



Data Management in Research Groups

Gabriela Schlau-Cohen
Haslam-Dewey Professor of Chemistry
MIT



THE
SCHLAU-COHEN
LAB



U.S. DEPARTMENT OF
ENERGY

Office of
Science

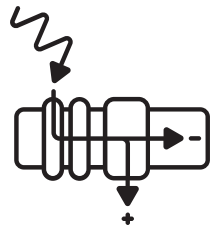
Spectroscopy and microscopy to explore the energetic and structural dynamics of biological systems

EFRC, Associate Director



1-3 PI research projects

Solar energy conversion



Optical spectroscopy

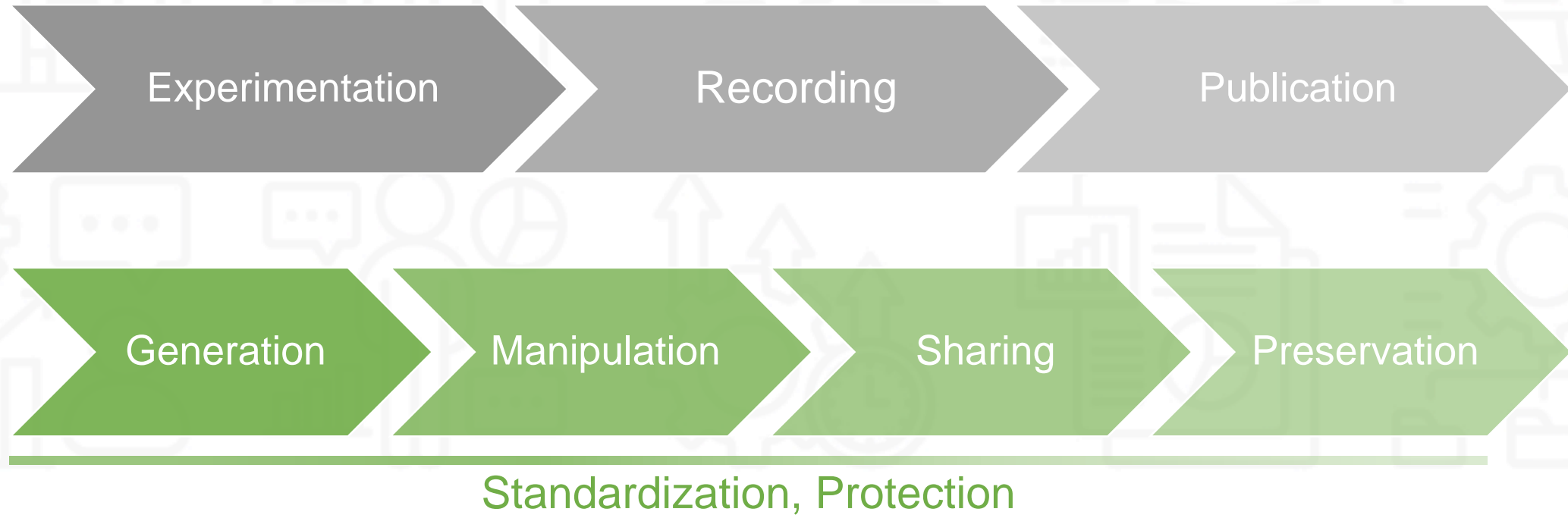


Biomolecular structure determination

Analysis of spectroscopic data

Data archiving

Scientific data life cycle in the research group



How do we link the components of the data life cycle?

How do we leverage state-of-the-art resources to advance scientific discovery?

Cloud-based repositories enable integrated data management

Biomolecular structure determination

Analysis of spectroscopic data

Data archiving



How do we link the components of the data life cycle?

How do we leverage state-of-the-art resources to advance scientific discovery?

Biomolecular structure determination: Protein Data Bank (PDB)

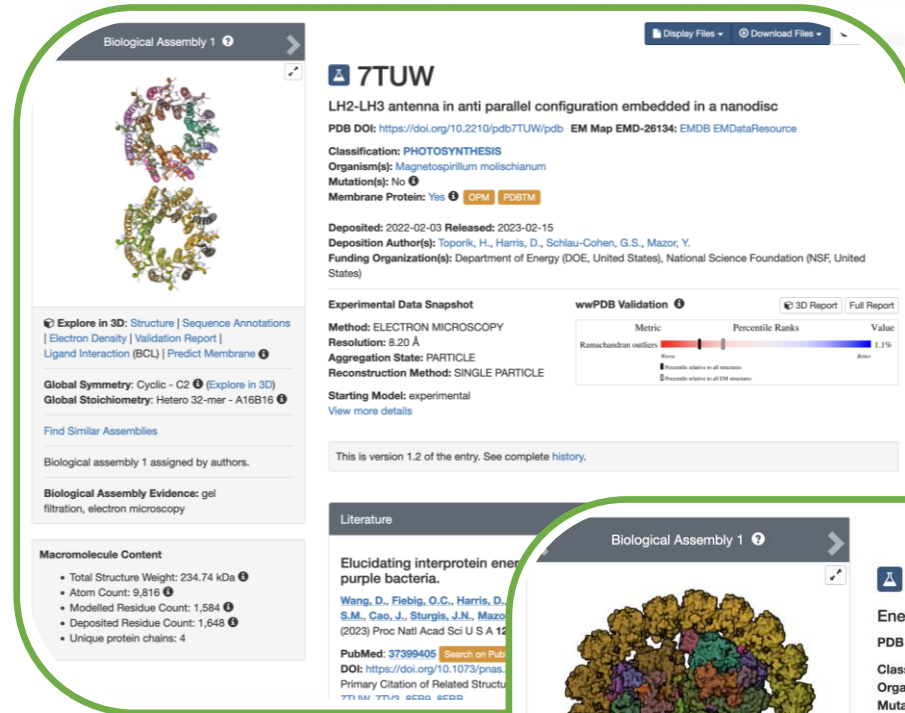


How do we link the components of the data life cycle?

How do we leverage state-of-the-art resources to advance scientific discovery?

Biomolecular structure determination: Protein Data Bank (PDB)

- Determination of energy-relevant biomolecular structures
- Structural model available to the community
- Robust UI for searching and visualization
- High level of quality control



Biological Assembly 1

7TUW
LH2-LH3 antenna in anti parallel configuration embedded in a nanodisc
PDB DOI: <https://doi.org/10.2210/pdb7TUW/pdb> EM Map EMD-26134: EMDB EMDataResource

Classification: PHOTOSYNTHESIS
Organism(s): Magnetospirillum molischianum
Mutation(s): No
Membrane Protein: Yes **OPM** **PDBTM**

Deposited: 2022-02-03 Released: 2023-02-15
Deposition Author(s): Toporik, H., Harris, D., Schlau-Cohen, G.S., Mazor, Y.
Funding Organization(s): Department of Energy (DOE, United States), National Science Foundation (NSF, United States)

Experimental Data Snapshot
Method: ELECTRON MICROSCOPY
Resolution: 8.20 Å
Aggregation State: PARTICLE
Reconstruction Method: SINGLE PARTICLE
Starting Model: experimental
[View more details](#)

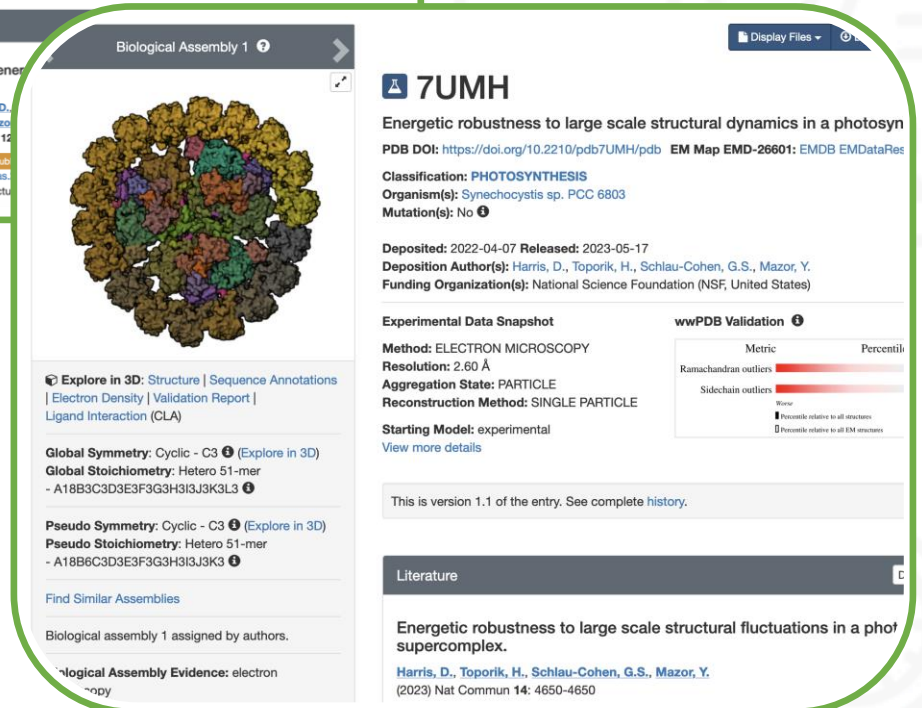
wwPDB Validation
Metric: Ramachandran outliers
Percentile Ranks: 1.1%
Value: 1.1%

Explore in 3D: Structure | Sequence Annotations | Electron Density | Validation Report | Ligand Interaction (BCL) | Predict Membrane

Global Symmetry: Cyclic - C2 (Explore in 3D)
Global Stoichiometry: Hetero 32-mer - A16B16

Macromolecule Content
• Total Structure Weight: 234.74 kDa
• Atom Count: 9,816
• Modeled Residue Count: 1,584
• Deposited Residue Count: 1,648
• Unique protein chains: 4

Literature
Elucidating interprotein energy transfer in purple bacteria.
Wang, D., Fiebig, O.C., Harris, D., S.M., Cao, J., Sturgis, J.N., Mazor, Y. (2023) Proc Natl Acad Sci U S A 120: 22111-22116
PubMed: 37399405 [Search on PubMed](#)
DOI: <https://doi.org/10.1073/pnas.2211111120>
Primary Citation of Related Structures: 7TUW 7TUW SC00 SC00



Biological Assembly 1

7UMH
Energetic robustness to large scale structural dynamics in a photosynthetic supercomplex.
PDB DOI: <https://doi.org/10.2210/pdb7UMH/pdb> EM Map EMD-26601: EMDB EMDataResource

Classification: PHOTOSYNTHESIS
Organism(s): Synechocystis sp. PCC 6803
Mutation(s): No

Deposited: 2022-04-07 Released: 2023-05-17
Deposition Author(s): Harris, D., Toporik, H., Schlau-Cohen, G.S., Mazor, Y.
Funding Organization(s): National Science Foundation (NSF, United States)

Experimental Data Snapshot
Method: ELECTRON MICROSCOPY
Resolution: 2.60 Å
Aggregation State: PARTICLE
Reconstruction Method: SINGLE PARTICLE
Starting Model: experimental
[View more details](#)

wwPDB Validation
Metric: Ramachandran outliers, Sidechain outliers
Percentile Ranks: 1.1%
Value: 1.1%

Explore in 3D: Structure | Sequence Annotations | Electron Density | Validation Report | Ligand Interaction (CLA)

Global Symmetry: Cyclic - C3 (Explore in 3D)
Global Stoichiometry: Hetero 51-mer - A18B3C3D3E3F3G3H3I3J3K3L3

Pseudo Symmetry: Cyclic - C3 (Explore in 3D)
Pseudo Stoichiometry: Hetero 51-mer - A18B6C3D3E3F3G3H3I3J3K3

Literature
Energetic robustness to large scale structural fluctuations in a photosynthetic supercomplex.
Harris, D., Toporik, H., Schlau-Cohen, G.S., Mazor, Y. (2023) Nat Commun 14: 4650-4650

Biomolecular structure determination: Protein Data Bank (PDB)

- Streamlined deposition
- Integrates with other resources designed to support new methodology
- Easy to coordinate with publication process
- Gold standard for repositories (similar to – and linkable with – Cambridge Structural Database, CSD)
- Supported by DOE, NSF, et al. since 1971

Wang, D. et al. *Proc Natl Acad Sci USA* **2023**, *120* (28), e2220477120.

EMD-43592
Single-particle
13.8 Å

3D View Gallery

Deposition: 01/02/2024
Map released: 12/06/2024
Last modified: 12/06/2024

EMD-43592
PDI-containing spoke of a hexagonal wireframe DNA origami

Sample Organism: *Escherichia coli*
Sample: PDI-containing DNA origami spoke

Deposition Authors: Harris D, Parsons MF, Gorman J, Schlau-Cohen GS, Bathe M

Sculpting photoproducts with DNA origami

Gorman J, Hart SM, John T, Castellanos MA, Harris D, Parsons MF, Banal JL, Willard AP, Schlau-Cohen GS, Bathe M (2024) *Chem*, 10, 1553–1575
PUBMED: 38827435

Gorman, J. et al. *Chem* **2024**, *10* (5), 1553–1575.

Harris, D. et al. *Nat Commun* **2023**, *14* (1), 4650.

Cloud-based repositories enable integrated data management

Biomolecular structure determination

- Mature databases are a model for **sharing, preservation, standardization**

Analysis of spectroscopic data

Data archiving



How do we link the components of the data life cycle?

How do we leverage state-of-the-art resources to advance scientific discovery?

Data Analysis: GitHub



How do we link the components of the data life cycle?

How do we leverage state-of-the-art resources to advance scientific discovery?

Data Analysis: GitHub

- Microsoft server that stores code in the cloud with a versioning language
- Big projects (e.g., alphafold) to small projects (e.g., Schlau-Cohen lab fluorescence lifetime analysis)
- Software packages accessible to scientific community
- Quality control individually managed at project level

The image displays two screenshots of GitHub repository pages, illustrating the structure and content of scientific software projects.

The top screenshot shows the repository `qudi-sclab` by `SchlauCohenLab`. The repository is public and has 6 branches and 0 tags. The file tree includes:

- `.github` (Initial commit, 8 months ago)
- `docs` (Initial commit, 8 months ago)
- `src/qudi` (Adding NI digital IO, last month)
- `tests` (Updating so it can work without qudi-lqo-modules, 6 months ago)
- `.gitignore` (Prototyping the SM2P qudi module, 8 months ago)
- `LICENSE` (Initial commit, 8 months ago)
- `LICENSE.LESSER` (Initial commit, 8 months ago)
- `README.md` (Initial commit, 8 months ago)
- `VERSION` (Initial commit, 8 months ago)
- `pyproj` (Initial commit, 8 months ago)
- `setup` (Initial commit, 8 months ago)

The bottom screenshot shows the repository `2D-FLC-code` by `PremashisManna`. The repository is public and has 1 branch and 0 tags. The file tree includes:

- `MatlabCodes` (uploaded codes, 4 years ago)
- `A_Technical_Note_on_2D-FLC.pdf` (uploaded files, 4 years ago)
- `IRF.mat` (uploaded codes, 4 years ago)
- `README` (Update README, 4 years ago)
- `simulated_data.mat` (uploaded codes, 4 years ago)

The `README` file in the bottom screenshot contains the following text:

```

This is 2D fluorescence lifetime correlation (2D-FLC) code originally written by Toru Kondo (toru.kondo.c@tohoku.ac.jp) while at Schlau-Cohen Lab at MIT.

Tutorial for the codes are in the pdf document named "A Technical Note on 2D-FLC.pdf"

All the matlab files can be found in the folder "MatlabCodes".

A synthetic data (simulated_data.mat) is also included with instrumentation response function (IRF.mat).

It is advisable to work with a simulated data first before analyzing the real data.

```

Data Analysis: GitHub

- Allows collaborative software development
- Easily linked within (multiple) publications
- Storage limited to code or small amounts of data
- Hosted by Microsoft server, other options are Gitea, self-hosted; Gitee, hosted by a Chinese company



Microsecond and millisecond dynamics in the photosynthetic protein LHCSR1 observed by single-molecule correlation spectroscopy

Toru Kondo^{a,1,2}, Jesse B. Gordon^a, Alberta Pinnola^{b,c}, Luca Dall'Osto^b, Roberto Bassi^b, and Gabriela S. Schlau-Cohen^{a,1}

^aDepartment of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139; ^bDepartment of Biotechnology, University of Verona, 37134



RESEARCH ARTICLE



Identification of distinct pH- and zeaxanthin-dependent quenching in LHCSR3 from *Chlamydomonas reinhardtii*

Julianne M Troiano^{1†}, Federico Perozeni^{2†}, Raymundo Moya¹, Luca Zuliani², Kwangyul Baek³, EonSeon Jin³, Stefano Cazzaniga², Matteo Ballottari^{2*}, and Gabriela S. Schlau-Cohen^{1*}

Biophysical *Journal*

Article



Membrane-dependent heterogeneity of LHCII characterized using single-molecule spectroscopy

Premashis Manna,¹ Thomas Davies,² Madeline Hoffmann,¹ Matthew P. Johnson,² and Gabriela S. Schlau-Cohen^{1,*}

¹Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts and ²Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield, United Kingdom

Cloud-based repositories enable integrated data management

Biomolecular structure determination

- Mature databases are a model for **sharing, preservation, standardization**

Analysis of spectroscopic data

- Emerging databases enable **generation, manipulation** in collaborative manner

Data archiving



How do we link the components of the data life cycle?

How do we leverage state-of-the-art resources to advance scientific discovery?

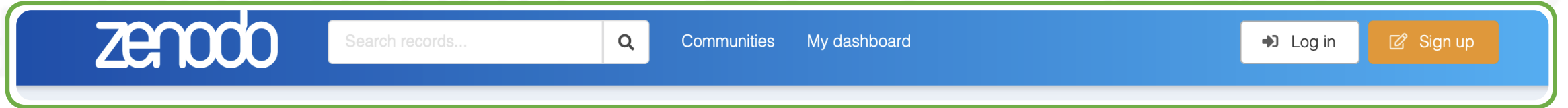
Data Archiving: Zenodo



How do we link the components of the data life cycle?

How do we leverage state-of-the-art resources to advance scientific discovery?

Data archiving: Zenodo



- Catch-all repository
- Projects from all over the world and every discipline
- No standardization, no quality control
- Challenging to search for specific data types

Data archiving: Zenodo

nature
chemistry

ARTICLES

<https://doi.org/10.1038/s41557-021-00841-9>

Check for updates

OPEN

Observation of robust energy transfer in the photosynthetic protein allophycocyanin using single-molecule pump-probe spectroscopy

Raymundo Moya¹, Audrey C. Norris¹, Toru Kondo^{2,3} and Gabriela S. Schlau-Cohen¹  

Data availability

The raw photon stream used to construct the single-molecule pump-probe traces and the corresponding fluorescence lifetime histograms are available at <https://doi.org/10.5281/zenodo.5541825>. [Source data](#) are provided with this paper.









Observation of robust energy transfer in the photosynthetic protein allophycocyanin using single-molecule pump-probe spectroscopy - single-molecule photon stream

Raymundo Moya¹; Audrey C. Norris¹; Toru Kondo²; Gabriela S. Schlau-Cohen¹ 

Show affiliations

Photon stream used in the article "Observation of robust energy transfer in the photosynthetic protein allophycocyanin using single-molecule pump-probe spectroscopy" to analyze single-molecule fluorescence emission. Detected emission for single-molecule pump-probe experiments with an associated instrument response function (IRF) and background fluorescence (BG). Each detected photon is described by its time within the collected photon stream and its time relative to the excitation laser. Data is organized by sample and by date. Also included is an .xlsx document with fitted timescales for all included molecules and Matlab structure titled "FinalDataAndStatistics.mat", which includes the final data, and statistics for the data used within the paper.

Files

SM2P_SMTTraces.zip	
 SM2P_SMTTraces.zip	
 SM2P_SMTTraces	
 APC_610nm_100fs	
 Day1	
 smUS115_converted.mat	53.1 kB
 smUS116_converted.mat	36.3 kB
 smUS117_converted.mat	71.3 kB

- Easy to incorporate in publishing process
- Make data available within the group, to collaborators and to scientific community

Cloud-based repositories enable integrated data management

Biomolecular structure determination

- Mature databases are a model for **sharing, preservation, standardization**

Analysis of spectroscopic data

- Emerging databases enable **generation, manipulation** in collaborative manner

Data archiving

- Existing databases enable **sharing, preservation** at scale without utility of **standardization**



How do we link the components of the data life cycle?

How do we leverage state-of-the-art resources to advance scientific discovery?

Cloud-based repositories enable integrated data management

Biomolecular structure determination

- Mature databases are a model for **sharing, preservation, standardization**

Analysis of spectroscopic data

- Emerging databases enable **generation, manipulation** in collaborative manner

Data archiving

- Existing databases enable **sharing, preservation** *at scale* without utility of **standardization**



How do we link the components of the data life cycle?

Repositories already can bridge many steps of cycle

How do we leverage state-of-the-art resources to advance scientific discovery?

Most useful when well supported by funding agencies

Spectroscopic studies of protein-protein association in model membranes



Graham Schmidt



Dihao Wang

Photosynthetic Systems

Chemical Sciences, Geosciences, & Biosciences Division
DE-SC0018097

Synthesizing Functionality in Excitonic Systems Using DNA Origami



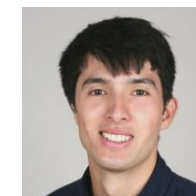
Adam Willard



Madi Scott



Mark Bathe



Jeff Gorman

Biomolecular materials

Materials Sciences & Engineering Division
DE-SC0019998

Energy capture and conversion in self-assembled chlorophyll analogues



Arup Kundu



Jonathan Lindsey
(NCSU)



Phuong Tran
(NCSU)

Solar Photochemistry

Chemical Sciences, Geosciences, & Biosciences Division
DE-SC0025243

Bioinspired light-escalated chemistry (BioLEC)



Greg Scholes
(Princeton)



Vicki Cleave



DE-SC0019370