Herman Skolnik Award Symposium 2017

Honoring David Winkler

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

Introduction

David Winkler, CSIRO Fellow, and professor at Latrobe Institute for Molecular Science, and Monash Institute of Pharmaceutical Sciences, Melbourne, Australia, received the 2017 Herman Skolnik Award for his seminal contributions to chemical information in the development of optimally sparse, robust machine learning methods for QSAR, and in leading the application of cheminformatics methods to biomaterials, nanomaterials, and regenerative medicine. A summary of his achievements has been published in the *Chemical Information Bulletin.* David was invited to present an award symposium at the Fall 2017 ACS National Meeting in Washington, DC. He invited six speakers:

L to R: Alex Tropsha, Johnny Gasteiger, Yoram Cohen; Tim Clark, David Winkler, Ceyda Oksel, Tudor Oprea

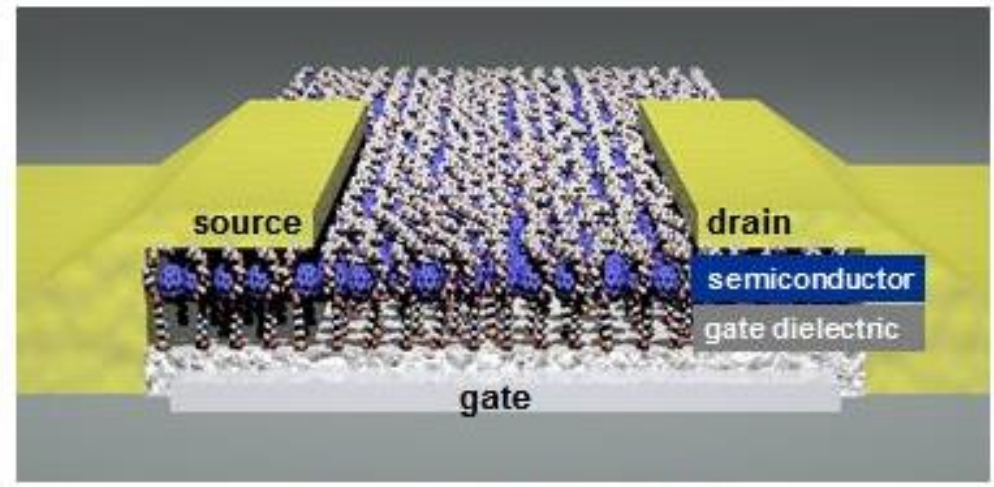
Approaching reality: simulating electronic devices

Tim Clark, of the University of Erlangen-Nürnberg, was the first speaker. The impact of modern hardware and software on simulations has not been an issue of doing things faster and faster, but rather one of doing calculations that we could not do before. *Ab initio* calculations can now be done on compounds with several hundred atoms, density functional theory calculations on a few thousand atoms, and semiempirical molecular orbital (MO) calculations on 100,000 atoms. Simulations of several microseconds are now standard.

Semiempirical (neglect of diatomic differential overlap, NDDO) molecular orbital (MO) calculations without local approximations are now possible for 100,000 atoms or more with the massively parallel semiEMPIRical molEcular-Orbital Program (EMPIRE) program,[1-3] which is freely available to academic groups. Calculation scales with approximately $N^{2.5}$. We are no longer limited to small or homogeneous, perfect systems, but can now include defects, dopants, impurities or domain boundaries in the calculations, or even calculate amorphous systems.

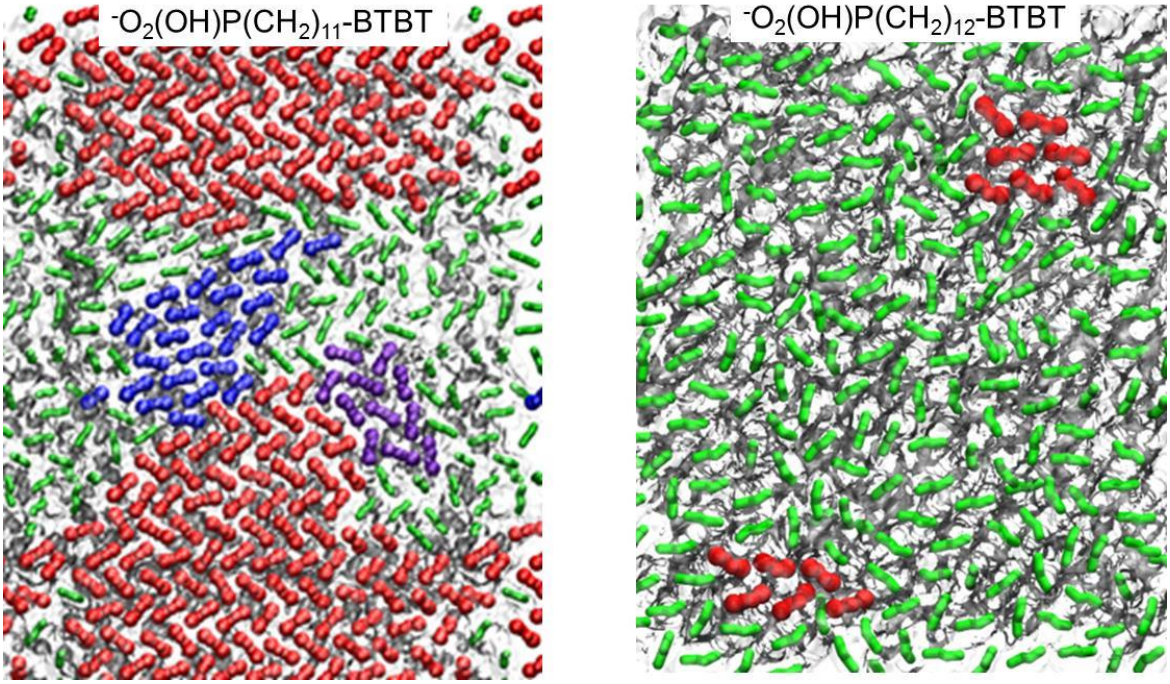
The results of such calculations can be used to simulate charge-transport through disordered monolayers. Clark's team has studied self-assembled monolayer field-effect transistors (SAMFETs) handling conformational freedom using classical atomistic molecular-dynamics (MD) simulations, electronic properties using very large scale semiempirical MO theory, and conductance by propagating single electrons or using diffusion quantum Monte-Carlo (DQMC) charge-transport simulations.[4-9]

The molecules that comprise the SAM contain insulating and semiconducting moieties, so that they serve as both gate dielectric and the active transistor channel in a device:



Tim's team has used simulations to describe and optimize complex systems of self-assembled monolayers on surfaces, not only to explain their morphology but also to predict molecular compositions and arrangements favorable for improved charge transport.[7] In more recent work,[10] they have constructed transistors based on SAMs of two molecules that consist of the organic p-type

semiconductor benzothieno[3,2-b][1]benzothiophene (BTBT), linked to a C_{11} or C_{12} alkylphosphonic acid. Both molecules form ordered SAMs, but the experiments show that the size of the crystalline domains and the charge-transport properties vary considerably in the two systems. Because of the angle of the head groups one can form crystalline domains and the other cannot. This can be reproduced with simple force field calculations.



The procedure for charge transfer simulations is as follows:

- Calculate the neutral system and use local properties as external potentials:
  - Local electron affinity[11,12] for electrons, local ionization energy[13] for holes
  - Cluster model or periodic-boundary conditions
- Monte-Carlo search for conductance paths
- DQMC simulations[14] for many electrons
- Propagate single charge carriers on these potentials to determine time scales.

Tim showed an MD simulation of the charge transport paths. For the transport calculations, the team employed a fully quantum mechanical description, namely Landauer transport theory.[9] In accord with experiment, they found an improved charge transport across BTBT-C_{11}-PA SAMs compared to BTBT-C_{12}-PA SAMs.

DQMC reproduces voltage/current curves (assuming that the number of Monte Carlo steps correlates with time) and reproduces experimentally observed hysteresis. It also revealed dimeric fullerene electron traps.[15] Density functional theory calculations indicate that van der Waals fullerene oligomers can form interstitial electron traps in which the electrons are even more strongly bound than in isolated fullerene radical anions. Spectroelectrochemical measurements on a bis-fullerene-substituted peptide

provide experimental support. The proposed deep electron traps are relevant for all organic electronics applications in which non-covalently linked fullerenes in van der Waals contact with one another serve as n-type semiconductors.

Finally Tim showed the results of simulations of hole-transport through a self-assembled monolayer substituted with a p-type organic semiconductor and with crystalline domains (see the work above on BTBT linked to a C_{11} or C_{12} alkylphosphonic acid). He illustrated hole transport through the monolayers. Hysteresis is not observed in this case. Tim also illustrated well-defined paths through the crystalline domains of the $O_2(OH)P(CH_2)_{11}$-BTBT material. The researchers have shown that structural order is particularly important for the electronic properties of semiconducting self-assembled monolayers, and they predict that semiconducting SAMs with a higher degree of crystallinity and larger crystalline regions will exhibit superior performance.

Applications of machine learning to materials and chemical property prediction

Alex Tropsha, of the University of North Carolina Chapel Hill, UNC Eshelman School of Pharmacy, is benefiting from the explosive growth of materials data. There are 160,000 entries in the Inorganic Crystal Structure Database (ICSD). There are numerous commercial and open experimental databases (NIST, MatWeb, MatBase etc.), and huge databases such as AFLOWLIB, Materials Project, and Harvard Clean Energy. The chemical space of possible materials is *huge* : about 10^{100} candidates.[16] The US government's Materials Genome Initiative recognizes the need for new high performance materials. The growth of materials databases and emerging informatics approaches offers the opportunity to transform materials discovery into data- and knowledge-driven rational design.

AFLOW is a globally available database of 1,688,245 material compounds, with over 167,136,255 calculated properties. The optimized geometries, symmetries, band structures, and densities of states available in the AFLOWLIB consortium databases have been converted into two distinct types of fingerprints: Band structure fingerprints (B- fingerprints), and Density of States fingerprints (D-fingerprints).[17] The framework is employed to query large databases of materials using similarity concepts, to map the connectivity of materials space (as a materials cartogram) for rapidly identifying regions with unique organizations and properties, and to develop predictive quantitative materials structure−property relationship (QMSPR) models for guiding materials design.

To represent the library of materials as a network (a material cartogram), the researchers considered each material, encoded by its fingerprint, as a node. Edges exist between nodes with similarities above certain thresholds (in this case, Tanimoto similarity and a threshold of 0.7). A materials map from B-fingerprints was made from 15,000 materials from ICSD, using DFT PBE calculations from AFLOWLIB. Four big clusters were observed: insulators, ceramics, and complex oxides; bimetals and polymetals; metallic and nonmetallic combinations; and small band gap semiconductors.

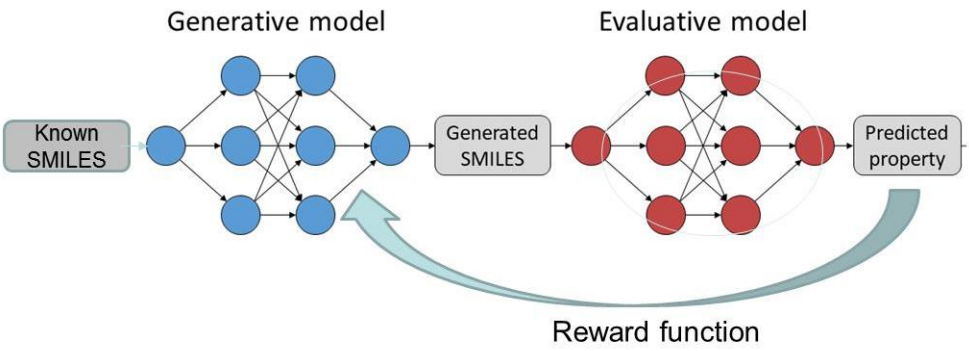
Novel descriptors (property-labeled materials fragments) not requiring prior DFT calculations have also been developed by Voronoi tessellation and neighbors search of crystal structures, followed by infinite periodic graph construction and property labeling, and generation of circular fingerprints..[18] Starting from only a crystal structure, regression models can be built to predict band gap energy, and thus electronic properties, or to predict thermo-mechanical properties such as bulk modulus, shear modulus, thermal expansion, heat capacity, and thermal conductivity. All the models are trained based on DFT-computed properties. Heuristic design rules can be extracted.

Material informatics has also been applied to design of a novel photocathode material for dye-sensitized solar cells (DSSCs).[19] By conducting a virtual screening of 50,000 known inorganic compounds, the researchers have identified lead titanate ($PbTiO_3$), as the most promising photocathode material. Notably, lead titanate is significantly different from the traditional base elements or crystal structures used for photocathodes. In experimental validation, the fabricated lead titanate DSSC devices exhibited the best performance in aqueous solution, showing remarkably high fill factors compared to typical photocathode systems. Currently, device performance is low, but it might be improved by designing a new dye.

Next Alex discussed applications of machine learning to designing chemicals with the desired physical and biological properties where compound structure is described only by its SMILES notation, and no other conventional chemical descriptors are used. The new approach developed in his lab is based on concepts from text mining that rely on neural networks to solve the problem of semantic similarity of texts.

The British linguist J. R. Firth is noted for drawing attention to the context-dependent nature of meaning. In particular, he is known for the 1957 quotation: “You shall know a word by the company it keeps”. To define the semantic similarity between two entities, Alex and his colleagues have made use of approaches embedded in Word2Vec, a neural network based approach to describe linguistic context of words developed at Google.[20] With Word2Vec, a network is trained using each word of a corpus of text and some configurable number of surrounding words. The model can be trained to either predict the surrounding context based on the current word, or to predict the current word from the context. Elena Tutubalina and Alex (manuscript in preparation) have performed drug clustering in semantic similarity space, using webmd.com, patient.info, drugs.com, amazon.com askapatient.com, and dailystrength.org as sources of user comments, and showed that drugs with similar pharmaceutical action do cluster together in the semantic similarity space.

Alex’s team has also experimented with *de novo* design of molecules with the desired properties using SMILES in Deep Reinforcement Learning:



Structural bias, physical properties, and biological activity have been used in proof of concept case studies of user-biased molecular design. In summary, Alex cited Confucius who said, "Without knowing the force of words, it is impossible to know more". Alex quipped "And remember: anything you say can, and will be used … for text mining!".

A nanoinformatics platform for environmental impact assessment of manufactured nanomaterials

Yoram Cohen of the University of California Center for Environmental Implications of Nanotechnology gave a talk co-authored by colleagues at the University of California. Nanoinfo.org is a nanoinformatics platform that supports the environmental impact assessment of engineered nanomaterials (ENMs) with a central database of ENM safety data and a toolkit for various exploration and analysis methods.[21] These methods include the estimation of environmental exposure levels of ENMs (MendNano), evaluation of environmental releases of ENMs (LearNano), analysis of high throughput toxicity data of ENMs (ToxNano), and predictive toxicity models, and analysis of the environmental impact of ENMs *via* Bayesian inference (NanoEIA).

NanoDatabank is a data repository of ENM properties, and experimental and simulation datasets of ENM toxicity and environmental fate and transport (F&T). It contains databases that include physicochemical properties; toxicological properties; experimental datasets of ENM toxicity, and F&T; and results of model simulations and estimation of ENM toxicity and F&T behavior, and physicochemical properties. It includes data for over 300 nanomaterials, and toxicity data for various cell lines, zebrafish and bacterial strains, from 325 publications. ToxNano is a high content data analysis tool (HDAT)[22,23] offering QSARs using random forest and Bayesian network toxicity models; analysis of knowledge evidence, and data visualization. MendNano (multimedia environmental distribution of nanomaterials) is a web-based modeling platform.[24,25] Nanoinf.org has 400 users from more than 50 countries.

As an example of work on the toxicity of nanomaterials, Yoram presented unpublished results on evaluating the body of evidence on quantum dots (QDs) *via* meta-analysis. QDs are very small semiconductor particles, only several nanometers in size, so small that their optical and electronic properties differ from those of larger particles. Many types of quantum dot will emit light of specific

frequencies if electricity or light is applied to them, and these frequencies can be precisely tuned by changing the dots' size, shape and material.

QD data were collected from 448 publications, reporting 2,703 samples, with 7 core types, 12 shell types, 13 surface modifications, 14 surface ligands, and 20 assay types. In the predictive toxicity model R^2 was about 0.81 for cell viability, and about 0.83 for IC_{50}. Yoram and his colleagues studied cause-effect relationships between cellular bioactivity and QD attribute. Median IC_{50} was ≤ 10 mg/L, for the surface ligands of type amphiphilic polymer, lipid, other hydrophobic, aminothiol, and other amphiphilic. It was uniformly distributed for silica. There was no correlation between surface charge and IC_{50}. The sensitivity distribution of IC_{50} for cell anatomical type suggests that more differentiated cells are more adversely affected by exposure to QDs. Toxicity is not governed by QD size alone: there is a wide range of IC_{50} for a given size, and toxicity can be high or low irrespective of the size. Core type affects toxicity, but the wide range of IC_{50} for a given core type suggests that there are other important attributes.

Bayesian network models can be useful for handling uncertainties, mixed attributes, and hidden conditional relationships since they provide rigorous and simple mathematical means of handling data uncertainty; they integrate graphical representation of the problem with probabilistic evaluation of variable relationships; they can incorporate prior knowledge based on data as well as expert opinion in a convenient representation of probability distributions; and they calculate the likelihood of specific scenarios based on prior knowledge.

Bayesian network model sensitivity analysis showed that QD toxicity is correlated with the most relevant (or significant) attributes tabulated below. The QD attributes identified in this study were consistent with previous analysis *via* random forest.[26]

Bayesian network for IC_{50}	Random forest for IC_{50}
Surface ligand	QD diameter
Shell	Surface ligand
QD diameter	Shell
Assay type	Assay type
Exposure time	Exposure time
Surface modification	Surface modification
Surface charge	Surface charge

Bayesian network for cell viability	Random forest for cell viability
Surface ligand	QD diameter
QD diameter	QD concentration
QD concentration	Surface ligand
Exposure time	Exposure time
Shell	Surface modification
Assay type	Assay type
Surface modification	Surface charge
Surface charge	

Bayesian networks for new explorations of association rules among various biological responses as a result of exposure to manufactured nanomaterials have also been demonstrated in zebrafish toxicity studies. Yoram and his co-workers used a nanomaterial biological interaction knowledge base of zebrafish phenotype data with 1,147 samples, and 11 biological responses (including mortality). The data included exposure to seven material types (carbon, cellulose, dendrimer, metal, (metal) oxide, polymeric, and semiconductor) of 0.8 –250 nm average primary size; concentration; number of embryos per experiment; and responses recorded for each exposure scenario.

The Bayesian network model for zebrafish mortality (percentage of dead embryos) had an R^2 of about 0.79. Sensitivity analysis of the key material properties and exposure conditions that correlate with Zebrafish mortality was carried out, and cause-effect relationships between zebrafish phenotypes and material properties and exposure conditions were investigated. Attribute significance was determined by exhaustive search of 13 attributes using bootstrapping. Mortality at 120 hours post-fertilization correlated with concentration used, core atomic composition, outermost surface, average particle size, surface charge, shell composition and purity. The significant attributes at 24 hours post-fertilization were the same but the ranking of the top four differed slightly.

The responsible development of beneficial manufactured nanomaterials requires a thorough understanding of their potential adverse environmental and human health impacts. This requires predicting the biological response of various receptors when exposed to these materials, along with an understanding of their fate and transport, and their range of likely exposure concentrations. Yoram's work helps to rank various nanomaterials with respect to their potential environmental impact.

Accurate and interpretable nano-QSAR models from genetic programming-based decision tree construction approaches

Ceyda Oksel of Imperial College London reported on the PhD work[27] she had done at the University of Leeds in collaboration with Xue Wang and David Winkler. Given the ever increasing use of ENMs, it is essential to assess properly all potential risks that may occur as a result of exposure to ENMs. The distinctive characteristics of ENMs that have made them superior to bulk materials for particular applications might also have a substantial impact on the level of risk they pose. Despite the clear benefits that nanotechnology can bring, there are serious concerns about the potential health risks associated with the production and use of ENMs, intensified by the limited understanding of what makes ENMs toxic and how to make them safe.

The involvement of computational specialists in nano-safety research has become more prominent since Registration, Evaluation, Authorization and restriction of CHemicals (the European Union's REACH regulation) promoted the use of *in silico* techniques such as QSAR for toxicity assessment. Data-driven models that decode the relationships between the biological activities of ENMs and their physicochemical characteristics provide an attractive means of maximizing the value of scarce, and expensive, experimental nanotoxicity data.

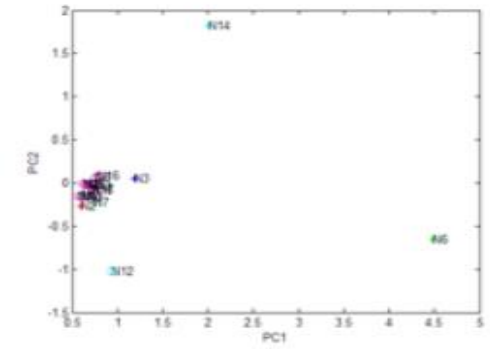
Nano-QSAR models can be used to predict the properties of new materials and to design safer materials. Leeds-based genetic programming-based decision tree (GPTree) approach[27] applies decision tree learning algorithms to identify the best combination of physicochemical properties to predict biological activity of ENMs. The trees are automatically constructed from the data. Decision trees have several advantages. They are able to deal with small, large and noisy datasets; they can detect nonlinear relationships (as well as linear ones); they allow input variables to be selected automatically; they are transparent; and they represent knowledge clearly (i.e., the models are interpretable).

GPTree begins with a random population of solutions and repeatedly attempts to find better solutions by applying genetic operators such as mutation and crossover. The first step is to construct a user-specified number of trees (usually a large number) starting from a random compound and randomly chosen descriptor. Once the initial population is generated, tournament selection is performed to identify the best tree to be used as a parent tree for genetic operators such as crossover. The best tree from the subset of trees is chosen by its fitness (e.g., accuracy). Genetic operators such as crossover and mutation are used to form the next generation of trees that are added or replace the current generation. These steps are repeated until the user-specified number of generations has been created. The decision tree model with the highest accuracy of classification for the training set is selected as the optimal decision tree model.

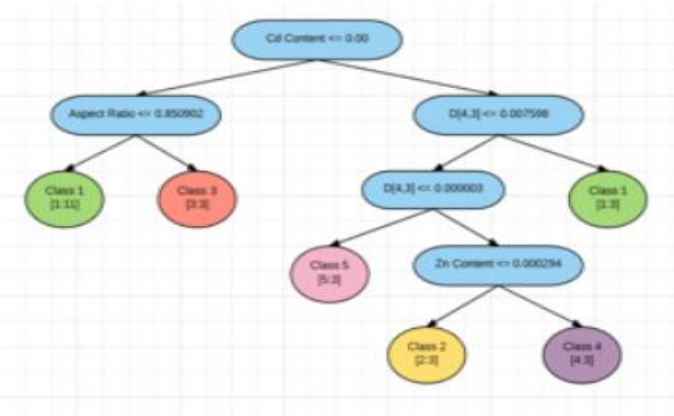
Ceyda demonstrated the application of genetic programming based decision tree construction algorithms to QSAR modeling of ENM toxicity by five case studies. The accuracy of the model predictions was satisfactorily high and clearly highly statistically significant relative to the classification rate due to chance.

In the first case study, a large set of in-house *in vitro* data (obtained in collaboration with Edinburgh University) was used. The dataset included a panel of 18 ENMs with varying structures (e.g., carbon-based materials and metal oxides), a set of *in vitro* cytotoxicity assays (e.g., LDH release, apoptosis, necrosis, viability, MTT and hemolytic effects), and several experimentally measured physicochemical properties (e.g., particle size and size distribution, surface area, morphology, metal content, reactivity and free radical generation). After a set of data preparation and scaling steps, a heat map of toxicity data combined with hierarchical clustering was constructed. As a second step, C-Visual Explorer (CVE) was used as a tool to create a parallel coordinate plot of the multivariate toxicity data. Similar to the heat map visualization results, the parallel coordinate plot showed that the aminated polystyrene latex beads and zinc oxide had the highest toxicity values in nearly all assays, followed by nanotubes that had medium to high toxicity values in viability and MTT assays.

Then, a dimensionality reduction technique, principal component analysis, was performed on all the toxicity data and the ENMs were divided into five categories according to their toxicity values. GPTree was used to identify potential descriptors contributing to the toxicity of four particular ENMs that were clearly separated from the main cluster formed by low-toxicity ENMs. It was concluded that high aspect ratio contributed to the toxicity of nanotubes, while the most likely factor driving the toxicity of zinc oxide was its high zinc content.

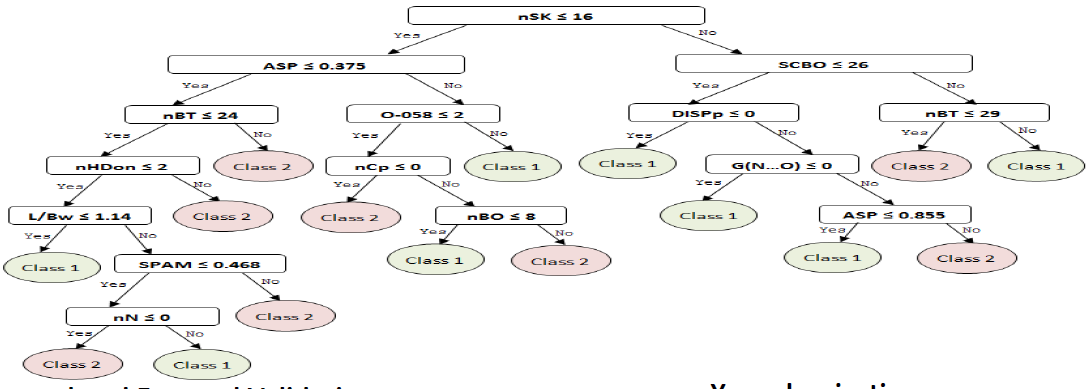


Class 1: 14 non-toxic NPs
Class 2: NP12 (Nickel Oxide)
Class 3: NP3 (Carbon Nanotubes)
Class 4: NP14 (Zinc Oxide)
Class 5: NP6 (Aminated Polystyrene Latex Beads)



GPTree Conclusions
- Important descriptors: aspect ratio, size measurements, Cd content and Zn content
- Class 3 (Nanotube): aspect ratio
- Class 4 (Zinc oxide): Zn content

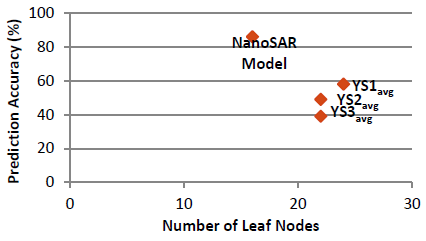
In the second case study, the cellular uptake of nanoparticles, 13 descriptors representing the hydrogen-bonding characteristics, functional group counts, molecular shape, composition and polarizability were found to be significant among a larger set of 147 chemically interpretable descriptors. The findings of GPTree analysis regarding the large contribution of lipophilicity, hydrogen bonding and molecular shape descriptors in the cellular uptake behavior of nanoparticles is consistent with earlier studies.



Internal and External Validation

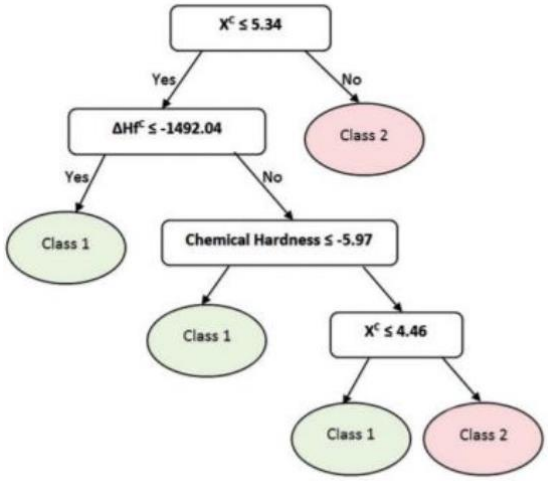
Internal Validation			External Validation		
Training Set	Predicted Class		Test Set	Predicted Class	
Actual Class	Nontoxic	Toxic	Actual Class	Nontoxic	Toxic
Nontoxic	39	1	Nontoxic	7	3
Toxic	0	47	Toxic	0	11
Sensitivity	98%		Sensitivity	79%	
Specificity	100%		Specificity	100%	
Accuracy	99%		Accuracy	86%	

Y-randomization



For a cytotoxicity to human keratinocytes dataset (the third case study),[28] the descriptors selected by GPTree were the enthalpy of formation of metal oxide nanocluster representing a fragment of the surface (ΔH_f^c), the Mulliken's electronegativity of the cluster, X^c, and the chemical hardness. The former two descriptors are consistent with the properties reported to be important for cytotoxicity of metal

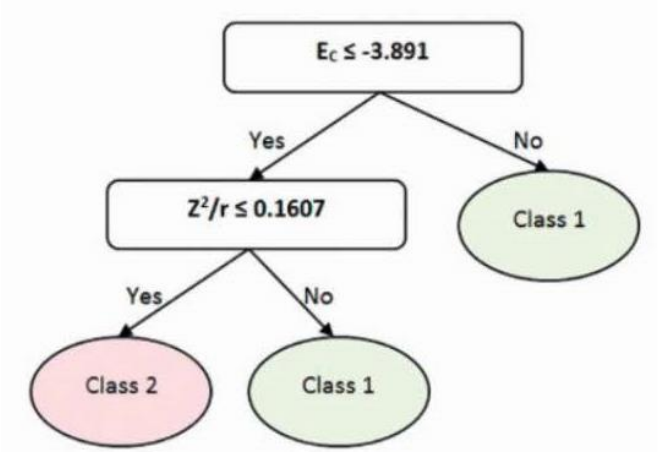
oxide nanoparticles. In addition, the chemical hardness corresponding to the reactivity was found to be an influential parameter on the cytotoxicity of nanoparticles.



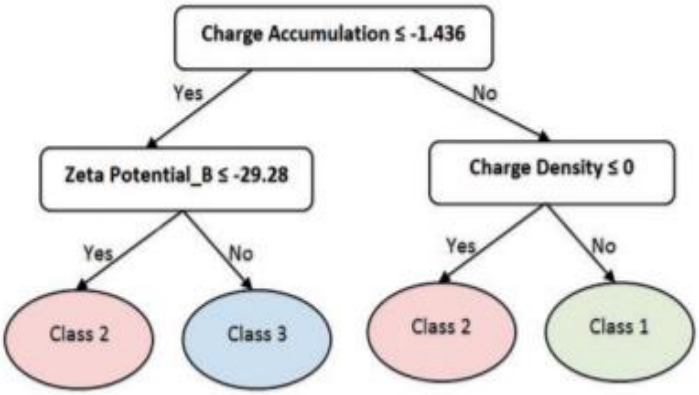
Training Set			Test Set		
	Predicted Class			Predicted Class	
Actual Class	Nontoxic	Toxic	Actual Class	Nontoxic	Toxic
Nontoxic	5	0	Nontoxic	4	0
Toxic	0	5	Toxic	0	4
Sensitivity	100%		Sensitivity	100%	
Specificity	100%		Specificity	100%	
Accuracy	100%		Accuracy	100%	

The descriptors selected by GPTree were used to develop a regression model which was statistically significant and had good predictivity (R^2 = 0.92, Q^2 = 0.72). A variable importance plot showed that X^c was twice as important as ΔH_f^c which was a little more important than η.

The data used in the fourth case study included a set of 27 descriptors, 23 ENMs, and a set of multi- and single-parameter toxicity screening assays. The descriptors selected by the GPTree model included nanoparticle conduction band energy, E_c, and ionic index of metal cation, Z^2/r. This finding is very consistent with past studies that identified these two descriptors as being important for the toxicity of metal oxide nanoparticles.



In the last case study, exocytosis of gold nanoparticles in macrophages, the optimal descriptors for predicting the exocytosis were the charge accumulation, zeta potential and charge density. These findings are in line with previous studies revealing an association between surface characteristics of gold nanoparticles, especially high positive surface charge, and their exocytosis patterns in macrophages.

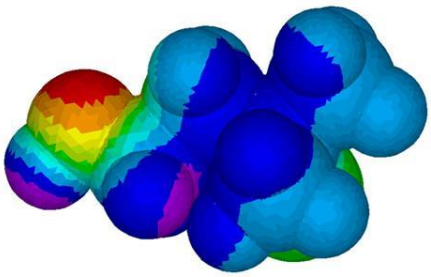


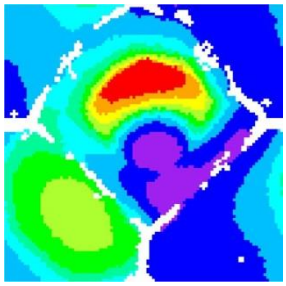
Ceyda concludes that the genetic programming based decision tree construction algorithm shows considerable promise in its ability to identify the relationship between molecular descriptors and biological effects of ENMs. Selected decision tree models yielded (external) prediction accuracies of 86-100%. Another statistical test (Y-randomization) was also performed to demonstrate the robustness of the selected models. This work is a first step in the implementation of a genetic programming based decision tree construction algorithm to nano-QSAR studies.

Self-organizing neural networks in chemistry

Johnny Gasteiger of the University of Erlangen-Nürnberg is skeptical about deep neural networks: they are good for getting funding but they are yet to be proven. Johnny illustrated some of the useful applications of shallow neural networks. Much like the human brain generates two-dimensional sensory maps of the environment, a Kohonen network (a self-organizing map) can generate two-dimensional maps of high-dimensional chemical data. Crucial for the success of the study of chemical problems by a self-organizing neural network is the representation of the chemical data.

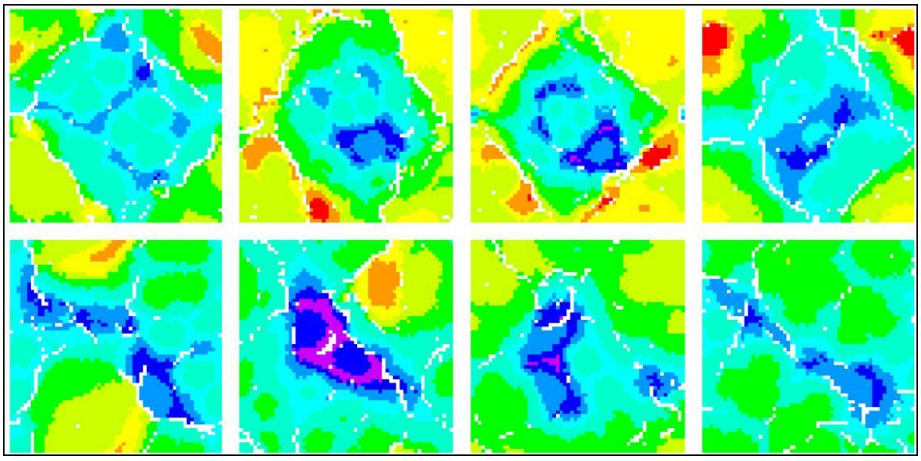
The shape and surface of molecules are very important: the entire electrostatic potential can be seen in a colored 3D model. Johnny has projected the 3D Cartesian coordinates of, for example, 2-chloro-4-hydroxy-2-methylbutane onto a Kohonen net to get a 2D map:





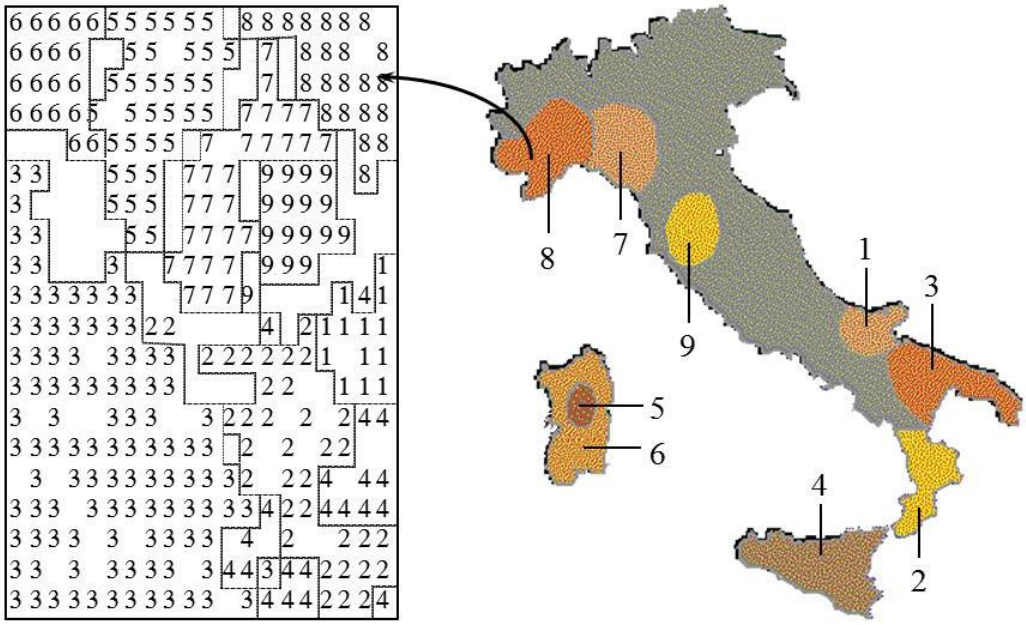
The neurotransmitter acetylcholine binds to two types of receptors, the muscarinic and the nicotinic receptor. Kohonen maps of the van der Waals surface of muscarinic agonists (muscarine, atropine,

scopolamine, pilocarpine) and nicotinic agonists (nicotine, (+)-anatoxin a, mecamylamine, pempidine) have also been produced by projecting points of the 3D surface on a 2D space.[29] Such maps allowed the total molecular electrostatic potential (MEP) of a compound to be represented in a single picture, instead of requiring a series of pictures as formerly. Johnny showed the maps of the MEPs of the eight compounds with muscarinic agonists in the top row and nicotinic agonists below.



The results showed that the MEP is important for the binding of these compounds to their receptors. The Kohonen maps reflect significant characteristics of the MEPs and can therefore be used in the search for biologically active compounds.

In analytical chemistry, neural networks have been used in the classification of Italian olive oils.[30] The classification was performed on a set of 572 Italian olive oils, from nine different regions, on the basis of an analysis of eight fatty acids. Kohonen learning was superior to a network using the back-propagation of errors. There were 250 oils in the training set and 322 in the test set; 312 of the 322 were correctly predicted. The nine Italian regions were nicely differentiated in the Kohonen map. What is, however, even more interesting is that the Kohonen map is reflecting the map of Italy. This emphasizes the power of unsupervised learning, discovering information that is hidden in the data. In this case, clearly, the different climates and the different soils are responsible for the separation of the regions of Italy in the self-organizing map:



Kohonen networks use unsupervised learning. Johnny next discussed examples of supervised learning. In one experiment the electronic properties located on the atoms of a molecule such as partial atomic charge, and electronegativity and polarizability values were encoded by an autocorrelation vector accounting for the constitution of a molecule.[31] Using the 49-dimensional vector of seven properties and seven distances, it is possible to distinguish between 112 dopamine agonists and 60 benzodiazepine receptor agonists even after projection into a Kohonen map. The two types of compounds can still be distinguished if they are buried in a dataset of 8,323 compounds of a chemical supplier catalog comprising a wide structural variety. The method can be used for searching for structural similarity, and, in particular, for finding new lead structures with biological activity.

Gasteiger's team has also worked on simulation of infrared spectra.[32] They developed an empirical approach to the modeling of the relationships between the 3D structure of a molecule and its IR spectrum based on a novel 3D structure representation, and a counterpropagation (CPG) neural network. The 3D coordinates of the atoms of a molecule are transformed into a structure code that has a fixed number of descriptors irrespective of the size of a molecule. The structure coding technique is referred to as radial distribution function (RDF) code.[33] 3D structures were transformed into radial codes (128 values) and put into a CPG network. IR spectra (128 absorbance values) were also input, and the network was trained. When IR spectra are simulated the fingerprint region is predicted well because of the representation of the 3D structure. A CPG network can be operated in reverse mode,[33] enabling the prediction of a structure code. The input of a query infrared spectrum into a trained CPG network provides a structure code vector, which represents the radial distribution function with 128 discrete values. This RDF code is then decoded to provide the Cartesian coordinates of a 3D structure.

Johnny concluded by mentioning his recent collaboration with David Winkler on dye solubility in carbon dioxide.[34] David has also worked on melting points of ionic liquids, fibrinogen adsorption to polymeric

surfaces, and normalized metabolic activity of polymeric biomaterials. Johnny encouraged David to continue to do good science.

Understudied proteins. Time to shift the paradigm

Tudor Oprea of the University of New Mexico believes that identifying novel targets as a precompetitive endeavor can lead to new therapeutic opportunities if academia and industry work together. Most protein classification schemes are based on structural and functional criteria. For therapeutic development, it is useful to understand how many data and what types of data are available for a given protein, thereby highlighting well-studied and understudied targets. Tudor and his co-workers classify proteins annotated as drug targets as "Tclin"; proteins for which *potent* small molecules are known as "Tchem"; proteins for which biology is better understood as "Tbio"; and proteins that lack antibodies, publications or National Center for Biotechnology Information *(*NCBI) Gene References Into Function (GeneRIFs) as "Tdark".

Tclin proteins are associated with drug mechanism of action (MoA). Tchem proteins have bioactivities in ChEMBL and DrugCentral, plus human curation for some targets. A Tbio protein lacks small molecule annotation, and is above the cutoff criteria for Tdark, or is annotated with a Gene Ontology (GO) molecular function or biological process leaf term(s) with an experimental evidence code, or has confirmed Online Mendelian Inheritance in Man (OMIM) phenotype(s). Tudor and his colleagues used name entity recognition software[35] from L. J. Jensen's lab to evaluate nearly 27 million abstracts to derive a publication score per protein. Tdark proteins ("understudied proteins") have little information available, and meet two of the following three criteria: a PubMed text mining score of less than five, three or fewer GeneRIFs, and 50 or fewer antibodies available according to antibodypedia. As external validation, Tdark proteins have statistically significantly lower values compared to the other three target development levels (TDLs) in terms of fewer GO terms, fewer patents, fewer National Institutes of Health (NIH) R01 grants, and fewer searches of the STRING-db database.

Tudor's first "take home message" was that there is a knowledge deficit: over 37% of the proteins remain understudied (the Tdark ones) and only about 10% of the proteome (Tclin and Tchem) can be targeted by potent small molecules. Are Tdark proteins underfunded because there is no scientific interest in this category, or is the lack of knowledge perpetuated by lack of funding? It is possible that the absence of high quality, well characterized molecular tools (i.e., antibodies or chemical probes) may be a root cause for this situation, but lack of tools leads to lack of interest, and lack of interest diminishes the probability of such tools being developed.

The patent literature is also of interest. Almost half of patent bioactivity data are never published elsewhere, and compounds may appear in patents two to four years before they appear in the literature. The SureChEMBL team has annotated the SureChEMBL patent corpus with gene and disease terms. Looking at patents between 2001 and 2013, they processed a set of 99 approved patents of interest to the Illuminating the Druggable Genome (IDG) consortium. These bioactivity data from 99

patents were manually extracted: 20,941 activity measurements for 11,358 compounds, and 1,134 assays. These data are already uploaded into ChEMBL 23. Data for seven IDG Phase 2 targets were uncovered by this patent data extraction exercise, data which progress TDLs of two targets (GPR6 and HCAR1) from Tbio to Tchem.

Anne Hersey of ChEMBL has estimated that more than 50% of the data from patents do not end up in peer-reviewed papers. IDG, Open Targets, BindingDB, and others could collectively, in a precompetitive manner, mine data from patents (if necessary, for only terminated projects, or out-of-patent drugs) and upload these data into ChEMBL and Pharos. Pharos[36] is the user interface to the Knowledge Management Center (KMC) for the IDG program funded by the NIH.

Approximately one-third of all mammalian genes are essential for life. Phenotypes resulting from knockouts of these genes in mice have provided insight into gene function and congenital disorders. The International Mouse Phenotyping Consortium (IMPC) has published research on the high-throughput discovery of novel developmental phenotypes.[37] They identified 2,788 genes with 8,241 significant phenotype calls in 25 major categories. The promise of the IMPC annotations is illustrated by examining the definite and clear links between human neurological and behavioral disorders (191 human genes) and the corresponding gene knockout mouse neurological and behavioral phenotypes. The majority of these links are for schizophrenia, Alzheimer's disease, epilepsy, and amyotrophic lateral sclerosis. Several rare diseases are also associated with these genes.

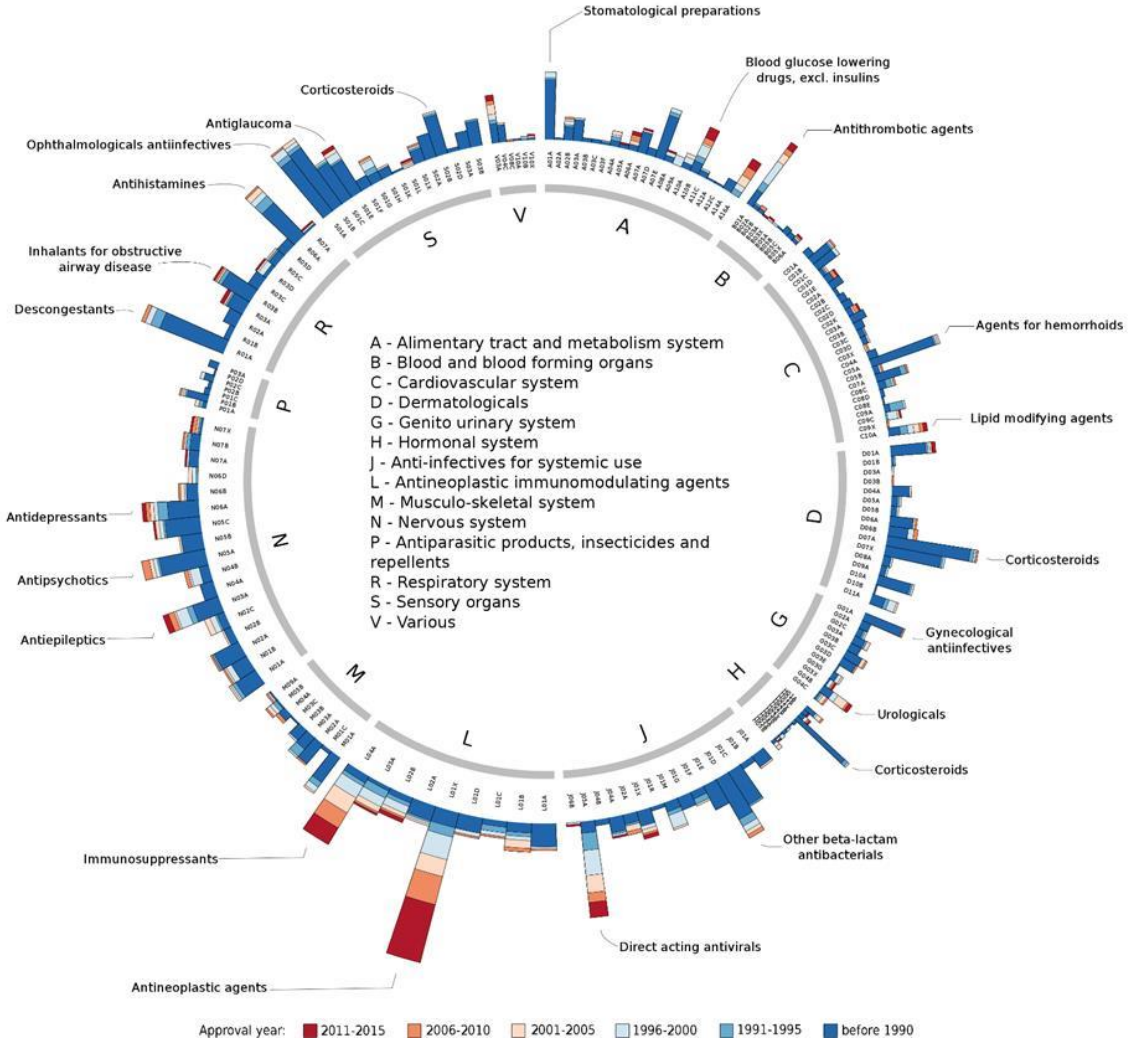
Of 119 Tdark genes prioritized by KMC to IMPC, 45 mouse lines were produced, with 41 phenotypes observed. Knockouts of the Tdark kinase Alpk3 have increased embryonic and perinatal lethality, with the surviving adults displaying severe heart defects. Of 482 Tbio genes submitted by KMC, 184 mouse lines were produced, with 145 phenotypes observed. Knockouts of the Tbio GPCR Adgrd1 display reproductive defects. (These are Tdark and Tbio statistics as of April 2017.) Tudor commented: "If you don't know very much to begin with, don't expect to learn a lot quickly."

Data from Cristian Bologa suggest that on average it takes 15-20 years for Tdark to bear fruit. The leptin receptor was Tdark in 1995, but led to an approved drug in 2014. The smoothened receptor was Tdark in 1997, and a drug was launched in 2012. Tudor gave several other examples. There is room for improvement in research funding. Text mining of all NIH grants for the period 2000-2015 suggests that 8,858 proteins received zero NIH funding. Of these, 6,051 are Tdark, and 2,616 are Tbio. This is to be expected, but 119 are Tchem and 72 are Tclin. Possible explanations could be old drug targets or research funded elsewhere. (Data from funding sources other than NIH are not available.) Pharma and academia could pay more attention to these 8,858 underfunded proteins.

Tudor's second take home message was that just because something is ignored it does not mean it lacks importance. Understudied proteins need funding and patience. Based on current evidence, IMPC has the most concerted Tdark exploration approach.

DrugCentral (http://drugcentral.org) is an open access online drug compendium[38] integrating structure, bioactivity, regulatory information, pharmacologic actions, and indications for active pharmaceutical ingredients approved by regulatory agencies. It integrates content for active ingredients with

pharmaceutical formulations, indexing drugs and drug label annotations, and complementing similar resources available online. Tudor's team used it initially to find how many drugs there are, but they also wanted to know how many drug targets there are. They have studied innovation patterns per therapeutic area:[39]



Drugs distributed by Anatomical Therapeutic Chemical (ATC) codes (levels 1-2). Concentric rings indicate ATC levels. Histograms represent the number of drugs distributed per year of first approval.

They have also examined the commercial impact of target classes by evaluating data from IMS Health on drug sales from 75 countries, aggregated over a five-year period (2011–2015). After excluding categories such as homeopathic medicines, they identified 51,095 unique products, and mapped them to 1,069 active pharmaceutical ingredients from DrugCentral, corrected by the number of active pharmaceutical ingredients (APIs) per product, then by the number of Tclin targets per API. The most lucrative target class from a therapeutic perspective was GPCR (27.42% market share). Tudor also tabulated the top 20

targets by revenue. His third take home message was that there are many unexplored opportunities. By his conservative estimate (about 15,000 disease concepts, and about 2500 unique drug indications), we address about 15% of human diseases with therapeutic agents.

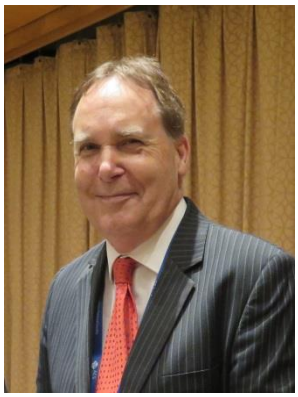
It has been said that the absence of a quantitative language is the flaw of biological research[40] or "the more facts we learn the less we understand". Again, when little is known, we should not expect knowledge to accumulate quickly. Separation by organ and cell is a conceptual fallacy. Medicine maintains this separation for necessity: by organ (e.g., cardiology or ophthalmology), and by disease category (e.g., oncology or infection). NIH Institutes are organized in a similar way. Many pharmaceutical companies are organized by therapeutic area. Yet genes, proteins and pathways do not observe such separation. The impact of this "mental divide" in science has yet to be understood.

A. B. Jensen *et al.* have studied disease correlations and temporal disease progression (trajectories)[41] on a large scale over 15 years, and grouped 1,171 significant trajectories into temporal patterns centered on a small number of early diagnoses that are central to disease progression. Hence it is important to focus on early diagnoses in order to mitigate the risk of adverse patient outcomes. The authors suggest such trajectory analyses may be useful for predicting and preventing future diseases of individual patients. Using data from the Cerner HealthFacts database, Tudor's team has found that the top diseases prior to Alzheimer's (over 5 years or more) are essential hypertension, hyperlipidemia, Type 2 diabetes mellitus, hypercholesterolemia, and coronary atherosclerosis. For renal failure, diseases over the previous five years are essential hypertension, heart failure, angina pectoris, chronic heart disease, and diabetes mellitus.

Diseases are concepts. They lack physical manifestation outside patients, so the search for cures has to be patient centered.[42] Animal models should be combined with mining of patient data. We ought to use electronic health record data to prioritize targets for further drug discovery. For example, we should get genes associated with diseases that *precede* Alzheimer's to investigate possible causality. Such priorities could be disease-specific, or phenotype-specific.

It is time to acknowledge that target prioritization for drug discovery is precompetitive knowledge. The pharmaceutical industry reward system is based on patents, which are awarded for *drugs*, not targets. Finding a good target leads to the "me-too" phenomenon. It is time to pool resources together on targets, team up with Open Targets and create a Target Selection Consortium, partnering industry with academia. "Double blind" studies could be cosponsored, to avoid the reproducibility crisis. IDG KMC is seeking new knowledge.

Sparse QSAR modeling methods for therapeutic and regenerative medicine



David Winkler's award address was co-authored by his colleague Frank Burden, now retired from CSIRO, and by co-workers at Imperial College London, King's College London, and the University of Nottingham, whose work is acknowledged in the literature references.

David's research concerns computational chemistry applied to a molecular level understanding of interactions of molecules and materials with biology. He has a strong interdisciplinary, translational research focus. His modeling, design and optimization of bioactive materials focuses on testing model predictions by subsequent experiments. He employs a range of computational tools including quantum chemistry, molecular dynamics and mechanics, molecular graphics, pharmacophore models, protein docking, and, in the case of this talk, quantitative structure-property relationship modeling. He is interested in the design of drugs and materials for therapeutic and regenerative medicine, especially control of stem cell fate, with a particular focus on the application of artificial intelligence (AI), machine learning, pattern recognition, complex systems science, evolutionary algorithms and adaptive learning.

His work has had commercial impact, including the transfer of neural network modeling technology to BioRAD Corporation; several field trials candidates with Du Pont and Schering Plough; and clinical trials of a radioprotectant drug for cancer radiotherapy patients (with Sirtex and the Peter Mac Cancer Institute). He developed core intellectual property (a novel antibacterial target in bacterial replisome) for the Betabiotics company spinoff, and discovered a new mechanism for strontium biomaterial-induced differentiation of mesenchymal stem cells to bone. He carried out a large project with Air Liquide Santé on using *in silico* methods to understand the surprisingly rich biological properties of noble gases. He discovered new antifibrotic and antihypertensive agents for Vectus Biosystems (allowing them to float on the stock market) and a first in class drug lead for myelofibrosis, which will be further developed by a new spin off company soon.

Winkler's research thinking was greatly influenced by complex systems science, which finds deep mechanistic similarities between areas of science that appear to have nothing in common. Concepts include nonlinear dynamical behavior, networks and their attractor states, self-organized criticality, chaos, and emergent properties. Complex systems science stimulates substantial lateral thinking and novel problem solving. Methods from other areas of science can provide novel solutions to problems in drug discovery; and methods developed for drug discovery can provide novel solutions to problems in other areas of science, such as biomaterials, gene expression, non-biological materials, and regenerative medicine.

QSAR was invented by Toshio Fujita (very recently deceased) and Corwin Hansch, and rapidly evolved into a method for optimization of drugs and agrochemicals. David and Toshio published a recent paper[43] on the two forms of QSAR: "explain" and "predict". Graham Richards' and Peter Andrews' seminal commercialization ventures influenced David to make translation a strong focus in his research.

The research for which David received the Skolnik award involved the application of modern computational and mathematical methods to optimizing the QSAR modeling process.[44] The first operation is to generate descriptors. Model quality is critically dependent on descriptors. Descriptors with low or no relevance to the property modeled degrade the model. Bad descriptors were a problem in early QSAR work, and there is still a major research need for good descriptors for materials. Next a subset of descriptors is chosen for the model in a context-dependent way. Choosing too many subsets can give chance correlations. In generating the relationship between the descriptors and the target property, model quality is less dependent on the modeling algorithm than on the descriptors, but there can be issues in overfitting, overtraining, ambiguity in network architecture, and subjective choices. The next operation is validating the performance of the model in predicting properties of new data. Here, cross validation and bootstrapping generate optimistic measures of performance, and an independent test set not used in training is best. The final operation is making new predictions from the model and synthesizing and testing new materials.

Descriptors are the last major research problem for QSAR. Many (such as DRAGON descriptors) are arcane; efficient, interpretable descriptors are needed. Descriptors specific to complex materials are essential but the field is embryonic. High throughput characterization data can augment computed descriptors.

There are advantages in removing irrelevant features. Least squares in multiple linear regression (MLR) has a Gaussian prior. This can be replaced with a Laplacian prior which effects the removal of uninformative weights by driving them to zero. Sparse Bayesian feature selection methods (feature selection using expectation maximization) identify a small number of relevant features very efficiently.[45]

There are many methods of varying sophistication in finding structure-activity relationships,[44] including simple linear statistical regression methods such as multiple linear regression; nonlinear regression methods using polynomials or nonlinear kernels, and nonlinear machine learning; bioinspired methods such as neural nets; support vector machines; and random forests. These have new applications in materials, nanotechnology, and regenerative medicine.

The universal approximation theorem states that neural networks can model any complex relationship given sufficient training data. Neural networks are very well suited to modeling of complex data, but they have problems such as overfitting and overtraining. They raise an ill-posed problem in statistics (instability), and optimum network architecture is ambiguous. The contribution of David and his co-workers is to develop very robust, self-optimizing sparse feature selection and neural network methods that overcome all these problems.[46] These methods have been shown to have performance similar to that of deep neural networks.

Sparse Bayesian modeling and feature selection, replacing the Gaussian prior with the Laplacian prior, is a general nonlinear modeling method[45,47-49] that automatically optimizes model complexity, prunes neural network weights to avoid overfitting, and prunes irrelevant descriptors to optimize the predictivity of a model. A sparsity-inducing Laplacian prior (LP) was introduced into Winkler's Bayesian

Regularized Artificial Neural Network algorithm (BRANN) creating BRANNLP.[47,49] Low relevance weights are set to zero, and descriptors are also pruned from the model if all weights are zero.

From selection and mapping, David turned to validation. Cross validation, bootstrapping, and other methods give an overly optimistic estimate of predictive power because the test set is not independent of the training set. An independent test set never seen by the model is the gold standard. Many measures of predictivity have been proposed. Test set validation is actually a simple problem in statistics; standard error of prediction, test set (SEP) is preferred over r^2 as it is less dependent on dataset size and model complexity.[46,50]

Methods from other areas of science can provide novel solutions to problems in drug discovery, and methods developed for drug discovery can provide novel solutions to problems in other areas of science. Implantable medical devices are an example. Bacterial adhesion and growth on biomaterial surfaces of joint prostheses, heart valves, shunts, vascular and urinary catheters, and intraocular lenses are serious problems in health care. There is a major unmet medical need for new coating materials for implantable and indwelling medical devices. David and his co-workers from Morgan Alexander's research team at the University of Nottingham have used machine learning methods to derive quantitative models relating the molecular structure of a polymer to the attachment of the bacteria to that polymer surface. These models can be used to screen large databases of new materials for those with low pathogen attachment.

Hook *et al.* have detected the attachment of selected bacterial species to 576 polymeric materials in a high-throughput microarray format.[51] In work by David and his colleagues, data from a large polymer microarray exposed to three clinical pathogens were used to derive robust and predictive machine learning models of pathogen attachment.[52] The BRANN models can predict pathogen attachment for the polymer library quantitatively. The models also successfully predict pathogen attachment for a second-generation library, and identify polymer surface chemistries that enhance or diminish pathogen attachment. A manuscript on work on multiple pathogen attachment models has been submitted.

Sparse feature selection methods have also identified a new mechanism for strontium biomaterial-induced differentiation of mesenchymal stem cells to bone. Strontium ranelate (Protelos) is a drug approved in the European Union for the treatment and prevention of osteoporosis. It reduces risk of vertebral and non-vertebral fractures in post-menopausal women. Although controversial, it is reported to have an anabolic *and* anti-catabolic effect on bone. Strontium ion's mechanism of action is not fully understood, but it is thought to up-regulate differentiation of osteoprogenitors or stimulate bone formation.[53-55]

David and his Imperial College co-workers,[56] Molly Stevens, Eileen Gentleman and Hélene Autefage, have evaluated the global response of human mesenchymal stem cells to strontium-substituted bioactive glasses using a combination of unsupervised biological and physical science techniques. Their objective analyses of whole gene-expression profiles, confirmed by standard molecular biology techniques, revealed that strontium-substituted bioactive glasses up-regulated the isoprenoid pathway, suggesting an influence on both sterol metabolite synthesis and protein prenylation processes.

In future, David hopes to see exploitation of new AI methods such as deep learning; improved descriptors for molecules that are effective and interpretable; exploitation of evolutionary methods of discovery aided by robotics; synergy of AI and evolutionary methods for adaptive evolution; adoption of *in silico* methods from drug discovery for materials and regeneration; development of autonomous or semiautonomous "closed loop" design methods; and more effective exploration of vast molecular or materials spaces.

Deep learning was predicted to be a breakthrough technology in 2013. Deep neural networks are not necessarily magic. According to the universal approximation theorem, a feed-forward network with a single hidden layer containing a finite number of neurons can approximate any continuous function, under mild assumptions on the activation function. This was first proved by Cybenko in 1989 for sigmoid activation functions. Hornik showed in 1991 that it is not the choice of the activation function, but the multilayer architecture itself which gives neural networks the potential of universal approximators.[46]

Deep learning methods have generated impressive improvements in image and voice recognition, and are now being applied to QSAR and QSAR modeling. A recent publication[46] describes the differences in approach between deep and shallow neural networks, compares their abilities to predict the properties of test sets for 15 large drug datasets, discusses the results in terms of the universal approximation theorem for neural networks, and describes how deep neural networks may ameliorate or remove troublesome "activity cliffs" in QSAR datasets. Materials space is vast and at least in some of its many dimensions, the fitness landscape is smooth. This allows adaptation, one step (one mutation) at a time. Evolution and machine learning can be combined in adaptive learning (the Baldwin effect).

A recent review discusses the problems of large materials spaces, the types of evolutionary algorithms employed to identify or optimize materials, and how materials can be represented mathematically as genomes.[57] It describes fitness landscapes and mutation operators commonly employed in materials evolution, and provides a comprehensive summary of published research on the use of evolutionary methods to generate new catalysts, phosphors, and a range of other materials. Another recent paper describes the materials genome in action.[58]

Machine learning methods have achieved wide applicability for example, in aqueous solubility of drugs[59]; polymers for stem cell growth;[60] cubane as a benzene isostere,[61] benign organic corrosion inhibitors;[62] markers for stem cell division;[63] materials for stem cell factories;[64] adverse effects of nanomaterials;[65] anticancer farnesyltransferase inhibitors;[66] and prediction of materials properties.[44]

In summary, AI tools developed for therapeutic medicine also work well for regenerative medicine. Neural networks are machine learning methods that are very applicable to (bio)materials design. The universal approximation theorem means that deep learning methods should not be superior to shallow neural networks for molecular design. Bayesian regularized neural networks can generate robust, predictive models of many types of materials and properties. Sparse Bayesian feature selection methods can reduce the dimensionality of problems, improve interpretability, and generate robust models with better predictivity. Evolutionary methods, combined with machine learning (adaptive evolution) can find effective materials quickly and efficiently.

Conclusion

Erin Davis, chair of the ACS Division of Chemical Information, formally presented the Herman Skolnik Award to David Winkler at a reception held in honor of David, following the symposium.

Erin Davis and David Winkler

References

(1) Hennemann, M.; Clark, T. EMPIRE: a highly parallel semiempirical molecular orbital program: 1: self-consistent field calculations. *J. Mol. Model.* **2014,** *20* (7), 2331.

(2) Margraf, J. T.; Hennemann, M.; Meyer, B.; Clark, T. EMPIRE: a highly parallel semiempirical molecular orbital program: 2: periodic boundary conditions. *J. Mol. Model.* **2015,** *21* (6), 144.

(3) Wick, C. R.; Hennemann, M.; Stewart, J. J. P.; Clark, T. Self-consistent field convergence for proteins: a comparison of full and localized-molecular-orbital schemes. *J. Mol. Model.* **2014,** *20* (3), 2159.

(4) Novak, M.; Jaeger, C. M.; Rumpel, A.; Kropp, H.; Peukert, W.; Clark, T.; Halik, M. The morphology of integrated self-assembled monolayers and their impact on devices - A computational and experimental approach. *Org. Electron.* **2010,** *11* (8), 1476-1482.

(5) Jedaa, A.; Salinas, M.; Jaeger, C. M.; Clark, T.; Ebel, A.; Hirsch, A.; Halik, M. Mixed self-assembled monolayer of molecules with dipolar and acceptor character. Influence on hysteresis and threshold voltage in organic thin-film transistors. *Appl. Phys. Lett.* **2012,** *100* (6), 063302/1-063302/4.

(6) Salinas, M.; Jaeger, C. M.; Amin, A. Y.; Dral, P. O.; Meyer-Friedrichsen, T.; Hirsch, A.; Clark, T.; Halik, M. The Relationship between Threshold Voltage and Dipolar Character of Self-Assembled Monolayers in Organic Thin-Film Transistors. *J. Am. Chem. Soc.* **2012,** *134* (30), 12648-12652.
(7) Jaeger, C. M.; Schmaltz, T.; Novak, M.; Khassanov, A.; Vorobiev, A.; Hennemann, M.; Krause, A.; Dietrich, H.; Zahn, D.; Hirsch, A.; Halik, M.; Clark, T. Improving the Charge Transport in Self-Assembled Monolayer Field-Effect Transistors: From Theory to Devices. *J. Am. Chem. Soc.* **2013,** *135* (12), 4893-4900.
(8) Bauer, T.; Schmaltz, T.; Lenz, T.; Halik, M.; Meyer, B.; Clark, T. Phosphonate- and Carboxylate-Based Self-Assembled Monolayers for Organic Devices: A Theoretical Study of Surface Binding on Aluminum Oxide with Experimental Support. *ACS Appl. Mater. Interfaces* **2013,** *5* (13), 6073-6080.
(9) Leitherer, S.; Jaeger, C. M.; Halik, M.; Clark, T.; Thoss, M. Modeling charge transport in C60-based self-assembled monolayers for applications in field-effect transistors. *J. Chem. Phys.* **2014,** *140* (20), 204702/1-204702/10.
(10) Schmaltz, T.; Gothe, B.; Krause, A.; Leitherer, S.; Steinrueck, H.-G.; Thoss, M.; Clark, T.; Halik, M. Effect of Structure and Disorder on the Charge Transport in Defined Self-Assembled Monolayers of Organic Semiconductors. *ACS Nano* **2017**, Ahead of Print.
(11) Ehresmann, B.; Martin, B.; Horn, A. H. C.; Clark, T. Local molecular properties and their use in predicting reactivity. *J. Mol. Model.* **2003,** *9* (5), 342-347.
(12) Clark, T. The local electron affinity for non-minimal basis sets. *J. Mol. Model.* **2010,** *16* (7), 1231-1238.
(13) Sjoberg, P.; Murray, J. S.; Brinck, T.; Politzer, P. Average local ionization energies on the molecular surfaces of aromatic systems as guides to chemical reactivity. *Can. J. Chem.* **1990,** *68* (8), 1440-1443.
(14) Bauer, T.; Jaeger, C. M.; Jordan, M. J. T.; Clark, T. A multi-agent quantum Monte Carlo model for charge transport: Application to organic field-effect transistors. *J. Chem. Phys.* **2015,** *143* (4), 044114/1-044114/9.
(15) Shubina, T. E.; Sharapa, D. I.; Schubert, C.; Zahn, D.; Halik, M.; Keller, P. A.; Pyne, S. G.; Jennepalli, S.; Guldi, D. M.; Clark, T. Fullerene Van der Waals Oligomers as Electron Traps. *J. Am. Chem. Soc.* **2014,** *136* (31), 10890-10893.
(16) Walsh, A. Inorganic materials: The quest for new functionality. *Nat. Chem.* **2015,** *7* (4), 274-275.
(17) Isayev, O.; Fourches, D.; Muratov, E. N.; Oses, C.; Rasch, K.; Tropsha, A.; Curtarolo, S. Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints. *Chem. Mater.* **2015,** *27* (3), 735-743.
(18) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **2017,** *8*, 15679.
(19) Moot, T.; Isayev, O.; Call, R. W.; McCullough, S. M.; Zemaitis, M.; Lopez, R.; Cahoon, J. F.; Tropsha, A. Material informatics driven design and experimental validation of lead titanate as an aqueous solar photocathode. *Mater. Discovery* **2016,** *6*, 9-16.
(20) Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. 2013, arXiv.org e-Print archive. https://arxiv.org/abs/1301.3781 (accessed September 4, 2017).
(21) Cohen, Y.; Rallo, R.; Liu, R.; Liu, H. H. In Silico Analysis of Nanomaterials Hazard and Risk. *Acc. Chem. Res.* **2013,** *46* (3), 802-812.
(22) Liu, R.; Jiang, W.; Walkey, C. D.; Chan, W. C. W.; Cohen, Y. Prediction of nanoparticles-cell association based on corona proteins and physicochemical properties. *Nanoscale* **2015,** *7* (21), 9664-9675.

(23) Liu, R.; Rallo, R.; Bilal, M.; Cohen, Y. Quantitative Structure-Activity Relationships for Cellular Uptake of Surface-Modified Nanoparticles. *Comb. Chem. High Throughput Screening* **2015,** *18* (4), 365-375.
(24) Liu, H. H.; Bilal, M.; Lazareva, A.; Keller, A.; Cohen, Y. Simulation tool for assessing the release and environmental distribution of nanomaterials. *Beilstein J. Nanotechnol.* **2015,** *6*, 938-951.
(25) Liu, H. H.; Cohen, Y. Multimedia Environmental Distribution of Engineered Nanomaterials. *Environ. Sci. Technol.* **2014,** *48* (6), 3281-3292.
(26) Oh, E.; Liu, R.; Nel, A.; Gemill, K. B.; Bilal, M.; Cohen, Y.; Medintz, I. L. Meta-analysis of cellular toxicity for cadmium-containing quantum dots. *Nat. Nanotechnol.* **2016,** *11* (5), 479-486.
(27) Oksel, C.; Winkler, D. A.; Ma, C. Y.; Wilkins, T.; Wang, X. Z. Accurate and interpretable nanoSAR models from genetic programming-based decision tree construction approaches. *Nanotoxicology* **2016,** *10* (7), 1001-1012.
(28) Gajewicz, A.; Schaeublin, N.; Rasulev, B.; Hussain, S.; Leszczynska, D.; Puzyn, T.; Leszczynski, J. Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles: Hints from nano-QSAR studies. *Nanotoxicology* **2015,** *9* (3), 313-325.
(29) Gasteiger, J.; Li, X. Representation of the electrostatic potentials of muscarinic and nicotinic agonists with artificial neuronal nets. *Angew. Chem., Int. Ed. Engl.* **1994,** *33* (6), 643-646.
(30) Zupan, J.; Novic, M.; Li, X.; Gasteiger, J. Classification of multicomponent analytical data of olive oils using different neural networks. *Anal. Chim. Acta* **1994,** *292* (3), 219-34.
(31) Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. Locating Biologically Active Compounds in Medium-Sized Heterogeneous Datasets by Topological Autocorrelation Vectors: Dopamine and Benzodiazepine Agonists. *J. Chem. Inf. Comput. Sci.* **1996,** *36* (6), 1205-1213.
(32) Schuur, J.; Gasteiger, J. Infrared Spectra Simulation of Substituted Benzene Derivatives on the Basis of a 3D Structure Representation. *Anal. Chem.* **1997,** *69* (13), 2398-2405.
(33) Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. Deriving the 3D structure of organic molecules from their infrared spectra. *Vib. Spectrosc.* **1999,** *19* (1), 151-164.
(34) Tarasova, A.; Burden, F.; Gasteiger, J.; Winkler, D. A. Robust modelling of solubility in supercritical carbon dioxide using Bayesian methods. *J. Mol. Graphics Modell.* **2010,** *28* (7), 593-597.
(35) Pletscher-Frankild, S.; Palleja, A.; Tsafou, K.; Binder, J. X.; Jensen, L. J. DISEASES: Text mining and data integration of disease-gene associations. *Methods (Amsterdam, Neth.)* **2015,** *74*, 83-89.
(36) Nguyen, D.-T.; Mandava, G.; Sheils, T.; Simeonov, A.; Southall, N.; Jadhav, A.; Guha, R.; Mathias, S.; Bologa, C.; Holmes, J.; Liu, G.; Mani, S.; Patel, J.; Sklar, L. A.; Ursu, O.; Waller, A.; Yang, J.; Oprea, T. I.; Brunak, S.; Jensen, L. J.; Fernandez, N.; Ma'ayan, A.; Rouillard, A. D.; Gaulton, A.; Hersey, A.; Karlsson, A.; Overington, J.; Liu, G.; Mehta, S.; Schurer, S.; Vidovic, D.; Mehta, S.; Patel, J.; Schurer, S.; Vidovic, D.; Sklar, L. A.; Waller, A. Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Res.* **2017,** *45* (D1), D995-D1002.
(37) Dickinson, M. E.; Flenniken, A. M.; Ji, X.; Teboul, L.; Wong, M. D.; White, J. K.; Meehan, T. F.; Weninger, W. J.; Westerberg, H.; Adissu, H.; Baker, C. N.; Bower, L.; Brown, J. M.; Caddle, L. B.; Chiani, F.; Clary, D.; Cleak, J.; Daly, M. J.; Denegre, J. M.; Doe, B.; Dolan, M. E.; Edie, S. M.; Fuchs, H.; Gailus-Durner, V.; Galli, A.; Gambadoro, A.; Gallegos, J.; Guo, S.; Horner, N. R.; Hsu, C.-W.; Johnson, S. J.; Kalaga, S.; Keith, L. C.; Lanoue, L.; Lawson, T. N.; Lek, M.; Mark, M.; Marschall, S.; Mason, J.; McElwee, M. L.; Newbigging, S.; Nutter, L. M. J.; Peterson, K. A.; Ramirez-Solis, R.; Rowland, D. J.; Ryder, E.; Samocha, K. E.; Seavitt, J. R.; Selloum, M.; Szoke-Kovacs, Z.; Tamura, M.; Trainor, A. G.; Tudose, I.; Wakana, S.; Warren, J.; Wendling, O.; West, D. B.; Wong, L.; Yoshiki, A.; McKay, M.; Urban, B.; Lund, C.; Froeter, E.; LaCasse, T.; Mehalow, A.; Gordon, E.; Donahue, L. R.; Taft, R.; Kutney, P.; Dion, S.; Goodwin, L.; Kales, S.; Urban, R.; Palmer, K.; Pertuy, F.; Bitz, D.; Weber, B.; Goetz-Reiner, P.; Jacobs, H.; Le Marchand, E.; El Amri, A.; El Fertak, L.; Ennah, H.; Ali-Hadji, D.; Ayadi, A.; Wattenhofer-Donze, M.; Jacquot, S.; Andre, P.; Birling, M.-C.; Pavlovic, G.; Sorg, T.; Morse, I.; Benso, F.; Stewart, M. E.; Copley, C.; Harrison, J.; Joynson,

S.; Guo, R.; Qu, D.; Spring, S.; Yu, L.; Ellegood, J.; Morikawa, L.; Shang, X.; Feugas, P.; Creighton, A.; Castellanos Penton, P.; Danisment, O.; Griggs, N.; Tudor, C. L.; Green, A. L.; Icoresi Mazzeo, C.; Siragher, E.; Lillistone, C.; Tuck, E.; Gleeson, D.; Sethi, D.; Bayzetinova, T.; Burvill, J.; Habib, B.; Weavers, L.; Maswood, R.; Miklejewska, E.; Woods, M.; Grau, E.; Newman, S.; Sinclair, C.; Brown, E.; Ayabe, S.; Iwama, M.; Murakami, A.; MacArthur, D. G.; Tocchini-Valentini, G. P.; Gao, X.; Flicek, P.; Bradley, A.; Skarnes, W. C.; Justice, M. J.; Parkinson, H. E.; Moore, M.; Wells, S.; Braun, R. E.; Svenson, K. L.; de Angelis, M. H.; Herault, Y.; Mohun, T.; Mallon, A.-M.; Henkelman, R. M.; Brown, S. D. M.; Adams, D. J.; et, a. High-throughput discovery of novel developmental phenotypes. *Nature (London, U. K.)* **2016,** *537* (7621), 508-514.

(38) Ursu, O.; Holmes, J.; Bologa, C. G.; Yang, J. J.; Mathias, S. L.; Nelson, S. J.; Oprea, T. I.; Knockel, J. DrugCentral: online drug compendium. *Nucleic Acids Res.* **2017,** *45* (D1), D932-D939.

(39) Santos, R.; Ursu, O.; Gaulton, A.; Bento, A. P.; Donadi, R. S.; Bologa, C. G.; Karlsson, A.; Al-Lazikani, B.; Hersey, A.; Oprea, T. I.; Overington, J. P. A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discovery* **2017,** *16* (1), 19-34.

(40) Lazebnik, Y. Can a biologist fix a radio? Or, what I learned while studying apoptosis. *Cancer Cell* **2002,** *2* (3), 179-182.

(41) Jensen, A. B.; Moseley, P. L.; Oprea, T. I.; Ellesoee, S. G.; Eriksson, R.; Schmock, H.; Jensen, P. B.; Jensen, L. J.; Brunak, S. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat. Commun.* **2014,** *5*, 4022.

(42) Horrobin, D. F. Opinion: Modern biomedical research: an internally self-consistent universe with little contact with medical reality? *Nat. Rev. Drug Discovery* **2003,** *2* (2), 151-154.

(43) Fujita, T.; Winkler, D. A. Understanding the Roles of the "Two QSARs". *J. Chem. Inf. Model.* **2016,** *56* (2), 269-274.

(44) Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. Quantitative Structure-Property Relationship Modeling of Diverse Materials Properties. *Chem. Rev. (Washington, DC, U. S.)* **2012,** *112* (5), 2889-2919.

(45) Figueiredo, M. A. T. Adaptive sparseness for supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003,** *25* (9), 1150-1159.

(46) Winkler, D. A.; Le, T. C. Performance of Deep and Shallow Neural Networks, the Universal Approximation Theorem, Activity Cliffs, and QSAR. *Mol. Inf.* **2017,** *36* (1-2), 1600118.

(47) Burden, F. R.; Winkler, D. A. Robust QSAR models using Bayesian regularized neural networks. *J. Med. Chem.* **1999,** *42* (16), 3183-3187.

(48) Burden, F. R.; Winkler, D. A. An Optimal Self-Pruning Neural Network and Nonlinear Descriptor Selection in QSAR. *QSAR Comb. Sci.* **2009,** *28* (10), 1092-1097.

(49) Burden, F. R.; Winkler, D. A. Optimal sparse descriptor selection for QSAR using Bayesian methods. *QSAR Comb. Sci.* **2009,** *28* (6-7), 645-653.

(50) Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J. Chem. Inf. Model.* **2015,** *55* (7), 1316-1322.

(51) Hook, A. L.; Chang, C.-Y.; Yang, J.; Luckett, J.; Cockayne, A.; Atkinson, S.; Mei, Y.; Bayston, R.; Irvine, D. J.; Langer, R.; Anderson, D. G.; Williams, P.; Davies, M. C.; Alexander, M. R. Combinatorial discovery of polymers resistant to bacterial attachment. *Nat. Biotechnol.* **2012,** *30* (9), 868-875.

(52) Epa, V. C.; Hook, A. L.; Chang, C.; Yang, J.; Langer, R.; Anderson, D. G.; Williams, P.; Davies, M. C.; Alexander, M. R.; Winkler, D. A. Modeling and prediction of bacterial attachment to polymers. *Adv. Funct. Mater.* **2014,** *24* (14), 2085-2093.

(53) Reginster, J. Y.; Seeman, E.; De Vernejoul, M. C.; Adami, S.; Compston, J.; Phenekos, C.; Devogelaer, J. P.; Curiel, M. D.; Sawicki, A.; Goemaere, S.; Sorensen, O. H.; Felsenberg, D.; Meunier, P. J. Strontium Ranelate Reduces the Risk of Nonvertebral Fractures in Postmenopausal Women with Osteoporosis: Treatment of Peripheral Osteoporosis (TROPOS) Study. *J. Clin. Endocrinol. Metab.* **2005,** *90* (5), 2816-2822.

(54) Meunier, P. J.; Roux, C.; Seeman, E.; Ortolani, S.; Badurski, J. E.; Spector, T. D.; Cannata, J.; Balogh, A.; Lemmel, E.-M.; Pors-Nielsen, S.; Rizzoli, R.; Genant, H. K.; Reginster, J.-Y.; Graham, J.; Ng, K. W.; Prince, R.; Prins, J.; Seeman, E.; Wark, J.; Reginster, J. Y.; Devogelaer, J. P.; Kaufman, J. M.; Raeman, F.; Ziekenhuis, J. P.; Walravens, M.; Pors-Nielson, S.; Beck-Nielsen, H.; Charles, P.; Sorensen, O. H.; Meunier, P. J.; Aquino, J. P.; Benhamou, C.; Blotman, F.; Bonidan, O.; Bourgeois, P.; De Vernejoul, M. C.; Dehais, J.; Fardellone, P.; Kahan, A.; Kuntz, J. L.; Marcelli, C.; Prost, A.; Vellas, B.; Weryha, G.; Lemmel, E. M.; Felsenberg, D.; Hensen, J.; Kruse, H. P.; Schmidt, W.; Semler, J.; Strucki, G.; Phenekos, C.; Balogh, A.; De Chatel, R.; Ortolani, S.; Adami, S.; Bianchi, G.; Brandi, M. L.; Cucinotta, D.; Fiore, C.; Gennari, C.; Isaia, G.; Luisetto, G.; Passariello, R.; Passeri, M.; Rovetta, G.; Tessari, L.; Badurski, J. E.; Hoszowski, K.; Lorenc, R. S.; Sawicki, A.; Diez, A.; Cannata, J. B.; Diaz Curiel, M.; Rapado, A.; Gijon, J.; Torrijos, A.; Padrino, J. M.; Roces Varela, A.; Bonjour, J. P.; Rizzoli, R.; Spector, T. D.; Clements, M.; Doyle, D. V.; Ryan, P.; Smith, I. G.; Smith, R. The effects of strontium ranelate on the risk of vertebral fracture in women with postmenopausal osteoporosis. *N. Engl. J. Med.* **2004,** *350* (5), 459-468.
(55) Meunier, P. J. Postmenopausal osteoporosis and strontium ranelate. Reply. *N. Engl. J. Med.* **2004,** *350* (19), 2002-2003.
(56) Autefage, H.; Gentleman, E.; Littmann, E.; Hedegaard, M. A. B.; Von Erlach, T.; O'Donnell, M.; Burden, F. R.; Winkler, D. A.; Stevens, M. M. Sparse feature selection methods identify unexpected global cellular response to strontium-containing materials. *Proc. Natl. Acad. Sci. U. S. A.* **2015,** *112* (14), 4280-4285.
(57) Le, T. C.; Winkler, D. A. Discovery and Optimization of Materials Using Evolutionary Approaches. *Chem. Rev. (Washington, DC, U. S.)* **2016,** *116* (10), 6107-6132.
(58) Thornton, A. W.; Simon, C. M.; Kim, J.; Kwon, O.; Deeg, K. S.; Konstas, K.; Pas, S. J.; Hill, M. R.; Winkler, D. A.; Haranczyk, M.; Smit, B. Materials Genome in Action: Identifying the Performance Limits of Physical Hydrogen Storage. *Chem. Mater.* **2017,** *29* (7), 2844-2854.
(59) Salahinejad, M.; Le, T. C.; Winkler, D. A. Aqueous Solubility Prediction: Do Crystal Lattice Interactions Help? *Mol. Pharm.* **2013,** *10* (7), 2757-2766.
(60) Epa, V. C.; Yang, J.; Mei, Y.; Hook, A. L.; Langer, R.; Anderson, D. G.; Davies, M. C.; Alexander, M. R.; Winkler, D. A. Modelling human embryoid body cell adhesion to a combinatorial library of polymer surfaces. *J. Mater. Chem.* **2012,** *22* (39), 20902-20906.
(61) Chalmers, B. A.; Xing, H.; Houston, S.; Clark, C.; Ghassabian, S.; Kuo, A.; Cao, B.; Reitsma, A.; Murray, C.-E. P.; Stok, J. E.; Boyle, G. M.; Pierce, C. J.; Littler, S. W.; Winkler, D. A.; Bernhardt, P. V.; Pasay, C.; De Voss, J. J.; McCarthy, J.; Parsons, P. G.; Walter, G. H.; Smith, M. T.; Cooper, H. M.; Nilsson, S. K.; Tsanaktsidis, J.; Savage, G. P.; Williams, C. M. Validating Eaton's Hypothesis: Cubane as a Benzene Bioisostere. *Angew. Chem., Int. Ed.* **2016,** *55* (11), 3580-3585.
(62) Winkler, D. A.; Breedon, M.; Hughes, A. E.; Burden, F. R.; Barnard, A. S.; Harvey, T. G.; Cole, I. Towards chromate-free corrosion inhibitors: structure-property models for organic alternatives. *Green Chem.* **2014,** *16* (6), 3349-3357.
(63) Huh, Y. H.; Noh, M.; Burden, F. R.; Chen, J. C.; Winkler, D. A.; Sherley, J. L. Sparse feature selection identifies H2A.Z as a novel, pattern-specific biomarker for asymmetrically self-renewing distributed stem cells. *Stem Cell Res.* **2015,** *14* (2), 144-154.
(64) Celiz, A. D.; Smith, J. G. W.; Langer, R.; Anderson, D. G.; Winkler, D. A.; Barrett, D. A.; Davies, M. C.; Young, L. E.; Denning, C.; Alexander, M. R. Materials for stem cell factories of the future. *Nat. Mater.* **2014,** *13* (6), 570-579.
(65) Epa, V. C.; Burden, F. R.; Tassa, C.; Weissleder, R.; Shaw, S.; Winkler, D. A. Modeling Biological Activities of Nanoparticles. *Nano Lett.* **2012,** *12* (11), 5808-5812.
(66) Polley, M. J.; Winkler, D. A.; Burden, F. R. Broad-Based Quantitative Structure-Activity Relationship Modeling of Potency and Selectivity of Farnesyltransferase Inhibitors Using a Bayesian Regularized Neural Network. *J. Med. Chem.* **2004,** *47* (25), 6230-6238.