (or generally gene products) they show bioactivity against [4]. While in this kind of study the links between the underlying structure connecting ligand chemical space and biological bioactivity space are of primary interest, the related terms 'chemical genetics' and 'chemical genomics'[5,6] focus more on the ability of small molecule chemistry to modulate biological systems in a specific and directed manner, similar to classical genetic approaches such as knock-out organisms. However, chemical tools can act as both protein activators and inhibitors, and they can be applied at various developmental stages and different concentrations; hence, they are often more flexible to use than their biological counterparts.

In recent years more and more knowledge about the chemistry of bioactive molecules has entered the public domain, in databases such as DrugBank [7], BindingDB [8], PDSP Ki [9] and so on, as well as (partially) cellular assay databases like PubChem Bioassay [10] and Chembank [11](see recent reviews such as [12] and [13] for a more comprehensive overview of these). Most recently, and as one of the biggest efforts of its kind, the European Bioinformatics Institute (EMBL-EBI) via a grant from the Wellcome Trust acquired the rights of the StARlite database previously sold by Inpharmatica, which added in excess of five hundred thousand bioactive chemicals with millions of binding and functional assay data points to the publicly available bioactivity knowledge in the form of the ChEMBL database [14] (discussed in a recent review [15]). Apart from databases linking chemical structures to protein targets, also phenotypic databases such as those capturing side effects of drugs are now becoming publicly available[16], increasing the amount of data available for chemogenomics studies that does not stop at the biochemical level but also extends this information to the organism level effects of chemical structures.

*Address correspondence to this author at the Unilever Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom; Tel: +44 (1223) 762 983; Fax: +44 (1223) 763-076; E-mail: andreas.bender@cantab.net

1568-0266/11 $58.00+.00

What is interesting to note is that publications containing the terms 'chemogenomic' or 'chemogenomics' in the Topic (as defined by Web of Science; containing publication title, abstract and keywords) has not seen the exponential growth one might have expected in recent years. This trend is visualized in Fig. (**1**); while citation data for 2010 are incomplete, it seems as if from 2005 to 2010 an approximate plateau phase of only around 25 publications per year were achieved. In absolute terms this is a relatively small number, and it would be an interesting subject of discussion what the reason would be – whether the concept did not work out in practice; whether people already switched to using other terms instead of these two; or whether the real boom of the area is still ahead of us. The future will certainly tell.

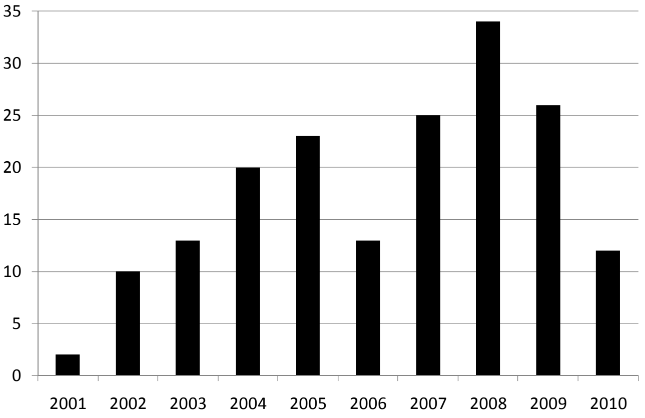


Fig. (1). Citations of 'chemogenomics OR chemogenomic' in Topic (Title, Abstract, Keywords) of Web of Science (as on 1 June 2010). While citation data for 2010 are incomplete, it seems as if from 2005 to 2010 an approximate plateau phase of around 25 publications per year were achieved.

Large, well indexed ligand-target-assay databases are at the very heart of chemogenomics studies which aim to relate ligand chemistry to the similarity of protein targets on a large scale; and it should be kept in mind that data completeness determines the analysis outcome [17,18]. Despite information not being perfect, it can still be of value in a more informed decision-making process. It was noted in pharmaceutical companies that information from related targets can be used to steer ligand discovery for an orphan target of interest in a productive direction, but at the same time it was by no means obvious in which way, and in which cases, this could be achieved. In parallel, the previously eminent paradigm of 'ligand selectivity' has recently been replaced in favor of a desired polypharmacology profile [19-21].Such a profile requires not only a set of targets of interest from the biological side that needs to be considered, but also ligand chemistry to be chosen so they can be bioactive against multiple targets in parallel. However, important problems are still unsolved for this paradigm, for example the question how a researcher defines a desired bioactivity in the first place; and how bioactivity profiles actually translate from animal systems to humans. This controlled polypharmacology is probably more easily achieved in some cases (e.g. a set of structurally related Class A GPCRs) than in case of others (e.g. inhibiting a protease and antagonizing a GPCR simultaneously). A recent study by one of the authors [21] was analyzing bioactive chemical space, as defined by the WOMBAT database, ECFP4 circular fingerprints and a Principal Component Analysis of the resulting Bayes Models per class, the result of which is visualized in Fig. (**2**). This figure can be interpreted as follows: PC1 and PC2 are the axes of maximum chemical ligand diversity; hence, the classes with the highest loadings along both axes have the most different chemistry in the set, here dopamine D2 receptor ligands and HIV integrase inhibitors. Along PC3 μ opioid receptor ligands possess the furthest distance from those two classes. And indeed, when analyzing ligand chemistry in more detail (see [21] for further details) we can see that D2 ligands always possess a tertiary nitrogen and often unsaturated six-membered rings; HIV integrase inhibitors typically possess catechol moieties and carboxylic acid groups, frequently they also are esters; and μ opioid receptor ligands show considerable diversity, from the complex morphine scaffold to much more diverse ligands that could also resemble enzyme inhibitors and GPCR ligands. Similar analyses, based sometimes on different datasets and different algorithms, were also performed by other groups such as at Pfizer[22,23], UCSF[24] and the University of Barcelona [25].

One concept that is implicit in chemogenomics approaches is that chemical space is reasonably well behaved, continuous and can be interpolated. With respect to the biological side, it is assumed that ligands active against one protein are more often than random also active against a related protein structure. Very much related to these concepts is the idea of 'affinity fingerprints' published by Kauvar *et al.* [26]. In its simplest form, affinity fingerprints make the assumption that affinity to novel targets can be approximated from known affinities to a set of ('orthogonal') proteins. Later this idea was extended to the computational domain with concepts such as docking-based fingerprints[27] and ligand based 'Bayes Affinity Fingerprints'[21], but what is interesting in the chemogenomics context is the following. As current chemogenomics thinking goes, only measures of protein similarity are related to measures of ligand similarity - and, hence, simply ligands of related receptors to an orphan receptor are used as a starting point for de-orphanization projects. However, this is only the simplest version of the affinity fingerprints imaginable. According to the affinity fingerprint concept, also bioactivity values against very *different* proteins could be used to make (approximate) predictions as to which ligands are active against a currently orphan receptor in a linear combination of affinities, or a yet unknown more complex function. Given that more and more data becomes available it is surely only a matter of time until the current chemogenomics concepts of 'similar targets bind similar ligands' get extended to also cover more complex relationships between ligand chemistry and target bioactivity space. We anticipate that research in this area will have huge benefits for experimental biologists as well as biological and medicinal chemist and will be transformative for the commercial life-science sector.

In a similar vein, the term proteochemometrics is certainly very dissimilar at first sight to chemogenomics; however, the underlying ideas of both concepts are not fundamentally different. Both of them attempt to relate ligand chemistry to a *set* of different proteins instead of a single protein (which would be more the domain of conventional

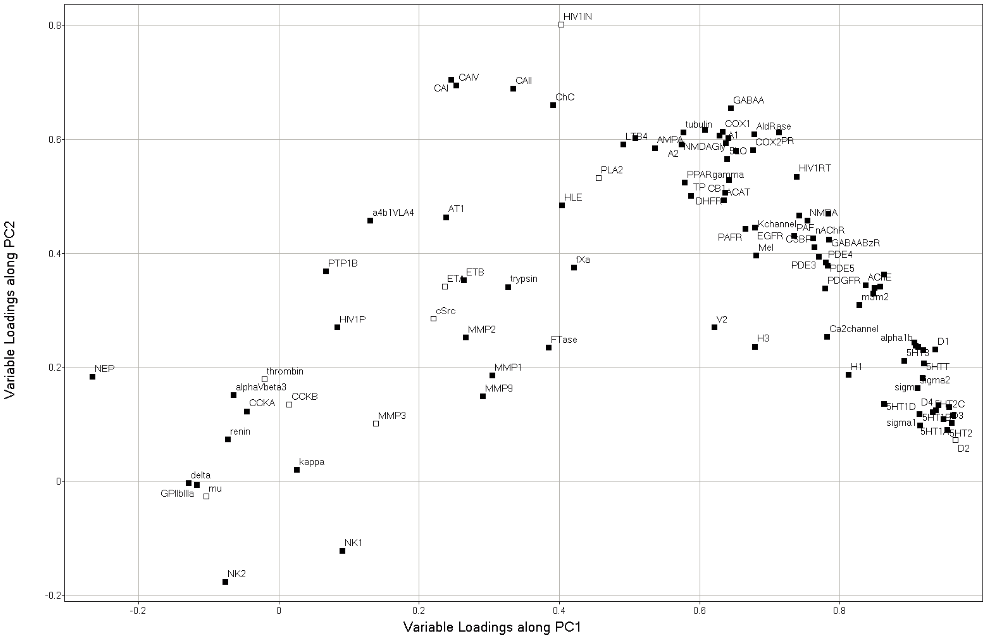


Fig. (2). Two first principal components of ligand bioactivity space. Inhibitors of HIV-1 integrase and ligands of the dopamine D2 receptor define the most distinct ligand classes from the chemical side. (Reprinted with permission from [21]).

structure-activity modeling), and the difference between both concepts are sometimes minor, with proteochemometric modeling being often more focused on a more defined set of targets with the aim to model bioactivity relationships quantitatively; however none of these characteristics are true in every case. One of the most comprehensive recent studies [28] was attempting to model all protein-ligand interaction space of enzymes, as taken from protein-ligand cocrystals. A model was generated on 826 pairs of proteins, druglike ligands and binding affinity or dissociation constant of the particular protein-ligand pair from AffinDB, PDB Bind, Binding MOAD and Protein-Ligand Database, and 542 entries from Brenda were used as a test set. Predictions for the external test set achieved an r^2 of 0.53 and a RMSEP of 1.5 (over a pKi range reported[28] to be from 0.7-11.0); hence, even when using very diverse protein data in this kind of modeling exercise affinity data can be modeled relatively reliably – which empirically underlines the validity of the concepts behind chemogenomic and proteochemometric modeling.

In this publication no comprehensive overview of all Chemogenomics studies can be given, and hence a set of recent reviews in the area should be mentioned [29-36]. In the following, we will now outline our recent advances in chemogenomics studies applied to receptor deorphanization, different representations of molecules in studies of this type, proteochemometrics studies performed in our group, as well as extensions of the chemogenomics concept into phenotype space.

CHEMOGENOMICS APPLIED TO RECEPTOR DEORPHANIZATION

As outlined in the introduction, the most straightforward analysis in the chemogenomics spirit (which is still dependent on quite a lot of variables such as the precise dataset chosen, the chemical representation used, and the distance metric employed) is to relate ligand similarity to protein similarity in a given bioactivity data set. Previous studies in the field exist, very early in the field of kinases [37] where structural binding site similarity was compared to ligand SAR; however only 58 ATP-site ligands were contained in the PDB at the time of the article. More recently this database has been extended significantly, with large-scale profiling data of kinases entering the public domain [38, 39]. Data of this type has also been used recently to construct a ligand-based kinase tree [40] as well as to enzyme families [41].

In the work performed in our group though GPCRs are the main target family of interest, and less chemogenomics work has yet been published in this area. One recent study [42] attempts indeed a (hypothetical) deorphanization of GPCRs, leaving all ligands of the receptor under consideration out of the training set and generating models using different kernel-based approaches. It was found that using a binding-pocket kernel that only takes the sequence of residues in the binding pocket into account, achieve 78.1% correct predictions, averaged over all classes, on the dataset, compared to 50% in case of random class assignments. However, what was not done was the analysis in which cases chemogenomics studies applied to GPCR ligands would succeed in practice, and to ‘re-draw’ the GPCR tree based on a chemogenomics analysis; this is precisely what we performed in our ongoing work in the group that is currently being validated prospectively.

Other related work has been published by Bock *et al.* [43] and Weill and Rognan [44]. Bock *et al.* [43] described receptors by using numerical values for the surface tension, isoelectric point, and accessible surface area for each amino acid of the receptor as well as a connectivity matrix, extended by atomic properties such as ionization energy, electron affinity and atom density on the ligand side and found that out of about 2,000,000 novel compounds only 2% are predicted to be active on a novel, orphan receptor. Models were generated by a Support Vector Machine on 5,319 receptor-ligand pairs from the PDSP database; however, no statistics for the models generated has been provided for a target deorphanization exercise.

In Weill and Rognan [44], a novel protein-ligand fingerprint, termed PLFP, was introduced that captures pharmacophoric properties of the ligands, as well as of the transmembrane ligand binding pocket. SHED, topological autocorrelation descriptors (DistFP) and MACCS keys were combined with SVM, Naïve Bayes and Random Forests to evaluate the influence of different chemical representations and model generation methods on predictive performance and models were evaluated on two external test sets containing a total of 60 data points. It was found that picking ligands for targets was usually easier to achieve than the reverse, predicting targets for ligands. Also, both global models, considering all GPCR ligand data in a single model, as well as a set of 19 local models, one for each subclass of GPCRs, were compared. Here it was found that DistFP with support vector machines on local models outperforms on average all other model generation methods – a hint that local bioactivity models may be required to model larger areas of bioactivity space.

We now extended the above study, and analyzed chemically in more detail in which cases a chemogenomics classification of GPCR ligands would be more likely to be successful in a deorphanization study than in others. Given the dependence of this type of study on a large number of factors (the diversity of chemical and biological space covered in the datasets; the measures used to establish similarities in both spaces; and the decision which data points of ‘neighboring’ sequences to include, just to name a few) we felt the need to establish some practical guidelines for future work in this area.

For this purpose, we used a receptor-based phylogenetic classification and compared it to a ligand-based classification that is based on exhaustive substructure mining. (Note that a full primary research article on this work has been published recently with more details of the results and a further interpretation which can be found in this reference [45].) For the protein side, a set of 44 amino acid residues likely to be involved in ligand binding of class A GPCRs were selected, based on previous work of Gloriam *et al.* [46]. This selection of residues is based on previous work by Surgand *et al.* [47] of 30 residues, but extends it by taking the recently published crystal structures of the human β2 and turkey β1 adrenoceptors, as well as that of the human adenosine A2A receptor and their residues involved in ligand binding into account to extend the previous set to 44 relevant residues. Based on those amino acid residues, a phylogenetic tree of class A GPCRs was constructed. An important nuance to classical sequence-based phylogenetic analysis is in the choice of the scoring matrix for amino-acid interchanges, since very different effects are anticipated for conserved molecular recognition effects compared to the different constraints observed in protein evolution. For example, glutamic acid and arginine are generally similar from a protein structure viewpoint, where they are regularly interchanged on the surface of proteins; however, when these amino acids are involved in the binding of a small molecule ligand they have very different properties. Again, looking to the future, we see investigating this area as being productive now that significant amounts of data are becoming available.

Complementary to these sequence-based classifications a substructure-based classification of the receptors from the ligand side was performed in parallel. Exhaustive substructure mining of GPCR ligands was performed as described in detail in the primary research article [48]. Structures were represented as labeled graphs with aromatic bonds being assigned a different bond type. In this study, the minimum support value was set to 30% of the number of ligands in each activity set, meaning that only substructures present in at least 30% of the ligands were considered in the further analysis. Substructures below 50 Dalton were discarded, since very small fragments are chemically not amenable to interpretation, which is easily the case for larger substructures of molecules since they are molecular representations accessible to the way of thinking of a chemist. To calculate the similarity between activity classes, the Pearson Correlation Coefficient between all feature frequencies in each activity class was calculated. The correlation coefficient was then transformed into a distance measure by subtracting the correlation coefficient from 1 and using linear scaling of all results to [0;1]. Tree construction might in principle be influenced by the order in which targets are provided to the tree constructor, and in order to investigate this effect the target input order was randomized 10 times and 10 new trees were generated and compared to each other. Only in very rare cases the trees generated were different though, supporting the robustness of the method employed.

The tree that was built based on the multiple sequence alignment as defined by Gloriam *et al.* [46] set is shown in Fig. (**3**). Four clusters are clearly defined in the tree, namely the aminergic receptors, adenosine receptors, prostanoid receptors, as well as the peptide-binding receptors. This clustering is very close to the ‘conventional’ results obtained from full-sequence phylogenetic trees; note that minor differences to the original tree presented by Gloriam *et al.* are visible since our target-based tree was only constructed using bioactivity classes for which sufficient ligand data was available. For comparison, the ligand-based receptor classification tree is provided in Fig. (**4**). Overall, to a large extent the target-based phylogenetic tree is conserved, with differences only visible for some of the receptor subclasses. It should be kept in mind that this may also partly be due to the chemistry tested against particular receptors – most scientists would test known ligands against one receptor subtype also against another receptor subtype, and hence due to this ‘selection bias’ some classes might move closer together than with more even sampling of chemical space screened against a receptor.

Except for the two purinergic receptors (P2Y1 and P2Y12) and the two glycoprotein hormone receptors (FSH and LH), all other receptor pairs represented by two subtype are clustered together. The adenosine receptors (ADORA1, ADORA2A, ADORA2B, ADORA3) group together with the A2B receptor being the most dissimilar from the other, which is consistent with previous results obtained by our group. The purinergic receptor P2Y12 is from the ligand-side found to be similar to the adenosine receptors, which is

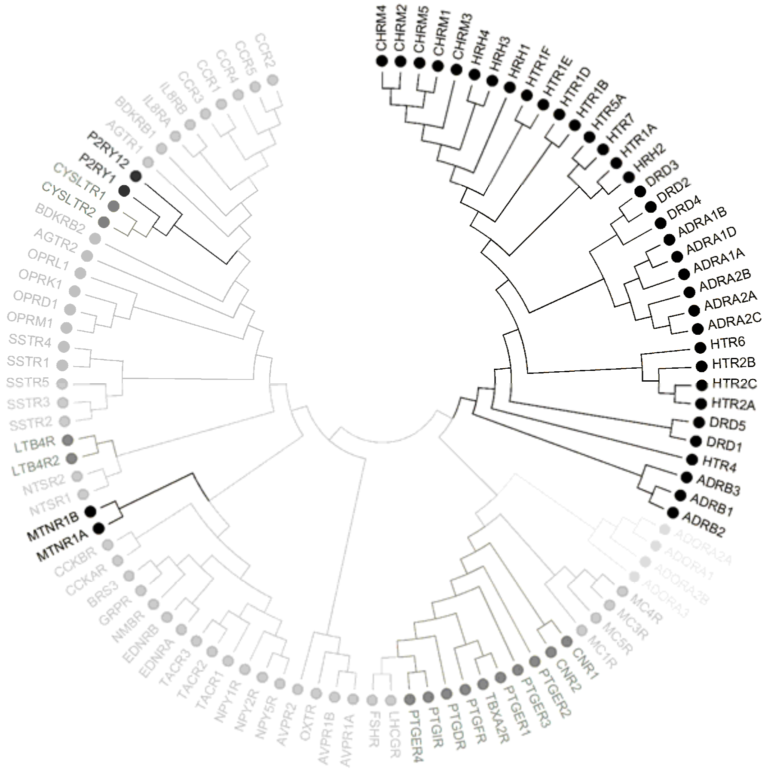


Fig. (3). Phylogenetic tree of human class A GPCRs based on the 44 residues identified relevant for binding by Gloriam *et al.* The color codes are as follows: black – receptor with aminergic ligands; dark grey – purinergic ligands and melatonin ligands; medium grey – lipid ligands; light grey – adenosine ligands. (Reprinted with permission from [45]; also see this reference for figure in color.)

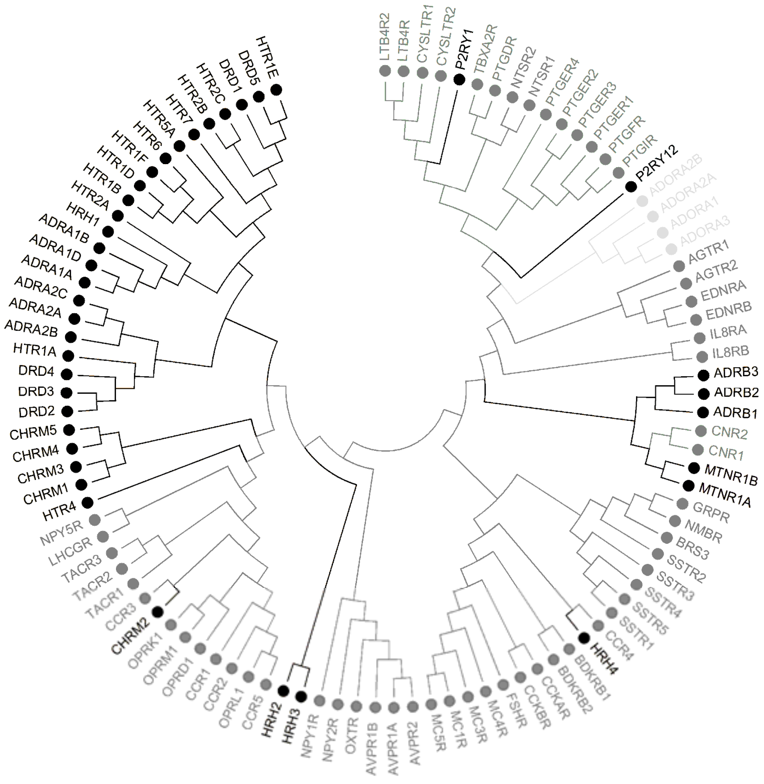


Fig. (4). Phylogenetic tree of class A GPCRs based on frequent substructure mining of GLIDA and ChEMBL data. The color codes are as follows: black – receptor with aminergic ligands; dark grey - purinergic ligands and melatonin ligands; medium grey – lipid ligands; light grey – adenosine ligands. (Reprinted with permission from [45]; also see this reference for figure in color.)

understandable due to the common purine core typical for ligands of both subfamilies. The muscarinic acetylcholine receptors M1, M3, M4, and M5 cluster together as one group, supporting the low subtype selectivity of muscarinic antagonists; however, the acetylcholine receptor M2 is found more distant from this cluster which may be the result of inclusion of allosteric ligands. Some clusters not present in the sequence-based tree are present in the ligand-based classification and they can often be well understood from the chemical point of view; e.g. the grouping of the eight prostanoid receptors displayed in Fig. (**3**). This cluster is based on the fact that most prostanoid receptor ligands are direct derivatives of the endogenous ligands, the so-called eicosanoids. These ligands form a very homogeneous group which is dominated by relatively long alkyl chains. The clustering of the leukotriene and cannabinoid receptors in this lipid cluster may seem strange at first; however, arachidonic acid is the common precursor for eicosanoids and two derivatives of arachidonic acid, anandamide and 2-arachidonyl-glycerol, both of which are endogenous ligands (‘endocannabinoids’) of the cannabinoid receptors and which explains the chemical similarity of ligands in this cluster, and hence, their relatedness from the chemical point of view.

While in the original work on this topic [45] the differences and commonalities between biological and chemical phylogenetic trees have been analyzed in detail, in the context of this review its applications to the deorphanization of GPCRs should be commented on in particular. In order to resemble a real-world setting, we performed a hypothetical ‘de-orphanization exercise’. To do this, we excluded in turn all ligands of each receptor in the dataset; we so ‘orphanized’ the receptor in this particular analysis run. Next, we ‘de-orphanized’ the receptor again by predicting its ligands by using a ligand-based bioactivity model derived from the closest neighbors of the receptor in sequence space. It was found that in 93% of the cases our hypothetical de-orphanization exercise was successful, with the model providing enrichment curves better than random (AUC > 0.5); and for 35% of receptors performance was even ‘good’ (as defined by PipelinePilot, with an AUC > 0.7). Sample plots for four receptors are shown in Fig. (**5**), namely for CHRM1 (muscarinic acetylcholine receptor M1), AGTR2 (angiotensin II receptor, type 2), P2RY1 (purinergic receptor P2Y, G-protein coupled, 1) and BRS3 (bombesin-like receptor 3). We could indeed find a rationale in which de-orphanization exercises are more likely to succeed than in others: The poor performance concerning the P2RY1 receptor is probably due to the nature of its ligands, since this set consists of a small number of highly similar ligands that all possess at least one phosphate group, a feature not found in other ligands in the database. (In fact, its most related sibling from the biological side, P2RY12 Fig. (**3**), moves far away in chemical space Fig. (**4**), illustrating the dissimilarity of the ligands) – and hence, in this case the deorphanization exercise fails. On the other hand, in case of CHRM1 very much related ligands are present in the dataset for its nearest neighbor, CHRM5 – hence, in this case the deorphanization exercise succeeds. Overall in this study our method was relatively successful to achieve this task since for 93% of the receptors studied performance better than random was achieved (AUC > 0.5), and for 35% of receptors even models with reasonable quality were obtained (AUC > 0.7).

CHEMICAL REPRESENTATIONS IN CHEMOGENOMICS APPROACHES

As in ligand-based virtual screening [49-51], the question how to represent chemicals in a chemogenomics study, as well as how to calculate the similarity or distance between different compound classes and hereby between receptors, is by no means obvious. Different groups use different approaches, from the pairwise comparison of compounds in each class followed by calculating expectation values [24] to calculating frequency vectors of circular fingerprints for each class and calculating the correlation coefficient between them [49,52]. In our previous work, with the aim to obtain a non-biased representation of chemical substructures, we used exhaustive substructure enumeration of databases to represent chemical structures which is outlined in the following and described in detail in a recent primary research article [48].

Molecular subgraph mining (also known as association learning) has two characteristics that render it very different from other molecular representations such as fingerprints and fragment-based representations: On the one hand it is a very demanding method to represent molecules due to the number of subgraphs that can be generated even for a drug-like small molecule; this number is in the order of millions for example for steroids, due also to the complex ring system present in them. On the other hand it is an unbiased representation of molecules (apart from the model assumption of ‘atoms’ and ‘bonds’) – no assumptions are made which molecular features are relevant for a certain property such as a bioactivity against a protein. This is the case in fingerprints (‘circular features’ are important e.g. for circular fingerprints, or ‘pairs of features are related to the property being considered’ in case of atom pairs, and so on) as well as key-based fingerprints, where precise chemical groups are predefined. Hence, in our work on the chemogenomics data mining of GPCR databases presented in detail recently [48] subgraph mining was employed to analyze the databases as hand.

In our particular case, molecular structures are represented as labeled graphs with only heavy atoms being considered in the analysis. Four types of chemical representation were used: the initial chemical structure representation with the atom and bond types unchanged, and three ‘Elaborate’ Chemical Representations (ECRs) that attach labels to atoms, as well as bonds, in order to generalize the chemistry encountered (similar to the typing of atoms in pharmacophores as lipophilic, charged, hydrogen bond donor/acceptor, etc.). [53,54] which is illustrated in Fig. (**6**). Atoms and bonds can both be typed; in our case aliphatic nitrogen, oxygen, and sulfur atoms were represented as aliphatic heteroatom by replacement with the symbol ‘No’ and an extra label was attached to nitrogen and oxygen atoms in order to indicate the type and number of bound hydrogen atoms. This captures information similar to hydrogen bond donor and acceptor properties of a heteroatom. The halogen atoms, Cl, Br, I, and F, were replaced by X, since they are partially negatively charged and, hence, can be isosteres in many

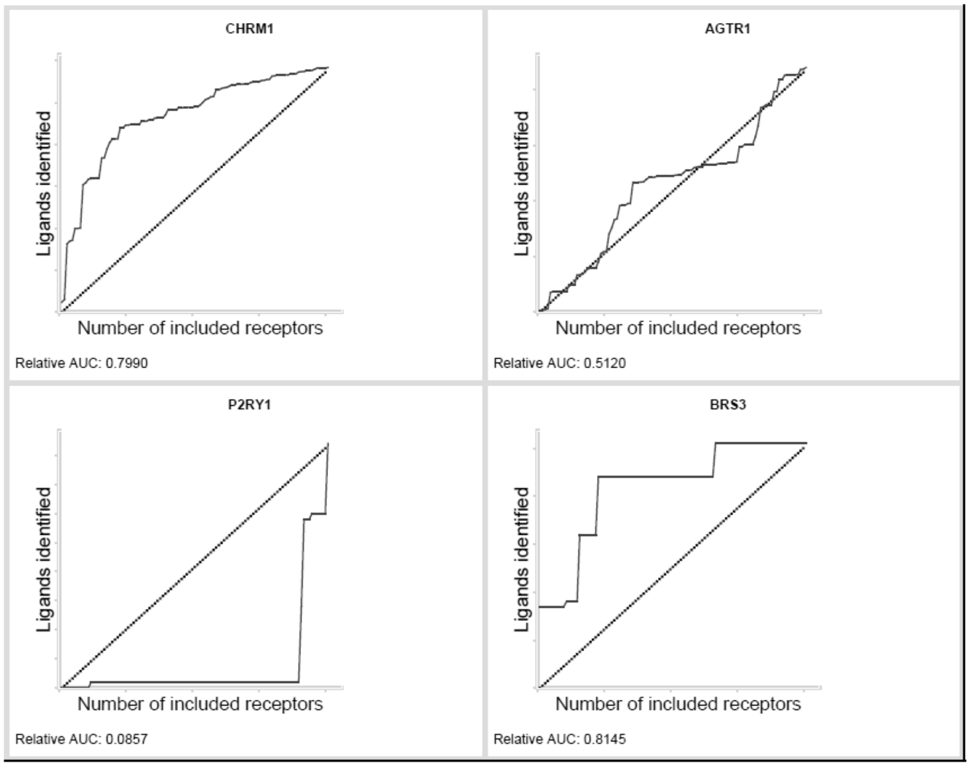


Fig. (5). Deorphanization enrichment for different sample bioactivity classes (for full receptor names see main text). Overall, for 93% of the receptors studied performance better than random was achieved (AUC > 0.5), while for 35% of receptors good-quality models were obtained (with an AUC > 0.7). (Reprinted with permission from [45])

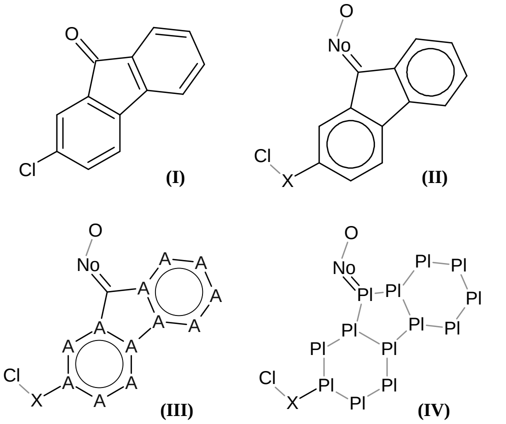


Fig. (6). A sample molecule in normal **(I)** and three elaborate chemical representations, the first one including aromatic bonds **(II)**, a second one including both aromatic atoms and bonds **(III)**, and one capturing planar ring systems **(IV)**. In the normal representation **(I)**, aromatic bonds are represented as alternating single and double bonds, while in the first elaborate representation **(II)**, a special aromatic bond type exists (representation **(III)** extends aromatic typing also to atoms). In both elaborate chemical representations, wildcards are used for heteroatoms ('No') and for halogens ('X'). (Reprinted with permission from [48])

(though, in particular in case of fluorine, not all) biological situations. Also one elaborate representation includes a special bond type for aromatic bonds, while the second representation also has a special type for aromatic atoms. Finally, the third representation offers a special type for planar ring systems, which has been successfully applied previously to predict the mutagenicity of compounds[54]. Overall, it was found that the elaborate representations indeed outperform the original representation of molecules (measured as features identified that better discriminate between GPCR ligands and background compounds), indicating that atom typing is beneficial in classifications of this type.

Algorithmically, the subgraph miner Gaston was used that was previously successfully applied to molecular datasets, details of which can be found in the original publications [53] and the significance of molecular substructures was measured as a the likelihood of a feature distribution occurring by chance, known also as the 'p-value' of a distribution (as defined in [54]) and the substructure with the lowest p-value was considered the most important one. Apart from the p-value of a substructure, an important parameter in frequent subgraph mining is the minimum support value, which is the fraction of molecules in a given dataset that should possess a particular substructure in order to be considered for further analysis. Lowering the minimum support will result in a larger number of substructures and vice versa – and given the large number (100,000s to millions) of substructures that can be potentially detected in a single molecule it is apparent how important a reasonable choice of this parameter is. The minimum support value was chosen empirically, resulting in practice in support values between 10% and 30% for the datasets used in this study.

As important as the molecular representation is also the choice of a suitable database, and in our study GPCR ligands were collected from the GLIDA and hGPCR-lig databases. The set from GLIDA consisted of 22,122 ligands for human, mouse and rat receptors, while hGPCR-lig contained 17,908 GPCR ligands from literature as well as the MDL Drug Data Report (MDDR) database. These two sets were compared against a control set, namely 15,993 compounds from ChemBridge's DIVERSet screening collection and each database was represented by a random set of 5,000 ligands, this limitation being due to the computational expense of the methods used here. Targets were arranged into a hierarchy of subfamilies, families, and classes which is shown in Fig. (**7**), and which originates from GPCRDB. For further details the reader is referred to the original study [48].

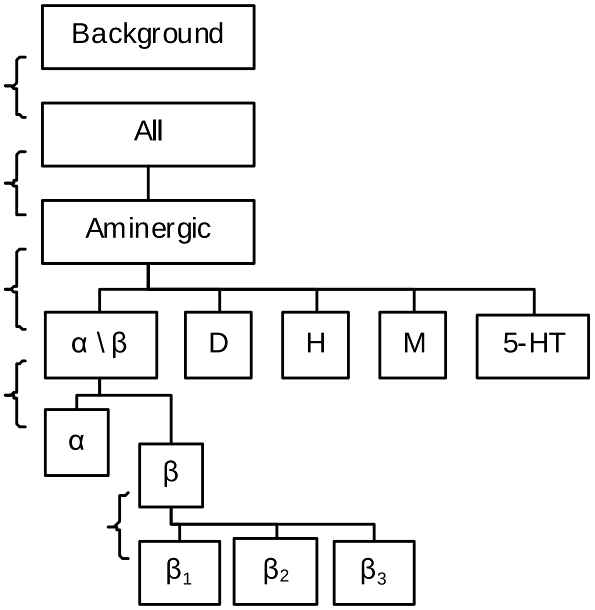
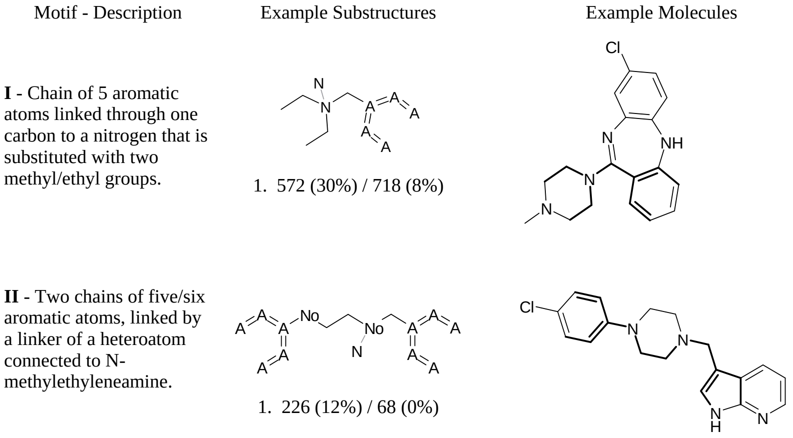


Fig. (7). Schematic drawing of the ligand bioactivity classes substructure mining was applied to (curly brackets indicate the sets that were compared against each other). 'Background' – The ChemBridge DIVERSet, 'All' – all GPCR ligands found in GLIDA, 'Aminergic' – all aminergic receptor ligands, 'α \ β' – Adrenoceptors, 'D' – Dopamine receptors, 'H' – Histamine, 'M' – Muscarinic Acetylcholine receptors, '5-HT' – Serotonin receptors, 'α' – α-adrenoceptors, 'β' – β-adrenoceptors, '$\beta_{1\text{-}3}$' – β-adrenoceptor subtypes 1 to 3. (Reprinted with permission from [48])

The different levels of the target hierarchy are all amenable to substructure mining, and while various studies have been performed which reproduce known data (such as the importance of positively charged nitrogens for class A GPCR ligands relevant for binding to an aspartate residue in the receptor), also more detailed results have been obtained from this analysis. Here only one example shall be commented on, namely for dopamine receptor ligands for which the outcome of the analysis is represented in Fig. (**8**).

For the dopamine receptor ligands, two types of specific substructures were identified Fig. (**8**) that are characteristic for this type of bioactivity. Here, the first characteristic substructure is present in 30% of the ligands and it consists of a chain of four or five aromatic atoms, connected to a tertiary nitrogen atom via a methyl linker. The second substructure, which is present in 12% of the ligands, consists of two aromatic chains which can be five or six atoms in length, and which are linked via a heteroatom (nitrogen or oxygen) connected to N-methylethyleneamine, while the terminal nitrogen of this linker may be substituted by an ethyl group. As can be seen in Fig. (**8**), the piperazine ring is part of those two characteristic substructures, although not in its entirety – the interpretation is that variations of this ring are possible when designing dopaminergic drugs, so that this feature does not remain static throughout the dataset. Similarly, the aromatic chains in both substructures are able to overlap with various types of aromatic systems; hence, it is the aromatic character of the system that is conserved in this feature, and not the precise nature of the ring system present. Another example of novel substructures conferring activity were fused 5, 6 bicyclic ring systems in serotonergic ligands; these may steer synthetic chemistry in novel directions when designing future bioactive molecules.

As for future research avenues, currently only the 2D graph of a molecule is considered for analysis, so any geometric information (such as chirality) is not taken into account. Also, the p-value is used to sort the substructures according to significance; however in practice measures such as enrichment of a compound bioactivity class might be of more relevance. Also it is important to also look at substructures frequent in both data sets when using the information


Motif - Description

Example Substructures

Example Molecules

I - Chain of 5 aromatic atoms linked through one carbon to a nitrogen that is substituted with two methyl/ethyl groups.

1. 572 (30%) / 718 (8%)

II - Two chains of five/six aromatic atoms, linked by a linker of a heteroatom connected to N-methylethyleneamine.

1. 226 (12%) / 68 (0%)

Fig. (8). Common motif and example substructures for most significant substructures of the dopamine receptor ligands, in aromatic atoms and bonds representation. An example drug that has motif I is clozapine, an antipsychotic agent used in the treatment of schizophrenia. Another example for motif I and also for motif II is compound L-745,870, a selective dopamine D_4 receptor antagonist. (Reprinted with permission from [48]) for designing novel drugs – while not conferring selectivity, those features might still be beneficial for compound affinity. Overall we can conclude from this study that our analysis is complementary to employing ‘privileged substructures’ in ligand design, since it is not restricted to existing scaffold structures. Apparently, elaborate chemical representations add substantial value when searching for structural features typical for active compounds. This enables the user in addition to detect bioisosteres of chemical groups which can be employed in the prospective design of ligands with a desired bioactivity profile.

PROTEOCHEMOMETRICS APPROACHES FOR EXTRAPOLATING BIOACTIVITY TO RELATED TARGETS

As mentioned above, the terms ‘chemogenomics’ and ‘proteochemometrics’ are not very different in nature, and from the experience of the authors the following statement is reasonable, that in general ‘proteochemometrics models’ are usually generated based on ‘chemogenomics data’. Proteochemometric models, as opposed to conventional SAR models, add a target descriptor in addition to the ligand descriptor to the model, in order to use bioactivity information from related targets to make better predictions where there are data points known for a target; but also to enable extrapolation of bioactivity data to novel targets – such as mutants of viral enzymes, or related receptor subtypes (for a recent review see [55] and for related research employing so-called ‘signature descriptors’ see [56]).

Originally this method was intended to improve prediction capabilities on a series of targets where data points were given [57, 58]. However, it is only a small step expanding this method to enable the prediction of specificity by inclusion of highly similar targets; in this case the model is then predicting bioactivities for targets where no data points are known. This idea is illustrated in Fig. (**9**) on a hypothetical dataset that consist of ligand A and very similar ligands B1 and B2, as well as target 1 and very similar targets 2A and 2B. In conventional QSAR models, for every target a separate bioactivity model needs to be constructed, and no extrapolation of bioactivities between targets is possible. However, proteochemometric modeling takes into account that targets 2A and 2B are similar (e.g. with respect to the shape and properties of the binding site), hence an approximate activity also for ligand B2 on target 2A can be predicted, and also for ligand B1 on target 2B. Both target 1 as well as ligand A are more dissimilar from the rest of the data, and here extrapolation is generally possible as a function of ligand as well as target similarity.

The observation that model extrapolation is possible not only as a function of ligand similarity, but also as a function of target similarity, extends the previous ‘applicability domain’ concept [59-62]known from QSAR to the biological domain. We can illustrate it with experimental data from our group on mutants of viral targets, namely of a set of 14 mutants of HIV reverse transcriptase (manuscript under preparation). Shown in Fig. (**10**) is the performance of proteochemometrics modeling in ‘leave-one-sequence-out experiments’, as measured by root mean squared error (RMSE). The model trained without sequence 8 is seen to perform inadequately in the validation, most likely due to the underestimated impact of the single backbone changing mutation that is not present in the other mutants. Hence, in proteochemometric modeling the ‘applicability domain’ concept previously established for the ligand side needs also be applied to the target side when considering the ability of models to extrapolate to related protein targets.

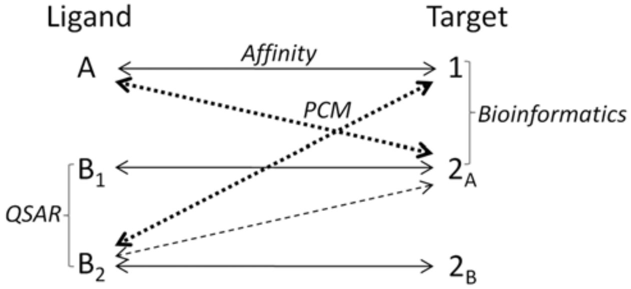


Fig. (9). Illustration of proteochemometric modeling on a hypothetical dataset that consist of ligand A and very similar ligands B1 and B2, as well as target 1 and very similar targets 2A and 2B. In conventional QSAR models, for every target a separate bioactivity model needs to be constructed, and no extrapolation of bioactivities between targets is possible. However, proteochemometric modeling takes into account that targets 2A and 2B are similar (e.g. with respect to the shape and properties of the binding site), hence an approximate activity for ligand B2 on target 2A can also be predicted, as well as (with probably more error) an activity of ligand B2 on target 1. Target 1 as well as ligand A are more dissimilar from the rest of the data, and here extrapolation is generally possible as a function of ligand- as well as target similarity.

Furthermore, we applied proteochemometric modeling to predict the activity of chemical compounds on the four adenosine receptors by addition of a target description, based on binding site similarity as this binding site is well known from a recent crystal structure [63]. We used SCFP4 fingerprints[64], which were shown to capture more information than many other methods with respect to their respective bioactivities[65], together with properties of 32 residues lining the binding site taken from the AAindex database, to describe our ligand-target complex. Support Vector Machines with real-valued output predictions as implemented in PipelinePilot Student Edition 6.1 were then trained on known Adenosine Receptor Ligands from ChEMBL. In order to validate the model, we spiked 4,556 random compounds from ZINC with 43 known high-affinity compounds from our in-house GIFT/LUF compound datasets, and ranked all ligands using the A2A output of the proteochemometric model predictions to select compounds to evaluate predictions on this external test set. The ranking of compounds is shown in Fig. (**11**), and it can be seen that the top 14 compounds are true positives (as measured against any Adenosine receptor subtype), while a total of 37 compounds from our in-house data set were among the 100 highest predicted compounds of the entire data set containing 4,599 data points. The three lowest predicted compounds were found back at position 249, 935 and 1081. The highest predicted compound was LUF5957 with a predicted pKi of 9.02 and an experimentally determined pKi of 9.14. Given the satisfying performance of this model, we employed it also to select novel compounds from supplier databases that

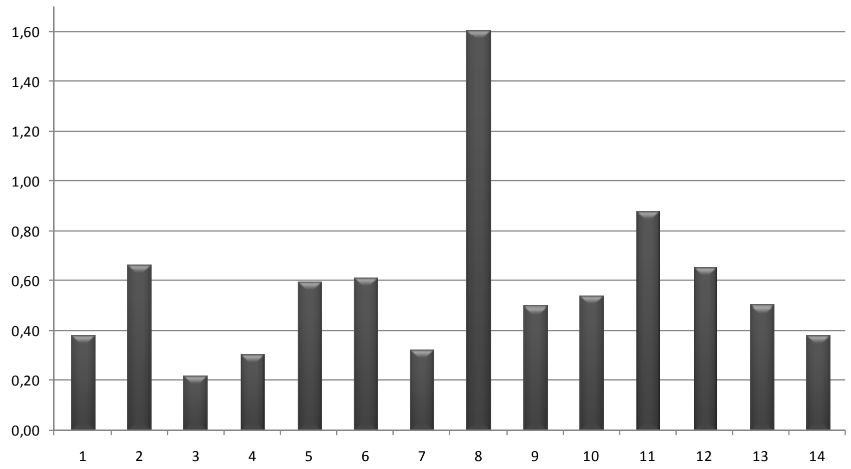


Fig. (10). Performance of proteochemometrics modeling in 'leave-one-sequence-out experiments', as measured by root mean squared error (RMSE). The model trained without sequence 8 is seen to perform inadequately in the validation, most likely due to the underestimated impact of the single backbone changing mutation that is not present in the other mutants. Hence, in proteochemometric modeling the 'applicability domain' concept previously established for the ligand side also needs to be applied to the target side when considering the ability of models to extrapolate to related protein targets.

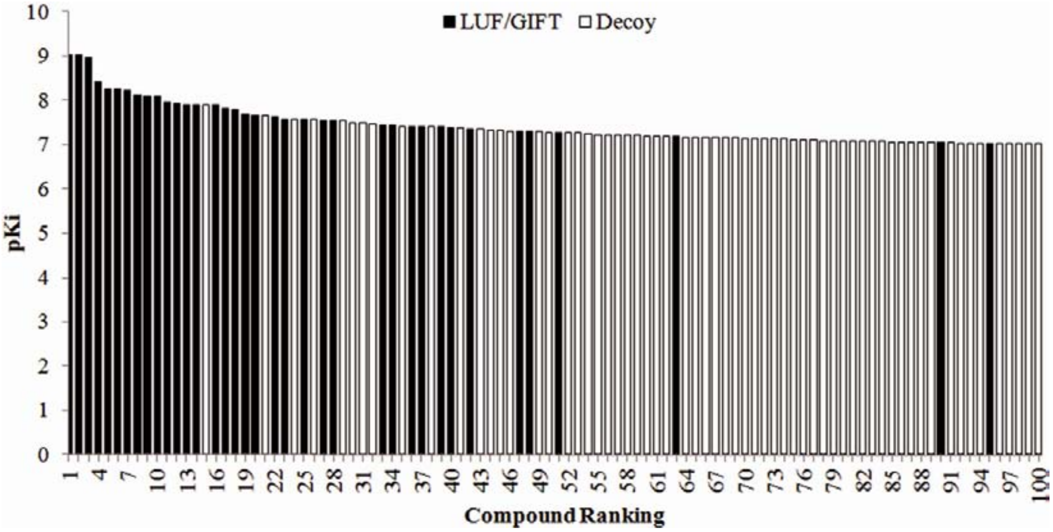


Fig. (11). The prediction of a random Zinc decoy set (1139 compounds per receptor) spiked with high affinity compounds active against any Adenosine receptor (43 in total) from our in-house database, using the A2A bioactivity model. Of the 43 high affinity compounds, 33 were predicted in the top 50 and 37 in the top 100. The highest predicted decoy compound ranked 14th with a pK_i of 7.90.

are currently being evaluated in-house in truly prospective testing mode.

LINKING PHENOTYPIC SPACE INTO CHEMOGENOMICS ANALYSES

Chemogenomics analyses in their original form are attempting to link ligand chemical space to target bioactivity space; however, in most cases until now this link has been restricted to bioactivity against a protein target typically reflected in a binding constant, inhibition constant and so forth. In practice though most often the modulation of a more complex biological system, such as on an organ or organism level, is of prime importance and there is no reason why chemogenomics analyses should not include more complex biological variables than the target similarity by itself. Recent ambitious analyses though are trying to link ligand chemical space, via protein target space, to phenotypic space and some examples of those applications shall be presented in the following.

One simple observation when dealing with either organisms or cell systems is that they can die. While on an organism level this can be a disturbing event, also on the cellular level this can lead to much distress on the side of the screener when dealing with cell-based assay systems. One very common assay type in pharmaceutical industry are reporter gene assays; and in cases where a decrease in signal (reporter gene expression) is used as a positive readout, without normalizing for the number of cells the signal is taken from cell death can lead to false-positive readouts, and compounds falsely flagged as 'hits' in a reporter gene assay screen. In our work at Novartis, we employed protein target prediction models [13,66] for a variety of purposes, and one of them was the analysis of frequent hitters in reporter gene assays[67], where we assumed a common, underling reason behind them hitting in a large number of assays in parallel. Hence, we employed target prediction models on compounds showing bioactivity in many assays, and it was indeed found that kinases involved in the cell cycle were very common targets of frequent hitters [67] – thus, providing a rational explanation, based on chemogenomics data, for an apparently positive signal in a phenotypic screen.

While cell-based phenotypes are very simple phenotypes, we can also start to make the crucial step to humans and adverse drug reactions observed upon compound administration. Recently publications have appeared which go back from observed adverse drug reactions to a common underlying mechanism of action [68], and in our studies we focused on a chemogenomics-based target prediction of compounds with an adverse drug reaction as annotated in the World Drug Index [52]. For every set of compounds with an adverse drug reaction, targets were predicted, and enriched targets in a particular compound class were then statistically associated with a particular adverse reaction; and hence, they might also be associated with this adverse reaction on a mechanistic level. An illustration is shown in Fig. (**12**), showing the correlation between ligand chemistry active against proteins and the ligand chemistry causing adverse drug reactions (phenotypic information). Both target space and phenotype space have thousands of dimensions, hence a chemogenomics analysis of the underlying data is potentially able to unearth novel relationships between mechanistic and phenotypic space, aiding both drug discovery as well as the analysis of adverse drug reactions.

A similar analysis was performed from a different type of data[69], namely high-content screening data, where parts of the cell are stained, and then geometrical features of the cell are observed after administration of a compound [70,71]. The essentials of the analysis are shown in Fig. (**13**), relating ligand chemistry to the phenotypic response observed, as well as the protein targets predicted to be hit by the ligand. An analysis of this type relates a phenotypic observation to an explanation (mode of action hypothesis), with the chemical structure being the link between both. It can clearly be observed that all three types of information (chemistry, phenotype, biology) are important to consider when studying behavior of a compound, since none of them is correlated well enough to any other.

More recently, also the concept of applying multiple interventions in parallel to biological systems in a coordinated manner has found considerable interest, both in the pharmaceutical area [72-74] as well as in more fundamental research [75,76]. This is based partly on our increasing understanding of single interventions, now making partially also possible to engineer desired combinations of interventions, but even more so this is due to the realization that multiple targets are needed to modulate biological networks in the desired manner due to effects such as redundant mechanisms [77,78], that render a cell more stable in more adverse conditions. From the chemical side, the concept of ‘chemical genetics’ is of much relevance here [5], where the biological concepts of genetics (such as gene knock-out experiments) are mimicked by compound application. This is not meant to resemble genetics completely, and in fact chemical genetics experiments differ much from their genetic siblings, since they allow interventions in a dose- and time-dependent manner. Also, protein surfaces even of inhibited proteins are still available to mediate protein-protein contacts, which is not the case if a protein is not expressed at all in the first place.

From the experimental side, the discovery of ligands with the desired bioactivity profile requires novel techniques such as diversity-oriented synthesis (DOS) to firstly explore chemistry potentially active against the desired targets [79,80], followed by more conventional optimization towards the required bioactivity profile. However, as known from drug discovery, serendipity is likely to play a prominent role in the area for the foreseeable future [81].

CONCLUSIONS

We are currently witnessing an ever increasing amount of publicly accessible bioactivity data, and this is coinciding with concepts such as polypharmacology which are currently being recognized for their importance in pharmaceutical research in academia as well as industry. The question of how to design bioactive matter against a single target is non-trivial; it is easy to imagine that designing chemistry with the

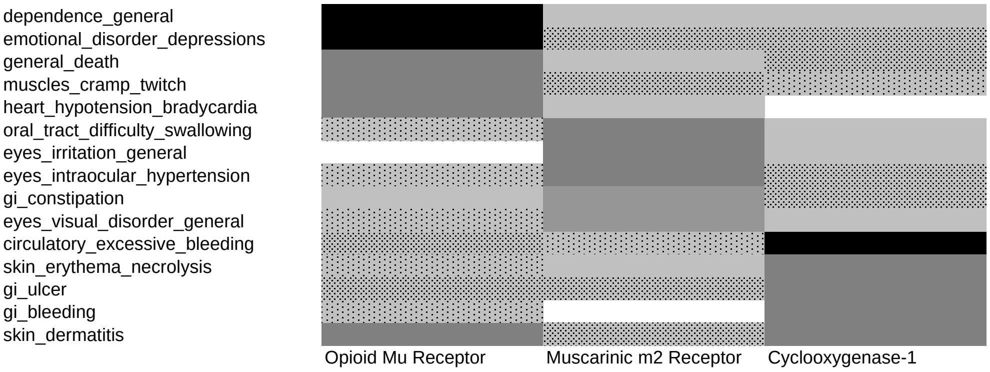


Fig. (12). Correlation between ligand chemistry active against proteins and the ligand chemistry causing adverse drug reactions (phenotypic information; darker colours indicate higher correlation). Both target space and phenotype space have thousands of dimensions, hence a chemogenomics analysis of the underlying data is potentially able to unearth novel relationships between mechanistic and phenotypic space, aiding both drug discovery as well as the analysis of adverse drug reactions.

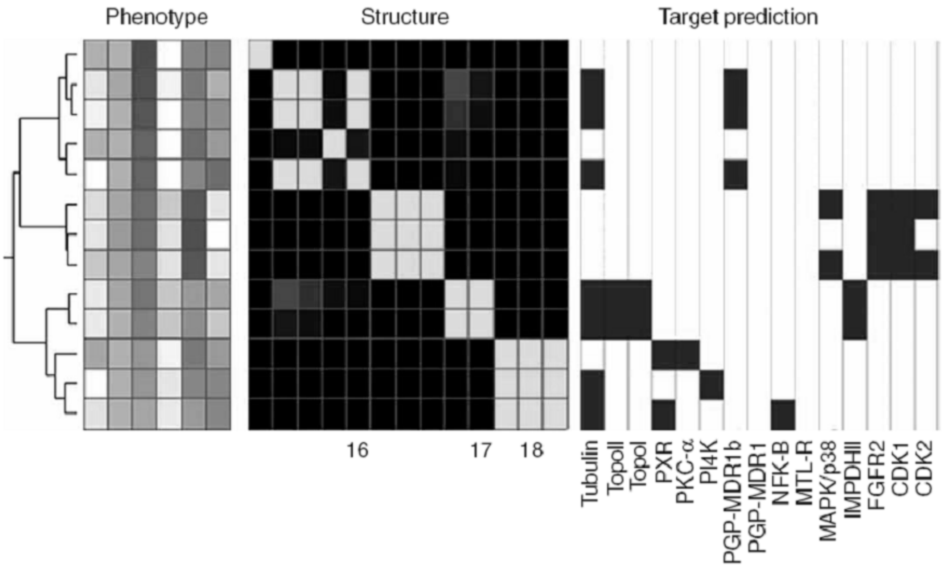


Fig. (13). Analysis of ligand chemical space, in relation to the phenotypic response observe in a microscopy-based high-content screen, as well as the protein targets predicted to be hit by the ligand (ligand numbers taken from the original publication[69]). An analysis of this type relates a phenotypic observation to an explanation (mode of action hypothesis), with the chemical structure being the link between both. It can clearly be observed that all three types of information (chemistry, phenotype, biology) are important to consider when studying the behaviour of a compound, since none of them is correlated well enough to any other. (Reprinted with permission from [69].)

right bioactivity profile against multiple targets is even more demanding. Hence, it becomes clear that we need the chemogenomics data at our disposal today to make wiser decisions regarding the design of bioactive matter in the future, by employing algorithms that have partly been developed already, but which for the most part still have to be conceived in the future. We have tens of millions of bioactivity data points available today – now we have to develop ways to make proper use of them.

ACKNOWLEDGMENTS

EvdH, APIJ and AB thank Dutch Top Institute Pharma (TI Pharma) for support of project number D1-105 during the receptor deorphanization and chemical representations parts of this work. AB thanks the Education Office of the Novartis Institutes for BioMedical Research for supporting the chemical representation and phenotype work of this manuscript and current funding. GJPvW thanks Tibotec for funding. DRS and WRJDG thank the EPSRC, BBSRC, MRC, Wellcome Trust and Francis and Augustus Newman Foundation for funding.

REFERENCES

[1] Bredel M.; Jacoby, E. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.*, **2004**, *5*, 262-275.

[2] Jacoby E. A novel chemogenomics knowledge-based ligand design strategy - Application to G protein-coupled receptors. *Quant. Struct.-Act. Relat.*, **2001**, *20*, 115-123.

[3] Caron P. R.; Mullican M. D.; Mashal R. D.; Wilson K. P.; Su M. S. Murcko M. A. Chemogenomic approaches to drug discovery. *Curr. Opin. Chem. Biol.*, **2001**, *5*, 464-470.

[4] Schuffenhauer A.; Floersheim P.; Acklin P. Jacoby E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 391-405.

[5] O' Connor, C. J.; Laraia, L.; Spring D. R. Chemical genetics. Chem. Soc. Rev., 2011, 40, DOI: 10.1039/C1CS15053G.

[6] Stockwell B. R. Chemical genetics: Ligand-based discovery of gene function. *Nat. Rev. Genet.*, **2000**, *1*,116-125.

[7] Wishart D. S.; Knox C.; Guo A. C.; Shrivastava S.; Hassanali M.; Stothard P.; Chang Z. Woolsey J. DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucl. Acids Res.*, **2006**, *34*, D668-672.

[8] Liu T.; Lin Y.; Wen X.; Jorissen R. N.; Gilson M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucl. Acids Res.*, **2007**, *35*, D198-201.

[9] Roth B. L.; Lopez E.; Beischel S.; Westkaemper R. B. Evans J. M. Screening the receptorome to discover the molecular targets for plant-derived psychoactive compounds: a novel approach for CNS drug discovery. *Pharmacol. Ther.*, **2004**, *102*, 99-110.

[10] PubChem - Part of the NIH Molecular Libraries Roadmap Initiative. http://pubchem.ncbi.nlm.nih.gov/.

[11] http://chembank.broad.harvard.edu/. C. ChemBank. http://chembank.broad.harvard.edu/.

[12] Williams A. J. Public chemical compound databases. *Curr. Opin. Drug Discov. Devel.*, **2008**, *11*, 393-404.

[13] Jenkins J. L.; Bender A., Davies J. W. *In silico* target fishing: Predicting biological targets from chemical structure. *Drug Discov. Today Technol.*, **2007**, *3*, 413-421.

[14] ChEMBL database, accessible at http://www.ebi.ac.uk/chembl/.

[15] Bender A. Databases: Compound bioactivities go public. *Nat. Chem. Biol.*, **2010**, *6*, 309-309.

[16] Kuhn M.; Campillos M.; Letunic I.; Jensen L. J. Bork P. A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.*, **2010**, *6*, 343.

[17] Mestres J.; Gregori-Puigjane E.; Valverde S.; Sole R. V. The topology of drug-target interaction networks: implicit dependence on drug properties and target families. *Mol. Biosyst.*, **2009**, *5*, 1051-1057.

[18] Mestres J.; Gregori-Puigjane E.; Valverde S.; Sole R. V. Data completeness--the Achilles heel of drug-target networks. *Nat. Biotechnol.*, **2008**, *26*, 983-984.

[19] Csermely P.; Agoston V. Pongor S. The efficiency of multi-target drugs: the network approach might help drug design. *Trends Pharmacol. Sci.*, **2005**, *26*, 178-182.

[20] Morphy R.; Rankovic Z. Designed multiple ligands. An emerging drug discovery paradigm. *J. Med. Chem.*, **2005**, *48*, 6523-6543.

[21] Bender A.; Jenkins J. L.; Glick M.; Deng Z.; Nettles J. H., Davies J. W. “Bayes Affinity Fingerprints” improve retrieval rates in virtual screening and define orthogonal bioactivity space: When are

multitarget drugs a feasible concept? *J. Chem. Inf. Model.*, **2006**, *46*, 2445-2456.

[22] Paolini G. V.; Shapland R. H.; van Hoorn W. P.; Mason J. S.; Hopkins A. L. Global mapping of pharmacological space. *Nat. Biotechnol.*, **2006**, *24*, 805-815.

[23] Fliri A. F.; Loging W. T.; Thadeio P. F.; Volkmann R. A. Analysis of drug-induced effect patterns to link structure and side effects of medicines. *Nat. Chem. Biol.*, **2005**, *1*, 389-397.

[24] Keiser M. J.; Roth B. L.; Armbruster B. N.; Ernsberger P.; Irwin J. J. Shoichet B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, **2007**, *25*, 197-206.

[25] Gregori-Puigjane E. Mestres J. A ligand-based approach to mining the chemogenomic space of drugs. *Comb Chem High Throughput Screen*, **2008**, *11*, 669-676.

[26] 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. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.*, **1995**, *2*, 107-118.

[27] Briem H.; Lessel U. *In vitro* and *in silico* affinity fingerprints: Finding similarities beyond structural classes. *Perspect. Drug Discov. Des.*, **2000**, *20*, 231-244.

[28] Strombergsson H.; Daniluk P.; Kryshtafovych A.; Fidelis K.; Wikberg J. E.; Kleywegt G. J. Hvidsten T. R. Interaction model based on local protein substructures generalizes to the entire structural enzyme-ligand space. *J. Chem. Inf. Model.*, **2008**, *48*, 2278-2288.

[29] Roche O.; Schneider P.; Zuegge J.; Guba W.; Kansy M.; Alanine A.; Bleicher K.; Danel F.; Gutknecht E. M.; Rogers-Evans M.; Neidhart W.; Stalder H.; Dillon M.; Sjogren E.; Fotouhi N.; Gillespie P.; Goodnow R.; Harris W.; Jones P.; Taniguchi M.; Tsujii S.; von der Saal W.; Zimmermann G. Schneider G. Development of a virtual screening method for identification of "frequent hitters" in compound libraries. *J. Med. Chem.*, **2002**, *45*, 137-142.

[30] Mestres J. Computational chemogenomics approaches to systematic knowledge-based drug discovery. *Curr. Opin. Drug Discov. Devel.*, **2004**, *7*, 304-313.

[31] Rognan D. Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.*, **2007**, *152*, 38-52.

[32] Klabunde T. Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br. J. Pharmacol.*, **2007**, *152*, 5-7.

[33] Marechal E. Chemogenomics: a discipline at the crossroad of high throughput technologies, biomarker research, combinatorial chemistry, genomics, cheminformatics, bioinformatics and artificial intelligence. *Comb. Chem. High Throughput. Screen*, **2008**, *11*, 582.

[34] Jacoby E. Mozzarelli A. Chemogenomic strategies to expand the bioactive chemical space. *Curr. Med. Chem.*, **2009**, *16*, 4374-4381.

[35] Bender A.; Young D. W.; Jenkins J. L.; Serrano M.; Mikhailov D.; Clemons P. A.; Davies J. W. Chemogenomic data analysis: prediction of small-molecule targets and the advent of biological fingerprint. *Comb. Chem. High-Throughput Screen*, **2007**, *10*, 719-731.

[36] Doddareddy M. R.; van Westen G. J. P.; van der Horst E.; Peironcely J. E.; Corthals F.; Ijzerman A. P.; Emmerich M.; Jenkins J. L.; Bender A. Chemogenomics: Looking at Biology through the Lens of Chemistry. *Stat. Anal. Data Mining*, **2009**, *2*, 149-160.

[37] Vieth M.; Higgs R. E.; Robertson D. H.; Shapiro M.; Gragg E. A. Hemmerle H. Kinomics-structural biology and chemogenomics of kinase inhibitors and targets. *Biochim. Biophys. Acta*, **2004**, *1697*, 243-257.

[38] Karaman M. W.; Herrgard S.; Treiber D. K.; Gallant P.; Atteridge C. E.; Campbell B. T.; Chan K. W.; Ciceri P.; Davis M. I.; Edeen P. T.; Faraoni R.; Floyd M.; Hunt J. P.; Lockhart D. J.; Milanov Z. V.; Morrison M. J.; Pallares G.; Patel H. K.; Pritchard S.; Wodicka L. M.; Zarrinkar P. P. A quantitative analysis of kinase inhibitor selectivity. *Nat. Biotechnol.*, **2008**, *26*, 127-132.

[39] Bantscheff M.; Eberhard D.; Abraham Y.; Bastuck S.; Boesche M.; Hobson S.; Mathieson T.; Perrin J.; Raida M.; Rau C.; Reader V.; Sweetman G.; Bauer A.; Bouwmeester T.; Hopf C.; Kruse U.; Neubauer G.; Ramsden N.; Rick J.; Kuster B.; Drewes G. Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nat. Biotechnol.*, **2007**, *25*, 1035-1044.

[40] Bamborough P.; Drewry D.; Harper G.; Smith G. K.; Schneider K. Assessment of chemical coverage of kinome space and its implications for kinase drug discovery. *J. Med. Chem.*, **2008**, *51*, 7898-7914.

[41] Bernasconi P.; Chen M.; Galasinski S.; Popa-Burke I.; Bobasheva A.; Coudurier L.; Birkos S.; Hallam R.; Janzen W. P. A Chemogenomic analysis of the human proteome: Application to enzyme families. *J. Biomol. Screen.*, **2007**, *12*, 972-982.

[42] Jacob L.; Hoffmann B.; Stoven V.; Vert J. P. Virtual screening of GPCRs: an *in silico* chemogenomics approach. *BMC Bioinformatics*, **2008**, *9*, 363.

[43] Bock J. R.; Gough D. A. Virtual screen for ligands of orphan g protein-coupled receptors. *J. Chem. Inf. Model.*, **2005**, *45*,402-1414.

[44] Weill N.; Rognan D. Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands. *J. Chem. Inf. Model.*, **2009**, *49*, 1049-1062.

[45] Van der Horst E.; Peironcely J. E.; Ijzerman A. P.; Beukers M. W.; Lane J. R.; van Vlijmen H. W. T.; Emmerich M. T. M.; Okuno Y.; Bender A. A novel chemogenomics analysis of G protein-coupled receptors (GPCRs) and their ligands: a potential strategy for receptor de-orphanization. *BMC Bioinformatics*, **2010**, *11*, 316.

[46] Gloriam D. E.; Foord S. M.; Blaney F. E.; Garland S. L. Definition of the G protein-coupled receptor transmembrane bundle binding pocket and calculation of receptor similarities for drug design. *J. Med. Chem.*, **2009**, *52*, 4429-4442.

[47] Surgand J. S.; Rodrigo J.; Kellenberger E.; Rognan D. A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors. *Proteins*, **2006**, *62*, 509-538.

[48] van der Horst E.; Okuno Y.; Bender A.; AP I. J. Substructure mining of GPCR ligands reveals activity-class specific functional groups in an unbiased manner. *J. Chem. Inf. Model.*, **2009**, *49*, 348-360.

[49] Bender A.; Jenkins J. L.; Scheiber J.; Sukuru S. C.; Glick M.; Davies J. W. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J. Chem. Inf. Model.*, **2009**, *49*, 108-119.

[50] Bender A.; Glen R. C. A Discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication. *J. Chem. Inf. Model.*, **2005**, *45*, 1369-1375.

[51] Bender A.; Glen R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.*, **2004**, *2*, 3204-3218.

[52] Bender A.; Scheiber J.; Glick M.; Davies J. W.; Azzaoui K.; Hamon J.; Urban L.; Whitebread S. Jenkins J. L. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure *ChemMedChem*, **2007**, *2*, 861-873.

[53] Kazius J.; Nijssen S.; Kok J.; Back T. Ijzerman A. P. Substructure mining using elaborate chemical representation. *J. Chem. Inf. Model.*, **2006**, *46*, 597-605.

[54] Kazius J.; McGuire R. Bursi R. Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.*, **2005**, *48*, 312-320.

[55] van Westen G. J. P.; Wegener J. K.; Ijzerman A. P.; Van Vlijmen H. W. T.; Bender A. Proteochemometric modeling as a tool for designing selective compounds and extrapolating to novel targets. *Med. Chem. Comm.*, **2011**, *2*, 16-30

[56] Faulon J. L.; Misra M.; Martin S.; Sale K., Sapra R. Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor. *Bioinformatics*, **2008**, *24*, 225-233.

[57] Lapinsh M.; Prusis P.; Gutcaits A.; Lundstedt T., Wikberg J. E. S. Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. *Biochim. Biophys. Acta-Gen. Subj.*, **2001**, *1525*, 180-190.

[58] Lapinsh M.; Prusis P.; Lundstedt T. Wikberg J. E. S. Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands. *Mol. Pharmacol.*, **2002**, *61*, 1465-1475.

[59] Dragos H.; Gilles M. Alexandre V. Predicting the predictability: a unified approach to the applicability domain problem of QSAR models. *J. Chem. Inf. Model.*, **2009**, *49*, 1762-1776.

[60] Tetko I. V.; Sushko I.; Pandey A. K.; Zhu H.; Tropsha A.; Papa E.; Oberg T.; Todeschini R.; Fourches D.; Varnek A. Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.*, **2008**, *48*, 1733-1746.

[61] Weaver S.; Gleeson M. P. The importance of the domain of applicability in QSAR modeling. *J Mol. Graph. Model.*, **2008**, *26*, 1315-1326.

[62] Schroeter T.; Schwaighofer A.; Mika S.; Laak A. T.; Suelzle D.; Ganzer U.; Heinrich N.; Muller K. R. Machine learning models for lipophilicity and their domain of applicability. *Mol. Pharm.*, **2007**, *4*, 524-538.

[63] Jaakola V. P.; Griffith M. T.; Hanson M. A.; Cherezov V.; Chien E. Y.; Lane J. R.; Ijzerman A. P.; Stevens R. C. The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science*, **2008**, *322*, 1211-1217.

[64] Rogers D., Hahn M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, **2010**, *50*, 742-754.

[65] Glen R. C.; Bender A.; Arnby C. H.; Carlsson L.; Boyer S.; Smith J. Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs*, **2006**, *9*, 199-204.

[66] Nettles J. H.; Jenkins J. L.; Bender A.; Deng Z.; Davies J. W.; Glick M. Bridging chemical and biological space: "target fishing" using 2D and 3D molecular descriptors. *J. Med. Chem.*, **2006**, *49*, 6802-6810.

[67] 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. Understanding False Positives in Reporter Gene Assays: *in silico* Chemogenomics Approaches to Prioritize Cell-based HTS Data. *J. Chem Inf. Model.*, **2007**, *47*, 1319-1327.

[68] Campillos M.; Kuhn M.; Gavin A. C.; Jensen L. J.; Bork P. Drug target identification using side-effect similarity. *Science*, **2008**, *321*, 263-266.

[69] 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. Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.*, **2008**, *4*, 59-68.

[70] Feng Y.; Mitchison T. J.; Bender A.; Young D. W. Tallarico J. A. Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds. *Nat. Rev. Drug Discov.*, **2009**, *8*, 567-578.

[71] Kummel A.; Gabriel D.; Parker C. N.; Bender A. Computational methods to support high-content screening: from compound selection and data analysis to postulating target hypotheses. *Exp. Op. Drug Disc.*, **2009**, *4*, 5-13.

[72] Jia J.; Zhu F.; Ma X.; Cao Z.; Li Y.; Chen Y. Z. Mechanisms of drug combinations: interaction and network perspectives. *Nat. Rev. Drug Discov.*, **2009**, *8*, 111-128.

[73] Lehar J.; Krueger A. S.; Zimmermann G. R.; Borisy A. A. Therapeutic selectivity and the multi-node drug target. *Discov. Med.*, **2009**, *8*,185-190.

[74] Lehar J.; Krueger A. S.; Avery W.; Heilbut A. M.; Johansen L. M.; Price E. R.; Rickles R. J.; Short G. F^{3rd}.; Staunton J. E.; Jin X.; Lee M. S.; Zimmermann G. R.; Borisy A. A. Synergistic drug combinations tend to improve therapeutically relevant selectivity. *Nat. Biotechnol.*, **2009**, *27*, 659-666.

[75] Lehar J.; Stockwell B. R.; Giaever G.; Nislow C. Combination chemical genetics. *Nat. Chem. Biol.*, **2008**, *4*, 674-681.

[76] Nelander S.; Wang W.; Nilsson B.; She Q. B.; Pratilas C.; Rosen N.; Gennemark P.; Sander C. Models from experiments: combinatorial drug perturbations of cancer cells. *Mol. Syst. Biol.*, **2008**, *4*, 216.

[77] Kitano H. A robustness-based approach to systems-oriented drug design. *Nat. Rev. Drug Discov.*, **2007**, *6*, 202-210.

[78] Kartal O.; Ebenhoh O. Ground state robustness as an evolutionary design principle in signaling networks. *PLoS One*, **2009**,*4*,e8001.

[79] Spandl R. J.; Bender A.; Spring D. R. Diversity-oriented synthesis; a spectrum of approaches and results. *Org. Biomol. Chem.*, **2008**, *6*, 1149-1158.

[80] Fergus S.; Bender A.; Spring D. R. Assessment of structural diversity in combinatorial synthesis. *Curr. Op. Chem. Biol*, **2005**, *9*, 304-309.

[81] Kubinyi H. Chance favors the prepared mind--from serendipity to rational drug design. *J. Recept. Signal. Transduct. Res.*, **1999**, *19*, 15-39.

Received: June 18, 2010 **Revised: August 22, 2010** **Accepted: September 02, 2010**