統合化推進プログラム 2026年度新規採択課題の紹介

MDデータの共有と再利用を支えるデータ・モデル基盤の構築

MDDB-AI hub

松永 康佑

埼玉大学大学院理工学研究科 准教授

理化学研究所計算科学研究センター AI4Science部門 ユニットリーダー

2026年10月5日 トーゴーの日シンポジウム2026 品川ザ・グランドホール

内容

① 研究開発の背景

- 分子動力学シミュレーションによる動態の計測
- MDデータの共有に関する世界的動向
- MDサロゲートモデル
- MDデータへのアクセス課題

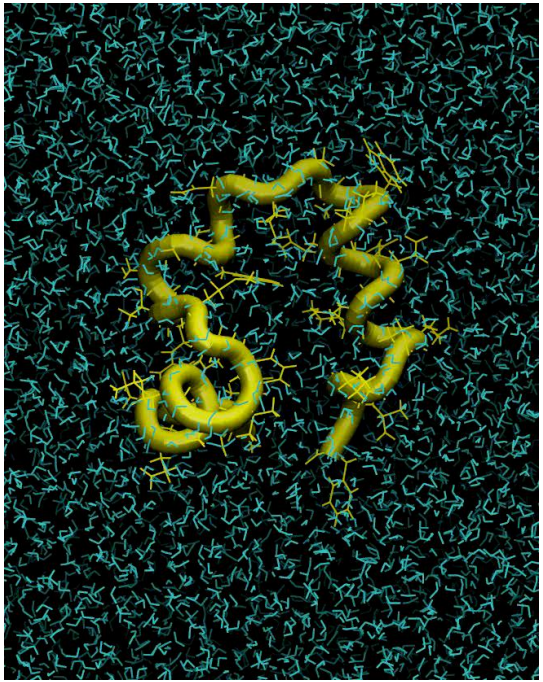
② 研究開発計画

- データベース全体像
- 研究体制
- スケジュール

③ 取り組み紹介

- バックエンド体制（HPCI共用ストレージ）
- フロントエンド

① 研究開発の背景: 分子動力学（MD）シミュレーションによる動態の計測

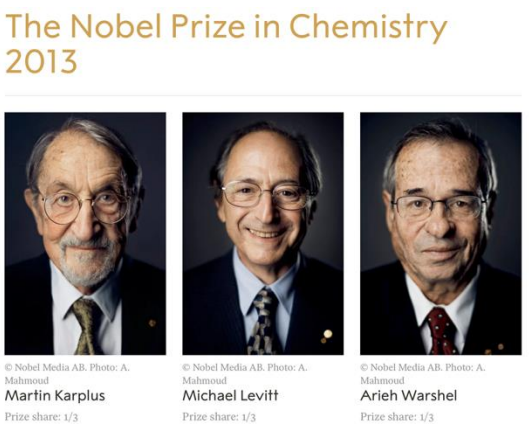


分子シミュレーションではまずコンピュータの中にタンパク質構造を作る。
次に、物理法則に従い原子を動かすことで原子の動きを観測する

ニュートンの第二法則

$$F = ma$$

原子の動きを観察

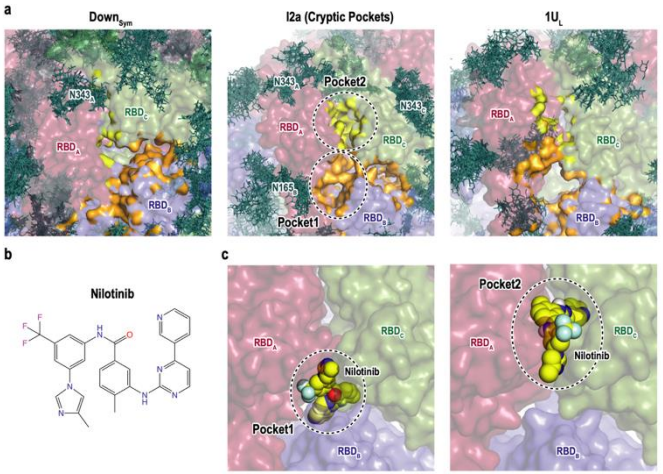


nobelprize.org

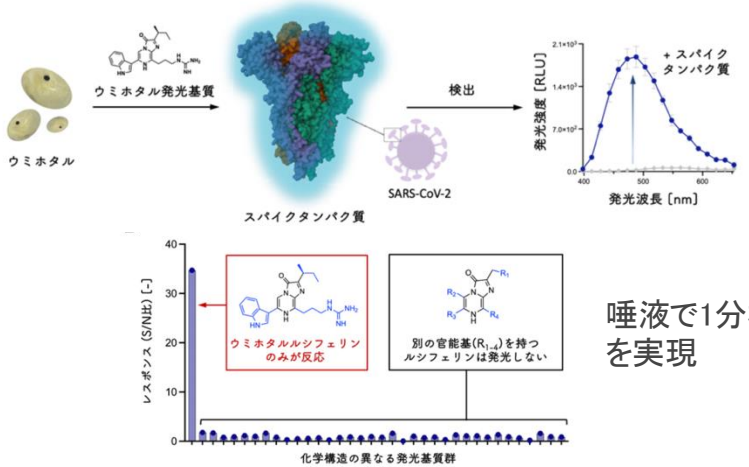
タンパク質の動き（動態）を計測する強力な手法
実験を補完するとともに、AIモデルの訓練データにもなる

① 研究開発の背景: 実験研究者が動態を計測することの利点

静的構造では見えない「Cryptic Pocket」をMD動態から発見： SARS-CoV-2 Spike Proteinでの実例



分担の杉田Gの成果例



唾液で1分検出を実現

「富岳」gREST計算で中間構造を予測

Spike ProteinのDown→Up構造変化の経路を「富岳」上のgREST法で計算し、中間体の構造に**Cryptic Pocketを発見**
(Dokainish et al., eLife 2022)

MDシミュレーションにより発見

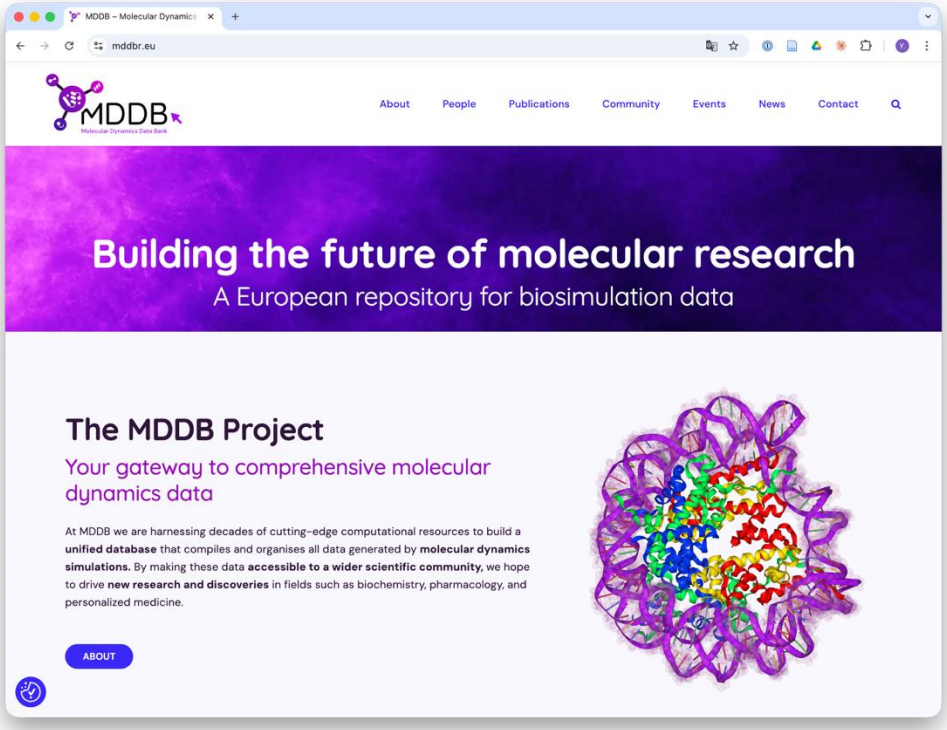
ウミホタル発光基質による迅速検出への応用

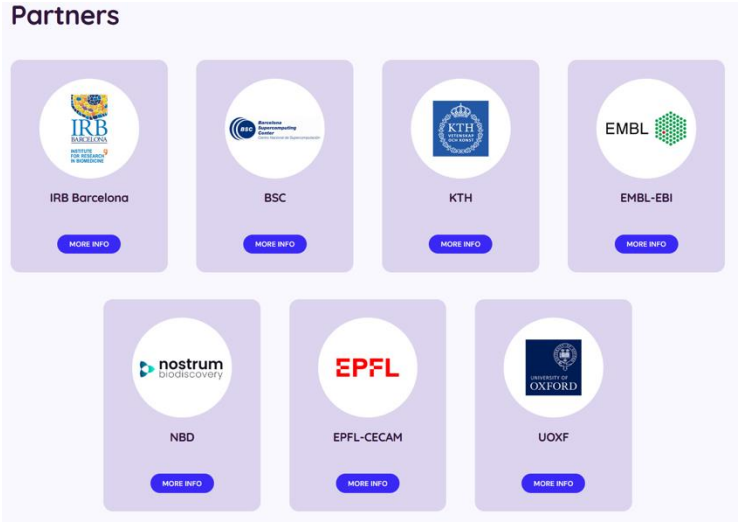
ウミホタル由来の発光基質がMD計算で見つけたCryptic Pocketに結合することを見出す（AIST 西原ら)。
(Nishihara et al., ACS Central Science 2024)

実験研究者がMDの結果を利用

MDDB-AI Hubがあれば: PDB ID入力→動態可視化→Cryptic Pocket発見→創薬ターゲットの拡充に貢献

① 研究開発の背景: MDデータの共有に関する世界的動向





R. Amaro *et al. Nature Methods, 22, 641-645 (2025)*
https://doi.org/10.1038/s41592-025-02635-0



DR. MODESTO OROZCO
Group Leader
Molecular Modelling and Bioinformatics



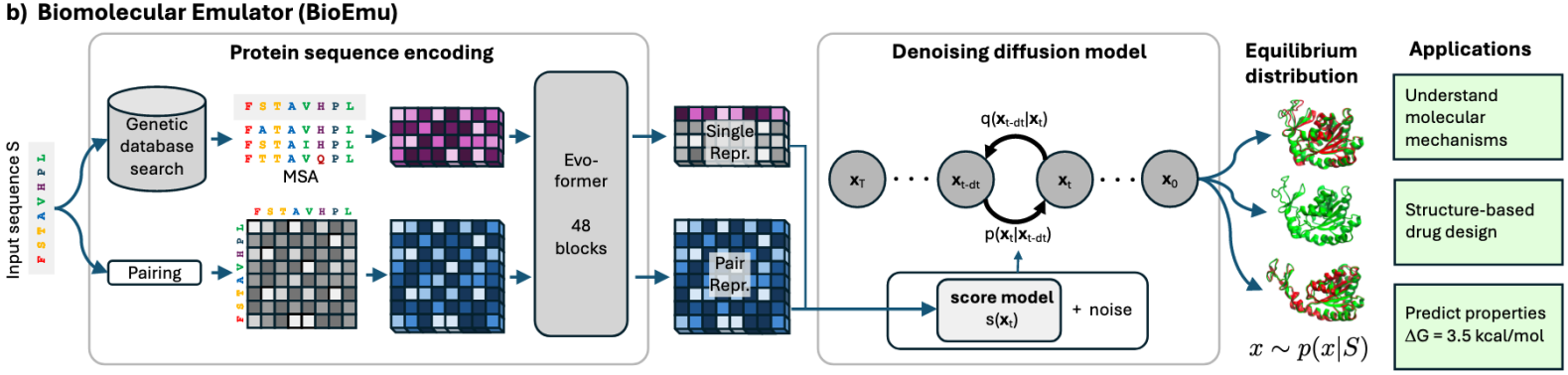
DR. ADAM HOSPITAL
Research Associate
Molecular Modelling and Bioinformatics

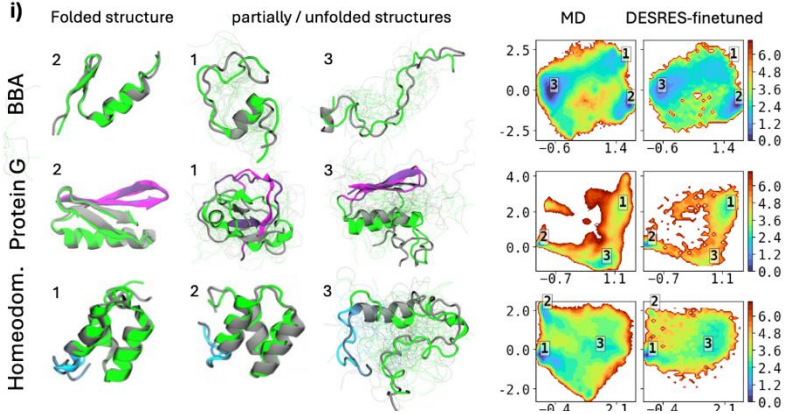
- 杉田、小林は代表メンバーであるOrozco (IRB Barcelona)らと技術的な連携について議論中
- HPCI共用ストレージの利用に対して理研、運用技術部門と共に技術的な検討を開始

① 研究開発の構想: MDを代替するAIサロゲートモデルの出現

MDサロゲートモデルの例(BioEmu)

S. Lewis et al., *Scalable Emulation of Protein Equilibrium Ensembles with Generative Deep Learning*. Science (2025)





- MDシミュレーションで数週間～数カ月かかる計算を、1GPUでわずか数時間で代替することができる
- しかし、精度・汎化性能にいまだ課題

MDDB-AI Hubがあれば:
訓練データの整備、モデルの取得、モデルの比較が可能となり、AIサロゲートモデルの開発の迅速化が可能に

① 研究開発の背景: MDデータへのアクセス課題

実験生命科学者に有用な「動態」

タンパク質の機能は「形」だけでなく「動き」で決まる。静的構造では見えない結合部位（Cryptic Pocket）やIDPの揺らぎが創薬や疾患理解の鍵。

計算科学者のシミュレーションデータ

コンピュータ上にタンパク質を置き、物理法則(F=ma)に従い原子を動かして「動き」を再現する手法。実験を補い、AI訓練データも生成できる。

AI研究者の必要なデータ・ベンチマーク

MDの動態をAIで高速再現するサロゲートモデル開発が活発（例 BioEmu: Science 2025）。

課題1: MD結果が実験研究者に届かない

MDの結果は計算科学者の手元にしかなく、PDB IDを入れて動態が見える仕組みがない。

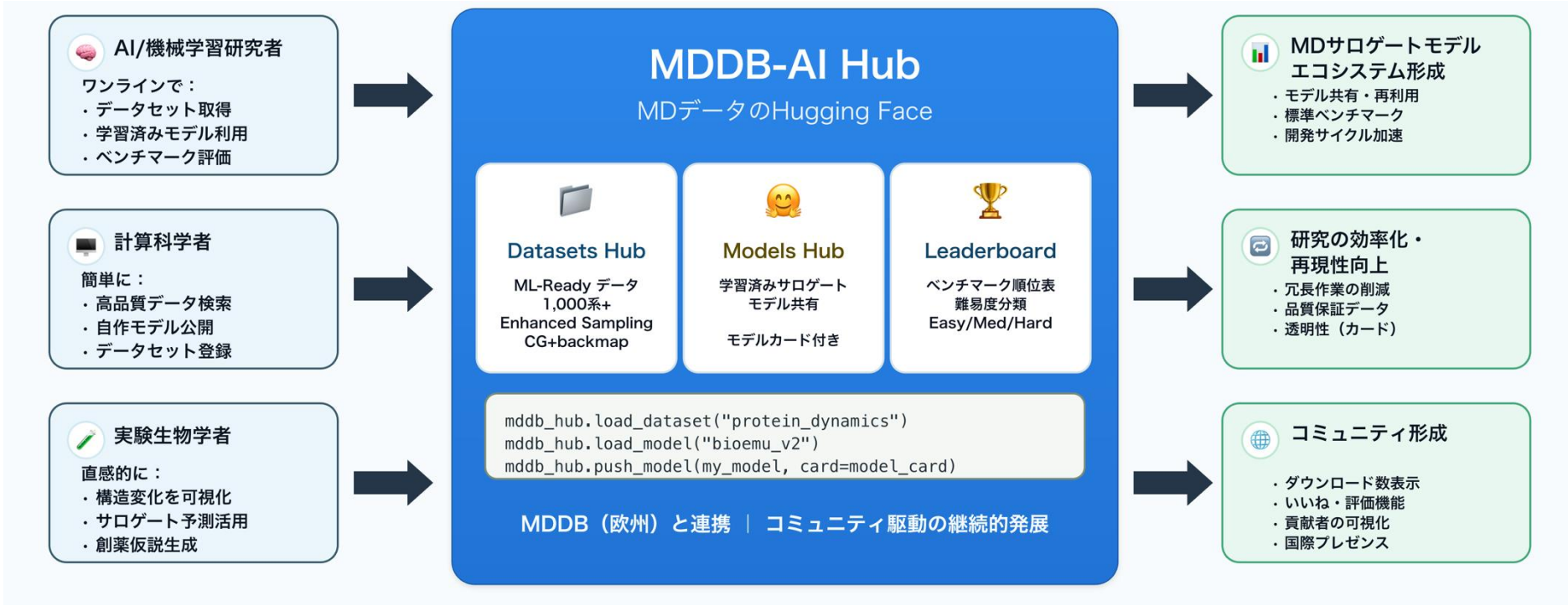
課題2: MDデータの共有基盤が必要

研究グループが独自にデータを保持し、再利用・共有がされていない。標準的な登録先が必要。

課題3: 標準ベンチマークがない

サロゲートモデルの性能を客観比較する標準指標がなく、モデル同士の優劣がつけられない。

② 研究開発計画：本課題 MDDB–AI hub の全体像



達成目標1	達成目標2	達成目標3	達成目標4
MDDB国際連携 ブランチとして稼働	実験生命科学者がWebから 動態を可視化(100名)	コアデータセット+ 独自データセットの公開	MDサロゲートモデル訓練基盤 モデル公開・ベンチマーク機能

② 研究開発計画: 研究体制

松永G（埼玉大）

全体総括
フロントエンド、MDサロゲート

- プロジェクト全体の統括
- MDサロゲートモデル向けデータ仕様・ML-Ready形式の策定
- Pythonクライアント mddb_hub、Web解析・構造可視化機能の開発
- ベンチマーク／リーダーボード設計・運用



杉田G（東京大）

MDデータ登録

- GENESISによる拡張サンプリングデータの整備
- REMD/REUS/gRESTデータのreweight処理・追加計算
- CGBack等を用いたマルチスケールデータ整備
- 搭載データの専門的品質評価



小林G（理研）

バックエンド、国際連携

- HPCI共用ストレージ・データ公開システムの設計・実装・運用
- テストベッド、α版、正式版の公開
- メタデータスキーマ設計とShibboleth認証によるアクセス管理
- 欧州MDDBとのAPI・メタデータ相互運用の実装



② 研究開発計画：スケジュール

2026年度（1年目）
基盤構築フェーズ

ユーザ視点:

動態の可視化

MDDBとの接続・ブランチ動作
杉田研・松永研がデータ登録準備

開発視点:

独自機能作成・構造可視化
動的挙動の統計・マルチスケール
Reweighting/Resampling導入

α版限定公開

2027年度（2年目）
機能拡充・ユーザ獲得

ユーザ視点:

計算科学者（数十名）がデータ登録
実験生命科学者（共同研究者）が
お試し利用

開発視点:

IDP・液滴の可視化
CG+backmappingパイプライン
サロゲートモデル導入準備

β版公開

2028年度（3年目）
公開・普及フェーズ

ユーザ視点:

学会・講義を通じて普及活動
生物学者約100名が利用
AI/ML研究者(数グループ)が
モデル訓練

開発視点:

ハンズオン開催・ユーザ拡大
サロゲート実証(3ターゲット)
DB論文投稿・国際ワークショップ

正式版一般公開

達成目標1	達成目標2	達成目標3	達成目標4
MDDB国際連携 ブランチとして稼働	実験生命科学者がWebから 動態を可視化(100名)	コアデータセット+ 独自データセットの公開	MDサロゲートモデル訓練基盤 モデル公開・ベンチマーク機能

③取り組み紹介 :バックエンド開発

小林グループ: MDDB日本ノード構築に向けた取り組み

MDDB（Molecular Dynamics DataBase）：
分子動力学(MD) シミュレーションデータの共有・公開・再利用を目的とした国際基盤
FAIR(Findable, Accessible, Interoperable, Reusable)原則準拠

欧州を中心に7拠点で運営中



R. Amaro *et al. Nature Methods, 22, 641-645 (2025)*
https://doi.org/10.1038/s41592-025-02635-0
MDDB https://mddbr.eu/
MDposit https://mdposit.mddbr.eu/

日本ノード構築に向けた取り組み

MDDB
Molecular Dynamics Data Bank

MDDB-AI HUB (Japan node)

HPCI High Performance Computing Infrastructure

HPCI Shared Storage

HPCI共用ストレージ
テストベッドを構築中
VM + MDDB Docker 環境

Japan　Asia-Pacific　Global

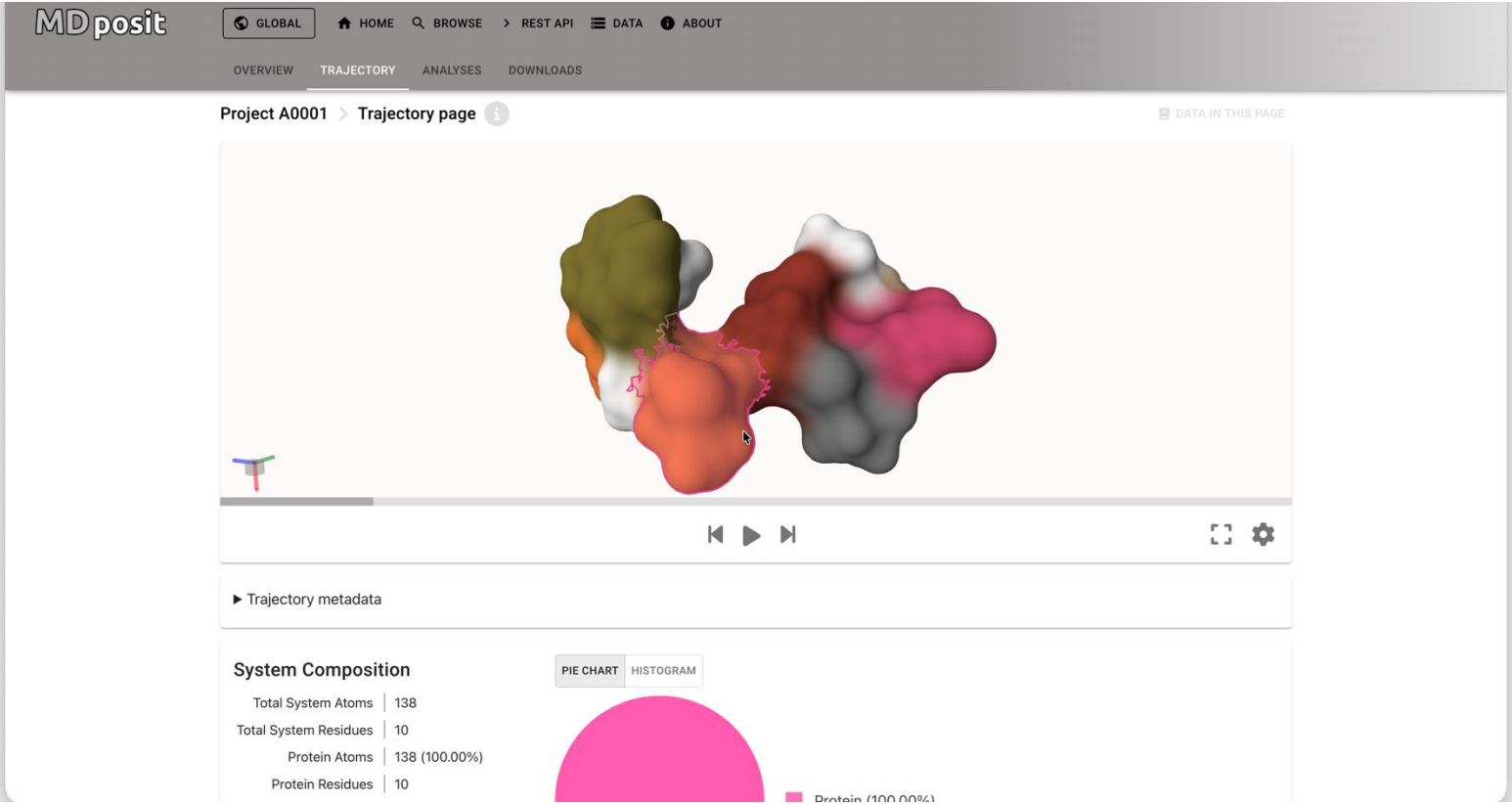
Technical achievements
- ローカル環境構築
- データ登録検証
- HPCIへの展開

詳細はポスター(小林G)にて！

Governance & Operation
- グローバルMDDBネットワークとの接続方式
- HPCI上のVMの管理体制
- 権限分担（運用側・開発側）
- 持続可能な運用モデル

③ 取り組み紹介：フロントエンド開発

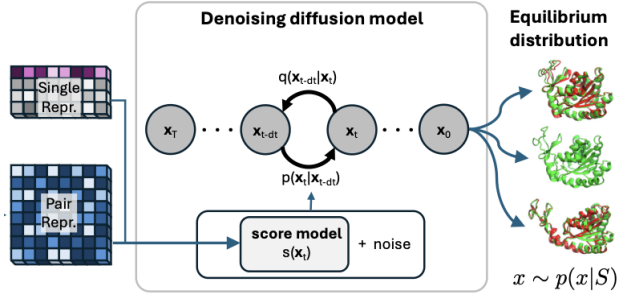
松永グループ: Molstarを用いたMDトラジェクトリの可視化実装



③取り組み紹介 :MDサロゲートモデルのリーダーボード

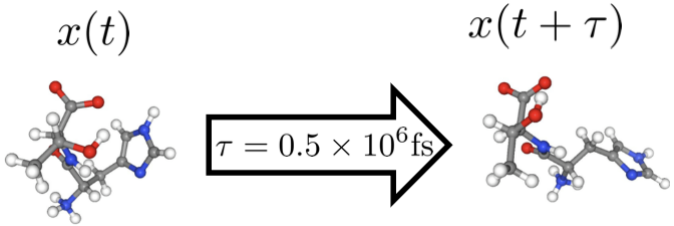
松永グループ: MDサロゲートモデルのルーブリックによる評価基盤の開発

EnsembleEmulator



Lewis et. al., 2026

MDEmulator



Klein et. al., 2023

評価軸	EnsembleEmulator	MDEmulator	…	何を測るか
構造妥当性	✔	✔		出力構造の物理的妥当性（結合長・角・立体衝突）
計算効率	✔	✔		計算時間
平衡分布	✔	✔		生成構造が平衡分布に従っているか
動力学	ー	✔		状態間遷移の速さ
構造被覆	✔	※系による		構造的に異なるPDB登録構造をカバーしているか

③ 進捗紹介：MDサロゲートモデルのリーダーボード

暫定的なリーダーボード

EnsembleEmulator

#	model_key	S (overall)	✓ eff	✓ validity	✓ equil	✓ coverage
1	boltz@boltz2	**0.761**	0.949	0.998	0.349	0.750
2	bioemu@bioemu-v1.1	**0.741**	0.947	0.506	0.722	0.788
3	colabfold@colabfold	**0.685**	0.880	0.778	0.329	0.751
4	bioemu@bioemu-v1.0	**0.668**	0.975	0.476	0.488	0.732
5	str2str@str2str	**0.627**	0.959	0.460	0.607	0.483

- ✓ Boltz は構造妥当性 ⭕
- ✓ Bioemu は平衡分布・構造被覆 ⭕

MDEmulator

#	model_key	S (overall)	✓ eff	✓ validity	✓ equil	✕ coverage	✓ kinetics
1	timewarp@timewarp-2aa	**0.823**	0.921	0.968	1.000	–	0.403
2	timewarp@timewarp-4aa	**0.746**	0.839	0.931	0.920	–	0.294
3	mdgen@mdgen-4aa	**0.697**	0.801	0.624	0.931	–	0.433
4	mdgen@mdgen-atlas	**0.540**	0.971	0.482	0.378	0.473	0.330

- ✓ Timewarp は構造妥当性 ⭕
- ✓ MDGen は動力学 ⭕
- ✓ 平衡分布：取りうる構造空間の広さが違うので単純比較できない

まとめ

- MDで得られるタンパク質の動態は、実験結果の解釈やAIモデルの学習に活用できる研究資源である。その価値を分野を越えて活かすには、計算条件とともにデータを共有し、専門家以外も理解し、再利用できる形に整えることが重要である

- MDDB-AI Hubは、MDデータの登録・検索・可視化と、MDサロゲートモデルの共有・評価を一体化し、実験・計算・AI研究者が同じ基盤でMDデータを活用できる「MDデータのHugging Face」を目指す