超大量ドッキングのための AI ドッキングシステムの構築
仕様書

国立研究開発法人理化学研究所

1　概要

国立研究開発法人理化学研究所（以下、「本研究所」と呼ぶ）HPC/AI 駆動型医薬プラットフォーム部門では、人工知能（AI）、分子シミュレーションや、実験を融合する新たな方法論の開発を目指し、そのための重要な手段として、HPC/AI 駆動型生命科学統合プラットフォームの構築を行なっている。

本調達では、スーパーコンピュータ「富岳」を利用した超大量（超高速）ドッキング（Ultra Large Scale Docking, ULSD）の実現を目指し、AI ドッキングシステムおよび大量ドッキングシステムの構築を行う。

2　目的

本仕様書は、ドッキングスコア（以下 DS）を予測する AI を用いて、良好な DS を有する化合物を、膨大な数の化合物から効率よく取得する AI ドッキングシステムの構築を目指すものである。また、そのために必要な大量ドッキングシステムの構築も実施する。

創薬等に用いることのできる合成・取得可能な化合物は近年増大しており、数十億規模の仮想ライブラリーが複数出現している。これらの膨大な化合物を全て、合成・入手・アッセイすることは現実的でないため、コンピュータを用いて合成・入手・アッセイする化合物を絞り込むことが必要になる。コンピュータを用いて化合物を絞り込む手段の一つとして、標的タンパク質の立体構造を利用したドッキングがある。しかしながら、膨大な数の化合物を全てドッキングすることは、実行時間および計算資源の観点から難しい。そこで、AI 技術を用いて、良好な DS を有することが期待される化合物を絞り込み、少ないドッキング数で良好な DS の化合物を取得する AI ドッキングシステムを構築する。

また、AI ドッキングシステムが良好な DS の化合物を効率よく取得しているかを確認するためには、大量化合物の DS を取得する必要があるため、このためのシステムも構築する。

3　仕様

AI ドッキングシステムおよび大量ドッキングシステムは、理研の保有する、富岳、創薬 DXPF Workflow(DXPF-WF)サーバー、創薬 DXPF WF サーバーに連結されたストレージサーバー上の化合物データベース、R&D 用 GPU(R&D-GPU)サーバーなどを利用したシステムとして動作するものとする。

AI ドッキングシステムおよび大量ドッキングシステムは DXPF-WF サーバーにユーザーがログインし、Jupyter notebook（以下 JN）を使ったユーザーインターフェイス（以下 UI）から、ドッキングのための設定を行い、富岳、DXPF-WF サーバー、R&D-GPU などを適宜利用してジョブを実行し、DXPF-WF サーバー上のジョブごとに設定されたディレクトリーあるいはデータベースに実行結果を格納する。

AI ドッキングシステムおよび大量ドッキングシステムのジョブスクリプトは JN により自動生成され、DXPF-WF サーバーで実行される。また、このジョブスクリプトにより富岳計算ノード上あるいは R&D-GPU サーバー上のドッキング、R&D-GPU サーバー上の DS 予測 AI（以下 DS-AI）構

築・予測などが起動・実行される。なお、どの処理がどのサーバーで実行されるかについては、監督員と協議の上変更することがある。

AI ドッキングシステムおよび大量ドッキングシステムのプロトタイプについては、理研で JN を用いて開発されたものが既に存在する。本業務ではこのプロトタイプを基に、必要に応じて、UI の改良、ドッキング用初期化合物セット設定プロセスの開発、ドッキングジョブのジョブ管理・エラー対応、DS-AI の改良、実行結果のデータベース登録・管理、大量化合物を使用した実行テストなどを行う。開発に必要なプロトタイプソフトウェア、実行環境、実行用アカウント、化合物データベース、富岳アカウント、富岳計算時間等は監督員から提供される。詳細については監督員と相談の上、決定する。

両システムにおいて、実行時間が長いものについては、必要に応じて DXPF-WF サーバーあるいは富岳計算ノード上でバッチジョブとして実行し、ジョブ終了時にメールで終了を知らせるシステムとすることがある。どの処理をバッチジョブにするかについては、監督員と協議の上決定する。

3.1 プロトタイプについて

現状のプロトタイプに関する情報（「･」で示す）および本件で改変する事項（「➢」で示す）は以下である。なお、改変する事項の詳細については監督員と相談の上、決定する。

3.1.1 AI ドッキングシステムプロトタイプ

- AI ドッキングタスク名 $NAME_{task}$（入力）：実施する AI ドッキングにタスク名をつける。$NAME_{task}$ はユーザーが JN UI を使用して入力する。各タスクは $NAME_{task}$_YYMMDDHHMM（YYMMDDHHMM は年月日時間）というディレクトリー下の適切な場所に入力、出力を格納・保存する。
- ドッキング用ソフトウェア：富岳計算ノード上あるいは R&D-GPU サーバー上で動作する AutoDock VINA（以下 ADV）。どちらの場合も、1 ノード中の 8 コアを使用して実行する。したがって、富岳計算ノードの場合 1 ノード(=48 コア)では 6 個の、DXPF-WF サーバーの場合 1 ノード(=40 コア)では 5 個の ADV が並列に流れる。現プロトタイプではドッキングを実行する部分が省略されており、既に実施した DS を使用するようになっている。
  - ➢ 本業務では、ドッキングを富岳計算ノードあるいは R&D-GPU サーバーで実施する処理を組み込む。
- タンパク質構造ファイル（ADV 入力）：ユーザーが準備したタンパク質構造の pdbqt 形式ファイル。UI からアップロードして使用される。
- 結合ポケット情報（ADV 入力）：ユーザーが事前に検討し、中心位置座標、直方体の各辺の長さを決めておく。ユーザーが UI から数値で入力。
  - ➢ 本業務では、結合ポケット情報について、中心位置座標、直方体の各辺の長さの情報を含むテキストファイルをアップロードし、使用するように修正する。
- ADV のドッキングポーズ探索パラメータ Exhaustiveness（ADV 入力）：ユーザーがユーザーインターフェイスから選択(default=8)。
- ADV のドッキングポーズ数 N_{pose}（ADV 入力）：ユーザーが UI から N_{pose} を数値で指定する。1 化合物 N_{pose} 結合ポーズ（DS 最小から N_{pose} 番目までの結合ポーズ）を出力する。
  - ➢ 本業務では、N_{pose}＝１で固定し、1 化合物 1 結合ポーズ（DS 最小のもの）を出力するよ

うに改変する。

- 初期ドッキング化合物リスト（ADV 入力）：最初にドッキングを行う化合物 ID のリスト。ユーザーが事前に化合物を指定・選択する。ユーザーインターフェイスからアップロードして使用される。このリストに何化合物（N_{dock}）を載せるかは、ユーザーが UI から数値で入力。ドッキングを繰り返す際もこの化合物数 N_{dock} が使用される。
  - 本業務では、N_{dock} 値をいくつかの N_{dock} 値（240,360,480,600,720）からメニューで選ぶ形に改修し、default=480 とする。
  - 本業務では、初期ドッキング化合物リストの作成方法は、ユーザーがユーザーインターフェイスを利用していくつかのオプションから選択する方法に改変する。

    【オプション案】

    ①初期ドッキング用化合物の重原子数 N_{heavy} および回転可能結合数 N_{rotb} の範囲、リスト化合物の数 N_{dock} をユーザーが指定し、化合物データベースから条件に合う化合物をランダムに選択する。

    ②ドッキング化合物の重原子数範囲(N_{heavy}, default= 30～35)、回転可能結合数範囲(N_{rotb}, default=2～4)が default でよい場合は事前に化合物を選択しておき、ユーザーが指定した数(N_{dock}, default=480)だけリストに載せる。
- ドッキング化合物リスト（DS-AI 出力、ADV 入力）：ドッキングの繰り返し 2 回目以降はドッキングしていない化合物の中で DS-AI が予測した DS が良いもの上位 N_{dock} 個を選択する。
- ドッキング用リガンド 3 次元構造（ADV 入力）：化合物データベース中に化合物 ID および pdbqt 形式のテキスト項目として格納されており、化合物 ID で検索（SQL）後、ADV 用に各化合物 ID 名の pdbqt ファイルを生成する。
- ドッキング繰り返し数 N_{iter}（入力）：N_{dock} 化合物のドッキングを繰り返す回数。AI ドッキングでは、$N_{dock} \times N_{iter}$ の化合物のドッキングを行うことにより、良好な DS を有する化合物を探索する。
  - 本業務では N_{iter} の default を 1000 に設定する。
- ドッキングに使用するノード数 N_{nodes}（入力）：ユーザーが UI から数値で入力する。
  - 本業務では、富岳計算ノードの場合、N_{nodes} の default を 64 に設定する。DXPF-WF サーバーの場合、N_{nodes} の default を 8 に設定する。
- 各 ADV が処理する化合物数は、JN がドッキングする化合物数 N_{dock} および使用するノード数 N_{nodes} に基づいて決定・配分する。
- 初期ドッキング後、「ドッキング→これまでに実施された DS に今回のドッキング結果を追加→全 DS を用いて DS-AI 予測モデルを構築→DS-AI 予測モデルを用いてドッキング未実施化合物の DS を予測→予測 DS が良いもの上位 N_{dock} を選択→ドッキング」を N_{iter} 回繰り返す。
- 各回の ADV 出力は、まず富岳ログインノード上に格納され、並列実施される全 ADV が完了した時点で DXPF-WF サーバーのストレージサーバーに全出力を移動する。
- DS および計算時間のリスト（ADV 出力、DS-AI 入力）：ドッキングした N_{dock} 化合物について、化合物 ID、DS、ドッキング所要時間、ドッキング結果のドッキングポーズ格納場所が記載されたリストを出力する。本リストは、ドッキング繰り返しごとに、DXPF-WF サーバー上に移動された ADV 出力に基づいて、DXPF-WF サーバーを用いて生成する。
- 予測モデル構築用 DS データ（DS-AI 入力）：初期ドッキングおよび各繰り返しドッキングの

結果得られたDSおよび計算時間のリストから、化合物IDとDSを抽出し、これまでドッキングが行われた全ての化合物のDSを統合する。予測モデル構築用DSデータはドッキングごとに、最新のドッキング結果を追加する形で更新される。更新されたDSデータの最終版はAIドッキングごとに作成されるデータベースに登録する。AIドッキング終了後は、AIドッキングごとに作成されるDSデータベースを用いて、MM-PBSA計算など次のステップに進む化合物を選定する。

- 予測モデル構築用フィンガープリントデータ（DS-AI入力）：予測したい化合物全てについて、あらかじめ、化合物IDに対応する化合物の2次元構造を表すSMILES表記を抽出し、SMILES構造に基づきRDKit（https://github.com/rdkit）を用いてMorgan fingerprint（以下FP）を生成する。生成したFPはFPS形式（https://chemfp.com/fps_format/）のファイルとして保存され、UIからアップロードされ、使用される。
  - 本業務では、これを次のように修正する：予測したい化合物全てについて、あらかじめ、化合物IDに対応する化合物の2次元構造を表すSMILES表記を抽出し、SMILES構造に基づきRDKitを用いてFPを生成する。生成したFPはデータベースに保存され、必要に応じてSQL等により抽出され、使用される。AIドッキングの各繰り返しでは、追加されたドッキング化合物について、化合物データベースより化合物IDに対応する化合物のFPを抽出し、使用する。
- DS-AI予測モデル構築：ドッキング繰り返しごとに、その時点までに蓄積されたDSデータを従属変数、FPを記述子として、scikit learn gradient boosting regressorを使用して予測モデルを構築する。
  - 現状のプロトタイプではgradient boosting regressorのハイパーパラメータ調整は実施していないが、本業務では、grid search等の手法を用いて、ドッキングごとのDS-AI予測モデル構築の際にハイパーパラメータ調整を行い、予測モデルを作成するように改変する。ハイパーパラメータ調整の詳細については監督員と協議の上決定する。
- DS予測（DS-AI出力）：予測したい化合物のFP情報を記述子としてDS-AI予測モデルによりDSを予測する。化合物のFP情報はアップロードしたFPSファイルから抽出する。
  - 本業務では、予測したい化合物のFP情報は、化合物データベースより化合物IDに対応するFPを抽出するよう改変する。

3.1.2 大量ドッキングシステムプロトタイプ

- 大量ドッキングタスク名 $NAME_{task}$（入力）：実施する大量ドッキングにタスク名をつける。$NAME_{task}$はJN UIよりユーザーが指定する。各タスクは $NAME_{task}$_YYMMDDHHMM（YYMMDDHHMMは年月日時間）というディレクトリー下の適切な場所に入力、出力を格納・保存する。
- ドッキング化合物リスト（ADV入力）：ドッキングする化合物リストにしたがって、富岳計算ノードで順次ドッキングジョブが流れるような自動処理プロセスがプロトタイプとして存在している。現状では、10万化合物程度のドッキングができるようになっている。
  - 本業務では、100万～200万化合物程度のドッキングジョブを自動的に実施できるように改変する。使用するノード数、ドッキングジョブあたりのコア数、1化合物の平均ドッキング処理時間、ドッキング結果のDXPF-WFサーバーへの移動時間、全化合物数

を考慮して、富岳 small job の制限時間（72 時間）内に収まるように全化合物を分割して処理できるようにする。分割処理ごとにドッキング化合物リストを作成して使用する。ドッキングする全化合物数、使用するノード数、１化合物の平均ドッキング処理時間 T_{dock} はユーザーが UI より入力するようにする。T_{dock} の default 値は 1 分とする。使用するノード数は、JN UI で 64 あるいは 128 から選択するようにし、default=128 とする。ドッキングジョブあたりのコア数は 8 で固定する。ドッキングする全化合物数の上限はノード数 64 の場合 100 万化合物、ノード数 128 の場合 200 万化合物とし、それ以上が指定された場合はユーザーにデータ分割を促すメッセージを表示し、処理を中止する。

  - 本業務では、ドッキングする化合物は既にデータベース化されているので、これを 100 万化合物に分割した化合物リストを別途作成しておき、それを利用する。

- タンパク質構造ファイル（ADV 入力）：ユーザーが準備したタンパク質構造の pdbqt 形式ファイル。UI からアップロードして使用される。
- 結合ポケット情報（ADV 入力）：ユーザーが事前に検討し、中心位置座標、直方体の各辺の長さを決めておく。ユーザーが UI から数値で入力。
  - 本業務では、結合ポケット情報について、中心位置座標、直方体の各辺の長さの情報を含むテキストファイルをアップロードするようにする。
- ADV のドッキングポーズ数 N_{pose}（ADV 入力）：ユーザーが UI から数値で入力。
  - 本業務では、N_{pose}＝１で固定する。
- DS および計算時間のリスト（ADV 出力）：ドッキングした全化合物の化合物 ID、DS、ドッキング所要時間、ドッキングポーズ格納場所が記載されたリストを出力する。本リストは、DXPF-WF サーバー上に移動されたドッキング出力ファイルに基づいて、DXPF-WF サーバーを用いて生成される。
  - 本業務では、ドッキングした全化合物を適切なサイズに分割し、化合物 ID、DS、ドッキング所要時間、ドッキングポーズ格納場所が記載されたリストを出力する。本業務では、上記の分割により作成された DS および計算時間のリストについて、これらをまとめるソフトウェアを別途作成する。

4　受託先業務従事者等要件

① 受託先業務従事者は、python あるいは Jupyter notebook（JN）に関する知識を有し、システム構築ができること。

② 受託先業務従事者は、富岳の使用実績があり、富岳のジョブ管理、富岳のファイル転送等に関する知識を有すること。また、富岳利用に関する課題ごとのファイル容量やジョブごとのファイル数などの諸制限にも知識を有し、それらを考慮したシステム構築ができること。

③ 受託先業務従事者は、Linux 上で稼働するリレーショナルデータベース、SQL 等の知識を有し、python, JN あるいは shell programing などによりデータベース操作ができること。

④ 受託先業務従事者は、ドッキングソフトウェアとそのリガンドデータの特徴をよく理解し、並列化処理の経験をもち、監督員とスムースにコミュニケーションがとれること。

⑤ 受託先業務従事者もしくは管理監督者は，ライフサイエンス研究に関して Journal of American Chemical Society, Journal of Molecular Biology レベルの学術雑誌に 3 報以

上掲載された経験を有しており、研究者とスムースなコミュニケーションをとれること。

⑥ 受託先業務従事者もしくは管理監督者は、科学技術に関する研究開発に1年以上従事した経験を有していること。

5 情報保護・管理について

受注者は本件実施において入手した情報管理を適切に行うこと。また、情報保護・管理に関する社内規程が整備されていること。

6 著作権、秘密保持

著作権、秘密保持については『別紙１』のとおりとする。

7 履行場所

作業を適切に実施するために、本仕様書の内容を請け負う企業内等での作業の他、理化学研究所の以下の場所において打ち合わせや監督員の指導を受けて作業を実施すること。

神奈川県横浜市鶴見区末広町 1-7-22

国立研究開発法人理化学研究所　横浜キャンパス　中央研究棟　C118/120/519

※監督員と協議の上、上記以外の場所となることもある。

8 納品と検収

1) 納期

2023 年 10 月 31 日

2) 納品物

本請負で納品すべきものは次の通りとする。

- 実装されたプログラム一式および操作マニュアル（電子ファイル）
- 仕様で定義された内容に関する報告書（電子ファイル）。
- 完了報告書

3) 検収条件

作業完了後、本研究所担当者の立会いのもと、検査を受けること。さらに、上記 2)に記載の納品物の提出とプログラムの動作確認をもって検収とする。

9 その他

1) 理化学研究所の計算資源等の利用

理研の情報セキュリティポリシー群を遵守したうえで、監督員の指導のもと，理化学研究所の計算資源等を利用することができる。

2) 理化学研究所内で業務を実施する際には担当者からの注意事項を遵守すること。

3) 本仕様書に記載なき事項については、当所担当者との協議の上決定すること。

10 監督員・検査員

監督員：

計算科学研究センター HPC/AI 駆動型医薬プラットフォーム部門

AI 創薬連携基盤ユニット 大田　雅照

検査員：

計算科学研究センター HPC/AI 駆動型医薬プラットフォーム部門 部門長

以上

（別紙 1）

著作権及び秘密保持に関する遵守事項

（著作権）

第 1 条 ソフトウェア等の本件成果物（以下、「ソフトウェア」という）の著作権（著作権法第 27 条及び第 28 条の権利を含む。但し、国立研究開発法人理化学研究所（以下、「甲」という）に原始的に帰属するものは除く。）は、本請負契約の検収日に、受注者（以下、「乙」という）より甲に移転する。乙は、甲及び甲が指定する者に対して、著作者人格権 を行使しないものとする。

2 ソフトウェアの作成以前から既に乙又は第三者がその著作権を有するモジュールやプログラム、デザイン等であって、ソフトウェアに含まれるモジュールやプログラム、デザイン等については、その著作権は、乙又は当該第三者に留保されることを確認する。乙は、当該モジュールやプログラム、デザイン等がソフトウェアに含まれている場合には、本請負契約の検収日に、その旨を申し出るものとする。

3 前項の定めに関わらず、請負契約において新たに改変が加えられたモジュールやプログラム、デザイン等については、原著作物の二次的著作物として扱い、二次的著作物についての著作権（著作権法第 27 条及び第 28 条の権利を含む）は、本請負契約の検収日に、乙 から甲に移転する。乙は、甲及び甲が指定する者に対して、著作者人格権及び原著作物に 係る著作権法第 28 条の権利を行使しないものとする。但し、甲は、当該モジュールやプロ グラム、デザイン等に大幅な改変を加える場合には、乙に対してその旨を申し出るものとし、乙の同意を得るものとする。

4 乙は、ソフトウェアに用いた第三者に係る権利について、甲が何ら第三者の著作権その他の権利を侵害せずに自由に利用できるように、適切な権利処理を行うものとする。万一、乙の責に帰する事由により第三者と甲との間に紛争を生じ又はそのおそれがある場合には、乙は自己の責任においてその対応と解決に当たるものとする。

（秘密保持）

第 2 条 乙は、本業務の履行に際して、甲から開示された資料等の秘密情報(以下，「秘密情報」とする)につき秘密を保持し、甲の書面による事前の同意を得た場合を除きこれを第三者に開示または漏洩してはならない。甲は、当該秘密情報を書面で開示する場合には、秘密である旨を表示するものとし、口頭等により開示する場合は開示する際に秘密である旨を明示し、秘密である旨を乙と確認するものとする。但し、次の各号のいずれかに該当する情報については秘密情報から除く。

（1）甲から開示を受ける前に公知であったかまたは開示を受けた後、乙の責によらず公知となった情報

（2）乙が開示を受ける前に保有していたことを証明できる情報

（3）乙が独自に開発したことを証明できる情報

（4）乙が正当な権原を有する第三者から開示を受けた情報

2 乙は、秘密情報を本業務の履行の為にのみ使用し、他の目的に使用しないものとする。

3 乙は、本業務の履行終了に当たり、甲から開示された資料等の処分につき甲の指示に従うものとする。

4 本条第１項の定めに関わらず、乙は、甲が認める範囲において、受注者の指定する者（以下、「丙」という）に秘密情報を開示できるものとする。

5 乙は、前項に基づき丙に開示した情報の取り扱いにつき、本条に基づき自己が負うと同等の義務を丙に課すものとし、その履行に一切の責任を負うものとする。

以　上