



PDBj
Protein Data Bank Japan



第44回日本分子生物学会 フォーラム 2F-02

“生命科学のデータベース活用法”

RDFグラフデータモデルによるwwPDBコアアーカイブの拡張

大阪大学蛋白質研究所
横地政志



Licensed under a Creative Commons 表示4.0国際ライセンス

©2021 日本蛋白質構造データバンク

日本蛋白質構造データバンク (PDBj)

蛋白質構造データバンク（Protein Data Bank、PDB）は生体高分子の構造データを集めた世界で唯一のデータアーカイブであり。日本では大阪大学蛋白質研究所のPDBjがアジア地域のデータ登録受付および処理を行っている。

主な検索機能：

- 日本語または英語のキーワード検索
- SQLクエリを使った詳細な検索
- SQLクエリビルダーを使って、画面上の選択によりSQLクエリを構築



The screenshot shows the PDBj website with a sidebar on the left containing links like Home, Data deposition (OneDep), Download, Standard format, and Quick links. The main content area includes 'About PDBj', 'Find the service you need' with various search filters (PDB, BMRB, EMDB, search, deposition, viewer, education/dictionary, NMR, electron microscopy, secondary structure, sequence, similarity, function prediction, chemical component, structure prediction, binding site, surface structure, 3D structure, genome, RDF, SPARQL, gene, drug, disease), and 'Latest news'.



国際組織「国際蛋白質構造データバンク」（worldwide Protein Data Bank、wwPDB）の下、5つのコアメンバー（RCSB PDB、PDBe、PDBj、BMRB、EMDB）により運営されている。

PDB 50周年記念イベント

1971年10月20日、生体高分子の構造データの単一のアーカイブであるPDBが設立されました。設立当初からPDBはオープンアクセスの文化を取り入れており、各研究コミュニティで利用されています。2021年、wwPDBは年間を通じて数多くの記念イベントを開催しました。

キックオフシンポジウム ↓

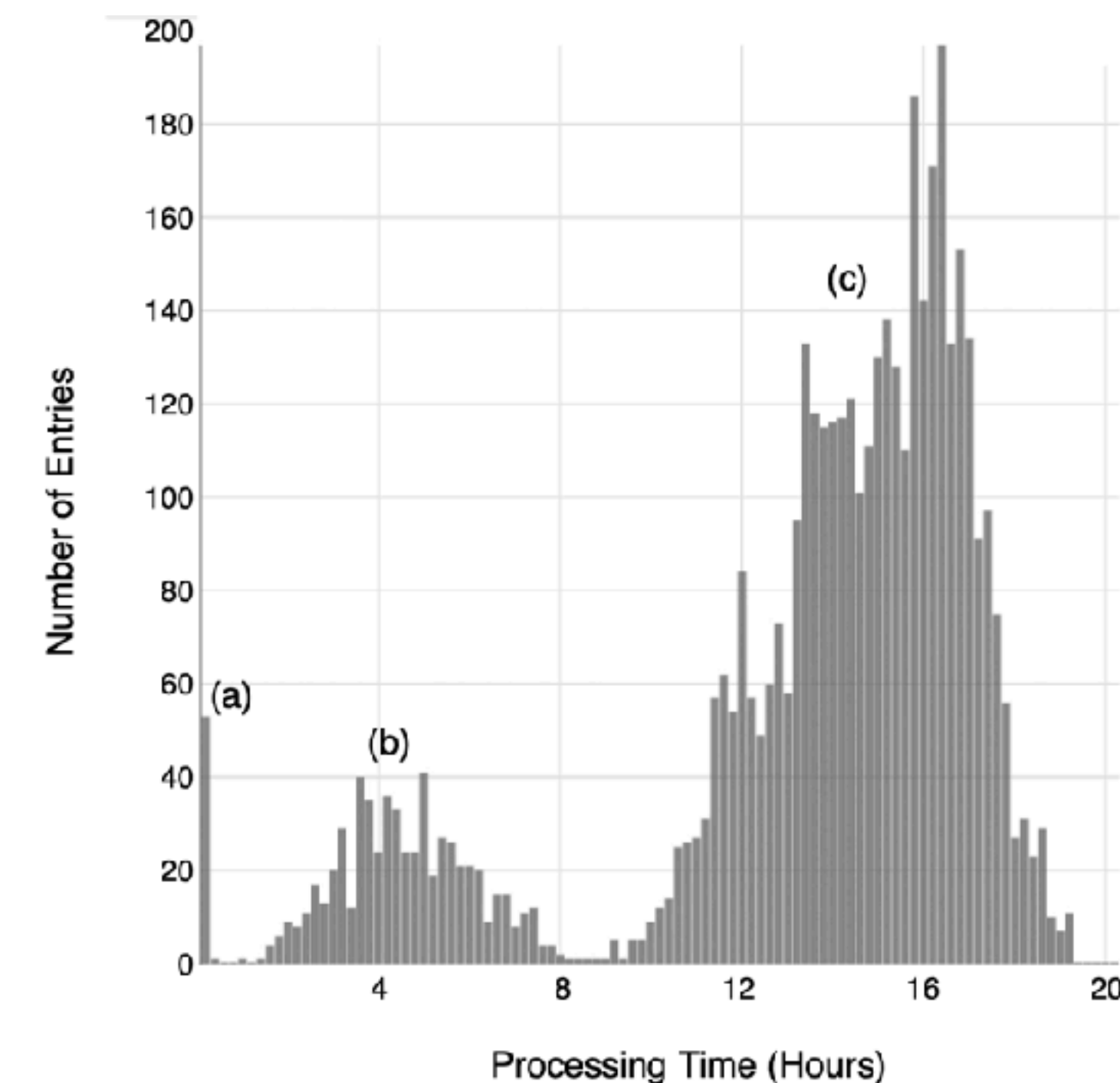
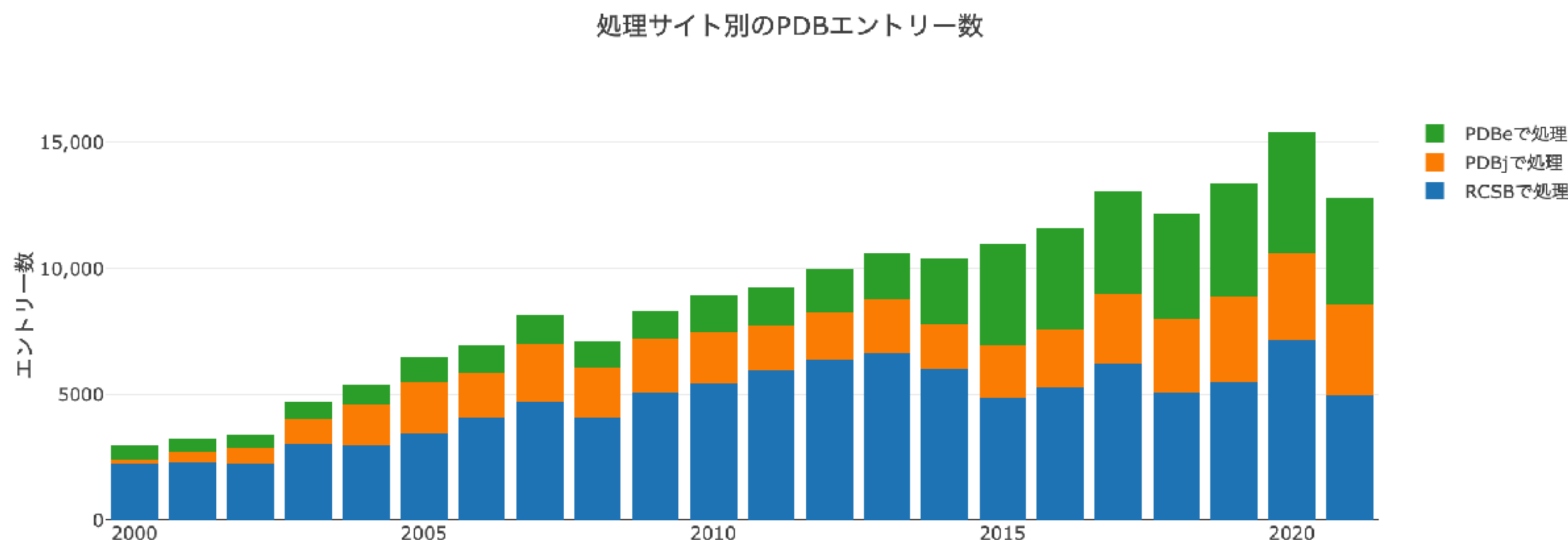
Special symposium celebrating the 50th anniversary of the Protein Data Bank (May 4-5)



**PDBj主催のシンポジウム →
日本生物物理学会サテライト
シンポジウム (11月24日)**



近年のPDBエントリーの登録状況



昨年は年間処理数が過去最高を記録した。本年も引き続き多くの登録が寄せられている。
現在、約18万5千件のPDBエントリーの半分は、過去8年以内に登録されたものである。
登録処理従事者（PDB Biocurator）に大きな負担がかかっている。

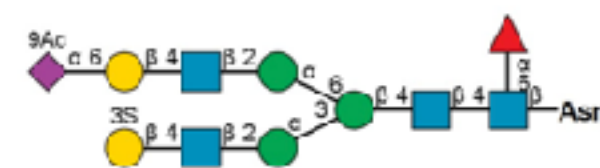
最近の活動

• 増大する構造データへの対応

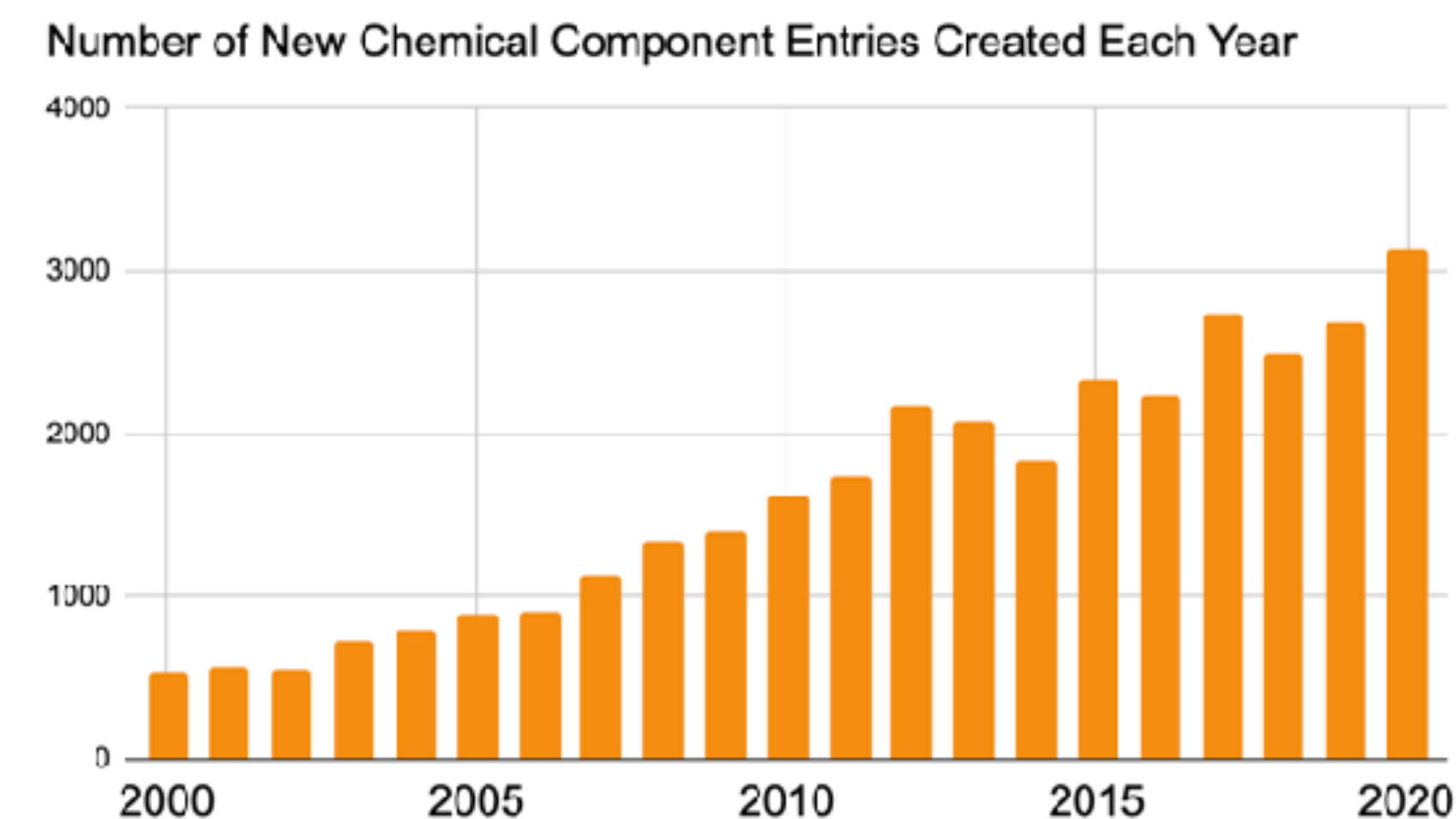
- アノテーションの効率化 (OneDep)
- CCD ID、PDB IDの拡張
- 登録PIによる公開済みエントリーの原子座標改訂 (バージョニング)

• 発見可能性、再利用可能性の向上

- 糖鎖分子の表現の改良 (分岐表現、SNFG画像)
- wwPDB検証レポート (低分子リガンド)
- wwPDB検証レポート (NMR制限情報の評価)



従来、PDB IDは4文字、化合物辞書 (Chemical Component Dictionary) のIDは3文字の英数字で識別されてきました。現在の登録件数の推移を想定すると、CCD IDは2~3年で枯渇することが予想されています。そのためwwPDBではPDB ID, CCD IDをそれぞれ8文字と4文字へ拡張する準備を進めています。



従来の4文字表記のPDB IDは、先頭に「PDB_0000」を付けることで拡張されます。なお、大文字小文字は区別しません。

例：PDB_00001ABC

最近の活動

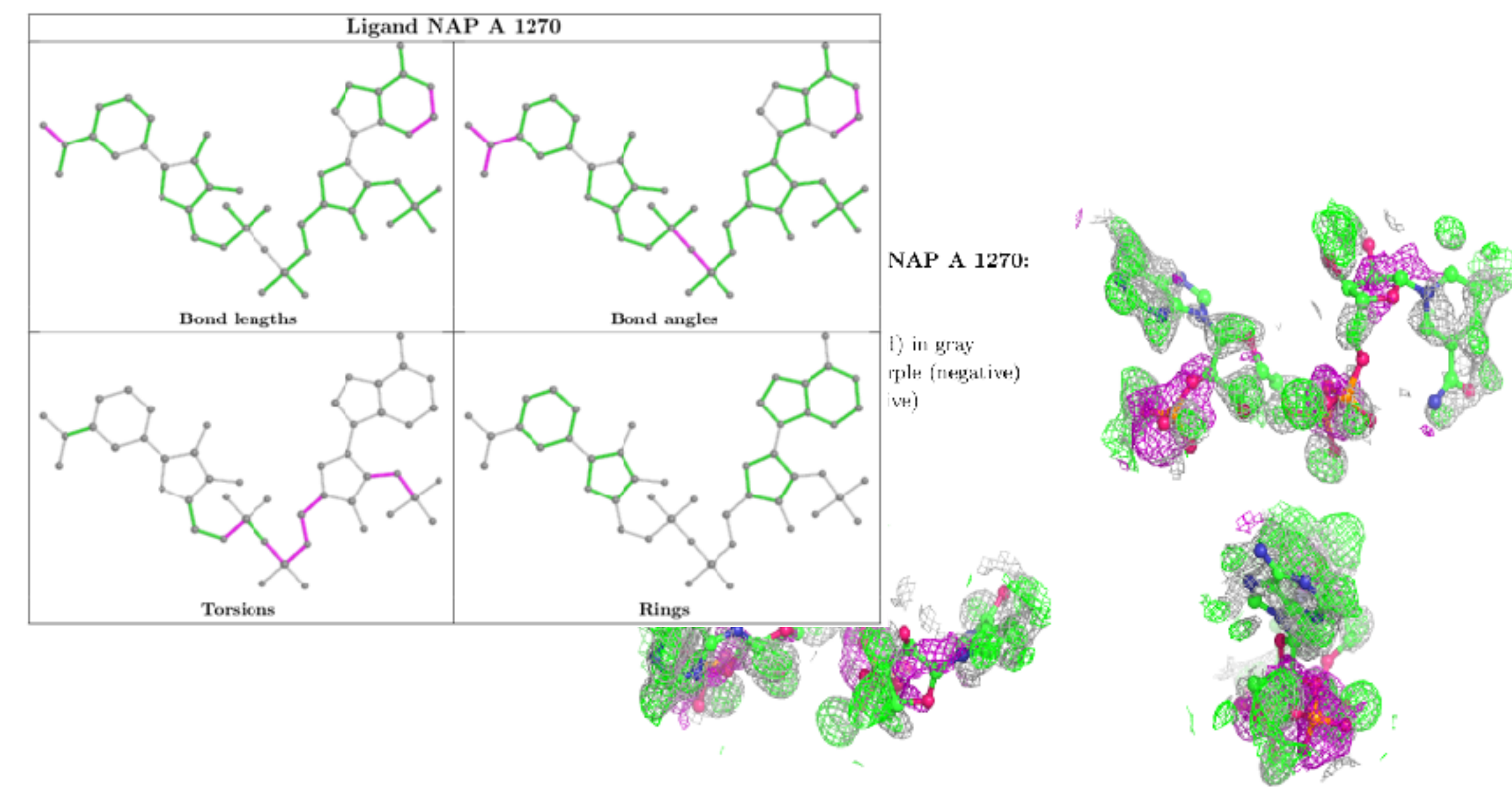
• 増大する構造データへの対応

- アノテーションの効率化 (OneDep)
- CCD ID、PDB IDの拡張
- 登録PIによる公開済みエントリーの原子座標改訂 (バージョニング)

• 発見可能性、再利用可能性の向上

- 糖鎖分子の表現の改良 (分岐表現、SNFG画像)
- [wwPDB検証レポート \(低分子リガンド\)](#)
- [wwPDB検証レポート \(NMR制限情報の評価\)](#)

LOI (Ligand of Interest) を指定したリガンド、幾何学的な構造検証に違反が発生したリガンドに関して、検証情報と周辺の電子密度マップを表す画像をwwPDBの検証レポート内に表示する。



最近の活動

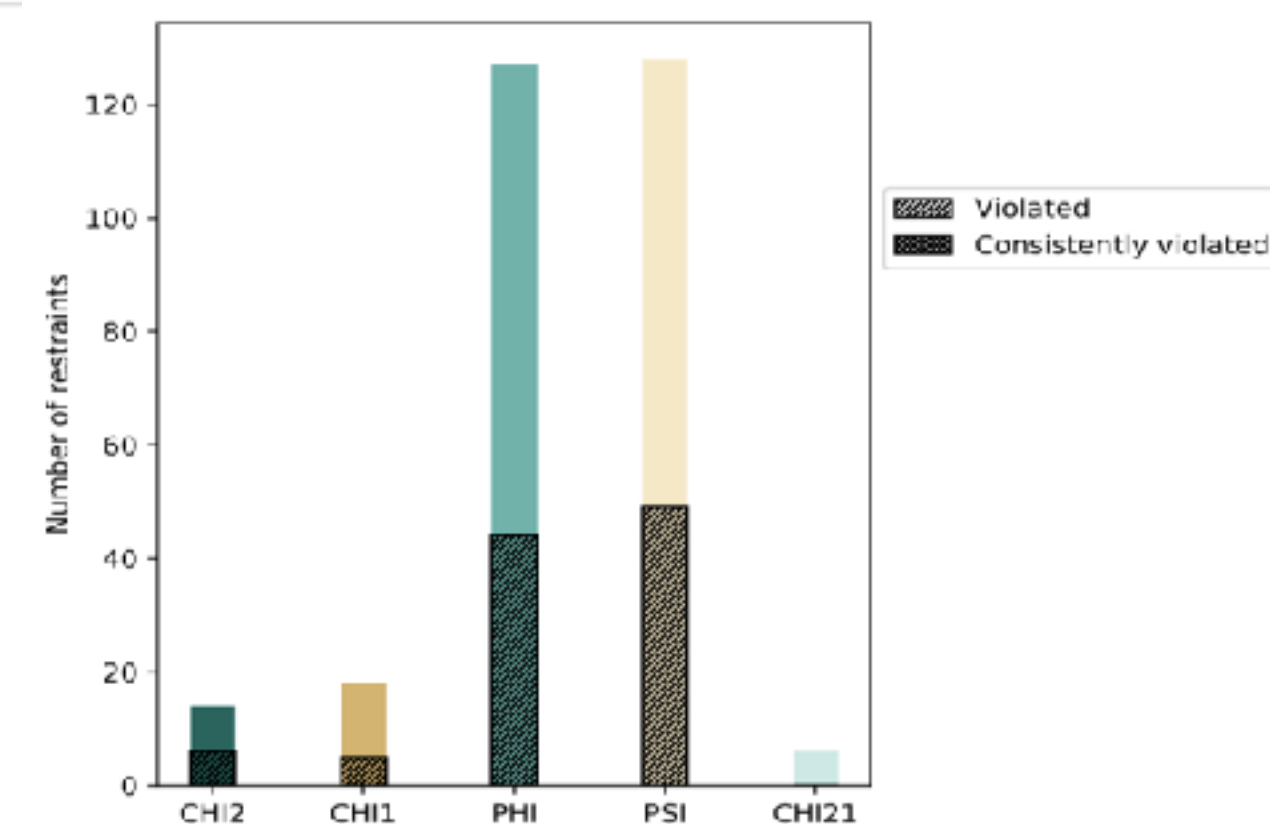
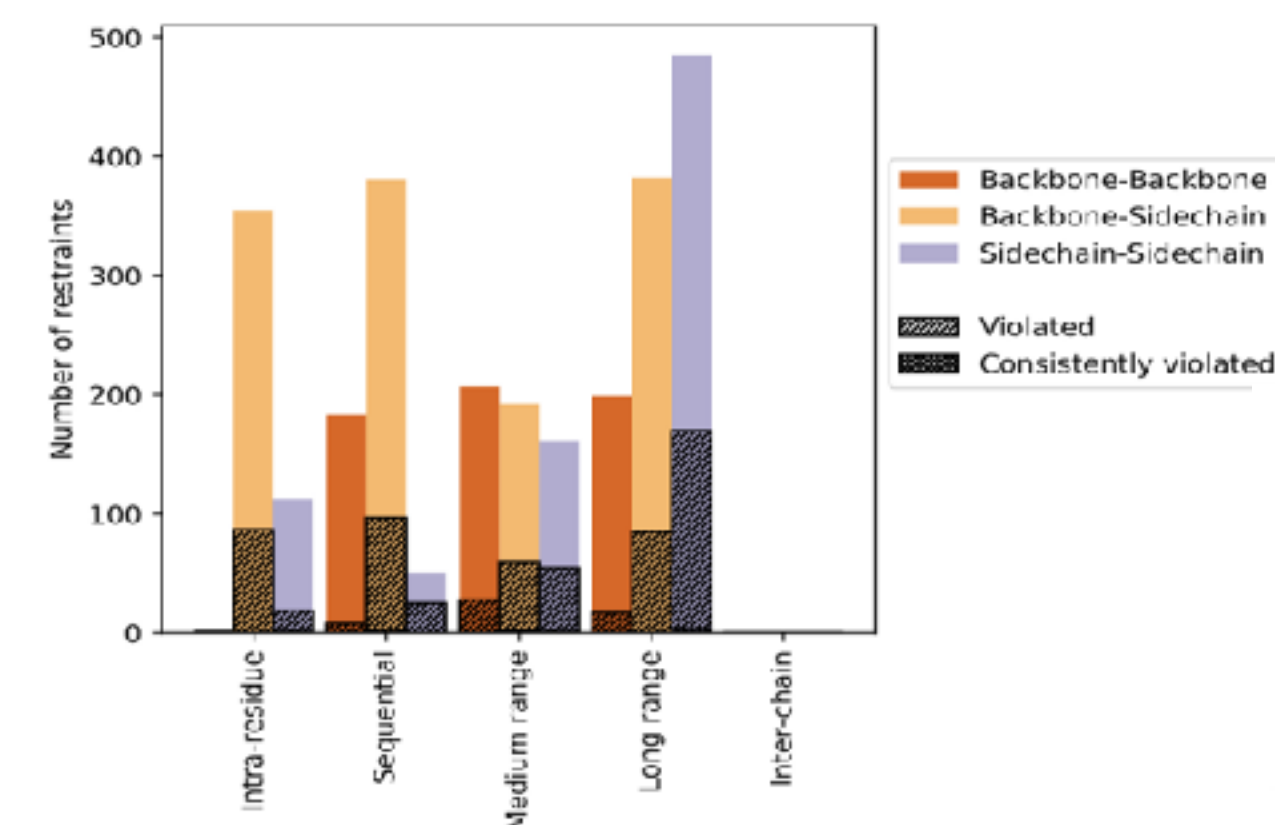
増大する構造データへの対応

- アノテーションの効率化 (OneDep)
- CCD ID、PDB IDの拡張
- 登録PIによる公開済みエントリーの原子座標改訂 (バージョニング)

発見可能性、再利用可能性の向上

- 糖鎖分子の表現の改良 (分岐表現、SNFG画像)
- wwPDB検証レポート (低分子リガンド)
- [wwPDB検証レポート \(NMR制限情報の評価\)](#)

NMR制限情報の検証結果をwwPDB検証レポートに追加。各グラフの網掛け、黒塗り部分が、制限情報と構造モデルの間の違反件数を表している。



COVID-19/SARS-CoV-2に対するPDBの取り組み



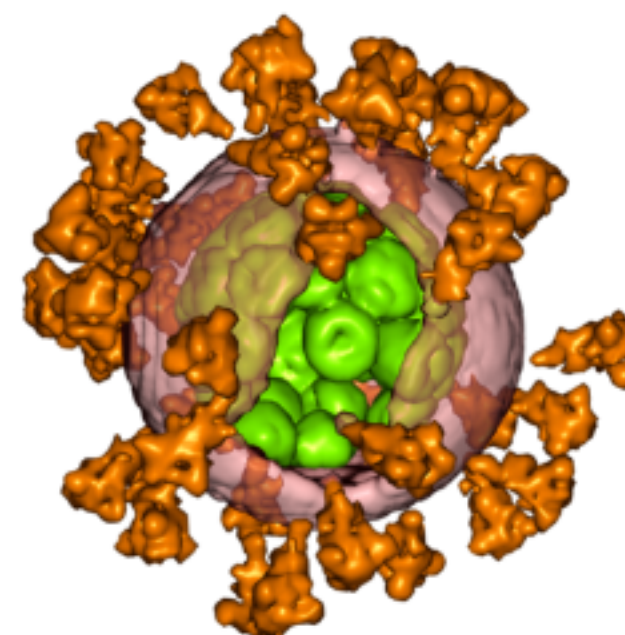
The screenshot shows the PDBj website interface. The top header includes the PDBj logo, navigation links in multiple languages (English, 日本語, 简体中文, 繁體中文, 한국어), and a search bar. The main content area is titled '新型コロナウイルスの構造情報' (Structural Information of Novel Coronavirus). It features a large image of the virus structure and a list of recent entries related to COVID-19, including proteases and RNA polymerase. The left sidebar contains navigation links such as 'ホーム', 'データ登録 (OneDep)', 'ダウンロード', '標準フォーマット', and 'クイックリンク'. The right sidebar displays search results and links to NMR, EM, and in Silico models.

- 特集ページの開設
- 2021年12月現在、SARS-CoV-2に関連するPDBエントリーは1640件。全体の0.9%に相当。
- On-HOLDステータスの登録者に、即時公開を提案

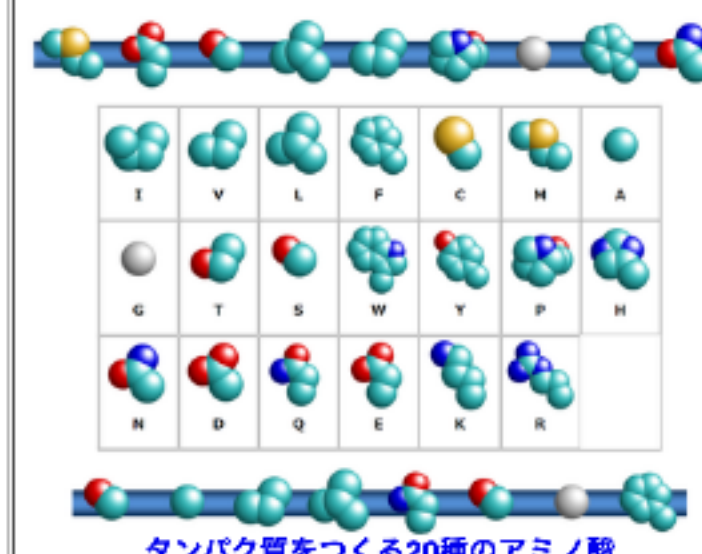
PDBjの教育分野への奉仕活動

- 「今月の分子」 ページ
- 「サイエンス・フェスタ」に出展
- 「PDBj入門」 サイト

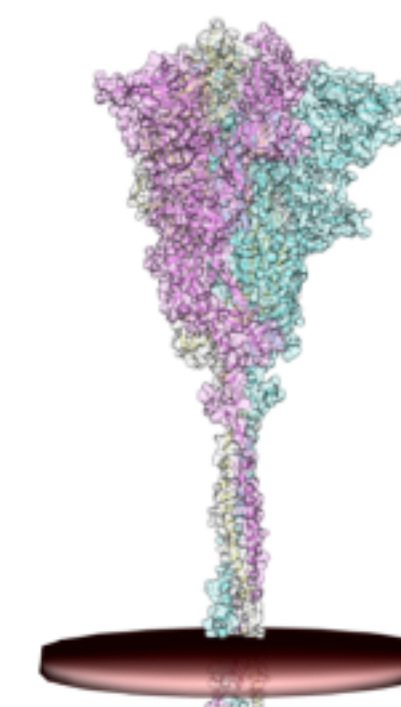
PDBj入門 コンテンツ抜粋



ウイルス粒子の形



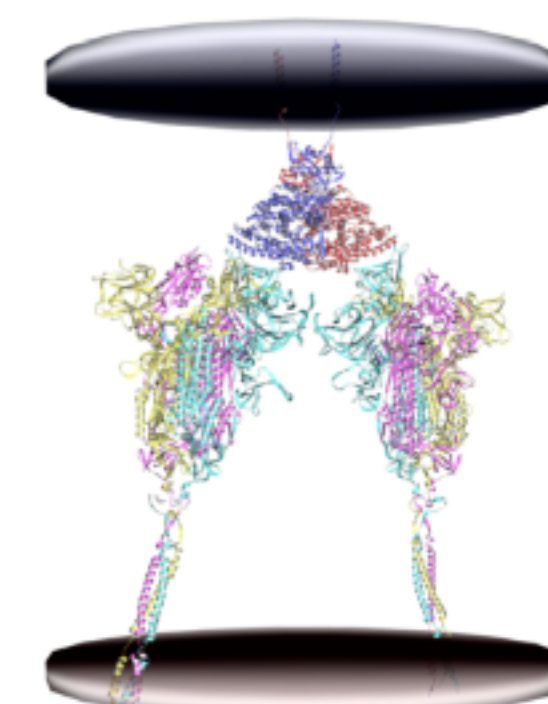
タンパク質をつくる20種のアミノ酸



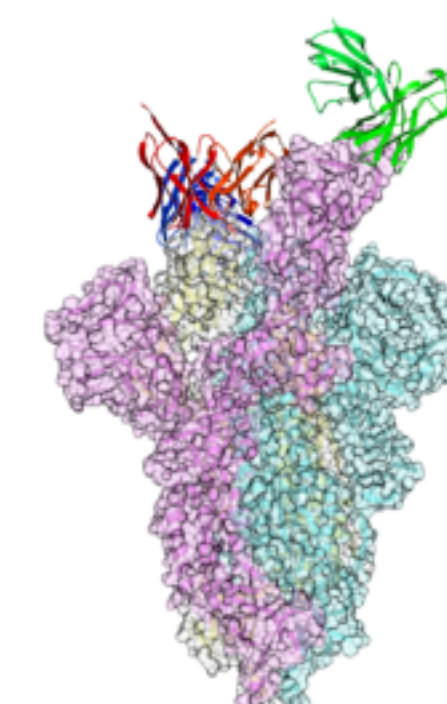
スパイクタンパク質の形



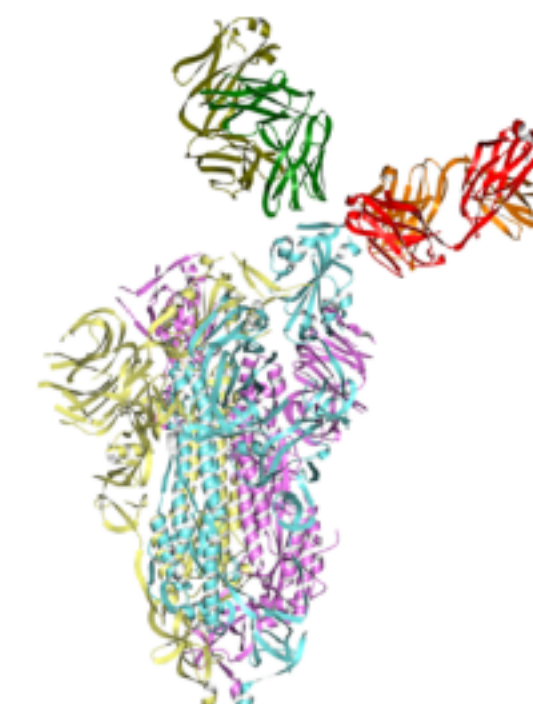
スパイクは閉じたり開いたりします



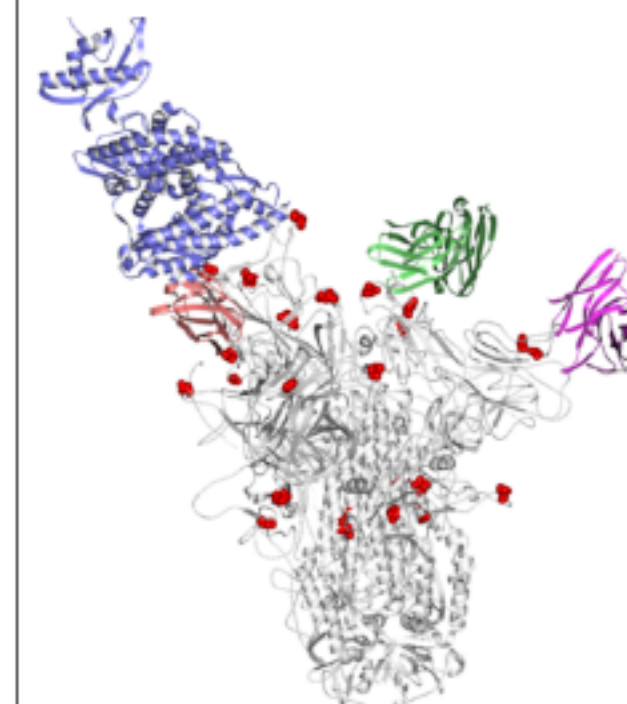
スパイクはヒトの受容体と結合します



ワクチンを打って作られる抗体はスパイクに結合します



抗体カクテル療法



変異株はスパイクのアミノ酸が少しちがいます



α、δ、ο変異株に対応



写真は2019年開催時のもの

Licensed under a Creative Commons 表示4.0国際ライセンス

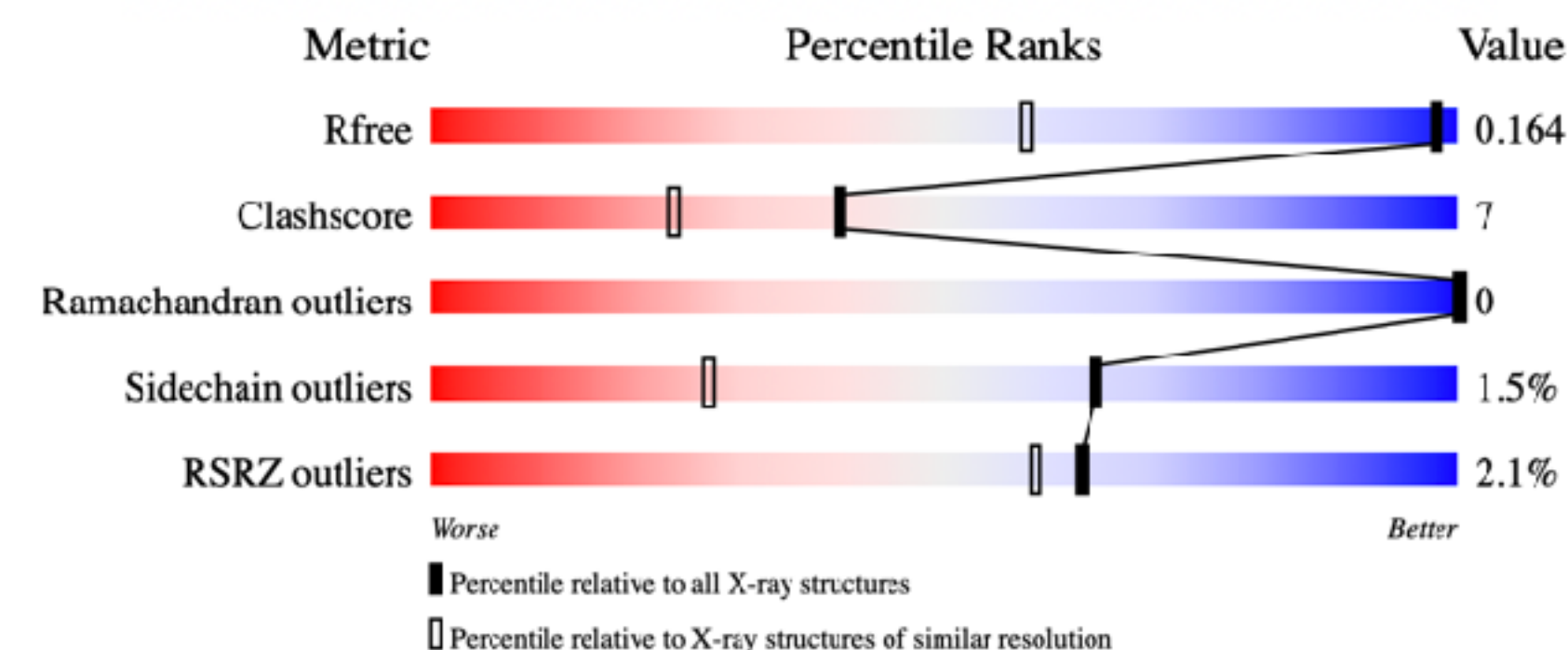
©2021 日本蛋白質構造データバンク

データ共有と利用促進に向けて

- 事前知識や実験データで構造モデルを評価



Full wwPDB X-ray Structure Validation Report [i](#)



- wwPDB検証レポートの拡張

- 低分子リガンド

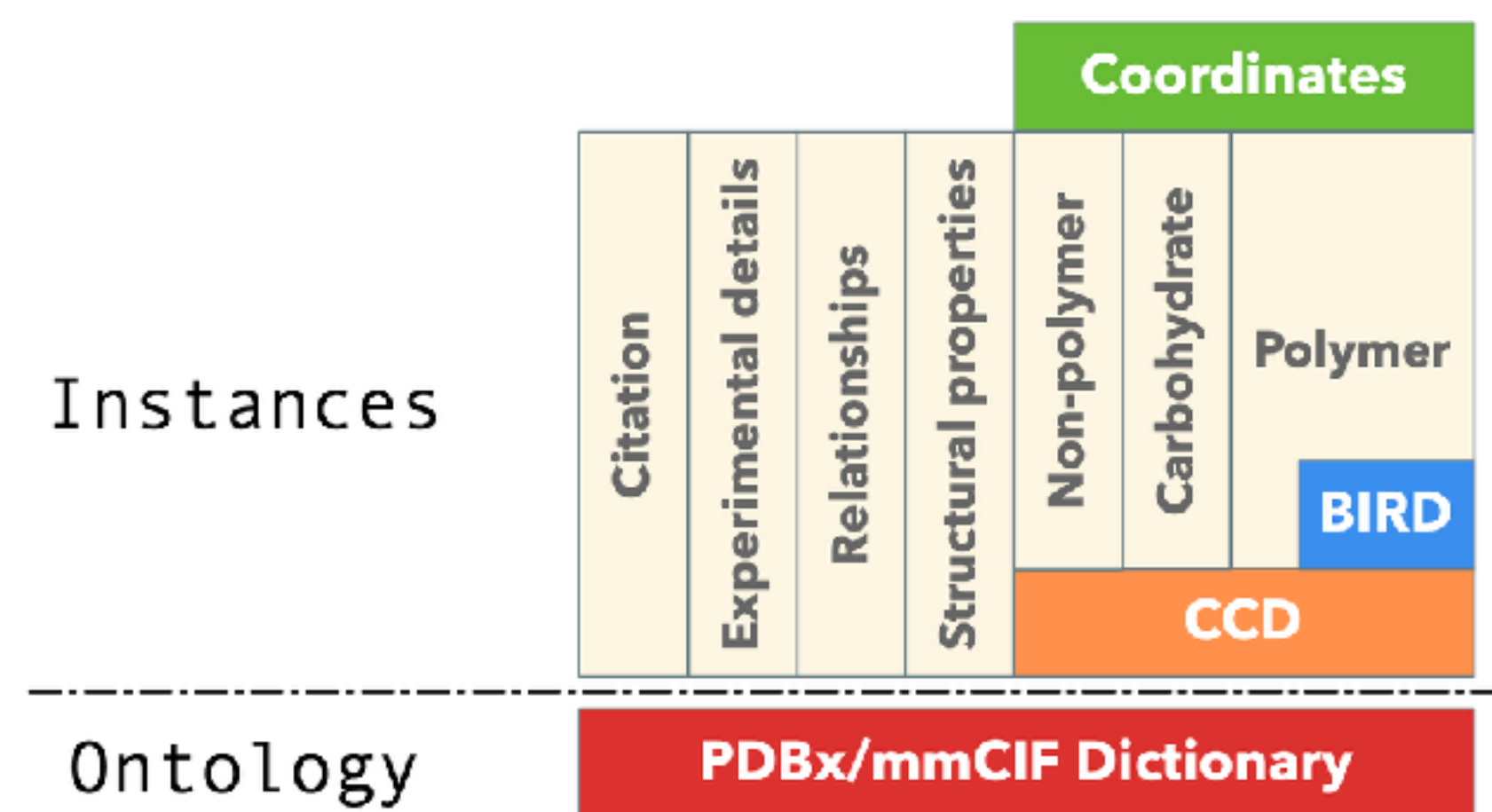
- EMマップ

- NMR制限情報

- 分子カテゴリーの適正な記述

- 非リボソームペプチドの記述の改良 (BIRD)

- 糖鎖分子の記述の改良 (Branched entity)

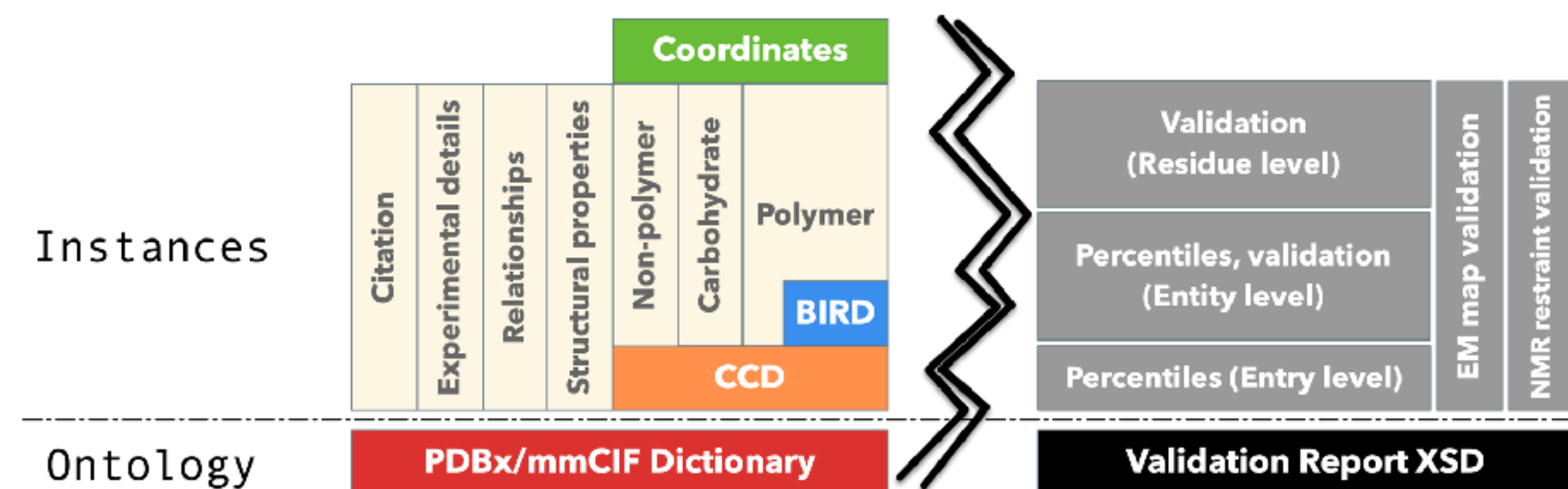


PDBコアアーカイブの模式図

PDBコアアーカイブの内部と外部の両方で連携を円滑化

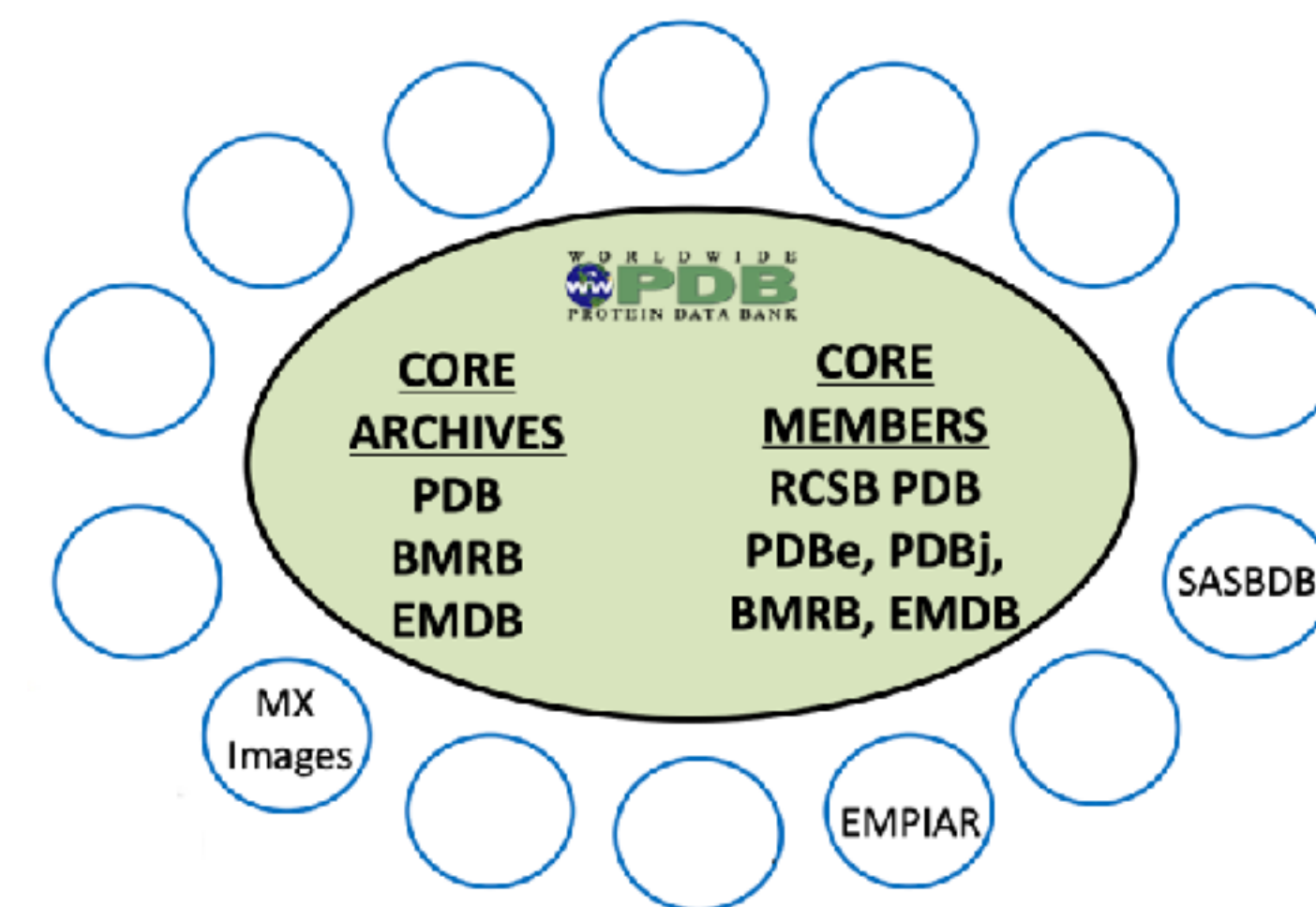
内部連携の課題

- PDBx/mmCIFと検証レポートのオントロジーの不一致



外部連携の課題

- PDBx/mmCIFの外部参照の基本姿勢
 - 外部参照識別子の永続性（URLの禁止）
 - wwPDBコアメンバーとPDB Federated Data Resources
 - 遺伝子オントロジーから検索に支障

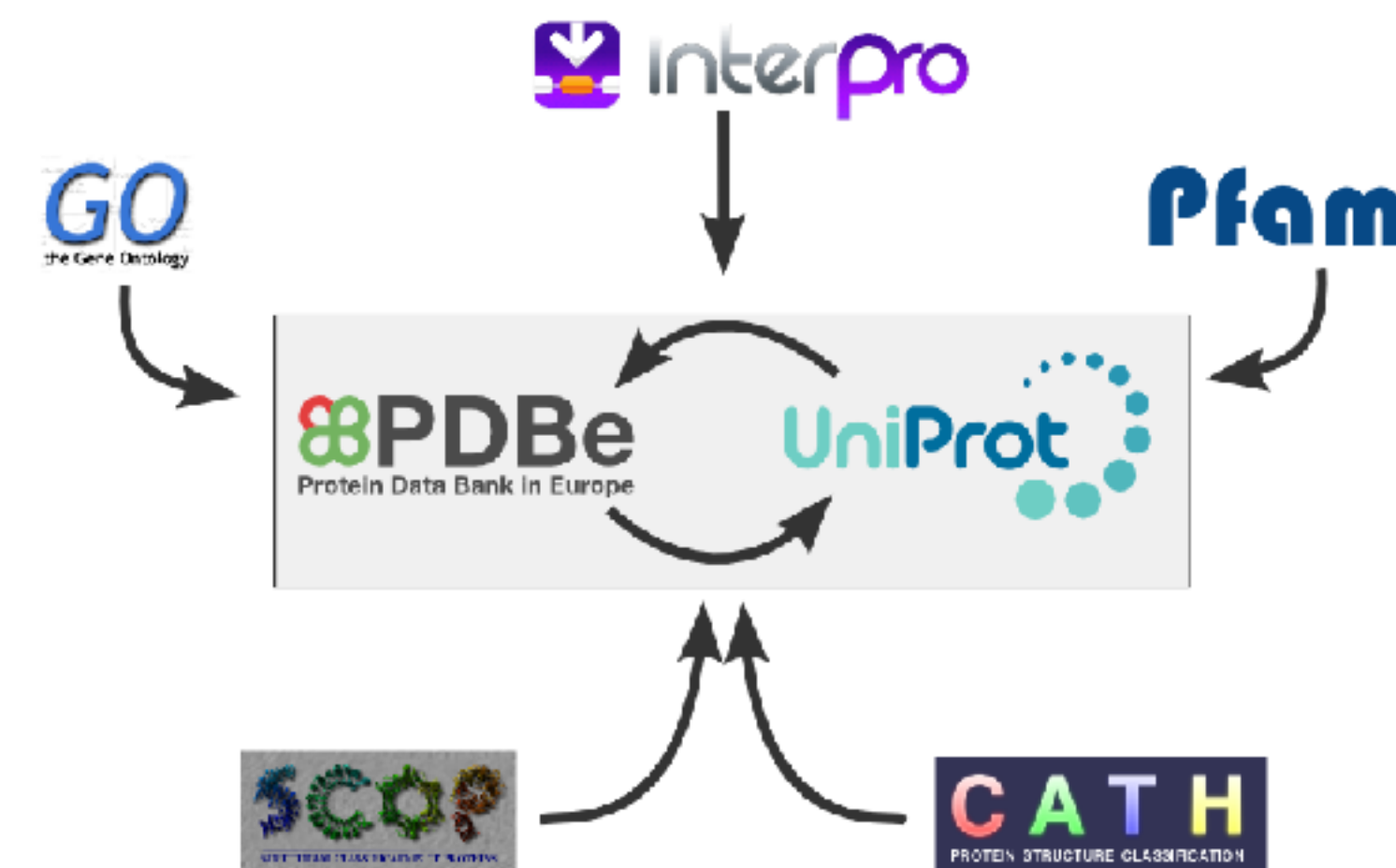


- 提案する解決策** — PDBx/mmCIFをマスターにしつつ、他のリソースと容易に繋がるWeb標準技術であるRDFを採用

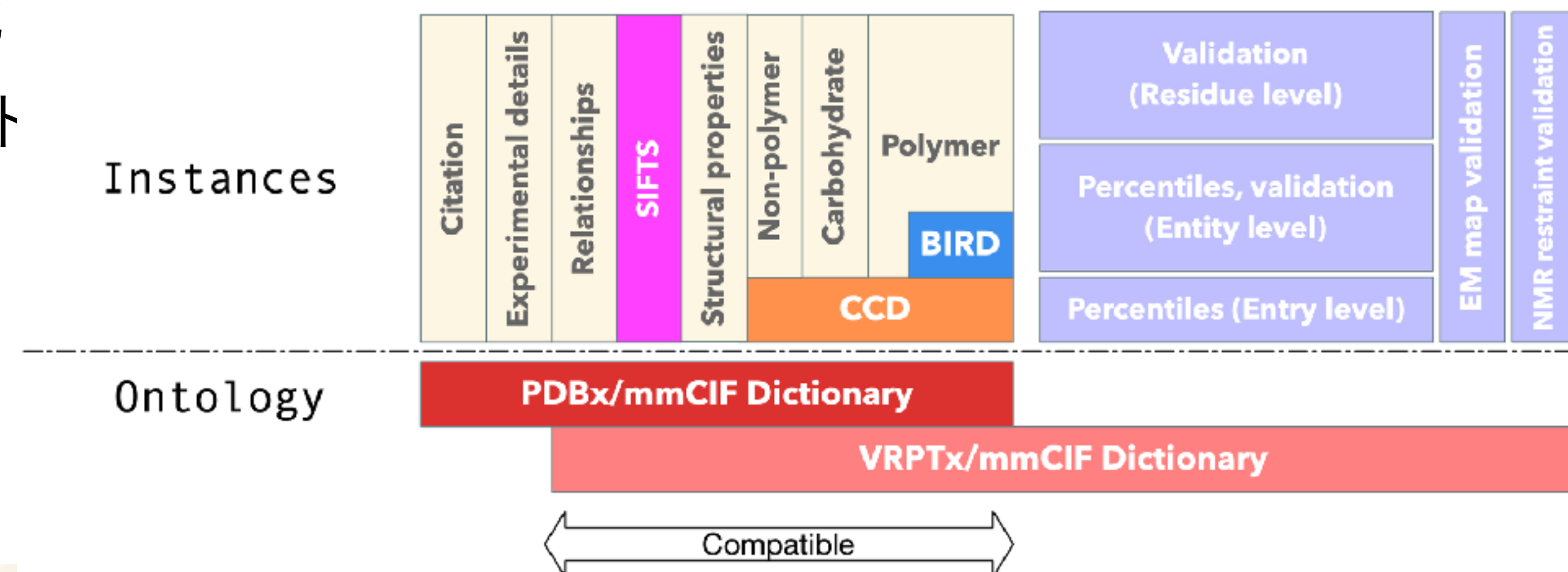
全てのPDBコアアーカイブのRDF化

SIFTS: Structure Integration with Function, Taxonomy and Sequence

SIFTSに含まれるデータリソース



検証レポートのPDBx/mmCIFによる再編成



発見可能性、相互運用性の向上

- エンティティアノテーションの強化
(PDBx/mmCIFのSIFTS関連データアイテム追加により実現)

- wwPDB検証レポートの再編成とRDF化
(PDBx/mmCIFを拡張した検証レポートの辞書の開発により実現)

- ペプチド様化合物辞書 (BIRD)のRDF化

RDFで外部リソースと繋がったPDBコアアーカイブ

- **PDB/RDF及びVRPT/RDF**

- BMRB, EMDB, SASBDB, DOI, ORCiD, PubMed, Taxonomy, UniProt, GenBank, EMBL, PIR, RefSeq, Norine, Enzyme, GO, InterPro, Pfam, CATH, SCOP, ENSEMBL, GlycoInfo

- **BIRD/RDF**

- Taxonomy, Enzyme, UniProt, Norine, KEGG, NADB, CCDC, PubChem, ChemSpider, CAS, ChEBI, MeSH, ChemIDplus, ChEMBL, BindingDB, DrugBank, ChemDB

- **chem_comp/RDF**

SARS-CoV-2のメインプロテアーゼに結合している化合物の RSRが0.1以下の構造を検索

```

PREFIX vrpto: <http://rdf.wwpdb.org/schema/pdbx-validation-v4.owl#>
SELECT DISTINCT ?pdb_id ?entity_id ?asym_id ?entity_type ?entity_src ?entity_desc ?fw
FROM <http://rdf.wwpdb.org/vrpto>
WHERE
{
  ?s vrpto:pdbx_sifts_unp_segments.unp_acc "P0DTD1" ; ←新型コロナウイルスのプロテアーゼ
  ... 省略
  vrpto:of_datablock ?datablock .
  BIND(SUBSTR(STR(?datablock), 27, 4) AS ?pdb_id) ←PDB IDの特定

  BIND(IRI(CONCAT(?datablock, "/entityCategory")) AS ?entity_category)
  ?entity_category vrpto:has_entity ?entity .
  ?entity vrpto:entity.id ?entity_id ;
    vrpto:entity.pdbx_description ?entity_desc ;
    vrpto:entity.src_method ?entity_src ;
    vrpto:entity.type ?entity_type ;
    vrpto:entity.formula_weight ?fw .
  ... 省略 (水、イオンなどの低分子化合物、酵素、RNA鎖の除去)

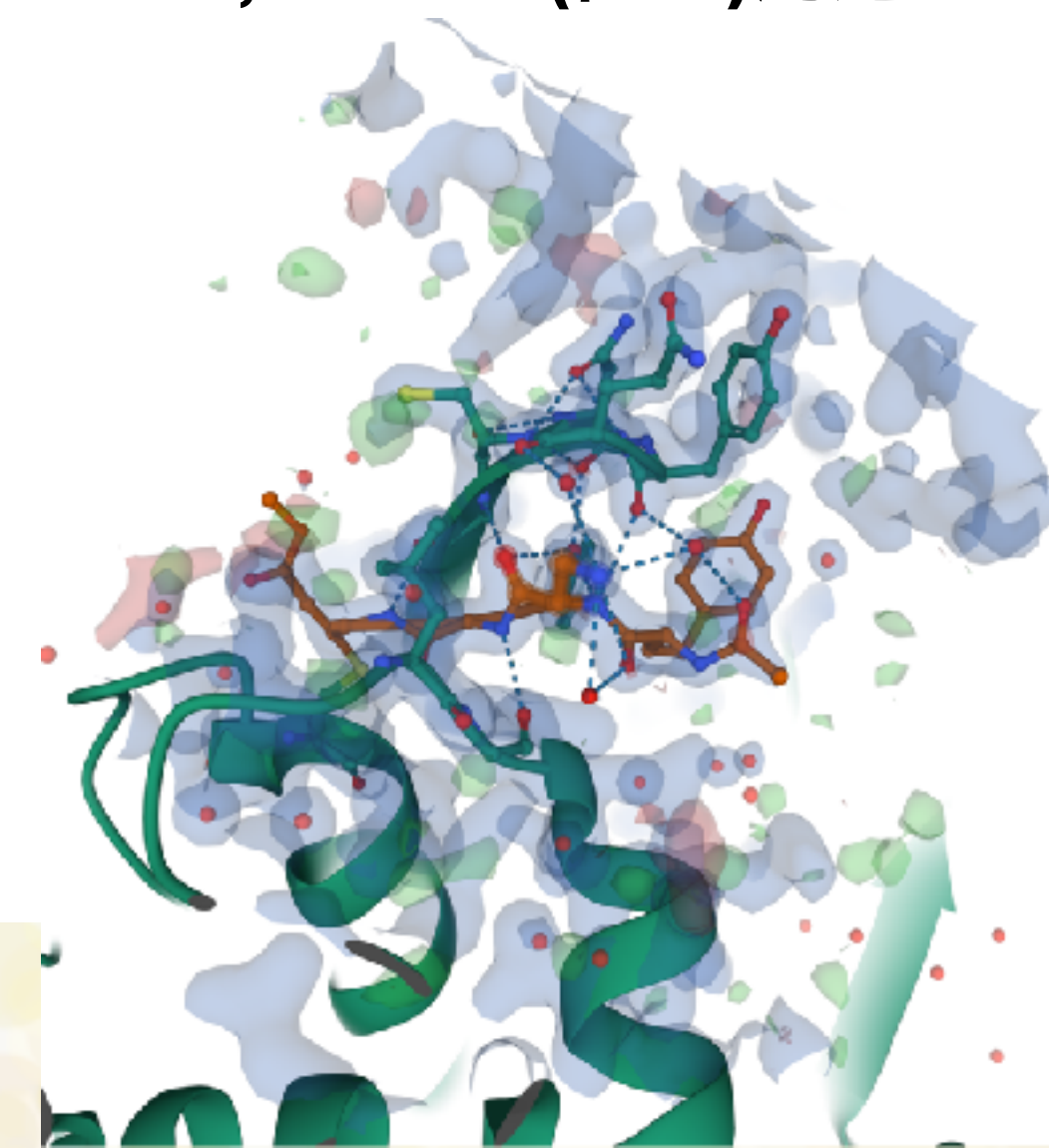
  BIND(IRI(CONCAT(?datablock, "/pdbx_dcc_mapCategory")) AS ?dcc_map_category)
  ?dcc_map_category vrpto:has_pdbx_dcc_map ?dcc_map .
  FILTER EXISTS {
    SELECT ?rsr WHERE {
      ?dcc_map vrpto:pdbx_dcc_map.auth_asym_id ?asym_id ;
        vrpto:pdbx_dcc_map.RSR ?rsr .
      FILTER(xsd:float(?rsr) <= 0.1) ←化合物のRSRが0.1以下の存在でフィルター
    } LIMIT 1
  }

```

SPARQL検索結果

pdb_id	entity_id	entity_type	entity_src	entity_desc	fw	
6WX4	2	I	polymer	syn	VIR251	479.527
6XA4	2	B	polymer	syn	inhibitor UAW241	387.581
6XFN	2	B	polymer	syn	UAW243	470.561
6XBH	2	C	polymer	syn	inhibitor UAW247	437.488
6XCH	2	C	polymer	syn	Leupeptin	429.578
7BQY	2	C	polymer	syn	IR,2Z)-4-(BENZYLOXY)-4-OXO-1-[[[	680.791
7CUT	2	B	polymer	syn	Z-VAD(OMe)-FMK	485.933
7M2P	2	B	polymer	syn	Inhibitor 18 in bound form	554.677
7MGS	2	B	polymer	syn	ER-ALA-VAL-LEU-GLN-SER-GLY-PH	965.085
7LBN	2	D	polymer	syn	Calpain I Inhibitor	385.542
7MGR	2	B	polymer	syn	A-VAL-LYS-LEU-GLN-ASN-ASN-GL	1029.167
7NEV	2	B	polymer	syn	LEUPEPTIN	429.578

PDB ID: 6WX4, VIR251(橙色)周辺の2Fo-Fcマップ



2021年に合意されたwwPDB定款

wwPDB 2021 Charter抜粋

2.2 File Formats

PDB Core Archive files containing atomic and multi-scale coordinates and metadata are distributed in PDBx/mmCIF as the primary format (defined in <https://mmcif.wwpdb.org/>).

Identical information is also distributed in secondary formats, including PDBML (<ftp://ftp.wwpdb.org/>) and **RDF** (<https://rdf.wwpdb.org/pdb>). The contents of these files are defined in **PDBML schema** (<https://pdbml.wwpdb.org/>), and **RDF schema** (<https://rdf.wwpdb.org/pdb>), respectively.

A subset of the **PDB Core Archive** is also distributed using the legacy PDB format, where possible.

(For the avoidance of doubt, all secondary format files, including so-called legacy PDB, PDBML, or **RDF** are products of the archive and not part of the **PDB Core Archive**.)

Quaternary structure files (characterizing the assembly observed in the experiment) are distributed in the PDBx/mmCIF or PDB legacy format.

<https://www.wwpdb.org/about/agreement>

- EMDBがコアメンバーとして加入
- PDB China / Indiaが準メンバーとして参加
- PDBML(XML)と並びRDFアーカイブが
secondaryフォーマットとする規定

今後の取り組み

- 各PDBサイトが生成している派生アーカイブ（例：PDBjのRDFアーカイブ）の共有化
- RDFで繋がったデータの応用（クエリ文構築支援）
- wwPDB検証レポートの改善を継続
 - EDマップと座標モデルの異常検出は、RSRZからRSCCZに変更
 - PDBに登録された過去のNMR実験データの標準化、再評価

データ基盤整備の目的

- データ利用可能性ステートメント(Data Availability Statement)に合致する標準を整備して、データの再利用、再分析を容易にすることで幅広い研究分野の発展につなげていくこと
- 実験データとよく一致するという意味で高品質の構造データの提供を通じて、データ駆動科学の促進（例えば、CASPコンテスト）
- 教育分野へのアウトリーチの取り組みを続け、一般社会の理解を得ること

One wwPDB team



A great respect for Dr. John D. Westbrook (1957-2021)

謝辞

- 大阪大学蛋白質研究所 (PDBj - DataOut)
 - 栗栖源嗣
 - Bekker Gert-Jan
 - 川端猛
 - 鈴木博文
 - 工藤高裕
 - 池川恭代
 - 岩田武史
- ライフサイエンス統合データベースセンター (DBCCLS)
 - 畠中秀樹
- 野口研究所 (GlyTouCan)
 - 山田一作
 - 松原正陽

質問、提案は、”ホーム/お問い合わせ” へ。

