イノベーションに情熱を。
ひとに思いやりを。

Daiichi-Sankyo

AWSを活用したデータ駆動型の創薬化学研究基盤の構築

CBI学会2022年大会
10/25/2022

第一三共株式会社　研究統括部　スマートリサーチ推進グループ
国本　亮

会社紹介（第一三共株式会社）

パーパス（存在意義）

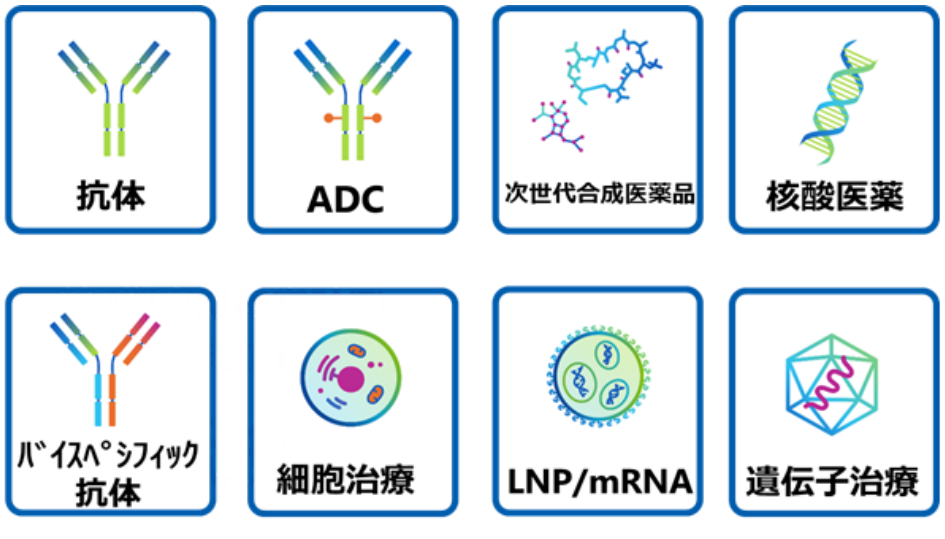
世界中の人々の健康で豊かな生活に貢献する

ミッション

革新的医薬品を継続的に創出し、多様な医療ニーズに応える医薬品を提供する

第一三共のマルチモダリティ戦略

Optimized modality



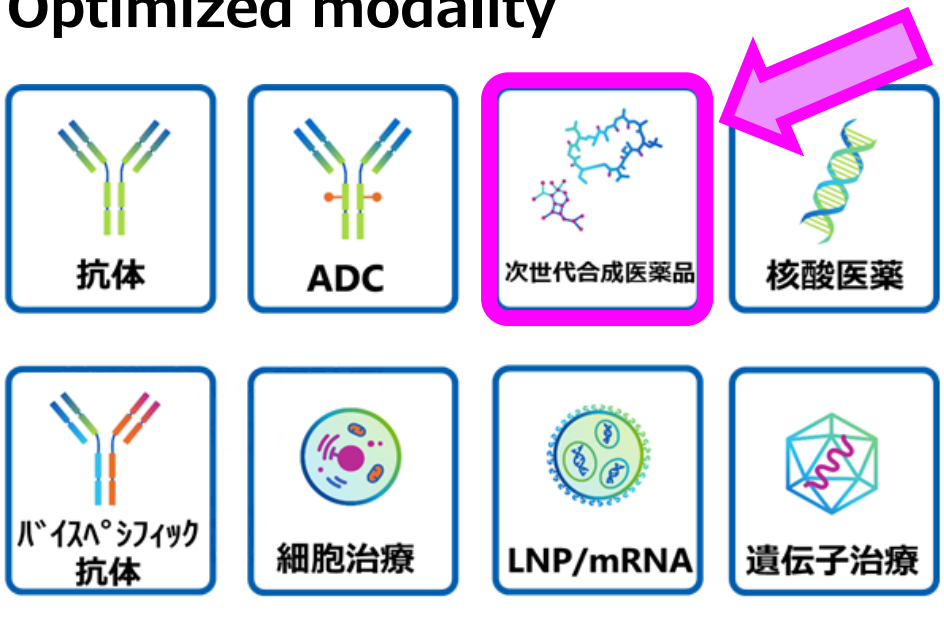
アンメット
メディカルニーズの
高い疾患

- **創薬標的や疾患に適したモダリティの選択に加え、新規モダリティの開発も同時に進めることにより、最適なモダリティを創製**
- **持続的成長を実現する鍵は、次の成長ドライバーの適切な評価と判断**
  - 適切な評価と優先順位付けを行い、ポテンシャルの高い医薬品候補を継続的に創出
  - 特定された有望な医薬品候補の開発を加速

モダリティ：創薬手段

第一三共のマルチモダリティ戦略

Optimized modality



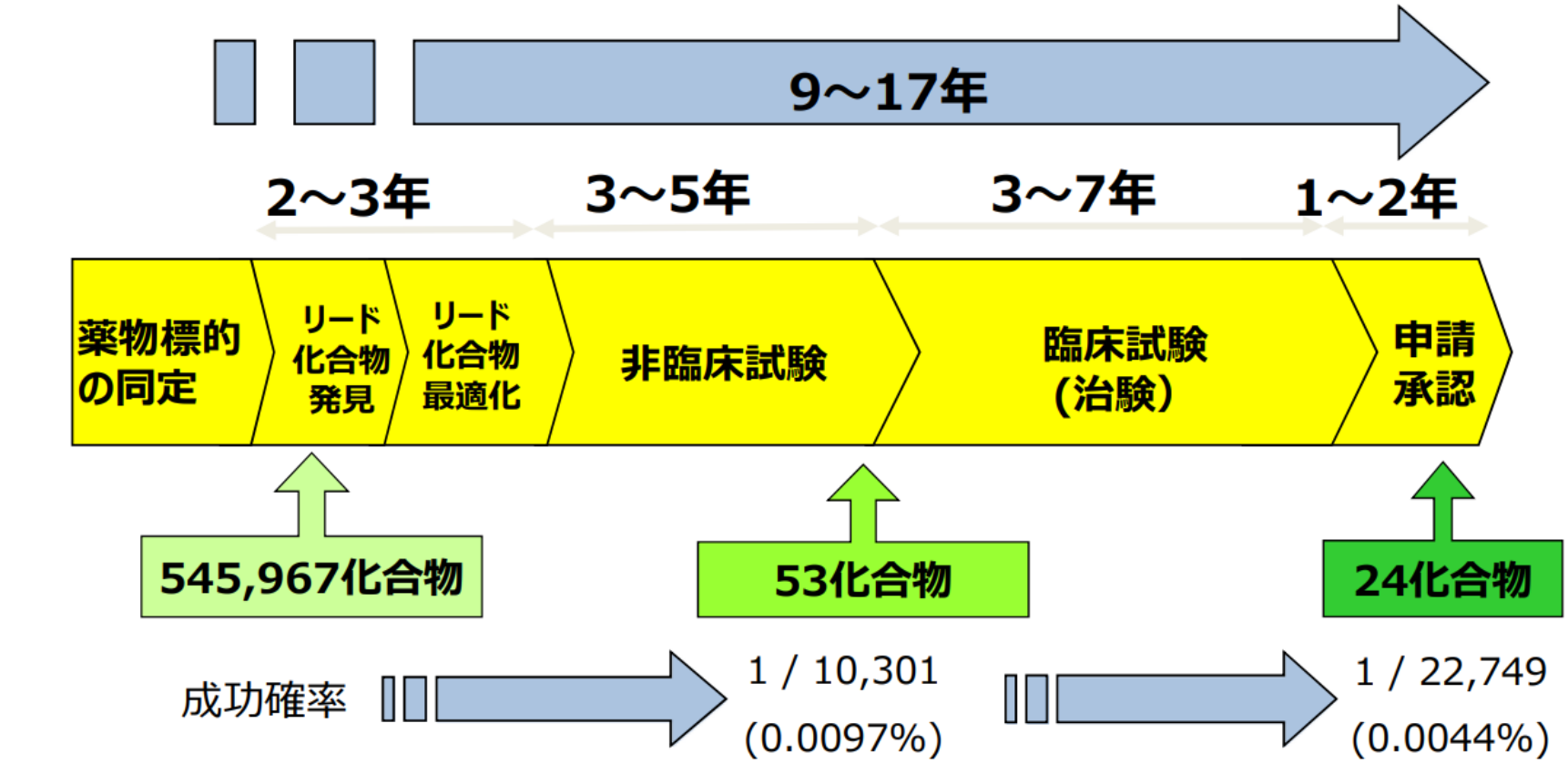
アンメット
メディカルニーズの
高い疾患

- **創薬標的や疾患に適したモダリティの選択に加え、新規モダリティの開発も同時に進めることにより、最適なモダリティを創製**
- **持続的成長を実現する鍵は、次の成長ドライバーの適切な評価と判断**
  - 適切な評価と優先順位付けを行い、ポテンシャルの高い医薬品候補を継続的に創出
  - 特定された有望な医薬品候補の開発を加速

モダリティ：創薬手段

医薬品開発に要する期間と成功確率

○ 医薬品の開発には10年以上の時間と数百億～数千億円規模の費用が必要。
○ 成功確率は年々低下（20年前:1/1.3万→現在:1/2.3万）し、難易度が上昇。

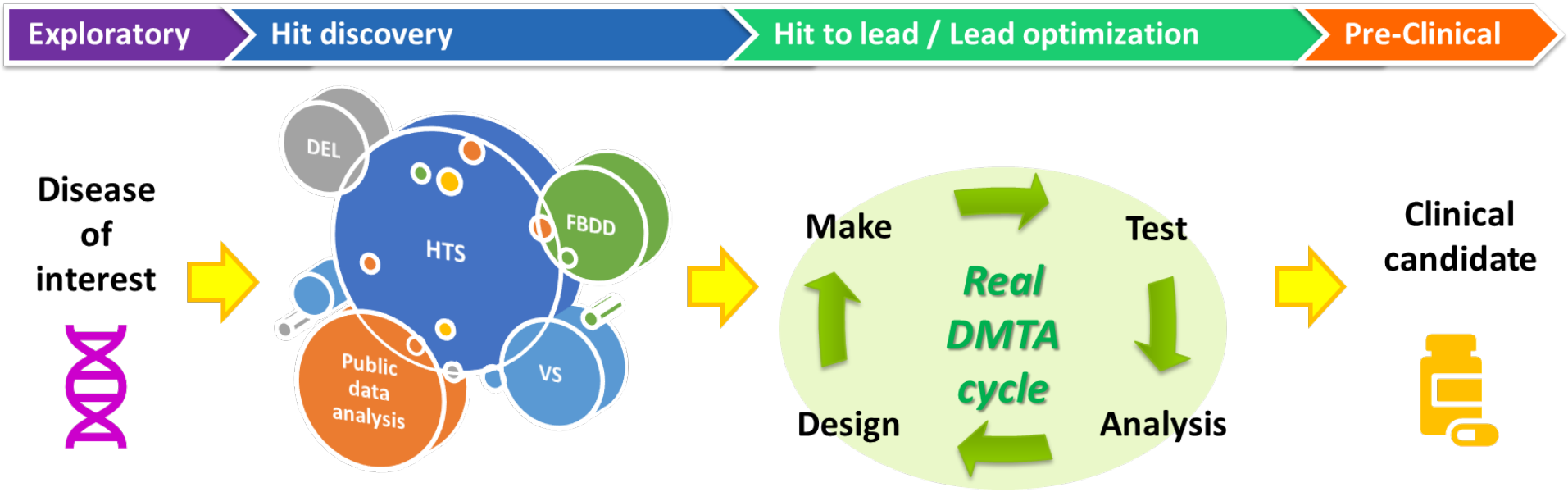


出典：日本製薬工業協会調べ

「医薬品産業ビジョン2021」（厚生労働省）（ https://www.mhlw.go.jp/content/10800000/000831974.pdf ）を加工して作成

創薬研究初期の基本的なワークフロー

疾患標的とモダリティを選定、タネを探してモノを磨く

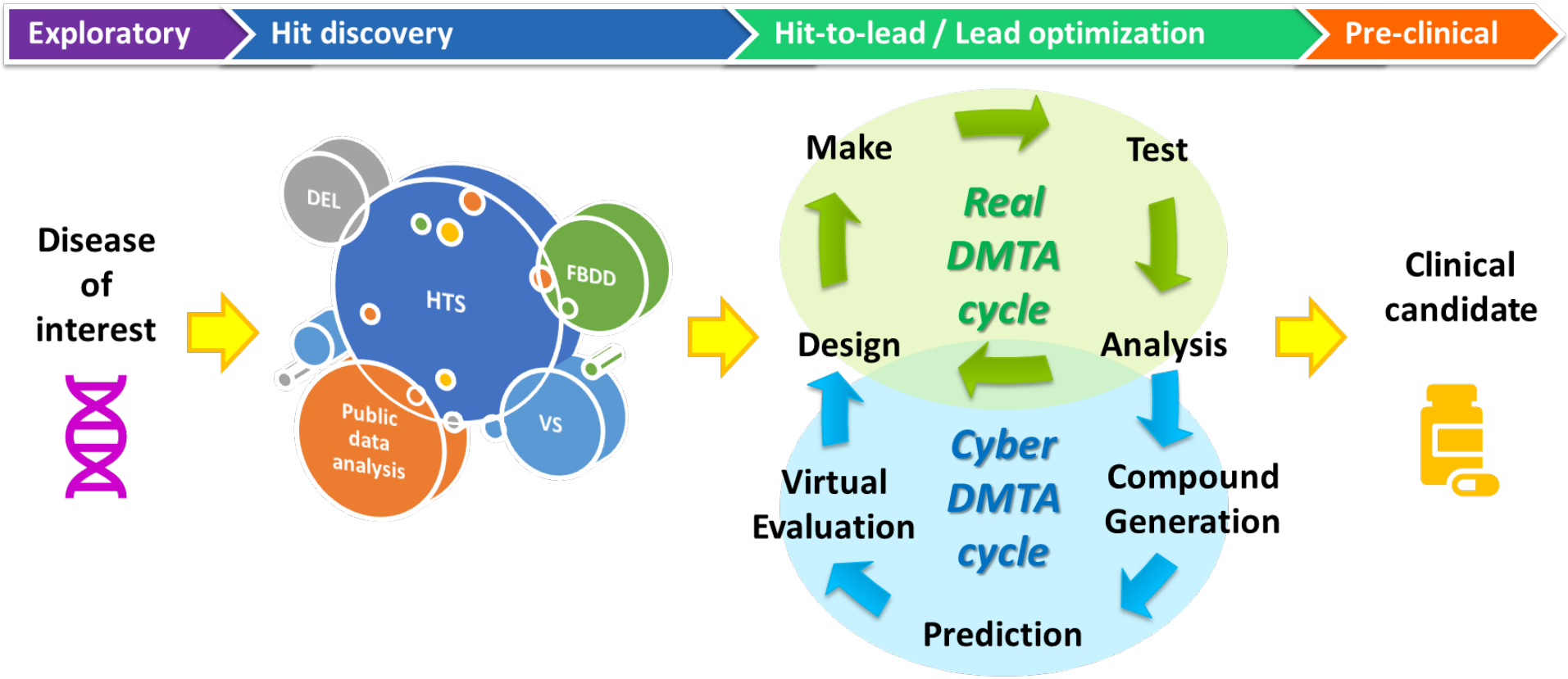


創薬化学（創薬バイオロジクス）研究は、DMTAサイクルが中心

Data-driven medicinal chemistry research workflow

Our Digital Transformation (DX):

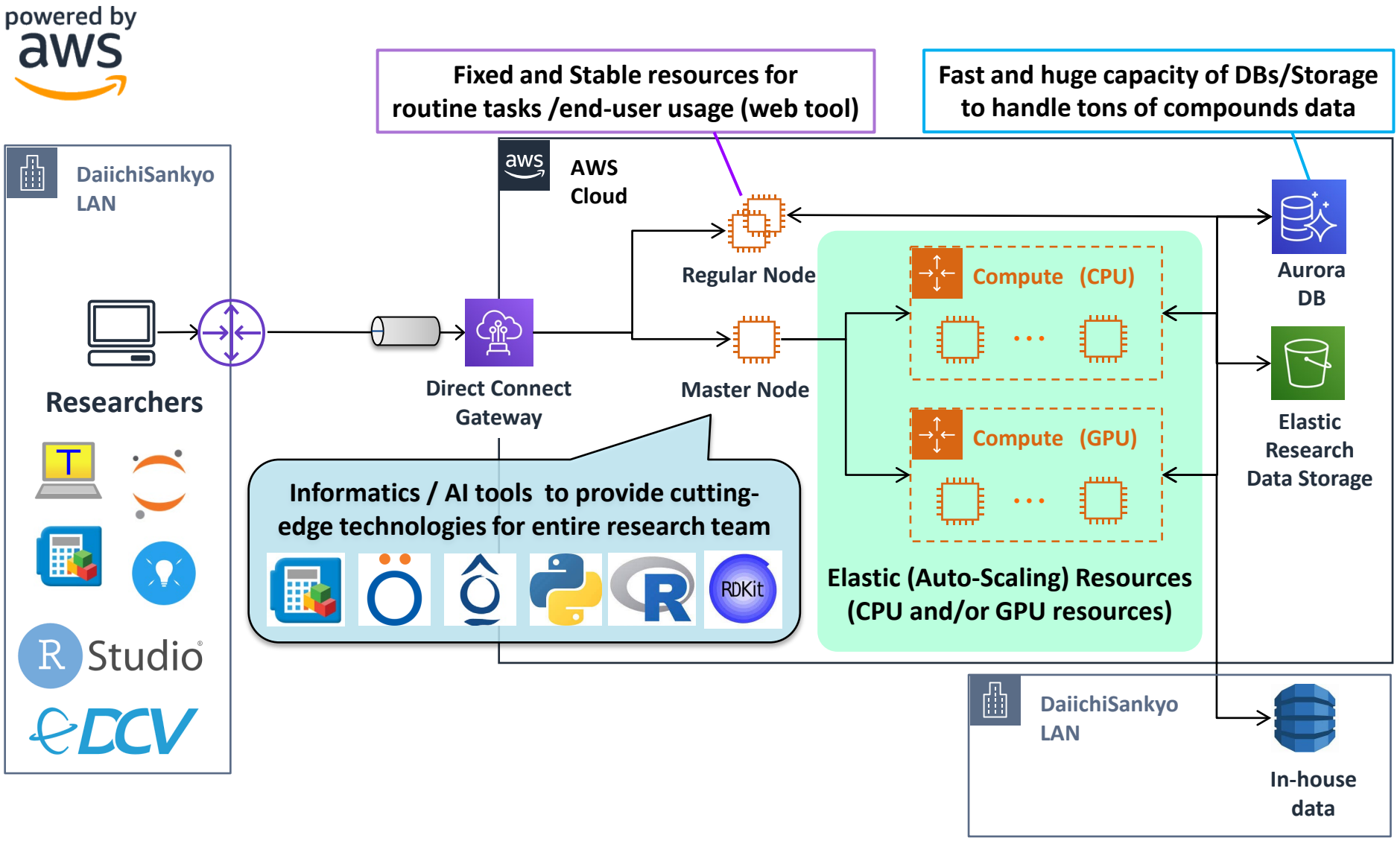
Provides a platform for medicinal chemists to make data-driven decisions



Hit discovery: Supporting the hit confirmation

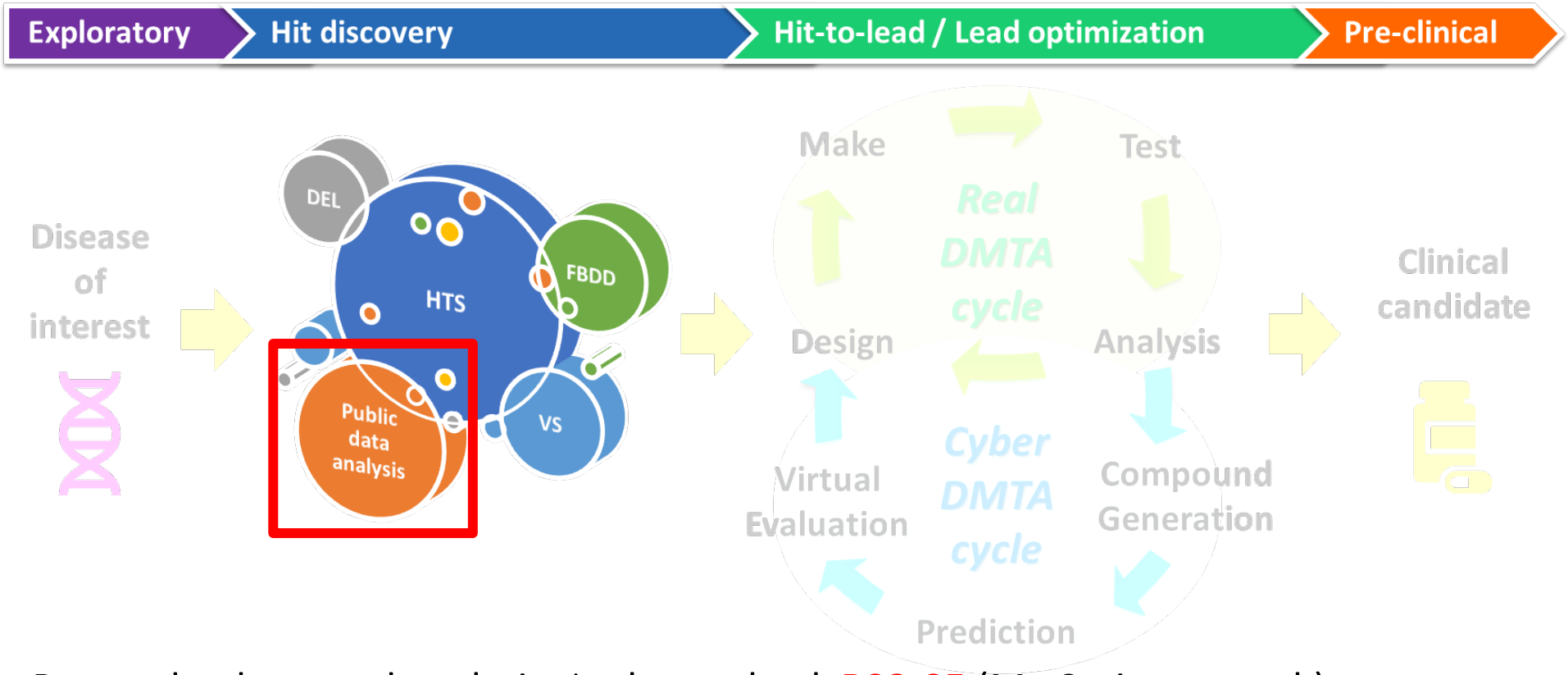
Hit-to-lead / Lead optimization: IP generation with data science

データ駆動型創薬を実施するためのクラウド環境



Public data analysis

Fast and accurate collection of existing public information is an important factor in drug discovery research



Patent database and analysis → please check **P03-05** (Mr. Serizawa et.al.)

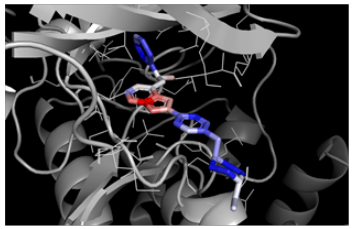
The database was built using Amazon Aurora

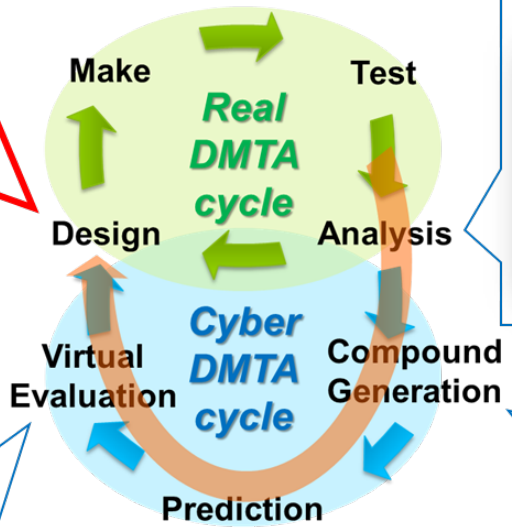
IP generation with data science

Data-driven decision making for the next compound to be made

Data-driven decision making with all involved scientists

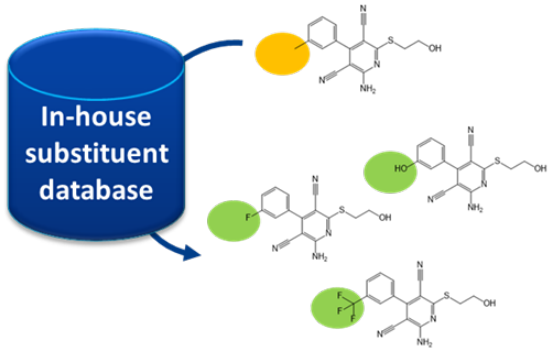
Structure based protein-ligand interaction analysis



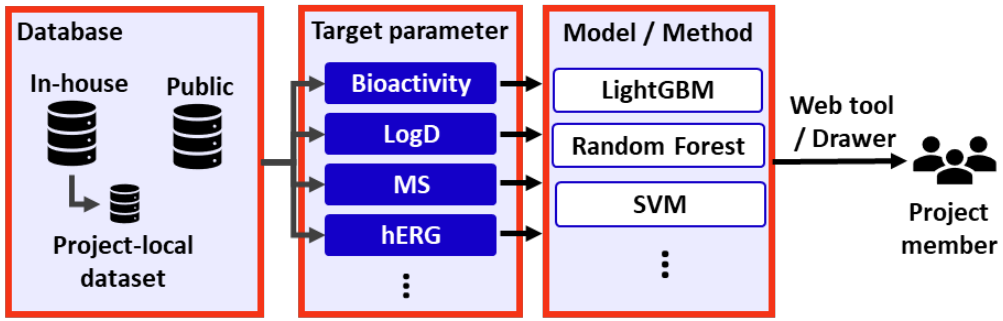


Current SAR visualization and analysis

Virtual compound generation by reagent-based substituent conversion



Machine learning based bioactivity and ADMET prediction



第一三共株式会社

SAR visualization and analysis

Data-driven decision making for the next compound to be made

Data-driven decision making with all involved scientists

DX team

CADD expert

MedChem

Structure based protein-ligand interaction analysis

Make

Test

Design

Analysis

Real DMTA cycle

Cyber DMTA cycle

Virtual Evaluation

Compound Generation

Prediction

Current SAR visualization and analysis

Virtual compound generation by reagent-based substituent conversion

In-house substituent database

Machine learning based bioactivity and ADMET prediction

Database

In-house

Public

Project-local dataset

Target parameter

Bioactivity

LogD

MS

hERG

Model / Method

LightGBM

Random Forest

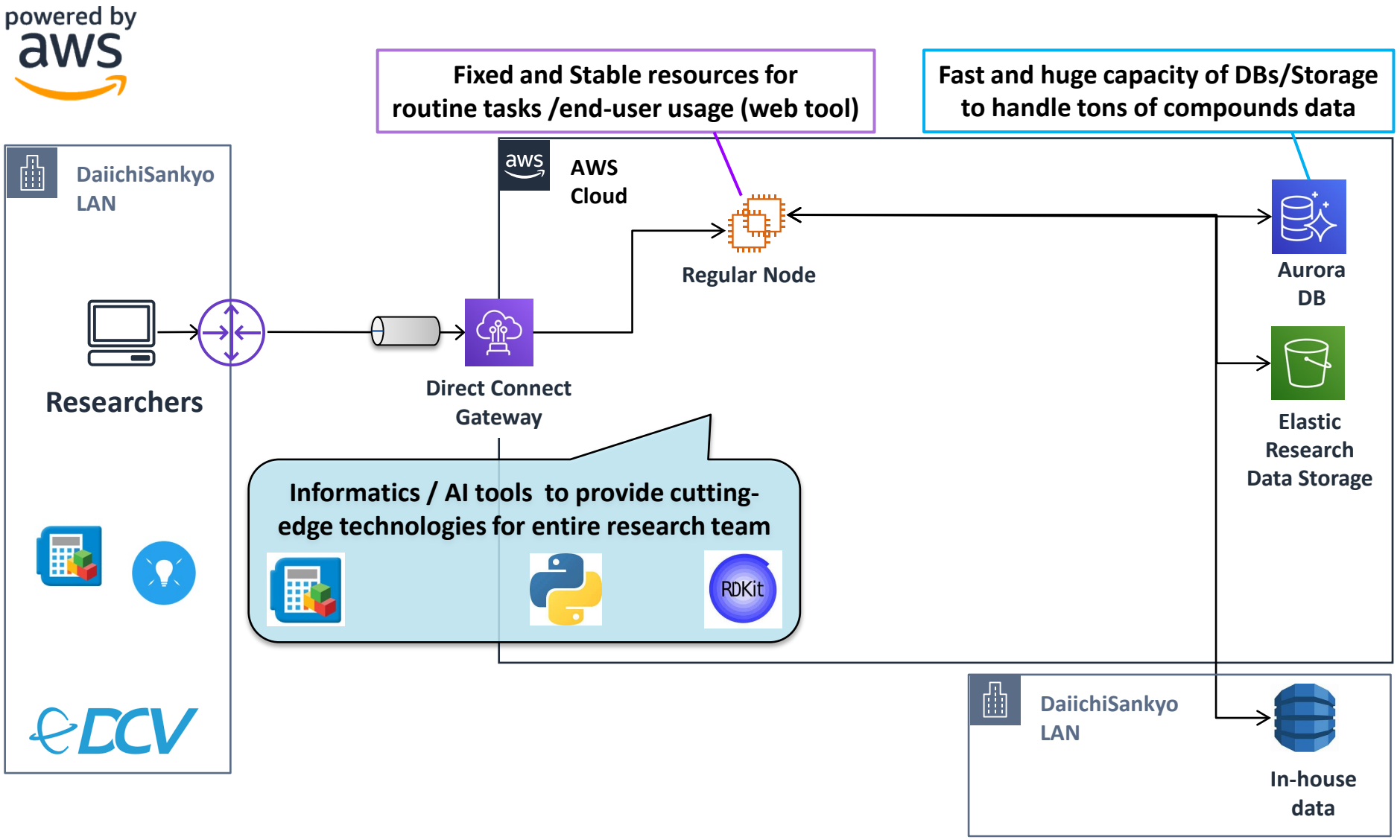
SVM

Web tool / Drawer

Project member

第一三共株式会社

Used resource on AWS



Data analysis scheme

Foster a culture of data collection

Discuss with Biologists

- What is this assay?
- What is raw data?
- Support for data registration

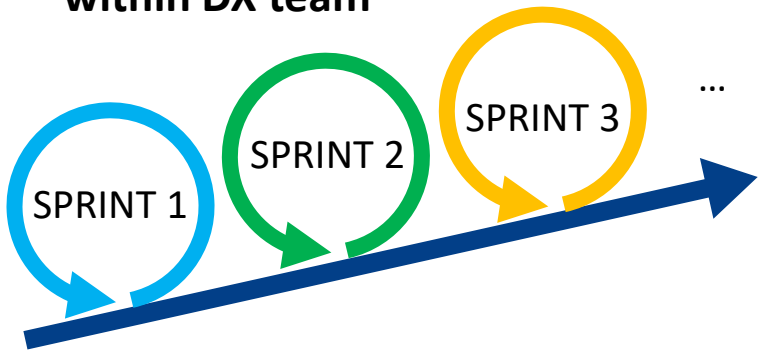
Discuss with IT team

- Data format in the in-house database
- Data security policy

Discuss with Biologists

- What is the assay lead-out?

1. **Identify project issues with PJ team**
2. **Analyze data & Provide tools within DX team**



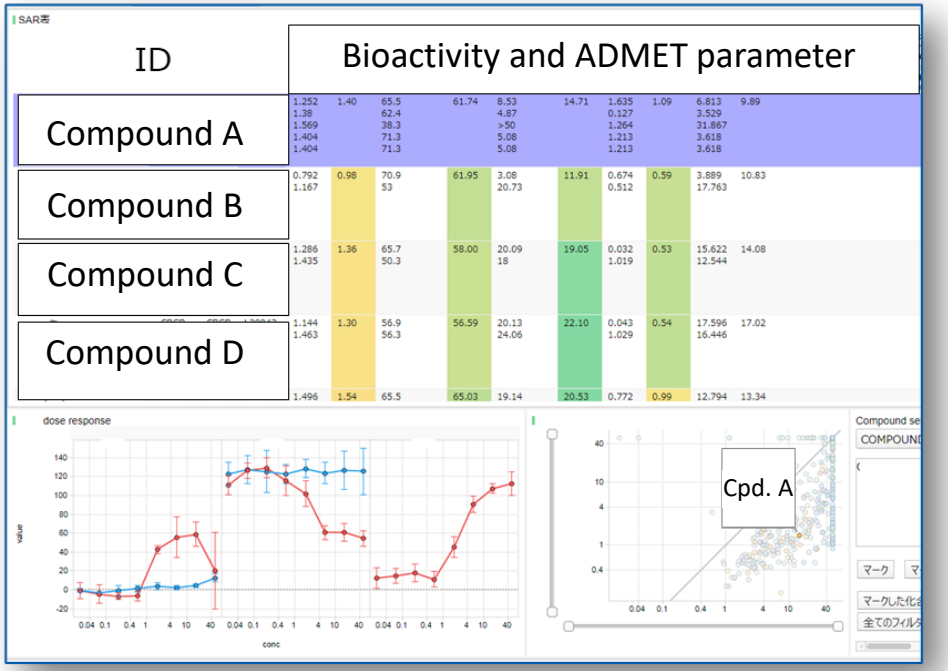
→ All assay data needed by Medicinal Chemists has been collected

* This DX team is mainly composed of medicinal chemists

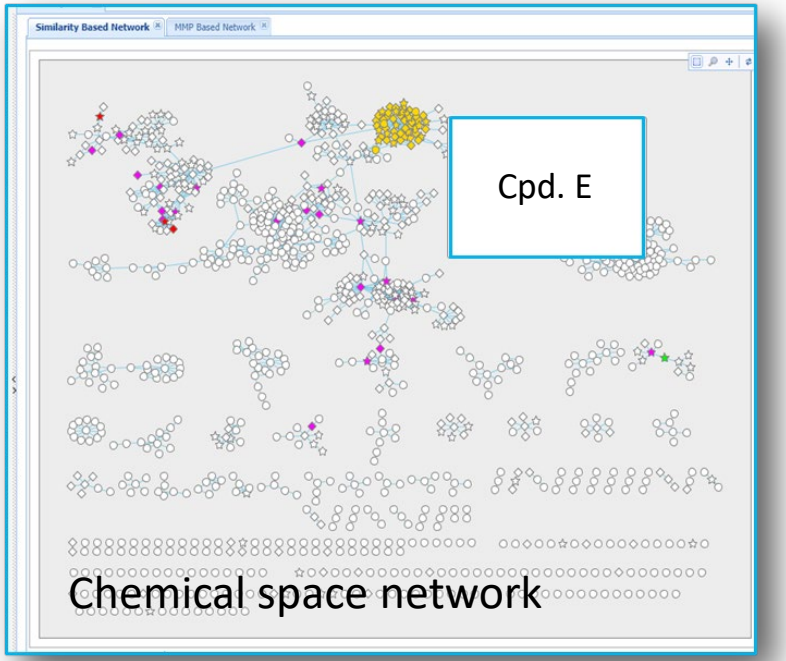
Data analysis and visualization 1

Foster a culture of data analysis

SAR table / plots



Network analysis



SAR table / plots: all properties in one table

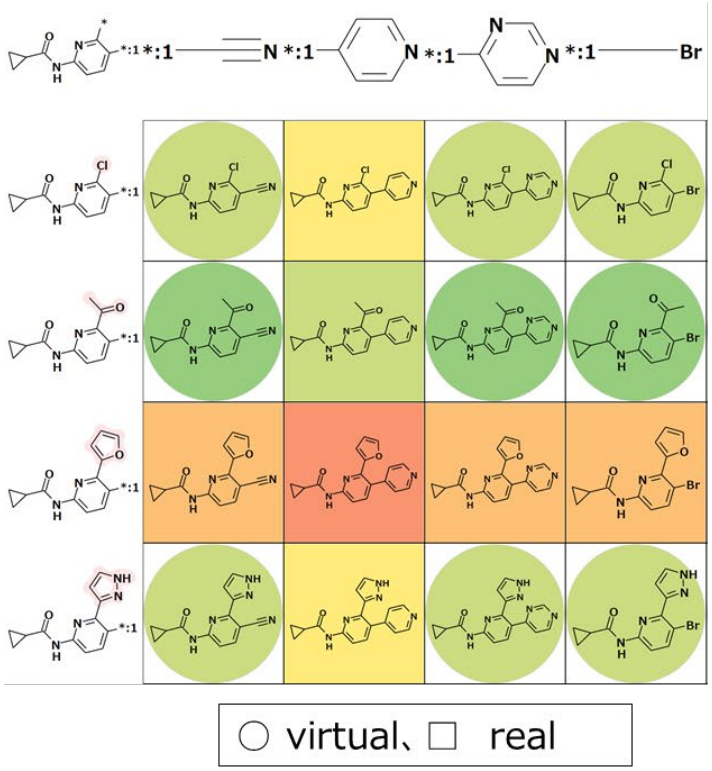
Network analysis: to understand the current status of synthetic deployment

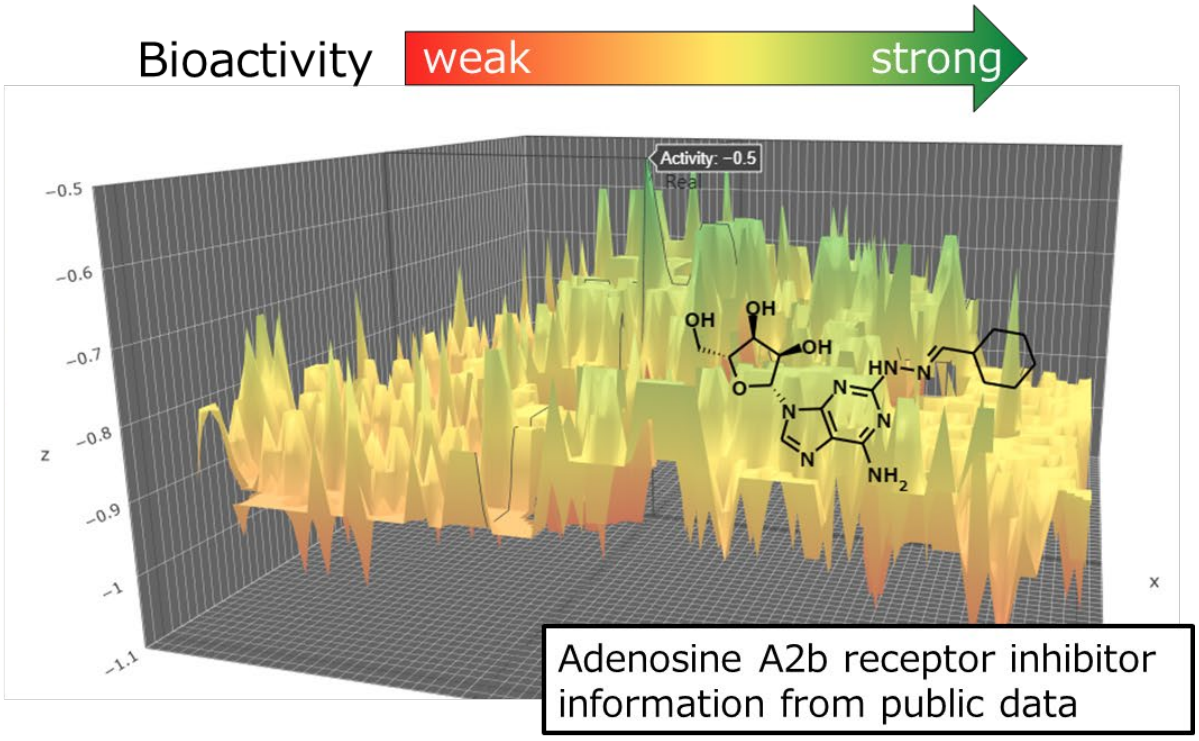
All project members share the same information and understand the issues

Data analysis and visualization 2

Foster a culture of data analysis

SAR Matrix





SAR Matrix: R-group table and virtual compound generation with fragment

Reference: A. Yoshimori et al., *ACS Omega* **2019**, *4*, 7061

Virtual compound generation

Data-driven decision making for the next compound to be made

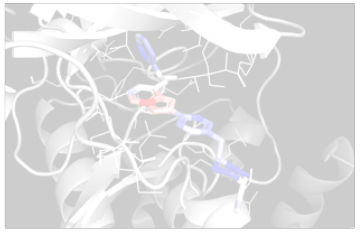
Data-driven decision making with all involved scientists

DX team

CADD expert

MedChem

Structure based protein-ligand interaction analysis



Make

Test

Design

Analysis

Real DMTA cycle

Cyber DMTA cycle

Virtual Evaluation

Compound Generation

Prediction

Current SAR visualization and analysis

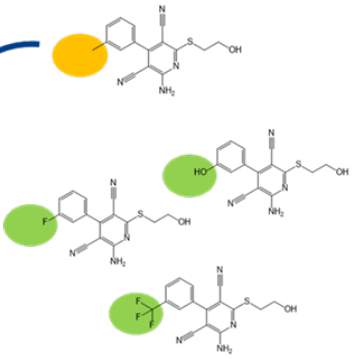
SAR table / plot

Network analysis

SARMM

Virtual compound generation by reagent-based substituent conversion

In-house substituent database



Machine learning based bioactivity and ADMET prediction

Database

In-house

Public

Project-local dataset

Target parameter

Bioactivity

LogD

MS

hERG

Model / Method

LightGBM

Random Forest

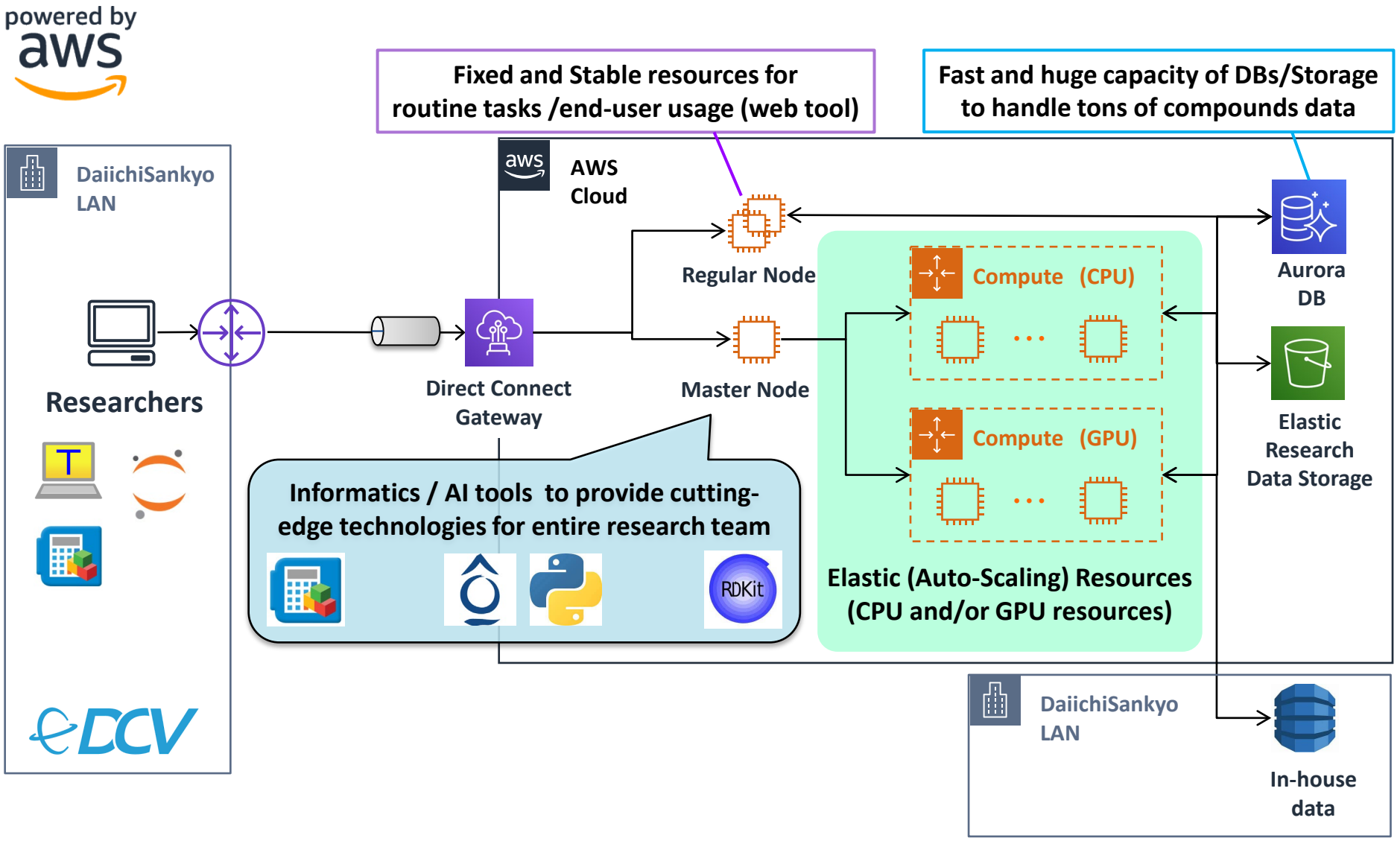
SVM

Web tool / Drawer

Project member

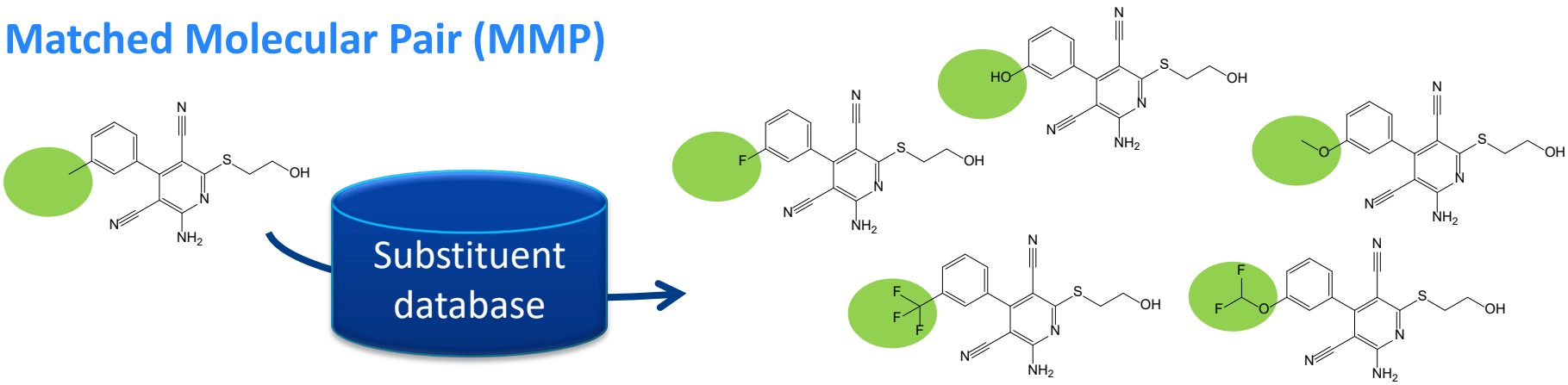
第一三共株式会社

Used resource on AWS



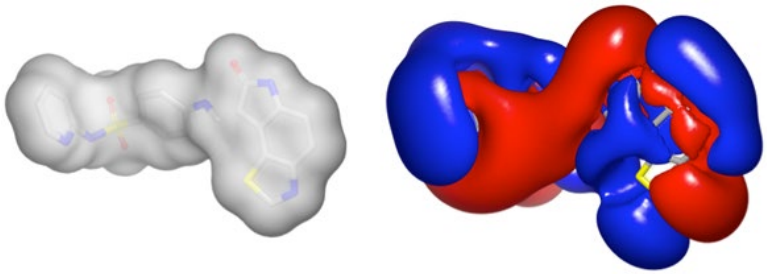
Conventional approaches

Matched Molecular Pair (MMP)

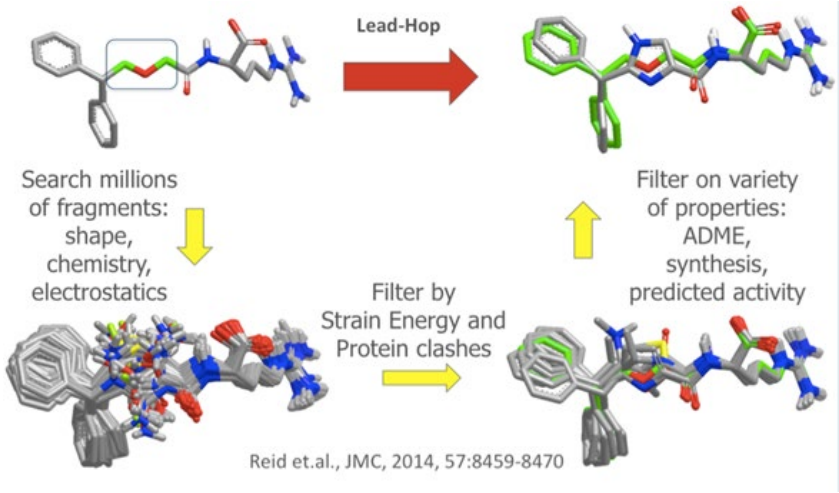


Chemical transformation → please check **P08-02** (Dr. Takeuchi et.al.)

Brood (Openeye)



Shape Electrostatic



Shape and electrostatic similarity searching with fragment replacement in 3D

https://www.eyesopen.com/news/2015/08/brood-v3.0

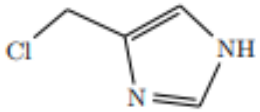
Recurrent Neural Network (RNN)

REINVENT (ref: arXiv:1704.07555v2)

化合物を文字列（Smiles）として学習し、一文字ずつ次に来る文字を予測していく

妥当な文字列を生成するモデル（Agent network）により、Smiles群を生成し、
それらの評価値を算出、損失関数を最小化するようにモデルを更新する（強化学習）

Graph:

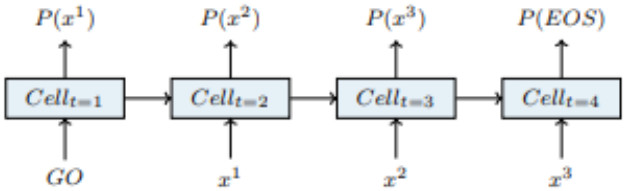


SMILES: ClCc1c[nH]cn1

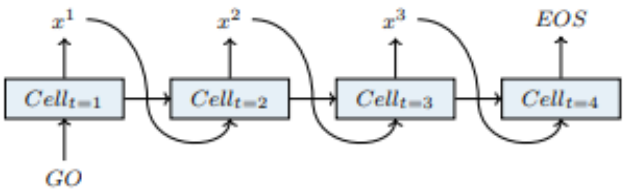
One-hot encoding:

	Cl	C	c	1	c	nH	c	n	1
C	0	1	0	0	0	0	0	0	0
c	0	0	1	0	1	0	1	0	0
n	0	0	0	0	0	0	0	1	0
1	0	0	0	1	0	0	0	0	1
nH	0	0	0	0	0	1	0	0	0
Cl	1	0	0	0	0	0	0	0	0

Learning the data



Generating sequences



Bioactivity and ADMET Prediction

Data-driven decision making for the next compound to be made

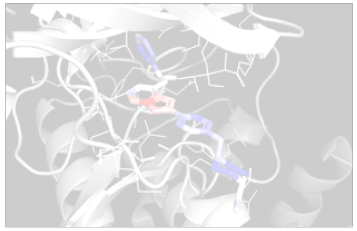
Data-driven decision making with all involved scientists

DX team

CADD expert

MedChem

Structure based protein-ligand interaction analysis



Make

Test

Real DMTA cycle

Design

Analysis

Cyber DMTA cycle

Virtual Evaluation

Compound Generation

Prediction

Current SAR visualization and analysis

SAR table / plot

Network analysis

SARMM

Virtual compound generation by reagent-based substituent conversion

In-house substituent database

Machine learning based bioactivity and ADMET prediction

Database

In-house

Public

Project-local dataset

Target parameter

Bioactivity

LogD

MS

hERG

Model / Method

LightGBM

Random Forest

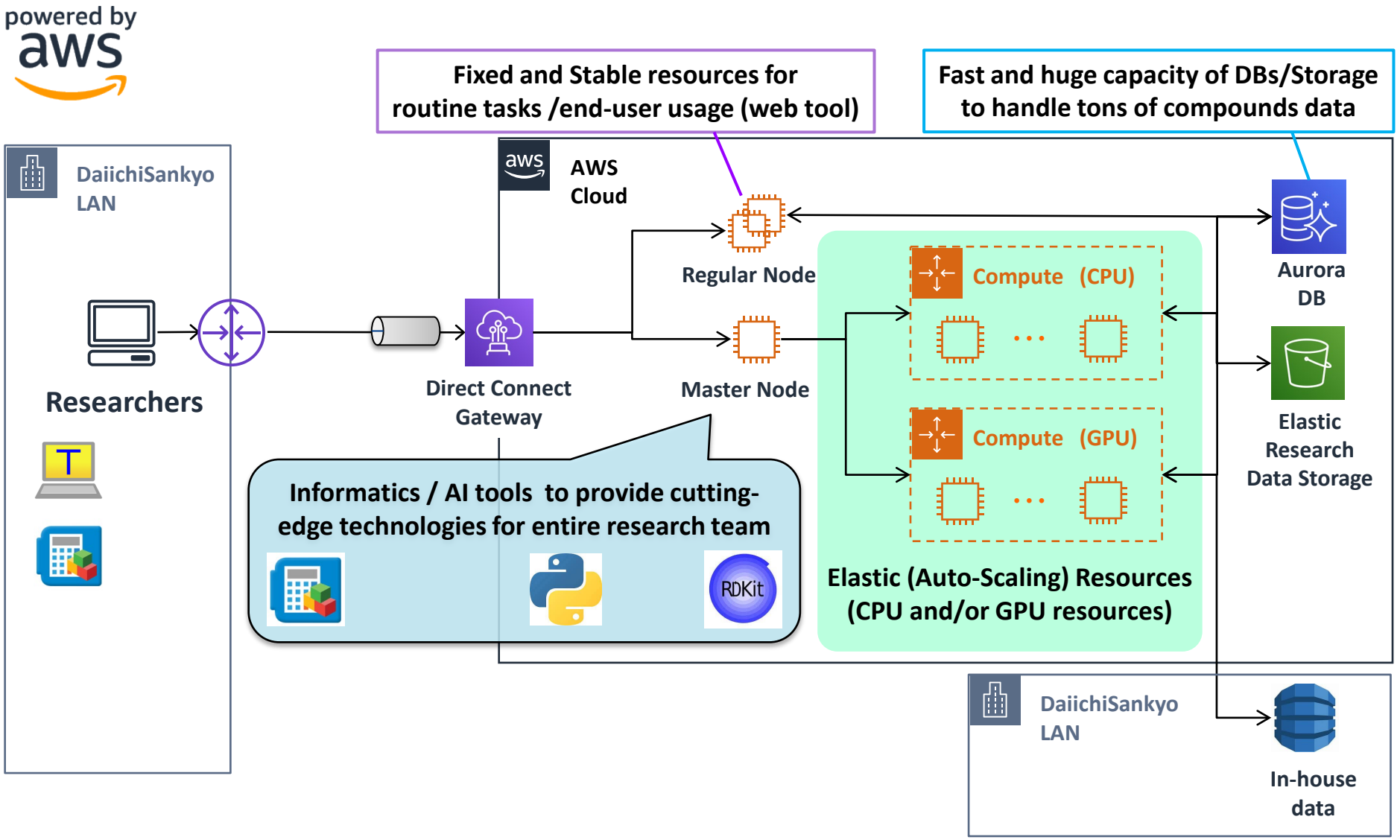
SVM

Web tool / Drawer

Project member

第一三共株式会社

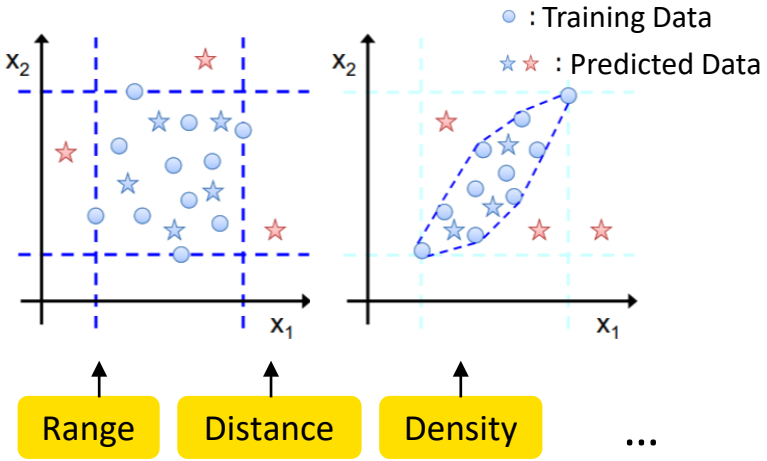
AWS resource



Parameter prediction with machine learning

Exploration of Applicability domain in AI / ML utilization

What is Applicability Domain (AD) ?



Parameters

- ✓ #CPD
- ✓ MW
- ✓ #Chemotype
- ✓ Property

... etc

The range of data for which the predictive model is expected to perform adequately

Goal for the collaboration

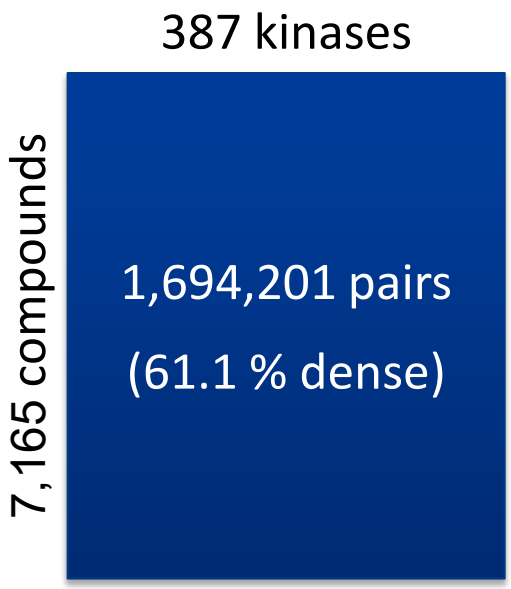
- ✓ Implementing various AI / ML methods in DS research site
- ✓ Development of the computing environment
- ✓ Establishment of AD analysis approaches
- ✓ **Evaluation of ADs by datasets and methods**

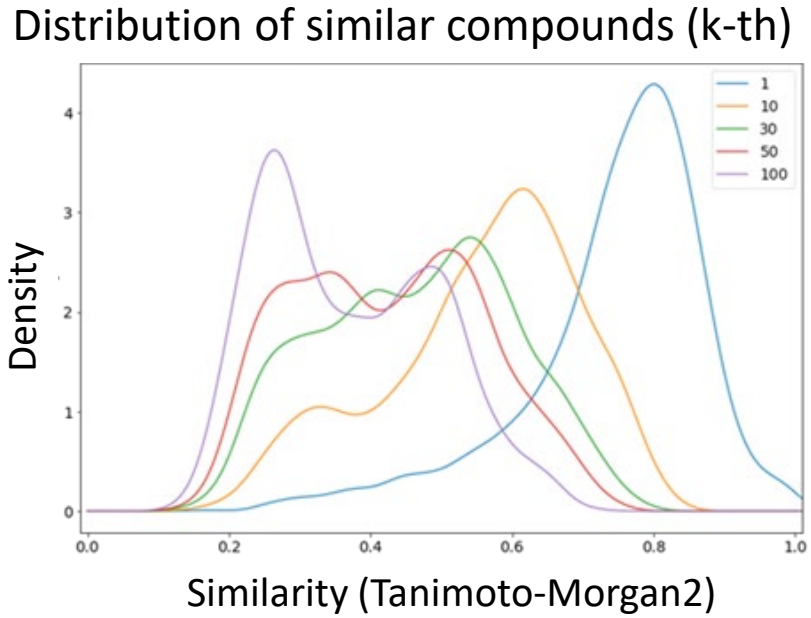
AI/ML prediction is not always a practical technology
-> Explicit knowledge of which situations AI/ML predictions are effective

Feasibility study with in-house data (Dataset)

Validation study of the performance of each prediction method (deep / non-deep) using in-house research data

- In-house kinase assay data (Active-site directed competition binding assays)





- Compound concentration: 10 µM & ATP concentration: 1 mM
- Hit definition: Compounds with residual activity rate ≤ 20 %
- Total hit rate: 17.7% Average hit rate: 18.11% (SD 2.98 %)

- In-house ADMET assay data → please check **P03-01** (Dr. Watanabe et.al.)

Feasibility study with in-house data (Methods)

For the purpose of performance validation, the well-known methods were used for comparison with general parameter sets

	Methoed
Graph layer (GNN)	GCNConv
	GraphConv
	SAGEConv
	GATConv
	ARMAConv
	SGConv
Conventional ML	LightGBM
	Random Forest
	Support Vector Classification
	k-Nearest Neighbor

Conditions

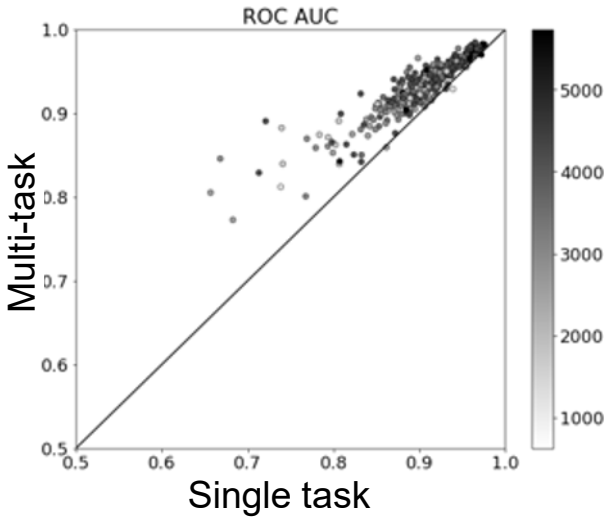
- Classification model (positive / negative)
- Random split
- Train : Valid : Test = 8 : 1 : 1 (5 trials)
- 72-dimensional atom feature for GNN
- Morgan 2 fingerprint for conventional ML
- Max of 100 epochs of training in each trial

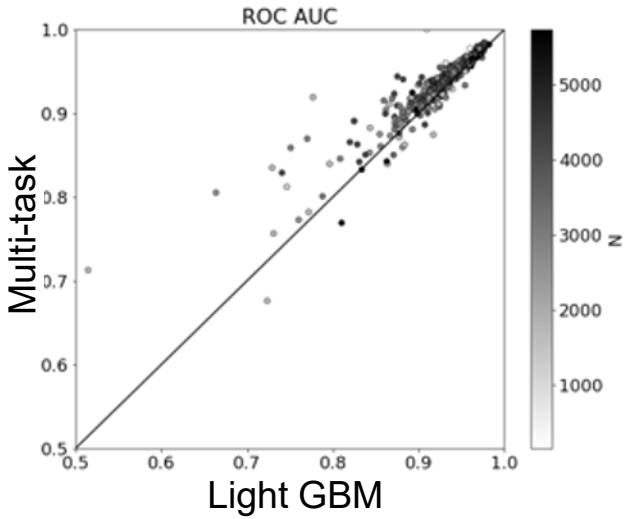
Feasibility study with in-house data (Result)

Prediction ROC-AUC score of each model

	Single task		Multi-task		Transfer learning	
	Validation	Test	Validation	Test	Validation	Test
ARMAConv	0.9514±0.0008	0.9004±0.0051	0.9389±0.0022	0.9295±0.0041	0.9440±0.0016	0.9327±0.0036
LightGBM	0.9521±0.0007	0.9190±0.0054	-	-	-	-

Comparison of per-target prediction ROC-AUC between models



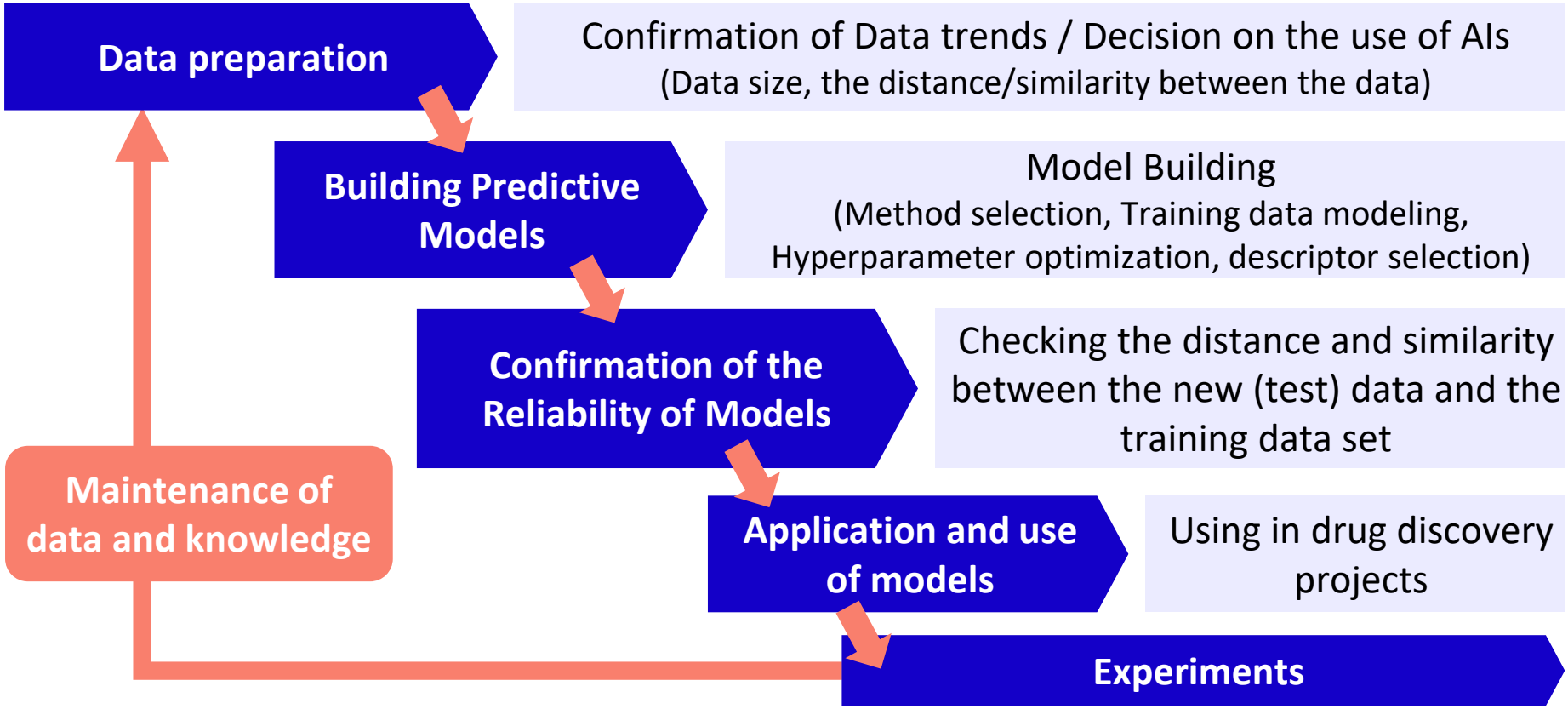


- **Conventional machine learning was more accurate than single task deep learning (DL)**
- **Accuracy was improved by multi-tasking / transfer learning, which are features of DL**

Feasibility study with in-house data (Conclusion)

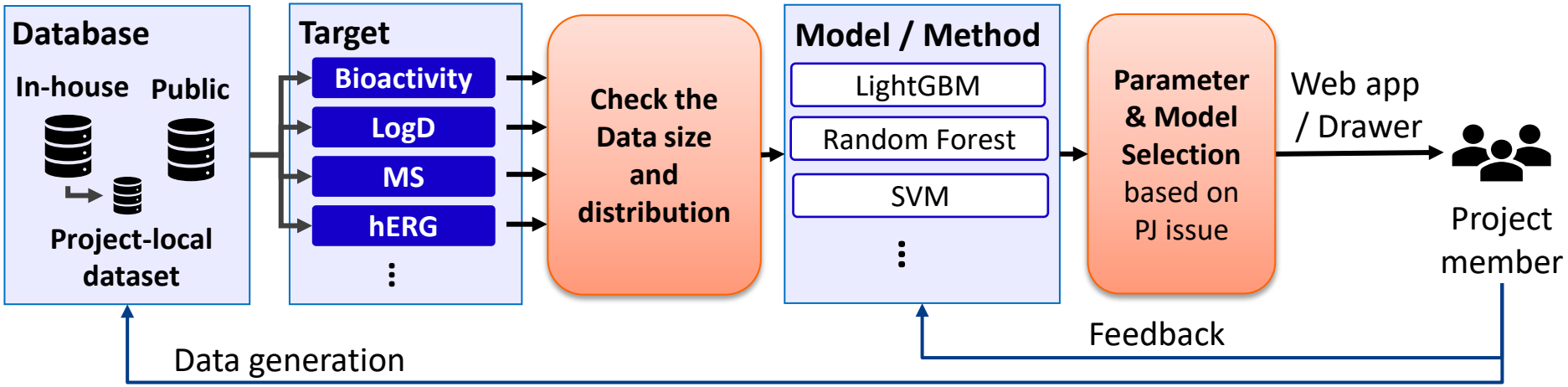
- In this data set, the prediction accuracy of the conventional method outperformed that of the GNN in the single task (**no big difference**)
- Multitasking and transition learning contribute to improved accuracy

→ In-house predictive models operate primarily with LightGBM



In-house auto ML workflow

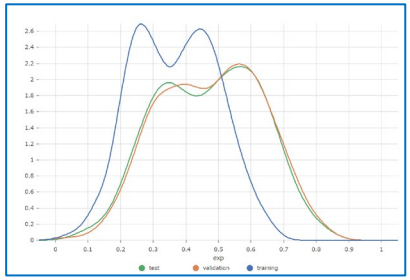
Life cycle of prediction model construction and validation



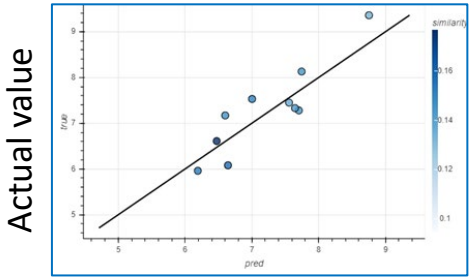
GUI tools to use the prediction models

- Web app for checking the data and model

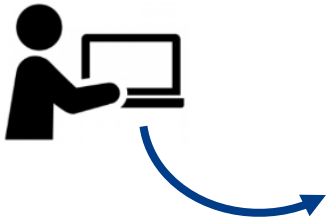
Data distribution



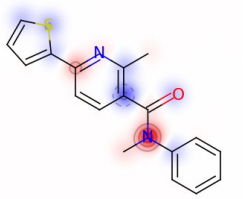
Prediction performance



- Predict new compounds



Web app or Drawer tool



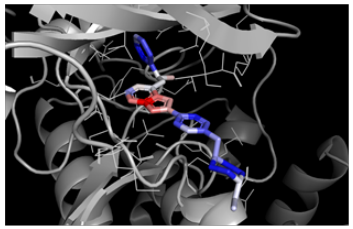
hERG: Negative
logD: XXXX
Solubility: YYYY
…

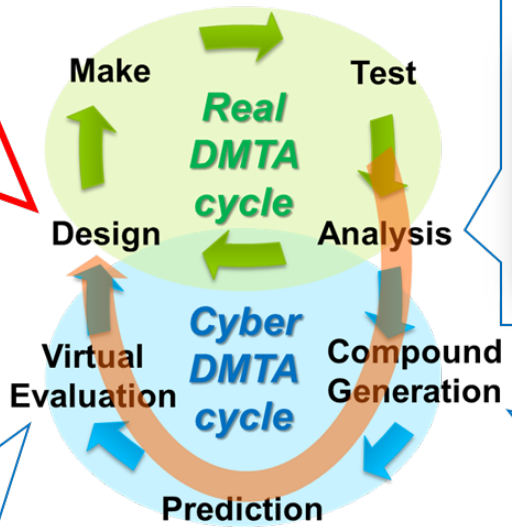
IP generation with data science

Data-driven decision making for the next compound to be made

Data-driven decision making with all involved scientists

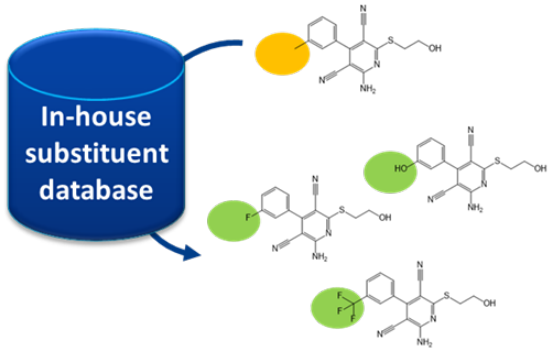
Structure based protein-ligand interaction analysis



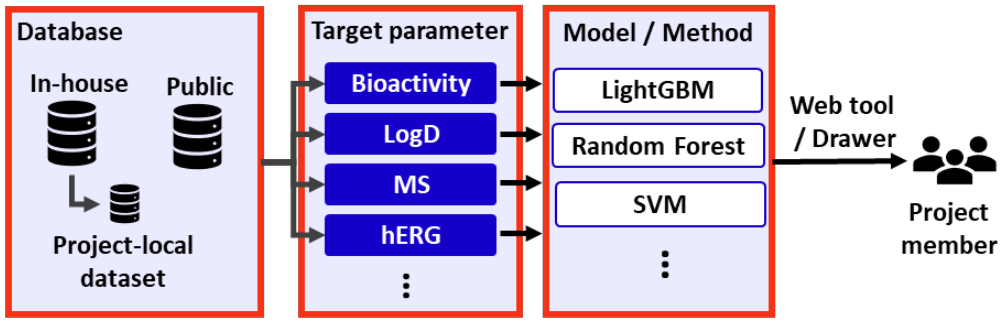


Current SAR visualization and analysis

Virtual compound generation by reagent-based substituent conversion



Machine learning based bioactivity and ADMET prediction

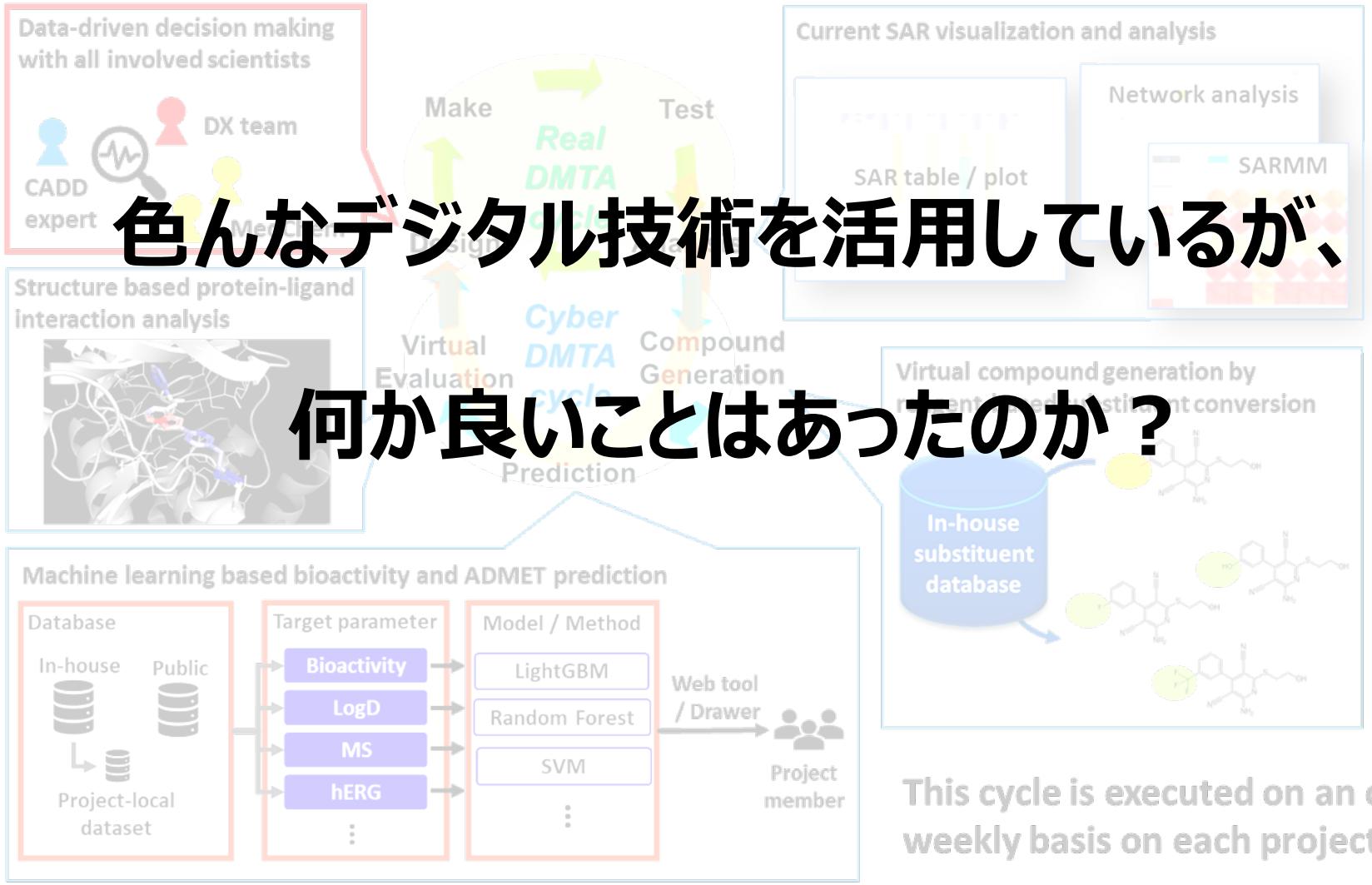


This cycle is executed on an ongoing weekly basis on each project

第一三共株式会社

Daiichi-Sankyo

Data-driven decision making for the next compound to be made



色んなデジタル技術を活用しているが、

何か良いことはあったのか？

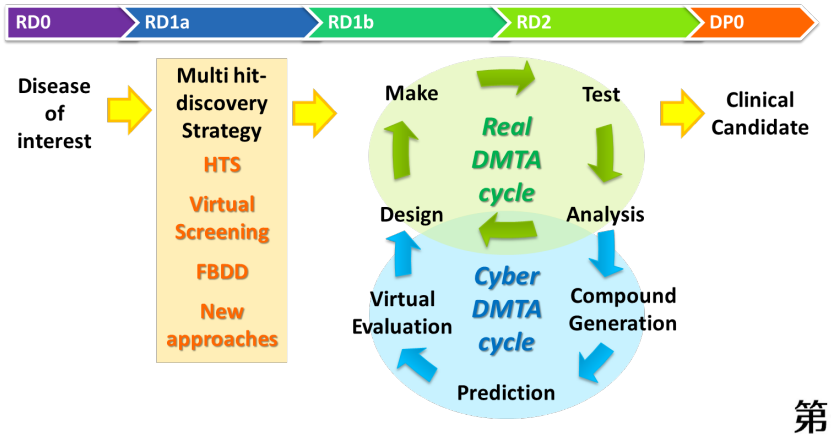
This cycle is executed on an ongoing weekly basis on each project

Evaluations from Medicinal chemists (2020.04-2021.09)

Survey on results and satisfaction with activities with DX

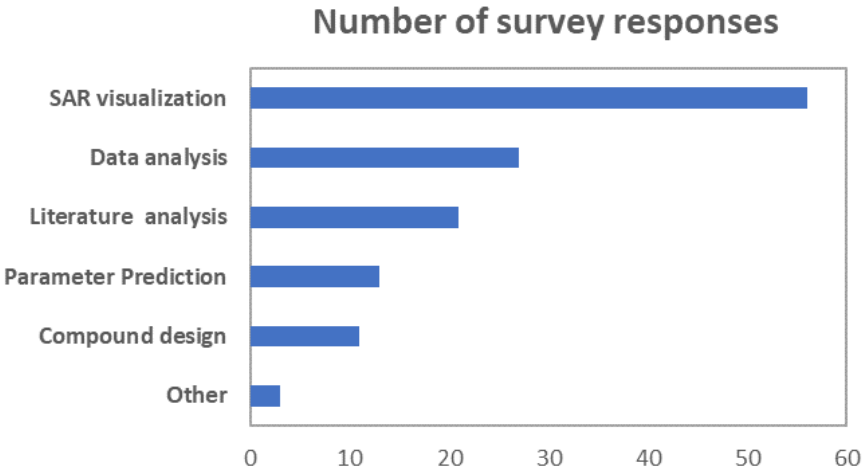
タイトル	テーマ名・プロジェクト名	グループ	時期	タスクリーダー	タスク内貢献者	D4 チーム貢献者	実施内容	カテゴリ	効果理由	効果	コスト	Before after D4 量的な効果	Before after D4 質的な効果	推進Gの体制に関するご意見	感想、他への適用可能性等

As a result of this 1.5 years, we've received more than 100 reports, thank you!



- Collecting the results of activities under each theme
- Visualization of effects and outcomes
- Gathering opinions from MedChem

The research benefit of DX for Medicinal Chemistry



Analyze the performance from surveys

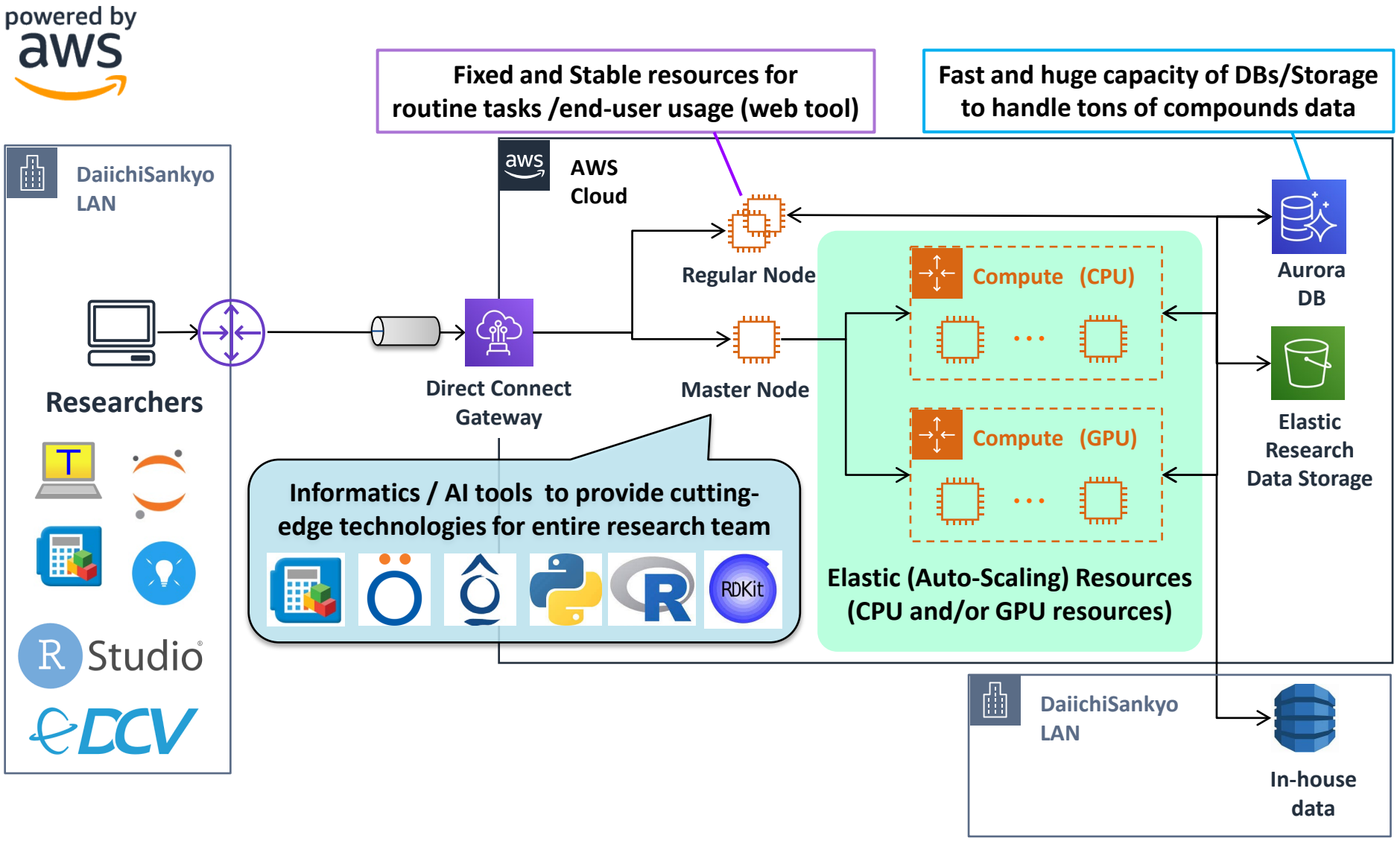
- ✓ SAR visualization and analysis most favorable
- ✓ Literature search and patent analysis are in high demand
- ✓ Prediction and structure generation need further study in the future

Compound design	#Reaction	Time reduction	# IP generation
SAR visualization	56	95%	-
Data analysis	27	65%	-
Literature analysis	21	90%	-
Parameter Prediction	13	-	-
Compound design	11	-	7 projects
Other	3	-	-

Contribution to IP generation is just the beginning

The overall efficiency of the Medicinal Chemistry Lab was increased by approximately 20%, considering the cost of the data scientist's work

データ駆動型創薬を実施するためのクラウド環境



Summary

- A platform / workflow for **conducting data-driven medicinal chemistry research** has been built by **DX team composed of Medicinal Chemists**
- Cyber DMTA parts are designed to support medicinal chemists
- **Visualized the impact of research DX**, for which outcomes are difficult to measure
  - ✓ Improved efficiency of data analysis operations by approximately 20% throughout Medicinal Chemistry Research Laboratory
  - ✓ Contributions to drug design are being made, but impact on IP generation has not been achieved yet

Advantages of AWS

- Excellent service and support
- Extensive examples of implementation based on experience

Acknowledgement

Daiichi Sankyo Co., LTD.

- Mr. Takayuki Serizawa
- Dr. Tsutomu Nagata
- Dr. Kosuke Takeuchi
- Dr. Toshiaki Watanabe
- Dr. Kazumasa Aoki

And all researchers of the Medicinal Chemistry Research Laboratory at DS

University of Bonn

- Prof. Dr. Jürgen Bajorath

ExaWizards Inc.

- Dr. Hirotomo Moriwaki
- Mr. Shin Saito
- Mr. Tomoya Matsumoto
- Dr. Kentaro Rikimaru
- Mr. Koji Hazama

AWS

- Mr. Takehiro Nakajima
- Dr. Daisuke Miyamoto

References

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) Moriwaki H, Saito S, Matsumoto T, Serizawa T, Kunimoto R. Global Analysis of Deep Learning Prediction Using Large-Scale In-House Kinome-Wide Profiling Data. *ACS Omega* **7**, 18374-18381 (2022)