Herman Skolnik Award Symposium 2018

Honoring Gisbert Schneider

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

Introduction

Gisbert Schneider, a full professor at ETH Zürich, holding the Chair for Computer-Assisted Drug Design, received the 2018 Herman Skolnik Award for his seminal contributions to *de novo* design of bioactive compounds, and the application of these innovative design concepts in both academia and industry. He is recognized as being a pioneer in the integration of machine-learning methods into practical medicinal chemistry, and for his coining the phrases "scaffold-hopping" and "frequent hitter". A summary of his achievements has been published in the *Chemical Information Bulletin.* Gisbert was invited to present an award symposium at the Fall 2018 ACS National Meeting in Boston, MA. There were 10 speakers in addition to Gisbert himself:

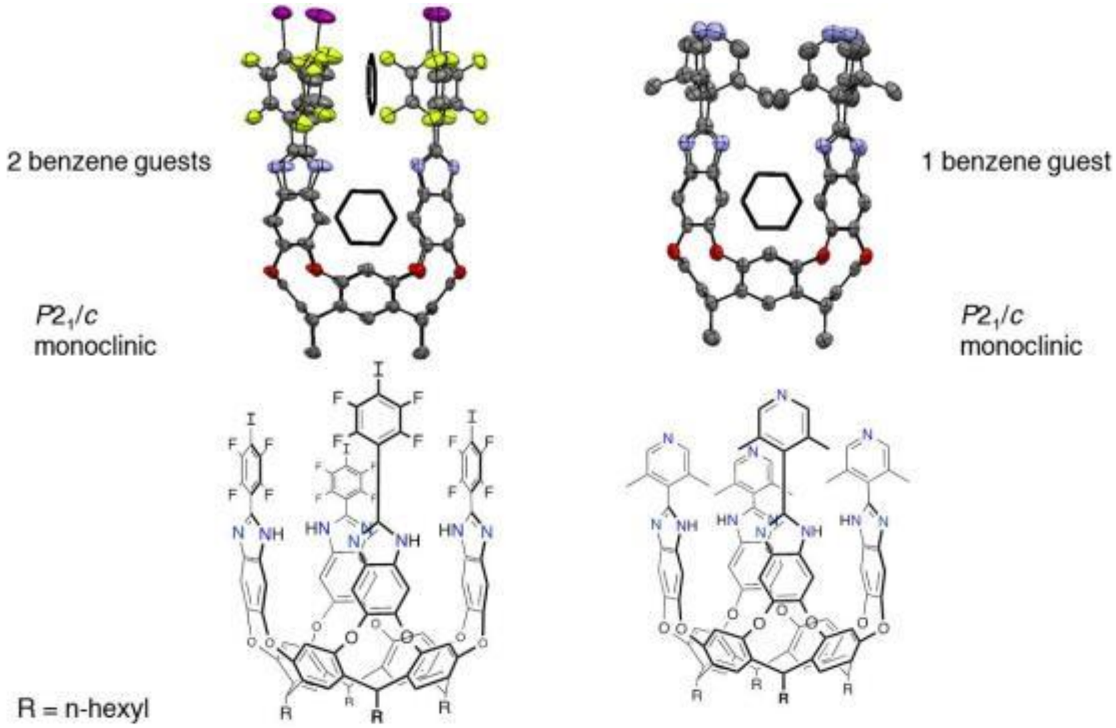
L to R: Ross King, Karl-Heinz Baringhaus, Gisbert Schneider, David Winkler, Jürgen Bajorath, Michael Schmuker, William Jorgensen, Yoshihiro Yamanishi, Kimito Funatsu, Alexandre Varnek, François Diederich

Molecular Recognition Studies to Advance Structure-Based Drug Design

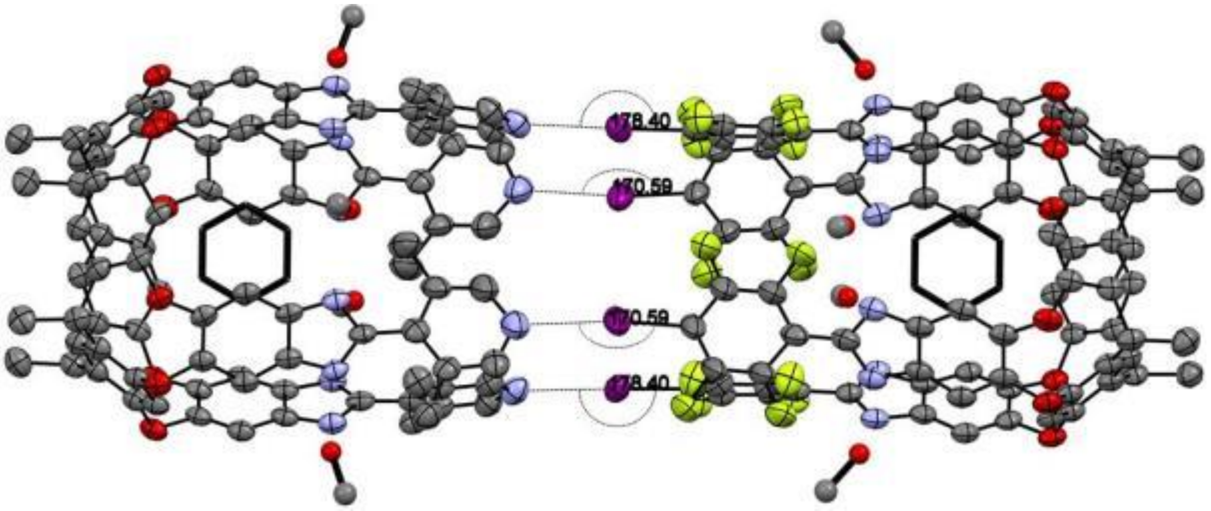
François Diederich of ETH Zurich was the first speaker. His team pursues a multidimensional approach toward deciphering and quantifying weak intermolecular interactions in chemical and biological systems. Experimental study in this research involves the investigation of protein-ligand interactions, synthetic host-guest complexation, and dynamic processes in designed unimolecular model systems, such as molecular torsion balances. It is complemented by computational analysis and exhaustive database mining in the Cambridge Structural Database and the Protein Data Bank (PDB). The findings from this comprehensive investigation greatly aid structure-based drug design.

The first part of François' talk concerned halogen-bonded and chalcogen-bonded supramolecular capsules. Rigorous geometrical requirements for halogen bonding (XB) have been established by both theory[1,2] and experiment.[3-8] The size of the σ-hole on iodine increases with decreasing hybridization state of the carbon atom of the XB donor. In parallel, the electronegative area decreases from $C(sp^3)$ to $C(sp^2)$, and changes to an electroneutral surface potential for C(sp). XB strength also increases if X is on electron-deficient heterocycles or fluoroarenes.[9] Halogen bonding between (iodoethynyl)benzene donors and quinuclidine in benzene affords Gibbs binding free energies (ΔG, 298 K) between -1.1 and -2.4 kcal/mol. The enthalpic driving force ΔH is compensated by an unfavorable entropic term $T\Delta S$ (due to the geometric constraints for XB).

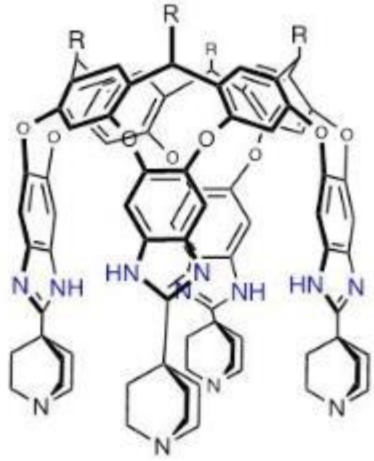
Multidentate halogen bonding involving hydrogen bonding,[10] metal coordination,[11] and ion pair interactions[12] has been reported. More recently, Dumele *et al.* have highlighted the formation of supramolecular capsules based solely on halogen bonding interactions[13] and have published details of their host-guest binding properties in solution. François showed some single XB hemispheres highly preorganized by four alcohol molecules bridging benzimidazole walls by hydrogen bonding:



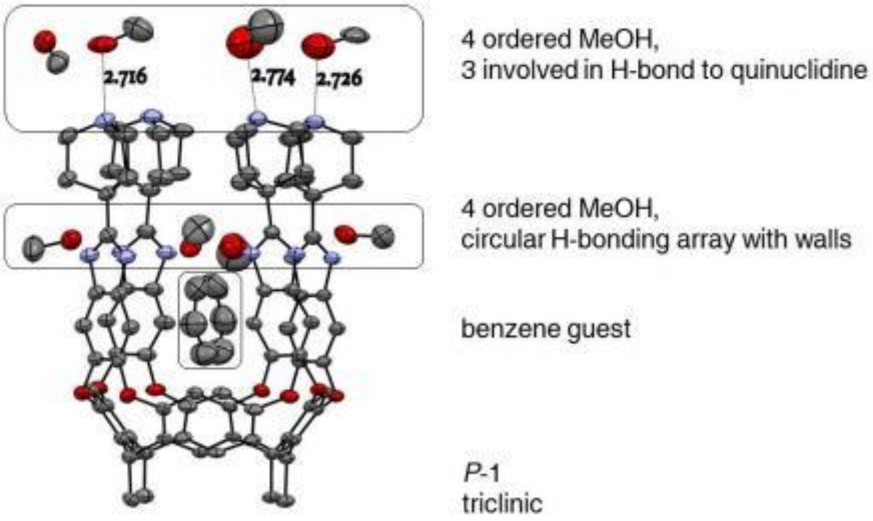
NMR binding data for the F, Cl, Br, and I cavitands as the XB donor showed association constants (K_a) of up to 5370 M^{-1}($\Delta G283$ K=-4.85 kcal mol^{-1}, for I), even in XB-competitive solvent, such as deuterated benzene, acetone, and methanol (70:30:1) at 283 K, where comparable monodentate model systems show no association. The thermodynamic profile showed that capsule formation was enthalpically driven.[13] The geometry of the highly organized capsules is shown by an X-ray crystal structure[14] which features the assembly of two XB hemispheres, geometrically rigidified by H-bonding to eight methanol molecules, and encapsulation of two benzene guests.



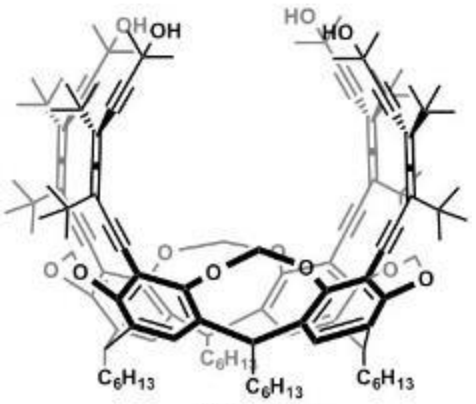
To enhance capsular association strength, tuning the XB donor is more efficient than tuning the XB acceptor, due to desolvation penalties in protic solvents, as shown for a tetraquinuclidine XB acceptor hemisphere:



Desolvation of strong XB acceptors leads to “slow exchange” on the NMR timescale. The X-ray crystal structure in this case reveals a 10-component assembly:

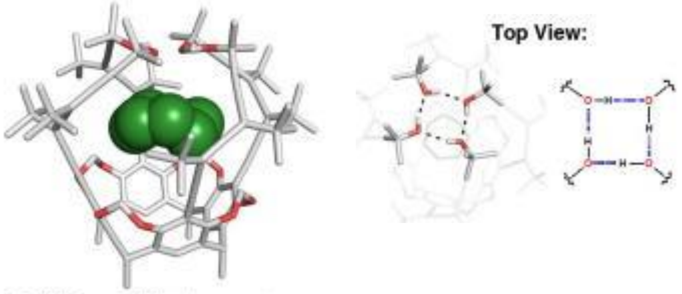

2.716
2.774
2.726
4 ordered MeOH,
3 involved in H-bond to quinuclidine
4 ordered MeOH,
circular H-bonding array with walls
benzene guest
P-1
triclinic

The second part of François’ talk concerned cycloalkane and cycloalkanol binding in an enantiopure cage compound. 1,3-Diethynylallenes (DEAs) are chiral building blocks with exceptional chiroptical properties.[15,16] In work by Gropp *et al.*,[17] four enantiopure DEAs with OH termini were attached to the rim of a resorcin[4]arene cavitand. Alleno-acetylenic cage (AAC) receptors are soluble in polar and apolar solvents, thermally and optically stable, and easy to synthesize. The lean, all-carbon backbones of the four alleno-acetylene walls shape a sizeable cavity:



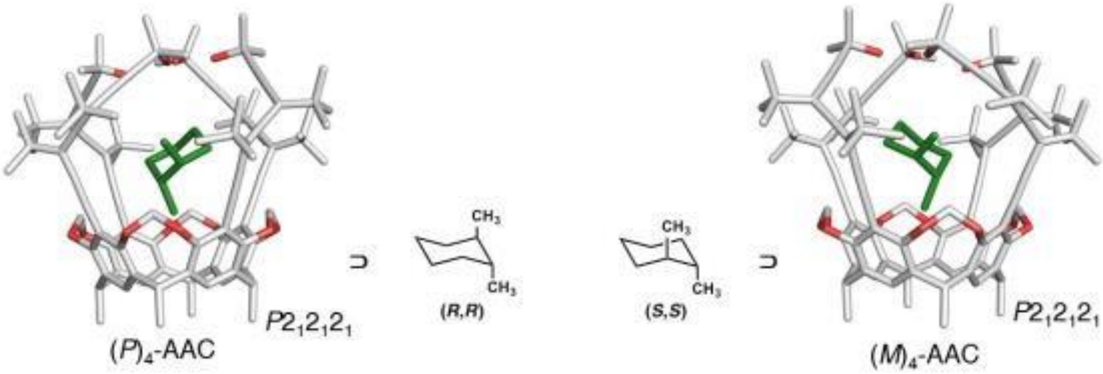
$(P)_4$-AAC

The system undergoes conformational switching between a cage form, closed by a circular H-bonding array:



$(P)_4$-AAC ⊂ cycloheptane $P2_12_12_1$

and an open form, with the tertiary alcohol groups reaching outwards. The cage form is predominant in apolar solvents, and the open conformation in small, polar solvents. Complete chiral resolution of (±)-*trans*-1,2-dimethylcyclohexane was found in the X-ray structures, with (P)4-AAC exclusively bound to the (*R*,*R*)-guest and (M)4-AAC to the (*S*,*S*)-guest.



The directionality of the circular H-bonding array is imposed by the absolute configuration of the alleno-acetylenic arms.

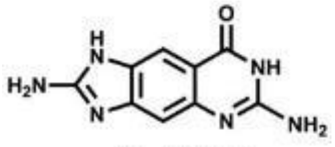
The crystals were grown in a protocol akin to that of Inokuma *et al.*[18] that does not require the crystallization of the sample: crystals of porous complexes are soaked in a solution of the target, such that the complexes can absorb the target molecules, and crystallographic analysis detects the absorbed guest structure along with the host framework. Occupancy was usually 100%. (*R,R*)- and (*S,S*)-*trans*-1,2-dimethylcyclohexane bind in the diaxial conformation with a remarkably small torsion angle. The diaxial conformation is the higher energy conformation.[19]

The dihedral angles $\vartheta_{a,a}$ (X-C(1)-C(2)-X/H) of the axial and diaxial conformers deviate substantially from 180°, down to 144°, accompanied by strong flattening of the ring dihedral angles. Theoretical calculations optimizing the structure of the isolated guest molecules[20] demonstrate that the noncovalent interactions with the host hardly affect the dihedral angles, validating that the host is an ideal means to study the elusive axial/diaxial conformers. Moving from *trans*-1,2-dimethylcyclohexane to *trans*-1,2-dihalocyclohexane results in enhanced diaxial binding since 1,3-interactions are less important. This work[20] also showed that (±)-*trans*-1,2-dihalocyclohexanes (X = Cl, Br) engage in significant halogen bonding interactions C-X···||| (acetylene) with the hosts. X-ray co-crystal structures of AACs further allowed for a detailed investigation, both experimental and theoretical, on the interplay between space occupancy, guest conformation, and chiral recognition based purely on dispersion forces, and weak C-X···π (X = Cl, Br, I) and C-X···||| (acetylene) contacts (X = Cl, Br). A review[21] has been published recently.

The final part of François' talk concerned water solvation of new carbohydrate-conjugated ligands at the active site of tRNA-guanine transglycosylase (TGT), which was investigated in collaboration with Prof. Gerhard Klebe at the University of Marburg. The intestinal disease shigellosis caused by *Shigella* bacteria affects over 120 million people annually. There is an urgent demand for new drugs as resistance against common antibiotics emerges. Bacterial tRNA-guanine transglycosylase (TGT) is a druggable target and controls the pathogenicity of *Shigella flexneri*. TGT recognizes tRNA only as a homodimer[22,23] and performs full nucleobase exchange at the wobble position G_{34}. A posttranslational tRNA modification from guanine to queuine (Q) is introduced in the wobble position in the anticodon of tRNA of all organisms (except yeast and archaebacteria), coding for the amino acids Asn, Asp, His, and Tyr. Eukaryotes need Q as nutrient but procaryote TGT introduces 7-aminomethyl-7-deazaguanine (PreQ1) instead of Q. If TGT is blocked, bacteria become apathogenic since the translation of the key virulence factor virF is blocked.[24-26]

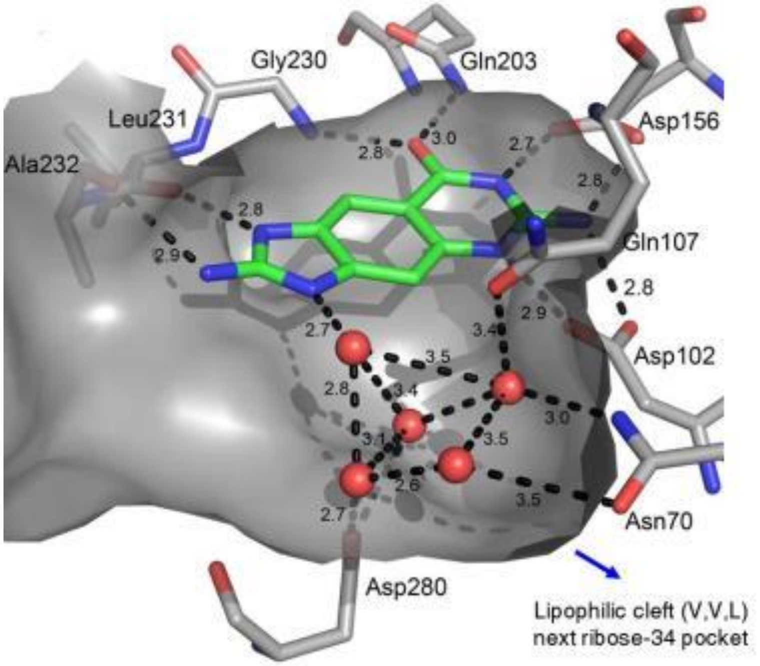
Prokaryotic TGT catalyzes replacement of guanine by PreQ1 at the wobble position of four specific tRNAs. The crystal structure of an intermediate has unexpectedly revealed that RNA is tethered to TGT through the side chain of Asp280. Thus Asp280, instead of the previously proposed Asp102, acts as the nucleophile for the reaction.[23] The crystal structure of the active site in *Zymomonas mobilis* TGT bound to tRNA after base exchange (PDB code 1Q2S) has been published.[23]

lin-Benzoguanines are strong binders at the base exchange site,[27] for example:

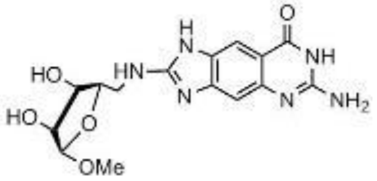


K_i 77 nM

A well conserved five-water cluster is located in the ribose 34 pocket and solvates Asn70 and Asp280. Free energy calculations[28] show that only two of these water molecules can be replaced without penalty. François showed an X-ray structure, PDB code 2Z7K:

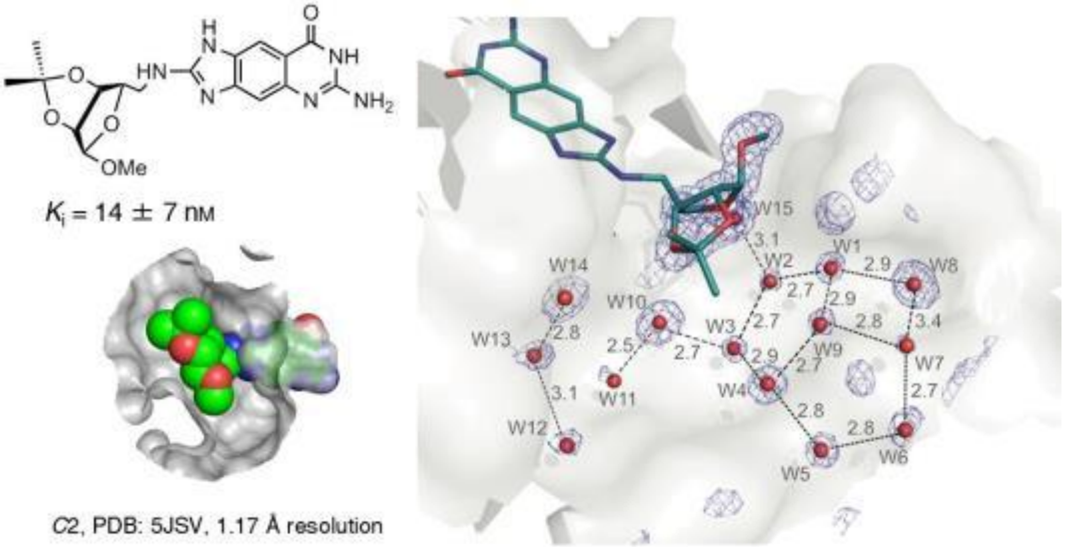


Movsisyan *et al.*[29] have reported the synthesis of sugar-functionalized *lin*-benzoguanines addressing the ribose-33 pocket of TGT from *Zymomonas mobilis*. Ligand binding was analyzed by isothermal titration calorimetry (ITC) and X-ray crystallography. Pocket occupancy was optimized by variation of size and protective groups of the sugars. In the following compound:



$K_i = 1120 \pm 110$ nM

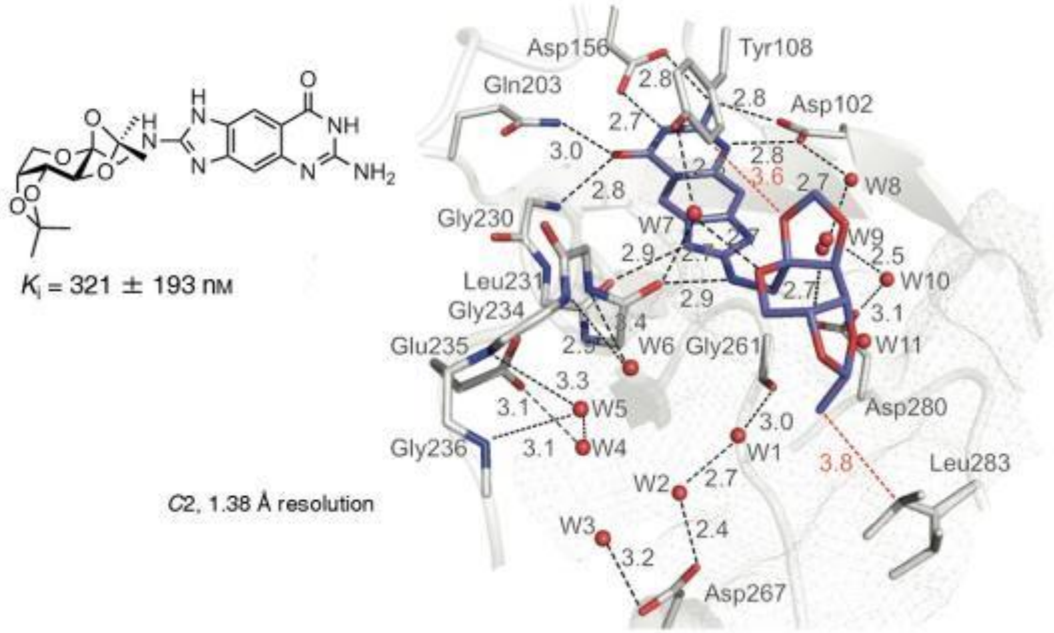
the ribose moiety is too small to fill the ribose-33 pocket properly and adopts multiple conformations. This is in agreement with ITC: a lower gain in ΔH^o is compensated by reduced entropic loss TΔS^o. In another case (below), a [5 + 5 + 4] tricyclic water cluster is clearly solved. Together with six other conserved waters W10 to W15, the protein in the complex is efficiently solvated. There are no direct H-bonding contacts from the water cluster to the ligand.



K_i = 14 ± 7 nM

C2, PDB: 5JSV, 1.17 Å resolution

In the case of a larger fructose bisacetonide (below), the tricyclic water cluster is completely absent. Its formation is prevented by the terminal fructose acetonide. Leu283 moves toward the ligand and establishes a dispersive contact (d(C···C) = 3.8 Å). The ligand is too large and Tyr108 also moves.



K_i = 321 ± 193 nM

C2, 1.38 Å resolution

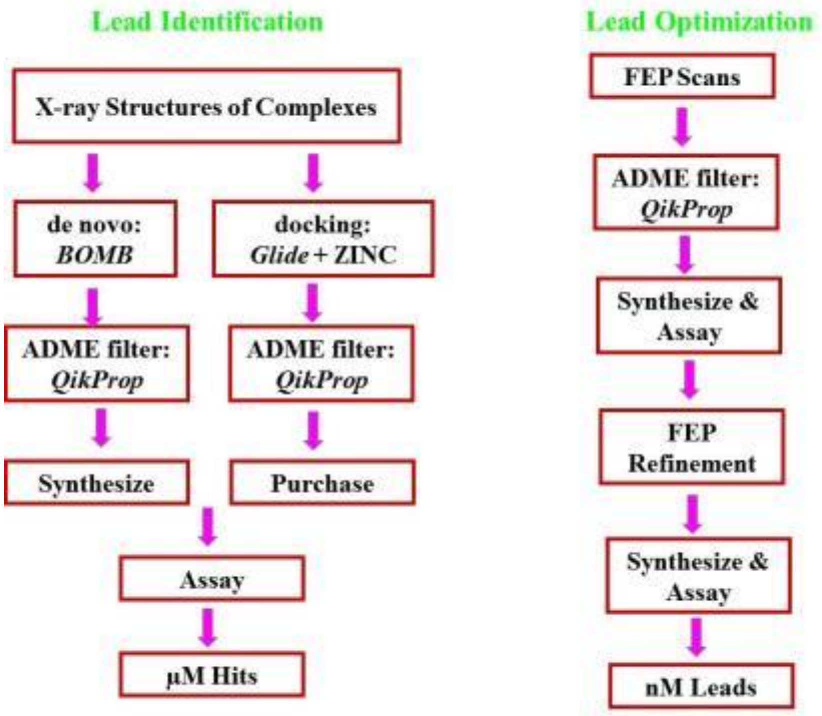
The ordered [5+5+4] water cluster in other complexes is formed between the sugar and protein hydrophobic surfaces to complete the optimal space filling of the 33-pocket. The situation is dramatically different for the complex above, where the solvent accessible surfaces of ligand and protein merge at the place where the tricyclic cluster is found in the other complexes.

Acetonide-protected ribo and psicofuranosyl derivatives are highly potent, benefiting from structural rigidity, good solubility and metabolic stability. The authors concluded that sugar acetonides have a significant, but not yet broadly recognized, value in drug development. François' team continues to work[30] on the sugar-type inhibitors of the TGT homodimer.

Gisbert had asked all the speakers to conclude with a statement about the challenges they faced. François said that his own desire was to learn about the energetics of the ordered water clusters solvating protein-ligand complexes. How much do such stable clusters contribute to the overall Gibbs free energy? What effect has the replacement of individual water molecules in such clusters by a polar ligand substituent? What are the thermodynamic quantities (ΔH and $T\Delta S$) for these water cluster formations? Is desolvation of low log*D* substrates (slow k_{on}) a general principle to achieve higher residency half-lives on target (slow k_{off}) of drugs?

Computer-aided discovery of enzyme inhibitors

William ("Bill") Jorgensen of Yale University started by describing the basics of binding. Stronger binding of a ligand to an enzyme leads to greater potency of the ligand as a potential drug. The Gibbs free energy of binding, $\Delta G_b = -RT \ln K_a = RT \ln K_d$. A K_d of 10^{-9} M arises from $\Delta G_b = -12.4$ kcal/mol, and the molecule is referred to as a "1-nM binder or inhibitor". A K_d of 10^{-9} M arises from $\Delta G_b = -12.4$ kcal/mol, and the molecule is referred to as a "1 nM binder or inhibitor". A K_d of 10^{-6} M arises from $\Delta G_b = -8.3$ kcal/mol, and the molecule is referred to as a1 µM binder or inhibitor". Screening hits are very rarely better than low-µM.[31] Lead optimization is needed to evolve them to low nM, to impart druglike properties, and to avoid toxicity and other liabilities. Lead identification and optimization can be automated:

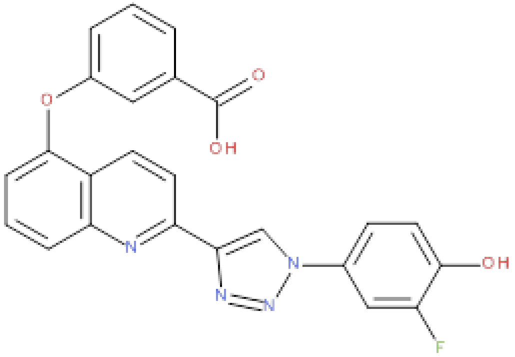


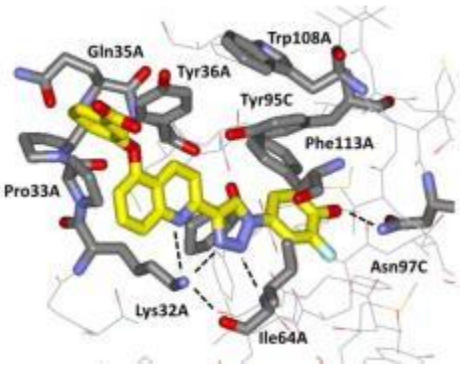
Bill has both computation and synthesis groups at Yale and he has collaborators at Yale who can carry out assays. X-ray crystallography is also done in-house. Underlying all the modeling is the representation of the inter and intramolecular energetics, with Optimized Potentials for Liquid Simulations-All-Atom *(*OPLS-AA) force fields, and the software used is Biochemical and Organic Simulation System (BOSS), Biochemical and Organic Model Builder (BOMB), and MCPRO. Derived from BOSS, MCPRO[32] performs Monte Carlo statistical mechanics simulations of peptides, proteins, and nucleic acids in solution. For *de novo* lead generation, the BOMB program builds combinatorial libraries in a protein binding site using a selected core and substituents. The OPLS-AA force field[33] is a well-proven model which has been continually tested and improved.[34-38] Bill's group offers a web-based service, LigParGen, that provides force field parameters for organic molecules or ligands.

Bill presented some applications of his group's software, centered on the design of inhibitors targeting macrophage migration inhibitory factor (MIF), and human immunodeficiency virus HIV-1 reverse transcriptase (HIV-1 RT). MIF is a cytokine released by T-cells and macrophages. It is also a keto-enol tautomerase. The MIF signaling cascade is initiated by binding to its receptor, CD74. MIF is a target for development of anti-inflammatory[39] and anti-cancer agents.[40] Strategies including *in silico* modeling, virtual screening, high-throughput screening, and screening of anti-inflammatory natural products have led to a large and diverse catalog of MIF inhibitors, as well as some understanding of structure-activity relationships.[40] MIF is also now of interest in neurological disorders.[41]

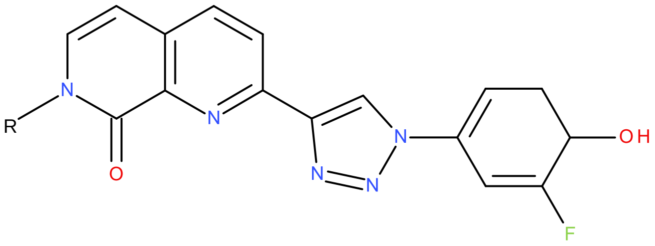
Bill's group prefers to measure K_i, the enzyme-inhibitor binding constant, rather than IC_{50}, which varies with substrate concentration and $K_{m.}$ K_i is determined by a *p*-hydroxyphenylpyruvate (HPP) tautomerase assay. In *de novo d*esign BOMB grows analogues from a core. It adds 1-4 substituents in five possible topologies; generates all conformations in the binding site; optimizes each with the host partly flexible;

and scores and outputs the best as a PDB structure or Z-matrix. Bill's team has reported on design, synthesis and protein crystallography of biaryltriazoles[42] as inhibitors of MIF. Compounds with K_i from 37 to 0.65 µM were synthesized, and a crystal structure 4WRB was obtained showing a binding site with a possible cation-π interaction with Lys32. Additional BOMB modeling encouraged pursuit of 5- and 8-phenoxyquinolinyl analogues (C8 example shown below). Activity was further enhanced by addition of a fluorine atom. The compound below had K_i = 0.014 µM (optimized from a C5 derivative with K_i = 3.0 µM). This compound is much more potent than others in the literature. Bill showed the related PDB structure 5HVS.





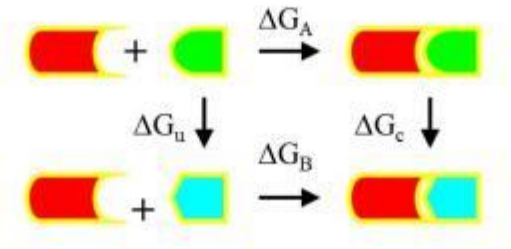
Development and use of fluorescence polarization assays using a fluorescein-labeled tracer provided direct binding data[43] and yielded K_ds in agreement with K_i. (This assay has also been used in Janus kinase 2 (JAK2 kinase)[44] studies.) Striking combinations of protein-ligand hydrogen bonding, aryl-aryl, and cation-π interactions are responsible for the high affinities. A new chemical series was then designed using this knowledge to yield two more strong MIF inhibitors. Coordination of the ammonium group of Lys32 in the active site using a 1,7-naphthyridin-8-one instead of a quinoline was investigated.[45] DFT calculations indicated potential benefits for an added hydrogen bond with the lactam carbonyl group, while free energy perturbation (FEP) results were neutral. Consistent with the FEP results, the naphthyridinones were found to have similar potency to the related quinolines in spite of the additional protein-ligand hydrogen bond.



R = CH_3; K_i = 0.075 µM   R = CH_2COOH; K_i = 0.090 µM

Phenols often show low bioavailability owing to glucuronidation and sulfation, although 7% of all approved oral drugs contain a phenol. Most of the potent MIF tautomerase inhibitors incorporated a phenol, which hydrogen-bonds to Asn97 in the active site. Results of structure-based and computer-aided design, starting from a 113 µM docking hit, have provided substituted pyrazoles as phenol alternatives[46] with potencies of 60-70 nM. Crystal structures (6CBG, 6CBH) of complexes of MIF with the pyrazoles highlight the contributions of hydrogen bonding with Lys32 and Asn97, and aryl-aryl interactions with Tyr36, Tyr95, and Phe113 to the binding. A systematic study of aqueous solubility[47] has been carried out. Currently, 14 of the Yale compounds are being tested by Charles River Laboratories for their ability to inhibit MIF-induced proliferation of glioma cells.[48]

Bill next turned to the topic of free energy perturbation (FEP)[49] which he has worked on since 1980. Since 2005 he has applied it to lead optimization. The thermodynamic cycle for relative free energies of binding is as follows.



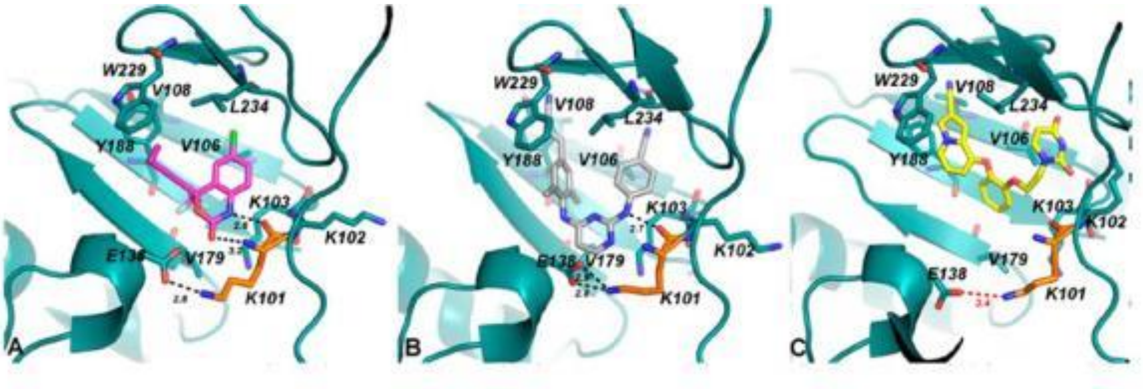
$$\Delta\Delta G_{binding} = \Delta G_B - \Delta G_A = \Delta G_c - \Delta G_u$$

FEP calculations are performed for the protein-ligand complex in water. Extensive Monte Carlo sampling is carried out for the protein, ligand, and water at 25 ^{o}C. The results identify modifications that should increase binding affinity. The software used is MCPRO.[32] During the 1980s, advances in the ability to perform computer simulations, and to calculate free energy changes, led to the expectation that such methodologies would soon show great utility for guiding molecular design. The 1980s also saw the rise of high-throughput screening and combinatorial chemistry, along with complementary computational methods for *de novo* design and virtual screening including docking. All these technologies appeared poised to deliver leads for any target, but realization of the expectations required significant additional effort and time. It was important to test if FEP could deliver in a prospective manner. Striking success has now been achieved for computer-aided drug lead generation and optimization.[50] FEP for drug lead optimization[51] is now accepted, with room for improvements: exciting challenges remain.

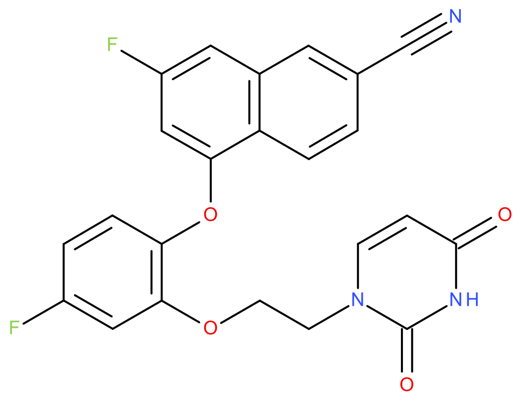
One application is optimizing the substituents on the phenol ring[42,43] of the MIF inhibitors described above, where the crystal structures confirm the position of F contacting Met101. Also a “heterocycle scan” was performed using FEP with Desmond[51] OPLS2.1. Experimentally, nothing has been found better than 1,2,3-triazole. A tetrazole and a thiadiazole are 1-5 µM inhibitors.

Most of the Yale FEP work has been on HIV-1 RT. Bill has reviewed[52] 10 years of efforts in his laboratory to discover anti-HIV agents. The work has focused on computer-aided design and synthesis of non-nucleoside inhibitors of HIV-1 reverse transcriptase (NNRTIs), with collaborative efforts on biological assaying and protein crystallography. MT-2 assays were carried out in Karen Anderson's laboratory. X-ray crystallography was carried out in Eddy Arnold's laboratory.[53] Monte Carlo and FEP simulations have often been executed to identify the most promising choices for heterocycles, linking groups, and substituents on rings.[54,55] Numerous design issues were successfully addressed including the need for potency against a wide range of viral variants, good aqueous solubility, and avoidance of electrophilic substructures. Computational methods including docking, *de novo* design, and FEP calculations made essential contributions. The result is novel NNRTIs with picomolar and low-nanomolar activities against wild-type HIV-1, and key variants that also show much improved solubility and lower cytotoxicity than recently approved drugs in the class.

Members of the catechol diether class are highly potent NNRTIs. Many NNRTIs, such as rilpivirine, bear a cyanovinylphenyl (CVP) group, an uncommon substructure in drugs that gives reactivity concerns. Bill's team used computer simulations to design bicyclic replacements[56] for the CVP group. FEP results for 18 alternatives were obtained and 30 new compounds were synthesized, including three with EC_{50} activity values of 10nM, 7nM and 0.38 nM. The team has also reported potent inhibitors active against HIV-1 RT with K101P, a mutation conferring rilpivirine resistance.[57] The results for K101P are explained by crystal structures: the Yale compounds do not make an H-bond with a K101 carbonyl.



The clinical benefits of NNRTIs are hindered by their unsatisfactory pharmacokinetic (PK) properties, and the rapid development of drug-resistant variants. Bill's team used computational and structure-guided design to develop two next-generation NNRTI catechol diether drug candidates, and published PK and humanized mouse studies.[58,59] One candidate is JLJ636:



Resistance associated with the Tyr181Cys mutation in HIV-1 RT has been a key roadblock in the discovery of NNRTIs. The team at Yale has reported covalent inhibitors of Tyr181Cys RT that can completely knock out activity of the resistant mutant and of the particularly challenging Lys103Asn/Tyr181Cys variant.[59] Conclusive evidence for the covalent modification of Cys181 was provided from enzyme inhibition kinetics, mass spectrometry, protein crystallography, and antiviral activity in infected human T-cell assays.

Massive computational docking experiments to identify noble gases target for new "atomic drugs"

Dave Winkler of the Monash Institute of Pharmaceutical Sciences and La Trobe University, Australia presented research carried out in collaboration with colleagues at the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia, and Air Liquide Santé in France. All chemists know that the noble gases are chemically inert, except to the most extreme reaction conditions. Paradoxically, several of these gases exhibit a wide range of fascinating biological effects, some of which have been demonstrated *in vivo* and in the clinic. Their most important physical properties are lipophilicity, size, polarizability, and arguably, their rarity. As they are such simple entities, modeling their properties and biological interactions is considerably more tractable than for small drugs where conformation, hydrogen bonding, charge, π interactions etc. are often the dominant interactions.

Dave and his colleagues have conducted an exhaustive review[60] of the literature on medical and pharmacological properties of noble gases. About 400 papers were studied, most of them published since 1980. The review was necessary for the authors to understand current experimentally verified interactions with proteins at atomic level, and to infer likely pharmacological effects of noble gases binding to human proteins from computational screening.

There is a strong correlation of anesthesia and other biological effects with solubility. Many potentially clinically useful properties are known or may emerge. It is impractical to carry out experiments to assess the effect of noble gases on every possible protein target; simulation predicts xenon binding sites and is the only feasible way to explore noble gas binding. Novel delivery systems are required to reduce cost and allow lighter noble gases to be used clinically.

Noble gases have many biochemical effects at the atomic level. Xenon affects a range of receptors involved in cell signaling; the N-methyl-D-aspartate receptor (NMDA receptor) is the receptor best studied experimentally. Two of the other molecular targets for xenon that have been identified[61] include the two-pore-domain potassium channel TREK-1, and the adenosine triphosphate-sensitive potassium channel (K(ATP)), but which of these targets are relevant to acute xenon neuroprotection is not clear.

Pharmacological effects of noble gases on cells, organs, and organisms include anti-apoptotic effects, cytoprotection, neuroprotection, analgesia, anticonvulsant effects, anesthesia, and effects on memory and addiction. Xenon is an ideal anesthetic. It is 1.5 times more potent than nitrous oxide; offers rapid induction and emergence; has low toxicity, and is devoid of teratogenicity. It protects neural cells from ischemic injury, and has many other clinical advantages. Its unique combination of analgesia, hypnosis, and lack of hemodynamic depression makes it a very attractive choice for patients.

Interestingly, NMDA receptors have recently been shown to play a major role in alcohol craving and relapse, and other addictions. A study of the effect of xenon on alcohol-seeking and relapse-like drinking behavior in rats[62] showed that even a brief exposure to xenon can induce an anti-reward effect lasting several days. Xenon may also usefully interfere with craving in human alcoholics, and potentially in other addictive settings.

Dave's team aimed to discover new, useful therapeutic applications for noble gases. Given the simplicity of the ligands, and the steady improvement of computational docking algorithms, it is feasible to conduct computational studies on a large number of proteins for their likely affinity for the noble gases. Dave hypothesized[60] that the results of simulations can identify which proteins bind the gases most tightly, and where the gases bind relative to ligand or cofactor binding sites, thus allowing proteins to be prioritized in terms of potential medical interest. Interpreting and visualizing data allows inferences of pharmacological pathways affected by noble gases. Confirmatory experiments will be run on the top candidates. Efficient delivery of noble gases by nanoparticles, to overcome issues with cost (for xenon), and the need for hyperbaric administration (for noble gases other than xenon and perhaps krypton), would greatly increase the added value.

The aims of the *in silico* screen were:

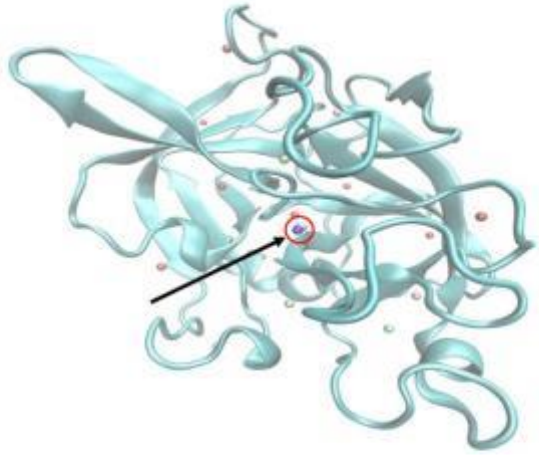
- to identify whether the xenon binding is in or very near the endogenous ligand binding pocket and could block ligand binding
- to quantify how tightly xenon binds, and how many xenon atoms the site will accommodate
- to identify sites with very good xenon binding where the binding may also be favorable for the smaller noble gases
- to identify the structural rules for noble gas binding sites
- to carry out a detailed analysis of interesting results using molecular dynamics (MD)
- to validate the predictions in laboratory experiments

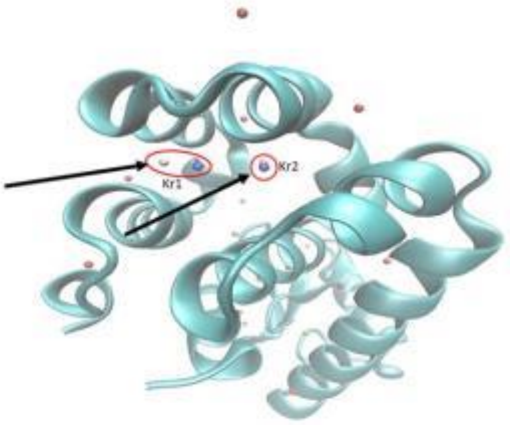
- to lead into experiments on efficient drug delivery.

An essential requirement for such a massive computational study is to validate the methods. If the methods can reliably locate the known experimental binding positions of noble gases in diverse proteins, a much larger computational study is more likely to yield useful new knowledge. To this end, a diverse subset of about 60 proteins in the PDB containing at least one xenon atom and having <50% sequence identity was analyzed to study how xenon binds to the protein, and how close xenon approaches site residues. In most of the cases, the closest approach is 3.5 – 4.5 Å. Activity originates from one of two mechanisms: xenon directly binds to the binding site, or xenon binds elsewhere causing an expansion in the cavity volume, which then presumably modifies the active site conformations or flexibility.

For docking, Dave used the AutoDock 4 package from the Scripps Institute. Liu *et al.* had previously found good agreement[63] between the xenon binding sites and energies for the NMDA receptor predicted by AutoDock and MD calculations. Force fields and energy cut-offs were fine-tuned by identifying all known xenon crystallographic binding sites; Andrew Warden of CSIRO carried out the DFT calculations.

There were 116 protein structures with xenon ligands in the validation set, 12 with krypton, and 4 with argon. Automated protein preparation was essential if the results were to be valid and useful: this was one of the most time-consuming parts of the project. The docking method was validated by quantifying how well simulations could predict binding positions in 132 diverse protein X-ray structures containing 399 xenon and krypton atoms. Dave and his co-workers found excellent agreement[64] between calculated and experimental binding positions of noble gases: 94% of all crystallographic xenon atoms were within 1 xenon van der Waals (vdW) diameter of a predicted binding site, and 97% lay within 2 vdW diameters; 100 % of crystallographic krypton atoms were within 1 krypton vdW diameter of a predicted binding site. Dave gave some examples, for example, porcine elastase (1C1M) with a single xenon atom in the structure (left) and T4 lysozyme mutant (1C6A) with two krypton atoms in the structure (right):





The PDB contains about 70,000 mammalian protein structures and about 130,000 structures in total. There is considerable redundancy in the PDB. Imposing sequence similarity allows redundancy to be removed, leaving representative structures for each protein. After these redundancy filters are applied the PDB contains 20,000 human protein structures at 100% similarity. Systematically mapping the energy of the noble gases at all grid locations (in 0.375 Å steps) in all 130,000 PDB structures is a formidable computational task. This task has now been completed and the resulting dataset is freely available on the CSIRO website. Dave encourages interested scientists to explore the dataset, perhaps by looking at how various noble gases bind to their favorite protein targets.

The team is now using automated and manual methods to analyze the massive dataset to find proteins with the most promising interactions with noble gases, primarily xenon and krypton. For these, they will explore the pharmacological effects of hits using the published literature; report on protein structure, and function, and known synthetic modulators of the protein's action, and infer pharmacological effects; and design experiments to validate the *in silico* predictions.

Proteins of special interest will be given detailed analysis using higher level modeling techniques: docking and molecular dynamics. Advanced docking can study in more detail the interactions of noble gases with specific sites in proteins, allow the binding partners to be characterized, and allow competition with natural ligands to be modeled. Molecular dynamics[63] can reveal the time course of interactions between noble gases with proteins, the trajectories of gases through proteins, the effect of the binding on protein structure, and so on. Delivery systems for noble gases are also important. Britton et al.[65] have demonstrated that *in vivo* xenon delivery *via* echogenic liposomes is neuroprotective. Microbubbles are another potential method for therapeutic noble gas delivery.

Dave concluded, as requested by Gisbert, with a list of grand challenges for molecular design. We need better mapping of structures into mathematical objects (deep learning); generative models that can predict real molecules with improved properties from models; use of evolutionary algorithms to explore more of the vast chemical space; and generation of autonomous systems.

Progression saturation analysis of analogue series using virtual candidate compounds

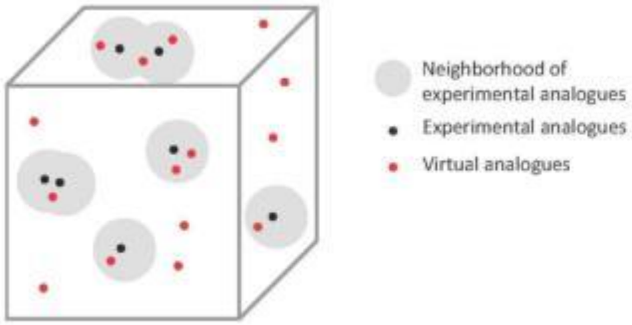
Jürgen Bajorath of the University of Bonn presented new work on lead optimization. In optimization, it is difficult to estimate when an analogue series might be saturated and synthesis of additional compounds would be unlikely to yield further progress. Only a few approaches are currently available to monitor series progression, and aid in decision making. Jürgen presented a new computational method to assess progression saturation of analogue series by comparing existing compounds to virtual candidates. In this method, virtual analogues are generated for an existing analogue series, and virtually extended series are projected into a chemical reference space. Chemical neighborhoods (NBHs) of existing analogues are defined and a dual scoring scheme is applied to quantify the saturation of analogue series focusing on neighborhoods. Different saturation categories are established on the basis of characteristic scores.

The NBH of each experimental analogue is defined on the basis of compound distance relationships in chemical space. NBH radii are adjusted for score calculations. The raw global saturation score quantifies chemical space coverage by existing and virtual analogues. The NBH radius of experimental analogues is set to the median of distances between each virtual analogue and its top 1% nearest virtual neighbors. The raw local saturation score quantifies the distribution of virtual candidates across the NBHs of active analogues. The NBH radius of active analogues is set to the median of pairwise distances between active analogues.

The raw global saturation score, *S*, is given by

$$(S) = \frac{|v_{exptl}|}{|V|}$$

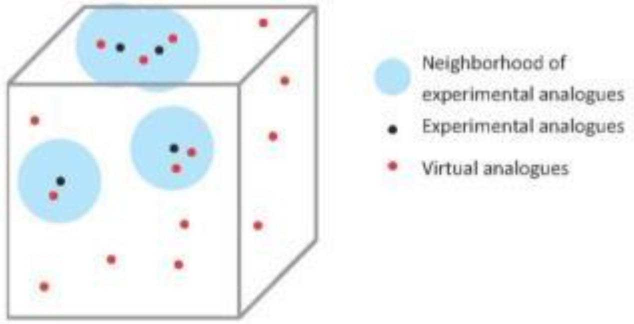
where *V* = set of virtual analogues, and v_{exptl} = set of virtual analogues in NBHs of experimental analogues. In the following example the raw global saturation score is 8/15.



The raw local saturation score, *A*, is given by

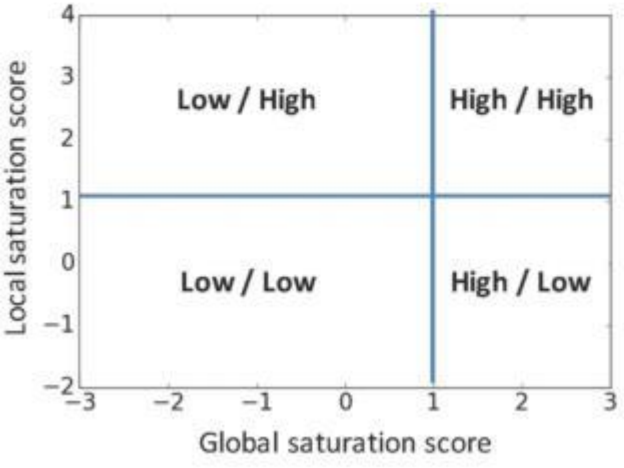
$$(A) = \frac{|A|}{|v_{active}|+1}$$

where *A* = set of active analogues, and v_{active} = set of virtual analogues in NBHs of active analogues. In the following example the raw local saturation score is 4/7.



The global saturation score increases with increasing numbers of virtual candidates falling into NBHs of experimental analogues, indicating extensive global chemical space coverage. The local saturation score increases with decreasing numbers of virtual candidates in NBHs of active analogues (“active NBHs”). For ensembles of series, raw global and local scores are converted into Z-scores.

A dual scoring scheme is obtained by combining global and local saturation scores (global/local). Combined scores define different saturation categories:

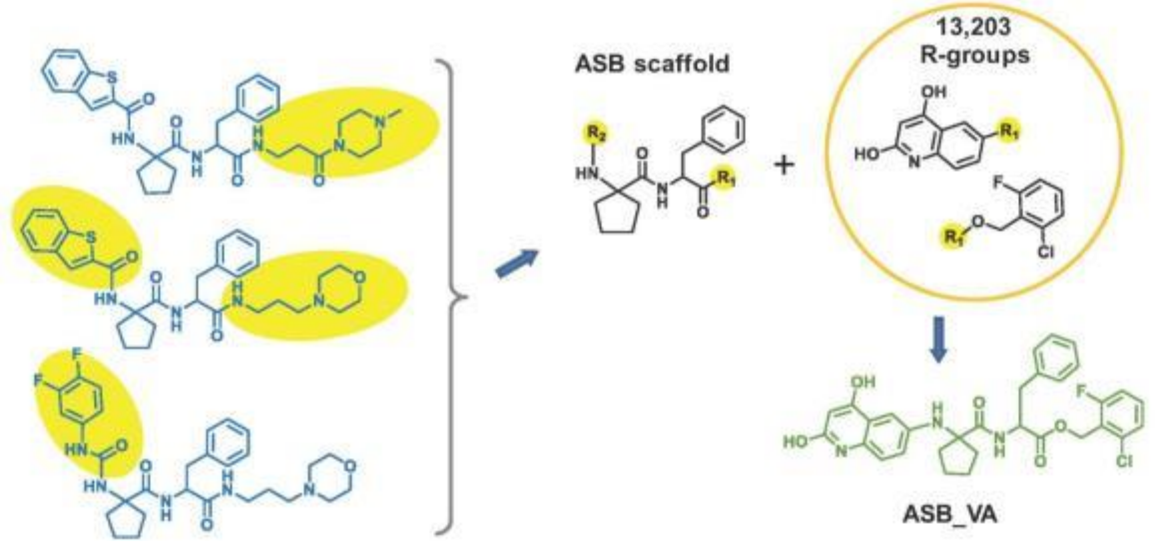


Early-stage series are categorized low/high: there is low chemical space coverage with only a few virtual candidates in active NBHs. Mid-stage series are categorized low/low: there is low chemical space coverage with many virtual candidates in active NBHs. Late-stage series are categorized high/low: there is extensive chemical space coverage with many virtual candidates in active NBHs. Saturated series are categorized high/high: there is extensive chemical space coverage with only a few virtual candidates in active NBHs.

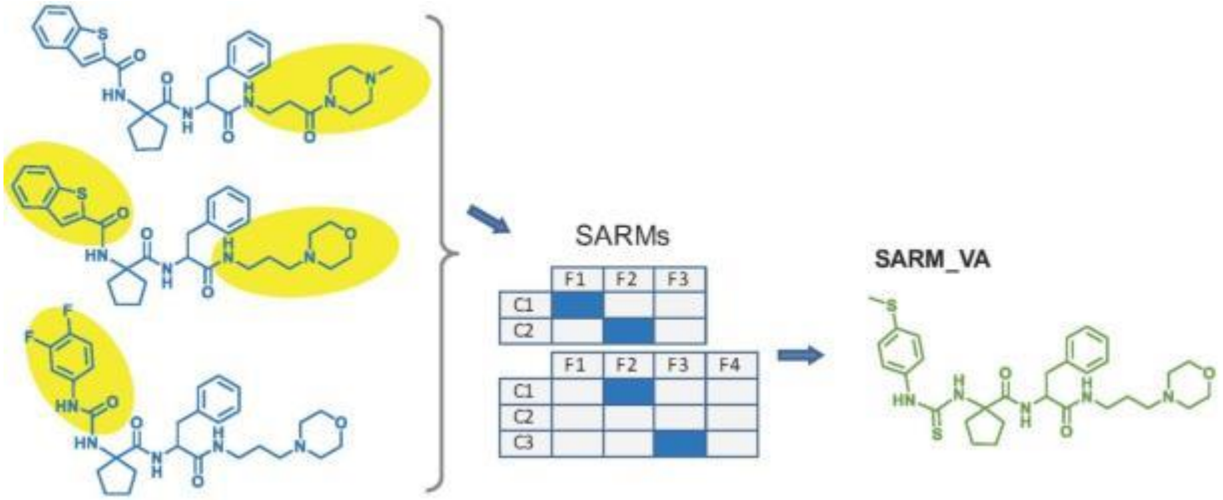
Jürgen carried out a proof of the concept. Using an in-house algorithm, 80 analogue series were extracted from PubChem Bioassays. A total of 1618 compounds (for 23 targets) had single substitution sites, and active and inactive analogues were included. Also, 64 analogue series were extracted from ChEMBL. In this case a total of 1422 compounds (for 62 targets) had multiple substitution sites, and only active analogues were included.

The first descriptor set included seven simple and chemically intuitive numerical descriptors: molecular weight, number of H-bond acceptors, number of H-bond donors, number of rotatable bonds, log*P*, aqueous solubility, and molecular surface area. The second descriptor set included seven “abstract” numerical descriptors (topological and shape indices, complex surface and charge descriptors, and synthetic feasibility index) with little correlation. Set one gave a seven-dimensional reference space; set one plus set two gave a fourteen-dimensional reference space.

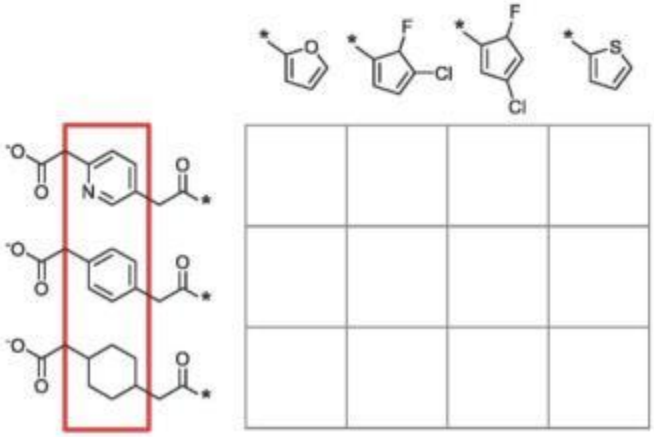
Analogue series-based (ASB) scaffolds extracted from ChEMBL were enumerated with 13,203 unique R-groups to generate “diverse” virtual analogues (ASB_VAs):



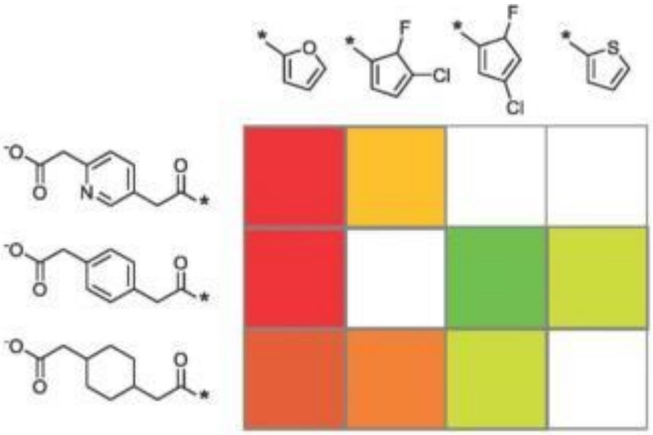
Series-specific design gave variable numbers of virtual analogues from SAR matrices (SARMs)[66], that is, "close-in"' virtual analogues (SARM_VAs):



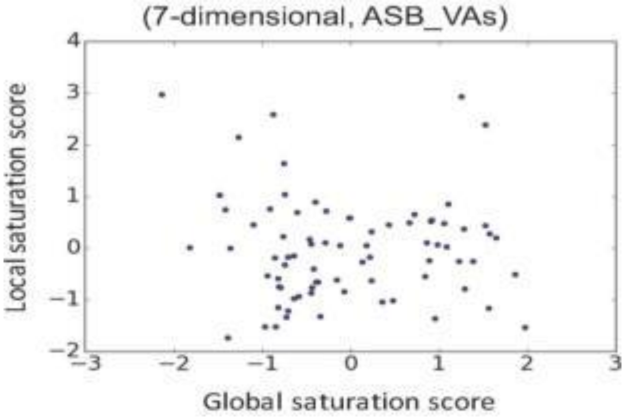
Jürgen briefly described the SARM methodology[66] for structural organization of compound sets. Systematic fragmentation of compounds yields matched molecular pairs (MMPs). In a dual fragmentation scheme, step 1 (standard fragmentation) gives cores and R-groups, and step 2 (core fragmentation) gives core MMPs. A matrix contains a subset of compounds with analogous core structures (that only differ at a single site). By design, SARMs are obtained only from series with multiple substitution sites.



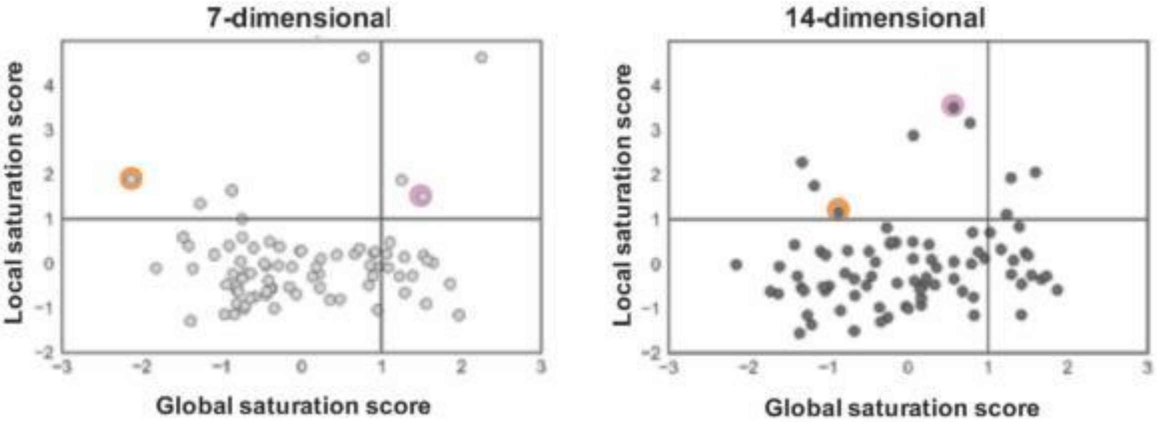
A matrix is filled with available core structures and R-groups (see below). Large series typically yield multiple SARMs. Cells are colored by increasing potency (from red to green below). “Empty” cells represent virtual analogues: unexplored (core plus R-group) combinations. These close-in virtual analogues form a “chemical space envelope” around compound series.



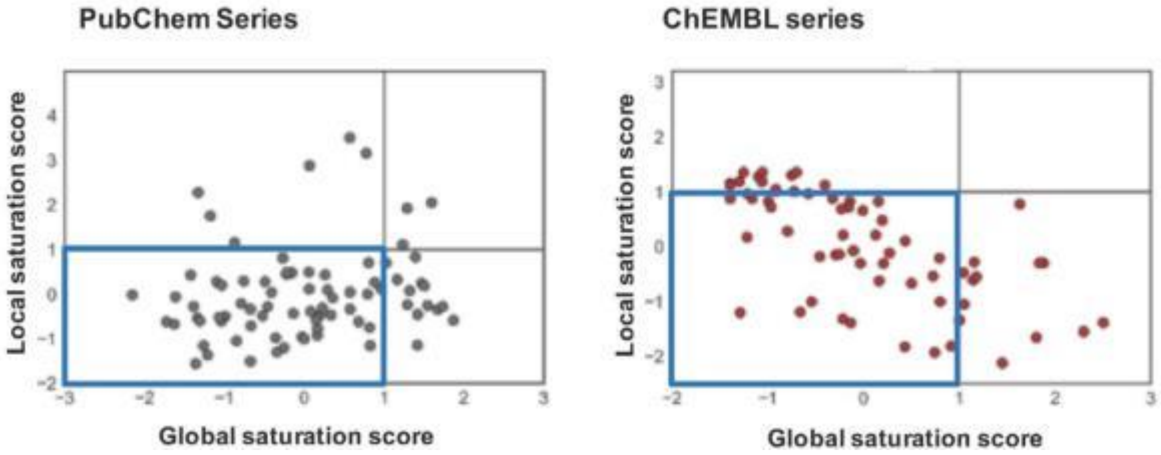
Jürgen presented his results. There was no detectable correlation between global and local saturation scores for the 80 PubChem series:



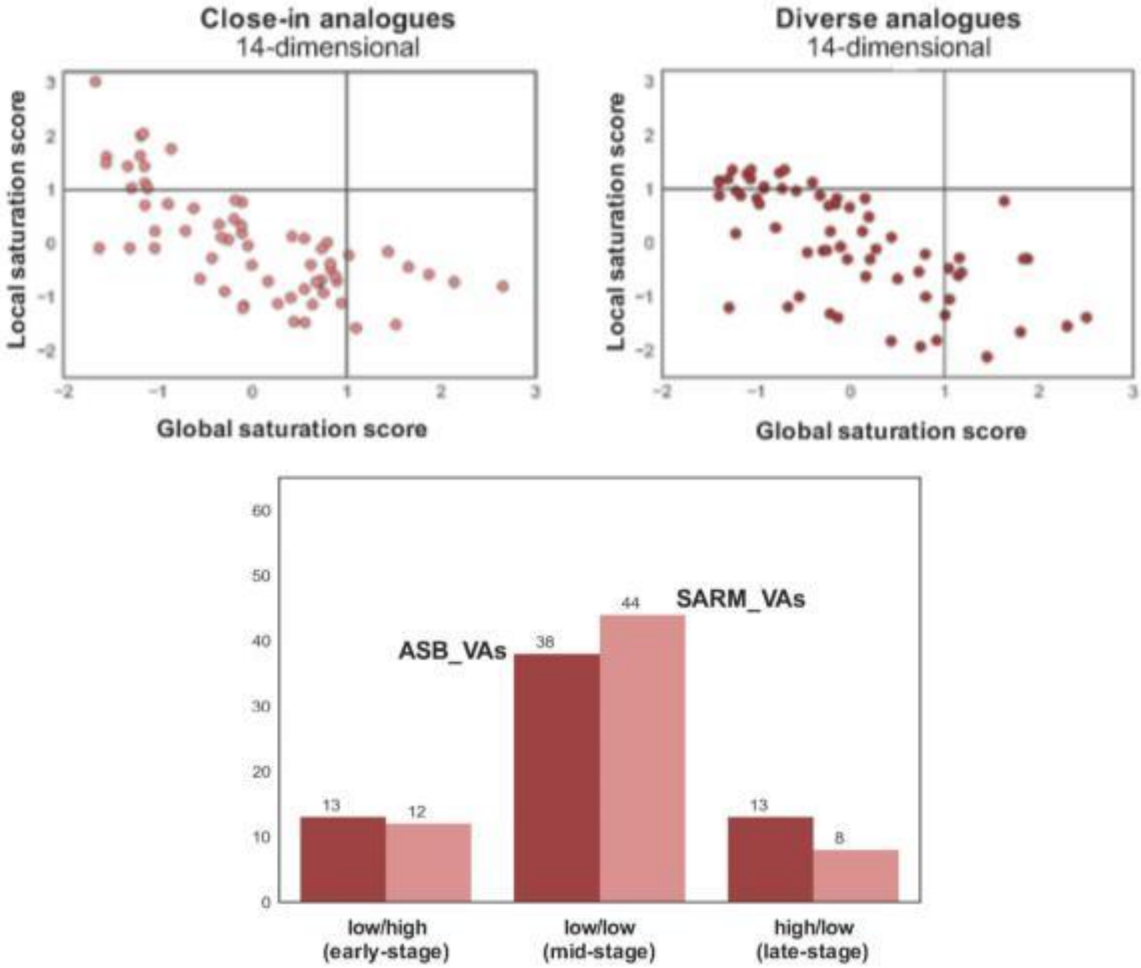
As regards reference space sensitivity, a comparison of 80 PubChem series in 7- and 14-dimensional chemical reference space (ASB_VAs) showed that 63 of 80 series retain their category assignment:



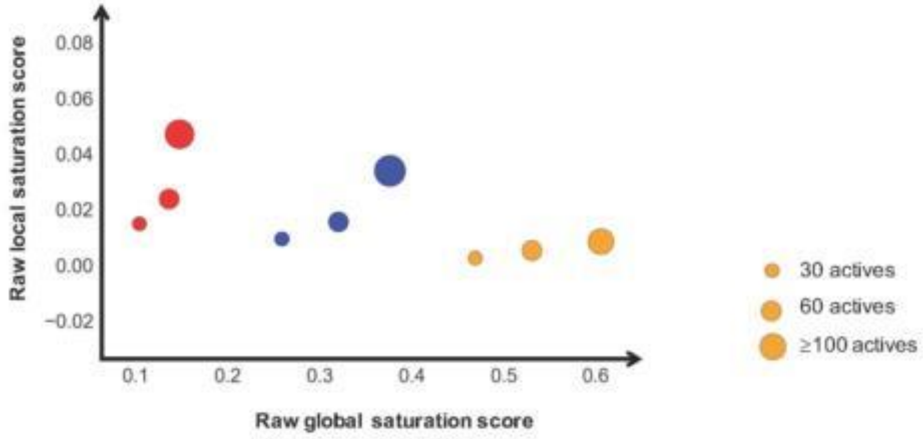
The category distributions of PubChem and ChEMBL series were similar as shown by the 14-dimensional ASB_VAs plot below. Both sets are dominated by series with mid-stage character.



Categorization of series is also only moderately influenced by varying analogue design strategy, although a small shift from late-stage towards mid-stage series is observed for SARM_VAs compared to ASB_VAs, for the ChEMBL series:



Jürgen presented three models for evolving series with more than 100 analogues (evaluated in cumulative increments). There was a consistent increase in saturation scores for subsets of increasing size, and the saturation level increases from the left to the right:



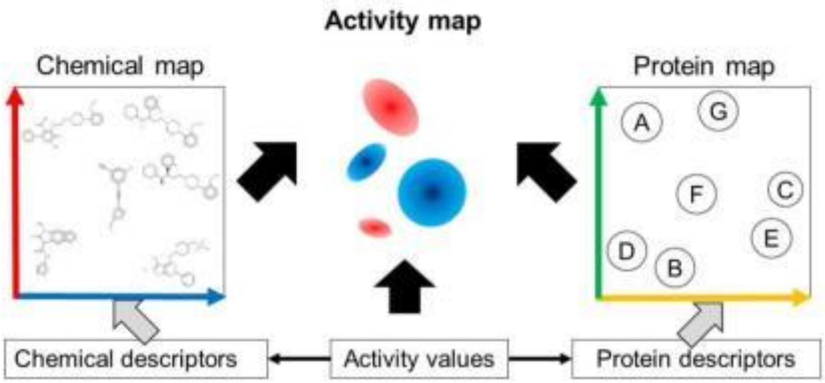
In summary, Jürgen's computational progression analysis of analogue series relies on a neighborhood concept and a virtual analogue-based scoring scheme. Progression states are categorized on the basis of combined global and local saturation scores. For proof-of-concept, large sets of analogue series were profiled. Categorization of series is only moderately influenced by varying space representations and analogue design strategies. Different saturation characteristics are detected for analogue series including the largest publicly available series. In response to Gisbert's question about grand challenges, Jürgen suggested rationalizing compound optimization, and demystifying AI approaches in chemistry beyond the hype.

Novel method proposing chemical structures with a desirable profile of activities based on chemical and protein spaces

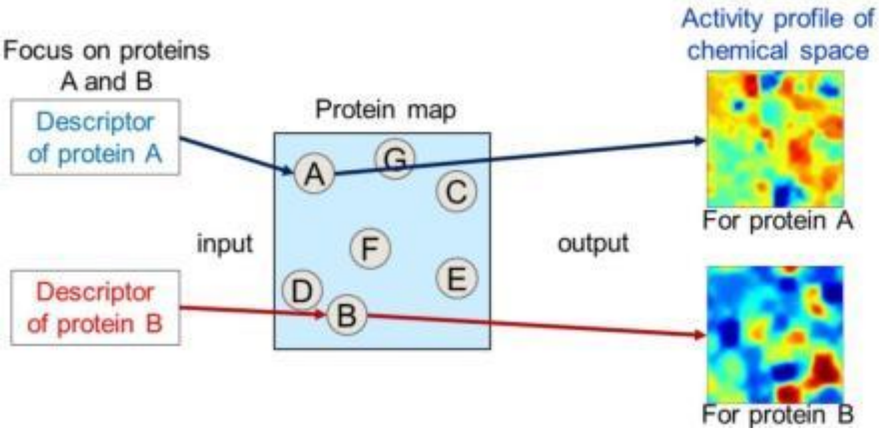
Kimito Funatsu of the University of Tokyo, Japan spoke about a new method for predicting relationships between compounds and proteins. Computational methods to generate models that zoom out from a single protein with a focused ligand set to a larger and more comprehensive description of compound-protein interactions, and that are valid in prospective experiments, are of importance in drug discovery.[67] Nevertheless, constructing a simple prediction model against one protein does not completely help drug design, because detecting chemical structures that act similarly against multiple proteins is necessary for preventing side effects of the potential drug. To tackle this problem Kimito's team proposes a new method that visualizes the chemical and protein spaces.

By visualizing chemical descriptor space in 2D (using a self-organizing map or generative topographic map) it is possible to determine an activity profile and to decide on target areas. By visualization of protein descriptor space, the activity profile of proteins (including an orphan protein) can be determined. Kimito's objectives were simultaneous visualization of chemical space and protein space, considering compound-protein interaction, and development of a structure search method. His strategy was to expand counterpropagation neural networks (CPNN)[68] into a multi-input and single-output system.

CPNN is a supervised learning method combined with a Kohonen neural network. Two maps can be created: a chemical map (a profile of chemical descriptors) and an activity map (a profile of activity values). In multi-input CPNN (MICPNN) the strategy is to create two maps of chemical space and protein space, and to associate activity values to the combination of two coordinates:



The activity profiles of chemical structures for several proteins can then be obtained:

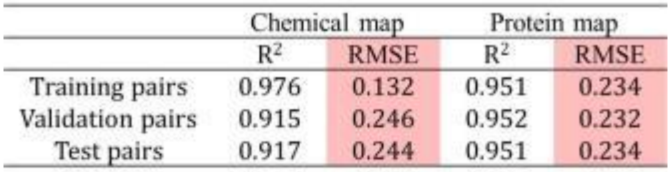


The procedure for drug discovery using MICPNN is to input descriptors of a protein and get the activity profile of compounds; to determine grids from the condition of activity value; and to obtain molecular descriptors corresponding to the grids. These descriptors can then be used in virtual screening, or in *de novo* design, to get new chemical structures.

G protein–coupled receptors (GPCRs), also known as seven-(pass)-transmembrane domain receptors, or 7TM receptors, are drug targets related to many diseases. A ligand binding site is known. In a case study Kimito used the GPCRSARfari dataset in ChEMBL.[69] This has 40,000 activity data points ($pK_i > 4.0$) for 21,272 compounds and 124 proteins in GPCR class A. Kimito used 20,000 as activity pairs for the MICPNN learning set, 10,000 as validation pairs for convergence determination, and 10,000 as test pairs to examine MICPNN.

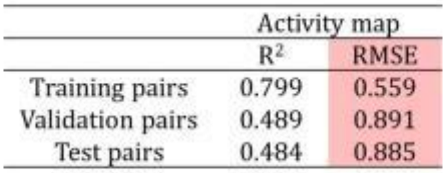
The chemical descriptors were 137 descriptors from Dragon 7.0. The 288 protein descriptors were z-scores[70] of 7TM domain regions obtained from GPCRdb.[71,72] The *z* score is the score vector obtained by applying principal component analysis to the physical property value of the amino acid; *z*3 (up to the third principal component) was used in this case study. Kimito presented these results:

Mapping accuracy:



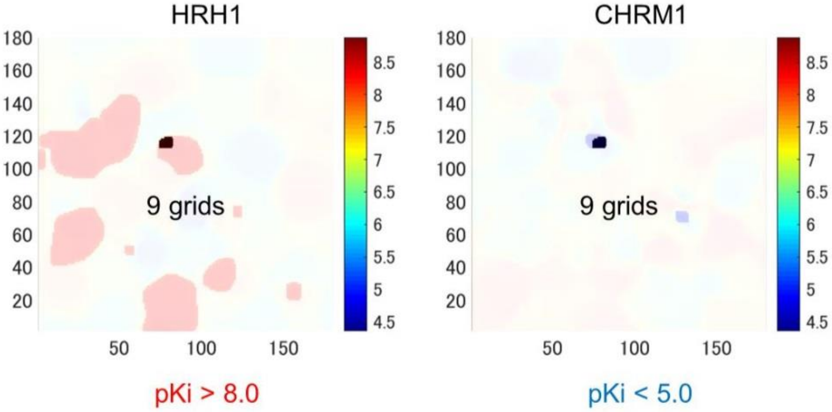
	Chemical map		Protein map	
	R^2	RMSE	R^2	RMSE
Training pairs	0.976	0.132	0.951	0.234
Validation pairs	0.915	0.246	0.952	0.232
Test pairs	0.917	0.244	0.951	0.234

Prediction accuracy:

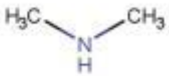
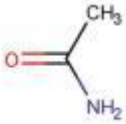


	Activity map	
	R^2	RMSE
Training pairs	0.799	0.559
Validation pairs	0.489	0.891
Test pairs	0.484	0.885

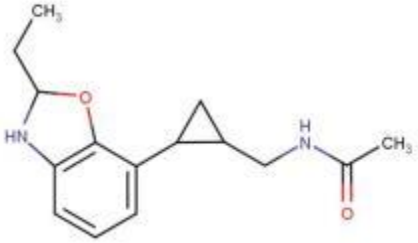
Having proposed MICPNN to visualize chemical space and protein space, and developed a chemical structure search process, Kimito showed that combination with a structure generation method[73] led to more effective chemical structure search. The target protein Kimito chose to illustrate structure search was histamine receptor H1 (HRH1), a class A GPCR related to diseases such as abnormality of airway function and atopic dermatitis.[74] HRH1 antagonists are bronchodilators, and anti-allergy and anti-nausea drugs. The structure of the human histamine H1 receptor complex with doxepin[75] has been published. First generation HRH1 antagonists have affinity for muscarinic acetylcholine receptors (CHRM),[76] and cause anticholinergic effects. So Kimito wanted to find compounds with high activity for HRH1 ($pK_i > 8$) and low activity for CHRM1 ($pK_i < 5$). Using these settings, he obtained limited areas on the Kohonen maps of HRH1 and CHRM1 when these two maps were overlapped:



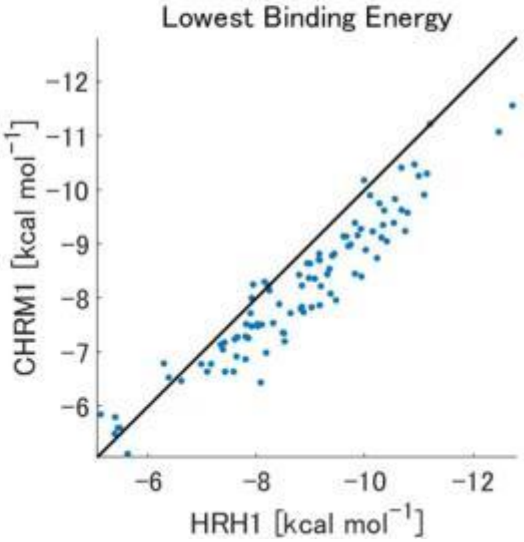
The nine pairs of descriptor values (and number of substructures) were:

Coordinate			…
(76, 114)	1	0	…
(77, 114)	1	0	…
(77, 115)	1	0	…
(78, 114)	1	0	…
(78, 115)	1	1	…
(79, 114)	1	0	…
(79, 115)	1	1	…
(79, 116)	1	1	…
(80, 115)	0	0	…

For testing the ability of the model Funatu used a GPCRSARfari dataset not used in modeling. One selected compound[77] (below) had predicted HRH1 activity of pK_i 8, and predicted CHRM1 activity of pK_i 4.88.



A docking simulation was run for the best 100 compounds, using AutoDock4.2.6,[78] and AutoDockTools-1.5.6,[78,79], and protein structures from the PDB. A plot of the binding energies indicated selectivity for HRH1:



Kimito showed docking simulations for ChEMBL343024, where the binding energy for HRH1 was -8.08 kcal mol^{-1}, and for CHRM1 was -6.44 kcal mol^{-1}.

It is a problem that interpretation of a quantitative structure-activity relationship (QSAR) model is dependent on descriptors. Kimito has used a graph convolution neural network in descriptor-independent QSAR analysis. It is possible to extract and visualize positive and negative features on substructure on each structure, and to summarize important substructure profiles for a specific target. This idea has been applied to the compound-protein interaction model: it is possible to extract, visualize and summarize important pairs of substructure profiles between compound and target.

Chemography: toward "universal" maps of druglike space

Alexandre (Sasha) Varnek of the University of Strasbourg, France had the goal of making a map encompassing all known compounds and activities in druglike chemical space. Oprea and Gottfries[80] were the first to use the term "chemography" in cheminformatics; Varnek's team continues to follow the theme.

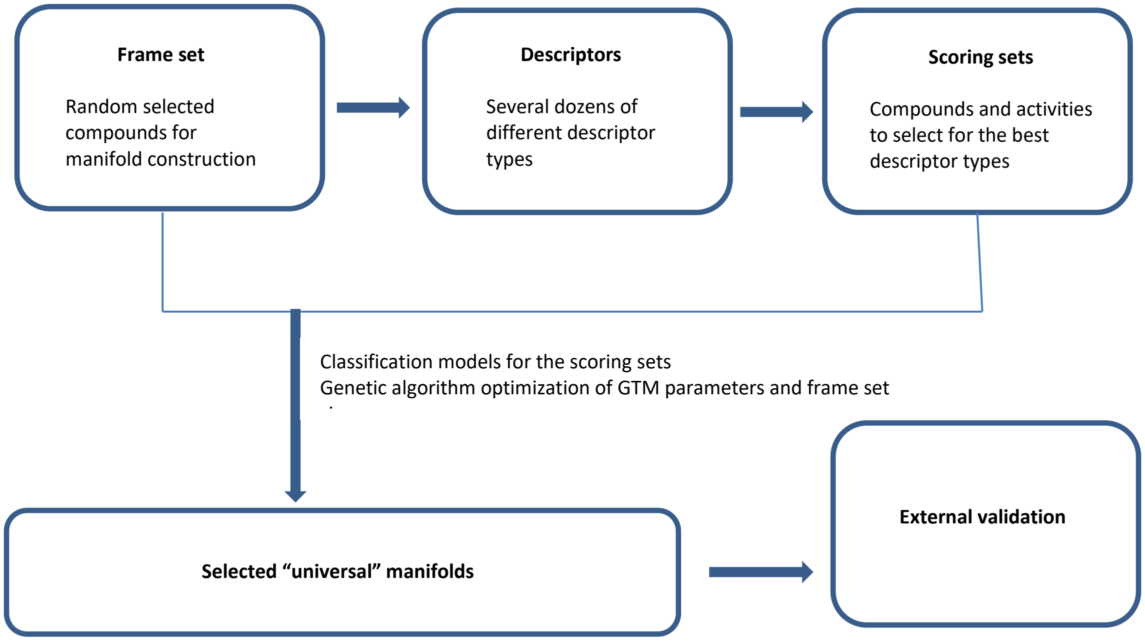
To go from multidimensional descriptor space to 2D latent space requires dimensionality reduction. Sasha uses *In Silico* Design and Data Analysis (ISIDA) descriptors[81-83] of which there are several hundred. A great many dimensionality reduction methods have been reported but Sasha chose to use Generative Topographic Mapping (GTM),[84-86] a probabilistic extension of self–organizing maps (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.[86]

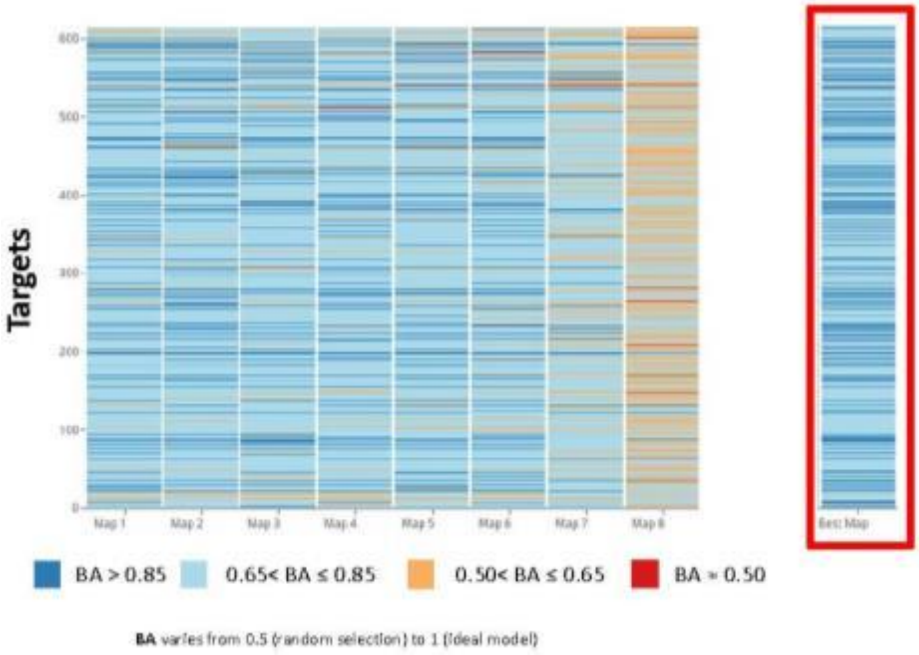
Sasha showed a probability density distribution in the latent space. Projection of an object on a GTM is described by the probability distribution ("responsibilities") over the lattice nodes. Using GTM, one can, for each molecule, evaluate the probability of finding it in a point on the grid. This probability distribution can be used to prepare an "activity landscape" or "class landscape" on a 2D map which, in turn, allows the user to make predictions of activities, or the probability that new compounds will be active or inactive.[87]

A "universal" map is expected to accommodate a variety of known chemotypes; to be able to distinguish different activity classes; and to separate actives and inactives within a given activity class. It needs to exhibit neighborhood behavior (NB). A given molecule may display several different activities, so several different descriptor spaces might be needed to construct the NB-compliant maps. This is equivalent to separating objects into different maps according to the color or the shape of the descriptors.

The choice of descriptors is a vital issue for chemical space construction. Optimal descriptors are supposed to be provided by the best GTM-based regression or classification models built on some “scoring” dataset(s). The more the activities that are used, the more universal is a map.[88] Descriptors leading to the largest scores are selected. Sasha showed the modeling workflow for producing universal maps:

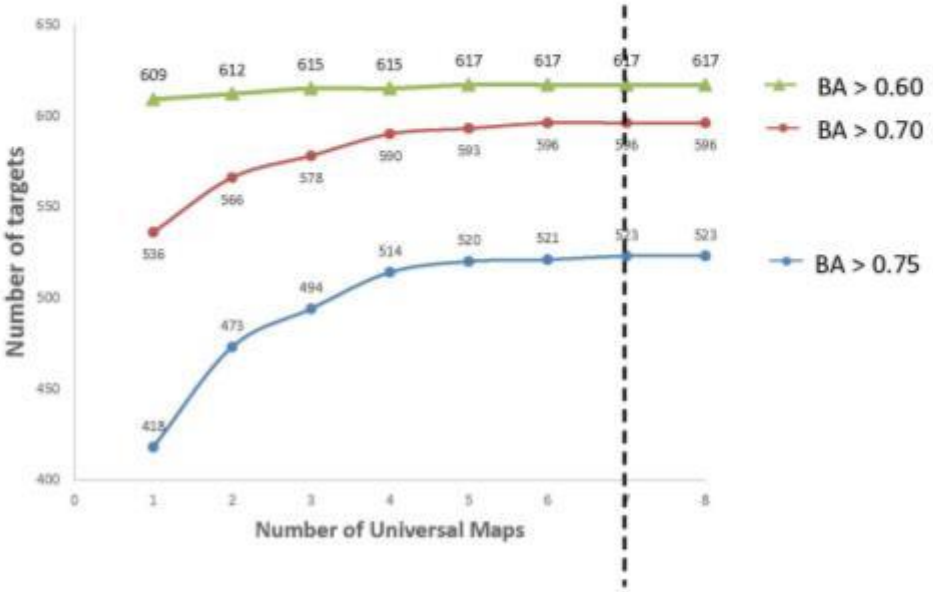

Frame set
Random selected compounds for manifold construction
Descriptors
Several dozens of different descriptor types
Scoring sets
Compounds and activities to select for the best descriptor types
Classification models for the scoring sets
Genetic algorithm optimization of GTM parameters and frame set
Selected “universal” manifolds
External validation

Frame sets from the ChEMBL database were used: five sets containing from 8,500 to 30,000 compounds. There were 100 ISIDA descriptor types. The scoring sets were 236 targets, including GPCRs, kinases, and nuclear receptors; and more than 30,000 compounds. From the scoring sets, universal manifolds were determined, and cycled through 382 validation datasets of over 100, 000 compounds, and the scoring sets, in order to find the best universal manifolds. All of the ChEMBL compounds were then projected on the manifolds. Eight universal maps described more than 1.5 million compounds and 618 targets from the ChEMBL database.

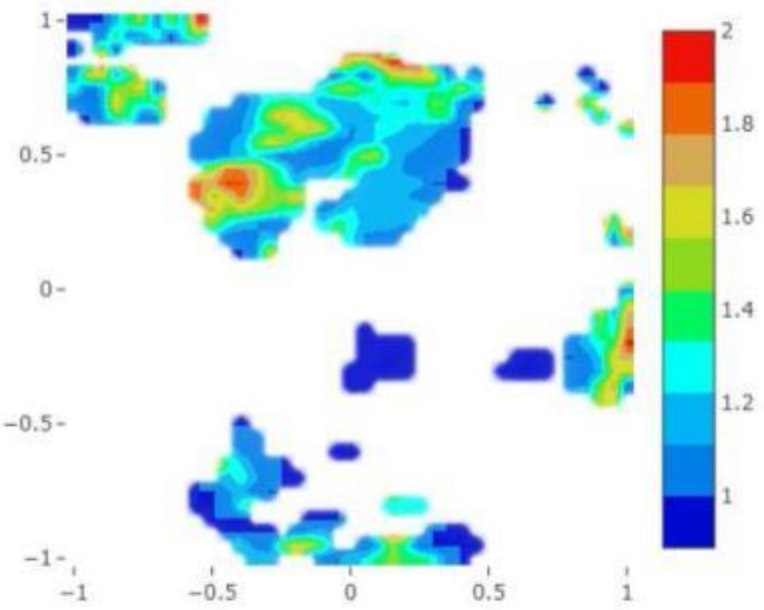


BA = balanced accuracy (defined as the mean of specificity and sensitivity)

Seven maps correctly predict the classes (active/inactive) for ligands against more than 600 targets:

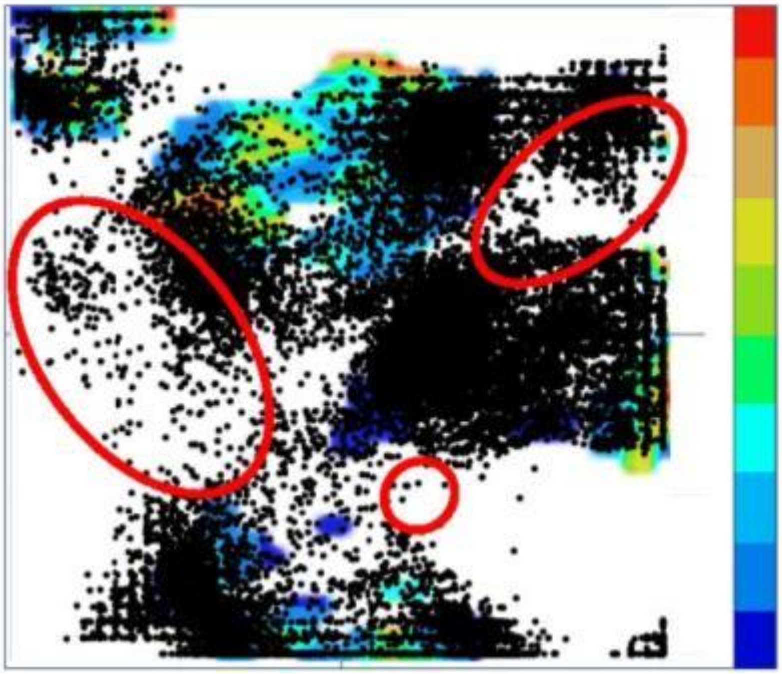


Universal maps can be applied to virtual screening. In a case study, Sasha's team virtually screened the Directory of Useful Decoys (DUD) using the activity landscape for adenosine receptor A2a. Molecules dropping into empty areas are not considered.



Red = active, blue = inactive. There were 1,303 actives in ChEMBL and 3,618

Sasha presented the results below (79 actives and 28,002 inactives found in DUD). The molecules in empty zones (schematically shown by red ellipses) were discarded in virtual screening



A consensus model performed better than its seven, constituent individual models. The number of molecules discarded by the GTM applicability domain ranged from 348 to 1643 for the seven maps, but was zero for a consensus model, showing that a consensus model provides better data coverage than the individual models.

Universal maps can also be used to detect privileged structures. Privileged substructures recur in compounds active against a given target, and are associated with biological activity. They are important in the design of novel bioactive compounds. The classical approach to detection of privileged substructures is scaffold-centric. In the GTM-based approach, structurally related structures are extracted from the responsibility distribution. Compounds with similar responsibility vectors are expected to be related. In recent work,[89] zones of a map preferentially populated by target-specific compounds were delineated. This helps to capture common substructures. On the basis of the common substructures, compounds were grouped together by GTM. Such privileged structural motifs were identified across three major target superfamilies including proteases, kinases, and G protein coupled receptors. Three universal maps are needed in order to extract the privileged structural motifs for the main classes of antivirals in ChEMBL.[87]

In summary, universal maps are able to support predictive models for a broad spectrum of biological activities and, thus, they could be used as a pharmacological profiling tool. Several (at least seven) maps are needed in order to achieve good performance in separating actives from inactives. Privileged substructures (not necessarily scaffolds) can be extracted from universal maps. Other cheminformatics applications with universal maps are virtual screening, target prediction, and drug repurposing.

Finally, Sasha touched briefly on *de novo* design using a deep learning and GTM combination. This is work reported in the master’s thesis of Sasha’s student Boris Sattarov. An autoencoder produces an efficient, dense representation of an input object by performing specific compression of learned data. A long short-term memory (LSTM) sequence-to-sequence autoencoder can perform SMILES reconstruction. An LSTM encoder takes a SMILES string, and produces real numbers (as latent variables) from which the LSTM decoder reconstructs the SMILES. A database of SMILES strings can then be used as input to the trained encoder, and a GTM built on the latent variables of the autoencoder. Novel structures from specific areas of the map are generated by the trained decoder, which produces SMILES strings of the novel compounds from latent variables.

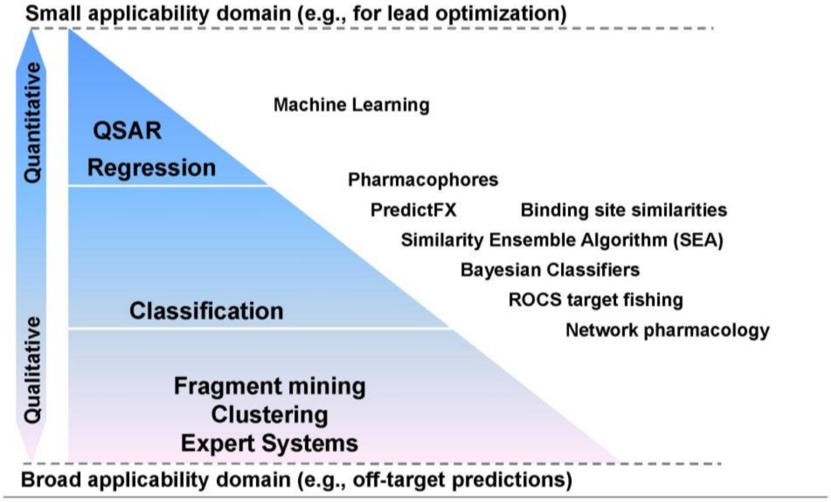
A case study was generation of analogues of A2a inhibitors. *De novo* structures were generated by sampling of GTM zones populated by existing ChEMBL structures. Validation was carried out by docking the ligands on the 2YDO PDB X-ray structure using the Sampler For Multiple Protein -Ligand Entities (S4MPLE)[90] tool. The distribution of the docking scores of the generated structures closely followed that of real actives.

Artificial Intelligence in drug design

Much of the afternoon session centered on artificial intelligence (AI). Representing five co-authors at Sanofi-Aventis in Germany, Karl-Heinz Baringhaus addressed the subject of AI in drug design. A recent review has analyzed the relative contributions of each of the steps in the drug discovery and development process to overall R&D productivity.[91] Karl-Heinz said that there are opportunities for AI in *all* areas of the drug discovery and development value chain. It has been estimated that the AI market in healthcare will be worth $6.6 billion in 2021. AI is a system for problem solving, making automated decisions, perceiving the environment, and taking actions. A "strong" AI system has the same capabilities as human thinking and reasoning; a "weak" AI system transfers a single human capability to a system (e.g., recognition of texts, voice, or images).

Drug design plays a pivotal role in lead identification and optimization, and AI has a role in this through deep learning, in particular for compound classification, *de novo* design and *in silico* profiling. For example, the majority of adverse drug reactions (ADRs) are dose-dependent, and ADRs can often be predicted on the basis of the pharmacology profiles of the candidate compound.[92]

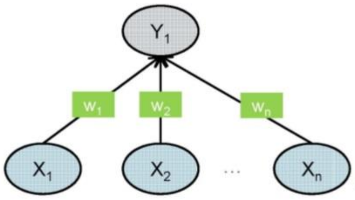
The application of AI in particular in the field of ADMET and antitarget modeling was the main focus of Karl Heinz's talk. He started with an overview of some predictive data mining methods:



Many 2D descriptors describing molecular topology have been used, as well as pharmacophore fingerprints, and 1D descriptors such as log*P*. Machine learning algorithms generate knowledge from experience: they learn iteratively from data without being explicitly programmed. The models generated can be used for prediction of novel instances. Classification methods and quantitative models have been developed. Different algorithms include linear regression, decision trees, random forest, support vector machine, etc. An example is building (Q)SAR with Cubist, in which both a decision tree and regression are combined.

Different QSAR models have different validity domains (e.g., hERG or CYP450). For a domain to be valid, predicted compounds must be similar to training molecules, and the descriptor range of predicted compounds must relate to the training set molecules. Karl-Heinz and his colleagues have reported some CYP450 2D-QSAR models, among others.[93]

Deep learning is a subsection of machine learning. It has been particularly effective due to the use of neural networks. It is based on analysis of large volumes of data (big data).



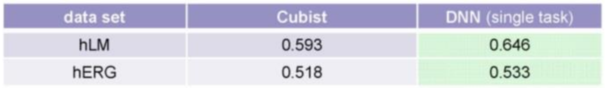
$$Y_1 = f(Z)$$

$$Z = \sum_{k=1}^{n} X_k w_k + b$$

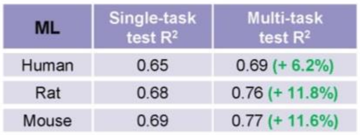
- Each **neuron** (*Y*) receives input from all other **input** units (*X*)
- Effect of each input is controlled by a **weight** (*w*) and **bias** (*b*)
- **Learning**: adapting the weights and biases to perform useful computations
- Neurons are activated via **activation function** (*f*)
- **Features** are extracted automatically

A single-task, deep neural net (DNN), aims to generate a model for just one activity. Molecules are described by a set of descriptors which are used as input for the network. Most often, fully connected and sequential deep neural nets with several hidden layers are applied for model building. Multitask deep networks have yet to be widely deployed in the pharmaceutical industry but recent work[94] has demonstrated that multitask deep networks are surprisingly robust, and can offer strong improvement over random forests. In a multitask network, a set of descriptors is used for model building of multiple activities in parallel, yielding multiple outputs (one for each input activity). Within model building all weights are shared.

Karl-Heinz compared Cubist and DNN results from a benchmark of internal prediction models of human liver microsome (hLM) stability and hERG inhibition. Application of multitask DNNs gave results superior to those for single-task DNNs. Multitask models capitalize on hidden trends.



data set	Cubist	DNN (single task)
hLM	0.593	0.646
hERG	0.518	0.533

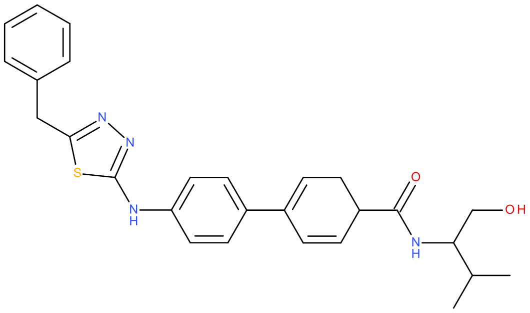


ML	Single-task test R^2	Multi-task test R^2
Human	0.65	0.69 (+ 6.2%)
Rat	0.68	0.76 (+ 11.8%)
Mouse	0.69	0.77 (+ 11.6%)

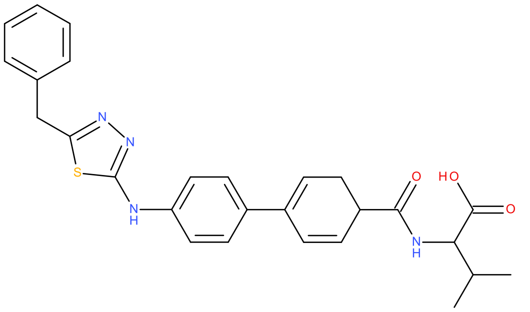
TensorFlow is an open source software library for high performance numerical computation originally developed by Google. It comes with strong support for machine learning and deep learning. Keras is a high-level neural networks API, written in Python and capable of running on top of TensorFlow.

Karl-Heinz gave a lead optimization example[93] concerning diacylglycerol O-acyltransferase 1 (DGAT1) inhibitors.

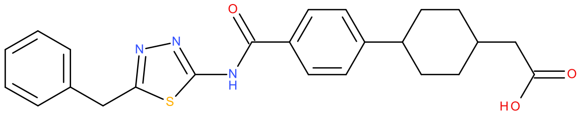
A lead structure had IC_{50} = 0.25 µM with metabolic lability = 53%:



A global model for metabolic lability, which is predictive for the chemical series of DGAT1 inhibitors in question, was successfully used to guide the lability optimization of the lead structure. A follow-up structure:



had IC_{50} = 0.03 µM, and metabolic lability = 1%, but it had poor permeability. After rescaffolding a permeable structure was found with IC_{50} = 0.04 µM, and low metabolic lability.



Karl-Heinz's final topic was *in silico* profiling of compounds and polypharmacology. Sanofi has numerous off-target profiling panels for receptors and enzymes, kinases and ion channels, covering important and promiscuous antitargets. In total, Sanofi has biological data available for about 3 million compounds against about 2,000 biological targets. These data were used for QSAR model generation in order to generate profiling models. Data for multiple targets (receptors, protein kinases, and ion channels) are retrieved, normalized, and filtered. Genetic algorithms are used for variable selection, in combination with Cubist regression trees as machine learning tool. Each *in vitro* profiling assay has an *in silico* counterpart, but only core panels have a large applicability domain. Forward prediction for external, experimental data after model generation has been successful (R^2 and $Q^2 > 0.6$) and a web application (CT+) for *in silico* profiling was rolled out.

Challenges in *in silico* profiling are the high number of experimental data for model building, and the building of reliable and predictive models, including validity domain estimation (VDE). Sanofi builds predictive models for about 400 targets of interest. All models are thoroughly validated, including VDE. The combination of the predictProfile models with CTlink, developed by Jordi Mestres' team,[95] is very powerful in selecting the most attractive hit series out of a high throughput screening run, and in the identification of potential side effects of lead series prior to subsequent optimization. Compound repurposing as well as polypharmacology of compounds can be addressed by both CTlink and the Sanofi models.

AI "leapfrogs" early off-target predictions and annotations, identification of series-related potential issues, prioritization of hit and lead series, and design of preclinical candidates. As Johann-Wolfgang von Goethe said "Es ist nicht genug, zu wissen, man muß auch anwenden; es ist nicht genug, zu wollen, man muß auch tun". ("It is not enough to know, you also have to apply; it is not enough to want, you also have to do".)

Robot Scientists automating drug design

Ross D. King, of the University of Manchester, United Kingdom, gave a talk on the automation of drug design. (The following talk, "Accelerating drug discovery through a fully automated design-make-test-analyze workflow", by Michael Kossenjans of AstraZeneca, was unfortunately withdrawn.) Ross said that the technology drivers behind his own work were improved computer hardware, improved data availability, and improved software (new machine learning methods, deep mining, etc.).

There have been multiple AI hype cycles, but this time it seems different. The speed of advance of AI, and especially machine learning, has surprised Ross. Machine learning is the core technology of Google, Facebook, Amazon, Tencent, Alibaba, Baidu, and many other systems. AI systems have superhuman scientific reasoning powers. They flawlessly remember vast numbers of facts; execute flawless logical reasoning, and optimal probabilistic reasoning; learn more rationally than humans, from vast amounts of data; extract information from millions of scientific papers, and so on. Yet AI systems have to be balanced with human abilities.

Science is a good application area for AI. Scientific problems are abstract, and restricted in scope, which suits AI, but also involves the real world. Nature is also honest and not trying to fool us, which also helps AI reasoning. Nature is also a worthy object of our study, unlike addictive advertising, and the generation of open scientific knowledge is a public good.

The first application of AI to science was the DENDRAL project in the 1960s and 1970s, the primary aim of which was to study hypothesis formation and discovery in science. Using knowledge of chemistry, it helped organic chemists to identify unknown organic molecules by analyzing their mass spectra. Meta-DENDRAL[96] was the project's machine learning system.

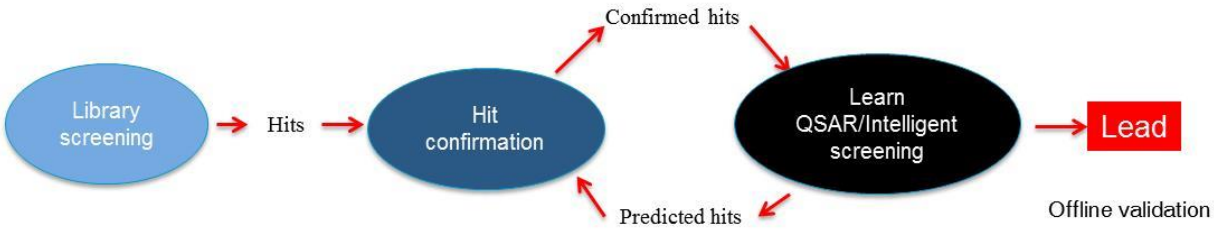
King's laboratory and collaborators[97] have developed "Robot Scientists": physically implemented robotic systems that apply techniques from artificial intelligence to execute cycles of automated scientific experimentation. A Robot Scientist can automatically execute cycles of: hypothesis formation, selection of efficient experiments to discriminate between hypotheses, execution of experiments using laboratory automation equipment, and analysis of results. The motivation for developing Robot Scientists is to understand science better, and to make scientific research more efficient.

Robot Scientists have the potential to increase the productivity of science. They can work more cheaply, faster, more accurately, and for a longer time than humans. They can also be easily multiplied. Thus they enable the high-throughput testing of hypotheses. They also have the potential to improve the quality of science by enabling the description of experiments in greater detail and semantic clarity.

The Robot Scientist "Eve" was designed to automate and integrate drug screening, hit confirmation, and QSAR development. Its combination of novel automation with synthetic biology assays enables faster and cheaper drug design.[98] Eve's focus was originally on malaria, and neglected tropical diseases such as shistosomiasis, leishmaniasis and Chagas disease. Hundreds of thousands of people die of these diseases, and hundreds of millions of people suffer infection. It is clear how to cure such diseases (unlike many diseases): kill the parasites. There is also little competition from the pharmaceutical industry. Enzymes targeted are dihydrofolate reductase (DHFR), N-myristoyl transferase (NMT), and phosphoglycerate kinase (PGK).

Eve uses graphs and standard cheminformatics methods to represent background knowledge. Eve uses induction (QSAR learning) to infer new hypotheses, active learning to decide efficient experiments, and an econometric model[98] to decide which compounds to test. Eve currently uses processes.[98] This has the advantages of being generative, and of outputting probabilities. Eve uses active learning to select compounds to test its hypotheses. This machine learning method can select its own examples, in this case from a set.

Standard chemical library screening is brute force, but Eve uses intelligent screening. In the standard "pipeline", the processes are not integrated, but in Eve they are automated and integrated:



Eve's hardware has acoustic liquid handling, high throughput 384-well plates, two industrial robot arms, an automated 60x microscope, liquid handlers, fluorescence readers, barcode scanners, a dry store, an incubator, a tube decapper, and other items.

Ross and his co-workers wanted to compare their AI-based screening against the standard brute-force approach. While simple to automate, standard screening is slow and wasteful of resources, since every compound in the library is tested. It is also unintelligent, as it makes no use of what is learnt during screening. To do the comparison, an econometric model was used.[98] Factors involved in the calculation of the utility of Eve were the number of compounds not assayed by Eve; the cost of the time to screen a compound using the mass screening assay; the cost of the loss of a compound in the mass screening assay; the number of hits missed by Eve; the cost of the time to screen a compound using a cherry-picking assay (a confirmation or intelligent assay); the cost of the loss of a compound in a cherry-picking assay; the utility of a hit; and the number of compounds assayed by Eve. The researchers demonstrated that the use of AI to select compounds economically outperforms standard drug screening.

Eve has discovered hits and leads against targets in multiple parasites, most notably that triclosan inhibits *Plasmodium* DHFR.[98] Triclosan is a simple compound, generally regarded to be safe: it is used in toothpaste. It targets both DHFR and fatty acid synthase II (FAS-II), well established targets. Activity has been demonstrated using multiple wet experimental techniques. Triclosan works against wild-type and drug-resistant *Plasmodium falciparum*, and *Plasmodium vivax*.

The goal of science is to increase our knowledge of the natural world through the performance of experiments. This knowledge should be expressed in formal logical languages which promote semantic clarity, which in turn supports the free exchange of scientific knowledge, and simplifies scientific reasoning. Robot Scientists provide excellent test beds for the development of methodologies for formalizing science. Using them it is possible to capture completely, and curate digitally, all aspects of the scientific process. The LABoratory Ontology for Robot Scientists (LABORS)[99,100] is designed to give the scientific community open access to the Robot Scientist experimental data and metadata.

The effect of AI on intellectual property is an interesting question. U.S. patent law is clear that only humans can invent patents. U.K. and European Union patent laws allow non-human inventors. Many patents in the United States are possibly invalid as the named inventor is incorrectly named as a human. AI-generated copyright works would seem to be in the public domain, except in a few countries such as the United Kingdom. Making AI systems legal entities is favored by companies to shield themselves from legal liabilities.

In chess there is a continuum of ability from novices up to grandmasters. Ross argues that this is also true in science, from the simple research of Eve, through what most human scientists can achieve, up to the ability of a Newton or Einstein. If you accept this, then just as in chess, it is likely that advances in computer hardware and software will drive the development of ever smarter Robot Scientists. In favor of this argument is the ongoing development of AI and laboratory robotics.

In response to Gisbert's request for a grand challenge, Ross said that, in his opinion, the deepest challenge in applying AI to science is finding radically new representations. Einstein, for example, noticed, in his concept of space-time, that the same physical phenomenon was described in two different ways, depending on what was moving.

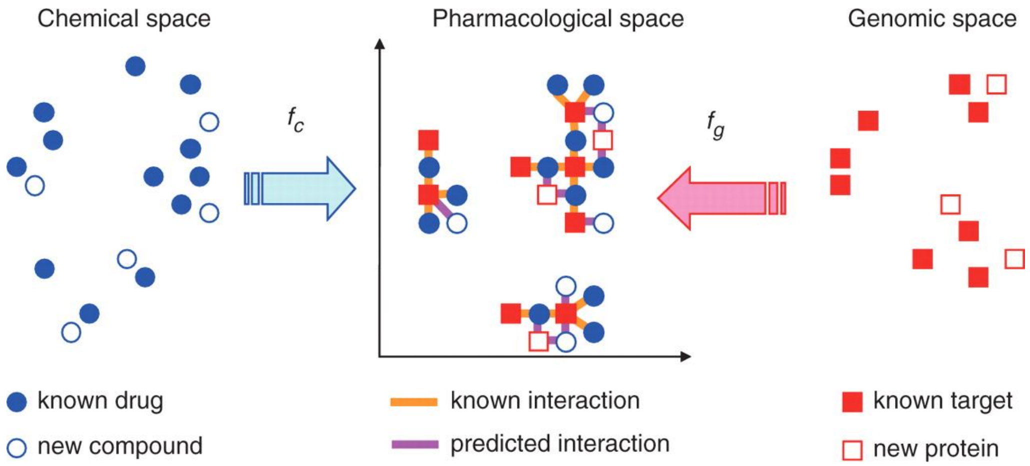
Ross believes that the collaboration between humans and Robot Scientists will produce better science than either humans or robots can alone. Even though a computer first beat the world chess champion over 20 years ago, teams of humans and computers still play better chess than either a human or a computer alone. Scientific knowledge will be primarily expressed in logic with associated probabilities, and published using the Semantic Web. Improved productivity of science leads to societal benefits: better food security, better medicines, etc.

Data-driven drug discovery and repositioning by machine learning methods

Yoshihiro Yamanishi of Kyushu Institute of Technology, Japan, spoke about drug repositioning. The identification of new indications of drugs, or in-house compounds, is an efficient strategy for drug development, and it has received much attention recently. A great deal of information (e.g., safety, pharmacokinetics, and manufacturing process) is available on existing drugs. The drug repositioning approach can increase the success rate of drug development, and reduce the cost in terms of time, risk, and expenditure. Well-known examples are sildenafil (Viagra) developed for angina, but now used in the treatment of erectile dysfunction, and pulmonary hypertension; and minoxidil (Riup, Rogaine) developed for hypertension but now used to treat alopecia.

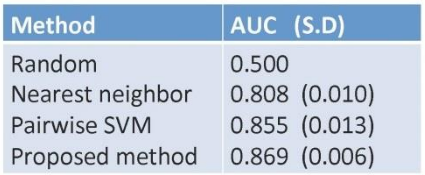
In the past, the approach to drug repositioning has been dependent on serendipity. The aim of the current work was computational drug repositioning based on biomedical big data (compounds, genes, proteins, and diseases). The researchers have developed novel machine learning methods to predict new associations between drugs and diseases, based on the molecular understanding of a variety of diseases: disease-causing genes, disordered pathways, environmental factors, and abnormal gene expression. Characteristic molecular features are often shared among different diseases, and networks of disease-disease relationships can be produced based on those molecular features. Yoshihiro and his co-workers set out to predict drug-protein-disease networks by machine learning. There are many approved drugs (about 60%) whose targets remain unknown. Using the networks, unknown interactions can be predicted. The genomic space (of proteins) is inter-related with chemical space in chemogenomics, with phenotypic space in phenomics, and with transcript space (drug-induced gene expression) in transcriptomics.

In the chemogenomics approach it is postulated that chemically similar drugs are predicted to interact with similar target proteins.[101-106] Compound similarity (chemical space) is measured based on descriptors for chemical structures, and protein similarity is calculated based on motifs, domains, and amino acid sequences. In the chemogenomics approach, Yoshihiro and his co-workers have used the generalized Jaccard coefficient for compound similarity, and the normalized Smith-Waterman score for protein similarity.[107] They investigated the relationship between drug chemical structure, target protein sequence, and drug-target network topology. They then developed a new supervised method to infer unknown drug-target interactions by integrating chemical space and genomic space into a unified space that they call “pharmacological space”.



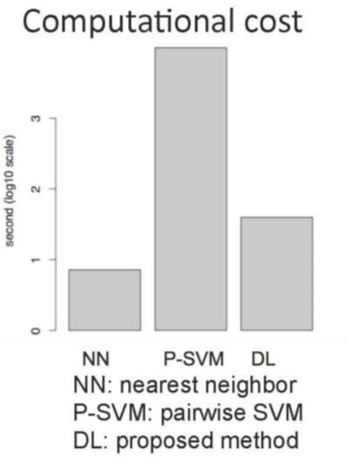
The supervised method is a two-step process. First, a model is learned to explain the "gold standard"'. Second, this model is applied to compounds and proteins absent from the "gold standard" in order to infer their interactions.[104,108,109] The novel method used is the bipartite graph learning method.[104,110] Details of the kernel-based distance learning algorithm[108,110] have been published.

The performance of the system was evaluated using a benchmark set of 6,769 interactions involving 1,874 drugs and 436 proteins from the Kyoto Encyclopedia of Genes and Genomes (KEGG), DrugBank, and Matador. Performance was superior to that of nearest neighbor and pairwise support vector machine (P-SVM) methods:



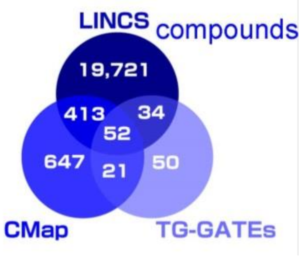
Method	AUC (S.D)
Random	0.500
Nearest neighbor	0.808 (0.010)
Pairwise SVM	0.855 (0.013)
Proposed method	0.869 (0.006)

AUC = Area under ROC curve
S. D. = standard deviation



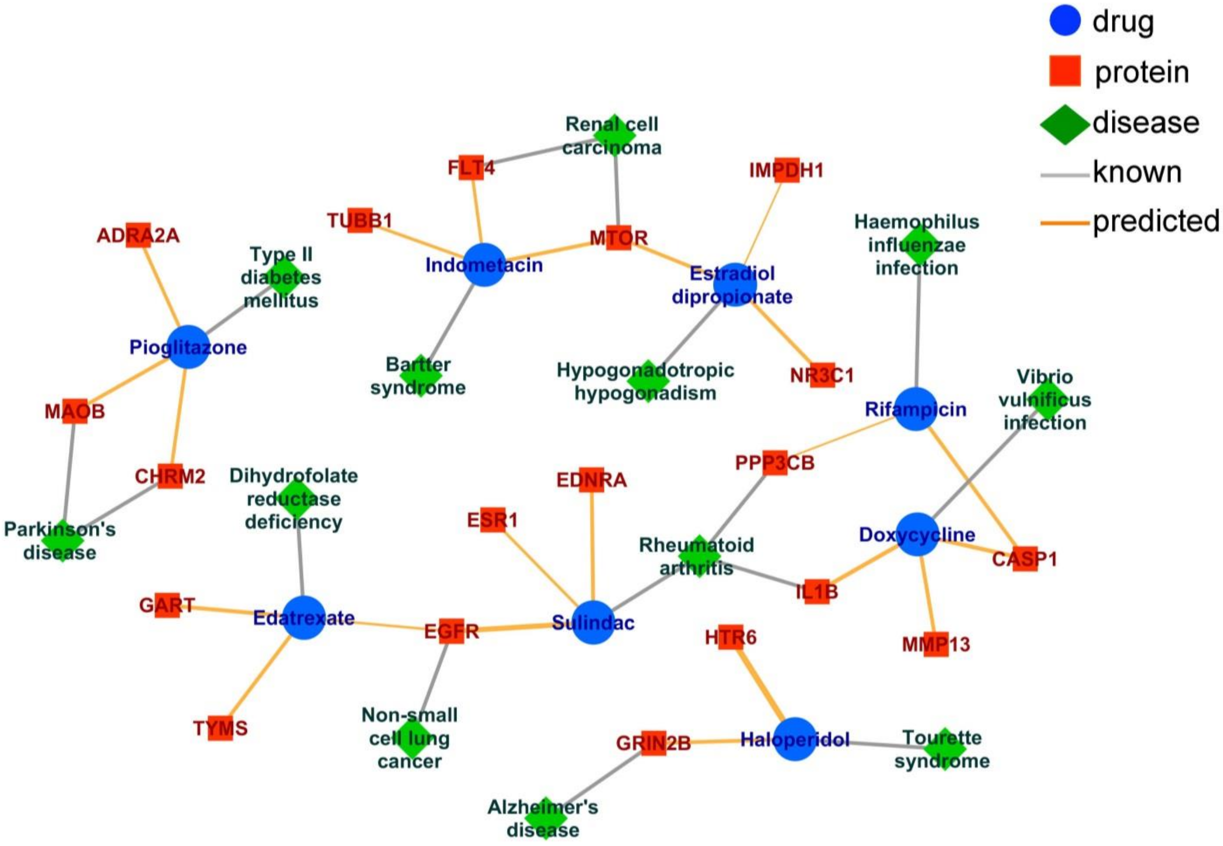
In the transcriptomics approach, it is postulated that drugs with similar gene expression patterns are likely to share common target proteins.[107,111-113] Each drug is represented by a gene expression profile in which each element is the ratio of drug treatment against control condition (ratio of gene expression value after drug treatment to gene expression value before drug treatment). The Library of Integrated Cellular Signatures (LINCS) is an NIH program which funds the generation of perturbational profiles across multiple cell and perturbation types, as well as read-outs, at a massive scale. The Connectivity Map (CMap) is a catalog of gene-expression data collected from human cells treated with chemical compounds. The Open Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System

(TG_GATEs) is a toxicogenomics database that stores gene expression profiles with chemical treatments. Gene expression profiles for 20,220 compounds, 22,277 genes, and 77 cells were analyzed.

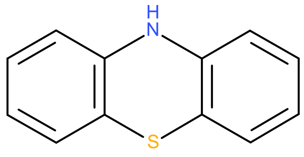


Performance evaluation on several benchmark datasets of different chemical diversities proved that the transcriptomics approach enables the prediction of target proteins without dependence on prior knowledge of chemical structures,[107] whereas the prediction accuracy of the chemogenomics approach is highly dependent on chemical structure similarities.

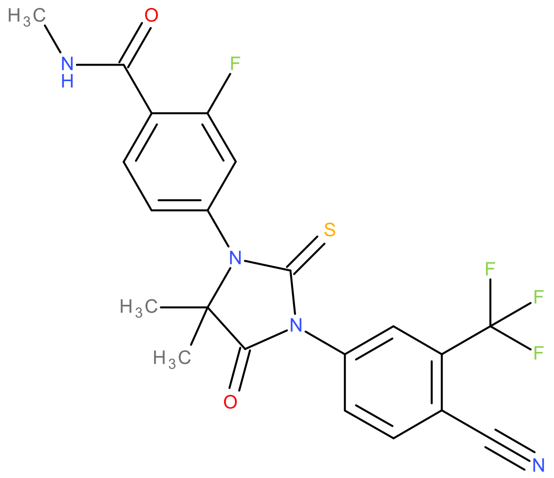
Yoshihiro and his team were able to perform large-scale prediction of new drug indications: from 8,270 drugs and 1,401 diseases, they predicted 196,048 new drug-disease associations, involving 6,301 drugs and 762 diseases. Yoshihiro showed a drug-protein-disease network:



An example of a transcriptomics-based prediction was the predicted indication of prostate cancer for phenothiazine, an antipsychotic drug.[113] The estimated protein was the androgen receptor (AR). A similar compound in the learning set was enzalutamide. The predicted inhibitory effect on AR was experimentally confirmed.



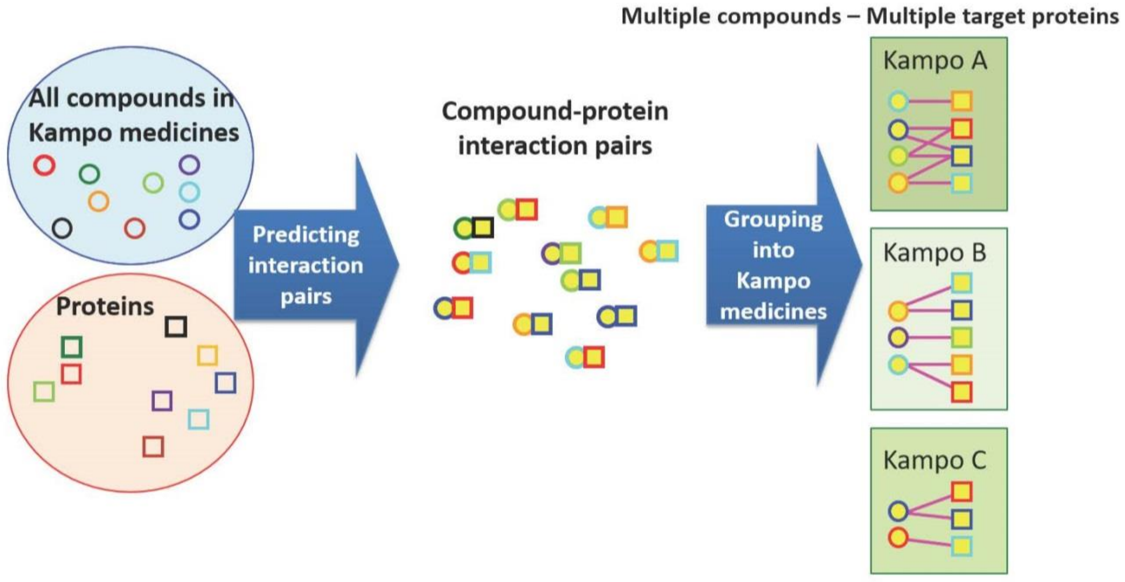
phenothiazine



enzalutamide

Yoshihiro and his co-workers have elucidated pathway activities from differentially regulated genes: they performed pathway enrichment analyses of regulated genes to reveal active pathways among 163 biological pathways.[113] They studied the relationship between identified pathways (activated and inactivated) and drug efficacy classes.

They have also worked on repositioning of natural medicines such as herbal medicines, crude drugs, and Kampo drugs. Target proteins and target pathways for Kampo drugs have been predicted from biomedical data by machine learning.[114] Daikentyuto (a Kampo drug for stomach ache) was predicted to work for colitis-associated colon cancer, and this was experimentally validated by mice models.

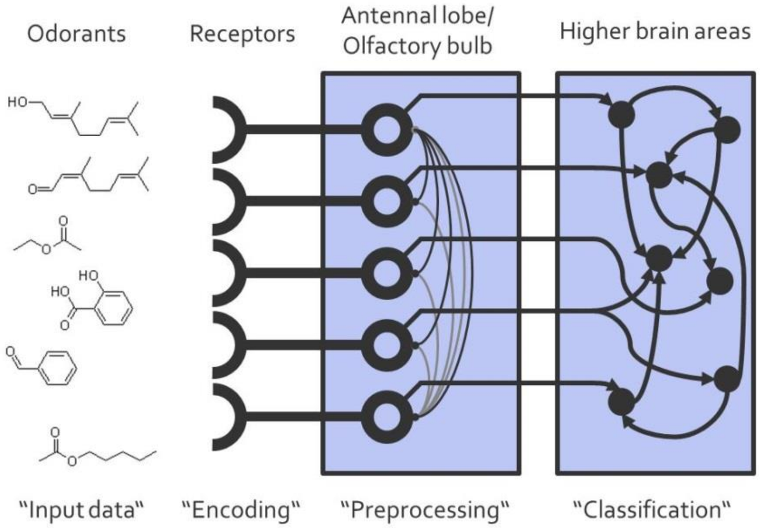

Multiple compounds – Multiple target proteins
All compounds in Kampo medicines
Proteins
Predicting interaction pairs
Compound-protein interaction pairs
Grouping into Kampo medicines
Kampo A
Kampo B
Kampo C

The methods proposed by Yoshihiro can predict potential target proteins, pathway activities, and new therapeutic effects of drug candidates, moving from organ-based disease classification to mechanism-based disease classification. Yoshihiro listed two challenging issues in the integration of cheminformatics and bioinformatics. They are computational methods (a) to take into account the biological systems (various molecular interaction networks) for target identification and drug screening, and (b) to use multilayered omics data for drug discovery in diseasome analysis, patient stratification, biomarker detection, target identification, and drug screening.

Pattern recognition on neuromorphic hardware inspired by the chemical sense

Michael Schmuker of the University of Hertfordshire, United Kingdom, introduced us to “odor space”. The sense of smell plays an important role in everyday life, but while we readily understand how to describe visual stimuli with light wavelengths and contrast, or sounds with frequency and amplitude, we lack understanding of the fundamental features of odors. In other words, we do not understand odor space[115] the notion of which describes different aspects of the sense of smell. There is chemical space, in which odorants are arranged along their physicochemical properties. The encoding of odors into neuronal responses defines the sensory space. The perceptual organization and meaning of odors is the basis for perception space. Finally, odors are embedded in a physical space that governs how odorants move from their sources to our noses.

Michael presented a diagram of the olfactory system from a data processing point of view:



In the encoding stage, olfactory receptors are broadly tuned and their “receptive fields” overlap: a class of receptor neurons responds to a wide range of odorants, and an odorant typically triggers responses in several different receptor neuron classes. The input space is vast, and large overlapping receptive fields ensure coverage.

In neurobiology, lateral inhibition is the capacity of an excited neuron to reduce the activity of its neighbors. Lateral inhibition disables the spreading of action potentials from excited neurons to neighboring neurons in the lateral direction. This creates a contrast in stimulation that allows increased sensory perception. It occurs primarily in visual processes, but also in tactile, auditory, and olfactory processing. Preprocessing through lateral inhibition in the data processing system above acts as a filter to enhance the representation of the input data. In this particular setting it counteracts the negative aspects of overlapping receptive fields.[116-118]

After the sensory input is processed in the antennal lobe (AL) network, projection neurons (PNs) convey the sensory code from the AL to higher brain regions. Michael and Gisbert have used a computational network model[119] of the olfactory system which serves as blueprint for a pattern recognition pipeline. They applied this model to predicting the smell of molecules and validated it on bioactivity data for druglike compounds, demonstrating that the computational model can be used not only for olfaction, but for any data in general.

The brain remains the ultimate benchmark for energy-efficient pattern recognition. Much of its efficiency is attributed to its massively parallel architecture and sparse messaging *via* action potentials, called spikes. The brain's architecture has long inspired engineers in building computing machines. Recently, major chip designers like IBM and Intel, and also academic research teams, have produced neuromorphic hardware systems, that is, silicon chips containing a large number of neuron-like small computing units, which communicate *via* short messages with precise timing. The Spikey chip is a neuromorphic chip developed at the University of Heidelberg.[120] It is a mixed signal system with analog neurons and digital routing. It is 10,000 times faster than biology, and supports a wide range of network topologies with a few hundreds of neurons. The challenge is to solve real-world computing tasks with spiking networks on the hardware system.

Michael and his colleagues have constructed a spiking neural network for the classification of multivariate data. The data are converted into spike trains using "virtual receptors" (VRs). Their output is processed by lateral inhibition and drives a winner-take-all circuit that supports supervised learning. VRs are conveniently implemented in software, whereas the lateral inhibition and classification stages run on accelerated neuromorphic hardware. When the network had been trained and tested on real-world datasets, the classification performance was on a par with a naive Bayes classifier (R_K=0.87 with 5-fold cross-validation, and 50 repetitions, compared to R_K=0.89 for naive Bayes).

An example is handwritten digit recognition,[121,122] using the Modified National Institute of Standards and Technology (MNIST) database. The network used by Michael and his colleagues outperformed naive Bayes through the preprocessing by lateral inhibition. The neuromorphic advantage is that lateral inhibition scales in constant time on hardware, whereas with conventional computers, scaling is with input dimensionality.

Today's pattern recognition runs on GPUs, performing matrix operations (linear algebra), involving dense communication, and very high bandwidth. Energy consumption is high: 300 W per GPU. Modern AI systems run on hundreds of GPUs plus hundreds of thousands of CPU cores (see, for example, the

OpenAI Five Dota agent). Compare these computers with the brain, which has amazing pattern recognition, reasoning, and control capabilities; massively parallel computation with lightweight compute units; and sparse communication *via* timed events (i.e., action potentials). Moreover the brain's entire energy consumption is only 20 watts.

GPU-based AI is incredibly powerful, but also incredibly power-hungry. Moore's law is to be abandoned, and alternative architectures are on the rise. Neuromorphic computing is event-based computing, and is power efficient. Big companies have launched neuromorphic chips: take, for example, IBM TrueNorth and Intel Loihi. Academic projects include SpiNNaker[122] at the University of Manchester, and BrainScaleS at the University of Heidelberg, both of them involved in the Human Brain Project.

SpiNNaker is a specialized, many-core system, with an ARM-based architecture. It consumes low power, is portable, and connects through a 100 Mbit Ethernet. Spikey is a mixed-signal system with analog neurons and digital routing (see above). It is low power, and portable, and has a USB 2 connection. GPU-enhanced Neuronal Network (GeNN) is a meta-compiler that generates optimized CUDA-kernels for spiking networks. GeNN compiles networks for accelerated execution on GPUs, which are high power, not portable, and connect by PCIExpress. On all neuromorphic platforms, the workstation controlling the hardware consumes the largest fraction of the power. For network simulation, a GPU is more efficient than a CPU for large networks; for small networks, the CPU simulation is the more efficient. SpiNNaker uses less power than a GPU. Moreover, power requirements are invariant to network size (up to the maximum size supported), which gives it an edge in power efficiency over both CPUs and CPUs.[122]

Michael's next topic was the physical component of odor space. In an open sampling system, where the chemosensory elements are directly exposed to the environment being monitored, the identification and monitoring of chemical substances presents a challenge due to the dispersion mechanisms of gaseous chemical analytes, namely diffusion, turbulence, and advection. Vergara *et al.*[123] examined electronic nose sensors in a turbulent wind tunnel. The sensors were at varying distance from the odor source. The authors made publicly available 18,000, 72-dimensional time-series recordings. The concentration of a gas decreases with distance from source, but in order to predict distance, the concentration at the source must be known. In a turbulent environment, gas intermittency also depends on source distance. Michael and his co-workers aimed to use spatiotemporal features of gas plumes for distance estimation, using Vergara's dataset.

They have demonstrated[124] that by appropriate signal processing, off-the-shelf metal-oxide sensors are capable of extracting rapidly fluctuating features of gas plumes that strongly correlate with source distance. They showed that with a straightforward analysis method it is possible to decode events of large, consistent changes in the measured signal, so-called "bouts". The frequency of these bouts predicts the distance of a gas source in wind-tunnel experiments with good accuracy. In addition, they found that the variance of bout counts indicates cross-wind offset to the centerline of the gas plume. The results offer an alternative approach to estimating gas source proximity that is largely independent of gas concentration. The analysis method employed demands very few computational resources, and is suitable for low-power microcontrollers.

While neuromorphic hardware generally delivers on the promise of energy efficiency, the challenge is to develop algorithms that harness the full potential of these chips. A natural choice for inspiration was the neural circuits that perform sensory computation in the brain. Of these, the olfactory system stands out for its capability to rapidly extract information from its extremely high-dimensional input that is chemical space. Based on previous work on olfactory sensory computation, Michael and his colleagues have used the computational architecture of the olfactory system as a template for a neuromorphic pattern recognition. They implemented this algorithm on several accelerated hardware platforms, and assessed recognition performance and power efficiency. Their work pioneers use cases for high-dimensional pattern recognition on neuromorphic architectures. The results highlight the energy-efficiency of neuromorphic hardware, and suggest a use in mobile and embedded platforms.

Rethinking molecular design

Gisbert Schneider concluded the symposium with the award address to which he gave the full title “Rethinking molecular design (…with artificial intelligence)”. Medicinal chemists need to know “what to make next”, but drug designers have to face the challenges of nonlinearity, errors, and incompleteness. The problem is difficult because the scientist is dealing with an adaptive, dynamic organism, but we *can* go from serendipity, narratives, and “gut feeling” to causality–driven engineering.[125] Algorithms for generating, scoring, and optimizing molecular structure are known as *de novo* drug design.[126] “*De novo*”, meaning “from the beginning” or “from scratch” is a misnomer since you cannot begin from scratch: you have to input *something*. But “*novo*” might also be correctly interpreted as “new”, “novel”, or “latest”.

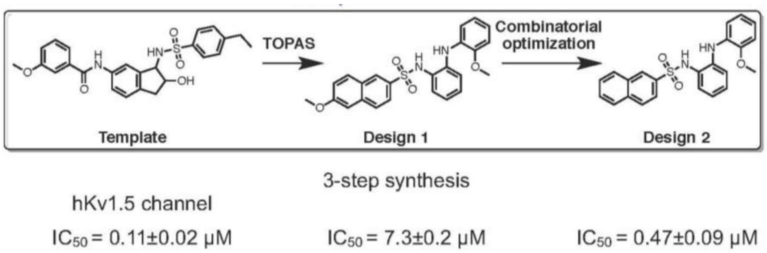
In the midst of the fourth industrial revolution, there is much excitement about the potential of artificial intelligence (AI) to further pharmaceutical research.[127] Essentially, an intelligent agent, human or machine, demonstrates an ability to solve problems, to learn from experience, and to deal with new situations. With regard to these three central criteria, certain machine learning modalities for *de novo* molecular design may be considered instances of AI.

Small-molecule drug discovery can be viewed as a challenging multidimensional problem in which various characteristics of compounds, including efficacy, pharmacokinetics and safety, need to be optimized in parallel to provide drug candidates. Recent advances in areas such as microfluidics-assisted chemical synthesis and biological testing, as well as AI systems that improve a design hypothesis through feedback analysis are now providing a basis for the introduction of greater automation into aspects of this process. This could potentially accelerate time frames for compound discovery and optimization and enable more effective searches of chemical space.[128]

In this context, the umbrella term “constructive learning” describes an entire class of problem–solving techniques, including generative deep networks, for which the ultimate learning goal is not necessarily to find the optimal model for the training data but rather to identify new instances (molecules) from within the applicability domain of the model which are likely to exhibit the desired properties. Several such systems have been designed, and developed to rationalize and articulate next steps in compound selection, synthesis, and testing.

The first automated ligand-based *de novo* design program was the TOPology-Assigning System (TOPAS). An evolutionary algorithm was developed for fragment-based *de novo* design of molecules.[129] This stochastic method aims at generating a novel molecular structure mimicking a template structure. A set of about 25,000 fragment structures serves as the building block supply. The fragments were obtained by a straightforward fragmentation procedure, applied to 36,000 known drugs. A strategy very similar to the Retrosynthetic Combinatorial Analysis Procedure, RECAP, developed by Hann and co-workers[130] was applied. At the time, only 11 reaction schemes were implemented for both fragmentation and building block assembly. This combination of drug-derived building blocks and a restricted set of reaction schemes proved to be a key for the automatic development of novel, synthetically tractable structures. In a cyclic optimization process, molecule architectures were generated from a parent structure by virtual synthesis, and the best structure of a generation was selected as the parent for the subsequent TOPAS cycle. Similarity measures were used to define fitness, based on 2D-structural similarity or topological pharmacophore distance between the template molecule and the variants.

First experimental proof[131] of the TOPAS concept was demonstrated by the successful *de novo* design of a new structural class of potent potassium channel inhibitors. Gisbert and his co-workers selected a known potent potassium channel blocking agent as the template molecule. Two molecules were synthesized based on the original design recommended by TOPAS. Electrophysiological measurement proved potassium channel blocking activity for both.

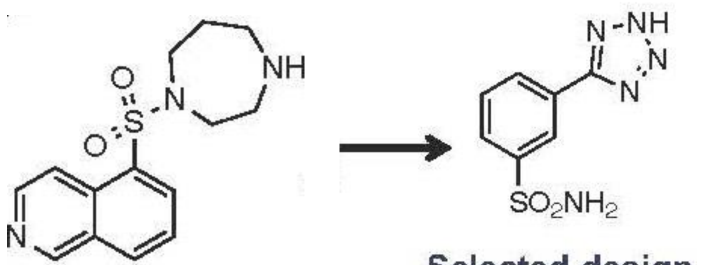


At that time, Gisbert coined the term “scaffold hopping” for the identification of isofunctional molecule structures with significantly different molecular backbones.[132]

Gisbert’s software called Design of Genuine Structures (DOGS)[133,134] features a further advanced ligand-based strategy for automated *in silico* assembly of potentially novel bioactive compounds. The quality of the designed compounds is assessed by a graph kernel method measuring their similarity to known bioactive reference ligands in terms of structural and pharmacophoric features. A deterministic compound construction procedure was implemented that explicitly considers compound synthesizability, based on a compilation of 25,144 readily available synthetic building blocks (from Sigma-Aldrich) and 58 established reaction principles, encoded as SMILES. This enables the software to suggest a synthesis route for each designed compound. DOGS performs ligand growing by reaction forecasting. Pseudo-reaction products are formed from a starting fragment, the most promising pseudo-reaction product is selected and full enumeration then takes place with the selected reaction scheme.

This controls the combinatorial explosion that would result from full enumeration of all 58 reactions with over 25,000 building blocks.

The design of a fragment-like inhibitor of death-associated protein kinase 3 (DAPK3)[135] was one of many successful applications of DOGS. The starting point was the DAPK3 inhibitor Fasudil. After 521 designs and predicted targets, the fourth-ranked selected design was chosen. (The top three looked less innovative.)



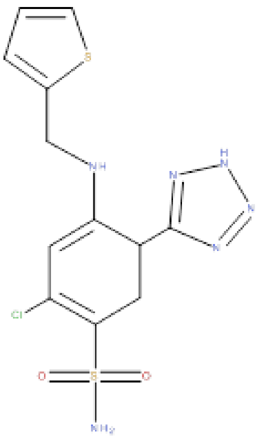
Selected design

Fasudil

IC_{50} = 52 µM, LE = 0.40

DAPK3 K_i = 1.2 µM, *LE* = 0.34

(LE = Ligand efficiency)

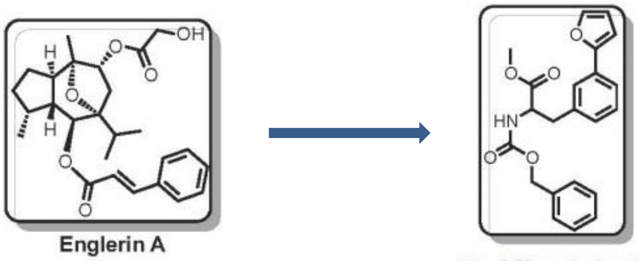


Azosemide

A new crystal structure (PDB 5A6N) of the inactive DAPK3 homodimer showed the fragment-like hit bound to the ATP pocket. Azosemide is an approved diuretic in Japan and it contains the designed structural framework. The researchers acquired a sample, tested it, and found an IC_{50} of 2 µM. Target prediction software based on machine learning models correctly identified additional macromolecular targets of the computationally designed compound, and of azosemide.

In work done in collaboration with Novartis,[136] Gisbert's team has applied ant colony optimization to combinatorial building block selection. By relying on publicly available structure-activity data, the researchers developed a predictive quantitative polypharmacology model for 640 human drug targets. By taking reductive amination as an example of a privileged reaction, they obtained novel subtype-selective and multitarget-modulating serotonin receptor antagonists, as well as ligands selective for the sigma-1 receptor, with accurately predicted affinities. Automated flow synthesis with inline analytics was carried out with reaction chips on the bench. The nanomolar potencies of the hits obtained, their high ligand efficiencies, and an overall success rate of up to 90 % demonstrate that this ligand-based computer-aided molecular design method may guide target-focused combinatorial chemistry. The results of this work with Novartis[136,137] suggest that seamless amalgamation of computational activity prediction and molecular design with microfluidics-assisted synthesis enables the swift generation of small molecules with the desired polypharmacology.

Gisbert and co-workers have also reported the computational *de novo* design of synthetically accessible chemical entities that mimic the complex sesquiterpene natural product (-)-Englerin A.[138] This natural product kills kidney cancer cells, but it is toxic. It also requires a 14-step synthesis,[139] and the target was unknown. Gisbert's team synthesized lead-like probes from commercially available building blocks and profiled them for activity against a computationally predicted panel of macromolecular targets. Both the design template (-)-Englerin A and its low-molecular weight mimetics presented nanomolar binding affinities, and antagonized the transient receptor potential calcium channel (TRPM8) in a cell-based assay, without showing target promiscuity or frequent-hitter properties. (Incidentally, Gisbert credits his wife and co-worker Petra with the actual invention[140] of the term "frequent hitter".) DOGS suggested a three-stage synthesis.



Englerin A

K_B = 0.4 μM

K_B (*S*) = 0.2 μM

Very recently Gisbert's team has reported a method for *de novo* design that uses generative recurrent neural networks (RNN) containing long short-term memory (LSTM) cells.[141,142] This computational model captured the syntax of molecular representation in terms of SMILES strings with close to perfect accuracy. The SMILES strings were from compounds in ChEMBL with nanomolar activity. The learned pattern probabilities can be used for *de novo* SMILES generation by fragment growing. This molecular design concept eliminates the need for virtual compound library enumeration. By employing transfer learning, the researchers fine-tuned the RNN's predictions for specific molecular targets. This approach enables virtual compound design without requiring secondary or external activity prediction, which could introduce error or unwanted bias. The results obtained advocate this generative RNN-LSTM system for high-impact use cases, such as low-data drug discovery, fragment-based molecular design, and hit-to-lead optimization for diverse drug targets.

By transfer learning, the general RNN model was fine-tuned on recognizing retinoid X and peroxisome proliferator-activated receptor (PPAR) agonists.[142] Five top-ranking compounds designed by the generative model were synthesized. Four of the compounds revealed nanomolar to low-micromolar receptor modulatory activity in cell-based assays. Apparently, the computational model intrinsically captured relevant chemical and biological knowledge without the need for explicit rules.

In another example (unpublished work) a first-in-class C-X-C chemokine receptor type 4 (CXCR4) agonist was designed by transfer learning from CXCR modulators in the literature. They were ranked by pharmacophore similarity according to chemically advanced template search (CATS) topological pharmacophores.[143] A one-stage synthesis was proposed an successfully implemented.

Scaffold hopping by automated computational *de novo* design works. *De novo* structure generation and optimization is almost a solved problem: new structures can be generated by explicit or implicit knowledge-based systems. Most of the designs are readily synthesizable. Nevertheless, many more applications are needed to assess the method. Scoring remains inherently difficult: a review on the "edge of chaos" discusses this.[125] Deep learning systems, and generative models, build on implicit medicinal-chemical knowledge. This mix of "mind and machine" will inspire and change drug design.[128]

Conclusion

Erin Davis, chair of the ACS Division of Chemical Information, formally presented the Herman Skolnik Award to Gisbert Schneider at the end of the symposium.

Erin Davis and Gisbert Schneider

References

(1) Clark, T.; Hennemann, M.; Murray, J. S.; Politzer, P. Halogen bonding: the σ-hole. *J. Mol. Model.* **2007,** *13* (2), 291-296.

(2) Riley, K. E.; Murray, J. S.; Politzer, P.; Concha, M. C.; Hobza, P. Br···O Complexes as Probes of Factors Affecting Halogen Bonding: Interactions of Bromobenzenes and Bromopyrimidines with Acetone. *J. Chem. Theory Comput.* **2009,** *5* (1), 155-163.

(3) Wash, P. L.; Ma, S.; Obst, U.; Rebek, J., Jr. Nitrogen-Halogen Intermolecular Forces in Solution. *J. Am. Chem. Soc.* **1999,** *121* (34), 7973-7974.

(4) Metrangolo, P.; Resnati, G. Halogen bonding: a paradigm in supramolecular chemistry. *Chem. - Eur. J.* **2001,** *7* (12), 2511-2519.

(5) Lucassen, A. C. B.; Karton, A.; Leitus, G.; Shimon, L. J. W.; Martin, J. M. L.; Van der Boom, M. E. Co-Crystallization of Sym-Triiodo-Trifluorobenzene with Bipyridyl Donors: Consistent Formation of Two Instead of Anticipated Three N···I Halogen Bonds. *Cryst. Growth Des.* **2007,** *7* (2), 386-392.

(6) Metrangolo, P.; Meyer, F.; Pilati, T.; Resnati, G.; Terraneo, G. Halogen bonding in supramolecular chemistry. *Angew. Chem., Int. Ed.* **2008,** *47* (33), 6114-6127.

(7) Cabot, R.; Hunter, C. A. Non-covalent interactions between iodo-perfluorocarbons and hydrogen bond acceptors. *Chem. Commun.* **2009,** *15*, 2005-2007.

(8) Sarwar, M. G.; Dragisic, B.; Salsberg, L. J.; Gouliaras, C.; Taylor, M. S. Thermodynamics of Halogen Bonding in Solution: Substituent, Structural, and Solvent Effects. *J. Am. Chem. Soc.* **2010,** *132* (5), 1646-1653.

(9) Dumele, O.; Wu, D.; Trapp, N.; Goroff, N.; Diederich, F. Halogen Bonding of (Iodoethynyl)benzene Derivatives in Solution. *Org. Lett.* **2014,** *16* (18), 4722-4725.

(10) Wyler, R.; de Mendoza, J.; Rebek, J., Jr. Formation of a cavity by dimerization of a self-complementary molecule via hydrogen bonds. *Angew. Chem., Int. Ed. Engl.* **1993,** *32* (12), 1699-1701.

(11) Fujita, M.; Nagao, S.; Ogura, K. Guest-Induced Organization of a Three-Dimensional Palladium(II) Cagelike Complex. A Prototype for "Induced-Fit" Molecular Recognition. *J. Am. Chem. Soc.* **1995,** *117* (5), 1649-1650.

(12) Oshovsky, G. V.; Reinhoudt, D. N.; Verboom, W. Triple-Ion Interactions for the Construction of Supramolecular Capsules. *J. Am. Chem. Soc.* **2006,** *128* (15), 5270-5278.

(13) Dumele, O.; Trapp, N.; Diederich, F. Halogen Bonding Molecular Capsules. *Angew. Chem., Int. Ed.* **2015,** *54* (42), 12339-12344.

(14) Dumele, O.; Schreib, B.; Warzok, U.; Trapp, N.; Schalley, C. A.; Diederich, F. Halogen-Bonded Supramolecular Capsules in the Solid State, in Solution, and in the Gas Phase. *Angew. Chem., Int. Ed.* **2017,** *56* (4), 1152-1157.

(15) Livingston, R. C.; Cox, L. R.; Gramlich, V.; Diederich, F. 1,3-Diethynylallenes: new modules for three-dimensional acetylenic scaffolding. *Angew. Chem., Int. Ed.* **2001,** *40* (12), 2334-2337.

(16) Rivera-Fuentes, P.; Nieto-Ortega, B.; Schweizer, W. B.; Lopez Navarrete, J. T.; Casado, J.; Diederich, F. Enantiopure, Monodisperse Alleno-acetylenic Cyclooligomers: Effect of Symmetry and Conformational Flexibility on the Chiroptical Properties of Carbon-Rich Compounds. *Chem. - Eur. J.* **2011,** *17* (14), 3876-3885, S3876/1-S3876/19.

(17) Gropp, C.; Trapp, N.; Diederich, F. Alleno-Acetylenic Cage (AAC) Receptors: Chiroptical Switching and Enantioselective Complexation of trans-1,2-Dimethylcyclohexane in a Diaxial Conformation. *Angew. Chem., Int. Ed.* **2016,** *55* (46), 14444-14449.

(18) Inokuma, Y.; Yoshioka, S.; Ariyoshi, J.; Arai, T.; Hitora, Y.; Takada, K.; Matsunaga, S.; Rissanen, K.; Fujita, M. X-ray analysis on the nanogram to microgram scale using porous complexes. *Nature* **2013,** *495* (7442), 461-466.

(19) Eliel, E. L.; Wilen, S. H.; Mander, L. N. *Stereochemistry of Organic Compounds*; Wiley: New York, NY, 1994.

(20) Gropp, C.; Husch, T.; Trapp, N.; Reiher, M.; Diederich, F. Dispersion and Halogen-Bonding Interactions: Binding of the Axial Conformers of Monohalo- and (±)-trans-1,2-Dihalocyclohexanes in Enantiopure Alleno-Acetylenic Cages. *J. Am. Chem. Soc.* **2017,** *139* (35), 12190-12200.

(21) Gropp, C.; Quigley, B. L.; Diederich, F. Molecular Recognition with Resorcin[4]arene Cavitands: Switching, Halogen-Bonded Capsules, and Enantioselective Complexation. *J. Am. Chem. Soc.* **2018,** *140* (8), 2705-2717.

(22) Romier, C.; Reuter, K.; Suck, D.; Ficner, R. Crystal structure of tRNA-guanine transglycosylase: RNA modification by base exchange. *EMBO J.* **1996,** *15* (11), 2850-2857.

(23) Xie, W.; Liu, X.; Huang, R. H. Chemical trapping and crystal structure of a catalytic tRNA guanine transglycosylase covalent intermediate. *Nat. Struct. Biol.* **2003,** *10* (10), 781-788.

(24) Okada, N.; Noguchi, S.; Kasai, H.; Shindo-Okada, N.; Ohgi, T.; Goto, T.; Nishimura, S. Novel mechanism of post-transcriptional modification of tRNA. Insertion of bases of Q precursors into tRNA by a specific tRNA transglycosylase reaction. *J. Biol. Chem.* **1979,** *254* (8), 3067-3073.

(25) Miles, Z. D.; McCarty, R. M.; Molnar, G.; Bandarian, V. Discovery of epoxyqueuosine (oQ) reductase reveals parallels between halorespiration and tRNA modification. *Proc. Natl. Acad. Sci. U. S. A.* **2011,** *108* (18), 7368-7372, S7368/1-S7368/32.

(26) McCarty, R. M.; Bandarian, V. Biosynthesis of pyrrolopyrimidines. *Bioorg. Chem.* **2012,** *43*, 15-25.

(27) Kohler, P. C.; Ritschel, T.; Schweizer, W. B.; Klebe, G.; Diederich, F. High-Affinity Inhibitors of tRNA-Guanine Transglycosylase Replacing the Function of a Structural Water Cluster. *Chem. - Eur. J.* **2009,** *15* (41), 10809-10817, S10809/1-S10809/18.

(28) Riniker, S.; Barandun, L. J.; Diederich, F.; Kraemer, O.; Steffen, A.; Gunsteren, W. F. Free enthalpies of replacing water molecules in protein binding pockets. *J. Comput.-Aided Mol. Des.* **2012,** *26* (12), 1293-1309.

(29) Movsisyan, L. D.; Schaefer, E.; Nguyen, A.; Ehrmann, F. R.; Schwab, A.; Rossolini, T.; Zimmerli, D.; Wagner, B.; Daff, H.; Heine, A.; Klebe, G.; Diederich, F. Sugar Acetonides are a Superior Motif for Addressing the Large, Solvent-Exposed Ribose-33 Pocket of tRNA-Guanine Transglycosylase. *Chem. - Eur. J.* **2018,** *24* (39), 9957-9967.

(30) Ehrmann, F. R.; Kalim, J.; Pfaffeneder, T.; Bernet, B.; Hohn, C.; Schaefer, E.; Botzanowski, T.; Cianferani, S.; Heine, A.; Reuter, K.; Diederich, F.; Klebe, G. Swapping Interface Contacts for inhibitors in the Homodimeric tRNA-Guanine Transglycosylase: An Option for Functional Regulation. *Angew. Chem., Int. Ed.* **2018,** *57* (32), 10085-10090.

(31) Hann, M. M. Molecular obesity, potency and other addictions in drug discovery. *MedChemComm* **2011,** *2* (5), 349-355.

(32) Jorgensen, W. L.; Tirado-Rives, J. Molecular modeling of organic and biomolecular systems using BOSS and MCPRO. *J. Comput. Chem.* **2005,** *26* (16), 1689-1700.

(33) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996,** *118* (45), 11225-11236.

(34) Robertson, M. J.; Tirado-Rives, J.; Jorgensen, W. L. Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *J. Chem. Theory Comput.* **2015,** *11* (7), 3499-3509.

(35) Vilseck, J. Z.; Tirado-Rives, J.; Jorgensen, W. L. Determination of partial molar volumes from free energy perturbation theory. *Phys. Chem. Chem. Phys.* **2015,** *17* (13), 8407-8415.

(36) Dodda, L. S.; Cabeza de Vaca, I.; Tirado-Rives, J.; Jorgensen, W. L. LigParGen web server: an automatic OPLS-AA parameter generator for organic ligands. *Nucleic Acids Res.* **2017,** *45* (W1), W331-W336.

(37) Dodda, L. S.; Vilseck, J. Z.; Tirado-Rives, J.; Jorgensen, W. L. 1.14*CM1A-LBCC: Localized Bond-Charge Corrected CM1A Charges for Condensed-Phase Simulations. *J. Phys. Chem. B* **2017,** *121* (15), 3864-3870.

(38) Yan, X. C.; Robertson, M. J.; Tirado-Rives, J.; Jorgensen, W. L. Improved Description of Sulfur Charge Anisotropy in OPLS Force Fields: Model Development and Parameterization. *J. Phys. Chem. B* **2017,** *121* (27), 6626-6636.

(39) Greven, D.; Leng, L.; Bucala, R. Autoimmune diseases: MIF as a therapeutic target. *Expert Opin. Ther. Targets* **2010,** *14* (3), 253-264.

(40) Trivedi-Parmar, V.; Jorgensen, W. L. Advances and Insights for Small Molecule Inhibition of Macrophage Migration Inhibitory Factor. *J. Med. Chem.* **2018**, Ahead of Print.

(41) Leyton-Jaimes, M. F.; Kahn, J.; Israelson, A. Macrophage migration inhibitory factor: A multifaceted cytokine implicated in multiple neurological diseases. *Exp. Neurol.* **2018,** *301* (Part B), 83-91.

(42) Dziedzic, P.; Cisneros, J. A.; Robertson, M. J.; Hare, A. A.; Danford, N. E.; Baxter, R. H. G.; Jorgensen, W. L. Design, synthesis, and protein crystallography of biaryltriazoles as potent tautomerase inhibitors of macrophage migration inhibitory factor. *J. Am. Chem. Soc.* **2015,** *137* (8), 2996-3003.

(43) Cisneros, J. A.; Robertson, M. J.; Valhondo, M.; Jorgensen, W. L. A Fluorescence Polarization Assay for Binding to Macrophage Migration Inhibitory Factor and Crystal Structures for Complexes of Two Potent Inhibitors. *J. Am. Chem. Soc.* **2016,** *138* (27), 8630-8638.

(44) Newton, A. S.; Deiana, L.; Puleo, D. E.; Cisneros, J. A.; Cutrona, K. J.; Schlessinger, J.; Jorgensen, W. L. JAK2 JH2 Fluorescence Polarization Assay and Crystal Structures for Complexes with Three Small Molecules. *ACS Med. Chem. Lett.* **2017,** *8* (6), 614-617.

(45) Dawson, T. K.; Dziedzic, P.; Robertson, M. J.; Cisneros, J. A.; Krimmer, S. G.; Newton, A. S.; Tirado-Rives, J.; Jorgensen, W. L. Adding a Hydrogen Bond May Not Help: Naphthyridinone vs. Quinoline Inhibitors of Macrophage Migration Inhibitory Factor. *ACS Med. Chem. Lett.* **2017,** *8* (12), 1287-1291.

(46) Trivedi-Parmar, V.; Robertson, M. J.; Cisneros, J. A.; Krimmer, S. G.; Jorgensen, W. L. Optimization of Pyrazoles as Phenol Surrogates to Yield Potent Inhibitors of Macrophage Migration Inhibitory Factor. *ChemMedChem* **2018,** *13* (11), 1092-1097.

(47) Cisneros, J. A.; Robertson, M. J.; Mercado, B. Q.; Jorgensen, W. L. Systematic study of effects of structural modifications on the aqueous solubility of drug-like molecules. *ACS Med. Chem. Lett.* **2017,** *8* (1), 124-127.

(48) Piette, C.; Deprez, M.; Roger, T.; Noel, A.; Foidart, J.-M.; Munaut, C. The Dexamethasone-induced Inhibition of Proliferation, Migration, and Invasion in Glioma Cell Lines Is Antagonized by Macrophage Migration Inhibitory Factor (MIF) and Can Be Enhanced by Specific MIF Inhibitors. *J. Biol. Chem.* **2009,** *284* (47), 32483-32492.

(49) Jorgensen, W. L.; Thomas, L. L. Perspective on Free-Energy Perturbation Calculations for Chemical Equilibria. *J. Chem. Theory Comput.* **2008,** *4* (6), 869-876.

(50) Jorgensen, W. L. Efficient Drug Lead Discovery and Optimization. *Acc. Chem. Res.* **2009,** *42* (6), 724-733.

(51) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015,** *137* (7), 2695-2703.

(52) Jorgensen, W. L. Computer-aided discovery of anti-HIV agents. *Bioorg. Med. Chem.* **2016,** *24* (20), 4768-4778.

(53) Bollini, M.; Frey, K. M.; Cisneros, J. A.; Spasov, K. A.; Das, K.; Bauman, J. D.; Arnold, E.; Anderson, K. S.; Jorgensen, W. L. Extension into the entrance channel of HIV-1 reverse transcriptase-Crystallography and enhanced solubility. *Bioorg. Med. Chem. Lett.* **2013,** *23* (18), 5209-5212.

(54) Jorgensen, W. L.; Ruiz-Caro, J.; Tirado-Rives, J.; Basavapathruni, A.; Anderson, K. S.; Hamilton, A. D. Computer-aided design of non-nucleoside inhibitors of HIV-1 reverse transcriptase. *Bioorg. Med. Chem. Lett.* **2006,** *16* (3), 663-667.

(55) Ruiz-Caro, J.; Basavapathruni, A.; Kim, J. T.; Bailey, C. M.; Wang, L.; Anderson, K. S.; Hamilton, A. D.; Jorgensen, W. L. Optimization of diarylamines as non-nucleoside inhibitors of HIV-1 reverse transcriptase. *Bioorg. Med. Chem. Lett.* **2006,** *16* (3), 668-671.

(56) Lee, W.-G.; Gallardo-Macias, R.; Frey, K. M.; Spasov, K. A.; Bollini, M.; Anderson, K. S.; Jorgensen, W. L. Picomolar Inhibitors of HIV Reverse Transcriptase Featuring Bicyclic Replacement of a Cyanovinylphenyl Group. *J. Am. Chem. Soc.* **2013,** *135* (44), 16705-16713.

(57) Gray, W. T.; Frey, K. M.; Laskey, S. B.; Mislak, A. C.; Spasov, K. A.; Lee, W.-G.; Bollini, M.; Siliciano, R. F.; Jorgensen, W. L.; Anderson, K. S. Potent Inhibitors Active against HIV Reverse Transcriptase with K101P, a Mutation Conferring Rilpivirine Resistance. *ACS Med. Chem. Lett.* **2015,** *6* (10), 1075-1079.

(58) Kudalkar, S. N.; Beloor, J.; Chan, A. H.; Lee, W.-G.; Jorgensen, W. L.; Kumar, P.; Anderson, K. S. Structural and preclinical studies of computationally designed non-nucleoside reverse transcriptase inhibitors for treating HIV infection. *Mol. Pharmacol.* **2017,** *91* (4), 383-391.
(59) Chan, A. H.; Lee, W.-G.; Spasov, K. A.; Cisneros, J. A.; Kudalkar, S. N.; Petrova, Z. O.; Buckingham, A. B.; Anderson, K. S.; Jorgensen, W. L. Covalent inhibitors for eradication of drug-resistant HIV-1 reverse transcriptase: From design to protein crystallography. *Proc. Natl. Acad. Sci. U. S. A.* **2017,** *114* (36), 9725-9730.
(60) Winkler, D. A.; Thornton, A.; Farjot, G.; Katz, I. The diverse biological properties of the chemically inert noble gases. *Pharmacol. Ther.* **2016,** *160*, 44-64.
(61) Banks, P.; Franks, N. P.; Dickinson, R. Competitive inhibition at the glycine site of the N-methyl-D-aspartate receptor mediates xenon neuroprotection against hypoxia-ischemia. *Anesthesiology* **2010,** *112* (3), 614-22.
(62) Vengeliene, V.; Bessiere, B.; Pype, J.; Spanagel, R. The Effects of Xenon and Nitrous Oxide Gases on Alcohol Relapse. *Alcohol.: Clin. Exp. Res.* **2014,** *38* (2), 557-563.
(63) Liu, L. T.; Xu, Y.; Tang, P. Mechanistic Insights into Xenon Inhibition of NMDA Receptors from MD Simulations. *J. Phys. Chem. B* **2010,** *114* (27), 9010-9016.
(64) Winkler, D. A.; Katz, I.; Farjot, G.; Warden, A. C.; Thornton, A. W. Decoding the Rich Biological Properties of Noble Gases: How Well Can We Predict Noble Gas Binding to Diverse Proteins? *ChemMedChem* **2018**, Ahead of Print.
(65) Britton, G. L.; Kim, H.; Kee, P. H.; Aronowski, J.; Holland, C. K.; McPherson, D. D.; Huang, S.-L. In Vivo Therapeutic Gas Delivery for Neuroprotection With Echogenic Liposomes. *Circulation* **2010,** *122* (16), 1578-1587.
(66) 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.
(67) Brown, J. B.; Niijima, S.; Okuno, Y. Compound-Protein Interaction Prediction Within Chemogenomics: Theoretical Concepts, Practical Usage, and Future Directions. *Mol. Inf.* **2013,** *32* (11-12), 906-921.
(68) Hecht-Nielsen, R. Counterpropagation networks. *Appl. Opt.* **1987,** *26* (23), 4979-83.
(69) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL database in 2017. *Nucleic Acids Res.* **2016,** *45* (D1), D945-D954.
(70) Kimura, T.; Miyashita, Y.; Funatsu, K.; Sasaki, S.-i. Quantitative Structure-Activity Relationships of the Synthetic Substrates for Elastase Enzyme Using Nonlinear Partial Least Squares Regression. *J. Chem. Inf. Comput. Sci.* **1996,** *36* (2), 185-189.
(71) Fredriksson, R.; Lagerstrom, M. C.; Lundin, L.-G.; Schioth, H. B. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol.* **2003,** *63* (6), 1256-1272.
(72) Isberg, V.; Mordalski, S.; Munk, C.; Rataj, K.; Harpsoee, K.; Hauser, A. S.; Vroling, B.; Bojarski, A. J.; Vriend, G.; Gloriam, D. E. GPCRdb: an information system for G protein-coupled receptors. *Nucleic Acids Res.* **2016,** *44* (D1), D356-D364.
(73) Mishima, K.; Kaneko, H.; Funatsu, K. Development of a New De Novo Design Algorithm for Exploring Chemical Space. *Mol. Inf.* **2014,** *33* (11-12), 779-789.
(74) Stojkovic, N.; Cekic, S.; Ristov, M.; Ristic, M.; Dukic, D.; Binic, M.; Virijevic, D. Histamine and Antihistamines / Histamin i antihistamini. *Acta Fac. Med. Naissensis* **2015,** *32* (1), 7-22.

(75) Shimamura, T.; Shiroishi, M.; Weyand, S.; Tsujimoto, H.; Winter, G.; Katritch, V.; Abagyan, R.; Cherezov, V.; Liu, W.; Han, G. W.; Kobayashi, T.; Stevens, R. C.; Iwata, S. Structure of the human histamine H1 receptor complex with doxepin. *Nature (London, U. K.)* **2011,** *475* (7354), 65-70.

(76) Church, M. K.; Maurer, M.; Simons, F. E. R.; Bindslev-Jensen, C.; van Cauwenberge, P.; Bousquet, J.; Holgate, S. T.; Zuberbier, T. Risk of first-generation H1-antihistamines: a GA2LEN position paper. *Allergy (Oxford, U. K.)* **2010,** *65* (4), 459-466.

(77) Maeda, I.; Hasegawa, K.; Kaneko, H.; Funatsu, K. Novel Method Proposing Chemical Structures with Desirable Profile of Activities Based on Chemical and Protein Spaces. *Mol. Inf.* **2017,** *36* (12), 1700075.

(78) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock and AutoDockTools: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009,** *30* (16), 2785-2791.

(79) Sanner, M. F. Python: a programming language for software integration and development. *J. Mol. Graphics Modell.* **1999,** *17* (1), 57-61.

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

(81) 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.

(82) 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.

(83) 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.

(84) Bishop, C. M.; Svensén, M.; Williams, C. K. I. GTM: The Generative Topographic Mapping. *Neural Comput.* **1998,** *10* (1), 215-234.

(85) Bishop, C. M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, 2006.

(86) 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.

(87) Klimenko, K.; Marcou, G.; Horvath, D.; Varnek, A. Chemical Space Mapping and Structure-Activity Analysis of the ChEMBL Antiviral Compound Set. *J. Chem. Inf. Model.* **2016,** *56* (8), 1438-1454.

(88) Sidorov, P.; Gaspar, H.; Marcou, G.; Varnek, A.; Horvath, D. Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds. *J. Comput.-Aided Mol. Des.* **2015,** *29* (12), 1087-1108.

(89) Kayastha, S.; Horvath, D.; Gilberg, E.; Guetschow, M.; Bajorath, J.; Varnek, A. Privileged Structural Motif Detection and Analysis Using Generative Topographic Maps. *J. Chem. Inf. Model.* **2017,** *57* (5), 1218-1232.

(90) Hoffer, L.; Horvath, D. S4MPLE - Sampler For Multiple Protein-Ligand Entities: Simultaneous Docking of Several Entities. *J. Chem. Inf. Model.* **2013,** *53* (1), 88-102.

(91) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discovery* **2010,** *9* (3), 203-214.

(92) Bowes, J.; Brown, A. J.; Hamon, J.; Jarolimek, W.; Sridhar, A.; Waldron, G.; Whitebread, S. Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nat. Rev. Drug Discovery* **2012,** *11* (12), 909-922.

(93) Baringhaus, K.-H.; Hessler, G.; Matter, H.; Schmidt, F. Development and applications of global ADMET models: in silico prediction of human microsomal lability. In *Chemoinformatics for Drug Discovery;* Bajorath, J., Ed.; Wiley-VCH: Weinheim, 2014; pp 245-265.

(94) Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model.* **2017,** *57* (8), 2068-2076.

(95) Vidal, D.; Mestres, J. In Silico Receptorome Screening of Antipsychotic Drugs. *Mol. Inf.* **2010,** *29* (6-7), 543-551.

(96) Buchanan, B. G.; Smith, D. H.; White, W. C.; Gritter, R. J.; Feigenbaum, E. A.; Lederberg, J.; Djerassi, C. Applications of artificial intelligence for chemical inference. 22. Automatic rule formation in mass spectrometry by means of the meta-DENDRAL program. *J. Am. Chem. Soc.* **1976,** *98* (20), 6168-6178.

(97) King, R. D.; Rowland, J.; Oliver, S. G.; Young, M.; Aubrey, W.; Byrne, E.; Liakata, M.; Markham, M.; Pir, P.; Soldatova, L. N.; Sparkes, A.; Whelan, K. E.; Clare, A. The Automation of Science. *Science* **2009,** *324* (5923), 85-89.

(98) Williams, K.; Sparkes, A.; Aubrey, W.; Bilsland, E.; Young, M.; Soldatova, L. N.; De, G. K.; Ramon, J.; de, C. M.; Oliver, S. G.; Sirawaraporn, W.; King, R. D. Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. *J. R. Soc. Interface* **2015,** *12* (104), 20141289.

(99) Soldatova, L. N.; Clare, A.; Sparkes, A.; King, R. D. An ontology for a robot scientist. *Bioinformatics* **2006,** *22* (14), e464-e471.

(100) Soldatova, L. N.; King, R. D. An ontology of scientific experiments. *J R Soc Interface* **2006,** *3* (11), 795-803.

(101) Faulon, J.-L.; Misra, M.; Martin, S.; Sale, K.; Sapra, R. Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor. *Bioinformatics* **2008,** *24* (2), 225-233.

(102) Jacob, L.; Vert, J.-P. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* **2008,** *24* (19), 2149-2156.

(103) Nagamine, N.; Sakakibara, Y. Statistical prediction of protein-chemical interactions based on chemical structure and mass spectrometry data. *Bioinformatics* **2007,** *23* (15), 2004-2012.

(104) Yamanishi, Y.; Araki, M.; Gutteridge, A.; Honda, W.; Kanehisa, M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* **2008,** *24* (13), i232-i240.

(105) Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L. H.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting new molecular targets for known drugs. *Nature (London, U. K.)* **2009,** *462* (7270), 175-181.

(106) Yabuuchi, H.; Niijima, S.; Takematsu, H.; Ida, T.; Hirokawa, T.; Hara, T.; Ogawa, T.; Minowa, Y.; Tsujimoto, G.; Okuno, Y. Analysis of multiple compound-protein interactions reveals novel bioactive molecules. *Mol. Syst. Biol.* **2011,** *7*, 472.

(107) Hizukuri, Y.; Sawada, R.; Yamanishi, Y. Predicting target proteins for drug candidate compounds based on drug induced gene expression data in a chemical structure-independent manner. *BMC Med. Genomics* **2015,** *8*, 82/1-82/10.

(108) Takarabe, M.; Kotera, M.; Nishimura, Y.; Goto, S.; Yamanishi, Y. Drug target prediction using adverse event report systems: a pharmacogenomic approach. *Bioinformatics* **2012,** *28* (18), i611-i618.

(109) Yamanishi, Y.; Kotera, M.; Moriya, Y.; Sawada, R.; Kanehisa, M.; Goto, S. DINIES: drug-target interaction network inference engine based on supervised analysis. *Nucleic Acids Res.* **2014,** *42* (W1), W39-W45.

(110) Yamanishi, Y. Supervised bipartite graph inference. In *Proceedings of the Conference on Advances in Neural Information and Processing System 21;* Koller, D., Schuurmans, D., Bengio, Y., Bottou, L., Eds.; MIT Press: Cambridge, MA, 2009; pp 1433–1440.

(111) Iskar, M.; Zeller, G.; Blattmann, P.; Campillos, M.; Kuhn, M.; Kaminska, K. H.; Runz, H.; Gavin, A.-C.; Pepperkok, R.; van Noort, V.; Bork, P. Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. *Mol. Syst. Biol.* **2013,** *9*, 662.

(112) Wang, K.; Sun, J.; Zhou, S.; Wan, C.; Qin, S.; Li, C.; He, L.; Yang, L. Prediction of drug-target interactions for drug repositioning only based on genomic expression similarity. *PLoS Comput. Biol.* **2013,** *9* (11), e1003315/1-e1003315/9, 9 pp.

(113) Iwata, M.; Sawada, R.; Iwata, H.; Kotera, M.; Yamanishi, Y. Elucidating the modes of action for bioactive compounds in a cell-specific manner by large-scale chemically-induced transcriptomics. *Sci. Rep.* **2017,** *7*, 40164.

(114) Sawada, R.; Iwata, M.; Yamanishi, Y.; Umezaki, M.; Usui, Y.; Kobayashi, T.; Kubono, T.; Hayashi, S.; Kadowaki, M.; Yamanishi, Y. KampoDB, database of predicted targets and functional annotations of natural medicines. *Sci. Rep.* **2018,** *8* (1), 11216.

(115) Schmuker, M.; Madany Mamlouk, A.; Pearce, T. C. Proceedings of the 1st International Workshop on Odor Spaces. *Flavour* **2013,** *3* (Suppl. 1), 1.

(116) Kasap, B.; Schmuker, M. Improving odor classification through self-organized lateral inhibition in a spiking olfaction-inspired network. In *Proceedings of the 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*; IEEE: Piscataway, NJ, 2013; pp 219–222.

(117) Yamagata, N.; Schmuker, M.; Szyszka, P.; Mizunami, M.; Menzel, R. Differential odor processing in two olfactory pathways in the honeybee. *Front. Syst. Neurosci.* **2009,** *3* (Dec.), No pp. given.

(118) Schmuker, M.; Yamagata, N.; Nawrot, M.; Menzel, R. Parallel Representation of Stimulus Identity and Intensity in a Dual Pathway Model Inspired by the Olfactory System of the Honeybee. *Front. Neuroeng.* **2011,** *4* (17), 1-13.

(119) Schmuker, M.; Schneider, G. Processing and classification of chemical data inspired by insect olfaction. *Proc. Natl. Acad. Sci. U. S. A.* **2007,** *104* (51), 20285-20289.

(120) Pfeil, T.; Grübl, A.; Jeltsch, S.; Müller, E.; Müller, P.; Petrovici, M. A.; Schmuker, M.; Brüderle, D.; Schemmel, J.; Meier, K. Six Networks on a Universal Neuromorphic Computing Substrate. *Front. Neurosci.* **2013,** *7* (11), 1-17.

(121) Schmuker, M.; Pfeil, T.; Nawrot, M. P. A neuromorphic network for generic multivariate data classification. *Proc. Natl. Acad. Sci. U. S. A.* **2014,** *111* (6), 2081-2086.

(122) Diamond, A.; Nowotny, T.; Schmuker, M. Comparing Neuromorphic Solutions in Action: Implementing a Bio-Inspired Solution to a Benchmark Classification Task on Three Parallel-Computing Platforms. *Front. Neurosci.* **2016,** *9* (491).

(123) Vergara, A.; Fonollosa, J.; Mahiques, J.; Trincavelli, M.; Rulkov, N.; Huerta, R. On the performance of gas sensor arrays in open sampling systems using Inhibitory Support Vector Machines. *Sens. Actuators, B* **2013,** *185*, 462-477.

(124) Schmuker, M.; Bahr, V.; Huerta, R. Exploiting plume structure to decode gas source distance using metal-oxide gas sensors. *Sens. Actuators, B* **2016,** *235*, 636-646.

(125) Schneider, P.; Schneider, G. De Novo Design at the Edge of Chaos. *J. Med. Chem.* **2016,** *59* (9), 4077-4086.

(126) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery* **2005,** *4* (8), 649-663.

(127) Schneider, G. Generative Models for Artificially-intelligent Molecular Design. *Mol. Inf.* **2018,** *37* (1-2), 1880131.

(128) Schneider, G. Automating drug discovery. *Nat. Rev. Drug Discovery* **2018,** *17* (2), 97-113.

(129) Schneider, G.; Lee, M.-L.; Stahl, M.; Schneider, P. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J. Comput.-Aided Mol. Des.* **2000,** *14* (5), 487-494.

(130) Lewell, X. Q.; Judd, D.; Watson, S.; Hann, M. RECAP-Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998,** *38* (3), 511-522.

(131) Schneider, G.; Clement-Chomienne, O.; Hilfiger, L.; Schneider, P.; Kirsch, S.; Bohm, H.-J.; Neidhart, W. Virtual screening for bioactive molecules by evolutionary De novo design. *Angew. Chem., Int. Ed.* **2000,** *39* (22), 4130-4133.

(132) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem., Int. Ed.* **1999,** *38* (19), 2894-2896.

(133) 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.

(134) Schneider, G. De novo design - hop(p)ing against hope. *Drug Discovery Today Technol.* **2013,** *10* (4), e453-60.

(135) Rodrigues, T.; Reker, D.; Welin, M.; Caldera, M.; Brunner, C.; Gabernet, G.; Schneider, P.; Walse, B.; Schneider, G. De Novo Fragment Design for Drug Discovery and Chemical Biology. *Angew. Chem., Int. Ed.* **2015,** *54* (50), 15079-15083.

(136) 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.

(137) 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.

(138) Friedrich, L.; Rodrigues, T.; Neuhaus, C. S.; Schneider, P.; Schneider, G. From Complex Natural Products to Simple Synthetic Mimetics by Computational De Novo Design. *Angew. Chem., Int. Ed.* **2016,** *55* (23), 6789-6792.

(139) Takahashi, K.; Komine, K.; Yokoi, Y.; Ishihara, J.; Hatakeyama, S. Stereocontrolled Total Synthesis of (-)-Englerin A. *J. Org. Chem.* **2012,** *77* (17), 7364-7370.

(140) Roche, O.; Schneider, P.; Zuegge, J.; Guba, W.; Kansy, M.; Alanine, A.; Bleicher, K.; Danel, F.; Gutknecht, E.-M.; Rogers-Evans, M.; Neidhart, W.; Stalder, H.; Dillon, M.; Sjoegren, E.; Fotouhi, N.; Gillespie, P.; Goodnow, R.; Harris, W.; Jones, P.; Taniguchi, M.; Tsujii, S.; von Saal, W.; Zimmermann, G.; Schneider, G. Development of a Virtual Screening Method for Identification of "Frequent Hitters" in Compound Libraries. *J. Med. Chem.* **2002,** *45* (1), 137-142.

(141) Gupta, A.; Müller, A. T.; Huisman, B. J. H.; Fuchs, J. A.; Schneider, P.; Schneider, G. Generative Recurrent Networks for De Novo Drug Design. *Mol. Inf.* **2018,** *37* (1-2), 1700111.

(142) Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Mol. Inf.* **2018,** *37* (1-2), 1700153.

(143) 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.