Journal of Infrared, Millimeter, and Terahertz Waves (2023) 44:795–813
https://doi.org/10.1007/s10762-023-00947-w

The dotTHz Project: A Standard Data Format for Terahertz Time-Domain Data

Jongmin Lee[1] · Chi Ki Leung[1] · Mingrui Ma[1] · Jasper Ward-Berry[1] · Supawan Santitewagun[1] · J. Axel Zeitler[1]

Received: 15 September 2023 / Accepted: 31 October 2023 / Published online: 13 December 2023
© The Author(s) 2023

Abstract
From investigating molecular vibrations to observing galaxies, terahertz technology has found extensive applications in research and development over the past three decades. Terahertz time-domain spectroscopy and imaging have experienced significant growth and now dominate spectral observations ranging from 0.1 to 10 THz. However, the lack of standardised protocols for data processing, dissemination, and archiving poses challenges in collaborating and sharing terahertz data between research groups. To tackle these challenges, we present the dotTHz project, which introduces a standardised terahertz data format and the associated open-source tools for processing of dotTHz files. The dotTHz project aims to facilitate seamless data processing and analysis by providing a common framework. All software components are released under the MIT licence through GitHub repositories to encourage widespread adoption, modification, and collaboration. We invite the terahertz community to actively contribute to the dotTHz project, fostering the development of additional tools that encompass a greater breadth and depth of functionality. By working together, we can establish a comprehensive suite of resources that benefit the entire terahertz community.

Keywords Terahertz · Time-domain spectroscopy · Open-source · Standardisation · CaTx · CaTSper

1 Introduction

Despite the unique characteristics and potential applications of terahertz radiation, its practical exploitation only began in the late 1980 s with the groundbreaking develop-

✉ J. Axel Zeitler
jaz22@cam.ac.uk

[1] Department of Chemical Engineering and Biotechnology, University of Cambridge, Philippa Fawcett Drive, Cambridge CB3 0AS, UK

ment of subpicosecond photoconductive antennas by Smith et al. [1]. These antennas played a pivotal role in overcoming the challenges associated with generating and accurately detecting terahertz radiation, which were the primary obstacles to its practical use. Building upon these advancements, Hu and Nuss [2] further emphasised the exceptional opportunities provided by terahertz time-domain imaging, extending its scope beyond spectroscopy. This capability has accelerated the expansion of terahertz technology into non-destructive testing applications, including art conservation, industrial product quality testing, and concealed explosive detection. Today, terahertz time-domain spectroscopy (THz-TDS) is widely applied in various fields spanning fundamental science to industrial engineering applications [3–5]. Following its introduction as a highly specialised tool by a small group of research laboratories, terahertz time-domain spectroscopy has evolved into a large field of study with a user base that ranges from expert scientists with decades of experience in time-domain technology to general laboratory technicians running samples on commercial turnkey THz-TDS instruments.

A distinct advantage of THz-TDS is its ability to simultaneously measure the amplitude and the phase information of the electric field. This distinguishes it from most infrared spectroscopy techniques since it allows the direct extraction of the complex refractive index and the complex dielectric constant without relying on the Kramers-Kronig relation. The working principle of THz-TDS involves acquiring a time-domain waveform followed by data processing to transform the time-domain data into a frequency-domain spectrum. This spectral information is heavily affected by the parameter settings in data acquisition and processing. Therefore, an in-depth understanding of the signal processing routine and the parameters used is essential to achieve repeatable and reproducible spectral analysis. While commercially available THz-TDS systems often provide a bundled software package for analysing the measured data, it is not always transparent what steps are carried out precisely, what assumptions are made, and what parameters are used. This lack of transparency can result in unintended variations in data analysis methodology and the resultant spectral data for measurements on instruments from different vendors for the same sample but measured or processed using different software [6].

As a result, many research groups in the terahertz time-domain field develop their own analysis tools. However, the use of a multitude of incompatible data structures complicates the exchange and application of these tools. In the case of custom-built spectrometers, simple ASCII text files are commonly used to store data for individual measurements. This approach requires manual differentiation between sample and reference data for each measurement. Furthermore, essential metadata, such as sample thickness, temperature, or concentration, is typically logged manually in laboratory notebooks and is not captured in the digital file, making it challenging to re-analyse old data or share it with colleagues from different groups. It is also a hurdle to accommodate increasingly common funder's mandates to make available all data that is associated with a publication. The open access requirements often stipulate for such data to be provided in a machine-readable, accessible, described, and re-usable format that ideally contains unmodified and complete data [7]. For commercial systems, some instruments utilise binary file structures with varying degrees of complexity. Still, the often proprietary nature of these file formats, combined with the undocu-

mented file architecture that can change between software package releases, makes exchanging information difficult and renders it impossible to re-analyse archived data once the software package has been updated. A standardised data format is needed to facilitate collaboration, reproducibility, and the long-term accessibility of terahertz spectroscopy data.

Our research group has utilised a set of in-house developed MATLAB script tools that have gradually evolved over decades. While these tools have provided us with excellent flexibility in data analysis, they have also resulted in redundant code and posed challenges in properly documenting the code and maintaining a comprehensive understanding of the algorithms. Additionally, with the growth of the group and the availability of more instruments, we have faced the increasing burden of managing large volumes of data.

To address these issues, we recently decided to enhance the usability of our tools with a graphical user interface (GUI) for more intuitive, interactive, and efficient analysis. However, when we shared these newly developed tools with collaborators, compatibility issues arose due to diverse data formats used by different commercial and home-built spectrometers. Similarly, collaborations among individuals and groups in the terahertz community are often limited to users of a specific TDS system or require laborious and manual data conversion to utilising existing signal processing routines. Such barriers hinder progress within the scientific community.

To overcome these limitations, we propose a solution by introducing a standardised dotTHz format for terahertz time-domain data, the Cambridge THz Converter (CaTx, see Section B) to facilitate the adoption of this data format, and the Cambridge THz Spectrum Analyser (CaTSper, see Section C) as a simple GUI-based processing platform for THz-TDS data analysis. Both software tools have been released as open source under the MIT licence [8, 9]. We are also actively developing additional tools that will be shared in due course. Moreover, comprehensive information including processing methods, step-by-step user guides, and inline code annotations can be accessed through the online documentation [10].

2 The dotTHz Data Format

2.1 Format Structure

Terahertz time-domain waveforms comprise a series of numeric values representing the amplitude of the electric field as a function of time. To extract the optical constants from such data for a specific sample, it is necessary to record both the time-domain waveform of the sample and a reference waveform, along with essential information about the measurement settings and the sample. This implies the need to manage and store at least a pair of data files for each measurement. For the sake of simple and efficient data management, the dotTHz project adopts the hierarchical data format version 5 (HDF5) [11]. The HDF5 format was initially developed by a collaboration between the US National Center for Supercomputing Applications (NCSA) and the US Department of Energy's Advanced Simulation and Computing Program (ASC) to

deal with extensive and complex data. By embracing the same principle, the dotTHz data format delivers the following key advantages to users:

1. Simple data structure for easy handling.
2. Logical data organisation for efficient data retrieval and referencing.
3. Direct attachment of essential metadata for convenient automated processing and analysis.
4. Ability to process specific subsets of data from large files.
5. Ability to store different types (e.g. time-domain waveforms, spatial coordinates, metadata) of data in a single dataset.
6. High-speed performance with contiguous and uncompressed datasets.
7. Wide platform support as an open-source format.
8. Easy data sharing with all information stored in a single file.

The dotTHz file follows a specific structure: for each measurement, a group of datasets corresponding to sample and reference measurements is stored together with the attributes that contain the metadata, as illustrated in Fig. 1. The attributes can have various forms, such as numeric value, numeric vector, and string (Table 1), enabling efficient extraction and referencing of information during subsequent analysis and data processing.

It is essential to emphasise that a single dotTHz file has the capability to accommodate multiple measurements. This enables the consolidation of data pertaining to a time series of measurements or variable temperature measurements of the same sample within a single file. Furthermore, this approach facilitates and simplifies the archiving and sharing of experimental data.

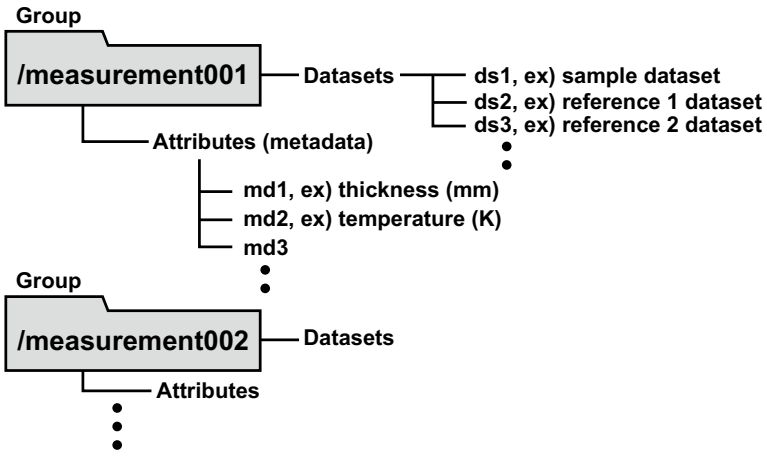


Fig. 1 The hierarchical structure of the dotTHz data format: multiple measurements with associated metadata can be stored in a single dotTHz file

Table 1 The dotTHz file datasets and attributes and the minimum requirement for a dataset as defined by this standard

No	Item	MATLAB format	HDF5 dataset/attribute	Requirement
1	Number	Numeric value	Optional prefix	Auto-numbering
2	Name	String scalar	Dataset name	Necessary
3	Description	String scalar	Attribute (description)	
4	Instrument profile	String scalar	Attribute (instrument)	
5	User profile	String scalar	Attribute (user)	
6	Date and Time	String scalar	Attribute (time)	
7	Mode	String scalar	Attribute (mode)	
8	Coordinate	Numeric vector	Attribute (coordinate)	
9	Metadata description	String scalar	Attribute (mdDescription)	
10	Metadata 1	Numeric value/vector, string scalar	Attribute (md1)	
11	Metadata 2	Numeric value/vector, string scalar	Attribute (md2)	
12	Metadata 3	Numeric value/vector, string scalar	Attribute (md3)	
13	Metadata 4	Numeric value/vector, string scalar	Attribute (md4)	
14	Metadata 5	Numeric value/vector, string scalar	Attribute (md5)	
15	Metadata 6	Numeric value/vector, string scalar	Attribute (md6)	
16	Metadata 7	Numeric value/vector, string scalar	Attribute (md7)	
17	dotTHz version	String scalar	Attribute (thzVer)	
18	Dataset description	String scalar	Attribute (dsDescription)	
19	Dataset 1	Numeric matrix	Dataset (ds1)	Necessary
20	Dataset 2	Numeric matrix	Dataset (ds2)	
21	Dataset 3	Numeric matrix	Dataset (ds3)	
22	Dataset 4	Numeric matrix	Dataset (ds4)	

2.2 Example Use Cases

In the following, we would like to outline a selection of representative use case scenarios of how we envisage the dotTHz file format being used in the terahertz community going forward.

2.2.1 THz-TDS Measurement of Pellet in Transmission

For a typical THz-TDS experiment of a single sample, the file will contain the time-domain waveform of the sample and one reference. The minimum metadata required will comprise of the sample thickness. It is expected that the metadata also contains a suitable identifier such as 'TX' to denote the measurement carried out in transmission.

Optionally, a single dotTHz file can contain the measurements of multiple samples and references or multiple measurements of the same sample and reference under varying conditions, such as a function of time or temperature for dynamic observations, and the conditions can be conveniently stored as additional metadata to facilitate subsequent analysis.

2.2.2 THz-TDS Measurement of Thin Film or Layered Structure

For measurements of thin films or multilayered materials, the metadata will contain information about each layer's thickness, either as individual values within multiple slots or as a single numeric vector within one slot, to facilitate compatibility with subsequent analysis tools.

2.2.3 THz Pump-Probe Measurement

Since terahertz pump-probe measurements require two references, three datasets can be used for each measurement as dataset 1, dataset 2, and dataset 3 for sample, reference, and pumped reference, respectively.

2.2.4 THz Time-Domain Imaging

Terahertz time-domain imaging (THz-TDI) datasets consist of terahertz measurement data, specifically sample, reference, and baseline measurements, along with associated coordinates and timestamps for location-dependent and time-dependent scanning, respectively. The coordinates and timestamps can be stored in the 'Date and Time' and 'Coordinates' attributes in Table 1. Typically, THz-TDI datasets have a considerably large size due to their raster scanning nature, and this size can be effectively reduced by eliminating redundant data. CaTx offers an option that only stores the differentiating coordinates or time attributes from each scan.

2.2.5 Potential Use Cases With Non-Time-Domain Data

The dotTHz Dataset space can be used for any matrix form of datasets, providing compatibility to non-time-domain data. However, it will be necessary to set up a minimum

outline for the dataset allocation for each application domain to keep its consistency and compatibility with subsequent analysis tools. The following is an example of two frequency-domain cases, and these can be updated for better applications along with analysis tool development.

Vector Network Analyser (VNA) Applications Four sets of S-parameter datasets can be stored in datasets 1 to 4, and each dataset will contain three rows for frequency, amplitude, and phase vectors.

Frequency-Modulated Continuous Wave (FMCW) Applications Similar to VNA measurement datasets, frequency, in-phase, and quadrature signals can be grouped as a dataset. While datasets can currently store up to four sets, this limitation is due to the current converter tool's display space and can be easily extended with minor modification of the tool.

3 Conclusion

The dotTHz project was initiated to reduce terahertz data analysis efforts and at the same time foster collaborations in the terahertz community. We have taken the initiative in designing and introducing CaTx and CaTSper, which aim to standardise the processing and analysis of terahertz data obtained from different terahertz instruments. These tools were successfully deployed as part of the data analysis routine in the group. We hope that the dotTHz format may facilitate the development of many other advanced data analysis tools within our community, building on the excellent work by many colleagues [12–14] as well as facilitating establishing databases, reference datasets and supporting standardised testing approaches of novel devices and technologies [15] in the future.

The dotTHz project is an ongoing endeavour, and additional open-source standardised terahertz analysis tools for different applications and data manipulation methods will be developed in the future. We invite researchers from the terahertz community to join and contribute to this development. We also strongly encourage scientists, engineers, and developers to download the tools from the online repository, thoroughly test them, make necessary modifications, and contribute back to enrich the dotTHz project. Through the dotTHz project, we aim to bring the terahertz community closer together, foster collaborations, and facilitate further advancements in the terahertz field. We firmly believe that by standardising and simplifying data analysis and processing, we can attract and encourage more individuals to explore the vast potential of terahertz technology and its numerous applications.

Appendix A: Data Processing Tools

In the absence of native support for the dotTHz files on THz-TDS instrument acquisition software, the first step of the workflow is to convert the existing data and metadata into the dotTHz data format using CaTx [8] (Fig. 2). The initial release of CaTx supports reading and converting the raw output data from commercial spectrometers provided by TeraView, Menlo Systems, and Toptica. This conversion process enables

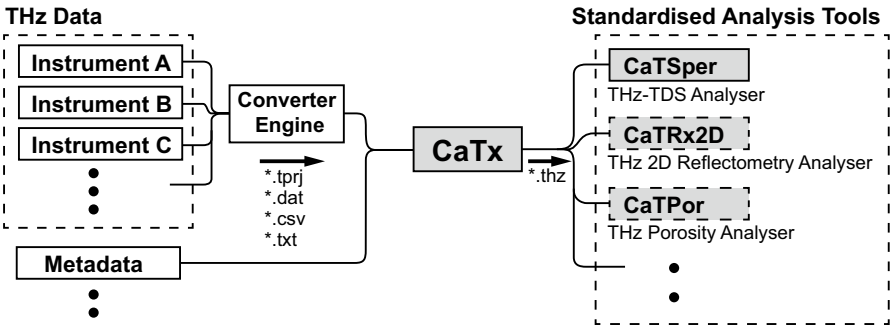


Fig. 2 A schematic diagram of the dotTHz project workflow. Terahertz data obtained from different spectrometers can be converted into a standardised dotTHz file, then analysed using standardised analysis tools

researchers to analyse, compare, and review THz-TDS datasets acquired and stored with different instruments and platforms. The tools were developed using MATLAB (MathWorks, Massachusetts, USA), and they require MATLAB to be installed before use. It is worth mentioning that all source codes are freely accessible without any licensing restrictions. We hope to encourage the development and sharing of Python versions of these tools as well. As the dotTHz project gains active support from the terahertz community, a comprehensive list of readily available tools will be accessible through the online community space.

Appendix B: Cambridge THz Converter (CaTx)

The Cambridge THz Converter (CaTx) is the core component of the dotTHz project. It streamlines and consolidates data from experiments that are stored across multiple files into a unified, standardised format. Its primary function is to combine time-domain sample waveforms, reference waveforms, and accompanying metadata, including sample thickness, temperature, or specific details about the sample's form. By structuring

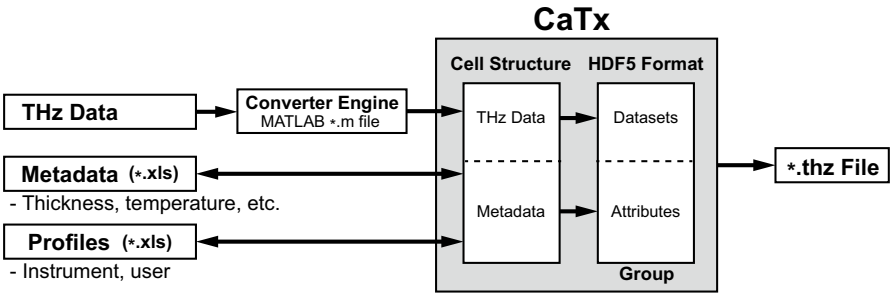
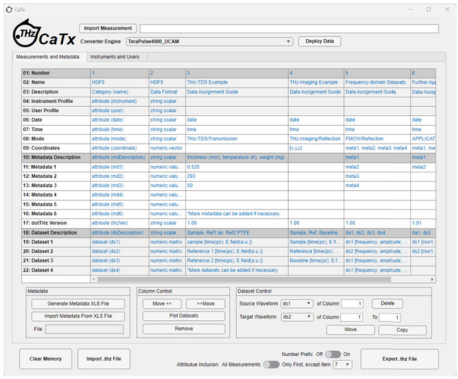
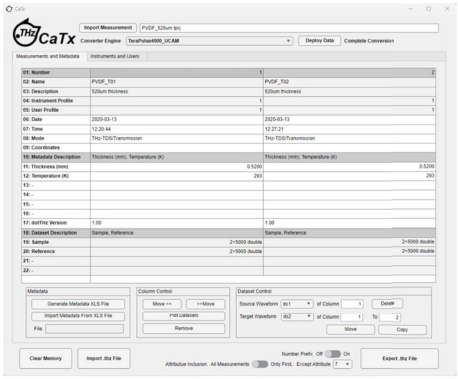


Fig. 3 An example of CaTx consolidating a dataset and its corresponding metadata and profile into a single dotTHz file. Multiple datasets and their metadata and profiles from different sources can be imported and stored as a single dotTHz file

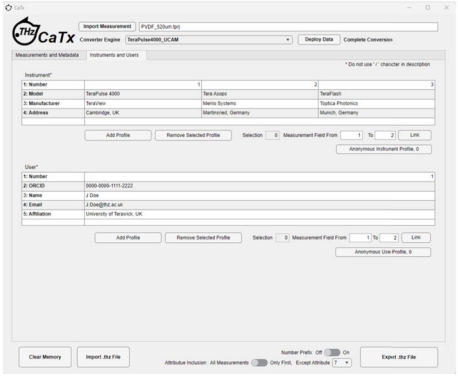
Fig. 4 Examples of CaTx's graphic user interfaces



(a) CaTx's launch interface (template mode) displays entry templates for each row with HDF5 format explanation for various applications.



(b) The measurements and metadata tab of CaTx displays the terahertz data and metadata that will be stored in the dotTHz file. Metadata can be imported from a Microsoft Excel spreadsheet or directly input into the displayed table.



(c) The instruments and users tab of CaTx allows details of the instrument and user to be added, and selected for specified datasets.

and packaging this information into a single file, researchers can easily access and analyse the data in a more organised and convenient manner (Fig. 3).

CaTx calls a converter engine function to accommodate the diverse time-domain waveforms of independent THz-TDS systems. To work with a specific system, users can select an appropriate engine from the converter engine list on the CaTx GUI. These engines are MATLAB script-based functions located in the `/Engines` sub-folder. While several converter engines are readily available in the online repository for major THz-TDS systems, researchers may need to modify existing engines or create new ones for their systems. In such case, they can refer to the online documentation or the inline annotations provided with the existing engines [10].

Operating CaTx is a straightforward process, requiring just a few clicks on the GUI. The converter will extract and organise the available data from the measurements into a table, as depicted in Fig. 4b. Users can import metadata from a Microsoft Excel spreadsheet or edit it directly via the GUI table.

The profile manager, accessible under the second tab titled 'Instruments and Users' in Fig. 4c, provides a convenient way to store and manage user and terahertz instrument information as attributes attached to their respective datasets. User profiles allow for the inclusion of ORCID, name, email, and affiliation details. Terahertz instrument profiles encompass information such as the instrument model, manufacturer, and location. While both profiles are optional, they can be easily updated using the corresponding Microsoft Excel spreadsheets. Finally, all datasets and their corresponding metadata can be exported in a single dotTHz file.

CaTx serves to provide a versatile platform for organising and converting data obtained from various projects and sources into a single, systematic, and simplified dotTHz file. This capability makes the converter highly valuable for processing terahertz data acquired from collaborators or from the literature, especially when the data is not stored in a familiar and compatible format. By enabling the consolidation of data into a standardised format, the converter streamlines the data integration process and enhances the efficiency of subsequent analysis.

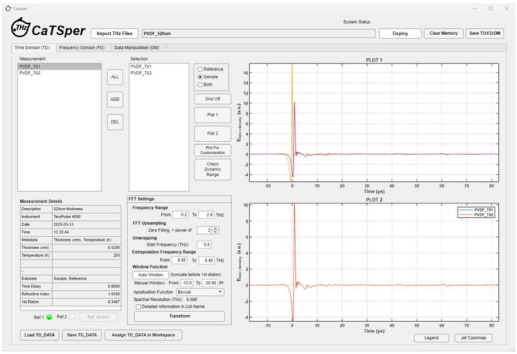
Appendix C: Cambridge THz Spectrum Analyser (CaTSper)

C.1 Overview

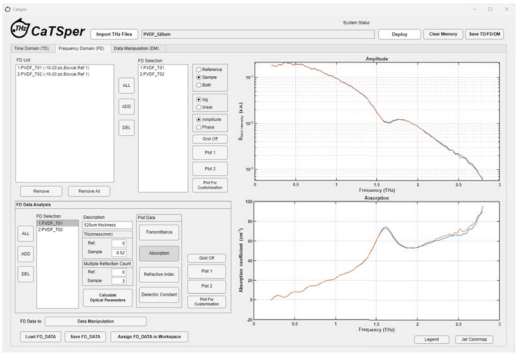
The Cambridge THz Spectrum Analyser (CaTSper) [9] has been designed as an analysis tool compatible with the dotTHz files. As with additional standardised analysis tools for other applications that are currently under development, CaTSper operates on the dotTHz files generated by CaTx, enabling terahertz time-domain data processing and analysis through a user-friendly interface.

CaTSper is organised into three sections, each dedicated to manipulating specific data formats: time-domain data, frequency-domain data, and extracted optical parameters. These sections are accessible through separate tabs on the GUI (Fig. 5), ensuring user consistency, convenience, and expandability. The following provides a brief overview of each tab's functionalities, while more detailed explanations can be found in the online documentation [10].

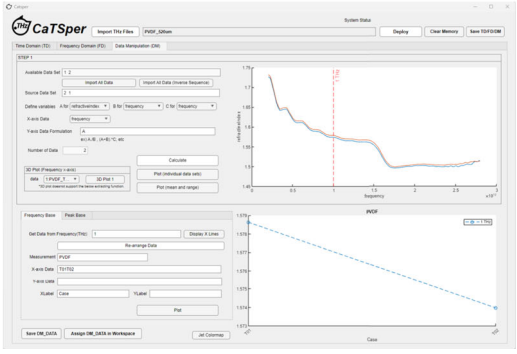
Fig. 5 Examples of CaTSper's graphic user interfaces



(a) The time domain tab of CaTSper displays the sample details and the time domain data. Sample details can be added and parameters such as frequency range, window function can be specified for fast Fourier transform.



(b) The frequency domain tab of CaTSper displays the Fourier transformed frequency-domain data, calculates and plots the optical parameters of transmittance, absorption, refractive index and dielectric constant.



(c) The data manipulation tab of CaTSper enables secondary data processing by using the selected variables and user-specified equations. Data from specific frequencies can also be extracted and plotted against a selected parameter.

- Tab 1: time domain
  - Plot time-domain waveforms
  - Fourier transform of the time-domain waveforms with user-defined truncation window and various window functions
- Tab 2: frequency domain
  - Plot frequency-domain spectra
  - Calculate optical parameters based on the spectrum and the associated metadata
  - Plot absorption coefficients, refractive indices, and dielectric constants
- Tab 3: data manipulation
  - Extract the data of interest
  - Calculate and display the extracted data in various forms
  - Save the extracted data in MATLAB.mat format

The user interface of CaTSper is designed to be self-explanatory, enabling users to easily modify parameters and processing options at each stage of the analysis. Furthermore, CaTSper allows for the convenient processing of the same dataset with different parameter settings, providing the ability to observe and evaluate the impact of parameter settings on the results. This workflow facilitates the real-time review of processed data, assisting users to make immediate adjustments to parameters and functions at each step before progressing to the next. If no further modifications are necessary, users can seamlessly progress through CaTSper workflow for data analysis. Processed data can be stored as a MATLAB m-file or assigned to the MATLAB workspace at any stage. In addition, as an open-source tool, CaTSper's terahertz data processing is completely transparent, allowing users to have full visibility into the analysis procedures and algorithms used.

In the initial release of CaTSper, the software supports a basic processing routine to extract the optical constants from simple pellet slabs. Further work has commenced to gradually extend the functionality to allow for more complex geometries, such as layered structures and thin films, to provide a more comprehensive tool. In particular, work is ongoing to implement the Nelly tool [16] into CaTSper.

C.2 Processing Steps

C.2.1 Time-Domain Analysis

The time delay Δt is the extra time needed for the THz pulse to traverse through the sample thickness H, compared to the THz pulse traversing the same thickness in the reference measurement (air, refractive index $n_{\mathrm{ref}} = 1$). The time-domain effective refractive index $n_{\mathrm{eff,TD}}$ of the sample is thus calculated by

$$n_{\mathrm{eff,TD}} = \frac{c\Delta t}{H} + 1$$

where c is the speed of light with a value of 3×10^8 ms $^{-1}$. n_{eff} is calculated to four significant figures.

The time delay due to one internal reflection occurring $\Delta t_{1\text{etl}}$ is also considered. For one internal reflection, the THz pulse additionally travels through two times the sample thickness, compared to the original Δt. $\Delta t_{1\text{etl}}$ is thus calculated by

$$\Delta t_{1\text{etl}} = \Delta t + \frac{2Hn_{\text{eff}}}{c}$$

C.2.2 Fourier Transform

The following and the user-selected processing options in CaTSper apply to both the reference and sample data.

Windowing

The time range in which relevant data needs to be Fourier transformed shall be specified. This can be done manually or via the auto window function. The auto window has a time range of $(-\Delta t_{1\text{etl}} + \Delta t, \Delta t_{1\text{etl}})$. This makes sure the reference and sample signals are equally spaced from the auto window's axis of symmetry. In addition, the reference and sample signal are respectively spaced from the start and end of the auto window function at a time equivalent to the additional time taken for one internal reflection.

Before performing a Fourier transform, it is crucial to consider that the Fourier transform assumes the time-domain signal to be periodic as it decomposes it into sinusoidal waveforms. However, it is typically non-periodic, and its ends are not smoothly connected to each other, resulting in high-frequency components in the frequency domain. Applying an apodisation function smoothes out these discontinuities at the signal's edges, reducing the high-frequency content and minimising spectral distortions. The following lists the available apodisation functions in CaTSper:

Boxcar Heaviside step function. The values of the selected data are not changed, and hence, the function is suitable for transient data.

Bartlett Symmetrical triangular function with zero as the two end values. The value at the triangular peak positively scales with the length of the data. The function length is the same as the data length. It gives little ripple in the results obtained after Fourier transform.

Blackman Summation of three cosine terms. The function is created with a length greater than the data length by one, and then the last value is removed from the function. It is suitable for applications where minimal leakage is required.

Hann Raised cosine. The two end values are at zero. The function length is one greater than the data length. It is suitable for random signals and is good against spectral leakage.

Hamming Raised cosine. The two end values are not at zero. The function length is one greater than the data length. After Fourier transform, the side lobes have a value lower than that of Hann, making Hamming suitable for optimising signal quality.

Taylor The MATLAB default settings are used. The coefficients in the function are not normalised. After Fourier transform, it gives a narrow main lobe with side lobe values that decrease monotonically. It is suitable for radar applications.

Triangular Symmetrical triangular function. If the length of the data has an odd value, the two end values are zero and the triangular peak is at one. If the length is instead even, the two end values are equal to the reciprocal of the length, and a plateau, instead of a triangular peak, is resulted. The function length is the same as the data length.

Fast Fourier Transform

The data is usually upsampled before Fourier transform. Upsampling approximates the situation when the signal is sampled at a higher rate. This is done by extending the data length, where the new length is determined by multiplying the original length of data by a power of two. The exponent is specified by the user and should have a value greater than zero. The additional entries created beyond the original data length are filled with zeros.

The augmented data is then respectively discrete Fourier transformed into frequency domain via the MATLAB built-in fast Fourier transform function. A N-by-N transformation matrix is multiplied with the data. N is the length of the augmented data, or the original data length if upsampling is not performed. After Fourier transform, the values in the frequency-domain data are divided by the original data length (before upsampling) for scaling.

Frequency Range and Spectral Resolution

The frequency-domain data will be trimmed according to the user-specified frequency range, which should be set based on considerations such as the instrument's signal-to-noise ratio, the range that gives relevant features. Values beyond the upper limit, which is the cutoff frequency, can be trimmed right after Fourier transform, but those below the lower limit are only trimmed after phase unwrapping, as otherwise erroneous values may result.

The spectral resolution ν_{res} of the frequency-domain data is defined by

$$\nu_{\text{res}} = \frac{1}{t_{\text{res}} N}$$

where t_{res} is the time resolution of the measured signal in time domain.

Amplitude and Phase Amplitude data are the scaled data obtained after fast Fourier transform. Phase data is obtained by unwrapping the frequency-domain data. The built-in MATLAB 'unwrap' function is adopted as it eliminates discontinuities between consecutive phases by adding multiples of $\pm 2\pi$ until the difference is less than π.

Due to the high signal-to-noise ratio at 0.8 THz, it is set as the starting point for unwrapping phase to reduce errors. This is instrument-specific, and one can change the value accordingly by accessing the 'TDSunwrap' function in the `CaTSper` source code.

Frequency domain data, that corresponds to frequencies greater than 0.8 THz, will be unwrapped in increasing values starting at 0.8 THz and vice versa for data corresponding to frequencies less than 0.8 THz.

A straight line is then fitted to unwrapped phase against a frequency from 0.05 to 0.4 THz. The intercept of the straight line at 0 THz gives the phase offset. The phase offset is then applied to all phase data for correction.

C.2.3 Frequency-Domain Analysis

Dynamic Range In CaTSper's DR Filter app, the user can first specify the cutoff frequency ν_{cutoff}. The noise floor $E_{\text{ref}}(\nu_{\text{cutoff}})$ is defined as the reference signal amplitude at ν_{cutoff}. The dynamic range DR can thus be defined as

$$\text{DR}(\nu) = \frac{E_{\text{ref}}(\nu)}{E_{\text{ref}}(\nu_{\text{cutoff}})}$$

where $E_{\text{ref}}(\nu)$ is the amplitude of the reference signal at frequency ν.

The upper limit frequency can also be specified in CaTSper's DR Filter app so that data at frequencies that are greater than the upper limit frequency will not be considered for analysis in the next steps.

Transmittance

Transmittance measures the fraction of the terahertz wave that is transmitted through the sample to the detector. The transmission amplitude T is defined as

$$T(\nu) = \frac{|E_{\text{sample}(\nu)}|}{|E_{\text{ref}(\nu)}|}$$

where E_{sample} is the frequency-domain amplitude of the sample measurement and E_{ref} is the frequency-domain amplitude of the reference.

The transmission phase ϕ_T is defined as

$$\phi_T(\nu) = \phi_{\text{ref}}(\nu) - \phi_{\text{sample}}(\nu)$$

where ϕ_{sample} is the frequency-domain phase of the sample measurement ϕ_{ref} is the frequency-domain phase of the reference measurement.

Refractive Index

Refractive index is a material property which measures the ratio between the speed of light in vacuum to that in the material. Both the refractive index of the reference n_{ref} and the medium n_{medium} are taken as one to match the methods in [17] for subsequent analysis. The frequency-domain effective refractive index $n_{\text{eff,FD}}$ of the sample can thus be calculated as

$$n_{\text{eff,FD}}(\nu) = \frac{c\phi_T(\nu)}{2\pi\nu H} + 1$$

where H is the sample thickness.

Absorption Coefficient

The absorption coefficient α quantifies the extent of loss in terahertz wave intensity through absorption. A higher value indicates higher absorption. The method by [17] is used to calculate α.

The reference factor is first determined using

$$\text{Reference factor}(\nu) = \frac{4n_{\text{medium}}n_{\text{ref}}}{(n_{\text{medium}} + n_{\text{ref}})^2}$$

As discussed earlier, both n_{medium} and n_{ref} take a value of one to match the methods in [17].

The sample factor is similarly defined as

$$\text{Sample factor}(\nu) = \frac{4n_{\text{medium}}n_{\text{eff,FD}}(\nu)}{\left(n_{\text{medium}} + n_{\text{eff,FD}}(\nu)\right)^2}$$

α is then calculated by

$$\alpha(\nu) = -\frac{2}{H}\ln\left(T(\nu)\frac{\text{Reference factor}(\nu)}{\text{Sample factor}(\nu)}\right)$$

In CaTSper's DR Filter app, the dynamic range of $\alpha(\nu)$ can be checked by the maximum absorption coefficient $\alpha_{max}(\nu)$, which can be calculated by

$$\alpha_{\text{max}}(\nu) = \frac{2}{H}\ln\left(\text{DR}\frac{\text{Reference factor}(\nu)}{\text{Sample factor}(\nu)}\right)$$

which references the method in [17].

Extinction Coefficient

Similar to α, the extinction coefficient κ is defined as the extent that terahertz wave can penetrate through the material. A higher value indicates a lower degree of penetration. κ is calculated using the Beer-Lambert Law

$$\kappa(\nu) = \frac{\alpha(\nu)c}{4\pi\nu}$$

Dielectric Constant

Permittivity measures the tendency of a material to be polarised by an electric field. The dielectric constant ε is defined as the ratio between the permittivity of the material to that of vacuum, which takes a value of one. The real and imaginary part of ε is calculated separately by

$$\text{Re}(\varepsilon(\nu)) = n_{\text{eff,FD}}(\nu)^2 - \kappa(\nu)^2$$

$$\text{Im}(\varepsilon(\nu)) = 2n_{\text{eff,FD}}(\nu)\kappa(\nu)$$

C.2.4 Data Manipulation

Finding Peaks

The MATLAB built-in function 'findpeaks' is used to identify peaks for a set of selected data (e.g. absorption coefficient α) against another (e.g. frequency). A peak is defined such that it has a value greater than its adjacent neighbours or has a value of infinity. A minimum peak prominence can be specified such that only peaks with prominence greater than that will be recorded (see MATLAB documentation for details on this parameter).

Acknowledgements We would like to express our sincere appreciation to Prince Bawuah from Menlo Systems, Germany, Sivaloganathan Kumaran, and Harvey Beere from the Semiconductor Physics Group at the Cavendish Laboratory, University of Cambridge, UK; Johanna Kölbel and Daniel Mittleman from Brown University, USA; Jean-Paul Guillet and Patrick Mounaix from the University of Bordeaux, France; Withawat Withayachumnankul from the University of Adelaide, Australia; Hungyen Lin from the University of Lancaster, UK; Riccardo Degl'Innocenti from Queen Mary University, UK; Emma MacPherson from the University of Warwick, UK; Martin Koch and Enrique Castro-Camus from the Philipps-University Marburg, Germany; and Andrew Burnett, University of Leeds, UK, for their valuable contributions, suggestions, ideas and encouragement. In particular, we would like to thank Jens Neu from the University of North Texas and Uriel Tayvah both formerly of the Schmuttenmaer Lab at the University of Yale, USA, for their enthusiasm and support in integrating the Nelly package (https://github.com/YaleTHz/nelly) into CaTSper in an upcoming release. Through our collaboration, the dotTHz project has been greatly enriched and the development of the compatible tools has been advanced.

Author Contribution JL: conceptualisation, software, writing—original draft. CKL: validation, data curation, writing—original draft, writing—online repository documentation, software—inline annotations. MM: validation, data curation, writing—online repository documentation. JW-B: validation, data curation. SS: validation, data curation, writing—review and editing. JAZ: conceptualisation, supervision, project administration, writing—review and editing

Availability of Data and Materials Online repository.

Code Availability Online repository.

Declarations

Competing interests The authors declare no competing interests.

Ethics Approval Not applicable

Consent to Participate Not applicable

Consent for Publication Not applicable

Conflict of Interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted

by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Smith, P.R., Auston, D.H., Nuss, M.C.: Subpicosecond photoconducting dipole antennas. IEEE Journal of Quantum Electronics 24(2), 255–260 (1988). https://doi.org/10.1109/3.121
2. Hu, B.B., Nuss, M.C.: Imaging with terahertz waves. Optics Letters 20(16), 1716 (1995). https://doi.org/10.1364/ol.20.001716
3. Leitenstorfer, A., Moskalenko, A.S., Kampfrath, T., Kono, J., Castro-Camus, E., Peng, K., Qureshi, N., Turchinovich, D., Tanaka, K., Markelz, A., Havenith, M., Hough, C., Joyce, H.J., Padilla, W., Zhou, B., Kim, K.-Y., Zhang, X.-C., Jepsen, P.U., Dhillon, S., Vitiello, M.S., Linfield, E.H., Davies, A.G., Hoffmann, M., Lewis, R., Tonouchi, M., Klarskov, P., Seifert, T., Gerasimenko, Y., Mihailovic, D.D., Huber, R., Boland, J., Mitrofanov, O., Dean, P., Ellison, B., Huggard, P., Rea, S., Walker, C., Leisawitz, D., Gao, J.R., Li, C., Chen, Q., Valusis, G., Wallace, V.P., MacPherson, E., Shang, X., Hesler, J., Ridler, N., Renaud, C., Kallfass, I., Nagatsuma, T., Zeitler, A., Arnone, D., Johnston, M., Cunningham, J.: The 2023 Terahertz Science and Technology Roadmap. Journal of Physics D: Applied Physics (2023). https://doi.org/10.1088/1361-6463/acbe4c
4. Dhillon, S.S.,Vitiello, M.S., Linfield, E.H., Davies, A.G., Hoffmann, M.C., Booske, J., Paoloni, C., Gensch, M., Weightman, P., Williams, G.P., Castro-Camus, E., Cumming, D.R.S., Simoens, F., Escorcia-Carranza, I., Grant, J., Lucyszyn, S., Kuwata-Gonokami, M., Konishi, K., Koch, M., Schmuttenmaer, C.A., Cocker, T.L., Huber, R., Markelz, A.G., Taylor, Z.D., Wallace, V.P., Zeitler, J.A., Sibik, J., Korter, T.M., Ellison, B., Rea, S., Goldsmith, P., Cooper, K.B., Appleby, R., Pardo, D., Huggard, P.G., Krozer, V., Shams, H., Fice, M., Renaud, C., Seeds, A., Stöhr, A., Naftaly, M., Ridler, N., Clarke, R., Cunningham, J.E., Johnston, M.B.: The 2017 terahertz science and technology roadmap. Journal of Physics D: Applied Physics 50(4), 043001 (2017). https://doi.org/10.1088/1361-6463/50/4/043001
5. Koch, M., Mittleman, D.M., Ornik, J., Castro-Camus, E.: Terahertz time-domain spectroscopy. Nat. Rev. Met. Pri. 3(1), 48 (2023). https://doi.org/10.1038/s43586-023-00232-z
6. Naftaly, M.: An international intercomparison of thz time-domain spectrometers. In: 2016 41st International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz), pp. 1–2 (2016). https://doi.org/10.1109/IRMMW-THz.2016.7758763
7. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K.: The FAIR guiding principles for scientific data management and stewardship. Sci. Dat. 3(1), 160018 (2016). https://doi.org/10.1038/sdata.2016.18
8. Terahertz Applications Group, University of Cambridge: CaTx (2023). https://github.com/dotTHzTAG/CaTx
9. Terahertz Applications Group, University of Cambridge: CaTSper (2023). https://github.com/dotTHzTAG/CaTSPer
10. Terahertz Applications Group, University of Cambridge: The dotTHz Project Documentation (2023). https://github.com/dotTHzTAG/Documentation
11. The HDF Group: The HDF Group's HPC program (2018). https://www.hdfgroup.org/2015/08/hdf-group-hpc-program/
12. Pupeza, I., Wilk, R., Koch, M.: Highly accurate optical material parameter determination with thz time-domain spectroscopy. Opt. Express 15(7), 4335–4350 (2007). https://doi.org/10.1364/OE.15.004335
13. Peretti, R., Mitryukovskiy, S., Froberger, K., Mebarki, M.A., Eliet, S., Vanwolleghem, M., Lampin, J.-F.: Thz-tds time-trace analysis for the extraction of material and metamaterial parameters. IEEE Transactions on Terahertz Science and Technology 9(2), 136–149 (2019). https://doi.org/10.1109/TTHZ.2018.2889227

14. Greenall, N.R.: Parameter Extraction and Uncertainty in Terahertz Time-Domain Spectroscopic Measurements (2017). https://etheses.whiterose.ac.uk/19045/
15. Vogel, T., Mansourzadeh, S., Nandi, U., Norman, J., Preu, S., Saraceno, C.J.: Photoconductive receivers at 1030 nm for high average power pulsed THz detection (2023)
16. Tayvah, U., Spies, J.A., Neu, J., Schmuttenmaer, C.A.: Nelly: A user-friendly and open-source implementation of tree-based complex refractive index analysis for terahertz spectroscopy. Analytical Chemistry 93(32), 11243–11250 (2021). https://doi.org/10.1021/acs.analchem.1c02132. PMID: 34355556
17. Jepsen, P.U., Fischer, B.M.: Dynamic range in terahertz time-domain transmission and reflection spectroscopy. Opt. Lett. 30(1), 29–31 (2005). https://doi.org/10.1364/OL.30.000029

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.