

1 **ETDB-Caltech: a blockchain-based distributed public database for electron**  
2 **tomography**

3

4

5

6 Davi R. Ortega<sup>1</sup>, Catherine M. Oikonomou<sup>1</sup>, H. Jane Ding<sup>1</sup>, Prudence Rees-Lee<sup>1</sup>, Alexandria<sup>^</sup>,  
7 Grant J. Jensen<sup>1,2,\*</sup>

8

9

10

11

12

13 <sup>1</sup> Division of Biology and Biological Engineering, California Institute of Technology, Pasadena,  
14 California, USA

15 <sup>2</sup> Howard Hughes Medical Institute, Pasadena, California, USA

16

17 \*Corresponding author

18 Email: [jensen@caltech.edu](mailto:jensen@caltech.edu) (GJJ)

19

20 <sup>^</sup>Membership of Alexandria is provided in the Acknowledgments.

21

## 22 **Abstract**

23 Three-dimensional electron microscopy techniques like electron tomography provide valuable  
24 insights into cellular structures, and present significant challenges for data storage and  
25 dissemination. Here we explored a novel method to publicly release more than 11,000 such  
26 datasets, more than 30 TB in total, collected by our group. Our method, based on a peer-to-peer  
27 file sharing network built around a blockchain ledger, offers a distributed solution to data  
28 storage. In addition, we offer a user-friendly browser-based interface, <https://etdb.caltech.edu>, for  
29 anyone interested to explore and download our data. We discuss the relative advantages and  
30 disadvantages of this system and provide tools for other groups to mine our data and/or use the  
31 same approach to share their own imaging datasets.

32

## 33 **Introduction**

34 Three-dimensional electron microscopy (3D EM) techniques produce large and information-rich  
35 datasets about biological samples. In electron tomography (ET), samples are imaged as they are  
36 tilted incrementally – typically 1-2 degrees between images. The resulting tilt-series of 2D  
37 projection images can then be computationally combined into a 3D reconstruction, or tomogram,  
38 of the sample with nanometer-scale resolution. ET has both biological [1] and materials science  
39 applications [2]. ET is frequently performed on frozen samples (cryo-ET) such as intact, small  
40 cells. Cryo-ET has revealed many details about cell ultrastructures that are inaccessible by other  
41 techniques, either because they cannot be purified intact or because they are not preserved by  
42 traditional EM sample preparations [3]. Another 3D EM technique, single particle analysis, also  
43 yields 3D information about cellular complexes [4].

44

45 Biological applications of 3D EM techniques are rapidly increasing, with an explosive rise in the  
46 number of datasets published [5] and excitement about the field (e.g. [6-8]). In addition,  
47 technological advances such as increased automation for higher-throughput data collection and  
48 movie acquisition with direct detectors are increasing the information content of datasets [9, 10],  
49 which makes management of these datasets a mounting challenge [11]. At the same time, public  
50 accessibility is of critical importance [12]. 3D EM techniques, while burgeoning, are still  
51 inaccessible to most cell biologists due to the expensive equipment (several million dollars to  
52 purchase and maintain, in a customized space) and specialized expertise required. In addition, the  
53 technology is still in a phase of active development, in both hardware and software. To facilitate  
54 software development efforts, programmers need access to large and varied test datasets.

55  
56 Public dissemination outlets for 3D EM datasets address two fundamentally different missions:  
57 (1) to provide curated, validated data for peer review and education [13]; and (2) to provide large  
58 quantities of possibly problematic data to facilitate biological discovery and software  
59 development. The first mission is well served by resources such as the Electron Microscopy Data  
60 Bank (EMDB) and the Cell Image Library. The EMDB, an invaluable community tool for  
61 deposition of 3D EM data [14], is part of the EMDatabank [15], a global resource for 3D EM  
62 managed by the worldwide Protein Data Bank (PDB) consortium [16]. Like its counterpart, the  
63 PDB [17], it is the standard repository for published structures, such as single particle  
64 reconstructions and subtomogram averages [18]. To encourage public access, the EMDB  
65 developed web-based visualization tools to interact with data [19, 20]. The Cell Image Library<sup>TM</sup>  
66 is an open-source catalog of curated images, animations and videos aimed at disseminating cell  
67 biology to the broader public [21]. Entries include light and electron microscopy imaging, as

68 well as correlated datasets. The resource includes datasets previously available as the Cell  
69 Centered Database (CCDB), an online repository of high-resolution, often 3D, light and electron  
70 microscopy data, including many electron tomograms [22-24].

71  
72 The second mission is currently served in a more piecemeal fashion, largely by initiatives from  
73 single labs and imaging centers to release a subset of their raw datasets for public access.  
74 Unfortunately, these resources often suffer from a lack of permanence due to lapsed maintenance  
75 of published websites. Recognizing the need for a centralized public repository of the raw EM  
76 datasets from which EMDB structures are derived, in 2016 the European PDB announced a sister  
77 site to the EMDB: the Electron Microscopy Public Image Archive, or EMPIAR [25]. EMPIAR  
78 collects tilt-series related to reconstructions deposited in the EMDB. It therefore offers an ideal  
79 resource for benchmarking software with verified, published datasets, but it is not designed for  
80 large-scale releases of unpublished, problematic and/or complicated datasets: datasets must be  
81 associated with an EMDB deposition; only tilt-series can be deposited (the resulting  
82 reconstructions are available in the EMDB, but associated files such as correlated light  
83 microscopy images or digital segmentations cannot be included); and much of the metadata is  
84 entered manually [26], a daunting task for a large batch of data.

85  
86 While releasing data of unverified quality may seem to be of dubious value, we would argue that  
87 it is necessary for the progress of the field. As pointed out by the developers of the CCDB, ET  
88 datasets that currently yield poor-quality reconstructions offer opportunities for developing better  
89 reconstruction methods [24]. Also, biological insights often come from unexpected places; as a  
90 single anecdotal example, years ago our lab collected electron tomograms of bacteria to study

91 chromosome segregation and observed novel tubes inside cells; we shared the images and a cell  
92 biologist made a connection to a secretion system he was studying, allowing us together to figure  
93 out its mechanism [27].

94

95 Since 2003, our lab has collected more than 30,000 ET datasets. Each dataset consists of a tilt-  
96 series of 2D TEM projection images and the resulting 3D tomographic reconstruction, as well as  
97 additional image, video, and segmentation files. Each dataset is 1-5 GB, and the full collection  
98 adds up to ~110 TB of data. To store and curate this volume of data for internal use by our  
99 group, we developed the Caltech Tomography Database, a central repository linked to a browser-  
100 based interface for lab members to browse, search, and download data [28]. To further  
101 streamline data handling, we integrated the internal Caltech Tomography Database with an  
102 automatic processing pipeline that uploads and processes datasets as they are acquired by the  
103 microscope [28]. The majority of our ET datasets come from cryo-preserved cells. They  
104 represent more than 100 unique species of bacteria, archaea, and eukaryotes and have led to  
105 dozens of publications about diverse aspects of cell ultrastructure. The nature of whole-cell  
106 imaging, though, means that these datasets are far from exhausted. While we collected them for a  
107 specific study, they contain information about many other aspects of cell biology that may be  
108 useful to other researchers.

109

110 While we have been sharing our data by publishing papers and depositing representative  
111 tomograms in the EMDB, we have also received many requests—from software developers,  
112 biologists, and EMPIAR—to share more of our data. We filled these individual requests, but  
113 wanted to explore a broader solution to enable our lab and others to share large amounts of data

114 of unverified quality in a persistent and decentralized fashion. The approach we describe here  
115 uses a distributed peer-to-peer file network tracked by an ownerless ledger (blockchain) system.  
116 We describe how we used this method to release more than 11,000 electron tomography datasets  
117 (excluding those that are still part of ongoing studies), representing 85 species and encompassing  
118 more than 30 TB. We discuss the advantages and drawbacks of our approach, and how it can be  
119 adopted by other groups that wish to share their own datasets.

120

## 121 **Results & Discussion**

122

### 123 *Approach*

124 In recent years, decentralized cryptographic ledgers, or blockchains, have been explored as a  
125 method to securely record data (typically cryptocurrency transactions, for which they were first  
126 conceived [29]). Rather than relying on a trusted central authority, blockchains employ a security  
127 model that builds consensus from a system of distributed users, none of whom necessarily need  
128 to trust one another. Originally developed to solve the problem of double-spending, blockchain  
129 technology has since been adapted to other uses. For instance, the Republic of Georgia uses the  
130 bitcoin blockchain to record land transfer titles, one of several countries using the cryptographic  
131 ledger to improve the security of property rights [30]. In the United States, blockchains have  
132 been proposed as a way for patients to control access to their digital medical records [31, 32].  
133 Blockchains are used by Nasdaq in the U.S. and stock exchanges in other countries to record  
134 private securities transactions [33].

135

136 In 2013, an anonymous developer announced a fork from a cryptocurrency called Litecoin to  
137 create a new cryptocurrency, FlorinCoin (FLO), whose ledger features a descriptive transaction  
138 comment line similar to that found on a traditional check. The text entered in this transaction  
139 comment is stored in the FLO blockchain along with the details of the transaction. Each  
140 comment can contain up to 528 characters [34]. In 2014, a company called Alexandria proposed  
141 to use this feature as a public record of information and developed an open source protocol  
142 termed the Open Index Protocol (OIP) [35]. They first used this protocol to record public social  
143 media status in the FLO blockchain and later, using a peer-to-peer distributed file-sharing  
144 network, they expanded the specifications of the protocol to register the metadata of videos and  
145 music in the FLO blockchain while storing the files in the peer-to-peer file-sharing network  
146 BitTorrent, allowing artists to prove ownership of these digital assets. From September 2017 to  
147 May 2018 FlorinCoin passed through a series of upgrades. It was renamed FLO, its code was  
148 updated to version 0.15 of Bitcoin (still retaining the sCrypt algorithm for proof-of-work), and  
149 the comment field was expanded to 1,040 characters. The current OIP specification (0.42) is  
150 optimized for the new FLO comment field size, encompasses a variety of data types, and uses a  
151 peer-to-peer file system called the InterPlanetary File System (IPFS) [36] to store files. File  
152 metadata is thereby cryptographically secured, and completely searchable, allowing anyone to  
153 discover and download the files from the IPFS.

154

155 We were curious to see if this blockchain-based data distribution model would be effective to  
156 openly and securely share our scientific imaging data. In the scheme, each dataset would be  
157 distributed to IPFS and its metadata recorded in the FLO blockchain. Any interested party,  
158 typically through a user-friendly front-end in their web browser, could query the blockchain for

159 datasets of interest and retrieve them from IPFS. We called the resulting distributed database the  
160 public Electron Tomography Database - Caltech (ETDB-Caltech), and its information flow is  
161 schematized in Figure 1.

162

163 **Figure 1. Information flow in the ETDB-Caltech file-sharing network.** Datasets hosted from  
164 a local server are distributed to IPFS, a network of seeding nodes that includes the local server.  
165 The associated metadata and locations of the files are recorded in the FLO blockchain using the  
166 OIP specification. Users can query this ledger to locate and retrieve desired files from the IPFS.  
167

168 We worked with Alexandria to develop a digital record type tailored to the metadata of our  
169 datasets that could be encoded easily in the FLO transaction comment. The result, Research-  
170 Tomogram, contains fields corresponding to the information we store about each dataset in our  
171 internal database. This information includes details about the user who collected the data,  
172 descriptions of the sample and its preparation, and data acquisition and processing parameters.  
173 Where appropriate, this information follows standard conventions for the 3D EM field [37]. We  
174 wrote a simple GoLang script to automatically read this information from the record in the  
175 internal lab database and translate it into an OIP Research-Tomogram record. If other groups  
176 want to adopt this approach, they can use a subset of these fields and/or add their own as  
177 necessary to match their local recordkeeping. Table 1 lists the currently available fields in the  
178 Research-Tomogram record.

179

180 **Table 1. Fields in the Research-Tomogram record.**

181

		Description
floAddress*		cryptographic key of publisher
info	title*	descriptive title of dataset (chosen at acquisition)
	description	notes about publication process of the record
	tags	searchable tags, e.g. "tomogram," "etdb," "jensen.lab"
details	date*	acquisition date



	NCBITaxID		NCBI taxonomy identifier
	artNotes		notes about the dataset
	scopeName		acquisition microscope, e.g. "Caltech Polara"
	speciesName*		species of cell imaged
	strain		information about the specimen strain
	tiltSingleDual		single-axis or dual-axis tilt acquisition scheme
	defocus		imaging defocus ( $\mu\text{m}$ )
	dosage		imaging electron dosage ( $e/\text{\AA}^2$ )
	tiltConstant		1: if constant angular increment; 0: if other method
	tiltMin		minimum of acquisition tilt range (degrees)
	tiltMax		maximum of acquisition tilt range (degrees)
	tiltStep		tilt increment (degrees)
	swAquisition		software used for acquisition
	swReconstruction		software used for reconstruction
	magnification		acquisition magnification (X)
	emdb		EMDB code if record is also available on EMDB
	microscopist		scientist who acquired tilt-series
	institution		e.g. "Caltech"
	lab		e.g. "Jensen Lab"
	sid		internal database identifier (laboratory specific)
storage	network*		e.g. "IPFS"
	files**	fname*	file name
		dname	name to be displayed in interface
		fsize	file size (bytes)
		type	e.g. "Tomogram" or "Image"
		subtype	e.g. "Tiltseries" or "Reconstruction"
		ctype	content type, e.g. "image/jpeg" or "video/mp4"
	location*		hash of file locations for retrieval
payment			payment information (N/A for this blockchain use)
timestamp*			time of publication to blockchain
type*			"Research"
subtype			"Tomogram"

182  
183  
184  
185

\* *mandatory field*

\*\**stores the indicated information for each file associated with the dataset*

186 As in other peer-to-peer networks, files can be chunked and hosted from multiple nodes in the  
187 network. Users who download a file and participate in IPFS can choose to host it in this fashion  
188 for other users. This feature makes the distribution model scalable; if many users are  
189 downloading a file, multiple seeds speed up those downloads, avoiding a bottleneck from a

190 single server. In our case, we expect relatively light file traffic, so at the current time, files are  
191 downloaded solely from our server, as in a traditional distribution model. In the rare event that a  
192 dataset is published in error, OIP offers the option of deactivating a published record. This action  
193 will not erase the metadata published in the blockchain, but the record will no longer be available  
194 to anyone using the OIP API to search the blockchain. In that case, if a user were interested in an  
195 unavailable tomogram, they would have to search the raw data in the blockchain, and hope that  
196 the files were still in the IPFS network.

197  
198 There are two ways that users can download our datasets. The first is through a direct query of  
199 the blockchain and IPFS. We built a command-line application that facilitates this approach; see  
200 *Materials & Methods* for details. To increase public accessibility, we added a second route: a  
201 browser-based front-end. This graphical interface, which can be found at <https://etdb.caltech.edu>,  
202 provides an intuitive, interactive experience for anyone to browse ETDB-Caltech datasets, view  
203 images and videos they contain, and download part or all of each dataset. A sample dataset  
204 display page is shown in Figure 2.

205  
206 **Figure 2. Sample entry page in the browser-based ETDB-Caltech interface.** A sample  
207 electron cryotomography dataset from a *Vibrio cholerae* cell is shown. An embedded video of  
208 the reconstruction appears at left and plays automatically. The metadata is shown at right. Files  
209 associated with the dataset are listed at the bottom of the page, where they can be downloaded  
210 individually.

211  
212 The ETDB-Caltech front-end offered us a chance to highlight scientific challenges for target user  
213 groups – cell biologists and software developers. We hope cell biologists will find novel features  
214 in the imaged cells, and identify those that remain mysterious. Electron tomograms contain a  
215 wealth of information, not all of which is currently interpretable; recently, for instance, we

216 published a paper describing some of the cellular features we have observed in our electron  
217 tomograms but could not identify [38]. We hope software developers will use the released  
218 datasets to improve image-processing algorithms. In particular, we hope the availability of these  
219 datasets contributes to the development of software that can: (1) more reliably find and track the  
220 fiducial markers used for alignment in tomographic reconstruction; (2) automatically and  
221 accurately segment the boundaries of cells; and (3) automatically segment large macromolecular  
222 complexes in cells. In addition to their usefulness to experts in the field, the datasets in ETDB-  
223 Caltech may be of interest to students and the general public. To welcome these users, we  
224 designed the front-end of ETDB-Caltech to be accessible and educational, with information  
225 about the data and technology, as well as a Featured Tomograms page highlighting various  
226 features of bacterial and archaeal cells that are visible in electron tomograms (Figure 3).

227

228 **Figure 3. Featured Tomograms page of the ETDB-Caltech interface.** Targeting students and  
229 others unfamiliar with ET data, the page highlights cellular features of bacteria and archaea  
230 visible by cryo-ET. Selecting a category takes the user to a page with a brief description of the  
231 structure and a few datasets containing examples.

232

### 233 ***Outlook***

234 Here we tested a new approach to publicly share a large amount of ET data. If our goal was  
235 simply to continue honoring requests from the community to make our datasets public, it would  
236 have been cheaper and easier to simply host the data from a local MySQL database, as we do for  
237 our internal group users. However, we also wanted to make a broader resource that could  
238 encompass data from many ET labs into a flexible repository that does not rely on a central  
239 authority. If ETDB is ultimately successful in enabling large-scale community data sharing, we  
240 believe it will complement (but never replace) the mission of curated repositories like EMDB

241 and EMPIAR by providing varied datasets with a wide range of quality and content for  
242 biological and technological projects.

243  
244 Compared to more centralized models of data storage, this dissemination model offers several  
245 attractive points. The first is flexibility. Multiple file types can be combined in a single OIP  
246 record, allowing, for example, light micrographs from correlative light and electron microscopy  
247 experiments and annotated segmentations to be included in EM datasets; this has been cited as a  
248 key feature lacking in some current repositories [12, 39]. Other file types from different imaging  
249 modalities can be accommodated with similar ease. The OIP specification of the Research-  
250 Tomogram record type requires few mandatory fields (Table 1). These fields can be adapted to  
251 the metadata collected by other groups, who may be using different internal databases (e.g. [40,  
252 41]). The flip side of this flexibility is that, compared to repositories of validated datasets like  
253 EMDB/EMPIAR [26], ETDB entries may be missing information like pixel size or contain  
254 errors in metadata. This caveat should be kept in mind when using the data in further studies;  
255 information critical to interpretation should be verified with the depositor.

256  
257 Another appealing feature of distributed file sharing is the distribution of storage and cost. 3D  
258 EM datasets are large, as reflected by EMPIAR, which has grown to accommodate >80 TB of  
259 stored data in 5 years [42]. These datasets are associated with only 168 studies [43]. The  
260 popularity of 3D EM methods, particularly cryo-ET [8], is growing rapidly: the number of  
261 entries in the EMDB has more than doubled over the last three years [5, 44]. There are currently  
262 more than 6,500 entries in the EMDB [44]; if each of these was associated with a similarly-sized  
263 dataset in EMPIAR, more than 3 PB of centralized storage space would be required. In a

264 distributed distribution model, each contributing lab is responsible for storing their own data,  
265 which they presumably already do. In our case, we could have implemented the system using our  
266 existing server, which hosts our internal database, at no added cost. For extra security, we chose  
267 to keep the server with the internal database behind a local firewall and mirror the relevant  
268 datasets on an additional server outside the firewall hosting ETDB. This second server, which is  
269 larger than necessary to accommodate additional applications and future growth, cost  
270 ~US\$7,000.

271  
272 In addition to the local server, files should be available from other nodes of the IPFS. This  
273 ensures data persistence in the event of, for instance, a local disk failure. Of course, how well  
274 this feature works depends on whether the system is widely adopted. In addition to users hosting  
275 IPFS nodes, institutions can also easily archive ETDB data through the IPFS. The more nodes  
276 are hosting a file in the IPFS, the higher the bandwidth for users to download it; this scalability is  
277 a major feature of peer-to-peer networks. Currently, however, the IPFS is still experimental and,  
278 like many new technologies, unstable. For that reason, we serve the files in our front-end directly  
279 from the IPFS node running on our local server, not through the full IPFS peer-to-peer network.  
280 However, IPFS is in rapid development and we expect soon to update the front-end to fetch and  
281 serve the files from the IPFS. Our command line application for bulk download, ETDB-  
282 downloads, already retrieves the files from the IPFS network.

283  
284 The maintenance of the ownerless ledger used to store the ETDB metadata, the FLO blockchain,  
285 depends on a distributed network of miners and users. This feature facilitates adoption as anyone  
286 can publish tomograms to the ETDB without having to seek permission from a central authority.

287 However, as in other cryptocurrencies, miners and users have an incentive to participate in the  
288 FLO network depending on a combination of factors including the costs of hardware and  
289 electricity, and the value of FLO in the cryptocurrency market. Although FLO has been in  
290 circulation for over 5 years, a relatively long time by cryptocurrency standards, its eventual  
291 success is difficult to predict. If FLO becomes an inviable option, it may be necessary to switch  
292 to a different ledger system in the future (Ethereum, Namecoin, and Bitcoin Cash are all capable  
293 of storing text). Note, however, that metadata already published remains accessible as long as at  
294 least one copy of the FLO blockchain exists; we host one ourselves.

295  
296 For us, the project took a few months to complete and the cost for the cryptocurrency  
297 transactions we used to publish 11,293 datasets was US\$17.89 (see *Materials and Methods*).  
298 Most of the development effort was invested in the user interface as well as the scripts to  
299 automatically upload datasets to the IPFS and the metadata to the FLO blockchain using OIP. If  
300 other groups wish to adopt the same approach to make their data public, they would only need to  
301 slightly modify these scripts (available on GitHub, see *Materials & Methods*) to match their  
302 internal database descriptors. Our front-end code is similarly available on GitHub so that other  
303 groups can easily adapt it to taste and use it to display: (1) their own data, (2) all ETDB datasets  
304 in the IPFS, or (3) a custom subset (e.g. data from a single species or technique). In addition,  
305 individuals interested in web applications for visualization and manipulation of tomograms can  
306 use the ETDB as a distributed database of content without needing to host any tomograms  
307 themselves. Outlets (e.g. science educators) can stream tomogram videos directly from the IPFS  
308 network.

309

310 Ultimately, we believe the relationship between the ETDB and curated central repositories like  
311 the EMDB is complementary. We will continue to support the invaluable mission of the EMDB  
312 and EMPIAR in safeguarding scientific data by submitting representative curated datasets we use  
313 in our publications. We hope that the ETDB can in turn help facilitate broader releases of large  
314 batches of electron tomography data for community use. If successful, the ETDB could even be  
315 integrated into centralized repositories by their hosting an IPFS node, enhancing accessibility of  
316 the data. The flexible features of this blockchain-based, distributed scheme of data sharing may  
317 also make it useful for other types of scientific data.

318

## 319 **Materials & Methods**

### 320 *ETDB-Caltech Distribution*

321 The ETDB-Caltech database is fed by a MySQL database (version 14.14 distribution 5.7.21)  
322 hosted on an Ubuntu Server (Artful Aardvark kernel version 4.3.0-37). The MySQL database  
323 contains the metadata of entries from the Caltech Tomography Database [28] that have been  
324 designated for publication. Associated files are stored in a RAID6 ext4 file system. Each night,  
325 the internal server hosting the internal Caltech Tomography Database executes a script to find  
326 datasets newly edited or marked for publication and copy them to the external ETDB-Caltech  
327 server, updating the MySQL database.

328

329 The ETDB-Caltech server runs a full node of the FLO blockchain, a node of the IPFS and a  
330 MySQL database. Upon changes in the MySQL database, a custom-built GoLang script (go-  
331 etdb, available on Github: <https://github.com/theJensenLab/go-etdb>) makes the new files  
332 publicly accessible through the InterPlanetary File System (IPFS, version 0.4.15-dev) [36]. The

333 IPFS daemon calculates a unique identifier to the dataset directory called a hash which is  
334 cryptographically dependent on the contents of the directory and makes the directory available to  
335 other nodes of the IPFS. This hash is combined with the metadata of each dataset and formatted  
336 according to Open Index Protocol (OIP, version 0.42) specification to create a JSON record (see  
337 Table 1). Each record generated this way is signed with a cryptographic key unique to the Jensen  
338 lab (the private key associated with public address  
339 FTSTq8xx8yWUKJA5E3bgXLzZqqG9V6dvnr) and published to the FLO blockchain by a  
340 daemon (OPId) on the server, attaching the record to the "floData" field of one or more  
341 transactions. The cost to publish the full set of 11,293 tomograms (at then-current rates of  
342 exchange) was US\$17.89.

343  
344 To search for ETDB-Caltech data, any user can use the cryptographic key given above to query  
345 the blockchain and retrieve matching ETDB records. This procedure is facilitated by an OIP  
346 daemon that scans and indexes the FLO Blockchain and exposes an Application Programming  
347 Interface (API) for public use. The API is accessible by a package (oip-js) deposited on the node  
348 package manager (npm). We also developed a command-line application for Unix-related  
349 environments (ETDB-downloads, manual available on Github:  
350 <https://github.com/theJensenLab/etdb-downloads/blob/master/userManual.md>) designed to allow  
351 users to download all or a subset of ETDB-Caltech datasets. Unlike the ETDB-Caltech website  
352 (see below), this application launches a temporary IPFS node and fetches the files from the IPFS  
353 network.

354

355 *ETDB-Caltech Interface*



356 The front-end was built using node.js (version 9.1), react (16.2.0), webpack (4.1.1), and Twitter  
357 Bootstrap. It uses the oip-js package (<https://github.com/oipwg/oip-js>) to connect to an  
358 OIPdaemon Representational State Transfer (REST) API, which scans the FLO blockchain for  
359 valid OIP records and indexes them into an internal database. Currently, oip-js queries  
360 OIPdaemon for a list of records with type "Research" and subtype "Tomogram" published by our  
361 lab (the private key associated with public address:  
362 FTSTq8xx8yWUKJA5E3bgXLzZqqG9V6dvnr). In the future, queries could also search for the  
363 cryptographic keys of different groups. Alternatively, records could be retrieved by a full-node  
364 search of the FLO blockchain (available on GitHub: <https://github.com/floblockchain/flo>) with  
365 OIPdaemon. Files are served for download from this interface directly from the IPFS node on the  
366 ETDB-Caltech server.

367  
368 The interface was designed to be easily navigable by scientists and non-scientists, and is  
369 optimized for viewing on all common web-enabled devices. We expect that in the future, some  
370 users and other labs may wish to customize this web interface. They can either copy and modify  
371 our template (available on GitHub: <https://github.com/theJensenLab/etdb-react>) or develop their  
372 own. While the Caltech ETDB interface displays only entries from our lab, other users may wish  
373 to build front-ends to display data from all labs sharing data using Open Index Protocol or to  
374 display only a subset of interest, for instance only those datasets corresponding to a particular  
375 species. In that case, instead of serving the files directly from the ETDB-Caltech IPFS node,  
376 those websites would use the peer-to-peer feature of the IPFS to search for the files in multiple  
377 nodes.

378

## 379 **Acknowledgments**

380 We thank members of the Jensen lab for helpful comments on the ETDB interface, as well as  
381 past and present lab members (Morgan Beeby, Ariane Briegel, Yi-Wei Chang, Songye Chen,  
382 Megan Dobro, Lu Gan, Gregory Henderson, Cristina Iancu, Andreas Kjær, Zhuo Li, Alasdair  
383 McDowall, Gavin Murphy, Martin Pilhofer, Rasika Ramdasi, Jian Shi, Poorna Subramanian,  
384 Matthew Swulius, William Tivol, Elitza Tocheva, Cora Woodward, Qing Yao, Zhiheng Yu, and  
385 Elizabeth Wright who generously allowed data they collected to be made public. We also thank  
386 other lab members whose data will be published in the future. The Alexandria team is composed  
387 of Devon Read James, Amy James, Jeremiah Buddenhagen, Sky Young, Ryan Chacon and  
388 Anthony Stewart. Thanks also to past Alexandria contributors Ryan Jordan, Ryan Taylor, and  
389 Joseph Fiscella for their work on the Open Index Protocol specification. This work was made  
390 possible through the support of the National Institutes of Health (grant R35 GM122588 to G.J.J.)  
391 and the John Templeton Foundation as part of the Boundaries of Life Initiative (grant 51250 to  
392 G.J.J.).

393

## 394 **References**

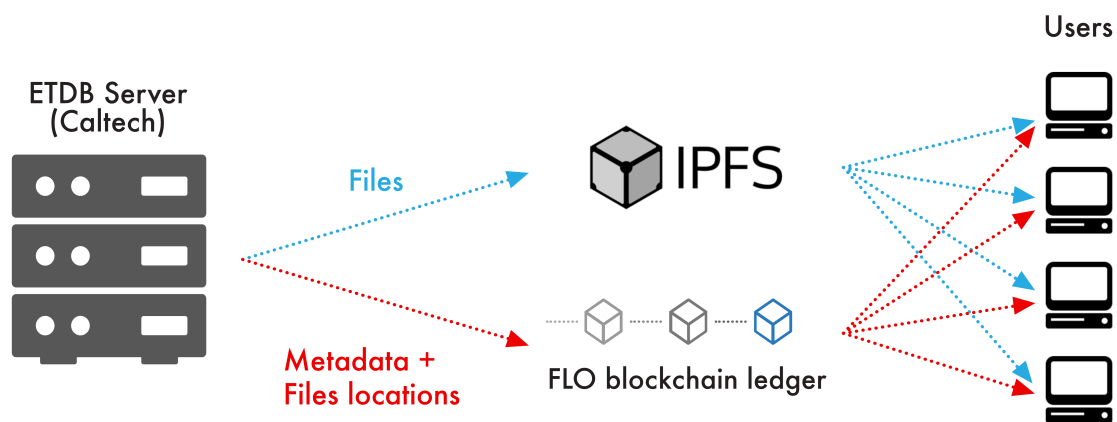
- 395 1. He, W. and He, Y. (2014). Electron tomography for organelles, cells, and tissues.  
396 *Methods Mol Biol* 1117, 445-83.
- 397 2. Ercius, P., Alaidi, O., Rames, M. J., and Ren, G. (2015). Electron Tomography: A Three-  
398 Dimensional Analytic Tool for Hard and Soft Materials Research. *Adv Mater* 27, 5638-  
399 63.
- 400 3. Oikonomou, C. M. and Jensen, G. J. (2017). Cellular Electron Cryotomography: Toward  
401 Structural Biology In Situ. *Annu Rev Biochem* 86, 873-896.
- 402 4. Elmlund, D., Le, S. N., and Elmlund, H. (2017). High-resolution cryo-EM: the nuts and  
403 bolts. *Curr Opin Struct Biol* 46, 1-6.
- 404 5. Patwardhan, A. (2017). Trends in the Electron Microscopy Data Bank (EMDB). *Acta*  
405 *Crystallogr D Struct Biol* 73, 503-508.
- 406 6. Callaway, E. (2015). The revolution will not be crystallized: a new method sweeps  
407 through structural biology. *Nature* 525, 172-4.
- 408 7. Prize, N. (2017). The 2017 Nobel Prize in Chemistry - Press Release (Nobelprize.org).

- 409 8. Marx, V. (2018). Calling cell biologists to try cryo-ET. *Nat Methods* 15, 575-578.
- 410 9. Frank, J. (2017). Advances in the field of single-particle cryo-electron microscopy over  
411 the last decade. *Nat Protoc* 12, 209-212.
- 412 10. Baldwin, P. R., Tan, Y. Z., Eng, E. T., Rice, W. J., Noble, A. J., Negro, C. J., Cianfrocco,  
413 M. A., Potter, C. S., and Carragher, B. (2017). Big data in cryoEM: automated collection,  
414 processing and accessibility of EM data. *Curr Opin Microbiol* 43, 1-8.
- 415 11. Patwardhan, A., Carazo, J. M., Carragher, B., Henderson, R., Heymann, J. B., Hill, E.,  
416 Jensen, G. J., Lagerstedt, I., Lawson, C. L., Ludtke, S. J., Mastronarde, D., Moore, W. J.,  
417 Roseman, A., Rosenthal, P., Sorzano, C. O., Sanz-Garcia, E., Scheres, S. H.,  
418 Subramaniam, S., Westbrook, J., Winn, M., Swedlow, J. R., and Kleywegt, G. J. (2012).  
419 Data management challenges in three-dimensional EM. *Nat Struct Mol Biol* 19, 1203-7.
- 420 12. Patwardhan, A., Ashton, A., Brandt, R., Butcher, S., Carzaniga, R., Chiu, W., Collinson,  
421 L., Doux, P., Duke, E., Ellisman, M. H., Franken, E., Grunewald, K., Heriche, J. K.,  
422 Koster, A., Kuhlbrandt, W., Lagerstedt, I., Larabell, C., Lawson, C. L., Saibil, H. R.,  
423 Sanz-Garcia, E., Subramaniam, S., Verkade, P., Swedlow, J. R., and Kleywegt, G. J.  
424 (2014). A 3D cellular context for the macromolecular world. *Nat Struct Mol Biol* 21, 841-  
425 5.
- 426 13. Berman, H. M., Burley, S. K., Chiu, W., Sali, A., Adzhubei, A., Bourne, P. E., Bryant, S.  
427 H., Dunbrack, R. L., Jr., Fidelis, K., Frank, J., Godzik, A., Henrick, K., Joachimiak, A.,  
428 Heymann, B., Jones, D., Markley, J. L., Moulton, J., Montelione, G. T., Orengo, C.,  
429 Rossmann, M. G., Rost, B., Saibil, H., Schwede, T., Standley, D. M., and Westbrook, J.  
430 D. (2006). Outcome of a workshop on archiving structural models of biological  
431 macromolecules. *Structure* 14, 1211-7.
- 432 14. Tagari, M., Newman, R., Chagoyen, M., Carazo, J. M., and Henrick, K. (2002). New  
433 electron microscopy database and deposition system. *Trends Biochem Sci* 27, 589.
- 434 15. Lawson, C. L., Baker, M. L., Best, C., Bi, C., Dougherty, M., Feng, P., van Ginkel, G.,  
435 Devkota, B., Lagerstedt, I., Ludtke, S. J., Newman, R. H., Oldfield, T. J., Rees, I., Sahni,  
436 G., Sala, R., Velankar, S., Warren, J., Westbrook, J. D., Henrick, K., Kleywegt, G. J.,  
437 Berman, H. M., and Chiu, W. (2011). EMDDataBank.org: unified data resource for  
438 CryoEM. *Nucleic Acids Res* 39, D456-64.
- 439 16. Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide Protein  
440 Data Bank. *Nat Struct Biol* 10, 980.
- 441 17. Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr., Brice, M. D., Rodgers,  
442 J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The Protein Data Bank. A  
443 computer-based archival file for macromolecular structures. *Eur J Biochem* 80, 319-24.
- 444 18. Editorial (2003). A database for 'em. *Nat Struct Biol* 10, 313.
- 445 19. Lagerstedt, I., Moore, W. J., Patwardhan, A., Sanz-Garcia, E., Best, C., Swedlow, J. R.,  
446 and Kleywegt, G. J. (2013). Web-based visualisation and analysis of 3D electron-  
447 microscopy data from EMDB and PDB. *J Struct Biol* 184, 173-81.
- 448 20. Salavert-Torres, J., Iudin, A., Lagerstedt, I., Sanz-Garcia, E., Kleywegt, G. J., and  
449 Patwardhan, A. (2016). Web-based volume slicer for 3D electron-microscopy data from  
450 EMDB. *J Struct Biol* 194, 164-70.
- 451 21. Orloff, D. N., Iwasa, J. H., Martone, M. E., Ellisman, M. H., and Kane, C. M. (2013).  
452 The cell: an image library-CCDB: a curated repository of microscopy data. *Nucleic Acids*  
453 *Res* 41, D1241-50.

- 454 22. Martone, M. E., Gupta, A., Wong, M., Qian, X., Sosinsky, G., Ludascher, B., and  
455 Ellisman, M. H. (2002). A cell-centered database for electron tomographic data. *J Struct*  
456 *Biol* 138, 145-55.
- 457 23. Martone, M. E., Zhang, S., Gupta, A., Qian, X., He, H., Price, D. L., Wong, M., Santini,  
458 S., and Ellisman, M. H. (2003). The cell-centered database: a database for multiscale  
459 structural and protein localization data from light and electron microscopy.  
460 *Neuroinformatics* 1, 379-95.
- 461 24. Martone, M. E., Tran, J., Wong, W. W., Sargis, J., Fong, L., Larson, S., Lamont, S. P.,  
462 Gupta, A., and Ellisman, M. H. (2008). The cell centered database project: an update on  
463 building community resources for managing and sharing 3D imaging data. *J Struct Biol*  
464 161, 220-31.
- 465 25. Iudin, A., Korir, P. K., Salavert-Torres, J., Kleywegt, G. J., and Patwardhan, A. (2016).  
466 EMPIAR: a public archive for raw electron microscopy image data. *Nat Methods* 13,  
467 387-8.
- 468 26. Henrick, K., Newman, R., Tagari, M., and Chagoyen, M. (2003). EMDep: a web-based  
469 system for the deposition and validation of high-resolution electron microscopy  
470 macromolecular structural information. *J Struct Biol* 144, 228-37.
- 471 27. Basler, M., Pilhofer, M., Henderson, G. P., Jensen, G. J., and Mekalanos, J. J. (2012).  
472 Type VI secretion requires a dynamic contractile phage tail-like structure. *Nature* 483,  
473 182-6.
- 474 28. Ding, H. J., Oikonomou, C. M., and Jensen, G. J. (2015). The Caltech Tomography  
475 Database and Automatic Processing Pipeline. *J Struct Biol* 192, 279-86.
- 476 29. Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System (White Paper).  
477 30. Shin, L. *The first government to secure land titles on the bitcoin blockchain expands*  
478 *project*. 2017 February 7, 2017 [cited 2018 August 7]; Available from:  
479 [https://www.forbes.com/sites/laurashin/2017/02/07/the-first-government-to-secure-land-](https://www.forbes.com/sites/laurashin/2017/02/07/the-first-government-to-secure-land-titles-on-the-bitcoin-blockchain-expands-project/)  
480 [titles-on-the-bitcoin-blockchain-expands-project/](https://www.forbes.com/sites/laurashin/2017/02/07/the-first-government-to-secure-land-titles-on-the-bitcoin-blockchain-expands-project/).
- 481 31. Cunningham, J. and Ainsworth, J. (2017). Enabling Patient Control of Personal  
482 Electronic Health Records Through Distributed Ledger Technology. *Stud Health Technol*  
483 *Inform* 245, 45-48.
- 484 32. Patel, V. (2018). A framework for secure and decentralized sharing of medical imaging  
485 data via blockchain consensus. *Health Informatics J.* 10.1177/1460458218769699,  
486 1460458218769699.
- 487 33. Bajpai, P. *How stock exchanges are experimenting with blockchain technology*. 2017  
488 June 12, 2017 [cited 2018 August 7]; Available from:  
489 [https://www.nasdaq.com/article/how-stock-exchanges-are-experimenting-with-](https://www.nasdaq.com/article/how-stock-exchanges-are-experimenting-with-blockchain-technology-cm801802)  
490 [blockchain-technology-cm801802](https://www.nasdaq.com/article/how-stock-exchanges-are-experimenting-with-blockchain-technology-cm801802).
- 491 34. FLO. *FLO*. [cited 2018 August 7]; Available from: <https://www.flo.cash/>.
- 492 35. *Open Index Protocol Wiki*. [cited 2018 August 7]; Available from: <https://oip.wiki/>.
- 493 36. Benet, J. (2014). IPFS - content addressed, versioned, P2P file system. *arXiv*  
494 arXiv:1407.3561.
- 495 37. Heymann, J. B., Chagoyen, M., and Belnap, D. M. (2005). Common conventions for  
496 interchange and archiving of three-dimensional electron microscopy information in  
497 structural biology. *J Struct Biol* 151, 196-207.
- 498 38. Dobro, M. J., Oikonomou, C. M., Piper, A., Cohen, J., Guo, K., Jensen, T., Tadayon, J.,  
499 Donermeyer, J., Park, Y., Solis, B. A., Kjaer, A., Jewett, A. I., McDowall, A. W., Chen,

- 500 S., Chang, Y. W., Shi, J., Subramanian, P., Iancu, C. V., Li, Z., Briegel, A., Tocheva, E.  
501 I., Pilhofer, M., and Jensen, G. J. (2017). Uncharacterized bacterial structures revealed by  
502 electron cryotomography. *J Bacteriol.* 10.1128/JB.00100-17.
- 503 39. Gutmanas, A., Oldfield, T. J., Patwardhan, A., Sen, S., Velankar, S., and Kleywegt, G. J.  
504 (2013). The role of structural bioinformatics resources in the era of integrative structural  
505 biology. *Acta Crystallogr D Biol Crystallogr* 69, 710-21.
- 506 40. Fellmann, D., Pulokas, J., Milligan, R. A., Carragher, B., and Potter, C. S. (2002). A  
507 relational database for cryoEM: experience at one year and 50 000 images. *J Struct Biol*  
508 137, 273-82.
- 509 41. Rees, I., Langley, E., Chiu, W., and Ludtke, S. J. (2013). EMEN2: an object oriented  
510 database and electronic lab notebook. *Microsc Microanal* 19, 1-10.
- 511 42. PDBe. *EMPIAR yearly data storage*. 2018 May 9, 2018]; Available from:  
512 [https://www.ebi.ac.uk/pdbe/emdb/statistics\\_empiar\\_yearly\\_size.html/](https://www.ebi.ac.uk/pdbe/emdb/statistics_empiar_yearly_size.html/).
- 513 43. PDBe. *EMPIAR entry releases*. 2018 [cited 2018 August 9]; Available from:  
514 [https://www.ebi.ac.uk/pdbe/emdb/statistics\\_empiar\\_entry\\_releases.html/](https://www.ebi.ac.uk/pdbe/emdb/statistics_empiar_entry_releases.html/).
- 515 44. PDBe. *EMDB map releases*. 2018 [cited 2018 August 9]; Available from:  
516 [https://www.ebi.ac.uk/pdbe/emdb/statistics\\_releases.html/](https://www.ebi.ac.uk/pdbe/emdb/statistics_releases.html/).
- 517

518 **Figures:**  
519 **Figure 1**



520

521 **Figure 2**


**Caltech**

[About](#)
[Browse Database](#)
[Featured Tomograms](#)
[Scientific Challenges](#)
[Contact](#)

## *Vibrio cholerae*

---



▶ 0:09 / 0:25 ⏮ ⏭ ⏴ ⏵

← Return to database

Share:

<b>Tilt Series date:</b> September 9th 2015
<b>Data Taken By:</b> Yiwei Chang
<b>Species / Specimen:</b> <i>Vibrio cholerae</i>
<b>Strain:</b> O395-N1
<b>Tilt Series Setting:</b> single axis, tilt range: (-60°, 60°), step: 1°, constant angular increment, dosage: 180eV/Å <sup>2</sup> , defocus: -8µm, magnification: 27500x.
<b>Microscope:</b> Caltech Polara
<b>Acquisition Software:</b> UCSFTomo
<b>Processing Software Used:</b> Raptor
<b>Notes:</b> Tilt series notes: Classical strain with ctxA deletion Cell harbors pMT5 plasmid (inducible toxT)

### Download files


---

#	Name	Size	Type	Download
1	20150909_AK_pMT15_10009.mrc	3.45 GB	Tilt series	<span style="border: 1px solid black; padding: 2px 10px;">DOWNLOAD</span>
2	20150909_AK_pMT15_10009_full.rec	534.53 MB	Reconstruction	<span style="border: 1px solid black; padding: 2px 10px;">DOWNLOAD</span>
3	keymov_yc2015-09-09-9.mp4	17.43 MB	Key movie	<span style="border: 1px solid black; padding: 2px 10px;">DOWNLOAD</span>
4	keymov_yc2015-09-09-9.flv	56.36 MB	Key movie	<span style="border: 1px solid black; padding: 2px 10px;">DOWNLOAD</span>
5	keyimg_yc2015-09-09-9.jpg	1.04 MB	Key image	<span style="border: 1px solid black; padding: 2px 10px;">DOWNLOAD</span>

522



## 523 Figure 3

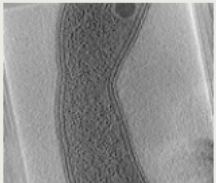
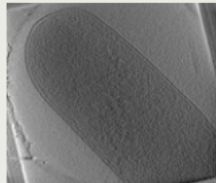
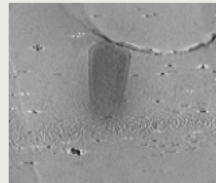
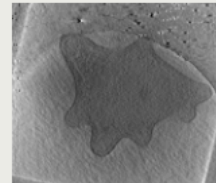
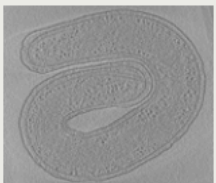
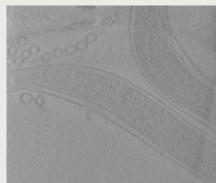
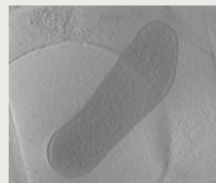
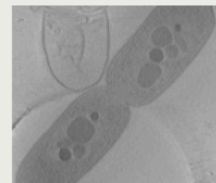
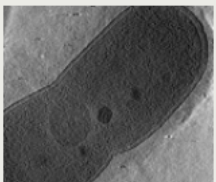
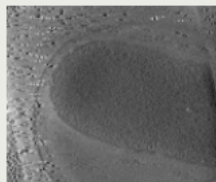
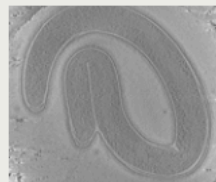
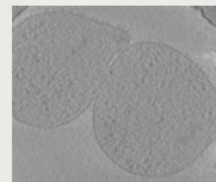
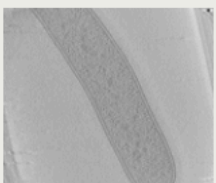

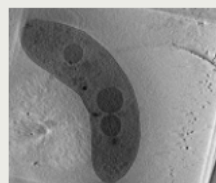

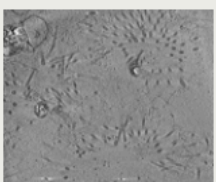
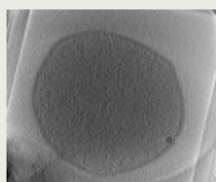
 [About](#) [Browse Database](#) [Featured Tomograms](#) [Scientific Challenges](#) [Contact](#)

### What's in a tomogram?

Many cell structures are visible by electron tomography.

Here are a few highlights from bacterial and archaeal cells.

Choose a category below to learn more and see examples.

 Cytoskeletal elements →	 Cell envelopes →	 Surface layers →	 Cell shapes →
 DNA →	 Outer membrane vesicles →	 Intracytoplasmic membranes →	 Carboxysomes →
 Storage granules →	 Gas vesicles →	 Flagella →	 Terminal organelles →
 Pili →	 Chemoreceptor arrays →	 Magnetosomes →	 Spores →
 Contractile injection machines →	 Phage →		

524