BINDS-PDBj講習会・蛋白研セミナー「データベース登録ノウハウ講習会」

創薬等先端技術支援基盤プラットフォーム
Basis for Supporting Innovative Drug Discovery and Life Science Research

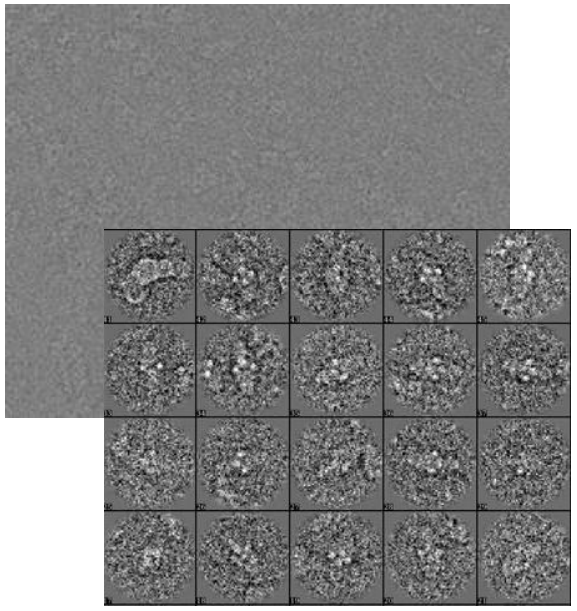
電顕2D画像アーカイブ　EMPIAR

川端　猛
大阪大学・蛋白質研究所・特任（招へい）准教授

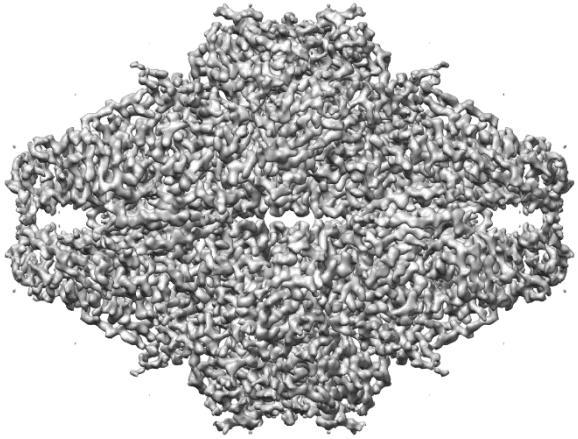
kawabata@protein.osaka-u.ac.jp

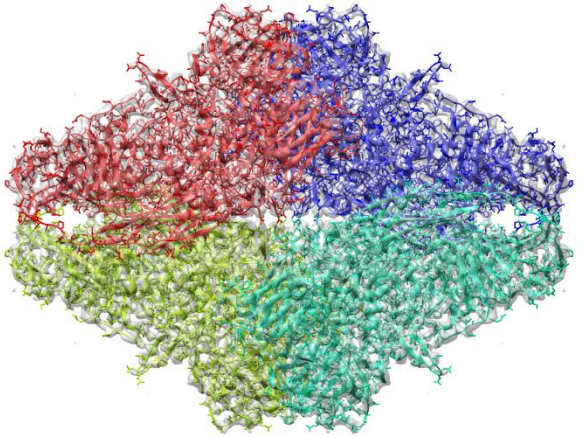
2021年1月20日（水）16:15-16:30
大阪大学　蛋白質研究所　本館2F大講義室　Zoomウェビナー開催

Data processing for EM Single Particle Analysis



Beta-galactosidase





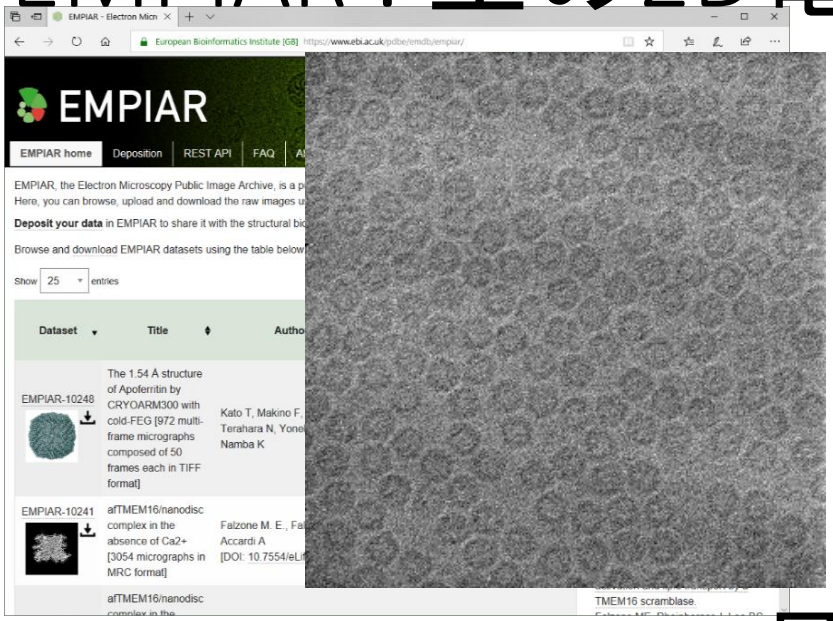
3D reconstruction → Atomic modeling

2D images	3D image	Atomic model
509 DM4 movies (7420 x 7676 x 38; 32 bit)	340 x 340 x 340 voxels	32878 atoms
EMPIAR-10013	emd_5995.map	pdb3j7h.ent
402.2GB	150 MB 3.2Å	2.6 MB
EMPIAR	**EMDB**	**wwPDB**

EMPIAR-PDBj

PDBj

EMPIAR：生の2D電顕画像のアーカイブ



Ardan Patwardhan (EBI)
によって創設

目的

1) 電顕の3Dマップの検証・再解析

2) 新しい画像解析手法・ソフトの開発促進

3) 教育・トレーニング

データサイズは膨大 (**713.3 TB**, 479エントリ) [2021/01/14]

円滑な配布のためにはミラーサイトが必要 (**EMPIAR-PDBj**).

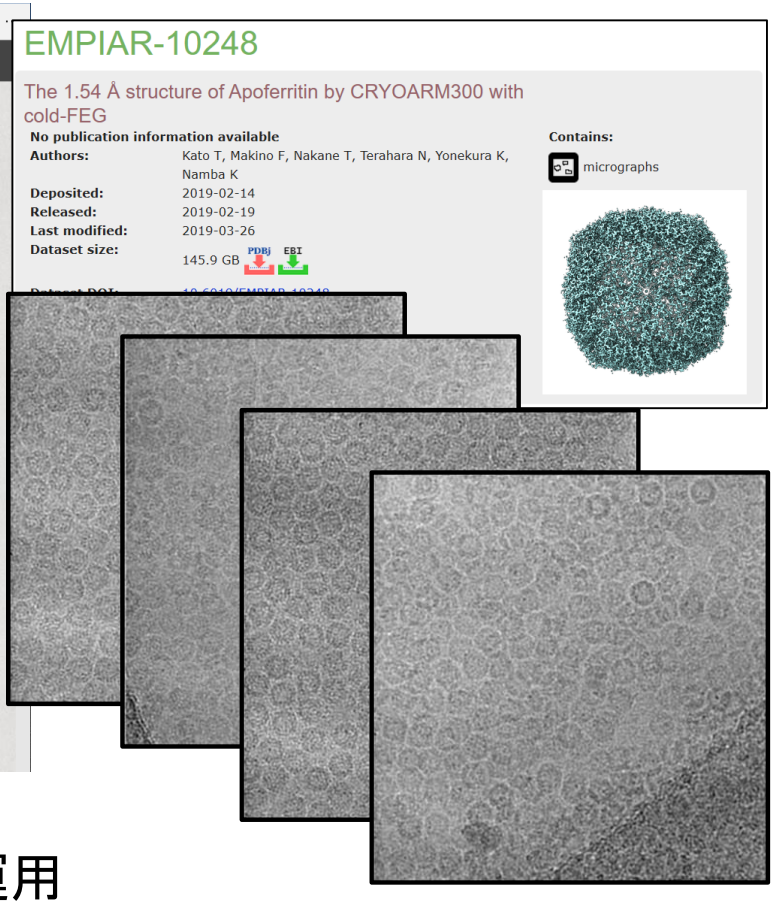
ミラーサイトEMPIAR-PDBj の運営開始

EMPIAR:　UKにある電顕の2D画像のデータベース

目的:　1)3Dマップの検証・再解析　2)ソフトの開発促進　3)教育

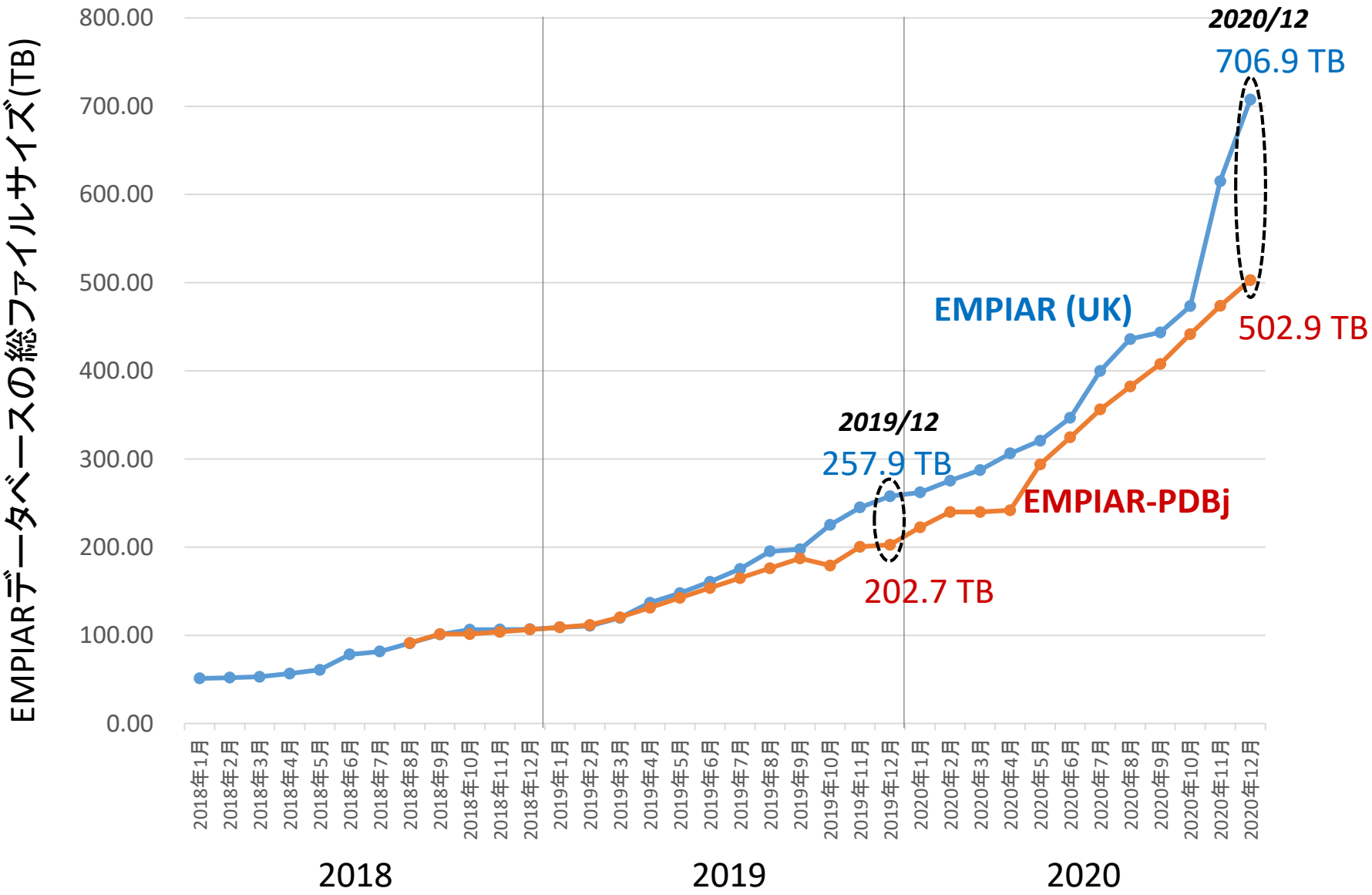
→サイズが大きいのでミラーサイトが必要

2021/01/14 現在
データエントリ数:479
全データサイズ : **約713.3 TB**

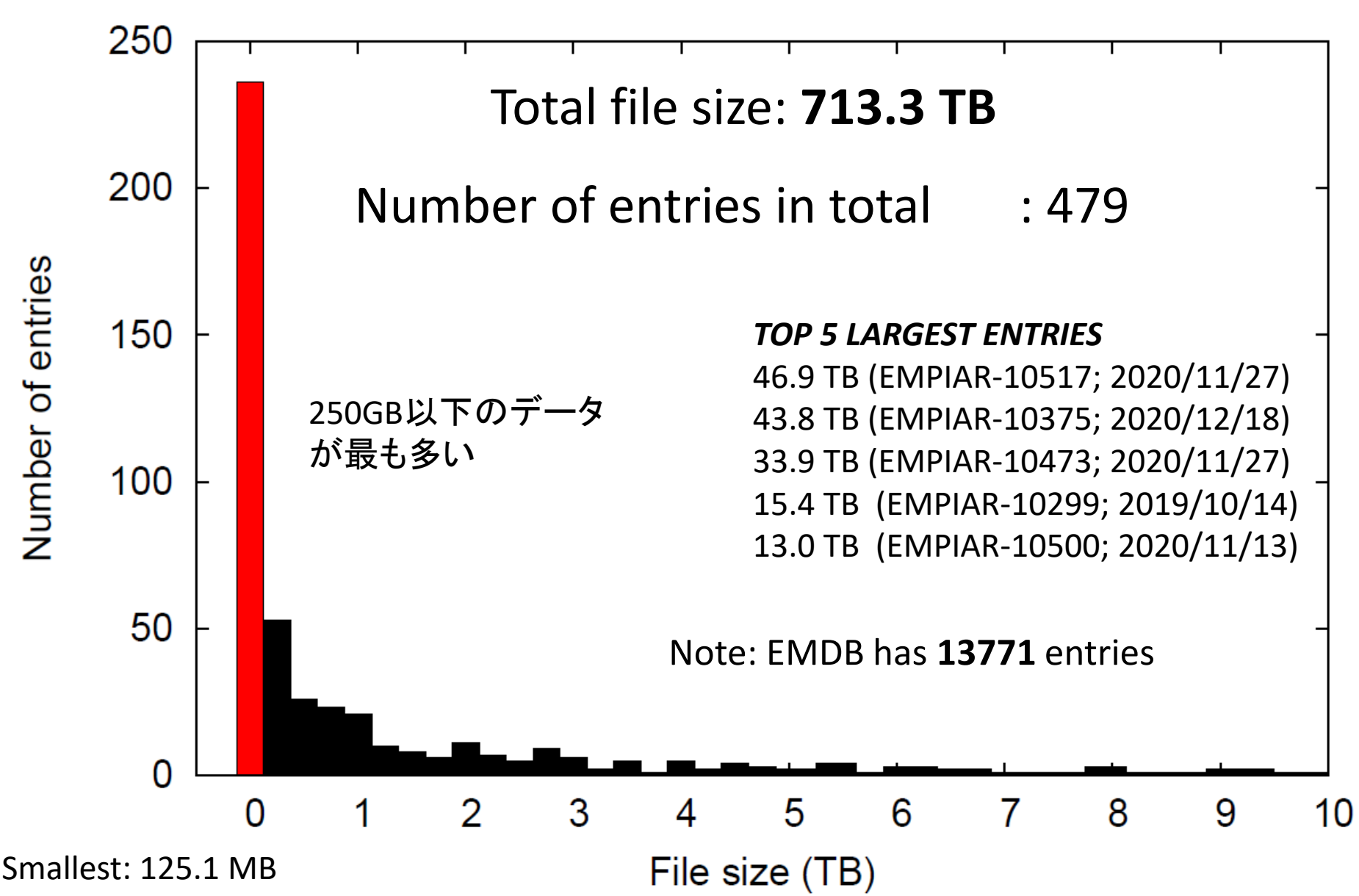


1) 日本サイトの一般公開開始（2018/12）
2) 高速ファイル転送ソフトウエア***Aspera***を導入・運用
3) 早稲田大学のコールドストレージにデータのバックアップ（由良研と共同）
4) 日本でのデータ登録(Deposition)支援の開始(2019/08)。ハードディスクの阪大への送付に対応。

EMPIARデータベースの増加



EMPIARのファイルサイズの分布



Smallest: 125.1 MB

EMPIAR-10127 Images for Picked Particles (**84.4 GB**)

169 MRCs (320 x 320 x ~100; 32 bit)

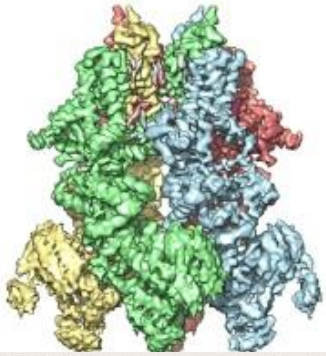
Human TRPM4 ion channel in a lipid nanodisc in a calcium-bound state

Publication:	Structure of the human TRPM4 ion channel in a lipid nanodisc
	Autzen HE, Cheng Y
	Science **359** 228-232 (2017)
	PMID: 29217581
	DOI: 10.1126/science.aar4510
Related PDB entry:	6bqv
Related EMDB entry:	EMD-7133
Deposited:	2017-11-30
Released:	2018-01-22
Last modified:	2018-01-22
Dataset size:	84.4 GB PDBj EBI
Dataset DOI:	10.6019/EMPIAR-10127

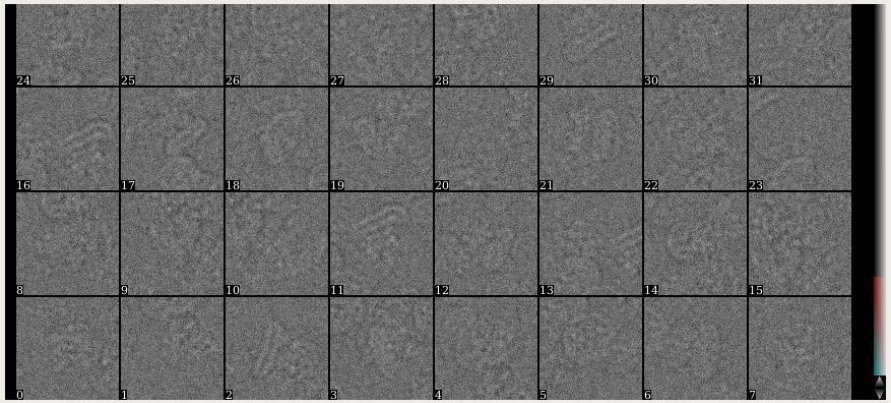
Contains: picked particles

2.2 Å

84.4 GB



Many small 2D images



✚ Image set

✚ particle stacks of TRPM4 particles post 2D clean-up

Category:	picked particles - multiframe - pro
Image format:	MRCS
No. of images or tilt series:	221100
Frames per image:	169
Image size:	(320, 320)
Pixel type:	32 BIT FLOAT
Details:	Particle positions in the corresponding .star files. Files have variable number of frames with maximum

169 MRCs (320 x 320 x ~100; 32 bit)

粒子画像群は、3Dマップの**FSC**による分解能の計算にまず必要

Multi-frame images (movies) (**12.4 TB**)

(7676 x 7420 x 38 x 1539; 32 bit)

EMPIAR-10061

2.2 A resolution cryo-EM structure of beta-
with a cell-permeant inhibitor

Publication: 2.2 A resolution cryo-EM st
with a cell-permeant inhibitor
Bartesaghi A, Merk A, Banerjee S, Matthies D, Wu X, Milne JL, Subramaniam S
Science **348** 1147-1151 (2015)
PMID: 25953817
DOI: 10.1126/science.aab1576

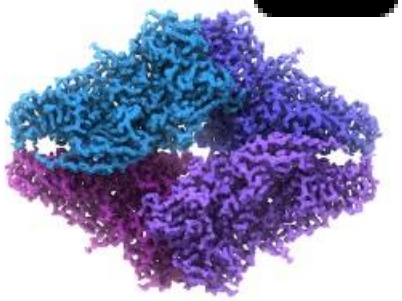
2.2 Å

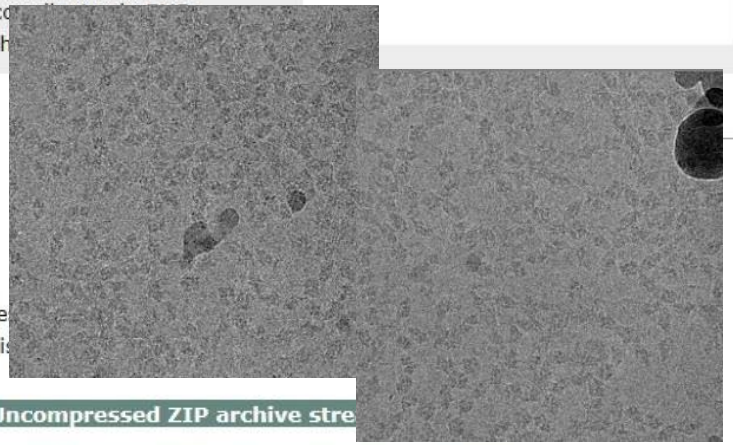
Related PDB entry: 5a1a
Related EMDB entry: EMD-2984
Deposited: 2016-04-11
Released: 2016-04-15
Last modified: 2016-05-13
Dataset size: 12.4 TB PDBj EBI

12.4 TB

Dataset DOI: 10.6019/EMPIAR-10061
Version history: 1 2016-05-13 Added the particle co
2984_boxes.tbz arch

micrographs





Averages of aligned movie frames

Category: micrographs - single frame
Image format: MRC
No. of images or tilt series: 1539
Image size: (7676, 7420)
Pixel type: 32 BIT FLOAT
Details: Micrographs are the average of frame
EMD-2984_boxes.tbz archive. There i
in EMAN's box format.

Micrographs 326.5 GB

Uncompressed ZIP archive stre

Raw movie frames

1539 MRC movies (7676 x 7420 x 38; 32 bit)

Category: micrographs - multiframe
Image format: MRC
No. of images or tilt series: 1539
Frames per image: 38
Image size: (7676, 7420)
Pixel type: 32 BIT FLOAT

Movies 12.1 TB

Uncompressed ZIP archive streamed via HTTP

Multi-frame images (movies) (**321 GB**)

(3710 x 3838 x 1338; 16 bit)

EMPIAR-1020

The first reconstruction of beta-galactosidase so

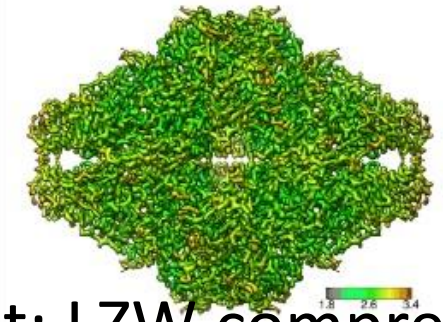
Publication:	First data of beta-galactosidase for validation of the state-of-the-art-cryo EM, named CRYOARM200 Kato T, Terahara N, Namba K
Related EMDB entry:	EMD-6840
Deposited:	2018-08-01
Released:	2018-08-15
Last modified:	2018-08-17
Dataset size:	321.4 GB PDBj EBI
Dataset DOI:	10.6019/EMPIAR-10204

2.6 Å

321.4 GB

Contains:

micrographs

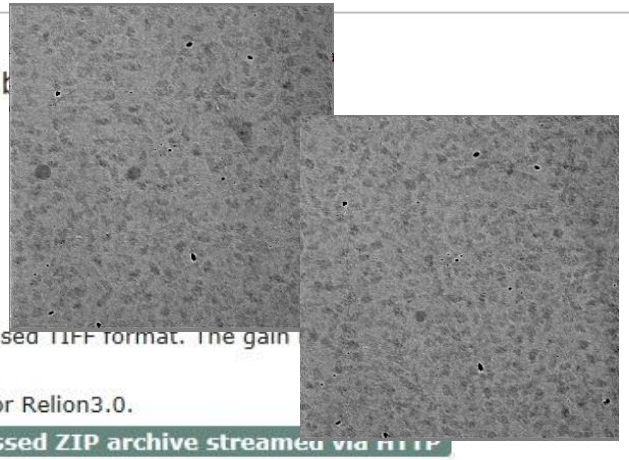


1338 TIFF movies (3710 x 3838 x 49; 16 bit; LZW compression) with a gain-reference file (K2-gain170629.mrc)

Image set

Unaligned multi-frame micrographs of beta-galactosidase recorded b

Category:	micrographs - multiframe
Image format:	TIFF
No. of images or tilt series:	1338
Frames per image:	49
Image size:	(3710, 3838)
Pixel type:	UNSIGNED 16 BIT INTEGER
Details:	Raw, non aligned micrograph moves in compressed TIFF format. The gain the same directory in MRC format. A part of this data was used as a tutorial data for Relion3.0.

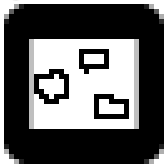
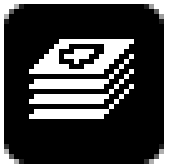


data 321.4 GB

Uncompressed ZIP archive streamed via HTTP

EMPIARには様々なデータが入っています

1) **Micrographs** or **Particle images**

2) **Multiple-frame** or **Single-frame**

(Movie) (Still image)

3) Gray scale depth: **32 bit** or **16 bit**

4) **Raw** or **Compressed** (*lossless*)

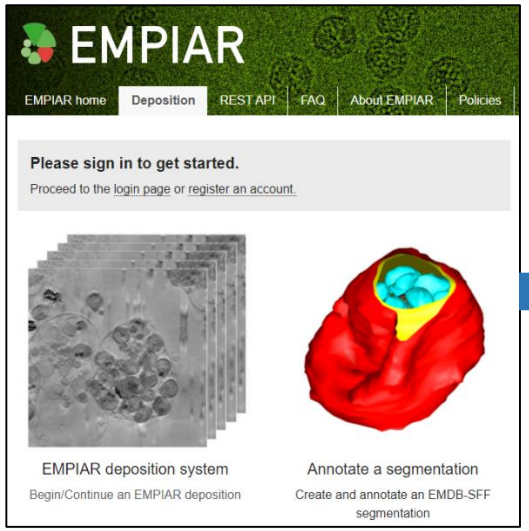
・どんなタイプのデータでも受け付けます

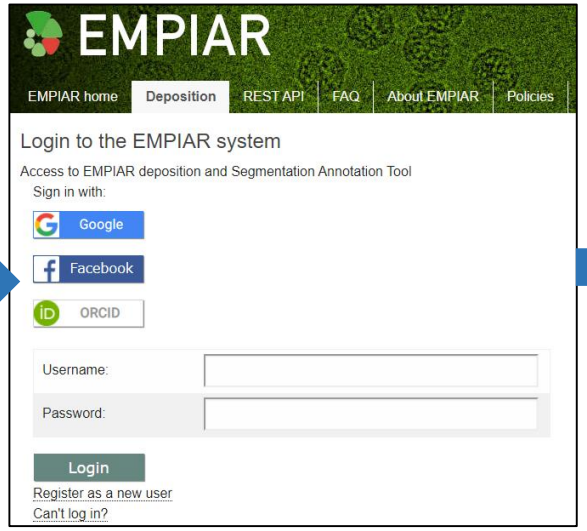
・ファイル名、ファイル数、ディレクトリ構成なども自由

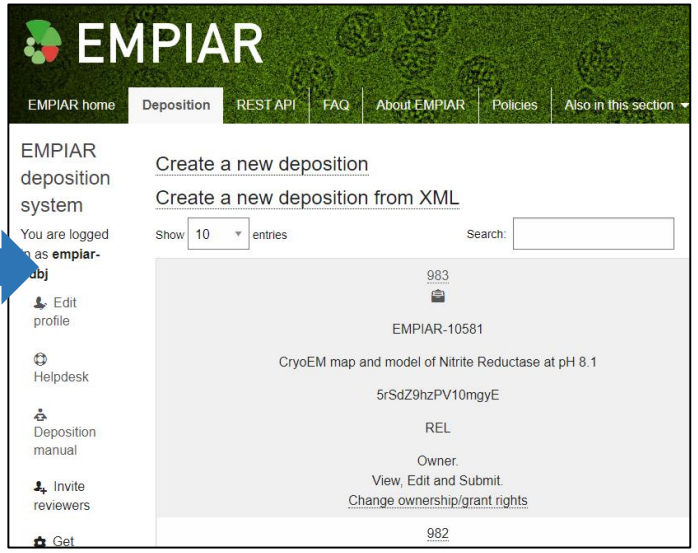
・できるだけrawがよいが、ファイルサイズとのバランス

EMPIARデータを登録する方法

[1]イギリスのEBIの登録サイトへ直接データを転送する方法







データが小さく、ネット環境がよく、長時間転送が可能であれば、スムーズに登録できるはず。

問題点：転送に時間がかかる。UKまで15時間/1TB。ネット環境が悪いと何度も失敗することも。

[2]データの入ったメディアを郵送・宅配便で大阪大学に送付する方法

阪大にHDDなどを郵送・宅急便で送る（１－２日）。UKへの転送は阪大で代行して行う

データの入ったメディアを郵送・宅配便で阪大に送付する方法

阪大スタッフがあなたの代わりに
UKへ転送します　(10TBなら1週間)

15 時間 / 1TB

ハードディスク (8-12 TB)

郵送・宅配便
(1-2 日)

東京

大阪

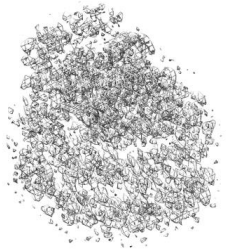
(1)登録者は 、ハードディスクを 阪大に
郵送か宅配便 で送る
(2)阪大スタッフがUKへネットワーク転送
(3)UKのスタッフがデータのチェック
(4)EMPIARデータベースで公開

HD送付の前に、電子メールでご連絡ください (empiar-help@protein.osaka-u.ac.jp)

阪大へのディスク送付で登録したエントリ

2019年度

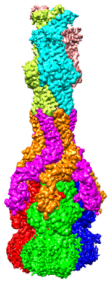
EMPIAR-10314



8.0 TB

Suga et al. Nat Plants 5, 626-636 (2019)

EMPIAR-10292



8.5 TB

Tsutsumi *et al*. ,Nat commu 10,1520 (2019)

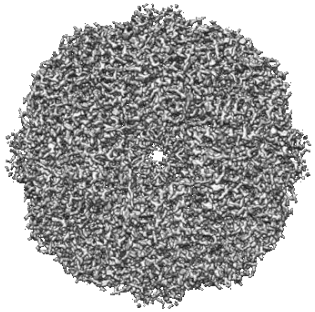
EMPIAR-10352

226 GB

Release after publication

2020年度

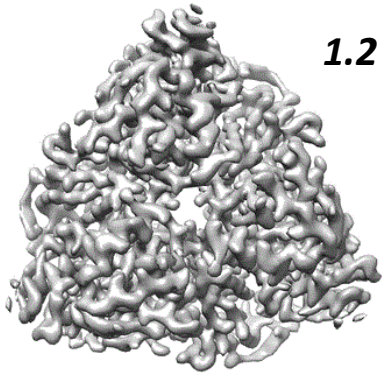
EMPIAR-10546



3.9 TB

Sato et al. J Struct Biol X 4, 100030 (2020)

EMPIAR-10580



1.2 TB

Adachi et al. bioRxiv (2020)

EMPIAR-10565

1.7 TB

Release after publication

EMPIAR-10569

1.5 TB

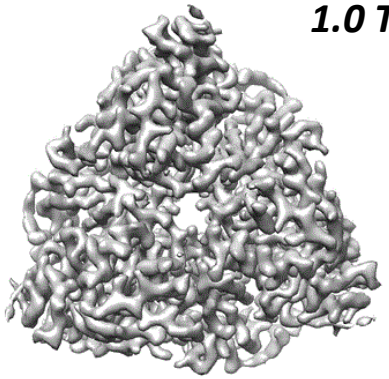
Release after publication

EMPIAR-10571

1.3 TB

Release after publication

EMPIAR-10581



1.0 TB

Adachi et al. bioRxiv (2020)

阪大経由でEMPIARデータを登録する方法

1. メタデータ記入用のjsonファイルをダウンロードして、著者名文献名などを入力。
   ※スキーマファイルを参考のこと。わからないところは空欄でも結構です。

2. 入力したjsonファイルを、empiar-help@protein.osaka-u.ac.jp宛にメールで送付。
   ※どのようなメディア(HDD, SSD、テープなど)でいつ送るつもりかも通知してください。

3. データの入ったメディア(HDD,SSD,テープなど)を郵送・宅配便で以下の住所に送付。

〒565-0871 大阪府 吹田市 山田丘3-2
大阪大学 蛋白質研究所 プロテインデータバンク研究室
川端 猛
Tel:06-6879-4311

※登録用の空のHDDを阪大から登録者の方に送付するサービスも行っています。データをコピーしていただいたあと着払いの宅急便で阪大に送り戻していただきます。登録用HDDの送付サービスを希望される方はメールでご連絡ください。

4. メディアが大阪大に到着後、こちらでデータの簡単なチェック
5. お預かりしたデータを、阪大のスタッフがイギリスのEBIに転送
6. イギリスEBIへの転送が完了しましたら、メディアを返送いたします

鋳型jsonファイルempiar_deposition.json (1)

※項目の意味はスキーマファイルempiar_deposition.schema.jsonに書かれています


```
{
 "title": "Vibrio cholerae toxin-coregulated pilus machine
 "release_date": "RE",
 "experiment_type": 3,
 "cross_references": [
  {
   "name": "EMD-8000"
  }
 ],
 "authors": [
  {
   "name": "('Smith', 'JW')",
   "order_id": 0,
   "author_orcid": "0000-0002-1825-0097"
  },
  {
   "name": "('Test', 'A')",
   "order_id": 1,
   "author_orcid": null
  }
 ],
 "corresponding_author": {
  "author_orcid": "0000-0002-1825-0097",
  "middle_name": "William",
  "organization": "Division of Biology, California Institute of Technology",
  "street": "1200 E California Blvd",
  "town_or_city": "Pasadena",
  "state_or_province": "California",
  "post_or_zip": "48984",
  "telephone": "012345",
  "fax": null,
  "first_name": "John",
  "last_name": "Smith",
  "email": "example@test.com",
  "country": "US"
 },
 "principal_investigator": [
```


“release_date” : RE - directly after the submission has been processed, **EP** - after the related EMDB entry has been released, **HP** - after the related primary citation has been published and **HO** - delay release of entry by one year from the date of deposition.

“experiment_type” : **1** - image data collected using soft x-ray tomography, **2** - simulated data, for instance, created using InSilicoTEM (note: we only accept simulated data in special circumstances such as test/training sets for validation challenges: you need to ask for and be granted permission PRIOR to deposition otherwise the dataset will be rejected), **3** - raw image data relating to structures deposited to the Electron Microscopy Data Bank, **4** - image data collected using serial block-face scanning electron microscopy (like the Gatan 3View system), **5** - image data collected using focused ion beam scanning electron microscopy, 6 - integrative hybrid modelling data."

鋳型jsonファイルempiar_deposition.json (2)


```
  "country": "US"
},
"principal_investigator": [
 {
  "author_orcid": "0000-0002-1825-0097",
  "middle_name": "William",
  "organization": "Division of Biology, Ca
  "street": "1200 E California Blvd",
  "town_or_city": "Pasadena",
  "state_or_province": "California",
  "post_or_zip": "48984",
  "telephone": "012345",
  "fax": null,
  "first_name": "John",
  "last_name": "Smith",
  "email": "test@example.com",
  "country": "US"
 }
],
"imagesets": [
 {
  "name": "Different tilt series for the Vibrio cholerae toxin-coregulated pilus machine revealed by electron cryotomography",
  "directory": "/data/micrographs",
  "category": "('T9', '')",
  "header_format": "('T1', '')",
  "data_format": "('T1', '')",
  "num_images_or_tilt_series": 16,
  "frames_per_image": 121,
  "frame_range_min": 1,
  "frame_range_max": 121,
  "voxel_type": "('T5', '')",
  "pixel_width": null,
  "pixel_height": null,
  "details": "You can write many details about your image data here !!",
  "image_width": 3838,
  "image_height": 3710
 }
],
```


"Category": **'T1'** corresponds to 'micrographs - single frame', **'T2'** - 'micrographs - multiframe', **'T3'** - 'micrographs - focal pairs - unprocessed', **'T4'** - 'micrographs - focal pairs - contrast inverted', **'T5'** - 'picked particles - single frame - unprocessed', **'T6'** - 'picked particles - multiframe - unprocessed', **'T7'** - 'picked particles - single frame - processed', **'T8'** - 'picked particles - multiframe - processed', **'T9'** - 'tilt series', **'T10'** - 'class averages', **'OT'** - other, in this case please specify the category in the second element.

"**header_format**" and "**data_format**": '**T1**' corresponds to 'MRC', '**T2**' - 'MRCS', '**T3**' - 'TIFF', '**T4**' - 'IMAGIC', '**T5**' - 'DM3', '**T6**' - 'DM4','**T7**' - 'SPIDER', '**OT**' - other, in this case please specify the header format in the second element in capital letters.,

"**voxel_type**": '**T1**' corresponds to 'UNSIGNED BYTE', '**T2**' - 'SIGNED BYTE', '**T3**' - 'UNSIGNED 16 BIT INTEGER', '**T4**' - 'SIGNED 16 BIT INTEGER', '**T5**' - 'UNSIGNED 32 BIT INTEGER', '**T6**' - 'SIGNED 32 BIT INTEGER', '**T7**' - '32 BIT FLOAT', '**OT**' - other, in this case please specify the header format in the second element in capital letters."

"Imagesets"の"details"は自由記述欄です。データ解析の際の留意点などはここに詳しく記載しください。

鋳型jsonファイルempiar_deposition.json (3)


```
"citation": [
 {
  "authors": [
   {
    "name": "('Smith', 'JW')",
    "order_id": 0,
    "author_orcid": "0000-0002-1825-0097"
   },
   {
    "name": "('Test', 'A')",
    "order_id": 1,
    "author_orcid": null
   }
  ],
  "editors": [],
  "published": true,
  "j_or_nj_citation": true,
  "title": "Architecture of the Vibrio cholerae toxin-coregulated pilus machine revealed by electron cryotomography",
  "volume": "2",
  "country": null,
  "first_page": "16269",
  "last_page": "16269",
  "year": 2017,
  "language": "English",
  "doi": "10.1038/nmicrobiol.2016.269",
  "pubmedid": "28165453",
  "details": "Three-dimensional in situ structure of a T4bP machine in its piliated and non-piliated states constructed from its tilt series.",
  "book_chapter_title": null,
  "publisher": null,
  "publication_location": null,
  "journal": "Nature microbiology",
  "journal_abbreviation": "Nat Microbiol",
  "issue": null
 }
]
}
```


出版論文の情報

まとめ

- EMDB, PDBに登録の際は、EMPIARの登録も是非！
- 電顕の2D画像の公開は、データの妥当性の検証、画像解析技術の進歩、教育啓蒙、に役立ちます。
- どんなデータ形式でも受け付けます（アーカイブなので）
- 単粒子解析でもトモグラフィーでもOK
- できるだけrawなデータ（できれば動画）をご提供ください。
- できるだけ詳しい説明書きを入れてください。

  (README.txtを入れる。Jsonの”imageset”:”details”に書く)
- 転送が大変な方は阪大が代行します。HDDを阪大にお送りください。