Photon-HDF5: an open file format for single-molecule fluorescence experiments using photon-counting detectors

A. Ingargiola, T. Laurence, S. Weiss, X. Michalet

Abstract

We introduce Photon-HDF5, an open and efficient file format to simplify exchange and long term accessibility of data from single-molecule fluorescence experiments based on photon-counting detectors such as single-photon avalanche diode (SPAD), photomultiplier tube (PMT) or arrays of such detectors. The format is based on HDF5, a widely used platform- and language-independent hierarchical file format for which user-friendly viewers are available. Photon-HDF5 can store raw photon data (timestamp, channel number, etc) from any acquisition hardware, but also setup and sample description, information on provenance, authorship and other metadata, and is flexible enough to include any kind of custom data.

The format specifications are hosted on a public website, which is open to contributions by the biophysics community. As an initial resource, the website provides code examples to read Photon-HDF5 files in several programming languages and a reference python library (*phconvert*), to create new Photon-HDF5 files and convert several existing file formats into Photon-HDF5. To encourage adoption by the academic and commercial communities, all software is released under the MIT open source license.

1. Introduction **2**

2. Current situation **3**

2.1 Photon-counting based experiments *3*

2.2 Examples of other formats *3*

3. Proposal of a new file format **4**

3.1 Design principles *4*

3.2 The parent HDF5 format *4*

3.3 Photon-HDF5 file format *5*

3.3.1 Photon-HDF5 structure 5

3.3.2 Additional file format features 7

3.3.3 Photon-HDF5 software tools 7

4. Scenarios of Photon-HDF5 usage **8**

Example 1 (Publication) *8*

Example 2 (Collaboration) *8*

Example 3 (Data Management) *8*

5. Conclusions & Perspectives **9**

1. Introduction

Single-molecule fluorescence techniques have facilitated ground-breaking developments in biophysics, as recently recognized by the Nobel committee (1). Techniques based on acquisition of long series of images obtained with ultra-sensitive cameras are the most popular among those, with numerous published software tools (2–4). A related set of single-molecule fluorescence techniques relying on photon-counting devices is used in studies of freely-diffusing single-molecules (5–7), studies of single-molecules in microfluidic mixers (8), optical or ABEL traps (9–11), or on immobilized molecules (12), where high temporal resolution (< 1 ms) is required. Among these techniques, single-molecule fluorescence resonant energy transfer (smFRET)(13), microsecond alternating laser excitation (µs-ALEX)(14), nanosecond ALEX (ns-ALEX)(15), also known as pulse interleaved excitation (PIE)(16), two-color coincident detection (TCCD)(17, 18), polarization anisotropy(19, 20), fluorescence correlation and cross-correlation spectroscopy (FCS, FCCS) (21, 22) or multi-parameter fluorescent detection (MFD)(23) are but a few in a constantly growing list.

Contrary to camera-based techniques, where most manufacturers provide a way to save data in a standardized format such as TIFF images, there is no such standardization for photon-counting hardware, where manufacturers of integrated systems have their own binary formats, and most systems are custom-built, including custom software generating oftentimes poorly documented binary files.

This situation creates a number of issues. First, it is rarely possible for an outsider (for instance a reviewer) to access the raw data of a published study without considerable coding work. Moreover, this assumes that the authors have provided a comprehensive description of their format. Worse, it is often difficult for different generations of students in a laboratory to decipher data file formats created by their predecessors. In summary, the proliferation of file formats impedes interoperability between different software, creates barriers to data sharing among researchers and undermines results reproducibility and long-term preservation of data. Moreover, with the growing trend by funding agencies to request data sharing and documentation, having multiple file formats will inevitably result in significant duplication of efforts by each group.

To help address these issues, we introduce Photon-HDF5(24), an open file format for single-molecule fluorescence experiments using photon-counting devices, including Time-Correlated Single-Photon Counting (TCSPC) data. The file format description comes accompanied with a reference Python library (*phconvert*)(25) to create Photon-HDF5 files from raw data or existing file formats, as well as examples(26) on how to read Photon-HDF5 files in Python, MATLAB and LabVIEW. Both format documentation and software are hosted on GitHub, a popular collaborative platform (27), with the aim of allowing the biophysics community to contribute to this project with comments, suggestions and code (28).

The article is organized as follows. In Section 2, we briefly review the current situation (characterized by a myriad of formats) and showcase a few successful examples of standardization in other scientific domains. In section 3, we examine requirements for an efficient, maintainable file format and introduce HDF5, a format on which Photon-HDF5 is based. We then describe the overall structure of Photon-HDF5 files and support software. In section 4 we illustrate the use of this format with a few scenarios, and, in

section 5, we conclude with a summary of Photon-HDF5 key features and with a call for the community to participate in shaping the evolution of the format.

2. Current situation

2.1 Photon-counting based experiments

Photon-counting based single-molecule fluorescence experiments can rapidly generate large amounts of data: in a typical experiment, count rates from a few kHz to several 10 kHz per detector and measurement durations of minutes are not uncommon and, increasingly, large number of individual detectors are encountered (29–31). Since each photon's information (time stamp, channel number, etc.) may occupy 64 or more bits per photon, most experiments store data in binary formats, which minimize disk space usage, at the expense of human readability. Because there is no natural or standardized structure for such a file format, each research group or company designing or putting together photon-counting instrumentation for single-molecule fluorescence experiments usually ends up developing its own custom binary file format.

In contrast with self-documenting text-based format, binary formats require detailed documentation of the file structure in order to be properly read (or written). Furthermore, extending a binary format to handle a new piece of information (such as modifications to the setup hardware) usually results in a new, incompatible format requiring specific handling and new, time-consuming code development. Since most researchers are, understandably, not inclined to invest significant resources in developing and documenting file formats, it is common to end up with a collection of poorly documented *ad hoc* binary formats within each individual research groups. The obvious risk is that, over the years, information about those file formats gets lost, effectively leading to data loss.

While vendors of photon-counting hardware and software have usually done a better job at designing and documenting file formats, their formats are nonetheless tightly bound to a specific hardware and therefore, generally not suitable to store data from a competitor's systems or from custom hardware used in individual laboratories. Additionally, although these formats can store some predefined metadata (such as for instance author, date of creation and comment), they are not extensible to accept additional metadata, necessary to describe a measurement in details.

The result of this situation is that it is difficult, if not impractical, for research groups not using identical hardware and software to exchange data to compare methods and results, and within each group, migration to new hardware or software is discouraged or results in incompatible data formats.

2.2 Examples of other formats

The need for standard file formats for domain-specific data sets is a common occurrence in many scientific fields. Most in the biophysics community will be familiar with the PDB file format for biomolecular structures maintained by the Protein Data Bank(32), or with the already mentioned TIFF format for images. The latter format, copyrighted by Adobe Systems, is only marginally adapted to the complexity of scientific data, but is used as a least common denominator export format by camera vendors in the absence of a better solution. The FITS format developed by the astronomy

community(33) (and supported by some camera manufacturers) is a good example of a powerful alternative, although its design principles are dated and it has not been adopted by the biophysics or microscopy communities.

In the single-molecule community, a text-based file format (SMD) was recently proposed to store data from a variety of surface-immobilized, camera-based experiments. These measurements consist of series of images, from which time traces (time binned data) are extracted for individual emitters identified in the series. These time traces have relatively short duration (limited by fluorophore bleaching) and therefore have minimal space requirements. The proposed SMD format based on JavaScript Object Notation (JSON)(34) is therefore an appropriate choice for these structured data sets of relatively small size. On the other end, using SMD for timestamp-based experiments, would be inefficient (due to its text-based nature) and result in irreversible information loss (due to the time binning).

3. Proposal of a new file format

3.1 Design principles

The design of a standard file format for long-term data preservation and to facilitate interoperability and data sharing, presents several (often contrasting) requirements.

Flexibility: the format must be capable of efficiently storing a variety of numeric data types and strings to represent both raw data and all metadata associated with an experiment. Moreover, the structure must be general enough to accommodate different types of time stamp-based single-molecule experiments. For instance, it should support experiments using continuous wave or pulsed excitation, one or many detection channels, one or more excitation wavelengths, etc. Since it would be impossible to predict all possible future technical developments, it should be easily extensible and customizable to accommodate them.

Ease of use: in order to encourage adoption and facilitate long-term maintainability, the file structure and complexity should be kept as low as possible. It should be possible for a user to quickly write a reader for the file format in any programming language and operating system of choice. Furthermore, there should be tools to facilitate creation of data files that complies with the specifications.

Small size: while disk storage is increasingly inexpensive, the advantages of a compact file format should not be overlooked. Smaller files are easier to backup, archive online and share as well as generally faster to read. The format must therefore support smart compression allowing fast read and write, and partial loading of large data sets.

3.2 The parent HDF5 format

HDF5 is an existing open hierarchical file format satisfying all requirements listed previously. In the following, we provide a brief overview of the HDF5 format. Detailed information can be found on the HDF Group website(35).

HDF5 is a general-purpose binary format designed to store numerical data sets, widely used in industry and in several scientific communities. HDF5 development is led by the non-profit HDF Group, which also maintains open-source platform-independent libraries to read and write HDF5 files in multiple languages.

HDF5 has a hierarchical (tree-like) data structure similar to files and folders in a file system: groups are the equivalent of folders, whereas data fields (e.g. arrays) are the equivalent of files. The main group of a HDF5 file is called "root" (indicated by the slash character "/") and can contain data fields or other groups. The HDF5 format is self-describing, meaning that a user does not need to know the hierarchical structure or the data types before reading the file. For example, in order to access a numerical array in a HDF5 file, the user only needs to specify its path: the array will be automatically loaded into memory using the appropriate data type. In other words, all the byte-level details (such as byte order and data type width) are transparently handled, greatly simplifying data reading and writing. Moreover, HDF5 transparently supports compression, allowing reducing file size while maintaining high reading speed without any added complexity for the user. Finally, each node in a HDF5 file (groups or data fields) can have any number of attributes, such as descriptions, which can be used to embed metadata to a dataset.

Libraries to operate on HDF5 files currently exist for the major programming languages used in scientific computing, such as C, Fortran, C++, Java, Python, R, MATLAB and LabVIEW. Furthermore, a free multi-platform graphical viewer and editor (HDFView (36)) allows users to quickly explore, modify or export content of HDF5 files. The availability of well-maintained open source libraries allows users to reliably manage HDF5 files without the need to deal with the underlying byte-level structure, effectively overcoming the drawbacks of binary formats highlighted in section 2.

HDF5 has been widely adopted in industry and in several scientific communities as the basis for specialized data formats. Because of its openness, popularity and long-term support, it represents a safe choice for long-term data preservation.

3.3 Photon-HDF5 file format

HDF5 is the foundation for the proposed format for photon-counting single-molecule fluorescence data, therefore called "Photon-HDF5" file format.

Photon-HDF5 is, essentially, a conventional structure to store timestamp-based single-molecule fluorescence data in HDF5 files (Fig. 1). Since Photon-HDF5 files are HDF5 files, the requirements of an open, efficient, well-documented and platform- and language-independent are automatically fulfilled.

We now briefly describe the main components of a Photon-HDF5 file. The interested reader is invited to visit the *photon-hdf5.org* website for further information and a complete description of the format and associated tools.

3.3.1 Photon-HDF5 structure

The Photon-HDF5 file structure contains 5 main groups (*photon_data*, *setup*, *sample*, *identity* and *provenance*) and two root fields: a general description string (*description*) and the acquisition duration (*acquisition_duration*). The general structure is illustrated in Fig. 2 (see also Fig. SI-1 and SI-2), while Fig. 1 shows a Photon-HDF5 file as it appears when opened with the HDFView visualizer.

a. The *photon_data* group

As visible in Fig. 1 (and detailed in Fig. SI-1), the *photon_data* group contains the raw data (*timestamps*, *detectors* and *nanotimes* arrays), as well as information needed to analyze it (in the *measurement_specs* subgroup). All arrays in *photon_data* have the same length, equal to the number of recorded photons. Information corresponding to the ith photon (its macrotime stamp, detector identity, and when applicable, the microtime measured by TCSPC hardware, or nanotime) is stored the ith value of each of these arrays.

Except for *timestamps*, the other arrays are present in a file only when needed, i.e. the *detectors* array is present only if there is more than one detector and the *nanotimes* array is present only if TCSPC timing data is recorded.

Inside *photon_data*, the *measurement_specs* group contains the measurement type and a type-dependent set of metadata fields. The currently supported measurement types are listed in Table SI-1. Each measurement type (specified in *measurement_type*) requires a particular set of fields which completely describe the experiment. For example, a single-laser single-molecule FRET data file will have a "smFRET" measurement type, and will include two numeric fields called *spectral_ch1* and *spectral_ch2* containing the donor and acceptor detector number.

Defining a new measurement type consists in choosing a "name" and a set of required fields in the *measurements_specs* group. Since measurement types are mutually exclusive, new measurement types can be introduced without interfering with the previously defined ones (see Fig. 2). This approach allows the community to gradually (and safely) extend the file format to support new measurement types.

The data contained in *measurement_specs* group provides all the necessary information needed to interpret the raw experimental data. Files not including a *measurement_specs* group are still valid Photon-HDF5 files but their content cannot be analyzed by a user without knowing additional measurement specifications.

b. The *setup* group

The *setup* group stores setup properties such as laser excitation wavelengths and powers, the number of detected spectral bands (e.g. 2-colors or 3-colors), the number of detected polarization states (e.g. none = 1 and 2 polarizations = 2) and the number of "split channels" (i.e. identical spectral or polarization channels split by non-polarizing beam-splitters).

c. Metadata groups

In addition to the previous groups, a Photon-HDF5 file may contain 3 other groups (*sample*, *identity* and *provenance*) which provide additional optional metadata (see Fig. 2, & SI-2).

The *sample* group contains a description of the sample (*sample_name*), the buffer (*buffer_name*) and the fluorophore names (*dye_names*). Storing this basic sample information with the data simplifies preservation of experimental details.

The *identity* group contains identification metadata such as dataset *author* (and its affiliation), Photon-HDF5 format version, software used to save the file (and its version), and optional dataset identifiers such as DOI or URL. This information is important when sharing datasets online, in order to keep track of authorship and provide proper credit.

As discussed in Section 4, a Photon-HDF5 file might be created by converting an original data file saved by an acquisition (or simulation) software outputting some custom file format. The *provenance* group is the place to store this information, such as the name and creation date of the original file, and the software used to save it.

3.3.2 Additional file format features

a. Field data types: Leveraging the self-describing nature of HDF5, instead of prescribing specific data types, Photon-HDF5 specifications only indicate whether a field is a scalar, an array or a string and whether the data type is integer or floating point. For examples, the *detectors* array (array of detector IDs) typically uses an unsigned 8-bit integer type, which is sufficient for the majority of setups. However, when using SPAD arrays with more than 256 pixels, a 16- or 32-bit integer data type can be used without any modification to the file format.

b. Custom user groups: For maximum flexibility, users can store arbitrary amounts of additional (custom) data in any position within the Photon-HDF5 hierarchy. To avoid incompatibilities with future versions of the format, we require that custom data be stored inside a group (or subgroup) named *user*.

c. Future Photon-HDF5 versions: In order to guarantee long-term accessibility to Photon-HDF5 files, future versions of the file format software will be backward-compatible. Specifically, previously defined fields will never be renamed or have their meaning modified. This ensures that software written to read an old version will be still capable to correctly read a newer version, although it will not be able to interpret new fields or measurement types.

3.3.3 Photon-HDF5 software tools

In order to facilitate the use and adoption of the Photon-HDF5 file format, we provide examples of how to read Photon-HDF5 files in several languages (currently Python, MATLAB and LabVIEW) (26).

We developed and maintain an open-source python library to create Photon-HDF5 files, *phconvert*, serving as reference implementation for the Photon-HDF5 format(25). When writing Photon-HDF5 files, *phconvert* validates the input data (making sure that all required fields are present and have correct name and type) and adds a description to each field (see Fig. 1 for an example).

In addition, *phconvert* includes a browser-based interface using Jupyter Notebooks(37, 38) to convert a few existing proprietary file formats into Photon-HDF5. The formats currently supported are HT3 (a file format created by PicoQuant for data acquired with their TCSPC hardware) and SPC/SET (a file format

created by Becker & Hickl for data acquired with their TCSPC hardware) as well as SM, a legacy file format developed by the Weiss Lab. Additional file formats can be easily added in the future. Examples of such data files and their converted Photon-HDF5 version are provided through the Photon-HDF5 website.

4. Scenarios of Photon-HDF5 usage

In this section, we illustrate a few use cases where employing the Photon-HDF5 format may lead to significant advantages.

Researchers performing (or interested in) photon-counting based single-molecule experiments can be classified in the following 3 categories:

a. Commercial acquisition hardware and software (Alice)
b. Custom acquisition hardware and software (Bob)
c. Mixture of a and b (Charlie)

In each case, Alice, Bob and Charlie use one or more custom or proprietary binary formats to store data from their experiments, and will use a custom or commercial software to analyze their data. We now examine 3 typical situations encountered by researchers.

Example 1 (Publication)

Alice wants to publish an article including her data so that as many researchers as possible can look at it (including the reviewers, who may not, however, have time to process it and might just be curious to take a peek at it).

The Photon-HDF5 format and HDFViewer will to the job for the reviewer and researchers using a different system, while the original format will allow only readers having the same commercial software to have a look at the data. If she uses one of the two most common commercial formats, she can use phconvert to create the Photon-HDF5 version. If not, writing code to create a Photon-HDF5 version should be simple thanks to the examples provided. Moreover, her code could be added to phconvert for the benefit of the entire community.

Example 2 (Collaboration)

Alice and Bob want to collaborate on a project, Alice doing the experiments and Bob performing some custom analysis with his software.

Since Bob doesn't use Alice's system, he would have to decipher the proprietary file format description of the vendor and extract the relevant data for his analysis. By asking Alice to first convert her file to the Photon-HDF5 format with phconvert, he will benefit from file size reduction, simple access to the data, and by incorporating the necessary code in his software, he will gain access to a variety of other files generated by other users of Photon-HDF5.

Example 3 (Data Management)

Charlie works with a student running some experiments on a commercial system, which he wants to compare to similar experiments he is carrying out on a new setup he has built with custom hardware and software.

Since Charlie needs to save new data, he needs a new format. But he also needs to read data generated by the student. The least effort consists in using a format which is efficient, well-documented and supported and in which his student's files can be easily converted. Photon-HDF5 naturally fits the bill, and additionally provides him with the same benefits enjoyed by Alice and Bob for publication and sharing.

While these scenarios do not directly involve manufacturers of single-photon counting hardware and their associated software, it is clear that manufacturers should take some interest in an effort such as Photon-HDF5. While it is not expected that they would abandon their proprietary format, their contribution to the definition of future versions of the format or to conversion tools is welcome and encouraged, as is that of anyone interested in the research community.

5. Conclusions & Perspectives

We have developed the Photon-HDF5 file format as a hardware-agnostic format for single-molecule fluorescence data based on photon-counting detectors. Photon-HDF5 is easy to use, customize and extend, in addition to inheriting all the advantages of the HDF5 format (open standard, self-described and efficient).

The format has been designed both for interoperability between different analysis software and long-term data archival. For this purpose, it includes a rich set of metadata to make self-contained measurement data files. Photon-HDF5 facilitates data sharing between research groups by reducing the efforts needed to decode the wide variety of binary formats used across laboratories.

While we have only discussed standard single-spot confocal experiments in the text, the format supports and is currently used for multi-spot experiments using multi-pixel SPAD arrays(24). Further extensions to accommodate future developments or new measurement types can be naturally incorporated.

Currently, the open source FRETBursts software package for smFRET burst analysis supports reading Photon-HDF5 files(39). Therefore, converting smFRET data to Photon-HDF5 already enables researchers to analyze a variety of data with an alternative software. Other software packages supporting Photon-HDF5 files are expected to be released in the near future.

The impact of Photon-HDF5 format will be determined by the extent of its adoption in the single-molecule community and its support by scientific publishers. To maximize the involvement of third parties, we embraced an open development model for both the format specification and the support software. Users are encouraged to send feedback (e.g. on perceived limitations or to add support for new measurement types) or other contributions. To facilitate the process, we include guidelines for contributors directly in the Photon-HDF5 reference documentation(24). Finally, by adopting an open consensus-based governance, we aim to empower the single-molecule community to determine the

future directions of Photon-HDF5 format in order to become a standard widely-useful tool for data sharing and preservation.

Author Contributions

AI designed the file format, implemented the supporting software, wrote the reference documentation, websites and manuscript; TL designed the file format, implemented reading examples, wrote the manuscript; SW wrote the manuscript; XM designed the file format, implemented reading examples, wrote the manuscript.

Acknowledgements

We thank Robert Boutelle for writing the example MATLAB script to read Photon-HDF5 files, Sangyoon Chung and Eitan Lerner for providing experimental data files. This work was supported in part by NIH Grant R01-GM95904 and by DOE Grant DE-FC02-02ER63421-00.

Figure 1: A Photon-HDF5 as seen with the HDFView visualizer. In contrast with other binary formats, Photon-HDF5 files (or any other HDF5 file) can be examined with an open source visualizer (HDFView). Note that the file structure (left pane) is immediately intelligible to the user, who can navigate it and select items for display or export. The time stamps array is selected in the left pane and its content displayed in the right pane. In the bottom pane, the TITLE attribute, among others, shows a description for the selected item (time stamps). Photon-HDF5 files include descriptions for every field facilitating interactive file exploration.

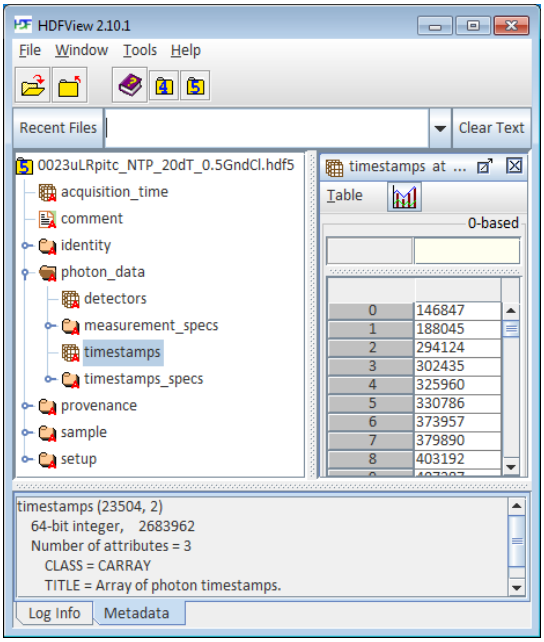
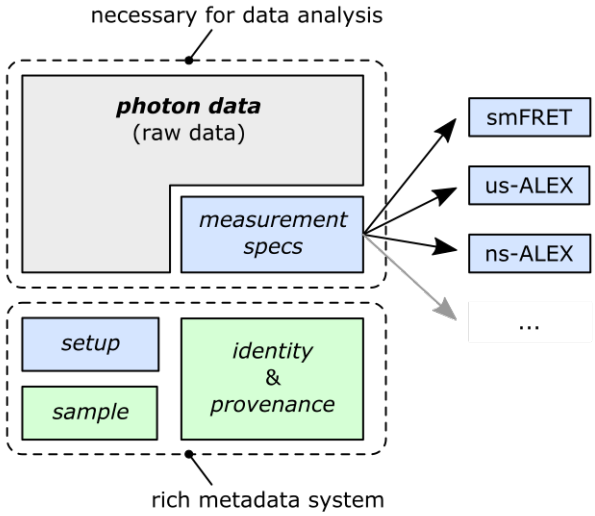


Figure 2. Photon-HDF5 file structure. The core *photon_data* group contains the raw data. A file containing only raw data is still a valid Photon-HDF5 file but it cannot be analyzed without additional information on what the data represents. The optional (but recommended) *measurement_specs* group identifies the type of measurement and specifies the associated metadata (i.e. role of each detector, measurement parameters, etc). When *measurement_spec* is present, a user has all the information to analyze the data. The other groups enhance the data file, providing important measurement metadata. *setup* contains information on the system used for the acquisition (e.g. excitation wavelengths, power, excitation scheme, etc). *sample* contains sample information (dyes, buffer, etc…). The last two groups *identity* and *provenance* provide author and file information.



References

1. Lippincott-Schwartz, J. 2015. Profile of Eric Betzig, Stefan Hell, and W. E. Moerner, 2014 Nobel Laureates in Chemistry. Proc. Natl. Acad. Sci. U. S. A. 112: 2630–2.

2. McKinney, S.A., C. Joo, and T. Ha. 2006. Analysis of single-molecule FRET trajectories using hidden Markov modeling. Biophys. J. 91: 1941–51.

3. Van de Meent, J.-W., J.E. Bronson, C.H. Wiggins, and R.L. Gonzalez. 2014. Empirical Bayes methods enable advanced population-level analyses of single-molecule FRET experiments. Biophys. J. 106: 1327–37.

4. Greenfeld, M., D.S. Pavlichin, H. Mabuchi, and D. Herschlag. 2012. Single Molecule Analysis Research Tool (SMART): an integrated approach for analyzing single molecule data. PLoS One. 7: e30024.

5. Fries, J.R., L. Brand, C. Eggeling, M. Köllner, and C.A.M. Seidel. 1998. Quantitative Identification of Different Single Molecules by Selective Time-Resolved Confocal Fluorescence Spectroscopy. J. Phys. Chem. A. 102: 6601–6613.

6. Dahan, M., A.A. Deniz, T. Ha, D.S. Chemla, P.G. Schultz, and S. Weiss. 1999. Ratiometric measurement and identification of single diffusing molecules. Chem. Phys. 247: 85–106.

7. Boukobza, E., A. Sonnenfeld, and G. Haran. 2001. Immobilization in Surface-Tethered Lipid Vesicles as a New Tool for Single Biomolecule Spectroscopy. J. Phys. Chem. B. 105: 12165–12170.

8. Lipman, E.A., B. Schuler, O. Bakajin, and W.A. Eaton. 2003. Single-molecule measurement of protein folding kinetics. Science. 301: 1233–1235.

9. Fields, A.P., and A.E. Cohen. 2011. Electrokinetic trapping at the one nanometer limit. Proc. Natl. Acad. Sci. U. S. A. 108: 8937–42.

10. Wang, Q., R.H. Goldsmith, Y. Jiang, S.D. Bockenhauer, and W.E. Moerner. 2012. Probing single biomolecules in solution using the anti-Brownian electrokinetic (ABEL) trap. Acc. Chem. Res. 45: 1955–64.

11. McHale, K., A.J. Berglund, and H. Mabuchi. 2007. Quantum dot photon statistics measured by three-dimensional particle tracking. Nano Lett. 7: 3535–9.

12. Yang, H., G. Luo, P. Karnchanaphanurach, T.-M. Louie, I. Rech, S. Cova, L. Xun, and X.S. Xie. 2003. Protein conformational dynamics probed by single-molecule electron transfer. Science. 302: 262–6.

13. Weiss, S. 1999. Fluorescence spectroscopy of single biomolecules. Science. 283: 1676–1683.

14. Kapanidis, A.N., N.K. Lee, T.A. Laurence, S. Doose, E. Margeat, and S. Weiss. 2004. Fluorescence-aided molecule sorting: analysis of structure and interactions by alternating-laser excitation of single molecules. Proc. Natl. Acad. Sci. U. S. A. 101: 8936–41.

15. Laurence, T.A., X. Kong, M. Jäger, and S. Weiss. 2005. Probing structural heterogeneities and fluctuations of nucleic acids and denatured proteins. Proc. Natl. Acad. Sci. U. S. A. 102: 17348–17353.

16. Müller, B.K., E. Zaychikov, C. Bräuchle, and D.C. Lamb. 2005. Pulsed interleaved excitation. Biophys. J. 89: 3508–22.

17. Orte, A., R. Clarke, S. Balasubramanian, and D. Klenerman. 2006. Determination of the fraction and stoichiometry of femtomolar levels of biomolecular complexes in an excess of monomer using single-molecule, two-color coincidence detection. Anal. Chem. 78: 7707–15.

18. Clarke, R.W., A. Orte, and D. Klenerman. 2007. Optimized threshold selection for single-molecule two-color fluorescence coincidence spectroscopy. Anal. Chem. 79: 2771–7.

19. Ha, T., T. Enderle, D.S. Chemla, P.R. Selvin, and S. Weiss. 1996. Single Molecule Dynamics Studied by Polarization Modulation. Phys. Rev. Lett. 77: 3979–3982.

20. Schaffer, J., A. Volkmer, C. Eggeling, V. Subramaniam, G. Striker, and C.A.M. Seidel. 1999. Identification of Single Molecules in Aqueous Solution by Time-Resolved Fluorescence Anisotropy. J. Phys. Chem. A. 103: 331–336.

21. Rigler, R., Ü. Mets, J. Widengren, and P. Kask. 1993. Fluorescence correlation spectroscopy with high count rate and low background: analysis of translational diffusion. Eur. Biophys. J. 22.

22. Schwille, P., F.J. Meyer-Almes, and R. Rigler. 1997. Dual-color fluorescence cross-correlation spectroscopy for multicomponent diffusional analysis in solution. Biophys. J. 72: 1878–1886.

23. Eggeling, C., S. Berger, L. Brand, J.R. Fries, J. Schaffer, A. Volkmer, and C.A.M. Seidel. 2001. Data registration and selective single-molecule analysis using multi-parameter fluorescence detection. J. Biotechnol. 86: 163–180.

24. Photon-HDF5 Home Page. http://photon-hdf5.org.

25. Ingargiola, A. phconvert - Library and converter for Photon-HDF5 files. http://photon-hdf5.github.io/phconvert.

26. Reading Photon-HDF5 in multiple languages. https://github.com/Photon-HDF5/photon_hdf5_reading_examples.

27. Photon-HDF5 project page on GitHub. https://github.com/Photon-HDF5.

28. Photon-HDF5: Guidelines for contributors. http://photon-hdf5.readthedocs.org/en/latest/contributing.html.

29. Rech, I., S. Marangoni, D. Resnati, M. Ghioni, and S. Cova. 2009. Multipixel single-photon avalanche diode array for parallel photon counting applications. J. Mod. Opt. 56: 326–333.

30. Laurence, T.A., S. Ly, F. Bourguet, N.O. Fischer, and M.A. Coleman. 2014. Fluorescence correlation spectroscopy at micromolar concentrations without optical nanoconfinement. J. Phys. Chem. B. 118: 9662–9667.

31. Michalet, X., A. Ingargiola, R.A. Colyer, G. Scalia, S. Weiss, P. Maccagnani, A. Gulinatti, I. Rech, and M. Ghioni. 2014. Silicon Photon-Counting Avalanche Diodes for Single-Molecule Fluorescence Spectroscopy. IEEE J. Sel. Top. Quantum Electron. 20: 1–20.

32. Berman, H., K. Henrick, and H. Nakamura. 2003. Announcing the worldwide Protein Data Bank. Nat. Struct. Biol. 10: 980.

33. FITS, The Astronomical Image Table Format. http://fits.gsfc.nasa.gov/.

34. Greenfeld, M., J.-W. van de Meent, D.S. Pavlichin, H. Mabuchi, C.H. Wiggins, R.L. Gonzalez, and D. Herschlag. 2015. Single-molecule dataset (SMD): a generalized storage format for raw and processed single-molecule data. BMC Bioinformatics. 16: 3.

35. HDF5 Homepage. https://www.hdfgroup.org/HDF5/.

36. HDFView Homepage. https://www.hdfgroup.org/products/java/release/download.html.

37. Shen, H. 2014. Interactive notebooks: Sharing the code. Nature. 515: 151–2.

38. OpenTechSchool: Introducing IPython Notebook. http://opentechschool.github.io/python-data-intro/core/notebook.html.

39. Ingargiola, A. FRETBursts - Burst analysis for smFRET experiments. http://tritemio.github.io/FRETBursts/.

SUPPORTING INFORMATION

Photon-HDF5: an open file format for single-molecule fluorescence experiments using photon-counting detectors

Antonino Ingargiola, Ted Laurence, Shimon Weiss, Xavier Michalet

1. Photon-HDF5 files examples

In this section, we describe a few examples of Photon-HDF5 files corresponding to different types of measurement, focusing for brevity on the mandatory fields, which are all found in */photon_data*. The full list of fields optionally present in the remaining groups (*/setup*, */sample*, */identity* and */provenance*) can be found in the reference documentation.

Photon-HDF5 files can be distinguished from other HDF5 files by the presence of two root node attributes *format_name* and *format_version*: the former must be the string "Photon-HDF5" and the latter is a string identifying the version. The measurement type is identified by reading the string in */photon_data/measurement_specs/measurement_type*, whose currently supported values are reported in Table SI-1.

1.1 us-ALEX

In us-ALEX smFRET data files, the following data fields are mandatory (their definition can be found on the photon-hdf5.org website):

1. Timestamps (array): */photon_data/timestamps*
2. Timestamps unit (scalar): */photon_data/timestamps_specs/timestamps_unit*
3. Detectors array (array): */photon_data/detectors*
4. Alternation period (scalar): */photon_data/measurement_specs/alex_period*
5. Donor detector ID (scalar): */photon_data/measurement_specs/detectors_specs/spectral_ch1*
6. Acceptor detector ID (scalar): */photon_data/measurement_specs/detectors_specs/spectral_ch*2
7. Donor excitation period range (2-elements array): */photon_data/measurement_specs/alex_excitation_period1*
8. Acceptor excitation period range (2-elements array): */photon_data/measurement_specs/alex_excitation_period2*

1.2 ns-ALEX

In ns-ALEX smFRET (also known as PIE) data files, in addition to the fields listed above for us-ALEX, the following additional data fields are mandatory mandatory (their definition can be found on the photon-hdf5.org website):

1. TCSPC nanotimes (array): */photon_data/nanotimes*
2. TCSPC bin width (scalar): */photon_data/nanotimes_specs/tcspc_unit*
3. TCSPC number of bins (scalar): */photon_data/nanotimes_specs/tcspc_num_bins*
4. Nanotimes time inversion (boolean): */photon_data/nanotimes_specs/time_reversed*
5. Laser pulse rate (scalar): */photon_data/measurement_specs/laser_repetition_rate*
6. Donor excitation TCSPC range (n-tuple): */photon_data/measurement_specs/ alex_excitation_period1*
7. Acceptor excitation TCSPC range (n-tuple): */photon_data/measurement_specs/ alex_excitation_period2*

Table SI-1: List of currently supported measurement types. Additional types can be added in the future based on user demand. The first column represents the string identifying the measurement type that is located at */photon_data/measurement_specs/measurement_type* in the Photon-HDF5 file. The second column is a brief description of the measurement type. The third column lists the required metadata fields in */photon_data/measurement_specs/* for each type of measurement.

Measurement type	**Description**	**Required metadata fields**
smFRET	2-colors smFRET with single wavelength excitation	• *detectors_specs/spectral_ch1* • *detectors_specs/spectral_ch2*
smFRET-usALEX	2-colors smFRET with 2 alternating CW excitation lasers	• *detectors_specs/spectral_ch1* • *detectors_specs/spectral_ch2* • *alex_period* • *alex_offset* • *alex_excitation_period1* • *alex_excitation_period2*
smFRET-usALEX-3c	3-colors smFRET with 3 alternating 3CW excitation lasers	• *detectors_specs/spectral_ch1* • *detectors_specs/spectral_ch2* • *detectors_specs/spectral_ch3* • *alex_period* • *alex_offset* • *alex_excitation_period1* • *alex_excitation_period2* • *alex_excitation_period3*
smFRET-nsALEX	2-colors smFRET with 2 interleaved pulsed lasers excitation and TCSPC recording. Also known as PIE.	• *detectors_specs/spectral_ch1* • *detectors_specs/spectral_ch2* • *laser_pulse_rate* • *alex_excitation_period1* • *alex_excitation_period2*

Figure SI-1: Partial diagram description of the Photon-HDF5 layout comprising the *photon_data* and *setup* groups, which contains the acquisition data and all the information needed for data processing. The symbol preceding each field indicates the data type of that field (e.g. array, string, scalar or boolean), as described in the legend.

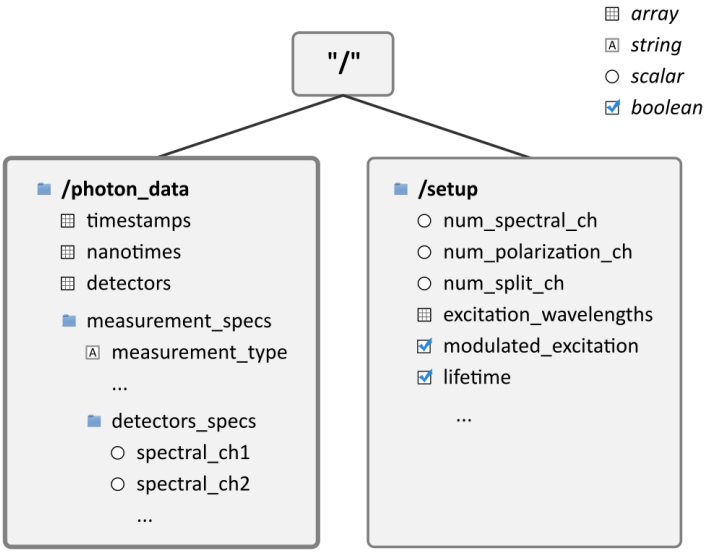
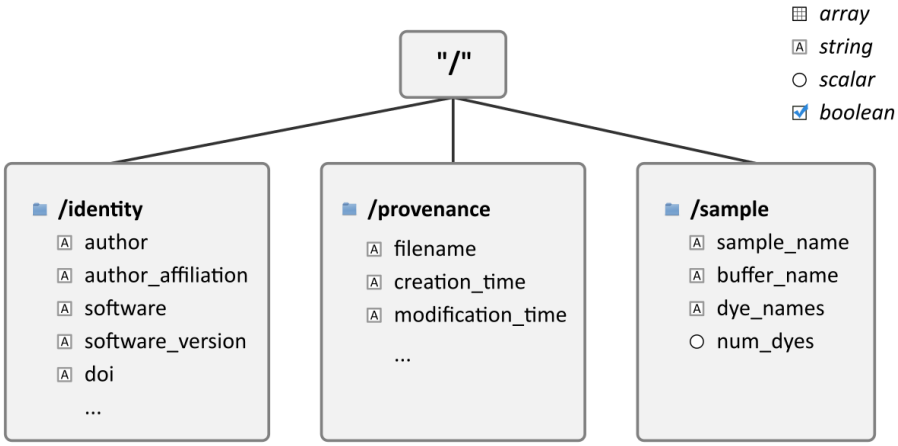


Figure SI-2: Partial diagram description of the Photon-HDF5 layout showing the *sample*, *identity* and *provenance* groups. These groups are not strictly necessary to analyze the data, but are important for long-term preservation, reproducibility and proper credit/citation when the file is shared.



2. Benchmarks

2.1 Overview

In this section, we report the results of two simple read- and write-speed benchmarks for two different data files: one based on a HT3 file (created by the MicroTime 200 hardware sold by PicoQuant) and one based on a SPC file (created by a SPC-630 board sold by Becker & Hickl). The benchmarks were coded in Python 2.7 and pytables 3.1 and run on a desktop PC with Intel® Core™ i5-3570K CPU, 16GB of RAM and a 7,200 rpm SATA hard drive (Western Digital WD2002FAEX). The quoted values are the result of a single execution after a system reboot (to eliminate perturbation from file-system cache).

It should be noted that the value reported are only indicative and cannot be generalized because compression performances (size and speed) can significantly vary for file to file. Furthermore HDF5 allows changing a parameter, the array chunk size (the size of an on-disk-contiguous chunk of data), which can affect (in particular, improve) performance. Here, we used the default value as set by the pytables library.

We benchmarked different compression levels ranging from 0 (no compression) to 9 (maximum compression) and 2 compression algorithms: zlib (HDF5 default) and blosc (a high-speed alternative compressor developed by the pytables team and not shipped by default with the HDF5 library). It is important to note that, when using languages other than python, using blosc requires compiling the blosc sources, which can represents a significant obstacle for the user. Therefore, we discourage the use of the blosc compressor for any file which needs to be shared. On the other hand, we included it in the benchmarks due to its impressive performance. For instance, in critical, real-time applications where the hard drive write-speed is a bottleneck (i.e. acquiring from large SPAD arrays), the blosc compressor could be used and the conversion to zlib postponed to a second, post-processing step.

2.2 Results

Read speeds were relatively uniform at all compression levels when using zlib, with values around 40 million photons per second (MP/s). In all examples, the read speed from compressed Photon-HDF5 format (compression level of 5) was around twice faster than reading from the native formats (the latter have an overhead due to byte decoding and rollover correction). It is worth noting that the read speed of the native formats depends on the specific decoding implementation, and C routines could potentially result in faster read speed (although we believe our implementation is reasonably fast).

Write speeds, on the contrary, were strongly affected by the compression level (5-10 MP/s using zlib5). Obviously, the fastest was to write uncompressed data (> 100 MP/s). Remarkably, blosc exhibited write speeds which were comparable to the uncompressed one, while significantly reducing the file size. For single-spot data files, the write speed was more than enough to save data in real time even when using the slow zlib compression. For setup using large detector arrays, it would be advisable to not use any compression or only blosc compression (and later convert the file to zlib compression for sharing).

Figure SI-3: Benchmark using a PicoQuant HT3 input file. In the first two graphs (speed), larger values are better, while in the last (size), smaller is better.

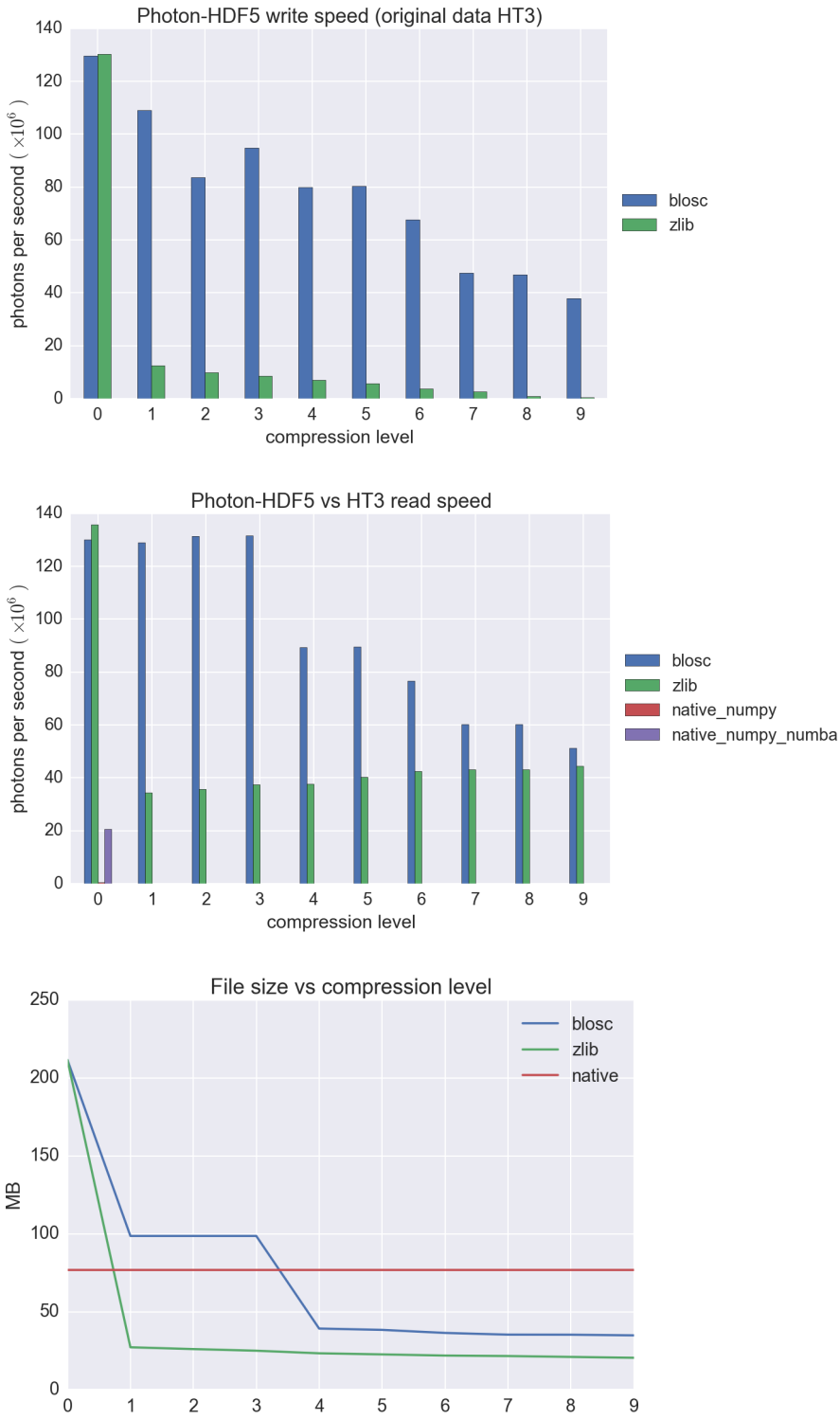
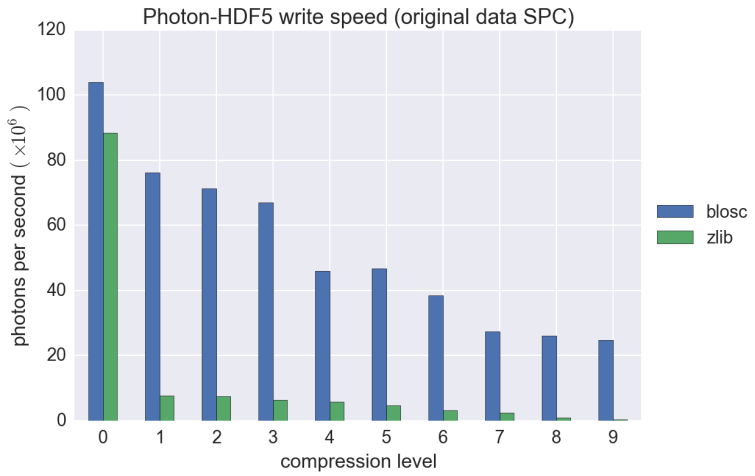
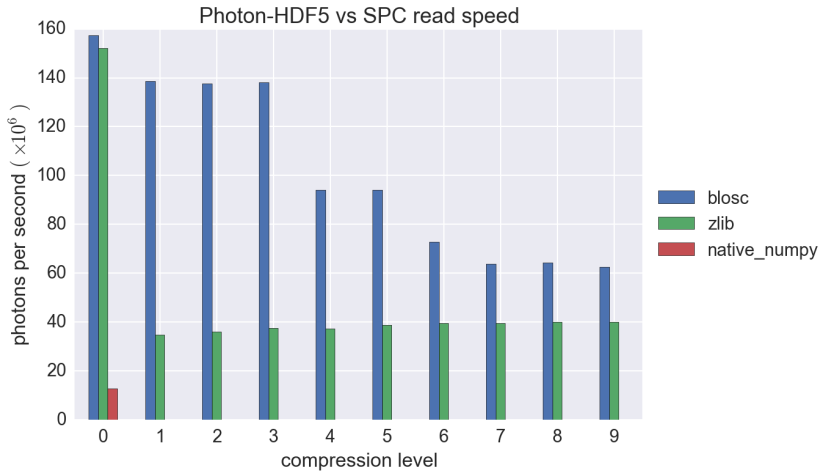
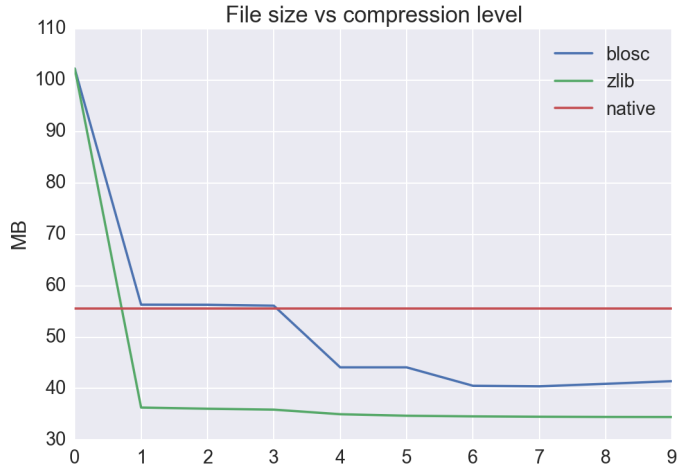



Figure SI-4: Benchmark using a Becker & Hickl SPC input file. In the first two graphs (speed), larger values are better, while in the last (size), smaller is better.