創薬DX Platformのための HPCとワークフロー

2024年8月6日

千葉峻太朗

R-CCS HPC/AI 駆動型医薬プラットフォーム部門

分子デザイン計算知能ユニット

自己紹介

計算の 計算による 計算のための科学

理化学研究所
計算科学研究センター（R-CCS）

理研の12研究センターの一つであると同時に
高性能計算科学における国家「知の拠点」センター

理研の他の研究センター
および国内外の大学・
研究機関との連携

R-CCSが橋頭保的な役割

計算のための科学

さまざまな科学分野との連携により
高性能計算を進化させる

光・量子・再構成可能コンピューティング・ニューロモルフィックコンピューティングなど、新しいコンピューティングの概念を支える材料やデバイスの開発

新しいコンピュータ・
アーキテクチャや
計算モデル

新しいデバイスの
ためのアルゴリズムや
プログラミングモデル

新しいコンピューティング
技術のための解析や
シミュレーション

新しいコンピューティング
技術による計算の高度化

相乗効果と融合

計算の科学

高性能計算の本質となる
コンピューティング技術の研究

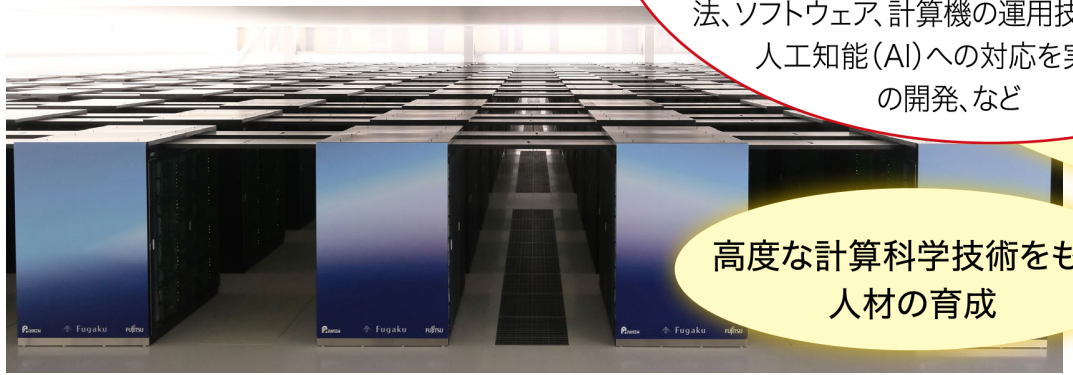
ポストムーア時代を見据えた新たなコンピューティング技術、アーキテクチャ、アルゴリズム等の開発やプログラミング手法、ソフトウェア、計算機の運用技術、ビッグデータ・人工知能（AI）への対応を実現する手法の開発、など

計算による科学

高性能計算を活用し
科学・社会の課題解決を目指す研究

高信頼性かつ高精度な解析・シミュレーションを用いた、生命科学、工学、気象・気候、防災・減災、物質科学、宇宙・素粒子物理学や社会科学などの研究、来るべきSociety5.0社会に向けた機械学習の応用開発など

スーパーコンピュータ「富岳」



高度な計算科学技術をもつ
人材の育成

産業界との連携

計算の科学


プログラミング環境
研究チーム
佐藤 三久


プロセッサ研究チーム
佐野 健太郎


高性能人工知能
システム
研究チーム
松岡 聡


大規模並列
数値計算技術
研究チーム
今村 俊幸


高性能ビッグデータ
研究チーム
佐藤 賢斗


次世代高性能
アーキテクチャ
研究チーム
近藤 正章

計算による科学


離散事象シミュレーション
研究チーム
伊藤 伸泰


量子系分子科学
研究チーム
中嶋 隆人


量子系物質科学
研究チーム
柚木 清司


粒子系生物物理
研究チーム
杉田 有治


粒子系シミュレータ
研究チーム
牧野 淳一郎


複合系気候科学
研究チーム
富田 浩文


複雑現象統一的解法
研究チーム
坪倉 誠


連続系場の理論
研究チーム
青木 保道


総合防災・減災
研究チーム
大石 哲


データ同化
研究チーム
三好 建正


計算構造生物学
研究チーム
Florence TAMA

2021年4月1日設置

HPC/AI駆動型医薬プラットフォーム部門


バイオメディカル
計算知能ユニット
奥野 恭史


創薬化学AIアプリケーション
ユニット
本間 光貴


分子デザイン計算
知能ユニット
池口 満徳


AI創薬連携基盤
ユニット
奥野 恭史

運用技術部門


施設運転技術
ユニット
塚本 俊之


システム運転技術
ユニット
宇野 篤也


チューニング技術
ユニット
庄司　文由


利用環境技術
ユニット
庄司 文由

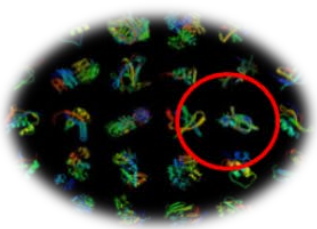
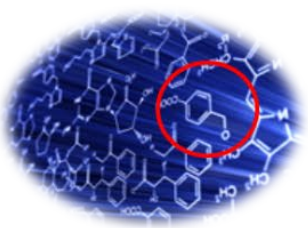
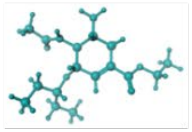
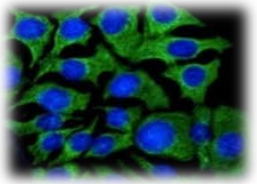

先端運用技術
ユニット
山本　啓二

背景：創薬の常識を覆したCOVID-19と欧米の底力

創薬の常識

医薬品開発の成功確率：2.5万分の1以下

（開発費用1200億円、開発期間約10年以上）


ターゲット探索 → リード探索 → リード最適化 → バイオアッセイ → 前臨床試験 → 臨床試験 → 承認 → 薬物治療

1年以内

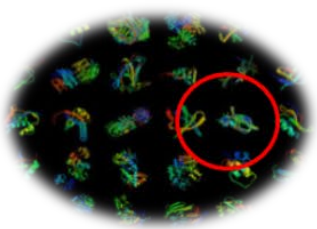
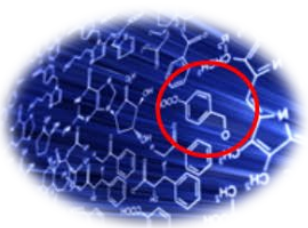
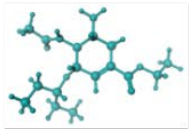
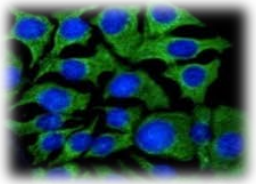
レムデシビル（米国）　　ワクチン（米国ファイザー）

デキサメタゾン（英国）　　ワクチン（米国モデルナ）

バリシチニブ（米国）　　ワクチン（英国アストラゼネカ）

AIやシミュレーションで少人数・低コストで創薬の超効率化を目指す

創薬の常識
医薬品開発の成功確率：2.5万分の1以下
（開発費用1200億円、開発期間約10年以上）

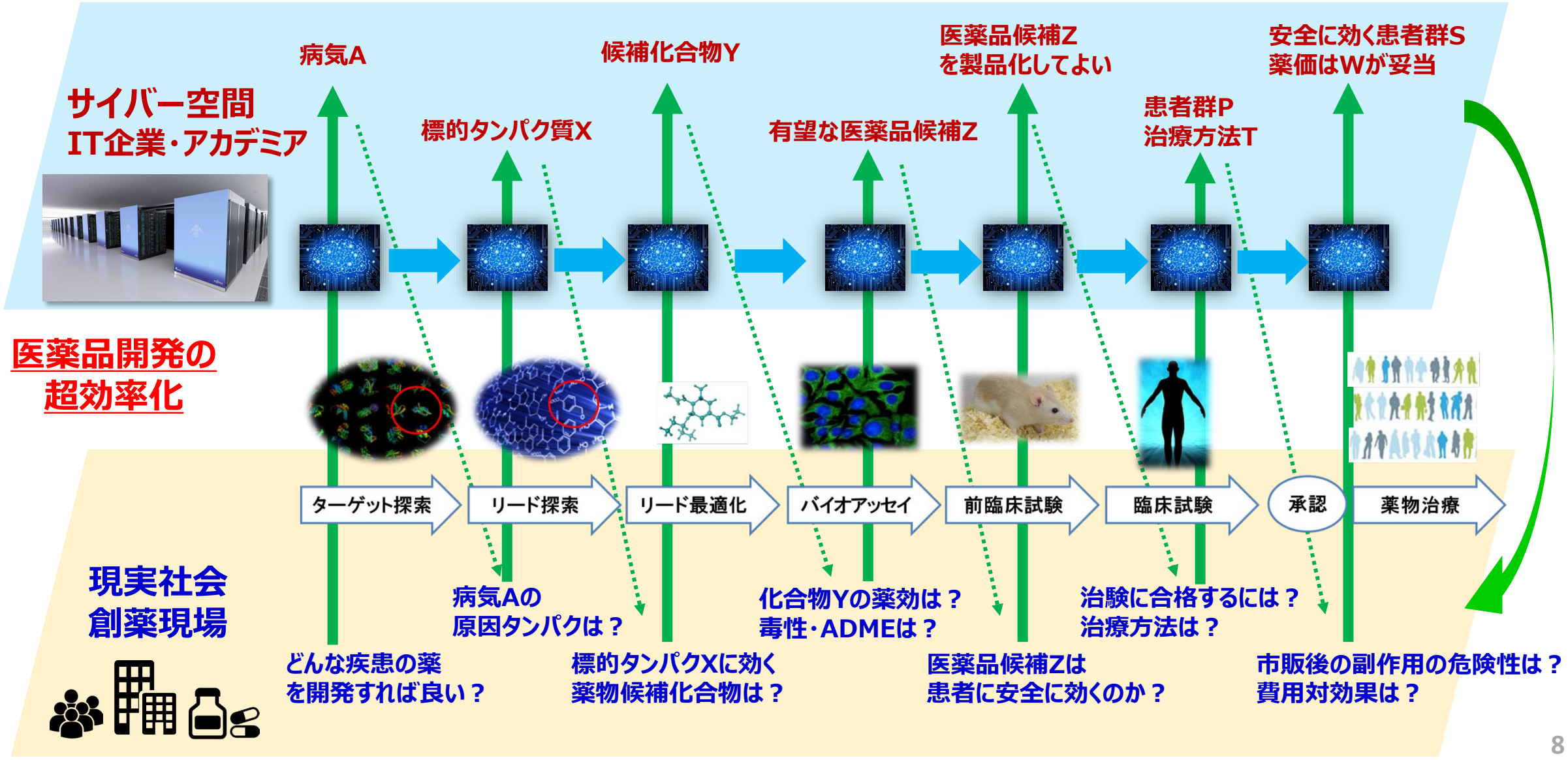

ターゲット探索 → リード探索 → リード最適化 → バイオアッセイ → 前臨床試験 → 臨床試験 → 承認 → 薬物治療




スパコンやAIによって加速する

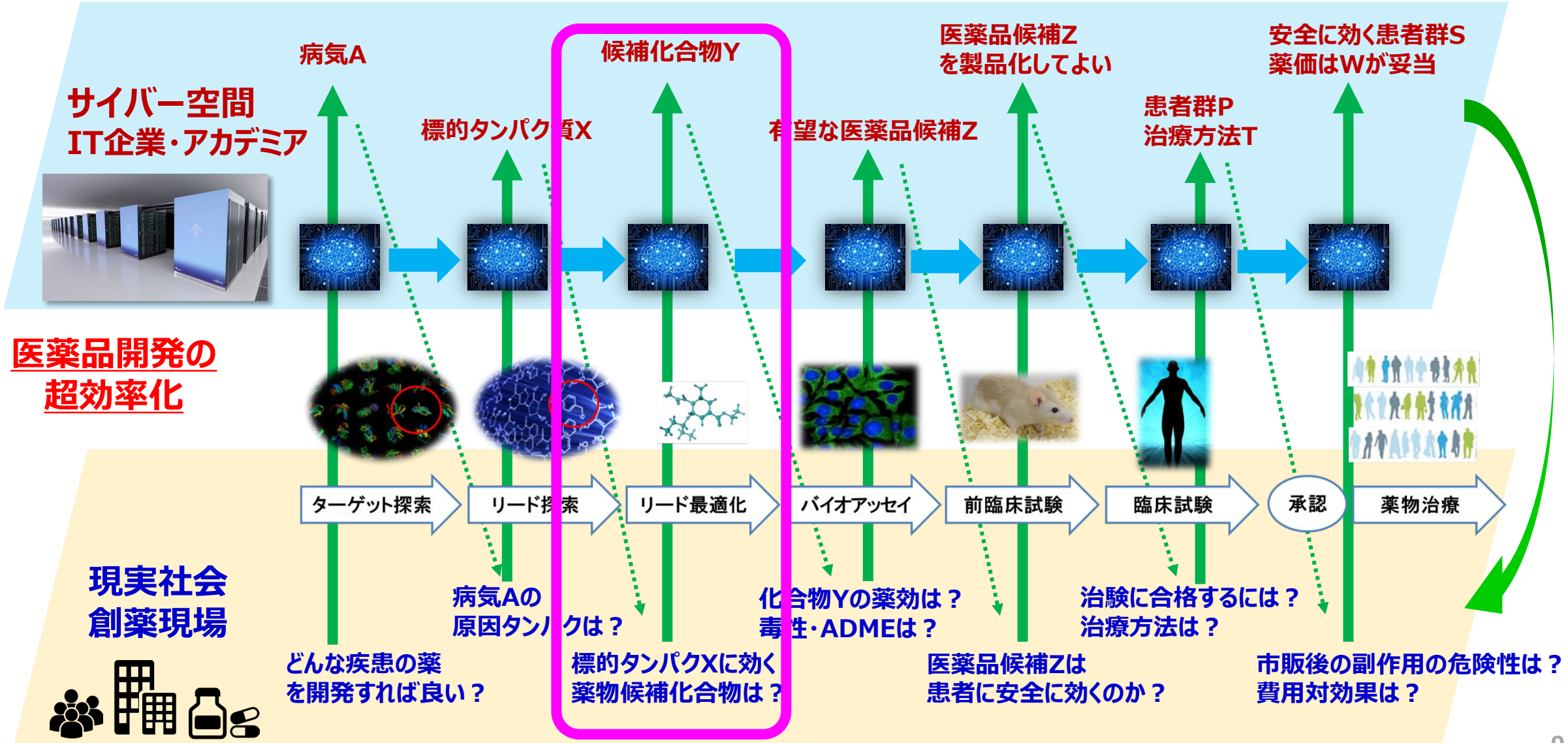
「富岳」創薬DXプラットフォーム

「富岳」を基軸として、AIとシミュレーションを融合させた革新的な創薬DXプラットフォームを構築し、創薬プロセスの超効率化を目指す。これにより、新薬やワクチン開発の少人数・低コスト・迅速化を実現する。



「富岳」創薬DXプラットフォーム

「富岳」を基軸として、AIとシミュレーションを融合させた革新的な創薬DXプラットフォームを構築し、創薬プロセスの超効率化を目指す。これにより、新薬やワクチン開発の少人数・低コスト・迅速化を実現する。



リード化合物創出　ワークフロー

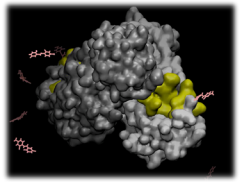
- 「富岳」を中心に、HPC/AIフローの自動化を図ることで、創薬の超効率化を実現
- リード化合物創出：標的タンパク質名を入力して、リード化合物を推定するHPC/AIフロー

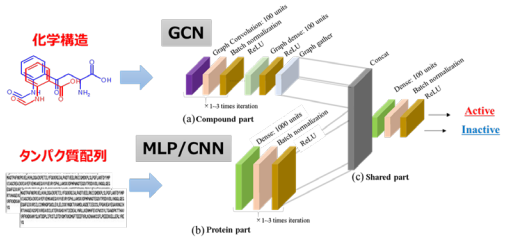
創薬標的分子の遺伝子名（例：タンパク質のアミノ酸配列）

AlphaFoldによる立体構造予測

数十億化合物
ライブラリー

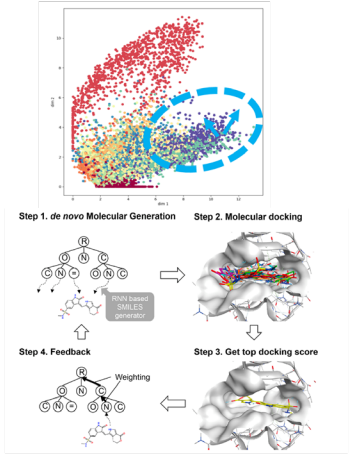
Dockingよる活性化合物探索



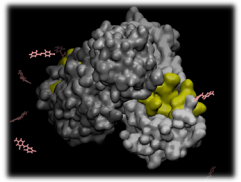


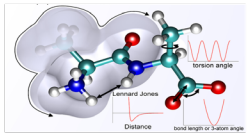
GCNによる活性･ADMET予測

生成モデルによるリード最適化

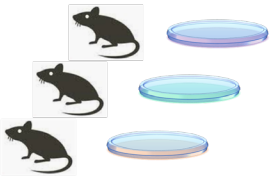


MDによるバーチャル評価





MM･QMによる
物理化学的安定性評価



新規薬剤分子の創出

世の中にあるossワークフローエンジン

（参考）OpensourceのWorkflow Framework

2022~2023年調査

JN: Jupyter Notebook, WM: Workflow Management
〇: 対応済み・わずかな手間で対応可, △: ある程度の手間で対応可, ×: 対応不可・相当手間をかけないと対応できない, ?:不明

名称	Owner	License	DX Platformで必要な機能						GitHub		備考
			富岳ジョブ投入対応	GUI/連携動作	中途再実行	マルチユーザ	フロー作成	フローGUI管理	☆	👁	
Airflow	Apache Airflow Community	Apache 2.0	×？ Job Schedulerとかは標準では無いAdd-onにあるか	△-× 本体にタスクのGUIを実現する機能は無い。別個に作成	〇 GUIでも操作可能	× （フローの実行権限の管理）	Python	〇	27,800	760	Pythonで制御するWMエンジン Apache Projectなので情報が豊富 Eco Systemも確立。多彩なadd-on ビジネス向けの定型Workflow構築が主目的なので、DXPFのようなフロー構築に適しているのか？
Prefect	Prefect Technologies	Apache 2.0 (open source版)	〇 PythonでSSHのWrapperを組めばよいか	△-× AirFlowに同じ	〇 GUIでも操作可能？	×	Python	〇	11,800	153	Airflowの代替を目標に開発されたWM Pythonの関数として定義したTaskをつなげてflowにして実行 軽量。シンプルで理解しやすい 動的なflow（Taskの結果で後続の作業が変わるflow）を作成することが可能
Ray	The Ray Team	Apache 2.0	×? Ray Clusterで管理するには、各々に専用の常駐プログラムが必要？	△-× AirFlowに同じ	?	×	Python	×	22,300	427	分散タスクを管理するRay Coreを中心としたシステム クラスタを管理するRay Cluster, MLフローをサポートするRay AIR等から構成される
Luigi	Spotify	Apache 2.0	×？ Job Schedulerとかは標準では無いAdd-onにあるか	△-× AirFlowに同じ	?	×	Python	〇	16,000	486	Pythonで制御するWMエンジン ビジネス向けの定型Workflow構築が主目的なので、DXPFのようなフロー構築に適しているのか？
Dask	Dask core developers	BSD3 Clause	×? Job Schedulerはあるが、富岳用にconfigできないっぽい	△-× AirFlowに同じ	?	×	Python	×	10,400	221	Pythonで制御するWMエンジン PythonのDataframeやndarrayを分散処理できる機能がある DXPFのように、HPC上のアプリケーションをPipelineでつなげたFlowを管理するという用途には適していない？

（参考）OpensourceのWorkflow Framework – 続き

2022~2023年調査

JN: Jupyter Notebook, WM: Workflow Management
〇: 対応済み・わずかな手間で対応可, △: ある程度の手間で対応可, ×: 対応不可・相当手間をかけないと対応できない, ?:不明

名称 c	Owner	License	DX Platformで必要な機能						GitHub		備考
			富岳対応	GUI/連携動作	中途再実行	マルチユーザ	フロー作成	フローGUI管理	☆	👁	
Cromwell	Broad Institute	BSD3 Clause	△ HPC制御機能をConfig可なのでいけそう	△-× AirFlowに同じ	〇? CallCaching機能が相当しそう	×	WDL	×	850	113	Javaで動くプログラム
AiiDA	MARVEL National Centre of Competence in Research,	MIT	△ HPC制御機能をConfig可なのでいけそう	△-× AirFlowに同じ	〇 各ステップの入力/出力をDBで記録・管理	×	Python	×	327	27	PythonからWMエンジンの操作を可能にするライブラリ群を提供
Covalent	Agnostiq	AGPL3	×? SSHを飛ばせるがJob Schedulingが無さそう	△-× AirFlowに同じ	?	×	Python	〇	173	10	見た目がWHEELに似ている
Makeflow	University of Notre Dame	GPLv2	△ 汎用Queue制御が使えそう	△-× AirFlowに同じ	△? “Make like”なので可能？	×	独自Makefile like	×	109	18	CCL Toolsという分散システムコラボツールの一部
Stream Flow	Università di Torino	LGPLv3	? SSHでいけるのか	?	?	×	CWL	×	23	7	ドキュメントがちょっとpoor
Jupyter Workflow	Università di Torino	LGPLv3	↑	JN連携	△ JNから動かすので可能？	×	↑	×	7	6	Stream FlowのJNの拡張 コア部分はStream Flowを使う Stream FlowをJNから使うためのUI的位置づけか

など多数

- Prefectを利用してDX Platformを構築

DXプラットフォームコンセプト：

- 様々なアプリを**一つのプラットフォーム**に実装し**全体の最適化**を図る
- **自由につなげた創薬シナリオを構築できる**
- 創薬アプリの**共通プラットフォームへ**

創薬DX Platformのワークフローエンジン

flow_trial.py

```
import requests
from prefect import flow, task

@task
def call_api(url):
    response = requests.get(url)
    print(response.status_code)
    return response.json()

@task
def parse_fact(response):
    fact = response["fact"]
    print(fact)
    return fact

@flow
def api_flow(url):
    fact_json = call_api(url)
    fact_text = parse_fact(fact_json)
    return fact_text

api_flow("https://catfact.ninja/fact")
```

タスク1の定義

タスク2の定義

関数を引数でつなげる形で、タスク1→タスク2のフローを定義

最初の引数を与えてフローを実行

Prefect（Open source版Apache-2.0）

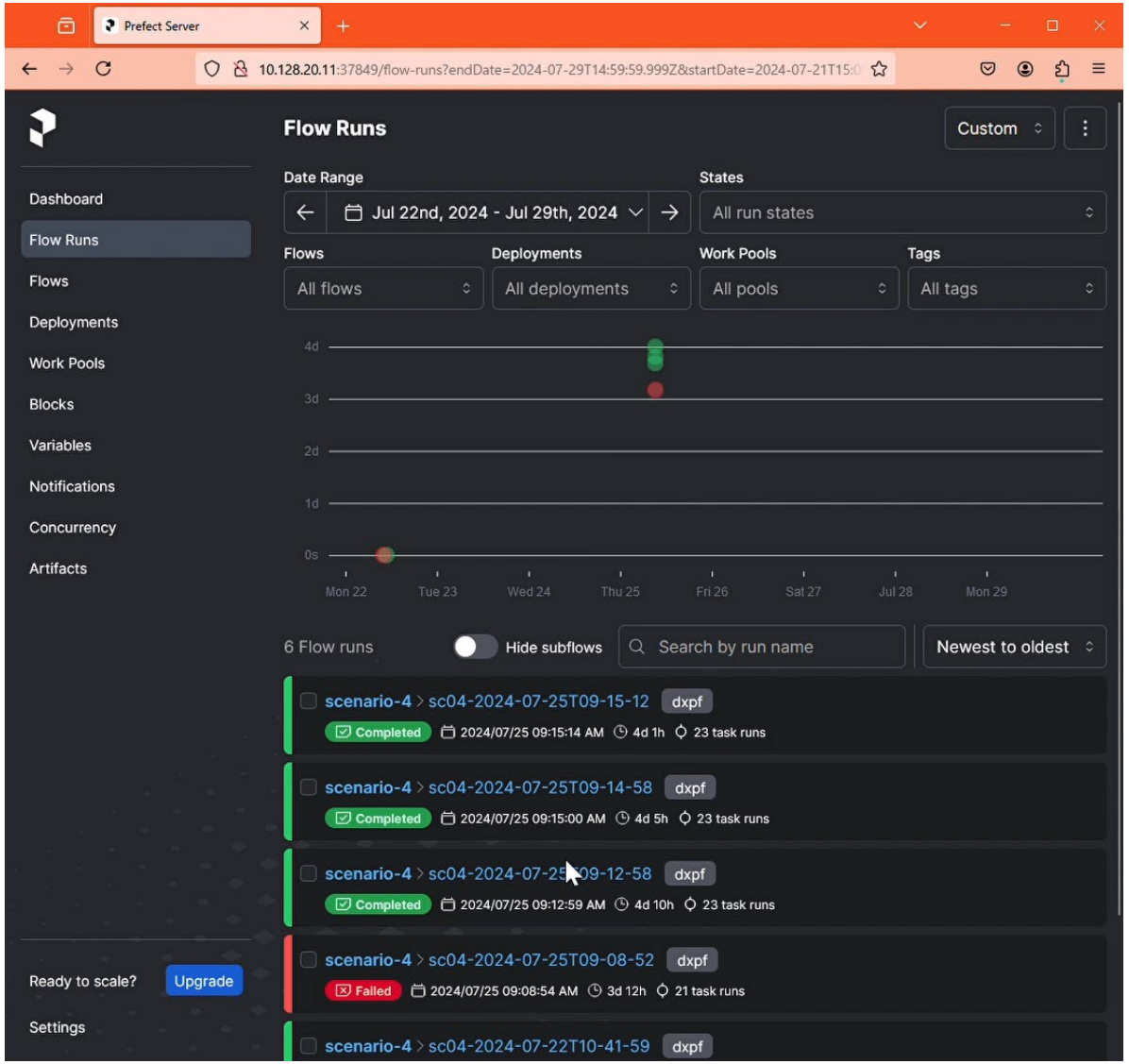
- 個別の機能をもった処理（**タスク**）を、**フロー**として複数つなげて実行するワークフローエンジン。全体として複雑なデータ処理を実現
- ワークフローエンジンとしてワークフローを効率的に実行する機能を提供する
  - 実行状況のモニタリング
  - タスクで発生したエラーの処理。
  - 保持した実行結果（キャッシュ）を利用した再実行
  - タスク間でのデータの受け渡しなど
- 「富岳」へのジョブ投入も可能
- GUIでフローに操作を加える場合は画面を構築する必要あり
- Prefect のPrefectの全体像やコンセプト、詳細https://docs.prefect.io/latest/
- Prefect には、Open source 版とCloud 版の二つのバージョンがある。
- 今回はOpen source版を利用。そのため、自前でサービスの立ち上げ・保守を実施する必要あり。

フロー例: 単純なドッキング

- 標的タンパク質分子に対して、
- 指定した化合物のドッキングを行い、
- 医薬品候補化合物の絞り込みを行う

計算（原因不明で）失敗した場合
タンパク質-医薬品候補化合物の間の結合力（ΔG）計算：MP-CAFEE

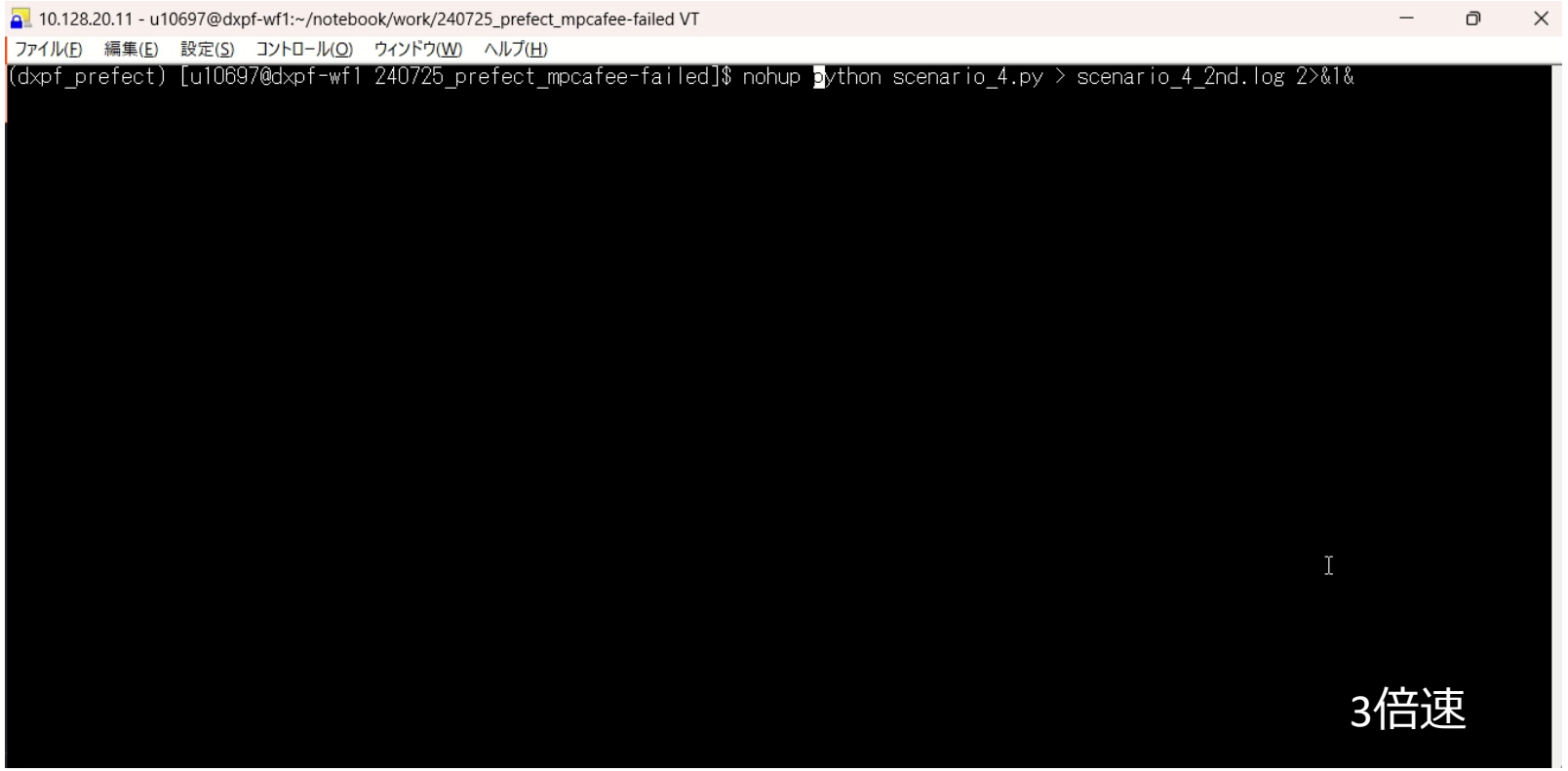
候補化合物分実行（10~100フロー）



- 多数のジョブの同時実行していると何らかのエラーで失敗することも多い
  - 例えば、通信切断、ほかユーザーの影響、一時的メモリ不足、原因特定困難（宇宙線…？）、…

タンパク質-リガンド間親和性（ΔG）計算のやりなおし

- まずはやり直してみる。
  - 各タスクの入出力は「ファイル渡し」になっているものが多い。また、各タスクの実装として、必要ファイルの存在有無でそのタスクの実行必要性を判定するように実装。
  - やり直しでダメなら入力ファイルなどを確認する。



候補化合物分実行（10~100フロー）

※そもそも「大量化合物→有望化合物への絞り込み」のような利用シーンでは、多少の失敗やエラーが起きても無視して別の化合物の処理を進めるように実装

今後の展開

AIとシミュレーションを融合させた革新的な創薬DXプラットフォームを構築し、創薬プロセスの超効率化を目指す。これにより、新薬やワクチン開発の超効率化を実現する。

連携の強化
- **他グループの開発アプリの実装・協力体制構築**
- **創薬から生命科学応用へ**

サイバー空間
IT企業・アカデミア

医薬品開発の超効率化

病気A
標的タンパク質X
有望な医薬品候補Z
医薬品候補Z
安全に効く患者群S
薬価はWが妥当
患者群P

新規アプリ開発
中分子・高分子デザインへの拡充

PFの臨床領域への拡大

ターゲット探索 → リード探索 → リード最適化 → バイオアッセイ → 前臨床試 → 治療

現実社会
創薬現場

病気Aの原因タンパクは？
化合物Yの薬効は？毒性・ADMEは？
治験に合格するには？治療方法は？
市販後の副作用の危険性は？費用対効果は？

現行PFの製薬企業・アカデミア創薬でのテスト・本格運用

外部資金

- MEXT「富岳」Society 5.0推進利用課題 (hp220284、hp230189)
- MEXT「富岳」成果創出加速プログラム「「富岳」で目指すシミュレーション・AI 駆動型次世代医療・創薬」(JPMXP1020230120、hp230216)
- AMED 創薬支援推進事業・産学連携による次世代創薬AI開発（DAIIA）JP22nk0101111