Herman Skolnik Award Symposium 2013

Honoring Richard D. (Dick) Cramer

Introduction

Dick Cramer is best known as the inventor of the technique of Comparative Molecular Field Analysis (CoMFA) and its introduction to the molecular and drug design fields. Early in his career, in the research group of E.J. Corey, Dick was involved with the first artificial intelligence methods to predict chemical synthesis, coining the acronym “LHASA” (Logic and Heuristics Applied to Synthetic Analysis) for the project. Dick has remained active in research and publishing at the forefront of his field. His work on “topomeric” descriptors, which allows CoMFA without tedious alignment of ligands, is proving a very successful tool in drug discovery. He currently serves as Senior Vice President, Science, and Chief Scientific Officer for Tripos, a Certara Company. Dick has also made major contributions to another entirely different field: baseball. He became interested in applying computers to baseball statistics and developed a program to feed detailed baseball statistics into the commentators’ box. He consulted with a number of major league teams, and is featured in the book *Moneyball* by Michael Lewis (recently made into a major motion picture). The award symposium covered all Dick’s fields of endeavor.

CoMFA

Since Dick is best known as the inventor CoMFA[1], it was fitting that the opening talk, by Bob Clark of Simulations Plus (bob@simulations-plus.com), outlined the history of CoMFA, citing eight articles in which he himself was a co-author.[1-8] CoMFA required the identification of the “bioactive conformer” and this was difficult in the era of combinatorial chemistry, so Dick and other colleagues came up with topomers and rules for conformer alignment, while Bob concentrated on traditional CoMFA.[9-17]

“The alignment problem” has many dimensions. One is aligning ligands to themselves (conformation), i.e., studying the relationships between substructures within an individual molecule. Another is aligning ligands to each other (“alignment”), i.e., studying the relationships between substructures in *different* ligand molecules. Finding an appropriate protein conformation and aligning the protein to the ligands are further dimensions. Bob favors the ligands’-eye view of protein binding[15] over the protein’s-eye view; given a basic pose obtained by docking or pharmacophore alignment, he likes to refine the alignment based on common substructures in the ligands and see how the protein adjusts to accommodate ligand variation.

At the spring 1998 ACS meeting, Bob spoke about making 3D QSAR both simple and robust. The literature background included seminal CoMFA publications,[1,18] papers on region selection methods,[19,20] and articles on descriptor transforms.[21,22] At that time there were concerns about “out of the box” CoMFA: the sensitivity of q^2 to changes in conformation and lattice alignment, and reproducibility from published applications. Approaches to dealing with the variability included avoiding alignment and grids altogether; better tempered molecular fields; “preventive medicines” such as inertial template alignment (which is somewhat related to topomers) and simple modified grid designs; and region focusing (weighting).

Adding grid points can reduce aliasing for unsampled field points, while removing grid points optimizes covariance between grid points. Bob discussed how to strike a reasonable balance using anisotropic spacing, and a face-centered cubic lattice to make CoMFA much less sensitive to alignment. He presented some plots of the effect of rotation on q^2 for the different lattices, and concluded that the sensitivity to positioning was less for the face-centered cubic grid, and the average performance was better as well.

Yvonne Martin (yvonnecmartin@comcast.net) presented a different perspective on the history of CoMFA. A molecule can be represented in 3D using shape, or electrostatic potential on a van der Waals surface or quantum chemical regions of high and low electron density, for example, but how do you convert these lovely 3D colored images into relevant descriptors for 3D QSAR? This is the problem that Dick started to address while he was working at Smith Kline & French (SK&F). He and Margaret Wise[23] described molecules by coarse steric and electrostatic energy maps calculated from the Boltzmann-weighted sum of the conformers of the compound. They derived descriptors using principal components analysis of the fields of the various molecules. The use of partial least squares (PLS) in solving underdetermined matrices was instrumental in helping Dick develop CoMFA. Svante Wold suggested this solution to Dick at the 1981 QSAR Conference. Most QSAR practitioners at that time did not know about PLS[24] or understand its power. CoMFA[1,25] was a descendent of DYLOMMS[23] combined with PLS, after Dick had left SK&F and associated with Garland Marshall, who had just founded Tripos. Dick's insight into the choice of fields for CoMFA is validated[26] by the observation that it well describes the traditional linear free energy descriptions, Hammett sigma constant and Taft E_s values.

Yvonne listed some key elements leading to innovation, each of which contributed to Dick's success. The four factors are: recognition that there is a problem, persistence in searching for a solution, creativity and insight in the search for a solution, and chance. The program GRID[27] is one example of an innovation, but its author Peter Goodford did not go on to invent CoMFA. He was aware of QSAR and the use of statistics in QSAR, but he did not focus on the problem of correlating the 3D properties of ligands with their biological potency. He failed to recognize the problem.

Yvonne's own team also missed the opportunity to invent CoMFA. Abbott had tested some compounds for diuretic activity and explained the progressive decrease in potency of these compounds as they occupy more and more new space compared to the most potent compounds. Extrapolating from the linear free energy relationship (LFER) explanation that the Taft E_s values are a function of the radius of the atom, her team wrote a program that generated 96 descriptors of shape as the length of vectors emanating from the first moment of inertia of the aligned molecules, and used statistics to derive the QSAR, but they did not find a good relationship with this dataset or others. What they missed was the correct description of molecules. Because they relied too much on the traditions from LFER, they failed on creativity and insight in the search for a solution.

Corwin Hansch's work leading to the invention of QSAR[28,29] started in 1948 with his collaboration with Robert Muir, a botanist who happened to have an office in the chemistry building. Hansch and Muir emphasized the Hammett sigma constant in their work on plant growth regulators. After a decade of struggling with the Hammett relationship, Hansch decided to investigate a possible relationship to partitioning into the cell. He found precedents in the work of Runar Collander,[30] and others. At this point, he hired Toshio Fujita: the second bit of luck (after the chance of meeting Muir) that led to QSAR. Neither the Hammett constant nor log *P* describes the SAR but Fujita suggested that perhaps both properties contribute to the SAR. He also recognized the additive nature of log *P*. Hansch suggested a parabolic function in log *P* to account for an optimum value. There now was the problem of how to fit the data to the proposed equation. Fortunately, there was a faculty member of the geology department, Donald McIntyre, who was fascinated by the possible influence of computers on research. He not only convinced a donor to give a computer to the chemistry department, but he also coded up the multiple regression equation. Chance was, however not the only factor in the invention of QSAR: there were 15 years of persistence behind the innovation.

Yvonne discussed a few examples of prominent scientists who could have invented QSAR, but did not. The Fieser group had evidence for the additive and constitutive nature of lipophilicity but they seemed to be unaware of earlier work on partitioning. What the Fieser group did not do was to recognize that there is a general problem in structure-activity relationships, that calculating lipophilicity would be a valuable exercise, and that multiple factors might contribute to potency. Brodie and Schanker studied drug absorption in 1960 but missed inventing QSAR mainly because they did not realize the general nature of the problem, because they did not know about Collander's work on octanol, and because they did not follow the LFER field, but especially they failed because they did not think to apply statistics to their relationships. So they failed on both persistence and insight.

Another example of the role of chance in innovation comes from Yvonne's own team.[31] They knew that they could not do CoMFA unless they knew how to choose conformations and how to align a diverse set of molecules. The only literature solutions required choosing the atoms to match. By chance, Yvonne read a paper by Brint and Willett[32] and realized that a pharmacophore is just a 3D maximal common substructure, but one in which the points are not atoms, but pharmacophore features. Chance, rather than persistence was the innovation factor here. Two other groups[33,34] worked on the alignment problem. Both provided means to select corresponding conformations, but as input they required the atoms or features that correspond in the various molecules, and this is not always obvious. They ignored part of the problem: recognition was the failure point in this case.

In conclusion, the fact that invention requires so many elements to coalesce does not negate the powerful role of persistent focus on attempting various solutions to the problem.

More on QSAR

The next speaker should have been Tony Hopfinger (hopfingr@gmail.com) but on the morning of the symposium he was taken ill. I had a copy of his slides and Dick Cramer valiantly attempted to present the paper in Tony's absence. Clearly anything I write in this article will be a poor reflection of what Tony might have said had he been there in person.

Tony worked with Dick, while Dick was at SK&F, to provide the structure generator eventually commercialized as ChemLab.[35] The two of them had an argument at an ACS meeting in Houston, Texas before the first CoMFA publication appeared. The issues were field *versus* overlap volume descriptors, conformation and alignment. Tony and Dick agreed to continue to disagree. Dick went on to gain fame from fields and CoMFA. Tony went on to develop Molecular Shape Analysis[36] and a find it a dead-end, but then, in an epiphany 4D-QSAR analysis was born.[37] The fourth "dimension" in the paradigm is *sampling* and includes the sampling of conformation, alignment, pharmacophore sites and entropy. The composite information coming from each of these sampled property sets is embedded in the resulting QSAR model.

The descriptors in 4D-QSAR analysis are the grid cell (spatial) occupancy measures of the atoms composing each molecule in the training set realized from the sampling of conformation and alignment spaces. A single "active" conformation can be postulated for each compound in the training set and combined with the optimal alignment for use in other molecular design applications including other 3D-QSAR methods. The influence of the conformational entropy of each compound on its activity can be estimated. Serial use of PLS, regression and a genetic algorithm (GA) is used to perform data reduction and identify the manifold of top 3D-QSAR models for a training set. The unique manifold of 3D-QSAR models is arrived at by computing the extent of orthogonality in the residuals of error among the most significant 3D-QSAR models in the general GA population. The models can be graphically represented by plotting the significant 3D-QSAR grid cells in space along with their descriptor attributes.

4D-QSAR is used to create and screen against 3D-pharmacophore QSAR models and can be used in receptor-independent or receptor-dependent modes. More recently Tony introduced a pseudo structure-based method, Membrane-Interaction QSAR analysis,[38,39] to estimate a wide range of ADME and toxicity endpoints based on interaction of test compounds with models of cellular membranes and a set of unique property descriptors.

The n-dimensional QSAR themes used in Tony' slides were conformation, alignment, spatial descriptors, the pharmacophore, whether or not to include the receptor, and what to do with conflicting or weird results. He reckons that he is now in a position to identify, probe, and think meaningfully about, but not perhaps solve, how to handle the many obstacles that have long plagued nD-QSAR analysis. Examples with respect to conformation are:

- How to completely explore conformations (MD, MC, or brute-force).
- How to handle receptor-independent and receptor-dependent searches.
- How to set limits on upper energies of ligand conformations and ligand receptor complexes.
- How to model large geometric changes in receptor geometry.

He wondered whether we should we let X-ray and NMR do the "heavy lifting" and let modeling come in for clean-up and refinement.

It is clear that there are still differences of opinion between Dick and Tony, but Tony concluded with a very fitting tribute to Dick as a colleague. This was in the form of the last two lines of a poem by William Butler Yeats, somewhat paraphrased: *"When I think where man's glory most begins and ends, I say my glory is to have such a good* <u>*and questioning*</u> *friend."*

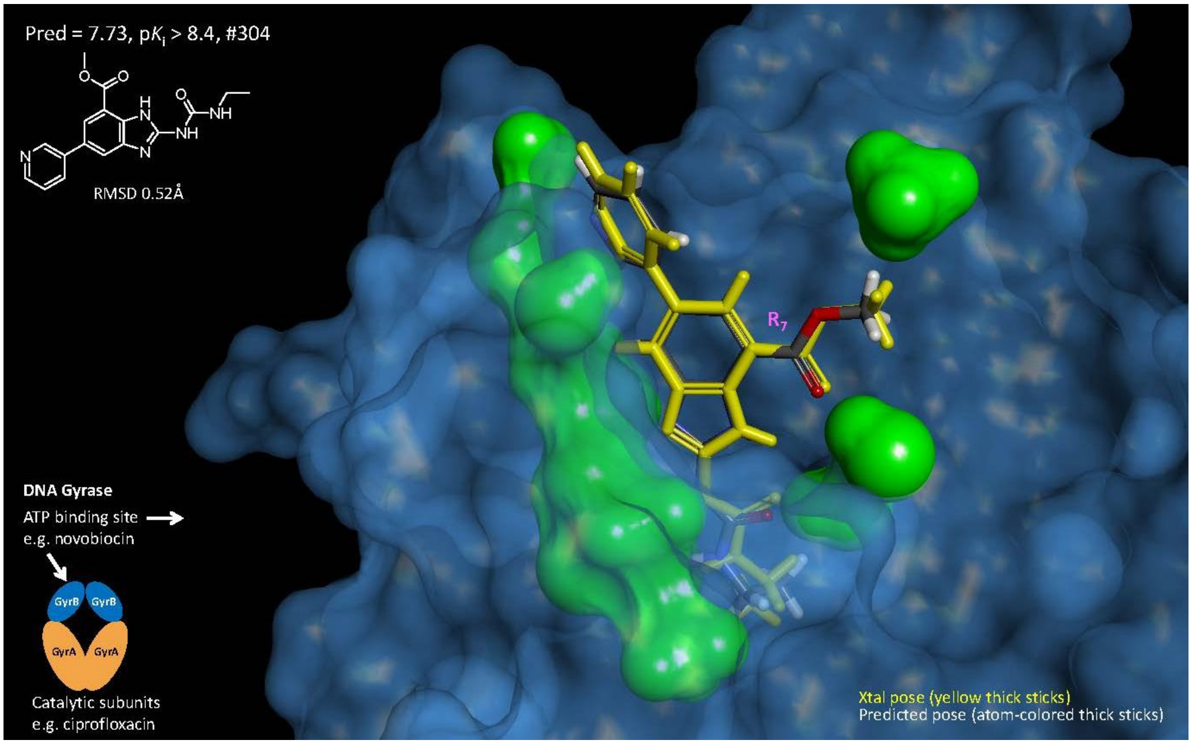
Ajay Jain (ajain@jainlab.org) of the University of California San Francisco has developed a family of 3D QSAR and docking approaches. Ajay showed Figure 6 of Dick's much-cited CoMFA paper.[1] It shows the major steric features of the QSAR for steroid binding to testosterone-binding globulin (TeBG). In this work Dick illuminated a new and exciting path for our field: his model predicted the right thing for the right reasons. Ajay's initial work with Compass[40,41] created a linkage between model and molecular pose. Compass involved a new representational scheme for capturing the 3D surface properties of small molecules that made it possible to address systematically the choice of the relative alignment and conformation (or pose) of competitive ligands including the detailed relationship of their hydrophobic shapes. A key insight was that the choice of pose should be directly governed by the function being used to predict binding affinity (essentially a direct analogy to physics where the lowest energy state is sought). The difficulty was that the function to predict activity was being induced at the same time as the pose choice. The Compass method overcame this problem, and was one of the foundational methods in establishing the field of multiple-instance learning.

Ajay showed a model of dihydrotestosterone (1D2S) binding to TeBG. If you make a small change to the steroid, to 1LHO, the alignment shifts a little. If you use estradiol (ILHV) the alignment flips. The protein moves too. These bidirectional relationships must be modeled in 3D QSAR. Because substituent modifications affect molecular pose, effects on activity will often be non-additive. Jain believes that it is vital to address the basic physical realities of protein ligand binding.

For QSAR as physical modeling, there must be a direct linkage between the model and molecular pose: if the model changes, the poses will as well; if substituents are changed, alignments will as well. Details of molecular shape and electrostatic properties have to matter to the model; non-additive behavior should be a natural consequence; and the models should have a direct relationship to physical protein binding pockets.

The QMOD approach[42] takes QSAR to a new level, by transforming the problem into one of molecular docking. A protein binding site is induced given SAR data using the multiple-instance machine learning paradigm developed for Compass. A skin is built around a small molecule pose, inducing a binding pocket that explains the data, so that you can predict the activity and geometry of new ligands. Model construction is fully automated. The agnostic Surflex-QMOD hypothesis for TeBG cares about the surfaces, not the atoms.

Ajay's student Rocco Varela has applied QMOD to 426 Vertex gyrase inhibitors.[43] He performed an iterative, temporal lead optimization exercise. A series of gyrase inhibitors with known synthetic order formed the set of molecules that could be selected for "synthesis." Beginning with a small number of molecules, based only on structures and activities, a model was constructed. Compound selection was done computationally, each time making five selections based on confident predictions of high activity and five selections based on a quantitative measure of three-dimensional structural novelty. Compound selection was followed by model refinement using the new data. Iterative computational candidate selection produced rapid improvements in selected compound activity, and incorporation of explicitly novel compounds uncovered much more diverse active inhibitors than strategies lacking active novelty selection. One of Rocco's models was chosen for the cover of the *Journal of Medicinal Chemistry*.



The pocket model actually looks like the experimentally determined gyrase pocket. QMOD is predicting the right thing for the right reason, just as Dick's CoMFA model predicted the right thing for the right reasons in 1988.

Drug Discovery

Bobby Glen of the University of Cambridge (rcg28@cam.ac.uk) moved on from adventures in "CoMFA-land" to adventures in drug discovery. He started by contrasting computation with reality. We do a calculation, but we do not know the correctness of our prediction until we see the results of the experiment, and the "experiment" may be the patient who takes the drug. Our objective is to mimic the real world of the patient as closely as possible, but describing molecules is difficult[44] so we make approximations, but when we do our calculations we need to think about what happens in the real world. We are interested in the properties of molecules not so much in what they *are* but in what they *do*.[45] In the real world a drug does lots of things, especially to sick people. A drug tested in a 25-year old male Olympic rower will have very different effects on a 67-year female patient with multiple chronic conditions.[46] Compounds show different physiological effects because of many different reasons and the multiple mechanisms are hard to model in a single structure activity relationship. Toxicity is often unexpected and is discovered in the clinic.

Computational methods are evolving to address the complexity of the process; the nature of drug discovery is now multivariate, and more and more data are becoming available. It is now possible to construct bioprints of molecules and their effects on multiple receptor systems. We can also introduce the effects of other biological systems such as transport and metabolism.[47,48]

Bobby's team has developed MetaPrint2D software (http://www-metaprint2d.ch.cam.ac.uk/metaprint2d) to predict the sites and products of metabolism.[49,50] In an example, Bobby input the SMILES for a partial agonist which has a main metabolite which is a full agonist. So, as the drug concentration lowers in blood, the remaining compound becomes more potent. In another example, the toxicity of acetaminophen (paracetamol) is predicted and the two relevant metabolic pathways are displayed. The primary pathway is glucuronidation which yields a relatively non-toxic metabolite, but at higher doses this pathway is saturated, and N-acetyl-p-benzoquinone is produced, causing liver damage.

A drug's activity can be modified by metabolism. Bobby showed the predicted metabolic pathways of promazine, and the predicted activities of some metabolites. It is also possible to predict "in reverse" and identify prodrugs. Bobby showed some biological effects of a promazine metabolite: effects which may possibly relate to phenotypic changes. The terminal metabolite thiodiphenylamine was predicted to be active against amine oxidase, cycloxygenase 1 and 2, and the sodium-dependent noradrenaline transporter.

Bobby concluded that drug discovery is developing holistic tendencies, driven by access to "Big Data", faster processing and, most of all, more complete algorithms, and, of course, more experimental validation. He paid tribute to Dick for being at the forefront of this revolution: he was one of the first to use multivariate data in CoMFA, and before that he was using multi-dimensional property visualization.

Tudor Oprea (toprea@salud.unm.edu) acknowledged that CoMFA had a huge influence in his own career. The lesson he learned from learned from Dick Cramer and Dave Patterson was "If you can't be right, be consistent". In his talk entitled "Think Local, Act Global" he said that in a Newtonian Universe it would be possible to predict the future, but we do not live in one. In chemical space, as in geography, maps need to be consistent. Tudor's chemical global positioning system, ChemGPS, makes a drugspace map by systematically applying conventions when examining chemical space, in a manner similar to the Mercator convention in geography. Chemography is the art of navigating in chemical space.[51,52] Rules are equivalent to dimensions (e.g., longitude and latitude), while structures are equivalent to objects (e.g., cities and countries). Selected rules include size, lipophilicity, polarizability, charge, flexibility, rigidity, and hydrogen bond capacity. Core structures include most marketed drugs with good oral permeability, as well as other biologically active compounds, while "satellites" are intentionally placed outside the chemical space of drugs, and include molecules having extreme values in one or more of the dimensions of interest. The map coordinates are *t*-scores extracted by principal component analysis (PCA) from 72 descriptors that evaluate the rules on a total set of 423 satellite and core structures. The PCA method, and ChemGPS, were inspired by Dick Cramer's BC(DEF) work.[53]

By successfully combining virtual and biomolecular screening, Tudor's team at the University of New Mexico discovered G-1, the first GPR30-specific agonist, capable of activating GPR30 in a complex environment of classical and new estrogen receptors.[54] They used a composite approach. 2D fingerprint technologies are really fast but they can lead you into a local trap: if you use a steroid as a query, the high-similarity hits will almost all be steroids. 3D technologies are not as fast as 2D, and require a choice of conformers but if you submit a rigid steroid as query, the chances are that you will find fewer steroids. 3D approaches include the ROCS shape-based method (http://www.eyesopen.com/rocs) and ALMOND (http://www.moldiscovery.com/soft_almond.php) based on pharmacophores. Cristian Bologa used a weighting scheme of 40% 2D (MDL and Daylight fingerprints), 40% shape, and 20% ALMOND with the intention of screening the top 100 hits, analyzing the primary hits and then fine-tuning the weighting scheme. In practice he got lucky: hits number 58 and 65 bound to ERα and ERβ, hit number 95 was G-1, and the hits were active in secondary assays. The team went on to identify a potent GPR30 antagonist.[55]

Data reliability can be a problem, but often goes unrecognized. In one previously used[56] dataset for human intestinal absorption (HIA), sulfasalazine was wrong because bacterial azo bond reduction occurs in the intestine and the measured HIA value was that of a metabolite. After removing two azo-containing drugs, as well as two drugs absorbed by paracellular mechanism, the bottom end of the sigmoidal curve describing Caco-2 absorption was removed, with little or no sigmoidal effect left.

The Biopharmaceutics Drug Disposition Classification System (BDDCS)[57] has four categories: class 1 high solubility and extensive metabolism, class 2 low solubility and extensive metabolism, class 3 high solubility and poor metabolism, and class 4 low solubility and poor metabolism. Tudor and his colleagues have compiled the BDDCS classification for 927 drugs.[58] They have also reported a computational procedure for predicting BDDCS class from molecular structures.[59] Transporter effects in the intestine and the liver are not clinically relevant for BDDCS class 1 drugs, but potentially can have a high impact for BDDCS class 2 (efflux in the gut, and efflux and uptake in the liver) and class 3 (uptake and efflux in both gut and liver) drugs. A combination of high dose and low solubility is likely to cause BDDCS class 4 to be under-populated in terms of approved drugs.[59] The model reported by Tudor and co-workers showed highest accuracy in predicting classes 2 and 3 with respect to the most populated class 1. For class 4 drugs a general lack of predictability was observed.

BDDCS has also been used to improve blood brain barrier predictions of oral drugs.[60] BDDCS class membership was integrated with *in vitro* P-gp efflux and *in silico* permeability data to create a classification tree that accurately predicted CNS disposition for more than 90% of 153 drugs in the dataset. Medicinal chemists are often taught that second generation antihistamines are successful due to log*P* optimization, which supposedly leads to little or no blood brain barrier (BBB) penetration. Tudor's team has shown that this is not true, since neither log*P* nor log*D* distribution differ between first generation (BBB penetrating) and second generation antihistamines.[61] They compared 64 H1R antagonists the log*P* and log*D* profiles of which overlap. The nine that are *effluxed* by P-gp include all second generation antihistamines. For these, P-gp becomes, *de facto*, a drug target.

In some work with Scott Boyer, Tudor examined the CEREP BioPrint dataset (http://www.cerep.fr/Cerep/Users/pages/ProductsServices/BioPrintServices.asp). The total number of potential activities was 371,448, whereas the total number of observed activities was 31,264, leading to a probability of 8.41% for observing bioactivity. They defined "biased targets" as those that exceed the 8.41% probability, and noticed that biased targets account for 76.81% of all activities in the CEREP dataset. Tudor and co-workers further looked at 871 chemicals measured in 131 DrugMatrix assays (http://ntp-server.niehs.nih.gov/?objectid=72016020-BDB7-CEBA-F3E5A7965617C1C1) and found that biased targets account for 83.34% of the activities in DrugMatrix. Tudor's pie-charts showed only partial overlap of chemicals and targets in CEREP and DrugMatrix; data-by-data comparison has revealed several molecular-target sets for which the overlap of bioactives in CEREP and DrugMatrix is zero, substantiating the need for accurate assay annotation and proper bioassay ontologies such as the work done by Stefan Schurer (http://bioassayontology.org).

If you have an assay, you have information; if you have two assays for the same target, you may have confirmation, or confusion. Do we really need big data when we often cannot handle small data? Human curation and attention to detail are needed before decision-making is well-served by experiment and model.

Sabermetrics

And so to something completely different: David Smith (dwsmith@retrosheet.org) focused on the philosophy of science and how this relates to many different kinds of inquiry including baseball research. Science is a procedure for study, largely independent of the topic under investigation. Nowadays this definition is being blurred in discussions of STEM disciplines, missing the point that science is special because of how questions are analyzed, not because of what is studied. Louis Pasteur said "*There are no such things as applied sciences, only applications of science*". Economics is an example of the application of scientific methods to an important area that is not a natural science. Since we are not defining science by what is studied, we need to define it in term of key features: definition of questions and proper criteria for evaluation. This is seen in the classical formulation of hypothesis, experiment and conclusion.

Not all areas of traditional science fit neatly into this paradigm. Astronomy, for example, is a scientific discipline but Copernicus depended primarily on observation, not on manipulation. Evolutionary biology, David's own discipline, is another example. In astronomy, the Copernican proposal of a heliocentric solar system made sense of a number of phenomena at a single stroke; there is no single observation that "proves" the theory. Before Darwin, biology was almost entirely descriptive and very fragmented. Darwin's proposal of natural selection provided the same sort of satisfying and unifying explanation that Copernicus did.

Natural history is the starting point for almost every scientific discipline: its observations eventually became organized and lent themselves to questions. At this point the study became scientific. Carl Linnaeus created classification systems for thousands of species of organisms, a feat of great organization but little analysis. The naturalist Alexander von Humboldt explored South America 30 years before Darwin, but his observations went further than mere cataloguing. The transition from natural history to science is perhaps best seen in the person of Charles Darwin who began his voyage on HMS Beagle as a naturalist, charged with collecting samples and making observations, but who after his return to England, spent 20 years organizing the material he had collected. He began to ask why certain patterns existed.

Evolution is often described as a historical science and so is baseball research. Evolutionary hypotheses and predictions are not about the future, but about an unknown past. In baseball, for decades the conventional wisdom was that the best batters were those who had the highest batting average, that is, the most hits per opportunity (at bat). Detailed study of modern events led to the hypothesis that reaching base by any means was of greater significance than base hits considered alone. Furthermore, advancing runners with extra base hits was historically undervalued. These two measures: reaching base (on base average) and advancing runners (slugging percentage) were combined to a single measure called OPS (on-base plus slugging) that was then used to examine baseball from 1901 through 2012. The results show a stronger correlation between runs per game and OPS than between runs per game and batting average. Note also that differences such as these are much easier to demonstrate when large datasets are available. Here we have a scientific result plus retrospective prediction.

The name given to this sort of work is Sabermetrics, a term based on "SABR", the acronym of the Society for American Baseball Research (http://sabr.org/), a national group of some 6000 members founded in 1971. When Dick and others began working on Sabermetrics in the late 1970s, detailed data were not readily available. Bill James then started to collect them and interesting studies could be carried out. For example, a stolen base is a valuable play that increases the chance of scoring, but the counterpart, a caught stealing, has a negative effect. A study showed that a stolen base attempt must be successful in at least two thirds of cases to be worth the risk. The importance of a first pitch strike has also been studied. If that first pitch is a strike because of a swing and a miss, then the pitcher usually has a good outcome, but if the first pitch is a foul ball, then the batter does better, and if that first pitch is hit into play, the batter does extremely well. A third example is clutch hitting: the assertion that some hitters increase their performance in tight situations. Dick did a sophisticated analysis (http://cyrilmorong.com/CramerClutch2.htm) to show that clutch hitting is an illusion. In addition to writing analysis software, Dick founded STATS, Inc., which began by gathering data for baseball studies but now covers other sports. Many baseball teams now use Sabermetrics and measures such as slugging are displayed on scoreboards.

David's group Retrosheet (http://www.retrosheet.org) has gathered play by play data for 165,000 of 185,000 games played since 1901 and has made it freely available on the Internet. Collection, digitization and publication of such data have led to a variety of Sabermetric analyses. Dick is one of the volunteers who examines images of old scorecards and converts them to digital form using specialized software. Baseball research offers unique opportunities to ask meaningful questions in a scientifically rigorous way and this is why professional scientists such as Dick and David are attracted to it.

Synthesis planning

In recognition of Dick's early work in the LHASA project, Todd Wipke (wipke@ucsc.edu) who published a seminal paper with Corey,[62] addressed the subject of synthesis planning at the award symposium. A much earlier paper by Corey's team[63] had a section entitled "Synthesis Plan" that discussed alternative disconnection plans: the reviewers did not like that section. In a later paper[64] Corey said "*the first task … should be an exhaustive analysis of the topological properties of the carbon network to define the range of possible precursors*". At that time Todd was generating all possible isomers of undecane and learning about 3D molecules and NMR. Corey and Wipke had different skill sets.

In 1967 the PDP-1 computer available at Harvard used 24K 18-bit words, drum storage, DECtape and paper tape, and the DECAL assembly language. It had three cathode ray tubes, a Rand tablet, and a joystick, and a Calcomp plotter for graphic output. The first synthesis planning program, Organic

Chemical Synthesis Simulation (OCSS), used *ab initio* mechanistic reactions in several steps to make a "name" reaction. Functional groups, rings, ring junctures, aromaticity, conjugation, and atom and bond classes were perceived. The logic-oriented approach uses clues in the target to predict a precursor: clues such as relationships of functional groups, functional group appendage relationships, ring sizes and ring junctures, and functional groups and rings. It was thus necessary to represent these entities. Chemists were excited by the Corey and Wipke publication;[62] there must be logic to synthesis planning because a computer can do it. Dick Cramer and Jeff Howe then joined the team[65,66] and Todd moved to Princeton.

Synthesis planning needed large programs and long term projects. Other problems included capturing reaction knowledge, granularity and consistency of the knowledge base, planning *versus* experimental detail, the shortage of trained, interested people, and the emergence of drug design as the new shiny toy. Chemists were turned off by predictions known to fail; they measured plans against empirical knowledge. Reaction databases were non-existent but chemists really wanted automated reaction retrieval. There was a drive toward specific representation, leading to a large number of transforms and to large synthesis trees.

The CGL computer at Princeton in 1969 used 64K 36-bit words and had a 5MB disk; it was a multi-user system. This was used for the Simulation and Evaluation of Chemical Synthesis (SECS) program[67] which featured interactive 3D energy minimization; an acoustic tablet for drawing and control; prediction of steric and electronic control; trigonal, tetrahedral, and trigonal bipyramidal stereochemistry; heterocyclic chemistry; metabolic reactions, and the ALCHEM language for transforms.

MDL's REACCS program was launched in 1980. It allowed reaction databases to be created. Classic reaction collections such as Theilheimer were digitized. These could be searched by structure, substructure, reaction centers, and even stereochemistry. Multi-step sequences were handled and full literature references were stored. REACCS enabled manual synthesis planning. Using selective databases such as Current Synthetic Methodology and Current Chemical Reactions it was found[68] that citation analysis with reaction substructure search allowed retrieval of reactions not even in the computer. Todd's team also worked on mining a large reaction database[69,70] to automate the building of the SECS knowledge base. Nowadays synthesis planning from a large reaction database is also available in Reaxys (http://www.elsevier.com/online-tools/reaxys).

Simply knowing the rules in chess does not make you a good chess player and the same can be said for SECS, where strategic control is necessary. A transform is the inverse of a synthetic reaction. Strategy is the problem solving method referring only to molecular structures. A goal is the result of applying a particular strategy to a particular problem, and refers to structure. Character is a type of structural change resulting from a transform. Wipke showed a symmetry example: three goals for breaking bonds in the retrosynthesis of beta-carotene. He also showed some QED predicate calculus of a strategy, and a topological goals chart. An important innovation was the separation of strategy from transforms.

Many companies used SECS. It was converted to a timesharing application and given a graphical GUI. A patent attorney wondered whether a synthesis produced by SECS were patentable. The program technology was adopted for other uses too. Students learned the logic of synthesis planning and synthesis papers included planning, just as Corey had anticipated in the 1960s.

Topomers

The final two papers in the symposium brought us up to date with Dick's current research interests. Bernd Wendt (Bernd.Wendt@certara.com) gave an overview of a wide array of topomer applications. To avoid duplication, I am leaving a detailed description of topomer technology until later. The first topomer application, in 1996 was ChemSpace,[6,71] used in library design for general screening. It was followed by the DBTOP shape similarity search tool for activity mining, topomer CoMFA,[72-75] AllChem,[76] a library of 10^{20} synthesizable structures, Quantitative Series Enrichment Analysis (QSEA)[77] for SAR mining, and, in 2013, Whole-Template CoMFA (WTC) to compare X-ray with template-based alignments.

Topomer shape similarity searching is very fast and increases the probability of finding active compounds. DBTOP for prospective selection of screening candidates by topomeric similarity was implemented as an automated workflow at Tripos Discovery Research leading to 308 selected compounds, and 11 successful "lead hops" in 13 assays.[8]

More recently, Bernd and his colleagues[78] identified a series of potent toluidinesulfonamide HIF-1 inhibitors, but the series was threatened by a potential liability to inhibit CYP2C9 which could cause dangerous drug–drug interactions. They then used structure-activity data from PubChem to develop a topomer CoMFA model that guided the design of novel sulfonamides with high selectivity for HIF-1 over CYP2C9 inhibition.

With Dick Cramer, Bernd examined the composition of 16 published QSAR datasets using Quantitative Series Enrichment Analysis (QSEA),[77] a procedure based on topomer technologies. QSEA allows the extraction of structure-activity relationships from large chemogenomic spaces starting from a single chemical structure. A heat map display in combination with topomer CoMFA and a novel series trajectory analysis revealed information for the assembly of structures into meaningful series. Global and local centroid structures can be determined from a similarity distance matrix and they build the origins for stepwise model building by increasing the similarity radius around the centroid nucleus. Bernd and Dick were able to determine whether compounds belonged to an emerging structure-activity relationship, and which compounds can be predicted within reliable limits.

QSEA has also been used in modeling off-target effects.[79] Queries were taken from the Jain set of marketed drugs to mine PubChem, ChemBank, and ChEMBL. SAR tables were constructed by assembling similar structures around each query structure that have an activity record for a particular target. QSEA was applied to these SAR tables to identify trends and to transform these trends into topomer CoMFA models. These models were able to highlight the structural trends associated with various off-target effects of marketed drugs, including cases where other structural similarity metrics would not have detected an off-target effect. One SAR trend identified was that fentanyl is inactive on hERG.

WTC is a current research project. Bernd and Dick took three datasets published by Brown and Muchmore (75 compounds tested against urokinase, 110 PTP-1B compounds and 123 Chk1-kinase compounds)[80] and aimed to develop CoMFA and CoMSIA[81] models for X-ray ligand poses and multitemplate aligned ligand poses, and then compare model robustness and interpretability and examine fluctuations of grid point interaction energies. In the three datasets having an experimental X-ray structure for every tested molecule, WTC alignment yielded CoMFA models which, compared to the "all-X-ray aligned" CoMFA models, provided equal or better statistical quality and seemingly superior interpretability and utility.

Dick Cramer's (Richard.Cramer@certara.com) award address homed in on Whole Template CoMFA. In theory, the primary cause of potency differences among ligands is steric and electrostatic field differences. Dick noted that when the goal is an informative comparison of ligand field differences, increasing ligand shape similarity is at least as productive as increasing physicochemical precision. As Tudor had observed earlier, if you cannot be sure of physical models, you can at least try to be consistent. Whole template CoMFA achieves ligand shape similarity by "copying" coordinates from any atom within a template ligand that "matches" a candidate's atom, and by using the topomer protocol to generate coordinates for the remaining "non-matching" atoms.

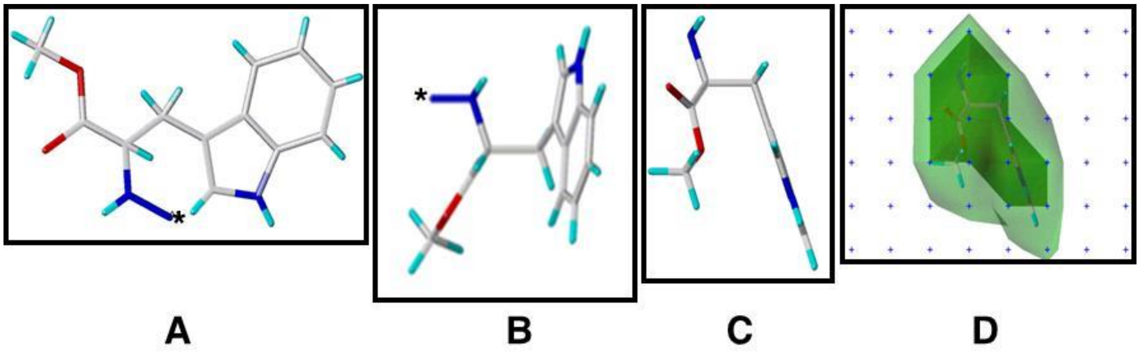
Dick has published four prospective "make and test" outcomes from topomer CoMFA.[82] Shape similarity is highly productive for a number of reasons. All QSARs seek to explain differences in training set potencies. For example, the difference could be down to the substitution of fluorine for hydrogen. The gridded fields of 3D-QSAR's descriptors directly and predictably express this: the differences in the local fields caused by changing hydrogen to fluorine may cause substantial change in the ligands' and the receptor's binding geometries. Furthermore, the frequency of chance correlation using PLS[83] is much lower than that for stepwise multiple regression, but perfect correlations involving descriptor subsets are not detected by PLS if the number of irrelevant descriptors is excessive. In CoMFA applications, the probability of chance correlation is usually negligible. Docking a small library moves the core around, producing field variation that is noise, because an invariant core cannot have caused changes in biological activity. With PLS, such noise tends to obscure the direct, certain, and causative field variation adjacent to the hydrogen or fluorine. Topomer generation rules were developed to produce alignments that are identical wherever the structures being compared are identical, or similar wherever the structural differences are slight. Topomer CoMFA focuses field variation and the resulting 3D-QSAR onto those direct, certain, and causative effects of 2D structure variation.

In WTC you identify the best matching "anchor bond" in the "candidate" (the test or training set structure to be aligned) and orient the candidate by overlay of its anchor bond onto that of the template. Anchor bond identification can be entirely automatic; template manual plus candidate automatic; or entirely manual. The best matching anchor bond includes all the candidate atoms that match a template atom. You then copy the coordinates of the matched template atoms to the matching candidate atoms and position the unmatched candidate atoms, by attaching their CONCORD-generated fragments and applying the topomer protocol.

To identify the candidate bond that best matches any template bond, the software considers, in order: every template, both "directions" of any bond, similarity in eight "localized" bond properties (or, within identical Murcko skeletons, identical location), and fraction of heavy atoms that match template atoms. Fully automatic identification involves combinatorial comparison of all pairings of plausible candidate and template bonds, where “plausible” means that one of the atoms defining the bond must not be carbon, or the bond type must be double or triple, or one of the atoms defining the bond must be in a ring and attached to at least three non-hydrogen atoms.

Atom matching uses breadth-first traversal, starting from a possible pairing of anchor bonds, in two passes: exact matching of atom and bond types (match score = 2), and skeleton matching only (the default, with match score = 1). Coordinate copying, using depth-first traversal, occurs if the atom is alicyclic or in rings whose atoms completely match, and hybridization agrees. It does not occur if the atom is in a ring and there are non-matching atoms in that ring. To modify or extend an outcome, a user can add templates.

A topomer is a single 3D model of a monovalent fragment constructed by a “black-box”. The only input is the “2D structure” of a single fragment (**A** below) embedded in 3D space by superposing the open valence (**B**), using valence geometries (bonds, angles, and rings) from CONCORD (**B**), and torsions, stereochemistry, and ring flips from canonical rules (**C**). The resulting strain energy is ignored. Several series can be combined in WTC to give a single 3D-QSAR, objectively based on all data, and X-ray interpretable.



Dick presented some initial WTC results that indeed combine diverse structures into a single predictive 3D-QSAR model, and are derived automatically. He used all the Factor Xa inhibitors in Bindingdb and showed that a combined WTC model was better than the single series WTC models. For example, the q^2 value ranged from -0.843, for 15 compounds that binding with PDB code NFX, to 0.602 for 21 1FJS structures; the combined model had q^2 0.616. Twelve, mainly poor datasets were combined into one good one. Results were even better for MAP kinase P38 alpha inhibitors.

The q^2 values for the combined models are probably to some extent artifacts since Bindingdb ligands are subsets, probably chosen to provide the most docking challenges for the least computation, and leave-one-out q^2 is too pessimistic when a unique structural change produces a strong effect on potency. Nevertheless, as it turns out, this contrast in results requires that the potency effect of a field at any particular lattice point be uniform, regardless of the great diversity of training set structures that produce different field intensities at that point. Thus the combined WTC models worked “for the right reason”.

One application area suggested for WTC is off-target prediction. Topomer applications in that field have already been published.[78,79,84] WTC allows any scientist to carry out 3D-QSAR modeling. Different project team members receive different benefits. Synthetic chemists can simultaneously consider the tradeoffs between synthetic costs and likelihood of therapeutic benefit. For the computer-aided molecular design practitioner an automatic protocol allows more attention on the most important issues such as training set composition and assessing validity of project-critical predictions. For project leaders, WTC allows more complete consideration of the dozens of relevant biological endpoints and the astronomical numbers of possible structural modifications.

In summary, WTC is a ligand alignment protocol for classical CoMFA that uses as input only 3D template(s) and a 2D SAR table, thus providing fast and convenient throughput; objectively determined models; application of crystallographic and/or pharmacophoric constraints; and structurally unlimited applicability. As output, it enables rapid, objective, structurally unlimited potency predictions that so far are reasonably accurate; contour maps that are more structurally informative; 3D database searching with potency predictions; and *de novo* design constrained by potency prediction. Its 3D-QSAR models can combine multiple series within a single model and be generated completely automatically.

Conclusion

The symposium was ably chaired by Brian Masek and Terry Stouch. After Dick's award address, Tony Williams, chair of the ACS Division of Chemical Information, formally presented the Herman Skolnik Award:

References

(1) 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.

(2) 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.

(3) Cramer, R. D.; Clark, R. D.; Patterson, D. E.; Ferguson, A. M. Bioisosterism as a Molecular Diversity Descriptor: Steric Fields of Single "Topomeric" Conformers. *J. Med. Chem.* **1996,** *39* (16), 3060-3069.

(4) Clark, R. D.; Cramer, R. D. Taming the combinatorial centipede. *CHEMTECH* **1997,** *27* (5), 24-31.

(5) Clark, R. D.; Ferguson, A. M.; Cramer, R. D. Bioisosterism and molecular diversity. *Perspect. Drug Discovery Des.* **1998,** *9/10/11* (3D QSAR in Drug Design: Ligand/Protein Interactions and Molecular Similarity), 213-224.

(6) Cramer, R. D.; Patterson, D. E.; Clark, R. D.; Soltanshahi, F.; Lawless, M. S. Virtual Compound Libraries: A New Approach to Decision Making in Molecular Discovery Research. *J. Chem. Inf. Comput. Sci.* **1998,** *38* (6), 1010-1023.

(7) Clark, R. D.; Brusati, M.; Jilek, R.; Heritage, T.; Cramer, R. D. Validating novel QSAR descriptors for use in diversity analysis. In *Molecular Modeling and Prediction of Bioactivity, Proceedings of the European Symposium on Quantitative Structure-Activity Relationships: Molecular Modeling and Prediction of Bioactivity , 12th, Copenhagen, Denmark, Aug. 23-28, 1998;* Gundertofte, K.; Jorgensen, F. S., Eds.; Kluwer Academic/Plenum Publishers: New York, NY, 2000; pp 95-100.

(8) Cramer, R. D.; Jilek, R. J.; Guessregen, S.; Clark, S. J.; Wendt, B.; Clark, R. D. "Lead Hopping". Validation of Topomer Similarity as a Superior Predictor of Similar Biological Activities. *J. Med. Chem.* **2004,** *47* (27), 6777-6791.

(9) Clark, R. D. Synthesis and QSAR of herbicidal 3-pyrazolyl α,α,α-trifluorotolyl ethers. *J. Agric. Food Chem.* **1996,** *44* (11), 3643-3652.

(10) Clark, R. D.; Leonard, J. M.; Strizhev, A. Pharmacophore models and comparative molecular field analysis (CoMFA). In *Pharmacophore Perception, Development, and Use in Drug Design;* Güner, O. F., Ed.; International University Line: La Jolla, CA, 1999; pp 153-167.

(11) Clark, R. D.; Sprous, D. G.; Leonard, J. M. Validating models based on large data sets. In *Rational Approaches to Drug Design. (Proceedings of the 13th European Symposium on Quantitative Structure-Activity Relationships, held 27 August-1 September 2000, in Dusseldorf, Germany.);* Holtje, H. D.; Sippl, W., Eds.; Prous Science: Barcelona, Spain, 2001; pp 475-485.

(12) Wolohan, P. R. N.; Clark, R. D. Predicting drug pharmacokinetic properties using molecular interaction fields and SIMCA. *J. Comput.-Aided Mol. Des.* **2003,** *17* (1), 65-76.

(13) Clark, R. D. Boosted leave-many-out cross-validation: the effect of training and test set diversity on PLS statistics. *J. Comput.-Aided Mol. Des.* **2003,** *17* (2-4), 265-275.

(14) Clark, R. D.; Fox, P. C. Statistical variation in progressive scrambling. *J. Comput.-Aided Mol. Des.* **2004,** *18* (7-9), 563-576.

(15) Clark, R. D. A ligand's-eye view of protein binding. *J. Comput.-Aided Mol. Des.* **2008,** *22* (6-7), 507-521.

(16) Clark, R. D. DPRESS: Localizing estimates of predictive uncertainty. *J Cheminform* **2009,** *1* (1), 11.

(17) Clark, R. D. Prospective ligand- and target-based 3D QSAR: state of the art 2008. *Curr. Top. Med. Chem. (Sharjah, United Arab Emirates)* **2009,** *9* (9), 791-810.

(18) Clark, M.; Cramer, R. D., III; Jones, D. M.; Patterson, D. E.; Simeroth, P. E. Comparative molecular field analysis (CoMFA). 2. Toward its use with 3D-structural databases. *Tetrahedron Comput. Methodol.* **1990,** *3* (1), 47-59.

(19) Cho, S. J.; Tropsha, A. Cross-Validated R2-Guided Region Selection for Comparative Molecular Field Analysis: A Simple Method To Achieve Consistent Results. *J. Med. Chem.* **1995,** *38* (7), 1060-1066.

(20) Norinder, U. Single and domain mode variables selection in 3D QSAR applications. *J. Chemom.* **1996,** *10* (2), 95-105.

(21) Kroemer, R. T.; Hecht, P. Replacement of steric 6-12 potential-derived interaction energies by atom-based indicator variables in CoMFA leads to models of higher consistency. *J. Comput.-Aided Mol. Des.* **1995,** *9* (3), 205-212.

(22) Lindgren, F.; Geladi, P.; Wold, S. Kernel-based pls regression; cross-validation and applications to spectral data. *J. Chemom.* **1994,** *8* (6), 377-389.

(23) Wise, M.; Cramer, R. D.; Smith, D.; Exman, I. Progress in three-dimensional drug design: the use of real time colour graphic and computer postulation of bioactive molecules in DYLOMMS. In *Pharmacochemistry Library, Vol. 6: Quantitative Approaches to Drug Design;* Dearden, J. C., Ed.; Elsevier: Amsterdam, The Netherlands, 1983; pp 145-146.

(24) Wold, S.; Martens, S.; Wold, H. The Multivariate Calibration Problem in Chemistry Solved by the PLS Method. In *Matrix Pencils: Proceedings of a Conference Held at Pite Havsbad, Sweden, March 22-24, 1982 (Lecture Notes in Mathematics);* Kagström, B.; Ruhe, A., Eds.; Srpinger Verlag: Heidelberg, Germany, 1983; pp 286-293.

(25) Cramer, R. D., III; Wold, S. B. Comparative molecular field analysis (CoMFA). US5025388A, 1991.

(26) Kim, K. H.; Martin, Y. C. Evaluation of electrostatic and steric descriptors for 3D-QSAR: the hydrogen ion and methyl group probes using comparative molecular field analysis (CoMFA) and the modified partial least squares method. *Pharmacochem. Libr.* **1991,** *16* (QSAR: Ration. Approaches Des. Bioact. Compd.), 151-154.

(27) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985,** *28* (7), 849-857.

(28) Hansch, C.; Fujita, T. ρ-σ-πAnalysis; method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* **1964,** *86* (8), 1616-1626.

(29) Fujita, T.; Iwasa, J.; Hansch, C. A new substituent constant, π, derived from partition coefficients. *J. Am. Chem. Soc.* **1964,** *86* (23), 5175-5180.

(30) Collander, R. Partition of organic compounds between higher alcohols and water. *Acta Chem. Scand.* **1951,** *5*, 774-780.

(31) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J. Comput.-Aided Mol. Des.* **1993,** *7* (1), 83-102.

(32) Brint, A. T.; Willett, P. Algorithms for the identification of three-dimensional maximal common substructures. *J. Chem. Inf. Comput. Sci.* **1987,** *27* (4), 152-158.

(33) Sheridan, R. P.; Nilakantan, R.; Dixon, J. S.; Venkataraghavan, R. The ensemble approach to distance geometry: application to the nicotinic pharmacophore. *J. Med. Chem.* **1986,** *29* (6), 899-906.

(34) Dammkoehler, R. A.; Karasek, S. F.; Shands, E. F. B.; Marshall, G. R. Constrained search of conformational hyperspace. *J. Comput.-Aided Mol. Des.* **1989,** *3* (1), 3-21.

(35) Pearlstein, R. A.; Malhotra, D.; Orchard, B. J.; Tripathy, S. K.; Potenzone, R., Jr.; Grigoras, S.; Koehler, M.; Mabilia, M.; Walters, D. E.; Doherty, D.; Harr, R.; Hopfinger, A. J. Three-dimensional structure modeling and quantitative molecular design using CHEMLAB-II. *New Methods Drug Res.* **1988,** *2*, 147-174.

(36) Rhyu, K. B.; Patel, H. C.; Hopfinger, A. J. A 3D-QSAR Study of Anticoccidial Triazines Using Molecular Shape Analysis. *J. Chem. Inf. Comput. Sci.* **1995,** *35* (4), 771-778.

(37) 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.

(38) Iyer, M.; Tseng, Y. J.; Senese, C. L.; Liu, J.; Hopfinger, A. J. Prediction and mechanistic interpretation of human oral drug absorption using MI-QSAR analysis. *Mol. Pharm.* **2007,** *4* (2), 218-231.

(39) Santos-Filho, O. A.; Hopfinger, A. J. Combined 4D-fingerprint and clustering based membrane-interaction QSAR analyses for constructing consensus Caco-2 cell permeation virtual screens. *J. Pharm. Sci.* **2008,** *97* (1), 566-583.

(40) Jain, A. N.; Koile, K.; Chapman, D. Compass: Predicting Biological Activities from Molecular Surface Properties. Performance Comparisons on a Steroid Benchmark. *J. Med. Chem.* **1994,** *37* (15), 2315-2327.

(41) Jain, A. N.; Dietterich, T. G.; Lathrop, R. H.; Chapman, D.; Critchlow, R. E., Jr.; Bauer, B. E.; Webster, T. A.; Lozano-Perez, T. Compass: a shape-based machine learning tool for drug design. *J. Comput.-Aided Mol. Des.* **1994,** *8* (6), 635-652.

(42) Jain, A. N. QMOD: physically meaningful QSAR. *J. Comput.-Aided Mol. Des.* **2010,** *24* (10), 865-878.

(43) Varela, R.; Walters, W. P.; Goldman, B. B.; Jain, A. N. Iterative Refinement of a Binding Pocket Model: Active Computational Steering of Lead Optimization. *J. Med. Chem.* **2012,** *55* (20), 8926-8942.

(44) Glen, R. C. Connecting the virtual world of computers to the real world of medicinal chemistry. *Future Med. Chem.* **2011,** *3* (4), 399-403.

(45) Orchard, S.; Al-Lazikani, B.; Bryant, S.; Clark, D.; Calder, E.; Dix, I.; Engkvist, O.; Forster, M.; Gaulton, A.; Gilson, M.; Glen, R.; Grigorov, M.; Hammond-Kosack, K.; Harland, L.; Hopkins, A.; Larminie, C.; Lynch, N.; Mann, R. K.; Murray-Rust, P.; Lo, P. E.; Southan, C.; Steinbeck, C.; Wishart, D.; Hermjakob, H.; Overington, J.; Thornton, J. Minimum information about a bioactive entity (MIABE). *Nat. Rev. Drug Discovery* **2011,** *10* (9), 661-669.

(46) Gleeson, M. P.; Modi, S.; Bender, A.; Robinson, R. L. M.; Kirchmair, J.; Promkatkaew, M.; Hannongbua, S.; Glen, R. C. The challenges involved in modeling toxicity data in silico: a review. *Curr. Pharm. Des.* **2012,** *18* (9), 1266-1291.

(47) Koutsoukas, A.; Simms, B.; Kirchmair, J.; Bond, P. J.; Whitmore, A. V.; Zimmer, S.; Young, M. P.; Jenkins, J. L.; Glick, M.; Glen, R. C.; Bender, A. From in silico target prediction to multi-target drug design: Current databases, methods and applications. *J. Proteomics* **2011,** *74* (12), 2554-2574.

(48) Koutsoukas, A.; Lowe, R.; KalantarMotamedi, Y.; Mussa, H. Y.; Klaffke, W.; Mitchell, J. B. O.; Glen, R. C.; Bender, A. In Silico Target Predictions: Defining a Benchmarking Data Set and Comparison of Performance of the Multiclass Naive Bayes and Parzen-Rosenblatt Window. *J. Chem. Inf. Model.* **2013,** *53* (8), 1957-1966.

(49) Kirchmair, J.; Williamson, M. J.; Tyzack, J. D.; Tan, L.; Bond, P. J.; Bender, A.; Glen, R. C. Computational Prediction of Metabolism: Sites, Products, SAR, P450 Enzyme Dynamics, and Mechanisms. *J. Chem. Inf. Model.* **2012,** *52* (3), 617-648.

(50) Kirchmair, J.; Howlett, A.; Peironcely, J. E.; Murrell, D. S.; Williamson, M. J.; Adams, S. E.; Hankemeier, T.; van, B. L.; Duchateau, G.; Klaffke, W.; Glen, R. C. How Do Metabolites Differ from Their Parent Molecules and How Are They Excreted? *J. Chem. Inf. Model.* **2013,** *53* (2), 354-367.

(51) Oprea, T. I.; Gottfries, J. Chemography: The Art of Navigating in Chemical Space. *J. Comb. Chem.* **2001,** *3* (2), 157-166.

(52) Oprea, T. I. Chemical space navigation in lead discovery. *Curr. Opin. Chem. Biol.* **2002,** *6* (3), 384-389.

(53) Cramer, R. D., III BC(DEF) parameters. 1. The intrinsic dimensionality of intermolecular interactions in the liquid state. *J. Am. Chem. Soc.* **1980,** *102* (6), 1837-1849.

(54) Bologa, C. G.; Revankar, C. M.; Young, S. M.; Edwards, B. S.; Arterburn, J. B.; Kiselyov, A. S.; Parker, M. A.; Tkachenko, S. E.; Savchuck, N. P.; Sklar, L. A.; Oprea, T. I.; Prossnitz, E. R. Virtual and biomolecular screening converge on a selective agonist for GPR30. *Nat. Chem. Biol.* **2006,** *2* (4), 207-212.

(55) Dennis, M. K.; Burai, R.; Ramesh, C.; Petrie, W. K.; Alcon, S. N.; Nayak, T. K.; Bologa, C. G.; Leitao, A.; Brailoiu, E.; Deliu, E.; Dun, N. J.; Sklar, L. A.; Hathaway, H. J.; Arterburn, J. B.; Oprea, T. I.; Prossnitz, E. R. In vivo effects of a GPR30 antagonist. *Nat. Chem. Biol.* **2009,** *5* (6), 421-427.
(56) Oprea, T. I.; Gottfries, J. Toward minimalistic modeling of oral drug absorption1. *J. Mol. Graphics Modell.* **2000,** *17* (5/6), 261-274.
(57) Wu, C.-Y.; Benet, L. Z. Predicting Drug Disposition via Application of BCS: Transport/Absorption/ Elimination Interplay and Development of a Biopharmaceutics Drug Disposition Classification System. *Pharm. Res.* **2005,** *22* (1), 11-23.
(58) Benet, L. Z.; Broccatelli, F.; Oprea, T. I. BDDCS Applied to Over 900 Drugs. *AAPS J.* **2011,** *13* (4), 519-547.
(59) Broccatelli, F.; Cruciani, G.; Benet, L. Z.; Oprea, T. I. BDDCS Class Prediction for New Molecular Entities. *Mol. Pharmaceutics* **2012,** *9* (3), 570-580.
(60) Broccatelli, F.; Larregieu, C. A.; Cruciani, G.; Oprea, T. I.; Benet, L. Z. Improving the prediction of the brain disposition for orally administered drugs using BDDCS. *Adv. Drug Delivery Rev.* **2012,** *64* (1), 95-109.
(61) Broccatelli, F.; Carosati, E.; Cruciani, G.; Oprea, T. I. Transporter-mediated efflux influences CNS side effects: ABCB1, from antitarget to target. *Mol. Inf.* **2010,** *29* (1-2), 16-26.
(62) Corey, E. J.; Wipke, W. T. Computer-assisted design of complex organic syntheses. *Science* **1969,** *166* (3902), 178-192.
(63) Corey, E. J.; Ohno, M.; Mitra, R. B.; Vatakencherry, P. A. Total synthesis of longifolene. *J. Am. Chem. Soc.* **1964,** *86* (3), 478-485.
(64) Corey, E. J. General methods for the construction of complex molecules. *Pure Appl. Chem.* **1967,** *14* (1), 19-37.
(65) Corey, E. J.; Wipke, W. T.; Cramer, R. D., III; Howe, W. J. Computer-assisted synthetic analysis. Facile man-machine communication of chemical structure by interactive computer graphics. *J. Am. Chem. Soc.* **1972,** *94* (2), 421-430.
(66) Corey, E. J.; Wipke, W. T.; Cramer, R. D., III; Howe, W. J. Techniques for perception by a computer of synthetically significant structural features in complex molecules. *J. Am. Chem. Soc.* **1972,** *94* (2), 431-439.
(67) Wipke, W. T.; Whetstone, P. Graphic digitizing in 3-D. In *Computer Graphics*; ACM: New York, NY, 1971; Vol. 5, p 10.
(68) Wipke, W. T.; Vladutz, G. An alternative view of reaction similarity: citation analysis. *Tetrahedron Comput. Methodol.* **1990,** *3* (2), 83-107.
(69) Yanaka, M.; Nakaura, K.; Kurumisawa, A.; Wipke, W. T. Automatic knowledge base building for the organic synthesis design program (SECS). *Prog. Clin. Biol. Res.* **1989,** *291* (QSAR: Quant. Struct.-Act. Relat. Drug Des.), 147-150.
(70) Yanaka, M.; Nakamura, K.; Kurumisawa, A.; Wipke, W. T. Automatic knowledge base building for the organic synthesis design program (SECS). *Tetrahedron Comput. Methodol.* **1990,** *3* (6A), 359-375.
(71) Cramer, R. D.; Poss, M. A.; Hermsmeier, M. A.; Caulfield, T. J.; Kowala, M. C.; Valentine, M. T. Prospective Identification of Biologically Active Structures by Topomer Shape Similarity Searching. *J. Med. Chem.* **1999,** *42* (19), 3919-3933.
(72) Cramer, R. D. Topomer CoMFA: A Design Methodology for Rapid Lead Optimization. *J. Med. Chem.* **2003,** *46* (3), 374-388.
(73) Jilek, R. J.; Cramer, R. D. Topomers: A Validated Protocol for Their Self-Consistent Generation. *J. Chem. Inf. Comput. Sci.* **2004,** *44* (4), 1221-1227.
(74) Cramer, R. D.; Cruz, P.; Stahl, G.; Curtiss, W. C.; Campbell, B.; Masek, B. B.; Soltanshahi, F. Virtual Screening for R-Groups, including Predicted pIC50 Contributions, within Large Structural Databases, Using Topomer CoMFA. *J. Chem. Inf. Model.* **2008,** *48* (11), 2180-2195.

(75) Cramer, R. D. R-group template CoMFA combines benefits of "ad hoc" and topomer alignments using 3D-QSAR for lead optimization. *J. Comput.-Aided Mol. Des.* **2012,** *26* (7), 805-819.

(76) Cramer, R. D.; Soltanshahi, F.; Jilek, R.; Campbell, B. AllChem: Generating and searching 10^{20} synthetically accessible structures. *J. Comput.-Aided Mol. Des.* **2007,** *21* (6), 341-350.

(77) Wendt, B.; Cramer, R. D. Quantitative Series Enrichment Analysis (QSEA): a novel procedure for 3D-QSAR analysis. *J. Comput.-Aided Mol. Des.* **2008,** *22* (8), 541-551.

(78) Wendt, B.; Mulbaier, M.; Wawro, S.; Schultes, C.; Alonso, J.; Janssen, B.; Lewis, J. Toluidinesulfonamide Hypoxia-Induced Factor 1 Inhibitors: Alleviating Drug-Drug Interactions through Use of PubChem Data and Comparative Molecular Field Analysis Guided Synthesis. *J. Med. Chem.* **2011,** *54* (11), 3982-3986.

(79) Wendt, B.; Uhrig, U.; Bos, F. Capturing structure-activity relationships from chemogenomic spaces. *J. Chem. Inf. Model.* **2011,** *51* (4), 843-851.

(80) Brown, S. P.; Muchmore, S. W. Large-Scale Application of High-Throughput Molecular Mechanics with Poisson-Boltzmann Surface Area for Routine Physics-Based Scoring of Protein-Ligand Complexes. *J. Med. Chem.* **2009,** *52* (10), 3159-3165.

(81) 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.

(82) Cramer, R. D. Rethinking 3D-QSAR. *J. Comput.-Aided Mol. Des.* **2011,** *25* (3), 197-201.

(83) Clark, M.; Cramer, R. D., III The probability of chance correlation using partial least squares (PLS). *Quant. Struct.-Act. Relat.* **1993,** *12* (2), 137-145.

(84) Nisius, B.; Goeller, A. H. Similarity-Based Classifier Using Topomers to Provide a Knowledge Base for hERG Channel Inhibition. *J. Chem. Inf. Model.* **2009,** *49* (2), 247-256.