Herman Skolnik Award Symposium 2015

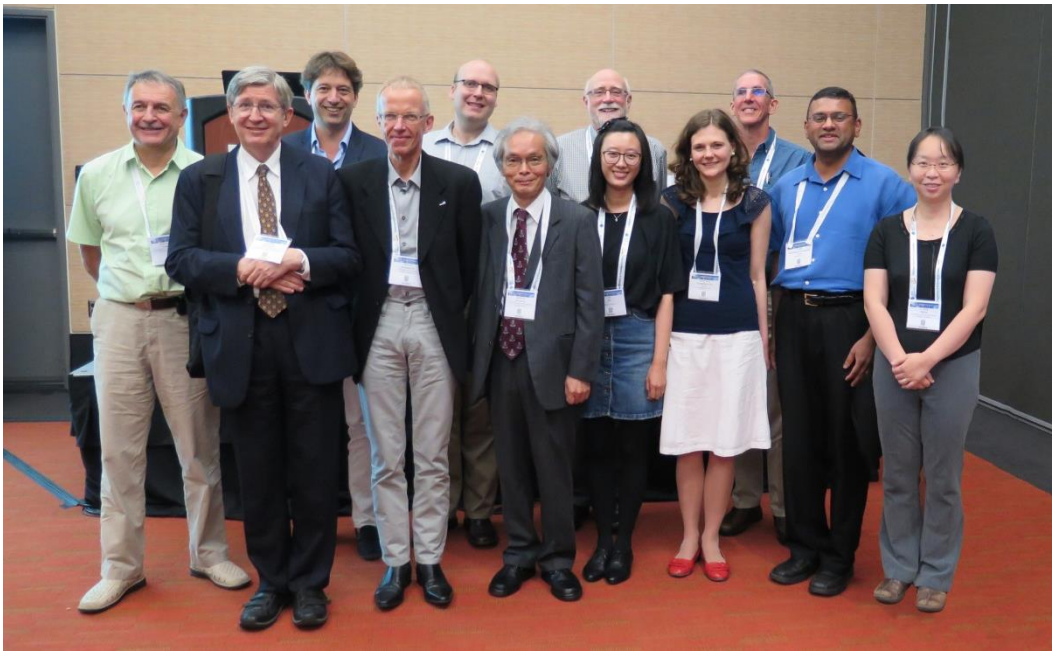
Honoring Jürgen Bajorath

A report by Wendy Warr (wendy@warr.com) for the ACS CINF *Chemical Information Bulletin*

Introduction

Veerabahu (Veer) Shanmugasundaram of Pfizer, who chaired the symposium, gave a brief introduction highlighting Jürgen's achievements. (A lengthier tribute has appeared at http://bulletin.acscinf.org/node/655.) Jürgen obtained his diploma (M.S.) and Ph.D. degrees (under Wolfram Saenger) in biochemistry from the Free University of Berlin. He then did postdoctoral studies with Arnie Hagler at Biosym in San Diego, focusing on DFT calculations of enzyme-inhibitor complexes. At Bristol-Myers Squibb he worked on protein modeling and structure-based design projects and developed his interests in bioinformatics and cheminformatics research. During his tenure at New Chemical Entities, he firmly established himself as a thought leader in cheminformatics. After 16 years in the United States, he returned to Germany where he is currently Professor and Chair of Life Sciences Informatics at the University of Bonn. Jürgen is a leader in the development and application of cheminformatics and computational solutions to research problems in medicinal chemistry, chemical biology and life sciences. He has done pioneering work in compound-centric data visualization and analysis in chemistry and is widely recognized for his seminal and prolific research work in several areas that are of interest to industry. His research interests include large-scale graphical SAR analysis, navigating high-dimensional space, multi-target modeling, machine learning and virtual screening.

The award symposium was divided into four sections. The first three speakers were "people Jürgen has looked up to": Tony Hopfinger, Gerry Maggiora and Peter Willett, all of them former Herman Skolnik Award winners. (Arnie Hagler was also to have been in this group but he was unable to attend.) The next speakers were Jürgen's colleagues and peers: Alexandre Varnek, Kimito Funatsu, Gisbert Schneider, Pat Walters and Veerabahu Shanmugasundaram. They were followed by some of Jürgen's present and past students: Ye Hu, Eugen Lounkine, and Anne Mai Wassermann. Finally Jürgen himself gave the award address.



Front row, L to R: Alexandre Varnek, Peter Willett, Jürgen Bajorath, Kimito Funatsu, Ye Hu, Anne Mai Wassermann, Veerabahu Shanmugasundaram, Jane Tseng
Back row, L to R: Gisbert Schneider, Eugen Lounkine, Gerry Maggiora, Pat Walters

Receptor-independent ligand activity models and receptor-dependent activity models

Jane Tseng presented the first talk on behalf of Tony Hopfinger of the University of New Mexico, who was unable to attend. In developing predictive methods to construct ligand-receptor binding models, most often to estimate IC_{50} values in the format of QSAR models, contributions from the receptor have been neglected. In the beginning, when protein-ligand structures were not available, the original goal of 4D-QSAR analysis[1] was to develop a methodology to complement Comparative Molecular Field Analysis (CoMFA).[2] In CoMFA, descriptors are calculated as grid point interactions between a probe atom and the target molecules and only one conformation of each compound is considered, not a conformational ensemble profile, as in the 4D-QSAR method. A new use of 4D-QSAR is to permit the parsing of information content arising from receptor-independent (RI) ligand activity models, as opposed to receptor-dependent (RD) models. To what extent is an RI ligand activity model (i.e., classic QSAR) of value in drug design applications?

4D-QSAR includes the conformational flexibility and the freedom of alignment by ensemble averaging in the conventional 3D descriptors found in traditional 3D-QSAR methods. Thus, the "fourth dimension" of the method is ensemble sampling of the spatial features of the members of a training set. In this approach, the descriptors are the occupancy frequencies of the different atom types in the cubic grid cells during the molecular dynamics simulation time, according to each trial alignment, corresponding to an ensemble averaging of conformational behavior.[3,4] The grid cell occupancy descriptors (GCODs) are generated for a number of different atom types (e.g., nonpolar, hydrogen bond acceptor, aromatic), called interaction pharmacophore elements (IPEs). The variable selection is made using a genetic function algorithm (GFA).[5] Multiple good QSAR models can be generated in the GFA step and the best model has to be established.

The 4D-QSAR methodology can be used in a receptor-dependent (RD) mode when the geometry of the receptor is available. In the RD-QSAR analysis, models are derived from the 3D structure of the

multiple ligand-receptor complex conformations. This approach provides an explicit simulation of the induced-fit process, using the structure of the ligand-receptor complex, where both ligand and receptor are allowed to be completely flexible by the use of molecular dynamics simulation. RD-QSAR is used to gather binding interaction energies, as descriptors, from the interaction between the analogue molecules and the receptor.[6] The RD-4D-QSAR approach[7] employs a novel receptor-pruning technique to permit effective processing of ligands with the lining of the binding site wrapped about them. Data reduction, QSAR model construction, and identification of possible pharmacophore sites are achieved by a three-step statistical analysis consisting of genetic algorithm optimization followed by backward elimination, multidimensional regression and ending with another genetic algorithm optimization.[8]

The paradigm of 4D-QSAR analysis does appear to afford identical and comparative model development capabilities for both RI and RD studies. Both numeric and actual spatial pharmacophore subtractions of RI- and RD-QSAR models developed from training sets in which receptor information is available can be performed and a general assessment of lost design information in an RI study can be made.

Jane presented results for six ligand-receptor systems for which both an RI- and an RD-4D-QSAR analysis model had been constructed. She presented tentative conclusions from comparisons of the RI- and RD-4D-QSAR models and their pharmacophore sites (GCODs). The RD models are about the same "quality" (in terms of r^2 values) as the RI models, but usually have fewer GCOD terms. The RD models usually contain one or more ligand-receptor based GCODs, but receptor-only based GCODs are not common in the RD models. Eye-ball selected clusters of "common" GCODs account for about 50% to 80% of the variance explained by the RI and the RD models.

Tony speculates that for ligand-receptor pharmacophore-based QSAR models, 20% to 40% of the targeted information in an RD-QSAR model is different from that of its corresponding RI-QSAR model. There are no discernible differences in atom-types or GCOD occupancy values, but RD GCODS are found near receptor walls whereas RI GCODs are found in "open" receptor space which is often most occupied by ligand atoms. This type of comparison of RI- and RD-QSAR models is only possible for datasets where explicit ligand-to-receptor binding occurs, and the identical pharmacophore generating methodology must be used on both the RI dataset and the corresponding RD dataset. The major finding of a 20 to 40% difference in a RI-4D-QSAR model and its corresponding RD-4D-QSAR model may be at odds with Dick Cramer's recent success[9] in correctly predicting 12 ligands from an RI ligand-receptor model.

Non-specificity of drug-target interactions

Gerry Maggiora of the Translational Genomics Research Institute, Tucson, gave this talk. System complexity, non-specificity, and biological reductionism are issues confronting drug discovery. Complex systems, such as biosystems, weather systems, and traffic systems, have numerous interacting component parts and unpredictable behavior. They are non-computable and have emergent properties. Emergent properties arise out of more fundamental entities and yet are irreducible with respect to them.

Biological systems are structurally and functionally complex. The central dogma of molecular biology, DNA makes RNA makes protein, is an overly simplistic concept.[10] An organism's phenotype is influenced by genomic and epigenetic phenomena, the latter being linked to a variety of biological

sensors that are able to sense their environments and influence the functions the system can carry out. The discovery of microRNA revealed that part of "junk DNA" is actually transcribed by the machinery in cells into bits of RNA that are fundamental controllers of life.

Biological systems have a hierarchical structure: population, organism, organs, tissues, cells, organelles, molecules. The reductionist approach seeks to decompose biological systems into their constituent parts in an effort to understand the biology induced by these parts. Moving up the biological hierarchy, function is reintroduced and the size and complexity of the systems tend to increase.[11]

The notion of specificity in biological systems has a long history. Notions of specificity and reductionism led to the single-target hypothesis which is still a major model in drug discovery. Adverse drug reactions and repurposed drugs imply a greater lack of specificity in biosystems than is generally assumed. The emerging field of polypharmacology addresses the interaction of drugs with multiple targets. A published analysis of the drug-target network[12] suggests a need to update the single drug-single target paradigm. Many analyses of drug-target interactions have been reported.[13-20] There are also many drug-target databases.[21-27]

Data quality is an issue for these databases: the data are obtained by different methodologies and different experimental protocols in different laboratories, and drug-target interactions are predicted. The data are inconsistent within and among databases. The data may not be complete, that is, it may not relate all selected compounds to all selected targets,[15] and drug-target space (all compounds against all targets) may not be completely covered. So, how promiscuous are drugs and how much biology is affected by the introduction of a single drug?

In the drug discovery landscape, the organismal level can be related to the molecular level by highly complex empirical models through to simpler mechanistic models; a mechanistic model relating a molecule to an organism is less physiologically relevant than an empirical model. Empirical models relate to phenotypic screening, mechanistic models to target-oriented screening.

Target-oriented drug discovery requires well validated targets. Sufficient details of the full mechanism of action are generally lacking, but, on the plus side, target-oriented screening is generally amenable to a high-throughput format. Screening hits are typically more limited than those obtained in phenotypic screens; it is unlikely that "inactive" regions of chemical space will be considered further; SAR typically neglects interactions with other targets; and follow-on phenotypic screens are required to assess biological efficacy.

Gerry gave as an example the use of imatinib in chronic myeloid leukemia. The target is Bcr-Abl kinase, a constitutively active product of the BCR/ABL fusion gene. Imatinib binds to the ATP binding region of Bcr-Abl kinase. The STITCH 4.0 database, however, shows that imatinib interacts with at least 10 different proteins, accounting perhaps for 24 or more adverse drug reactions observed for imatinib. Gerry also showed a network of imatinib interactions with the 52 proteins that interact in any fashion with imatinib. The Drug2Gene database records 41 proteins associated with imatinib binding. BCR/ABL behavior is complex[28,29] and drug resistance, both *de novo* and secondary, is observed. It is a scientific aphorism that in an experiment it is difficult to find what you are *not* looking for.

Lessons can be learned from metabolic engineering: numerous metabolic engineering studies show that metabolic networks cannot be regulated by perturbing a single network component. Manipulating single genes or gene products does not affect phenotype; or the genes' influence on phenotype does not arise in a simple, obvious fashion. Introduction of any xenobiotic into a biosystem affects multiple, and in many cases diverse targets, so there is a significant degree of "mechanistic uncertainty" in target-oriented drug discovery.

Phenotypic screening[30] has thus re-emerged. Phenotypic methods, which rely much less on mechanistic details, can provide a more robust platform. They employ a phenomenological (empirical) approach that is function-based rather than mechanism-based. They are hypothesis-driven, and similar to statistically based systems models. Phenotype-based approaches are closer to "intrinsic biology" with increased likelihood of finding viable leads. They typically generate a greater number of diverse hits than target-based screens. Functional responses are ideally, but not always, related to disease states. Phenotype-based approaches are particularly useful in cases where the biology is not clearly understood. Phenotypic screens are inherently multi-target screens but are target- and mechanism-agnostic. Promiscuity may be a virtue in phenotypic screens. The use of high-throughput formats is limited, but improvements are on the way. Determining the mechanism of action may be an issue, and incorrect target determination can cause significant problems. Gerry made two final observations: if your only tool is a hammer, all problems begin to look like nails; and the bigger the hammer, the easier it is to pound the nails.

Molecular similarity approaches in cheminformatics

Peter Willett of the University of Sheffield outlined the early history of molecular similarity, and presented a bibliometric analysis. As Rouvray[31] noted "Similarity is ubiquitous in scope, interdisciplinary in nature, and seemingly boundless in its ramification". Mendeleev's 1869 discovery of the Periodic Table is often cited as the first example of similarity concepts in chemistry, but there are many other historical examples.[32] Computational measures of similarity are of great importance for cheminformatics, as a result of the "similar property principle", which states that structurally similar molecules have similar properties. There are many exceptions to the principle but it is still a useful rule-of-thumb. It is generally ascribed to a book by Johnson and Maggiora,[33] but Johnson and Maggiora had earlier ascribed it to a 1980 work by Wilkins and Randic.[34] The principle was in fact widely understood, even if not expressed in explicit form, much earlier than that, all the way back to 1868.[35] Analogous similarity relationships in geography[36] and social networks[37] have been referenced in recent publications in cheminformatics, the latter in studies of chemical space networks by Jürgen Bajorath. The cluster hypothesis underlying document clustering[38] spurred Peter's own studies of chemical clustering, given the analogies between cheminformatics and information retrieval.

The similarity principle provides not only a rationale for using similarity techniques in cheminformatics but also a way of validating them, for example in comparison of measures for similarity searching where benchmark datasets of actives and inactives are used to evaluate the relative effectiveness of different measures on the extent to which nearest neighbors of known actives are also active. There are analogous validation approaches in clustering and diversity applications, for example, all the molecules in a given cluster should have broadly similar properties.

The earliest example of clustering chemical databases was work[39] at ICI Pharmaceuticals Division in which fragment-based similarities were used to cluster around a known active if there were at least some number of nearest neighbors above a similarity threshold. Common structural features in such clusters were identified. Adamson and Bush[40,41] were the first to use 2D substructure searching features in a comparison of the effectiveness of similarity measures for single-linkage clustering. Fingerprint-based measures are still the most common 40 years later.

At Sheffield University extensive comparative studies of a wide range of similarity measures and clustering methods were carried out by Willett and Winterman,[42-44] using the Adamson-Bush evaluation procedures. Fragment occurrences were found to be slightly better than incidences. The Tanimoto coefficient was found to be the most effective coefficient of those tested, and it is still the standard for similarity applications in cheminformatics. Ward's hierarchic agglomerative method[45] has since proved to be the preferred clustering method, but the non-hierarchic, nearest-neighbor method of Jarvis and Patrick[46] was for years a cost-effective alternative, given the algorithmic complexity of Ward's method.

The use of the similar property principle for ligand-based virtual screening was initially studied in the mid-1980s at Lederle Laboratories,[47] the Upjohn Company, and Pfizer in the United Kingdom together with Sheffield University.[44,48] The use of substructure-searching fragments and simple association coefficients is effective and efficient in operation, and is a simple enhancement of existing database software; there was therefore a rapid take-up, and 30 years later this is still a standard approach to virtual screening. Many other 2D and 3D approaches are now available, but they are still less widely used. Perhaps the main enhancement since the initial work is the use of data fusion methods, as first studied at Merck,[49-51] and at Sheffield.[52-55]

Developments in combinatorial chemistry and high-throughput screening in the early 1990s spurred interest in the selection of diverse sets of compounds, [56,57] but work on compound selection had been undertaken several years previously at Upjohn, and at Pfizer in the United Kingdom together with Sheffield University, based directly on the similarity measures that had been developed previously for clustering and similarity searching. Methods included cluster-based selection,[43] and dissimilarity-based selection to optimize a diversity index.[48,58] The latter, using the Kennard-Stone algorithm, is now widely implemented as MaxMin.[59]

Peter concluded his talk with a bibliometric analysis of the literature of molecular similarity, as reflected in the Web of Science database. He found 86,663 citations to 2,980 articles on molecular similarity, with an *h*-index of 114 and a mean of 29.1 citations per article. The distribution of author contributions is highly skewed: Jürgen Bajorath is the most prolific author, with 95 of the 2,980 articles (Peter himself is close behind with 88), but there are 6,579 singleton contributions. As regards organizations, Sheffield has published largest number of articles (111), but there are 1,014 singletons amongst the 1,767 distinct organizations. Ten organizations, including five private-sector ones (AstraZeneca, GlaxoSmithKline, Merck, Novartis, and Pfizer) have 50 or more articles. Thirty of the 2,980 articles have 250 or more citations; the top five are by Allen *et al.*,[60], with 1400 citations, Klebe *et al.*,[61] Willett *et al.*,[62] Tropsha *et al.*,[63] and Bemis and Murcko.[64]

The citations appeared in 3,977 distinct publications, with the most frequent being *J. Chem. Inf. Model.* (2724), *J. Med. Chem.* (2075), *Bioorg. Med. Chem.* (1019), *Bioorg. Med. Chem. Lett.* (986), *J. Comput.-Aided Mol. Design* (734), *Eur. J. Med. Chem.* (695), *Mol. Inf.* (645), *J. Mol. Graphics Modell.*

(461), *PLoS One* (429), and *J. Am. Chem. Soc.* (372). The citing journals come from 202 distinct Web of Science subject categories: the methodological tools developed by the molecular similarity community are thus clearly of very broad applicability. Jürgen Bajorath has 11,037 citations to his 452 articles (120 of them in *J. Chem. Inf. Model*.) with an *h*-index of 48 and a mean of 24.4 citations per article. Of these, 16 have 100 citations or more, the top five[65-69] illustrating Jürgen's contributions to multiple fields of chemistry and the life sciences.

Generative topographic mapping

Alexandre (Sasha) Varnek, of the University of Strasbourg, France, described this tool for chemical space analysis. There are many ways of visualizing chemical space. In descriptor-based chemical space, where a D-dimensional vector represents each molecule, two popular approaches are used: similarity network graphs, and dimensionality reduction techniques which transfer the objects from the D-dimensional chemical space into a latent space of 2 or 3 dimensions.

Principal component analysis (PCA) and self-organizing Kohonen maps (SOM) are commonly used for exploration of large chemical spaces but both have drawbacks. PCA processes nonlinear data poorly. SOM is a nonlinear method and due to its topology-preserving character, it provides more information-rich plots than PCA, but it suffers from its purely empirical nature and it lacks solid statistical foundations.

Generative Topographic Mapping (GTM)[70,71] is a probabilistic extension of SOM. GTM relates the latent space with a 2D "rubber sheet" (or manifold*)* injected into the high-dimensional data space. The visualization plot is obtained by projecting the data points onto the manifold and then letting the rubber sheet relax to its original form. GTM generates a data probability distribution in both initial and latent data spaces. GTM can thus be used not only to visualize the data, but also for structure-property modeling tasks.[72]

Sasha showed a probability density distribution in the latent space. Projection of an object on a GTM is described by the probability distribution ("responsibilities") over the lattice nodes. Using GTM, one can, for each molecule, evaluate the probability of finding it in a point on the grid. There are two possibilities: one can use the responsibilities as molecular descriptors which can be used for predictions, or one can prepare an "activity landscape" to make predictions.

In the course of this project, Sasha's team has developed several utilities named ISIDA[73-75]/GTM (where ISIDA stands for *In Silico* Design and Data Analysis descriptors). They allow QSAR models to be created by GTM, and optimized and visualized, and the activity can be mapped. Chemical space maps can be used as a virtual screening tool. Sasha showed a GTM activity landscape of the stability of Lu^{3+} complexes with organic molecules;[76] strong and weak binders were clearly differentiated.

An activity landscape can be used directly to predict activities of test compounds using the distribution of responsibilities. In particular, in each node the product of activity landscape value for the training set and responsibility of the given test compounds is calculated by summation over all nodes of the map. It has been shown[76] that the performance of GTM-based regression models is similar to that obtained with four popular machine-learning methods (random forest, k-NN, M5P regression tree and PLS) and ISIDA fragment descriptors. By comparing GTM activity landscapes built both on predicted and experimental activities, one may visually assess the model's performance and identify the areas in the chemical space corresponding to reliable predictions.

Sasha reported some work on a GTM-based model's applicability domain.[77] The Biopharmaceutics Drug Disposition Classification System (BDDCS), based on solubility and degree of metabolism, is used by agencies such as FDA for granting biowaivers. Sasha and his co-workers have described the modeling in two-dimensional latent space for the four classes of the BDDCS using VolSurf descriptors. Three new definitions of the applicability domain (AD) of models were suggested: one class-independent AD which considers the GTM likelihood, and two class-dependent ADs considering either the predominant class in a given node of the map or informational entropy. The class entropy AD was found to be the most efficient for the BDDCS modeling. The predominant class AD can be directly visualized on GTM maps, which helps the interpretation of the model.

Sasha's team has also studied a database of more than 2 million compounds containing 37 subsets coming from catalogs of 36 chemical suppliers, and the NCI database. The researchers focused both on the parameters able to characterize the whole dataset, and on the analysis of individual libraries, to see how they covered the chemical space, to what extent they overlap, and which library has compounds possessing a particular activity profile. GTM incremental learning[78] is a solution for such large datasets.

Sasha showed a GTM of the entire database built on MOE descriptors. Each data point represented a molecule and the data were colored according to molecular weight. The left hand side of the map was populated by light molecules and the right-hand one by heavier molecules. Instead of using each data point, you can use a data density distribution function represented by the ensemble of cumulated responsibilities. The density maps can also be built for the individual libraries. You can color the same GTM map by different properties or activities to visualize different property landscapes. Superposition of different activity landscapes helps you to select areas populated by compounds with particular activity profiles. The data coverage can be measured by normalized Shannon's entropy calculated directly from the responsibilities. Surprisingly, the small ASINEX library covers the entire latent space more uniformly than the large Enamine library.

Sasha concluded with a few details of Stargate GTM (S-GTM), in which one GTM connects activity space and descriptor space. S-GTM can be used to predict a pharmacological profile and to discover structures corresponding to a given pharmacological profile. The method has been applied to a set of eight GPCR activities.

Development of a knowledge-generating platform from drug discovery through to production

Kimito Funatsu of the University of Tokyo described a knowledge-unifying platform driven by big data. While massive amounts of quantitative data have accumulated across the pipeline of drug discovery, all the way from a candidate's initial discovery up to its production process, knowledge of and data analysis for each of the discovery and production processes has remained isolated. The big data in Kimito's project consist of a large virtual library containing chemical structures of drug candidates, interaction data between many proteins and many drug candidates, and plant operating data and product quality data. The objectives are: automated generation of a huge virtual library, discovery of new drugs, and acquisition of synthetic routes from the library; construction of a mathematical model derived from many proteins versus many compounds together with other biological information, and extraction of a guide for drug discovery; and knowledge extraction for

process monitoring and control, plus development of the automated construction of a soft sensor model and a model maintenance system for process monitoring.

Prof. Okuno's group at Kyoto University is working on ligand-target information. Problems in the threefold, chemical-target-phenotype model of drug discovery include a shortage of experimental compound-protein interaction data, and compound-phenotype association data; and a lack of information on direct associations between target protein and phenotype. From mathematical models (logistic regression, PLS and SVM), predictions can help to fill the gaps. In previous work for predicting compound-protein interactions using information about chemical structures and protein sequences, the researchers used SVM, which trains up to 250,000 interactions but it is hard to learn larger scale data because of memory and computation time limits. They are trying to apply deep learning to train millions of interactions. In the prediction of protein-phenotype associations, they have compared the performance of PLS and SVM with that of logistic regression. They have demonstrated useful accuracy and high speed, but the number of proteins with positive weights is limited. In future work they aim for large-scale prediction of associations for all possible combinations between compounds and phenotypes, and they plan interaction prediction using deep learning, learning from a bigger dataset of interactions, and interpreting a deep belief network derived from big data.

Dr. Taiji's group at RIKEN is working on a very large scale virtual library (billions of compounds) with a synthetic route for all compounds, for assessment of synthetic feasibility. The massive generation of chemical structures using transformational rules involves rewriting of the transform-oriented synthesis planning (TOSP) generator[79] to allow parallel-processing, and validation of the transform and fragment data.

Kimito's part of the joint project concerns a soft sensor for monitoring and controlling a chemical plant. In chemical plants, efficient and stable production is required, keeping the quality of chemicals high. Operators have to monitor the operating condition of the plants and control process variables. NIR spectra, temperature, and pressure are easy to measure online. Concentration and density are difficult to measure online[80] and are predicted in this project, using a statistical model, from the NIR spectra, temperature, and pressure input to the sensor. Until recently, application of soft sensors online has not been possible because of low predictive accuracy and complex maintenance of the sensor.

Problems in soft sensor analysis include data reliability and selection; outlier detection and noise treatment; deciding on an appropriate regression method; overfitting; nonlinearity among process variables; variable selection; dynamics in the modeling process; model interpretation; model validation; applicability domain and predictive accuracy; model degradation; model maintenance; and detection and diagnosis of abnormal data. The predictive ability of soft sensors depends on the quality of database, but the amount of data in such a database is limited, so database monitoring is essential for highly predictive soft sensors. Data measured in plants are not fully exploited in process control. Soft sensors express relationships between process variables, so an efficient control method using a soft sensor model is required.

Since the predictive performance of adaptive models depends on databases, Kimito's group has proposed a database monitoring index (DMI),[81] to monitor the database and a database monitoring method using the DMI. The DMI proposed is based on similarity between two data. The more similar

two data are, the smaller DMI is. New data are stored when the minimum DMI value exceeds a threshold. Through the analysis of simulation data and real industrial data, the researchers have confirmed that databases can be appropriately managed and the predictive accuracy of adaptive soft sensor models increased by using the proposed method.

The three research groups aim to establish a platform which allows them to unify knowledge about different processes, and to advance research into improved and optimized systems that view pharmaceutical development from a comprehensive, correlated, and high-level perspective.

Enabling drug discovery by computational molecular design

Gisbert Schneider, of ETH, Zürich, Switzerland, gave a talk on *de novo* drug design and target prediction. The computer-based design of drug candidates is a complementary approach to high-throughput screening; *de novo* design[82] supports drug discovery projects by generating novel pharmaceutically active agents with desired properties in a cost- and time-efficient manner. An example is the identification of novel cannabinoid-1 receptor inverse agonists for the treatment of obesity.[83] A recent publication[84] reviews software for *de novo* drug design with a special emphasis on fragment-based techniques that generate druglike, synthetically accessible compounds.

The software Design of Genuine Structures (DOGS)[85,86] features a ligand-based strategy for automated *in silico* assembly of potentially novel bioactive compounds. The construction procedure explicitly considers compound synthesizability, based on a compilation of 25,144 available synthetic building blocks and 58 established reaction principles, with 25 regioselective variants. This enables the software to suggest a synthesis route for each designed compound. The quality of the designed compounds is assessed by a graph kernel method[87,88] measuring their similarity to known bioactive ligands in terms of structural and pharmacophoric features. Virtual intermediates are compared with a template. The scoring method does not just rely on substructure similarity, and the pharmacophore comparison is very permissive compared with graph similarity. The origin of the idea was patent beating.

Combinatorial *de novo* design can also be coupled with microfluidic synthesis and analytics.[89-91] Gisbert's team has recently reported[92] a multi-objective *de novo* design study driven by synthetic tractability and aimed at the prioritization of computer-generated 5-HT_{2B} receptor ligands with accurately predicted target-binding affinities. Gaussian process models were built for 974 proteins annotated in ChEMBL, and the team designed and synthesized structurally novel, selective, nanomolar, and ligand-efficient 5-HT_{2B} modulators. The results suggest that amalgamation of computational activity prediction and molecular design with microfluidics-assisted synthesis enables the swift generation of small molecules with the desired polypharmacology. In another example, Fasudil, a not very active, but ligand efficient Rho-kinase inhibitor was used as a template in DOGS to design a fragment-like candidate that was made and tested, and could be grown into Azosemide, approved for treatment of hypertension in Japan.

Gisbert's team has also developed an approach to target prediction. Several computational tools for predicting macromolecular targets of new chemical entities were publicly available, but none of these methods was explicitly designed to predict target engagement by *de novo* designed molecules, so the researchers devised self-organizing map-based prediction of drug equivalence relationships (SPiDER),[93] that merges the concepts of self-organizing maps, consensus scoring, and statistical analysis to identify targets for both known drugs and computer-generated molecular scaffolds. Some

15,000 drugs and druglike compounds are used as the basis for clustering and 11 targets per compound are predicted on average. The approach results in confident predictions.

The targets of natural products are largely unknown, which hampers rational drug design and optimization. Gisbert's team has developed and validated a computational method for the discovery of such targets. The technique does not require three-dimensional target models and may be applied to structurally complex natural products. The algorithm dissects the natural products into fragments and infers potential targets by comparing the fragments to drugs with known targets. Kohonen self-organizing maps and chemically advanced template search (CATS) topological pharmacophores[94] are used.[95,96]

Of the 210,213 structures in the *Dictionary of Natural Products*, 31% are fragment-like and 69% have large structures. Gisbert's system confidently predicted targets for 36% of the fragment-like products and 22% of the large ones. Sparteine is a deadly class 1a Na+ channel blocker with high ligand efficiency. Gisbert predicted that it interacted with the kappa opioid receptor. The fragment-like, synthetically tractable structures goitrin, isomacroin and graveolinine were input to SPiDER for target inference. Five out the six confidently predicted targets were correct, unreported targets, and the molecules were profoundly dissimilar to the most similar reference compound. Graveolinine shows dual target engagement (5-HT_{2B} and COX2) and could lead to polypharmacological tool compounds for example, for migraine.[97] In a prospective validation, it has been shown that fragments of the potent antitumor agent archazolid A contain relevant information regarding its polypharmacology. Biochemical and biophysical evaluation confirmed the predictions.[96] These results obtained with SPiDER corroborate the practical applicability of the approach to natural product "de-orphaning". DOGS and SPiDER lead from complex natural products to synthesizable new chemical entities.

Integrating public data sources into the drug discovery workflow

Pat Walters of Vertex Pharmaceuticals discussed two examples of work carried out at his company. The first was high-throughput screening (HTS) data analysis. Bench scientists want to be able to find hints of SAR, that is, to identify scaffold classes and related classes, and visualize activity distributions. They want to find out what is known about the activity of the compound class from both internal data and the literature, and about the properties and pharmacokinetics of the class. They want additional information, if any, from the literature about patents, properties, and synthesis. Pat listed the guiding principles for a system that meets these requirements. The first is to keep things simple: analysis tools should be intuitive, and molecules should be organized in a "medicinal chemistry driven" fashion. The second is to make the results visually compelling with a data dashboard, and one click access to details. Above all, the system must enable answers to critical questions.

The workflow involves partitioning actives into scaffold classes; profiling each scaffold class according to activity distribution, emerging SAR, selectivity, properties and ADME, and literature information; and prioritizing scaffolds for further exploration in "analogue by catalog", and exploratory chemistry.

Pat ran through an example, with screenshots, of identifying three ring scaffolds, keeping the most frequently occurring scaffolds, and displaying them in the HTS Viewer, with or without molecule details. Related scaffolds are identified by scaffold similarity and arranged by similarity in the HTS

Viewer. Activity on the HTS target is viewed by way of boxplot distributions in which there is an adjustable activity cutoff; boxplots provide an easy comparison of related scaffolds. To evaluate selectivity, activity against other targets is studied. Users can perform a general or target class specific analysis. Selective series are identified from plots of number of active assays against numbers of assays tested. Users can drill down to activity details and compare activity and selectivity for related scaffolds by displaying boxplots alongside the activity scatterplots. "Thermometer plots" show distributions of properties related to ADME. These plots can be displayed in yet another column alongside the scaffolds, boxplots and activities.

Scientists want to answer many literature questions. What biological activities are known for this compound class? Have related compounds been in clinical trials? Is this compound class mentioned in patents? Is the class well characterized (by physical properties)? Has the synthesis of the class been reported? There are many external sources of biological activity, physical properties, synthesis, drug data, and patents. While these databases provide a wealth of information, the data are often not in a format that is easily accessible to the bench scientist. In addition, scientists may be unaware of these resources, or may not know how to access and integrate the data. While it is tempting simply to integrate large amounts of public data into in-house systems, software developers must be careful to inform, without overwhelming, the target audience.

Vertex applications, with substructure search included, link to SciFinder via the SciFinder API, and use an internal database for Reaxys, ChEMBL, and Thomson Reuters Integrity. Pat showed screenshots of the HTS Viewer links to Reaxys, ChEMBL, and Integrity data. In each case links allow the user to jump directly to the underlying data. Vertex and CAS collaborated to provide a direct link to SciFinder using the SciFinder API. The Vertex system also addresses numerous other considerations such as identification of potential false positives and negatives; compound purity; "efficiency" of hits; filtering out undesirable compounds from assays; replicates and statistics; and hit follow-up.

Pat's second example of work carried out at Vertex concerned patent informatics. IPedia is a platform for information sharing. Vertex used to have an in-house system for capturing data from the patent offices and chemical structures were entered manually. The release of SureChEMBL has changed all that, but unfortunately SureChEMBL's automated process extracts *all* structures including reagents, solvents and the like.

Can SureChEMBL be used to find interesting structures? Pat's team took 30 drug patents (based on work done at AstraZeneca),[98] and eliminated three which were not in ChEMBL. They looked if the drug structure was in the SureChEMBL curated set, and tried to develop heuristics to identify the key compounds. The number of structures per SureChEMBL patent was a minimum of 11, and a maximum of 916, with a median of 161. The drug structure was found in 19 of the 27 patents. Structures were classified as interesting if they were 0.8 similar to the drug, by Tanimoto coefficient and MDL keys, and as boring if they were less than 0.8 similar. There were 1598 interesting structures and 4357 boring ones. Descriptors for structures were frequency of occurrence in SureChEMBL, location in the patent, number of heavy atoms, molecular weight, and number of neighbors at Tanimoto 0.8. The team built a simple recursive partitioning model using the ctree method in the "party" package in R 3.0.2. Simple heuristics proved to be very effective (accuracy 0.91 and kappa 0.77). Neighbor counts identified interesting structures: interesting compounds have

more neighbors. This example again illustrates how public data can be used to advantage in drug discovery projects.

"Close-in" analogue prioritization using SAR matrices

Veer Shanmugasundaram described work done at Pfizer in collaboration with Jürgen Bajorath. Jürgen has published papers on heterogeneous SAR,[99] activity cliffs versus selectivity cliffs,[100] SAR monitoring using activity landscapes,[101] and molecular mechanism based network-like similarity graphs.[102] Visualizations include network-like similarity graphs, SAR matrices, ligand-target differentiation maps, and bipartite matched molecular series graphs. Veer's talk related to SAR matrices,[103] which are designed to highlight different SAR patterns in large compound data sets. They provide chemically intuitive organization of analogue series, and easy identification of activity cliffs, providing immediate suggestions for compound design.

The SAR matrix data structure organizes compound data sets according to structurally analogous matching molecular series in a format reminiscent of conventional R-group tables. An intrinsic feature of SAR matrices is that they contain many virtual compounds that represent unexplored combinations of core structures and substituents extracted from compound datasets on the basis of the matched molecular pair formalism. These virtual compounds are candidates for further exploration but are difficult, if not impossible to prioritize on the basis of visual inspection of multiple SAR matrices.

Pfizer therefore worked with Jürgen to develop a compound neighborhood concept as an extension of the SAR matrix data structure that makes it possible to identify preferred virtual compounds for further analysis. On the basis of well-defined compound neighborhoods, the potency of virtual compounds can be predicted by considering individual contributions of core structures and substituents from neighbors. SAR-rich matrices are prioritized based on SAR patterns, property variance, and size and dimension of matrices and confidence values can be included in the matrix visualization. In extensive benchmark studies, virtual compounds have been prioritized in different datasets on the basis of multiple neighborhoods yielding accurate potency predictions.[104]

A retrospective analysis was carried out using six large sets of different G-protein coupled receptor antagonists extracted from ChEMBL for which K_i values were available. Matrix-pattern-based, matrix pattern based weighted by similarity, and analysis of variance models were compared with Jürgen's nearest neighbor analysis. The prediction accuracy (r^2) was best for analysis of variance models (between 0.7 and 0.84). Depending on the algorithmic fragmentation scheme, single-cut matrices (i.e., one exocyclic bond in a compound is systematically deleted to yield key and value fragments), dual-cut (two exocyclic bonds are simultaneously deleted), and triple-cut matrices (three exocyclic bonds are deleted) are separately generated. Veer showed boxplots of the mean error and type of matrices on the ChEMBL datasets and scatterplots of the distribution of absolute error and neighborhood similarity. The SAR matrices can be adapted for visualization in Spotfire DXP. This environment offers a structure-data viewer, filters, dynamic interactive visualizations, and a direct connection to the Pfizer database.

In summary, a visual examination of SAR using an adaptation of SAR Matrices in DXP provides a way to view, mine and interrogate single, double and triple-cut matrices dynamically, and study SAR trends quickly. Several methods that prioritize virtual compounds "to fill" close-in analogue space ranging from nearest neighbor methods, and similarity weighting to ANOVA analysis all appear to

perform equally well. Predictions based on single-cut matrices are as valuable as those with more complex double- and triple-cut matrices.

The second part of Veer's talk concerned series progression. The problem was to see if Pfizer could develop some diagnostic methods to evaluate if they were adding SAR information as chemical series progressed, and to determine whether more "close-in" analogues should be made, or whether new lead series should be identified. The strategy was a chronological analysis of "SAR information content" using SAR matrices, and using that to distinguish productive and unproductive series.

SAR matrices with a minimum of two compounds per series and a minimum of two series per matrix were used. SAR matrices were classified as old if a matrix in the previous year had the same cores and real compounds, expanded if a matrix in the previous year was a subset (had a subset of cores and real compounds), and new if no matrix in the previous year was a subset. Veer showed a number of plots of series progression, and of raw discontinuity score against average potency. He concluded that monitoring changes in SAR information content in multiple series could provide some interesting diagnostics in evaluating series progression. Matrices with increased discontinuity are considered to provide rich SAR information. The appearance of new matrices with increased SAR discontinuity or expansion of current matrices provides clear signals to evaluate series progression.

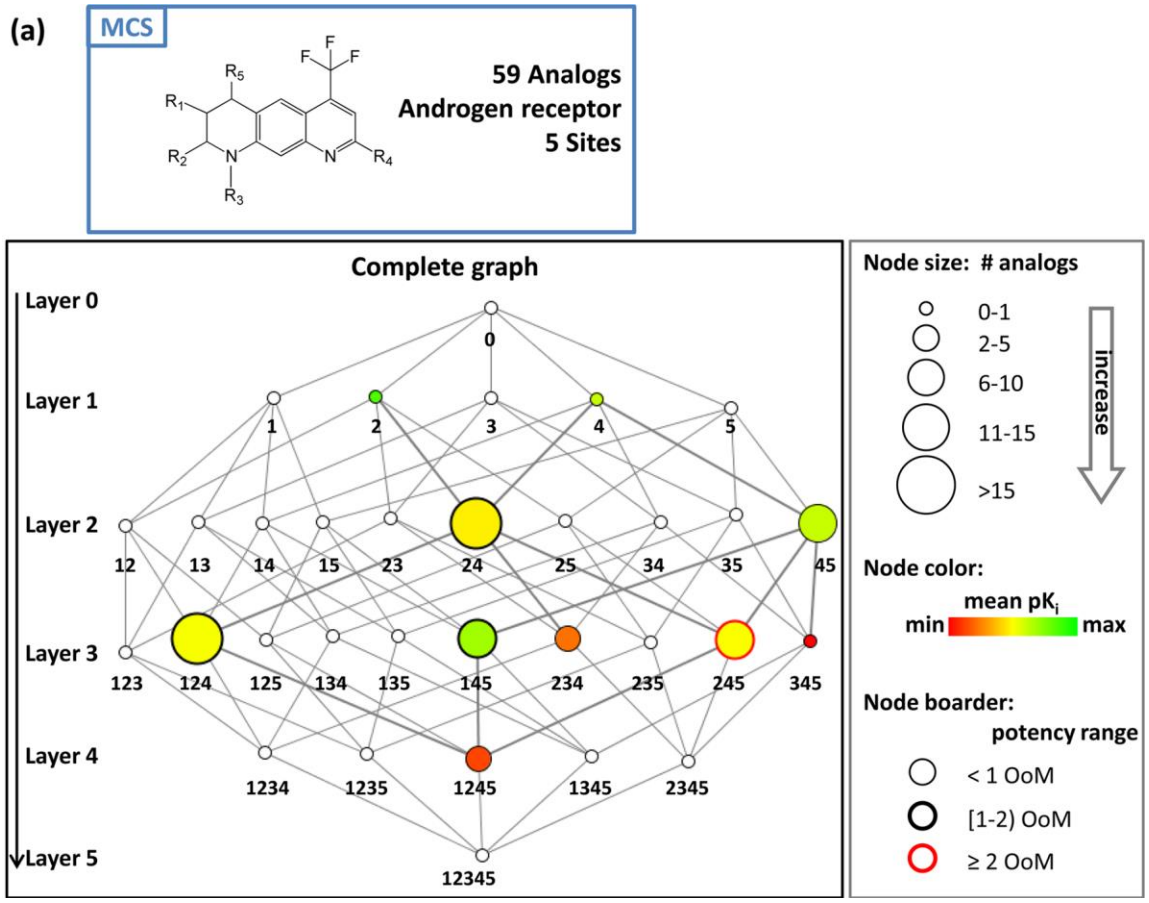
Graphical analysis of analogue series and associated SAR information

Ye (Pauline) Hu, one of Jürgen's current students in Bonn, presented AnalogExplorer. Analogue series are compounds sharing the same molecular scaffold or maximum common substructure (MCS). The conventional data format for them is a standard R-group table with all substituents and associated potency values. This is difficult to use for large and structurally heterogeneous series, so, as rapidly increasing amounts of SAR data become available, graphical approaches have been introduced to explore structure-activity relationships (SARs) contained in compound data sets. Exemplary MCS-based visualization methods include SAR maps, and the combinatorial analogue graph (CAG). In SAR maps analogous compounds contain substituents at two different sites. In the matrix format each cell represents a compound with corresponding substituents, and cells are colored by potency values.[105] In a CAG,[106] nodes are pairs of compounds with variations at one, two, or maximally three sites, colored by SAR discontinuity scores. Edges are the relationships between substitution sites.

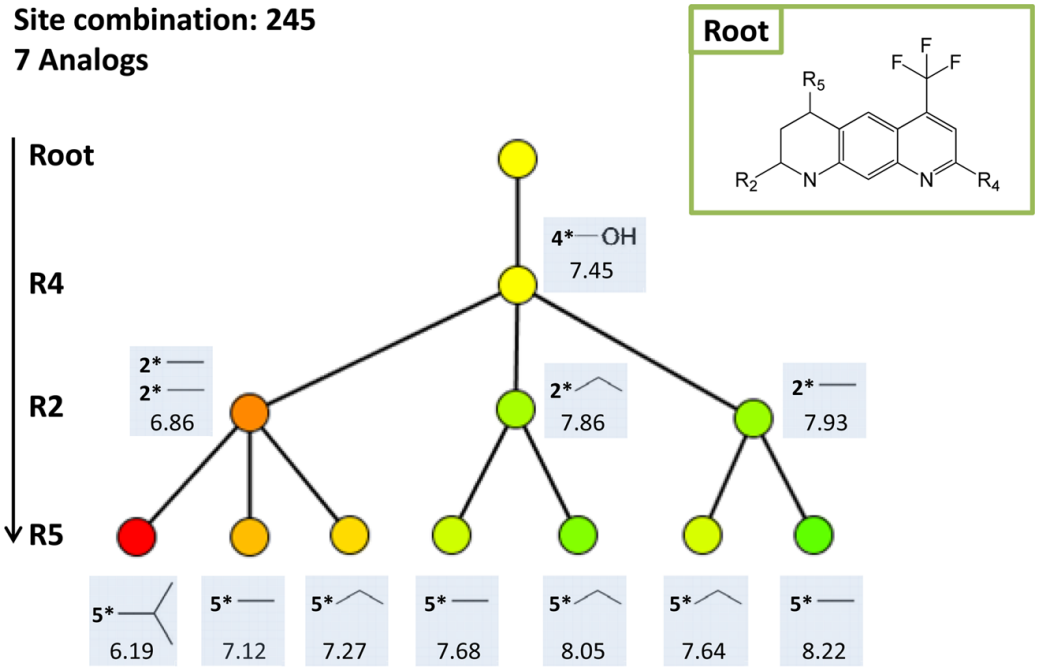
AnalogExplorer[107] uses a compound-based approach, rather than a compound pair or substituent based approach. At a global level it explores substitution sites and site combinations, deconvolutes a series into subsets of analogues having varying R-groups at the same site(s), and prioritizes analogues at specific site(s) that have desired activity. At a local level it represents a subset of analogues at given substitution site(s) on the basis of R-groups.

For graphical analysis, an analogue series is organized into subsets on the basis of the MCS. Each subset comprises compounds having varying R-groups at the same substitution site or site combination. Mapping of an analogue to the MCS of the series determines its subset membership. Each analogue of a series belongs to one and only one subset. An example is given below. In the complete graph, each node represents a substitution site or site combination, and all compounds with varying R-groups at the given site(s). The root node 0 corresponds to a compound having no R-group at any site. The node 1 represents analogues that only contain R-groups at R1, the node 12

analogues with R-groups at R1 and R2, and so on. Only a subset of these nodes is populated with analogue subsets. Nodes are scaled in size according to the number of analogues of the subset they represent and are colored according to the mean potency values of their analogues. The border thickness of nodes reflects the potency range covered by analogues comprising the corresponding subset.



Nodes are connected via edges according to subset relationships among the substitution sites, that is, if a substitution site defining a node is a subset of other site combinations. Therefore, edges in the graph reflect hierarchical relationships between nodes in adjacent layers. In an AnalogExplorer reduced graph, empty nodes, indicating unexplored sites or site combinations, and edges between them are omitted for ease of interpretation. Another graphical component, termed R-group tree, is designed to represent a subset of analogues with given substitution site(s). An R-group tree of node 245, can be constructed, for example:



Pauline presented graphs for four different applications. The first application was a single analogue series of 52 histamine H4 receptor antagonists, with 5 sites. The majority of site combinations were are associated with subsets of potent analogues and four site combinations were associated with activity cliffs. The second application was a single series of 38 analogues with two targets (tyrosine protein kinase ABL and tyrosine protein kinase SRC) leading to two graphs. Application three was multiple series for the target tyrosine protein kinase SRC. Five graphs were made for five qualifying series that were available in the target set: 22 analogues with four sites, 44 analogues with seven sites, 13 analogues with six sites, 22 analogues with six sites, and 22 analogues with four sites. The fourth application concerned four (out of five) structurally related series targeting tyrosine protein kinase SRC. A matched molecular pair calculation reduced these to a core from a combination of four of the scaffolds. Pauline showed complete and reduced graphs for 101 analogues with four sites. A Java implementation of AnalogExplorer routines is made freely available via the ZENODO open access platform.

Various ways to define molecular similarity

Eugen Lounkine of Novartis described work done with three different types of fingerprints. The concepts of molecular fingerprints and molecular similarity have matured and found innumerable applications, but nowadays Novartis does not just use chemical similarity to find compounds that biologically will behave the same; rather, the company directly builds on biological profiles to assess biosimilarity. From a medicinal chemistry point of view, one of the primary goals of HTS hit list assessment is the identification of chemotypes with an informative SAR. A common way to prioritize them is molecular clustering of the hits. Typical clustering techniques, however, rely on a general notion of chemical similarity or standard rules of scaffold decomposition and are thus insensitive to molecular features that are enriched in biologically active compounds. This hinders SAR analysis because compounds sharing the same pharmacophore might not end up in the same cluster and thus are not directly compared to each other by the medicinal chemist. Similarly, common chemotypes that are not related to activity may contaminate clusters, distracting from important chemical motifs.

Eugen and his colleagues have projected bioactivity onto chemical fingerprints; they have combined molecular similarity and Bayesian models, and introduced an activity-aware clustering approach, and a feature mapping method for the elucidation of distinct SAR determinants in polypharmacological compounds. They found that activity-aware clustering grouped compounds sharing molecular cores that were specific for the target or pathway at hand, rather than grouping inactive scaffolds commonly found in compound series. Weighted clusters often spread across many conventional clusters and there were large clusters that both methods agreed on.[108]

Eugen and his colleagues have also developed a tool that compares compounds solely on the basis of their bioactivity: the chemical biological descriptor called high-throughput screening fingerprint (HTS-FP). Data are aggregated from 234 Novartis biochemical and cell-based assays and can be used to identify bioactivity relationships among the in-house collection of about 1.8 million compounds. A similarity metric was derived combining both the numerical correlation of the activity *z*-scores, using the Pearson correlation coefficient, and the number of assays in common between the compounds.[109] HTS-FPs have been useful in both virtual screening and scaffold hopping. They are valuable not only because of their predictive power but mainly because they relate compounds solely on the basis of bioactivity. One challenge is the sparse nature of the HTS-FP matrix: the number of biologically annotated compounds still covers only a minuscule fraction of chemical space. To overcome this problem, Novartis has introduced Bioturbo similarity searching[110] that uses chemical similarity to map molecules without biological annotations into bioactivity space and then searches for biologically similar compounds in this reference system.

In addition, capturing the rich descriptions of compound-induced phenotypes from the literature gives yet another molecular fingerprint: the literature fingerprint. A naïve Bayesian model looks for themes around hits. Similarity search around a reference compound finds other compounds mentioned in the same biological or clinical context. By similarity searching a collection of terms, tool compounds for a phenotype of interest could be found. Novartis is carrying out exploratory annotation of phenotypic hits by text mining of abstracts and curated sources (e.g., ChEMBL provides a reference for each compound activity). MeSH terms for PubMed articles are filtered for informative terms. By data mining within and across projects, “signatures” derived from fingerprints can be found. A signature is a fingerprint template endowed with meaning. The meaning is encoded by relating the signatures database to the phenotype database.

Novartis has developed network algorithms to build and navigate heterogeneous similarity networks from the three types of fingerprint. Eugen gave an example starting with a selection of painkillers, connected only by literature relationships. In the first expansion of the network all the seed compounds have neighbors, but they come from different similarity measures: more painkillers (quercetin and doxorubicin) are added from the literature, diclofenac analogues from chemical similarity, and ibuprofen similars from HTS-FP. Next, the neighbors themselves are connected among each other, sometimes with more than one method. Distinct clusters emerge as interesting neighbors of neighbors (connectors) are added: morphine analogues, NSAIDs, and oncology pain management. Pairs that are connected by more than one method can be identified. These voting schemes are intuitive in graphs, and harder to formalize in conventional approaches. Alternatively, one can use degree, number of distinct edge types, etc. A flow algorithm is use to distribute scores. This standard graph neighborhood scoring algorithm is intuitive to carry out and visualize, and easily scalable. Eugen also showed networks of glitazones, antidepressants, and statins and warfarin.

Graph representations provide a unique opportunity to combine distinct similarity domains in an interoperable way.

Could "inactive" compounds be good starting points for drug discovery?

Anne Mai Wassermann, now at Pfizer, talked about work done at Novartis while she was one of Jürgen's postdoctoral students. For a long time the paradigm for screening library design has been diversity. Chemical diversity has been used as a surrogate for biodiversity, but biological fingerprints themselves could be used.[111,112] What should then be done about the inactives? A compound inactive in a great many screens might be a good, selective lead from another screen. An analysis across more than 200 Novartis biochemical and cellular HTS assays showed that 112,872 compounds (14%) were consistently inactive in 100 assays. A permutation experiment showed that this was not a random chance effect. NIH Molecular Libraries campaigns also have many genuine inactives. The term "dark chemical matter" (DCM) has been coined for such compounds.

An analysis of 1,273 "dark" and 1,257 active compounds proved that intrinsic compound solution quality is not a factor in the inactivity, but it did reveal bad news about the quality of screening collections. Analysis of the properties of a Novartis set of compounds showed that DCM compounds are more soluble and less hydrophobic than actives, and they are smaller and have fewer rings. When the structural differences, if any, between DCM and actives were studied by multidimensional scaling it turned out that dark compounds and actives are not too different; dark compounds are not outliers in either Novartis or PubChem collections. Active compounds with dark substructures have lower hit rates; the nearest neighbors of actives near DCM tend to be more selective.

Anne Mai displayed some dark substructures; chemists thought that they looked fairly "innocent". She also showed some dark natural products that looked as if they should be active. Could dark compounds be valuable hits and potential tool compounds? Perhaps it could be that they seem inactive at typical screening concentrations. Novartis carried out an analysis of 34 additional high throughput screens in each of which at least 60,000 dark compounds were tested; previously active compounds yielded many more hits in these 34 screens than dark compounds, but, while 87% of the dark hits were hits in only one of the 34 screens, only 57% of previously active compounds showed this (i.e., 43 % of the hits from this compound class hit in more than one of the 34 screens). The difference between active and dark compounds was even greater when natural product compounds were tested at 10 micromolar rather than 1 micromolar in 37 cancer cell lines. This fits with the Novartis hypothesis about concentration.

Experiments were next carried out covering broad biology. Of 1,408 compounds (704 dark and 704 active) submitted to 40 reporter gene assays, 92 actives but only 24 dark compounds were hits at 4 micromolar concentration. When a dark compound is active it is more selective. In a gene expression panel, 61 genes were measured for 188 dark compounds and 164 actives, at 1 micromolar and 10 micromolar concentrations. The results again supported the concentration hypothesis. The mechanism of action of a dark compound was elucidated in yeast HIP profiling; 200 dark compounds were tested against 6,000 heterozygous yeast strains, each with a different gene copy deleted. Some initial SAR studies with an antifungal panel have suggested a compound and analogues that were *in vitro* highly potent against *C. neoformans*, which causes fungal meningitis and encephalitis. The compound was clean against a human safety panel.

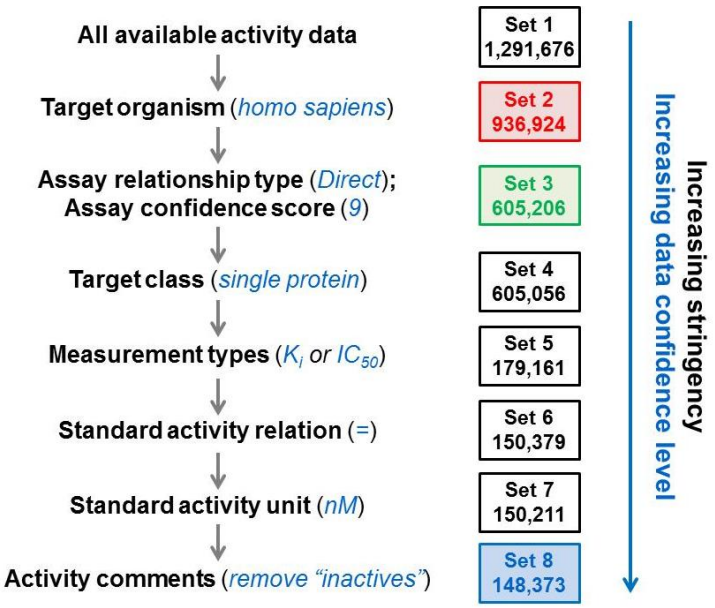
These experiments demonstrate that, when tested for the right phenotype or target, DCM can elicit strong biological responses. Consequently, Novartis believes that DCM is not generally biologically inert, and concludes that that their reduced promiscuity makes compounds from DCM a valuable resource for selective biological probes, and starting points for drug discovery programs.

Complexity and heterogeneity of data for chemical information science

Finally, Jürgen gave his award address. Similar to the situation in biology a few years ago, we currently witness the advent of the big data era in medicinal chemistry. UniChem, for example, now links 91 million compounds; there are 61 million compounds in PubChem. Increasing amounts of bioactivity data are available. ChEMBL has 13.5 million activity annotations for 1.5 million compounds and 10,774 biological targets. BindingDB has 1.1 million binding records for 495,128 compounds and 7,030 protein targets. PubChem has 60.7 million compounds, 1.15 million assays and 206,541 confirmatory assays. DrugBank records 7,759 drugs, 1,602 approved drugs, and 4,300 protein targets. What an opportunity all these data offer!

Rapidly growing compound numbers and volumes of activity data require elaborate infrastructures for deposition, curation, and organization, but the need for such infrastructures only partly reflects the challenges associated with big data phenomena. The "5Vs" are cited as criteria[113] for big data: volume, velocity, variety, veracity, and value. Jürgen believes that the increasing complexity and heterogeneity of compound data are additional challenges for computational analysis and knowledge extraction, and are probably even greater challenges than mere data volumes. To illustrate his point, Jürgen tabulated some data for trimeprazine and promethazine (closely related anti-allergic agents) in DrugBank, ChEMBL, BindingDB, and PubChem. Data incompleteness also applies in this example.

Another criterion that could be added is data confidence. Jürgen took compound datasets from ChEMBL 18 to illustrate varying confidence levels:



Jürgen presented a ligand-centric view of promiscuity and the impact of data confidence. Evidence is mounting that polypharmacological drug behavior is often responsible for therapeutic efficacy, suggesting the consideration of new drug development strategies. Target promiscuity of compounds

is at the origin of polypharmacology. For many bioactive compounds, multiple target annotations are available, indicating that compound promiscuity is a general phenomenon, but careful analysis of compound activity data reveals that the degree of apparent promiscuity is strongly influenced by data selection criteria and the type of activity measurements that are considered.[19] The average promiscuity rate of Jürgen's set 1 from ChEMBL was 6.7. The rate fell as confidence level increased;[114] the promiscuity rate of set 8 was only 1.5.

Jürgen's team has also studied compound promiscuity over time. Using sets 2, 3 and 8 from ChEMBL 20, they found that there has only been a minor increase in promiscuity over a great many years.[115] For the years 2004-1014, the promiscuity rate for set 2 has risen from about 1.8 to 2.5; for set 8 it has remained steady at about 1.5. It is interesting that approved drugs are more promiscuous. The promiscuity rate for a set 2 equivalent of approved drugs has risen from 5.9 in 2000 to 24.4 in 2014; for a set 8 it has risen from 1.9 to 3.7. The promiscuity rate of imatinib is particularly interesting: on the basis of low-confidence data, it has risen from 7 in 2004 to 690 in 2014! The high-confidence set 8 figure for 2014 is 27.

Global average promiscuity across five target families, GPCR class A, ion channels, kinases, nuclear receptors, and proteases is only 1.5 for sets of type 8 in ChEMBL 20. For example, one might have expected kinase inhibitors to be more promiscuous but they do not appear to be any more so than average if high confidence data levels are considered. Global average promiscuity does not vary a great deal around 1.5 as molecular weight and lipophilicity are varied, except in the case of compounds with molecular weight less than or equal to 200, where promiscuity is about 2.2.[115]

Ye Hu and Jürgen have also taken a target-centric view of promiscuity, derived from compound activity data.[116] The ability of target proteins to bind structurally diverse compounds and compounds with different degrees of promiscuity was systematically assessed on the basis of activity data and target annotations. Intuitive first- and second-order target promiscuity indices (TPIs) were introduced to quantify these binding characteristics and relate them to each other. TPI_1, the first-order target promiscuity index is calculated as the number of unique scaffolds of all compounds active against a given target; it indicates the ability of a target to interact with structurally diverse compounds. TPI_2, the second-order target promiscuity index, is the average degree of promiscuity of all compounds active against the target; it reflects the tendency of a target to interact with specific and promiscuous compounds.

The average TPI_1 value over all targets is 77 (for K_i data) and 61 (for IC_{50} data). This is not surprising: it is well known that many targets bind structurally diverse compounds. Only about 18% of all targets interact with compounds having no other reported activity ("pseudo-specific" compounds); here the TPI_2 value is 1. Most targets bind varying numbers of promiscuous compounds.

Targets that interact with compounds that are structurally diverse (more than 120 distinct scaffolds), but with no other reported activities, have high TPI_1 and low TPI_2. Examples are leukotriene A4 hydrolase and C-X-C chemokine receptor type 3. Targets that interact with compounds that are structurally homogeneous and preferentially promiscuous have low TPI_1 and high TPI_2. Examples are group IID secretory phospholipase A2 and matrix metalloproteinase 16. TPI_2 values establish the promiscuity profiles of target families; Jürgen showed some pie-charts of TPI_2 values for various target families.[116]

We are entering the big data era in chemical information science: compounds and activity data volumes, heterogeneity, and complexity are increasing. Data heterogeneity and inconsistency across databases is observed. Compound data mining offers significant opportunities for pharmaceutical R&D, but ensuring high data confidence and integrity is important. Promiscuity is the molecular basis of polypharmacology. Degrees of promiscuity vary with data confidence. Compound- and target-centric views of promiscuity can be taken.

Conclusion

After Jürgen's award address, Rachelle Bienstock, chair of the ACS Division of Chemical Information, formally presented the Herman Skolnik Award to Jürgen Bajorath:

References

(1) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997,** *119* (43), 10509-10524.

(2) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988,** *110* (18), 5959-5967.

(3) Andrade, C. H.; Pasqualoto, K. F. M.; Ferreira, E. I.; Hopfinger, A. J. 4D-QSAR: Perspectives in Drug Design. *Molecules* **2010,** *15* (5), 3281.

(4) Albuquerque, M. G.; Hopfinger, A. J.; Barreiro, E. J.; de Alencastro, R. B. Four-dimensional quantitative structure-activity relationship analysis of a series of interphenylene 7-oxabicycloheptane oxazole thromboxane A2 receptor antagonists. *J. Chem. Inf. Comput. Sci.* **1998,** *38* (5), 925-938.

(5) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994,** *34* (4), 854-866.

(6) Santos-Filho, O. A.; Hopfinger, A. J.; Cherkasov, A.; Bicca de Alencastro, R. The receptor-dependent QSAR paradigm: an overview of the current state of the art. *Med. Chem.* **2009,** *5* (4), 359-366.

(7) Pan, D.; Liu, J.; Senese, C.; Hopfinger, A. J.; Tseng, Y. Characterization of a Ligand-Receptor Binding Event Using Receptor-Dependent Four-Dimensional Quantitative Structure-Activity Relationship Analysis. *J. Med. Chem.* **2004,** *47* (12), 3075-3088.
(8) Pan, D.; Tseng, Y.; Hopfinger, A. J. Quantitative Structure-Based Design: Formalism and Application of Receptor-Dependent RD-4D-QSAR Analysis to a Set of Glucose Analogue Inhibitors of Glycogen Phosphorylase. *J. Chem. Inf. Comput. Sci.* **2003,** *43* (5), 1591-1607.
(9) Cramer, R. D. Template CoMFA Generates Single 3D-QSAR Models that, for Twelve of Twelve Biological Targets, Predict All ChEMBL-Tabulated Affinities. *PLoS One* **2015,** *10* (6), e0129307.
(10) Otter, T. Toward a New Theoretical Framework for Biology. In *Genetic and Evolutionary Computation Conference (GECCO) Workshop on Self-organization in Evolutionary Algorithms* Seattle, WA, June 26-30, 2004 http://www.cdres.com/content/GeccoTowardANewTheoreticalFrameworkForBiology_Otter_2004.pdf.
(11) Maggiora, G. M. The reductionist paradox: are the laws of chemistry and physics sufficient for the discovery of new drugs? *J. Comput.-Aided Mol. Des.* **2011,** *25* (8), 699-708.
(12) Yildirim, M. A.; Goh, K.-I.; Cusick, M. E.; Barabasi, A.-L.; Vidal, M. Drug-target network. *Nat. Biotechnol.* **2007,** *25* (10), 1119-1126.
(13) Paolini, G. V.; Shapland, R. H. B.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006,** *24* (7), 805-815.
(14) Vogt, I.; Mestres, J. Drug-target networks. *Mol. Inf.* **2010,** *29* (1-2), 10-14.
(15) Mestres, J.; Gregori-Puigjane, E.; Valverde, S.; Sole, R. V. Data completeness-the Achilles heel of drug-target networks. *Nat. Biotechnol.* **2008,** *26* (9), 983-984.
(16) Jalencas, X.; Mestres, J. On the origins of drug polypharmacology. *MedChemComm* **2013,** *4* (1), 80-87.
(17) Hu, Y.; Bajorath, J. Promiscuity profiles of bioactive compounds: potency range and difference distributions and the relation to target numbers and families. *MedChemComm* **2013,** *4* (8), 1196-1201.
(18) Hu, Y.; Bajorath, J. How Promiscuous Are Pharmaceutically Relevant Compounds? A Data-Driven Assessment. *AAPS J.* **2013,** *15* (1), 104-111.
(19) Hu, Y.; Bajorath, J. Compound promiscuity: what can we learn from current data? *Drug Discovery Today* **2013,** *18* (13-14), 644-650.
(20) Hu, Y.; Bajorath, J. Activity profile relationships between structurally similar promiscuous compounds. *Eur. J. Med. Chem.* **2013,** *69*, 393-398.
(21) Kuhn, M.; Szklarczyk, D.; Pletscher-Frankild, S.; Blicher, T. H.; von Mering, C.; Jensen, L. J.; Bork, P. STITCH 4: integration of protein–chemical interactions with user data. *Nucleic Acids Res.* **2014,** *42(Database issue):D401-407*.
(22) Roider, H. G.; Pavlova, N.; Kirov, I.; Slavov, S.; Slavov, T.; Uzunov, Z.; Weiss, B. Drug2Gene: an exhaustive resource to explore effectively the drug-target relation network. *BMC Bioinf.* **2014,** *15*, 68.
(23) von Eichborn, J.; Murgueitio, M. S.; Dunkel, M.; Koerner, S.; Bourne, P. E.; Preissner, R. PROMISCUOUS: a database for network-based drug-repositioning. *Nucleic Acids Res.* **2011,** *39* (Suppl. 1), D1060-D1066.
(24) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008,** *36* (Database Iss), D901-D906.
(25) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012,** *40* (D1), D1100-D1107.
(26) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.* **2007,** *35* (suppl 1), D198-D201.

(27) Wang, Y.; Suzek, T.; Zhang, J.; Wang, J.; He, S.; Cheng, T.; Shoemaker, B. A.; Gindulyte, A.; Bryant, S. H. PubChem BioAssay: 2014 update. *Nucleic Acids Res.* **2014,** *42* (D1), D1075-D1082.
(28) Salesse, S.; Verfaillie, C. M. BCR/ABL-mediated Increased Expression of Multiple Known and Novel Genes That May Contribute to the Pathogenesis of Chronic Myelogenous Leukemia. *Mol. Cancer Ther.* **2003,** *2* (2), 173-182.
(29) Kim, T. M.; Ha, S. A.; Kim, H. K.; Yoo, J.; Kim, S.; Yim, S. H.; Jung, S. H.; Kim, D. W.; Chung, Y. J.; Kim, J. W. Gene expression signatures associated with the in vitro resistance to two tyrosine kinase inhibitors, nilotinib and imatinib. *Blood Cancer J.* **2011,** *1*, e32.
(30) Moffat, J. G.; Rudolph, J.; Bailey, D. Phenotypic screening in cancer drug discovery - past, present and future. *Nat. Rev. Drug Discovery* **2014,** *13* (8), 588-602.
(31) Rouvray, D. H. The evolution of the concept of molecular similarity. In *Concepts and Applications of Molecular Similarity;* Johnson, M. A., Maggiora, G. M., Eds.; John Wiley & Sons: New York, 1990; pp 15-42.
(32) Rouvray, D. H. Definition and role of similarity concepts in the chemical and physical sciences. *J. Chem. Inf. Comput. Sci.* **1992,** *32* (6), 580-586.
(33) *Concepts and Applications of Molecular Similarity*. Johnson, M. A.; Maggiora, G. M., Eds.; John Wiley & Sons: New York, 1990.
(34) Wilkins, C. L.; Randic, M. A graph theoretical approach to structure-property and structure-activity correlations. *Theor. Chim. Acta* **1980,** *58* (1), 45-68.
(35) Crum Brown, A. On the connection between chemical constitution and physiological action. *J. Anat. Physiol.* **1868,** *2* (2), 224-242.
(36) Tobler, W. R. A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* **1970,** *46*, 234-240.
(37) McPherson, M.; Smith-Lovin, L.; Cook, J. Birds of a feather: homophily in social networks. *Ann. Rev. Sociol.* **2001,** *27*, 415-444.
(38) van Rijsbergen, C. J. *Information Retrieval*; Butterworth: London, 1979.
(39) Harrison, P. J. A method of cluster analysis and some applications. *Appl. Stat.* **1968,** *17*, 226-236.
(40) Adamson, G. W.; Bush, J. A. Method for the automatic classification of chemical structures. *Inf. Storage Retr.* **1973,** *9* (10), 561-568.
(41) Adamson, G. W.; Bush, J. A. Comparison of the performance of some similarity and dissimilarity measures in the automatic classification of chemical structures. *J. Chem. Inf. Comput. Sci.* **1975,** *15* (1), 55-58.
(42) Willett, P.; Winterman, V. A comparison of some measures for the determination of inter-molecular structural similarity: measures of inter-molecular structural similarity. *Quant. Struct.-Act. Relat.* **1986,** *5* (1), 18-25.
(43) Willett, P.; Winterman, V.; Bawden, D. Implementation of nonhierarchic cluster analysis methods in chemical information systems: selection of compounds for biological testing and clustering of substructure search output. *J. Chem. Inf. Comput. Sci.* **1986,** *26* (3), 109-118.
(44) Willett, P.; Winterman, V.; Bawden, D. Implementation of nearest-neighbor searching in an online chemical structure search system. *J. Chem. Inf. Comput. Sci.* **1986,** *26* (1), 36-41.
(45) Ward, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963,** *58*, 236-244.
(46) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared nearest neighbors. *IEEE Transactions on Computers* **1973,** *C-22*, 1025-1034.
(47) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985,** *25* (2), 64-73.
(48) Bawden, D. Browsing and clustering of chemical structures. In *Chemical Structures;* Warr, W. A., Ed.; Springer Verlag: Berlin, 1986; pp 145-150.

(49) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996,** *36* (1), 118-127.
(50) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Comput. Sci.* **1996,** *36* (1), 128-136.
(51) Sheridan, R. P. Chemical similarity searches: when is complexity justified? *Expert Opin. Drug Discovery* **2007,** *2* (4), 423-430.
(52) Ginn, C. M. R.; Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W. Similarity Searching in Files of Three-Dimensional Chemical Structures: Evaluation of the EVA Descriptor and Combination of Rankings Using Data Fusion. *J. Chem. Inf. Comput. Sci.* **1997,** *37* (1), 23-37.
(53) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of molecular similarity measures using data fusion. *Perspect. Drug Discovery Des.* **2000,** *20*, 1-16.
(54) Willett, P. Enhancing the effectiveness of ligand-based virtual screening using data fusion. *QSAR Comb. Sci.* **2006,** *25* (12), 1143-1152.
(55) Willett, P. Combination of Similarity Rankings Using Data Fusion. *J. Chem. Inf. Model.* **2013,** *53* (1), 1-10.
(56) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring diversity: experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.* **1995,** *38* (9), 1431-1436.
(57) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **1996,** *39* (16), 3049-3059.
(58) Lajiness, M. S.; Johnson, M. A.; Maggiora, G. M. Implementing drug screening programs using molecular similarity. In *QSAR: Quantitaive Structure-activity Relationships in Drug Design;* Fauchere, J. L., Ed.; Alan R. Liss: New York, 1989; pp 173-176.
(59) Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graphics Modell.* **1998,** *15* (6), 372-385.
(60) Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C. F.; Mitchell, E. M.; Mitchell, G. F.; Smith, J. M.; Watson, D. G. The development of versions 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* **1991,** *31* (2), 187-204.
(61) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *J. Med. Chem.* **1994,** *37* (24), 4130-4146.
(62) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998,** *38* (6), 983-996.
(63) Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003,** *22* (1), 69-77.
(64) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996,** *39* (15), 2887-2893.
(65) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discovery* **2004,** *3* (11), 935-949.
(66) Aruffo, A.; Farrington, M.; Hollenbaugh, D.; Li, X.; Milatovich, A.; Nonoyama, S.; Bajorath, J.; Grosmaire, L. S.; Stenkamp, R.; et, a. The CD40 ligand, gp39, is defective in activated T cells from patients with X-linked hyper-IgM syndrome. *Cell (Cambridge, Mass.)* **1993,** *72* (2), 291-300.
(67) Linsley, P. S.; Greene, J. L.; Brady, W.; Bajorath, J.; Ledbetter, J.; Peach, R. Human B7-1 (CD80) and B7-2 (CD86) bind with similar avidities but distinct kinetics of CD28 and CTLA-4 receptors. *Immunity* **1994,** *1* (9), 793-801.
(68) Foy, T. M.; Aruffo, A.; Bajorath, J.; Buhlmann, J. E.; Noelle, R. J. Immune regulation by CD40 and its ligand gp39. *Annu. Rev. Immunol.* **1996,** *14*, 591-617.

(69) Sica, G. L.; Choi, I.-H.; Zhu, G.; Tamada, K.; Wang, S.-D.; Tamura, H.; Chapoval, A. I.; Flies, D. B.; Bajorath, J.; Chen, L. B7-H4, a molecule of the B7 family, negatively regulates T cell immunity. *Immunity* **2003,** *18* (6), 849-861.
(70) Bishop, C. M.; Svensén, M.; Williams, C. K. I. GTM: The Generative Topographic Mapping. *Neural Comput.* **1998,** *10* (1), 215-234.
(71) Chupakhin, V.; Marcou, G.; Gaspar, H.; Varnek, A. Simple Ligand-Receptor Interaction Descriptor (SILIRID) for alignment-free binding site comparison. *Comput. Struct. Biotechnol. J.* **2014,** *10* (16), 33-7.
(72) Kireeva, N.; Baskin, I. I.; Gaspar, H. A.; Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol. Inf.* **2012,** *31* (3-4), 301-312.
(73) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J. Comput.-Aided Mol. Des.* **2005,** *19* (9/10), 693-703.
(74) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA - platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr. Comput.-Aided Drug Des.* **2008,** *4* (3), 191-198.
(75) Horvath, D.; Bonachera, F.; Solov'ev, V.; Gaudin, C.; Varnek, A. Stochastic versus Stepwise Strategies for Quantitative Structure−Activity Relationship GenerationHow Much Effort May the Mining for Successful QSAR Models Take? *J. Chem. Inf. Model.* **2007,** *47* (3), 927-939.
(76) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. GTM-Based QSAR Models and Their Applicability Domains. *Mol. Inf.* **2015,** *34* (6-7), 348-356.
(77) Gaspar, H. A.; Marcou, G.; Horvath, D.; Arault, A.; Lozano, S.; Vayer, P.; Varnek, A. Generative Topographic Mapping-Based Classification Models and Their Applicability Domain: Application to the Biopharmaceutics Drug Disposition Classification System (BDDCS). *J. Chem. Inf. Model.* **2013,** *53* (12), 3318-3325.
(78) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. Chemical Data Visualization and Analysis with Incremental Generative Topographic Mapping: Big Data Challenge. *J. Chem. Inf. Model.* **2015,** *55* (1), 84-94.
(79) Funatsu, K. Computer-aided synthesis design and reaction prediction. *Kagaku Kogyo* **2007,** *58* (2), 124-129.
(80) Masuda, Y.; Kaneko, H.; Funatsu, K. Multivariate Statistical Process Control Method Including Soft Sensors for Both Early and Accurate Fault Detection. *Ind. Eng. Chem. Res.* **2014,** *53* (20), 8553-8564.
(81) Kaneko, H.; Funatsu, K. Database monitoring index for adaptive soft sensors and the application to industrial process. *AIChE J.* **2014,** *60* (1), 160-169.
(82) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery* **2005,** *4* (8), 649-663.
(83) Alig, L.; Alsenz, J.; Andjelkovic, M.; Bendels, S.; Benardeau, A.; Bleicher, K.; Bourson, A.; David-Pierson, P.; Guba, W.; Hildbrand, S.; Kube, D.; Luebbers, T.; Mayweg, A. V.; Narquizian, R.; Neidhart, W.; Nettekoven, M.; Plancher, J.-M.; Rocha, C.; Rogers-Evans, M.; Roever, S.; Schneider, G.; Taylor, S.; Waldmeier, P. Benzodioxoles: Novel Cannabinoid-1 Receptor Inverse Agonists for the Treatment of Obesity. *J. Med. Chem.* **2008,** *51* (7), 2115-2127.
(84) Hartenfeller, M.; Schneider, G. De novo drug design. *Methods Mol. Biol. (N. Y., NY, U. S.)* **2011,** *672*, 299-323.
(85) Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G. DOGS: reaction-driven de novo design of bioactive compounds. *PLoS Comput. Biol.* **2012,** *8* (2), e1002380.
(86) Spaenkuch, B.; Keppner, S.; Lange, L.; Rodrigues, T.; Zettl, H.; Koch, C. P.; Reutlinger, M.; Hartenfeller, M.; Schneider, P.; Schneider, G. Drugs by Numbers: Reaction-Driven De Novo Design of Potent and Selective Anticancer Leads. *Angew. Chem., Int. Ed.* **2013,** *52* (17), 4676-4681.

(87) Rupp, M.; Proschak, E.; Schneider, G. Kernel approach to molecular similarity based on iterative graph similarity. *J. Chem. Inf. Model.* **2007,** *47* (6), 2280-2286.
(88) Rupp, M.; Schneider, G. Graph Kernels for Molecular Similarity. *Mol. Inf.* **2010,** *29* (4), 266-273.
(89) Reutlinger, M.; Rodrigues, T.; Schneider, P.; Schneider, G. Multi-Objective Molecular De Novo Design by Adaptive Fragment Prioritization. *Angew. Chem., Int. Ed.* **2014,** *53* (16), 4244-4248.
(90) Reutlinger, M.; Rodrigues, T.; Schneider, P.; Schneider, G. Combining On-Chip Synthesis of a Focused Combinatorial Library with Computational Target Prediction Reveals Imidazopyridine GPCR Ligands. *Angew. Chem., Int. Ed.* **2014,** *53* (2), 582-585.
(91) Rodrigues, T.; Schneider, P.; Schneider, G. Accessing New Chemical Entities through Microfluidic Systems. *Angew. Chem., Int. Ed.* **2014,** *53* (23), 5750-5758.
(92) Rodrigues, T.; Hauser, N.; Reker, D.; Reutlinger, M.; Wunderlin, T.; Hamon, J.; Koch, G.; Schneider, G. Multidimensional De Novo Design Reveals 5-HT2B Receptor-Selective Ligands. *Angew. Chem., Int. Ed.* **2015,** *54* (5), 1551-1555.
(93) Reker, D.; Rodrigues, T.; Schneider, P.; Schneider, G. Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proc. Natl. Acad. Sci. U. S. A.* **2014,** *111* (11), 4067-4072.
(94) Reutlinger, M.; Koch, C. P.; Reker, D.; Todoroff, N.; Schneider, P.; Rodrigues, T.; Schneider, G. Chemically Advanced Template Search (CATS) for Scaffold-Hopping and Prospective Target Prediction for Orphan' Molecules. *Mol. Inf.* **2013,** *32* (2), 133-138.
(95) Grabowski, K.; Baringhaus, K.-H.; Schneider, G. Scaffold diversity of natural products. Inspiration for combinatorial library design. *Nat. Prod. Rep.* **2008,** *25* (5), 892-904.
(96) Reker, D.; Perna, A. M.; Rodrigues, T.; Schneider, P.; Reutlinger, M.; Monch, B.; Koeberle, A.; Lamers, C.; Gabler, M.; Steinmetz, H.; Muller, R.; Schubert-Zsilavecz, M.; Werz, O.; Schneider, G. Revealing the macromolecular targets of complex natural products. *Nat. Chem.* **2014,** *6* (12), 1072-1078.
(97) Rodrigues, T.; Reker, D.; Kunze, J.; Schneider, P.; Schneider, G. Revealing the Macromolecular Targets of Fragment-Like Natural Products. *Angew. Chem., Int. Ed.* **2015,** *54* (36), 10516-10520.
(98) Tyrchan, C.; Bostroem, J.; Giordanetto, F.; Winter, J.; Muresan, S. Exploiting Structural Information in Patent Specifications for Key Compound Prediction. *J. Chem. Inf. Model.* **2012,** *52* (6), 1480-1489.
(99) Wawer, M. J.; Jaramillo, D. E.; Dancik, V.; Fass, D. M.; Haggarty, S. J.; Shamji, A. F.; Wagner, B. K.; Schreiber, S. L.; Clemons, P. A. Automated structure-activity relationship mining: connecting chemical structure to biological profiles. *J. Biomol. Screening* **2014,** *19* (5), 738-748.
(100) Peltason, L.; Hu, Y.; Bajorath, J. From Structure-Activity to Structure-Selectivity Relationships: Quantitative Assessment, Selectivity Cliffs, and Key Compounds. *ChemMedChem* **2009,** *4* (11), 1864-1873.
(101) Iyer, P.; Hu, Y.; Bajorath, J. SAR Monitoring of Evolving Compound Data Sets Using Activity Landscapes. *J. Chem. Inf. Model.* **2011,** *51* (3), 532-540.
(102) Iyer, P.; Stumpfe, D.; Bajorath, J. Molecular Mechanism-Based Network-like Similarity Graphs Reveal Relationships between Different Types of Receptor Ligands and Structural Changes that Determine Agonistic, Inverse-Agonistic, and Antagonistic Effects. *J. Chem. Inf. Model.* **2011,** *51* (6), 1281-1286.
(103) Wassermann, A. M.; Haebel, P.; Weskamp, N.; Bajorath, J. SAR Matrices: Automated Extraction of Information-Rich SAR Tables from Large Compound Data Sets. *J. Chem. Inf. Model.* **2012,** *52* (7), 1769-1776.
(104) Gupta-Ostermann, D.; Shanmugasundaram, V.; Bajorath, J. Neighborhood-Based Prediction of Novel Active Compounds from SAR Matrices. *J. Chem. Inf. Model.* **2014,** *54* (3), 801-809.
(105) Agrafiotis, D. K.; Shemanarev, M.; Connolly, P. J.; Farnum, M.; Lobanov, V. S. SAR Maps: A New SAR Visualization Technique for Medicinal Chemists. *J. Med. Chem.* **2007,** *50* (24), 5926-5937.

(106) Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Exploration of Structure-Activity Relationship Determinants in Analogue Series. *J. Med. Chem.* **2009,** *52* (10), 3212-3224.
(107) Zhang, B.; Hu, Y.; Bajorath, J. AnalogExplorer: A New Method for Graphical Analysis of Analog Series and Associated Structure-Activity Relationship Information. *J. Med. Chem.* **2014,** *57* (21), 9184-9194.
(108) Lounkine, E.; Nigsch, F.; Jenkins, J. L.; Glick, M. Activity-Aware Clustering of High Throughput Screening Data and Elucidation of Orthogonal Structure-Activity Relationships. *J. Chem. Inf. Model.* **2011,** *51* (12), 3158-3168.
(109) Petrone, P. M.; Simms, B.; Nigsch, F.; Lounkine, E.; Kutchukian, P.; Cornett, A.; Deng, Z.; Davies, J. W.; Jenkins, J. L.; Glick, M. Rethinking Molecular Similarity: Comparing Compounds on the Basis of Biological Activity. *ACS Chem. Biol.* **2012,** *7* (8), 1399-1409.
(110) Wassermann, A. M.; Lounkine, E.; Glick, M. Bioturbo Similarity Searching: Combining Chemical and Biological Similarity To Discover Structurally Diverse Bioactive Molecules. *J. Chem. Inf. Model.* **2013,** *53* (3), 692-703.
(111) Petrone, P. M.; Wassermann, A. M.; Lounkine, E.; Kutchukian, P.; Simms, B.; Jenkins, J.; Selzer, P.; Glick, M. Biodiversity of small molecules - a new perspective in screening set selection. *Drug Discovery Today* **2013,** *18* (13-14), 674-680.
(112) Wawer, M. J.; Li, K.; Gustafsdottir, S. M.; Ljosa, V.; Bodycombe, N. E.; Marton, M. A.; Sokolnicki, K. L.; Bray, M.-A.; Kemp, M. M.; Winchester, E.; Taylor, B.; Grant, G. B.; Hon, C. S.-Y.; Duvall, J. R.; Wilson, J. A.; Bittker, J. A.; Dancik, V.; Narayan, R.; Subramanian, A.; Winckler, W.; Golub, T. R.; Carpenter, A. E.; Shamji, A. F.; Schreiber, S. L.; Clemons, P. A. Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. *Proc. Natl. Acad. Sci. U. S. A.* **2014,** *111* (30), 10911-10916.
(113) Lusher, S. J.; McGuire, R.; van Schaik, R. C.; Nicholson, C. D.; de Vlieg, J. Data-driven medicinal chemistry in the era of big data. *Drug Discovery Today* **2014,** *19* (7), 859-868.
(114) Hu, Y.; Bajorath, J. Influence of Search Parameters and Criteria on Compound Selection, Promiscuity, and Pan Assay Interference Characteristics. *J. Chem. Inf. Model.* **2014,** *54* (11), 3056-3066.
(115) Hu, Y.; Jasial, S.; Bajorath, J. Promiscuity progression of bioactive compounds over time. *F1000Res* **2015,** *4* (Chem. Inf. Sci.), 118.
(116) Hu, Y.; Bajorath, J. Quantifying the tendency of therapeutic target proteins to bind promiscuous or selective compounds. *PLoS One* **2015,** *10* (5), e0126838.