

"Chem-Bio Informatics Society (CBI) Annual Meeting 2024"

Modeling studies and databases for generating diversity

2024
10.28^{MON} - 10.31^{THU}

Tower Hall Funabori
 4-1-1Funabori, Edogawa-ku, Tokyo

President: Kenji Mizuguchi
 (Osaka Univ.)

Organizing Committee Chair: Yayoi Natsume
 (NIBIOHN)

Program Chair: Masakazu Sekijima
 (Tokyo Inst. of Tech.)

Mathematical formulas visible in the background include:
 $e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$
 $f(x) = a_0 + \sum_{n=1}^{\infty} \left(d_n \frac{n\pi x}{2} \right)$
 $(1+x)^n = 1 + \frac{nx}{1!} + \frac{n(n-1)x^2}{2!} + \dots$
 $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$
 $\cos\alpha + \cos\beta = 2\cos\frac{\alpha+\beta}{2}\cos\frac{\alpha-\beta}{2}$
 $a^2 + b^2 = c^2$
 $A = \pi r^2$

Chemical structures visible in the background include:
 $\text{H}_2\text{C}=\text{CH}_2$
 $\text{H}-\text{C}-\text{CH}_3$
 COOH
 NH_2
 $\text{H}-\text{C}-\text{H}$
 COOH

Chem-Bio Informatics Society(CBI)
 Annual Meeting 2024

Abstracts

Abstracts

Oral Presentation

001-01

Unveiling the Potential of RNA-Targeted Small Molecule Therapies: Innovations in Computational and Biophysical Approaches

Ella MORISHITA^{*1}, Amiu SHINO¹, Maina OTSU¹, Koji IMAI¹, Kaori FUKUZAWA²

¹ Basic Research Division, Veritas In Silico Inc.

² Graduate School of Pharmaceutical Sciences, Osaka University

(* E-mail: ecm@vi14si.com)

RNA molecules possess intricate structures that allow them to carry out a variety of roles in human biology and disease, making them attractive targets for small molecule therapies. The FDA-approval of risdiplam, an RNA splicing modulator to treat spinal muscular atrophy, exemplifies the potential of targeting RNA structures with small molecules. However, challenges persist, including the need for advanced computational and biophysical techniques essential for efficient drug discovery. To address these challenges, we are pioneering efforts to develop computational and biophysical technologies that would deepen our understanding of RNA structures and their interactions with small molecules [1].

Initially, we identify RNA structures suitable for targeting with small molecules using our proprietary RNA secondary structure prediction and structural analysis software. Subsequently, we employ quantitative fluorescence resonance energy transfer (qFRET) to screen small-molecule libraries against the target RNA. After primary screening with qFRET, we conduct orthogonal assays, including (1) biolayer interferometry (BLI) for determining binding kinetics; (2) 1D nuclear magnetic resonance (1D NMR) for identifying key residues on the RNA target that bind small molecules, and (3) isothermal titration calorimetry (ITC) for measuring the thermodynamic parameters of binding.

Upon identifying hit small molecules, we determine their 3D structures in complex with the target RNA using NMR or X-ray crystallography. Following successful 3D structure determination, we analyze the important interactions in the RNA–small molecule complex using the fragment molecular orbital (FMO) method. This information guides medicinal chemists in designing derivatives with enhanced pharmacological activity. Finally, we determine the correlation between the experimentally and computationally derived binding energies and exploit the correlation to rationally design derivatives of the hit molecules.

By integrating state-of-the-art computational analysis with rigorous biophysical

characterization, our approach offers a comprehensive framework for the discovery and optimization of RNA-targeted small molecules, driving innovation in drug development especially for the treatment of diseases of unmet medical needs. In this presentation, I will illustrate our successful efforts in identifying fluoroquinolone compounds that bind to an RNA stem loop structure, A4G, and demonstrate the potential of FMO-guided design to develop more potent RNA-targeted small molecules [2].

[1] Morishita, E. C. Discovery of RNA-targeted small molecules through the merging of experimental and computational technologies, *Expert Opin Drug Discov*, 2023, 18, 207–226.

[2] Shino, A.; Otsu, M.; Imai, K.; Fukuzawa, K.; Morishita, E. C. Probing RNA–small molecule interactions using biophysical and computational approaches, *ACS Chem Biol*, 2023, 18, 2368–2376.

001-02

Computational Chemistry and Structure-Based Molecular Design for a Cyclic Peptide Drug Discovery Platform

Atsushi MATSUO*

Research Division, Chugai Pharmaceutical Co.,Ltd

(* E-mail: matsuoats@chugai-pharm.co.jp)

The construction of a technological platform for molecules inhibiting intracellular protein-protein interactions (PPIs) can be a game changer for drug discovery. Although small molecules and antibodies are major modalities, they are not suitable for target proteins that lack a deep cavity for small molecule binding or proteins located in intracellular areas beyond the reach of antibodies. One potential solution is to use membrane-permeable mid-sized cyclic peptides (defined here as those with a molecular weight of 1000-2000 g/mol). We have established such a platform, and our most advanced molecule, the pan-RAS inhibitor LUNA18¹, is currently in clinical trials.

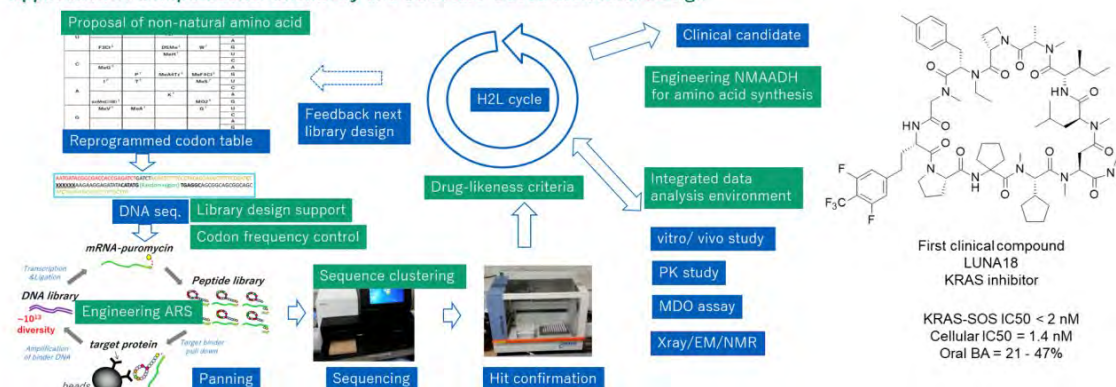
In this presentation, we will explain the details of the computational chemistry and structure-based molecular design used in the development of our cyclic peptide drug discovery platform². We will also present new insights that have not yet been disclosed, such as detailed protein engineering, sequence clustering, and the development of a data analysis environment. Through this presentation, we will demonstrate that computational chemistry and structure-based molecular design are critical for the development of new drug modality platform technologies.

References

- 1) Tanada, M., et al., J. Am. Chem. Soc. 2023, 145, 16610
- 2) Ohta, A., et al., J. Am. Chem. Soc. 2023, 145, 24035

Overview of Cyclic Peptide Discovery Platform

Application of Computational Chemistry and Structure-Based Molecular Design



001-03

Machine learning approach to analyze DNA-encoded library screening data for hit identification

Syunya SUZUKI *, Kazuma KAITOH, Yoshihiro YAMANISHI

Graduate School of Informatics, Nagoya University
(* E-mail: suzuki.syunya.h7@s.mail.nagoya-u.ac.jp)

DNA-encoded library (DEL) is a new technology for hit compound screening, where each compound in the DEL has a DNA tag whose sequence identifies the structure of the compound. Typically, a DEL is composed of millions or billions of compounds, and DEL is expected to contribute to reduction of the cost and time for identifying hit compounds in the pharmaceutical industry. Each compound in the DEL has a central scaffold that is directly linked to a DNA tag and associated side chain structures. Amplifying the DNA tags of compounds that interact with the target protein using PCR and reading them using next-generation sequencing enable us to detect compound-target protein interactions. However, compounds often interact not only with the target protein but also with the matrix that immobilizes the target protein. At the stage of amplifying and reading the DNA tags, it is not possible to distinguish between compounds that interact with the target protein and those with the immobilizing matrix. Therefore, frequently observed false positive hits are a serious obstacle in the DEL screening.

In this study, we proposed a machine learning approach to distinguish true positive hits and false positive hits in the DEL screening. We constructed a discrimination model to extract the substructures involved in false positive compounds based on the results of DEL screening in which the target protein was immobilized on a fixed support and the results of screening in which only one fixed support was used. The proposed method successfully identified compounds that interacted with the target protein and those with the immobilizing matrix separately, and the use of the Shapley value of the discriminant model contributed to the extraction of the substructures involved in the interaction with the target molecule and those with the immobilizing matrix. The proposed approach is expected to be useful for distinguishing false positive hits in the DEL screening analysis and for designing DELs consisting of compounds that avoid interactions with the immobilizing matrix.

001-04

Scaffold-retained molecule generation considering gene expression profiles with deep learning

Yuki MATSUKIYO ^{*1,2}, **Yuko SAKAJIRI** ², **Kaho YAMABE** ³, **Saki OHSHIMA** ³, **Chika TOHZAWA** ⁴, **Tomokazu SHIBATA** ¹, **Ryusuke SAWADA** ⁵, **Takuya OKADA** ^{3,4}, **Hisashi MORI** ^{6,7}, **Naoki TOYOOKA** ^{3,4}, **Yoshihiro YAMANISHI** ²

¹Department of Bioscience and Bioinformatics, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology

²Department of Complex Systems Science, Graduate School of Informatics, Nagoya University

³Graduate School of Pharma-Medical Sciences, University of Toyama

⁴Faculty of Engineering, University of Toyama

⁵Department of Pharmacology, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University

⁶Department of Molecular Neuroscience, Faculty of Medicine, University of Toyama, ⁷Research Center for Idling Brain Science, University of Toyama

(* E-mail: matsukiyo.yuki566@mail.kyutech.jp)

In the lead optimization process, molecular substructures are optimized while a molecular scaffold (i.e., core structure) is retained to improve molecular properties. Conventionally, the lead optimization is performed by experimental methods that are resource- and time-consuming. To perform this process more efficiently, deep learning models such as recurrent neural network (RNN) and variational autoencoder (VAE) have been utilized for scaffold-retained molecule generation (i.e., generating molecular structures while retaining a scaffold). However, previous scaffold-retained molecule generation methods primarily used chemical information and did not consider biological information [1-3]. Some previous molecule generation methods integrate comprehensive biological information within the cell into the molecule generation by using gene expression profiles, but they are specifically for molecule generation from scratch [4,5]. Therefore, to the best of our knowledge, there is no study on scaffold-retained molecule generation using gene expression profiles.

In this study, we present a novel computational method to generate molecules from gene expression profiles in a scaffold-retained manner. The proposed method consists of two deep learning models (i.e., VAE and RNN). The VAE was used to extract latent vectors from gene expression profiles and the RNN generated new chemical structures from the latent vectors in a scaffold-retained manner. We also carried out docking simulations of generated molecules to

obtain molecules with desirable binding affinity to a target protein. We used the proposed method to design a novel inhibitor of glutaminase 1 (GLS1), an attractive target for anticancer treatments [6]. We synthesized and experimentally evaluated some generated molecules, revealing that one of them is a very promising novel GLS1 inhibitor. These findings highlight the great potential of gene expression data-driven molecule generation in the lead optimization process.

- [1] He, J. et al. *J.Cheminform.* 2021, 13, 26.
- [2] Langevin, M. et al. *J. Chem. Inf. Model.* 2020, 60, 5637–5646.
- [3] Kaitoh, K. and Yamanishi, Y. *J. Chem. Inf. Model.* 2022, 62, 2212-2225.
- [4] Méndez-Lucio, O. et al. *Nat. Commun.* 2020, 11.
- [5] Kaitoh, K. and Yamanishi, Y. *J. Chem. Inf. Model.* 2021, 61, 4303-4320.
- [6] Okada, T. et al. *Bioorg. Med. Chem. Lett.* 2023, 93, 129438.

001-05

Development of a Molecular Generative model via the Decoupled Setting on Multi-objective Bayesian Optimization

Takamasa SUZUKI ^{*1}, Nobuaki YASUO², Masakazu SEKIJIMA¹

¹School of Computing, Tokyo Institute of Technology

²Tokyo Tech Academy for Convergence of Materials and Informatics (TAC-MI), Tokyo Institute of Technology

(* E-mail: suzuki.t.dq@m.titech.ac.jp)

The cost of developing a new drug has risen dramatically year after year. To address that problem, studies on molecular generative models that can generate compounds with novel structures have been raised in recent years. Thousands of computer-aided drug discovery (CADD) methods have been practiced in the pharmaceutical industry. Existing generative models achieve multi-objective optimization with the weighted sum or product. In other methods, generative models consider the Pareto optimality to avoid the problem of linear aggregation. However, in calculating the Pareto frontier, all candidates are required to have all objective values.

In this study, we have developed a new deep learning-based multi-objective de novo molecular generative model, which could simultaneously optimize molecules with the "decoupled setting" in entropy-based multi-objective Bayesian optimization (MBO) frameworks. MBO has several approaches, such as entropy-based, hypervolume-based, and scalar-based methods. Entropy-based methods maximize the entropy of the distribution of candidates as an acquisition function and enable to evaluation of objective function separately. The "decoupled setting" realized avoiding costs of repeatedly calculating high computationally complicated objective functions. With the setting, it is not necessary for all candidate points to be calculated in all objective functions. The setting decreases computational costs and makes molecule searches efficient. The proposed method is one of the multi-objective models and optimizes multiple objective functions, binding affinity, and drug-likeness. This method guides the discovery of high-efficiency drugs.

001-06

Optimizing Multitask Learning with Evolutionary Metrics for Enhanced QSAR-based Natural Product Activity Prediction

Donny RAMADHAN ^{*1, 2}, **Kenji MIZUGUCHI**¹

¹Laboratory for Computational Biology, Institute for Protein Research, Osaka University

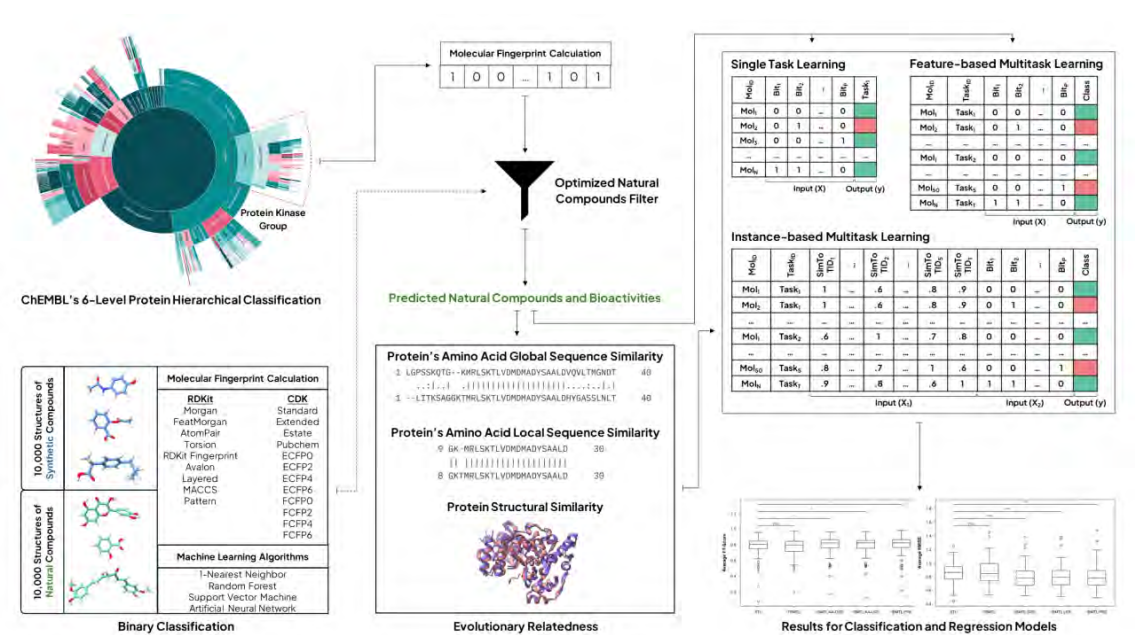
²Research Center for Pharmaceutical Ingredients and Traditional Medicine, National Research and Innovation Agency (BRIN)

(* E-mail: donny.ramadhan@protein.osaka-u.ac.jp)

Natural products exhibit a wide range of structural diversity with their relatively high degree of three-dimensionality, which could play an essential role in their interactions with drug targets. Given the limited availability of bioassay data for pure natural products in public databases, applying multitask learning (MTL) models in quantitative structure-activity relationship (QSAR) studies is expected to enhance the prediction of natural product activity. The effectiveness of transferred information in MTL depends on the relatedness of the tasks combined. However, only a few studies have examined this task-relatedness for use in MTL models, especially for QSAR studies. This research explores the effects of various evolutionary metrics used as input features on the performance of MTL models using limited datasets of natural product biological activities. We curated datasets from the ChEMBL database that comprise the biological activities of natural products against drug targets in the protein kinase group. These datasets were initially filtered using binary classification to identify predicted natural products and their activities. A total of 94 and 86 target proteins were used for classification and regression models, respectively. A single-task learning (STL) model, using Avalon fingerprints (1024 bits) as input features and an artificial neural network, served as the control for predicting the activity of natural compounds for each protein. Subsequently, feature-based multitask learning (FBMTL) was conducted by training the dataset on all proteins within a protein class and predicting the activity of all compounds for each protein. Instance-based multitask learning (IBMTL), a type of FBMTL, incorporated additional input features; in our study, we utilized three types of evolutionary metrics: global sequence similarity, local sequence similarity, and structural similarity of proteins. The results indicate that by leveraging evolutionary relatedness, IBMTL demonstrates statistically significant improvements across all performance parameters of classification and regression models compared to STL, despite using limited datasets of natural

products and their bioactivities. FBMTL, on the other hand, fails to show similar improvement in performance.

Keywords: evolutionary relatedness, feature-based multitask learning, instance-based multitask learning, single-task learning



002-01

Deep learning of new morphological characteristics of blood vessel in breast cancer.

Tomoyasu SUGIYAMA *¹, **Masayoshi FUJISAWA**², **Koichiro DOI**¹, **Tomonari KASAI**³, **Hiroyuki KAMEDA**⁴, **Toshiaki OHARA**²

¹School of Bioscience and Biotechnology, Tokyo University of Technology

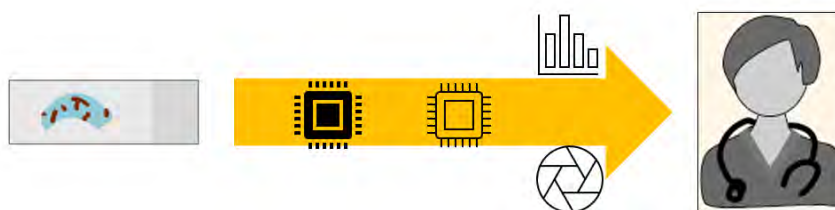
²Department of Pathology and Experimental Medicine, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University

³Neutron Therapy Research Center, Okayama University

⁴Tokyo University of Technology

(* E-mail: tsugiyama@stf.teu.ac.jp)

Angiogenesis has an important role in cancer prognosis. We have previously associated the morphology of capillary, e.g., C-shaped and excessively branched capillaries, with poor prognosis in breast cancer (Pathol. Int. 2024; 74:394-407). The new morphological characteristics identified in capillaries are of interest as reliable biomarkers for taking into consideration of therapeutic interventions. The methodology, however, requires expertized works of microscopic observation to find those capillaries in tumor tissues. This study aimed to develop artificial intelligence (AI) models for identifying and classifying intra-tumoral capillaries in immunohistochemical images based on the morphology using pix2pix with conditional generative adversarial networks and visual geometry group 16 with convolutional neural networks. We first applied an AI model to extract and visualize capillaries from an immunohistochemical image of CD31 and HE stains. Then, other AI model classified capillaries in each small area out of original size of image. Thus, the AI models visualized regions of C-shaped and excessively branched capillaries as abnormal areas. Subsequently, the AI model calculated abnormal scores for the probability of classification. As a result, C-shaped and excessively branched capillaries showed high score compared to normal capillaries. The workflow of making AI models could be useful for the AI prediction of capillaries which are prognostic in breast cancer.



O02-02

Enhancing Drug-Target Interaction Prediction using Large Language Models and Low-Rank Adaptation

Rintaro YASHIRO *, Nobuyuki YASUO, Masakazu SEKIJIMA

Tokyo Institute of technology

(* E-mail: yashiro.r.ab@m.titech.ac.jp)

Prediction of Drug-Target Interaction (DTI) is crucial in drug discovery applications such as drug repositioning, hit compound discovery, and side effect investigation. Experimental methods for DTI prediction are often costly and time-consuming, leading to the development of in silico prediction approaches. Databases like DrugBank and the Therapeutic Target Database (TTD) are commonly used for these in silico predictions. DrugBank, a comprehensive resource of drug information annotated from PubMed, is essential for creating training datasets when developing DTI prediction models.

However, the manual extraction of relevant DTI information from databases like PubMed, which sees millions of new articles added annually, is impractical. Therefore, automating the extraction of DTIs has become a critical need in the field.

In this study, we fine-tuned the Llama3-8B and Llama3-70B models, state-of-the-art large-scale language models known for their exceptional natural language processing capabilities, to develop a model that automatically extracts DTI information from article data. We also evaluated the impact of different prompts on the model's performance and its ability to generalize. The results demonstrate significant improvements in both extraction accuracy and generalization capability.

002-03

Exploring Synthetically Accessible Chemical Spaces with Product-of-Expert Chemical Language Models

Shuya NAKATA ^{*1}, Yoshiharu MORI ^{1, 2}, Tanaka SHIGENORI ¹

¹Graduate School of System Informatics, Kobe University

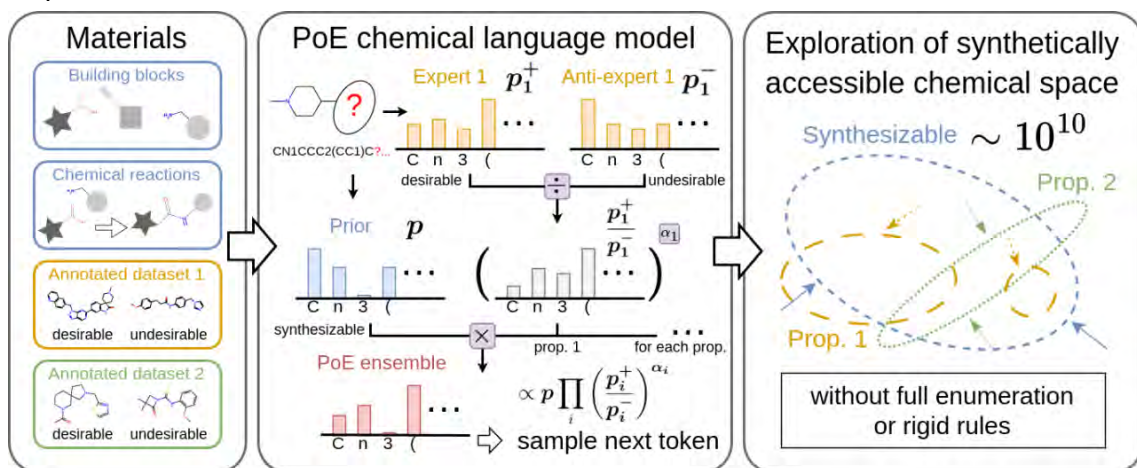
²KQCC, Keio University

(^{*} E-mail: nakata@landscape.kobe-u.ac.jp)

Ultra-large synthesis-on-demand chemical spaces have emerged as a valuable resource for drug discovery. Chemical language models that enable direct compound generation have the potential to accelerate the exploration of these vast spaces. However, existing methods are not designed to explore specific chemical spaces and often overlook synthetic accessibility. To address these limitations, we introduce product-of-experts (PoE) chemical language models, a modular and scalable approach to exploring synthetically accessible chemical spaces. This method combines a *prior* model pre-trained on a chemical space of interest with *expert* and *anti-expert* models fine-tuned using external property-specific datasets, allowing for controlled compound generation within the target space.

We constructed an ultra-large chemical space of 10^{10} compounds for evaluation and found that chemical language models can implicitly learn the underlying generative process, i.e., the possible combinations of building blocks and reactions. Furthermore, we demonstrated that the proposed approach effectively guides compound generation toward desired properties while ensuring synthetic accessibility.

The presentation will detail our method and results, and discuss some implications for future studies.



002-04

Feature Design of Molecular 3D Structures for Fast Approximate Nearest Neighbor Search

Kotaro KAMIYA *

SyntheticGestalt KK

(* E-mail: k.kamiya@syntheticgestalt.com)

Background:

In drug discovery and molecular design, the ability to efficiently compare and search for similar molecular structures is crucial. Machine learning techniques have shown promise in this area, often utilizing 3D representations of molecules that encompass their shape and physicochemical properties, such as charge density and electrostatic potential. However, directly comparing these 3D representations can be computationally expensive, particularly when dealing with large datasets, limiting the scalability of applications like virtual screening.

Challenges and Proposed Solution:

A naive approach to nearest neighbor search, where every molecule is compared to every other molecule, quickly becomes infeasible as the dataset grows. This bottleneck hinders the ability to explore vast chemical spaces efficiently.

To overcome this challenge, we propose leveraging Approximate Nearest Neighbor (ANN) search techniques. ANN algorithms, such as Hierarchical Navigable Small Worlds (HNSW), offer a way to find approximate nearest neighbors with significantly reduced computational cost. These algorithms excel at finding similar items in high-dimensional spaces, making them well-suited for molecular structure comparison.

However, implementing ANN for molecular structures presents two key challenges:

Infinite Dimensions: Molecular properties like electrostatic potential are continuous functions in 3D space, making them inherently infinite-dimensional. ANN algorithms, on the other hand, typically operate on finite-dimensional vector representations.

Alignment Invariance: Meaningful comparisons of molecular structures require considering their alignment in 3D space. A molecule's orientation and position should not affect its similarity to another molecule. Incorporating alignment invariance into ANN search adds another layer of complexity.

To address these challenges, we propose a feature design approach that focuses

on creating alignment-invariant representations of molecular structures that are suitable for ANN search. We specifically target 3D molecular graphs, where atoms are represented as nodes and bonds as edges. This representation is both chemically intuitive and computationally efficient.

Our approach involves exploring various techniques for generating alignment-invariant features from 3D molecular graphs. These techniques include persistent homology, which captures topological features of the molecular shape, and a neural network architecture designed for 3D molecular graphs. By extracting and combining these features, we aim to create a compact and informative representation that can be readily used with ANN algorithms.

002-05

Clmpy: A platform for Chemical Language Model comparable training and structure generation ability

Shumpei NEMOTO ^{*}, Yasuhiro YOSHIKAI, Tadahaya MIZUNO, Hiroyuki KUSUHARA

Graduate school of pharmaceutical sciences, The University of Tokyo

(^{*} E-mail: nemo88@g.ecc.u-tokyo.ac.jp)

Chemical Language Models (CLM) are natural language models trained on chemical structures. Since the groundbreaking research by Gomez et al. in 2016, CLMs have rapidly evolved, utilizing string-based representations of compounds (such as SMILES and InChI) as input [1]. A key strength of CLM is their ability to leverage sophisticated Natural Language Model technologies. By utilizing neural machine translation architectures, CLM enable representation learning of diverse chemical structures without the need for any auxiliary tasks. This capability has been applied in various cheminformatics tasks such as descriptor generation for Quantitative Structure-Activity Relationship (QSAR) studies. However, the mechanisms by which these models recognize and learn diverse chemical structures remain unclear. Furthermore, many studies construct models in unique environments, leading to a lack of standardized comparison. Here, we aimed to develop a platform that enables the training and comparison of multiple CLM, by leveraging our experiences in this field [2, 3]. We implemented four model structures (GRU, GRU-VAE, Transformer, and Transformer-VAE) with two tokenizers (conventional tokenizer and Simplified Feature Learning (SFL)), resulted in 8 different models [4]. When trained on 30 million compounds obtained from the public compound database ZINC, all models achieved high translation accuracy exceeding 80%, with minimal differences in precision between models. These results demonstrate that our platform enables high-fidelity comparisons across different model architectures. In comparing the constructed models, we evaluated the accuracy rates for structures with chirality. Despite comparable overall translation accuracy, the GRU models outperformed the Transformer models in this aspect. This result aligns with previous reports suggesting that vanilla Transformer have a weakness in recognizing chirality [3]. Aiming for public and free accessibility, we have packaged the platform as a Python module. This package allows users to specify arbitrary parameters for training in their own environments, facilitating model comparisons under various conditions. It supports both interactive environments like Jupyter notebooks and command-line execution.

This research enables high-precision inter-model comparisons of CLM based on neural machine translation. We anticipate that this will deepen our understanding of chemical structure recognition by CLM and contribute to advancements in in silico drug discovery. Future work will utilize this platform to compare the predictive accuracy of downstream tasks (such as toxicity prediction) using different CLM and to analyze differences in the learning processes of each CLM.

- [1] Gomez-Bómbarelli R.; et al, ACS Cent. Sci., 2018, 4, 268-276
- [2] Nemoto S.; et al, J. Cheminform., 2023, 45, 15
- [3] Yoshikai Y.; et al, Nat. Commun., 2024, 15, 1197
- [4] Lin X.; et al, Brief. Bioinform., 2020, 21(6), 2099-2111

002-06

Fragment descriptors in predictive modeling for molecules and reactions

Pavel SIDOROV *

Hokkaido University, Institute for Chemical Reaction Design and Discovery (ICReDD)

(* E-mail: pavel.sidorov@icredd.hokudai.ac.jp)

The application of machine learning methods in chemistry requires encoding chemical structures as vectors of numbers (features or descriptors) to then train predictive models on them. Currently, there are many types of descriptors that are used in the field, from physico-chemical parameters to those derived from 2D or 3D structures of molecules. However, calculation of such parameters may be complicated and render the modeling process too slow. Fragment descriptors encode molecules as number of occurrences of different substructures, allowing to directly translate the chemical structure into numerical format. Moreover, these are derived from 2D structures (molecular graphs), which allows for cheap and fast modeling. In this presentation, we show two examples of the application of fragment descriptors in predictive modeling in chemistry. First, the prediction of absorption spectra of photoswitches was facilitated by using 2D fragment descriptors in collaboration with Dr Hashim (RIES). In a benchmark with other 2D descriptors fragments have shown superior performance. Second, a collaboration with Dr Tsuji (WPI-ICReDD, List group) involved modeling and design of novel potent catalysts in organic synthesis. While traditionally costly 3D calculations are used for such tasks, we managed to predict a new highly selective catalyst by using simple 2D fragments for both catalysts and reactions. Moreover, these descriptors allow to use methodologies for model interpretation, facilitating the rational and guided design of new compounds with desired properties. For example, in the latter case, the ColorAtom technique was used to explain which substructures are most important for ensuring high selectivity of new catalysts.

003-01

Construction of Flavivirus database and therapeutic antibody discovery using machine learning algorithm

Piyatida NATSRITA *, Kenji MIZUGUCHI

Laboratory for Computational Biology, Department of Biological Sciences, Osaka University

(* E-mail: u131573g@ecs.osaka-u.ac.jp)

Flavivirus infection responsible for approximately 50 - 100 million apparent diseases and 300 million infections per year. Member of the Flavivirus group include Zika (ZIKV), Japanese encephalitis (JEV), West Nile (WNV), Yellow Fever (YFV), and dengue viruses, including 4 serotypes (DENV-1, DENV-2, DENV-3, DENV-4). Several studies hypothesized that cross-reactivity among distinct serotypes and other flaviviruses is a major problem of severe caused by antibody dependent enhancement (ADE) phenomenon. In this study, we aim to develop a novel dataset of CDR-H3-epitope sequences together with IC50 values and sequence-based ML approach to predict the potential neutralizing antibodies against Flavivirus towards the characterization and analysis of our obtained sequences in terms of neutralizing levels, cross-reactivities, and important features for broad neutralization. Firstly, we generated the dataset of CDR-H3 sequences together with epitope sequences and labeled with IC50 values toward the characterization and analysis of our obtained sequence in terms of neutralizing levels. In a total of 3,767 pairs including 1,366 high neutralizing activity ($IC \leq 10$ ng/ μ L) and 2,400 low neutralizing activity ($IC > 10$ ng/ μ L). In the dataset, we found 541 cross-reactive antibodies, and 826 non-cross-reactive antibodies. From ML analysis, we found 20 important features including

chiral carbon and aromaticity. The larger pool of CDR-H3-epitope-IC50 data lead to empower a ML model high-throughput screening performance for sequence classification. Further effort will focus on different encoding method comparison, antibody repertoire-level Flavivirus screening and classification. The potential antibody candidates against Flavivirus will be evaluated by performing molecular docking and MD simulations of each candidate with DENV, ZIKV, YFV, WNV, and TBEV to determine the location of binding, binding affinity, and stability. This finding might be useful for further development of therapeutic antibodies in the future.

003-02

Digital Transformation on Small Molecule Optimization Research at Astellas Pharma

Mori KENICHI *¹, Kenji NEGORO²

¹Modality Informatics, ResearchX, DigitalX, Astellas Pharma Inc.

²Platform Sciences & Modalities, Discovery Intelligence, Applied Research & Operations, Astellas Pharma Inc.

(* E-mail: kenichi-mori@astellas.com)

Astellas Pharma has been developing a digital transformation platform to accelerate and improve the quality of the Design-Make-Test-Analysis cycle in lead compound optimization in the small molecule drug discovery research process, including the introduction of AI and robots, to enable researchers to advance their research more efficiently, since the end of 2019. In addition to developing the platform, we have also been engaged in various activities to enable researchers to use this platform daily in their research projects. As a result, we have reached a phase where almost all medicinal chemistry researchers involved in synthesis in drug discovery research projects are using the platform. Furthermore, various success cases have been observed in actual drug discovery research use cases. On the other hand, we have also seen challenges especially in the prediction and compound design with three-dimensional structures. Therefore, we try to solve these challenges by making the most of the opportunities of recent technological innovations in the field of generative AI and a GPU supercomputer. In this presentation, we will introduce our efforts in digital transformation of drug discovery research, particularly focusing on small molecule modalities, related activities and success cases, challenges, and efforts to solve problems using cutting-edge technologies.

003-03

Open Molecule Generator: A Multipurpose Molecule LLM

David JIMENEZ BARRERO *

Elix, Inc

(* E-mail: david.jimenez@elix-inc.com)

The latest advancements in large language models (LLMs) have significantly expanded the range of possibilities, paving the way for new opportunities in the field of generative molecular drug design. Modern paradigms in the realm of generative models outlined the concept of Foundation Models[1]; often large-scale, pre-trained machine learning models that serve as a general-purpose tool for a wide range of tasks. These models are typically trained on vast amounts of diverse data and can be fine-tuned, adapted, or extended to perform specific tasks with relatively small amounts of additional data. Furthermore, leveraging techniques such as Low Rank Adaptation (LoRA)[2], fine-tuning can be made cost efficient. In this project, we created a large molecule-oriented Foundation Model, pre-trained with a semantically curated construction of the ChEMBL[3] dataset in order to be able to handle several downstream tasks that might require physico-chemical, structural, or activity-related information of molecules. To achieve this goal, each molecule in the pre-training dataset contains information such as molecular SMILES, physico-chemical properties, semantic description of substructures, and known activity with protein targets. The dataset was tokenized and the model trained with 6.5B tokens using next-token prediction. Once trained, it was tested in a Multi-Constrained Molecular Generation task, where the model aims to generate molecules that satisfy up to 26 physico-chemical requirements. The highest error achieved in Mean Absolute Error (MAE) was for the Molecular Weight property, with a relatively small deviation of 18.07 Da. from the target. Other interesting properties like logP and Quantitative Estimate of Drug-Likeness[4] achieved a MAE of 0.5 and 0.05 respectively. We expect to extend its functionality via fine-tuning to more specialized tasks.

[1] Competition and Markets Authority (2023). AI Foundation Models: Initial Report,
https://assets.publishing.service.gov.uk/media/65081d3aa41cc300145612c0/Full_report_.pdf

[2] Edward J. Hu et al., LoRA: Low-Rank Adaptation of Large Language Models.

arXiv:2106.09685

[3] Anna Gaulton et al., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 2012 Jan; 40(Database issue): D1100–D1107.

[4] G. Richard Bickerton et al., Quantifying the chemical beauty of drugs. *Nat Chem.* 2012 Jan 24; 4(2): 90–98.

003-04

Application of machine learning to single-cell RNA sequencing provides the candidate drugs against drug-tolerant persister cells in colorectal cancer

Yosui NOJIMA *¹, Ryoji YAO², Takashi SUZUKI¹

¹Center for Mathematical Modeling and Data Science, Osaka University

²Department of Cell Biology, Japanese Foundation for Cancer Research

(* E-mail: nojima@sigmath.es.osaka-u.ac.jp)

The inactivation of the *APC* gene is a crucial early event in colorectal cancer (CRC) development. Familial adenomatous polyposis (FAP) is a hereditary syndrome characterized by numerous adenomas in the colon, which significantly increase CRC risk due to an autosomal dominant mutation in the *APC* gene. The carcinogenesis mechanism in FAP mirrors that of sporadic CRC.

Previously, we evaluated the efficacy of anticancer drugs using organoids derived from benign and malignant tumors in FAP patients. The results showed that organoids from malignant tumors were resistant to the MEK inhibitor trametinib, likely due to the presence of drug-tolerant persister (DTP) cells in the cancer tissues.

Single-cell RNA sequencing (scRNA-Seq) is a powerful technology that allows for high-resolution analysis of gene expression in individual cells, offering new insights into cancer biology. Machine learning (ML), a branch of artificial intelligence, leverages statistical techniques to learn from data and is increasingly used in cancer diagnosis, prognosis, and treatment.

In this study, we conducted scRNA-Seq on FAP-derived organoids and built ML models using public data to identify DTP cells resistant to MEK inhibitors. Additionally, we identified candidate drugs against DTP cells in FAP organoids using public drug sensitivity data and validated the effects of these drugs using a cell viability assay.

This is the first study to demonstrate that ML models can identify DTP cells and propose a novel strategy for identifying candidate drugs against DTP cells.

003-05

Predicting Novel Therapeutic Target Molecules Using Neural Networks: Validation and Applicability to Unknown Diseases

Hayato TSUMURA ^{*1}, Narumi HATANO¹, Mayumi KAMADA², Ryosuke KOJIMA¹, Hiroaki IWATA³, Yasushi OKUNO^{1, 4}

¹Graduate School of Medicine, Kyoto University

²School of Frontier Engineering, Kitasato University

³Faculty of Medicine, Tottori University

⁴RIKEN Center for Computational Science(RCCS)HPC/HPC- and AI-driven Drug Development Platform Division

(* E-mail: tsumura.hayato.22x@st.kyoto-u.ac.jp)

Identifying therapeutic target molecules related to disease causation and progression is a crucial and challenging step in the early stage of drug discovery. Although the space of potential target molecules is vast, the number of experimentally verifiable molecules is limited. Thus, computational approaches, especially machine learning, are increasingly adopted to discover novel therapeutic target molecules.

Gene expression profiles obtained from directional perturbations such as knockdown or overexpression of genes encoding drug target molecules can reflect the functionality of drugs that inhibit or activate these targets. A previous method has demonstrated that gene expression profiles representing the perturbation response of candidate target genes and similarity scores between diseases are beneficial in predicting new target-disease relationships. However, this approach was limited to diseases with known therapeutic target molecules, and there is room for improvement in prediction accuracy (ROC-AUC: 0.63 for inhibitory target prediction, 0.65 for activatory target prediction).

This study presents a neural network-based approach to predict novel therapeutic targets applicable to diseases regardless of with and without known targets. The validation results show a significant improvement in prediction accuracy compared to previous methods (ROC-AUC: 0.93 for inhibitory target prediction, 0.82 for activatory target prediction). Furthermore, A leave-one-out validation strategy was employed to examine the model's predictive performance on each disease, thereby assessing the robustness and generalizability of our approach. The results demonstrate that our model can predict therapeutic target molecules for diseases not included in the training phase. Moreover, our investigation of the predicted candidate target molecules has identified literature supporting their association with diseases. The results

indicate the potential of our model to be applied to diseases with previously unknown target molecules.

The model proposed in this study has the potential to predict novel therapeutic target molecules for rare and novel diseases, which have been challenging for conventional approaches.

003-06

Design of Receptor Selective Cell-Penetrating Peptides Using Deep Learning and Simulations

Iori YAMAHATA *¹, Hayashi SHUTO², Koseki JUN³, Shimamura TEPPEI^{1, 2}

¹Nagoya University Graduate School of Medicine

²Institute of Science Tokyo

³National Institute of Advanced Industrial Science and Technology

(* E-mail: yamahata.iori.x1@s.mail.nagoya-u.ac.jp)

Despite advances in intracellular drug targeting, the cellular membrane permeability of large molecule drugs remains a major challenge. Cell-penetrating peptides (CPPs) offer promising solutions but are hindered by low cell selectivity. This study developed a novel method to design CPP sequences with enhanced receptor selectivity, focusing on receptor-mediated endocytosis.

Our approach consisted of two main steps:

1. Construction of a deep generative model to create CPP-like sequences:

We developed a CPP-specific model based on EvoDiff, using Low-Rank Adaptation (LoRA). The model was trained on 1,082 known CPP sequences and efficiently generated diverse CPP candidate sequences.

2. An in silico optimization cycle to enhance selectivity:

This cycle, implemented on a supercomputer, involved:

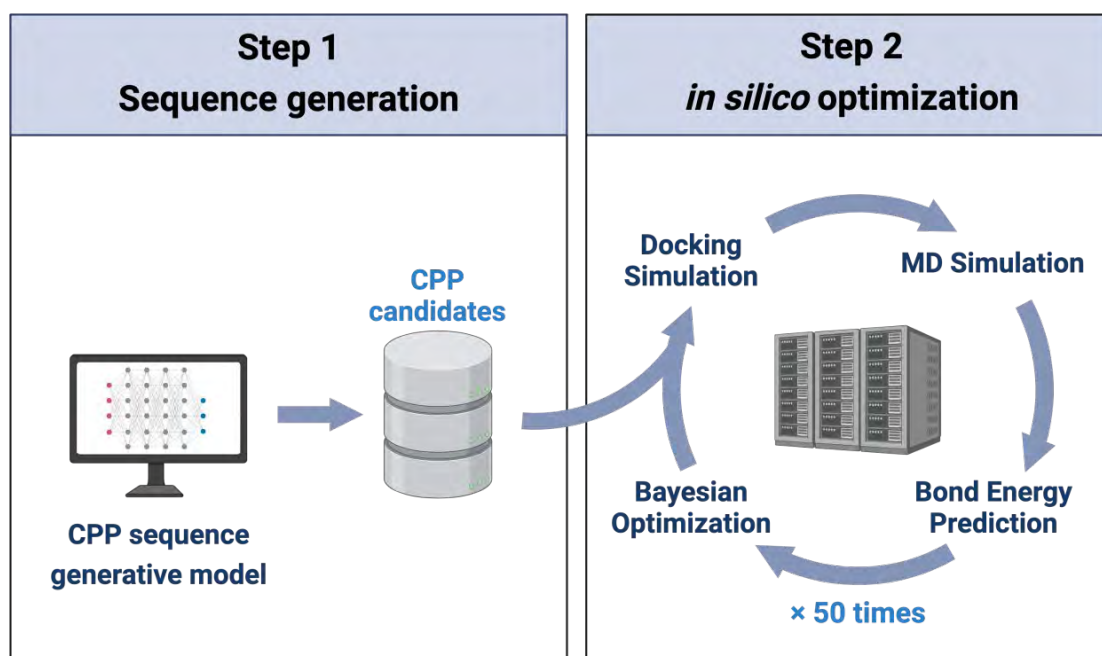
- a) Docking simulations between candidate peptides and target receptors
- b) Binding energy calculations using molecular dynamics simulations
- c) Training of deep learning models to predict binding energies
- d) Bayesian optimization for selecting sequences for the next cycle

As a proof of concept, we applied our method to design CPPs that selectively internalize via CXCR4 while minimizing interaction with NRP1. We simulated 2,000 sequences per cycle using DiffDock for molecular docking and MM/GBSA for binding energy estimation.

The optimization cycle progressively improved sequence quality, as indicated by increasing hypervolume metrics. We successfully identified several CPP sequences with high predicted selectivity for CXCR4. In silico validation showed improved binding energies compared to known CPPs.

This computational approach represents a significant advance in the rational design of selective CPPs. The next crucial step will be to validate these findings through in vitro experiments, which will assess the actual receptor selectivity and internalization efficiency of the designed CPPs.

Our method demonstrates the power of combining deep learning, molecular simulations, and optimization techniques in addressing complex biological challenges. Upon experimental validation, this approach could potentially accelerate targeted drug delivery and be applicable to other areas of peptide drug discovery.



004-01

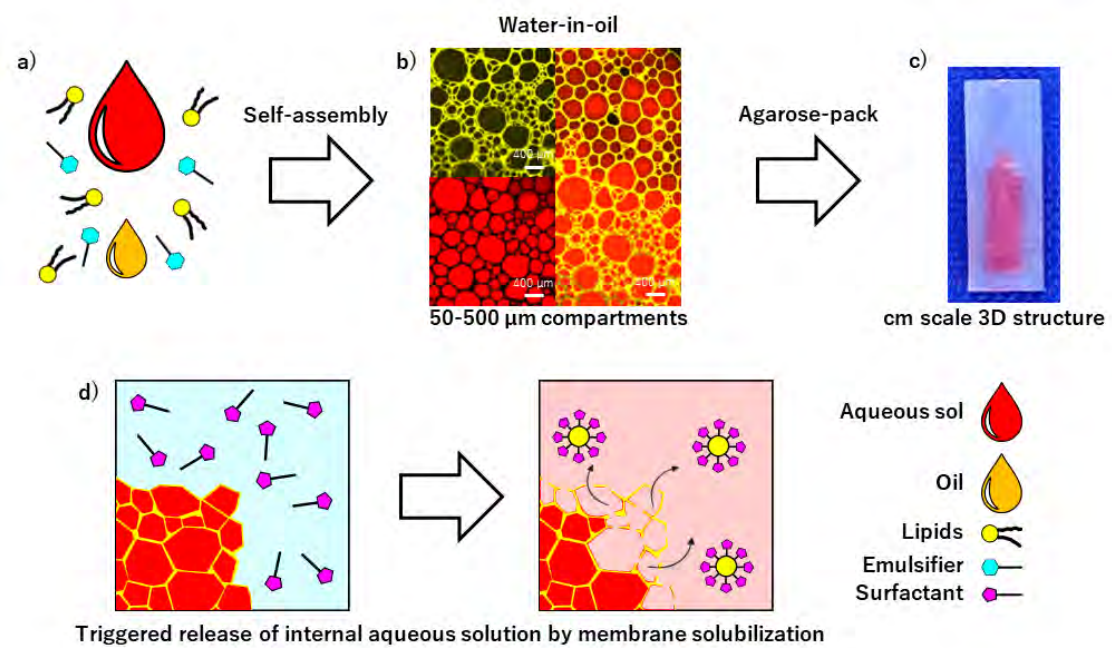
Biomimetic Multicellular Lipid-Based Membranes for Stimulus-Responsive Drug Delivery

Tsuyoshi INABA *, Richard James ARCHER, Shin-ichiro M. NOMURA

Molecular Robotics Laboratory, Department of Robotics, Grad.Sch.Eng.,Tohoku Univ.

(* E-mail: tsuyoshi.inaba.q7@dc.tohoku.ac.jp)

Drug delivery systems have garnered attention as a method to utilize existing drugs more efficiently, raising expectations for effective treatments with fewer side effects. However, medications applied to the skin, such as patches and ointments, still rely on uncontrolled diffusion for delivery. This study proposes the use of bio-inspired lipid-hybrid membranes to structurally control the release of model drugs through cell-like compartmentalization. Using biocompatible amphiphilic molecules, we demonstrated the facile self-assembly of aqueous-based micro-compartments into tightly packed centimeter-scale “multicellular” systems within a portable polysaccharide matrix. This lipid-based micro-compartmentalization strategy allows for the separation and storage of multiple drugs, potentially enabling stepwise or sustained delivery in response to environmentally triggered membrane solubilization. Membrane solubilization and drug release was conducted with a surfactant which was highly dependent on solution concentrations of salt (NaCl), giving an environmental trigger for release. Encapsulation of multiple model drugs with lipid-membrane based spatial separation demonstrated time delayed drug release with environmental sensitivity. This research could lead to the development of highly customizable medical patches for smart drug delivery.



004-02

Physical Reservoir Computing Device Using Active Matter Composed of a Swarm of Biomolecular Motors.

YIMING GONG ^{*1}, Gikyo **USUKI**², Sidak Singh **GREWAL**¹, Kazuki **SADA**^{2,3}, Arif Md. Rashedul **KABIR**⁴, Marie **TANI**¹, Masatoshi **ICHIKAWA**¹, Nathanael **AUBERT-KATO**⁵, Ibuki **KAWAMATA**¹, Akira **KAKUGO**¹

¹Graduate School of Science, Kyoto University

²Graduate School of Chemical Sciences and Engineering, Hokkaido University

³Faculty of Science, Hokkaido University

⁴Faculty of Science and Engineering, Macquarie University

⁵Department of Information Sciences, Ochanomizu University

(* E-mail: gong.yiming.66v@st.kyoto-u.ac.jp)

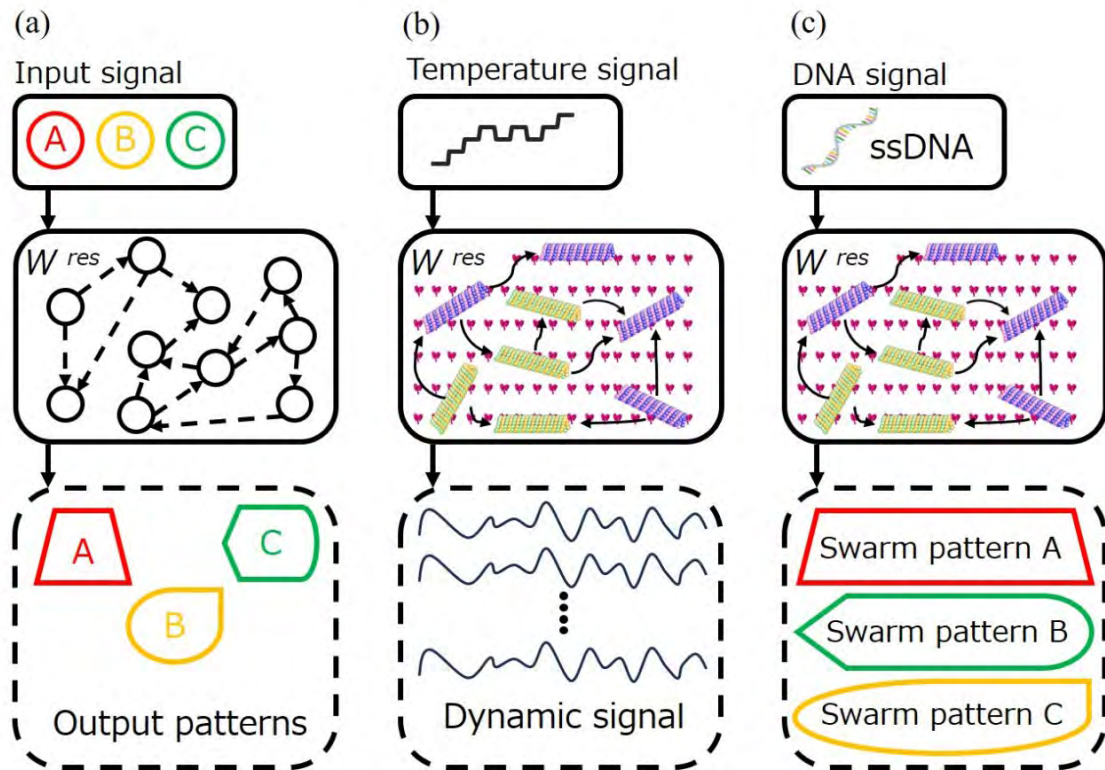
Reservoir Computing (RC) is a pioneering computational architecture that regards the dynamical system (called reservoir) response against inputs as information processing resources. In the case of in silico realization of RC, computational models such as Recurrent Neural Networks without optimized training can be used as a reservoir. If the model has sufficiently complex dynamics, one can map the readout of the reservoir to desired outputs by merely optimizing a small network (output layer), which is placed downstream of the model (Figure 1a). As a result, RC has low training overhead compared with training the whole model using machine learning algorithm.

The concept of RC can be extended from in silico to in vitro realization, which can directly receive physical inputs such as molecules or temperature, and process information for practical application in a physical system. The framework for using complex real-time physical systems as a reservoir is called Physical Reservoir Computing (PRC). Standard problems that PRC can address include memorizing temporal behavior of input signal and classifying inputs into pre-defined patterns.

Among physical systems, swarming, which is characterized by the collective movement and self-organization of agents, has emerged as a suitable candidate of reservoirs. Swarming is known for its capabilities of displaying highly order emergence from local interactions of distributed agents. If we can use chemical information as inputs for a swarming system, typical memory and classification functions of PRC have a potential to develop applications such as retrieving health history and conducting diagnosis.

This research aims to demonstrate a PRC device using biomolecular motors as a reservoir which can respond to inputs such as temperature and DNA signals.

For this purpose, we conjugated MTs with DNA, which can interact with each other through DNA hybridization to form swarming structures. We first employ temperature as time-dependent inputs that can affect the dynamical system by influencing the interaction between MTs (Figure 1b). Since MT interactions are controlled by DNA, higher temperature will result in melting DNA hybridization and dissociation of swarming MTs, and vice versa. This temperature-dependent behavior allows the system to respond to time series input of temperature for the memory demonstration of PRC. Next, we employ different types of DNA signals as input to evaluate the performance of the MT swarming PRC for classification problem. Using different single-stranded DNA complementary to the MT-conjugated DNA, the system can form different types of swarm patters. This implies that we can categorize the input by observing the patterns, once the correspondences between inputs and output patters are trained (Figure 1c). The aim of this research to harness MT swarming as PRC system, will lead us to perform practical computational tasks such as diagnostics using molecular system in the future.



004-03

GAN-Based Multi-Axis Resolution-Enhanced 3D Visualization of Giant Vesicles

Soichiro HIROI ^{*1, 2}, **Taro TOYOTA**^{1, 3}, **Akihiko KONAGAYA**²

¹Department of Basic Science, Graduate School of Arts and Science, The University of Tokyo

²Molecular Robotics Research Institute, Co., Ltd.

³Universal Biology Institute, The University of Tokyo

(* E-mail: hiroi@molecular-robot.com)

In the rapidly evolving field of molecular visualization, accurately representing the 3D morphology of complex structures remains a significant challenge [1]. Giant vesicles (GVs), with their dynamic nature and sensitivity to environmental factors [2], present a particularly intriguing subject for high-resolution imaging and virtual reality (VR) representation.

Previously, we developed a GAN-based model to enhance the clarity of GV microscopy images for VR visualization [3]. Building upon this foundation, our current study introduces substantial improvements in both dataset construction and model architecture to address the persistent issues of image blurring and loss of fine structural details across multiple axes.

In this study, we have reconstructed our synthetic dataset to be more robust and capable of capturing intricate membrane details and vesicle contours in both lateral (XY) and axial (Z) dimensions. This refined dataset serves as a stronger foundation for our deep learning model. Additionally, we have developed an improved generative model with two key enhancements:

1. Implementation of specialized loss functions tailored to visualize vesicle shapes and membrane states more effectively in 3D space.
2. Integration of multi-axis resolution enhancement techniques, including Z-axis interpolation, to more accurately capture and represent the 3D structure of GV.

[1] Konagaya, A.; Gutmann, G.; Zhang, Y. Co-creation environment with cloud virtual reality and real-time artificial intelligence toward the design of molecular robots. *J. Integr. Bioinform.* 2022, 20220017.

[2] Lipowsky, R. Remodeling of Membrane Compartments: Some Consequences of Membrane Fluidity. *J. Biol. Chem.* 2014, 395, 253–274.

[3] Hiroi, S.; Toyota, Y.; Konagaya, A. Deep Learning-Based Deconvolution of Confocal Laser scanning Fluorescence Microscopy Images for Enhanced Visualization of Giant Vesicles, Proceedings of the CBI Annual Meeting 2023, 002-02.

004-04

Investigation on heterogeneous pairs of cell-sized liposomes formed in a microfluidic device

Haruto OBUCHI ^{*1}, ZHANG YITING², Shogo HAMADA³, Keita ABE⁴, Akihiro INADA⁵, Teijiro ISOKAWA⁵, Kaoru UESUGI⁶, Hironori SUGIYAMA⁷, Satoshi MURATA⁴, Taro TOYOTA^{1, 8}

¹Graduate School of Arts and Sciences, Department of Basic Science, University of Tokyo

²College of Science, Rikkyo University

³School of Computing, Tokyo Institute of Technology

⁴School of Engineering, Tohoku University

⁵School of Engineering & Graduate School of Engineering, University of Hyogo

⁶Department of Mechanical Systems Engineering, College of Engineering, Ibaraki University

⁷School of Engineering, University of Tokyo

⁸Universal Biology Institute, University of Tokyo

(* E-mail: haruto-obuchi342@g.ecc.u-tokyo.ac.jp)

Liposomes, which are closed lipid bilayer membranes, have drawn much attention in cosmetics and pharmaceuticals [1,2] . The construction of chemical artificial intelligence (CAI) has recently garnered significant attention in efforts to improve the technologies of diagnostic medicines and artificial organs [3] . The unit of CAI is a complex of three types of functionalized cell-sized liposomes, i.e. sensor, processor, and actuator liposomes, which can mutually transfer chemical information through intermembrane transducer molecules. It remains as a challenge to arrange such three different liposomes in a specific order. We have developed a microfluidic device (MFD) which can simultaneously array multiple cell-sized liposomes to form heterogeneous pairs, which are regarded as signal senders and receivers, and evaluate interactions between the liposomes.

The MFD trapped more than one hundred liposome pairs of uniform size in the microstructures. Interestingly, we observed that under a constant flow pressure, liposomes containing phosphatidylglycerol (liposome A) were forced out of the microstructures preferentially over those without it (liposome B). Therefore, the regulation of the flow pressure for introducing these liposomes enabled us to increase the ratio of number of AB-type liposome pair to those of the others (AA, BA, and BB). Based on this phenomenon, we developed a rearrangement method of liposome pairs in the MFD, by which the formation of the unit of CAI

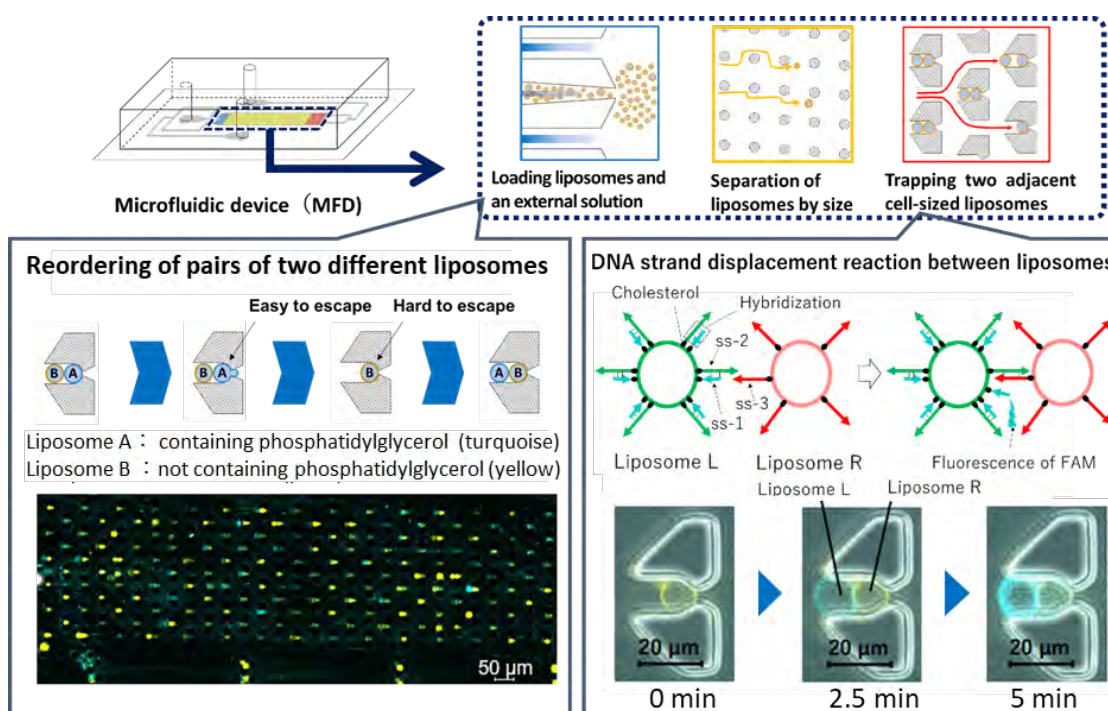
would be realized.

Next, we examined intermembrane DNA strand displacement reaction at the heterogeneous pairs of cell-sized liposomes by the MFD. We separately prepared two kinds of liposome, Liposome L and R, which beard DNA-cholesterol conjugates within the membrane. Liposome L carried hybridized DNA-cholesterol conjugates, ss-1 and ss-2. Conjugate ss-1 has a single strand DNA tagged with a fluorophore called FAM, and ss-2 has a single strand DNA complementary to that of ss-1 and quenches FAM. Liposome R beard conjugate ss-3 the DNA of which is complementary to that of ss-2 and has a longer sequence than ss-1. The sequential introduction of liposome L followed by liposome R allowed the formation of numerous heterogeneous pairs of liposomes L and R in the MFD. An increase in the fluorescence intensity of liposome L in the pair was observed, which means that the DNA strand displacement reaction between hybridized ss-1/ss-2 and ss-3 restored the FAM fluorescence. The current results will lead a novel statistical validation method of two different liposomes carrying DNA-cholesterol conjugates. Furthermore, the results obtained in this study suggest that on-membrane DNA reactions between adjacent cell-sized liposomes can contribute to the construction of a computational system.

[1] Y. Jiang., et al. 2024. *Pharmaceutics*. 16, 34.

[2] Y. Rahimpour and H,Hamishehkar.2012. *Expert Opin. Drug Deliv.* 9,4, 443–455.

[3] S. Murata., et al. 2022. *Adv. Funct. Mater.* 32, 37, 2201866.



004-05

Integrated web user interface for DNA nanotechnology including coarse-grained molecular dynamics simulation

Ibuki KAWAMATA *

School of Science, Kyoto University

(* E-mail: kawamata.ibuki.8p@kyoto-u.ac.jp)

DNA nanotechnology is an expanding field of research that aims to fabricate nanoscale architectures and machineries in nanoscale precision using DNA as a building block. The advantage of using DNA is its programmability by which one can rationally design the interaction among DNA molecules by defining base sequences of DNA. Among several design strategies to build nanoscale objects, DNA origami methodology is receiving much attention thanks to its versatility and robustness. In a standard DNA origami structure, circular single-stranded DNA with more than 7000 nucleotides and more than 200 short single-stranded DNAs are used as materials. Because it is not practical to design a DNA structure by human trails and errors using pen and pencil, computer aided design tools and standardized file formats play important roles. Although many useful file formats and software have been developed, file conversion between different formats or graphical user interface to operate software such as molecular dynamics simulation is lacking. Here, we built a set of web user interface to accelerate the computational design, visualization, simulation, and analysis of DNA nanostructures. In the presentation, our research using the software will be discussed along with the introduction of the web user interface.

O04-06

Development of a Supervised Deep Learning Method for DNA Sequence Estimation from DNA Images

Hirotaka KONDO *¹, Akinori KUZUYA^{1, 2, 3}, Akihiko KONAGAYA³

¹Organization for Research and Development of Innovative Science and Technology, Kansai University

²Department of Chemistry and Materials Engineering, Kansai University

³Molecular Robotics Research Institute, Co., Ltd.

(* E-mail: kondo@kansai-u.ac.jp)

Atomic force microscopy (AFM) has attracted widespread attention as a technique for observing the shape and structure of DNA molecules on a nanoscale. However, extracting DNA sequence information directly from AFM images is difficult due to AFM's noises and resolution limitations, making it difficult to read nucleotide sequence precisely. Various methods have been attempted in recent years to overcome these challenges, using deep learning to remove noises and increase resolution [1].

In this study, we propose a new method to estimate sequences from DNA images using supervised machine learning. We generated 360 images of a DNA model in units of 1 degree using a molecular image viewer (VMD). Then, we aligned the pixel images so that the center of each image to create molecular surface images. Supervised machine learning was performed on between these images and the corresponding sequence information. In particular, we used a modified T with a large bulging functional group for T discrimination and a modified C for C discrimination to recognize the position of each modified DNA. Learning in groups of one mutation, two mutations, and multiple mutations resulted in accurate sequence discrimination.

Supervised machine learning was also performed on a grayscale image, which was more similar to the AFM image, and could be also useful to discriminate DNA sequences.

[1] Xianran HU, Qing LIU, Gregory GUTMANN, Masayuki YAMAMURA, Akinori KUZUYA, Akihiko KONAGAYA: High-Resolution AFM Imaging of DNA Structures: An Approach via Cycle GANs and Virtual Reality Integration, CBI 2023 conference, p.180 (2023).

O05-01

AlphaFold protein 3D structures enhance genome-wide scale compound-protein interaction prediction with deep learning

Yuga MORIYAMA ^{*1}, **Sae OKAMOTO**², **Tomokazu SHIBATA**², **Ryusuke SAWADA**³, **Yoshihiro YAMANISHI**¹

¹Graduate School of Informatics, Nagoya University

²Department of Bioscience and Bioinformatics, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology

³Graduate School of Medical and Dental Sciences, Okayama University

(* E-mail: moriyama.yuga.i8@s.mail.nagoya-u.ac.jp)

Keywords : protein 3D structures, ligand binding site, compound-protein interaction, graph neural network

In the early stages of drug discovery, the identification of compounds that regulate therapeutic target proteins of the disease is an important issue. However, experimental methods are costly and time-consuming. To solve the problem, machine learning plays a key role in compound-protein interaction (CPI) prediction. CPIs are greatly influenced by 3D structures of proteins; thus, it is ideal to consider protein 3D structures in the CPI prediction. However, most previous studies on CPI prediction have been using amino acid sequences as protein features, because only 14% of the proteins encoded in the human genome have fully determined 3D structures [1]. Recently, AlphaFold2 has made it possible to obtain genome-wide protein 3D structures from amino acid sequences, and comprehensive protein 3D structures are a useful resource for pharmaceutical research. Thus, the use of AlphaFold2 protein 3D structures may enhance the accuracy of CPI prediction.

In this study, we developed a deep learning-based method for genome-wide scale CPI prediction using 3D structures of proteins. First, compound structures were converted to graphs with atoms as nodes and bonds as edges, and the graphs were input to the graph neural network (GNN) to construct feature vectors of compounds. Second, protein structures are converted to three-dimensional interaction (3Di) alphabetic strings based on the distances and angles between amino acids with a discrete variational autoencoder, and the 3Di alphabets with 20 states representing the geometrical arrangement of amino acids were used to construct feature vectors of proteins. In addition, only protein 3D structures of the ligand-binding pockets were transformed into graphs with amino acids as nodes and spatial proximities as edges, and the

resulting pocket-constraint graphs were input to the GNN to construct feature vectors of proteins. These feature vectors were input to a fully connected neural network to predict CPIs. In the results, we confirmed that the protein structure information was more useful than the protein sequence information and our proposed method achieved the highest accuracy on benchmark datasets. The proposed method is expected to be useful for various applications in drug discovery.

[1] Ryusuke, S. et al. *iScience*, 27, 6, 110032 (2024).

[2] Jhon, J. et al. *Nature*, 596, 583–589 (2021).

005-02

The effect of food on pharmacokinetics of acotiamide for the treatment of functional dyspepsia

Kazuyoshi YOSHII *, Ryotaro IWAKIRI, Ryoko TODA

Ethical drug research, ZERIA Pharmaceutical Ltd., Co.

(* E-mail: kazuyoshi-yoshii@zeria.co.jp)

It is known that the pharmacokinetics of drugs can be altered by the meal itself or the physiological changes caused by the meal. In drug development, changes in pharmacokinetics due to meals can influence the design of clinical trials, making it important to evaluate the effect of meals in advance. Japanese guidelines recommend clinical trials to assess the effect of meals using the final formulation, and at that timing, various data are acquired, thus considering mechanism-based investigations such as physiologically based pharmacokinetic models is deemed important. Initially, machine learning models were examined for the purpose of predicting the effects of meals. Due to the low prediction accuracy of machine learning models using structural features, it was inferred that predicting the effects of meals based solely on structural information is difficult at the current time. Next, the impact of physiological changes caused by meals on pharmacokinetics was investigated using pharmacokinetic models. Within the range of compounds studied, it was found that changing the physiological parameters of the pharmacokinetic model alone was not sufficient to predict the effects of meals. Lastly, a case study was conducted using the pharmacokinetic model to investigate the effects of meals on the functional dyspepsia treatment drug, acotiamide. Compared to fasting administration, pre-meal administration of acotiamide reported an increase in C_{max} , and post-meal administration reported a decrease in AUC. The causes of these changes were considered to be acotiamide's promotive effect on gastric emptying and the inhibitory effect of meal components on absorption, hence a sensitivity analysis of related parameters was conducted. The sensitivity analysis showed that by altering the gastric transit time and membrane permeation rate, it was possible to simulate the effects of meals on acotiamide. Therefore, at the current stage, a case-by-case investigation using pharmacokinetic models is considered effective for understanding the effects of meals on the pharmacokinetics of drugs. Acotiamide is administered before meals to relieve symptoms of functional dyspepsia such as postprandial fullness, and pre-meal administration shows positive effects, suggesting that taking it before meals could lead to more effective symptom relief.

O05-03

Computational exploration of bipolar disorder multi-omics data in the quest for novel drug targets

Flora R. AIGBE *, Kosuke HASHIMOTO, Kenji MIZUGUCHI

Laboratory for Computational Biology, Institute for Protein Research, Osaka University

(* E-mail: u383788b@ecs.osaka-u.ac.jp)

Bipolar disorder (BD) is a difficult to manage psychiatric mood disorder. Despite available management options, it is associated with higher mortality rates with increasing incidence. This is partly due to its obscure pathophysiology and consequent lack of adequate drug targets. Potential drugs for proposed new targets such as Ca²⁺-calmodulin-dependent protein kinase kinase-2 raise safety concerns due to cellular metabolic implications. The polygenic nature of BD is still under investigation; studies continue to implicate new genes. Progress has been slowed by the difficulty in modelling BD or obtaining relevant data. Computer-aided approaches have been limited by focus on single omics data type (per study), such as genomic or transcriptomic data alone. This study was designed to systematically evaluate publicly available multi-omics data to identify robust signatures that will aid in potential new target(s) identification. Whole genome microarray (GSE62191), RNAseq transcriptomic (GSE53239) and DNA methylation epigenomic (GSE112179) data were obtained from the Gene Expression Omnibus (GEO) database. These data sets were analyzed using R's weighted gene co-expression network analysis (WGCNA) respectively. BD-associated genes or probes from this step were then used for differential expression analysis (DEA) using limma R package and subsequent evaluation to derive a list of DE genes that overlap across all the different data types, to be used for further analyses.

Using WGCNA, clusters or modules of significantly correlated (Pearson's $r \geq 0.5$; Student's t-test $p < 0.05$) co-expressed genes or probes with BD were identified from all the data sets. Of these genes, upregulated and downregulated DE genes were selected based on Benjamini-Hochberg's adjusted P value less than 0.05 for each data type respectively. About 474 genes from the whole genome microarray data, 5,249 transcripts from the RNAseq transcriptome data and 6,512 gene-mapped CpG methylation sites, or their respective probes, were found to be differentially expressed in the BD group relative to the control group. These were then sorted for overlapping genes across all the data types. Of these, at least 1 unannotated probe and 40 genes were identified to overlap across all

3. Among the annotated 40 overlapping genes are ARHGEF10L, with a functionally related BD-associated gene, ARHGEF7, and DGKD, which has been previously linked to BD. Findings so far show the potential of WGCNA followed by DEA to help identify potential new players in BD across multiple independent data sets and aid new target identification.

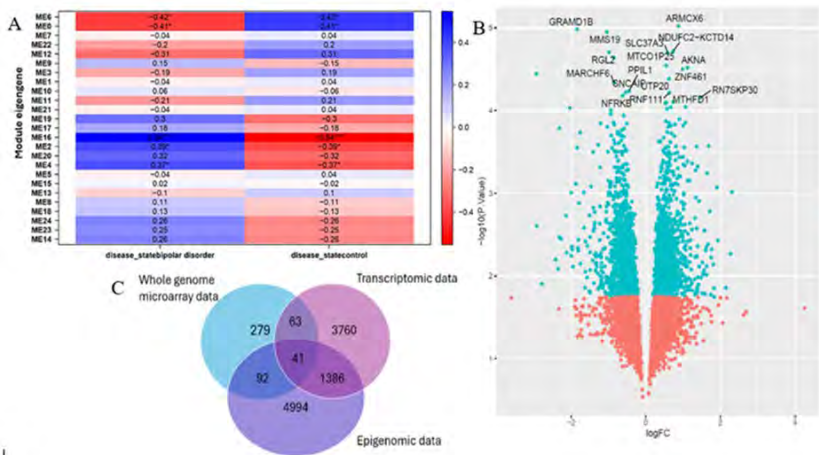


Figure 1: Representative visualization of WGCNA, DE analysis and overlapping genes across omics data types. A- Heat map of module and disease trait relationship for whole genome microarray data (WGM), B-Volcano plot of differentially expressed WGCNA derived BD associated genes for transcriptomic data, C- Venn diagram showing overlapping differentially expressed BD associated genes across WGM, transcriptomic and epigenomic data sets.

005-04

Learning the Language of Life: Feasibility of Using LLMs to Understand Latent Characteristics of Proteins from Residue Structural Environments

Nina HOLSMOELLE *, Kenji MIZUGUCHI, Gert-Jan BEKKER

Laboratory for Computational Biology, Institute for Protein Research, Osaka University

(* E-mail: n.holsmoelle@gmail.com)

In this project, we explored the feasibility of using Large Language Models (LLMs) to extract knowledge such as dynamic structural information from static protein structures. To uncover latent characteristics inherent within static configurations, we combined machine learning techniques with one-dimensional representations of local protein structures, assuming an inherent structural logic that is learnable.

Categorizing local environments of amino acids, we constructed datasets from the Protein Data Bank (PDB) and employed a Masked Language Model (MLM) for feature learning. From the analysis of the trained model's high-dimensional residue representations, we concluded that the model has indeed been able to successfully acquire a partial understanding of the amino acids' characteristics and structural properties.

The goal of this research is to advance our understanding of protein interactions and functions, which holds significant implications for medical and health-related issues, such as drug design and disease treatment. With our findings, we hope to introduce a novel methodology for integrating static and dynamic protein data, paving the way for innovations in protein modelling and biomedical applications.

005-05

SGCRNA: A Novel Tool for Gene Co-Expression Network Analysis Using Spectral Clustering

Tatsunori OSONE *, Takeshi TAKARADA

Department of Regenerative Science, Okayama University
(* E-mail: osone@okayama-u.ac.jp)

Weighted Gene Co-Expression Network Analysis (WGCNA) is a potent methodology that is capable of identifying functional modules and pivotal genes through the analysis of gene co-expression patterns. The utilisation of WGCNA for the structural analysis of gene expression data enhances the comprehension of intricate biological processes, thereby leading to its widespread adoption in various studies. Since its inception in 2005, numerous advanced tools, including applications to single-cell RNA-seq, have been developed. However, the scale-free nature of networks, a premise of WGCNA, was widely accepted two decades ago when WGCNA was first proposed. Recently, it has been reconsidered that only a limited number of networks exhibit these properties. Additionally, from a usability perspective, WGCNA necessitates manual parameter adjustments in at least two steps. Biologically, although correlations are considered, the ratios between genes are not, and negative correlations are fundamentally disregarded. To address these four points, a novel method has been developed. By applying this method to publicly available data from bulk RNA-seq, single-cell RNA-seq, and spatial transcriptome analyses, biologically meaningful results have been obtained in all cases. Compared to the outcomes derived from conventional methods, the new method has demonstrated higher utility and increased depth of analysis. These findings suggest that the new method can produce biologically significant results within a reasonable execution time without relying on arbitrary parameter adjustments. As this study lacks wet lab validation, it is necessary to empirically verify whether the hub genes and their putative controlling transcription factors predicted by the new method are accurate.

O05-06

Exploring cancer treatment candidates targeting chromatin remodeling factors

Reiko WATANABE *¹, Junsoo SONG¹, Ui AYAKO², Mizuguchi KENJI¹

¹Laboratory for Computational Biology, Institute for Protein Research, Osaka University

²Institute of Development, Aging and Cancer, Tohoku University

(* E-mail: reiko-watanabe@protein.osaka-u.ac.jp)

The SWI/SNF chromatin-remodeling family comprises various protein complexes that regulate gene expression during cellular development and influence the DNA damage response in an ATP- and complex-dependent manner. Recent genome sequencing of various cancer cells has revealed frequent mutations in SWI/SNF components, particularly in ARID1A, a variant subunit in the BRG1-associated factor (BAF) complex of the SWI/SNF family. ARID1A mutations or loss of function are commonly observed in several cancers, including ovarian, endometrial, and gastric cancers, and are contributing to cancer development by disrupting normal cell growth regulation, promoting genomic instability, and enhancing tumor progression. In this study, we aimed to explore potential therapeutic targets for cancer treatment based on two strategies. First, we proposed an innovative patient stratification strategy considering both ARID1A protein expression level and its function and identified differentially expressed genes (DEGs) between groups. Our method highlights transcriptional variations of tumor immune microenvironment, which were hard to detect with stratification based on only mutation information. Second, synthetic lethality (SL) was considered to search potential therapeutic targets. Drugs targeting SL partners of ARID1A could potentially treat cancers with ARID1A loss while sparing normal cells. Although many wet-lab methods have been developed to screen for SL pairs, the number of known SL pairs is currently a very small fraction of all candidate pairs due to the large number of human gene combinations. By using computational prediction tools that can effectively reduce the search space for SL pairs, potential SL pairs for ARID1A were collected and screened based on prognosis in tumor patients. These approaches can propose potential targets in cancer therapeutics and by integrating molecular stratification and synthetic lethality strategies, it can be possible to develop more effective and less toxic therapeutic interventions tailored to the specific genetic makeup of cancer patients.

006-01

3D structure-based chemical foundation model to predict the bioactivity and toxicity

Tsuyoshi KIMURA *, Yoshihiro YAMANISHI

Graduate School of Informatics, Nagoya University

(* E-mail: kimura.tsuyoshi.p9@s.mail.nagoya-u.ac.jp)

It is extremely challenging to identify drug candidate compounds with desired properties [1]. The number of possible organic compounds is estimated to exceed 10^{60} ; thus, experimental validation of all possible compounds is infeasible due to time and cost constraints. Machine learning approach plays a key role in the compound screening. However, labeled datasets of compounds with known properties are small in most cases, which makes it difficult to generalize the machine learning model well with the small labeled datasets. To address the problem, the foundational model has been receiving much attention. Most previously developed foundational models use a self-supervised pretraining on large unlabeled datasets such as SMILES strings or molecular graphs, followed by fine-tuning on smaller and labeled datasets. However, molecular properties such as bioactivity and toxicity depend heavily on the 3D structures of compounds; thus, it is difficult to accurately predict the molecular properties with foundation models that ignore the 3D structures of compounds. In this study, we propose a 3D structure-based chemical foundation model to predict various molecular properties (e.g., bioactivity and toxicity) of drug candidate compounds. Our foundational model is pre-trained based on 3D structures of compounds with an optimization technique that is more accurate than RDKit, and is fine-tuned for various molecular property prediction tasks. The optimization of compounds involves a supervised learning to match randomly generated 3D conformations with labeled conformations [2]. The pre-training task involves a self-supervised learning, where the inputs are 3D structures with randomly missing atoms and the structures are reconstructed from the 3D coordinates of the remaining atoms. In the results, we confirmed that our proposed method worked equally or better than existing methods, even with an extremely smaller dataset. The proposed model is expected to reduce the necessary computational resources and be applicable to complex molecular property prediction tasks.

[1] Regine Bohacek, *et al. Med. Res. Rev.*, **1996**.

[2] Gengmo Zhou, *et al. The Eleventh International Conference on Learning Representations*, **2023**.

006-02

Comparison Between Word Embeddings and Molecular Descriptors by Clustering and Distribution Analysis from Antioxidant Articles

Yuto MATSUMOTO *, Hiroaki GOTOH

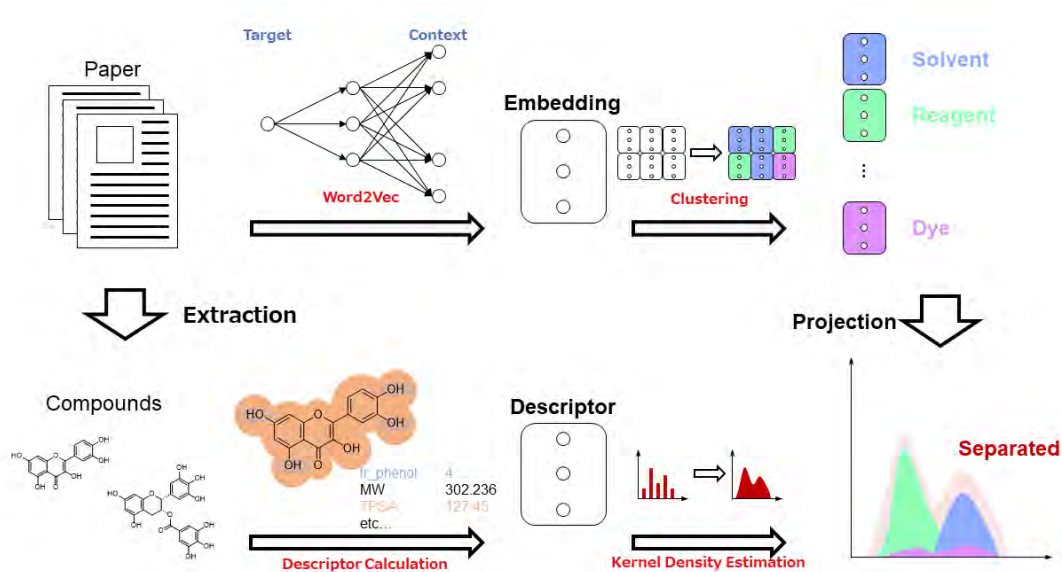
Graduate School of Engineering, Yokohama National University
(* E-mail: matsumoto-yuuto-sh@ynu.jp)

In chemistry, antioxidant, which protect oxidation to organism, such as cell wall, is very important factor in the view of immunity. However, antioxidant effect is emerged by various mechanisms and various reaction, so exploration of antioxidant is very difficult without many trials or filtering. On the other hand, the large amount of previous study about antioxidant was published but only parts of those are used by each chemist. Focusing on this untouched information to predict antioxidant capacity and seek important factor of prediction or emerging mechanisms, we process antioxidant journals by natural language processing(NLP).

For examples of NLP in chemistry, there are ChemDataExtractor and few-shot learning using generative pre-trained transformer models. On the other hand, most of the currently reported models focus on information extraction and qualitative prediction, and there are few examples in which NLP models are combined with molecular descriptors, which are quantitative values calculated from the structure of compounds, and using to predict. so our purpose sets quantitative prediction, we compared between descriptors and embeddings, and investigated projection between those in this study.

We obtained 1744 articles of *Antioxidants*, which is the journal published by MDPI corporation, from 1996 to 2020 and used the Word2Vec model to obtain embeddings. Among the embeddings, 1294 words meaning compounds were clustered. In the compounds classified by the clustering, we identified features in the usage of each cluster, such as being used as a solvent, a reagent, involved in the mechanism of action in the body, and used as a target for antioxidant capacity assays. In addition, we identified structural features of the compounds, such as the abundance of flavonoids, in some clusters. To analyze in detail the correlation of this classification on structural features, kernel density estimation (KDE) was performed on 208 standardized molecular descriptors, which has calculated by RDKit, over each cluster. This confirmed the distribution of molecular descriptors by cluster based on the mean, variance, and graph shape of the KDEs. In this visualization, clusters often had isolated means or low variances, and simple modal curve even though we did not use molecular

descriptors under clustering methods. The results confirm the dependence of each cluster on molecular descriptors. Two of the clusters tended to have a specific shape with respect to the classified compounds and their molecular structures. This study revealed that there is a correspondence between linguistic analysis using embeddings and the meaning of structures in the compound space represented by molecular descriptors. We plan to use this method to clarify the structural effects on activity from the results of linguistic processing of various journals.



O06-03

Reconstructable latent representation of molecules by Graph Transformer VAE

Yasuhiro YOSHIKAI *, Tadahaya MIZUNO, Hiroyuki KUSUHARA

Laboratory of Molecular Pharmacokinetics, The University of Tokyo Graduate School of Pharmaceutical Sciences

(* E-mail: yoshikai-yasuhiro701@g.ecc.u-tokyo.ac.jp)

Obtaining latent representation of molecules is one of the key processes in data science for chemistry, as it enables various downstream tasks like molecular property prediction.

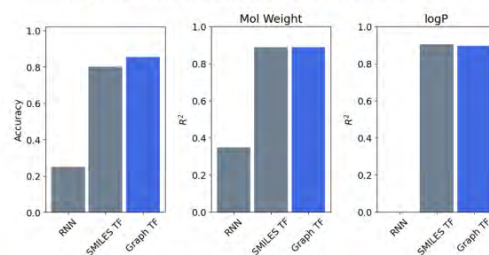
One of the major representation learning architectures is autoencoder, which learns to encode molecules to latent representations and then decode back to their original structures. This architecture has the advantage that the model can be trained in an unsupervised way, without auxiliary information. However, existing representations like those obtained by unsupervised representation learning have difficulty in restoring the original structure, and existing restorable descriptors, such as those generated by Variational Autoencoder (VAE), have not been well studied about their reconstruction performance.

We first examined the reconstruction performance of molecule of several existing restorable descriptors and found that many of those actually have little reconstruction ability. Besides, the restored molecules often show different molecular properties, such as molecular weight or logP. To address this, we developed a molecular representation which can reconstruct original molecules with high accuracy. Our model is based on VAE, and utilizes graph Transformer in Encoder. The developed representation showed high accuracy in reconstructing the property and structure of original molecules. Notably, we found that decreasing the weight of the KL divergence term of VAE in the reconstruction loss improves the reconstruction performance, while degrading the continuity of the latent space. These results are expected to provide the foundation of unsupervised learning of molecules, and contribute to the improvement and proper usage of restorable latent representation.

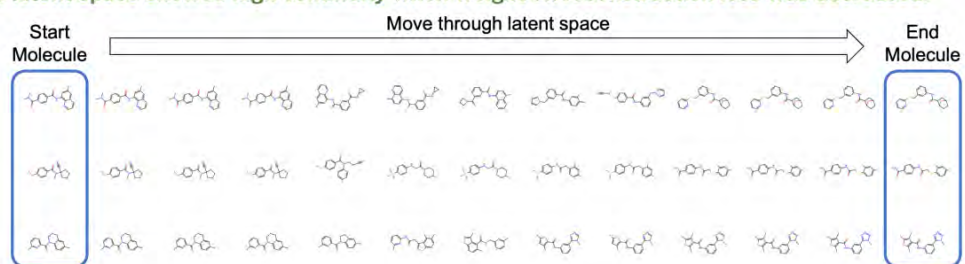
① Graph-based VAE model was developed



② The model showed high reconstruction performance



③ The latent space showed high continuity when weight on reconstruction loss was decreased.



O06-04

Towards the Design of Natural Product Biosynthetic Gene Clusters Using Natural Language Processing Technology

Tomoki KAWANO ^{*1}, **Taro SHIRAISHI**^{1, 2}, **Maiko UMEMURA**³, **Tomohisa KUZUYAMA**^{1, 2}

¹Graduate School of Agricultural and Life Sciences, The University of Tokyo

²Collaborative Research Institute for Innovative Microbiology, The University of Tokyo

³Bioproduction Research Institute, National Institute of Advanced Industrial Science and Technology

(* E-mail: kawnao-tomoki030@g.ecc.u-tokyo.ac.jp)

Natural products, predominantly synthesized by microorganisms and plants, play crucial roles in modern medicine. These compounds are typically produced by biosynthetic gene clusters (BGCs), which are groups of genes collectively responsible for the synthesis of specific metabolites. With the advent of genome sequencing technologies, a vast amount of genomic data has been accumulated, revealing numerous unknown BGCs.

This study explores the application of natural language processing (NLP) techniques, specifically the RoBERTa algorithm, to understand and generate BGCs. We hypothesized that the evolutionary process of BGCs is recorded in genomic sequences as positional information, analogous to tree rings. By treating functional domains within genes as tokens and gene clusters as sentences, we aimed to capture the relationships between domains using NLP models.

We prepared four datasets of varying complexity: 1) BGCs from the antiSMASH database, 2) complete genomes of Actinobacteria, 3) all bacterial complete genomes, and 4) bacterial and fungal genomes. These datasets were used to train RoBERTa models.

Our results demonstrated that:

1. RoBERTa models successfully learned the relationships between functional domains in BGCs.
2. The models accurately predicted masked domains within known BGCs from the MIBiG database, with over 60% of domains correctly predicted as the top

choice and about 90% within the top 10 predictions.

3. The model trained on the most diverse dataset (bacterial and fungal genomes) showed enhanced ability to provide biologically plausible alternatives, potentially useful for exploring novel BGCs.
4. The models could predict the compound class of BGCs with high accuracy, especially for well-studied classes like NRPSs and PKSs.
5. We demonstrated the potential for generating novel BGCs by masking and predicting domains within existing clusters, as exemplified with the cyclooctatin biosynthesis pathway.

Our current models are not specifically trained for BGC generation. As a future direction, we are exploring seq2seq training methods where plausibly less evolved BGCs are used as input to predict more evolved BGCs as output. This approach could lead to the development of models capable of generating evolved BGCs, potentially opening new avenues for the discovery of novel natural products and their biosynthetic pathways.

This study represents a significant step towards leveraging the vast amount of accumulated genomic data for the rational design of novel biosynthetic pathways and the discovery of new natural products, potentially revolutionizing the field of natural product drug discovery.

Table: Statistical values of prediction results for each model

Rank	Model 1 (BGC)	Model 2 (Actinobacteria)	Model 3 (Bacteria)	Model 4 (Bacteria+fungi)
Mean	164.9	286.7	128.9	65.5
Median	1	1	1	1
Minimum	1	1	1	1
Maximum	19658	19696	19679	19634
#1	67.5%	62.4%	68.5%	64.6%
Within 10	88.1%	85.1%	88.8%	87.0%
Above 1000	2.23%	3.54%	1.76%	1.30%

We evaluated the prediction accuracy of four models using 2,492 BGCs. For each BGC, we masked internal tokens and had the models predict them. The table shows the statistical values of the ranks assigned by each model to the correct tokens. A smaller average rank indicates more accurate predictions.

O06-05

Design of Novel Compounds Through Protein-Ligand Interaction-Based Generative Methods

Mami OZAWA¹, Shogo NAKAMURA¹, Nobuaki YASUO², MASAKAZU SEKIJIMA *¹

¹Department of Computer Science, Tokyo Institute of Technology

²TAC-MI, Tokyo Institute of Technology

(* E-mail: sekijima@c.titech.ac.jp)

The generation of new compounds with protein-ligand interactions is an important issue in structure-based drug design. In this study, we propose a model for generating new compounds called “IEV2Mol” that incorporates the interaction energy vector (IEV) between the protein and ligand obtained from docking simulations. This IEV quantitatively captures the strength of each interaction type, such as hydrogen bonds, electrostatic interactions, and van der Waals forces, and unlike the conventional interaction fingerprint (IFP), it reflects the strength of the interaction. IEV2Mol, by integrating IEV into an end-to-end variational autoencoder (VAE) framework that learns chemical space from SMILES representations and minimizes SMILES reconstruction error, can generate compounds with the desired interactions more accurately.

To evaluate the effectiveness of this model, we conducted benchmark tests comparing it with randomly selected compounds, an unconstrained VAE model (JT-VAE), and an RNN model based on interaction fingerprints (IFP-RNN). The results showed that the compounds generated by IEV2Mol had a significantly higher rate of retaining the binding mode of the query structure than the other methods.

The IEV2Mol proposed in this study is expected to contribute to the efficiency of compound generation based on interaction energy in the design of new compounds for target proteins. In addition, the source code and trained models for IEV2Mol, JT-VAE, and IFP-RNN used in this study are available under the MIT license.

O06-06

Analysis of the usefulness of AlphaMissense score for predicting protein function. -Evaluation by *GLA*, the causative gene of Fabry disease-

Yuji SAKAHASHI ^{*1}, **Yohei MIYASHITA**^{2, 3}, **Yasuki ISHIHARA**², **Osamu YAMAGUCHI**^{1, 4}, **Yoshihiro ASANO**^{2, 3}

¹Omics Research Center, National Cerebral and Cardiovascular Center

²BioBank, National Cerebral and Cardiovascular Center

³Department of Genomic Medicine, National Cerebral and Cardiovascular Center

⁴Department of Cardiology, Pulmonology, Nephrology, and Hypertension, Ehime University

(* E-mail: sakahashi.yuji@ncvc.go.jp)

With the development of next-generation sequencing technology, Variant of Uncertain Significance (VUS) is accumulating. Missense variants, which account for many VUS, tend to have limited training data due to insufficient functional analysis of the mutant protein, making it difficult to create highly accurate in silico pathogenicity prediction models. In this regard, AlphaMissense was announced in 2023 as a new algorithm for predicting the pathogenicity of missense variants. AlphaMissense is a pathogenicity prediction tool based on protein structure information from AlphaFold and has high prediction accuracy for known pathogenic variants. However, because AlphaMissense makes predictions based on protein structural information, its predicted impact on protein function or correlation with clinical phenotypes remain controversial.

In this study, we focused on *GLA*, the gene responsible for Fabry disease, and evaluated the relationship between the *in vitro* enzyme activity measurement results of α GAL mutants as a protein function and the AlphaMissense prediction results. From a list of 2,850 *GLA* variants predicted by AlphaMissense, we used 633 variants for which *in vitro* enzyme activity data have been published in several previous studies. The enzyme activity of the mutants was evaluated using pctWT, which represents the activity of the α GAL mutant when the activity of the α GAL wild type is taken as 100%. First, the correlation between AlphaMissense score and pctWT was evaluated for 633 variants. The results showed a negative correlation with a correlation coefficient of -0.57. Next, machine learning models were constructed to predict changes in enzyme activity of α GAL mutants using the AlphaMissense score. Three models (Logistic regression, SVM, and XGBoost) were created to classify variants with pctWT reduced to less than 5% as "Severe-LOF" and those with pctWT between 5% and 100 as "Mild-LOF". To build the models, we used the features from

AlphaMissense and the PDB file of α GAL predicted by AlphaFold. The analysis showed that the ROC-AUC for all three models was around 0.8 (Figure 1). These results indicate that the AlphaMissense score is a useful indicator for predicting changes in α GAL enzyme activity. This may be because 1) since α GAL is an enzyme, changes in protein conformation are closely related to changes in enzyme activity, and 2) there is sufficient experimental data on mutants that can be used to build and validate the model. Based on the above, we are currently measuring the enzyme activity of unknown variants of α GAL mutants for which enzyme activity data has not been obtained in previous studies and evaluating the agreement with the model prediction results.

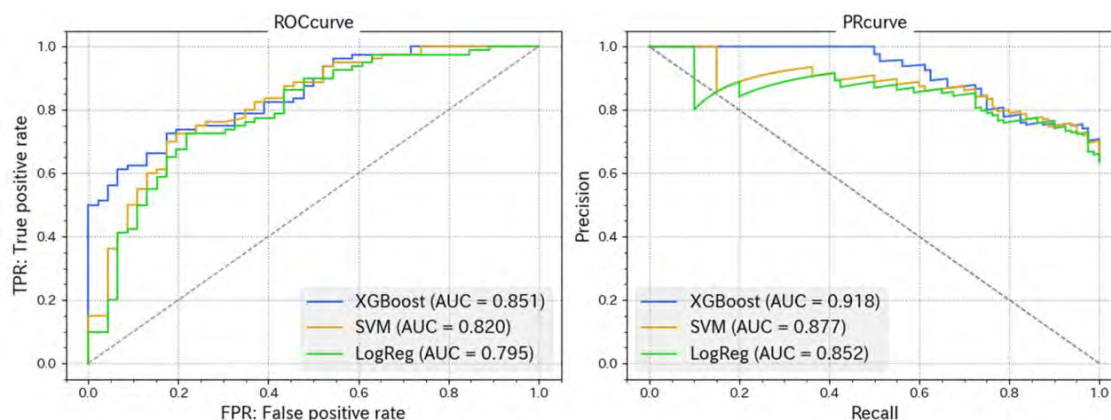


Figure 1: Results of a machine learning model to predict enzyme activity variation in α GAL mutants.

007-01

SynthFormer: A Customizable Framework for Virtual Synthesis-Based Molecule Generation

Joshua OWOYEMI *, Tasuku ISHIDA

Elix, Inc.

(* E-mail: joshua.owoyemi@elix-inc.com)

Generating new molecules with specific properties is a crucial aspect of drug discovery and materials science. Traditional methods, such as virtual screening, are limited in their ability to efficiently explore the vast chemical space. Recent advances [1] in machine learning have led to the development of generative models for molecule design, but these often overlook the crucial aspect of synthesizability, hindering their practical application. We present SynthFormer, a novel framework for molecule generation and optimization that addresses synthesizability by directly incorporating chemical reactions and building blocks into the design process. This approach ensures that the generated molecules are not only novel and possess desirable properties but are also synthetically accessible, bridging the gap between computational design and experimental realization. The framework's adaptability allows for customization based on specific requirements, including the choice of chemical reactions, building blocks, and optimization algorithms and objectives, making it a valuable tool for various domains within chemistry and materials science. We evaluate the framework by exploring the de novo generation of selected patented compounds and show that the framework is able to suggest similar compounds while proposing feasible synthetic routes to achieve the generated molecules. We also compare the performance of multiple optimization approaches such as Monte Carlo Tree Search [2] and Reinforcement Learning [3] while utilizing the framework for the rediscovery of known compounds.

[1] Xiangru Tang, Howard Dai, Elizabeth Knight, Fang Wu, Yunyang Li, Tianxiao Li, Mark Gerstein. 2024. A survey of generative AI for de novo drug design: new frontiers in molecule and protein generation. *Briefings in bioinformatics*, 25(4), bbae338. <https://doi.org/10.1093/bib/bbae338>

[2] Shoichi Ishida, Tanuj Aasawat, Masato Sumita, Michio Katouda, Tatsuya Yoshizawa, Kazuki Yoshizoe, Koji Tsuda, Kei Terayama. ChemTSv2: Functional molecular design using de novo molecule generator. 2023. *WIREs Comput Mol Sci.*; 13(6):e1680. <https://doi.org/10.1002/wcms.1680>

[3] Sai Krishna Gottipati, Boris Sattarov, Sufeng Niu, Yashaswi Pathak, Haoran

Wei, Shengchao Liu, Karam J. Thomas, Simon Blackburn, Connor W. Coley, Jian Tang, Sarath Chandar, and Yoshua Bengio. 2020. Learning to navigate the synthetically accessible chemical space using reinforcement learning. In Proceedings of the 37th International Conference on Machine Learning (ICML'20), Vol. 119. JMLR.org, Article 344, 3668–3679.

007-02

Relationship between cyclic peptide structure in solution and membrane permeability

Takashi MATSUMOTO *, Takashi SATO

Rigaku Corporation

(* E-mail: t-matumo@rigaku.co.jp)

Cyclic peptide drugs are attracting attention as a new modality that have complementary characteristics of small molecule drugs and biopharmaceuticals and are thought to have the potential to penetrate the limitations associated with each. In developing cyclic peptide drugs, three-dimensional structural information is essential for optimizing their interactions with target molecules and improving their functions. Solution NMR-based structural analysis has been the mainstream method, but sometimes, it becomes challenging because functional association and/or aggregation cause severe changes in inherent flexibility and relaxation time. On the other hand, crystal structures give clear-cut atomic coordinates. Still, packing artifacts are known to occur in crystal structures, especially for flexible peptide molecules; thus, there is always a question of whether the obtained fine structure reflects its functional states.

Small-angle X-ray scattering (SAXS) has been used to obtain protein structural information in the solution. However, the small-angle data represents long-range information such as the outer shapes of the proteins and/or protein-protein interactions and the lack of more detailed information in a finer distance range. To see more detail, such as the displacement of a loop region in a domain structure, we may need higher angle datasets that contain shorter-range information than the established SAXS dataset. However, the scattering signal decreases exponentially in solution through the scattering angle since the target molecules are vigorously moving in the solution, and those motions randomize distance vectors between fine structures.

Even in such a challenging landscape, if we can obtain the scattering data in good quality up to a sufficient resolution to observe the detailed characteristics and conformational movements, it will help to understand the function and conformational characters in the solution, which we have not known yet. We are developing technologies that enable that observation and abbreviating MAXS (middle angle x-ray scattering), which employs scattering data in a broader range up to higher q area to obtain more detailed structural information of proteins and especially open the doors to the analysis of peptides in solution, that possesses molecular size under the order of domain flexibility.

Here, we present solution conformations by two low molecular weight cyclic peptides, Polymyxin B and Cyclosporine A, by X-ray solution scattering. Those peptides were measured in well-dispersed conditions in pure water and anhydrous ethanol, respectively, and a detailed conformation of those peptides has been obtained utilizing MAXS data.

A comparison of the conformations of those peptides in solution and crystal states revealed changes in the structural nature and the usability of X-ray scattering for those cyclic peptides. We are extending the analysis to more complex systems where other spectroscopic methods have severe limitations.

007-03

Efficient docking simulation-based generation of bioactive compounds with deep generative models

Hideto HOSHINO *, Li CHEN, Yamanishi YOSHIHIRO

Graduate School of Informatics, Nagoya University
(* E-mail: hoshino.hideto.g7@s.mail.nagoya-u.ac.jp)

The identification of bioactive compounds that regulate the function of a therapeutic target protein is important in the drug development, but conventional experimental methods are costly and time-consuming. Thus, deep learning-based structure generators have been studied as a more efficient method[1]. Most of the previous studies use a quantitative structure-activity relationship (QSAR) model trained on chemical structures and the corresponding bioactivities in the reward function of structure generator. However, if there is insufficient bioactivity data used for training a QSAR model, the accuracy of the QSAR model tends to be low, and the quality of the newly generated compounds generated by the structure generator tends to be poor. A possible solution is to perform docking simulations[2] for calculating the binding affinity with the three-dimensional structure of a therapeutic target protein in the reward function of the structure generator, but it requires huge computational costs. In this study, we developed an efficient docking simulation-based structure generator that generates new bioactive compounds with high binding affinity to a therapeutic target protein, which was made possible by incorporating a binding affinity QSAR model into a pure transformer encoder-based generative adversarial network (TenGAN)[3]. First, docking simulations against a given target protein were performed with pre-selected compounds. Next, a QSAR model was trained to predict binding affinity scores from the chemical structures using the binding affinity scores calculated by the docking simulations. Finally, new compounds with high binding affinity were generated using the predicted binding affinity as a reward with reinforcement learning in the structure generator. As a case study, we showed the usefulness of the proposed method in the design of new bioactive compounds for various target proteins such as epidermal growth factor receptor (EGFR). The introduction of the binding affinity QSAR model eliminated the need for docking simulations during the GAN training, which enabled a significant reduction in computational cost. For example, the proposed method requires just one day for generation of 5,000 compounds, while conventional methods require approximately 280 days. The proposed method is expected to be useful for rapid structure design of bioactive

compounds for any target proteins for which three-dimensional structures are available.

- [1] Kaitoh et al, Journal of Chemical Information and Modeling, 61, 4303-4320, 2021
- [2] Jerome et al, Journal of Chemical Information and Modeling, 61, 3891-3898, 2021
- [3] Li et al, International Conference on Artificial Intelligence and Statistics, 238, 361-369, 2024

007-04

Optimization of Generator Reward Function Settings for Non-covalent KRAS Inhibitors

Casey J. GALVIN *, Masakazu ATOBE

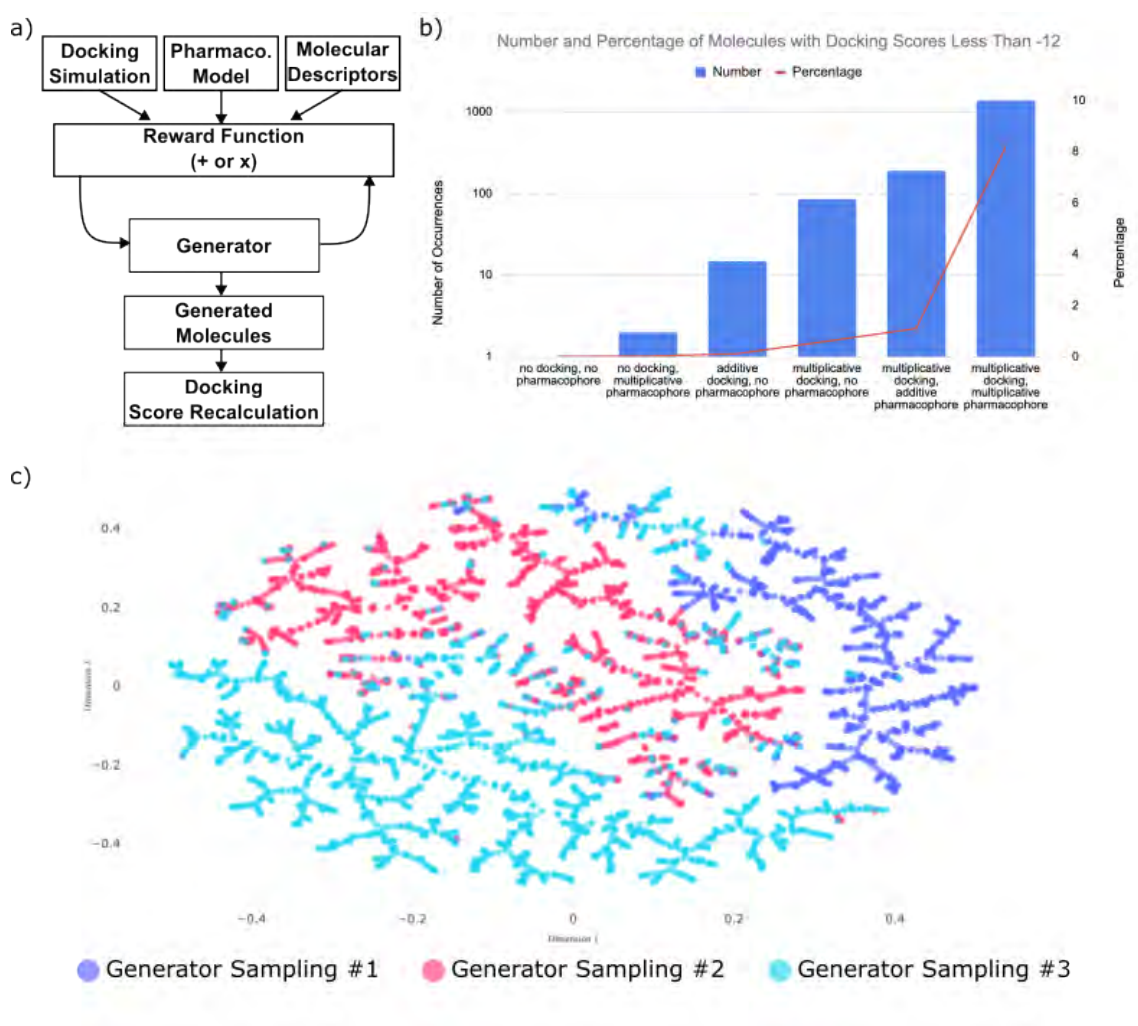
Elix, Inc.

(* E-mail: casey.galvin@elix-inc.com)

KRAS mutations are linked to various cancers, however the first inhibitors were only introduced in 2021 [1]. These inhibitors bind covalently to the cysteine residue specific to G12C, prompting the need for non-covalent inhibitors for other KRAS mutants [2]. This study combines generative models and computational chemistry techniques to design such inhibitors. We incorporate docking simulations and pharmacophore models based on a single crystal structure of a covalent KRAS inhibitor into a reward function that guides a generative model to optimize molecular design under a data-scarce regime. Incorporating either technique into the generator reward function improved the number of molecules achieving docking scores below our target threshold (lower is better), with the best results achieved by combining both pharmacophore models and docking scores as multiplicative components. This approach led to approximately 10% of the molecules (over 1,000 total) exhibiting high docking performance, which is an order of magnitude or more greater than generators without docking and/or pharmacophore model components. This finding demonstrates the strength of a multi-faceted reward function. Clustering based on molecular scaffold of high-performing molecules generated by identical reward functions run multiple times revealed distinct chemical space exploration across the different runs. This finding supports the need for multiple generator samplings to properly surface potential molecules of interest. This study highlights specific strategies in designing non-covalent pan-KRAS inhibitors by optimizing reward functions that integrate docking simulations and pharmacophore modeling. The synergy between these rewards as multiplicative factors underscores the importance of multi-component reward functions in advancing drug discovery, and identifies a strategy of overcoming the data scarcity typical of early drug discovery campaigns.

[1] Mullard, A. The KRAS Crowd Targets Its next Cancer Mutations. *Nature Reviews Drug Discovery* 2023, 22 (3), 167–171.
<https://doi.org/10.1038/d41573-023-00015-x>.

[2] Kim, D.; Herdeis, L.; Rudolph, D.; Zhao, Y.; Böttcher, J.; Vides, A.; Ayala-Santos, C. I.; Pourfarjam, Y.; Cuevas-Navarro, A.; Xue, J. Y.; Mantoulidis, A.; Bröker, J.; Wunberg, T.; Schaaf, O.; Popow, J.; Wolkerstorfer, B.; Kropatsch, K. G.; Qu, R.; de Stanchina, E.; Sang, B.; Li, C.; McConnell, D. B.; Kraut, N.; Lito, P. Pan-KRAS Inhibitor Disables Oncogenic Signalling and Tumour Growth. *Nature* 2023, 619 (7968), 160–166. <https://doi.org/10.1038/s41586-023-06123-3>.



007-05

Development of Scaffold and Fragment Definition Algorithms with a Case Study on Chemical Library Analysis

Kazuma KAITOH ^{*}, Yoshihiro YAMANISHI

Graduate School of Informatics, Nagoya University

(^{*} E-mail: kaitoh@i.nagoya-u.ac.jp)

One of the key roles of chemoinformatics in drug discovery is the analysis of chemical libraries. By integrating insights from these analyses into the design of new chemical libraries, the identification of hit and lead compounds can be significantly enhanced. A common approach to analyzing chemical libraries involves dividing compounds into substructures. Scaffold extraction, which identifies the core structures of compounds, is a widely employed method. Bemis-Murcko scaffold (BM scaffold)^[1] is a well-established algorithm for scaffold extraction that defines ring systems as scaffolds. However, the BM scaffold has a notable limitation: it cannot define scaffolds for compounds lacking ring structures. This limitation can result in the extraction of scaffolds that do not accurately represent the structural reality of certain compounds.

Beyond scaffold extraction, several algorithms exist for fragmenting compounds, such as the Breaking of Retrosynthetically Interesting Chemical Substructures (BRICS)^[2]. However, these algorithms occasionally fail to generate appropriate fragments for specific compounds. In this study, we developed K scaffold algorithm for scaffold extraction and kBRICS algorithm for fragment extraction. These methods were applied to the ChEMBL 34 database to analyze the characteristics of the resulting scaffolds and fragments. K scaffold algorithm extends BM scaffold rules by incorporating graph structures within compounds, while kBRICS enhances BRICS by adding rules based on organic chemical reactions.

When applied to the 2,180,814 organic compounds in ChEMBL 34, BM scaffold failed to define scaffolds for 19,338 compounds, whereas the K scaffold successfully defined scaffolds for all compounds. Additionally, kBRICS generated fragments for a greater number of compounds compared to BRICS, rBRICS^[3], and pBRICS^[4]. Detailed analyses of these findings will be presented during the conference.

[1] Bemis, G. W.; Murcko, M. A, The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, 39, 2887-2893.

[2] Degen, J.; *et. al.*, On the Art of Compiling and Using 'Drug-Like' Chemical

Fragment Spaces. *ChemMedChem* **2008**, 3, 1503-1507.

[3] Zhang, L.; *et. al.*, r-BRICS – A Revised BRICS Module that Breaks Ring Structures and Carbon Chain. *ChemMedChem* **2023**, e202300202.

[4] Vangala, S. R.; *et. al.*, pBRICS: A Novel Fragmentation Method for Explainable Property Prediction of Drug-Like Small Molecules. *J. Chem. Inf. Model.* **2023**, 63, 5066-5076.

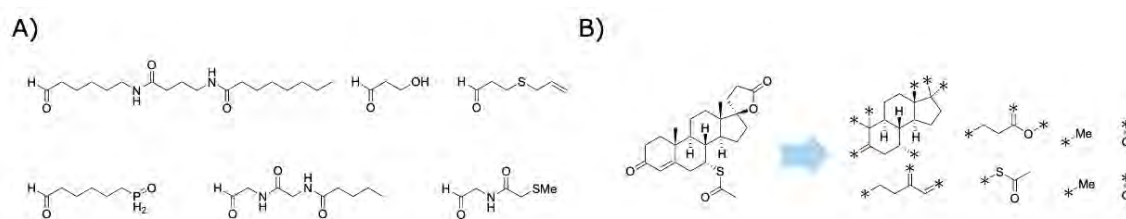


Figure A) Examples of K scaffold. B) An example of fragments by kBRICS.

007-06

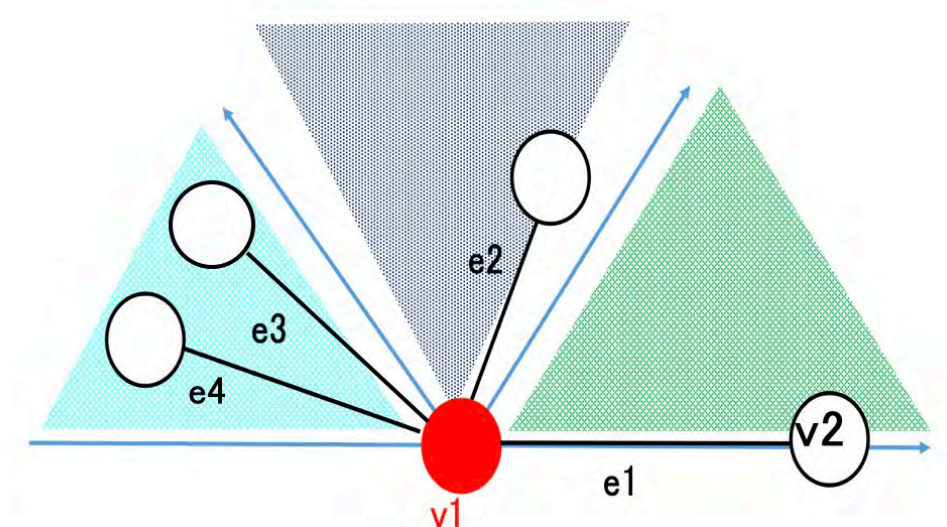
Deep Learning-Based Protein-Protein Interaction Prediction Considering Angle Information

Yukinobu MATSUNO *, Masakazu SEKIJIMA

Department of Computer Science, Tokyo Institute of Technology

(* E-mail: matsuno.y.af@m.titech.ac.jp)

Many proteins are known to control cellular functions through protein-protein interactions (PPIs). Therefore, understanding the structure of protein complexes is valuable information for structure-based drug discovery. Although experimental methods such as X-ray crystallography exist, determining the structure of protein complexes requires enormous costs and expenses. As a result, computational methods called docking have been developed to predict the structure of protein complexes. Evaluating the structures obtained by docking using only the docking score function has low accuracy, so a method called reranking is used to re-evaluate the docked structures. As a reranking method, a technique has been developed that uses graph neural networks (GNN) to learn about protein-protein interactions using graph structures. However, existing GNNs for reranking have a problem in that they do not fully capture three-dimensional structural features because they do not consider information about edge angles. In this study, we aimed to improve the prediction accuracy by using an architecture that considers angle information. As a result, we succeeded in achieving higher accuracy with the model that included angle information compared to the model without angle information.



008-01

Quantum Chemistry-Based Protein-Protein Docking without Any Empirical Parameters

Takeshi ISHIKAWA*

Graduate School of Science and Engineering, Kagoshima University
(* E-mail: ishi@cb.kagoshima-u.ac.jp)

Protein-protein interactions (PPIs) have recently garnered attention as prime targets for pharmaceuticals, making them one of the most vital research areas in the life sciences. Specifically, within the realm of protein science, significant efforts have been devoted to developing protein-protein docking methods to predict the three-dimensional structures of protein complexes. A vital aspect of protein-protein docking is the criterion for judging the appropriateness of the complex structure, which is also referred to as the scoring function, and most existing scoring functions utilize empirical parameters, such as molecular force fields.

Recently, we developed "visualization of the interfacial electrostatic complementarity (VIINEC)," a method for analyzing PPIs based on full quantum chemical calculations[1-3]. In VIINEC, the interface of proteins within complexes is first identified using the electron density (EDN), following which electrostatic complementarity between the proteins is analyzed by plotting the electrostatic potential (ESP) over the interface. Applying VIINEC to 17 complexes found that ESP patterns of the two proteins at the interface are inverted in positive and negative and match each other like a puzzle. These results suggest that VIINEC can predict complex structures by seeking structures with high electrostatic complementarity.

This study developed a novel protein-protein docking approach based on VIINEC[4]. It can be regarded as a protein-protein docking without any empirical parameters as it solely relies on the EDN and ESP obtained from full quantum chemical calculations. The scheme of our protein-protein docking is given in Figure 1. Performance evaluations were conducted using 53 complexes in a benchmark set for protein-protein docking. The results demonstrated a success rate of 75.4% for dockings employing bound state structures and 17.0% for those employing unbound state structures. These performance metrics are comparable to those of existing docking methods.

- [1] T. Ishikawa*, Chem. Phys. Lett., 761 (2020) 138103
- [2] H. Ozono, T. Ishikawa*, J. Chem. Theory Comput., 17 (2021) 5600
- [3] H. Ozono, K. Mimoto, T. Ishikawa*, J. Phys. Chem. B, 126 (2022) 8415
- [4] S. Kousaka, T. Ishikawa*, J. Chem. Theory Comput., 20 (2024) 5164

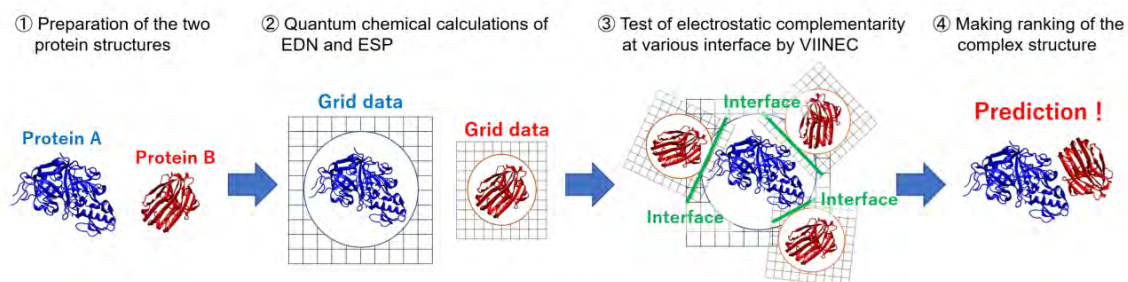


Figure 1: Scheme of our protein-protein docking method

008-02

Prediction and Analysis of Protein-Ligand Complexes Through R-value Analysis of High-Temperature Molecular Dynamics Simulation

Mochammad Arfin Fardiansyah NASUTION ^{*1, 2, 3}, **Gert-Jan BEKKER**¹,
Suyong RE³, **Chioko NAGAO**^{1, 2, 3}, **Kenji MIZUGUCHI**^{1, 2, 3}

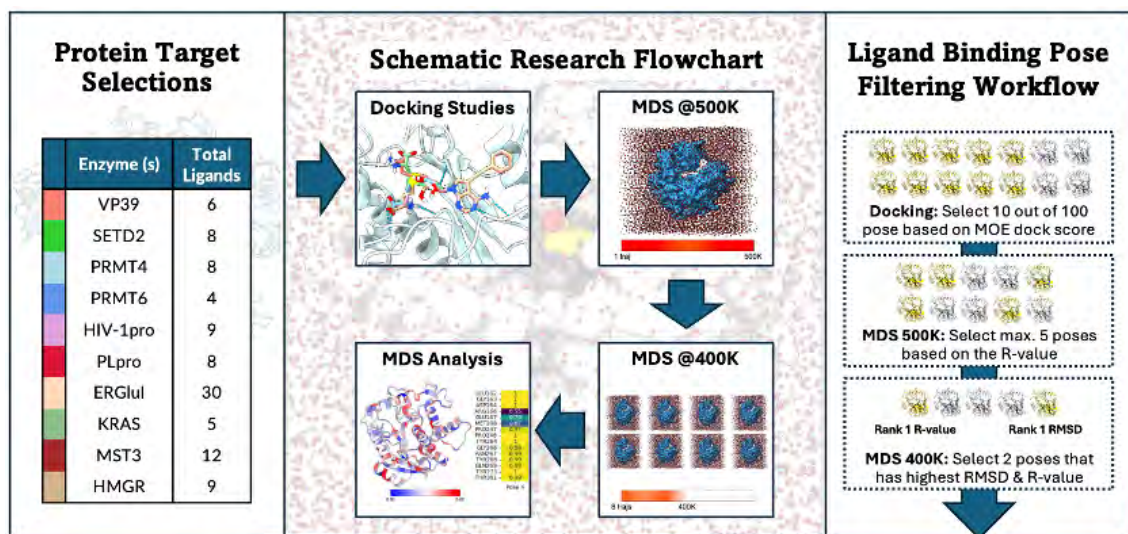
¹Laboratory for Computational Biology, Institute for Protein Research, Osaka University

²Department of Chemistry, Graduate School of Science, Osaka University

³Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition

(* E-mail: mohammad.arfin@protein.osaka-u.ac.jp)

Docking methods remain the primary approach for predicting protein-ligand interactions but often face reliability issues. While molecular dynamics (MD) simulations can refine docking poses, their high computational cost limits widespread use. To address this issue, we applied a methodology that combines classical docking and high-temperature MD simulations in 10 different enzyme systems, using MOE 2022.10 and GROMACS 2023.3 software, respectively. We utilized R-value-based and protein dynamics analyses to filter and refine the binding configurations produced from docking studies based on ligand binding stability. Our approach found a non-linear relationship between the R-value and RMSD score, where a lower RMSD (indicating similarity to the crystal structure) often corresponds to a better R-value (indicating more stable binding). Notably, 72 out of 98 (73.5%) predicted binding poses that showed the highest stability under 400K MD simulations also had high similarity (RMSD < 3.0 Å). Additionally, our methodology can be used to validate experimental results and explore new potential binding modes. In this study, when applied to the SARS-CoV-2 PLpro system, we found that although 5 out of 8 ligands exhibited high dissimilarity compared to their respective crystal structures, the predicted binding poses retained high stability under our MD simulations. However, the dissimilar regions of the ligand were not well supported by the experimental electron density. Combined, this suggests that the binding modes found by our analysis for these complexes might be the more likely ones. Here, we have shown that our protocol can be used as an effective in-silico approach for analyzing and predicting protein-ligand complexes, without resorting to highly computationally expensive methods such as dynamic docking.



008-03

Conditional structure prediction of protein-compound complex

Atsuhiko TOMITA *

Drug Discovery, Preferred Networks, inc.
(* E-mail: atomita@preferred.jp)

Deep learning based protein structure prediction, such as Alpha Fold2 [1], has made it possible to predict protein structures with high accuracy. Most of these methods use multiple sequence alignment (MSA) as input and predict the protein structure. By manipulating this input MSA, researchers have solved various problems. For example, it is known that proteins can adopt various conformational states in vivo, but conventional prediction methods had the problem of predicting conformations that were biased toward a particular state. Then, it was reported that by directly modifying the input MSA, the biased conformation can be relaxed and the conformation of other states can be stochastically obtained [2]. In addition, another application of MSA engineering was reported: a method for predicting the structure of protein-compound complexes by extending MSA features to compounds other than amino acids [3]. With the advent of such methods, structure prediction can now be utilized in the drug discovery area, including low-molecular-weight compounds. However, the prediction accuracy for protein-compound complexes is not sufficient, and improved methods are still needed.

To improve this problem, we propose a method to optimize MSA features indirectly by incorporating external knowledge into the network of protein-compound complex prediction.

Recently, in order to efficiently predict a protein structure in a specific state, methods have been investigated that indirectly modify MSA features toward the desired state using user-defined constraints [4]. These methods sample the target state more efficiently than conventional methods that directly modify the MSA to sample the structure stochastically. We applied these indirect MSA modification methods to the prediction of 3D structures of protein-compound complexes.

We incorporated external knowledge such as pharmacophores and similarity to other protein-compound complexes into the structure prediction. As a result, we optimized the MSA feature to generate a complex structure that is consistent with both the structural validity and the introduced external knowledge. The

results showed improvement in the prediction of the complex structure that were incorrect in the prediction without external knowledge. In this presentation, we report that our approach is effective in predicting the structure of protein-compound complexes, a field where few effective improvement methods have been reported so far.

- [1] Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021)
- [2] Diego, A. et al. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife* 11:e75751. (2022)
- [3] Bryant P. et al. Structure prediction of protein-ligand complexes from sequence information with Umol. *Nature Communications* 15, Article number: 4536 (2024)
- [4] Xie, T. et al. Conditioned Protein Structure Prediction. *bioRxiv* 2023.09.24.559171 (2023)

O08-04

Integrating Mathematical Modeling and Molecular Dynamics Simulations to study the effect of EGFR Mutations in Lung Cancer

Ai SHINOBU ^{*2}, Hayate TAKAGISHI¹, Noriaki OKIMOTO³, Makoto TAIJI³, Mariko OKADA¹

¹Institute for Protein Research, Osaka University

²WPI-PRIME, Osaka University

³Center for Biosystems Dynamics Research, RIKEN

(* E-mail: shinobu.ai.prime@osaka-u.ac.jp)

Understanding disease mechanisms at a molecular level is crucial for the development of targeted therapies. Molecular dynamics (MD) simulations offer atomic-level insights into these mechanisms. However, the interconnected nature of proteins necessitates a comprehensive approach through systems biology and network modeling. Here, we propose a methodology that integrates mathematical modeling, experimental data, and MD simulations to investigate the effects of mutations in the EGFR signaling system, which was identified as a key component in lung cancer pathogenesis.

Using six variants of lung cancer-derived H1299 cell lines (WT, L858R, E709G, G719S, S768I, L861Q), we conducted time-series experiments to capture molecular activity data within the EGFR pathway. Our findings indicate that the phosphorylation intensity of the Shc adaptor protein, which binds directly to EGFR did not correlate with EGFR expression levels and was reduced in mutants compared to the wild type (WT). We thus determined the relative affinity and rate coefficients for the binding of EGFR to Shc peptide with both molecular dynamics simulations and by parameter fitting to experimental data using a system of ODEs.

We observed a trend in the data that can be explained by structural and mechanistic insights from the simulations. However, to fully understand these trends, further refinement is needed in both the mathematical modeling and the methods used for affinity calculation in MD simulations.

008-05

Large-Scale MD-Based CPI Prediction Using Supercomputer Fugaku

Natsuki KANAZAWA ^{*1}, Shigeyuki MATSUMOTO¹, Shuntaro CHIBA², Yuta ISAKA², Kiyoshi TAKEMURA², Mitsugu ARAKI¹, Biao MA², Takao OTSUKA¹, Hiroaki IWATA¹, Kei TERAYAMA³, Yasushi OKUNO^{1, 2}

¹Graduate School of Medicine, Kyoto University

²RIKEN Center for Computational Science

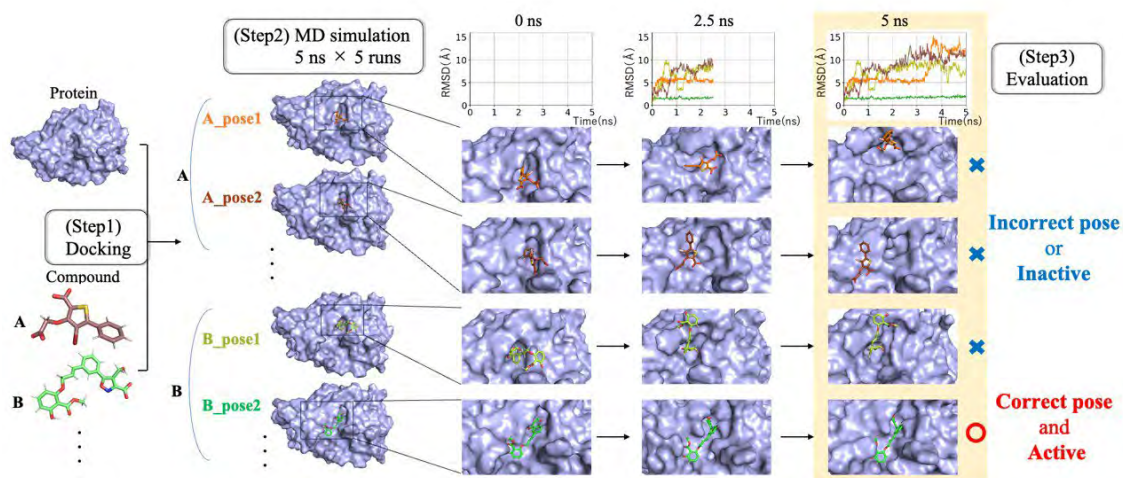
³Graduate School of Medical Life Science, Yokohama City University

(* E-mail: kanazawa.natsuki.82a@st.kyoto-u.ac.jp)

In the early stages of drug discovery, compound screening is conducted to select compounds that interact with target proteins from a vast compound library. To accelerate this process, docking simulations, a computational method based on the 3D structure of proteins, are often employed. However, they typically ignore the dynamics property essential for the protein-compound interaction, resulting in low accuracy for predicting binding poses and affinities.

In contrast, molecular dynamics (MD) simulations are a powerful computational method that accurately capture the dynamic behavior of proteins. MD simulations of compound-protein complexes can calculate binding and dissociation processes and affinities with high precision. Nevertheless, due to the high computational cost, applying MD simulations to large-scale screenings involving thousands to tens of thousands of interactions has been challenging and has not been previously attempted.

In this study, aiming to enable large-scale screenings using MD simulations, we developed a method to evaluate the binding stability of drug candidates through extensive short-time MD calculations using the supercomputer “Fugaku.” This method involves performing short-time MD simulations on tens of thousands of compound-protein complexes generated by docking simulations and comprehensively evaluating their binding stability based on dynamic behavior. Our results demonstrate that it is possible to accurately assess interaction activity with short-time MD calculations on the order of nanoseconds and that this method is applicable to large-scale screenings involving tens of thousands of compounds. Additionally, we developed an automated execution system that allows users with limited computational science expertise to run the entire workflow of this method. In the near future, we aim to further refine this method to develop a practical tool for drug discovery.



008-06

Progress of data collection in FMO database and efforts to evaluate structural qualities of biological macromolecules using quantum chemical interaction energy analysis

Chiduru WATANABE ^{*1}, Kikuko KAMISAKA¹, Chie TAKEMOTO¹, Norihiko TANI¹, Tomohiro SATO¹, Toru SENGOKU², Yoshio OKIYAMA³, Teruki HONMA¹

¹Center for Biosystems Dynamics Research, RIKEN

²Graduate School of Medicine, Yokohama City University

³Graduate School of System Informatics, Kobe University

(* E-mail: chiduru.watanabe@riken.jp)

Elucidating biomolecular interactions such as protein–ligand, protein–protein, and nucleic acid interactions is vital information for drug discovery and structural biology. Our group has been focusing on quantum mechanics (QM), which incorporates the effects of donating and withdrawing electrons and charge transfer and can appropriately deal with the CH/π and π–π interactions. Fragment molecular orbital (FMO) method [1] enables us to efficiently perform *ab initio* QM calculations for large biomolecules. The benefit of this fragmentation scheme is the availability of interfragment interaction energy (IFIE) and pair interaction energy decomposition analysis (PIEDA).

Since 2014, our group has been developing an FMO database of quantum chemical calculations of biological macromolecules [2]. In constructing the database, we continue developing and improving automated FMO calculation protocols for its data collection [3]. We have collected FMO calculation data focusing on apoproteins, kinases, nuclear receptors, antigen – antibodies, COVID-19-related proteins, and others. This year, we are collecting data considering drug discovery modalities and AI, focusing on antibodies, nucleic acids, and AlphaFold structures. In this presentation, we will introduce the status of the collection of these FMO data, IFIE/PIEDA analysis for various interactions meaningful for structural folding and molecular recognition, and quality evaluation efforts for FMO calculations and their initial structures of biological macromolecules.

Acknowledgment

The authors thank Mr. Yuya Seki of TechnoPro R&D Company for FMO calculation support. This research was done as a part of activities of the FMO Drug Design Consortium (URL: <https://fmodd.jp>). This research was partially supported by

Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from AMED under Grant Number JP24ama121030. The authors acknowledge JSPS KAKENHI Grant No. 23K18192. The results of FMO calculations were obtained using the Fugaku (project IDs: hp240114 and ra000017) and the Hokusai (project ID: RB230116).

References

- [1] Kitaura, K. *et al.*, *Chem. Phys. Lett.* **1999**, 313, 701–706.
- [2] Takaya, D. *et al.*, *J. Chem. Inf. Model.* **2021**, 61, 777–794.
- [3] C. Watanab *et al.*, *CBI J.* **2019**, 19, 5–18.

009-01

Towards the Construction of Next-Generation Molecular Robots with Quick Motion and Information Processing

Shin-ichiro M. NOMURA *

Department of Robotics, Graduate school of Engineering, Tohoku University
(* E-mail: shinichiro.nomura.b5@tohoku.ac.jp)

Molecular robotics is a pioneering field of engineering that enables the development of robots capable of efficient operation at the micrometer to nanometer scales by designing and constructing sensors, control circuits, and actuators at the molecular level. This technology has the potential to bring about new innovations in various fields, including medicine and environmental monitoring, such as drug delivery and environmental sensing. Lipid vesicles, or liposomes, have been studied as a promising body for these robots to maintain functional separation and targeted operation of molecular systems. These vesicles, at a microscale size, are expected to perform significant functions[1]. However, these vesicles, which typically carry a payload of approximately 10^{-15} L, contain only a few molecules even at pM concentrations, limiting the processing capabilities that can be achieved by individual vesicles.

This limitation has led to growing interest in multicellular molecular robots that can handle and coordinate multiple types of artificial cellular compartments. Recently, we successfully developed an artificial multicellular platform that spontaneously and stably formed structures with diameters exceeding several centimeters[2]. This platform is currently being studied as a macroscopic drug delivery system.

In conjunction with this, we have discovered a novel mechanism as a molecular sensor, where single-stranded DNA (ssDNA) with specific base sequences can be transmitted across lipid membranes into the interior of artificial cells through hybridization with cholesterol-modified ssDNA—a system we have termed “Chabashira”[3]. With further refinement, this mechanism is expected to enable molecular sensing and processing between the internal environment of the molecular robot and its external surroundings.

As the size of these platforms increases beyond the centimeter scale, the effectiveness of simple diffusion as a means of stirring diminishes. To address this, we have found a phenomenon in which a porous body, utilizing the Marangoni effect at the gas-liquid interface, can achieve high-speed motion (~ 30 mm/s) to effectively stir molecules within an aqueous solution[4]. We are currently trying to integrate these novel component technologies to construct a

“next-generation molecular robot” that operates at high speeds in response to environmental stimuli while performing molecular information processing. In this presentation, we discuss the development and prospects of this research.

- [1] Y. Sato et al., *Sci Robot* 2017, 2, DOI 10.1126/scirobotics.aal3735.
- [2] R. J. Archer et al., *Langmuir* 2023, 39, 4863–4871.
- [3] K. Yoshida et al., *ChemRxiv* 2024, DOI 10.26434/chemrxiv-2024-571kp.
- [4] R. Archer et al., *ChemRxiv* 2024, DOI 10.26434/chemrxiv-2024-3kb8h.

009-02

Experimental validation of a modified Whiplash PCR for profiling temporal and coexistence patterns of nucleic acids

Ken KOMIYA *, Chizuru NODA

Institute for Extra-cutting-edge Science and Technology Avant-garde Research (X-star), Japan Agency for Marine-Earth Science & Technology (JAMSTEC)

(* E-mail: komiyak@jamstec.go.jp)

Molecular pattern classification by molecular reaction is of great interest both for biological and medical applications and for basic molecular systems science. Whiplash PCR (WPCR) is a unique molecular reaction system, which implements state machines through repeated DNA extension by DNA polymerase. In WPCR, each state machine executes successive state transitions according to a set of transition rules encoded with a single DNA molecule. Parallel operation of a vast number of state machines, that is MIMD (multiple-instruction multiple-data) computation, can be performed in a single reaction tube [1]. Displacement-WPCR (D-WPCR), in which each transition step is driven by primer binding and extension, was then proposed to overcome the drawbacks of low reaction efficiency and high temperature condition over 80°C [2]. However, as only two rounds of state transitions have been reported so far [3], its feasibility and usefulness has not yet been explored. In this presentation, we will report the experimental optimization of the reaction conditions and propose the design for allowing the use of genetic sequences toward biological applications. We will also discuss the challenges and prospects of D-WPCR.

[1] Komiya, K.; Sakamoto, K.; Kameda, A.; Yamamoto, M.; Ohuchi, A.; Kiga, D.; Yokoyama, S.; Hagiya M. DNA polymerase programmed with a hairpin DNA incorporates a multiple-instruction architecture into molecular computing, *BioSystems*, **2006**, 83, 18–25.

[2] Rose, J.A.; Komiya, K.; Yaegashi, S.; Hagiya, M. Displacement Whiplash PCR: Optimized architecture and experimental validation, *Lecture Notes in Computer Science*, **2006**, 4287, 393–403.

[3] Komiya, K.; Yamamura, M.; Rose, J.A. Experimental Validation and Optimization of Signal Dependent Operation in Whiplash PCR, *Natural Computing*, **2010**, 9, 207–218.

009-03

Development of a sustainable database for middle molecules using AI-driven data curation

Kazuyoshi IKEDA ^{*1, 2}, Tomoki YONEZAWA², Masanori OSAWA², Tsubasa NAGAE³, Kentaro TOMII³

¹RIKEN Center for Computational Science, RIKEN

²Faculty of Pharmacy, Keio University

³Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology

(* E-mail: kazuyoshi.ikeda@riken.jp)

Medium molecules, including peptides, nucleic acids, and other medium-sized synthetic compounds, have attracted considerable attention as promising lead molecules for drug targets that are difficult to address using traditional drug discovery approaches (i.e., small molecules). These molecules have unique binding properties and are considered favorable candidates for modulating protein-protein interactions (PPIs). We aim to construct an interaction database of middle molecules (peptides, non-peptides, and nucleic acids) to accumulate information on target molecules that are difficult to discover for drug discovery. In our database, target interaction sites of middle molecules can be identified based on ligand binding site similarity data, and our AI technologies can predict interactions between targets and ligands with high accuracy.

We have undertaken a multifaceted approach to systematically collecting and curating medium-sized molecule data from open public databases. These include data on diverse classes of medium molecules, such as cyclic peptides, oligonucleotides, and peptidomimetics. In addition, we complement the interaction data by screening for PPI targets using our medium-sized compound library at Keio University.

In this study, we developed an automatic data curation method using AI that efficiently retrieves and integrates data from literature. In particular, we attempted to improve efficiency by applying large-scale language models (such as GPT) to the chemical curation of compounds. We also developed a protocol to predict interactions between compounds and targets and curate them interactively. Although these technologies currently have accuracy limitations, we will discuss their usefulness in improving efficiency compared to conventional methods.

009-04

Designing an Information Infrastructure for the Integration and Utilization of Multimodal Bioactivity Information

Shuichiro MAKIGAKI *, Mayumi KAMADA

Faculty of Future Engineering, Kitasato Institute
(* E-mail: makigaki.shuichiro@kitasato-u.ac.jp)

In the research and development of data science and artificial intelligence, data preparation and processing are the starting points and crucial factors for the accuracy of subsequent data analysis and predictive models. For drug discovery, researchers are being called upon to integrate their proprietary compound and bioactivity data with existing large-scale database information.

However, integrating databases has many challenges, especially when including natural compounds with complex structures and properties. The measurement data related to compounds are highly diverse, and their structures and properties differ on synthetic processes, making it difficult to manage them in a unified format. Many of these multimodal data are not integrated because of their multidimensionality and multi-layered property, and comprehensive utilization still needs to be fully realized. Moreover, when integrating different datasets, if unique identifiers for the data are common, they can be linked together. However, when dealing with private libraries or newly synthesized compounds, compounds that do not exist in existing databases cannot be linked.

This study proposes an approach to integrating these complex compound data. We start by using structural similarity clustering and common substructure alignment to identify and evaluate structural change processes. Then, we integrate databases by considering the relationships of fragment insertion, substitution, and deletion between compounds and their specific substructure relationships. To treat these complex relationships, we employ the Resource Description Framework (RDF). Although RDF is known as a graph data model and a component of semantic web technology, its adoption is also progressing in life sciences databases. As an application example, we demonstrate database integration by combining a subset of NPAtlas with ChEMBL using the proposed approach. We will provide specific examples of what kinds of searches can be performed and discuss the database's utility and the effectiveness of our proposed approach.

Through this study, we will formulate a new data model for integrating multimodal bioactivity data with public databases and for inter-modal collaboration. By integrating original data with existing databases, we aim to enable the use of individual activity information that was difficult to achieve in traditional chemical biology, leading to the expansion of the compound latent space.

009-05

Leveraging LLMs for Quantum Chemistry: A Comparative Study of Input File Generation for Gaussian, DFTB+, and ORCA

Gergely JUHASZ ^{*1}, Johannes Mario MEISSNER², Ilya KULYATIN²

¹School of Science, Chemistry, Tokyo Institute of Technology

²ResearchCopilot

(^{*} E-mail: juhasz@chem.titech.ac.jp)

Computational tools are now essential in chemical research and development, especially for predicting the properties of new compounds, screening processes, and studying the relationship between structure and properties. However, using these tools effectively can be challenging because it requires a deep understanding of how to choose the right methods, set up models, and adjust parameters. This knowledge often comes from extensive study of the literature, established benchmarks, and experiences shared among colleagues.

In this presentation, we explore how large language models (LLMs) can assist in planning quantum chemistry simulations and generating input files. Specifically, we tested recent versions of OpenAI's GPT, Claude, and open-source LLMs to generate input files for popular quantum chemistry software packages such as Gaussian, DFTB+, and ORCA. We compare these LLMs to see how well they help select computational methods and create accurate, grammatically correct input files. Our study examines the strengths and weaknesses of these models, helping to identify where they can be useful and where expert input is still necessary.

The potential impact of these tools goes beyond just making processes faster. LLMs could make quantum chemistry more accessible to researchers who don't have a strong background in computational methods, allowing them to conduct advanced research. Additionally, LLMs could help make it easier to reproduce calculations from published studies, addressing a common problem in the field where replicating results can be difficult due to the complexity of setting up input files. By integrating LLMs into computational chemistry workflows, we could create a more inclusive research environment where more scientists can use advanced computational techniques and ensure their work is reproducible.

009-06

Structure and stability of glycan interaction network on the HIV envelope glycoprotein

Suyong RE *

Artificial Intelligence Center for Health and Biomedical Research (ArCHER),
National Institutes of Biomedical Innovation, Health and Nutrition (NIBIOHN)
(* E-mail: suyongre@nibiohn.go.jp)

Many glycoproteins and glycolipids are present on the cell surface, playing roles not only in intercellular communication but also in various diseases, including cancer and infections. On the cell surface, numerous glycans are thought to form cluster structures through dynamic association and dissociation, potentially influencing the recognition by receptor proteins. However, there is no experimental method to directly identify these structures, and the concept remains hypothetical.

Here, we focus on the HIV envelope glycoprotein, which is one of the most extensively studied system. The spike proteins of enveloped viruses in general hold a high proportion of high-mannose-type glycans, which are thought to form clusters. These glycan clusters are considered to play a role in binding with lectins and neutralizing antibodies [1], and identifying their shape and stability could aid in the development of drugs targeting glycans. The HIV envelope glycoprotein is an extreme case of this and makes it as a good model to investigate glycan cluster structures. In this work, based on the electron microscopy structure of the trimeric BG505 SOSIP.664, which preserves the native-like HIV envelope structure (PDB ID: 5ACO) [2], a structural model was constructed with a total of 60 glycans (45 of which were high-mannose-type) attached according to the experimental data [3]. Molecular dynamics simulations were performed over microseconds using GENESIS program package [4,5]. The simulations reveal the variety of conformations and interactions of surface glycans. The results show that glycosylation unevenly shields the protein surface and has only a minor impact on protein dynamics, which is consistent with our previous work on the Lassa envelope glycoprotein [6]. Further analysis of the glycan network suggests that high-mannose-type glycans tend to form clusters in the gp120 domain. These cluster structures change when the glycan at N332, critical for binding neutralizing antibodies, is absent. These results suggest that the glycan clusters are potential antibody targets and there are critical glycans that regulate the stability of these clusters.

References:

- [1] Pritchard, L. K. et al. *Nature Commun.* 2015, 6: 7479.
- [2] Lee, J. H. et al. *Structure* 2015, 23: 1943–1951.
- [3] Behrens, A-H. et al. *Cell Rep.* 2016, 14: 2695–2706.
- [4] Jung, J.; Mori, T. et al., *WIREs Comput. Mol. Sci.* 2017, 39: 2193–2206.
- [5] Kobayashi, C.; Jung, J. et al. *J. Comput. Chem.*, 2017, 38: 2193–2206.
- [6] Re, S.; Mizuguchi, K. *J. Phys. Chem. B*, 2021, 125:2089-2097.

Abstracts

Poster Presentation

P01-01

Hepatitis C Virus Drug Resistance Mechanism: Docking and Molecular Dynamics Study of NS5A-Drug Complex

Yaxuan WANG^{*1}, Ai TOYODOME², Seiichi MAWATARI², Midori TAKEDA³, Masanori IKEDA³, Takeshi ISHIKAWA¹

¹Graduate School of Science and Engineering, Kagoshima University

²Graduate School of Medical and Dental Sciences, Kagoshima University

³Joint Research Center for Human Retrovirus Infection, Kagoshima University

(* E-mail: k4340100@kadai.jp)

Chronic hepatitis C virus (HCV) infections can lead to hepatocellular carcinoma, and the virus's high genotypic variability hinders effective vaccine development. Inhibitors against the NS5A protein, such as Daclatasvir, Ledipasvir, Velpatasvir, Elbasvir, and Pibrentasvir, are used as antiviral agents. However, they are highly resistant to Q24K/L28M/R30Q/A92K mutations, with Pibrentasvir showing resistance to R30E but not other mutations[1,2]. In this study, we modeled the HCV NS5A protein structure to understand the resistance mechanism of these antiviral agents.

Our study focuses on the N-terminal drug resistance mutation site of the NS5A protein. Since no crystal structure, including this mutation site, has been reported, homology modeling was performed to create a three-dimensional structure model, and then docking simulations were performed using AutoDock Vina[3] to create complex models with the inhibitors. However, our models failed to adequately explain the drug resistance mechanism. Next, to mimic the human body's environment more closely, we attempted molecular dynamics (MD) simulations for NS5A protein with biological membranes. The genmixmem program[4] was used for modeling, and GROMACS[5] was used for MD simulations. A 100 ns trajectory of the membrane-protein system was obtained (Figure 1) and the inhibitors attached to the 60 ns structure, by which membrane-protein-inhibitor systems were created. Now, we are proceeding with the MD simulations for these systems.

[1] R. Bartenschlager, et al., The molecular and structural basis of advanced antiviral therapy for hepatitis C virus infection, *Nature Reviews Microbiology*, 11 (2013) 482.

[2] A. Toyodome, et al., Analysis of the susceptibility of refractory hepatitis C virus resistant to nonstructural 5A inhibitors, *Sci. Rep.*, 14 (2024) 16363.

- [3] O. Trott and A. J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, 31 (2010) 455.
- [4] T. Lu, J. Wang, and L. Xu, genmixmem program (<http://sobereva.com/245>).
- [5] M. J. Abraham, et al., GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *Software X*, 1-2 (2015) 19.

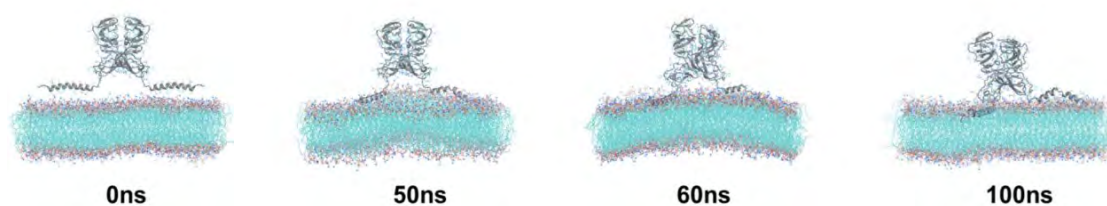


Figure 1: Trajectory of the membrane-protein system

P01-02

Generation of Structural Ensemble of Linear Diubiquitin Based on PCS Experiments

Yoshiki YUGAMI^{*1}, **Xue-Ni HOU**², **Sotaro FUCHIGAMI**³, **Hidehito TOCHIO**², **Kei MORITSUGU**¹

¹Department of Science, Osaka Prefecture University

²Graduate School of Science, Kyoto University

³Graduate School of Pharmaceutical Sciences, Shizuoka University

(* E-mail: sfc05146@st.osakafu-u.ac.jp)

Ubiquitin is a protein consisting of 76 amino acids, and the polyubiquitin modification of proteins is involved in various functions including protein degradation, DNA repair, and signal transduction. Polyubiquitin chains can form diverse structures through the polymerization of ubiquitin via seven Lys residues (K6, K11, K27, K29, K33, K48, K63) and the N-terminal residue (M1; linear), and can function by binding to linkage-specific substrates. In our previous work, pseudo contact shifts (PCS) using lanthanoids were applied to linear diubiquitin and the PCS data was found to be reproduced by using a series of representative structures. This study aims to establish a method for deriving the structure ensemble and the associated free energy landscape of linear diubiquitin that is in good agreement with the PCS data by (1) all-atom molecular dynamics (MD) simulations and more simply, (2) coarse-grained MD with machine learning technique.

In (1), the initial model of linear diubiquitin was taken from PDB: 3b0a, and fully solvated in a rectangular box. Using eight replicas with varying interactions between the two ubiquitin, structural sampling simulation was performed for 500 ns using gREST module of GENESIS. The PCS data were calculated for each of the resulting structural ensemble. In this study, a weight was assigned to each structure and optimized using the steepest descent method to match the PCS experimental data with Tm tags on both D39 and G47 distal ubiquitin, thus extracting the structure ensemble consistent with the experiment. In (2), coarse-grained MD was carried out using cafemol to generate an extensive structural ensemble by connecting extended (PDB: 2w9d) and compact (PDB: 3axc) structures. The obtained Ca-atom structures were converted to all-atom models using the software cg2all, and PCS data were calculated. Similar to (1), the weights for all the structure were calculated and the 2D free energy landscape was derived along the distance and the torsion angle of the two ubiquitin.

While the PCS data from a single crystal structure did not match with the experiment, the agreement from the PCS data using the 400,000 gREST structures was also insufficient. The optimization of the weights for the MD structures was found to drastically improve the agreement. The free energy landscape was also changed considerably, showing that combining PCS experiment can avoid the convergence problem of the structure sampling and remove the artifacts from the MD force field. It was also demonstrated that combining coarse-grained MD with reconstructing the all-atom model can also generate the accurate structure ensemble to match the PCS experiment with much less computational cost.

P01-03

Investigation of the utility of steered MD in the prediction of binding affinity: a case study of HSP90

Chisato KANAI ^{*1}, Enzo KAWASAKI¹, Atsushi YOSHIMORI²

¹INTAGE Healthcare Inc.

²Institute for Theoretical Medicine, Inc.

(* E-mail: kanai@intage.com)

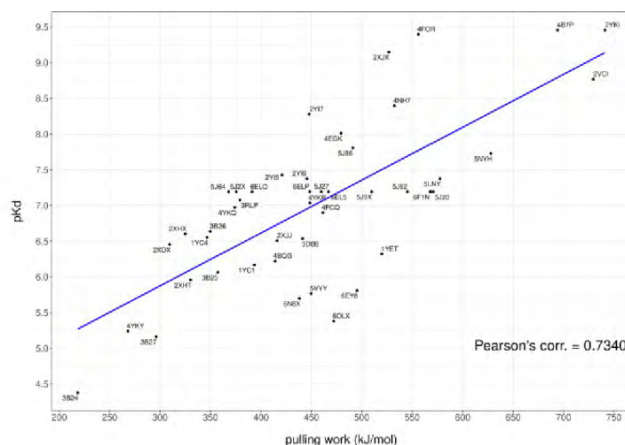
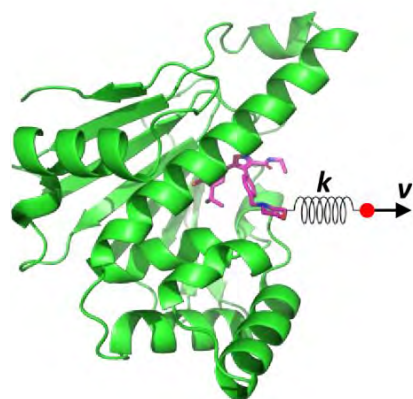
In recent years, various models for de novo design using AI-generated compound structures have been proposed and widely used in drug discovery research. We offer Deep Quartet, a service that designs compound structures using a pharmacophore model based on the three-dimensional structure of proteins as the scoring function. Deep Quartet has been applied to various proteins where a large number of compound structures has been generated [1]. Even after applying ADME filters such as the Rule of 5 and synthetic accessibility scoring, thousands of compounds often remain as candidates. This situation is believed to be common to other de novo design systems using generative AI.

In actual drug discovery research, the number of in silico-designed compounds that can be synthesized is, typically, only a few, at most a few dozen. With this in mind, we investigated whether Steered MD could be used as a final selection method from the thousands of compounds designed by generative AI. Steered MD simulates the process of pulling a ligand out of its bound state with a protein using MD simulations, calculating the work energy required to dissociate the ligand from the protein. Although the results are from non-equilibrium MD simulations, previous studies [2, 3] have reported some correlation between ligand pulling work and binding free energy. An advantage of Steered MD is its significantly lower computational cost compared to binding free energy methods (FEP, TI, Umbrella Sampling, MMPBSA). Even if the prediction accuracy of the binding strength is slightly reduced, the low computational cost is still attractive, making it feasible to perform calculations for over a thousand compounds. In this poster presentation, we report the results of our evaluation of the binding strength of compounds based on Steered MD, using the protein HSP90 as an example.

[1] YOSHIMORI, Atsushi, et al. Design and synthesis of DDR1 inhibitors with a desired pharmacophore using deep generative models. ChemMedChem, 2021, 16.6: 955-958.

[2] VUONG, Quan Van; NGUYEN, Tin Trung; LI, Mai Suan. A new method for navigating optimal direction for pulling ligand from binding pocket: application to ranking binding affinity by steered molecular dynamics. *Journal of chemical information and modeling*, 2015, 55.12: 2731-2738.

[3] HO, Kiet; TRUONG, Duc Toan; LI, Mai Suan. How good is jarzynski's equality for computer-aided drug design?. *The Journal of Physical Chemistry B*, 2020, 124.26: 5338-5349.



P01-04

Cross-reactivity of T cell receptors against HCoV through three-dimensional structure prediction

Ao KIKUCHI ^{*1}, Toru EKIMOTO¹, Tsutomu YAMANE², Shuntaro CHIBA², Kanako SHIMIZU³, Shin-ichiro FUJII³, Mitsunori IKEGUCHI^{1, 2}

¹Graduate School of Medical Life Science, Yokohama City University

²Center for Computational Science, RIKEN

³Center for Integrative Medical Science, RIKEN

(* E-mail: w235417a@yokohama-cu.ac.jp)

In cellular immunity, killer T cells play a major role in eliminating virus-infected cells and preventing the spread and severity of infection. One of the crucial steps in triggering the immune response is the recognition of the viral antigen peptide-human leukocyte antigen (HLA) complex (pHLA) presented on the surface of infected cells by the T cell receptor (TCR) of the killer T cell via the formation of a complex of pHLA and TCR. Most TCRs have antigen specificity, but there are also cross-reactive TCRs that recognize multiple types of pHLA. Recently, several antigen peptides derived from human coronavirus (HCoV) with high affinity to HLA-A*24:02, the most common HLA type in Japanese, were identified, and it was revealed that there are killer T cells that cross-react with these peptides (Shimizu, K. et al. Commun Biol. 2021). This suggests that the killer T cells that worked with seasonal coronaviruses may also work when infected with novel coronaviruses, and therefore prediction of TCR cross-reactivity is expected to lead to elucidation of cross-reaction mechanisms and development of therapeutic and preventive measures to enhance immune responses. However, there is no method to predict TCR cross-reactivity with high accuracy. In this study, we aimed to develop a method to classify TCRs that react to HCoV-derived peptide-HLA-A*24:02 complex (pHLA) based on their three-dimensional structures. From the pHLA and TCR sequences, we predicted a structure of the pHLA-TCR complex using TCRmodel2, an AlphaFold2-based structure prediction tool (Yin, R. et al. Nucleic Acids Res. 2023) and constructed a method for classifying the reactivity of the TCR using the confidence score of the binding interface estimated from the predicted structure (Jumper J et al. Nature. 2021) to classify TCR reactivity. As a result, we succeeded in classifying about 90% of 28 types of TCR-pHLA reactivity including cross-reactivity.

P01-05

Prediction Method for Protein-Bound Conformation of Macrocycles

Shoya HAMAUE *, Isao YASUMATSU, Ayako MORITOMO, Mizuki TAKAHASHI, Hiroyuki HANZAWA

Modality Research Laboratories I, Research Function, R&D Division, Daiichi Sankyo Co., Ltd.

(* E-mail: syoya.hamaue@daiichisankyo.com)

In recent years, so-called “Beyond Rule of 5 (bRo5)” modalities have been recognized in the pharmaceutical industry to possess new therapeutic potential. Examples of synthetic bRo5 modalities are peptides, nucleic acids, macrocycles, and heterobifunctional molecules such as Targeted Protein Degraders (TPDs). These modalities offer advantages in addressing medical needs that were difficult to achieve with small molecules, such as targeting Protein-Protein Interactions (PPIs) and achieving high target specificity. However, due to their large molecular weight and complexity, it is generally considered challenging to accurately predict their 3D structures, biochemical activities, and physicochemical properties using computational chemistry methods. Therefore, it is necessary to devise new computational methods to overcome these challenges and improve prediction accuracy.

In this study, we selected macrocycles as an example of bRo5 modalities. Macrocycles can adopt a vast number of conformations in solution, and thus computational methods for accurate prediction of the protein-bound conformations of macrocycles have not yet been fully established. Then, we aimed to establish a versatile computational approach for predicting their protein-bound conformation. To accomplish this objective, we evaluated the computational methods that can reproduce the bound conformation of crystal structures in a variety of cases. We curated 51 complex crystal structures from the Protein Data Bank (PDB), considering the diversity in the ring size of the macrocycles, the type of targets, and the pocket shapes. To thoroughly explore the extensive range of possible conformations of the macrocycles, we examined the conformational search conditions for it alone using three different modeling tools. As a result, we found it effective in predicting protein-bound conformations to explore expansion of the conformational space by using multiple modeling tools. Additionally, we established potential energy criteria to exclude unstable conformations from the explored conformational ensemble. Finally, we performed docking studies on the conformations (11~4,223) that are filtered from generated conformations (33~7,158) to verify whether they

could bind to the target proteins with the predicted conformations.

As an outcome of this study, we successfully identified a set of conditions that accurately reproduced the protein-bound conformations for 46 out of 51 complexes. This methodology is expected to enable the prediction of protein-bound conformations of macrocycles without any previous knowledge of their bound form, regardless of ring size, target categories, and pocket shapes. In the future, we would like to investigate whether this methodology can be applied to other bRo5 modalities.

P01-06

Enhanced Prediction of Antigen-Antibody Complex Structures through Aggressive Structural Refinement by AlphaFold2

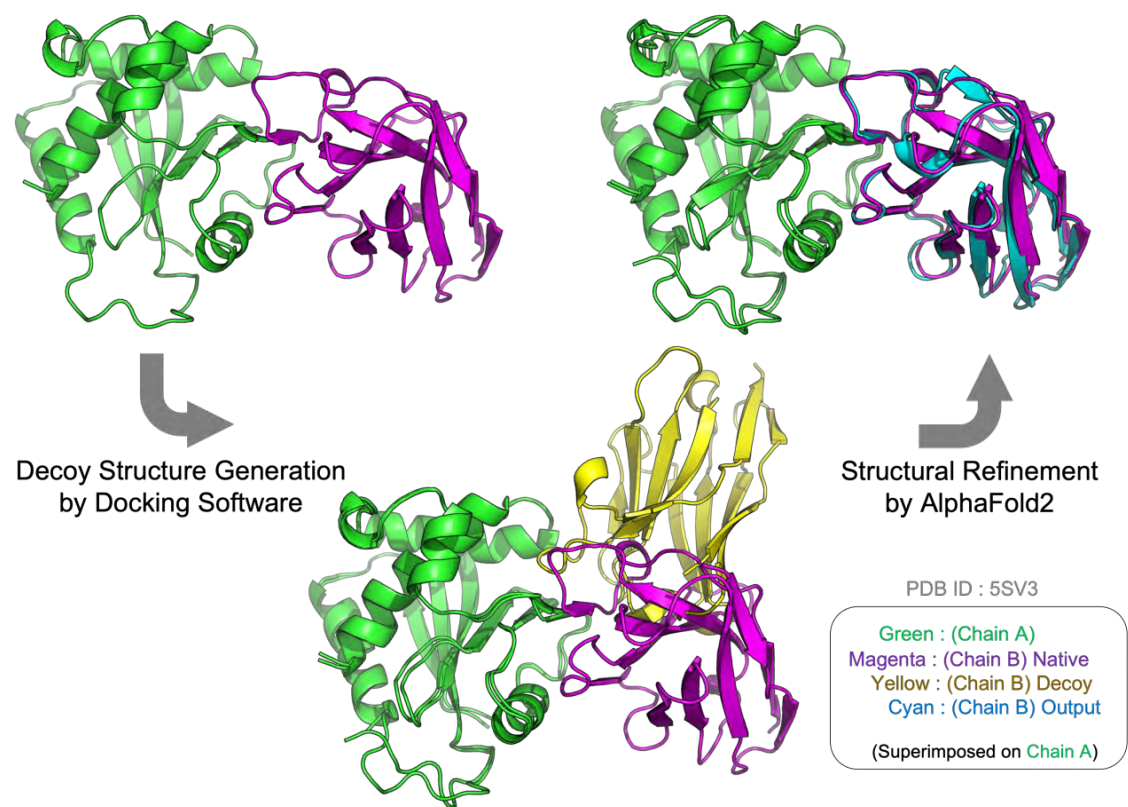
Seiya TANAKA ^{*1}, Masaki KOYAMA¹, Hiroki ONODA², Leonard CHAVAS^{1,2}, George CHIKENJI¹

¹Department of Applied Physics, Graduate School of Engineering, Nagoya University

²Synchrotron Radiation Research Center, Nagoya University

(* E-mail: tanaka.seiya.a1@s.mail.nagoya-u.ac.jp)

Accurate prediction of protein structures is crucial for elucidating their biological functions. AlphaFold2, a cutting-edge machine learning neural network algorithm, has significantly enhanced the accuracy of monomer structure predictions by leveraging co-evolutionary information. However, predicting protein complex structures, particularly Antigen-Antibody complexes, remains a formidable challenge largely due to the absence of co-evolutionary signals at the interaction interfaces. Recent progress has integrated AlphaFold2 with physics-based docking programs to address this gap, leading to success rates that surpass those observed with AlphaFold2 alone in complex structure prediction. Despite these advancements, success rates have considerable room for improvement. The current study introduces a methodology enhancing the precision of complex structure predictions. Our approach employs a physics-based protein docking program to generate multiple decoys for input into AlphaFold2 for quality refinement and confidence evaluation. A novel aspect of our method is the modification of AlphaFold2 to allow aggressive structural refinement. We demonstrate that our approach enables significant conformational adjustments in certain complexes, achieving closer proximity to native structures. This structural refinement performed by AlphaFold2 improves the accuracy of predicting complex structures, particularly for Antigen-Antibody complexes, making a significant methodology advancement.



P01-07

Deep-Learning model for Predicting the Replacement of Water Molecule upon Ligand Binding

Yuki ITO¹, Masateru OHTA², Mitsunori Ikeguchi^{2,3}, Takashi YOSHIDOME *¹

¹Department of Applied Physics, Graduate School of Engineering, Tohoku University

²AI-Driven Drug Discovery Collaborative Unit, HPC- and AI-Driven Drug Development Platform Division, Center for Computational Science, RIKEN

³Graduate School of Medical Life Science, Yokohama City University

(* E-mail: takashi.yoshidome.b1@tohoku.ac.jp)

In drug discovery, the performance of docking software is often limited due to the exclusion of water molecules located at the interface between proteins and ligands from the input data. A proposed solution involves incorporating only those water molecules that remain bound during the protein-ligand binding. Although molecular dynamics (MD) simulations can in principle be possible to incorporate the water molecules, they are notoriously time-consuming. Thus, a fast and accurate method is required for predicting the water molecules that should be incorporated into the drug discovery.

To address this challenge, here the following protocol for predicting the water molecules that should remain in the binding pocket is proposed. First, the hydration structure around a protein is computed using our deep-learning model “gr Predictor” [1] enabling the prediction of the hydration structure in approximately a minute while MD requires a few hours to obtain the hydration structure. Next water molecules are placed using the obtained hydration structure and the program suite “placevent” [2]. Finally, a convolutional neural network (CNN) is implemented to predict each water molecule as either “displaceable” or “non-displaceable”. Upon testing the model on unknown data, it achieved an accuracy of 0.6971 and a recall of 0.584. We will also show a prediction result of replaced and non-replaced water molecules in a holo structure using the apo structure.

[1] K. Kawama, Y. Fukushima, M. Ikeguchi, M. Ohta, and T. Yoshidome, J. Chem. Inf. Model., 62, 4460 (2022).

[2] D.J. Sindhikara, N. Yoshida, F. Hirata, J. Comput. Chem., 33, 1536 (2012).

P01-08

Comprehensive docking simulations using AlphaFold2-based human olfactory receptors for odor prediction

Hirotsada KANESHIRO ^{*1}, Masakazu SATO¹, Airi TANAKA¹, Shuya NAKATA¹, Yoshiko AIHARA², Hirotaka Nishioka KITO³, Yoshiharu MORI⁴, Shigenori TANAKA¹

¹Laboratory of Computational Science, Graduate School of System Informatics, Kobe Univ., Kobe, Japan

²Grad. Sch. of Agri. Sci., Kobe Univ., Kobe, Japan

³Fac. of Sci. and Eng., Kindai Univ., Osaka, Japan

⁴KQCC, Keio. Univ., Yokohama, Japan

(* E-mail: 238x017x@stu.kobe-u.ac.jp)

When we smell, we can distinguish many odors, but it is difficult to predict an odor from its molecular structure information.

The fact that there is no clear classification of smells, such as the five tastes in the sense of taste, also makes it difficult to predict smells.

Therefore, the purpose of this study is not to make absolute odor prediction, but to predict which odor molecules outside the database are closest to which molecules in the database by using odor molecules and odor descriptors registered in the database ATLAS.

One simple approach to this comparison of odorant molecules is to compare the structural similarity of odorant molecules, but in this study, odor similarity is examined based on docking scores with human olfactory receptors.

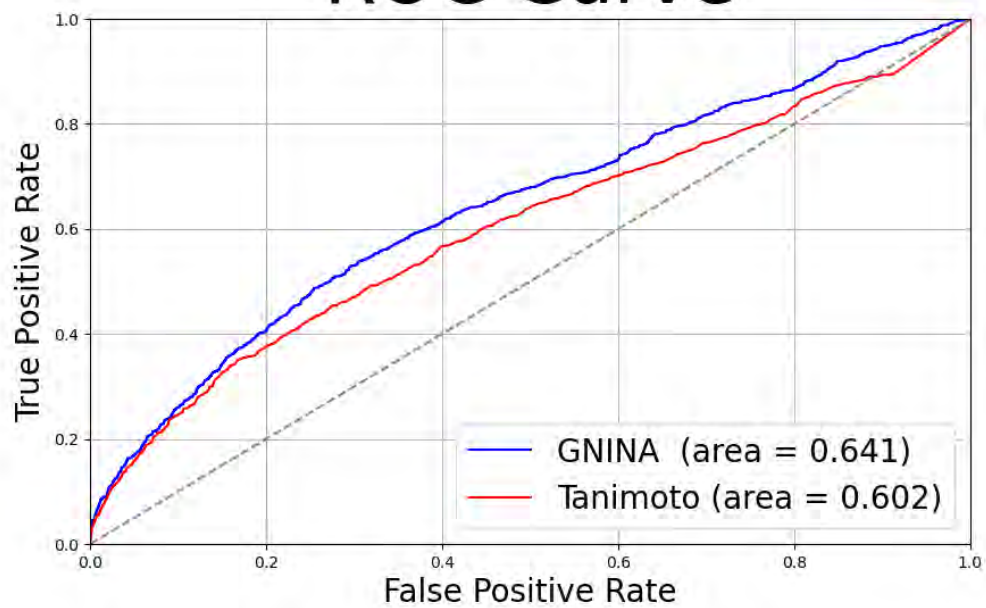
Specifically, using OR52c (PDBID:8HTI), which has recently been structurally determined by cryo-EM, as a template, we generated approximately 400 human olfactory receptor structures using AlphaFold2 and performed comprehensive docking using GNINA with approximately 130 odor molecules from ATLAS.

The scores obtained at this time were compared with a score (Tanimoto coefficient) that expresses the structural similarity of the molecules.

the results were found to be superior to those obtained using an index that evaluates the similarity of molecular structure.

The results are shown in the following figure.

ROC Curve



P01-09

Generative Model for Protein Structural Ensembles Enhanced by Molecular Dynamics Simulation Data

Shinji IIDA *¹, Yutaka SAITO^{1, 2, 3}

¹School of Frontier Engineering, Kitasato University

²Graduate School of Frontier Sciences, The University of Tokyo

³Artificial Intelligence Research Center, AIST

(* E-mail: iida.shinji@kitasato-u.ac.jp)

The function of proteins is closely related to their three-dimensional structure. Proteins recognise other molecules through their three-dimensional structure, performing functions such as catalysing molecular reactions, transport, and transmitting biological signals. The three-dimensional structure of proteins provides useful information for designing molecules that regulate protein function.

Even though static protein structure prediction has become accurate, protein structures fluctuate, and predicting three-dimensional structures whilst considering these fluctuations remains challenging. Molecular dynamics (MD) simulations are known to be an effective means of studying structural ensembles. However, when applying MD simulations to targets that form diverse structures, they can become trapped in stable states and requires a huge amount of computational time, which makes MD simulations ineffective to obtain a structural ensemble.

To alleviate the cost of structural ensemble generation, we build a generative model that produces realistic, protein structural ensembles without extensive MD simulations: i. We created training data for structural ensembles by independently performing all-atom MD simulations. ii. We then performed continual pre-training for Pepflow, a diffusion model developed by another group [Abdin, O.; Kim, P. M. Nat. Mach. Intell. 2024, 1–12.], to expand its applicability domain. While Pepflow primarily used partial structures from PDB as training data, it also incorporated a small amount of MD data. In this study, we expanded the training dataset by conducting MD simulations for 8,000 peptides.

We evaluated the Pepflow models with or without our continual pre-training through the reproducibility of probability distribution with respect to structural quantities of protein structure, such as dihedral angles, bond distance, principal

components. For example, Figure 1 indicates a Ramachandran plot of an alanine in a three amino-acid peptide. It demonstrates that the distribution of dihedral angles (middle in Figure 1) was in agreement with that obtained by a MD simulation (left in Figure 1), whereas the distribution of the original Pepflow (right in Figure 1) failed to generate metastable states.

The MD-data enhanced Pepflow would have the potential to improve peptide docking and design and to provide various initial structures of peptides that may enhance the structural sampling coverage of MD simulations.

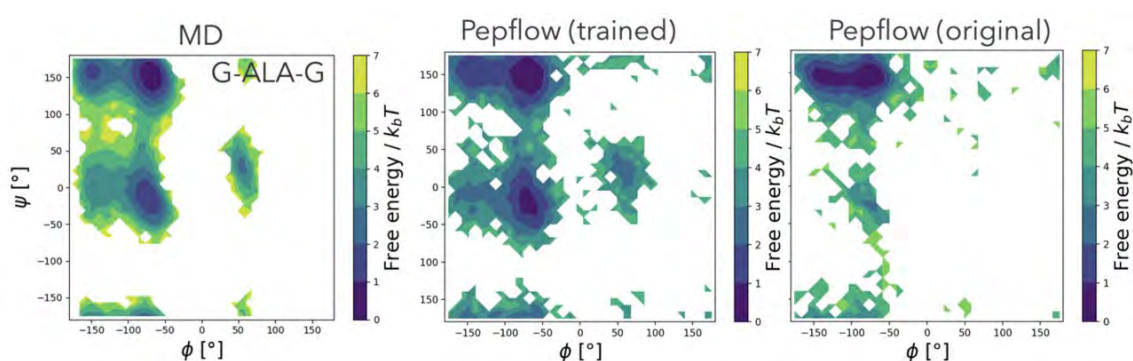


Figure 1. Ramachandran plot of an alanine in a peptide. Left: All-atom MD. Middle: Trained Pepflow without continual pre-training. Right: Original Pepflow

P01-10

Epicatechin n -mers ($n \geq 5$) adopt more compact conformations than catechin n -mers

Toshiaki UEDA ^{*1}, KAWAHARA TAKESHI^{1, 2}, MAKABE HIDEFUMI^{1, 2}, UMEZAWA KOJI^{1, 2}

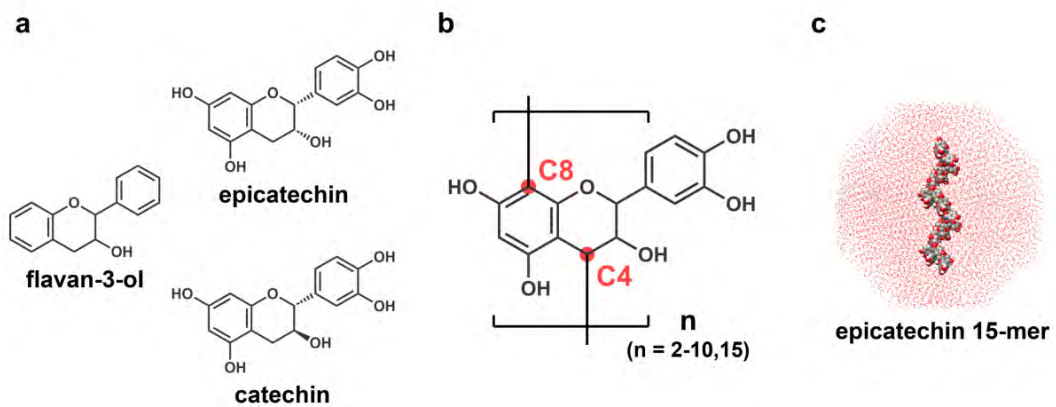
¹Graduate School of Science and Technology, Shinshu University

²Institute for Biomedical Sciences, Shinshu University

(* E-mail: 24bs104a@shinshu-u.ac.jp)

Epicatechin and catechin belong to flavan-3-ol of flavonoids. Epicatechin is a cis-trans isomer of catechin (Fig. a). These oligomers (n -mers, Fig. b) are connected by the inter-flavan (C4-C8) bonds and are found in natural products. The long epicatechin n -mers ($n \geq 5$) have an anti-invasive activity against cancer cells. In contrast, the long catechin n -mers ($n \geq 5$) do not. It suggested that the conformations of the long epicatechin and catechin n -mers are different. The purpose of our study was to characterize conformational properties of epicatechin and catechin n -mers. The conformational ensembles of epicatechin and catechin n -mers ($n = 2-10$, and 15) were calculated with an all-atom model in explicit solvents (Fig. c) by an enhanced-sampling method, multicanonical molecular dynamics simulation. The 300K conformational ensembles were analyzed.

The end-to-end distances were calculated to understand the molecular lengths of epicatechin and catechin n -mers. The results showed that the lengths of the long epicatechin n -mers were shorter than those of the long catechin n -mers. The residue-residue contacts and solvent-accessible-surface area per residue were analyzed to find intramolecular interaction. In the long n -mers ($n \geq 5$), epicatechin residues made the contacts with higher probability than catechin, and epicatechin residues were buried inside the molecule. These results showed that intramolecular interactions of epicatechin residues may lead to the compact structure in the long epicatechin n -mers compared to catechin.



P01-11

The Computational Study on the Secondary Structure Formation of Nascent Peptides Inside the Ribosome Tunnel with Biomolecular Environments Mimicking Model

Takunori YASUDA *¹, **Rikuri MORITA**², **Yasuteru SHIGETA**², **Ryuhei HARADA**²

¹Doctral Program in Biology, Institute of Life and Environmental Sciences,
University of Tsukuba

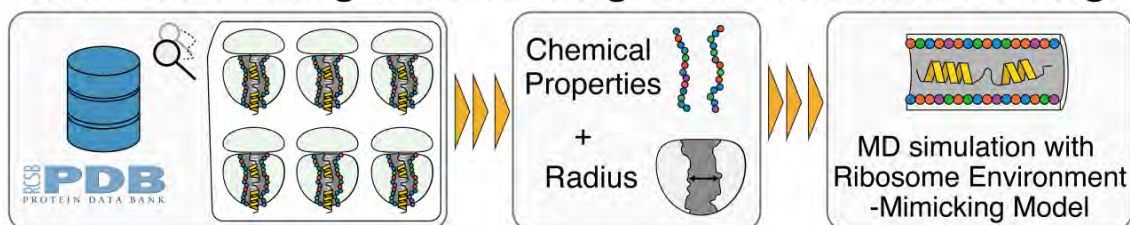
²Center for Computational Sciences , University of Tsukuba
(* E-mail: takunoriyasuda@gmail.com)

The conformation of proteins is determined not only by their sequence but also by surrounding environment. Therefore, investigating the conformations in the cellular environment is crucial for understanding protein function in cells. To investigate protein configurations within diverse biomolecular environments, several approaches that combine molecular dynamics (MD) simulations with simplified models have been proposed. Specifically, carbon nanotubes and hydrophobic cages have been well established as models of ribosome-tunnels or protein chaperons, respectively. However, recent findings suggest that these uniform hydrophobic models may not adequately capture the effects within each biomolecular environment. Based on these facts, it is necessary to generate spherical and cylindrical models based on a variety of chemical properties corresponding to the components within target biomolecules. Therefore, we developed a new open-source tool called Biomolecular Environment-Mimicking Model Generator (BEMM-GEN).

Furthermore, as an application of BEMM-GEN, we focused on the ribosomal tunnel. Inside the ribosome tunnel, a nascent peptide forms its specific α -helix. Despite the deeply relationship between such temporally α -helix formation and biological functions, a comprehensive analysis of the ribosome tunnel environment's impact on the α -helix formation has not been conducted yet. Therefore, we generated a computational model called Ribosome Environment-Mimicking Model (REMM) by considering the radius and components of experimentally determined ribosome structures. Using an enhanced all-atom molecular dynamics simulation, we investigated the properties of the nascent peptides in the experimental structures inside the Carbon nanotube (CNT), in addition to the REMM. Herein, we adopt the CNT as a reference model and compared the ability of both models to replicate the α -helix of the nascent peptides. Finally, we elucidated the mechanism of ribosome tunnel-specific α -

helix formation by analyzing the nascent peptides inside each model.

Reveal the Effect of Ribosome Tunnel-Environments on Protein Configuration During Co-Translational Folding



P01-12

Kinetic Analysis of Membrane Permeation Process of Cyclic Peptides Using Markov State Models with Molecular Dynamics Simulations

Kei TERAURA ^{*1}, **Masatake SUGITA** ^{1, 2}, **Keisuke YANAGISAWA** ^{1, 2}, **Yutaka AKIYAMA** ^{1, 2}

¹Department of Computer Science, School of Computing, Institute of Science Tokyo

²Middle-Molecule IT-based Drug Discovery Laboratory (MIDL), Institute of Science Tokyo

(* E-mail: terakura@bi.c.titech.ac.jp)

Cyclic peptides are gaining attention as novel pharmaceuticals because they can target intracellular protein-protein interactions, which have been difficult to address with small molecules or antibody drugs. However, the poor membrane permeability of cyclic peptides is a significant bottleneck in drug development. To design membrane-traversing cyclic peptides, understanding the membrane permeation mechanisms of cyclic peptides, especially permeation by passive diffusion, is crucial for their design. Understanding the mechanisms is challenging due to the diversity of cyclic peptide conformations and the long timescale of membrane permeation. To address these issues, research has been conducted using molecular dynamics simulations in addition to experimental studies. These simulation-based studies have yielded a certain degree of predictive accuracy and insights. Nevertheless, the discussions have only been based on one-dimensional or two-dimensional reaction coordinates, such as the overall peptide position and angle relative to the membrane, and local changes in three-dimensional structures during the membrane permeation process have not been considered.

In this study, we analyzed the membrane permeation process of cyclic peptides using a Markov State Model (MSM) based on multidimensional reaction coordinates, including position, orientation, and detailed conformation of the peptide. MSM is a powerful tool that quantifies long-timescale behavior using multiple short simulations and suggests key features of functional dynamics. We applied MSM to understanding the membrane permeation mechanism of cyclic peptides. We conducted approximately 1,000 unbiased simulations of 100 ns each across more than 15 peptides. The initial conformations were selected from previous research data [1]. The eigenvectors obtained from the MSM indicated several motions that represent potential bottlenecks in the membrane

permeation process. However, some challenges still remain, such as converging timescales.

[1] Sugita, M.; Fujie, T.; Yanagisawa, K.; Ohue, M.; Akiyama, Y. Lipid Composition Is Critical for Accurate Membrane Permeability Prediction of Cyclic Peptides by Molecular Dynamics Simulations. *J. Chem. Inf. Model.* 2022, 62, 4549–4560, DOI: 10.1021/acs.jcim.2c00931

P01-13

Predicting Lysine Reactivity: Insights from Constant-pH MD Simulations and Experimental Correlation

Osamu ICHIHARA *

Schrödinger KK

(* E-mail: osamu.ichihara@schrodinger.com)

The selective functionalization of lysine residues in proteins is a key strategy in the development of bioconjugates. However, the reactivity of lysine residues varies significantly depending on their local environment, influencing the success of such modifications. In this study, we focus on RNase A as a model system, utilizing Schrödinger 's constant-pH molecular dynamics (MD) simulation tool implemented in the Desmond program to predict the pKa values of lysine residues and correlate these with experimentally determined reactivity data from Xi Chen et al. (Bioconjugation Chemistry 2012, 23, 500-508).

Our results reveal a strong correlation between the predicted pKa values and the reactivity of lysines in RNase A, with some discussion extending to Lysozyme C and Somatostatin. This correlation underscores the potential of constant-pH MD simulations as a predictive tool for identifying reactive lysine sites, which could streamline the design of site-selective functionalization strategies. While this study primarily focuses on RNase A, the insights gained suggest potential applications in Antibody-Drug Conjugates (ADCs), where precise conjugation is critical. Although still an emerging possibility, this approach could contribute to more efficient and targeted ADC development. Our findings highlight the value of integrating computational predictions with experimental data to advance protein engineering, bioconjugation strategies, and potentially, ADC design.

P01-14

Protein Tertiary Structure Prediction with Fine-tuned AlphaFold2 for Ligand Virtual Screening

Yuki YASUMITSU *, Masahito OHUE

School of Computing, Institute of Science Tokyo

(* E-mail: yasumitsu.y.aa@m.titech.ac.jp)

Virtual Screening (VS) is a method for selecting drug candidate compounds from a large number of compounds using a computer.

Structure-Based Virtual Screening (SBVS) is a method to perform VS based on the 3D structure of proteins.

Compared to ligand-based methods, SBVS does not use known experimental information on the target protein, and thus can discover highly novel drug candidate compounds.

In general, it is known that the drug-bound holo structure is more accurate than the drug-unbound apo structure for SBVS.

The 3D structure of a target protein is necessary for SBVS, but when the 3D structure is unknown, it is necessary to predict the 3D structure. Homology modeling has been used to predict the 3D structure using homologous proteins with known structures, but the use of predicted structures based on machine learning models is now being considered.

In a previous study applying AlphaFold2, a protein conformation prediction model, to SBVS, the screening performance of the predicted structures was found to be inferior to that of the holo structure and comparable to that of the apo structure.

In this study, we propose a method for fine tuned AlphaFold2 using a dataset of holo structures to build a model that predicts a structure suitable for SBVS.

We also attempted to improve screening performance by using the holo structure in the template structure used by AlphaFold2.

Screening performance was then evaluated by performing docking simulations on the DUD-E data set.

P01-15

Dynamic Relationship Between the Entrance to the Ligand Binding Site and the Dimer Interface in MAO-B

Yoshitaka TADOKORO *¹, **Shota SHIMOGOCHI**¹, **Ryota KIYOOKA**¹, **Masaki OTAWA**³, **Naoyuki MIYASHITA**^{1, 2}

¹Graduate School of Biology-Oriented Science and Technology, KINDAI University

²Faculty of Biology-Oriented Science and Technology, KINDAI University

³School of Physical Sciences, SOKENDAI

(* E-mail: tadoyoshi@miyashita-lab.net)

Monoamine oxidase B (MAO-B) is located on the outer membrane of mitochondria and is known to catalyze the oxidation of dopamine [1-2]. It is associated with neurological diseases such as Parkinson's disease and other neurological disorders. Recently, the development of Positron Emission Tomography (PET) tracers for disease diagnosis has garnered attention [3-4]. PET tracers targeting MAO-B have also been developed [4]. MAO-B typically adopts a dimeric conformation in the mitochondrial membrane. The ligand entrance gates of the MAO-B dimer are small and appear "semi-closed" in the structure determined by X-ray crystallography. The detailed mechanisms governing the entrance gate to the binding site remain unclear. To investigate the dynamics of the entrance gate for ligand entry into the binding site, we performed molecular dynamics simulations of MAO-B in the mitochondrial membrane, both with and without PET tracers. We found large entrance gates in the MAO-B dimer, and these gates exhibited asymmetry within the dimer. The dynamics of these large entrance gates were linked to the dynamics of the dimer interface, with the slight asymmetry of the dimer interface dynamics contributing to the asymmetric dynamics of the entrance gates.

[1] C. Binda, et al., "Insights into the mode of inhibition of human mitochondrial monoamine oxidase B from high-resolution crystal structures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 17, pp. 9750-9755, 2003.

[2] A. F. Brooks, et al., "In vivo metabolic trapping radiotracers for imaging monoamine oxidase-A and -B enzymatic activity," *ACS Chemical Neuroscience*, vol. 6, no. 12, pp. 1965-1971, 2015.

[3] J. S. Fowler, et al., "Selective reduction of radiotracer trapping by deuterium substitution: Comparison of carbon-11-L-deprenyl and carbon-11-deprenyl-D2

for MAO B mapping," *Journal of Nuclear Medicine*, vol. 36, no. 7, pp. 1255-1262, 1995.

[4] R. Harada, et al., "18F-SMBT-1: A selective and reversible PET tracer for monoamine oxidase-B imaging," *Journal of Nuclear Medicine*, vol. 62, no. 2, pp. 253-258, 2021.

P01-16

Conformational study of macrocyclic peptides in solvent by MD simulations to improve their membrane permeability

Ekishin YANAGI ^{*1}, **Patricia BRANDL**², **Stephanie M. LINKER**², **Sereina RINIKER**², **Takayuki KATOH**¹, **R. H. P. van NEER**³, **Hiroaki SUGA**¹

¹Department of Chemistry, School of Science, The University of Tokyo

²Computational Chemistry, Institute of Molecular Physical Science, Department of Chemistry and Applied Biosciences, ETH Zurich

³National Center for Advancing Translational Sciences, National Institutes of Health

(* E-mail: ekishin@g.ecc.u-tokyo.ac.jp)

Short macrocyclic peptides, composed of 8 to 15 amino acids, are attracting attention as protein-protein interaction (PPI) inhibitors. Their small and constrained structure allows for both a unique combination of conformational rigidity and flexibility. This enables peptides to fit easily into flat and large surface areas of target proteins, resulting in selective and strong interactions. However, membrane permeability decreases as their molecular size increases, which has been the major bottleneck for therapeutic peptide development targeting intracellular proteins.

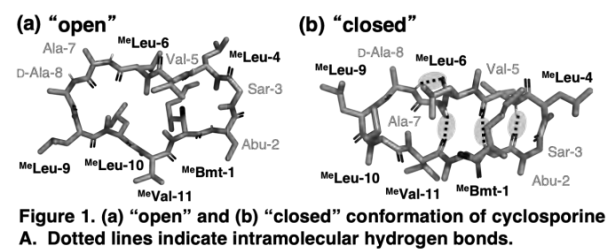
Despite many peptides not being membrane-permeable, there are some notable exceptions, such as cyclosporine A (CsA), known as possessing high membrane permeability. Membrane-permeable peptides are said to have the following characteristics.

1. Having few charged amino acids, since the surface of the membrane is hydrophobic.
2. Including N-methyl amino acids; CsA also contains many N-methyl amino acids.
3. Having conformational flexibility between “open” (no intramolecular hydrogen bonds) and “closed” (intramolecular hydrogen bonds are formed and polar groups are maximum shielded); computational studies have demonstrated that CsA has this feature (Figure 1).

As the conditions for membrane-permeable peptides are still not fully understood, conformational analysis by Molecular dynamics (MD) simulations is crucial to our further understanding.

The aim of this research is to establish a generic methodology to improve the membrane permeability of bioactive macrocyclic peptides with poor membrane permeability by combining wet and dry experiments. To achieve this purpose,

we chose SaD3, which showed strong binding affinity and inhibitory activity against *Staphylococcus aureus* co-factor-independent phosphoglycerate mutase (Sa-iPGM), as a starting point (Neer et al. ACS Chem. Biol. 2022, 17). SaD3 has two charged amino acids and one N-methyl amino acid and is not membrane-permeable. As Sa-iPGM is an intracellular target, improving the membrane permeability of SaD3 could make SaD3 a powerful drug for infectious diseases. To this end, we ascertained whether SaD3 shows conformational flexibility between “open” and “closed”. Also, we investigated how conformation would change when the charged and N-methyl amino acids were replaced by other amino acids. MD simulations in two solvents were performed for SaD3, SaD3-E12A (Glu at position 12 was changed to Ala), and SaD3-dNMe (MeSer at position 9 was changed to Ser) (Figure 2). The results showed that the SaD3 tend to take both “open” and “closed” conformation. Additionally, we found that the substitution of even a single amino acid could significantly change the conformation of the peptides.



name	Peptide sequence
SaD3	Ac _y QVTVWWA ^{Me} SPWEDC-NH ₂
SaD3-E12A	Ac _y QVTVWWA ^{Me} SPW ^{Ala} DC-NH ₂
SaD3-dNMe	Ac _y QVTVWWA SPWEDC-NH ₂ <div style="text-align: center;"> S </div>

Figure 2. Sequence of peptides to be simulated

P01-17

High-precision and Efficient Prediction of Intermolecular Interaction Energies Using Deep Learning on Quantum Chemical Calculation Data

Yudai KOBAYASHI *, Natsuki KANAZAWA, Shigeyuki MATSUMOTO, Takao OTSUKA, Yasushi OKUNO

Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University

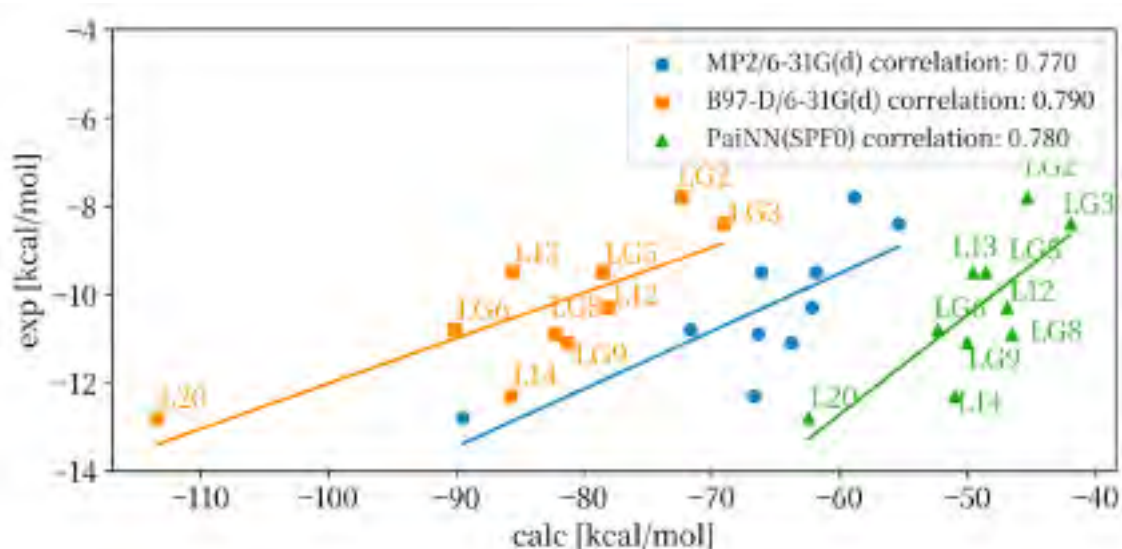
(* E-mail: kobayashi.yudai.72r@st.kyoto-u.ac.jp)

Understanding local interactions between drug molecules and target proteins is crucial in life sciences. While quantum mechanics (QM) calculations offer precise insights, they are computationally expensive for large-scale systems or extensive compound screening. Recent advances in artificial intelligence, particularly deep learning applied to QM calculations, have shown promise in addressing this challenge. We developed a QM-AI model based on PaiNN [1], an improved version of SchNet [2], trained on a neutral subset of the Solvated Protein Fragments (SPF) dataset [3], to predict intermolecular interaction energies in protein-ligand systems, specifically applying it to FK506-binding protein and ten ligand compounds. Figure 1 shows the correlations between the experimental binding affinities and the computed interaction energies. Our PaiNN(SPFO) model (green triangles) showed comparable performance to previous QM calculations using multilayer fragment molecular orbital (FMO) method by FMO2-MP2/6-31G(d) (blue circles) and FMO2-B97-D/6-31G(d) (orange squares) [4]. The correlation coefficients between the experimental and the computed values were $R=0.780$ for PaiNN(SPFO), $R=0.790$ (MP2), and $R=0.770$ (B97-D) for the QM calculations. These results demonstrate that our PaiNN(SPFO) model can predict FKBP-ligand interactions with accuracy comparable to QM calculations. It's remarkable that our model, trained on small multi-molecular systems, can effectively represent complex many-body interactions in large protein-ligand systems. Next, we focused on the computational efficiency of our QM-AI model. Typically, the QM computational times increase exponentially with the system size. Previous studies have indicated that the QM computational times for FKBP-ligand interactions were approximately 6 hours for the smallest ligand, LG2 (1712 atoms), and about 30 hours for the largest ligand, L20 (1792 atoms). In contrast, our QM-AI model, PaiNN(SPFO), has computed the FKBP-ligand interaction energies in 5.33 seconds for FKBP-LG2 and 5.76 seconds for FKBP-L20, demonstrating remarkable computational efficiency for such protein-ligand systems. Our

proposed QM-AI model demonstrates the potential for rapid and efficient calculation of intermolecular interaction energies in protein-ligand systems.

References:

1. Schütt, K. T., Unke, O. T. & Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. (2021).
2. Schütt, K. T. et al. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. (2017).
3. Unke, O. T. & Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *Journal of Chemical Theory and Computation* 15, 3678–3693 (2019).
4. Otsuka, T., Okimoto, N., Taiji, M., Assessment and Acceleration of Binding Energy Calculations for Protein-Ligand Complexes by the Fragment Molecular Orbital Method. *J. Comput. Chem.*, 36, 2209-2218 (2015).



P01-18

Mechanisms of Type-51 R-body conformational changes revealed by in silico methods

Hiroaki OHEDA ^{*1}, Toru EKIMOTO¹, Tsutomu YAMANE², Kosuke KIKUCHI³, Koki DATE³, Takafumi UENO³, Mitsunori Ikeguchi^{1, 2}

¹Yokohama City Univ.

²RIKEN

³Tokyo Institute of Technology

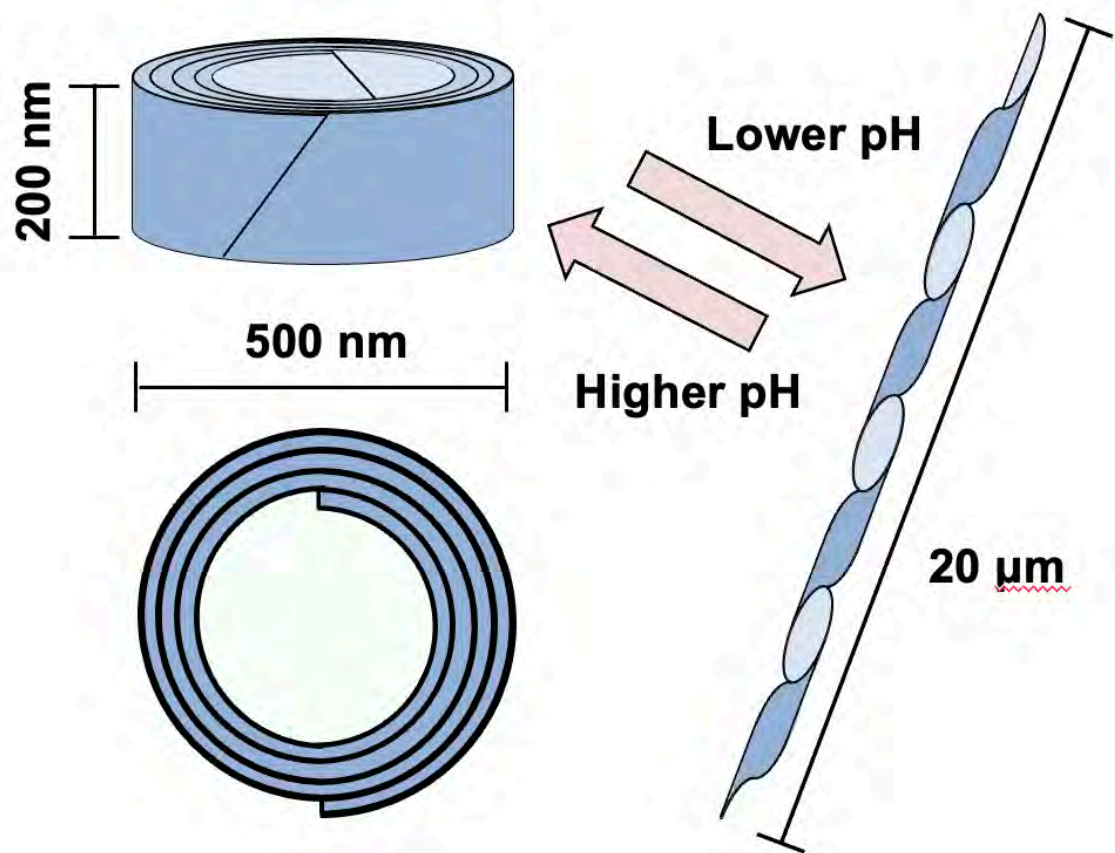
(* E-mail: w235411b@yokohama-cu.ac.jp)

Type 51 Refractile Body (R-body) is a cylindrical and coiled protein polymer found in the cytoplasm of *Caedibacter taeniospiralis*, an intracellular symbiotic bacterium residing in *Paramecium*. The R-body is primarily composed of two major proteins, RebA and RebB. This polymer exhibits reversible structural changes: it can switch from a cylindrical form, approximately 200 nm in height, to a needle-like structure extending up to 20 μ m in length, depending on the pH.

Recent studies utilizing solid-state NMR (SS-NMR) have revealed that RebA and RebB, the main components of the R-body, are predominantly composed of α -helices. These studies also identified several regions in the proteins that act as helix breakers. Despite these advancements in understanding its partial secondary structure, the unique coiled arrangement of the polymer significantly complicates the comprehensive analysis of its structure and function.

The goal of this study is to clarify the polymerization mechanisms and pH-dependent conformational changes of the R-body through various in silico methods. Molecular dynamics (MD) simulations were conducted based on a monomer structure predicted by AlphaFold2, which confirmed the presence of kinks at several residues. These kinked regions correspond to the helix breaker regions identified in previous SS-NMR studies. Furthermore, MD simulations under different pH conditions highlighted several key residues likely involved in the pH-dependent structural transitions of the R-body.

Future research will focus on constructing a comprehensive model that integrates these monomeric findings with experimental data. This will be followed by MD simulations aimed at further elucidating the mechanisms behind the dynamic structural transformations of the entire R-body.



P01-19

Molecular simulation analysis for nucleic acids

Kenji YAMAGISHI ^{*1}, Hiroyuki TSUKADAH², Haruto NARITA¹, Shota MYOCHIN¹, Hisae YOSHIDA¹, Seiichiro ISHII¹, Masahiro SEKIGUCHI¹, Takeshi ISHIKAWA³, Taiichi SAKAMOTO⁴

¹Nihon University

²Nissan Chemical CORPORATION

³Kagoshima University

⁴Chiba Institute of Technology

(* E-mail: yamagishi.kenji@nihon-u.ac.jp)

A nucleic acid is a macromolecule composed of nucleotide chains. The most common nucleic acids, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), are biopolymers essential to life. We studied the structure and function of the nucleic acids using molecular simulations. In this presentation, we provide an overview of our research project.

1. Aptamer Aptamers are short, single-stranded oligonucleotides that bind to specific target molecules such as proteins, nucleic acids, and small molecules. An aptamer that binds to the Fc portion of human Immunoglobulin G (IgG) was identified using the SELEX technique. We have been conducting research aimed at clarifying the mechanism of the high affinity and specificity of the aptamer against IgG using fragment molecular orbital calculations and molecular dynamics simulations.

Ref [1] H. Yoshida, T. Ishikawa, T. Sakamoto, K. Yamagishi, et al., Chem. Phys. Lett., 738, 136854 (2020).

2. Antisense Oligonucleotides Antisense oligonucleotides (ASOs), composed of single-stranded DNA-like oligonucleotides, represent a key category of oligonucleotide therapeutics, and constitute the majority of oligonucleotide therapeutics approved by 2024. We have been conducting molecular dynamics simulations of the RNase H complex with ASO/RNA duplex structure to elucidate the structural dynamics and interactions.

3. Modified Nucleoside Various chemically modified nucleoside analogs have been developed to improve the properties and performance of natural nucleic acids. We performed molecular dynamics calculations of the novel thymidine analogs to investigate the changes in nucleoside sugar puckering and molecular fluctuations caused by chemical modifications.

Ref [2] Y. Zhou, S. Ishii, K. Yamagishi, Y. Ueno, et al., RSC Adv., 10, 41901 (2020).

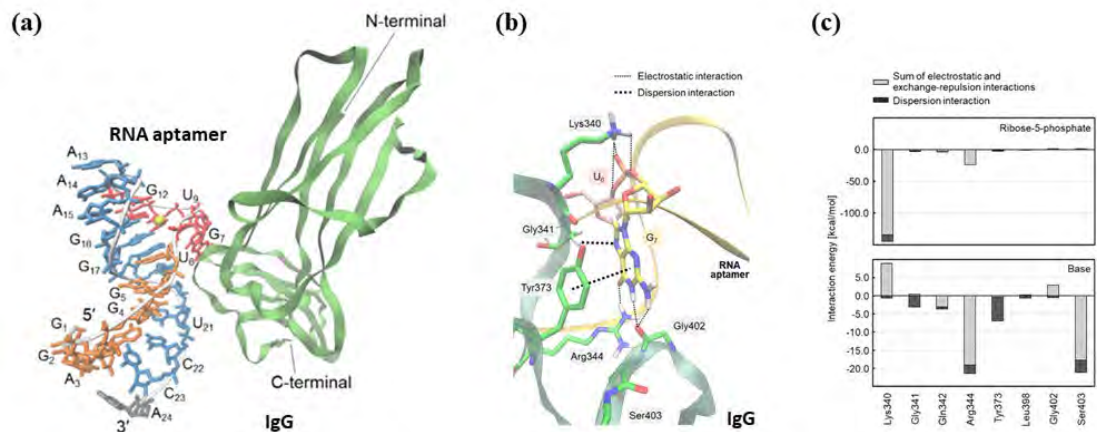


Figure 1. Interaction analysis between RNA aptamer and IgG using FMO calculation

P01-20

Induced-Fit Posing (IFP): A new pose prediction tool for hit to lead stage of drug discovery

Samuel TOBA *, Chiharu KONDA

OpenEye, Cadence Molecular Sciences
(* E-mail: samuel.toba@eyesopen.com)

Accurate prediction of binding poses is a cornerstone of structure-based drug design, essential for developing effective therapeutic agents. The accuracy of these predictions determines the efficiency of identifying and optimizing lead compounds, ultimately impacting the success rate of drug discovery projects. Accurate binding pose prediction is particularly achievable during the lead optimization phase, where the molecules of interest often share significant structural similarities with known crystallographic ligands. This similarity simplifies the docking process, allowing for more reliable predictions. However, the hit-to-lead stage introduces complexities that can challenge the reliability of these predictions due to structural diversity.

The introduction of Induced-Fit Posing (IFP) addresses these challenges by incorporating flexibility into the docking process, allowing for more accurate predictions of how diverse ligands bind to target proteins. Through a multi-step process involving pruning, hypothesis generation, MD simulation, and consensus scoring, IFP significantly enhances pose prediction accuracy. Retrospective cross-docking studies validate its effectiveness, showing a substantial improvement in successful predictions. By adopting IFP, researchers can better navigate the complexities of the hit-to-lead phase, predict compounds with diverse chemotypes, incorporate protein binding site flexibilities into their models, and advance their drug discovery efforts with greater confidence.

P01-21

Generation of a suitable structure for SBDD by AlphaFold2 via Genetic Algorithm Parameter Search

Keisuke UCHIKAWA *, Kairi FURUI, Masahito OHUE

Department of Computer Science, School of Computing, Institute of Science
Tokyo

(* E-mail: uchikawa.k.ac@m.titech.ac.jp)

Structure-based virtual screening (SBVS), which utilizes protein conformational information, has gained significant attention in recent years due to its potential to discover highly novel drug candidate compounds more effectively than other methods. However, a major challenge lies in the variability of screening accuracy depending on the conformation of the target protein. In this study, we focused on the combination of AlphaFold2, a representative method for protein structure prediction, and SBVS. We explored a method to improve the accuracy of SBVS using predicted structures by optimizing the parameters used in AlphaFold2 with a genetic algorithm.

Specifically, we used only shallow MSA (multiple sequence alignment) for prediction with AlphaFold2, and then explored how to introduce mutations based on docking scores using a genetic algorithm. As a result, we obtained predicted structures for CXCR4 that exhibited SBVS performance significantly surpassing that of the PDB structure. For KIF11, although the performance was slightly inferior to the PDB structure, we were able to generate predicted structures with performance that could not be achieved by the standard predictions of AlphaFold2. These results suggest that the application range of SBVS can be expanded by utilizing predicted structures from AlphaFold2.

P02-01

Investigation of the Allosteric Binding Sites of ERK2 by Metadynamics Simulation

Hajime SUGIYAMA ^{*1}, Seisuke HASEGAWA², Mayu YOSHIDA², Takayoshi KINOSHITA²

¹Mitsubishi Chemical Corp.

²Osaka Metropolitan University

(* E-mail: hajime.sugiyama.ma@mcgc.com)

ERK2 (extracellular signal-regulated kinase 2) is a member of the mitogen-activated protein kinase family. STAT3 (signal transducer and activator of transcription 3), which plays an essential role in normal glucose homeostasis, and regulates cell proliferation, differentiation, and various other cellular responses. Based on previous studies on the ERK2/STAT3 pathway, the authors focused on ERK2 as a candidate target for diabetes treatment and identified small molecule compounds with inhibitory activity through in silico screening[1]. The crystal structures of ERK2 in complex with the compounds revealed that one previously reported inhibitor binds to a novel allosteric site in close proximity to the TXY motif of ERK2[2], while the other inhibitors bind to the KIM site[3,4]. To estimate the origin of the difference in binding sites between the inhibitors, we performed a computer simulation analysis. Simulations were performed for each binding site of a single molecule of ERK2 in water, and their stable binding modes and binding free energies were calculated. A metadynamic simulation approach was used to predict the binding state from the dissociation state of the inhibitors without a priori assumption of the binding modes. The resulting binding modes to the KIM site and to the novel site showed different characteristics, reflecting the structural flexibility derived from the composition of the surrounding residues. Furthermore, the calculated binding free energies differed from the experimental results, with higher affinity on the KIM site than on the novel side. These results provide one hypothesis that the inhibitor interacts with a multichain binding site composed of ERK2 dimers, observed in crystal structure, rather than a single ERK2 binding site.

[1] T. Kinoshita, et al, Bioorg. & Med. Chem. Lett. 26: 955–958 (2016)

[2] M. Yoshida, et al, Biochem. Biophys. Res. Commun. 593: 73–78 (2022)

[3] H. Sugiyama, et al, Bioorg. & Med. Chem. Lett. 93: 129431 (2023)

[4] S. Hasegawa, et al, Biochem. Biophys. Res. Commun. 704: 149707 (2024)

P02-02

PairMap: An Intermediate Insertion Approach to Improve Accuracy in Relative Free Energy Perturbation Calculations of Distant Compound Transformations

Kairi FURUI ^{*1}, Takafumi SHIMIZU², Yutaka AKIYAMA¹, Roy S. KIMURA², Yoh TERADA², Masahito OHUE¹

¹School of Computing, Institute of Science Tokyo

²Alivexis, Inc.

(* E-mail: furui@li.c.titech.ac.jp)

Accurate prediction of the difference in binding free energy between compounds is crucial for reducing the high costs associated with drug design and lead optimization.

Relative binding free energy perturbation (RBFEP) calculations are effective for small structural changes; however, large topological changes pose significant challenges for calculations, leading to high errors and difficulties in convergence. To address such issues, we propose a new approach---PairMap---that focuses on introducing appropriate intermediates for complex transformations between 2 compounds. PairMap generated intermediates exhaustively, determined the optimal conversion paths, and introduced thermodynamic cycles into the perturbation map to improve accuracy and reduce computational cost.

PairMap succeeded in introducing appropriate intermediates that could not be discovered by existing simple approaches by comprehensively considering intermediates.

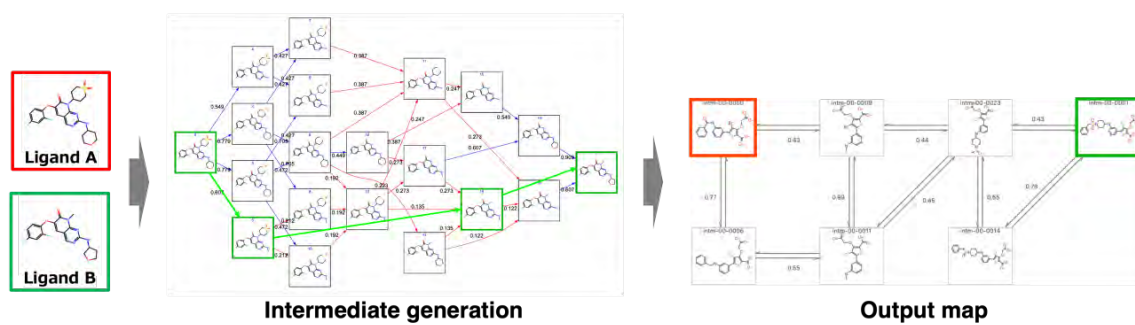
Furthermore, we evaluated the accuracy of the prediction of binding free energy using 9 compounds selected from Wang *et al.*'s benchmark set, which included particularly complex transformations. The perturbation map generated by PairMap achieved excellent accuracy with a mean absolute error of 0.93 kcal/mol compared to 1.70 kcal/mol when using the perturbation map generated by the conventional Flare FEP intermediate introduction method.

Moreover, in a scaffold hopping experiment conducted with the PDE5a target involving complex transformations, PairMap provided more accurate free energy predictions than ABFEP calculations, yielding more reliable results compared to experimental data.

Additionally, PairMap can be utilized to introduce intermediates into congeneric series, demonstrating that complex links on the perturbation map can be resolved with minimal addition of intermediates and links.

In conclusion, PairMap overcomes the limitations of existing methods by

enabling RBFEP calculations for more complex transformations, further streamlining lead optimization in drug discovery.



P02-03

Fragment Molecular Orbital Calculations for Zinc-Containing smHDAC8

Siyun WANG *, Sota TANAKA, Shuhei MIYAKAWA, Yu-Shi TIAN, Daisuke TAKAYA, Kaori FUKUZAWA

Graduate School of Pharmaceutical Sciences, Osaka university

(* E-mail: wangsiyun001121@outlook.com)

Background

Fragment Molecular Orbital (FMO) calculations have been proven useful in drug design and protein-ligand binding analysis. However, metal ions contained in biomolecule systems limit the application of this method due to the difficulty in both fragmentation and complex electronic structure of transition metal ions. This study used a zinc (Zn^{2+})-containing protein *Schistosoma mansoni* Histone Deacetylase 8 (smHDAC8) to investigate the methodology and attempted to provide a protocol for FMO calculation for metalloprotein.

Methods

smHDAC8 contains a catalytic Zn^{2+} active site for ligand binding. 14 active ligands of smHDAC8 have been reported by Martin Marek *et al.* [1]. Initially, we selected one complex as a template structure (PDB ID: 6hsh). Other complexes were constructed by replacing the coordinates of ligands and performing structural optimization using Amber10:EHT force field. All the structures were prepared by MOE. After structure preparations, energy minimizations for side chains were conducted using a constraint of tether 1.0. Subsequently, Zn^{2+} and surrounding residues were constructed as one large fragment (i.e., merging code). FMO calculations were performed using the 6-31G* basis set. The energy analysis was conducted in a "super-molecule" manner:

$$E_{\text{bind}} = E_{\text{protein}} + E_{\text{ligand}} - E_{\text{complex}}$$

Finally, the binding energies of ligands were compared with their biological activities measured as IC_{50} values.

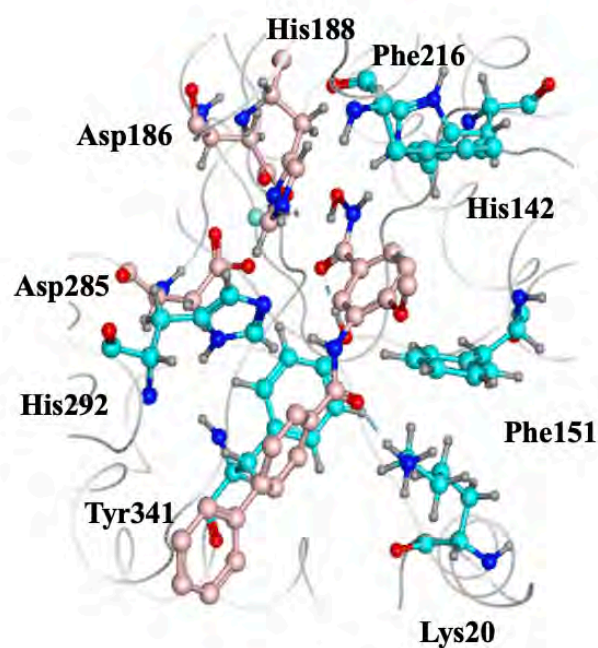
Results and Discussion

FMO calculations for all these complex models succeeded, indicating an executable approach for metalloprotein. The interactions between the key ligand and the protein observed in the experiment were validated using IFIE analysis performed with FMO, and the results were consistent. Additionally, it was found that the residues His142, Phe151, and Phe216 form π - π interactions with the ligand, which are also crucial for the ligand function.

Reference

[1] Marek M, Shaik TB, Heimbürg T, et al. Characterization of Histone

Deacetylase 8 (HDAC8) Selective Inhibition Reveals Specific Active Site Structural and Functional Determinants. J Med Chem. 2018;61(22):10000-10016. doi:10.1021/acs.jmedchem.8b01087



**Energy minimization
for side chains**

FMO calculation

**Correlation between
IFIE and pIC₅₀**

P02-04

Interaction Analysis between pHLA and TCR using MD Simulation and Fragment Molecular Orbital Calculation

Suzu ITAMI ^{*1}, **Yoshiki ARITSU**², **Mizuki KITAMATSU**³, **Chihiro MOTOZONO**², **Norihito KAWASHITA**³

¹Graduate School of Science and Engineering, Kindai University

²Joint Research Center for Human Retrovirus infection, Kumamoto University

³Faculty of Science and Engineering, Kindai University

(* E-mail: 2333310130r@kindai.ac.jp)

Keywords: Fragment Molecular Orbital Method, Molecular Dynamics, SARS-CoV-2

The T cell receptor (TCR) on cytotoxic T cells recognizes peptide fragments derived from viral proteins presented on HLA class I molecules on virus-infected cells as antigens. T cells that sense virus-infected cells play an important role in controlling viral infection by killing these infected cells. In mRNA vaccines used for novel coronavirus infections, it has become clear that vaccine-induced T cells are also involved in controlling viral infection. In future vaccine boosters, selective induction and reactivation of functional T cells may be important for controlling infection and preventing severe disease.

The 9-residue peptide NF-9 (NYNYLYRLF), derived from the SARS-CoV-2 spike protein, forms a complex (pHLA) with HLA, which is recognized by the TCR. The delta/omicron BA.5 variant also has an NF9-5R (NYNYRYRLF) containing a Leu5Arg mutation, and pHLAs with this mutation have been found not to bind to the TCR^[1]. In this study, MD simulations and Fragment Molecular Orbital (FMO) calculations^[2] were performed on pHLA before and after the mutation to elucidate the molecular mechanism by which with the Leu5Arg mutation evades TCR recognition.

Based on the crystal structure of HLA/NF-9, two structures, WT and NF9-5R, were created using MOE 2020. Force field were applied using Ambertools22 and GROMACS 2021.5 was used for MD simulations and their analysis. Simulations of 50 ns were run three times for each structure, sampling every 1 ns between 10-50 ns obtaining 41 structures per simulation. FMO calculations were performed on the sampled structures, and IFIE (Inter-Fragment Interaction Energy) and PIEDA (Pair Interaction Energy Decomposition Analysis) component values^[3] were calculated using ABINIT-MP Open Ver. 1 Rev. 22 at a calculation level of MP2/6-31G*.

MD simulation results showed that the RMSF of C α atoms of the 4th-6th residues in NF-9 was larger in NF9-5R, indicating that the mutation made it more prone to fluctuations. The average distance between hydroxyl group of Tyr6 and carbonyl group of Asn3 in NF9-5R was about 2 Å smaller than in WT. FMO calculations showed that the average ES values of Tyr4 and Tyr6 were -4.9 ± 6.4 kcal/mol in WT and -12.6 ± 8.8 kcal/mol, indicating that the mutation results in stronger electrostatic interactions. These results suggest that Tyr6 of NF9-5R is more likely to form hydrogen bonds with Asn3, thus making it less likely to face the TCR side and no longer bind to the TCR.

This study was conducted as part of the activities of the FMO Drug Design Consortium (FMO DD), and the supercomputer "Fugaku" was used for MD simulations and FMO calculations (project numbers: hp230131, hp240162).

- [1] Motozono C., *et al.*, *Cell Host & Microbe*. **29**(7), 1124-1136.e11, (2021).
- [2] Kitaura K., *et al.*, *Chem Phys Lett*. **313**(3-4), 701-708, (1999).
- [3] Fedorov D. G., *et al.*, *J. Comput. Chem*. **28**(1), 222-237, (2007).

P02-05

NNP-based Force Field Optimization to Improve RBFEP Performance

Junya YAMAGISHI *, Yunoshin TAMURA

Preferred Networks

(* E-mail: jyamagishi@preferred.jp)

Accurate binding affinity prediction techniques, such as free energy perturbation (FEP), have become very effective tools for enhancing small molecule drug discovery. However, since FEP is based on molecular dynamics simulations, it is known that the accuracy of binding affinity prediction depends on the accuracy of the molecular force field (FF). Our validation studies have shown that the accuracy of the existing FFs for complicated drug-like compounds is not sufficient. It is necessary to improve the accuracy of FF for more accurate prediction of binding affinity.

There are several challenges to optimizing FF parameters. First, each FF parameter, including bonded terms and partial charges, interferes with each other. This means that there are no universal parameters that can be applied to multiple molecules: ideally, tailor-made FF parameters should be made for each molecule.

The second is related to accurate energy and force calculations used as a reference for optimizing FF parameters, which are typically performed by quantum mechanical (QM) calculations. QM calculations of drug-like compounds ($M_w < 500$) with thousands of conformations are very time-consuming and impractical when tailoring FF parameters.

To overcome these challenges, we applied a new protocol to tailor-make FF parameters for each drug-like molecule using neural network potential (NNP) instead of QM calculations. Preferred Potential (PFP) [1] was used as the NNP, whose accuracy has been verified for drug-like molecules. Because PFP can predict accurate energies and forces thousands of times faster than QM calculations, we were able to generate tailor-made FF parameters for each full-sized drug-like molecule. In this presentation, we will show the accuracy of PFP for drug-like molecules and the performances of FF parameter optimization using PFP. We will also show comparisons of the results of relative binding free energy perturbation (RBFEP) performed with our service, called P-FEP[2], using existing and optimized FF parameters.

- [1] Takamoto, S., Shinagawa, C., Motoki, D. et al. Towards universal neural network potential for material discovery applicable to arbitrary combination of 45 elements. Nat Commun 13, 2991 (2022).
- [2] <https://tech.preferred.jp/ja/blog/pfep-launch/>

P02-06

SpatialPPI 2.0: Enhancing Protein-Protein Interaction Prediction Through Distance Matrix Analysis Using Link Regression in Graph Attention Networks

WENXING HU *, Masahito OHUE

Tokyo Institute of Technology
(* E-mail: perseids2032@gmail.com)

Protein-protein interactions (PPIs) play a pivotal role in a wide array of biological processes, and accurately predicting these interactions is essential for advancing our understanding of cellular functions. In this study, we introduce a novel computational approach that leverages link regression within Graph Attention Networks (GATs) to predict spatial distances between two independent protein structures. Our method is designed to enhance the prediction of PPIs by focusing on the distance between residues rather than the conventional approach of contact map prediction.

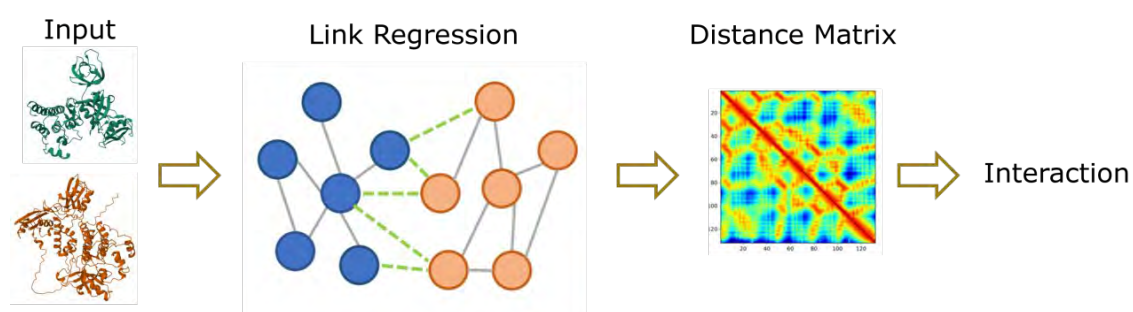
We trained our model on the Protein-Protein Interfaces Prediction dataset from the ATOM3D project and validated its effectiveness on AlphaFold-predicted datasets. The uniqueness of our approach lies in its utilization of link regression to estimate inter-residue distances, which can be further processed to refine the prediction of protein-protein interfaces. Unlike existing methods that employ link classification, our model directly predicts the distance between residues, providing a more nuanced understanding of potential interaction sites.

Traditional GNN-based models typically use a Siamese network to independently extract features from each protein structure. In contrast, our method innovatively combines two protein structures with variable links between them, allowing the model to dynamically update the distances and capture the influence of residues across the interacting proteins. This approach offers a more holistic view of protein interactions, as it considers the complex interdependencies between residues in both proteins.

Comparative analysis with existing methods such as D-SCRIPT, which uses sequence-based approaches and contact maps as intermediate predictions, demonstrates the potential of our model to improve the accuracy and reliability of PPI predictions. Our method also shows promise when compared to other GNN-based models like PIPR and GNNGL-PPI, highlighting the advantages of

integrating spatial distance prediction with graph-based learning.

The potential applications of our findings are far-reaching, extending to any domain where accurate PPI prediction is critical. This includes drug discovery, where understanding PPIs can inform the design of inhibitors or enhancers, and in systems biology, where mapping interaction networks can elucidate the pathways underlying various diseases. This study underscores the potential of GAT-based models with link regression in advancing the field of structural bioinformatics, offering a new direction for PPI prediction methodologies.



P02-07

Development of RIKEN Natural Products Depository Database

Xingmei OUYANG *¹, Hiroyuki HIRANO¹, Hiroyuki OSADA^{1, 2}

¹RIKEN Center for Sustainable Resource Science, RIKEN

²Institute of Microbial Chemistry

(* E-mail: xingmei.ouyang@riken.jp)

The RIKEN Natural Products Depository (NPDepo) is a public depository of chemical compounds. Currently, the NPDepo contains 66,313 pure compounds, about one-third of which are natural products and their derivatives. It has provided a chemical library at the requests of domestic and international researchers. The chemical library is an indispensable resource for exploring useful bio-active compounds.

To promote effective utilization of the NPDepo compounds, an effective structure database system must be constructed for this library. However, natural products remain difficult to encode by conventionally used molecular fingerprinting methods because of their complex and diverse structures. The database system for such a diverse range of compounds, including natural compounds, was insufficient to perform similar structure searches and clustering. We have studied the characteristics and steric features of complex fused ring systems of natural compounds and designed the NPDepo Informatics System (NIS) with 16 fragment types as parameters, focusing on evaluation and application. We compared the similarity search results of our system with results from the previously used fingerprinting methods available in RDKit (MACCS key, Morgan, RDKit, Topological Torsions, AtomPairs).

The results of our developed system show that it is a fast and effective tool for searching for structural analogues of hit compounds in biological evaluations and for clustering natural compounds in complex fused ring systems in the NPDepo library into general fused ring system structure groups.

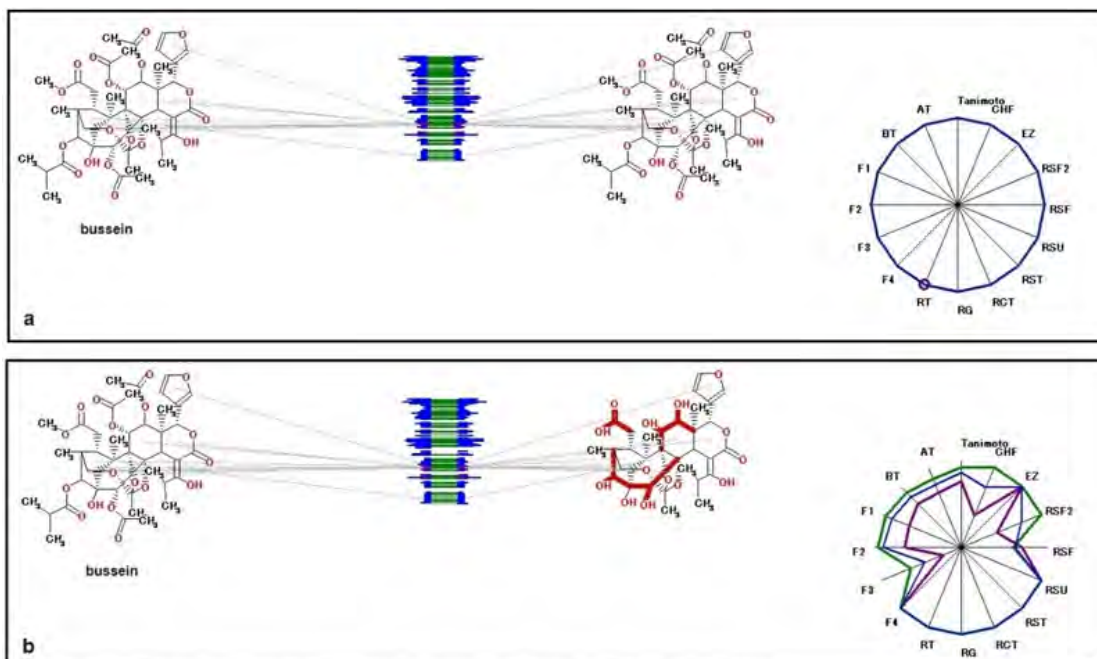


Fig.1 Results of bussein's similarity searching in NPDepo database were shown by the NPDepo Informatics System(NIS) with sixteen parameters. a) a compound which is completely consistent with bussein is found; b) one of compounds which are partially consistent with bussein is shown.

P02-08

Machine learning based prediction of quantum mechanical interaction energy between amino acid residues using fragment molecular orbital method

Tomohiro SATO ^{*1}, Watanabe CHIDURU¹, Okiyama YOSHIO²

¹Center for Biosystems Dynamics Research, RIKEN

²Graduate School of System Informatics, Kobe University

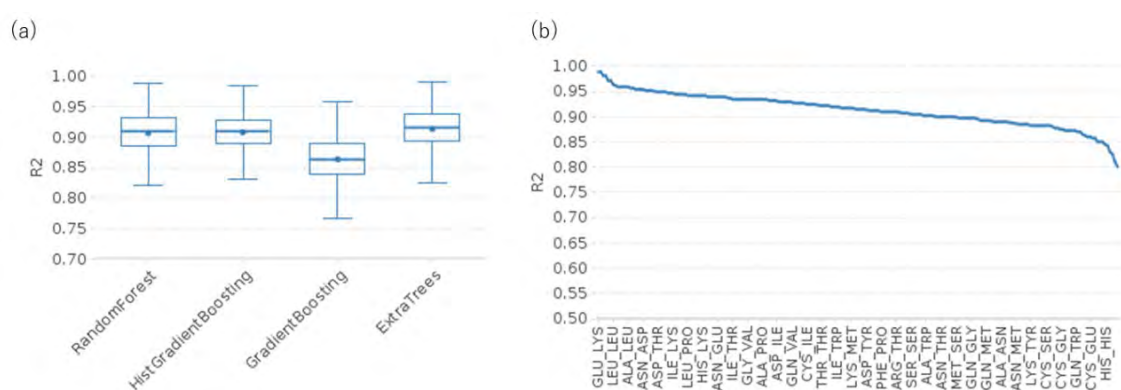
(* E-mail: tomohiro.sato@riken.jp)

Recently, various attempts for the application of quantum mechanics (QM) calculation to biomolecules were reported due to the development of methods to accelerate QM calculation for large molecular systems such as QM/MM or fragment molecular orbital (FMO) method [1, 2]. However, the computation cost is still high to apply QM calculation to high-throughput screening or molecular dynamics in which more than thousands of calculations were required. In this study, we created machine learning models to emulate interfragment interaction energies (IFIEs) calculation between amino acid residues by learning those data in FMO database (FMO DB, URL: <https://drugdesign.riken.jp/FMODB/>) [3,4]. Thus, the model can be used as the alternative to conventional molecular force fields to evaluate inter/intra protein interactions for MD calculation or evaluation of antibody-antigen interaction.

The regression models of IFIEs were built using random forest, extra trees, gradient boosting, and histogram-based gradient boosting based on the FMO dataset of 6,946 apoproteins, including 20,835 entries registered in FMO DB. For each of the amino acid fragment pairs within 6 Å, which are not directly connected, the pairwise distances of heavy atoms in respective fragments were used as the explanatory variables to encode the geometric arrangement of the fragments, and learned with corresponding IFIEs. Among the machine learning techniques, the extra trees regressor recorded the highest prediction performance with R^2 of 0.907 and RMSE of 2.032 in an average of all the 210 combinations of amino acid residues (Fig). The models provided excellent performances in amino acid pairs forming strong electrostatic interactions like glutamate-lysine pair and aspartate-lysine ($R^2=0.990$, 0.989 by extra trees, respectively), and relatively low performances in cysteine-related pairs for which relatively small amount of structural data are available, such as non-disulfide bonded cysteine pair and cysteine-histidine pair ($R^2=0.811$, 0.800). Structural preparation procedure and consistency between the predicted IFIEs and their energy elements decomposed by PIEDA analysis are also to be assessed.

References

1. Warshel, A.; Levitt, M. J. Mol. Biol. 1976, 103, 227–249.
2. Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M., Chem. Phys. Lett. 1999, 313, 701–706.
3. Watanabe, C.; Watanabe, H.; Okiyama, Y.; Takaya, D.; Fukuzawa, K.; et al., CBIJ. 2019, 19, 5–18.
4. Takaya, D.; Watanabe, C.; Nagase, S.; Kamisaka, K.; Okiyama, Y.; et al., J. Chem. Inf. Model. 2021, 61, 777–794.



P02-09

Computational assessment of the binding mode of Verteporfin, an inhibitor targeting the YAP-TEAD protein-protein interaction

Yurika Ikegami ^{*1}, **Genki Kudo** ², **Takumi HiraO** ¹, **Ryunosuke Yoshino** ^{3, 4}, **Takatsugu Hirokawa** ^{3, 4}

¹Degree programs in Comprehensive Human Sciences, Graduate School of Comprehensive Human sciences Doctoral Program in Medical Sciences, University of Tsukuba

²b Doctoral Program in Physics Degree Programs in Pure and Applied Sciences, Graduate School of Science and Technology, University of Tsukuba

³Transborder Medical Research Center, University of Tsukuba

⁴Division of Biomedical Science, Faculty of Medicine, University of Tsukuba

(* E-mail: s2230361@s.tsukuba.ac.jp)

Purpose:

Inhibiting overexpressed yes-associated protein (YAP) through YAP inhibitors is expected to improve the prognosis of patients with malignant tumor. Verteporfin (Verteporfin: VP) has been reported to inhibit the formation of the YAP-TEA domain family member (TEAD) complex. Previous docking studies have suggested that VP bind to the WW domain of YAP or the TEAD interaction surface of YAP1,2). However, the precise binding poses have not been elucidated. Additionally, it remains unclear which of the four VP isomers is the most effective against YAP. In this study, we re-evaluate the interaction between VP and YAP using in silico techniques such as docking simulation and AlphaFold2(AF2), a highly accurate protein structure prediction method. Furthermore, we aim to identify the isomer with the strongest inhibitory activity against YAP to enhance its efficacy as an adjuvant-like drug.

Methods:

We performed the in silico techniques in the following steps; (1) The YAP-TEAD complex was modeled using ColabFold, (2) Docking simulations were conducted between the four VP isomers, named Ia-1, Ia-2, Ib-1, and Ib-2 and YAP using Glide, and (3) The binding pose clusters were analyzed. The docking sites were determined at the YAP-TEAD interaction interface and the druggable pocket, generating two grids at the docking sites. The docking results were rescored using binding free energy calculations using MM/GBSA method. Clustering of the binding modes was performed using Protein-Ligand Interaction Fingerprint (PLIF).

Results:

As the results of docking calculation and clustering analysis, the binding poses were classified into nine clusters. The Ia-2 isomer showed the lowest binding free energy. The pose generated from the grid centered on the druggable pocket formed hydrogen bonds with Gln200, Thr197, and Lys254; however, since these residues are not involved in protein-protein interactions (PPI) with TEAD, PPI inhibition did not occur. On the other hand, another pose generated from the grid centered on the YAP-TEAD interaction interface formed hydrogen bonds with Met86 and Arg87 of YAP. This result suggests that VP competitively binds to Met86 in place of TEAD residues. Therefore, this binding pose of Ia-2 is considered the mechanism by which the YAP-TEAD interaction is inhibited.

Conclusions:

Ia-2 tends to form more stable hydrogen bonds with Arg and Lys in YAP compared with the other isomers. The combination of carboxyl methyl chirality in Ia-2 creates a stable binding pose with Met86 and Arg87 of YAP. The binding mode suggested by in silico analysis indicates that if VP was to be used as an anti-cancer drug in the future, administering the isolated Ia-2 might be more effective than administering the Visudyne which is a mixture of the four VP isomers.

References:

- (1) Kandoussi I. et al. *Bioinformation*. 2017 Jul 31;13(7):237-240.
- (2) Wei C & Li X. *Mol Med Rep*. 2020 Nov;22(5):3955-3961.

P02-10

Case studies of deep learning-based molecular docking program in medicinal chemistry

Kazuya OSUMI *, Naoya UKEGAWA, Tomohide MASUDA

Pharmaceutical Research Laboratories, Toray Industries, Inc.

(* E-mail: kazuya.osumi.v5@mail.toray)

Molecular docking is a computational procedure in which the non-covalent bonding of molecules, such as a protein receptor and a ligand, is predicted. The scoring function in the docking process is responsible for evaluating the correctness of the pose of the molecule in the binding site and predicting its binding affinity. Several recent efforts, such as GNINA, have demonstrated success in combining 3D grid representations with convolutional neural networks (CNNs). This allows the model to learn its own representation of the protein-ligand interaction in order to determine what constitutes a strong binder. In this study, we report the examination of the following two cases of utilizing GNINA.

1) Retrospective analysis: Classification of the ligand function

In the research on ROR γ t inhibitors, a slight modification of the terminal substituent switched the function of the ligand from an antagonist to an agonist. There have been several reports of similar cases, and the function of the ligand has been explained by X-ray co-crystal structure analysis and MD simulations. Therefore, we examined whether it is possible to identify the ligand function more conveniently by docking simulation. We docked each ligand to inactive and active states of the proteins and compared their docking poses and scores. As a result, it was suggested that when using GNINA, the function of the ligand can be identified by the difference in docking scores for both the inactive and active forms of the protein.

2) Prospective prediction: Selection of compounds expected to have high affinity

In the research on kinase inhibitors targeting the CNS, it was inferred from the initial SAR of the hit compound and the known complex structures that there is a space where the substituent can be extended beyond the terminal amino group of the hit compound. Therefore, we constructed a virtual library of approximately 3,000 compounds derived from the combination of the amino group of the hit compound and diverse set of carboxylic acids. After filtering by druglikeness and CNS-MPO scores, we selected 96 compounds expected to have high affinity by docking simulation using GNINA. As a result, 21 compounds

were obtained with improved inhibitory activity than the hit compound, suggesting that the selection of compounds by GNINA's docking score is effective.

- [1] Francoeur, P. G., *et al.*, *J. Chem. Inf. Model*, **2020**, 60, 4200-4215.
- [2] McNutt, A., *et al.*, *J. Cheminformatics*, **2021**, 13, 43.
- [3] Osumi, K., *et al.*, *The 39th Medicinal Chemistry Symposium*, **2022**.
- [4] René, O., *et al.*, *ACS Med. Chem. Lett.* **2015**, 6, 276–281.
- [5] Yukawa, T., *et al.*, *J. Med. Chem.* **2019**, 62, 1167– 1179.

P02-11

Binding Affinity Prediction Through Unsupervised Learning of Protein-Ligand MD Trajectories

Kodai IGARASHI ^{*}, Masahito OHUE

School of Computing, Institute of Science Tokyo

(^{*} E-mail: igarashi@li.c.titech.ac.jp)

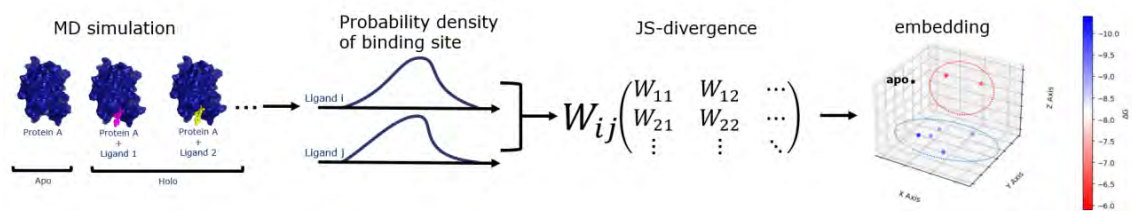
Prediction of binding affinity is an important technique that provides a useful basis for drug and enzyme design. However, prediction based on biochemical experiments and structural analysis requires much time and effort. Therefore, effective drug design is expected to be achieved by predicting binding affinity using bioinformatics.

Molecular dynamics simulation and machine learning have been proposed as common methods, but these methods have problems such as limited computation time and available data sets. There are also methods that use molecular dynamics simulations and deep learning to compare protein dynamics for a set of ligands for a specific protein and predict binding affinity. However, the computational cost of molecular dynamics simulations and deep neural networks is high.

In this study, we propose a method to predict binding affinity for a specific protein by comparing protein dynamics using molecular dynamics simulations and Jensen-Shannon divergence, which reduces the computational cost of deep neural networks.

Specifically, we extract trajectory data of protein-ligand complexes generated by MD simulations and estimate the probability distribution of binding site residue trajectories. The similarity of the probability distributions among different ligands is then compared by Jensen-Shannon divergence. The points reflecting the obtained distance matrix are then projected into a two-dimensional space to represent the dynamics of the different ligands.

The results show that the accuracy of the proposed method is comparable to that of previous studies and reduces the computational cost by deep neural networks. We also investigated the effect of the initial structure used in the simulation on the simulation and on the prediction accuracy of binding affinity for the difference between the crystal structure and the modeling structure.



P02-12

Preprocessing of FMO calculations and practical visualization of interaction energies for drug design

Hirofumi WATANABE *¹, Yuji TANAKAMARU²

¹WithMetis Co., Ltd.

²Malo21st

(* E-mail: h.watanabe314@gmail.com)

Based on quantum mechanics, the fragment molecular orbital (FMO) method provides computational evaluations of protein-ligand interactions through inter-fragment interaction energy (IFIE) analysis and its decomposition analysis (PIEDA).

We have been developing an FMO-based drug design software package called Q-AIR. First, Grinder, one of the modules of Q-AIR, preprocesses FMO calculations. This module has several preprocess functions, such as adding missing atoms in protein and protonation of both protein and ligand, as well as minimization near complemented regions. We will discuss the influence of protonation or complementation and appropriate preprocessing for obtaining a better correlation between activities and interactions. Next, Flair viewer in Q-AIR provides visual analysis between a ligand and surrounding amino acid residues and insights for modifications of compound structures. At the last CBI annual meeting, we reported adding a 2D diagrammatic depiction while keeping information on the 3D positions and distances of amino acid residues surrounding the ligand. This year's update for the viewer is the improvement of the algorithm of 2D-diagrammatic depiction. With this improvement, we can get stable and easy-to-understand depictions of any situation regarding a ligand and its surrounding residues.

P02-13

Prediction of quantum mechanical interactions between the ligand and each amino acid residue in protein-ligand complexes

Ryosuke KITA ^{*1}, **Hiromu MATSUMOTO**¹, **Chiduru WATANABE**², **Daisuke TAKAYA**⁷, **Yu-Shi TIAN**⁷, **Masateru OHTA**³, **Naoki TANIMURA**⁴, **Koji OKUWAKI**⁵, **Mitsunori IkeGUCHI**^{3, 6}, **Kaori FUKUZAWA**⁷, **Teruki HONMA**², **Tsuyohiko FUJIGAYA**¹, **Koichiro KATO**¹

¹Department of Applied Chemistry, Kyushu university

²RIKEN Center for Biosystems Dynamics Research

³RIKEN Center for Computational Science

⁴Mizuho Research & Technologies, Ltd.

⁵Engineering Technology Division, JSOL Corporation

⁶Graduate School of Medical Life Science, Yokohama City University

⁷Graduate School of Pharmaceutical Sciences, Osaka University

(* E-mail: ryosuke092413@gmail.com)

Evaluating protein-ligand interactions through molecular simulations plays a crucial role in efficiently identifying promising drug candidates from millions of compounds in computational drug discovery. Quantum mechanical (QM) calculations have been expected as an ideal method for evaluating protein-ligand interactions due to their high computational accuracy. However, the high computational cost made it difficult to apply to large molecular systems like proteins. To address this challenge, the fragment molecular orbital (FMO) method was developed, enabling the calculation of protein-ligand interaction energies at the QM level. The FMO method not only allows QM calculations of the entire protein but also enables detailed interaction analysis by extracting protein-ligand interaction energies as inter-fragment interaction energies (IFIE). However, even using the supercomputer Fugaku, FMO method requires several hours of computation time per structure, and further reduction of computational cost has been desired. Therefore, we considered it necessary to develop a machine learning model that predicts IFIE without requiring explicit FMO calculations.

This study aims to develop a machine learning model capable of predicting IFIE between ligands and each amino acid residue constituting the protein within seconds per structure using a standard personal computer. We considered the following methods for descriptors used to numerically represent molecular structures and properties, and machine learning algorithms that use these descriptors as input to predict IFIE between each amino acid residue

constituting the protein and the ligand, enabling the model to handle any ligand. For descriptors, we employed Atom-Centered Symmetry Functions (ACSF) [1], a method that can vectorize the surrounding environment of individual ligand constituent atoms. As a machine learning algorithm, we used High-Dimensional Neural Network Potentials (HDNNP) [2], which can learn IFIE as a sum of interactions acting on individual ligand constituent atoms. Although these methods have been developed as methodologies for machine learning force fields, they have not been used for interaction prediction before.

The target protein selected was CDK2. For each of the 80 ligands known to bind to CDK2, 20 binding poses were generated, and descriptors as input were calculated for each amino acid residue (11,9250 points). The model was trained to minimize the error between the IFIE values (true values) calculated using Fugaku and the predicted values. As a result, a trend was observed between the true values and predicted values (Fig. 1), successfully constructing a model capable of rapidly predicting QM interactions between each amino acid residue constituting the protein and the ligand, using 3D atomic coordinates as input.

Reference

- [1] J. Behler, J. Chem. Phys. 134, 074106 (2011).
- [2] J. Behler, M. Parrinello, Phys. Rev. Lett. 98, 146401 (2007).

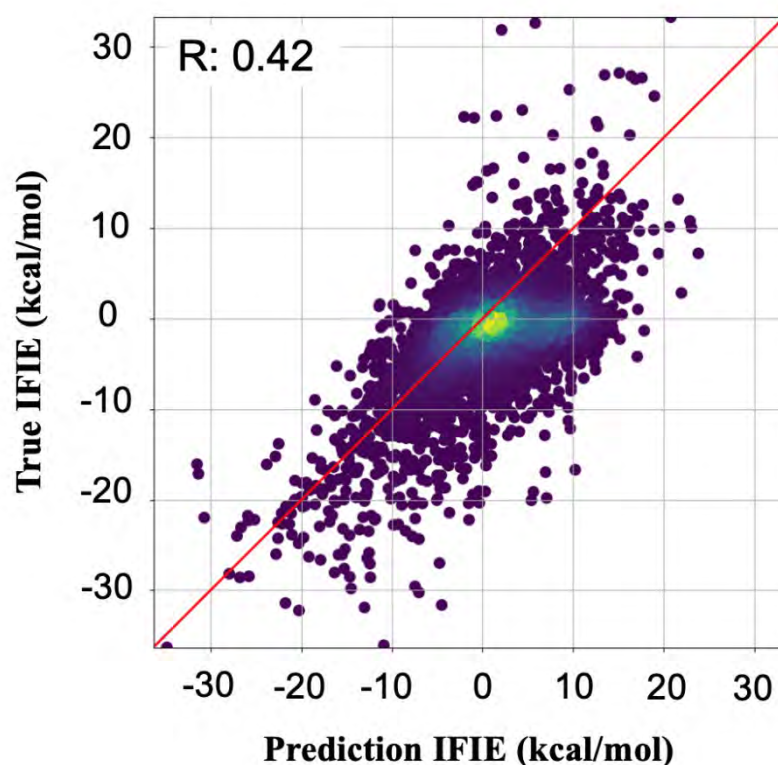


Fig.1. True-Prediction IFIE Plot

P03-01

Disease Prediction from Small Sample Gut Microbiome Data

Daiki SAKAI *, Takuji YAMADA

Yamada Lab, Institute of Science Tokyo, School of Life Science and Technology
(* E-mail: sakai.d.ab@m.titech.ac.jp)

Recent studies have shown that various diseases are associated with the gut microbiome. In particular, diseases such as colorectal cancer and inflammatory bowel disease have been studied extensively and many study cohorts have been published. The construction of machine learning models from these data has also enabled the prediction of diseases from gut microbiome data. In contrast, rare diseases with fewer patients are less studied and have fewer published data. Due to the insufficient data, it has been difficult to apply machine learning to such rare diseases. In such small data situations, a learning method called transfer learning is often applied. Transfer learning is a learning method that aims to improve performance on a target task by using learned knowledge from a similar source task. Although there are studies that have applied transfer learning to diseases prediction from gut microbiome, most of them are validated only on target tasks with sufficient amount of data. Thus, the effectiveness of transfer learning for rare diseases with a small number of patients is not well understood. In this study, we investigated the effectiveness of transfer learning in diseases prediction from gut microbiome data, focusing on situations in which the target task data is extremely small. For model construction and validation, we obtained relative abundance table of multiple study cohorts from curatedMetagenomicData. A transfer learning model was constructed using study cohorts with a large amount of data as source data and a rare disease cohort with a small amount of data as target data. We have compared the performance of transfer learning models with the baseline models. As a result, the performance metric improvement in the transfer learning model was observed for some combinations of source and target diseases.

P03-02

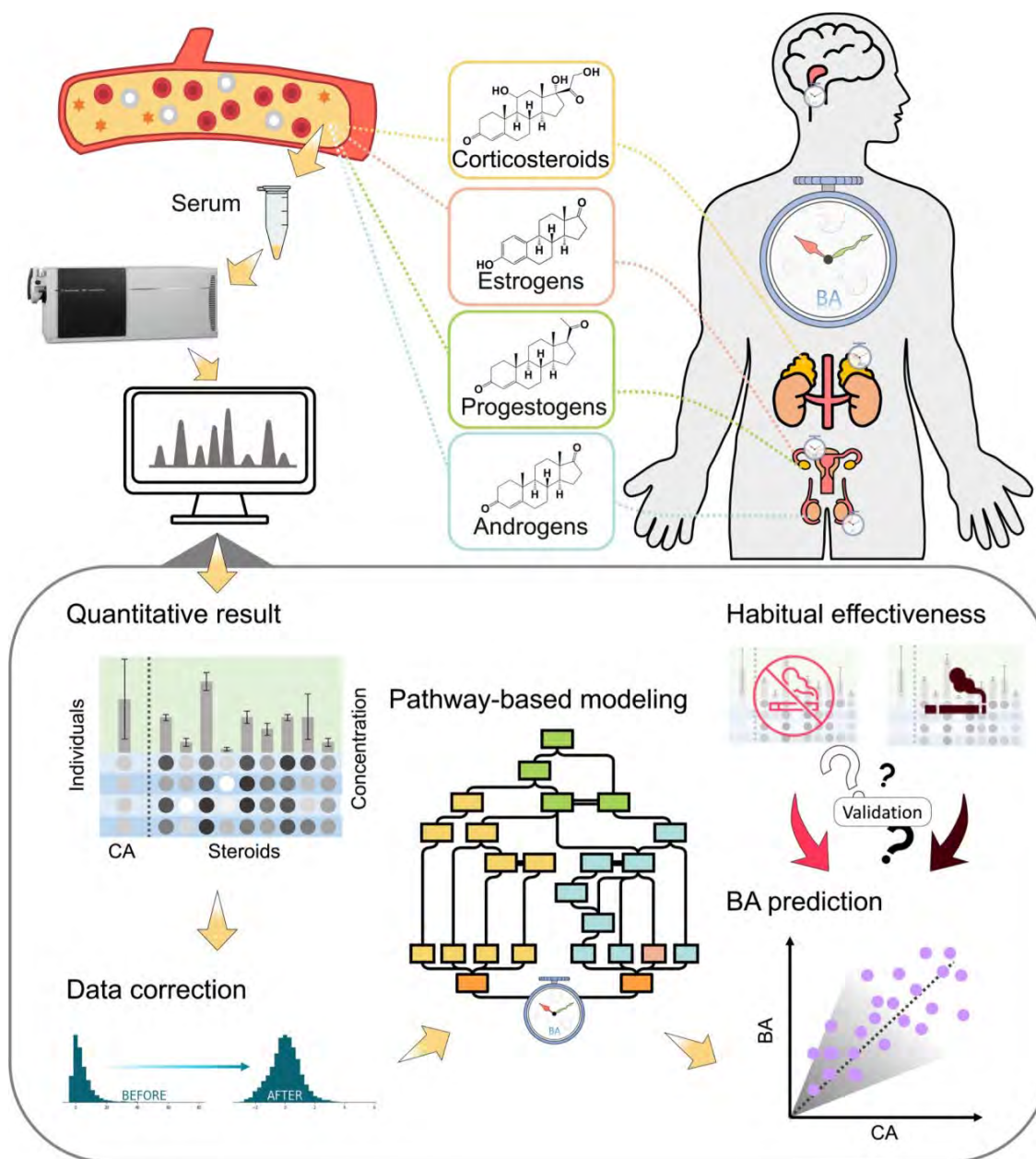
Biological Age Prediction Using a Deep Neural Network Based on Steroid Metabolic Pathways

Zi WANG *, Qiuyi WANG, Kenji MIZUGUCHI, Toshifumi TAKAO

IPR, Osaka University

(* E-mail: tokumunoboushi@yahoo.co.jp)

Aging, a multifaceted process marked by the progressive accumulation of cellular damage, necessitates a deeper understanding of biological aging mechanisms for effective intervention in age-related diseases. Previous research has established that steroid hormones, integral to a comprehensive system of physiological transformations, exhibit significant metabolic interactions. Additionally, stress-related corticosteroid hormones and sex hormones have been found to strongly correlate with the degree of physiological aging. Consequently, utilizing steroid hormone interactions for BA estimation offers a more accurate and comprehensive approach. In this study, we introduce a novel method for BA prediction by developing a DNN model based on steroid metabolic pathways for both sexes. The model was constructed using 22 steroid profiles derived from 98 individual serum samples, analyzed with an in-house Liquid Chromatography-Mass Spectrometry/Mass Spectrometry (LC-MS/MS) method known for its high sensitivity and precision. To ensure the robustness and reliability of our models, we applied a rigorous data correction method to eliminate experimental batch effects and individual physiological variations, and tested the model on an additional 50 serum samples. Our approach provides deeper insights into aging heterogeneity and biological processes by developing a pathway-based DNN model that accounts for the expansion of aging heterogeneity over time and the biological interpretability of molecular interactions. By examining differences in metabolic pathways between sexes, we highlight the distinct effects of stress-related and sexual steroids on BA. Furthermore, our findings suggest that smoking habits may influence the BA of males, primarily through their impact on stress-related steroids. This study offers a more nuanced understanding of the aging process and holds promise for more accurate and comprehensive BA predictions compared to traditional methods.



P03-03

A deep learning model for predicting chemical-induced rat hepatocellular necrosis using transcriptome data.

Kouki MAEBARA ^{*1}, Kyoko ONDO², Tomoaki TOCHITANI², Toru USUI², Izuru MIYAWAKI², Kaori AMBE¹

¹Department of Regulatory Science, Graduate School of Pharmaceutical Sciences, Nagoya City University

²Preclinical Research Unit, Sumitomo Pharma Co., Ltd.

(* E-mail: c202050@ed.nagoya-cu.ac.jp)

Open TG-GATEs is a toxicity database which includes transcriptome data from animal experiments on chemicals [1]. Recently, transcriptome data is expected to be utilized for toxicity evaluation of chemicals including pharmaceuticals. However, since transcriptome data contains a large amount of information, appropriate data pre-processing is required to utilize it. The DeepInsight method is capable of converting high-dimensional data into images [2], and it could be used for processing transcriptome data. In this study, we tried to construct a deep learning model to predict chemicals that cause hepatocellular necrosis in rats, an important histopathological finding when evaluating hepatotoxicity, from imaged liver transcriptome data.

This study used animal study data and transcriptome data published on Open TG-GATEs. Chemicals that showed histopathological findings related to hepatocellular necrosis in a 28-day repeated-dose rat study were designated as positive, and chemicals that did not show hepatocellular necrosis were designated as negative. The expression data of 31,099 genes observed in the liver of rats in the single-dose study of these chemicals were imaged using the DeepInsight method and used as explanatory variables in the prediction model. A pre-trained convolutional neural network (CNN) model was constructed to determine whether or not each compound induced hepatocellular necrosis in rats.

For model construction, 127 chemicals (25 positive, 102 negative) were randomly split 4:1 into training data and validation data, and their prediction performance was evaluated using the hold-out method. After five trials with different splitting patterns in the validation data, the mean and standard deviation of the evaluation indices ROC-AUC, f1 score, sensitivity, and specificity were 0.752 (0.052), 0.531 (0.069), 0.800 (0.112), and 0.705 (0.084), respectively. These results suggest that the liver transcriptome data obtained from a single-dose study in rats might predict hepatocellular necrosis induced

in a 28-day repeated-dose study.

[1] <https://dbarchive.biosciencedbc.jp/en/open-tggates/desc.html>

[2] Sharma, et al., Sci Rep., 9, 11399 (2019).

P03-04

Reaction-Aware Molecular Optimization Using Conditional Transformer and Reinforcement Learning

Shogo NAKAMURA ^{*1}, **Nobuaki YASUO** ², **Masakazu SEKIJIMA** ³

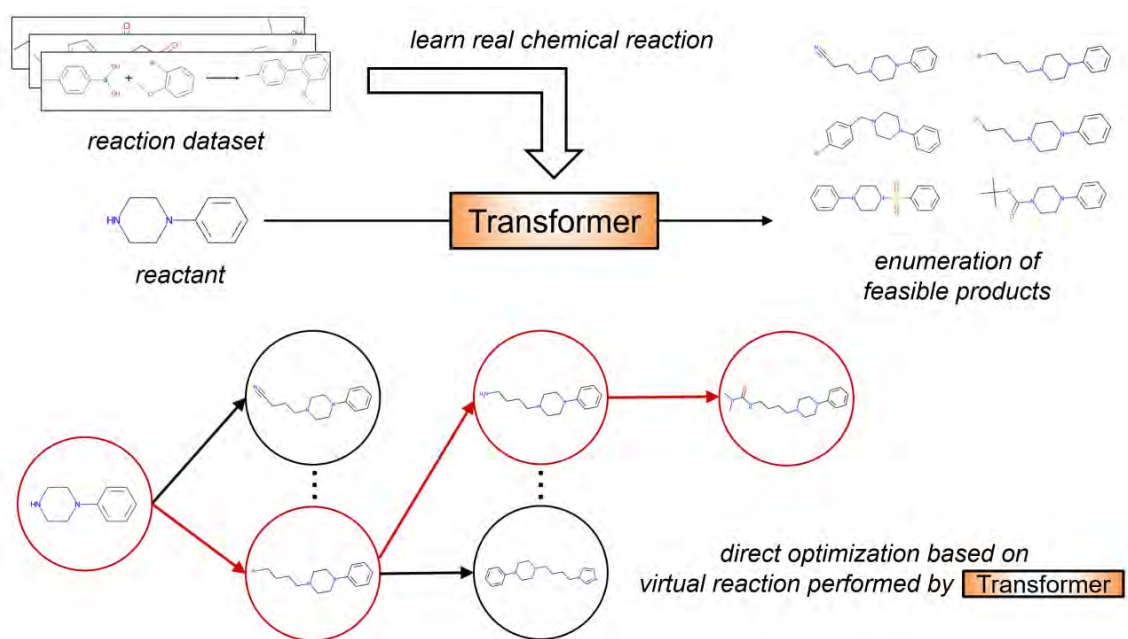
¹Department of Life Science and Technology, Tokyo Institute of Technology

²Academy for Convergence of Materials and Informatics (TAC-MI), Tokyo Institute of Technology

³Department of Computer Science, Tokyo Institute of Technology

(* E-mail: nakamura.s.bu@m.titech.ac.jp)

Designing molecules with desirable properties is an important research issue in drug discovery. Recent advances in deep learning have led to the development of molecular generation models to obtain compounds with desired properties. However, existing compound discovery models often ignore the key issue of ensuring the feasibility of organic synthesis. To address this issue, we propose TRACER (molecular optimization using a conditional Transformer for reaction-aware compound exploration with reinforcement learning). The core of TRACER is a Conditional Transformer model trained on a dataset of chemical reactions. By explicitly training on chemical reactions, the model can predict realistic products from a given reactant under the constraints of the reaction type specified by the graph convolutional network. Molecular optimization results in the activity prediction model targeting the dopamine receptor D2 showed that TRACER effectively generated compounds that scored highly. The structure-wide Transformer model captures the complexity of organic synthesis and allows exploration within the vast chemical space while accounting for real-world reactivity constraints. The source code, activity prediction models, and curated datasets are available in our public repository (available at <https://github.com/sekiyima-lab/TRACER>).



P03-05

Computational determination of SMARTS molecular query containment relationships

Seiji MATSUOKA *¹, Minoru YOSHIDA^{1, 2, 3}

¹Center for Sustainable Resource Science, RIKEN

²Office of University Professors, The University of Tokyo

³Collaborative Research Institute for Innovative Microbiology, The University of Tokyo

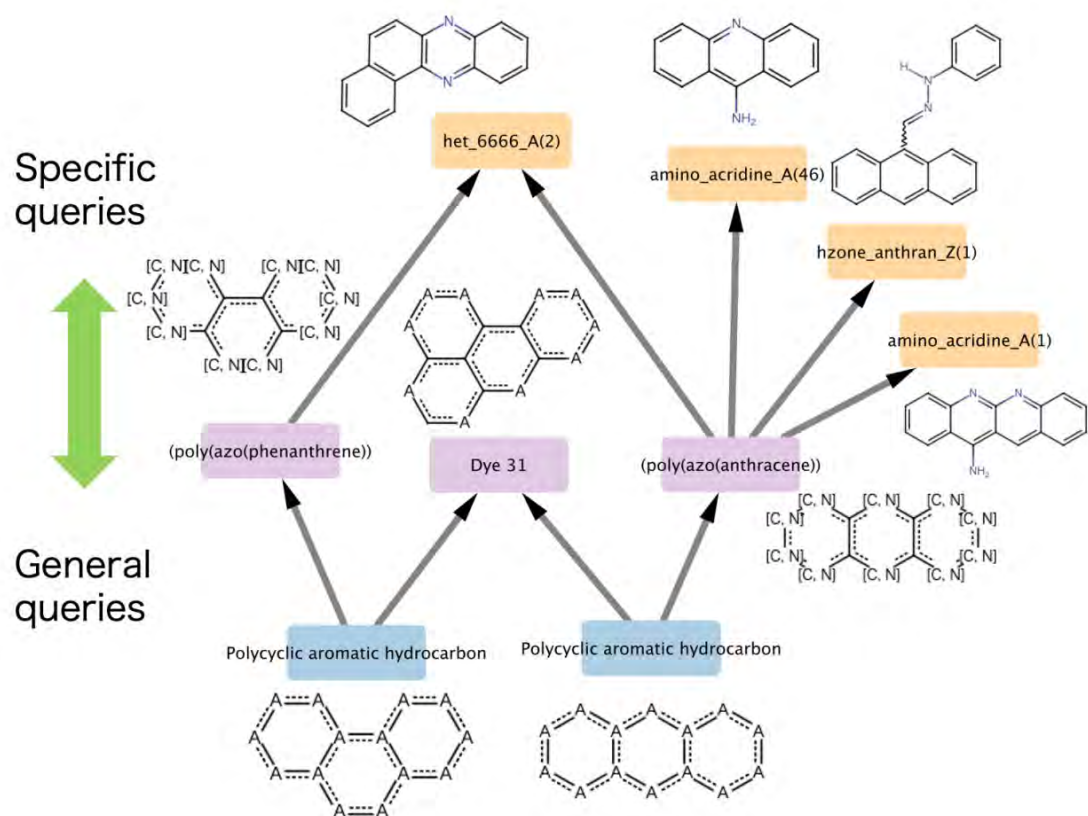
(* E-mail: smatsuoka@riken.jp)

Describing molecules as substructure patterns is convenient for focusing on particular substructures and functional groups that are responsible for a distinctive function of the molecules. For example, patent claims often cover a group of compounds that meet certain criteria, expressed as substructure patterns called Markush structures. A database of substructure patterns in a machine-readable format is promising for chemical information systems, such as those used for patent or regulatory chemical searches. However, this involves significant challenges. To search substructure patterns using a substructure pattern query, it is necessary not only to match substructures but also to determine the containment relationships of each atom/bond attributes, including logical expressions.

We have developed an algorithm for determining containment relationships of SMARTS, a commonly used molecular query language. This algorithm consists of a substructure matching based on the VF2 algorithm and containment relationship determination of atom/bond attributes by using truth tables that can handle logical operators and recursive expressions in SMARTS. This is implemented as part of MolecularGraph.jl, a molecular graph modeling toolkit, and is available as open-source software.

This method also allows a substructure pattern dataset to be represented as a directed acyclic graph (DAG) by obtaining containment relationships for all combinations of substructure patterns. This capability was demonstrated through the systematic visualization of relationships in ChEMBL structural alerts dataset used in medicinal chemistry. Additionally, a dataset of frequently occurring functional groups and substructures was constructed to generate containment relationships in a DAG. This enables the functional annotation of molecules as a subgraph of that DAG. Such an approach can potentially be

applied to the development of molecular descriptors that capture features of molecular functions.



P03-06

Development of New data analysis platform for medicinal chemist in Daiichi Sankyo

Takayuki SERIZAWA *, Kosuke TAKEUCHI, Kosuke MINAGAWA, Kan SHIRAISHI, Shunya MAKINO

Group 1, Modality Research Laboratories, Daiichi Sankyo Co., Ltd.

(* E-mail: takayuki.serizawa@daiichisankyo.com)

From 2018, Daiichi Sankyo has built new cheminformatics team and designed Data Driven Drug Discovery (D4) group engaged in close interactions with other researchers participating in the Design-Make-Test-Analyze (DMTA) cycle[1] such as medicinal chemists, CADD team, and pharmacologists

From our D4 contribution analysis we found that SAR visualization/analysis is one of the important D4 activities, because medicinal chemists need to analyze SAR data and decide what to make next with their hypothesis.

However , there are few tools that can offer flexible SAR visualization/analysis environments for users and supporting idea sharing among them.

In order to overcome the issue, we built new data analysis platform with Datagrok[2]. The platform can quickly visualize complex SAR data and run lots of cheminformatics tasks such as clustering molecules, MMP analysis and applying QSAR prediction. Besides, medicinal chemists can share their design idea and progress of the design in the same platform, so that they can prioritize their ideas before starting synthesis. From technical point of view, the platform can offer flexible development environment with multiple programming languages such as python, R, and JavaScript for cheminformatics team.

Here we would like to present our internal effort to new SAR analysis platform with Datagrok[2].

1. Drug Discov. Today, 27 (8) (2022), pp. 2065-2070
2. <https://datagrok.ai/>

P03-07

Data augmentation method of chimeric protein sequences for fine-tuning of protein language models

Kei YOSHIDA *, Shoji HISADA, Atsushi OKUMA, Taketo KAWARA, Takuya YAMASHITA, Yoshihito ISHIDA, Daisuke ITO, Hiroko HANZAWA, Shizu TAKEDA

Research & Development Group, Hitachi, Ltd.

(* E-mail: kei.yoshida.qp@hitachi.com)

Chimeric antigen receptor-T (CAR-T) cell therapy has produced remarkable clinical responses, especially in cancer treatments. CAR-T cells can recognize and bind specific antigens on the surface of target cells via chimeric antigen receptors (CARs), and finally kill target cells.

CARs are chimeric proteins generated by fusion of functional protein domains which recognize target antigens or perform intracellular signal transduction. Recently, some research groups are trying to optimize CARs to enhance the therapeutic effect of the therapy with computational tools. There are two steps for optimization. First, learning the relationship between protein sequence representations (numerical encodings) and its cellular functions, and second designing new sequences based on the learning results.

Protein language models (PLMs) trained on a huge number of protein sequences are attracting attention as an approach to summarize the protein sequences into representations. In the case of target protein engineering tasks, PLMs fine-tuned on task-specific data have become a powerful tool for getting representations which is useful for state-of-the-art predictive methods. However, there are no PLMs which produce optimal representations from CARs because the sufficient volume of sequence data needed for fine-tuning of PLMs is not available.

In this study, we propose a data augmentation method to increase CAR sequences for efficient fine-tuning of PLMs. In the method, we utilize homologous sequences in each CAR domain and combine them to generate target CAR-like sequence. We fine-tuned ESM-2 using the method and predicted the cytotoxic activity of hundreds of anti-CD19 CAR (28/28/28z) mutants. The results indicate that fine-tuned ESM-2 achieved better predictive performance than the original ESM-2. We conclude that our method will be a reliable approach for representing CAR sequences and lead to designing CARs more efficiently.

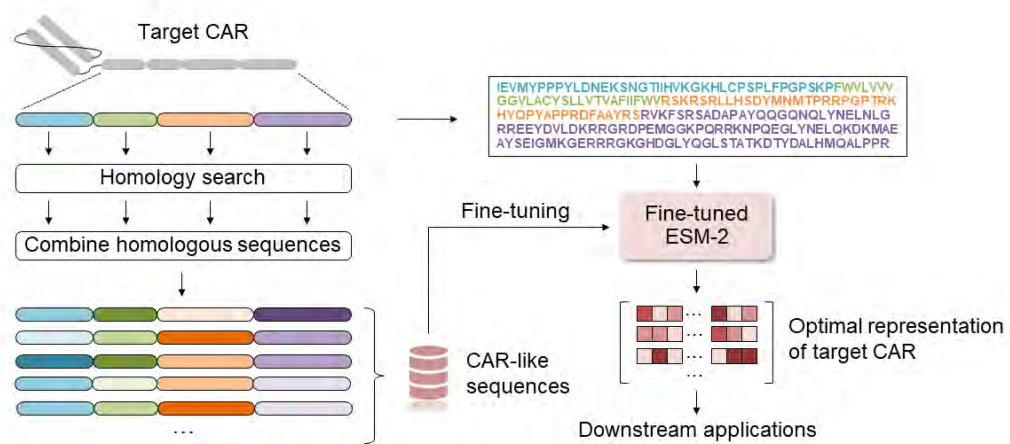


Figure | Representation of chimeric antigen receptor (CAR) sequences with fine-tuned EMS-2

P03-08

Predicting Chemical Roles Using Natural Language Processing on Database Descriptions

Yuya KOIDE *¹, Yuto MATSUMOTO², Hiroaki GOTOH²

¹College of Engineering Science, Yokohama National University

²Graduate School of Engineering, Yokohama National University

(* E-mail: koide-yuya-mc@ynu.jp)

With the development of natural language processing technology, it is possible to efficiently utilize the enormous amount of document data that experimental data and other data aggregate, and to predict the properties of compounds directly from the document data.

We clustered the embedded representations of compounds obtained by analyzing 1744 papers with the Word2Vec model and confirmed the structural features in each cluster. Furthermore, we identified the distribution of chemical descriptors in each cluster.

In this study, we further developed this previous study and introduced an approach to predict the presence or absence of various chemical roles of compounds by analyzing the descriptive text of compounds using natural language processing, focusing on databases containing aggregated article information.

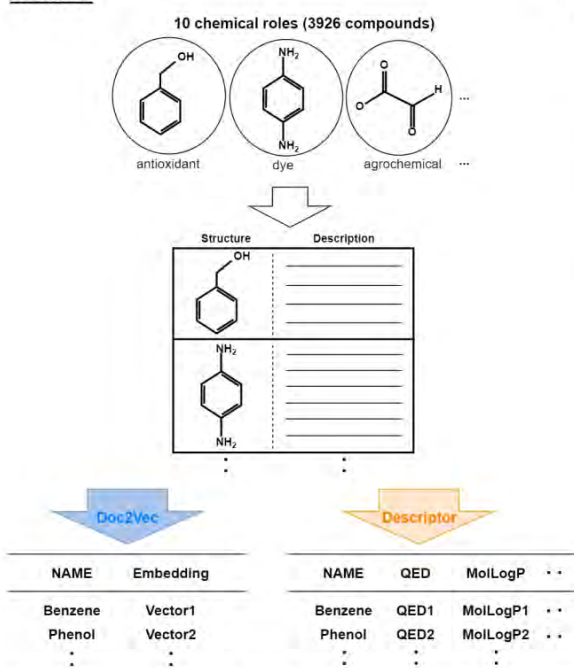
In this study, from a total of 41891 compounds registered as 3stars in ChEBI, a database of chemical compounds, 10 chemical roles (antioxidant, anti-inflammatory agent, allergen, dye, toxin, flavoring agent, agrochemical, volatile oil, antibacterial agent, insecticide) were extracted to a dataset of 3926 compounds. Then, embeddings of the compounds were obtained by natural language processing of the descriptions of those compounds obtained from PubChemAPI using the Doc2Vec model. Then, the presence or absence of each chemical role of the compounds was predicted by logistic regression using the obtained embeddings as input, and these results were compared with the F1 scores predicted from the 32 chemical descriptors. These 32 chemical descriptors were obtained from 208 chemical descriptors obtained by RDKit, extracting only those that were continuous values and those with a correlation coefficient of 0.95 or less between each descriptor.

The obtained embeddings of the compounds were found to be distributed for each chemical role. Furthermore, the prediction of the presence or absence of each chemical role by the embeddings was better than that by the chemical descriptors for many tasks. On the other hand, the prediction accuracy of dyes

and volatile oils by this method was comparable to that by chemical descriptors, suggesting that compounds belonging to dyes and volatile oils may have similar structural characteristics. In addition, the prediction accuracy of toxin was low both by this method and by chemical descriptors, suggesting that compounds belonging to toxin include compounds with diverse structures and characteristics. Based on these results, we are considering what kind of substructure influences the presence or absence of each chemical role.

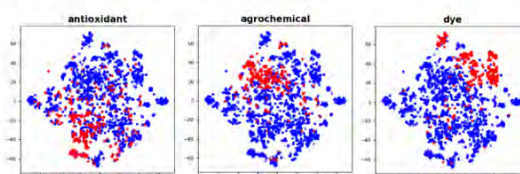
By analyzing the explanatory text of a group of compounds using Doc2Vec, it is possible to comprehensively analyze and understand the compounds, including their characteristics and behaviors that cannot be captured by their structures and physical properties. The embeddings obtained from the analysis can also be used as new descriptors that utilize the properties of the compounds.

Method

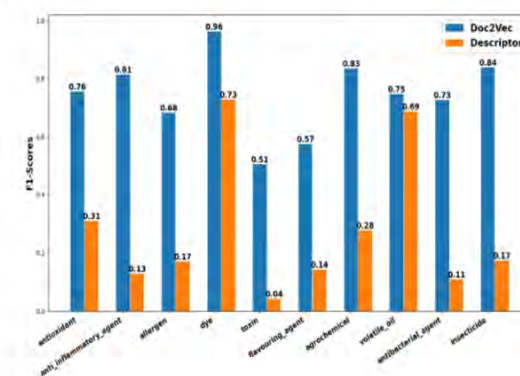


Result

Visualization



Prediction



P03-09

Large-scale single nucleus RNA-seq analysis of Lewy body diseases subtypes

Supakorn PONGPAKDEE *, Kosuke HASHIMOTO, Kenji MIZUGUCHI

Osaka University

(* E-mail: u734967b@ecs.osaka-u.ac.jp)

Single cell/nucleus data analysis become advancing to incorporate with biomarker discovery in neurodegenerative diseases, particularly on Lewy body diseases (LBDs) that compose of Parkinson's disease (PD), Dementia with Lewy body (DLB) and Parkinson's disease dementia (PDD). While these diseases share a common neuropathological hallmark overlapping in cortical and subcortical brain area by aggregated α -synuclein protein, called Lewy body (LB), and associate with the prevalence of dementia in Alzheimer's disease (AD), the clinicopathological progressive trajectory of LB between brainstem and olfactory bulb tract initially represents differently among LBDs heterogeneity, suggesting diverse LB pathology distributions. Although several studies of transcriptomic profiling for LBDs have highlighted potential molecular therapeutic targets, the underlying mechanisms in specific cell types to distinguish LBDs remain elusive. This study aims to identify distinct genes marker for LBDs subtypes using large-scale transcriptomic single nucleus RNA-seq (snRNA-seq) data analysis. Postmortem snRNA-seq data derived from substantia nigra (SN) and midbrain was retrieved from eight public datasets, including unaffected control (CTL), PD, PDD and DLB samples (n=66, 55, 19 and 4, respectively). Two of eight datasets were used as reference to transfer cell-type annotations, and all datasets were integrated into a single matrix using previous probabilistic deep learning model (scVI). Overall, we obtained a single expression matrix containing approximately 800,000 cells with 18,632 genes. These cells were classified into seven major cell types: neuron, oligodendrocyte, oligodendrocyte progenitor cell, astrocyte, dopaminergic neurons, microglia and endothelial cell. The differential expression (DE) and gene ontology (GO) enrichment analysis revealed significant gene sets associated with mitochondrial function, extracellular organization and membrane protein co-translation, between disease conditions. Due to limited data availability, additional collection is necessary to enhance the representation of various subtypes in the dataset.

P03-10

SAR analysis and visualization utilizing a fragment-based approach: Application to a public data analysis of Targeted Protein Degradator

Hiroyuki HAKAMATA *, Kosuke TAKEUCHI, Toshiaki WATANABE

Modality Research Laboratories I, R&D Division, Daiichi Sankyo Co., Ltd.

(* E-mail: hiroyuki.hakamata@daiichisankyo.com)

Structure-Activity Relationship (SAR) represents chemical landscape between chemical structures and their biological activities, which plays a central role in medicinal chemistry. Visualization of various information such as potency, ADMET and in vivo experiment data often provides medicinal chemists with new insights. In small molecule drug discovery, several visualization techniques such as R-group decomposition and followed by network analysis have been established to date [1]. Using these techniques effectively enables medicinal chemists to come up with next compound designs to be synthesized, which can lead to an acceleration in lead optimization campaign.

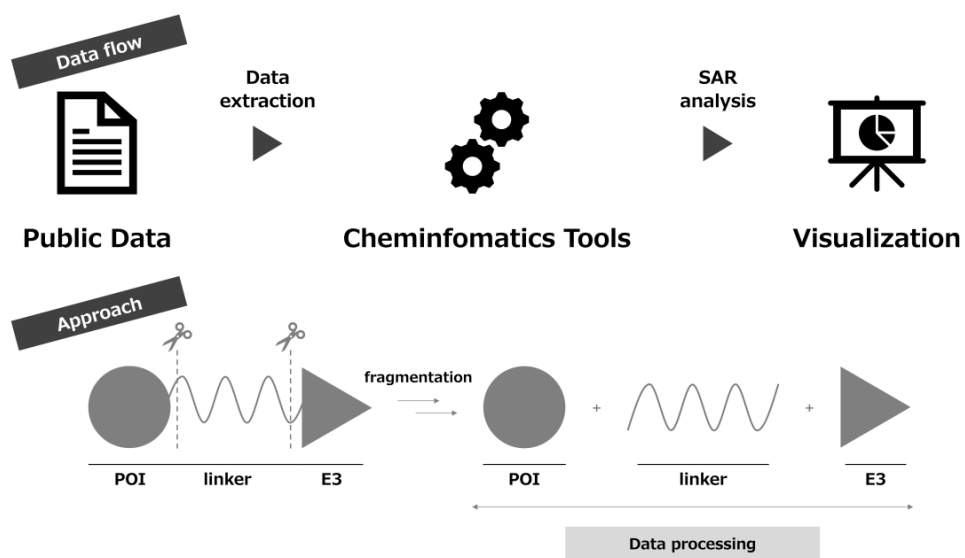
Targeted Protein Degradator (TPD) is a middle-sized molecule composed of three fragments: Protein of Interest (POI) ligand, linker, and E3 ligand. With its degradation MOA instead of inhibition demonstrated by traditional small molecules, it is highly expected to be one of promising modalities. However, due to generally beyond rule of 5 chemical-space, there are numerous options on chemical modification to aim at sweet spot between activity and ADMET property. Besides, its unique MOA makes comprehensive data analysis complicated.

Since starting the examination of "Data-Driven Drug Discovery" (D4) in 2018, we have been leveraging cheminformatics methods such as the development of SAR tables and extracting information from public data, leading to improvement of operational efficiency in multiple drug discovery projects [2].

In this presentation, we will introduce the SAR analysis method using public data on TPD as a case study and discuss workflow for data extraction, molecular fragmentation, assigning physicochemical property values, and visualization.

References

- [1] Dagmar Stumpfe, Jürgen Bajorath, Recent developments in SAR visualization, *Med. Chem. Commun.*, 2016, 7, 1045-1055.
- [2] Ryo Kunimoto, Jürgen Bajorath, Kazumasa Aoki, From traditional to data-driven medicinal chemistry: A case study, *Drug Discovery Today*, 2022, 27, 2065-2070.



P03-11

CycPeptMP: Development of Membrane Permeability Prediction Model of Cyclic Peptides with Multi-Level Molecular Features and Data Augmentation

Jianan LI ^{*1}, Keisuke YANAGISAWA^{1, 2}, Yutaka AKIYAMA^{1, 2}

¹Department of Computer Science, School of Computing, Institute of Science Tokyo

²Middle-Molecule IT-based Drug Discovery Laboratory (MIDL), Institute of Science Tokyo

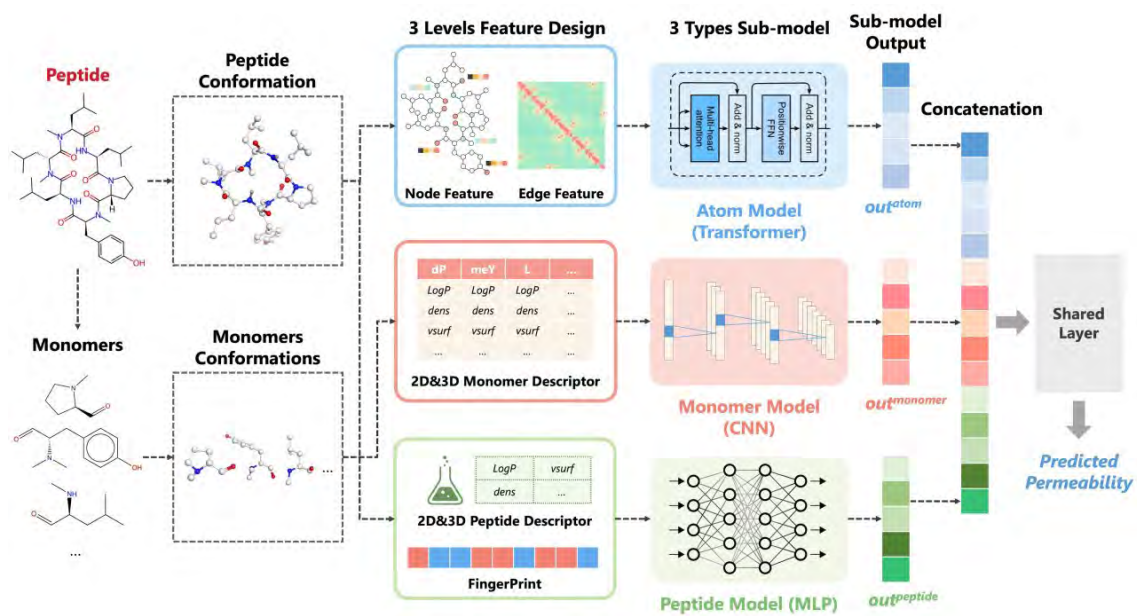
(* E-mail: li@bi.c.titech.ac.jp)

Cyclic peptides are versatile therapeutic agents that boast high binding affinity, minimal toxicity, and the potential to engage challenging protein targets. However, the pharmaceutical utility of cyclic peptides is limited by their low membrane permeability, an essential indicator of oral bioavailability and intracellular targeting. Current machine learning-based models of cyclic peptide permeability show variable performance owing to the limitations of experimental data. Furthermore, these methods only use features derived from the whole molecule that have traditionally been used to predict small molecules and ignore the unique structural properties of cyclic peptides.

We developed CycPeptMP: an accurate and efficient method to predict cyclic peptide membrane permeability [1]. We designed features for cyclic peptides at the atom-, monomer-, and peptide-levels and seamlessly integrated these into a fusion model using deep learning technology. Additionally, we applied various data augmentation techniques to enhance model training efficiency using data from CycPeptMPDB [2], the latest database we constructed. The fusion model exhibited excellent prediction performance for the logarithm of permeability, with a mean absolute error of 0.355 and correlation coefficient of 0.883. Ablation studies demonstrated that all feature levels contributed and were relatively essential to predicting membrane permeability, confirming the effectiveness of augmentation to improve prediction accuracy.

[1] Li J, Yanagisawa K, and Akiyama Y. CycPeptMP: Enhancing Membrane Permeability Prediction of Cyclic Peptides with Multi-Level Molecular Features and Data Augmentation. *Brief Bioinform*, 2024, accepted.

[2] Li J, Yanagisawa K, Sugita M, Fujie T, Ohue M, and Akiyama Y. CycPeptMPDB: A Comprehensive Database of Membrane Permeability of Cyclic Peptides, *J Chem Inf Model*, 63(7): 2240–2250, 2023.



P03-12

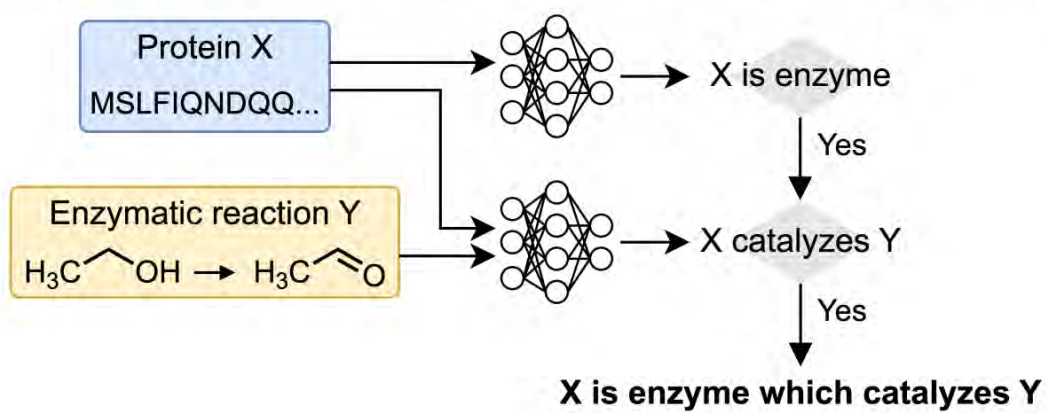
Deep learning-based enzyme screening to identify orphan enzyme genes

Keisuke HIROTA *, Takuji YAMADA

School of Life Science and Technology, Institute of Science Tokyo
(* E-mail: hirota.k.221@gmail.com)

Progress in sequencing technology has yielded large numbers of protein sequences, including enzymes which are important proteins that catalyze specific chemical reactions. However, many proteins of unknown function are encoded in genomes from bacteria to humans, and many functions of living organisms remain unsolved. As for enzymes which are important proteins that catalyze specific chemical reactions, many enzymes are without sequence information in databases. Those enzymes whose activities are known but whose sequences are unknown are called orphan enzymes. This gap between known proteins and enzyme reactions suggests that some proteins of unknown function might be orphan enzymes. However, most existing tools to predict enzymatic function rely on EC numbers and have limited applicability to orphan enzymes, which are often not sufficiently annotated with EC numbers. Moreover, previous studies that assign protein sequences to enzymatic reactions based on reaction similarity cannot handle proteins of unknown function. Therefore, a computational tool is needed to evaluate the correspondence between any enzyme reaction and protein sequence before costly experimental validation. In this study, we propose a deep learning-based framework for enzyme screening that takes protein sequences and enzyme reactions as inputs and evaluates their correspondence. The proposed method differs from previous studies in two aspects. First, our approach performs a two-step evaluation: classification of enzymes and non-enzymes to handle any given protein including non-enzyme proteins, and prediction of their correspondence with enzymatic reactions of interest. Second, the proposed method embeds enzymatic reactions into vector representations, enabling it to handle reactions that cannot be annotated with enzymatic reaction labels, such as EC numbers. This study then aims to establish a computational tool to evaluate the correspondence between any given enzymatic reaction and protein sequence and to enable a fast and large-scale search for orphan enzyme candidates.

Integrated evaluation of multiple deep learning model predictions



P03-13

Data-driven design of visible-light photoswitches using structural features

Said BYADI *, Pavel SIDOROV

Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University

(* E-mail: saidbyadi@icredd.hokudai.ac.jp)

In the current study, we present a exhaustive computational approach to predict two properties, λ_{max} (the wavelength with maximum light absorption) and $t_{1/2}$ (the thermal half-life of a metastable photoisomer), of visible-light photoswitches by using quantitative structure-property relationship modeling (QSPR). Photoswitches, which undergo reversible changes in structure and properties when exposed to light, have important applications in materials science and biology. Traditional methods for predicting these properties rely on time-consuming density functional theory (DFT) calculations, which led to the need for more efficient computational techniques.

To address this, we developed machine learning (ML) models leveraging a robust dataset of azobenzenes and azoheteroarenes collected from literature sources, comprises 798 unique compounds with measured absorption maxima and 134 compounds with measured half-lives. The ML models utilize structural descriptors (including CircuS fragments, Morgan fingerprints, and other structural and topological parameters) derived directly from 2D representations of the compounds, allowing for faster modeling processes. We successfully conducted a rigorous benchmark investigation to identify the most relevant structural descriptors for predicting λ_{max} and $t_{1/2}$. To build and validate our models we used the Descriptors and Optimization tools (DOTools) platform as a powerful Python library for calculation of chemical descriptors and hyperparameters optimization of three methods SVM Random forest and XGboost [*]. Our selected descriptors incorporated molecular fingerprints and fragment counts, which were used for models' training. The best-performing model was validated by repeated 10-fold cross-validation and demonstrated similar predictive precision to density functional theory (DFT) calculations, but with significant reduction of inference time. The machine learning method employed in this study was Support Vector Machines (SVM), chosen for its ability to handle smaller datasets with high accuracy. The best predictive accuracy for the absorption maximum (λ_{max}) was achieved using models based on CircuS

fragments.

Our study demonstrates the potential of QSPR modeling in predicting the key properties of photoswitches with a good precision. This advancement exhibits a significant step toward fast and efficient design of functional materials, with new implications on diverse scientific and technological applications.

[*]: ChemRxiv: <https://chemrxiv.org/engage/chemrxiv/article-details/6694790901103d79c508aaea>. DOI 10.26434/chemrxiv-2024-23v3c.

P03-14

Recent Developments of FMO DB in 2024: Efforts Towards Utilization of FMO data

Kikuko KAMISAKA ^{*1}, Chiduru WATANABE¹, Daisuke TAKAYA², Teruki HONMA¹

¹Center for Biosystems Dynamics Research, RIKEN

²Graduate School of Pharmaceutical Sciences, Osaka University

(* E-mail: kikuko.kamisaka@riken.jp)

Our group has been developing FMO DB (<https://drugdesign.riken.jp/FMO DB/>)[1], which has centrally managed the results of calculations using the fragment molecular orbital (FMO) method since 2017. FMO data and user numbers have steadily increased.

The FMO method, a type of quantum chemical calculation, is a technique that precisely analyzes the behavior of electrons and, through Pair Interaction Energy Decomposition Analysis (PIEDA), accurately handles dispersion forces, such as CH- π , π - π , and Cation- π interactions, enabling high-precision predictions of intermolecular and intramolecular interaction energies. FMO DB provides researchers with the analysis results of this FMO method, creating an environment in which they can quickly access reliable data without having to perform calculations themselves. This improves research efficiency and enables more accurate interaction analysis.

Since 2024, FMO DB has started collaborating with PDBj (<https://pdbj.org/>), which provides 3D structure information data on biopolymers obtained by structural biology experiments. Mutual links between entries registered in FMO DB and PDBj have been established, thereby allowing researchers to access 3D structural information of biopolymers and detailed interaction energies based on those structures. An environment is being established for the mutual use of both databases. In addition, the Web API, which had been developed as a beta version, has been made public, allowing researchers and developers to access FMO DB's calculation data through the API and use them in their research. This mutual collaboration and API release will promote the utilization of FMO data in structural biology and drug discovery research[2] enabling effective approaches to elucidate and deepen our understanding of molecular functions through FMO-based interaction analysis. Meanwhile, the BioStation Viewer software, which can visualize and analyze FMO calculation results, has added a function to directly load data from FMO DB, making it even easier to use FMO calculation data. This presentation introduces the current status and recent efforts of

FMODB as outlined above.

Acknowledgment

The authors thank Prof. Genji Kurisu and Dr. Gert-Jan Bekker of Osaka University for their support in collaborating with PDBj, as well as Dr. Kazumi Tsuda, Dr. Shu Koyama of Science & Technology Systems, Inc., and Mr. Akifumi Kato of Scorpion Tech LLC for technical support. This research was done as part of activities of the FMO Drug Design Consortium (<https://fmodd.jp/top-en/>) and was partially supported by Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from AMED under Grant Number JP24ama121030. FMO calculations were performed using Fugaku (project IDs: hp240114 and ra000017).

[1] Takaya, D. et al., J. Chem. Inf. Model. 2021, 61, 2, 777–794.

[2] Watanabe, C. et al., J. Phys. Chem. Lett. 2023, 14, 15, 3609–3620.

P03-15

Development of the data management system to acquire the strategic data for AI

Miwa SATO *¹, Mari OHTA¹, Shion HOSODA¹, Akira KIMURA², Takahiro MIMORI², Michiaki HAMADA², Daisuke KIGA², Kazuhide AIKOH¹, Miaomei LEI¹, Tanabe MAIKO¹, Ito KIYOTO¹, Akihiko KANDORI¹

¹Center for Exploratory Research, Research and Development Group, Hitachi, Ltd.

²Department of Electrical Engineering and Bioscience, Faculty of Science and Engineering, Waseda university

(* E-mail: miwa.sato.jr@hitachi.com)

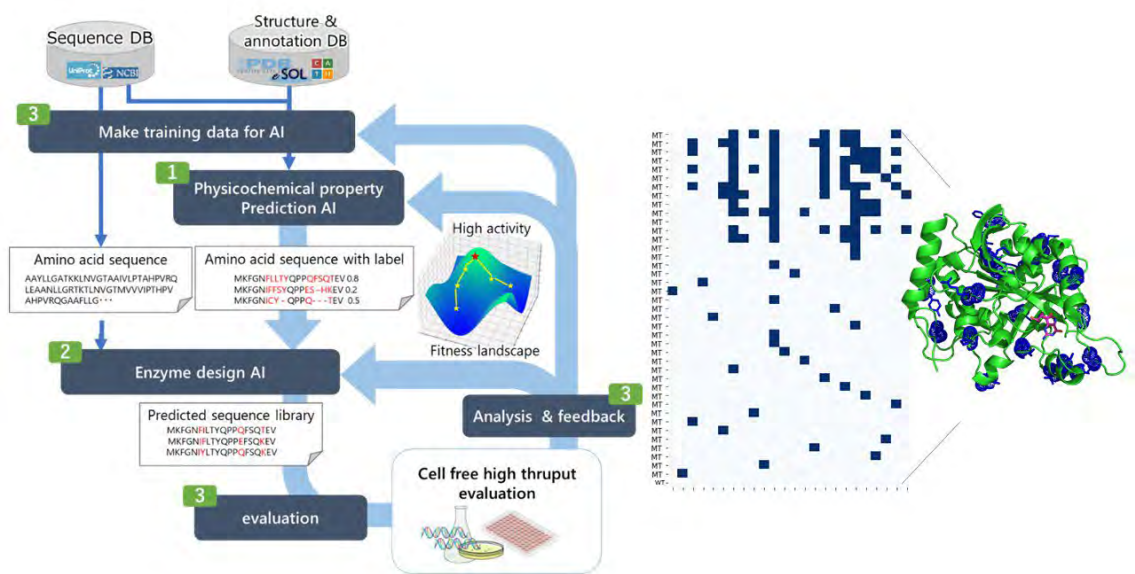
Greenhouse gases are one of the main causes of climate change, and reducing their emissions is a global challenge. To achieve carbon neutrality, we are working on the development of substance production technologies that utilize biological functions. In the development of enzymes responsible for substance production, it is necessary to design their amino acid sequences appropriately. Artificial modification requires many trials due to the complexity of biological functions, so the DBTL cycle is repeated: design the enzyme sequences (Design), build the actual enzyme (Build), evaluate the desired function (Test), interpret the results (Learn), and feed back to the enzyme design again. The DBTL cycle is repeated. Since there are countless combinations of amino acids and the search space is enormous, AI is expected to speed up the process.

However, there are challenges in utilizing AI, such as uniform interpretation of data and information, as well as collection and organization of the large amount of training data required for AI development. In addition, there is still a gap between the number of sequences predicted by the generative AI and the number of sequences that can actually be experimentally verified, making it difficult to experiment with all the sequences predicted by the AI. To effectively proceed with the Build/Test phase of the DBTL cycle, it is necessary to evaluate and select sequences that will be beneficial for AI training.

To solve these data issues, we constructed the data management system that strategically acquires the necessary data for the AI to realize an efficient enzyme development cycle (DBTL cycle) through sequence design by the AI and experimental validation. Issues for improving AI performance were identified and addressed, by building a prototype, we obtained prospects for enzyme improvement through the linkage of AI and wet experiments.

This research is based on results obtained from a project JPNP14004

commissioned by the New Energy and Industrial Technology Development Organization (NEDO).



P03-16

Enhancing Biological Insights with TargetMine: Integration of Genomic Region Annotations

Yi-An CHEN ^{*1}, Lokesh Pati TRIPATHI^{1, 2}, Kenji MIZUGUCHI^{1, 3}

¹Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition

²Laboratory for Transcriptome Technology, RIKEN Center for Integrative Medical Sciences

³Institute for Protein Research, Osaka University

(* E-mail: chenyan@nibiohn.go.jp)

Integrating diverse biological data into comprehensive, easily accessible platforms is crucial for advancing research and enabling discoveries. TargetMine is a robust data warehouse that has been meticulously developed and continuously updated over the past decade to support the evolving needs of the biological research community. It integrates a vast collection of data types, facilitating seamless access and analysis for researchers.

Recent enhancements to TargetMine include the incorporation of detailed genomic region annotations. These annotations encompass exons, introns, promoters, and enhancers, providing researchers with critical insights into the genome's regulatory elements and structural components. By integrating these genomic annotations, TargetMine significantly enhances the capacity for genomic studies, enabling researchers to explore complex biological questions with greater precision.

TargetMine's sustained commitment to data integration and regular updates ensures that it remains a vital resource for the research community. Our platform consolidates data from various sources and provides sophisticated data mining and analysis tools, empowering researchers to generate new hypotheses and drive scientific innovation.

TargetMine is freely available at <https://targetmine.mizuguchilab.org/>.

P03-17

Natural product-like compound generation with chemical language models

Koh SAKANO *, Kairi FURUI, Masahito OHUE

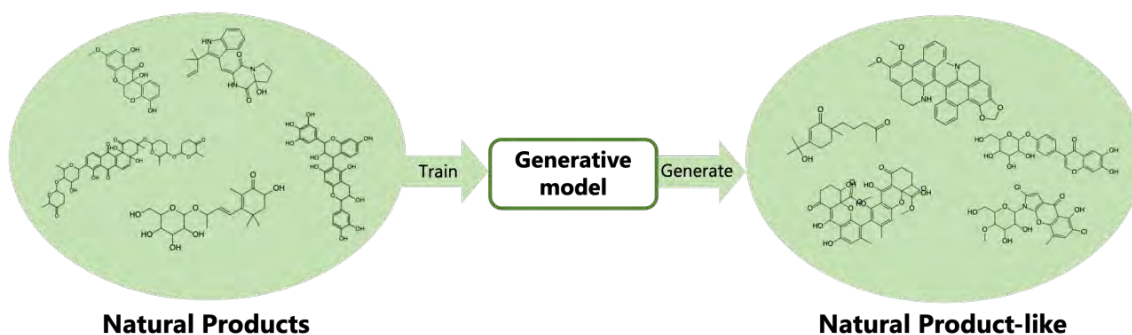
School of Computing, Institute of Science Tokyo

(* E-mail: sakano@li.c.titech.ac.jp)

Natural products are substances created by organisms in nature, often known for their biological activity and diverse structures. While drug development using natural products has been a common practice for many years, these compounds' complex structures pose significant challenges in determining their structure and synthesizing them. When compared to the more efficient high-throughput screening of synthetic compounds, natural products drug discovery tends to be avoided in terms of the cost.

In recent years, deep learning-based methods have been applied to the generation of molecules. Particularly, chemical language models, applications of natural language processing technology to the field of chemistry, have made remarkable progress. In this study, we fine-tuned pre-trained chemical language models on a natural product dataset and generated natural product-like compounds.

A total of 100 million molecules were generated, and the results showed that the distribution of the generated compounds was similar to that of natural products. The effectiveness of the generated compounds as drug candidates was also evaluated. This study proposes a method to explore the vast chemical space and reduce the time and cost of natural product drug discovery.



P03-18

Development of an Integrated Machine Learning Model Incorporating Compound-Protein Information for Design and Prediction of Small-Molecule Modulators of PPIs

Tsubasa NAGAE ^{*1, 2}, **Kohei SODA**^{2, 3}, **Kazuyoshi IKEDA**^{4, 5}, **Masashi TSUBAKI**², **Kentaro TOMII**^{1, 2, 3}

¹Graduate School of Medical Life Science, Yokohama City University

²Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology

³Graduate School of Frontier Sciences, The University of Tokyo

⁴Center for Computational Science, RIKEN

⁵Faculty of Pharmacy, Keio University

(* E-mail: w235434d@yokohama-cu.ac.jp)

Background

Protein-protein interactions (PPIs) have emerged as promising targets in drug discovery, offering potential to address previously undruggable diseases. PPI modulators, including both inhibitors and stabilizers, often exhibit unique properties that diverge from traditional drug-like molecules. This brings both opportunities and challenges in drug development.

Existing indicators and computational methods used for conventional drug targets are often inapplicable to PPI modulators due to their unique properties. This has led to the development of specific indicators for PPI-targeting compounds. However, these PPI-focused indicators have significant limitations: they mainly emphasize inhibitors, often struggle to accurately evaluate stabilizers, and frequently lack integration of specific protein target information. Given these challenges, the field of small-molecule modulators of PPIs urgently requires novel computational approaches for more effective design and prediction of both inhibitors and stabilizers.

Methods

We constructed a novel dataset of PPI stabilizers and inhibitors, including information on their target protein pairs. Data collection involved database mining and literature review, with each entry comprising a triplet of compound information (SMILES) and amino acid sequence information for the target protein pair. For stabilizers, where existing databases were insufficient, we extracted candidates based on structural information of protein-ligand-protein triplets from the Protein Data Bank (PDB) to augment the dataset.

In our machine learning model, we treated stabilizer entries as positive examples and inhibitor entries as negative examples. We primarily used representations generated by compound language models and protein language models as input features to capture the characteristics of PPI modulators.

Results

We constructed a dataset containing over 4,000 entries of triplets. Statistical data analysis revealed a slight bias towards negative examples, which we addressed by curating a balanced dataset for machine learning purposes. Then, we built a machine learning model to classify PPI stabilizers and inhibitors, and evaluation of its performance confirmed that classification was possible with satisfactory accuracy.

This study suggests the possibility of a new approach to designing small-molecule modulators of PPIs. Our model, which incorporates both compound and protein target information, may contribute to deepening our understanding of PPI modulator prediction and design.

P03-19

REALM: Region-Empowered Antibody Language Model for Antibody Property Prediction

Toru NISHINO *¹, Noriji KATO¹, Takuya TSUTAOKA¹, Yuanzhong LI¹, Masahito OHUE²

¹Bio Science & Engineering Laboratory, FUJIFILM Corporation

²Institute of Science Tokyo, School of Computing

(* E-mail: toru.nishino@fujifilm.com)

To reduce the manufacturing costs of antibody drugs, it is crucial to predict antibody property from antibody sequences.

Recently emerged protein language models (pLM) can build property prediction models based solely on fine-tuning with a small amount of antibody property data.

However, accurate prediction of antibody property with pLM is challenging because pretraining of protein language models primarily focuses on learning antibody co-evolution from large antibody sequence database.

In this study, we propose Region-Empowered Antibody Language Model (REALM), an antibody language model pretrained from scratch with novel pretraining strategy, to incorporate not only co-evolution but also region information of antibodies.

Region information within the variable region of antibodies, particularly loop structures such as complementary determining regions (CDR) and strand structures, is important for understanding the characteristics of antibodies.

Moreover, we proposed a strategy for determining masking positions that enables antibody structural information to be more appropriately embedded in the protein language models.

We evaluate our proposed REALM using a dataset of three assays: hydrophobicity, thermal stability, and specificity.

The evaluation results show that REALM improves the accuracy of the two assays, hydrophobicity and thermal stability, compared to the previous antibody language model.

In addition, we analyze the internal behavior of our antibody language model. We show that the proposed REALM enables focusing on residues regarding important region information.

P03-20

Generalized Molecular Representation for Drug Discovery via Molecular Graph Latent Diffusion Autoencoder

Daiki KOGE ^{*1}, **Naoaki ONO** ^{2, 3}, **Takashi ABE** ¹, **Shigehiko KANAYA** ^{2, 3}

¹Graduate School of Science and Technology, Niigata University

²Data Science Center, Nara Institute of Science and Technology (NAIST)

³Division of Information Science, Graduate School of Science and Technology, Nara Institute of Science and Technology (NAIST)

(* E-mail: daiki-ko@ie.niigata-u.ac.jp)

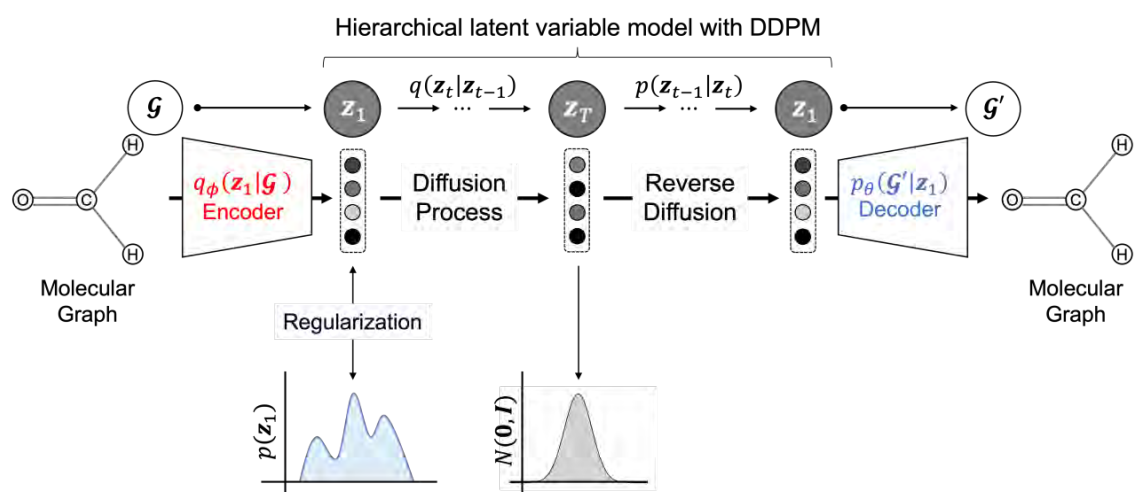
In drug discovery using machine learning, since labeled data with specific molecular properties are limited, it is essential to construct a prediction model with high generalization performance (i.e., high prediction performance for previously unseen data) from limited labeled data. Modeling meaningful latent representations numerically from chemical structures can facilitate generalization of molecular property predictions. Such a molecular representation is designed by feature engineering and representation learning approaches. Although the space of all possible organic compounds is very enormous, a molecular representation to generalize for the entire compound space may accelerate drug discovery. Variational autoencoder (VAE) is one of the representative deep learning methods for constructing the molecular representation. Although the VAE is an appropriate method for representing a chemical structure as a single latent variable vector, the standard distribution used as a prior distribution of latent variables oversimplifies the molecular representation and may affect the generalization performance of the property prediction.

This study aims to learn a molecular representation that improves the generalization performance of a molecular property prediction model using a deep generative model that integrates a graph transformer autoencoder and denoising diffusion probabilistic model (DDPM). Our proposed model maximizes the marginal likelihood with a smooth and multi-modal probability distribution generated by DDPM as the prior distribution.

We constructed prediction models for quantum chemical properties, such as HOMO energy, physicochemical properties such as solubility, and biochemical properties, such as biological activity. We analyzed the generalization performance of our model and several existing models using the widely applicable information criterion (WAIC) and the widely applicable Bayesian information criterion (WBIC). Our method demonstrated higher generalization

performance compared to the several existing methods. Additionally, this method efficiently identified desirable molecules in a chemical structure search experiment using Bayesian optimization with a Gaussian process regression model.

The results confirm that our method is effective in constructing a prediction model with high generalization performance from limited data.



P03-21

Data utilization and DX talent development on in-house KNIME platform

Toshiyuki OHFUSA ^{*1}, **Takanobu ARAKI**², **Ikumi KURIWAKI**², **Ayato SUGIYAMA**¹, **Kazuya NAGAOKA**¹, **Kenichi MORI**¹, **Takamune YAMAMOTO**³, **Kenji NEGORO**², **Kota TOSHIMOTO**⁴, **Shinji SOGA**⁵, **Takuya SHIMOMURA**⁶, **Tomomi YOKOYAMA**⁷, **Hiroko TAMURA**⁶

¹Modality Informatics, ResearchX, DigitalX, Astellas Pharma Inc.

²Platform Sciences & Modalities, Discovery Intelligence, Applied Research & Operations, Astellas Pharma Inc.

³Research Informatics, ResearchX, DigitalX, Astellas Pharma Inc.

⁴Systems Pharmacology, Advanced Translational Science & Management, Non-Clinical Biomedical Science, Applied Research & Operations, Astellas Pharma Inc.

⁵Biologics Engineering, Discovery Intelligence, Applied Research & Operations, Astellas Pharma Inc.

⁶Pharmaceutical Developability Labs, CMC Research, Astellas Pharma Inc.

⁷Technology Exploration 1, CMC Research, Astellas Pharma Inc.

(* E-mail: toshiya.ohfusa@astellas.com)

In recent years, various modalities have been developed for drug discovery approaches which target undruggable targets. These modalities include "bRO5" (beyond rule of 5) chemical modalities, such as bifunctional compounds as protein degraders and peptide molecules, as well as various biological modalities. Additionally, coupled with advances in automation equipment and the use of new technologies in the drug discovery field, such as generative AI, the amount of data generated at each step of the drug discovery process is steadily increasing.

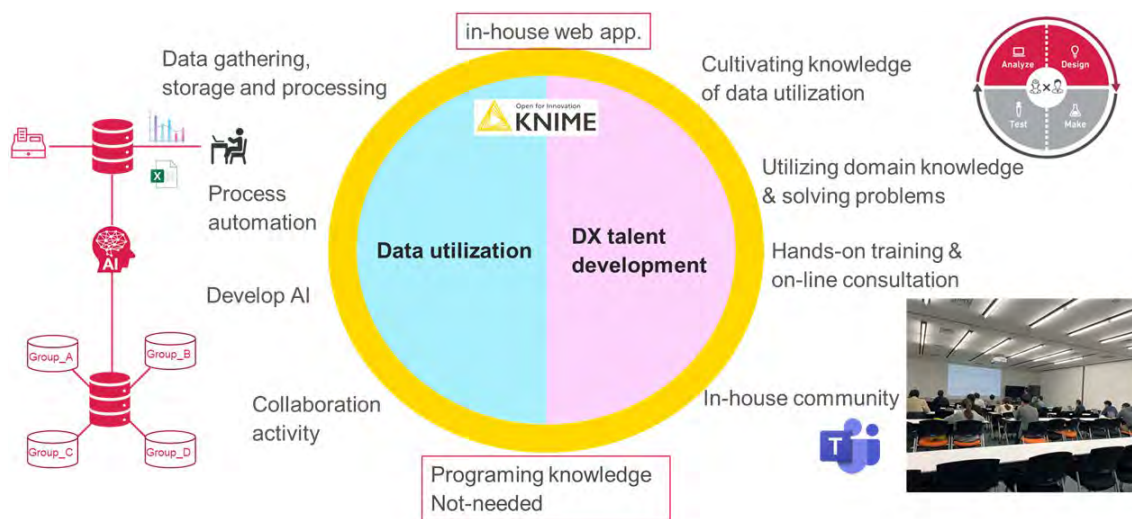
The increased data can be utilized for molecular predictions and simulations, leading to reduced operating costs. Moreover, enabling data-driven decision-making to improve the quality and efficiency of drug discovery is believed to expedite the delivery of new drugs to patients.

However, many wet lab researchers are not accustomed to handling large amounts of data, especially in CUI (Character User Interface) environments. Data interface issues pose a hurdle to learning in various fields and contribute to inefficiencies in data utilization within each laboratory. Therefore, our cross-departmental DX promotion team is building a foundation platform for DX activities to increase the number of individuals capable of handling data in various wet lab tasks. This platform utilizes KNIME, a GUI (Graphical User Interface)-based no-code/low-code tool that is easy for beginners to

understand and learn. The developed tool can be shared with other members as a web application by deployment on the KNIME Server platform installed on a shared server.

Furthermore, as an educational activity for members of each department, the DX team has established an in-house community to facilitate study sessions, information sharing, and provide a platform for questions and consultations. KNIME enables researchers in each department to utilize data effectively by leveraging their domain knowledge. Consequently, they can approach problem-solving from a field perspective in the development of various modalities, leading to significant improvements in decision-making speed and work efficiency. As part of data utilization efforts, automation of regular report creation within each organization, automatic data aggregation across organizations, and the development of prediction models have increased, in turn resulting in significant time and effort savings. Additionally, this platform is widely used not only in the discovery research department but also in the CMC research department and manufacturing department, with an increasing number of application developers in each department.

In this poster presentation, we will showcase the utilization of data and DX human resource development using the in-house KNIME platform.



P03-22

Astellas's Digital Transformation for Small Molecule Drug Discovery Research

Takanobu ARAKI ^{*1}, **Ikumi KURIWAKI**¹, **Wataru HAMAGUCHI**¹, **Toshiyuki OHFUSA**², **Kazuya NAGAOKA**², **Arina AFANASEVA**², **Natnael HAMDA**², **Kenichi MORI**², **Kenji NEGORO**¹

¹Platform Sciences & Modalities, Discovery Intelligence, Applied Research & Operations, Astellas Pharma Inc.

²Modality Informatics ResearchX, DigitalX, Astellas Pharma Inc.

(* E-mail: takanobu.araki@astellas.com)

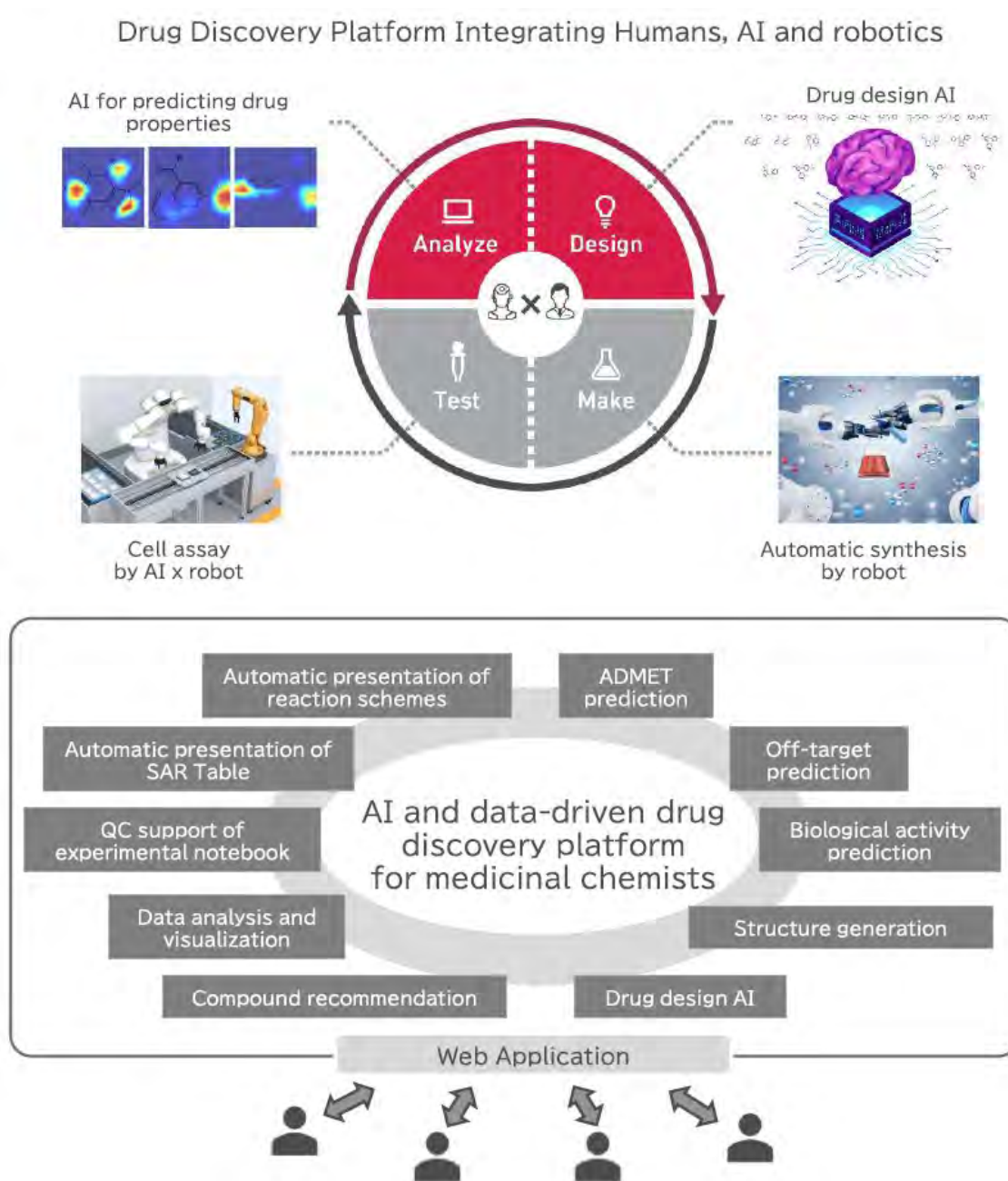
Recent years have seen remarkable progress in digital technologies such as AI, simulation, and robotics across various industries. These technologies are utilized to drive digital transformation efforts with the goal of revolutionizing business operations and generating new value. The situation for pharmaceuticals reflects that in other industries. Digital technologies are expected to improve productivity in the drug discovery research process, allowing pharmaceutical companies to deliver better drugs to patients as quickly as possible.

The optimization process, which accounts for the majority of small-molecule drug discovery research, is called the Design-Make-Test-Analysis (DMTA) cycle. Researchers iterate this cycle, while integrating feedback from experiment results into the process, to refine their hit compounds into clinical candidates. As part of our digital transformation efforts to accelerate drug optimization research, we have installed AI/robot technologies such as ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) and activity predictions, chemical structure generation and robot synthesis into each step of the DMTA cycle (Figure 1). This collaborative system between humans and AI/robots has enabled us to create high-quality drug candidates in a shorter period of time, successfully achieving an approximately 70% reduction in the time from a hit compound to a clinical candidate.

To further improve the system we developed, it is important to expand the user base and develop talent. The Dx tools on the platform were developed by medicinal chemists using the low-code/no-code platform KNIME®. The user-friendly tools were well received by other medicinal chemists, resulting in more users. Some users even expressed interest in creating their own tools. In response, KNIME® workshops were held to train medicinal chemists to become tool developers as well as to improve their data handling and analysis skills.

These efforts induced the trainees to generate new ideas for improving operational processes. The tools developed in collaboration with them enhanced the efficiency of research operations.

In this poster presentation, we will introduce an overview of our drug discovery platform which integrates humans, AI and robotics, and tools that enhance the productivity of medicinal chemists.



P03-23

Unraveling Microbiome Complexity: A Knowledge Graph Approach to Functional Interpretation in Drug Discovery

Hirokazu NISHIMURA *¹, **Taku HIRATA**¹, **Maaly NASAR**⁴, **Mark STREER**⁴,
Michael HUGHES⁴, **Sachiko FURUYA**², **Fumihiko OONO**³, **Ryuuta SAITOU**¹

¹Discovery Technology Laboratories, Mitsubishi Tanabe Pharma Corporation

²Oncology & Immunology Unit, Mitsubishi Tanabe Pharma Corporation

³Business Development Department, Mitsubishi Tanabe Pharma Corporation

⁴SciBite Ltd.

(* E-mail: nishimura.hirokazu@ma.mt-pharma.co.jp)

The human microbiome holds significant potential for drug discovery and personalized medicine. Advances in metagenomic analysis have significantly enhanced our ability to detect changes in microbial communities. However, a major technical challenge remains in the functional interpretation of these changes. Understanding how specific microbes and their metabolites influence human physiology is still complex and unresolved.

To address this challenge by clearly understanding the relationships between the microbiome and disease based on the molecular mechanism, we have developed a comprehensive knowledge graph database that links documented microbes and their metabolites to human biological functions. Initially, First, we searched MEDLINE and PMC for literature related to microbes and their metabolites, retrieving approximately 80,000 relevant articles. Utilizing SciBite's named entity recognition (NER) and extraction engine, TERMite, we extracted 10 types of nodes (Microbe, Metabolite, Indication, Gene, Pathway, Gene Ontology, etc.), totaling 538,527 nodes, and mapped 6,852,832 edges. Machine Learning (ML) algorithms and Large Language Models (LLM) were then employed to classify and score these edges.

Our knowledge graph database offers several key advantages. First, it integrates and organizes data from various studies, providing a unified platform for researchers to explore microbial functions and their impacts on human health. Second, the graphical representation of data allows for intuitive visualization of relationships, making it easier to identify potential causal links between microbiome alterations and physiological outcomes. These advantages support hypothesis generation for elucidating disease mechanisms and identifying therapeutic targets.

P03-24

Age prediction from DNA methylation data using machine learning

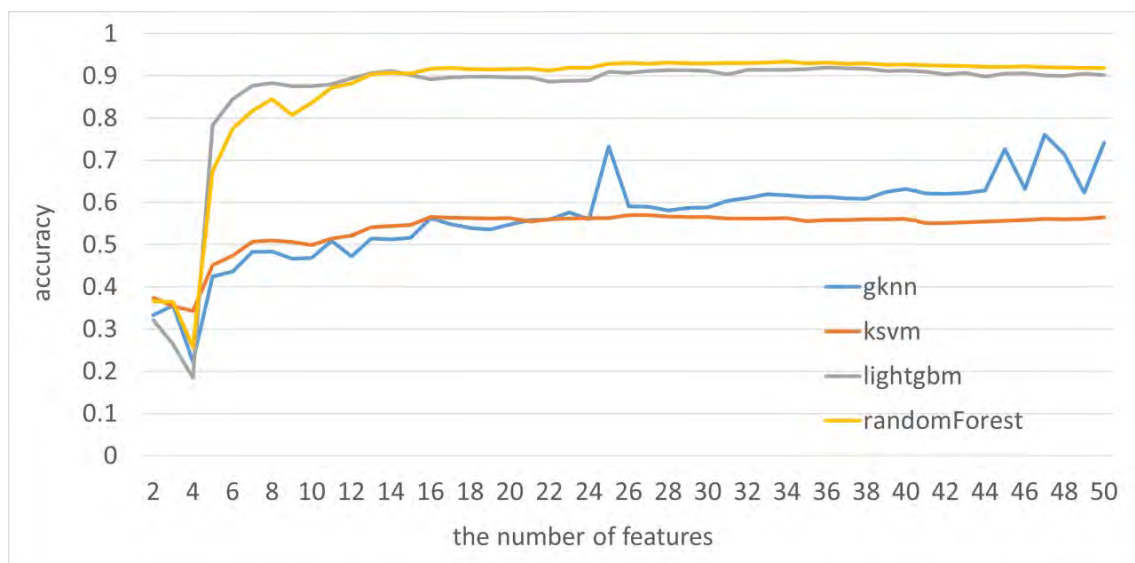
Nagisa MATSUO ^{*1}, Kenji SATOU²

¹Graduate School of Natural Science and Technology, Kanazawa University

²Institute of Transdisciplinary Sciences for Innovation, Kanazawa University

(* E-mail: matsuo31@stu.kanazawa-u.ac.jp)

Gene expression in living organisms is greatly influenced by the methylation status of DNA. Since the methylation status of DNA changes with age, it is thought that by measuring this, it is possible to estimate a person's biological age (an age that indicates the degree of the body aging, separate from chronological age). In a previous study, Horvath selected 353 CpG probes from the large number of CpG probes present on DNA to predict biological age, which is highly correlated with chronological age, in various cell types using these methylation levels. In addition, a recent study by Galkin et al. reported that a biological age prediction method using deep learning is effective for large-scale DNA methylation data. In this study, we examine the prediction accuracy of various machine learning algorithms using the same DNA methylation data as in the Galkin's previous study. Since deep learning does not always achieve the highest accuracy in the field of machine learning that deals with classification and regression problems, it is important to examine how effective other machine learning methods are in predicting biological age from methylation data. As a result of experimenting with four machine learning methods and two importance measures in feature selection, we were able to achieve the highest accuracy (correlation coefficient) of 0.9334 by combining 34 features and random forest. Although the accuracy was slightly lower than the best score 0.94 achieved by previous research using deep learning with 1,000 features, it was confirmed that random forests can also achieve equally accurate predictions with only 34 selected features. Further analysis about the role of these selected CpG probes will be conducted and the results will be presented in poster session.



P03-25

Exchange System for Glycan Textual Notations Development to Integrate Various Glycan Databases and Improve Search Accuracy

Hiromitsu SHIMOYAMA *, Masaaki MATSUBARA, Issaku YAMADA

Glycoinformatics, The Noguchi Institute
(* E-mail: shimoyama@noguchi.or.jp)

Various notation formats have been developed to describe glycan structures, and then, databases with different notation exist independently. It is time-consuming to change the notation according to the databases to deposit and search glycans. The purpose of our study is development of a program that automatically converts a glycan notation according to databases to reduce above tasks and enable us to access databases easily, accurately, and widely. Each glycan notation has their own limitations, for example, IUPAC cannot to represent ambiguous linkage patterns. However, to access the database comprehensively and automatically, even such a structure should be kept internally. Therefore, Web3 Unique Representation of Carbohydrate Structures (WURCS), which can uniquely represent complex glycan structures, was chosen to design the internal objects. As WURCS is complicated, glycan structures are expected to be obtained in a simple way such as IUPAC and GlycoCT, and the program stores the input internally and converts it according to that of databases. For example, an information of queries with an ambiguous pattern obtained in GlycoCT format would be simplified for IUPAC databases but may be fully utilized for other databases. As the internal data object are repeatedly used for different interconversion, the object related code can be simplified. Then, such program would be easy to maintain.

P03-26

Drug discovery study integrating compound generative AI and molecular docking

Noriaki OKIMOTO³, Makoto TAIJI³, Mariko OKADA¹

¹Institute for Protein Research, Osaka University

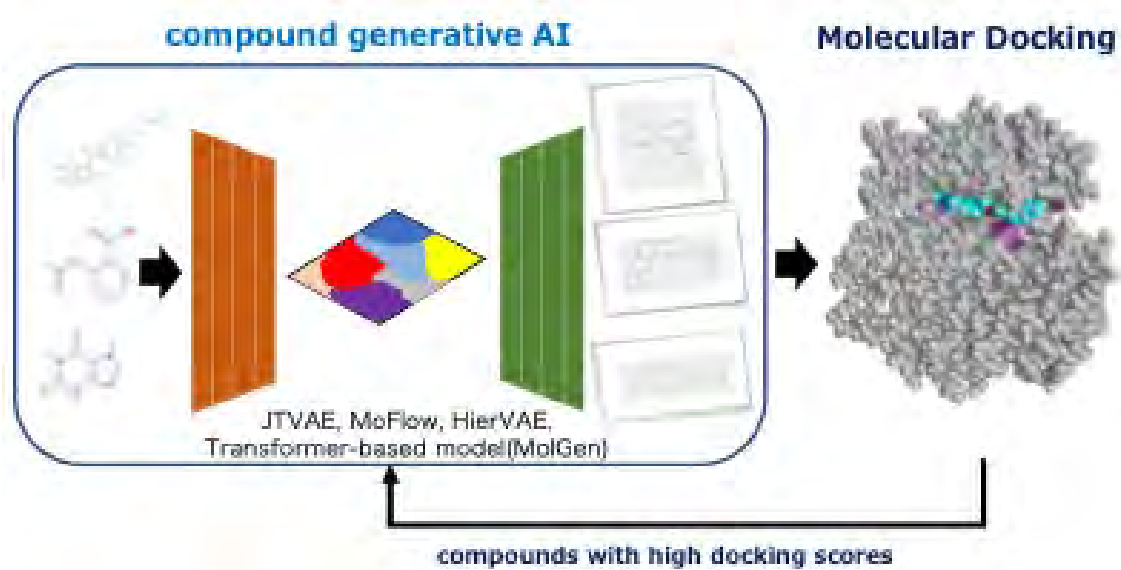
²WPI-PRIME, Osaka University

³Center for Biosystems Dynamics Research, RIKEN

(* E-mail: okimoto@riken.jp)

In recent years, AI for compound generation has gained significant attention as a crucial technology in drug discovery. This study aims to advance drug discovery by integrating AI-driven compound generation with molecular docking. The AI-driven system combines VAE, Flow, and GPT models to generate potential inhibitors targeting specific proteins, and its drug discovery capabilities are evaluated to assess its effectiveness in discovering novel compounds.

In this research, we focused on discovering inhibitors targeting thymidine phosphorylase (TP). TP is an enzyme involved in the catabolism of thymidine and is an important factor in promoting cellular aging, as well as serving as a target for cancer suppression. Additionally, several TP inhibitors have been identified. Using this drug discovery system, we generated compounds with high docking scores, selected those with high structural similarity from a commercial chemical library, and performed experimental evaluations. These evaluations revealed that several compounds successfully demonstrated TP inhibitory activity. We are currently optimizing these active compounds and continuing validation in close collaboration with experimental studies. We will provide the details on the day.



P03-27

Spike separation of high-gamma power in ECoG using peak detection

Masato SAKAGAMI *¹, Masashi KINOSHITA², Mitsutoshi NAKADA², Kenji SATOU³

¹Graduate School of Natural Science and Technology, Kanazawa University

²Department of Neurosurgery, Kanazawa University

³Institute of Transdisciplinary Sciences for Innovation, Kanazawa University

(* E-mail: zdnmasato@stu.kanazawa-u.ac.jp)

Introduction

The brainwaves measured by attaching electrodes directly to the cortical surface are referred to as electrocorticogram (ECoG). Brainwaves exhibit distinct properties based on their frequency ranges. Especially, increase in high-gamma power (HGP) above 60 Hz are often interpreted as indicative of local brain activity associated with specific tasks. In this study, we aimed to analyze the increase of HGP in ECoG data during cognitive tasks within a short time frame of less than 0.1 seconds. Through this process, spike-like signals were frequently found in power above approximately 100 Hz. In noise separation for brainwaves, visual inspection by experts and low-pass filters are commonly used. However, due to the averaging effect of low-pass filters, simple application of them to this case does not work well. Therefore, in this study, we aim to analyze the increase of HGP in more detail by attempting to separate spikes using a method that utilizes peak detection.

Methods

In this study, we used ECoG data from several patients who underwent awake craniotomy at Kanazawa University Hospital in the past. The data were collected while the patients performed specific tasks. The ECoG were measured at a sampling rate of 500 Hz using 16 electrodes arranged in a grid approximately 3 cm square. The cognitive tasks consisted of either a naming task, in which a patient called the name of an illustrated object, or an emotion task, in which a patient identified the emotion shown in a facial photograph.

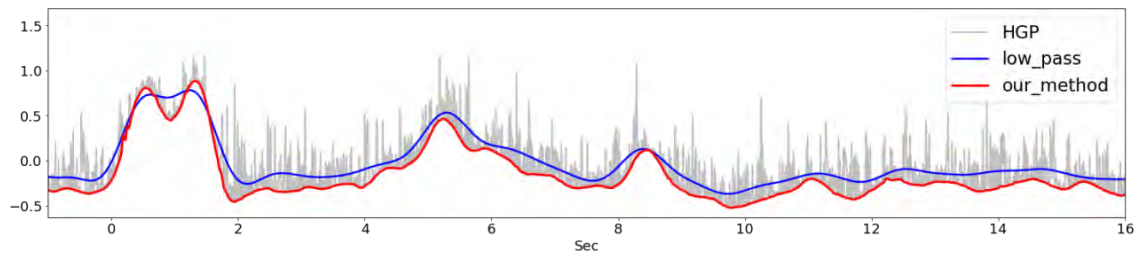
For preprocessing the data, the following steps were carried out: 1) Perform a Fourier transform on 1-second segments of the brainwave data, with a 0.002-second shift. 2) Perform z-transformation for each frequency greater than 60 Hz to normalize the power across frequencies. 3) Calculate the average across frequencies.

Additionally, for spike separation, the following method was used: 4) Detect

peaks of local minimum values in the signal. 5) Since some of the detected peaks have extremely high values far from neighbors, local “maximum” peaks among them are repeatedly detected and removed. 6) Finally, apply spline interpolation to the detected peaks, then reconstruct the signal with the spike components removed.

Results and Discussions

As shown in the figure, this method was able to separate the spike components more appropriately in comparison with the low-pass filter. As a result, the task-related increase in HGP could be identified in more precise time of onset. This suggests that the proposed method may be useful for separating spike components in HGP analysis.



P03-28

Estimation of transmission routes of the COVID-19 BA.1.1.2 variant using McAN and 3D graph visualization

Masafumi SAITO ^{*1}, **Yoshinori TAKAHASHI**², **Yasunori IWATA**³,
Kenji SATOU⁴

¹Graduate School of Natural Science and Technology, Kanazawa University

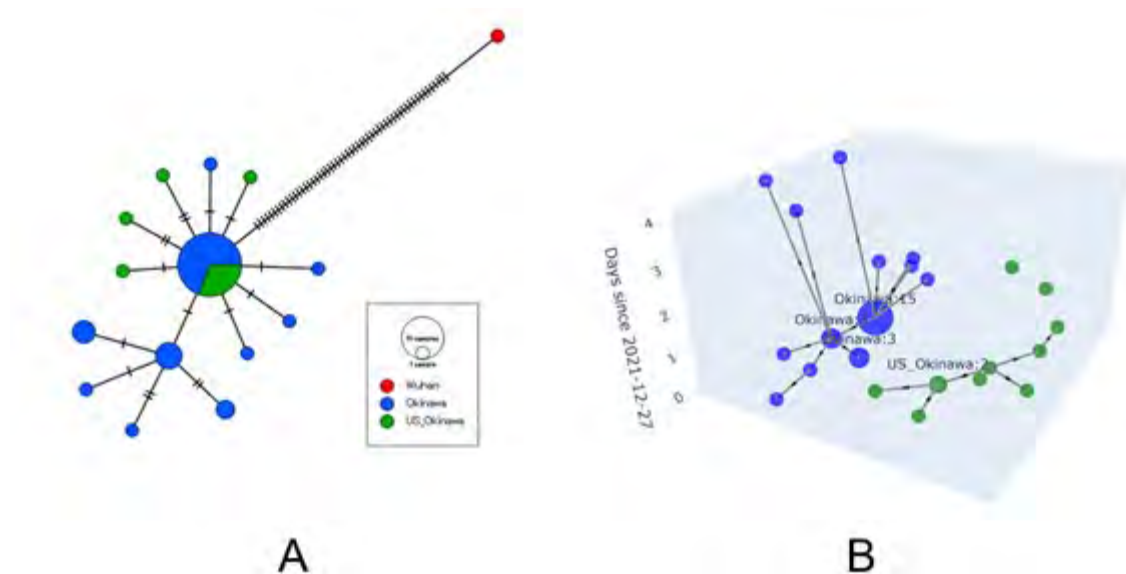
²Department of general medicine and infectious diseases, JA Toyama Koseiren Takaoka Hospital

³Department of Nephrology and Rheumatology, Kanazawa University

⁴Institute of Transdisciplinary Sciences for Innovation, Kanazawa University
(* E-mail: saito562943@gmail.com)

In 2019, the COVID-19 quickly spread worldwide and caused many infection clusters. To prevent such clusters and future outbreaks, it is important to track the transmission routes of infections and identify how the virus spreads in these clusters. Previous studies have reported that phylogenetic analysis on the whole genome sequences of viruses can reveal detailed transmission routes. In this study, we analyzed the transmission routes of an Omicron variant (BA.1.1.2) infection cluster that appeared in Okinawa Prefecture in Japan at the end of December 2021. This cluster is considered to have started from infections among US military personnel at a base in Okinawa and then spread further. We conducted a phylogenetic analysis using the genome sequences of BA.1.1.2 variant from 42 infected individuals found in Okinawa between December 27 and December 31, 2021. Among these, 11 strains came from US military personnel, and 31 strains were from community infections in Okinawa. By comparing these genome sequences with the reference genome sequence from Wuhan, China, we identified the genetic relationships among the viruses and estimated the local transmission routes. To understand the transmission routes, we used two methods: the median-joining network and McAN haplotype network analyses. The median-joining network method calculates genetic distances between viruses based on their mutations and constructs a network to visualize the transmission routes. This method is widely used for tracking transmission. The McAN method is a new algorithm that considers specific mutations and infection dates to create the network. The result of the median-joining network analysis is shown in Figure A on the left. Each node represents one or more viruses, with green nodes indicating infections in US military and blue nodes indicating community infections. The number of marks on the lines between nodes shows the number of mutations. The central large node clusters

strains with very similar genetic distances, making it hard to distinguish between community and military infections. It was also difficult to determine if the infections started from US military personnel. In contrast, the McAN analysis result in Figure B on the right shows that infections by US military personnel and community infections formed separate clusters, suggesting no direct relationship. This result indicates that the median-joining network and McAN methods might lead to different conclusions when analyzing infection clusters. Furthermore, our study suggests that McAN can build networks considering mutation types and infection dates, and visualize detailed transmission routes, including their timelines, using a three-dimensional directed graph.



P03-29

A framework for enhanced de novo protein design using deep learning and bayesian optimization

Shuto HAYASHI ^{*1}, **Jun KOSEKI**², **Teppei SHIMAMURA**^{1, 3}

¹Institute of Science Tokyo

²Cellular and Molecular Biotechnology Research Institute, National Institute of Advanced Industrial Science and Technology

³Nagoya University Graduate School of Medicine

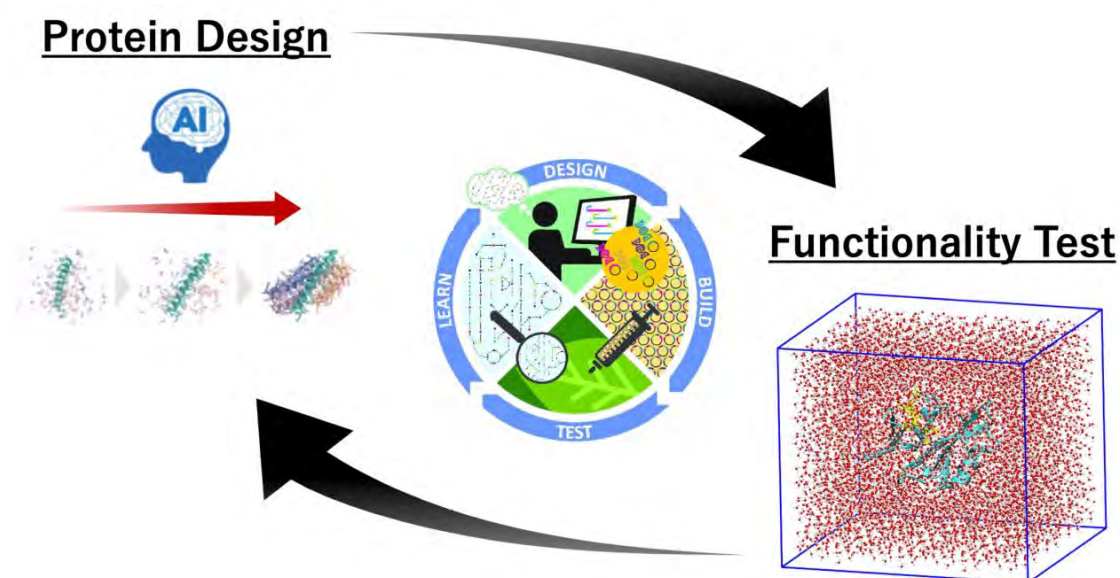
(* E-mail: s-haya.csb@tmd.ac.jp)

The field of de novo protein design has experienced significant progress in recent years, particularly with the advent of deep learning techniques. These innovations have enhanced our ability to create custom-designed proteins for diverse applications, ranging from therapeutics to novel materials. However, despite these notable improvements, the functionality of computationally designed proteins often remains inferior to their naturally occurring counterparts or those developed through conventional expert-driven methodologies.

To address this problem, we introduce a design-build-test-learn (DBTL) cycle framework tailored for the development of proteins with enhanced functionality. The framework consists of two main phases: an initial pre-DBTL phase and an iterative DBTL cycle phase. In the pre-DBTL phase, we employ a combination of two deep learning-based protein design methods, specifically RFdiffusion and ProteinMPNN, to generate a diverse pool of potential functional proteins. To further enrich the diversity of the pool, we implement a combinatorial assembly strategy, which allows for the exploration of a broader sequence space.

Following the initial phase, highly functional proteins in the candidate pool are identified through the DBTL cycles, which comprise three key components: in silico evaluation, deep learning-based prediction, and multi-objective Bayesian optimization. During the in silico evaluation, we utilize MD simulations and MM/GBSA to assess protein characteristics, including binding free energies and structural stabilities. The results of the evaluation are then used to train an ensemble of neural networks. This ensemble model can be used not only to predict protein functionalities, but also to infer the uncertainty of the prediction, directly from amino acid sequences. Using the trained ensemble model as a surrogate model, we implement a multi-objective Bayesian optimization algorithm to propose promising protein candidates for subsequent rounds of evaluation.

To validate our framework, we applied it to the design of inhibitors targeting BCAT1, a key enzyme implicated in cancer cell metabolism. Our results demonstrate the efficacy of the pre-DBTL phase in generating proteins capable of inhibiting cancer cell proliferation. Furthermore, through the iterative DBTL cycle, we were able to identify proteins with substantially enhanced inhibitory effects, showcasing the power of our approach in optimizing protein functionality. This study represents a significant advancement in the field of computational protein design. By integrating deep learning-driven design with iterative optimization and in silico evaluation, our framework offers a powerful tool for the development of highly functional proteins. The versatility of our platform opens up new avenues for the development of custom-designed proteins across a wide spectrum of applications, from enzyme engineering for industrial biotechnology to the creation of novel biomaterials.



P03-30

Directional Graph Modelling for Solution Design and Experiment Automation

Yusuke SAKAI *

Center for Biosystems Dynamics Research, RIKEN

(* E-mail: yusuke.sakai@riken.jp)

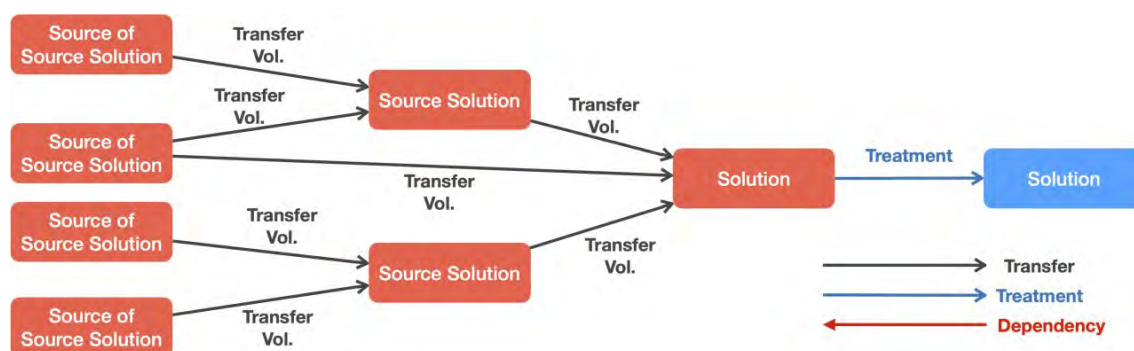
The popularisation of smart analytical equipment and liquid handling devices has made laboratory automation crucial for enhancing research efficiency. Automation advances science by improving reproducibility, enabling high-throughput investigations, and enhancing work-life balance. Research automation now spans experimental procedures, experiment design, record-tracking, data analysis, and interpretation.

This presentation focuses on automating experimental stages in DNA nanotechnology and synthetic biology, showcasing an in-house relational database (RDB) solution and automated liquid handling devices. DNA origami, which involves hundreds of components, demands laborious mixing and careful inventory management of numerous oligo DNAs. As research advances, the number of DNAs can tens of thousands, with countless combinations. To address these challenges, we developed a tool called 'SolutionDesignerRDB (SDRDB)' to automate these processes, reducing reliance on researchers' diligence.

SDRDB, build using Claris FileMaker, is a low-code, GUI-rich relational database application. The solution design within the app is modelled as a directed graph, where each node represents an individual solution, and the directional edges represent the volume of liquid transferred. The sample preparation processes, such as liquid handling, heating, and separating, are similarly modelled, with each node representing a specific state of an individual solution, and the edges showing their dependency (process order) and procedural details. The main features of the app include: (1) an RDB that supports multi-stage mixing and essential experimental procedures, allowing systematic and dynamic configuration of sample preparation, (2) a user-friendly GUI for designing solution compositions within the database, either in batches or individually, (3) scripts to generate protocol files for specific liquid handling devices, and (4) scripts for importing oligo DNA lists generated by cadnano into the database for each DNA origami structure (the primary goal of the system). Due to its low-code design, users can easily adapt scripts and the GUI to fit their specific research needs with minimal effort. SDRDB's solution design and integration with liquid handling systems help overcome bottlenecks in research involving

complex mixture designs, not just in DNA nanotechnology.

We are seeking funding and collaborators to migrate the system to a free web application using PostgreSQL or equivalent and JavaScript for broader accessibility. We also aim to re-implement the system in a native Graph Database model, renaming it SolutionDesiGnerDB (SDGDB) for enhanced data analysis, or NoSQL (SolutionDesiGnerDB, SDNDB) for multimodal extension. Additionally, we intend to integrate Bayesian inference functions for automatic sample series formulation, upgrading the app to ExperimentDesignerRDB (EDRDB).



P04-01

Analysis of Kinase Binding Specificity of Staurosporine using the Fragment Molecular Orbital Method

Ruri MIHATA ^{*1}, Riko HIGASHINO², Mayu KITANO³, Shuhei MIYAKAWA¹, Shi Yu TIAN¹, Daisuke TAKAYA¹, Takayoshi KINOSHITA³, Shigenori TANAKA², Kaori FUKUZAWA¹

¹Graduate School of Pharmaceutical Sciences, Osaka University

²Graduate School of System Informatics, Kobe University

³Graduate School of Science, Osaka Metropolitan University

(* E-mail: mihata-r@phs.osaka-u.ac.jp)

Introduction

Controlling the selectivity of protein kinases, which are widely distributed in the human body, is a critical issue in drug design. Structural features of kinases, such as the P-loop, essential for ATP binding, and the DFG motif, crucial for the expression of activity, are well known. Staurosporine (STU) exhibits low selectivity and acts as an inhibitor of various protein kinases. This study investigated the detailed interactions between STU and target kinases using computational methods of molecular dynamics (MD) and fragment molecular orbital (FMO), providing insights to understand the binding specificity and selectivity of STU.

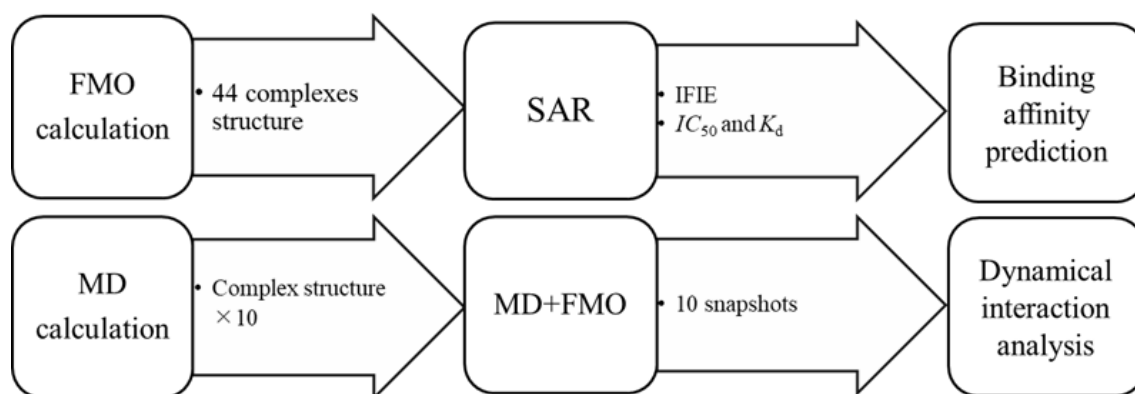
Methods

Employing complex structures of STU and 44 kinds of kinases from Protein Data Bank (PDB) based on X-ray crystallography, FMO calculations were conducted to analyze Inter-Fragment Interaction Energy (IFIE). Structure-activity relationships (SAR) between IFIE and kinase inhibitory activity IC_{50} , as well as kinase binding activity K_d , were investigated. In addition, ten MD runs of 100 ns between ALK kinase and STU were performed. For ten snapshots from each MD trajectory, FMO calculations were performed (MD+FMO). MD calculations were performed using GROMACS under ff14SB for protein force field, TIP3P for water force field, GAFF2 for ligand force field and FMO calculations were conducted using ABNIT-MP at a calculation level of MP2/6-31G*. The calculation flow is shown in fig.

Results and Discussion

The SAR analysis was performed on 44 different kinases with low correlation coefficient ($R^2 = 0.17$), and classification of kinases based on P-loop and DFG-motif did not improve the correlation. Therefore, to remove the influence of crystal packing, MD calculations were performed. The MD results showed that

while the RMSD of the overall structure was suppressed to $2.05 \pm 0.22 \text{ \AA}$, the RMSD of the P-loop was large at $2.95 \pm 0.77 \text{ \AA}$, suggesting that the structure was relaxed in an aqueous environment. Additionally, a comparison was made between the IFIE from the crystal structure and the dynamically averaged IFIE from MD+FMO. A significant difference was observed in the interaction between STU and Asp1203, where the IFIE from the crystal structure was -106.1 kcal/mol , while the averaged IFIE from MD+FMO was -43.5 kcal/mol . These findings suggest that crystal packing affects the interaction energy of STU and kinase, and comprehensive MD+FMO study is a promising approach to understanding kinase binding specificity.



P04-02

Dynamical Interaction Energy Analysis of Elastase in Each Reaction State: Insights from Molecular Dynamics and Fragment Molecular Orbital Calculations

Shuhei MIYAKAWA^{*1}, Shi Yu TIAN¹, Daisuke TAKAYA¹, Takayoshi KINOSHITA³, Shigenori TANAKA², Kaori FUKUZAWA¹

¹Graduate School of Pharmaceutical Sciences, Osaka University

²Graduate School of System Informatics, Kobe University

³Graduate School of Science, Osaka Metropolitan University

(* E-mail: miyakawa-s@phs.osaka-u.ac.jp)

[Introduction]

Elastase, a serine protease classified into the chymotrypsin family, has been variously studied in molecular biology and its catalytic triad consists of His57, Asp102, and Ser195 residues. In this study, classical molecular dynamics (MD) and fragment molecular orbital (FMO) calculations were conducted to investigate the dynamical interactions of three states: apo, enzyme-substrate complex and tetrahedral intermediate.

[Methods]

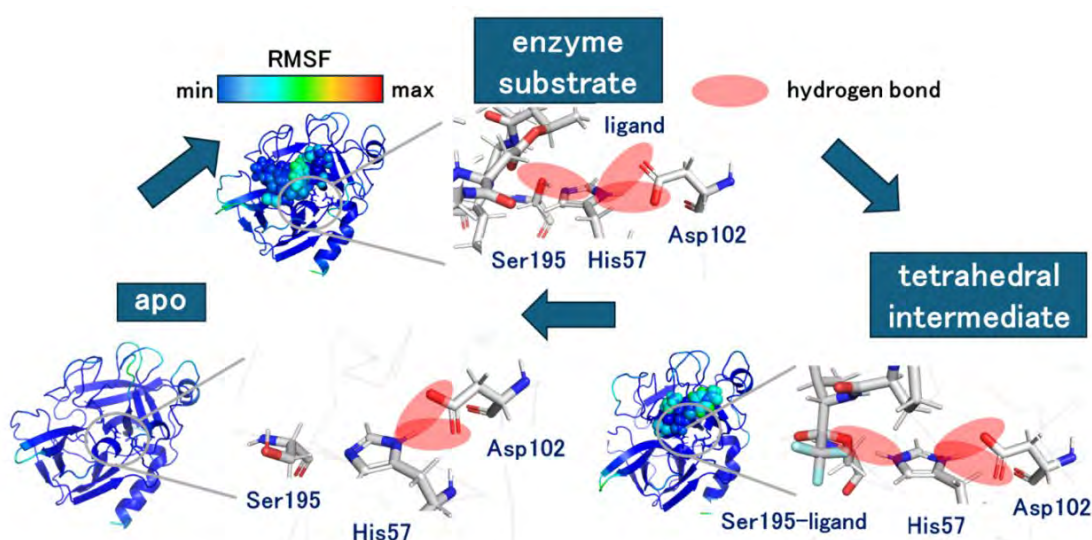
MD calculations by GROMACS were performed for 100 ns × 3 runs on each of the structures: apo and the tetrahedral intermediate structures obtained by neutron crystallography¹, and the enzyme-substrate complex structure obtained by X-ray crystallography², followed by trajectory analysis. FMO calculations were performed on a total of 900 structures extracted from each trajectory at 1 ns interval using the ABINIT-MP program at the MP2/6-31G* level. The inter-fragment interaction energy (IFIE) obtained from the FMO calculation, and pair interaction energy decomposition analysis (PIEDA) were utilized to analyze the hydrogen bond network of the enzyme reaction site. Here, PIEDA component consists of electrostatic energy (ES), exchange repulsion energy (EX), charge transfer energy (CT+mix), and dispersion energy (DI).

[Result & Discussion]

The three reaction states of elastase indicated little structural changes, revealed by the RMSD within 1.8 ± 0.2 Å and the low RMSF values as shown in the figure.

However, in the apo state, Ser195, a nucleophilic residue that attacks the substrate, showed in a different conformation from the other states, indicating that the structure of Ser195 did not form a hydrogen bond with His57 in the absence of the substrate. In contrast, PIEDA analysis revealed that the main components of the interaction between Ser195 and His57 were ES and CT+mix, indicating that they formed a stable hydrogen bond in the enzyme-substrate state. The hydrogen bond network at the reaction site was further strengthened in the tetrahedral intermediate state. We will analyze dynamical interactions of the substrates and surrounding residues of elastase for different state of its catalytic cycle.

1. Tamada, T., Kinoshita, T., Kurihara, K., Adachi, M., Ohhara, T., Imai, K., Kuroki, R., & Tada, T. (2009). Combined High-Resolution Neutron and X-ray Analysis of Inhibited Elastase Confirms the Active-Site Oxyanion Hole but Rules against a Low-Barrier Hydrogen Bond. *Journal of the American Chemical Society*, 131(31), 11033–11040. <https://doi.org/10.1021/ja9028846>
2. Kinoshita, T., Kitatani, T., Warizaya, M., & Tada, T. (2005). Structure of the complex of porcine pancreatic elastase with a trimacrocyclic peptide inhibitor FR901451. *Acta Crystallographica Section F*, 61(9), 808 – 811. <https://doi.org/10.1107/S1744309105026047>



P04-03

Development of the Cryptic Site searching method with Mixed-solvent molecular dynamics and Topological data analyses methods

Jun KOSEKI ^{*1}, **Motono CHIE** ^{1,2}, **Yanagisawa KEISUKE** ^{3,4}, **Kudo GENKI** ⁵, **Yoshino RYUNOSUKE** ^{6,7}, **Hirokawa TAKATSUGU** ^{6,7}, **Imai KENICHIRO** ^{1,8}

¹Cellular and Molecular Biotechnology Research Institute, National Institute of Advanced Industrial Science and Technology

²Computational Bio Big-Data Open Innovation Laboratory, National Institute of Advanced Industrial Science and Technology

³Department of Computer Science, School of Computing, Tokyo Institute of Technology

⁴Middle Molecule IT-based Drug Discovery Laboratory, Tokyo Institute of Technology

⁵Physics Department, Graduate School of Pure and Applied Sciences, University of Tsukuba

⁶Division of Biomedical Science, Faculty of Medicine, University of Tsukuba

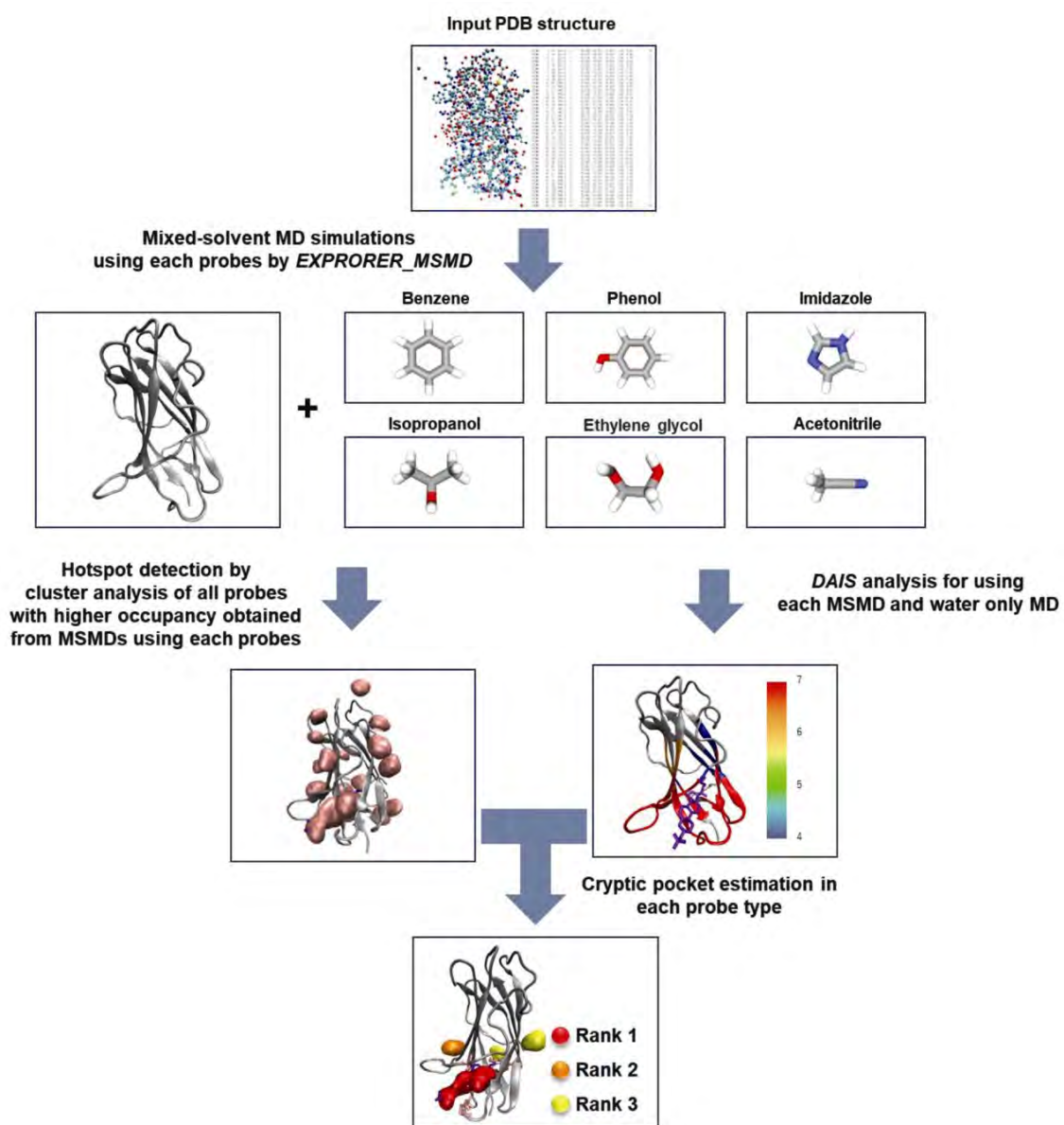
⁷Transborder Medical Research Center, University of Tsukuba

⁸Global Research and Development Center for Business by Quantum-AI Technology, National Institute of Advanced Industrial Science and Technology

(* E-mail: jun.koseki@aist.go.jp)

Some functional proteins change their structure to give rise to a binding site only when a binding molecule approaches them. Such binding sites are called cryptic sites and important targets to expand the scope of drug discovery. However, it's still difficult to predict cryptic sites correctly. Therefore, we propose a method to correctly detect cryptic sites using topological data analysis and mixed-solvent molecular dynamics (MSMD) simulation. To detect hotspots, we employed MSMD simulations using six probes with various chemical properties (Benzene, Isopropanol, Phenol, Imidazole, Acetonitrile, and Ethylene glycol). Then, the possibility of cryptic site was then ranked using our topological data analysis method, the Dynamical Analysis of Interaction and Structural changes (DAIS). For nine target proteins with cryptic sites, the proposed method significantly outperformed the accuracy of the recent machine learning method, Pocketminer. We can detect six of the nine cryptic sites at hotspot Rank 1. In our method, the MSMD simulations with six different probes were employed to search for hotspots showing "ligandability" on protein surfaces, and the DAIS was used for ranking to the possibility of cryptic sites based on estimation of "structural changeability" of protein. The synergistic combination enables to

predict cryptic sites with highly accuracy.



P04-04

Analysis of HS-AFM images of proteins combining MD simulation and machine learning

Katsuki SATO *, Takaharu MORI

Department of Chemistry, Tokyo University of Science

(* E-mail: 1324572@ed.tus.ac.jp)

High-speed atomic force microscopy (HS-AFM) has been widely used for real-time, direct observation of protein conformational changes. Typical resolution of the HS-AFM images is ~ 0.15 nm in the vertical direction, while 2–3 nm in the lateral direction, making the identification of the protein conformation difficult. To solve this problem, we developed a new algorithm combining molecular dynamics (MD) simulation and convolutional neural network (CNN). In the method, MD simulation is first carried out to sample conformational changes of the target protein, and then pseudo-AFM images are generated from each MD snapshot as training data set of CNN. After training the CNN using the pseudo-AFM images, it is applied to the experimental AFM images. To investigate the performance of our method, we selected a protein that undergoes a large conformational change. We performed the MD simulations of the protein to sample various conformations, followed by the analysis of an “artificial” experimental AFM image of the known state. The results demonstrated that the well-trained CNN could identify a conformational state of the target protein. The detailed results will be shown at the poster presentation.

P05-01

Multi-Task Deep Learning using Graph Convolutional Networks for Predicting the Unbound Fraction in Human, Mouse, and Rat Plasma

Harutoshi KATO *, Yuki DOI, Akira SASAKI

DMPK Research Laboratories, Mitsubishi Tanabe Pharma Corporation

(* E-mail: kato.harutoshi@mb.mt-pharma.co.jp)

The unbound fraction in plasma (f_u), calculated from plasma protein binding (PPB), is a crucial pharmacokinetic parameter that have significant impact on pharmacological and toxicological effects of a drug. The computational prediction of f_u is considered to be effective in drug discovery cycle as it can reduce the evaluation period and support drug design. In this study, we aimed to develop a prediction model for the f_u in human, mouse, and rat using multi-task deep learning with graph convolution networks (GCN).

The f_u data of approximately 5000 compounds each in human, mouse, and rat were used as the proprietary internal dataset for model development and validation. In addition, human f_u data of approximately 2000 compounds published were added as an external dataset to expand the chemical space of the training dataset. As regression model, several deep learning models were constructed by single-task and multi-task learning using GCN, and the prediction performance of the models were evaluated using the test datasets.

As a result, we confirmed that the multi-task deep learning model for predicting f_u using the internal dataset (human, mouse, rat) and external dataset (human) as training dataset achieved the best prediction performance based on the coefficient of determination (R^2) on the test dataset, with values of 0.65, 0.78, and 0.89 for human, mouse, and rat, respectively.

P05-02

Enhancing the Reliability of Machine Learning Predictions through Quantitative Evaluation of the Applicability Domain: A Case Study of Multi-Task Prediction Model of Unbound Fraction in Human, Mouse, and Rat Plasma

Yuki DOI *, Harutoshi KATO, Akira SASAKI

DMPK Research Laboratories, Mitsubishi Tanabe Pharma Corporation

(* E-mail: doi.yuuki@ma.mt-pharma.co.jp)

The process of drug development is inherently time-consuming and costly. Therefore, it would be beneficial to employ machine learning (ML) techniques to reduce the time and cost required for each stage of the process by predicting certain outcomes. While the accuracy of ML models (referred to as “activity models” in this study) is typically validated using test datasets during model development, assessing the reliability of predictions in real-world scenarios remains challenging. The inappropriate use of activity models can result in erroneous decisions, thereby undermining the trust in these models and reducing the potential applications of these models. The objective of this study is to develop an “error model” to predict the assurance of an activity model’s output by leveraging metrics that have been demonstrated to be correlated with the reliability of prediction (DA metrics). This approach aims to enhance the reliability of ML predictions.

The activity model utilized was a multi-task deep learning model that predicted the unbound fraction in human, mouse, and rat plasma. The DA metrics employed include Similarity, Local Error, and PREDICTED, as reported in the literature [1, 2]. The error model was developed using these DA metrics with Random Forest to classify whether the prediction error would be within a two-fold range. The probability predicted by the error model, indicating whether the prediction error is within two-fold, is referred to as the Confidence Score. The actual prediction error of the activity model was then compared with the Confidence Score. Furthermore, the impact of DA metrics on the Confidence Score was analyzed using SHAP (SHapley Additive exPlanations).

For compounds with a Confidence Score below 0.5, the proportion within 2-fold error was less than 50%. In contrast, for compounds with a Confidence Score above 0.5, the proportion within 2-fold error was 75% or greater. Additionally, as the threshold of Confidence Score for including the calculation of accuracy

was increased, the R^2 value was increased and RMSE value was decreased. SHAP analysis revealed that an increase in Similarity metrics and a decrease in Local Error metrics were associated with higher Confidence Scores.

These findings indicate that the Confidence Score is valuable tool for enhancing the reliability of activity model's predictions and maximizing their appropriate application in drug development.

[1] Sheridan, R. P., Using random forest to model the domain applicability of another random forest model. *J Chem Inf Model* **2013**, 53, 2837-50.

[2] Sheridan, R. P., The Relative Importance of Domain Applicability Metrics for Estimating Prediction Errors in QSAR Varies with Training Set Diversity. *J Chem Inf Model* **2015**, 55, 1098-107.

P05-03

Development of tools to enhance the extracting process of ADME activity information from the Common Technical Document (CTD)

Masataka KURODA ^{*1, 2}, **Reiko WATANABE**³, **Hitoshi KAWASHIMA**¹, **Yasuhiko HASHIDA**⁴, **Atsuo MATSUEDA**⁴, **Kenji MIZUGUCHI**³

¹National Institutes of Biomedical Innovation, Health and Nutrition

²Mitsubishi Tanabe Pharma Corporation

³Institute for Protein Research, Osaka University

⁴Chinou Jouhou Shisutemu Inc.

(* E-mail: m-kuroda@nibiohn.go.jp)

Many types of data need to be included in the application for a pharmaceutical product. In particular, absorption, distribution, metabolism and excretion (ADME)-related data are abundant and of high quality because they are generated using quality-controlled experimental protocols. The generation, publication and use of such data are important in drug discovery and development to build predictive models of compound profiles or databases. The Common Technical Document (CTD) provides a common format between Japan, the US and the EU, for the registration of a human pharmaceutical product to facilitate documentation and speed up the process. However, while there are standards for the overall document structure, the data representation varies from drug to drug because it is up to the applicant to decide how to write the individual data. In addition, the CTD is distributed in PDF format, and although it can be viewed, data extraction and formatting will be non-trivial. In this study, we developed a dedicated tool, *CTDCurator*, with the aim of reducing the effort required for data formatting in particular and improving the quality of the data, such as the accuracy and consistency of words and units. In the initial stages, tasks such as detecting experimental values and conditions, registering newly appearing words and units, standardizing words and units, and exception handling are always required, and this tool has made it possible to output formatted data in less time than manual work. In addition, the quality improvements achieved by this tool can be applied to other data extraction tasks.

P05-04

Improving the performance of prediction models for small datasets of cytochrome P450 inhibition with deep learning

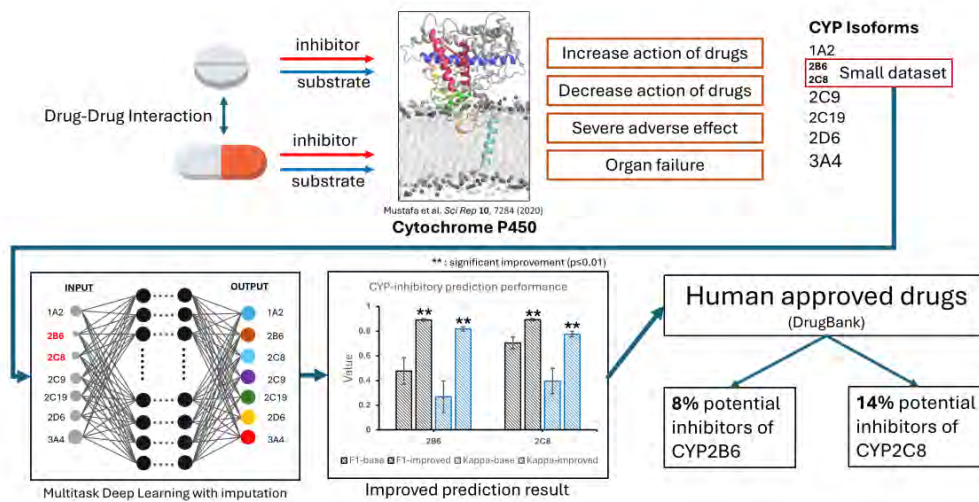
ELPRI EKA PERMADI ^{*1, 2}, **Reiko WATANABE**¹, **Kenji MIZUGUCHI**¹

¹Institute for Protein Research, Osaka University, Japan

²Research Center for Pharmaceutical Ingredients and Traditional Medicine, National Research and Innovation Agency, Indonesia

(* E-mail: elpri@protein.osaka-u.ac.jp)

The human cytochrome P450 (CYP) is the major enzyme that metabolizes drugs, xenobiotics, and toxins. It is known that drug-drug interactions and drug-induced CYP inhibition can lead to adverse events. Thus, identifying potential CYP inhibitors is crucial for safe drug administration, especially for the least known CYP isoforms. However, CYP2B6 and CYP2C8 are currently more difficult to collect sufficient amounts of data than the five major CYPs, i.e., CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4/5, and it made difficult to build a predictive model with sufficient accuracy. This study aims to develop and validate a deep learning model for predicting cytochrome P450 (CYP) inhibition by focusing on isoforms with limited data availability using related data from major CYP isoforms with larger data. Additionally, we explored the inhibitory activity of approved drugs against CYP enzymes based on the constructed prediction models for CYP inhibition. Initially, a comprehensive dataset of around 12 thousand data points targeting seven CYP isoforms was compiled from public databases. Then, we constructed single task, fine-tuning, and multitask models incorporating data imputation of predicted data. We highlighted the potential of multitask deep learning models with predicted data imputation, achieving significant improvement ($p \leq 0.01$) in CYP inhibition prediction compared to the single task model. In addition, three multitask models trained on data imputed with predictions were successfully applied to identify 8% and 14% of human-approved drugs that potentially inhibit CYP2B6 and CYP2C8, respectively. Utilizing multitask learning with imputation of the missing values is useful for improving the performance of the CYP small dataset. Furthermore, discovering the potential inhibitors of CYP2B6 and CYP2C8 may help the practitioner to prevent drug adverse events by avoiding the combination of drugs.



P05-05

Addressing Common Metabolism Problems in Drug Discovery with *in Silico* Methods

Sumie TAJIMA ^{*1}, Daniel A. BARR², Shiori TAKASE¹, Mario ÖEREN², Peter A. HUNT², Tomáš CHRIEN², Tamsin E. MANSLEY², Matthew D. SEGALL²

¹HULINKS Inc.

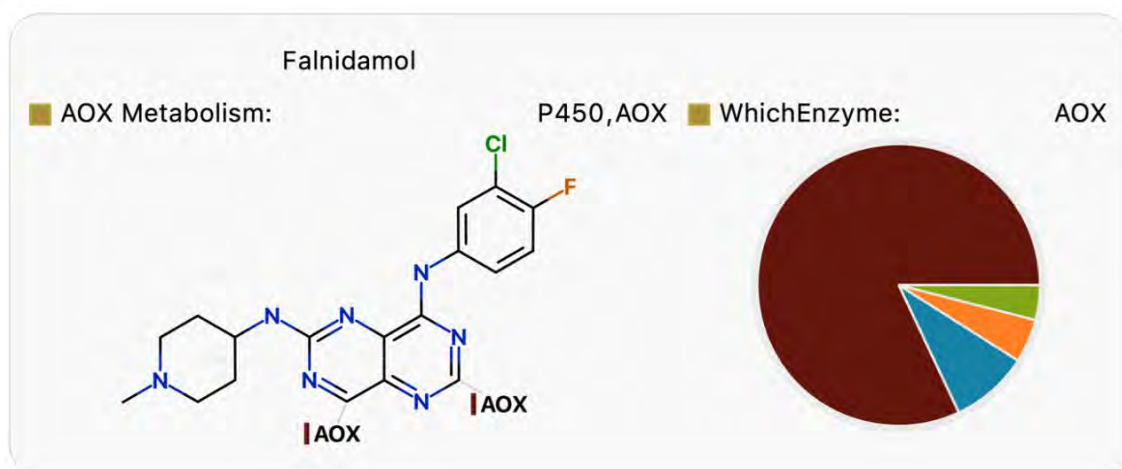
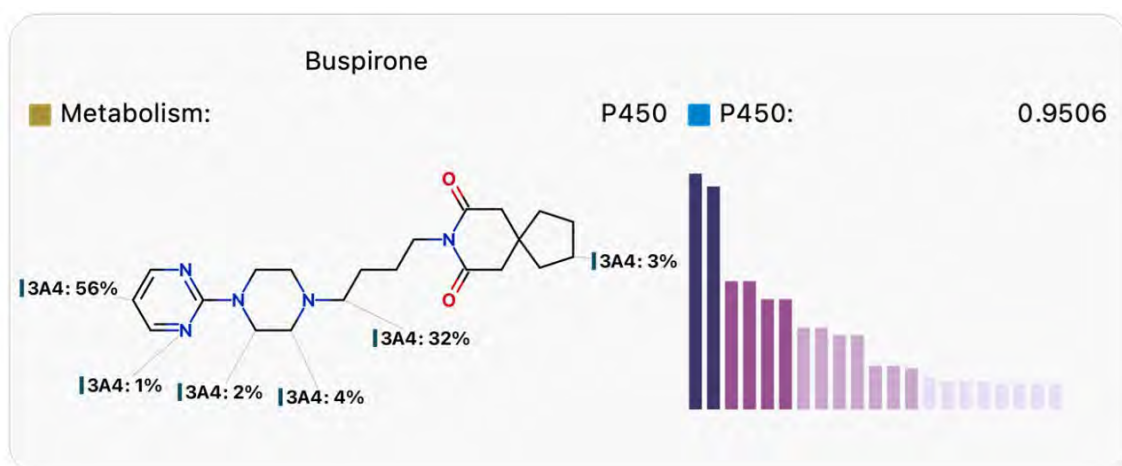
²Optibrium Limited

(* E-mail: tajima@hulinks.co.jp)

In silico metabolism prediction can address critical questions to guide lead optimization. Using several case studies, we demonstrate the application of these models to address design challenges involving metabolic (in)stability, the formation of reactive and/or toxic intermediates, and to mitigate the risk of genetic polymorphisms and drug-drug interactions. In addition, we illustrate how these models can inform the selection of *in vitro* and *in vivo* pre-clinical experiments to avoid surprises in late-stage trials. Furthermore, accurate predictions of metabolite profiles early in the discovery process provide essential guidance for drug design. Optibrium's mechanistic metabolism models cover metabolism by P450, AOX, FMO, UGT, and SULT enzymes¹⁻⁵. By combining these models, metabolic pathway analysis proposes the most likely metabolites with greater precision than other methods, assisting in metabolite identification studies and enabling potentially active, reactive, or toxic metabolites to be identified⁶.

References

- 1) Mario Öeren, Peter J. Walton, James Suri, David J. Ponting, Peter A. Hunt and Matthew D. Segall, (2022) *J. Med. Chem.* **65**(20) pp. 1406-1408
- 2) Mario Öeren, Sylvia C. Kaempfer, David J. Ponting, Peter A. Hunt and Matthew D. Segall, (2023) *J. Chem. Inf. Model.* **63**(11) pp. 3340-3349
- 3) Mario Öeren, Peter J. Walton, Peter A. Hunt, David J. Ponting and Matthew D. Segall, (2021) *J. Comput.-Aided Mol. Des.* **35**(4) pp. 541-555
- 4) Jonathan D. Tyzack, Peter A. Hunt and Matthew D. Segall, (2016) *J. Chem. Inf. Model.* **56**(1) pp. 2180-2193
- 5) Peter A. Hunt, Matthew D. Segall & Jonathan D. Tyzack, (2018) *J. Comput.-Aided Mol. Des.*, **32** pp. 537-546
- 6) Mario Öeren, Peter A. Hunt, Charlotte E. Wharrick, Hamed Tabatabaei Ghomi and Matthew D. Segall, (2023) *Xenobiotica* DOI: 10.1080/00498254.2023.2284251



P05-06

In silico prediction of total clearance, volume of distribution, and half-life with deep learning

Ryoko TERADA *, Reiko WATANABE, Kenji MIZUGUCHI

Institute for Protein Research of Osaka University

(* E-mail: u264709j@ecs.osaka-u.ac.jp)

In drug discovery, in silico screening with artificial intelligence (AI) is being put to practical use. Predicting compound profiles comprehensively based on compound structures in the early stages of drug discovery is expected to contribute further to the efficiency of drug discovery by reducing the cost and time required for experiments. Various parameters describe pharmacokinetics, which change from when a drug is administered to the body until it is excreted. It is known that total clearance (CL_{tot}), volume of distribution in steady state (V_{dss}), and half-life (T_{1/2}) derived from these two parameters have a large effect on the blood concentration profile, and these can be used to estimate the dose and dosing interval. Although the prediction of these parameters has been conducted for a long time, the construction of prediction models is challenging because complex interactions of multiple phenomena affect the value of those parameters, and the amount of publicly available human clinical data is limited. Recently, AI technologies such as deep learning have been rapidly developing, and multi-task learning, fine-tuning, and other techniques have made it possible to improve efficiency by utilizing data similarity.

In this study, we aim to build regression models that can predict CL_{tot}, V_{dss}, and T_{1/2} by deep learning with public data, taking advantage of the similarity and relevance of the three parameters for building a multi-task model. First, after the intensive manual curation, we created the datasets of over 1,200 compounds with chemical structure information and their three parameter values from the ChEMBL database and the literature. Then, we built the prediction models with graph convolution networks, and the constructed single- and multi-task models were evaluated by using RMSE and R square. We also analyzed multi-task models, adding related tasks such as fraction unbound in plasma (f_{u,p}) and mean residence time (MRT). Multi-task models showed better accuracy than single-task models, especially at T_{1/2}. We also found that the combination of tasks in multi-task learning affected their final accuracy differently. We will discuss the appropriate model setting for predicting CL_{tot}, V_{dss}, and T_{1/2}.

P05-07

Unbound Fraction Optimized Method for Predicting Human Pharmacokinetic Clearance: Advanced Allometric Scaling Method and Machine Learning Approach

Yuki UMEMORI *¹, Koichi HANDA², Saki YOSHIMURA¹, Michiharu KAGEYAMA¹

¹Translational Science, Discovery DMPK, Axcelead Tokyo West Partners

²Discovery Science, Drug Discovery Chemistry, Axcelead Tokyo West Partners

(* E-mail: yuki.umemori@axcelead-twp.com)

Accurate prediction of human pharmacokinetic (PK) parameters, particularly human clearance (CL), to estimate human dosing in the drug discovery phase is crucial for enhancing the success rate of drug development. Among the various methods, Single Species Scaling (SSS) using rat PK data and the unbound fraction in plasma (fu) from both human and rat, known as SSS fu rat, has been widely employed. However, the datasets used in the SSS fu rat are limited to about 200 compounds, and leaving the accuracy with external datasets unclear; allometric scaling models have been built without consideration of separated dataset traditionally. Recent advances have also employed machine learning models to predict human CL, leveraging structural information.

In this study, we prepared 200 training and 62 external test compounds that was obtained experimentally. Using these data we investigated the conventional SSS rat; a new method, the Unbound Fraction Optimized SSS (UFO SSS) fu rat, which predicts using the SSS rat method for compounds with low fu and a newly calculated allometric equation for remaining compounds; a Random Forest (RF) machine learning model using ECFP4; a consensus model of UFO SSS and RF. We first analyzed the training dataset of 200 compounds and found that compounds with an fu value of less than 0.03 in either human or rat exhibited poor predictive accuracy using the SSS fu rat. This value (0.03) was used as the threshold value of fu in UFO SSS rat. To investigate our approach, we compared the predictive performance of each model with 62 external test datasets. For SSS fu rat, UFO SSS fu rat, RF, and the consensus model, the percentages of compounds within 2-fold error were 37.1%, 41.9%, 40.3%, and 41.9%; the percentages of compounds exceeding 5-fold error were 29.0%, 21.0%, 21.0%, and 16.1%; the Geometric Mean Fold Error were 5.5, 4.8, 2.6, and 2.6, respectively. Hence, we conclude that the consensus model achieved the best overall performance, demonstrating its superior predictive accuracy.

We can stress that these models were validated by the external dataset of 62

compounds, presenting a novel approach that enhances the accuracy of human clearance predictions in the drug discovery phase. By combining an allometric method optimized with unbound fraction data and machine learning techniques, our approach provides a more reliable prediction method for human pharmacokinetics, ultimately contributing to the success of drug development.

P06-01

Cell State Analysis of Immune Cells in the Tumor Microenvironment with Deep Learning

Jiaxin LI*¹, Artem LYSENKO^{2,3}, Tatsuhiko TSUNODA^{2,3}

¹Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo

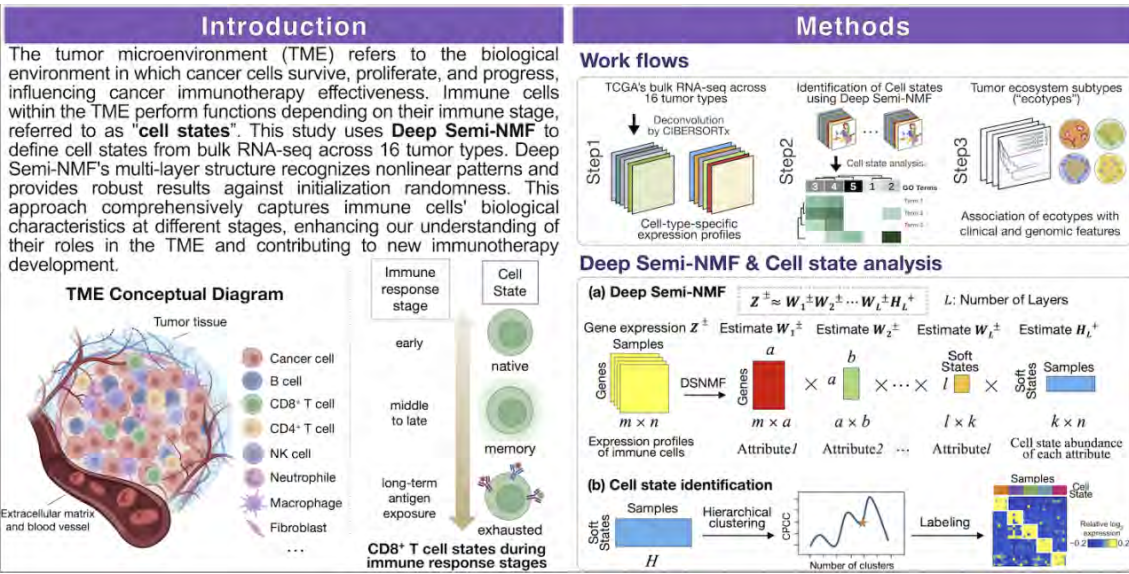
²Department of Biological Sciences, Graduate School of Science, The University of Tokyo

³RIKEN Center for Integrative Medical Science
(* E-mail: 1530221035@edu.k.u-tokyo.ac.jp)

Immunotherapy is a revolutionary advancement in cancer treatment, targeting the body's immune system to fight malignancies more effectively. However, its success varies among patients due to the complexity of the tumor microenvironment.

EcoTyper, a machine learning tool, uses gene expression data to identify distinct immune cell states, offering insights into the tumor's immunological landscape. This study enhances EcoTyper with Deep Semi-NMF, refining cellular process analysis by extracting features at multiple hierarchical levels. This method is effective for complex cancer datasets, like those from 6000 patients across 16 cell types in TCGA. Deep Semi-NMF's layered approach abstracts general cellular characteristics in the first layer, then delves into specific immune cell interactions in subsequent layers. This allows for more nuanced identification of cell states, improving immune response classification. More immune cell states are identified compared to traditional NMF techniques. Each new cell state is correlated with specific clinical outcomes, linking molecular insights to therapeutic impacts. This is crucial for understanding how different immune states affect immunotherapy efficacy, providing pathways to enhance patient response rates.

The insights from the Deep Semi-NMF enhanced EcoTyper model extend into areas of immune evasion and personalized treatment strategies. By understanding how certain immune cell states correspond to patient susceptibility or resistance to immunotherapy, clinicians can tailor treatments to leverage the immune system more effectively, offering hope for improved survival rates and quality of life for patients.



P06-02

A Novel Endometrial Cancer Patient Stratification Considering ARID1A Protein Expression and Function with Effective Use of Multi-omics Data

Junsoo SONG *¹, Ayako UI², Kenji MIZUGUCHI¹, Reiko WATANABE¹

¹Laboratory for Computational Biology, Institute for Protein Research, Osaka University

²Institute of Development, Aging and Cancer, Tohoku University

(* E-mail: u920213a@ecs.osaka-u.ac.jp)

Conventional patient stratification based solely on mutations or mRNA expression often fails to reflect functional activity accurately. Since proteins serve as the cell's functional molecules, their expression directly correlates with cellular states. Consequently, stratifying patients based on protein expression offers significant advantages; however, several challenges remain. First, the disparity in data volume across different omics fields complicates efficient data utilization. Proteomics suffers from limited data availability due to its technical complexity and relatively low throughput. Second, neither mutation nor protein expression alone can serve as a perfect indicator of functional activity for certain proteins such as ARID1A. ARID1A, a DNA-binding subunit of the SWI/SNF family, is the most frequently mutated gene in this complex, particularly associated with ovarian and endometrial cancers. A comprehensive consideration of ARID1A mutations, activity, and protein expression is essential for developing therapeutic strategies. Research indicates that most ARID1A loss-of-function mutations are heterozygous, and protein expression is detectable in all samples. Thus, additional direct information about ARID1A activity is necessary for more precise patient stratification based on ARID1A activity.

To address these issues, we propose an innovative patient stratification strategy. We developed a machine learning model to supplement insufficient protein expression data of ARID1A. This model was trained using publicly available multi-omics data, integrating information from multiple databases. By estimating the transcriptional regulation of genes directly targeted by ARID1A, we inferred the activity of ARID1A in tumor tissue. With fully supplemented proteomics data and activity labels, patients were stratified into three groups (High, Mid, Low) considering both function and protein expression of ARID1A in a patient. We identified differentially expressed genes (DEGs) between groups and Gene Set Enrichment Analysis (GSEA) revealed that our method highlights transcriptional variations of tumor immune microenvironment, which were ambiguous with stratification based on mRNA expression. Further investigation

of the extracted DEGs may reveal genes that help predict the effectiveness of immune checkpoint inhibitor (ICB) treatment in ARID1A-deficient patients.

P06-03

Single-Cell Transcriptome Analysis Reveals Roles of GABA Receptors in the Connectivity of Dorsal-Ventral Motor Neurons in *C. elegans*

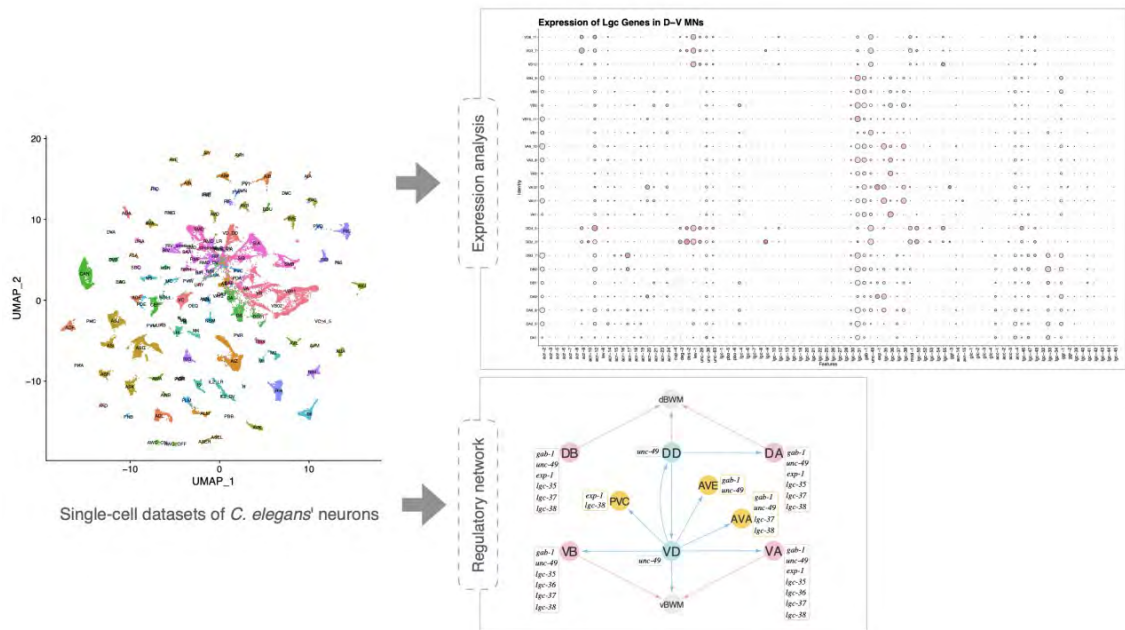
Xingran WANG *, Kosuke HASHIMOTO, Kenji MIZUGUCHI

Laboratory for Computational Biology, Institute for Protein Research, Osaka University

(* E-mail: u323248e@ecs.osaka-u.ac.jp)

The cys-loop receptors, encoded by *lgc* genes, play a pivotal role in chemical neurotransmission and are targets for drugs treating neurological disorders. While mammals possess approximately 45 *lgc* genes, *C. elegans* with a simple nervous system has 102 *lgc* genes. Many of these genes are absent in the human genome, posing intriguing questions about their functions. Among these 102 *lgc* genes, only about 20% have been characterized to date, leaving the functions of the majority yet to be elucidated.

In this study, we employed single-cell transcriptome analysis to delineate the expression patterns of cys-loop receptors within *C. elegans* neurons. The dataset encompasses 70,296 neurons from L4 stage larvae, representing all 118 canonical neuron classes. Using R Seurat package, we discovered that *lgc* genes are primarily expressed in motor neurons, with particularly high levels in dorsal-ventral motor neurons (D-V MNs). Notably, D-V MNs show elevated expression of the seven GABA receptors identified to date in *C. elegans*, which include both inhibitory and excitatory subtypes. Each D-V MN type demonstrates distinct expression profiles among these GABA receptor subtypes. Moreover, both excitatory and inhibitory GABA receptors can be co-expressed on the same D-V MN, adding complexity to their roles in locomotion beyond traditional neuron-muscle junctions. By integrating the expression and the connectivity of D-V MNs, we found that mixed-type GABA receptors enable interactions with defecation interneurons (AVL and DVB). These interneurons regulate the defecation motor program by modulating D-V MNs, which in turn affects body wall and intestinal muscles. This study highlights the intricate roles of GABA receptors in neural connectivity and lays a robust foundation for future research to elucidate the underlying mechanisms of GABAergic regulation in motor behaviors.



P06-04

Impact of Intramolecular Hydrogen Bonds on Permeability Glycoprotein Mediated Transportation

Yulong GOU ^{*1}, Suyong RE², Kenji MIZUGUCHI¹, Chioko NAGAO¹

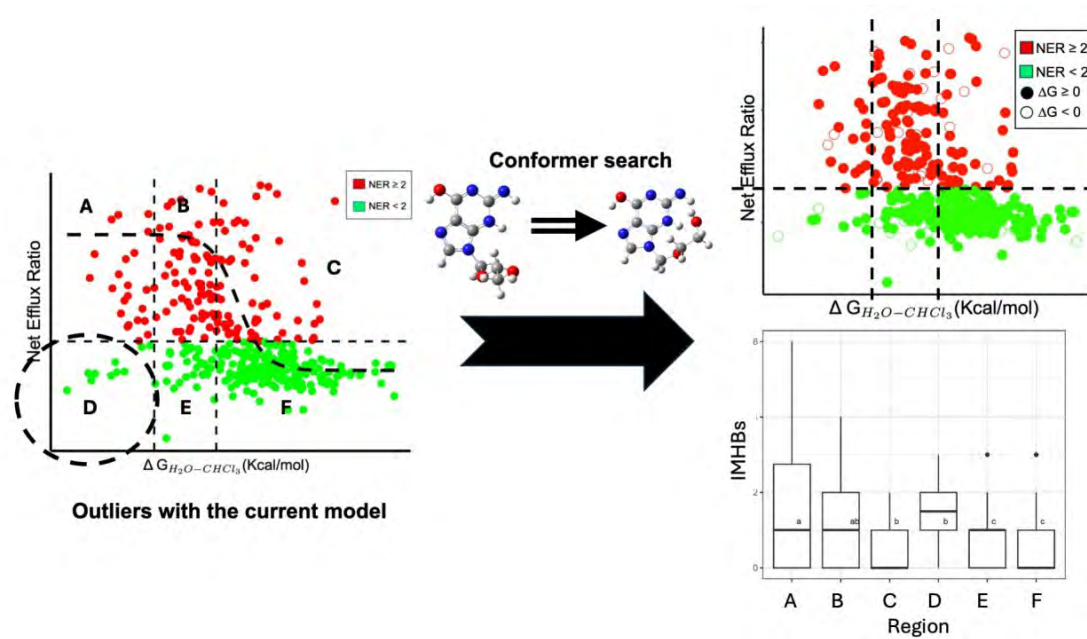
¹Laboratory for Computational Biology , Osaka University, Insitute for Protein Research

²National Institute of Biomedical Innovation, Health and Nutrition

(* E-mail: gouyl96@protein.osaka-u.ac.jp)

Permeability glycoprotein (P-gp) is widely expressed in many cells related to pharmacokinetics and physical barriers in human body. P-gp pumps diverse substances and toxins out of cells through ATP-mediated conformational changes. Overexpression of P-gp lead to multidrug resistance. The efflux ratio (ER) is often used to determine the ability of P-gp-mediated efflux in drug discovery. Due to the lack of large-scale data, a de novo computational method based on solvation free energy was developed¹, which is useful for predicting whether a drug compound is a P-gp substrate or not. In this method, the net ER is related to the solvation free energy by a sigmoidal function. Recently, we evaluated this method, using in-house data set of 397 compounds, and found that there are non-negligible outliers, although this method provides a good result on several compounds². The outliers typically involve the compounds with the large solvation free energy, the low efflux ratio, and the high potential to donate hydrogen bond. Since the intermolecular hydrogen bonds (IMHBs) can affect the solvation free energy, we re-evaluated the method by rigorously considering the IMHBs. We employed the RDKit³ to generated 100 conformers for each of 397 compounds and obtained the optimal structures. Many of the obtained structures contain IMHB regardless of substrates or non-substrates. Notably, the outliers have more IMHB than other compounds. We found that accounting for IMHBs in the calculation of solvation free energies decrease the values compared to the original evaluation without the conformer search, which slightly reduces the number of outliers. Considering the non-negligible outliers, explicitly accounting for the specific interactions between the compounds and P-gp could further improve the prediction model.

1. Gunaydin H. et al. ACS Med.I Chem. Lett. 2013, 4 (1), 108-112.
2. Gou Y. et al. ACS Med.I Chem. Lett. 2023, 15 (1), 54-59.
3. RDKit: Open-source cheminformatics; <http://www.rdkit.org>



P06-05

Improved Method of Predicting Protein Allosteric Site Based on Atomistic Bond-to-bond Interaction by Using GNN

Chaowen OU *, Takashi ISHIDA

School of Computing, Tokyo Institute of Technology

(* E-mail: ocw24680@gmail.com)

Allosteric effect is a fundamental aspect of protein function, where the binding of a ligand at a site distinct from the active site—known as an allosteric site—induces a conformational change that affects protein activity. This mode of regulation presents a promising avenue for drug discovery, as allosteric modulators offer the potential for high specificity and reduced side effects compared to orthosteric drugs, which target the active site directly. However, identifying allosteric sites is challenging due to the complexity and dynamic nature of protein structures.

Traditional methods for discovering allosteric sites have relied on extensive experimental techniques, such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. While these methods provide valuable insights, they are often time-consuming and resource-intensive. Recent advances in machine learning, particularly deep learning, have revolutionized the field of computational biology. Graph convolutional networks (GCNs) have emerged as powerful tools for modeling complex interactions within biological systems.

The key innovation of our method lies in the use of energy-weighted graphs to model the strength and distribution of atomic interactions within proteins. By focusing on the energetic properties of bonds, we aim to identify regions of the protein structure that are critical for allosteric signaling. Our model integrates features from both the protein graph and candidate pocket features, allowing for a comprehensive analysis of potential allosteric sites.

To train and validate our model, we utilized the Allosteric Site Database (ASD), a curated collection of experimentally verified allosteric sites. We applied rigorous preprocessing steps to ensure the quality and relevance of the data, resulting in a dataset that is well-suited for machine learning applications. Our approach addresses the challenge of class imbalance—a common issue in biological datasets—by employing a sampling strategy that enhances the representation of positive allosteric site samples.

The evaluation of our method demonstrates its superior performance compared to existing models, such as PASSer2.0, across various metrics, including accuracy, precision, recall, and F1 score. The proposed method not only

achieves higher accuracy but also significantly improves recall, indicating its enhanced ability to detect true positive allosteric sites.

In conclusion, our research presents a novel and effective approach for predicting protein allosteric sites using deep learning techniques. By integrating graph-based protein representations with energy-weighted bond interactions, we provide a powerful tool for identifying potential drug targets. This work advances the computational prediction of allosteric sites and underscores the importance of considering atomic-level interactions in understanding protein function and regulation.

P06-06

Development of RNA velocity method using numerical integration of ordinary differential equations

Yuki KOBAYASHI *, Tomonari MATSUDA

Department of Environmental Engineering, Graduate School of Engineering,
Kyoto University

(* E-mail: kobayashi.yuki.58w@st.kyoto-u.ac.jp)

RNA velocity is one of the trajectory inference methods of single-cell transcriptome analysis, which uses the number of mature/immature mRNA and kinetics about transcription/splicing/degradation of mRNA. RNA velocity method has the potential to recover real time series by describing the dynamics with differential equations. However, any current methods have their own challenges in dealing with the differential equations: requirement of analytical solutions, machine learning of the differential equations and so on. To solve the problems, we develop a method using numerical integration of ordinary differential equations. It is expected that more realistic dynamics can be used even when analytical solutions are difficult to obtain. In addition, simulations also use numerical integration but have often been performed qualitatively. However, our method enables quantitative evaluation and modification of simulations, and it is expected to promote knowledge-based in silico analysis.

Integration of observations and numerical simulations is named as data assimilation, which has been developed in geosciences. In this study, we formulated the method using the four-dimensional variational method (4D Var), which can integrate observations from multiple times at once. Here, we derive 4D Var using maximum likelihood estimation of normal distribution for observation errors.

In this study, we assume that the gene expression kinetics is the same as current RNA velocity methods and use pseudo data generated from the kinetics of a single gene. The kinetics of mRNA are described by the parameters of transcription rate, degradation rate, expected number of mature/immature mRNA, and time interval for each cell. Applying 4D Var, we should decide numerical integration method, observation operator, and optimization method according to the experimental setting. The numerical integration method is used as the fourth order Runge–Kutta scheme. The observation operator is a matrix that returned the number of mature/immature mRNA. In addition, Observation

data are generated by adding a normally distributed random number into the true values of mature/immature mRNA. Optimization is performed by increment method and conjugate gradient method. As a result, the curve fitting on the gene expression space stretched by mature/immature mRNA was successful, and it was confirmed that the other unobservable parameters could be successfully estimated like current methods.

One of the current issues of our methods is the high initial value dependence of the analytical results. Therefore, optimization methods and initial value setting using the analytical solution will be developed. In addition, mRNA dynamics assumed in this study are very simple, but the superiority of our method is to handle complex dynamics. Therefore, we will model the dynamics of gene expression in more detail and our methods can be performed more accurately than with current RNA velocity methods.

P06-07

Compound Retrosynthesis Analysis Using Consensus Estimate

Akira SHINOHARA *, Takashi ISHIDA

Department of Computer Science, School of Computing, Tokyo Institute of Technology

(* E-mail: shinohara.a.ae@m.titech.ac.jp)

Research on compound synthesis methods is one of the important themes in organic chemistry. Retrosynthesis analysis, a method that designs synthetic routes by repeatedly performing chemically rational cleavages until the target compound becomes easily and inexpensively obtainable compounds, is a very useful analysis for designing synthetic routes. Therefore, to improve the prediction accuracy of each step in multi-step retrosynthesis analysis, many studies have been conducted on single-step retrosynthesis analysis, which only considers one-step retrosynthesis reactions. Single-step retrosynthesis analysis can be broadly categorized into two types based on the use of templates: "template-based" and "template-free" methods.

Template-based methods perform well for predictions that reference templates, but they lack generalization ability and require time and effort to create templates. As a result, since 2017, the development of template-free methods using machine learning has been frequently conducted. While template-free methods resolve the disadvantages of template-based approaches, they tend to have slightly lower prediction accuracy. Furthermore, within each method, there are differences in techniques such as the use or non-use of Atom-Mapping, the utilization of SMILES, the use of substructures, and the use of graph structures. Both methods have their pros and cons, making it difficult to develop a model that serves as a compromise between the two.

In this research, we propose a method that utilizes consensus estimate to improve the accuracy of single-step retrosynthesis analysis. we select several template-based and template-free models and obtain their respective prediction results. Then, for the compounds predicted by all models, we adjust their rankings by methods such as taking the average of their ranks across all models, and perform re-ranking. When comparing the prediction accuracy obtained through this process with that of the original models and the most accurate model reported in the literature, we confirmed an improvement in accuracy.

P06-08

Development of docking simulation with high-speed graph neural network scoring function

Kohei HOASHI ^{*}, Takashi ISHIDA

Department of Computer Science, School of Computing, Tokyo Institute of Technology

(^{*} E-mail: hoashi.k.aa@m.titech.ac.jp)

Docking simulation is a primary method for narrowing down candidate compounds when developing new drugs. It predicts the binding poses between a protein and a ligand and their binding affinity. The binding pose is optimized in the simulations using search algorithms to minimize its binding affinity. The binding affinity is generally calculated using a scoring function from the binding poses.

There are now two types of scoring functions: the classical method and machine learning (ML)- based. The classical methods use a manually designed liner equation, including several terms representing chemical and physical properties involved in binding. In contrast, ML-based methods directly output predicted binding affinity via ML models. Accuracy is one of the scoring function's most critical factors, but execution speed is also essential because it is executed many times during docking. Many ML-based methods have been reported to offer superior prediction accuracy. However, classical methods are considered faster due to their lower computational requirements.

Although using ML-based scoring methods will improve the accuracy of docking simulation, existing docking simulations use a classical scoring method because of the calculation speed. Recently, several ML-based scoring methods that are fast enough for docking simulation have been proposed. GenScore is an ML-based method using a mixture density network. The mixture density network makes it possible to express the score for each pair without using ML. The parameters used in the mixture density network are learned by a graph neural network.

In this research, we used GenScore as a scoring function and AutoDock Vina as a docking engine because it is one of the most widely used docking tools. Docking using the proposed methods involves two steps. First, GenScore calculates the parameters of the mixture density network using the protein and

ligand structures as input. This step is performed only once for each new protein-ligand pair. Subsequently, the calculated parameters are loaded into AutoDock Vina, and docking is executed using GenScore's score calculation.

We use the Posebusters benchmark dataset for evaluation, which has contained complexes from PDBbind since 2021. This dataset doesn't overlap with GenScore's training dataset. This evaluation differs from GenScore in terms of timing. They used GenScore for re-docking and re-ranking. However, we use it for docking. It often causes the appearance of binding poses far away from the native pose.

P06-09

Investigation of the trends and the potential in drug development for rare and intractable diseases based on the KEGG NETWORK

Mao TANABE *, Makoto HIRATA, Ryuichi SAKATE

Laboratory of Rare Disease Information and Resource Library, Center for Intractable Diseases and ImmunoGenomics (CiDIG), National Institutes of Biomedical Innovation, Health and Nutrition

(* E-mail: mtanabe@nibiohn.go.jp)

For many rare and intractable diseases (RIDs), the pathophysiological mechanisms still remain unexplained and there are few drugs for the treatment of these diseases. An understanding of approved drugs is important to improve drug development. In DDrare (Database of Drug Development for Rare Diseases) [1], the targets of drugs in clinical trials are mapped to the KEGG PATHWAY to be grasped on molecular networks. In this study, to understand the relationship between drug targets and disease genes more precisely, we mapped them to the KEGG NETWORK (networks) defined as functionally meaningful segments of pathways [2,3]. We found that disease genes tended to be included in networks characteristic for each disease group, whereas drug targets were mapped to networks common to many disease groups. The number of drugs targeting the networks containing disease genes was small in every disease group. However, because several studies have recently addressed that the drugs that target proteins with direct genetic evidence of disease association and their molecular partners are more likely to be approved [4,5], we confirmed the results using the KEGG NETWORK and integrating the risk genes obtained from the latest GWAS data. The results were clearer and more detailed than those of previous studies. Considering the findings in literature, the fact that a drug targeting a network with disease genes has been approved suggests the perturbation of the network by specific causes, but not always by the disease genes in the network. The knowledge acquired in this study shows the possibility that the perturbed network, and in turn precise targets of drug for a RID can be found by obtaining GWAS data, other omics data, or the information about environment factors of the disease, and by mapping them to the functionally meaningful segments of pathways such as the KEGG NETWORK.

[1] DDrare: Database of Drug Development for Rare Diseases,
<https://ddrare.nibiohn.go.jp/>

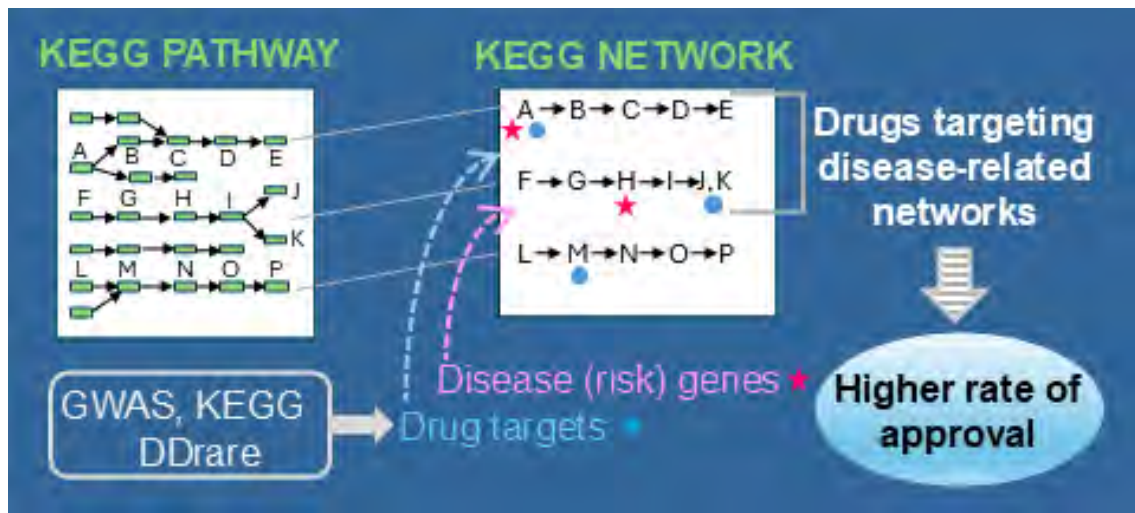
[2] KEGG: Kyoto Encyclopedia of Genes and Genomes,

<https://www.kegg.jp/kegg/>

[3] Tanabe M., Hirata M., and Sakate R., Trends in drug development for rare and intractable diseases based on the KEGG NETWORK. NAR Molecular Medicine. 2024; 1

[4] Nelson M.R., Tipney H., Painter J.L., *et al.* The support of human genetic evidence for approved drug indications. Nat Genet. 2015; 47:856-60.

[5] Okada Y., Wu D., Trynka G., *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 2014; 506:376-81.



P06-10

Prediction of medium components for bacteria using deep Learning

Ryui SATO *, Takuji YAMADA

Department of Life Science and Technology, Institute of Science Tokyo
(* E-mail: sato.r.bh@m.titech.ac.jp)

Isolation and culture of microorganisms, especially bacteria, is an important process for the application and use of the species. However, most of the known bacteria are either difficult to culture or difficult to isolate and culture. With the recent development of metagenomic analysis technology, the species and functions of bacteria have been inferred by extracting DNA of all bacteria from the environment without isolation and culture. While bacterial genome information can now be obtained, the establishment of culture methods is still an urgent task for research and utilization of unknown bacteria or bacteria with low abundance in samples.

The objective of this study is to identify bacteria-selective culture media from genomic information of bacteria with unknown culture conditions. This study is expected to elucidate the selection and estimation of culture conditions and the relationship between bacterial metabolism and culture medium components.

In this study, we are creating a database that links metabolic pathways based on bacterial culture conditions and functional gene information obtained from genome information. We are also developing a deep learning-based method for predicting culture media components that visualizes the correspondence between bacterial gene information and medium components as inputs. In previous studies, bacteria were treated only in terms of phylogenetic and ecological similarity of species based on 16SrRNA, and therefore, the metabolic pathways of bacteria and the genes involved in them were not mentioned. The proposed method differs from previous studies in that it focuses on the functional genes of bacteria. This allows us to predict the culture medium composition with respect to bacteria that have the same metabolic function but are at different positions in the phylogenetic tree, which cannot be predicted by phylogenetic and ecological similarity of species.

P06-11

Elucidation of Stabilization Mechanisms of Intrabodies Based on Statistical Thermodynamics

Koki HATTORI ^{*1}, **Yuto HOSHI**¹, **Hiroshi YUKAWA**^{1, 2}, **Satoshi OGASAWARA**¹, **Masahiro KINOSHITA**^{1, 3}, **Takeshi MURATA**¹, **Satoshi YASUDA**¹

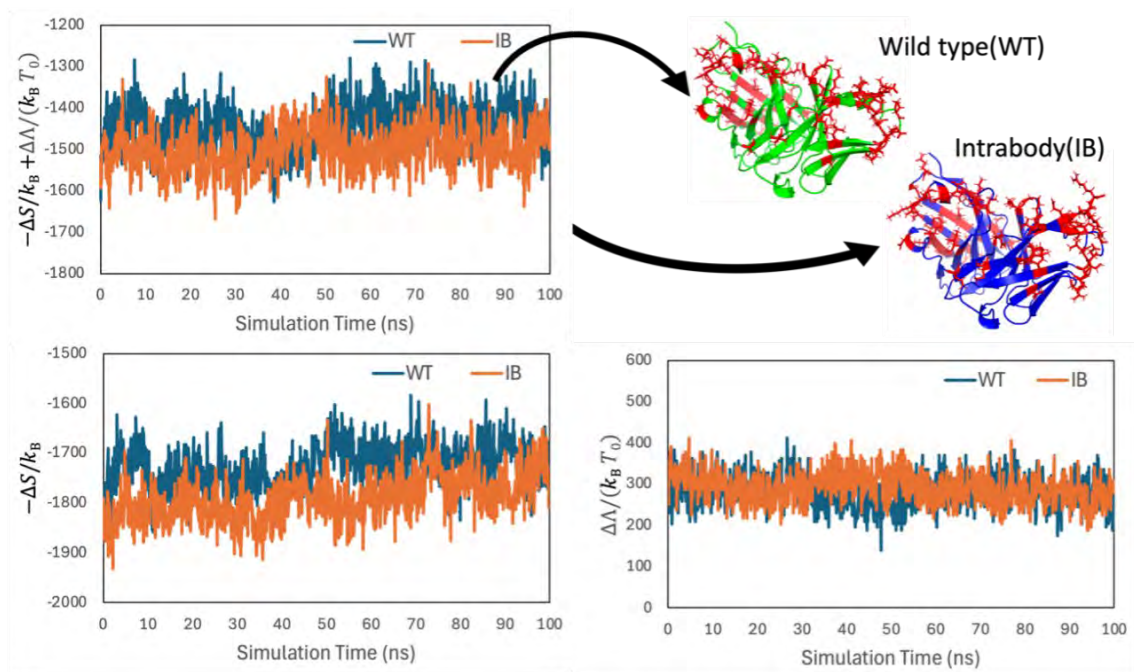
¹Graduate School of Science and Engineering, Chiba University

²National Institutes for Quantum Science and Technology

³Institute of Advanced Energy, Kyoto University

(* E-mail: 23wm2615@student.gs.chiba-u.jp)

Single-chain variable fragments (scFvs), which are recombinant antibody fragments consisting of only the variable light chain and variable heavy chain domains covalently connected to one another by a short polypeptide linker, are expected to be used in pharmaceuticals and other applications. Recently, stabilized multimutant scFvs, known as intrabodies, constructed using knowledge-based methods have been reported, but the details of their stabilization mechanisms have remained unclear. In our earlier study, Kinoshita et al. developed the free energy function (FEF) that can accurately evaluate the thermal stability of proteins. Our FEF consists of the entropy term S and the energy term Λ . S represents the water entropy change arising from the excluded volume effect and is calculated using an original method based on statistical thermodynamic theory. Λ , on the other hand, represents the energy change due to dehydration and is calculated by counting hydrogen bonds within proteins and between water and proteins. In this study, we have elucidated the stabilization mechanism of intrabodies by calculating the free energy changes upon folding using FEF for four intrabodies and their wild-type structures. The structure models of the intrabodies and the wild type were constructed using AlphaFold2. Fluctuations of the antibody were taken into account using molecular dynamics simulations. The results of this study are as follows: Intrabodies have a larger water entropy gain upon folding than the wild types. This indicates that closer packing of side chains is achieved by the amino acid mutations of intrabodies. We have also identified which amino acid mutations among the multimutants play a particularly important role in stabilization. Based on these findings, we are now investigating further stabilizing mutants of intrabodies.



P07-01

Development of Pre-Fragment-Based MMP Analysis

Toshiaki WATANABE *, Osamu IWAMOTO, Hiroyuki HAKAMATA

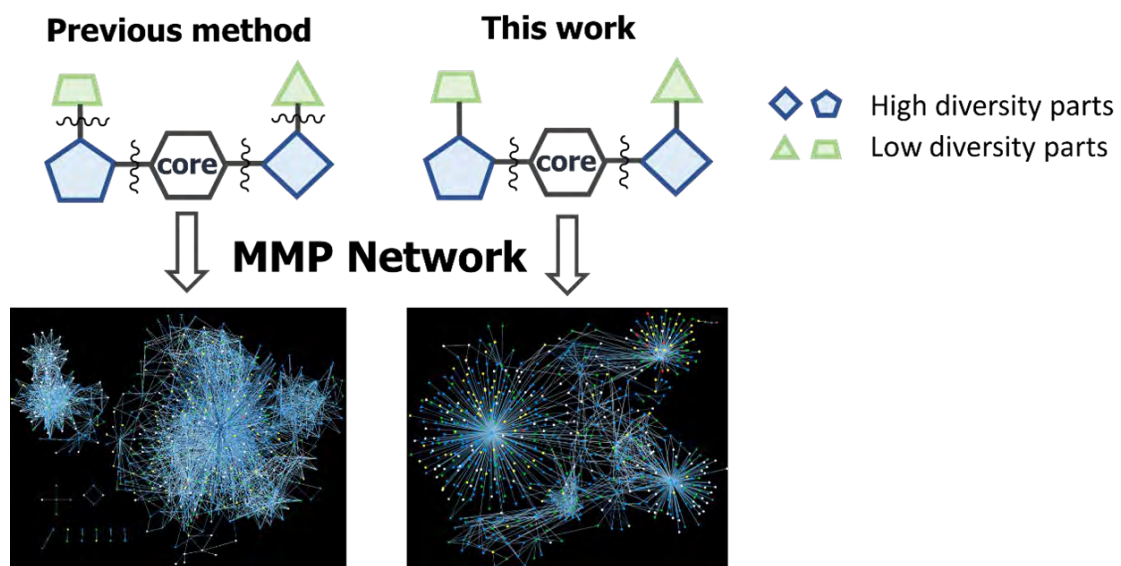
R&D Division, DAIICHI SANKYO CO., LTD.

(* E-mail: toshiaki.watanabe@daiichisankyo.com)

Matched Molecular Pairs (MMPs) are defined as pairs of compounds that differ at only one specific point in their molecular structure.^{1,2} Structure-activity relationship (SAR) analysis can be performed by examining the differences in pharmacological activity and ADMET properties associated with the structural changes between these compound pairs. However, MMPs are typically generated by fragmenting molecules according to uniform rules, which may also cleave less frequently transformed (less diverse) moieties. This often results in the generation of a large number of MMPs, which complicates the post-processing steps required to extract meaningful insights. To address this issue, we propose a method that enables more efficient SAR analysis by allowing cleavage at arbitrary positions based on synthetic synthons. This approach improves the relevance and manageability of the generated MMPs.

To demonstrate the applicability of our method, we conducted two case studies: (1) application to patent analysis and (2) application to modalities other than small molecules. In these drug discovery projects, our approach facilitated more efficient SAR analysis, improved patent key compound prediction, and enabled SAR visualization of medium-sized molecules.

1. Ed Griffen, Andrew G. Leach, Graeme R. Robb, and Daniel J. Warner. Matched Molecular Pairs as a Medicinal Chemistry Tool, *J. Med. Chem.*, 2011, 54, 22, 7739-7750
2. Mathias Wawer and Jürgen Bajorath. Local Structural Changes, Global Data Views: Graphical Substructure–Activity Relationship Trailing. *J. Med. Chem.*, 2011, 54, 8, 2944-2951



P07-02

Discovery of a new histone deacetylase 8 inhibitor using machine learning-aided drug screening

Yasunobu YAMASHITA *, Atika NURANI, Yuuki TAKI, Yuri TAKADA, Yukihiro ITOH, Takayoshi SUZUKI

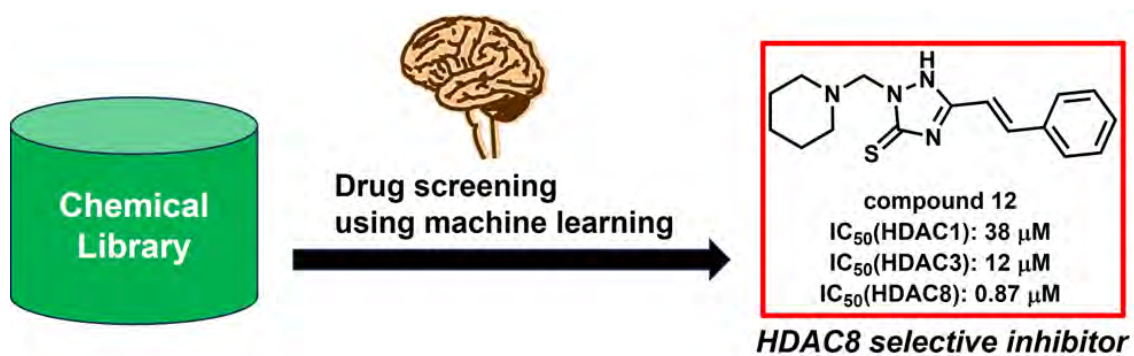
SANKEN, Osaka University

(* E-mail: yyamashita@sanken.osaka-u.ac.jp)

Histone deacetylase 8 (HDAC8) is a zinc-dependent enzyme that catalyzes the deacetylation of non-histone proteins and is involved in cancer development. HDAC8 inhibitors are promising candidates as anticancer agents. However, most reported HDAC8 inhibitors contain a hydroxamic acid moiety, which is often associated with mutagenicity. Therefore, we used machine learning for drug screening to identify non-hydroxamic acid HDAC8 inhibitors. In this study, we established a prediction model based on the random forest (RF) algorithm for screening HDAC8 inhibitors, as it exhibited the best predictive accuracy on a training dataset augmented with data generated by the synthetic minority over-sampling technique (SMOTE). Using the trained RF-SMOTE model, we screened the Osaka University library for compounds and selected 50 virtual hits. However, none of these initial 50 hits showed HDAC8-inhibitory activity. In a second screening, utilizing an RF-SMOTE model retrained with a dataset including these 50 inactive compounds, we identified non-hydroxamic acid compound 12 as an HDAC8 inhibitor with an IC₅₀ of 0.87 μ M. Interestingly, its IC₅₀ values for HDAC1 and HDAC3 inhibitory activity were 38 μ M and 12 μ M, respectively, demonstrating that compound 12 has high selectivity for HDAC8. Through the use of machine learning, we expanded the chemical space for HDAC8 inhibitors and identified non-hydroxamic acid 12 as a novel HDAC8 selective inhibitor.

Reference

1) A. Nurani, Y. Yamashita, Y. Taki, Y. Takada, Y. Itoh, T. Suzuki. Chem. Pharm. Bull. 2024, 72, 173–178.



P07-03

Open Source Program Github and Its Application in Drug Discovery

Kiyoshi HASEGAWA *, Yuya SEKI, Yu LIU, Yukiyo ITO

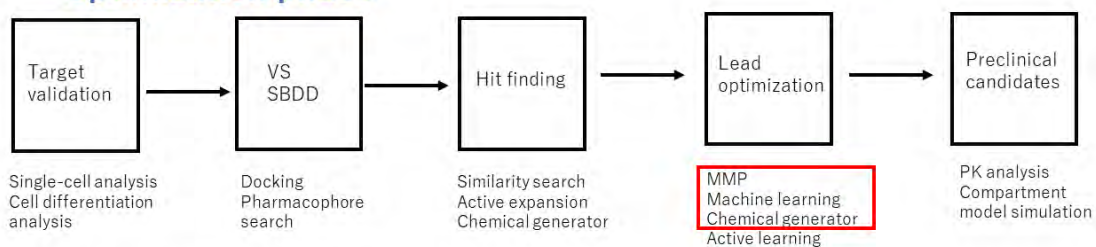
Division of Informatics Promotion, TECHNOPRO R&D company
(* E-mail: Hasegawa.Kiyoshi@technopro.com)

GitHub is an open-source program aimed at validating the paper and further improving its algorithm when researchers have submitted to journal. It was recently recognized that Github programs are useful tools in drug discovery toward pre-clinical phase. For example, in cheminformatics, this includes generating compound structures within proteins and predicting the activity and physical properties of compounds with chemical interpretations. By combining these two approaches, it is possible to obtain new chemical structures with improved compound profiles. In bioinformatics, this includes single-cell analysis and cell differentiation analysis with time series. Additionally, it encompasses the generation of sequences for active antibodies and peptides.

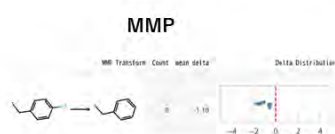
We focus on the optimization phase to customize MMP (matched molecular pairs), machine learning and chemical generator tools in cheminformatics field. First of all, all possible MMP transformation rules are extracted from RDKit and Pandas libraries. Then, each MMP transformation is validated on two aspects. One is how to fit the prediction values from chemical graph convolution model to the observed values. DeepChem library is used for building chemical graph convolution model. Second is whether the shaded colors derived the attention values on chemical structure is matched to SAR (structure-activity relationships) and chemist's intuitions. The attention values are calculated from the graph convolution weights and the contributions of chemical fragments to activities. Third, if above two criteria would be passed, the remaining MMP transformations are processed to chemical generator framework. Keeping the core structure, MMP transformations are used to generate possible chemical libraries.

This optimization strategy will be applied to other ADME properties other than hERG inhibition data. Also, this strategy is further extended to multi-optimization solutions when the MMP transformations from possible ADME properties would be prepared in advance.

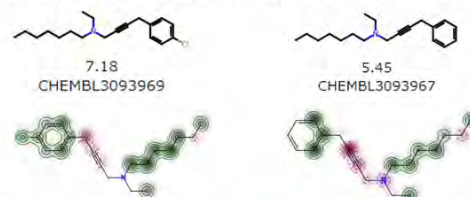
MMP, Machine learning and Chemical generator in lead optimization phase



(Example) public hERG data



Chemical graph convolution with attentions



One MMP transformation is selected.

This MMP transformation is valid.

Chemical generator using the valid MMP transformations

P07-04

Development of accurate *in silico* screening protocol based on protein structural fluctuation and drug binding mode

Hiroto TERADA *, Kei MORITSUGU

Graduate School of Science, Osaka Metropolitan University

(* E-mail: sg24479j@st.omu.ac.jp)

Structure-based *in silico* screening is widely used as an important tool in drug discovery. It is often the case that a single crystal structure is used as a receptor, then questions arise about how suitable the receptor structure is for a diverse compound library and whether the structures obtained from molecular dynamics (MD) simulations are useful for improved accuracy relative to using a single structure. In this study, we examined comprehensive evaluations of how to utilize the crystal structure ensemble, what kinds of MD simulations are good to explore structural variations, how to process these structural ensembles using structural clustering, and how to quantify the binding affinity of each inhibitor using available docking scores.

In this study, we used as two test systems, Type III inhibitors of MEK1, a phosphorylation enzyme involved in the MAPK signaling pathway, and Type I inhibitors of EGFR, a receptor tyrosine kinase involved in signaling pathways related to cell proliferation. From all co-crystal structures available in the KLIFS database, we chose the inhibitors with experimental IC₅₀ values for evaluating the *in silico* screening methods. To evaluate the usefulness of the MD simulation, we performed 1-microsecond simulations of both MEK1 (PDB: 1s9j) and EGFR (4jrv), and for both apo and drug-bound holo forms. The structural clusterings to obtain a set of representative structures were then calculated by various patterns. Comprehensive docking simulations were then carried out using Autodock Vina for all inhibitors against the obtained receptor structures. Various patterns of binding affinity scores were attempted by combining thus derived docking scores, and the prediction accuracy was examined.

The correlation coefficient between the experimental IC₅₀ values and the binding affinity scores of the inhibitors showed that using a single co-crystal structure underwent inaccurate prediction. In contrast, averaging the docking scores derived from multiple structures significantly improved the correlations. Additionally, it was found that using MD structures of the holo form resulted in better correlations rather than the apo form. The contact analysis of the inhibitor during MD simulation showed that the lead skeleton appropriately constrained the movement of the receptor protein's binding site, allowing the generation of

structures that accommodate various compound modifications. This highlights the strength of the MD simulation in holo form to generate suitable receptor structures for docking simulation. Furthermore, assuming the case of no crystal structure available, the protocol of making structural modeling, docking the lead skeleton of the inhibitors, and performing the associated holo-form MD simulations was conducted, indicating comparable accuracy with the result using a co-crystal structure.

P07-05

Development of Prediction Models for Membrane Permeability of Cyclic Peptides using 3D Descriptors obtained from Molecular Dynamics Simulations and 2D Descriptors

Masatake SUGITA ^{*1, 2}, **Yudai NOSO**¹, **Takuya FUJIE**^{1, 2}, **Jianan LI**¹, **Keisuke YANAGISAWA**^{1, 2}, **Yutaka AKIYAMA**^{1, 2}

¹School of Computing, Institute of Science Tokyo

²Middle Molecule IT-Based Drug Discovery Laboratory (MIDL), Institute of Science Tokyo

(* E-mail: sugita@bi.c.titech.ac.jp)

Improving membrane permeability is an critical issue in cyclic peptide drug discovery. Although membrane permeability prediction has been performed based on molecular dynamics simulations,[1] it is computationally expensive. Alternatively, machine learning models can predict membrane permeability at negligible cost, but it requires a larger dataset. However, only 7334 experimental values of membrane permeability are available at the time we started our research and the number of data has not increased significantly. Therefore, we developed a machine learning protocol using 3D descriptors obtained from molecular dynamics simulations in combination with 2D descriptors, to generate a universal model with a realistic computational cost. We targeted 252 peptides across four datasets. Several 3D descriptors are obtained from the predicted conformations outside the membrane, at the water/membrane interface, and in the membrane based on the replica exchange with solute tempering/replica exchange umbrella sampling method. Simple learning algorithms such as random forest and support vector machine were used. The best prediction performance was obtained using XGBoost, with a correlation coefficient $R = 0.77$ and mean absolute error = 0.46. The important descriptors included those representing hydrophilicity and hydrophobicity of the peptide, as well as the conformational differences between inside and outside the membrane, the degree of freedom of the peptide, and the approximate shape of the peptide at the membrane center. In addition, the ability of the model to predict membrane permeability of peptides with different chemical structures from the training data was confirmed by excluding one of the four data sets and then creating a new training model using the protocol developed in this study to predict the excluded data. The results showed the model's generic nature with $R = 0.49$ and RMSE = 0.85. In such situations, it is difficult to predict membrane permeability using only 2D descriptors,

demonstrating that descriptors based on conformations obtained from MD are essential. We also added new dataset consisting of 20 peptides and performed external validation.

[1] Masatake Sugita et al., J. Chem. Inf. Model., 62, 18, 4549-4560 (2022)

P07-06

***De novo* PROTAC linker design to enhance cell membrane permeability based on a data-driven method**

Yuki MURAKAMI ^{*1}, Shoichi ISHIDA¹, Yosuke DEMIZU^{1, 2}, Kei TERAYAMA¹

¹Graduate School of Medical Life Science, Yokohama City University

²Division of Organic Chemistry, National Institute of Health Science

(* E-mail: w245513c@yokohama-cu.ac.jp)

Proteolysis targeting chimeras (PROTACs) have garnered significant interest as next-generation therapeutics capable of degrading disease-related proteins of interest (POIs) [1]. PROTACs are chimeric molecules consisting of an E3 ligase-binding moiety, a POI ligand, and a linker. The linker structures of PROTACs significantly influence biodegradation efficiency and physicochemical properties including cell membrane permeability [2]. Optimizing the linker structures is crucial for improving both biodegradation efficiency and physicochemical properties. Recently, many machine learning-based methods that generate PROTAC linkers with various improved properties, such as linker length, logP, and three-dimensional binding conformations, have been developed [2]. However, a PROTAC linker design method to improve cell membrane permeability, which is one of limitations of PROTACs [3], remains undeveloped. Here, we developed a machine learning-based PROTAC linker design method to improve cell membrane permeability. To evaluate the cell membrane permeability of PROTACs with the designed linker, we constructed a prediction model using a machine learning approach based on public experimental data. We designed PROTAC linkers by combining molecular generative models [4,5], and the prediction model to improve cell membrane permeability. At this conference, we report both the proposed method and the results.

[1] Tsai Jonathan M.; *et al.*, *Targeted protein degradation: from mechanisms to clinic*, *Nature Reviews Molecular Cell Biology*, 2024, 1-18.

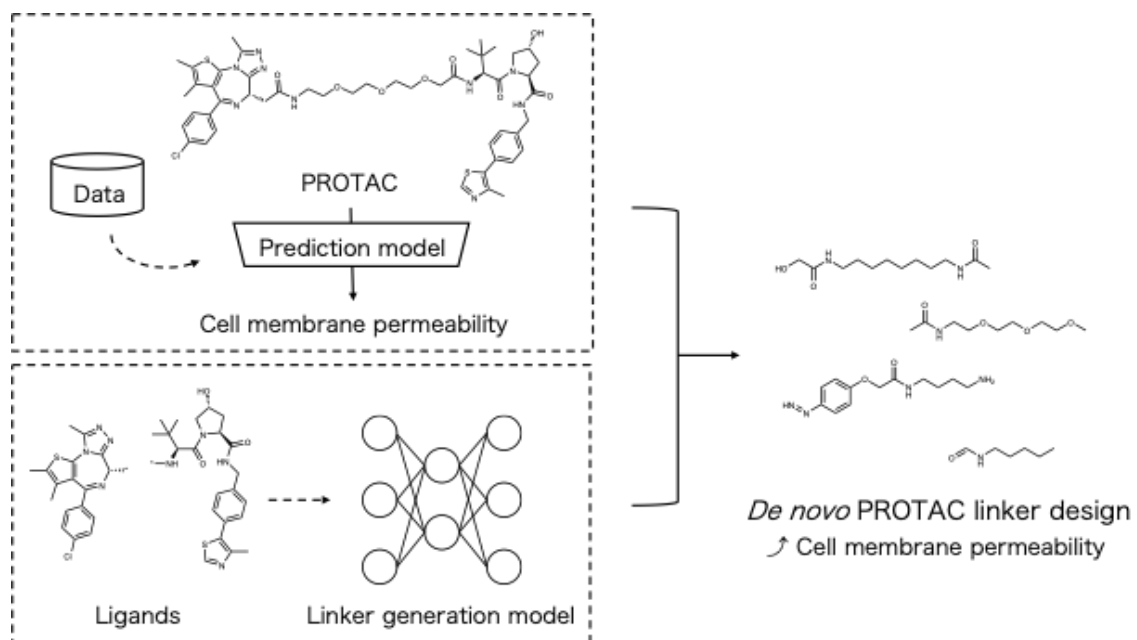
[2] Dong Yawen; *et al.*, *Characteristic roadmap of linker governs the rational design of PROTACs*, *Acta Pharmaceutica Sinica B*, 2024.

[3] Apprato Giulia; *et al.*, *Exploring the chemical space of orally bioavailable PROTACs*, *Drug Discovery Today*, 2024, 103917.

[4] Ishida Shoichi; *et al.*, *ChemTSv2: Functional molecular design using de novo molecule generator*, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2023, 13, e1680.

[5] Zheng Shuangjia; *et al.*, *Accelerated rational PROTAC design via deep*

learning and molecular simulations, *Nature Machine Intelligence*, 2022, 4, 739-748.



P07-07

Scaling up Binding Free Energy Calculations: Integrating Free Energy Perturbation (FEP) and Active Learning to Prioritize Compound Designs

Yunoshin TAMURA *, Junya YAMAGISHI, Mizuki TAKEMOTO

Preferred Networks

(* E-mail: ytamura@preferred.jp)

Binding free energy calculation measures the change in free energy when a compound binds with its target protein. It's a commonly used technique for designing drug candidates with strong affinity. P-FEP is one of the relative free energy perturbation (RBFEP) implementations [1], which calculates the binding free energy differences between compounds within a group sharing a common scaffold. The performance of P-FEP has been demonstrated on a number of benchmark sets. One of the challenges with RBFEP is its significant computational load, requiring the use of high-performance computing systems. However, even with such resources, applying FEP to thousands or tens of thousands of compounds within a finite time remains a significant issue.

As a potential solution to this problem, the pool-based active learning has been proposed [2]. This study demonstrated the effectiveness of P-FEP combined with the pool-based active learning protocol targeting proteins involved in the treatment of disease. The targets were selected based on whether the 3D structures were available, compounds had common or similar scaffold, and their activities were diverse.

Before applying the active learning, we calculated the binding free energy of the compounds whose SAR has been available. And good correlation between experimental values and calculated values were confirmed. Initially in the active learning, we prepared a compound pool using machine learning-based and chemical reaction-based methods. Then we repeated sampling and labeling based on the active learning strategy using the Gaussian process regression model. Utilizing our approach, we were able to narrow down compounds with strong binding affinity. In this presentation, we will discuss the effectiveness of our approach in prioritizing compounds for synthesis and evaluation from a large compound pool.

[1] <https://tech.preferred.jp/ja/blog/pfep-launch/>

[2] Thompson J. et al. Optimizing active learning for free energy calculations. *Artif. Intell. Life Sci.* 2022,2,100050.

P07-08

A Dirichlet diffusion model for generation of high-quality antimicrobial peptide sequences

Koichi OKI ^{*1}, **Shuto HAYASHI** ², **Jun KOSEKI** ³, **Teppei SHIMAMURA** ^{1, 2}

¹Graduate School of Medicine, Division of Systems Biology, Nagoya University

²Medical Research Institute, Institute of Science Tokyo

³Cellular and Molecular Biotechnology Research Institute, National Institute of Advanced Industrial Science and Technology (AIST)

(* E-mail: oki.koichi.c5@s.mail.nagoya-u.ac.jp)

The misuse of antibiotics has led to the rise of drug-resistant bacteria, projected to become the leading cause of death globally by 2050. Antimicrobial peptides (AMPs), which function differently from traditional small-molecule drugs, have gained attention due to their ability to delay bacterial resistance. AMPs interact with bacterial membranes due to their amphipathic nature, causing cell lysis. While machine learning has been explored for generating novel peptides, the discrete nature of peptide sequences makes feature extraction and quality generation challenging. Current AMP generative models, including autoencoders, diffusion models, and transformers, tackle this issue by mimicking original sequence characteristics. However, these models often label peptides in datasets using a fixed threshold for Minimum Inhibitory Concentration (MIC), leading to low-resolution MIC data. For more effective peptide generation, MIC should be treated as continuous values.

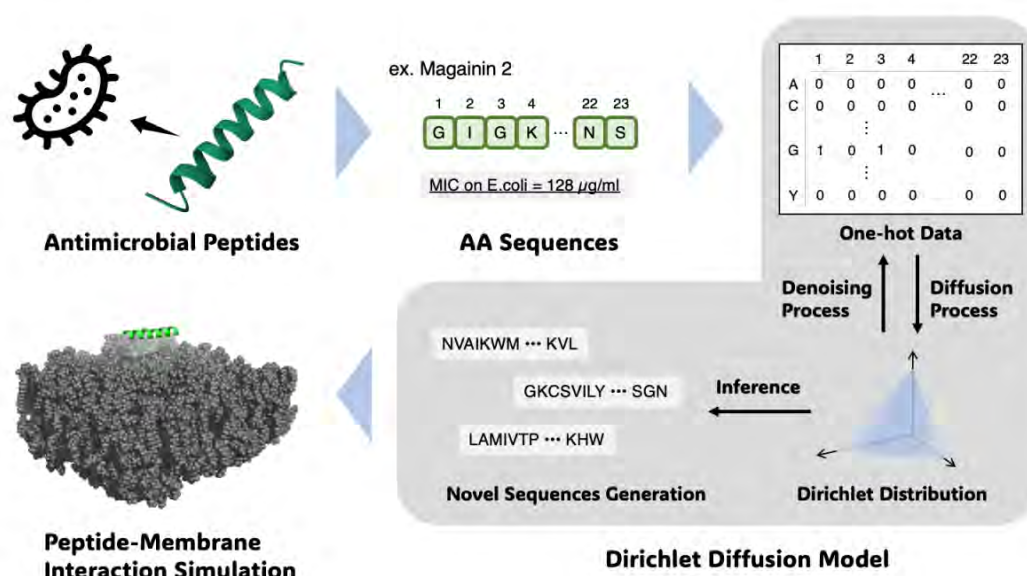
We developed a deep learning-based method for generating high-quality AMP sequences using a Dirichlet diffusion score model, known for its strong performance in discrete data generation. Our model incorporates continuous MIC values during training and sequence generation, establishing a framework that uses molecular dynamics (MD) simulations to verify AMP and bacterial membrane interactions.

The model was trained on sequences with confirmed antimicrobial activity from the DBAASP database. The generated AMPs retained key characteristics of the original sequences, such as amino acid composition and physicochemical property distributions. By using continuous MIC values and guiding the model to generate sequences with lower MIC values, our method efficiently produces highly active peptides compared to existing models.

We also conducted MD simulations on the generated AMP candidates. Although AMPs primarily exert their effects through interactions with bacterial membranes, simulating a cell's lipid bilayer is computationally expensive. We

combined coarse-grained and all-atom simulations, enabling high-throughput and effective evaluation of many generated peptides.

This study marks the first application of the Dirichlet diffusion score model in AMP sequence generation and shows that integrating continuous MIC values with multiscale simulation techniques can enhance peptide design significantly. Further validation through experiments with actual bacteria or animals, including assessments of toxicity and hemolytic effects, is necessary to confirm the quality of the generated AMPs.



P07-09

Development of a Massive Fluorogenic Probe Library Based on Bayesian Optimization toward the Discovery of Novel Biomarker Enzymes

Daiki ISHIMOTO *¹, Ryo TACHIBANA¹, Yasuteru URANO^{1, 2}

¹Laboratory of Chemistry and Biology, Graduate School of Pharmaceutical Sciences, The University of Tokyo

²Department of Chemical Biology and Molecular Imaging, Biomedical Engineering, Radiology and Biomedical Engineering, Graduate School of Medicine, The University of Tokyo

(* E-mail: d-ishimoto2001@g.ecc.u-tokyo.ac.jp)

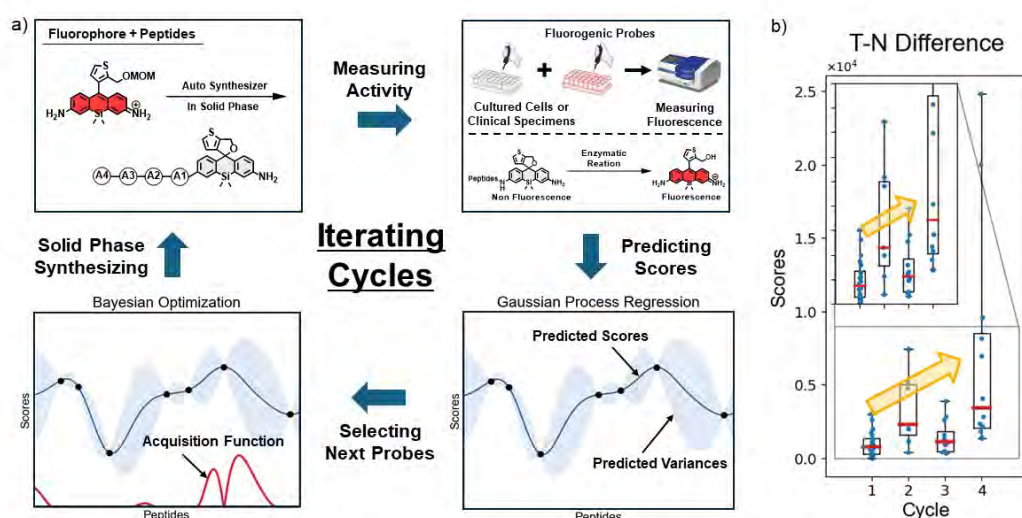
Drastic treatment for many types of cancer is to resect whole tumor regions. However, it is difficult to distinguish between tumor and non-tumor tissues with unaided human eyes and cancer tissues left in patients' body cause recurrence and metastasis which lead to poor prognosis. For these reasons, clinical needs for cancer-selective visualizing tools are growing and among those are fluorogenic probes. These fluorogenic probes are non-fluorescence and enzymatic reactions at target sites turn fluorescence on. In previous research, a fluorogenic probe library was constructed to find novel tumor-selective biomarker enzymes and as a result, promising probes and biomarker enzymes were found for tumor-selective visualization.

However, present ways of developing target-specific probes have many problems. For example, since we must synthesize all candidate probes, varieties of target enzymes and probes' structure are limited. In addition to this practical problem, tumor-selective enzymes need to be known well in advance to find its selective probes. Considering how many enzymes and its substrates are, establishing conclusive methods for searching target-specific enzyme-substrate pairs would be difficult. Here, we propose one solution to this challenge by using Machine Learning techniques: Gaussian Process(GPR) and Bayesian Optimization(BO).

BO leverages results of GPR, which is used for predicting scores of imaginary probes. In BO, Acquisition Function plays a key role in deciding which probes are more likely to show better scores among imaginary probes. There are many Acquisition Functions like Probability Improvement, Expected Improvement and Predicted Entropy Search as major examples. They commonly use predicted scores to estimate to what extent each imaginary probe improves scores and regression accuracy. Next, we synthesize suggested probes by BO and evaluate its scores in experiments. After that, BO are applied again to recalculate

imaginary probes' improvements. By iterating this process, predicting and evaluating, we could reach probes that have desired properties without evaluating a numerous amount of candidate probes.

In this research, we employed a fluorophore with bright fluorescence in the visible wavelength region: HMRR (hydroxymethyl rhodamine red). HMRR with tetra peptides consist of 20 essential amino acids are adapted as candidate probes which have 160,000 (204) kinds in total. It is totally impossible to synthesize and evaluate all these probes in real and these high-dimensional probes have yet to be explored well. Making use of Machine Learning techniques introduced above, finding novel biomarker-probe pairs becomes easier from a great variety of candidates with a minimum number of real experiments. We hope our research contributes to explain how Machine Learning system works in Chemical Biology.



- a) An overall View of this research. Fluorogenic probes are synthesized at first and their scores are applied to Machine Learning to predict scores of imaginary peptides. Iterating this process leads to findings of novel target-selective peptides and biomarker enzymes.
- b) Demonstrating how our concept works in trial to find Colorectal cancer specific peptides using Colorectal cancer specimens' lysate.

P07-10

Virtual validation and the efficient learning methods exploration in federated learning (FL) for drug development research

Ziwei ZHOU ^{*1}, Reiko WATANABE¹, Masataka KURODA^{2, 3}, Kenji MIZUGUCHI¹

¹Institute for Protein Research, Osaka University

²National Institutes of Biomedical Innovation, Health and Nutrition, ³Mitsubishi Tanabe Pharma Corporation

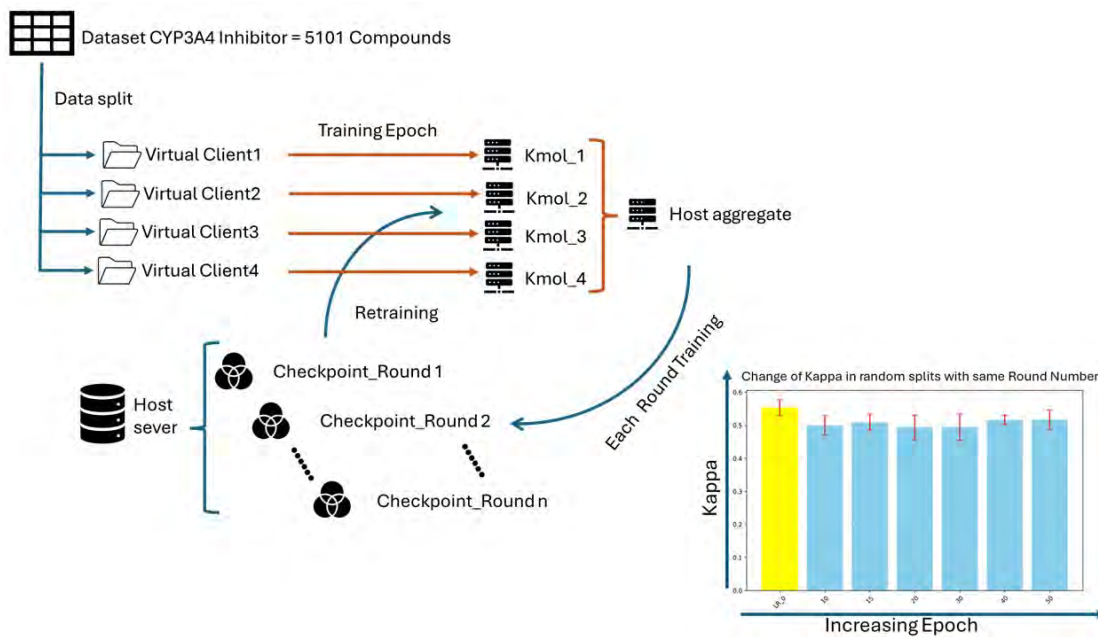
(* E-mail: u631859f@ecs.osaka-u.ac.jp)

Applying Machine Learning (ML) methods in drug development requires large-scale and high-quality experimental data, which is hard to obtain from public sources. It is expected that collaboration with multiple pharmaceutical companies will significantly facilitate data collection. However, sharing data with another research group needs to involve the risk of data leaking and even causing economic losses. Intellectual property issues are a major barrier to business collaborations and public-private partnerships. Federated Learning (FL) is an ML method that allows multi-source data join training via the client's server, which solves intellectual property problems to a certain extent. There is a possibility that differences in chemical space, number of data sets, and data heterogeneity among each client may worsen the consistency and convergence of the training model. Although it is essential to consider appropriate learning conditions in FL, the impact of data differences among clients on FL has not yet been fully explored.

This research aims to validate the feasibility of FL and identify efficient learning methods. The study is based on CYP3A4 inhibition data with a 10 μ M threshold containing 5101 compounds. The datasets were split using three different segmentations: random splits and splits showing different chemical spaces with and without unifying the number of data sets to simulate real situations. Then classification models for CYP3A4 inhibition were constructed using graph convolutional networks at virtual FL in different training conditions such as hyperparameter setting in epoch, round numbers and other settings.

It showed that hyperparameters will largely influence the performance of FL. In the best hyperparameters setting used in full dataset training, FL models with random split could reach the equivalent performance of local training. We will discuss how we can optimize the learning condition in FL.

Our finding would make it possible to present more suitable learning conditions for future FL implementation and provide a possible method to boost model training speed via FL methods.



P07-11

Structure and Interaction Analysis of Nucleic Acid Encapsulated ssPalm Lipid Nanoparticles by Multiscale Simulation

Naoko KONAMI ^{*1}, **Koji OKUWAKI** ², **Hiroki TANAKA** ³, **Hidetaka AKITA** ⁴, **Kenjiro HIGASHI** ⁴, **Takayuki FURUISHI** ⁵, **Etsuo YONEMOCHI** ⁶, **Kaori FUKUZAWA** ¹

¹Graduate School of Pharmaceutical Sciences, Osaka University

²JSOL Corporation

³Graduate School of Pharmaceutical Sciences, Tohoku University

⁴Graduate School of Pharmaceutical Sciences, Chiba University

⁵Juntendo University Faculty of Pharmacy

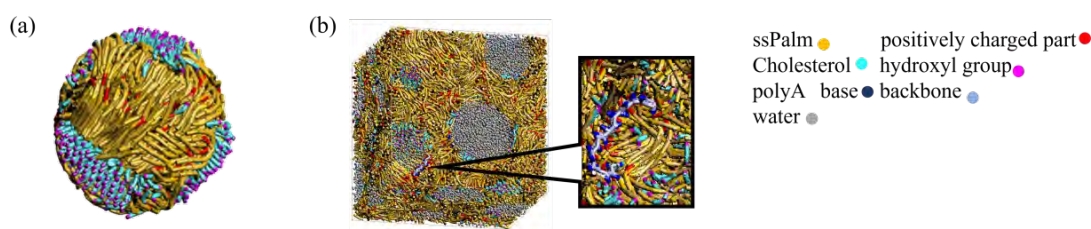
⁶School of Pharmacy at Narita, International University of Health and welfare

(* E-mail: konami-n@phs.osaka-u.ac.jp)

ssPalm is an SS-cleavable and pH-activated lipid-like material for nucleic acid delivery. Depending on the linker structure of oleic acid scaffold and tertiary amine, the transfection efficiency increases in the order of ssPalmO-Ph (SSOP), ssPalmO-Bn (SSOB), and ssPalmO-P4C2 (SSOC). In this study, we analyzed the structure of ssPalm-LNPs and the interaction between lipids by molecular simulation in order to understand the influence of lipid linker structure and composition on LNP structure. The molecular structures of ssPalm, cholesterol (Chol), and polyA were divided into coarse-grained particles, and the interaction parameters for the Dissipative Particle Dynamics (DPD) method were calculated using the Fragment Molecular Orbital (FMO) method. Two models were created using the FMO-DPD method with the COGNAC program: (a) LNP models and (b) LNP core models, to evaluate particle structure in water and the structure of encapsulated nucleic acids within the LNP, respectively. The LNP model comprised ssPalm/Chol at a 60/40 molar ratio with 91 vol% water, while the core model contained ssPalm/Chol at a 75/25 molar ratio along with polyA and 20 vol% water. Both models were simulated for 1 million steps within a 100,000-particle simulation box. Then, all -atom models were constructed by the reverse mapping approach based on the DPD structure. Molecular Dynamics (MD) calculations were performed for 100 ns using the Amber10:EHT force field in the AMBER program. Finally, FMO calculations of about 30,000 atoms were conducted for the MD structure after 100 ns at the MP2/6-31G level using the ABINIT-MP program. In all three ssPalm-LNP structures, Chol was not miscible with ssPalm, and clusters of Chol were formed on the LNP surface. The hydroxyl groups of Chol faced the outer aqueous phase of the particles. FMO calculations show that SSOP and SSOB with benzene rings have energetically more stable

interactions between ssPalm compared to SSOC without benzene rings. It was suggested that the stabilization of LNPs by introducing benzene rings is involved in enhancing transfection efficiency. In addition, an endo-aqueous phase was formed in the LNP core, and Chol clusters were also observed at an interface between lipid and water. PolyA was presented at the interface, surrounded by charged portions of ssPalm, while avoiding Chol clusters. Multi-scale simulations combining DPD, MD, and FMO enable the structure and inter molecular interaction of LNPs at the atomic level, providing insights for enhancing their activity.

Structures after DPD simulation of ssPalm-LNP



Structures after DPD simulation of (a) the LNP model (ssPalm/Chol = 60/40 mol% + 91 vol% water) and (b) the LNP core model (ssPalm/Chol = 75/25 mol% + polyA + 20 vol% water). Water molecules are not shown in (a) for clarity.

P07-12

Natural-Product Screening Toward Discovery of Anti-Aging Glutaminase-1 Inhibitors. An Electronic-Structure Informatics Study

Mio YOKOYAMA *, Mizuki IWASAKI, Yusuke TATEISHI, Manabu SUGIMOTO

Kumamoto University

(* E-mail: 245d8785@st.kumamoto-u.ac.jp)

Accumulation of the senescent cells in aging within the human body can trigger various diseases, leading to death. Recently, the persistence of these senescent cells is linked to Glutaminase-1 (GLS1). Thus, developing GLS1 inhibitors is considered one of the significant challenges in anti-aging [1]. To develop more highly active inhibitors and/or to discover unique compounds that have been unknown so far, in silico screening is expected to play an important role. Aiming to discover unknown scaffolds of GLS1 inhibitors, we herein apply Electronic-Structure Informatics (ESI), which has been suggested by the authors' group [2]. ESI is an informatics focusing on information obtained through electronic-structure (quantum chemistry) calculations. The molecular descriptors therein have been derived on some theoretical bases. The ESI descriptor set does not directly contain descriptors that reflect structural features. Because they represent features related to electronic structure, they are expected to provide molecules with unexpected scaffolds that differ from known GLS1 inhibitors in the course of inhibitor screening. We explore the natural product (NP) database for screening, anticipating the potential of NP.

We have developed a machine learning model to screen for potential GLS1 inhibitors by calculating ESI descriptors for 260 GLS1 inhibitor molecules obtained from ChEMBL. These descriptors were utilized as explanatory variables in constructing a regression model with the activity value (pIC_{50}) as the objective variable. It was found that the regression model using the extra trees regressor is successful where a coefficient of determination (R^2) was 0.777 for the test set of molecules corresponding to 20 % of molecules among the 260 GLS1 inhibitors. Because the reproducibility of the present model is considered reasonably good, we have constructed the final regression model for compound screening by including all the molecules. In applying this final model, we screened 2647 molecules from a NP database called KampoDB [3] to identify highly active GLS1 inhibitors. The KampoDB is a database for traditional Japanese medicines, so our approach herein corresponds to "drug-repositioning".

We could have found several potential candidates for GLS1 inhibitors in the KampoDB. For structural modifications (optimization), we took the molecular

anatomy-and-remodeling approach: a candidate molecule was decomposed into fragments, and their contributions to pIC_{50} were analyzed using the regression model, which we call “NP anatomy”. Then, some fragments were replaced with others to enhance the inhibitory activity. We call this latter approach “NP remodeling”. In the presentation, we will show the detailed results of compound screening and our molecular optimization.

- [1] Y. Johmura, *et al.*, *Science*, **371**, 265-270 (2021).
- [2] M. Sugimoto *et al.*, *Chem. Lett.*, **50**, 849-852 (2021).
- [3] R. Sawada *et al.*, *Sci. Rep.*, **8**, 11216 (2018).

P07-13

DiffInt: Integrating Explicit Hydrogen Bond Modeling into Diffusion Models for Structure-Based Drug Design

Masami SAKO ^{*1}, Nobuaki YASUO², Masakazu SEKIJIMA¹

¹Department of Computer Science, Tokyo Institute of Technology

²Academy for Convergence of Materials and Informatics (TAC-MI), Tokyo Institute of Technology

(* E-mail: sako.m.ab@m.titech.ac.jp)

Structure-based drug design is a crucial approach in the drug discovery process and aims to effectively create molecules that bind to specific target proteins. In recent years, significant progress has been made in applying deep learning methods to molecule generation, making it possible to generate ligand molecules directly in the protein pocket with 3D information. However, these studies have not been able to incorporate protein-ligand interaction information, making it difficult to efficiently generate ligand molecules with high binding affinity.

In this study, we realize pharmacophore modeling that preserves hydrogen bond between the protein and ligand molecules in the structure-based drug design via deep learning model. By introducing "interaction particles" that explicitly represent hydrogen bonds between the protein and ligand molecules, it becomes possible to generate ligand molecules with the desired hydrogen bonds retained. The model combines an E(3)-equivariant graph neural network with a diffusion model framework. The diffusion process gradually adds noise to the input ligand molecule, whereas the inverse process generates a new molecule through denoising. Both the protein pocket structure and the interacting particles remain fixed as conditions throughout these processes, leading to the generation of molecules with specific desired interactions.

The model has been trained on 100,000 protein-ligand complexes in the CrossDocked dataset and evaluated by generating the ligand molecules 100 times for each protein pocket of a test set consisting of 100 proteins. Hydrogen bond reproducibility and hydrogen bond energies estimated from docking simulations outperform existing models.

P07-14

QUBO Problem Formulation of Fragment-Based Protein – Compound Flexible Docking

Keisuke YANAGISAWA ^{*1, 2}, **Takuya FUJIE**³, **Kazuki TAKABATAKE**⁴, **Yutaka AKIYAMA**³

¹Department of Computer Science, School of Computing, Institute of Science Tokyo

²Middle Molecule IT-Based Drug Discovery Laboratory (MIDL), Institute of Science Tokyo

³Ahead Biocomputing, Co., Ltd.

⁴Toshiba Digital Solutions Corporation

(* E-mail: yanagisawa@c.titech.ac.jp)

Protein – ligand docking plays a significant role in structure-based virtual screening. In recent years, the size of compound libraries has been growing explosively, resulting in a demand of faster docking methods. One of the promising approaches is quantum annealing, which has been attracting attention. Quantum annealing efficiently solves combinatorial optimization problems and is suitable for docking because docking is an optimization problem searching for the best scoring pose. In quantum annealing, it is important to formulate this docking as a quadratic unconstrained binary optimization (QUBO) problem to obtain solutions, and thus, such formulations have been attempted. However, most previous studies did not consider the internal degrees of freedom of the compound that is mandatory and essential.

In this study, we formulated fragment-based protein–ligand flexible docking, considering the internal degrees of freedom of the compound by focusing on fragments (rigid chemical substructures of compounds) as a QUBO problem. [1] We introduced four factors essential for fragment – based docking in the Hamiltonian: (1) interaction energy between the target protein and each fragment, (2) clashes between fragments, (3) covalent bonds between fragments, and (4) the constraint that each fragment of the compound is selected for a single placement (Figure A).

We also implemented a proof-of-concept system and conducted redocking for the protein–compound complex structure of Aldose reductase using SQBM+, which is a simulated quantum annealer. The predicted binding pose reconstructed from the best solution was near-native (Figure B, RMSD = 1.26 Å), which can be further improved (Figure C, RMSD = 0.27 Å) using conventional energy minimization. The results indicate the validity of our QUBO

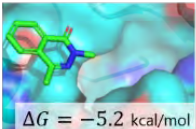
problem formulation.

[1] K Yanagisawa, T Fujie, K Takabatake, Y Akiyama. QUBO Problem Formulation of Fragment-Based Protein-Ligand Flexible Docking. Entropy 26, 397, 2024. DOI: 10.3390/e26050397

A) Four factors to construct the Hamiltonian

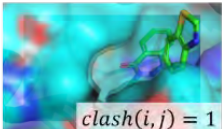
$$H = A \sum_i \Delta G_i x_i + B \sum_{i,j} b_{ij} x_i x_j + C \sum_{i,j} c_{ij} x_i x_j + \frac{D}{2} \sum_k \left(\sum_{f_i=k} x_i - 1 \right)^2$$

Interaction energy
scores of fragments



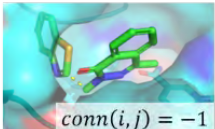
$\Delta G = -5.2 \text{ kcal/mol}$

Clashes
between fragments




$clash(i, j) = 1$

Covalent bonds
between fragments



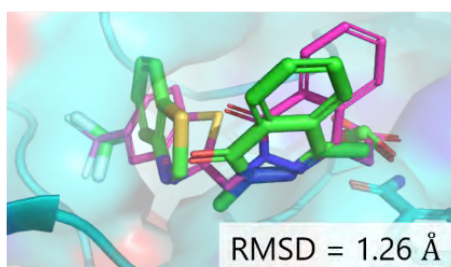
$conn(i, j) = -1$

The constraint
each fragment is selected
for a single placement



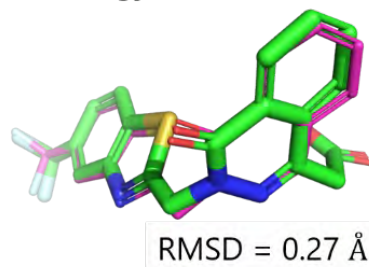
or or ...

B) The redocking result



Green: The best output of our method
 Purple: Co-crystallized ligand structure
 Cyan: protein structure (ALDR)

C) After energy minimization



Green: Energy-minimized structure
 Purple: Co-crystallized ligand structure

P07-15

Acquisition of Bias Information for Protein-Ligand Docking by Mixed-Solvent Molecular Dynamics

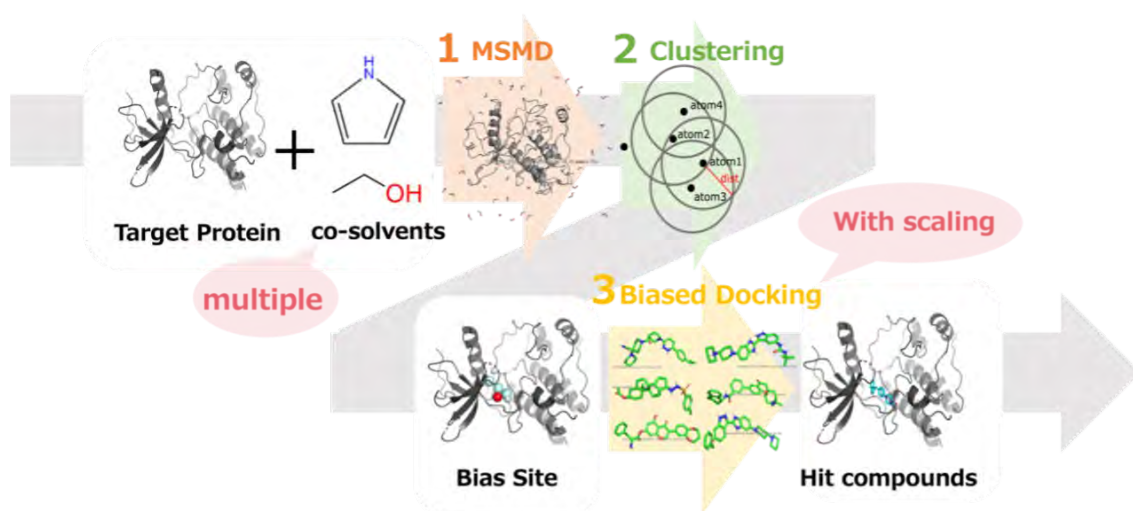
Kaho AKAKI ^{*}, Keisuke YANAGISAWA, Yutaka AKIYAMA

School of computing, Institute of Science Tokyo

(^{*} E-mail: akaki.k.aa@m.titech.ac.jp)

Virtual screening (VS) is a widely used method to computationally select drug candidates from a large number of compounds. In VS, protein-ligand docking is often performed to estimate binding affinities and binding modes. However, the accuracy of docking is not sufficient. Biased docking is a technique to introduce additional energy scores into docking aiming to improve the accuracy. One of the methods to estimate an adequate bias is mixed-solvent molecular dynamics, or MSMD. MSMD involve MD in the presence of explicit water molecules mixed with probe molecules or functional group fragments such as for hotspot detection, binding site identification, and binding free energy estimation. Arcon et al. shows that biased docking with interactions information from MSMD can provide better results [1]. However, they only used ethanol as a probe molecule. In this study, we additionally utilized pyrrole as well as ethanol as probe molecules, and constructed four types of bias information: (1) aromatic rings, (2) hydrogen atoms to be hydrogen bond donors, (3) oxygen and nitrogen atoms to be hydrogen bond acceptors, and (4) nitrogen atoms that are not hydrogen bond acceptors. In addition, we introduced scaling for the strength of the bias. We evaluated our method on seven target proteins. Using MSMD, we calculated the interaction energies of multiple atom types and protein surfaces for every target, and obtained bias information for protein-ligand docking. The results indicated that the VS accuracy was improved for four of the seven target proteins with the bias obtained from ethanol and five of the seven target proteins with the bias obtained from pyrrole.

[1] Arcon JP, et al. JCIM, 59(8), 3572-3583, 2019.



P07-16

Development of a compound pre-screening method based on docking of fragments

Masayoshi SHIMIZU *, Keisuke YANAGISAWA, Yutaka AKIYAMA

School of Computing, Institute of Science Tokyo

(* E-mail: shimizu@bi.c.titech.ac.jp)

Structure-based virtual screening (SBVS) is a computational technique to select compounds using 3D structural information of target proteins and compounds. Recently, more than several hundred million compounds are registered in compound database [1], and thus, a method to select from a large number of compounds at high speed is required. Protein-ligand docking is often used in SBVS, but it has high computational complexity and development of faster methods is highly demanded. One of the ways to realize faster calculation is to focus on the fragments of compounds. Since, for example, 28 million compounds can be represented by only 263 thousand fragments [2], fragment-based ligand docking approach enables evaluation of compounds with a fragment set that has fewer number of types than that of the compound set. However, fragment-based docking still spends several CPU core years to evaluate hundreds of millions of compounds.

In this study, unlike conventional fragment-based methods that the compounds are used as input (Figure A), a pre-selected representative fragment set is used as input (Figure B). We proposed a compound pre-screening method that searches for possible conformations of known compounds based on the relative distance between spatial arrangement of pairs of representative fragments and selects those compounds as candidates (Figure C).

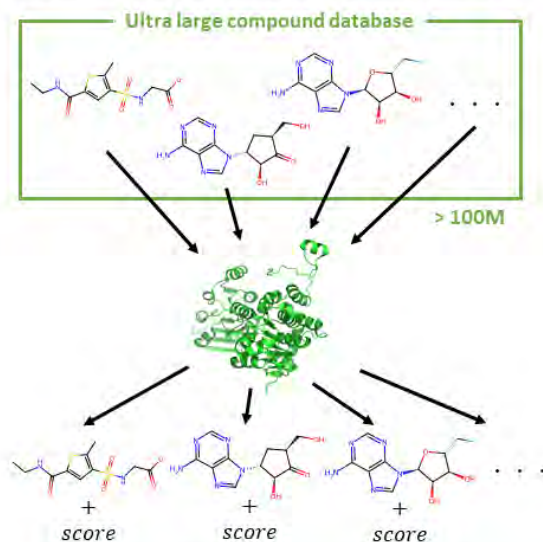
In computational experiments with six targets in DUD-E, we confirmed that pre-screening was up to 60 times faster than existing protein-ligand docking tools like REStretto [3]. On the other hand, the prediction accuracy still needs to be improved. A possible approach to improve accuracy might be to consider arrangements of more than three fragments.

[1] BI Tingle, et al., JCIM 63, 1166-1176, 2023.

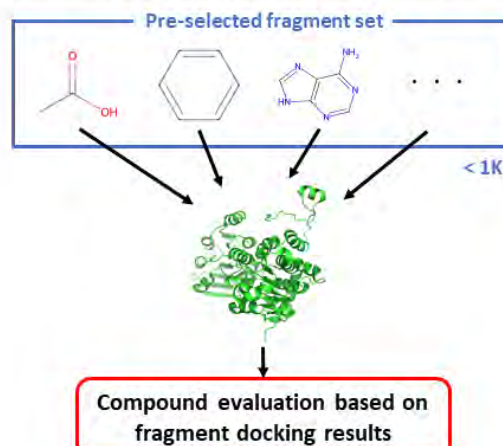
[2] K Yanagisawa, et al., Bioinformatics 33, 3836-3843, 2017.

[3] K Yanagisawa, et al., ACS omega 7, 30265-30274, 2022.

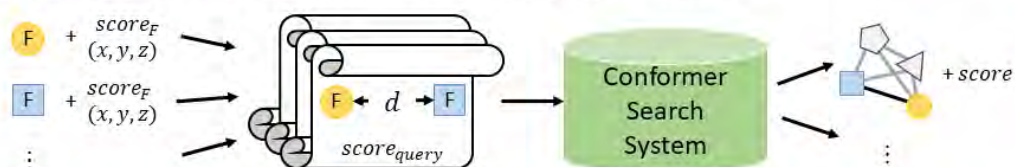
A. Traditional method (input is **compounds)**



B. Proposal method (input is **fragments)**



C. How to evaluate compounds based on fragment docking results



P07-17

Report on Participation in the Tox24 Challenge: Construction of a High-Accuracy QSAR Predictive Model for Transthyretin Activity

Yuma IWASHITA *, Kyosuke KIMURA, Tomoya KOMASAKA, Koki SHISHIDO, Taichi NAKAMURA, Mizuho ASADA, Yoshihiro UESAWA

Laboratory of Medical Molecular Analysis, Meiji Pharmaceutical University
(* E-mail: m246208@std.my-pharm.ac.jp)

【Objective】 The Tox24 Challenge [1, 2] is a QSAR competition aimed at advancing computational methods for predicting the in vitro activity of compounds. Specifically, the challenge is to predict the measured affinity of various compounds for transthyretin (TTR). Sponsored by AIDD, AiChemist, Chemical Research in Toxicology, and ICANN2024; submission deadline is August 31, 2024. Participants have to upload their predictive values of the affinity of diverse compounds for TTR to a dedicated website where their performance and ranking will be then displayed in real time on a leaderboard [3]. Our laboratory has assembled a model development team for this competition. This presentation outlines our dataset preparation and machine learning techniques used for constructing our TTR activity prediction model.

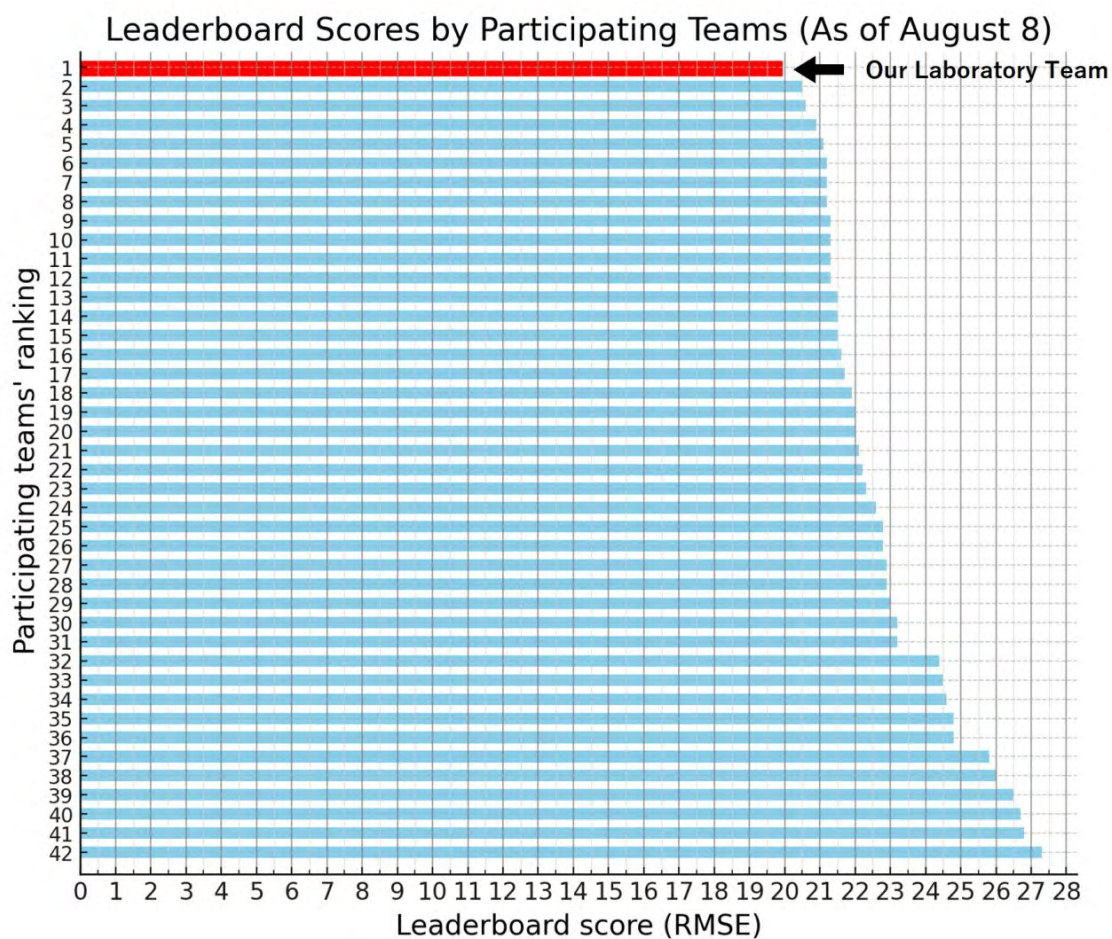
【Methods】 The dataset was obtained from the dedicated website on OCHEM [4]. The compound data were categorized into three sets: training (1012 compounds), leaderboard (200 compounds), and blind (300 compounds) sets. The training dataset included measured TTR activities for model development. The leaderboard dataset lacked measured values; participants submitted predicted activities for the leaderboard dataset to the organizers, who then evaluated accuracy based on the root mean square error (RMSE). The measured values for the dataset will be released by the organizers on August 15 and will be used to further develop the model. The blind dataset will be used for the final ranking of predictive models and will remain undisclosed until the competition ends. We devised methods to adjust molecular descriptors based on the SMILES data provided by the organizers and implemented rigorous validation to ensure the accuracy of our machine learning models.

【Results and Discussion】 The proposed model achieved an RMSE of 19.9 on the leaderboard dataset, securing the top position at the time of this report. The final winning model based on the blind dataset will be announced on September 19, 2024.

References :

1) Tetko IV. Tox24 Challenge. Chem Res Toxicol. 2024 Jun 17;37(6):825-826.

- 2) <https://e-nns.org/icann2024/challenge/>
- 3) <https://ochem.eu//challenge/show.do?render-mode=full>
- 4) <https://ochem.eu/home/show.do>



P07-18

Lead generation of a V-ATPase inhibitor using molecular generative AI

Taiyo TOITA ^{*1}, Kano SUZUKI^{2, 3}, Shoichi ISHIDA¹, Akira KATSUYAMA^{4, 5}, Satoshi ICHIKAWA^{4, 5}, Masateru OHTA⁶, Mitsunori IKEGUCHI^{1, 6}, Takeshi MURATA^{1, 2, 3}, Kei TERAYAMA^{1, 7, 8}

¹Graduate School of Medical Life Science, Yokohama City University

²Graduate School of Science, Chiba University

³Membrane Protein Research Center, Chiba University

⁴Faculty of Pharmaceutical Science, Hokkaido University

⁵Center for Research and Education on Drug Discovery, Hokkaido University

⁶HPC- and AI-driven Drug Development Platform Division, RIKEN Center for Computational Science

⁷RIKEN Center for Advanced Intelligence Project

⁸MDX Research Center for Element Strategy, Tokyo Institute of Technology

(* E-mail: w245418f@yokohama-cu.ac.jp)

Vancomycin-resistant *Enterococcus faecium* (VRE) is a bacterium that causes nosocomial infections. VRE is resistant to many antibiotics, narrowing treatment options and spreading globally. Because of the health risks and limited treatment options caused by VRE, the World Health Organization (WHO) has listed VRE as one of the priority pathogens requiring new antimicrobials [1]. VRE grows predominantly under alkaline conditions after antibiotic administration by specifically expressing a Na⁺-transporting V-ATPase [2, 3]. We have discovered a hit compound that binds specifically to Na⁺-transporting V-ATPase and exhibits inhibitory activity. This compound contributes to inhibiting VRE growth by binding between the multiple subunits called the V_o region of V-ATPase. However, although this inhibitor is active in the small intestine, it is not effective in the large intestine. Therefore, there is room for improvement in the binding affinity and membrane permeability of this inhibitor. In this study, we performed lead generation for the V-ATPase inhibitor with a focus on improving binding affinity considering membrane permeability. For lead generation, we employed ChemTSv2 [4], a molecular generative AI based on recurrent neural networks and an exploration system with Monte Carlo tree search. When performing lead generation for the V-ATPase inhibitor, we considered several conditions to generate the structure. To fit into the small-sized binding pocket, we trained the generative AI using various datasets with limited molecular weight. In addition, we designed some functions that evaluate docking scores and interactions with surrounding residues to improve binding

affinity. To keep the location of the part of the known inhibitor that contributes to the activity, we filtered by RMSD of the common structure. Here, we report the methods and the results.

- [1] Willyard, C. The drug-resistant bacteria that pose the greatest health threats, *Nature*, 2017, 543, 15
- [2] Murata, T. et al. Intracellular Na⁺ regulates transcription of the ntp operon encoding a vacuolar-type Na⁺-translocating ATPase in *Enterococcus hirae*, *J. Biol. Chem*, 1996, 271, 23661-23666
- [3] Murata, T. et al. The ntpJ gene in the *Enterococcus hirae* ntp operon encodes a component of KtrII potassium transport system functionally independent of vacuolar Na⁺-ATPase, *J. Biol. Chem*, 1996, 271, 10042-10047
- [4] Ishida, S. et al. ChemTSv2: Functional molecular design using de novo molecule generator, *WIREs Comput. Mol. Sci*, 2023, 13, e1680

P07-19

Exploring the Power of Structural Biology on Degradar Discovery

Yifan HU *, Wenjun GUI, Jiaquan WU

Biortus Biosciences Co. Ltd, Biortus Biosciences Co. Ltd
(* E-mail: yifan.hu@biortus.bio)

The emergence of protein degradation as a novel therapeutic modality has garnered significant attention within the drug discovery community. Unlike traditional small molecule drug discovery, where structure-based drug design (SBDD) is extensively utilized for the rational design of hit and lead compounds, the application of structural biology to degrader discovery has thus far been limited to elucidating the binding modes of selected lead PROTACs or molecular glues. This limitation primarily stems from two challenges: the lack of robust protocols for assembling stable degradation complexes, and the suboptimal resolution of these complex structures as determined by either X-ray crystallography or cryo-electron microscopy (cryo-EM).

In this study, we presented a comprehensive analysis of the CRBN E3 ligase-based PROTAC/molecular glue system using both cryo-EM and X-ray crystallography. We successfully resolved the cryo-EM structure of the CRBN-DDB1-PROTAC-target ternary complex at a resolution of 3.1 Å. Additionally, we resolved the X-ray crystallography structure of the CRBN-DDB1-molecular glue-target ternary complex at a resolution of 2.7 Å. Given that the target protein is identical, with the only variable being the compound (PROTAC or molecular glue), we conducted a detailed comparative analysis of the cryo-EM and X-ray crystallography structures, particularly focusing on the binding interfaces of these ternary complexes.

This study highlights several key findings. First, while the high-resolution cryo-EM and X-ray crystallography structures show considerable similarity in the DDB1 domain, they exhibit significant differences in the conformation of the target protein. Second, the binding poses of the PROTAC and molecular glue in both structures provided atomic-level insights into their interactions with CRBN and the target protein. Third, we identified a novel interface between CRBN and the target in the cryo-EM structure that enhances the cooperativity of ternary complex formation; this interface was absent in the X-ray crystallography structure, potentially due to the shorter length of the molecular glue compared to the PROTAC. Finally, we demonstrated the importance of selecting

appropriate DDB1 truncations for achieving high-resolution structures. Specifically, the cryo-EM structure benefited from the deletion of the BPB domain in DDB1, which stabilized the ternary complex, while the full-length DDB1 used in X-ray crystallography improved crystallization success.

Our findings establish cryo-EM and X-ray crystallography as complementary techniques for the structural determination of E3 ligase-PROTAC/molecular glue-target ternary complexes, offering advantages in both resolution and speed.

P07-20

Constructing a machine learning model for discriminating Urotensin-II receptor inhibitors and its application

Kentaro KAWAI *¹, **Momoko KYUTA**², **Runa MINATO**², **Shoki HOSHIKAWA**¹, **Reiko KONISHI**², **Kazuyuki SATO**¹, **Kohji KOMORI**², **Ko KAWADA**², **Akira MUKAI**²

¹Laboratory for Medicinal Chemistry, Faculty of Pharmaceutical Sciences, Setsunan University

²Laboratory for Clinical Pharmacology and Therapeutics, Faculty of Pharmaceutical Sciences, Setsunan University

(* E-mail: kentaro.kawai@pharm.setsunan.ac.jp)

Rare adverse reactions of cardiotoxicity have been reported with Remdesivir (RDV), a drug approved for COVID-19. Although the detailed mechanism of occurrence was unknown, it was reported that the Urotensin-II receptor (UT2R) pathway is involved in RDV cardiotoxicity and UT2R inhibition reduces the occurrence of cardiotoxicity. We have therefore conducted an analysis of US Food and Drug Administration (FDA)-approved drugs using the machine-learning model for UT2R activity to derive new candidate compounds for the treatment of RDV-induced cardiovascular events.

Compounds acting on UT2R were extracted from the ChEMBL database. Compounds with a pChEMBL value of 6 or higher were labeled as active. However, as there were few compounds with a pChEMBL value of less than 6, a random sampling of data from among the ChEMBL listed compounds was used to select compounds as inactive compounds. Each compound was converted to six fingerprints, which were trained using four classification models, including SVM and random forests. The created models were then used to discriminate FDA-approved drugs, and compounds with the potential to show binding properties to UT2R were obtained.

The ROC-AUC and Cohen's Kappa of 24 models combining machine learning algorithms and fingerprints showed high values. The FDA-approved drugs were analyzed, and drugs that were suggested to bind to UT2R were found. FAERS was used to analyze cardiotoxicity-related conditions, and the relationship between the drugs and adverse events such as arrhythmia was investigated.

P07-21

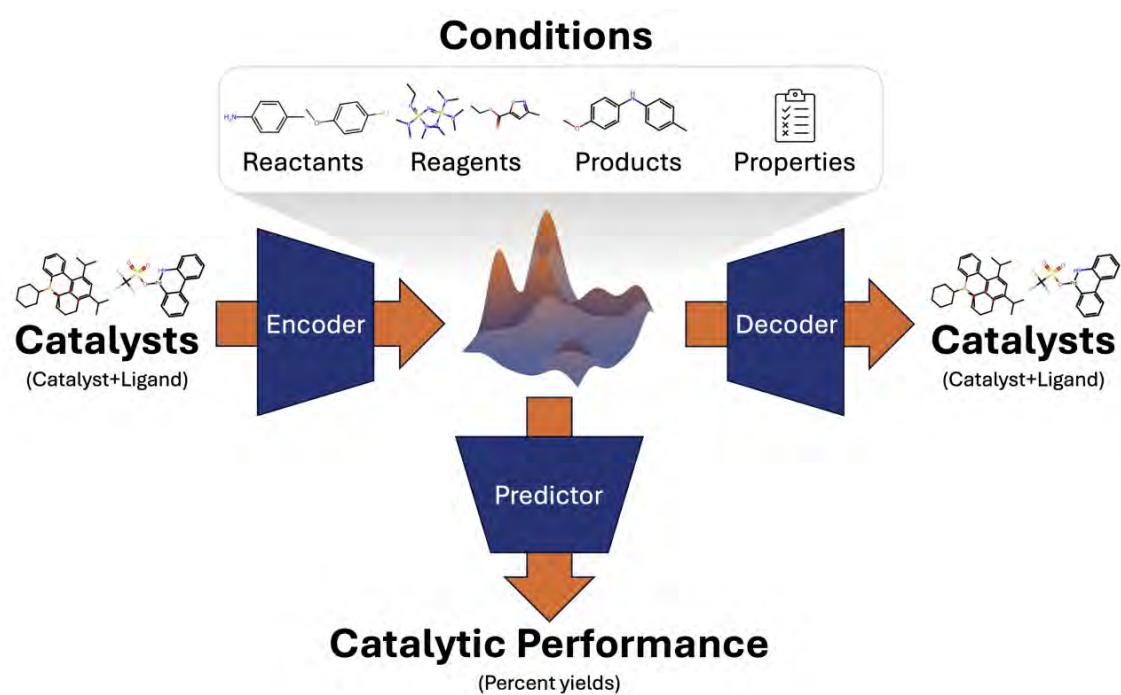
Reaction-conditioned variational autoencoder model for catalyst generation and catalytic performance prediction

Apakorn KENGKANNA*, Masahito OHUE

School of Computing, Institute of Science Tokyo

(* E-mail: kengkanna@li.c.titech.ac.jp)

Catalysts are materials that accelerate the rate of a chemical reaction without being permanently consumed. The structures of catalyst, particularly metal-ligand complexes, crucially determine the catalytic activity of chemical reactions. Designing effective catalysts is the key process for optimizing catalytic reactions. Recent approaches, including generative models, have been proposed to design new catalysts. However, these methods are developed for a specific reaction class without considering reaction components and conditions, and rely on predefined fragment categories, limiting novel discoveries outside the existing chemical space. Here, we present a reaction-conditioned generative model based on a variational autoencoder for generating catalysts and ligands and predicting their catalytic performance. This model is pre-trained on a large available reaction database, allowing for broader knowledge acquisition and fine-tuning for downstream reaction classes. The model shows comparable performance in catalytic activity prediction tasks, as well as demonstrates a promising way to generate possible catalysts and ligands for given reaction conditions. This work will help facilitate the optimization of ligands for catalysis and further enhance catalyst design and discovery processes in both the chemical and pharmaceutical industries.



P07-22

Drug discovery research utilizing BROOD: A Fragment Replacement and Molecular Design tool

KOSUKE MINAGAWA *, SHUNYA MAKINO, KAN SHIRAISHI

Modarity Research Laboratories 1, Daiichi Sankyo Co., Ltd.

(* E-mail: kosuke.minagawa@daiichisankyo.com)

BROOD, a tool developed by OpenEye, proposes structures that have similar shape and electrostatic properties to a specific substructure based on information from the ChEMBL database. It is a useful tool for exploring the chemical and property space around hit or lead compounds.

Since launching the "Data-Driven Drug Discovery" (D4) strategy in 2018, we have been actively working on improving the operational efficiency of drug discovery research through information extraction from public data and the utilization of cheminformatics methods [1]. As part of these efforts, we have incorporated structure generation into the drug discovery process and shared ideas with medicinal chemists.

The process from BROOD structure generation to actual compound synthesis varies depending on the project's circumstances. Specifically, factors such as the presence of target protein structure information, the structural transformation site (side chain or skeleton), or whether overtake approach or not. Depending on these differences, it is possible to propose adequate compounds for medicinal chemists to find compounds to be synthesized easily and quickly.

In this presentation, we introduce various scenarios of the process from structure generation to actual compound synthesis.

References

[1] Ryo Kunitomo, Jürgen Bajorath, Kazumasa Aoki, From traditional to data-driven medicinal chemistry: A case study, *Drug Discovery Today*, 2022, 27, 2065.

P07-23

A small molecule inhibitor that binds to the unstable state of its target kinase DYRK1A demonstrates slowly dissociation from the complex

**Sora SUZUKI ^{*1}, Koji UMEZAWA², Gaku FURUIE¹, Daichi NAKAMURA³,
Ninako KIMURA¹, Masato YAMAKAWA¹, Yuto SUMIDA^{3, 4}, Takashi
NIWA^{3, 4, 5}, Takamitsu HOSOYA^{3, 4}, Kii ISAO^{1, 2}**

¹Department of Agriculture, Graduate School of Science and Technology, Shinshu University

²Institute for Biomedical Sciences, Shinshu University

³Laboratory for Chemical Biology, RIKEN BDR

⁴Laboratory of Chemical Bioscience, Institute of Biomaterials and Bioengineering, Tokyo Medical and Dental University

⁵Graduate School of Pharmaceutical Sciences, Kyusyu University

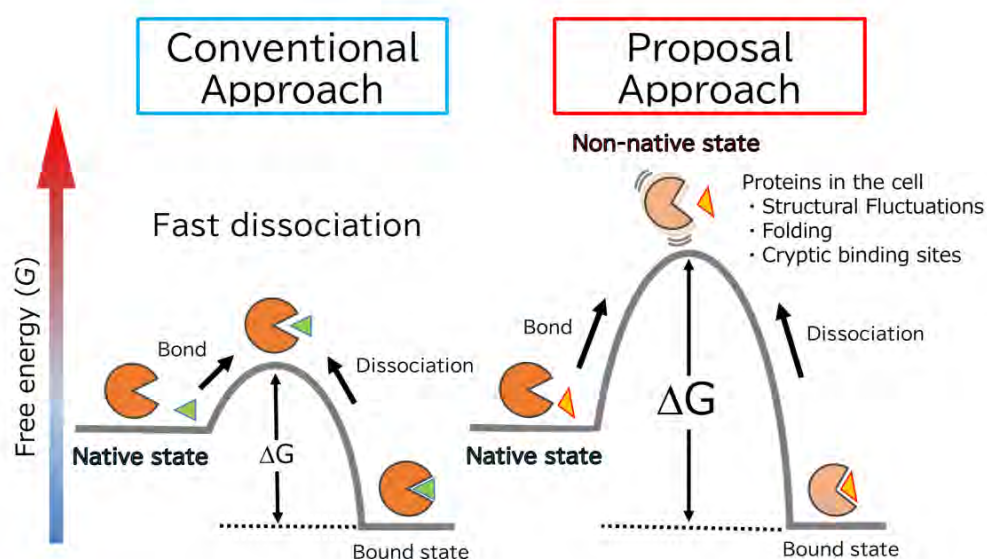
(* E-mail: 23as104a@shinshu-u.ac.jp)

In small molecule drug discovery, inhibitors that dissociate slowly from their target proteins have been being highly sought. The inhibitory effect of such inhibitors is maintained even when their concentration in the body decreases due to metabolism and excretion. As a result, the frequency of administration can be reduced, thereby alleviating side effects.

In this presentation, we report that small molecule inhibitors that selectively target the non-native states of proteins dissociate slowly from their complexes. The native state is the stable conformation of the protein, characterized by the lowest free energy. In contrast, the non-native state is an unstable conformation with a partially denatured protein structure and higher free energy. Therefore, when small molecule inhibitors selectively bind to the non-native state, the free energy significantly decreases compared to binding to the native state. Generally, for a small molecule inhibitor to dissociate from the complex, the complex must obtain the lowered free energy from the surrounding environment. This implies that a significant amount of free energy is required from the environment for the inhibitor to dissociate from the complex. However, since the frequency of the complex obtaining such a large amount of free energy from the environment is low, dissociation inevitably becomes slow.

We validated this theory using the kinase DYRK1A, which is involved in neurological disorders, and a selective small molecule inhibitor for its non-native state, FINDY. Experimental results showed that FINDY did not dissociate from

its complex with DYRK1A. Furthermore, we analyzed the binding structures of an inhibitor for the native state of DYRK1A and a structural analogue that is selective for the non-native state. By comparing these two binding structures, we successfully identified the structural characteristics of inhibitors that exhibit slow dissociation. This study provides new options for inhibitor design and discovery in small molecule drug discovery targeting kinases.



We provide new options for inhibitor design and discovery in small molecule drug discovery targeting kinases.

P07-24

Correlation Analysis of Excipient Modulated Viscosity of Monoclonal Antibody and Molecular Surface Patch Properties

Yoshiro KIMURA *

Life Science Dept., MOLSIS Inc.

(* E-mail: kimura.yoshiro@molsis.co.jp)

When an antibody drug is administered by injection, a high-concentration solution of 100 mg/mL or more is required because of the limited amount administered at one time. However, the formulations with such high concentrations can result in a high viscosity. Subcutaneous injection typically requires viscosities below 15–20 cP. Various excipient molecules, such as saccharides and amino acids, are used to adjust the viscosity. This experimental process is expensive and time-consuming, so *in silico* techniques are expected to improve efficiency.

The molecular surface patch analysis functionality implemented in MOE[1], which is the molecular modeling and simulation software, has been reported to be useful for protein-protein interaction analysis[2] and estimating physical properties such as protein retention times in hydrophobic interaction chromatography[3][4]. The patch analysis has also been used in studies of antibody viscosity [5], but to our knowledge no studies considering excipients have been done to date. Therefore, we decided to perform molecular surface patch analysis on a system including an antibody and its interacted excipient molecules and investigate the correlation between various physical properties obtained by this and viscosity. If they are correlated, it will be possible to predict viscosity, screen excipients, and analyze the factors behind the increase in viscosity.

We investigated the correlation between the experimental viscosity of an antibody with some excipients and various physical properties obtained by molecular surface patch analysis and found that the surface area of the positive and negative charge patches was highly correlated with the viscosity. In this presentation, we will introduce the details of the calculation method and the comparison with the experiment.

[1] Molecular Operating Environment (MOE), 2022.02; Chemical Computing Group ULC, 1010 Sherbrooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2022.

[2] Depetris, R. S.; et al. *Proteins: Struct., Funct., Bioinf.* 2022, 90 (4), 919–

935.

[3] Thorsteinson, N.; et al., *mAbs*, 2021, 13 (1), 1981805.

[4] Jetha, A.; et al., *mAbs*, 2018, 10 (6), 890-900.

[5] Armstrong, G. B.; et al., *Comput. Struct. Biotechnol. J.*, 2024, 23, 2345-2357.

P07-25

Predicting Antibody Stability pH Values from Amino Acid Sequences: Leveraging Protein Language Models for Formulation Optimization

Takuya TSUTAOKA *¹, Noriji KATO¹, Toru NISHINO¹, Yuanzhong LI¹, Masahito OHUE²

¹Bio Science & Engineering Laboratory, FUJIFILM Corporation

²School of Computing, Institute of Science Tokyo

(* E-mail: takuya.tsutaoka@fujifilm.com)

Monoclonal antibodies (mAbs) offer significant therapeutic benefits; however, their formulation requires careful optimization to prevent instability, such as aggregation and thermal degradation. Standard practices for determining optimal formulation conditions rely on time-consuming and costly wet lab experiments. Therefore, we developed a machine learning-based approach to predict the optimal pH value for stabilizing mAbs using only their amino acid sequences. Briefly, amino acid sequences were input into a protein language model to extract features, which were then used in a regression model to predict the pH values. We compiled an original dataset of 56 commercially available mAbs and obtained their pH values from publicly available FDA documentation. The performance of our approach was evaluated using a 10-fold cross-validation method, assessing the correlation coefficient between the predicted and actual pH values. Due to the absence of directly relevant methods, we established a baseline by comparing various combinations of elements, including different antibody domains, protein language models, and regression models. We also conducted feature engineering to enhance the predictive performance by incorporating structural information and descriptors. Our approach achieved a high Pearson correlation coefficient of 0.88. This result complements that of wet lab experiments and highlights the potential of increasing the efficiency and cost-effectiveness of optimizing the conditions for mAb formulation.

P07-26

Development of a Platform for Crystal Structure Prediction of Drug Molecules

Okimasa OKADA ^{*1}, **Yuya KINOSHITA**², **Koki NISHIMURA**², **Aaron NESSLER**³,
Hiroomi NAGATA¹, **Michael SCHNIEDERS**³, **Kaori FUKUZAWA**⁴, **Etsuo YONEMOCHI**⁵

¹Mitsubishi Tanabe Pharma Corporation

²Takeda Pharmaceutical Company Limited

³University of Iowa

⁴Osaka University

⁵International University of Health and Welfare

(* E-mail: okada.okimasa@mc.mt-pharma.co.jp)

In the design and quality control of drugs, crystalline polymorphism differs in physical properties such as solubility and physical/chemical stability, and affects the quality of drugs. Therefore, polymorphism is one of the important control items for quality control of drugs. It is generally reported that crystalline polymorphism is present in 80% of pharmaceutical compounds. Pharmaceutical companies perform screening experiments to select the optimal crystal form for each drug, but it is not possible to cover all crystallization conditions within a limited time period. Therefore, the crystal form is determined by screening experiments under certain crystallization conditions. However, many pharmaceutical companies have experienced troubles where a new stable crystal form suddenly appears in the manufacturing process or during long-term storage.

Therefore, reducing the possibility to have new stable crystal forms is an important issue for pharmaceutical companies with a mission to stably supply drugs to patients. One potential solution is computational crystal structure prediction. In this research, we aim to develop a platform for automatic prediction calculation of stable crystal structures. In this system, a stable crystal structure is output from a 2D molecular structure by creating a polarizable force field, generating a stable candidate crystal structure for each space group, narrowing down the crystal structure by lattice energy and density, removing duplicates, and high-precision lattice energy calculation by density functional theory.

P07-27

Automated molecular modeling and property assessment for ADCs

Takashi Ikegami *

Life Science Dept., MOLSIS Inc.

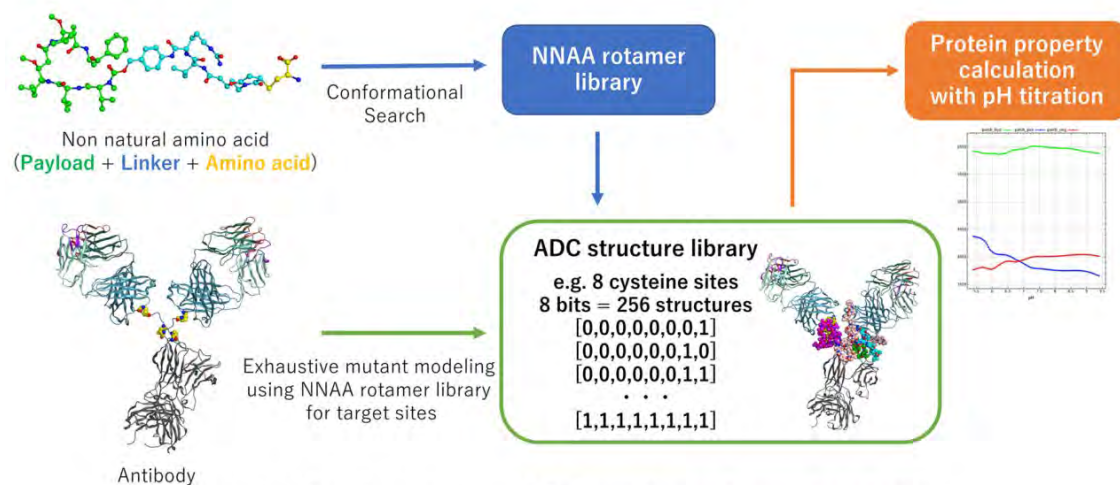
(* E-mail: ikegami.takashi@molsis.co.jp)

Antibody-drug conjugates (ADCs) are known as a class of key biopharmaceutical drugs in oncology. ADCs have a cytotoxic drug (payload) bonded to an antibody via a linker. Monoclonal antibody of ADC selectively binds to a target protein on the surface of a tumor cell and ADC delivers payloads into the tumor cell to cause cell death. In the development of ADCs, drugs are typically conjugated to solvent-exposed cysteine or lysine residues in the antibody. Because cysteine in antibodies is less common than lysine, cysteine is often used to minimize the distribution of the number and location of drugs. Since there are 8 cysteine residues in an antibody with inter-chain disulfide bonds, a total of 256 possible combinations of ADC structures can exist if chemical modifications are performed exhaustively. The composition ratio of the number of drug molecules conjugated to an antibody in an ADC is called the drug-antibody ratio (DAR). Drugs are often hydrophobic, and ADCs with high DAR can significantly alter their physical properties and cause experimental problems such as aggregation of ADCs. Therefore, it is important to predict ADCs with appropriate DAR among many combinations using in silico techniques. To estimate the effects in the physical properties of ADCs depending on the position and number of drugs, we propose a workflow that exhaustively model the 3D structure of ADCs and estimate their physical properties on Molecular Operating Environment (MOE)[1]. The workflow is shown in the figure:

1. Define amino acids covalently bound to the drug via the linker as non-natural amino acid (NNAA) and construct the side-chain rotamer libraries by conformational analysis.
2. Construct mutant library (ADC structure library) by replacing an amino acid (e.g. cysteine) with the NNAA at a specified position in the antibody. The stable side-chain conformation of NNAA is automatically selected from the rotamer library.
3. Calculate physical properties on the obtained ADC library, taking into account pH and conformational changes in ADCs.

This workflow will make it easier to find ADCs with the desirable physical properties and improve the efficiency of ADC design.

[1] *Molecular Operating Environment (MOE)*, 2022.02; Chemical Computing Group ULC, 1010 Sherbrooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, **2024**.



Workflow for automated ADC modeling and property calculation

P07-28

Validation of the reproducibility of hit-to-candidate using ChemTS

Tomoki YONEZAWA *¹, **Masateru OHTA**², **Shoichi ISHIDA**³, **Kei TERAYAMA**³, **Teruki HONMA**², **Kazuyoshi IKEDA**^{1, 2}

¹Faculty of Pharmacy, Keio University

²R-CCS, RIKEN

³Graduate School of Medical Life Science, Yokohama City University

(* E-mail: yonezawa-tm@pha.keio.ac.jp)

Various structure generation methods have been developed, and their application to drug design is expected to enhance the efficiency of the drug development process. In this study, we have attempted to generate the chemical structures of marketed drugs and clinical candidates using AI to verify the applicability of structure generation to drug discovery. To make this verification practical, we aimed to reproduce the structure development process from hit compounds to approved drugs or clinical candidates.

Structure development from hit compounds often involves multi-objective optimization to improve not only the primary activity against the target, but also physical properties such as solubility and membrane permeability, as well as pharmacokinetic properties. Among structure generation AIs, ChemTS generates structures while performing reinforcement learning to improve the target parameters. We used a predictive model of physicochemical and ADME properties, and set the predicted value as the reward function. This enables the search for structures with improved predicted values. We combined the main activity evaluation by docking or 3D shape similarity with the predictive model of physicochemical and ADME properties such as solubility, membrane permeability, and metabolic stability, and used it as a reward to generate structures using ChemTS.

To find applicable examples of structure development that could be evaluated using the predictive model, we employed the following approach. A review paper published in the Journal of Medicinal Chemistry provided examples of successful structure development of approved drugs and clinical candidates from hit compounds. Hit-drug and hit-candidate pairs were extracted from the paper, and related activity information were also obtained from ChEMBL. When generating the structure of DORAVIRINE, an approved HIV reverse transcriptase inhibitor, we successfully reproduced its structure with docking, membrane permeability, solubility, and metabolic stability as rewards. In a subsequent verification test with TEPOTINIB, a c-Met inhibitor, we attempted to generate its

structure with docking, membrane permeability, and metabolic stability as rewards. We successfully generated similar structures of highly active compounds other than TEPOTINIB. Based on these verification results, we will discuss the strengths and weaknesses of structure generation using ChemTS.

P07-29

Automated Hit-to-Lead Optimization Using the SINCHO Protocol and ChemTS

Genki KUDO ^{*1}, **Shota NAKAJIMA**², **Yudai ICHIKAWA**³, **Takumi HIRAO**⁴,
Ryunosuke YOSHINO^{4, 5}, **Hitoshi KAMIJIMA**², **Takatsugu HIROKAWA**^{4, 5}

¹Pure and Appl. Sci., Grad. Sci. Tech., University of Tsukuba

²Research Institute of Systems Planning, Inc.

³Med., Med. and Health Sci., University of Tsukuba

⁴Faculty Med., University of Tsukuba

⁵TMRC., University of Tsukuba

(* E-mail: s2330052@u.tsukuba.ac.jp)

The Hit-to-lead process in drug discovery and development involves optimizing a hit compound, which initially has low affinity and selectivity, into a lead compound with high affinity and selectivity. For the rational lead compound design, the 3D structural information of the target protein is crucial. This information has become more accessible with the advent of AlphaFold technology and advances in crystal structure analysis. Nevertheless, in the structure-based drug design, the utilization of this valuable information is limited because the hit-to-lead process still heavily relies on the trial-and-error approach of medicinal chemists. Therefore, a computational method to support and replace this manual process is needed.

In this study, we introduce an automatic hit-to-lead system using SINCHO protocol [1,2], which we developed, and ChemTS [3], a de novo molecular generator. This system begins with the 3D structure of a protein-hit compound complex. Based on the structure, the SINCHO protocol identifies the protein pocket and R-point pair that have the potential to improve affinity through substructure modification. The appropriate molecular weight and logP for the modified substructure are also predicted from the SINCHO results. Then, the lead compound candidates are designed using ChemTS, aligning with these predicted properties.

We will provide details of this system and present the case study.

Reference

- [1] Kudo G, et al. J. Chem. Inf. Model. 2024;64(11):4475-4484.
- [2] Kudo G, et al. Bioinformatics. 2023;39(4):btad212.
- [3] Yang X, et al. Sci. Technol. Adv. Mater. 2017;18(1):972-976.

P07-30

Application of Amino-Acid Mapping: Activity Prediction for Drug Discovery

Yuka MATSUMOTO ^{*1}, **Akito SABURI**¹, **Kyosuke TSUMURA**², **Issei DOI**²,
Yasushi HIKIDA¹

¹Imaging & Informatics Laboratory, Fujifilm Corporation

²Analysis Technology Center, Fujifilm Corporation

(* E-mail: yuka.b.matsumoto@fujifilm.com)

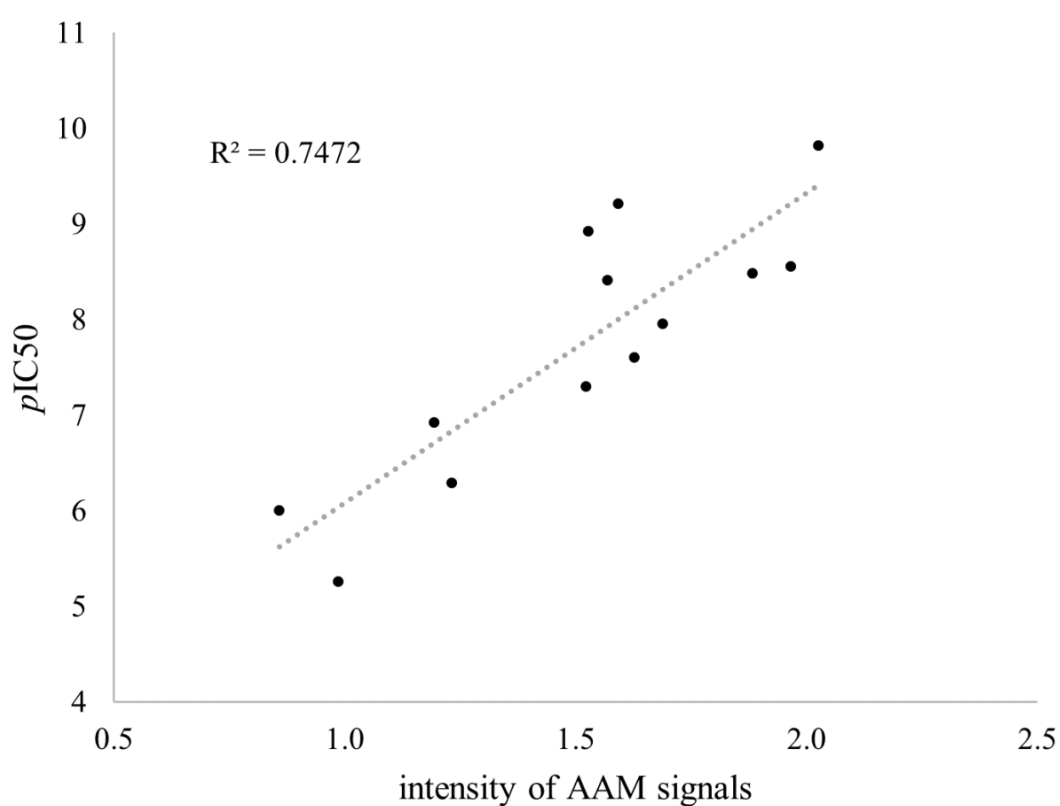
In drug discovery, *in silico* technologies have been widely used to calculate the interactions between compounds and proteins based on their chemical structures, and furthermore, to predict the biological activities of compounds. We have developed Amino-Acid Mapping (AAM) descriptor which can characterize the interactions between compounds and nearby amino acid residues, and demonstrated that the similarities of AAM profiles to those of a known active compound corresponds to the binding energies calculated based on FMO method and their pIC₅₀ values towards protein tyrosine kinase SYK inhibitor [1-3].

Herein, we report that the intensities AAM signals themselves would inherently contain information about the biological activities of compounds. Regarding the inhibitory activities towards SYK, we confirmed that compounds with higher intensities of AAM tend to have higher pIC₅₀ values. Although the overall intensities AAM signals showed a weak correlation with pIC₅₀ value, the intensities of AAM signals partitioned by atomic groups of the compounds exhibited a good correlation with pIC₅₀ value (Fig.).

Furthermore, the strength of the correlation between the intensities of AAM signals and pIC₅₀ values varied depending on the type of amino acids. Among them, the intensities of AAM signals of Asp, which makes a key interaction between SYK, were included in the group with a strong correlation with pIC₅₀ values. This result indicated that AAM descriptor can provide insights into chemical motifs of the ligands and the amino acid residues, which are involved in the binding, without any structural information of the target protein.

Therefore, the AAM descriptor can be used not only to improve the similarity with biologically active compounds but also to predict their biological activities. It was also demonstrated that our AAM methodology has the potential to discovery and design high activity compounds based on the structural information. As another application, the correlation between the intensities of AAM signals partitioned by atomic groups and pIC₅₀ values can be applied to provide structural insights related to the biological activities.

- [1]Mao Tanabe et al., bioRxiv (2023), doi:
<https://doi.org/10.1101/2023.07.03.547598>
- [2]Fujifilm started a CRO service based on AI-AAM, i.e., drug2drugs®. See the following URL for a detailed description: https://labchem-wako.fujifilm.com/jp/custom_service/products/95323.html
- [3]Jun Nakabayashi et al., CBI Annual Meeting 2023, P03-07.



P07-31

Efficient Single Step Synthesizable Molecular Design using Wasserstein Autoencoder

Jinzhe ZHANG ^{*1}, **Jiawen LI** ^{1, 3}, **Mizuki TAKEMOTO** ¹, **Ryuichiro ISHITANI** ^{1, 2, 4}

¹Drug Discovery, Preferred Networks Inc

²Department of Computational Biology and Medical Sciences, The University of Tokyo

³Division of Computational Drug Discovery and Design, Medical Research Institute,, Tokyo Medical and Dental University

⁴Department of Biological Sciences, The University of Tokyo

(* E-mail: jzhang@preferred.jp)

De novo molecular design algorithms address the inverse design problem by creating chemical structures that optimize a given set of desired properties. However, these generative models often overlook the synthetic accessibility of the generated candidates, leading to increased time and cost at the synthesis stage. Generative models that do account for synthetic accessibility frequently underperform in molecular generation tasks.

We propose a Wasserstein autoencoder-based generative model that designs chemical compounds with desired properties while ensuring all generated compounds can be synthesized using single step chemical reaction from a predefined set of possible reaction types, based on a given set of building blocks. By crafting a smoother latent space, we demonstrate that our method outperforms non-synthesizability-aware models in sampling efficiency while maintaining synthesizability. Additionally, our model surpasses existing synthesizability-aware models in terms of optimizing target properties.

This approach facilitates the screening of Bespoke library[1] and the rapid synthesis and testing of designed candidates during the quick prototyping stage of molecular research.

[1] Kaplan, Anat Levit, et al. "Bespoke library docking for 5-HT2A receptor agonists with antidepressant activity." Nature 610.7932 (2022): 582-591.

P07-32

Quantitative Assessment of Protein–Ligand Activity Prediction from 3D Docking Poses for Urate Transporter 1

MARTIN *¹, Mochammad Arfin Fardiansyah **NASUTION**², Ziwei **ZHOU**²,
Xingran WANG², Reiko **WATANABE**², Kenji **MIZUGUCHI**², Ichigaku
TAKIGAWA^{1, 3}

¹WPI-ICReDD, Hokkaido University

²Institute of Protein Research, Osaka University

³Institute for Liberal Arts and Sciences, Kyoto University

(* E-mail: martin@icredd.hokudai.ac.jp)

Urate transporter 1 (URAT1), responsible for reabsorbing over 90% of uric acid in the kidneys and thereby preventing its accumulation in the blood—which can lead to gout—is a crucial target for the development of new anti-hyperuricemic medications. Currently, lesinurad are the only URAT1 inhibitor approved by the Food and Drug Administration, highlighting the need for additional treatments. Machine learning (ML) can enhance structure-based virtual screening for discovering novel URAT1 inhibitors by predicting protein-ligand interactions. This study provides quantitative assessment of the prediction performance of the latest ML models for URAT1 using a unique dataset of 3D URAT1-ligand structures generated by AlphaFold2 and Smina, a fork of AutoDock Vina v1.1.2, as input data and pIC50 as label data. The dataset includes high-active URAT1 inhibitors, low-active URAT1 inhibitors, and simulated inactive structures generated by DeepCoy (Imrie et al., 2021) with properties matching those of the high-active inhibitors. The ratio of active to inactive compounds is maintained at 1:100. The systematic evaluations with respect to AUC, Enrichment Factor and Normalized Enrichment Factor are reported for the ML methods, such as supervised ML with PLEC fingerprints (Wójcikowski et al, 2019), convolutional neural networks (McNutt et al, 2021), and geometric interaction graph neural network (GIGN) (Yang et al, 2023).

P07-33

Development of an efficient compound 3D conformer search system based on relative position of fragments

Tomoya SAITO *, Keisuke YANAGISAWA, Yutaka AKIYAMA

Department of Computer Science , Institute of Science Tokyo

(* E-mail: t_saito@bi.c.titech.ac.jp)

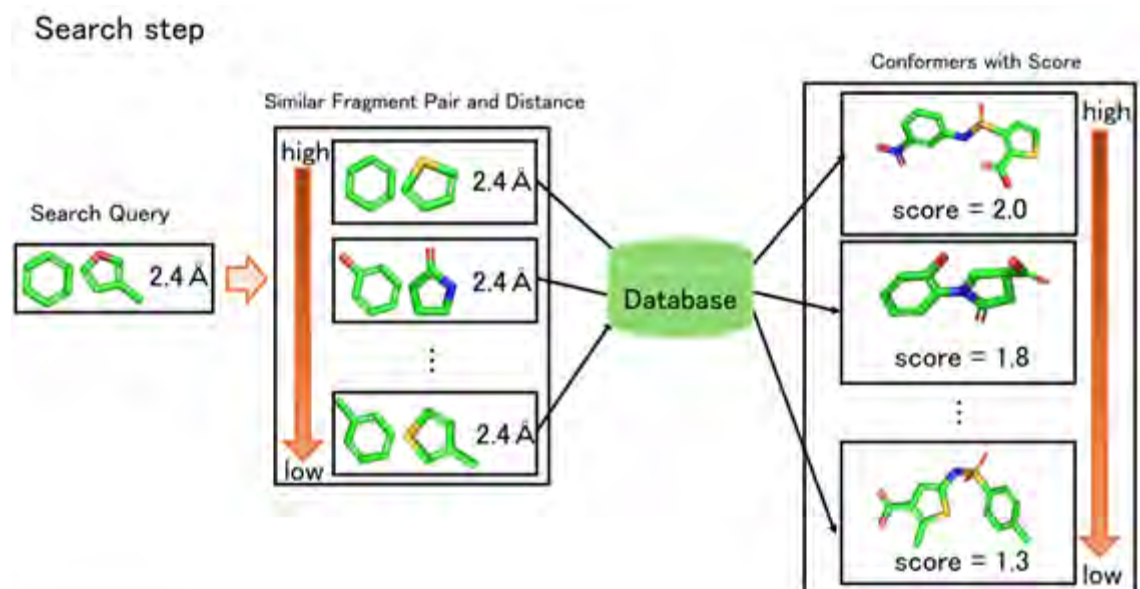
Fragment-based docking has recently been investigated as a mean to speed up large-scale docking calculations. In the context of docking calculations, a fragment is defined as a substructure of a compound that has no internal degrees of freedom. Fragment-based docking estimates the docking score and binding pose of the entire compound by combining the results of fragment-wise docking calculations. It uses fragments as the elements of a compound, and since these fragments are common to other compounds, the results of the fragment-wise calculation can be reused. However, when reconstructing the feasible conformations of a compound from fragments, it is necessary to take into account collisions and possible covalent bonding between fragments, which is a bottleneck in fragment-based docking. For example, eHiTS [1] reconstructs it by solving the NP-hard maximum clique finding problem.

To address the issue, we propose fragment-based conformer database and search system rather than conformer reconstruction algorithms. This system retrieves adequate conformations, thus avoiding the need to consider spatial collisions.

In the database preparation step, pre-generated 3D conformers are decomposed into fragments with positions, and all pairs of these fragments with Euclidean distance information are registered in the database. In the search step, users input a query consisted of two fragments with distance constraints between them, and the database outputs registered conformers which match the query with a reasonable error margin.

For highly accurate and precise search, it is ideal to distinguish all fragment types. However, with hundreds of thousands of fragment types, it is not practical to record every possible fragment pair. To solve this problem, we introduced a similarity search method that clusters similar fragments as a group. This approach allows us to reduce the number of fragment types in the database and enhance recall. Additionally, we are working on improving precision by ranking the search results based on the fragment similarities.

[1] Zsolt Zsoldos, Darryl Reid, Aniko Simon, Sayyed Bashir Sadjad, A.Peter



P07-34

Molecular Properties Prediction by Contrastive Learning Using Graph Neural Network

Koshiro AOKI *, Apakorn KENGKANNA, Masahito OHUE

School of Computing, Institute of Science Tokyo

(* E-mail: aoki.k.as@m.titech.ac.jp)

Molecular properties are the chemical, biological and physical characteristics of a compound. Being able to predict the properties is useful in the search for the new drugs and machine learning models with molecular representation have been evolved. Particularly, molecular graph representations are the more naturalistic representation of compounds and molecular representation learning with GNN which can use graph representations as the input has achieved success. Despite this progress, it is difficult to learn all of the vast chemical space with machine learning due to the lack of labeled data.

In recent years, to address this problem, self-supervised learning has been investigated with large unlabeled data. In this study, we constructed the GNN contrastive learning model and verified the effects of augmentation strategies in contrastive learning.

We attempted to improve the performance of molecular properties prediction by using four augmentation strategies in contrastive learning. Prediction performance was evaluated on both classification and regression tasks. In addition to MoleculeNet, which are the molecular properties prediction benchmarks and widely used, performance was also evaluated on other biological data sets.

P08-01

Predicting clinical laboratory test result related to urine tests in patients with type 2 diabetes mellitus with renal complications using clinical trial data

Hiroki ADACHI *, Takuya SEKIYAMA, Taku SAKAUE, Yasuo SUGITANI

Biometrics Department, Chugai Pharmaceutical Co., Ltd.

(* E-mail: adachi.hiroki16@chugai-pharm.co.jp)

Clinical laboratory tests are medical tests that analyze a patient's blood, urine etc., and use the results to understand the patient's characteristics and medical condition. Clinical laboratory test results are an important tool for improving the quality of medical care, as they are used to detect diseases at an early stage, determine the progression of a disease, and even determine the effectiveness of treatment. We focused on the prediction of urinalysis values that are rarely measured in routine medical care. It is known that patients with type 2 diabetes mellitus with renal complications have higher urinalysis scores due to decreased renal function. We created prediction models using urinalysis scores from clinical trial data of patients with type 2 diabetes mellitus with renal complications conducted in the past as the objective variable and the results of key laboratory tests as explanatory variables to evaluate the prediction accuracy and improve prediction accuracy. We selected clinical trials that included patients with type 2 diabetes mellitus with renal complications. We were able to create a prediction model for both multilevel and binary classification of urinary protein in a specific study that included patients with type 2 diabetes mellitus with renal complications. For multilevel classification of urine protein scores, a model using Random Forest was created and an accuracy of AUC 0.8090 obtained. For binary classification, a model using Support Vector Classifier was created and an accuracy of AUC 0.8974 was obtained. The clinical studies are designed to have a regular schedule of patient laboratory tests. Compared to actual medical care, it is thought that the tests are performed more rigorously and that there are fewer missing values. We believe that it is necessary to verify the applicability of the predictive model based on the data in actual clinical practice, and we plan to conduct such verification using actual clinical data.

P08-02

Machine learning models for predicting cross-reactivity of beta-lactam antibiotic allergy

Shoki HOSHIKAWA ^{*1}, Keisuke FUKUI², Yukiko KARUO¹, Atsushi TARUI¹, Masaaki OMOTE¹, Kazuyuki SATO¹, Kentaro KAWAI¹

¹Faculty of Pharmaceutical Sciences, Setsunan University

²Nozaki Tokushukai Hospital

(* E-mail: shoki.hoshikawa@setsunan.ac.jp)

In clinical practice, not a few patients have a penicillin allergy, and pharmacists use cross-allergy tables for beta-lactam antibiotics to propose changes to the medication. However, even when switching to a drug that is generally contraindicated, there are cases where no allergic symptoms occur, and there are things that cannot be explained using the cross-allergy table for beta-lactam antibiotics. In clinical practice, there are cases where, from the perspective of treating infectious diseases, it is necessary to reluctantly change to a drug from a different class with a broad spectrum of action that is not a beta-lactam antibiotic, but there is a need in the field to refrain from using drugs with as broad a spectrum of action as possible from the perspective of drug resistance. Therefore, we investigated the possibility of using AI to suggest drugs that can be used in the same class of antimicrobial agents as a substitute for narrow-spectrum beta-lactam drugs. Specifically, we evaluated the prediction ability of a leave-one-out approach to evaluating the combination of two drugs (in four categories: recommended, caution, principal contraindication, and contraindication) in 35 beta-lactam antimicrobial agents. Here, three fingerprints (MACCS, Morgan, and Topological fingerprints) and four machine learning methods (SVM, lightGBM, Random Forest, and k-nearest neighbor) were used for the study. As a result, the highest accuracy of 0.98 was achieved when predicting the recommended drug combinations using Morgan fingerprints with lightGBM. On the other hand, the lowest accuracy of 0.89 was obtained when predicting the drug combinations of the caution category using MACCS fingerprints with SVM. We then focused on the compounds for which the prediction did not work well and evaluated the reasons from the perspective of structural similarity. We also evaluated the relationship between the combinations in each category and structural similarity using Tanimoto similarity of fingerprints. Drug combinations that are recommended by AI while having low similarity may be useful in clinical drug selection.

P09-01

Modular photostable fluorescent DNA blocks for tracking collective movements of motor proteins

Ryota SUGIE ^{*1}, **Tomoki KITA**², **Shinsuke NIWA**^{2,3}, **Yuki SUZUKI**¹

¹Department of Applied Chemistry, Graduate School of Engineering, Mie University

²Graduate school of Life Sciences, Tohoku University

³Frontier Research Institute for Interdisciplinary Sciences(FRIS), Tohoku University

(* E-mail: 423m328@m.mie-u.ac.jp)

Active intracellular transport is carried out by the collective action of multiple motor proteins. The properties of this process depend not only on the number of motor proteins responsible for transporting the cargo but also on the specific types of motor proteins involved. DNA nanotechnology has provided an attractive approach for modeling such a complex system by allowing different types of motors to be linked in a programmable manner. The movements of the assembled complex are often tracked using total internal reflection fluorescence (TIRF) microscopy; however, the blinking and photobleaching of fluorescent dyes limit the duration of imaging, which in turn restricts detailed analysis of the collective movements. In this study, we developed a connectable photostable DNA nanostructure, designated as the fluorescence-labeled tiny DNA origami block (FTOB). The FTOB is a 4-helix bundle of approximately 8.4 nm in size, in which five (or six) fluorescent dyes have been integrated, resulting in its minimal blinking and photobleaching properties. By designing a pair of connector DNAs, FTOB can be heterodimerized after the attachment of the motor protein of interest via the ALFA-tag/NbALFA system, thereby enabling the formation of a mimic of a cargo transport complex with a selected combination of motor proteins. We believe that our modular FTOB system would serve as a novel tool for the investigation of cooperative cargo transport at the single molecule level.

P09-02

Size-Selective Capturing of Exosomes Using DNA Tripods

Ryosuke IINUMA ^{*1, 2}, **Xiaoxia CHEN**³, **Takeya MASUBUCHI**^{2, 4}, **Takuya UEDA**^{2, 5}, **Hisashi TADAKUMA**^{2, 3, 6}

¹JSR Life Sciences Corporation

²Graduate School of Frontier Science, The University of Tokyo

³School of Life Science and Technology, ShanghaiTech University

⁴Department of Cell and Developmental Biology, School of Biological Sciences, University of California San Diego

⁵Graduate School of Science and Engineering, Waseda University

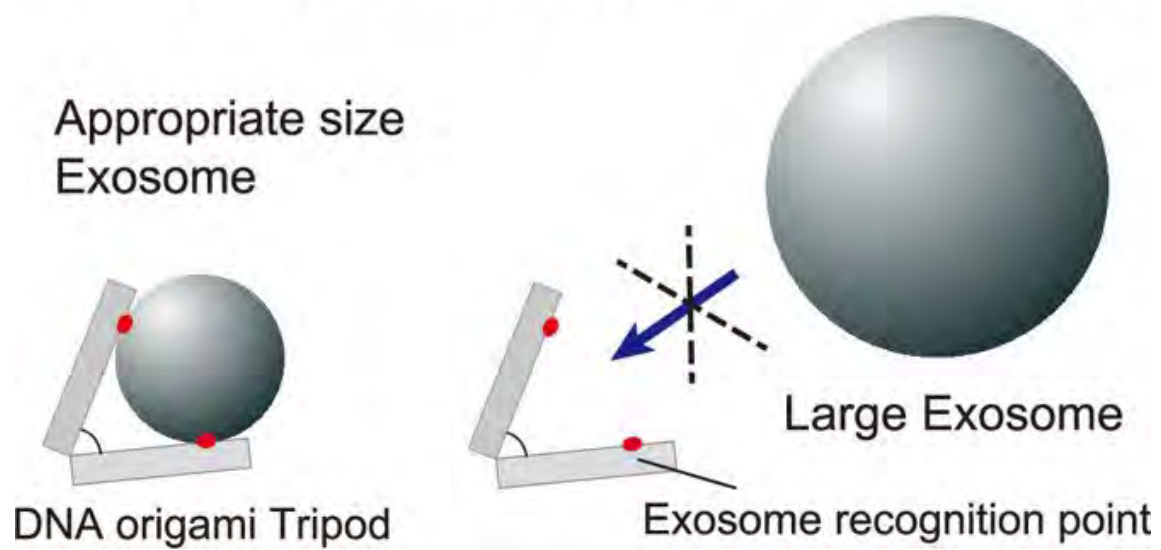
⁶Gene Editing Center, ShanghaiTech University

(* E-mail: Ryosuke_Iinuma@jls.jsr.co.jp)

Fractionating and characterizing target samples are fundamental to the analysis of biomolecules. Extracellular vesicles (EVs), containing information regarding the cellular birthplace, are promising targets for biology and medicine. However, the requirement for multiple-step purification in conventional methods hinders analysis of small samples. Here, we apply a DNA origami tripod with a defined aperture of binders (e.g., antibodies against EV biomarkers), which allows us to capture the target molecule.

For the first step, we verify the concept of size selective capture of target molecules using magnetic beads as the model, and confirm that the DNA tripod only captures beads whose size fits inside the defined aperture. Similarly, using exosomes from HT-29 cell line as a model, we show that our tripod nanodevice can capture a specific size range of EVs with cognate biomarkers from a broad distribution of crude EV mixtures. Next, we demonstrate that the size of captured EVs can be controlled by changing the aperture of the tripods. Finally, we applied the principle to a solid-supported capture system for EVs. This simultaneous selection with the size and biomarker approach should simplify the EV purification process and contribute to the precise analysis of target biomolecules from small samples.

Simultaneous selection by size and biomarker of exosome



P09-03

Anisotropic Swarming of DNA Modified Microtubules Under UV Light

Chung Wing CHAN *, Marie TANI, Masatoshi ICHIKAWA, Ibuki KAWAMATA, Akira KAKUGO

Division of Physics and Astronomy, Graduate School of Science, Kyoto University
(* E-mail: chan.wing.76t@st.kyoto-u.ac.jp)

Active matter systems, which covert free energy to generate their own motion and forces. The collective motion of these self-propelling agents often spontaneously displays emergent transport behaviors influenced by environments. Such systems can exhibit swarming behaviors, offering advantages like robustness and flexibility. Inspired by these phenomena, swarming in various active matter systems in a programmable manner is explored from artificial particles to living systems. Despite advancements in materials and control strategies, achieving programmable self-assembly in micro-scale swarming robots remains challenging. Molecular robots, such as microtubules (MTs)-kinesin systems, show potential in overcoming this challenge. In our study, we demonstrate the swarm of MTs can be controlled by UV and visible light by conjugating photo-responsive DNA(p-DNA) to MTs, presenting new pathways for optical control of swarming behavior. To obtain more physical insights from the experimental system, we also combine computer simulation with Vicsek type model, and effective hydrodynamic theory to investigate the swarming dynamics of p-DNA-conjugated MTs under optical patterns. Our preliminary result suggests that the motion direction can be oriented by the gradient of light intensity. Based on it, by manipulating optical patterns, our research sheds light on controlling MTs swarm, advancing the understanding of collective dynamics under external stimulation.

P09-04

De novo protein design of suitable binders for DNA origami-based devices

Jielin WANG¹, Peiqi HUANG¹, Hisashi TADAKUMA ^{*1, 2}

¹School of Life Science and Technology, ShanghaiTech University

²Gene Editing Center, ShanghaiTech University

(* E-mail: tadakuma@iqb.u-tokyo.ac.jp)

DNA origami-based nanodevices integrated with protein are powerful tools. Multi-binder systems are a promising approach to capture, mark, and treat target molecules and cells. Using commercially available binders (antibodies), we have recently shown the potential of DNA origami-based multi-binder systems, where we have successfully captured exosomes in a size- and biomarker-specific manner. However, commercially available binders (e.g., antibodies and nanobodies) are mostly strong binders with high k_{on} and low k_{off} , while suitable kinetic characteristics highly depend on the target and aim. Moreover, to capture, mark, and treat multi-protein complexes, multiple binders recognizing specific target regions are required to ensure that each protein is recognized and targeted individually. Here, we try to overcome these challenges by designing binders using de novo protein design methods, attempting to obtain weak binders (high k_{off} binders). This approach would allow each protein to readily dissociate due to the binders' low affinity, while the entire complex could be captured by multiple binders. We will present our latest results regarding dry (design) and wet (biochemical experiments) approach.

P09-05

Over the Membrane: Study of Nucleic Acid Sequence Transfer Using Cholesterol-Modified DNA

Rinka AOKI *, Keita ABE, Satoshi MURATA, Hideaki MATSUBAYASHI, Shinichiro M NOMURA

Molecular Robotics Laboratory, Department of Robotics, Division of Mechanical Engineering,, Graduate School of Engineering, Tohoku University
(* E-mail: aoki.rinka.r6@dc.tohoku.ac.jp)

Molecular robotics is a pioneering engineering field that focuses on designing and constructing sensors, control circuits, and actuators at the molecular level, thereby enabling the development of robots that operate effectively on scales ranging from micrometers to nanometers. Specifically, DNA-based information-processing technologies are gaining attention as control circuits for molecular robots. However, because these systems operate in homogeneous aqueous solutions, effective compartmentalization is essential to maintain functional separation and targeted operation of these molecular systems. To this end, micrometer-sized lipid vesicles (liposomes) were used. Liposomes, which encapsulate water-soluble substances, are expected to function as the "body" of molecular robots, but the lipid membrane poses a challenge by hindering the passage of macromolecules, such as DNA, making communication with the external environment difficult.

Our laboratory's previous research[1] has shown that the hybridization of cholesterol-modified ssDNA enables the transmission of ssDNA with specific base sequences across membranes into liposomes. We refer to this mechanism as the "Chabashira" mechanism. In this study, we focused on optimizing the spatial arrangement of cholesterol within the membrane, leading to the development of a refined 'Chabashira.' Specifically, we tried to improve the system so that all sequences can be delivered across the membrane, rather than just the terminal sequences.

We evaluated its behavior on giant liposomes membrane by assessing the delivered DNAs by the system and their distribution, thereby confirming that molecular communication across the membrane was indeed achieved.

This mechanism is expected to increase the complexity of control circuits in nanodevices. Additionally, it has the potential to bring about new innovations in fields such as medicine and environmental monitoring, including drug delivery and environmental sensing.

[1] K. Yoshida, K. Abe, Y. Sato, I. Kawamata, R. J. Archer, H. T. Matsubayashi, S. Hamada, S. Murata, S. Nomura, ChemRxiv 2024, DOI 10.26434/chemrxiv-2024-571kp. This content is a preprint and has not been peer-reviewed.

P09-06

Multi-reconfigurable DNA nanolattice guided by a combination of external stimuli

Yuri KOBAYASHI *, Yuki SUZUKI

Department of Applied Chemistry, Graduate School of Engineering, Mie University

(* E-mail: 423m323@m.mie-u.ac.jp)

The advancements in structural nucleic acids nanotechnology have enabled the construction of various stimuli-responsive nanomachines through molecular self-assembly. These efforts have now extended to the development of multi-reconfigurable nanodevices exhibiting more complex motions, for which operations of distinct movable parts in a combinatorial and reversible manner are required. Here, we report a multi-reconfigurable DNA origami lattice actuator capable of transforming into unique shapes depending on the combination of different types of external cues. The structure consists of nine frames, each composed of a rigid 4-helix bundle connected by flexible single-stranded DNAs. Each frame (except for the central frame) has two bridge strands that form tetraplex structures, such as i-motif or G-quadruplex, in response to pH changes or K⁺. By regulating tetraplex formations not only with chemical cues but also with their complementary suppressor strands, each frame shape is reversibly reconfigured, enabling the interconversion of different configurations of the lattice actuator. Our simple yet modular design approach will facilitate the development of intelligent biomaterials that exhibit specific transformations in response to combinations of different types of external stimuli.

P09-07

Construction of dual-responsive circular DNA origami nanoactuator

Ryoya SAKAGUCHI *, Yuki SUZUKI

Department of Applied Chemistry Molecular Biotechnology Laboratory, Mie University

(* E-mail: 424M325@m.mie-u.ac.jp)

The properties and functions of a material depend not only on its molecular composition but also the arrangement of its constituent molecules. A platform that enables the manipulation of the relative positions and postures of multiple molecules with nanoscale precision in two- and three-dimensional space has the potential to facilitate the development of novel materials whose functions can be switched on demand. As such a platform, we here report a construction of DNA origami two-dimensional nanoactuator with a circular shape. The circular DNA origami nanoactuator is designed based on a ringed 6-helix bundle having an intrinsic right-handed twist along its axis. The origami structure comprises repeated units of a transformable module, each containing i-motif (iM)-forming bridge strands. Upon iM formation induced by a pH change, the bridge strands contract, causing the module to bend. The cumulative effect of this bending results in writhing of the circular structure. The modules are also designed to be actuated by the addition of an anti-iM strand, which hybridizes with the iM-forming bridge to form duplex DNA. The stiff duplex DNA forces the modules to bend in the direction opposite to that induced by the iM formation, thereby causing another type of transformation from a relaxed circular to a compacted wavy circular shape. Our study will pave the way for the construction of DNA origami nanomachines and nanodevices that exhibit two-dimensional compression and extension.

P09-08

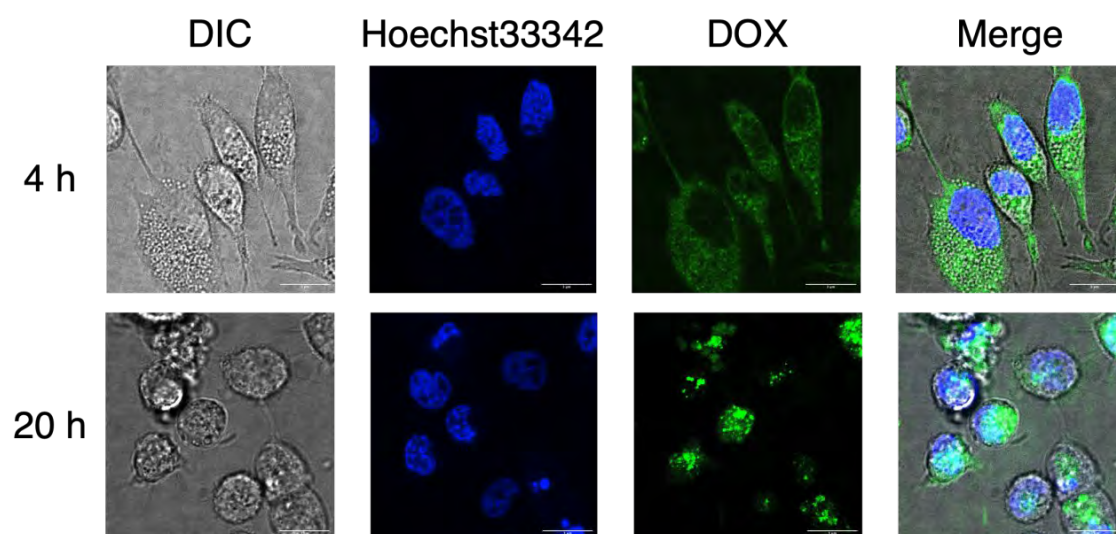
Evaluation of anticancer activity and investigation of cellular uptake mechanism of drug-loaded DNA Origami dendrimers for application to drug delivery system

Koichi TANIMOTO *, Yuki MINAMIDE, Maria HASHIMOTO, Haruki TANAKA, Yuki MANO, Akinori KUZUYA

Department of Chemistry and Materials Engineering, Kansai University
(* E-mail: k769965@kansai-u.ac.jp)

In recent years, drug delivery systems (DDS) have been actively developed as a technology for efficient and safe delivery of drugs to the site of disease. Nanocarriers are used to deliver therapeutic agents to target cells in DDS. One of the possible nanocarriers is DNA Origami, which can create arbitrary structures by folding long single-stranded circular DNA with many short single-stranded DNAs [1]. We have developed "DNA Origami dendrimers," which have four generations of four-branched dendritic structure from the center. This structure can bind doxorubicin (DOX), a popular anticancer drug, to the stems. In this study, we delivered DOX-loaded DNA Origami dendrimers into human cervical carcinoma (HeLa) cells and evaluated anticancer activity and cellular uptake mechanism.

As a result, we confirmed that red fluorescence from Ethidium Homodimer-1, which stains dead cells, was not observed after 4 hours of incubation in the Live/Dead Assay, but many cancer cells were dead after 20 hours. This may be due to the gradual release of DOX from DNA Origami dendrimers over time. This is also supported quantitatively by MTT Assay. In investigation of the cellular uptake mechanism, it was confirmed that green fluorescence from FAM modified with DNA Origami dendrimers co-localized with red fluorescence from LysoTracker Red DND-99, which stains lysosomes around the cell nucleus. This suggests that DNA Origami dendrimers were internalized via the endocytosis pathway. Similar results were obtained with DOX-loaded DNA Origami dendrimers, showing cellular uptake via the endocytosis pathway, and then only DOX was released from lysosomes and transported to the nucleus.



P10-01

Decision-making model to enhance subjective well-being through individualized lifestyle modifications based on counterfactual explanation

Yunosuke MATSUDA ^{*1}, Satoshi WATANABE¹, Yoji OKUGAWA¹, Nobuyuki NAKANISHI¹, Keishi MATSUMOTO¹, Shinya HAYASAKA²

¹BATHCLIN Corporation

²Faculty of Human Life Sciences, Tokyo City University

(* E-mail: matsuda_yunosuke@bathclin.co.jp)

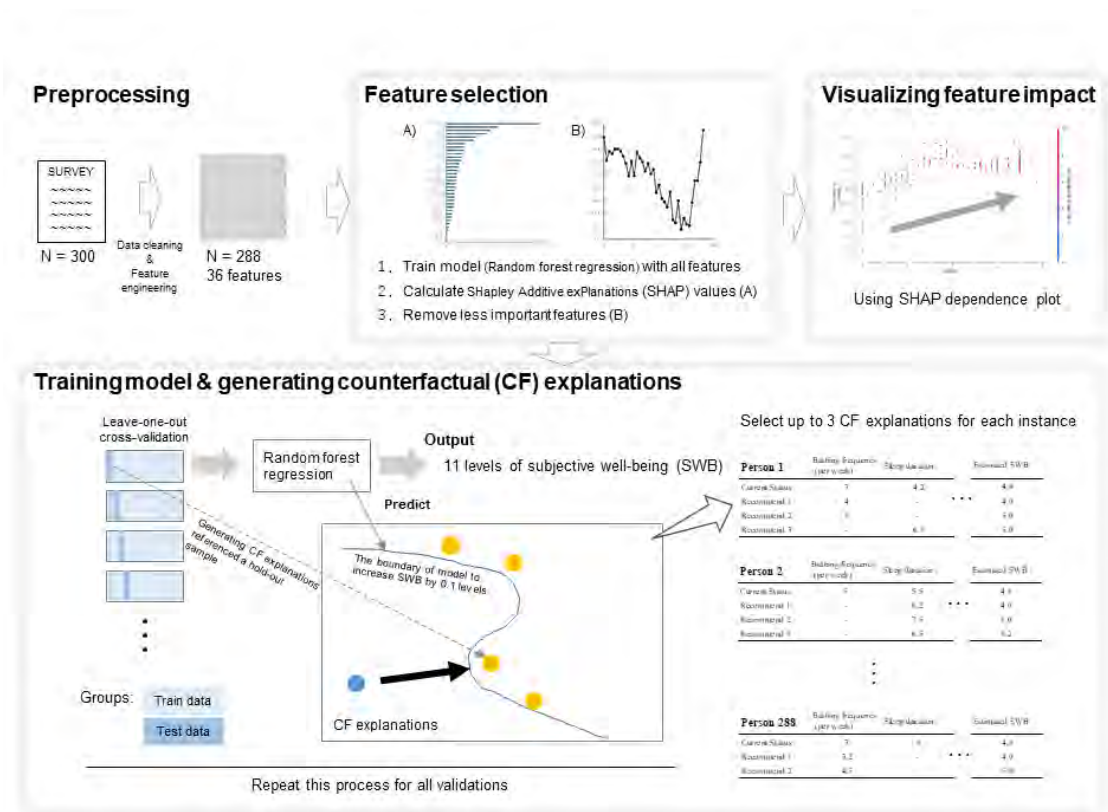
Subjective well-being (SWB), an individual's judgment about their overall well-being, is of broad interest, spanning both popular writing and scientific work. Some studies have reported that SWB is a multifaceted concept influenced by various factors, including economic stability, health, and personal growth. Numerous research reports have documented methods to enhance SWB, such as improving sleep quality. However, it is important to note that these methods are not universally applicable to all individuals. Therefore, we examined the feasibility of supporting individualized lifestyle modifications that increase SWB utilizing counterfactual (CF) explanation. CF explanation, a type of explainable machine learning technique, describes the changes needed in a sample to flip the outcome of the prediction. We developed a decision-making model based on 2022 survey data on lifestyle and quality of life, including an 11-point scale question on SWB. This decision-making model includes the SWB prediction model and employs CF explanation.

Using SHapley Additive exPlanations (SHAP) values for feature selection, we identified seven attributes (e.g., childhood memories of bathing, marital status, and household income) and seven lifestyle features (e.g., frequency of bathtub bathing, sleep duration, and the difference between bath time on weekdays and weekends) as effective predictors. These 14 features were used to train random forest regression for SWB prediction. Leave-one-out cross-validation was used to evaluate the model performance and generate CF explanations for each hold-out sample.

SHAP dependence plots calculated from the model tended to be consistent with previous research (e.g., predicted SWB peaked at approximately 7 hours of sleep duration and gradually increased with frequency of bathtub bathing from 0 to 4 per week). We obtained 859 CF explanations and found that frequency of bathtub bathing was included mainly. Focusing on CF explanations in the low

frequency of bathtub bathing group, it was found that there was a difference depending on the sleep duration of instances. It was observed that some CF explanations recommended increasing not the frequency of bathtub bathing but the sleep duration for short sleepers (≤ 6 hours), even under low frequency of bathtub bathing conditions. Moreover, most CF explanations recommended healthy lifestyle modifications for short or long sleepers (≥ 8 hours) and instances of low frequency of bathtub bathing. Thus, this model allowed for both individualized support and previous standardized support.

In this study, CF explanations generated by our model were diverse and highly valid. However, the trained data include concerns over the impact on SWB by potentially unobserved variables and the lack of certainty in the direction of causality. Clarifying causal relationships or limiting the output of the model to effects observed in intervention studies would make the model more valuable and predictive.



P11-01

Evaluation of Data-Driven Drug Discovery Approaches: Utilizing Redmine Ticket Management System for Tracking and Analyzing Activities

Kosuke TAKEUCHI *, Takayuki SERIZAWA

Group1, Modality Research Laboratory, DAIICHI SANKYO CO., LTD.

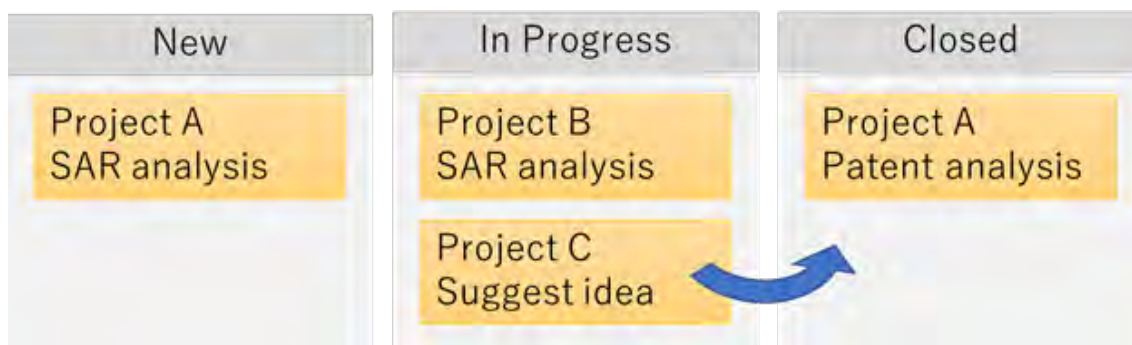
(* E-mail: kosuke.takeuchi@daiichisankyo.com)

In Daiichi Sankyo Co., Ltd. (DS), a new team, the Data-Driven Drug Discovery (D4) group, was formed, consisting of chemoinformaticians and data scientists. Our main mission was to accelerate the traditional experimental design-make-test-analyze (DMTA) cycle through data science. As a result, we achieved approximately 20% increase in medicinal chemistry project time efficiency[1]. As the next step in our aim, we focused on making concrete and quantitative contributions helping to obtain drug candidates in internal research projects, while maintaining research time efficiency at the same level. To measure D4 activity quantitatively, a new system was developed on Redmine[2], a highly flexible and customizable open-source software (OSS). We have already been using Redmine as a task management system through tickets, so a new feature was appended to record D4 activity on each ticket.

In our presentation, we will show the recording system and our preliminary results. Compared to the previous report[1], we found that we continued committing SAR analysis and patent analysis. Additionally, we also found that there was an increase in proposing compound ideas for research projects. By customizing a task management system, it has become easier to monitor our daily efforts and to analyze which D4 activities are working well retrospectively and quantitatively.

[1] Kunimoto R, Bajorath J & Aoki K. From traditional to data-driven medicinal chemistry – a case study. Drug Discov Today 27, 2065-2070, 2022.

[2] <https://www.redmine.org/>



P11-02

Japanese Food Ontology Development

Chihiro HIGUCHI ^{*1, 2}, **Ai OYA**¹, **Michihiro ARAKI**^{1, 3, 4, 5}

¹Artificial Intelligence Center for Health and Biomedical Research (ArCHER), National Institutes of Biomedical Innovation, Health and Nutrition (NIBIOHN)

²School of Medical and Dental Sciences, Tokyo Medical and Dental University

³Graduate School of Medicine, Kyoto University

⁴Graduate School of Science, Technology and Innovation, Kobe University

⁵National Cerebral and Cardiovascular Center

(* E-mail: higuchi@nibiohn.go.jp)

Nutrition is a key component of good health. However, many individuals may face challenges in obtaining sufficient nutrition due to factors such as allergies that affect their health. To conduct nutritional research that yields reliable and comparable results, it is necessary to use uniform terminology and accurate food descriptions. Consequently, there is a demand for a computer-readable Japanese food ontology that can precisely represent the characteristics and relationships of various foods.

The National Health and Nutrition Survey (NHNS) is a data collection initiative aimed at assessing the nutritional intake and lifestyle of the Japanese population, with the ultimate goal of improving their health. The NHNS involves physical measurements and blood tests from 3,412 households across 300 districts in Japan. The survey covers 1,630 Japanese food items, classified into three tiers: large categories (e.g., cereals, legumes, vegetables), medium categories (e.g., rice and its processed products, wheat and its processed products), and small categories (e.g., flour products, bread products).

We employed the Web Ontology Language (OWL) to describe the NHNS data, utilizing its hierarchical classification scheme and appropriate Uniform Resource Identifiers (URIs). This ontology is published as an alpha version of FGNHNS on BioPortal (<https://bioportal.bioontology.org/ontologies/FGNHNS>). Our future work includes enhancing the ontology by adding Wikidata information, linking with FoodOn, integrating with the Standard Tables of Food Composition in Japan, linking with the Agricultural Vocabulary System, and incorporating food allergy information.

In developing this ontology, we have been leveraging basic models such as

Large Language Models (LLMs), and the environment surrounding LLMs has recently evolved further. In this poster presentation, we will introduce our current efforts based on the latest updates.

P11-04

Development of a model to predict the severity of systemic lupus erythematosus using LIFE Study data

Kiyohiro TOYOFUKU ^{*1}, Koichiro KATO^{1, 3}, Haruhisa FUKUDA²

¹Department of Applied Chemistry, Graduate School of Engineering, Kyushu University

²Department of Health Care Administration and Management, Graduate School of Medical Sciences, Kyushu University

³Center for Molecular Systems, Kyushu University

(* E-mail: toyofuku.kiyohiro.864@s.kyushu-u.ac.jp)

Systemic lupus erythematosus (SLE) is an autoimmune disease that causes inflammation and tissue damage in various organs throughout the body. It is one of the designated intractable diseases, and it is estimated that there are approximately 60,000 to 100,000 patients throughout Japan.

Although advances in steroids and immunosuppressive drugs have greatly improved the prognosis of life, quality of life (QOL) is still a problem due to organ damage caused by SLE and complications associated with long-term steroid use. Therefore, it is considered important to prevent worsening of the disease and to maintain a stable disease state for a long period of time with minimal steroid use, with the goal of preventing organ damage and improving health-related QOL. In Japan, SLE has been treated with immunosuppressive agents (tacrolimus and mizoribine), but in recent years, important drugs for SLE treatment such as hydroxychloroquine, mycophenolate mofetil (immunosuppressive agent), and belimumab/anifrolumab (biological agent) have become covered by insurance one after another, making it possible to treat SLE as in Europe and the US. This has made it possible to provide SLE treatment similar to that in Europe and the United States. Currently, Japan has a large number of therapeutic agents to choose from, including conventional therapies, but the treatment methods suitable for Japanese patients and the optimal treatment strategy to prevent worsening of the disease (relapse) and to maintain a stable state (remission) have not yet been clarified. Therefore, we planned a data-driven investigation to find out which factors are important for the maintenance of disease severity and remission in SLE. By conducting this study, we believe that new aspects of SLE in the Japanese population will be clarified, which will help improve patient outcomes and maintain QOL.

To collect the data of SLE patients, we use a database produced by the Longevity Improvement & Fair Evidence (LIFE) Study. In this study, SLE patients were

identified from LIFE Study data using ICD10 codes and extracted by linking information on specific health checkups and medications. Since patients with severe SLE generally receive high doses of steroids and immunosuppressive drugs, we scored SLE severity based on the drugs administered. Next, these data were arranged into a time series, and patients were classified into three severity classes based on the shape of the time series. Then, a random forest classification model was constructed using patient information (e.g., laboratory values) as explanatory variables and the severity classes calculated from the time series as objective variables, and explanatory AI (SHAP) was used to search for important factors. The results suggest that there is an association between blood LDL cholesterol level and the risk for severe disease.

P11-05

Exploring unexpected factors related to glaucoma onset in diabetes patients using LIFE Study data

Kaito SASAKI ^{*1}, Qiao HE², Koichiro KATO^{1, 4}, Haruhisa FUKUDA³

¹Department of Applied Chemistry, Graduate school of Engineering, Kyushu University

²Department of Systems Life Sciences, Graduate School of Systems Life Sciences, Kyushu University

³Department of Health Care Administration and Management, Graduate school of Medical science, Kyushu University

⁴Center for Molecular Systems, Kyushu University, Kyushu University

(* E-mail: sasaki.kaito.864@s.kyushu-u.ac.jp)

It is generally said that glaucoma is one of the main cause of blindness in Japan. According to the survey from Japan Glaucoma Society (JGS), one in twenty people over 40 years old suffers from glaucoma. In addition, nine out of ten glaucoma patients don't go to ophthalmology and are not cured. This means older people are vulnerable to glaucoma and its patients are difficult to be aware of its symptoms. One of the main reasons why glaucoma patients have difficulty in recognizing its sign is that progress of glaucoma is slow. Appearance of invisible places and narrowing visible field are the typical symptoms of glaucoma. The symptoms slowly emerge, and the patients are hard to notice them in the initial phase. Thus, most glaucoma patients take delay treatment of glaucoma and in some case, become blindness. Moreover, although glaucoma patients receive proper treatments, symptoms of glaucoma can't improve completely. The purpose of treatment is to stop or slow the progress of glaucoma, not to restore vision. On the other hand, various diseases and medications are found to be associated with the development of glaucoma. In particularly, a meta-analysis shows a strong association with the development of glaucoma and diabetes. It is estimated that there are approximately 10 million diabetes patients in Japan. As mentioned above, its patients are more likely to contract glaucoma than non-diabetics. In contrast, there are patients with diabetes who do not develop glaucoma. The factors that influence this difference are currently unknown. Therefore, to improve quality of life of diabetics and curtail medical expenses towards glaucoma treatments, it is significant to identify the factors that affect the development of glaucoma among diabetic patients. In this study, we will conduct a data-driven study using the LIFE Study data, a database that integrates health-related data on health, medical care, long-term care, and

administration held by local governments on a per-resident basis, to identify factors correlated with the development of glaucoma in diabetic patients. We extracted diabetes patients during 2017-2018 from specific health checkups dataset based on the definition of inspection values (fasting blood pressure ≥ 126 mg/dL and HbA1c $\geq 6.5\%$) and medication of diabetes drug. Each extracted patient was tied to specific health checkup information (blood pressure, cholesterol level etc.) for explanatory variable and labeled with the presence or absence of glaucoma development for the objective variable. A model was then constructed to classify whether or not the patient developed glaucoma using LightGBM and SHAP (one of the explainable AI methods). Analysis of the predictive rationale for the classification model suggested an association between low HbA1c and glaucoma development. In addition to this analysis, factor exploration is also underway in survival analysis methods.

P11-06

Survival Analysis of Chronic Kidney Disease Using Multi-Regional Data from the LIFE Study

Hiromu MATSUMOTO ^{*1}, Tomohiro RYU², Koichiro KATO^{1, 3}, Fukuda HARUHISA⁴

¹Department of Applied Chemistry, Graduate School of Engineering, Kyushu University

²Department of Chemistry, Graduate School of Science, Kyushu University

³Center for Molecular Systems, Kyushu University

⁴Department of Health Care Administration and Management, Graduate School of Medical Sciences, Kyushu University

(* E-mail: matsumoto.hiromu.238@s.kyushu-u.ac.jp)

By late 2022, the number of dialysis patients in Japan had reached approximately 350,000, highlighting the severity and prevalence of chronic kidney disease (CKD). Given the irreversible nature of renal function decline, early detection and prevention are crucial. However, CKD is often asymptomatic in its early stages, making timely identification challenging. Consequently, there is a critical need to develop predictive models that can identify early indicators of CKD and more effectively target at-risk populations.

Accurate prediction of kidney function trajectories, particularly glomerular filtration rate (GFR) and estimated GFR (eGFR), is crucial for the timely diagnosis and classification of chronic kidney disease (CKD), diagnosed when eGFR falls below 60 mL/min/1.73m². While existing machine learning models show potential in predicting eGFR, they heavily depend on prior eGFR data. Developing models that can predict eGFR decline independently of historical measurements is crucial for enhancing predictive accuracy and understanding the factors driving renal function decline.

To address these challenges, we utilized data from the Longevity Improvement & Fair Evidence (LIFE) Study, a large-scale, multi-regional cohort study that integrates health-related data. This comprehensive database allowed for survival analysis across 14 municipalities, enabling an examination of the impact of regional factors on survival outcomes alongside various health metrics.

Diabetic patients with fasting blood glucose levels of 126 mg/dL or higher and eGFR between 60 and 70 mL/min/1.73m² were classified as high-risk and targeted for tracking in this study. The survival analysis dataset was split into an 8:2 train-test ratio, ensuring a balanced distribution of patients who crossed the eGFR threshold of 60 mL/min/1.73m² during follow-up and those who did

not. The train set contained 8,563 records from 2,354 patients, and the test set included 2,192 records from 389 patients. We incorporated explanatory variables such as biometric measurements (e.g., height, weight, blood pressure), self-reported lifestyle factors (e.g., exercise, sleep habits), and regional residence as a dummy variable. Survival analysis was conducted with the decline in eGFR as the primary endpoint, specifically defining the event as reaching eGFR levels below 60 mL/min/1.73m². The Cox proportional hazards model and Random Survival Forest were employed for model construction.

The survival analysis using the Cox proportional hazards model identified risk factors consistent with those previously reported. Furthermore, the results suggested that regional residence may be associated with the rate of progression of CKD. Ongoing work includes further refinement of these findings through advanced analyses, including the use of the SurvSHAP tool, which provides interpretability of survival models, to gain deeper insights into the explanatory variables.

P11-07

Data-driven search for diseases whose patient numbers are associated with weather variability using LIFE study data

Kensei ORITA *¹, SUN KIM¹, Koichiro KATO^{1, 2}, Haruhisa FUKUDA³

¹Department of Applied Chemistry, Graduate School of Engineering, Kyushu University

²Center for Molecular Systems, Kyushu University

³Department of Health Care Administration and Management, Graduate School of Medical Sciences, Kyushu University

(* E-mail: orita.kensei.423@s.kyushu-u.ac.jp)

In recent years, the climate has been changing rapidly, and these variations have become increasingly noticeable year by year. The relationship between weather and health status has long been studied, but only a limited number of diseases, such as sinusitis and asthma, have been identified due to the huge amount of data required. Therefore, this study overcame the lack of data by using a database produced by the Longevity Improvement & Fair Evidence (LIFE) study, a longitudinal cohort database that collected and linked administrative claims data for residents of participating municipalities.

In this study, monthly data on temperature, humidity, and precipitation were obtained from the Japan Meteorological Agency website for weather data. For medical data, monthly data on the number of patients with various diseases were calculated using claims data from the LIFE study. The number of patients was adjusted for age and then analyzed by sex. Seven years of data from April 2015 to March 2022 were used for the analysis.

A vector autoregressive (VAR) model was employed for the time series analysis. This model was chosen because it captures the dynamic interdependencies between multiple time series data and allows for a comprehensive analysis of the impact of changing weather patterns on health status.

The results of the analysis suggest an association between some time series of weather data and those of the patient counts. For example, the time series of average temperature and that of the number of patients in schizophrenia could be related. The results of this study may contribute to the development of models based on weather data to predict the number of patients with various diseases, which could lead to improved health care and public health measures in local communities. In the future, more practical models will be developed by analyzing a wider range of data sets and examining their applicability to other regions.

P11-08

Analysis of interactions between fatty acid membranes with pH-dependent phase structures and nucleic acid monomers using Molecular Dynamics simulation

Ryoji ABE ^{*1}, Taren GINTER¹, Kosuke FUJISHIMA^{1, 2}

¹Department of Life Science and Technology, Institute of Science Tokyo

²Earth-Life Science Institute, Institute of Science Tokyo

(* E-mail: fujiki.r.aa@m.titech.ac.jp)

Fatty acids are amphipathic molecules with simple structures and have attracted attention as prebiotic membrane compartments. Furthermore, nucleic acids are molecules that serve as the basis for replication, and their synthesis in prebiotic environments and experiments has been confirmed. Recent experimental verification suggests the possibility of coevolution between fatty acids and nucleic acids. Fatty acid vesicles have also been shown to form at a different pH range in the presence of nucleic acid monomers, suggesting that nucleic acids reduce the pH sensitivity of fatty acids. Moreover, two different diffusion rates were observed for nucleic acid monomers in the presence of fatty acid vesicles, suggesting that nucleic acid monomers are adsorbed to the fatty acids.

Despite these experimental results, the detailed mechanism of fatty acid-nucleic acid interaction has yet to be elucidated. However, a thorough understanding of this interaction is essential in clarifying the prebiotic coevolution of fatty acid membranes and nucleic acids. Therefore, to understand this detailed mechanism, we conducted all-atom molecular dynamics simulations using GROMACS and the CHARMM36 force field to analyse the interaction.

We conducted simulations for 200 ns on a system comprising a fatty acid membrane and nucleic acid monomers in an aqueous solution. The fatty acid membranes were 1:1 to 2:1 mix of protonated and deprotonated oleic acid (18 carbons, 1 degree of unsaturation, cis) based on the conditions necessary for vesicle formation. We also calculated the radial distribution functions of heteroatoms in nucleic acids for each oxygen atom of oleic acid to interpret the results.

Our analysis demonstrated the formation of hydrogen bonds between oleic acid and the nucleobases and sugars. We observed differences in the formation of hydrogen bonds based on the positions of the hydrogen donors and acceptors of the nucleobases. We also found that the probability of forming hydrogen bonds varies depending on the position of the hydroxyl group in the sugar moiety. Furthermore, we noted differences in forming hydrogen bonds with nucleic acids

for fatty acid membranes with different protonation composition ratios. This study partially reproduced the experimental results of previous studies and clarified the contribution of hydrogen bonds in the interaction between the fatty acid membrane and nucleic acid monomers. In addition, the hydrogen bonds formed between the nucleic acid and the fatty acid membrane differ depending on the type of nucleobase and the protonation composition ratio of the fatty acid, suggesting coevolutionary selectivity between fatty acids and nucleic acids. Further simulations are being performed to systematically understand the interactions between fatty acids and nucleic acids and clarify the coevolution of fatty acids and nucleic acids related to the origin of life.

P11-09

A Virtual Reality Platform for Molecular Dynamics Based on Unity Engine

Yuhui ZHANG ^{*1, 2}, Gregory GUTMANN², Akihiko KONAGAYA²

¹School of Computing, Tokyo Institute of Technology

²Molecular Robotics Research Institute, Co., Ltd.

(* E-mail: zhang.y.av@m.titech.ac.jp)

We present a novel virtual reality (VR) platform for Molecular Dynamics (MD) built with the Unity engine. While primarily aimed at advancing Molecular Robotics research, our system's versatility allows for broader applications across various scientific fields. Our system supports common MD file formats, including PDB files as well as GRO and DCD trajectories, ensuring seamless integration with existing workflows. Leveraging Unity's rendering features, we have implemented a highly efficient multi-platform rendering system capable of handling approximately 1 million atoms in real-time. A collection of tools is also provided to enhance the user experience of their interaction with the virtual world.

Furthermore, it may connect to our simulation server for real-time interaction with ongoing simulations. While still a work in progress, our system aims towards a collaborative experience, where multiple users can engage in a shared virtual environment, facilitating real-time interaction and discussion. Our system is designed to integrate with popular molecular dynamics software, streamlining the transition between traditional and VR-based workflows. Our VR platform offers the MD community immersive, real-time scientific exploration and analysis opportunities.

P11-10

From Computer-Assisted Routine/Repeated 'Automation' to AI-Assisted Future-Oriented 'Autonomous (Intelligent/Creative)': Division and Impact of 'Automation' and 'Autonomous' in Research Contents

Kohtaro YUTA *

In Silico Data,Ltd.

(* E-mail: k-yuta@insilicodata.com)

Preface : The era has always continued to develop along with technological advancements and changes, and this progress has evolved in a way that supplements the functions that humans need or find difficult. For hundreds of years, from the agricultural era to the present, there have been fields where technology cannot be applied. That is human intellectual and creative abilities. The latest large-scale generative AI has the revolutionary power to step into the human intellectual and creative abilities, which were previously impossible to apply with conventional technology, and change the era. In the near future, in addition to the current 'automation,' 'autonomous (intelligent/creative)' will advance in all fields, and the era will transition.

Yuta believes that in the near future, research contents will be developed under new technology (AI), and proposes that research contents should be classified into conventional 'automation' parts and near-future 'autonomous (intelligent/creative)' parts and addressed accordingly.

1.Classification of Research Contents into 'Automation' and 'Autonomous (Intelligent/Creative)'

Research contents can be broadly classified into two types: 'Automation' research and 'Autonomous (Intelligent/Creative)' research.

Autonomous research involves contents where the application of computers is difficult or impossible. These contents heavily depend on the abilities of researchers and include advanced activities such as various ideas, creativity, insights, judgments, and decisions.

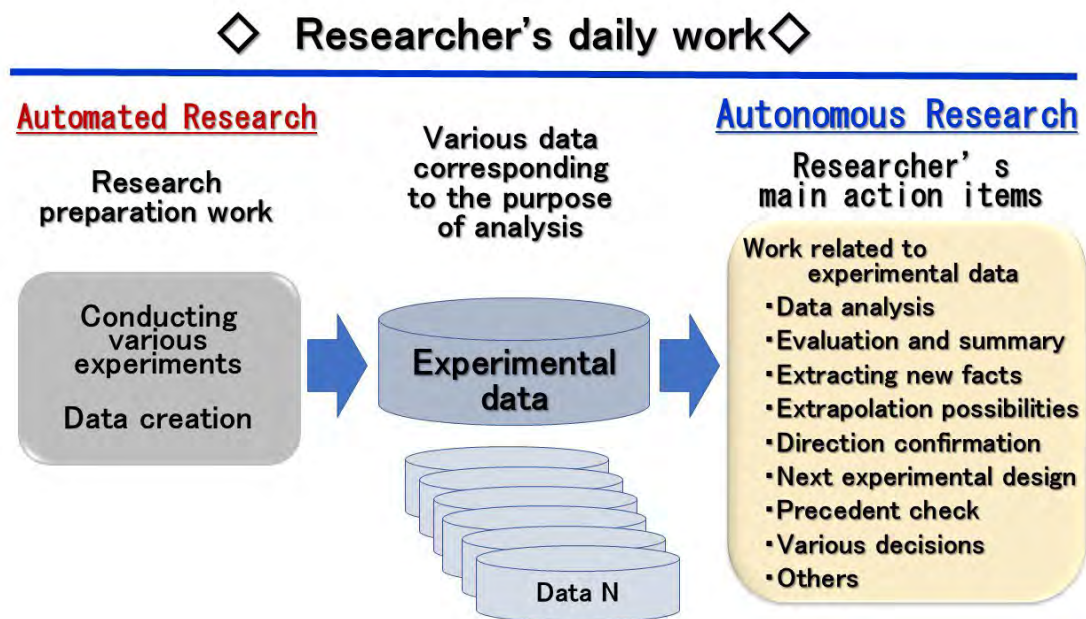
2. The Proportion of 'Automation' Research and 'Autonomous' Research in Research Contents

The flow of research contents can be divided into three stages: the preparation stage before starting the research, the process of collecting experimental data after starting the research, and the process of summarizing the experimental data to verify hypotheses or discover new facts, and then compiling and presenting the results. Looking at this flow, the content of research contents in

the first and last stages falls under 'autonomous' research. On the other hand, the research contents in the second stage are mainly 'automation' research supported by computers.

Summary :

To adapt to research focused on “autonomy” in the new era, it is important to classify and evaluate the content of “automation” and “autonomy” within the research itself. As a result of classifying the research content, it was found that 70% to 80% of the research content is related to “autonomy.” To improve the efficiency and quality of research, it is necessary to improve “autonomous” research, which could not be enhanced by the application of conventional technology. Large-scale generative AI supports this.



Chem-Bio Informatics Society Annual Meeting 2024 Abstracts
April 1, 2025

Editor in chief :	Kenji Mizuguchi, Yayoi Natsume
Production Director :	Akihiko Konagaya
Production Staff :	Naomi Komiyama, Mari Shiozuka, Sae Kishi Megumi Takazawa, Masumi Fujita
Published by :	Chem-Bio Informatics Society (CBI) Annual Meeting 2024 Kyowa Create 1st Building 3rd Floor 3-11-1 Shibaura, Minato-ku, Tokyo, JAPAN

E-mail : cbi2024@cbi-society.org
URL: <https://cbi-society.org/taikai/taikai24/>
Copyright© 2024CBI Society All rights reserved.
ISBN : 978-4-910628-15-8