

¹ gediDB: A toolbox for processing and providing Global Ecosystem Dynamics Investigation (GEDI) L2A-B and L4A-C data

⁴ **Simon Besnard**  ¹, **Felix Dombrowski**  ², and **Amelia Holcomb**  ³

⁵ 1 GFZ Helmholtz Centre Potsdam, Potsdam, Germany ² University of Potsdam, Potsdam, Germany ³
⁶ Department of Computer Science, University of Cambridge, Cambridge, UK ¶ Corresponding author

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Open Journals](#) 

Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

License

Authors of papers retain copyright and release the work under a

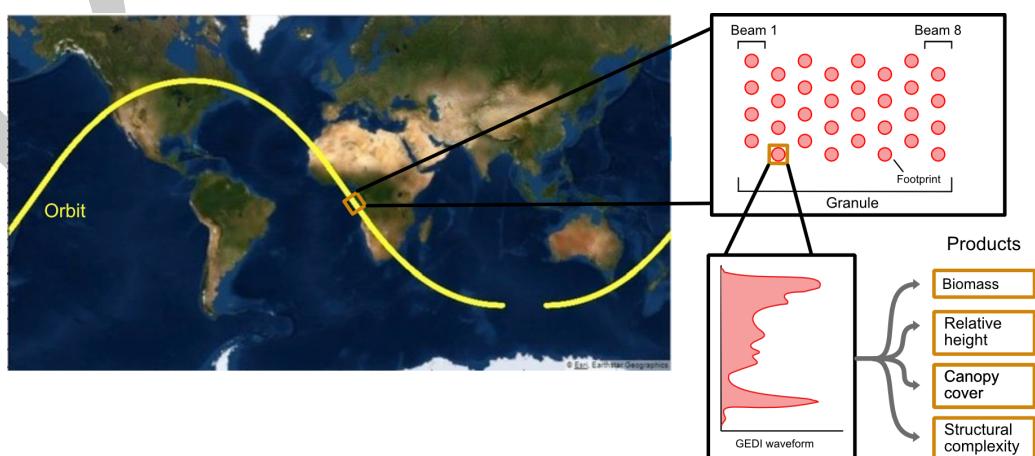
Creative Commons Attribution 4.0 International License ([CC BY 4.0](#))

⁷ Abstract

⁸ The Global Ecosystem Dynamics Investigation (GEDI) mission provides spaceborne LiDAR
⁹ observations that are essential for characterising Earth's forest structure and carbon dynamics.
¹⁰ However, GEDI datasets are distributed as complex HDF5 granules, which pose significant
¹¹ challenges for efficient, large-scale data processing and analysis. To overcome these hurdles,
¹² we developed gediDB, an open-source Python standardised Application Programming Interface
¹³ (API) that streamlines both the processing and querying of GEDI Level 2A–B and Level 4A–C
¹⁴ datasets. Built on the optimised multidimensional array database TileDB, gediDB enables
¹⁵ operational-scale processing, rapid spatial and temporal queries, and reproducible LiDAR-based
¹⁶ analyses of forest biomass, carbon stocks, and structural change.

Statement of Need

²⁰ High-volume LiDAR datasets from the Global Ecosystem Dynamics Investigation (GEDI)
²¹ mission ([R. Dubayah et al., 2020](#)) (Fig. 1) have become essential for quantifying forest
²² dynamics, estimating biomass, and analysing carbon cycling. The open availability of GEDI's
²³ spaceborne LiDAR data has enabled forest structural analysis at near-global scales. However,
²⁴ practical use remains hindered by the complexity of raw HDF5 granules, the absence of scalable
²⁵ infrastructure for efficient access, and a lack of standardised tools for large-scale spatial and
²⁶ temporal subsetting.

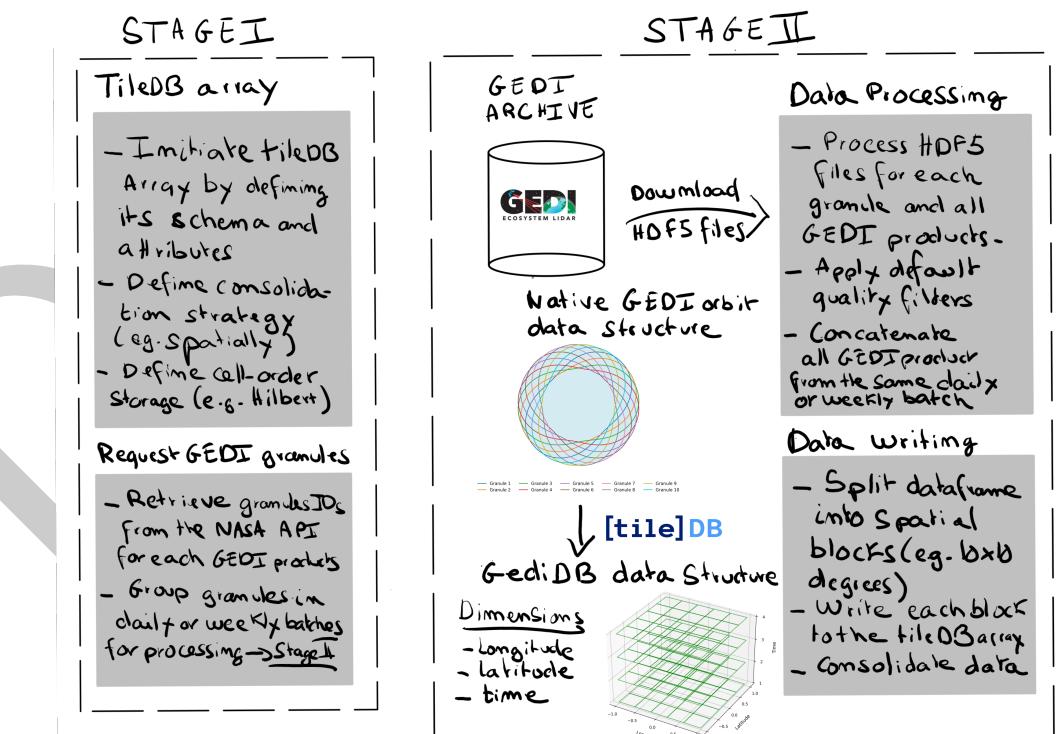


²⁵
²⁶ *Fig. 1: A schematic representation of the GEDI data structure. Credits: Amelia Holcomb's*
²⁷ *PhD dissertation*

Several efforts in the broader NASA LiDAR community have tackled similar challenges. For example, SlideRule (Shean et al., 2023) provides a scalable, cloud-based framework for processing ICESat-2 photon data, enabling users to query and transform complex satellite LiDAR datasets into analysis-ready forms. This illustrates a common pattern: while raw LiDAR missions deliver highly relevant observations, the data formats and scales make direct scientific use difficult without specialised infrastructure. For GEDI, existing services such as NASA's GEDI Subsetter via the Multi-Mission Algorithm and Analysis Platform (MAAP) (Daniels et al., 2025) offer useful access for small to moderate-scale extractions, but they are not designed for operational-scale workflows or integration into reproducible pipelines.

gediDB addresses this gap by providing a robust, scalable Python-based API for unified access to GEDI Level 2A (R. Dubayah, Hofton, et al., 2021), 2B (R. Dubayah, Tang, et al., 2021), 4A (R. O. Dubayah et al., 2022), and 4C (De Conto et al., 2024) products. Built on the TileDB storage engine (TileDB, Inc., 2025), it enables fast querying of multidimensional arrays by spatial extent, temporal range, and variable selection. Seamless integration with geospatial libraries such as xarray (Hoyer & Hamman, 2017) and geopandas (Jordahl et al., 2020) ensures compatibility with reproducible workflows, from local machines to cloud and high-performance computing environments. By leveraging TileDB's advanced spatial indexing, gediDB simplifies and accelerates GEDI data access and analysis (see Fig. 2).

The increasing use of GEDI in global applications, such as canopy height mapping (Pauls et al., 2024), disturbance assessment (Holcomb et al., 2024), and forest degradation monitoring (Bourgoin et al., 2024), highlights the need for efficient, scalable tooling. By streamlining data access and enabling large-scale, reproducible workflows, gediDB fills this role for the GEDI community, complementing efforts like SlideRule in the ICESat-2 domain and supporting broader ecological monitoring and policy-relevant research.



54 Core functionalities

55 Extensive documentation and tutorials are available at <https://gedidb.readthedocs.io>, offering
56 clear setup instructions, configuration guidance, and workflow examples. Users can access a
57 globally processed GEDI dataset directly, avoiding the need for local downloads, as detailed in
58 the [database documentation](#).

59 Data processing framework

60 The gediDB package centres on two primary modules that streamline GEDI data ingestion and
61 access:

- 62 ■ **GEDIProcessor**: Ingests raw GEDI granules and transforms them into structured TileDB
63 arrays (Fig. 3). The process includes filtering, standardisation, and spatio-temporal
64 chunking to ensure high-performance querying.
- 65 ■ **GEDIProvider**: Enables flexible access to GEDI data using spatial and temporal filters
66 and variable selection. Output is compatible with Python libraries such as xarray and
67 pandas ([Reback et al., 2020](#)).

68 Configurable and reproducible workflows

69 Custom configuration files define the TileDB schema and data retrieval parameters, supporting
70 reproducibility and adaptability across diverse computing environments.

71 Robust data downloading

72 The API connects directly to NASA's Common Metadata Repository (CMR) and includes
73 robust retry logic and error handling to ensure consistent, fault-tolerant data acquisition.

74 High-performance data storage

75 GEDI data are stored as sparse TileDB arrays optimised for fast spatial and temporal queries.
76 The array structure accommodates large, multidimensional datasets efficiently (Fig. 3).

77 Parallel processing capabilities

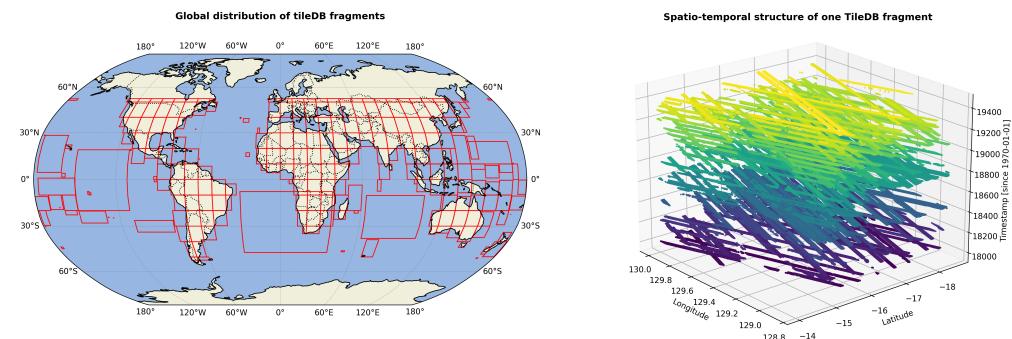
78 gediDB supports parallelised downloading, processing, and storage using libraries such as
79 Dask ([Rocklin, 2015](#)) and concurrent.futures, enabling high-throughput workflows on HPC
80 systems.

81 Advanced querying functionality

82 gediDB offers flexible querying capabilities, including bounding-box, temporal range, and
83 nearest-neighbour queries. Both scalar and profile-type variables are supported.

84 Rich metadata integration

85 Comprehensive metadata—covering provenance, units, variable descriptions, and versioning—is
86 embedded within the TileDB arrays, ensuring transparency and reproducibility.



87

88 Fig. 3: Illustration of the global GEDI data storage schema using TileDB arrays.

89 Performance benchmarks

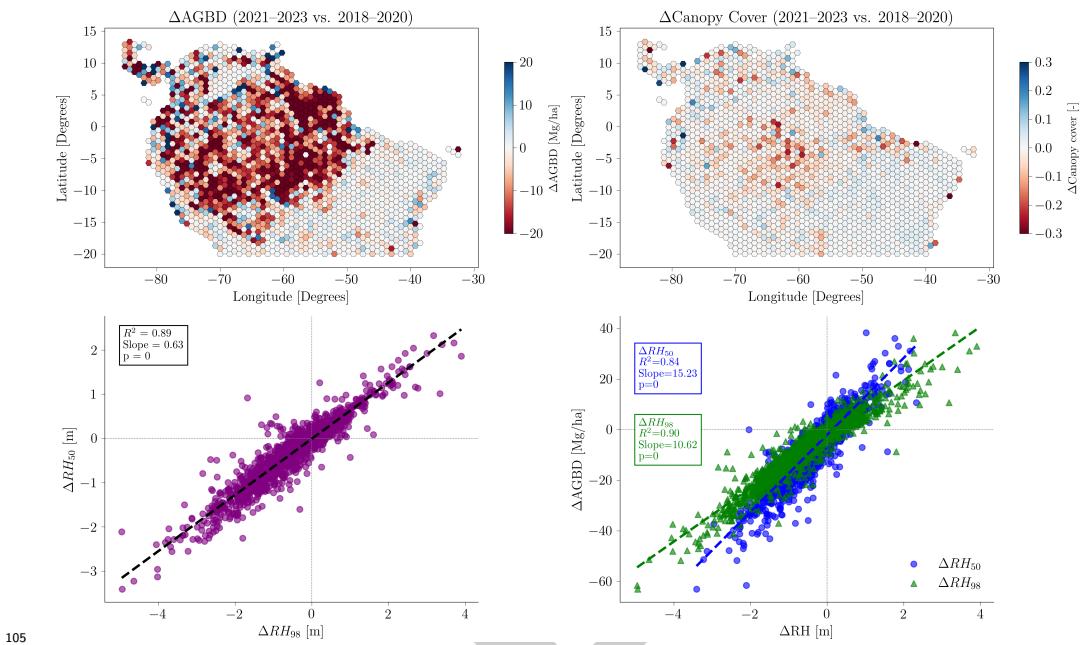
90 To evaluate the efficiency of gediDB, we benchmarked representative research scenarios and
 91 compared them to equivalent queries performed with NASA's GEDI Subsetter via the Multi-
 92 Mission Algorithm and Analysis Platform (MAAP). This comparison highlights not only absolute
 93 performance, but also practical trade-offs between the two approaches. The table below reports
 94 query times (in seconds) for varying spatial and temporal extents.

Scenario	Spatial extent	Time range	Variables queried	Query time (gediDB)	Query time (MAAP)
Local-scale query	1° × 1° bounding box	1 month	relative height metrics, canopy cover	1.8	TBD
Regional-scale query	10° × 10° bounding box	6 months	relative height metrics, biomass, plant area index	17.9	TBD
Continental-scale query	Amazon Basin	1 year	canopy cover, biomass	28.9	TBD

95 Benchmarks for gediDB were performed on a Linux server with dual Intel® Xeon® E5-2643 v4
 96 CPUs (12 cores, 24 threads), 503 GB RAM, and NVMe SSD (240 GB) + HDD (16.4 TB)
 97 storage. All queries ran on NVMe-backed data to ensure high I/O throughput.

98 Example use cases

99 We used gediDB to analyse aboveground biomass and canopy cover dynamics across the
 100 Amazon Basin (Fig. 4). The workflow extracted variables including aboveground biomass,
 101 canopy cover, and relative height (RH) metrics across large spatial extents and multiple years.
 102 Data were aggregated within a 1°×1° hexagonal grid to support spatiotemporal analysis of
 103 forest structural change. The analysis pipeline was implemented entirely in Python using
 104 geopandas and xarray, making it fully reproducible from data extraction to visualisation.



105
106 *Fig. 4: Visualisation of changes in aboveground biomass density (AGBD) (top left panel) and*
107 *canopy cover (top right panel) between 2018–2020 and 2021–2023, aggregated to a $1^\circ \times 1^\circ$*
108 *hexagonal grid over the Amazon Basin. The bottom left panel shows the relationship between*
109 *changes in ΔRH_{50} and ΔRH_{98} , with each point representing a hexagon. The bottom right*
110 *panel shows the relationship between changes in canopy height metrics (ΔRH_{50} and ΔRH_{98})*
111 *and $\Delta AGBD$, with each point representing a hexagon. This highlights how vertical canopy*
112 *structure dynamics relate to biomass change across the region.*

113 A key advantage of gediDB is that large-scale extractions can be performed directly within
114 Python workflows, eliminating the need for manual downloads or interactive tools such as
115 MAAP. For example, the following snippet retrieves biomass, canopy cover, and RH metrics
116 for the Amazon Basin:

```

import geopandas as gpd
import gedidb as gdb

# Instantiate provider with S3 backend
provider = gdb.GEDIProvider(
    storage_type="s3",
    s3_bucket="dog.gedidb.gedi-l2-l4-v002",
    url="https://s3.gfz-potsdam.de"
)

# Load region of interest (Amazon Basin)
roi = gpd.read_file("amazon_basin.geojson")

# Query GEDI data as xarray dataset
ds = provider.get_data(
    variables=["agbd", "cover", "rh_98", "rh_50"],
    query_type="bounding_box",
    geometry=roi,
    start_time="2018-01-01",
    end_time="2024-01-01",
    return_type="xarray"
)

```

117 Future development

118 Planned future developments for gediDB are designed to improve usability and extend the
119 package's scope for both researchers and operational users:

- 120 ■ **Compatibility with upcoming GEDI product releases:** ensures long-term sustainability
121 of the toolbox as new mission data become available, avoiding version lock-in for users
122 building workflows on gediDB.
- 123 ■ **Improved performance and flexibility in querying profile variables:** will make it easier for
124 users to analyse canopy structure profiles (e.g., RH metrics) at scale, which are currently
125 among the most data-intensive GEDI products.
- 126 ■ **Support for direct HDF5 access from AWS S3:** will enable gediDB to operate directly on
127 cloud-hosted GEDI granules (e.g., on NASA's MAAP infrastructure), avoiding the need
128 for local downloads and reducing storage overhead. This work corresponds to [Issue #15](#)
129 ("Allow direct h5 read w/o download"), which aims to let the GEDIProcessor access
130 HDF5 files in an S3 bucket without intermediate steps.
- 131 ■ **Expanded documentation and tutorials:** will benefit new users by lowering the entry
132 barrier, providing clear end-to-end examples, and connecting scientific use cases to code
133 snippets.
- 134 ■ **Strengthened testing for reliability and maintainability:** supports developers and long-
135 term users by ensuring that changes do not break existing workflows, and by increasing
136 trust in the reproducibility of analyses built on gediDB.

137 Development progress and discussion of these features are tracked openly through the project's
138 [GitHub issues](#) and roadmap.

139 Conclusion

140 gediDB enhances the usability of GEDI LiDAR datasets by addressing challenges of data com-
141 plexity, scalability, and reproducibility. Built on TileDB, it enables efficient data management,
142 fast querying, and integration into geospatial workflows—facilitating large-scale analysis of
143 forest dynamics and carbon cycling. Its open-source, community-driven design supports ongoing
144 progress in remote sensing and environmental science.

145 Acknowledgements

146 The development of gediDB was supported by the European Union through the [FORWARDS](#)
147 and [NextGenCarbon](#) projects. We also acknowledge funding for 3D-ABC by the Helmholtz
148 Foundation Model Initiative, supported by the Helmholtz Association. We would also like to
149 thank the R2D2 Workshop (March 2024, Potsdam) for providing the opportunity to meet and
150 discuss GEDI data processing. We recognise using OpenAI's ChatGPT and Grammarly AI
151 tools to enhance the manuscript's sentence structure, conciseness, and grammatical accuracy.

152 References

- 153 Bourgoin, C., Ceccherini, G., Girardello, M., Vancutsem, C., Avitabile, V., Beck, P., Beuchle,
154 R., Blanc, L., Duveiller, G., Migliavacca, M., & others. (2024). Human degradation of
155 tropical moist forests is greater than previously estimated. *Nature*, 631(8021), 570–576.
156 <https://doi.org/10.1038/s41586-024-07629-0>

- 157 Daniels, C., French, J., Adhikari, S., Bhusal, A., Mandel, A. I., & Kirkland, S. (2025). *MAAP-project/gedi-subsetter: 0.10.0* (Version 0.10.0). Zenodo. <https://doi.org/10.5281/zenodo.1512227>
- 160 De Conto, T., Armston, J., & Dubayah, R. O. (2024). *GEDI L4C Footprint Level Waveform Structural Complexity Index, Version 2*. ORNL DAAC, Oak Ridge, Tennessee, USA. <https://doi.org/10.3334/ORNLDAAAC/2338>
- 163 Dubayah, R. O., Armston, J., Kellner, J. R., Duncanson, L., Healey, S. P., Patterson, P. L.,
164 Hancock, S., Tang, H., Bruening, J. M., Hofton, M. A., Blair, J. B., & Luthcke, S. B.
165 (2022). *GEDI L4A Footprint Level Aboveground Biomass Density, Version 2.1*. ORNL
166 DAAC, Oak Ridge, Tennessee, USA. <https://doi.org/10.3334/ORNLDAAAC/2056>
- 167 Dubayah, R., Blair, J. B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt,
168 G., Kellner, J., Luthcke, S., Armston, J., Tang, H., Duncanson, L., Hancock, S., Jantz, P.,
169 Marselis, S., Patterson, P. L., Qi, W., & Silva, C. (2020). The global ecosystem dynamics
170 investigation: High-resolution laser ranging of the earth's forests and topography. *Science of Remote Sensing*, 1, 100002. <https://doi.org/10.1016/j.srs.2020.100002>
- 172 Dubayah, R., Hofton, M., Blair, J., Armston, J., Tang, H., & Luthcke, S. (2021). *GEDI L2A Elevation and Height Metrics Data Global Footprint Level V002*. NASA EOSDIS Land
173 Processes Distributed Active Archive Center. https://doi.org/10.5067/GEDI/GEDI02_A.002
- 176 Dubayah, R., Tang, H., Armston, J., Luthcke, S., Hofton, M., & Blair, J. (2021). *GEDI L2B Canopy Cover and Vertical Profile Metrics Data Global Footprint Level V002*. NASA
177 EOSDIS Land Processes Distributed Active Archive Center. https://doi.org/10.5067/GEDI/GEDI02_B.002
- 180 Holcomb, A., Burns, P., Keshav, S., & Coomes, D. A. (2024). Repeat GEDI footprints measure
181 the effects of tropical forest disturbances. *Remote Sensing of Environment*, 308, 114174.
182 <https://doi.org/10.1016/j.rse.2024.114174>
- 183 Hoyer, S., & Hamman, J. J. (2017). Xarray: N-d labeled arrays and datasets in python.
184 *Journal of Open Research Software*, 5(1), 10. <https://doi.org/10.5334/jors.148>
- 185 Jordahl, K., Bossche, J. V. den, Fleischmann, M., Wasserman, J., McBride, J., Gerard,
186 J., Tratner, J., Perry, M., Badaracco, A. G., Farmer, C., Hjelle, G. A., Snow, A. D.,
187 Cochran, M., Gillies, S., Culbertson, L., Bartos, M., Eubank, N., maxalbert, Bilogur,
188 A., ... Leblanc, F. (2020). *Geopandas/geopandas: v0.8.1* (Version v0.8.1). Zenodo.
189 <https://doi.org/10.5281/zenodo.3946761>
- 190 Pauls, J., Zimmer, M., Kelly, U. M., Schwartz, M., Saatchi, S., Ciais, P., Pokutta, S., Brandt,
191 M., & Gieseke, F. (2024). *Estimating canopy height at scale*. <https://doi.org/10.48550/arXiv.2406.01076>
- 193 Reback, J., McKinney, W., jbrockmendel, Van den Bossche, J., Augspurger, T., Cloud, P.,
194 Hawkins, S., Gfyoung, Sinhrks, Klein, A., Roeschke, M., & Tratner, W. (2020). Pandas-
195 dev/pandas: pandas. Zenodo. <https://doi.org/10.5281/zenodo.3509134>
- 196 Rocklin, M. (2015). Dask: Parallel computation with blocked algorithms and task scheduling.
197 *Proceedings of the 14th Python in Science Conference*, 130–136. <https://doi.org/10.25080/majora-7b98e3ed-013>
- 199 Shean, D., Swinski, J. p., Smith, B., Sutterley, T., Henderson, S., Ugarte, C., Lidwa, E., &
200 Neumann, T. (2023). SlideRule: Enabling rapid, scalable, open science for the NASA
201 ICESat-2 mission and beyond. *Journal of Open Source Software*, 8(81), 4982. <https://doi.org/10.21105/joss.04982>
- 203 TileDB, Inc. (2025). *Tiledb: Modern database engine for complex data based on multi-dimensional arrays*. <https://github.com/TileDB-Inc/TileDB-Py>