Herman Skolnik Award Symposium 2019, Honoring Kimito Funatsu

A Report for the *Chemical Information Bulletin* by Wendy Warr

Introduction

Prof. Kimito Funatsu was selected to receive the 2019 Herman Skolnik Award for his contributions to structure elucidation, *de novo* structure generation, and applications of cheminformatics methods to materials design and chemical process control. His seminal contributions include the conceptualization and implementation of algorithms and expert systems for structure elucidation and chemical synthesis design, systems which have been extensively applied in the pharmaceutical industry. In recent years, he has increasingly focused on inverse QSAR analysis, including *de novo* structure generation, and the development of the soft sensor methodology for chemical process control. The latter approach represents another example of ground-breaking research with immediate practical and industrial application potential.

Kimito has secured large amounts of funding from the chemical and pharmaceutical industries to drive large-scale collaborative projects at the interface between academia and industry, most recently in the context of the CREST Program on Big Data Applications, funded by the Japan Science and Technology Agency. With more than 200 peer-reviewed publications, and a plethora of presentations and conference contributions, Kimito is among the core of leaders of the chemical information and informatics field worldwide.

He obtained his doctoral degree in physical organic chemistry from Kyushu University in 1983, and joined Prof. Shinichi Sasaki's group at Toyohashi University of Technology in 1984. During his time with that group, he worked on a variety of cheminformatics applications including the structure elucidation system CHEMICS,[1] the organic synthesis design systems artificial intelligence for planning and handling organic synthesis (AIPHOS)[2] and knowledge base-oriented synthesis planning system (KOSP),[3] and other systems in the areas of *de novo* design, and chemogenomics. In 2004, he moved to the University of Tokyo to continue research in these areas as a full professor, and there he expanded into material design and soft sensors for monitoring and controlling chemical plants.[4] In addition to his professorship, he is the research director of the Data Science Center at the Nara Institute of Science and Technology (NAIST).

Kimito initiated the tradition of organizing biannual international cheminformatics schools in Japan. He also initiated the Computer-aided Chemistry Forum for scientific communication and practical training in cheminformatics, and established the Japanese Society of Cheminformatics. His relentless community service efforts also include his tenure as the President of the Division of Chemical Information and Computer Sciences of the Chemical Society of Japan (2004–2014). He has received several awards in recognition of his many contributions, including awards from the Japan Information Center of Science and Technology in 1988, from the Society of Computer Chemistry Japan in 2003, and from the Society of Chemical Engineering in 2017.

Kimito was invited to present an award symposium at the Fall 2019 ACS National Meeting in San Diego, CA. There were 10 speakers, in addition to Kimito himself.

L to R: Shigehiko Kanaya, Yoichi Zushi, Yukihiko Uchi, Yuya Takeda, Yoshihiro Yamanishi (back row), Kimito Funatsu, Kenji Hori (back row), Gisbert Schneider (back row), Kiyoshi Hasegawa, Manabu Sugimoto, Jürgen Bajorath (pictured as insert)

Monitoring progress in lead optimization

Jürgen Bajorath of the University of Bonn presented a computational method termed Compound Optimization MOnitor (COMO)[5] that helps to determine if further optimization progress can be expected for a given analogue series (AS) or if sufficient numbers of analogues have been generated. In COMO, virtual analogues (VAs) are used to populate the chemical space around an AS; chemical neighborhoods (NBHs) of analogues are defined, and VAs falling inside and outside NBHs are determined; potency distributions of analogues are analyzed; lead optimization relevant properties of analogues are evaluated; and multiple scores are calculated to quantify AS progression.

To produce VAs, the core of an AS is decorated with more than 16,000 substituents extracted from ChEMBL, applying 12 retrosynthetic (RECAP)[6] rules. The VAs are sampled using RECAP-rule-compliant substituents and hydrogen atoms. A large number of analogue series from different sources have been studied, and alternative chemical space representations and virtual analogues of different designs have been explored.[7]

Coverage of chemical space around ASs has been estimated by defining NBHs of experimental analogues and screening these NBHs with virtual compounds.[7,8] To evaluate compound distributions in AS-centered chemical space and across NBHs of analogues, distances are calculated. The distance between two compounds is given by the Euclidean distance between two multidimensional vectors encoding molecular properties. VAs may or may not map to NBHs of existing analogues, and VAs may be located in overlapping NBHs (Figure 1).

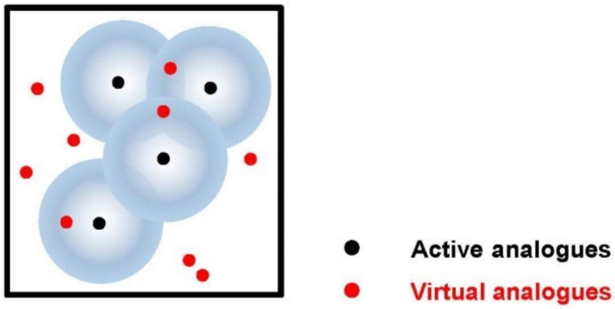


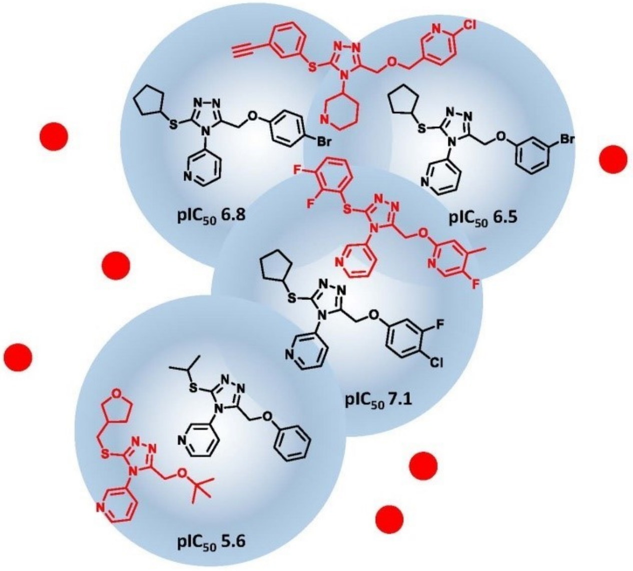
Figure 1. Neighborhood analysis

To evaluate lead optimization progress, it must be determined how extensively chemical space around a given AS is covered, and how densely an AS samples the covered space (chemical saturation), and whether analogues display significant potency variations (SAR progression). Varying potency of structural analogues indicates SAR discontinuity.

A COMO scoring scheme[5] was developed for profiling ASs that addresses the questions of chemical saturation and SAR progression. The chemical space coverage score (*C*) quantifies the VA coverage of all NBHs; the coverage density score (*D*) measures the VA coverage of overlapping NBHs; and the chemical saturation score (*S)* combines *C* and *D* (2*CD*/*C*+*D*).

The SAR progression score (*P*) quantifies potency variations of analogues sharing VAs in their NBHs. It is a VA-dependent measure of local SAR discontinuity. The SAR heterogeneity score (*H*) relates the potency distribution of analogues with overlapping NBHs to the mean potency of the AS. It is a VA-independent measure of global SAR heterogeneity. The multiproperty score (*M*) evaluates multiple compound properties. It is independent of the COMO scoring formalism, and includes a “traffic light” score for each individual ADME property of any active analogue.

Figure 2 shows exemplary analogues of a series of ATPase inhibitors, and virtual analogues falling into their NBHs. There are nine VAs, four NBHs, and three VAs in NBHs. Two of the VAs are in overlapping NBHs.



Coverage (3/9)	*C* = 0.33
Density 1-1/(6/3)	*D* = 0.50
Saturation	*S* = 0.40
SAR Progression	*P* = 0.66
SAR Heterogeneity	*H* = 0.69

Figure 2. Exemplary analogues belonging to a series of ATPase inhibitors (black), and their NBHs (light blue), and virtual analogues (red) falling into the NBHs.

Finally, Jürgen presented some unpublished work on analogue series profiling, in which his team studied 72 ASs, extracted from ChEMBL 24, with 1-6 substitution sites and 50-148 analogues. It was found that *S* and *P* scores are largely insensitive to varying VA counts. The standard VA population size was 3000 VAs. With increasing NBH radii, saturation increases and differentiates ASs (for 72 ASs and 3000 VAs). The standard NBH radius threshold (the distance threshold for the top 1% of closest pairwise distances between virtual analogues) was 1.0 (1st percentile). *P* scores change very little because they largely depend on the potency distribution of existing analogues falling into *overlapping* neighborhoods. A plot of *P* score against *S* score, with points sized by AS size and colored by *M* score, shows that *S* and *P* scores are uncorrelated and do not scale with AS size. It also shows that scoring differentiates between series with different chemical saturation and SAR progression characteristics.

COMO also includes a compound design component. VAs serve a dual role as diagnostic compounds and candidates for lead optimization. Virtual analogues generated for chemical saturation analysis provide a pool of candidates for synthesis. COMO can be combined with machine learning to produce predictive models for VA selection. Support vector regression is being used for potency prediction of VAs. The methodology is easily expandable to include multiple optimization of relevant properties. Practical applications are underway.

Electronic-structure informatics using 3D descriptors of molecules

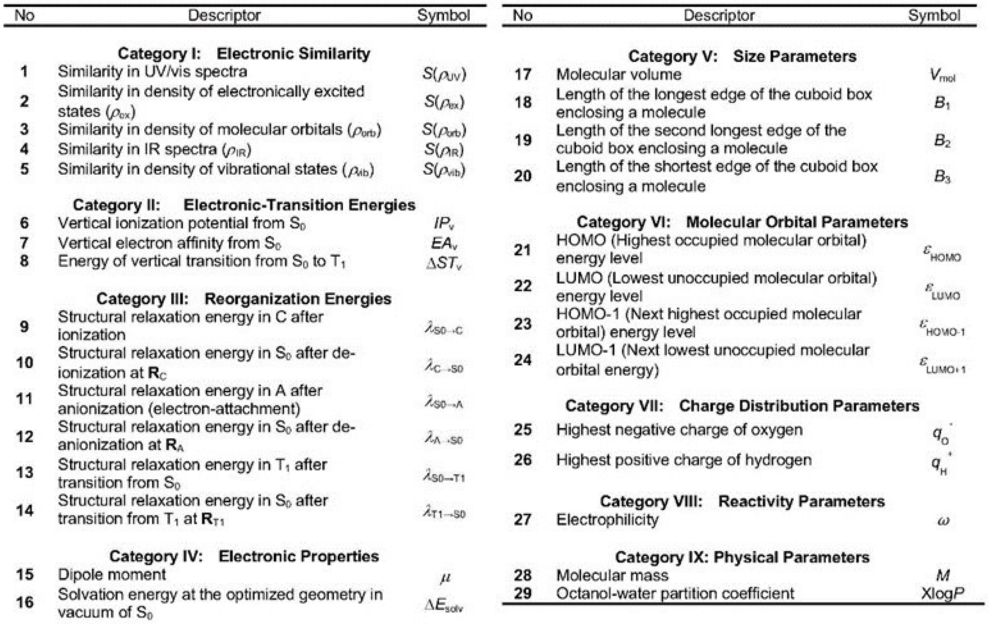
Manabu Sugimoto of Kumamoto University began by paying tribute to Kimito Funatsu and other great scientists who have inspired us. He explained that electronic -structure informatics is a discipline which obtains chemical information from electronic structures and the responses of molecules. It has been applied by great scientists to structure-reactivity relationships, structure-activity relationships, and structure-property relationships.

Years ago, Manabu, working with Prof. Hiroshi Nakatsuji, and studied metal NMR chemical shifts with an *ab initio* molecular orbital method[9] to establish a reliable method of calculating those shifts and to clarify electronic mechanisms and origins of the shifts. They found that both ground and excited states are important for quantitative structure-property relationships.

The Hubbard-Holstein model is a simple model to describe the behavior of solid crystalline materials by considering the action of electrons and phonons on a lattice. It is a combination of the Hubbard model (electrons on a lattice) and the Holstein model (phonons and their interactions with electrons on a lattice). The Hamiltonian for the Holstein model has the same first term as the Hubbard model (i.e., the electron hopping) but has two new terms concerning phonons. Electron-phonon coupling is important in the understanding of various chemistries, such as hole transport materials,[10] organometallic chemistry,[11] and organic light emitting diodes.[12]

Both molecular structure and molecular properties are related to electronic structure, so, there is the possibility of doing cheminformatics as quantum chemists would. Energy is the secret. In a chemical reaction, the energy falls from the reactant state, then rises to the transition state, and then falls again as the products are formed. The theory of chemical equilibrium tells us that biological activity can be related to “energy changes”.

Until recently, most of the descriptors that Manabu and his colleagues have been applying correspond to spectroscopic features of molecules. This set of descriptors has limitations in describing three-dimensional features related to molecular recognition. Therefore, for efficient cheminformatics modeling, Manabu now suggests three dimensional descriptors, which represent topological features of interaction energy surfaces and molecular orbitals, and coarse-grained descriptions of three dimensional features of molecules (Figure 3).



No	Descriptor	Symbol
	Category I: Electronic Similarity	
1	Similarity in UV/vis spectra	$S(\rho_{UV})$
2	Similarity in density of electronically excited states (ρ_{ex})	$S(\rho_{ex})$
3	Similarity in density of molecular orbitals (ρ_{orb})	$S(\rho_{orb})$
4	Similarity in IR spectra (ρ_{IR})	$S(\rho_{IR})$
5	Similarity in density of vibrational states (ρ_{vib})	$S(\rho_{vib})$
	Category II: Electronic-Transition Energies	
6	Vertical ionization potential from S_0	IP_v
7	Vertical electron affinity from S_0	EA_v
8	Energy of vertical transition from S_0 to T_1	ΔST_v
	Category III: Reorganization Energies	
9	Structural relaxation energy in C after ionization	$\lambda_{S0\to C}$
10	Structural relaxation energy in S_0 after de-ionization at $\mathbf{R}_C$	$\lambda_{C\to S0}$
11	Structural relaxation energy in A after anionization (electron-attachment)	$\lambda_{S0\to A}$
12	Structural relaxation energy in S_0 after de-anionization at $\mathbf{R}_A$	$\lambda_{A\to S0}$
13	Structural relaxation energy in T_1 after transition from S_0	$\lambda_{S0\to T1}$
14	Structural relaxation energy in S_0 after transition from T_1 at $\mathbf{R}_{T1}$	$\lambda_{T1\to S0}$
	Category IV: Electronic Properties	
15	Dipole moment	μ
16	Solvation energy at the optimized geometry in vacuum of S_0	ΔE_{solv}
	Category V: Size Parameters	
17	Molecular volume	V_{mol}
18	Length of the longest edge of the cuboid box enclosing a molecule	B_1
19	Length of the second longest edge of the cuboid box enclosing a molecule	B_2
20	Length of the shortest edge of the cuboid box enclosing a molecule	B_3
	Category VI: Molecular Orbital Parameters	
21	HOMO (Highest occupied molecular orbital) energy level	ε_{HOMO}
22	LUMO (Lowest unoccupied molecular orbital) energy level	ε_{LUMO}
23	HOMO-1 (Next highest occupied molecular orbital) energy level	ε_{HOMO-1}
24	LUMO-1 (Next lowest unoccupied molecular orbital energy)	ε_{LUMO+1}
	Category VII: Charge Distribution Parameters	
25	Highest negative charge of oxygen	q_O^-
26	Highest positive charge of hydrogen	q_H^+
	Category VIII: Reactivity Parameters	
27	Electrophilicity	ω
	Category IX: Physical Parameters	
28	Molecular mass	M
29	Octanol-water partition coefficient	X$\log P$

Figure 3. 3D descriptors

Electronic states to be considered are neutral spin-singlet ground (S_0) state; neutral spin-singlet excited (S_n) states; lowest neutral spin-triplet excited (T_1) state; ionized (cation) state; and electron-attached (anion) state. Molecular size is also important, for example, in ligand-target docking. The main contributors to intermolecular interaction energy are intermolecular distance, excitation energy, and transition dipole moment. Again, Manabu emphasized that both ground and excited states are important for quantitative structure-property relationships. Solvation-desolvation energy is important in ion exchange phenomena.

Manabu briefly described four applications of these 3D descriptors. Toshi Ideo has studied polyphenols as fatty acid synthase (FASN) inhibitors and has obtained a coefficient of determinants (R^2) of 0.9098 between experimental and predicted logIC_{50}. For some terpene antibacterial reagents, an R^2 of 0.763 was obtained for experimental *versus* predicted log*MIC* (minimum inhibitory concentration). The descriptors were rather less successful in predicting the biological activity of some chemicals regulating food intake.

The fourth application, carried out by Alga Manggara, concerned acute aquatic toxicity of 33 alkylphenols. Toxicity is related to both chemical and physical descriptors. The model predicts the concentration of a substance that inhibits 50% of the growth (IGC_{50}) of a *Tetrahymena pyriformis* population within a designated period. A dataset from Cronin *et al.*[13] was used. A correlation matrix for the 29 descriptors evaluated for the alkylphenols showed strong correlation between certain pairs. Some correlating descriptors were eliminated to give five descriptor sets with the following results:

- Electronic + Size (R^2 = 0.950)
- Electronic + Size + (MO + Charge + Reactivity + Physical) version 1 (R^2 = 0.929)
- Electronic + Size + (MO + Charge + Reactivity + Physical) version 2 (R^2 = 0.946)
- Electronic (R^2 = 0.798)
- Electronic + (MO + Charge + Reactivity + Physical) (R^2 = 0.820)

Different regression models were obtained for the five sets because *B*, *M*, and *EA* (see Figure 3) are mutually correlated. The size of the alkyl group is important to the mechanism of action: smaller IGC_{50} leads to larger *pIGC*, larger *B,* larger *M* and smaller *EA* (lower LUMO).

In summary, Manabu described a new set of quantum chemical descriptors of molecules designed to describe three-dimensional features. He presented some applications showing reasonable correlations with experiments. In work on acute aquatic toxicity of alkyl phenols, the regression models were different for different descriptor sets because of strong correlation between descriptors.

Fast evaluation of potential synthesis routes using DFT calculations on the basis of Transition State Database (TSDB)

Four research groups are involved in a Japan Science and Technology Agency (JST) Core Research for Evolutionary Science and Technology (CREST) project "Development of a knowledge-generating platform driven by big data in drug discovery through production processes". Makoto Taiji's team at Riken first makes a very large scale virtual library (VLSVL) of drug candidate molecules, and adds synthetic routes and physicochemical properties. Yasushi Okuno's group at Kyoto University takes information from the Riken group, studies ligand protein interactions, and passes back potential drug candidates to Riken. Kenji Hori of Yamaguchi University works on the rapid evaluation of the feasibility of synthetic routes, and shares routes and physicochemical properties with the Riken group. Kimito Funatsu's group at Tokyo University is concerned with process control and the operation of chemical plants. They interact both with the Riken group and with Kenji's group.

Kenji spoke about his own contributions. He described a procedure for *in silico* screening for synthesis routes. Systems such as Kimito's transform-oriented synthesis planning system (TOSP),[14] and knowledge base-oriented synthesis planning system (KOSP)[3,14] usually offer many routes for any one target molecule. It may be hard to decide which route to try first and to confirm whether, in practice, the selected route actually produces the desired target. *In silico* screening addresses these problems.

Initially, organic chemists do a preliminary screening to select fewer than 10 potential reactions. These are the input for the *in silico* screening process, where transition state (TS) searches of the main and side reactions are carried out, and the Gibbs free energy of activation ($\Delta G^{\ddagger}$) is calculated. The reactions selected by this process are analyzed in more detail by estimation of solvent effects and study of the effectiveness of the route. The output is a set of ranked synthesis routes for experimental study. Kenji's team has reported several successful implementations of this process.[15-17]

The benefit of *in silico* screening is to exclude experiments which are unlikely to produce the target, dramatically decreasing the number of experiments needed and shortening the time spent on synthesis route development. Since it takes a long time to search TSs for reactions, Kenji aimed to shorten the CPU times for TS optimizations. Information on similar reactions is extremely useful in locating TSs, and, to obtain it, a database including TS information is needed.

Kenji's team has gathered information on chemical reactions such as molecular names, keywords, optimized coordinates, and log files of quantum mechanical calculations and has constructed a Quantum Mechanical Calculations Results Database (QMRDB). The data in QMRDB are used to construct another database, the Transition State Database (TSDB).[17] The PostgreSQL program is used for data handling, and the Open Babel program for molecular structure retrieval. The databases are searched in a Web browser.

The team has developed a cloud system for managing the databases and theoretical calculations. The user enters a SMILES string in a program called *c*Structure, on a Windows client, to search TSDB for mechanisms which use a similar reactant, or produce a similar product. The Tanimoto coefficient, TS coordinates, a chemical equation, and $\Delta G^{\ddagger}$ and ΔG values are returned by the database server in the cloud. *i*Structure, a program on the client, adds substituents at accurate positions and produces the input for the Gaussian09 program, which is run on the server. A procedure has been developed in the *i*Structure program to create an initial structure for TS optimization for complicated targets, starting by downloading simple TS coordinates.

Kenji outlined how quantum chemistry assisted synthesis route development can contribute to synthetic chemistry in the 21st century. Synthesis planning systems usually offer many routes for a given molecule. *In silico* screening can drastically reduce the many possible routes to a few of the most feasible for experimental work. If the target is indeed obtained by actual synthesis, further work can be carried out, such as physicochemical property calculation. If the target is not obtained, feedback on the experimental results is passed to the *in silico* screening system.

In an example, target molecules from the VLSVL are sent to the Okuno team for deep learning predictions and docking calculations. The resulting drug candidates are passed to the *in silico* screening system. Ranked synthesis suggestions from that system are tested in synthetic experiments, and bioassays are carried out on compounds successfully synthesized. Active compounds are submitted for further investigation; data on inactive compounds are fed back into the docking system.

Molecules in the VLSVL were created by applying name reactions to molecules in a library of druglike molecules. Corresponding mechanisms were first examined to confirm whether or not TSs for the reactions existed. It is necessary to evaluate the toxicity of candidate molecules from the VLSVL using medicinal chemistry knowledge, and toxicity predictions are required to select potential candidates in the VLSVL and thus reduce the computational time needed for screening. It is very easy to construct initial structures of transition states using the *i*Structure program, but alternative synthesis routes have to be created when the synthesis route for the VLSVL is confirmed not to produce the target molecules. The TOSP and KOSP programs are used to suggest alternatives. Kenji presented two specific examples where the calculation of $\Delta G^\ddagger$ and ΔG for a transition state successfully supports the choice of alternative routes.

The design of new functional molecules is easy, but the development of their real synthesis routes is very difficult. Effort should be devoted to reducing the time wasted at this stage. Kenji gave three reasons why the synthetic routes from synthesis planning systems are not guaranteed to produce the targets: the precursors are much more complicated than the targets themselves; the number of synthesis routes diverges in multistep routes; and the reaction may not produce the target as the main product. A check based on cheminformatics can suggest a route that is likely to produce the desired target. It is possible to produce all the plausible products for a given set of reactants by using an appropriate reaction SMARTS.

Kenji closed by presenting an innovation cycle for developing functional molecules (Figure 4). AI molecular design suggests targets with the desired physicochemical properties and has a strategy for improving the functions of molecules. Synthesis routes from synthetic planning systems are produced but are not guaranteed to produce the targets. At the end of the cycle, results of the measurement of physicochemical properties of the targets are fed back into the AI system. The missing link is the connection of AI molecular design and *in silico* synthesis route development to the measurement of physicochemical properties of targets.

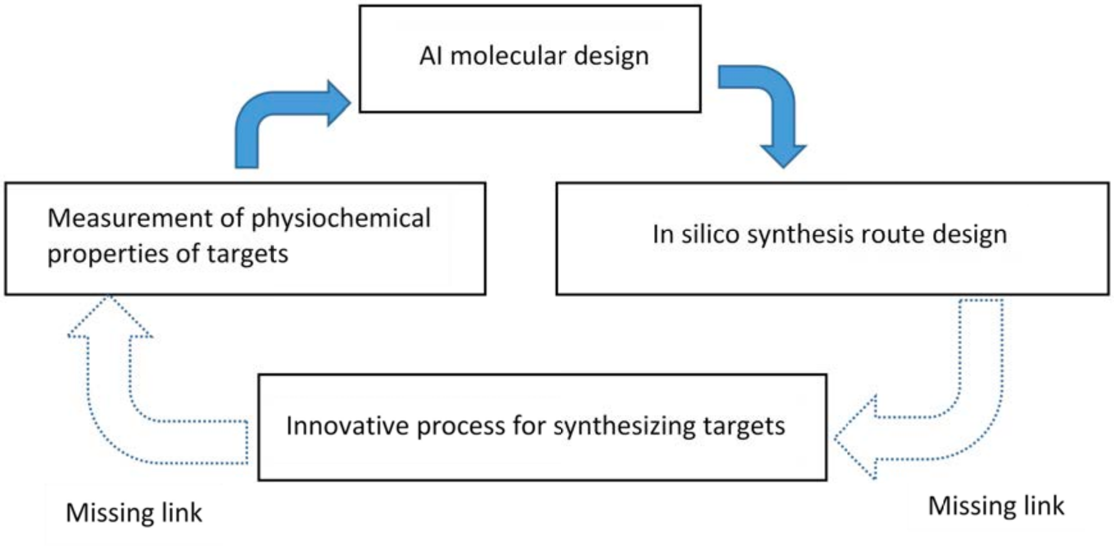

AI molecular design
In silico synthesis route design
Innovative process for synthesizing targets
Measurement of physiochemical properties of targets
Missing link
Missing link

Figure 4. Innovation cycle for developing functional molecules

The missing link involves theoretical calculations, transition state searches, judgment on the basis of calculated $\Delta G^{\ddagger}$, and ranking synthesis routes. Experimental work can then confirm a synthesis route, and increase the yield of the reaction. The results of the experimental work are fed back into the theoretical calculation system. This whole subcycle, an innovative process for synthesizing targets, is the missing link which connects AI molecular design (followed by route development) to the measurement of physicochemical properties of targets.

Development using materials informatics in Japanese companies

The Materials Genome Initiative (https://www.mgi.gov) is a U.S. multi-agency initiative designed to create a new era of policy, resources, and infrastructure that support U.S. institutions in the effort to discover, manufacture, and deploy advanced materials twice as fast, at a fraction of the cost. When it was launched in 2011, the impact in Japan was significant, and similar efforts began in Japan. The discipline of materials informatics is perceived as new by some people, but Kimito Funatsu has long been working on the application of cheminformatics methods to materials design and chemical processing, and has laid the foundations of materials informatics.

Yukihiko Uchi of Asahi Kasei Corporation described work done in his own company in collaboration with Kimito's team (http://www.mssj.jp/conf/62/program/2P-33.html). They have developed a structure prediction method for unknown compounds using two types of gas chromatography simultaneously plus mass spectrometry (GC x GC/MS) and quantitative structure-retention relationship (QSRR) inverse analysis models. Two-dimensional gas chromatography (GC x GC) is a gas chromatography technique that uses two different columns with two different stationary phases. The basic assumption of QSRR is that the retention time of GC has some correlation with various physical properties of compounds.[18]

In forward analysis, a correlation model is built from known structures and their retention times. In inverse analysis, humans decide on the substructures that might represent key peaks in the MS of an unknown compound, and a structure generation program is used to generate plausible full structures. A correlation model is then used to predict the retention times for those structures, for comparison with the observed retention times. The quality of the decisions made about substructures depends on the experience and ability of the individual making the decision.

In work done by Yukihiko's colleagues, the first GC column has a nonpolar stationary phase (for boiling point separation), and the second column has a medium polarity phase (for polarity separation). There are two different retention times from the two columns. In forward analysis, molfiles of standard compounds are given Dragon6 descriptors (http://talete.mi.it/index.htm), the two retention times for each compound are measured, and a model is built for each type of retention time, using ensemble partial least squares regression. Excellent agreement was obtained between predicted and observed nonpolar retention times (R^2 = 0.94); the agreement was not quite so good for medium polarity retention time (R^2 = 0.54).

The structure generation algorithm used in inverse analysis is Chemish (http://www.cheminfonavi.co.jp/chemish/), developed in Kimito's laboratory. Molfiles for the candidate structures output by Chemish are given Dragon6 descriptors; a correlation model is used to predict the retention times for those structures; those retention times are compared with the measured ones; and candidate structures with times close to those of the unknown compound are ranked in order of probability.

Yukihiko presented two verification examples. The first is shown in Figure 5. The correct candidate was ranked 23rd among 359 structures and had retention times of 36.7 minutes and 5.2 seconds. (The first and second choices had retention times of 37.6 minutes and 4.7 seconds; and 37.1 minutes and 5.0 seconds.)

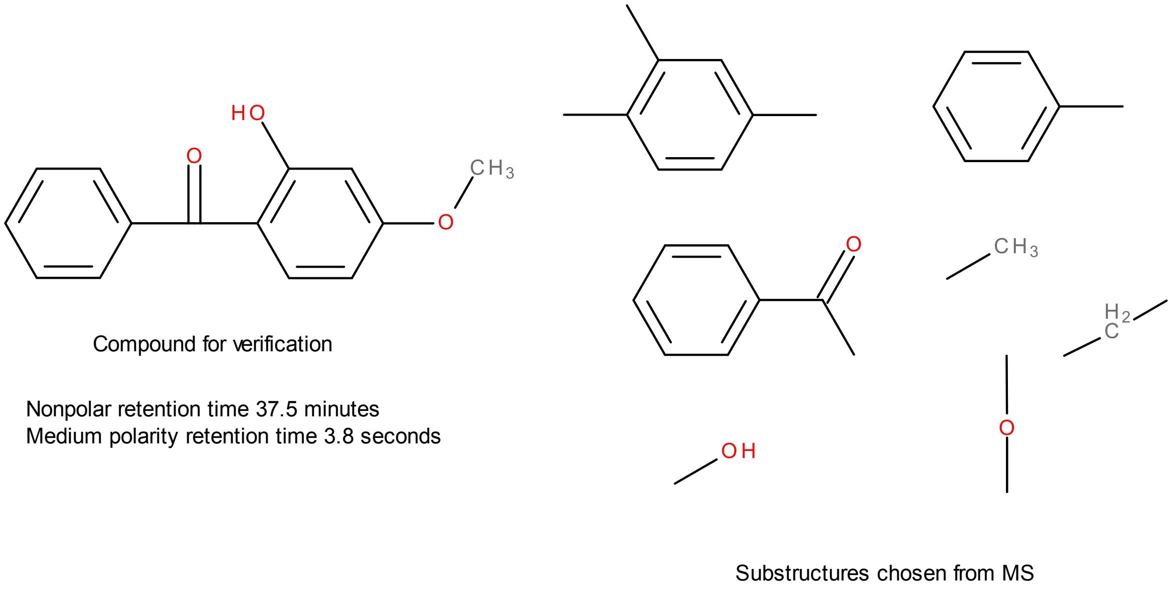

HO
O
CH3
O
Compound for verification
Nonpolar retention time 37.5 minutes
Medium polarity retention time 3.8 seconds
O
CH3
H2C
OH
O
Substructures chosen from MS

Figure 5. Inverse analysis verification example.

A second verification example is shown in Figure 6. There were 72 candidate structures. The correct candidate was ranked third with retention times of 47.0 minutes and 4.8 seconds. (The candidates ranked first and second had retention times of 44.7 minutes and 4.5 seconds; and 46.3 minutes and 4.5 seconds.)

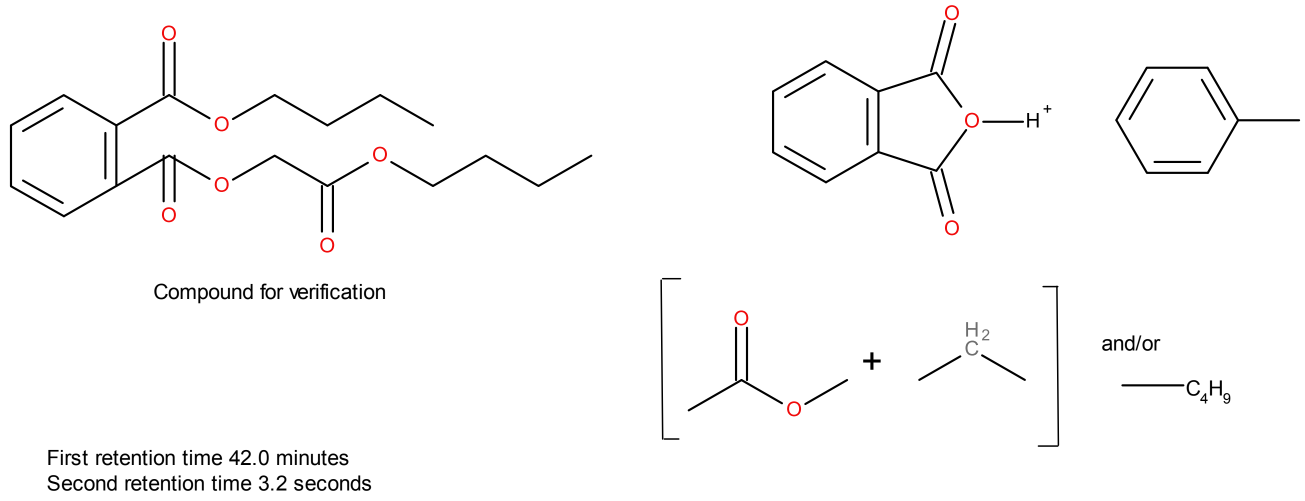

Compound for verification
First retention time 42.0 minutes
Second retention time 3.2 seconds
and/or
C_4H_9

Figure 6. Second verification example.

In this study, QSRR is a technology that has potential but still needs improvement. The accuracy of the prediction model is not good enough when using the medium polarity column, and the accuracy of the substructure generation depends on the ability of an individual.

In materials development at Asahi Kasei Corporation, 10 types of raw materials have been selected from 80 possible types and, until now, have been advanced based on human intuition and experience. There are innumerable combinations to determine the ratio of the amounts of materials. It is possible to reduce the number of trials and errors by using informatics technology such as the inverse QSAR technology described here. Using cheminformatics is a response to the era of data-driven material informatics.

Prediction and control of a vacuum deposition process by a data-driven method

Yoichi Zushi and Yuya Takeda of Kaneka Corporation discussed two different examples: the development of a soft sensor in a thin-film photovoltaic (PV) deposition process, and the development of a prediction and control method of an organic, light-emitting diode (OLED) deposition process.

A thin-film solar cell (Figure 7) is made by depositing one or more thin layers of photovoltaic material on a substrate, such as glass, plastic, or metal. Amorphous silicon is a non-crystalline, allotropic form of silicon and the most well-developed thin film technology to date. A new attempt to fuse the advantages of bulk silicon with those of thin-film devices is thin-film polycrystalline (PC) silicon on glass. These types of thin-film cell are mostly fabricated by a technique called plasma-enhanced chemical vapor deposition (CVD). Post-processing such as laser treatment or sputtering follows.

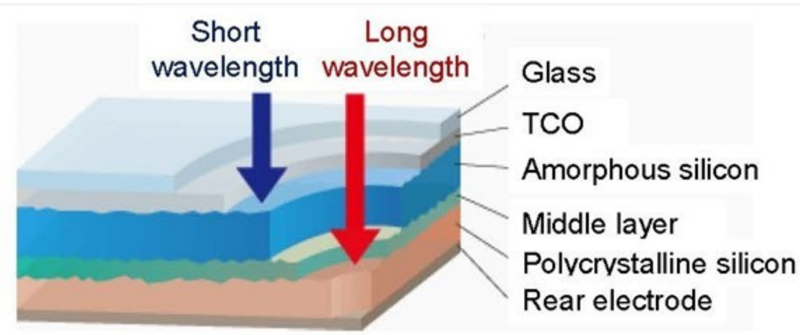


Figure 7. Thin-film silicon photovoltaic cell

PV efficiency greatly depends on the deposition conditions of silicon layers. Material gas flow rate, chamber pressure, and radio frequency power are each controlled by a device, but substrate temperature is indirectly adjusted with a heater and is affected by CVD operating status. The temperature falls sharply as polycrystalline silicon follows amorphous silicon. It is important to control the substrate temperature during deposition. In practice, the substrate temperature is difficult to measure online because there is a vacuum inside the process chamber, and there is no space for sensor installation, etc. A soft sensor was thus developed.

The modeling conditions are as follows. The response variable is substrate temperature. The data are acquired by pasting a thermocouple on the substrate, connecting it with the temperature logger, inputting it into the process chamber, and following the temperature during a few hours for the deposition batch. Explanatory variables (about 300 of them) are related to the dynamic characteristics of measurable material gas flow rate, chamber pressure, radiofrequency power, and panel heater temperature. Each process data item affects the substrate temperature with a time delay. Substrate temperature data are acquired under changing deposition conditions.

The first prediction model was created based on the time from the start of the batch, but an accurate prediction model could not be obtained due to the small amount of data. The data were therefore divided according to the features of each of the process steps, and multiple prediction models were created. The modeling method was partial least squares. The substrate temperature was then predicted accurately over time. Excellent agreement was obtained between predicted and observed values (R^2 = 0.993, RMSE = 0.90).

An online monitoring system was developed for the predicted substrate temperature. In short, the team developed a substrate temperature soft sensor during deposition with a small amount of data. The substrate temperature is predicted online, and the substrate temperature is stabilized by monitoring and control.

The second part of this presentation concerned a prediction and control method for an OLED device deposition process. The performance of OLED devices depends on the deposition process of organic materials. More than 24 hours are needed from the start of the deposition process until quality inspection, so quick feedback is prevented. There are quality specifications such as brightness, driving voltage, and color etc., but they have a trade-off relationship, so it is difficult to decide on operating conditions. The researchers developed a prediction and inverse analysis method to decide on operating conditions. The objective was to take the data from the quality inspection process, and feed them into an AI-based system to produce quality prediction and monitoring data and operating conditions that could be used in the next organic materials deposition experiment; the cycle then could be reiterated.

In the model, the explanatory variables (*x*), such as temperature and pressure, are a function of the response variables (*y*), such as voltage and brightness. Methods such as partial least squares, gradient boosting, support vector regression, ElasticNet, and random forest were used to build the model. In the inverse analysis, conditions (*x*) that satisfy the quality requirements (*y*) with the model were obtained, as were constraints on the conditions such as boundary and time. Multiobjective optimization using a genetic algorithm was then used to propose new operating conditions.

Machine learning models were obtained that were sufficiently accurate. The speaker showed good straight line plots for predicted *versus* experimental values for a number of *y* variables in training and test set prediction. In the inverse analysis and multiobjective optimization using a genetic algorithm, if the manipulated variables were randomized all at once, in some cases, the researchers failed to obtain values of *x* when *y* satisfied the requirements. This was because the OLED lighting device has a layered structure (anode, organic layer, lighting layer, organic layer, lighting layer, organic layer, cathode), and there are strongly related variables in each layer. A new method was therefore used where the variables were grouped by layer. This method worked much better. The core technology of this system can be used in other cases, and the prediction and inverse analysis method will be used for other processes in future work at Kaneka.

Designing synthesizable, bioactive compounds with chemistry-savvy machine intelligence

Gisbert Schneider of ETH Zurich, Switzerland, recently co-authored a paper[19] with Kimito Funatsu and others, which stated that QSPR and QSAR are shifting from a mere prediction of property or activity towards design. Gisbert started his talk by summarizing three approaches[20,21] to the issue in drug design of "what to make next". The chemist uses expert knowledge and intuition, with the knowledge represented both explicitly and implicitly. The "rational machine" uses rules and chemical transformations, where the knowledge representation is explicit. The "intuitive machine" uses distribution sampling, with implicit (probabilistic) knowledge representation.

Gisbert's first *de novo* design approach for small molecules, the TOPology-Assigning System (TOPAS),[22,23] was based on (al)chemical transformations. The World Drug Index was fragmented by RECAP[6] into 25,563 unique building blocks which could be recombined to make new molecules.

Later Schneider's team developed a reaction-based *de novo* design system, Design of Genuine Structures (DOGS).[24,25] The compound construction procedure explicitly considers compound synthesizability, based on a compilation of 25,144 readily available synthetic building blocks and 58 established reaction principles.[26] This enables the software to suggest a synthesis route for each designed compound. A combinatorial explosion in the structure generator is prevented by machine learning models, heuristics, and intuition. DOGS has been used successfully in the *de novo* design of small molecules as natural product mimetics. Further applications of DOGS have been published recently.[27]

In an article in the Toronto *National Post* published on May 30, 2019, and updated on June 6, 2019, Joseph Brean quotes David Gunkel, a philosopher of robotics and ethics at Northern Illinois University. Gunkel said "We are now at a point where we have AI [systems] that are not directly programmed. They develop their own decision patterns."

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.[28,29] 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 "deep-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 general RNN model was fine-tuned on recognizing retinoid X and peroxisome proliferator-activated receptor (PPAR) agonists.[29] 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.

A very recent development is a bidirectional RNN-LSTM model. In the past, SMILES strings were generated from the generation point towards the right; now the method works as in Figure 8. The novelty of valid SMILES generated was 92 ± 2 % for the unidirectional RNN-LSTM model, and 97 ± 3 % for the bidirectional RNN-LSTM model.

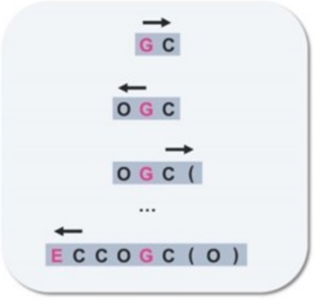


Figure 8. Bidirectional RNN-LSTM model

Schneider's team is now exploring the capability of "chemistry-savvy" machine intelligence to generate synthetically accessible molecules in Design of Innovative NCEs Generated by Optimization Strategies (DINGOS).[30] This is a virtual assembly method that combines a rule-based approach (predicting building blocks for synthesis) with a machine learning (neural network) model trained on successful synthetic routes described in chemical patent literature. This unique combination enables a balance between ligand similarity based generation of innovative compounds by scaffold hopping and the forward-synthetic feasibility of the designs. In a prospective proof-of-concept application, DINGOS successfully produced sets of *de novo* designs for four approved drugs that were in agreement with the desired structural and physicochemical properties. Target prediction indicated more than 50% of the designs to be biologically active. Four selected, computer-generated compounds were successfully synthesized in accordance with the synthetic route proposed by DINGOS. The results of this study demonstrate the capability of machine learning models to capture implicit chemical knowledge from chemical reaction data, and suggest feasible syntheses of new chemical matter.

Valid chemical structures with explicit synthesis routes are produced, in a synthesizable chemical space, with implicit reactivity scoring. The reaction forecasting involves node expansion with DINGOS, navigation with Monte Carlo Tree Search, and use of a reward, or "scoring function" (e.g., similarity to the template, or activity prediction). Gisbert coined the term DinGO in order to allude to the game "GO" played in DeepMind's Alpha Go Zero.[31] DinGO plays the "what if?" game.

Another of Gisbert's projects uses deep convolutional neural networks (CNNs). His team has published[32] a hybrid CNN approach for molecular pattern recognition in drug discovery. Using self-organizing map images of molecular pharmacophores as input,[33,34] CNN models were trained to identify C-X-C chemokine receptor type 4 (CXCR4) modulators with high accuracy. The machine learning classifier identified first-in-class, synthetic CXCR4 full-agonists. Additional macromolecular targets of the small molecules were predicted *in silico* and tested *in vitro*, revealing modulatory effects on dopamine receptors, and chemokine receptor type 1 (CCR1). These results positively advocate the applicability of molecular image recognition by CNNs to ligand-based, virtual compound screening, and demonstrate the complementarity of machine intelligence and human expert knowledge.

Gisbert concluded with some comments on the applicability of AI. We *can* expect several things from AI-driven molecular design: readily synthesizable, inspiring designs; similar success with rule-driven and data-driven AI; and better decisions for "failing early" and "choosing wisely". We cannot expect drugs from scratch, or flawless prediction models.[20,21]

Activity landscape and its application to molecular design

Kiyoshi Hasegawa of Chugai Pharmaceutical Co. and his colleagues have applied an activity landscape technique to mouse, rat and human clearance data in order to select lead compounds from huge, high-throughput screening datasets. A naïve Bayesian method with ECFP_6 fingerprints from Pipeline Pilot was used. The frequencies of active and inactive states for each substructure were counted, and structural descriptors with biased frequency were selected. An in-house set of mouse clearance data was used. Mouse liver microsome clearance, CL_{int} was measured (in µL/min/mg protein). The dataset was split into 25,000 assay results for the training set and 1000 for the test set. The threshold between stable and unstable was set to 30 µL/min/mg protein to maximize power of discrimination between stable and unstable. The prediction accuracy for the test set was 78% (true 76%, false 80%).

Activity landscape representations of different types of compound sets were calculated from potency data and pairwise compound distances in chemical space.[35] From the fingerprints, a coordinate-free chemical reference space was generated by calculation of pairwise compound distances (dissimilarities). The set of all pairwise distances defines this reference space. Then, multidimensional scaling was used to project these molecules from the coordinate-free reference space onto an *x*/*y*-plane on the basis of the chemical dissimilarities. The 2D map was then color-graded by a geographic method.

The prediction model was applied to compounds that had already been through high throughput screening, including 22 classes and 819 actives. The mouse stability of all of the primary actives was predicted, and chemical classes were identified in which all members of the class were predicted to be unstable. In this experimental validation, the result was 97% correctness for unstable compounds. These compounds fell into seven classes with no stable compounds.

Kiyoshi has built a website where the structures of up to three species can be input for prediction of stability or instability, with a prediction score. A structure is displayed, colored to show the metabolically labile and stable atoms. A talk about this was given by J.T. Metz *et al.* at the SciTegic Users Group Meeting in 2007. The colored activity landscape can be viewed with Pipeline Pilot, using an interactive link from a circle in a scatter plot to see the chemical structure and its data.

Another issue is the gap between the enzyme and cell activities. This phenomenon is often encountered in drug discovery: the cell activity of the molecule is not high even though the enzyme activity is high. Comparing two activity landscapes of the enzyme and cell activities, it is possible to investigate which molecular skeleton is a promising target for lead optimization. Since the promising chemical space can be easily detected, libraries to fill that space can be designed. Kiyoshi has constructed a prediction model for cell activity from enzyme activity, and log*D* models using Simulations Plus ADMET Predictor (https://www.simulations-plus.com/software/admetpredictor/). He showed activity landscape displays for two selected molecules (Figure 9).

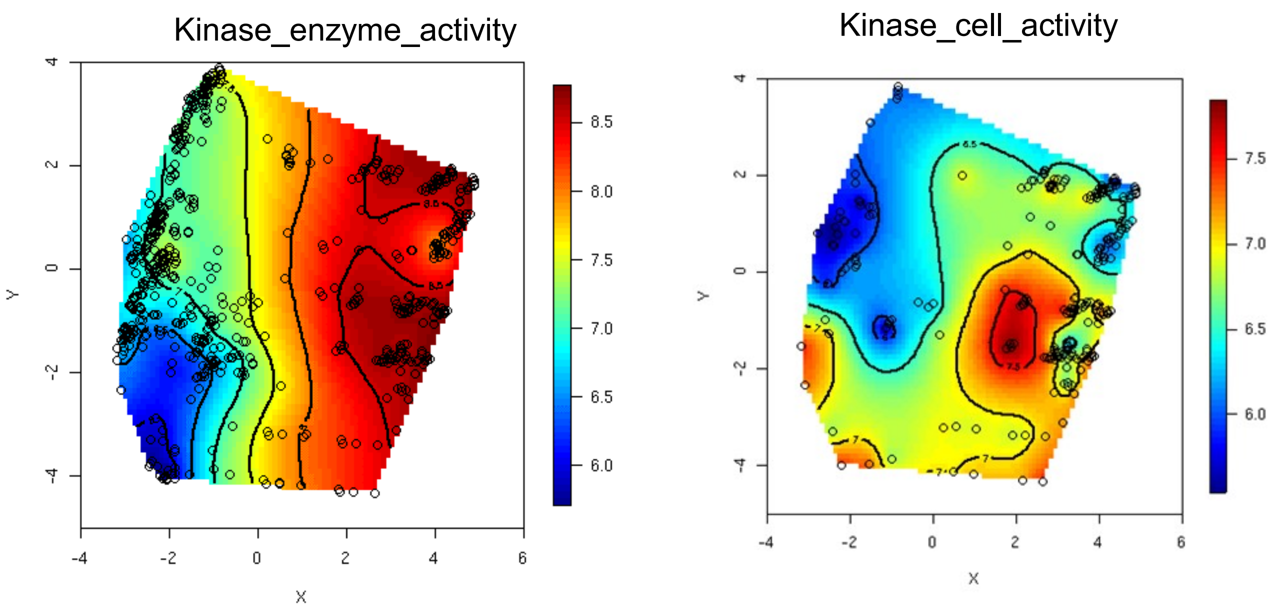


Figure 9. Activity landscapes for molecules selected from virtual library

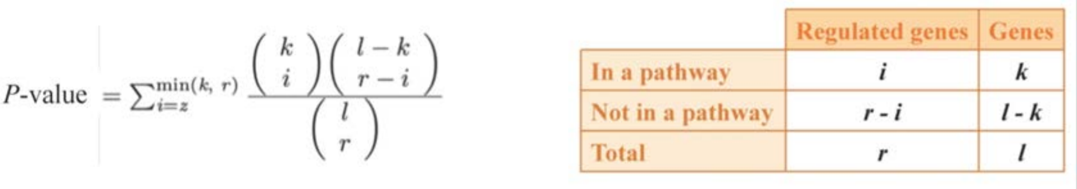
Data-driven drug discovery and medical treatment by machine learning

Drug repositioning involves identification of new therapeutic effects of existing drugs and of compounds that failed to be approved in the past. It is an efficient strategy for drug development that has attracted much attention. A great deal of information on existing drugs is available (e.g., information on safety, manufacturing processes, and pharmacokinetics). Drug repositioning can increase the success rate of drug development, and reduce the cost in terms of time, risk, and expenditure. A well-known example is sildenafil (Viagra), which was developed as a treatment for angina but was repositioned to treat erectile dysfunction and pulmonary hypertension.

Yoshihiro Yamanishi at the Kyushu Institute of Technology, and his colleagues, have developed novel machine learning methods that can be used 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. For example, the abnormal expression of phosphodiesterase type 5 (PDE5) is observed in both erectile dysfunction and pulmonary hypertension. Networks of drug-disease relationships can be produced by machine learning methods based on molecular features of drugs and diseases.

Yoshihiro has proposed a pathway-based drug discovery approach. A traditional approach is to search for drugs that regulate a single biomolecule, but, in this approach, molecular interactions between biomolecules are not taken into account. In pathway-based drug discovery[36] the approach is to search for drugs that regulate a pathway; molecular interactions are considered by using pathway information. Integration of drug-induced gene expression data with molecular network analysis can lead to prediction of new therapeutic effects of drug candidates.

Activated and inactivated pathways are identified from drug-induced gene expression signatures. The up- and down-regulated genes in the signatures are mapped onto many biological pathway maps, and the enrichment of the up- and down-regulated genes in each pathway is evaluated by pathway enrichment analysis. The 163 biological pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG, https://www.genome.jp/kegg/) were used. Now, $z = | G_{drug} \cap G_{pathway} |$ where G_{drug} denotes a set of up- or down-regulated genes in a signature induced by a drug, and $G_{pathway}$ denotes a set of genes in a pathway map. Assuming that z follows a hypergeometric distribution,[36] the probability of observing an intersection of size z between $G_{pathway}$ and G_{drug} is computed as in Figure 10. The gene expression values in the signature of each drug are represented with a feature vector.



$$P\text{-value} = \sum_{i=z}^{\min(k,\, r)} \frac{\binom{k}{i}\binom{l-k}{r-i}}{\binom{l}{r}}$$

	Regulated genes	Genes
In a pathway	i	k
Not in a pathway	$r - i$	$l - k$
Total	r	l

Figure 10. Pathway enrichment analysis

Yoshihiro has collaborated[36] with Kenzaburo Tani, at the University of Tokyo, on pathway-based drug discovery for cancers. They analyzed chemically induced gene expression data of 1112 drugs on 66 human cell lines and explored drugs that inactivate cell cycle pathways, activate p53 signaling pathways, and activate apoptosis pathways. They performed a large-scale prediction of potential anticancer effects for all the drugs and experimentally validated the results. They successfully identified several potential anticancer drugs.

Natural medicine (e.g., Kampo in Japan) is popular, but the mechanism of action of these treatments is unclear. In “ordinary” medicine the mode-of-action is based on the interaction of one compound with one target. In natural medicine, multiple compounds interact with multiple targets, and the target proteins may work cooperatively. Perhaps pathway analysis on compound-induced transcriptome data would be helpful in such cases.

There have been numerous publications[37-43] recently on identification of the modes of action of drugs, and prediction of drug therapeutic indications. It is very difficult and expensive to observe gene expression profiles experimentally for all combinations of drugs and human cell lines, so large parts of drug-induced gene expression data are unknown or unobserved. Connectivity Map (CMap), in which genes, drugs, and disease states are connected by virtue of common gene-expression signatures, was scaled up, as part of the NIH Library of Integrated Network-Based Cellular Signatures (LINCS) Consortium.[44]

A novel gene expression profiling method, L1000,[44] was used in the LINCS program, and it has opened the door to the large-scale analysis of drug-induced transcriptome data (drug profiles). However, there are far more unknown or unobserved values than known ones. This can be connected to disease profiles: disease-specific transcriptome data on highly and lowly expressed gene profiles. Alzheimer’s disease, asthma, atopic dermatitis, breast cancer, cystic fibrosis, inflammatory bowel disease, dengue, adrenoleukodystrophy, and many more diseases have been studied. For example, Wang *et al.* have reported a crowdsourcing project to annotate and reanalyze a large number of gene expression profiles from Gene Expression Omnibus (GEO).[45] A cleaned database of extracted signatures was used to visualize and analyze these signatures on the CRowd Extracted Expression of Differential Signatures (CREEDS).

Previous methods for missing value imputation or data completion[46-52] are applicable to *matrix-structured* data. Yoshihiro and co-workers have proposed a method applicable to *tensor-structured* data: Tensor-Train Weighted OPTimization (TT-WOPT).[53] They applied TT-WOPT to drug-induced transcriptome data: 16 cell lines, 261 drugs, and 978 genes represented by a 261 × 978 × 16 tensor. As a baseline method, they also tested the CP-WOPT algorithm,[54] which is a previously established tensor decomposition method applicable to data completion tasks. In the cross-validation experiments for performance evaluation, they randomly added artificial missing values to the original data before imputation. The relative standard error (RSE) between the original tensor and the one with imputed values was measured. In the case of artificial missing rates of 10% for the cell lines in total, RSE for

TT-WOPT was 0.0694 compared with RSE = 0.0750 for a nearest neighbor approach. In only one of the cell lines was the RSE for the nearest neighbor method (0.0415) better than RSE for TT-WOPT (0.0416). TT-WOPT also works well for 50% and 90% missing rates.

In Yoshihiro's work, the original drug-induced transcriptome data are subjected to tensor decomposition to get a new version, including imputed data. From the latter a drug indication prediction can be made. For comparison, three existing transcriptome-based drug repositioning methods were compared, with and without tensor decomposition: inverse signature,[38] XSum,[41] and multitask learning.[55] A benchmark dataset of 353 associations (261 drugs and 46 diseases) was used. The area under the receiver operating characteristic curve (AUC) was measured. Tensor decomposition contributed to more accurate prediction of drug indications in most cases. AUCs were more than twice as large for the multitask learning method for all 16 cell lines. Moreover, tensor decomposition is more effective in cell lines with high missing rates.

Yoshihiro gave two examples of predicted indications which have been confirmed with independent resources. Amodiaquine is an antimalarial drug. A predicted indication was pituitary adenomas, and this was confirmed by the literature.[56] Niclosamide was originally an anthelminthic. A predicted indication of adult T-cell leukemia was confirmed by the literature.[57]

Yoshihiro concluded that machine learning methods can predict new therapeutic effects of drug candidate compounds. Pathway analysis is useful for mode-of-action identification and drug discovery, and tensor decomposition for omics data contributes to enhancing the performance of drug indication prediction. Such methods will speed up the delivery of necessary drugs to patients.

Integrated cheminformatics and bioinformatics data science

Shigehiko Kanaya of the Nara Institute of Science and Technology and his colleagues have constructed a species-metabolite database for plants, the KNApSAcK Core Database,[58] which contains (as of April 2019) 51,179 metabolite entries, 22,944 species entries, and 116,315 metabolite-species pair entries. This sort of database is useful because it allows the systematic analysis of large numbers of organic compounds with known and unknown structures in metabolomics. Shigehiko's team has also developed a search engine for the database, making it possible to search for metabolites based on an accurate mass, molecular formula, metabolite name, or mass spectrum in several ionization modes. Various other databases can also be accessed on the KNApSAcK website (http://kanaya.naist.jp/KNApSAcK_Family/), and the search engine can be downloaded (Figure 11).

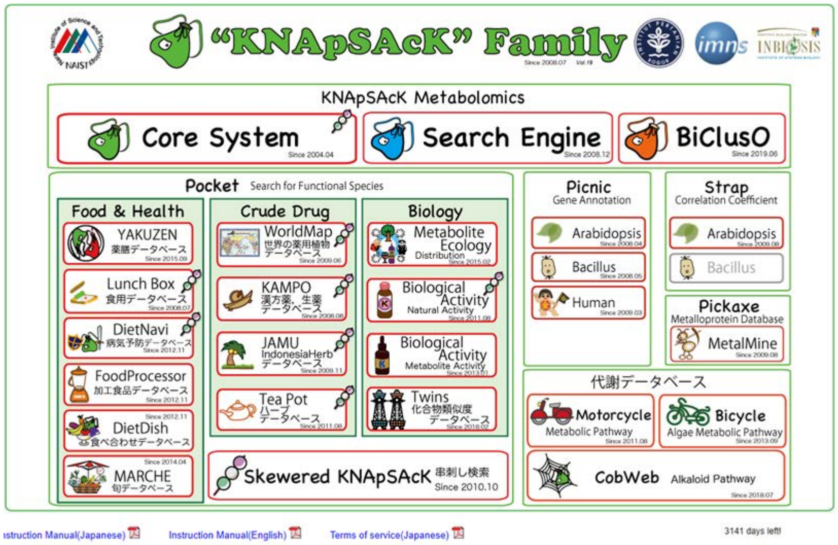

"KNApSAcK" Family
Since 2008.07 Ver.19
KNApSAcK Metabolomics
Core System Since 2004.04
Search Engine Since 2008.12
BiClusO Since 2019.06
Pocket Search for Functional Species
Food & Health
YAKUZEN 薬膳データベース Since 2015.09
Lunch Box 食用データベース Since 2008.07
DietNavi 病気予防データベース Since 2012.11
FoodProcessor 加工食品データベース Since 2012.11
DietDish 食べ合わせデータベース Since 2012.11
MARCHE 旬データベース Since 2014.04
Crude Drug
WorldMap 世界の薬用植物データベース Since 2009.06
KAMPO 漢方薬、生薬データベース Since 2008.08
JAMU IndonesiaHerb データベース Since 2009.11
Tea Pot ハーブデータベース Since 2011.08
Biology
Metabolite Ecology Distribution Since 2015.02
Biological Activity Natural Activity Since 2011.08
Biological Activity Metabolite Activity Since 2013.01
Twins 化合物類似度データベース Since 2018.02
Skewered KNApSAcK 串刺し検索 Since 2010.10
Picnic Gene Annotation
Arabidopsis Since 2008.04
Bacillus Since 2008.05
Human Since 2009.03
Strap Correlation Coefficient
Arabidopsis Since 2009.08
Bacillus
Pickaxe Metalloprotein Database
MetalMine Since 2009.08
代謝データベース
Motorcycle Metabolic Pathway Since 2011.08
Bicycle Algae Metabolic Pathway Since 2013.09
CobWeb Alkaloid Pathway Since 2018.07
ıstruction Manual(Japanese)
Instruction Manual(English)
Terms of service(Japanese)
3141 days left!

Figure 11. KNApSAcK screen shot

Shigehiko believes that data science can be created by integrating disciplines including theoretical understanding in various research fields, informatics (for data systemization, model construction, and prediction), and statistics (for validation). Specifically, chemistry and chemical physics could be overlapped with chemical information (molecular structures), and chemometrics validation. An example is the overlap of deuterium isotope effects in solvolysis reactions,[59-61] AIPHOS,[2] and computer-aided structure elucidation (CHEMICS[1] and soft sensors[4,62]). Kimito Funatsu sits at the center of the overlap of such systems.

The role of data science is in moving from vertical relationships to horizontal relationships (Figure 12), standardizing mining techniques. The KNApSAcK family (Figure 11) is an example, allowing the understanding of biology based on natural products databases.

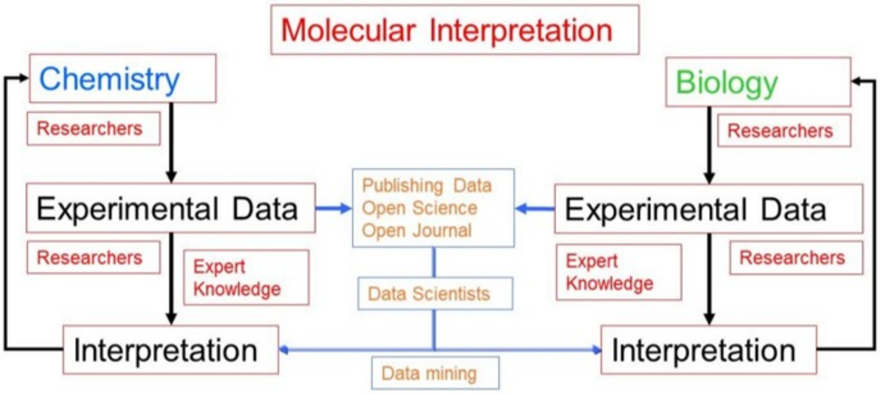


Figure 12. Data science

There are about 20,000 alkaloids in the KNApSAcK database, but few of their biosynthesis pathways are fully identified. Shigehiko and his co-workers have constructed a model to predict the precursors of alkaloids based on multi-graph convolutional neural networks (MGCNN).[63] It is sometimes difficult for current fingerprint representations to emphasize specific features for target problems efficiently. It is advantageous to allow the model to select the appropriate features according to data-driven decisions. By encoding a molecule as an abstract graph, applying "convolution" on the graph, and training the weight of the neural network framework, the neural network can optimize feature selection for the training problem. By incorporating the effects from adjacent atoms recursively, graph convolutional neural networks can extract the features of latent atoms that represent chemical features of a molecule efficiently. The researchers trained the network to distinguish the precursors of 566 alkaloids, which are almost all of the alkaloids with known biosynthesis pathways, and showed that the model could predict starting substances with an average accuracy of 97.5%. The prediction of pathways contributes to understanding of alkaloid synthesis mechanisms and the application of graph based neural network models to similar problems in bioinformatics would therefore be beneficial.

Development of data-driven chemistry in chemistry and chemical engineering

Cheminformatics has been applied to various areas of chemistry: molecular design, materials design, organic synthesis design, structure elucidation and process control. Kimito Funatsu presented an overview of these applications during his research life. In pursuit of a desired function, a novel compound, material, or device is required. The first step in producing one is to decide what to make. Designing the molecule or material may involve modeling, inverse analysis, or data analysis. The next step is deciding how to make the product, and this may involve synthesis design or product prediction. A production process is then needed to make a commercial product. Reaction, separation, and refinement are carried out in the chemical plant, and, in order to produce the product with the desired property, quality control of the process is important. The fourth step is analysis, which may require structure determination.

Finally, after the products are provided to the public, methods for recycling and reuse are also required. These five units are important subjects in cheminformatics. Knowledge to support them is created from many kinds of data and information.

This knowledge has to be organized, by data modeling, and used for prediction and design. Structure-property (or activity) relationship models can be constructed, and candidate molecules or materials can be generated that satisfy the desired property, by "inverse analysis" (see Figure 13). The generation of candidate structures controlled by the model is the driving force for *de novo* design (in drug discovery), design of highly functional polymers (including monomer design), and catalyst design. Developing a structure generator is challenging; even for a molecular formula as simple as C_6H_6, there are 217 possible isomers.

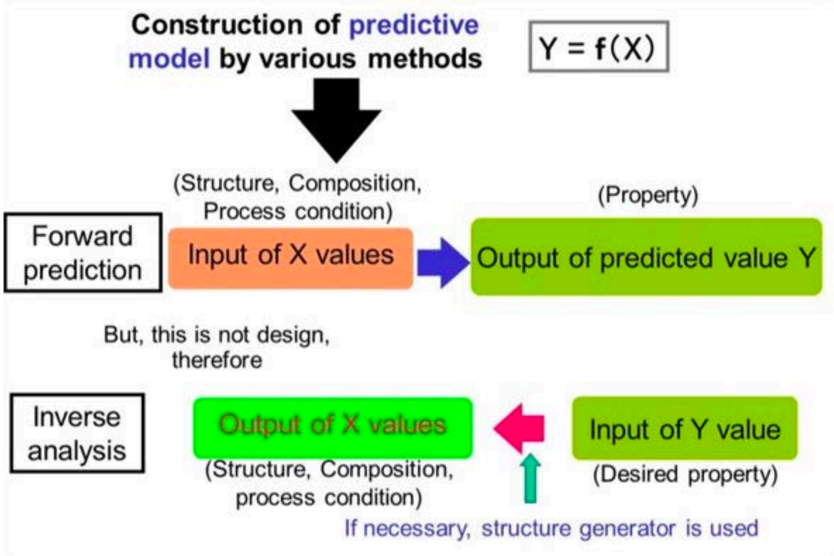

Construction of predictive model by various methods
Y = f(X)
(Structure, Composition, Process condition)
(Property)
Forward prediction
Input of X values
Output of predicted value Y
But, this is not design, therefore
Inverse analysis
Output of X values
Input of Y value
(Structure, Composition, process condition)
(Desired property)
If necessary, structure generator is used

Figure 13. Novel molecule and material design

In the first stage of development of new drugs, various lead compounds with high activity are required. To design such compounds, Kimito and co-workers have focused on chemical space defined by structural descriptors. New compounds close to areas where highly active compounds exist will show the same degree of activity. Visualization of chemical space is useful for understanding activity distribution in chemical space and determination of the target area for structure search by activity distribution. Structures in chemical space are described by many descriptors, giving rise to high dimensional chemical space. This is projected onto a 2D plane by generative topographic mapping. The activity is displayed as a heat map on this 2D plane. Thus target areas for structure search can be assigned.

Kimito's team has developed a new *de novo* design system[64] to search a target area. First, highly active compounds are manually selected as initial seeds. Then, the seeds are entered into the system, and structures slightly different from the seeds are generated and pooled. Next, seeds are selected from the new structure pool based on the distance from target coordinates on the map. Activity distribution and druglikeness can be visualized on the same map, and the target area selected by considering overlap. The initial *de novo* design system for exploring chemical space (DAECS) was modified[65] to enable the user to select a target area to consider properties other than activity, and improve the diversity of the generated structures by visualizing the druglikeness distribution and the activity distribution, generating structures by substructure-based structural changes, including addition, deletion, and substitution of substructures, as well as the slight structural changes used in DAECS. Through case studies using ligand data for the human adrenergic alpha2A receptor and the human histamine H1 receptor, it has been shown that the modified DAECS can generate high diversity druglike structures, and the usefulness of the modification of the DAECS has been verified.

In the recent study,[65] where the target protein was the histamine H1 receptor, the training data were 522 structures and pK_i values selected from ChEMBL, and the descriptors were 142 fingerprints from PubChem. The training data for construction of the discriminant model were 1000 structures from BIOVIA Comprehensive Medicinal Chemistry (http://accelrys.co.jp/products/databases) and 1000 non-drug structures from the BIOVIA Available Chemicals Directory (http://accelrys.co.jp/products/databases). The modeling method was support vector machine. Kimito showed visualization with structure generation (Figure 14) and some generated structures (Figure 15).

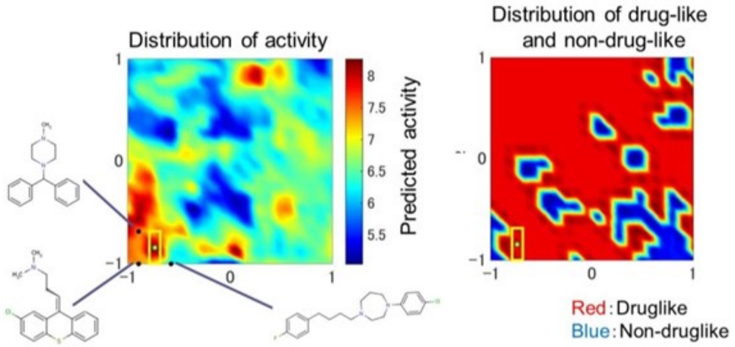


Figure 14. Structure generation in high activity and druglike areas

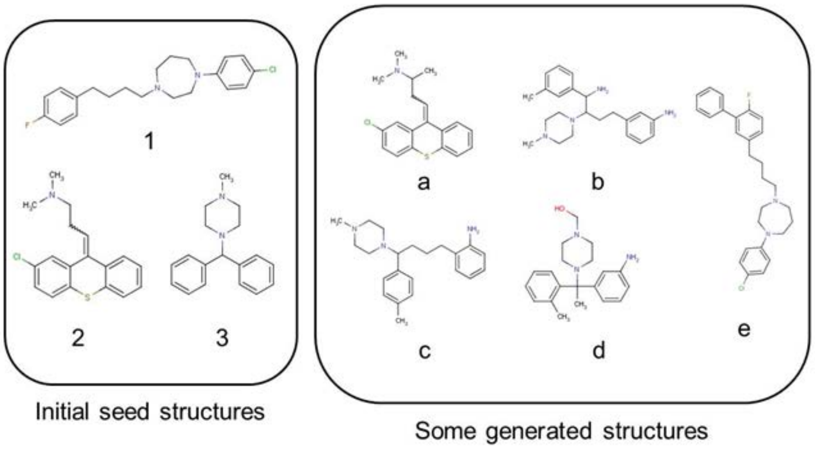


Figure 15. Structure generation.

Kimito next discussed polymer alloys, a class of polymer blends where addition of a second polymer is tailored to provide controlled morphology and thus specific performance characteristics. Polymer alloys can be produced by mixing, melting, and crystalizing a mixture of multiple polymers, then by molding, melting, and crystalizing the mixture. Data items include the properties of each component polymer, and the mixing conditions, molding conditions, and alloy properties. There can be 100-200 data items.

Kimito has also worked on the design of more efficient polymeric optical films (Figure 16) that manage the polarization of light. He explained the mechanism of polarizing transmittance and reflection, and its relationship to polymer orientation in machine direction (MD), and polymer orientation in transverse direction (TD).

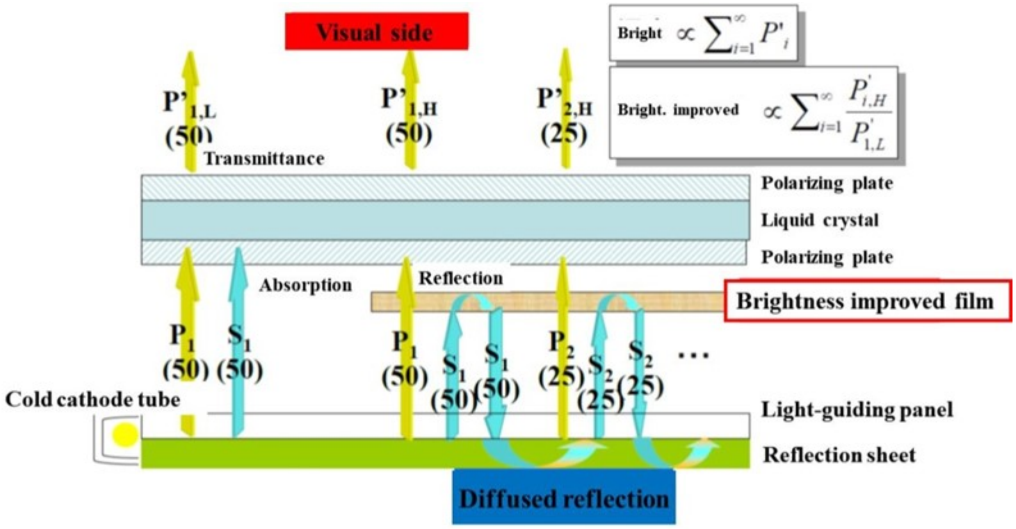


Figure 16. Brightness-improved film

His objective was to construct a quantitative model of the properties of light improved film and to design a more efficient film. He aimed to optimize process conditions such as extrusion and to achieve brightness (in cd/m^2) ≥5400, MD transmittance ≥82%, and TD transmittance ≤20%. Object variables were brightness, MD transmittance, and TD transmittance. Explanatory variables were composition (percentages of polyethylene naphthalate, polyethylene terephthalate, and polystyrene), percentages of three compatibilizing agents, and process conditions (stretching temperature, extrusion machine ID (1 or 2), stretching magnification, and thickness). The number of samples was 26. The results of partial least square analysis were excellent: brightness $R^2 = 0.916$, $Q^2 = 0.682$, MD transmittance $R^2 = 0.977$, $Q^2 = 0.920$, TD transmittance $R^2 = 0.930$, $Q^2 = 0.746$.

Kimito's final topic was process control. In operating chemical plants, operators have to monitor the operating condition of the plants and control process variables. So, process variables such as temperature, pressure, liquid level, and concentration of products need to be measured online, but none of them is easy to measure online because of technical difficulties, large measurement delays, high investment cost, and so on. In order to cope with this problem, soft sensors are widely used in chemical plants. Soft sensors are inferential models constructed between easy-to-measure variables, such as temperature and pressure, and variables that are difficult to measure online, such as concentration or property. By inputting temperature and pressure variables to soft sensor models, the soft sensor can estimate property and concentration variables online with high accuracy. Thus, the process operator can obtain, and use the predicted values for process control in real time.

Kimito returned to his initial theme of the steps in cheminformatics, the first step being what to make and the second being design. In the design step, a structure-property relationship model is constructed for inverse analysis, to generate candidate structures or materials. It is important to incorporate process parameters into the modeling step. How to make the material is considered at the same time. This is an important concept in materials design, because the property of materials is strongly affected by process conditions, even for the same starting materials. Eventually, a production process is needed to make a commercial product. Here process control, namely quality control of the product, is particularly important. In this step, the quality of the product is monitored by a soft sensor online, and the quality is controlled by operating process parameters. Simultaneous consideration of quality control can realize integrated treatment of materials design, examination of process conditions, and quality control. Kimito emphasizes this concept as process informatics.

Conclusion

Elsa Alvaro, chair of the ACS Division of Chemical Information, formally presented the Herman Skolnik Award to Kimito Funatsu at the end of the symposium.

Kimito Funatsu, upon receipt of the Herman skolnik Award, with Elsa Alvaro

References

(1) Funatsu, K.; Sasaki, S.-i. Recent Advances in the Automated Structure Elucidation System, CHEMICS. Utilization of Two-Dimensional NMR Spectral Information and Development of Peripheral Functions for Examination of Candidates. *J. Chem. Inf. Comput. Sci.* **1996,** *36* (2), 190 -204.

(2) Funatsu, K.; Sasaki, S. Computer-assisted organic synthesis design and reaction prediction system, "AIPHOS". *Tetrahedron Comput. Methodol.* **1988,** *1* (1), 27-37.

(3) Satoh, K.; Funatsu, K. A novel approach to retrosynthetic analysis using knowledge bases derived from reaction databases. *J. Chem. Inf. Comput. Sci.* **1999,** *39* (2), 316-325.

(4) Funatsu, K. Process Control and Soft Sensors. In *Applied Chemoinformatics;* Engel, T., Gasteiger, J., Eds.; Wiley-VCH: Weinheim, Germany, 2018; pp 571-584.

(5) Vogt, M.; Yonchev, D.; Bajorath, J. Computational Method to Evaluate Progress in Lead Optimization. *J. Med. Chem.* **2018,** *61* (23), 10895-10900.

(6) 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.

(7) Yonchev, D.; Vogt, M.; Stumpfe, D.; Kunimoto, R.; Miyao, T.; Bajorath, J. Computational Assessment of Chemical Saturation of Analogue Series under Varying Conditions. *ACS Omega* **2018,** *3* (11), 15799-15808.

(8) Kunimoto, R.; Miyao, T.; Bajorath, J. Computational method for estimating progression saturation of analog series. *RSC Adv.* **2018,** *8* (10), 5484-5492.

(9) Nakatsuji, H.; Sugimoto, M. Theoretical study on the metal NMR chemical shift. Molybdenum complexes. *Inorg. Chem.* **1990,** *29* (6), 1221-5.

(10) Sakanoue, K.; Motoda, M.; Sugimoto, M.; Sakaki, S. A Molecular Orbital Study on the

Hole Transport Property of Organic Amine Compounds. *J. Phys. Chem. A* **1999,** *103* (28), 5551-5556.

(11) Sugimoto, M.; Yamasaki, I.; Mizoe, N.; Anzai, M.; Sakaki, S. Acetylene-insertion reactions into Pt(II)-H and Pt(II)-SiH3 bonds. An ab-initio MO study and analysis based on the vibronic coupling model. *Theor. Chem. Acc.* **1999,** *102* (1-6), 377-384.

(12) Sugimoto, M.; Sakaki, S.; Sakanoue, K.; Newton, M. D. Theory of emission state of tris(8-quinolinolato)aluminum and its related compounds. *J. Appl. Phys.* **2001,** *90* (12), 6092-6097.

(13) Cronin, M. T. D.; Aptula, A. O.; Duffy, J. C.; Netzeva, T. I.; Rowe, P. H.; Valkova, I. V.; Wayne Schultz, T. Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to Tetrahymena pyriformis. *Chemosphere* **2002,** *49* (10), 1201-1221.

(14) Funatsu, K. Computer-aided synthesis design and reaction prediction. *Kagaku Kogyo* **2007,** *58* (2), 124-129.

(15) Hori, K.; Sadatomi, H.; Miyamoto, A.; Kuroda, T.; Sumimoto, M.; Yamamoto, H. Towards the development of synthetic routes using theoretical calculations: an application of in silico screening to 2,6-dimethylchroman-4-one. *Molecules* **2010,** *15*, 8289-8304.

(16) Yamamoto, H.; Yamaguchi, T.; Yoshimura, K.; Sumimoto, M.; Hori, K. Towards in silico Synthetic Route Development. Computer Aided Organic Synthesis Developments of Target Compounds. *J. Synth Org. Chem.* **2012,** *70* (7), 722-730.

(17) Hori, K.; Sumimoto, M.; Murafuji, T. Quantum chemistry-assisted synthesis route development. *AIP Conf. Proc.* **2015,** *1702* (1, International Conference of Computational Methods in Sciences and Engineering (ICCMSE-2015)), 090019/1-090019/3.

(18) *Structure and Retention in Chromatography: A Chemometric Approach,* Kaliszan, R., Ed.; Harwood Academic Publishers: Amsterdam, The Netherlands, 1997.

(19) Schneider, G.; Funatsu, K.; Okuno, Y.; Winkler, D. De novo Drug Design - Ye olde Scoring Problem Revisited. *Mol. Inf.* **2017,** *36* (1-2), 1700030.

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

(21) Schneider, G. Mind and machine in drug design. *Nat. Mach. Intell.* **2019,** *1* (3), 128-130.

(22) 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.

(23) 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.

(24) 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.

(25) Schneider, G. De novo design - hop(p)ing against hope. *Drug Discovery Today Technol.* **2013,** *10* (4), e453-e460.

(26) Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K.-H.; Schneider, G.; Jacoby, E.; Renner, S. Probing the Bioactivity-Relevant Chemical Space of Robust Reactions and Common Molecular Building Blocks. *J. Chem. Inf. Model.* **2012,** *52* (5), 1167-1178.

(27) Schneider, G.; Clark, D. E. Automated de novo drug design: are we nearly there yet? *Angew. Chem., Int. Ed.* **2019,** *58* (32), 10792-10803.

(28) 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.

(29) 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.

(30) Button, A.; Merk, D.; Hiss, J. A.; Schneider, G. Automated de novo molecular design by hybrid machine intelligence and rule-driven chemical synthesis. *Nat. Mach. Intell.* **2019,** *1* (7), 307-315.

(31) Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; Chen, Y.; Lillicrap, T.; Hui, F.; Sifre, L.; van den Driessche, G.; Graepel, T.; Hassabis, D. Mastering the game of Go without human knowledge. *Nature* **2017,** *550* (7676), 354-359.

(32) Bruns, D.; Gawehn, E.; Kumar, K. S.; Schneider, P.; Baumgartner, M.; Schneider, G. Identification of synthetic activators of cancer cell migration by hybrid deep learning. *ChemBioChem* **2019**, Ahead of print.

(33) 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.

(34) Schneider, P.; Mueller, A. T.; Gabernet, G.; Button, A. L.; Posselt, G.; Wessler, S.; Hiss, J. A.; Schneider, G. Hybrid Network Model for "Deep Learning" of Chemical Data: Application to Antimicrobial Peptides. *Mol. Inf.* **2017,** *36* (1-2), 1600011.

(35) Peltason, L.; Iyer, P.; Bajorath, J. Rationalizing Three-Dimensional Activity Landscapes and the Influence of Molecular Representations on Landscape Topology and the Formation of Activity Cliffs. *J. Chem. Inf. Model.* **2010,** *50* (6), 1021-1033.

(36) Iwata, M.; Hirose, L.; Kohara, H.; Liao, J.; Sawada, R.; Akiyoshi, S.; Tani, K.; Yamanishi, Y. Pathway-Based Drug Repositioning for Cancers: Computational Prediction and Experimental Validation. *J. Med. Chem.* **2018,** *61* (21), 9583-9595.

(37) Lamb, J.; Crawford, E. D.; Peck, D.; Modell, J. W.; Blat, I. C.; Wrobel, M. J.; Lerner, J.; Brunet, J.-P.; Subramanian, A.; Ross, K. N.; Reich, M.; Hieronymus, H.; Wei, G.; Armstrong, S. A.; Haggarty, S. J.; Clemons, P. A.; Wei, R.; Carr, S. A.; Lander, E. S.; Golub, T. R. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science (Washington, DC, U. S.)* **2006,** *313* (5795), 1929-1935.

(38) Dudley, J. T.; Sirota, M.; Shenoy, M.; Pai, R. K.; Roedder, S.; Chiang, A. P.; Morgan, A. A.; Sarwal, M. M.; Pasricha, P. J.; Butte, A. J. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci. Transl. Med.* **2011,** *3* (96), 96ra76.

(39) Kosaka, T.; Nagamatsu, G.; Saito, S.; Oya, M.; Suda, T.; Horimoto, K. Identification of drug candidate against prostate cancer from the aspect of somatic cell reprogramming. *Cancer Sci.* **2013,** *104* (8), 1017-1026.

(40) Cheng, J.; Xie, Q.; Kumar, V.; Hurle, M.; Freudenberg, J. M.; Yang, L.; Agarwal, P. Evaluation of analytical methods for connectivity map data. *Pac. Symp. Biocomput.* **2013**, 5-16.

(41) Cheng, J.; Yang, L.; Kumar, V.; Agarwal, P. Systematic evaluation of connectivity map for disease indications. *Genome Med* **2014,** *6* (12), 540.

(42) van Noort, V.; Schoelch, S.; Iskar, M.; Zeller, G.; Ostertag, K.; Schweitzer, C.; Werner, K.; Weitz, J.; Koch, M.; Bork, P. Novel Drug Candidates for the Treatment of Metastatic Colorectal Cancer through Global Inverse Gene-Expression Profiling. *Cancer Res.* **2014,** *74* (20), 5690-5699.

(43) 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.

(44) Subramanian, A.; Narayan, R.; Corsello, S. M.; Peck, D. D.; Natoli, T. E.; Lu, X.; Gould, J.; Davis, J. F.; Tubelli, A. A.; Asiedu, J. K.; Lahr, D. L.; Hirschman, J. E.; Liu, Z.; Donahue, M.; Julian, B.; Khan, M.; Wadden, D.; Smith, I. C.; Lam, D.; Liberzon, A.; Toder, C.; Bagul, M.; Orzechowski, M.; Enache, O. M.; Piccioni, F.; Johnson, S. A.; Lyons, N. J.; Berger, A. H.; Shamji, A. F.; Brooks, A. N.; Vrcic, A.; Flynn, C.; Rosains, J.; Takeda, D. Y.; Hu, R.; Davison, D.; Lamb, J.; Ardlie, K.; Hogstrom, L.; Greenside, P.; Gray, N. S.; Clemons, P. A.; Silver, S.; Wu, X.; Zhao, W.-N.; Read-Button, W.; Wu, X.; Haggarty, S. J.; Ronco, L. V.; Boehm, J. S.; Schreiber, S. L.; Doench, J. G.; Bittker, J. A.; Root, D. E.; Wong, B.; Golub, T. R. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell (Cambridge, MA, U. S.)* **2017,** *171* (6), 1437-1452.e17.

(45) Wang, Z.; Monteiro, C. D.; Jagodnik, K. M.; Fernandez, N. F.; Gundersen, G. W.; Rouillard, A. D.; Jenkins, S. L.; Feldmann, A. S.; Hu, K. S.; McDermott, M. G.; Duan, Q.; Clark, N. R.; Jones, M. R.; Kou, Y.; Goff, T.; Woodland, H.; Amaral, F. M. R.; Szeto, G. L.; Fuchs, O.; Schussler-Fiorenza Rose, S. M.; Sharma, S.; Schwartz, U.; Bausela, X. B.; Szymkiewicz, M.; Maroulis, V.; Salykin, A.; Barra, C. M.; Kruth, C. D.; Bongio, N. J.; Mathur, V.; Todoric, R. D.; Rubin, U. E.; Malatras, A.; Fulp, C. T.; Galindo, J. A.; Motiejunaite, R.; Juschke, C.; Dishuck, P. C.; Lahl, K.; Jafari, M.; Aibar, S.; Zaravinos, A.; Steenhuizen, L. H.; Allison, L. R.; Gamallo, P.; Segura, F. d. A.; Dae Devlin, T.; Perez-Garcia, V.; Ma'ayan, A. Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nat. Commun.* **2016,** *7*, 12846.

(46) Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R. B. Missing value estimation methods for DNA microarrays. *Bioinformatics* **2001,** *17* (6), 520-525.

(47) Oba, S.; Sato, M.-a.; Takemasa, I.; Monden, M.; Matsubara, K.-i.; Ishii, S. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* **2003,** *19* (16), 2088-2096.

(48) Ouyang, M.; Welsh, W. J.; Georgopoulos, P. Gaussian mixture clustering and imputation of microarray data. *Bioinformatics* **2004,** *20* (6), 917-923.

(49) Bø, T. H.; Dysvik, B.; Jonassen, I. LSimpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Res.* **2004,** *32* (3), e34-e34.
(50) Kim, H.; Golub, G. H.; Park, H. Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics* **2004,** *21* (2), 187-198.
(51) Cai, Z.; Heydari, M.; Lin, G. Iterated local least squares microarray missing value imputation. *J. Bioinf. Comput. Biol.* **2006,** *04* (05), 935-957.
(52) Wang, X.; Li, A.; Jiang, Z.; Feng, H. Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme. *BMC Bioinf.* **2006,** *7*, 32.
(53) Iwata, M.; Berenger, F.; Sawada, R.; Hamano, M.; Yamanishi, Y.; Yuan, L.; Yuan, L.; Zhao, Q.; Tabei, Y.; Zhao, Q.; Akiyoshi, S.; Yamanishi, Y. Predicting drug-induced transcriptome responses of a wide range of human cell lines by a novel tensor-train decomposition algorithm. *Bioinformatics* **2019,** *35* (14), i191-i199.
(54) Acar, E.; Dunlavy, D. M.; Kolda, T. G.; Mørup, M. Scalable tensor factorizations for incomplete data. *Chemom. Intell. Lab. Syst.* **2011,** *106* (1), 41-56.
(55) Sawada, R.; Iwata, M.; Yamato, H.; Yamanishi, Y.; Tabei, Y.; Yamanishi, Y. Predicting inhibitory and activatory drug targets by chemically and genetically perturbed transcriptome signatures. *Sci. Rep.* **2018,** *8* (1), 156.
(56) Salentin, S.; Adasme, M. F.; Heinrich, J. C.; Haupt, V. J.; Daminelli, S.; Zhang, Y.; Schroeder, M. From malaria to cancer: Computational drug repositioning of amodiaquine using PLIP interaction patterns. *Sci. Rep.* **2017,** *7* (1), 1-13.
(57) Xiang, D.; Yuan, Y.; Chen, L.; Liu, X.; Belani, C.; Cheng, H. Niclosamide, an anti-helminthic molecule, downregulates the retroviral oncoprotein Tax and pro-survival Bcl-2 proteins in HTLV -1-transformed T lymphocytes. *Biochem. Biophys. Res. Commun.* **2015,** *464* (1), 221-228.
(58) Afendi, F. M.; Okada, T.; Yamazaki, M.; Hirai-Morita, A.; Nakamura, Y.; Nakamura, K.; Ikeda, S.; Takahashi, H.; Altaf-Ul-Amin, M.; Darusman, L. K.; Saito, K.; Kanaya, S. KNApSAcK Family Databases: Integrated Metabolite-Plant Species Databases for Multifaceted Plant Research. *Plant Cell Physiol.* **2012,** *53* (2), e1.
(59) Funatsu, K.; Fujio, M.; Tsuno, Y. Deuterium isotope effects on solvolyses of aralkyl systems. Acetolyses of p-methoxy and unsubstituted 1-phenyl-2-propyl tosylates. *Mem. Fac. Sci., Kyushu Univ., Ser. C* **1981,** *13* (1), 125-34.
(60) Funatsu, K.; Fujio, M.; Tsuno, Y. Deuterium isotope effects on solvolyses of aralkyl systems. II. Solvent dependence of α-deuterium isotope effects in the solvolysis of 2-adamantyl tosylate. *Mem. Fac. Sci., Kyushu Univ., Ser. C* **1982,** *13* (2), 391-6.
(61) Funatsu, K.; Kimura, M.; Furukawa, T.; Fujio, M.; Tsuno, Y. Deuterium isotope effects on solvolyses of aralkyl systems. III. Deuterium isotope effects on the acetolysis of β-arylethyl tosylates. *Mem. Fac. Sci., Kyushu Univ., Ser. C* **1984,** *14* (2), 343-54.
(62) Funatsu, K. Soft Sensors: Chemoinformatic Model for Efficient Control and Operation in Chemical Plants. *Mol. Inf.* **2016,** *35* (11-12), 549-554.
(63) Eguchi, R.; Ono, N.; Hirai Morita, A.; Katsuragi, T.; Nakamura, S.; Huang, M.; Altaf-Ul-Amin, M.; Kanaya, S. Classification of alkaloids according to the starting substances of their biosynthetic pathways using graph convolutional neural networks. *BMC Bioinf.* **2019,** *20* (1), 1-13.
(64) 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.
(65) Takeda, S.; Kaneko, H.; Funatsu, K. Chemical-Space-Based de Novo Design Method To Generate Drug-Like Molecules. *Journal of Chemical Information and Modeling* **2016,** *56* (10), 1885-1893.

Wendy A. Warr
Wendy Warr & Associates
wendy@warr.com