HYPERION: Hyperspectral Penetrating-type Ellipsoidal Reconstruction for Terahertz Blind Source Separation

Chia-Hsiang Lin\textsuperscript{1,2,6}, Yi-Chun Hung\textsuperscript{3,6}, Feng-Yu Wang\textsuperscript{4,5}, and Shang-Hua Yang\textsuperscript{4,†}

\textsuperscript{1}Department of Electrical Engineering, National Cheng Kung University
\textsuperscript{2}Miin Wu School of Computing, National Cheng Kung University
\textsuperscript{3}Department of Electrical and Computer Engineering, University of California, Los Angeles
\textsuperscript{4}Department of Electrical Engineering, National Tsing Hua University
\textsuperscript{5}Department of Life Science, National Tsing Hua University
\textsuperscript{6}Contributed equally

April 5, 2022

1 Abstract

Terahertz (THz) technology has been a great candidate for applications, including pharmaceutic analysis, chemical identification, and remote sensing and imaging due to its non-invasive and non-destructive properties. Among those applications, penetrating-type hyperspectral THz signals, which provide crucial material information, normally involve a noisy, complex mixture system. Additionally, the measured THz signals could be ill-conditioned due to the overlap of the material absorption peak in the measured bands. To address those issues, we consider transmitted (penetrating-type) signal mixtures and aim to develop a \emph{blind} hyperspectral unmixing (HU) method without requiring any information from a prebuilt database, such as predefined material composition or material information. The proposed HYperspectral Penetrating-type Ellipsoidal ReconstructION (HYPERION) algorithm is unsupervised, not relying on collecting extensive data or sophisticated model training. Instead, it is developed based on elegant ellipsoidal geometry under a very mild requirement on data purity, whose excellent efficacy is experimentally demonstrated.

2 Introduction

The recent advances in remote sensing technologies have revolutionized the ways people interact with the world in communication, manufacturing, molecular science, life science, healthcare,
geography, and astronomy [1–3]. Among all remote sensing methods that perceive information non-invasively, electromagnetic (EM) wave has become one of the most critical tools due to their promising nature of the high-speed operation, long traveling distance, and unique light-matter interactions in specific spectral bands [4]. With the use of the EM wave, unveiling multi-functional behaviors inside objects with deep-subwavelength spatial precision and tera-frame-per-second imaging speed has also been well-demonstrated nowadays [5]. Terahertz (THz) spectrum, located between microwave and infrared EM spectrum, has recently aroused extensive attention for identifying a great variety of materials, including molecules, proteins, explosives, chemical mixtures, and charged particles in a remote distance [6–9]. The chemical compositions and the molecular structures of materials can directly map their unique molecular dynamics in the THz frequency regime. This enables the THz spectroscopy tools to unveil fruitful material information inside optically opaque objects without any labels. Due to the non-ionizing, non-destructive nature of THz wave, THz spectroscopy systems are of great interest for medical, pharmaceutical, non-destructive evaluation, and industrial inspection [10, 11]. Different systems based on THz spectroscopic modalities have further extended the application fields. THz near-field systems such as THz scanning near-field microscopy [12] and THz scanning tunneling microscopy [13] significantly shrink spatial resolution down to atomic level, which becomes powerful tools to perform near-field studies of advanced materials and resolving structural information of biological samples [13]. Material dynamics down to tens of femtosecond temporal resolution can be further determined through time-resolved THz spectroscopy [14]. In addition, hyperspectral THz imaging systems map out not just water content inside biological samples but chemical distribution inside a sealed package [10, 11].

Although THz spectroscopy has shown its great promises in the past decades, many materials, chemical compounds, and biomolecules have not yet been fully characterized in the THz frequency range, unlike in visible light and infrared spectrum. Furthermore, different THz spectroscopic modalities, measuring conditions, or sample preparation methods would introduce considerable discrepancies between samples with the same material [15, 16]. The typical way to perform THz spectroscopy for material identification is to establish THz material characteristics of pure substances locally, followed by material information extraction [17]. This way, a ppb-level sensing capability has been demonstrated by evaluating the spectral shifts, and amplitude changes of the material absorption spectrum in the THz regime [18]. However, in many real-world scenarios, such as identifying chemical portions of malignant tumors or in situ quantifying glucose in a body fluid sample, it is still challenging to purify and extract pure substances for further THz material dataset establishment. This physical constraint significantly limits the practical use of THz spectroscopy systems in a wide range of applications. We want to ask here: Can we separate material signatures of mixture chemicals without measuring or even knowing their pure substances? This is a blind spectral unmixing problem waiting to be answered, especially in THz spectroscopy, THz hyper-
spectral imaging, and THz biophotonics fields. Recently, research groups have implemented THz spectral unmixing methods for material blind separation – hard modeling factor analysis (HMFA) [19], nonnegative matrix factorization (NMF) [20], and independent component analysis (ICA) [21]. However, all the conventional THz spectral unmixing methods still rely on properties of material information, the size of a measured dataset, and the signal-to-noise ratio of the dataset, which severely limits their scope of practical use. HMFA uses the correlation of mixture signals to unmix sources, which requires curve fitting for the massive amount of material information, such as spectral peak amplitudes, peak locations, and spectral profiles. The blind unmixing performance of the nonconvex NMF is sensitive to initial parameters and the signal-to-noise ratio, leading to difficulties for practical uses. One of the requirements in ICA is the statistical independence of the sources, causing inaccurate amplitudes of THz unmixed spectrum compared with the ground truth. Here, we proposed an unsupervised blind separation algorithm, HYperspectral Penetrating-type Ellipsoidal ReconstructION (HYPERION), which needs neither collecting big data nor sophisticated model training. In this article, we demonstrated a comprehensive study on HYPERION for non-invasive, non-destructive material separation and identification. To the best of our knowledge, HYPERION has shown superior performance among the state-of-the-art THz spectral unmixing methods regarding separation accuracy (i.e., spectral angle error, mean square error) and noise immunity testing with mixture materials. Due to the superior blind separation capability, we have further demonstrated highly accurate material blind separation mapping through HYPERION THz hyperspectral imaging. It should be noted that there is no prior knowledge of pure substance information, no supervised model training, no need to establish THz material dataset for pure substances, and no restricted requirement on the data purity \( \gamma \). Our proposed blind separation modality, HYPERION, could potentially make paradigm shifts of THz technology for chemical sensing, industrial inspection, material identification, biomedical sensing, and imaging in the near future.

3 Results

Assumption evaluation

Our method, HYPERION, is designed based on Löwner-John ellipsoid (LJE) and the linear mixing model ((Supplementary Note 14)), where the THz data follows a convex geometry structure (see “Methods” for the details). By utilizing the information of the LJE defined as the maximum-volume ellipsoid inscribed in the convex geometry, HYPERION has the efficacy to blindly unmix sources from the data with low data purity [22]. To validate the linear mixing model on the transmitted THz signals, we have measured and characterized mixture tablets composed of D-lactose monohydrate, D-sucrose, and L-tyrosine through a broadband THz time-domain spectroscopy (THz-TDS) system. The three chemicals are selected due to their abundant information within THz spectral bands,
including spectral absorption lines and varying absorption spectral trends. The obtained ternary
dataset contains the three pure-substance THz spectra (D-lactose monohydrate, D-sucrose, and L-
tyrosine) and six mixture spectra composed by each pair of the chemicals with a ratio of 3:7, which
is designed to test the data purity criterion. The details of measurement protocol and environment
condition are elaborated in Supplementary Note 1.

We compared the THz spectra simulated by the convex combinations of the three chemicals with
their corresponding THz measured spectrum with the ternary dataset. All samples are prepared
based on the standardized protocol (see “Methods” for the details). As shown in Fig. 1(a), all
the simulated spectra are highly aligned with the measured spectra, including the spectral features
and the locations of spectral lines. We then project high dimensional data of the ternary dataset
into a low dimensional space through principal component analysis (PCA). The PCA evaluation
result provides solid evidence that the necessary criteria of holding the linear mixing model are
well-preserved (Supplementary Note 2). To further explore the linear mixing model in terms of
spectrum similarity and accuracy, we use spectral angle mapper (SAM) and root-mean-square
error (RMSE) [23, 24] to evaluate the diminutive differences between simulated spectrum and
measured spectrum (Supplementary Note 3). The SAM indicator is based on the projection angle
between two spectra in comparison, which explicitly reveals the spectral shape similarity without
mixing the influence of spectrum offset. The RMSE indicator is the square root of the quadratic
mean of differences between the two spectra. It includes both shape similarity and the effects
of the band-wise differences among the measured spectral range. The SAM and RMSE of this
ternary dataset are less than 5 degrees and 6 cm\(^{-1}\) as shown in Fig. 1(b), which demonstrates
both the excellent spectral shape similarity as well as the slight deviation between the measured
and simulated mixture spectra. Based on the validation of qualitative and quantitative analysis,
the assumption of the linear mixing model is experimentally verified. Note that the linear mixing
model has been extensively used in remote sensing for reflected (reflecting-type) signals but never for
transmitted (penetrating-type) signals as far as we know; thus, the experiment is needed. We then
transform the ternary dataset from the THz transmission spectrum to the material absorption
spectrum of each measured sample (Supplementary Note 4), which directly presents broadband
material information, including resonant frequencies, spectral features, and light-matter interaction
levels in the THz range. Additionally, we extract and present material absorption spectrum between
0.2 THz to 1.75 THz since the ASOPS THz-TDS system dynamic range in power spectra of the
selected materials approximately decreases to about 5 dB at 1.75 THz that start to mismatch with
the linear mixing model. As shown in Fig. 1(c), all unmixed pure substance absorption spectra
are highly aligned with the measured ones, including the spectral peak locations and the material
absorption profiles. The unmixed D-glucose and L-tyrosine values are slightly deviated from the
measured values above 1.5 THz since the signal-to-noise ratio (SNR) of the THz spectroscopy
system gradually decreases from 60 dB at 0.3 THz to 5 dB at 1.75 THz.

Nevertheless, at low-SNR spectral regions (1.2 - 1.6 THz), HYPERION still provides great unmixing capability - less than 3.51 cm\(^{-1}\) deviated absorption coefficient values from the measured ones. For the spectral regimes with a noticeable unmixing difference, for example, the measured values of L-tyrosine absorption peak are 15.35 cm\(^{-1}\) and 4.78 cm\(^{-1}\) differences from the unmixed values at 0.949 THz and 1.295 THz, respectively. Due to the overlap of water vapor and material absorption peaks, the linear mixing model is not well-preserved. Since HYPERION takes all the spectral information into account by the LJE instead of specific spectral components, even in non-linear matter-matter or light-matter interaction spectral regions, the unmixed material absorption spectrum based on HYPERION still demonstrates great spectral accuracy of each source spectral signature but just a relatively higher mismatch in absorption coefficient values.

Despite the nonlinear interaction and the low SNR levels in certain bands, HYPERION offers an accurate unmixing performance – 10.63 degrees with SAM and a 2.54 cm\(^{-1}\) with RMSE of overall unmixed material absorption spectrum has been achieved. This SAM value presents high spectral shape similarity of unmixed absorption coefficients, preserving rich material spectral profiles and details among a wide frequency range. In another perspective, the extraordinarily low RMSE shows both high unmixing accuracy and spectral similarity of the unmixed absorption coefficient, providing various types of material spectral features for further chemical identification and non-destructive evaluation applications.

**Quinary Dataset with Pure Substances**

HYPERION has several advantages, including support of complex mixture systems, milder data purity requirement, excellent spectral accuracy on unmixed sources, and remarkable noise immunity, due to the simultaneous use of all spectral information by LJE. To further investigate the performance of HYPERION in broad application scope, we design three experimental modules to verify HYPERION characteristics in terms of unmixing capability of complex mixture system, low data purity dataset, and noisy environment condition based on the quinary dataset (see “Methods” for the details). The quinary dataset contains the five pure chemicals (D-glucose, D-lactose monohydrate, L-tyrosine, L-histidine, and D-sucrose) with rich material information in THz spectral range and ten mixture tablets. The five pure chemicals are specifically selected to validate HYPERION performance under different unmixed spectral scenarios - identifying single/multiple spectral peaks and overlaid spectral profiles simultaneously. Within the spectral range of 0.2 - 1.7 THz, L-tyrosine and L-histidine have a single absorption peak at 0.95 and 0.78 THz, respectively. D-lactose monohydrate contains two absorption peaks located at 0.52 THz and 1.3 THz. D-sucrose and D-glucose present the high spectral similarity of their THz absorption spectra, which is used to evaluate ill-conditioned levels of HYPERION. Furthermore, D-lactose is also selected to investigate
Figure 1. Unmixed results of the ternary dataset by HYPERION. (a) The three subplots at the top panel are the measured spectra of pure substances (D-glucose, L-tyrosine, and D-lactose monohydrate). Gray arrows indicate the absorption lines of water vapor. The bottom two rows of subplots are the mixture spectra with different mixing ratios; D-glucose, D-lactose monohydrate, and L-tyrosine are denoted as G, L, and T, respectively. The simulated lines (black) are computed based on the linear mixing model and the measured spectra of the pure substances. (b) The spectral angle mapper (SAM) and root mean square error (RMSE) between the simulated and measured spectra. G:L:T represents the proportion of D-glucose, D-lactose monohydrate, and L-tyrosine. (c) The comparison between the unmixed material absorption spectrum and the measured absorption spectrum upon the ternary dataset with pure substances. The unmixed and measured material absorption spectrum are shown as solid lines and dotted lines, respectively.
the unmixing performance of HYPERION on the pharmaceutical application since D-lactose is the common ingredient as an excipient in tablets [25, 26]. The ten mixture tablets are composed of each pair of the five chemicals with a 5:5 ratio, which satisfies the milder requirement of the data purity for a quinary case [27]. Identical experimental setup, tablet preparation, data acquisition, and data preprocessing protocol are well-followed as the previous ternary dataset establishment.

Comparing the unmixed and measured absorption spectra shows that HYPERION has great efficacy on the different spectral scenarios. As shown in Fig. 2, the five unmixed absorption spectra (solid lines) based on the quinary dataset demonstrate a high correlation with the measured absorption spectra (dotted lines) within a broad frequency range. In terms of spectral peak locations, unmixed absorption spectra with varieties of material properties are precisely resolved. The spectral peak deviation of all unmixed absorption spectra is less than 10 GHz, close to the spectral resolution limit of the asynchronized optical sampling (ASOPS) THz-TDS system (see “Methods” for the details) at protocol settings. Additionally, the increasing trends and the characteristics of the absorption spectra are well unmixed since HYPERION utilizes the convex geometry of the whole measured frequency bands for the blind separation instead of specific bands information, such as peak regimes. Those mismatched values between unmixed and measured absorption spectra mainly come from nonlinear light-matter interaction [28], and limited SNR levels at higher frequencies.

To address the ill-conditioned case, HYPERION transforms the data matrix into the preconditioned space, where the similar material absorption spectra can be easily separated (see “Methods” for the details). As shown in Fig. 2, D-glucose and D-sucrose absorption spectra are highly correlated except for the slightly different slope within the 0.2 THz – 1.2 THz frequency range. Under this condition, HYPERION again shows the unmixing capability of a SAM of 9.57 degrees and an RMSE of 2.30 cm$^{-1}$, demonstrating its resilience to ill-conditioned scenarios with the complex mixture system. Considering different spectral characteristics among the quinary dataset, HYPERION still reaches 12.15 degrees and 2.88 cm$^{-1}$ in SAM and RMSE, respectively. Compared with less complicated material systems (e.g., ternary system), HYPERION maintains a similar accuracy level of unmixed material spectrum based on complex mixture systems (e.g., quinary system). Additionally, HYPERION is featured for its low computation time due to the convexity nature of the LJE (see “Methods” for the details). The computation time of HYPERION on this dataset is less than 3 seconds under a general personal laptop (Intel Core i5 with 8 GB memory).

**Quinary Dataset without Pure Substances**

In most on-site blind source separation scenarios, such as pharmaceutical inspection, biomarker analysis, remote sensing, and chemical identification, the existence and identification of pure substances remain unknown, which generally involve highly mixed chemicals and imply a low data
Figure 2. Unmixed material absorption spectra of the quinary dataset with pure substances. The comparison between unmixed material absorption spectrum (solid line) and measured material absorption spectrum (dotted line) upon the quinary dataset with pure substances. The subset of the figure shows that the D-glucose and D-sucrose absorption spectra are highly correlated. The material absorption spectrum difference between D-glucose and D-sucrose within 0.2 - 1.2 THz is dominated by different slopes (gray region) while it is dominated by shape within 1.2 - 1.75 THz.
Figure 3. Comparison of unmixed material absorption spectra with and without pure substances of the quinary dataset. Quinary dataset (D-glucose, D-lactose monohydrate, D-sucrose, L-tyrosine, L-histidine) with and without pure substances unmixed by HYPERION are indicated by solid and dotted lines, respectively. The values of spectral angle mapper (SAM) and root-mean square error (RMSE) of dataset with and without pure substances are also shown (with/without) to evaluate the spectral shape similarity.

purity. To this end, it is crucial to extract source signals from mixtures with low data purity [22].

Here, we have evaluated the HYPERION based on the same quinary dataset but excluded all pure substances. There is no prerequisite information from all mixture tablets in this testbed, including spectral information, the existence of pure substances, and spectra composition. To address this type of dataset with scarce information, HYPERION fits a simplex to the data convex geometry in the preconditioned space. When the optimal simplex is constructed, the corners of the optimal simplex correspond to the unmixed sources (viewed as vectors) in the preconditioned space (see “Methods” for the details). As shown in Fig. 3, the difference of unmixed absorption spectra among cases with and without pure substances is almost negligible. This is because LJE has been demonstrated to provide comparable performance with the pure substance case [29], as long as the requirement of data purity is satisfied (Supplementary Note 5). The slight difference in performance comes from the measurement noise of the dataset and the number of measured datasets – 15 versus 10 measured tablets. Without the need for pure substance information and its measurement on-site, HYPERION has the potential to expand the THz spectroscopy application scopes, opening up the door to industrial applications in non-invasive sensing, chemical identification, and in vivo biomarker extraction.
Noise Immunity Evaluation

To evaluate the efficacy of HYPERION with THz spectroscopy systems in general, we further investigated the noise immunity performance of HYPERION. In this study, we apply additive white Gaussian noise (AWGN) to the measured THz time-domain electric field signals with a 0.001% - 0.1% standard deviation (SD) range (Supplementary Note 10). Moreover, the quinary THz dataset without pure substances is chosen to better evaluate real-world scenarios – low data purity and noisy condition. In those conditions, the data convex geometry will be distorted, resulting in the inaccuracy of unmixed material absorption spectra. To overcome the issue, HYPERION includes a general regularizer in the objective function to accommodate the distortion of convex geometry. With the general regularizer and the regularization parameter, $\lambda$, HYPERION does not force unmixed spectra to form the theoretically required regular structure in the noiseless case. Instead, it encourages the regular structure (see “Methods” section for the details). Based on the design of the regularizer, HYPERION is well-suited for the penetrating-type THz data, which is normally ill-conditioned and noisy. As shown in Fig. 4(a), the SAM and RMSE of HYPERION unmixing results maintain at low levels while the noise SD is less than 0.1%. In the cases of noisy environment (e.g., top six highest SD conditions), HYPERION performs slightly inferior since the SNR of THz signals in the high-frequency regime drop dramatically at higher noise levels, which causes severe distortion of the data convex geometry and leads to a deteriorating impact of the unmixing performance. To further evaluate the noise immunity capability of HYPERION among unmixed chemicals and frequency ranges, the unmixed absorption spectra of the five pure substances are demonstrated at 0.1% noise SD (Fig. 4(b)-(f)). Under this severe noise condition, the unmixed absorption spectra through HYPERION still show great alignment with the noise-free ground truth in low-frequency bands. Within the 0.2 – 1 THz frequency range, the SAM and RMSE of the D-glucose absorption spectrum are 12.32 degrees and 2.50 cm$^{-1}$, respectively. In higher frequency bands, a noticeable difference of every material absorption spectrum starts to show. Although the global trends and features are mostly well-preserved, the high-frequency spectral fluctuations due to the severe additive noise could be an issue for high-precision chemical identification applications.

In sum, HYPERION demonstrates its noise immunity feature with distorted convex geometry datasets. By utilizing this powerful feature, HYPERION not only can fit in a strictly controlled data acquisition environment but also be well suitable for real-world blind detection schemes under a noisy environment.

Comparison with commonly used unmixing methods

To further evaluate the unmixing performance of the geometry-based HYPERION, we introduce the commonly used unmixing methods as a comparison: statistic-based nonnegative independent component analysis (nICA), algebra-based nonnegative matrix factorization (NMF), model-based
Figure 4. The performance of HYPERION under different noise standard deviation (SD). (a) The root mean square error (RMSE) and spectral angle mapper (SAM) of HYPERION under different noise SD. (b-f) The unmixed absorption spectra and ground truth material absorption spectra of L-histidine, D-sucrose, D-glucose, D-lactose monohydrate, and L-tyrosine within 0.2 - 1.75 THz under SD = 0.1%.
hard modeling factor analysis (HMFA), and geometry-based successive projection algorithm (SPA) [20, 21, 30–33]. nICA is one of the modified versions of ICA, where it imposes the nonnegative constraint on unmixed sources. With the nonnegative constraint, nICA can unmix the spectra of nonnegative values and lead to better performance than ICA since the fraction of incident radiation absorbed by the material is always nonnegative, leading to the nonnegative values of a material absorption spectrum. NMF is the unmixing method for blind separation problems designed according to the non-convex optimization approaches (Supplementary Note 13). HMFA is the unmixing method based on peak fitting and has demonstrated its efficacy in mid-infrared and THz bands [19, 32]. SPA uses the selection and projection in the vector space to unmix the blind sources. The comparison among HYPERION, nICA, NMF, HMFA, and SPA under different noise SD is shown in Fig. 5. From the unmixed material spectra and the RMSE values, HYPERION has outperformed the THz unmixing methods under a broad noise SD range from 0.001% to 0.1%. As expected, it comes from the three designed properties of HYPERION. First, HYPERION filters out most noise in the affine fitting step, where the data is projected into the lower dimensional convex hull by the singular value decomposition approach. Second, HYPERION uses convex geometry to find the LJE, which can further address the ill-conditioned nature of THz source signals. Third, the embedded general regularizer provides better accommodation to the noisy data by soft fitting for the distortion of convex data geometry in the objective function. As shown in Fig. 5(a, e), although nICA can resolve the spectral absorption peaks, the unmixed material absorption spectra by nICA are severely compressed. The reason is that the nICA method assumes all material absorption spectra are independent in the complex mixture system [31]. However, the summation of the chemical composition (in %) must equal to one, making it impossible for the sources to be statistically independent, and this sum-to-one constraint leads to an assumption of nICA independent assumption. In Fig. 5(b), NMF shows decent unmixing capability because it can converge to the local optimum [34, 35]. However, NMF is relatively sensitive to noisy conditions, probably due to its non-convexity nature. As a result, the unmixing performance of NMF on the THz dataset bounces dramatically at different unmixing trails and noise levels. Consequently, it is challenging to accurately resolve unmixing material spectra under a low SNR spectral regime. As shown in Fig. 5(f), the unmixed absorption spectra through NMF deviate more from the ground truth signal compared to HYPERION while frequency increases. In the case of HMFA (Fig. 5(c, g)), it is not capable of resolving the ill-conditioned data (i.e., D-Glucose and D-Sucrose) nor the spectral trends since the peak fitting approach inherently has difficulties in differentiating overlapped features. As shown in Fig. 5(d, h), SPA has the inferior unmixing performance in the absorption peak regions since the pure-pixel assumption does not hold in those datasets [36]. The convergence time of these methods is also provided in Supplementary Table. 8. Overall, HYPERION delivers superior performance than the four THz unmixing modalities in resolving
broadband spectral features, unmixing ill-condition data, reducing noise impact, and preserving spectral details in a noisy environment.

Application

In the sections above, HYPERION has demonstrated superior features on unmixing penetrating-type THz material spectra. Those accurately unmixed spectra can benefit several applications. To demonstrate one of the potential applications, we propose the combination of the raster-scanning THz-TDS system and HYPERION (THz HYPERION) to visualize the “secret recipes” of arbitrary objects at a remote distance. We use the unmixed spectra from the quinary dataset without pure substances and prepare a small test set of 15 chemical tablets with different compositions based on D-lactose monohydrate, D-glucose, and L-tyrosine. The 15 tablets are mounted on the paper board covered by a copper foil, which is utilized to block THz signals encoding the material information of the paper board. Every tablet is measured by the THz-TDS system with the configurations described earlier. The images in visible light and THz band images of the test set are shown in Fig. 6(a) and 6(b), respectively. The unmixed material spectra by the quinary dataset without pure substances are used to restore the chemical compositions of the 15 tablets in the test set by the following convex optimization problem:

\[
\min_{r_i \in \mathbb{R}^3} \|x_i - Ar_i\|_1 \quad \text{s.t.} \quad r_i \geq 0, \quad 1^T r_i = 1,
\]

where \(i = 1, 2, \ldots, 15\). \(x_i\) and \(r_i\) are the measured THz material absorption spectra of the test set and the optimization variables to estimate the corresponding ratio of compositions, respectively. Columns of \(A\) are the unmixed spectra from the quinary dataset without pure substances. The \(L^1\) norm is chosen to effectively decrease the influence of the unmixed spectra deviation in the absorption peak regions. Since equation (1) is a convex problem, the solver CVX [37] is adopted for the following demonstration (Supplementary Note 6). The qualitative comparisons between the ground truth and the estimated material composition map are shown in Fig. 6(c, d). As expected, the majority of tablet material compositions are highly aligned with the ground truth except very few tablets (i.e. 10th, 11th and 12th tablets). This experimental result indicates that THz HYPERION is applicable for estimating the arbitrary composition of complex mixture systems while the pure material spectra are accurately unmixed. In Fig. 6(e), we further evaluate the unmixed material composition map qualitatively. Most unmixed results present less than 20% inaccuracy of each composition difference compared with ground truth. The inaccuracy of estimating material composition is caused by the number of spectra in the quinary dataset, the high absorption spectrum similarity of D-lactose monohydrate and D-glucose, and the increased
Figure 5. The performance comparison among HYPERION, NMF, HMFA and SPA under different noise SD. (a-d) The performance comparison between HYPERION with NMF, nICA, HMFA and SPA under different noise SD, respectively. The noise-free values are indicated with dotted lines. The unmixing performance is evaluated by root mean square error (RMSE) between the unmixed absorption spectra and the ground truth. (e-h) Five unmixed absorption spectra by NMF, nICA, HMFA and SPA are compared with the ground truth.
system noise in the high-frequency regime. It is worth mentioning that the accuracy of unmixed spectra would be greatly enhanced while the size of the measured dataset is increased. To this extent, the quality of unmixed material signatures can be significantly elevated by extending the measurement area and the pixel number of the multi-material objects under test. Furthermore, different regularizers can be further designed to match with HYPERION for specific application purposes. Overall, our demonstration shows the potential for some blind separation mapping applications, including pharmaceutical analysis, functional imaging, and space exploration.

4 Discussion

In this paper, we propose HYperspectral Penetrating-type Ellipsoidal ReconstructION (HYPERION) to blindly unmix the sources (i.e., the THz absorption spectra of pure substances) from transmitted THz signals, which are usually noisy and low data purity. In HYPERION, affine fitting and simplex fitting are utilized to address the noisy and low data purity issues, respectively. In affine fitting, the noise is filtered out by projecting the data into the lower dimensional convex hull; in simplex fitting, HYPERION is encouraged to fit a regular simplex to the data convex geometry in the preconditioned space, where a relatively mild data purity is required (i.e., $\gamma > \frac{1}{\sqrt{q-1}}$).

Additionally, the THz spectra with the overlapped material absorption peak in the measured THz bands, such as D-glucose and D-sucrose in 0.2 - 1.75 THz bands, can be ill-conditioned and lead to inferior performance. To address the ill-conditioned case, HYPERION transforms the transmitted THz signals into the preconditioned space, where the LJE information can easily separate the similar THz spectra. Based on the affine and simplex fittings and the transformation based on LJE information, HYPERION is capable of handling the ill-conditioned THz signals with noise and low data purity.

To evaluate the unmixing efficacy of HYPERION upon the ill-conditioned dataset with low data purity, we selected the quinary dataset without pure substances for the qualitative and quantitative analysis. In the ill-conditioned case, the THz material absorption spectra of D-glucose and D-sucrose are well-unmixed with RMSE of 2.59 cm$^{-1}$ and 3.08 cm$^{-1}$ compared with the ground truth spectra, respectively. In the low data purity case, even though the dataset only contains the highly mixed tablets (i.e., 5:5 for every pair of the pure substances), HYPERION still demonstrates the comparable unmixing performance of the dataset with pure substances. Note that, compared to the 1:9 mixing condition, the 5:5 mixing condition is indeed highly mixed. Additionally, for the noisy condition, we have also applied AWGN with 0.001% - 0.1% SD range to the quinary dataset without pure substances. In the SD range under 0.04%, HYPERION delivers similar unmixing performance; in the SD range higher than 0.04%, the unmixing performance is slightly inferior since the SNR of THz signals drops dramatically in the high-frequency regime. With the detailed
Figure 6. Blind chemical mapping by THz HYPERION. (a) The optical and (b) THz images of 15 tablets with different proportions composed of D-glucose, D-lactose monohydrate, and L-tyrosine, respectively. The THz image is scanned with the step size of 0.25 mm; the colormap shows the normalized field strength of the time-resolved transmitted THz signal. (c) The ground truth and (d) the experimental compositions of tablets. The color basis is blue, red, and green as shown in the ground truth 1st, 11th, and 15th points, respectively. The color channels are linear combination of the selected color basis by the corresponding ratio of the chemical composition. 1st, 11th, and 15th points are corresponding to D-glucose, D-lactose monohydrate and L-tyrosine, respectively. (e) The comparison between the estimated proportions by THz HYPERION and the ground truth (D-lactose monohydrate/L-tyrosine/D-glucose).
evaluation, HYPERION reveals the great unmixing efficacy on complex mixture systems under noisy conditions.

With the support of complex mixture systems, HYPERION is capable of contributing to the applications, including pharmaceutical analysis, biomedical diagnosis [38], and art conservation [39]. Among those applications, pharmaceutical analysis is much more challenging since it requires identification of drugs and monitoring the impurity of drugs [40]. In this sense, we have demonstrated how HYPERION can further help analyze drug compositions and impurity levels without prior drug recipe information. In the demonstration, a material absorption spectrum dataset, composed of the mixtures of D-lactose monohydrate, D-glucose, and L-tyrosine with distinctive material composition, is adopted to simulate the different impurity levels of drugs. Upon the dataset, THz HYPERION can precisely estimate most of the mixture composition and deliver less than 20% inaccuracy of each composition compared with ground truth.

THz hyperspectral imaging is a great extent for HYPERION for future work since the LJE can be more accurately estimated from a large number of spectra. The task-oriented regularizers can also be designed for the spatial relation of pixels, such as spatial continuity and self-similarity. To this extent, more precise unmixed material absorption spectra and complex material mapping with a wide composition range can be well-resolved. Furthermore, the more efficient optimizer can significantly decrease the computation time, which utilizes a large number of spectra in hyperspectral images. In addition to the above future work for the more precise estimation of unmixed spectra and material mapping, the required information of tablet thickness in HYPERION can be relaxed by combining with the methods estimating the complex refractive index based on THz-TDS systems [17, 41]. By relaxing the required information of tablet thickness, HYPERION can further contribute to the applications where the sample thickness cannot be measured and obtained.

5 Methods

Löwner-John ellipsoid (LJE) for transmitted (penetrating-type) THz signals

In complex mixture systems, the material absorption spectra of the overlapped material absorption peaks in the measured THz bands can be quite similar (i.e., D-sucrose and D-glucose). It is the so-called ill-conditioned HU problem, having drawn attention from very recent machine learning literature [42]. As far as we know, the most effective solution for addressing the ill-conditioned HU is based on the Löwner-John ellipsoid (LJE) theory [42], which elegantly exploits the data convex structure (i.e., the convex hull of signature vectors contains the data vectors; cf. equation (2)). However, almost all the HU theories (including LJE) were developed for the reflecting-type signals, probably because the need for the HU techniques mainly comes from the satellite hyperspectral remote sensing, for which the hyperspectral signals are reflected from the objects to the satellite
sensors [43]. To apply the LJE theory on the transmitted THz signals, our first task is to reveal the data convex structure of the THz signals.

Let \( m_i(t) \), \( i = 1, \ldots, n \) be the impulse response function of the \( i \)th material that the input THz signal \( x(t) \) penetrates through, and let \( \otimes \) denote the convolution operator. The transmitted THz signal measured at time \( t \) can then be modeled as

\[
y(t) = x(t) \otimes m_1(t) \otimes \cdots \otimes m_n(t).
\]

By taking the Fourier transform on both sides, we then have

\[
Y(f) = X(f) M_1(f) \cdots M_n(f),
\]

where \((Y, X, M_i)\) are the Fourier transforms of \((y, x, m_i)\), respectively, and \( f \) is the frequency index. If there are \( k \) frequency samples \( f_1, \ldots, f_k \) in the THz spectral regions, then \((Y, X, M_i)\) can be considered as \( k \)-dimensional vectors (e.g., \( Y = [Y(f_1), \ldots, Y(f_k)]^T \)). Let us define the standardized data as

\[
Z(f) = \log \left( \frac{|Y(f)|}{|X(f)|} \right) = \sum_{i=1}^{n} \log(|M_i(f)|).
\]

Also, the materials \( M_1, \ldots, M_n \) may not be distinct, we assume that there are \( q \) distinct materials in the set \( \{M_1, \ldots, M_n\} \), and let \( N(i) \) be the set of indices corresponding to the \( i \)th distinct material in the set \( \{M_1, \ldots, M_n\} \). By defining \( H_i(f) = \prod_{j \in N(i)} M_j(f) \), we can further simplify the expression as

\[
Z(f) = \sum_{i=1}^{q} \log(|H_i(f)|).
\]

Here, we neglect the energy loss between the interfaces of materials since it is relatively small compared to energy loss in the lossy medium (Supplementary Note 4). According to the derived approximation of material absorption coefficients (Supplementary Note 4), we have

\[
\log(|H_i(f)|) = \frac{1}{2} \alpha_i(f)d_i, \quad i = 1, \ldots, q,
\]

thereby leading to the standardized data representation

\[
Z(f) = \sum_{i=1}^{q} \frac{1}{2} \alpha_i(f)d_i,
\]

where \( \alpha_i(f) \) is the material absorption coefficient of the \( i \)th distinct material at frequency \( f \), and \( d_i \) is the penetration depth of the \( i \)th distinct material.

Naturally, we define the hyperspectral signature of the \( i \)th material as

\[
s_i \triangleq [\alpha_i(f_1), \ldots, \alpha_i(f_k)]^T \in \mathbb{R}^k, \quad i = 1, \ldots, q.
\]

Let \( d_{i}^n \) be the penetration depth of the \( i \)th distinct material in the \( n \)th sample data. It is elegant to observe that if we have the information of the value of \( l_n \triangleq \frac{1}{2}(d_{1}^n + \cdots + d_{q}^n) \), simply normalizing the \( n \)th standardized data \( Z_n \triangleq [Z_n(f_1), \ldots, Z_n(f_k)]^T \) by \( l_n \) can reveal the desired convex structure, i.e., \( \frac{Z_n}{l_n} \in \text{conv}\{s_1, \ldots, s_q\} \). If the incident angle is zero, \( l_n \) is nothing but half of the thickness of the \( n \)th sample. Therefore, in the proposed HYPERION algorithm, the \( \ell \) samples \( Z_1, \ldots, Z_\ell \)
are allowed to have different thickness $2l_1, \ldots, 2l_\ell$. Based on different thickness of samples, the normalized/standardized data $\mathbf{x}_n = \left[ \frac{z_n(f_1)}{l_n}, \ldots, \frac{z_n(f_k)}{l_n} \right]^T$ have the desired data convex structure

$$\mathbf{x}_1, \ldots, \mathbf{x}_\ell \in \text{conv}\{s_1, \ldots, s_q\}. \quad (2)$$

The aim is to recover the THz signatures $s_1, \ldots, s_q$ from the preprocessed THz transmitted signals $\mathbf{x}_1, \ldots, \mathbf{x}_\ell$ for material identification.

Due to the convex structure in equation (2), HYPERION can use the information of LJE, defined as the maximum-volume ellipsoid inscribed in the standardized/normalized penetrating-type THz signals as shown in Fig. 7, for the source separation with mild requirement of data purity (Supplementary Note 5).

To be precise, the standardized/normalized transmitted THz data matrix $\mathbf{X}$ can be explicitly written as

$$\mathbf{X} \triangleq [\mathbf{x}_1, \ldots, \mathbf{x}_\ell] = \mathbf{PT}\Sigma \in \mathbb{R}^{k \times \ell},$$

where $\mathbf{P} \in \mathbb{R}^{k \times q}$ with $[\mathbf{P}]_{ij} \triangleq \alpha_j(f_i)$, $\mathbf{T} \in \mathbb{R}^{q \times \ell}$ with $[\mathbf{T}]_{ij} \triangleq \frac{1}{2}d_i^j$, and $\Sigma \in \mathbb{R}^{\ell \times \ell}$ is the diagonal matrix with the $i$th diagonal entry being $\frac{1}{l_i}$. Then, computing the LJE of the THz signals $\mathbf{X}$ can be proven to be a convex optimization problem [44] as below:

$$\left( \mathbf{F}^\ast, \mathbf{c}^\ast \right) = \arg \max_{\mathbf{F} \in \mathbb{S}_++^{q-1}, \mathbf{c} \in \mathbb{R}^{q-1}} \log \text{det}(\mathbf{F})$$

$$\text{s.t.} \quad \|\mathbf{Fb}_i\| \leq h_i - \mathbf{b}_i^T \mathbf{c}, \ \forall i = 1, \ldots, H,$$

where $\mathbb{S}_++^{q-1}$ is the positive semidefinite (PSD) cone, the halfspace parameters $\{(\mathbf{b}_1, h_1), \ldots, (\mathbf{b}_H, h_H)\}$ come from the $\mathcal{H}$-polytope representation of the transmitted THz data $\mathbf{X}$, i.e., $\mathcal{H}(\mathbf{X}) = \mathcal{H}(\mathbf{PT}\Sigma) \equiv \mathcal{H}(\text{conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}) = \bigcap_{i=1}^H \{\mathbf{x} \mid \mathbf{b}_i^T \mathbf{x} \leq h_i\} [45]$, and the optimal argument $(\mathbf{F}^\ast, \mathbf{c}^\ast)$ gives the desired maximum-volume ellipsoid inscribed in the THz data $\mathbf{X} \equiv \mathcal{H}(\mathbf{X})$ (i.e., LJE) as

$$\mathcal{E}(\mathbf{F}^\ast, \mathbf{c}^\ast) \triangleq \{\mathbf{F}^\ast \mathbf{\alpha} + \mathbf{c}^\ast \mid \|\mathbf{\alpha}\| \leq 1\} \subseteq \mathbb{R}^{q-1}.$$

Note that we have assumed w.l.o.g. that the dimension of $\text{conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}$ is no greater than $q-1$, as implied by equation (2). The solving details for equation (3) are discussed in the Supplementary Note 7.

LJE $\mathcal{E}(\mathbf{F}^\ast, \mathbf{c}^\ast)$ for solving the transmitted THz unmixing problem

As aforementioned, THz data is ill-conditioned. As reported in [42], the information embedded in the LJE $\mathcal{E}(\mathbf{F}^\ast, \mathbf{c}^\ast)$ is critical in preconditioning the data $\mathbf{X}$; specifically, LJE yields the precondi-
Figure 7. Illustration of HYPERION. The illustration is taken in the ternary case for comprehensive visualization. The points, \( m_1, m_2 \) and \( m_3 \) are the unknown THz sources of the complex mixture system. The LJE is first obtained by the convex hull of data points in the original space. The transformation between the obtained LJE and a Euclidean ball, \( F \), and the center of the LJE, \( c \) are then utilized to transform data points to the preconditioned space for addressing the ill-conditioned case. Since the signatures of precondition operator, \( C(m_1), C(m_2) \) and \( C(m_3) \), should form corners of a simplex with provable regular structure in the noise-free case, HYPERION is encouraged to fit a regular simplex to the data convex geometry. To extract the unmixed signatures, the corners of the constructed simplex are transformed back to the original space by the inverse function of precondition operator.
tioned data
\[ C(X) = (F^*)^\dagger (PT\Sigma - c^*1_q^T) \in \mathbb{R}^{(q-1)\times \ell}, \]
whose corresponding (preconditioned) THz signatures are easily verified to be \( \tilde{S} = (F^*)^\dagger \left( [s_1, \ldots, s_q] - c^*1_q^T \right) \); note that the precondition operator \( C(\cdot) \) is applied columnwisely as \( C(v) = (F^*)^\dagger (v - c^*) \) for any given (column) vector \( v \). Remarkably, estimating the preconditioned THz signatures \( \tilde{S} \) is much more friendly because, according to the LJE theory [42, Theorem 1], columns of \( \tilde{S} \) will form a regular simplex centered at the origin whenever the data purity \( \gamma > \frac{1}{\sqrt{q-1}} \). As \( \tilde{S} \) is obtained, the THz signatures can be simply recovered as \([s_1, \ldots, s_q] = (F^*)^\dagger \tilde{S} + c^*1_q^T\). Therefore, we can focus on estimating \( \tilde{S} \) from the preconditioned THz data \( C(X) \) next.

According to the regular simplex structure, \( \tilde{S} \) can be characterized as [42, Equation (11)]
\[ \tilde{S} = \alpha U^T S_0, \tag{4} \]
where \( S_0 \) forms any unit-volume regular simplex centered at the origin \( 0_{q-1} \), with closed-form expression available in [42, Proposition 2]; the unitary matrix \( U \) (i.e., \( U^TU = I_{q-1} \)) and the scalar \( \alpha > 0 \) (with closed-form expression available in [42, Equation (12)]) are for rotating and scaling \( S_0 \) to fit \( \tilde{S} \).

As the transmitted THz data is not just ill-conditioned but also noisy, the regular structure may not be strictly satisfied. Therefore, unlike [42], we did not force the regular structure. Instead, we just use such information to design a regularizer \( \phi(\tilde{S}) \triangleq \| \tilde{S} - \alpha U^T S_0 \|_F^2 \), by minimizing which we just encourage the regular structure (rather than forcing it).

To finish the design of the HU criterion for the transmitted THz data, we need to design the data-fitting term. To this end, we need the following lemma, whose proof is given in Supplementary Note 8.

**Lemma 1** The precondition operator \( C \) satisfies the relation of \( C(X) = \tilde{S}T\Sigma \). \( \square \)

By Lemma 1, we naturally design the data fitting term as \( \| C(X) - \tilde{S}T\Sigma \|_F^2 \), which, together with the regularizer \( \phi \), yields the following HU criterion for transmitted THz signals:
\[ \min_{\tilde{S}, T, \Sigma, U} \| C(X) - \tilde{S}T\Sigma \|_F^2 + \lambda \phi(\tilde{S}) + I_{\text{unitary}}(U) \tag{5} \]
\[ \text{s.t. } T \geq 0, \ \Sigma \geq 0, \ 1_q^T T = 1_q^T, \]
where \( \lambda := 1 \) is the regularization parameter, and \( I_{\text{unitary}}(\cdot) \) is the indicator function of the set of unitary matrices. An algorithm for solving equation (5) is provided in Supplementary Note 9. Once equation (5) is solved, the column vectors of the optimal solution \( \tilde{S}^* \) then serves as the estimates of the THz signatures in the preconditioned space as shown in Fig. 7. Based on the
problem formulation and the solving detail, the algorithm, termed HYperspectral Penetrating-type Ellipsoidal ReconstructION (HYPERION), has been completed. Remarkably, HYPERION does not use any information about the pattern of resonant peaks of the signatures and is designed under a fully unsupervised setting.

ASOPS THz-TDS System

In the asynchronized optical sampling (ASOPS) THz-TDS system (Menlo TERA ASOPS, Menlo Systems, Germany) as shown in Fig. 8, two asynchronized Er-doped fiber femtosecond lasers are fed into an InGaAs/InAlAs THz photoconductive antenna emitter and an LT-InGaAs/InAlAs THz detector. The power, bandwidth, and dynamic range of the THz-TDS system are up to 60 µW, less than 4.5 THz, and greater than 80 dB. The generated THz radiation from the THz photoconductive antenna emitter then consecutively travels through two convex THz lenses with a focal length of 10 cm. The first convex lens is used for THz beam collimation, and the second convex lens is to focus the THz radiations on the tablets. The diameter of the focused beam is approximately 1.5 mm of full width at half maximum (FWHM). After penetrating through tablets, the transmitted THz waves contained material absorption information in THz range, then traveling through two identical convex THz lenses to the THz photoconductive antenna detector. THz photoconductive antenna detector is used to retrieve the time-resolved THz electric field and convert the electric field signal to the photocurrent. The connected transimpedance amplifier (TIA) then amplifies the induced photocurrent to voltage signals. The bandwidth and the gain of the TIA are 1.8 MHz and $10^6$ (V/A), respectively. Each measurement contains a 100 ps time-domain trace with a 5 fs temporal resolution. To further increase the signal-to-noise ratio (SNR) of the system, we average 1,000 time-domain traces, which can effectively decrease the time-domain additive white Gaussian noise to 0.015% according to the law of large numbers (Supplementary Note 10). With the setting above, the system offers a dynamic range of over 65 dB from 0.1 THz to 3 THz.

Tablet Preparation

In this study, all tablets were composed of D-glucose, D-lactose monohydrate, L-histidine, L-tyrosine, and D-sucrose powders. D-glucose, D-lactose monohydrate, and D-sucrose powders were from Merck & Co., Inc. (Kenilworth, NJ, USA). L-histidine, L-tyrosine powders were from Acros Organics (Geel, Belgium). Each chemical was ground and mixed uniformly by the ball mill, ensuring that the Mie scattering is greatly reduced in the measured spectral characteristics. Followed by the pulverization, the chemical powders were poured into a tablet die made by a high-speed steel (HSS) for subsequent high-pressure compression. Powders were compressed by a hydraulic press (Specac, Orpington, U.K.) to form 3 mm-thick tablets (total mass: 0.3 g; thickness: 3.05 ± 0.02 mm; see Supplementary Table 6 for the details) under the pressure of 1,000 kg/cm² for 15 seconds.
Figure 8. Experimental setup. The THz photoconductive antenna (PCA) emitter and detector are excited by two asynchronized femtosecond lasers with repetition rate ($f_{\text{rep}}$) of 100 MHz + 10 Hz and 100 MHz. The repetition rate difference ($\Delta f$) of two lasers is 10 Hz. The generated THz beam is focused to a spot with a diameter of 1.5 mm by THz convex lenses (focal length: 10 cm), interacting with chemical tablets and being detected by the THz photoconductive antenna detector. The detected THz electric field is converted to photocurrent and is further amplified by a transimpedance amplifier (TIA). The computer converts the amplified analog signal to the digital data and presents the measured THz signals in both time and frequency domains.
Tablets were then mounted on a 0.6 mm thick polylactic acid (PLA) plate. The copper foil is covered around the tablets to block THz signals encoding the PLA plate material information. Additionally, the copper foil prevents the low-frequency diffraction effect caused by the edge of the PLA plate.

6 Supplementary Information

Supplementary Note 1: Measurement Protocol

Each spectrum is obtained by fast Fourier transform (FFT) of a 100 picoseconds (ps) THz time-domain trace based on an asynchronous optical sampling (ASOPS) THz-TDS system (see Supplementary Note 11). To increase the system dynamic range, 1,000 spectra of each mixture are measured and averaged, providing more than 60 dB dynamic range at 0.3 THz and 25 dB dynamic range at 1.75 THz. All the measurements are conducted at room temperature (23°C), one atmosphere (atm) pressure, and 60% humidity. Under this condition, the dataset contains water vapor absorption lines at 0.56 THz, 0.75 THz, 0.99 THz, 1.10 THz, 1.16 THz, 1.21 THz, 1.23 THz, 1.41 THz, 1.60 THz, 1.66 THz, and 1.72 THz within 0.2 - 1.75 THz frequency range (Supplementary Fig. 9), which contribute nearly 2.3% of the nonlinear region in the measured frequency bands.

Supplementary Note 2: Linear Mixing Model Validation

Whether the dataset follows the linear mixing model is essential since HYPERION utilizes the convex geometry of the model to unmix the material signatures. To validate the linear mixing model upon the dataset, principle component analysis (PCA) is one of the most suitable tools since data information can still be reserved after projecting to a lower-dimensional space [46]. To be precise, we project the non-pure mixture signatures to the optimal 2 dimensional (2D) space, which is constructed by the two eigenvectors corresponding to the two largest eigenvalues. Additionally, only non-pure mixture points are selected for the calculation of the 2D space since the assumption of pure substances is not held. After obtaining the optimal 2D plane, the data points of pure substances are further projected to the optimal 2D plane. In the evaluation of the linear mixing model, the quinary dataset with pure substances is selected due to the material variety. As shown in Supplementary Fig. 10, the high-dimensional data points are projected onto the two-dimensional plane. Except for the extreme cases, most of the data points in every subplot form the triangle in the projected 2D space, where three points lie at the vertex and three points fall at the center of the side. The misaligned data points are caused by the inaccurate estimation of the two-dimensional plane, which is caused by the measurement noise and nonlinear light-matter interaction. Based on the projected data points, the measured data points are highly aligned with the linear mixing model. Additionally, the triangles involving D-glucose and D-lactose are narrower
Figure 9. The THz frequency-domain spectrum of air. Water vapor absorbs THz wave at 0.56 THz, 0.75 THz, 0.99 THz, 1.10 THz, 1.16 THz, 1.21 THz, 1.23 THz, 1.41 THz, 1.60 THz, 1.66 THz, and 1.72 THz within 0.2 - 1.75 THz frequency range. The absorption lines are indicated by gray dotted lines.

since the signatures of the two chemicals are the ill-conditioned case. It is worthwhile to mention that the alignment accuracy between the model and measured data may be further enhanced by modeling the nonlinear matter-matter interaction into the linear mixing model.

Supplementary Note 3: Spectral Angle Mapper (SAM) and Root Mean Square Error (RMSE)

Spectral angle mapper (SAM) is typically used to evaluate the shape similarity between two spectra. The definition is as a following equation:

$$\theta(s_1, s_2) = \cos^{-1}\left(\frac{s_1 \cdot s_2}{\|s_1\|_2 \cdot \|s_2\|_2}\right),$$

where $s_1$ and $s_2$ are spectra.

Root mean square error (RMSE) does not only focus on the shape similarity but the band-wise differences. RMSE is defined as a following equation:

$$d(s_1, s_2) = \sqrt{\frac{\|s_1 - s_2\|_2^2}{n}},$$

where $n$ is the dimension of the spectrum.
Figure 10. Principle component analysis (PCA) projection of the quinary dataset. Data points from the quinary dataset with pure substances are projected on a two-dimensional projection plane by PCA. Five pure substances are D-glucose, D-lactose monohydrate, D-sucrose, L-tyrosine, and L-histidine, which are abbreviated as G, L, S, T, and H.
Supplementary Note 4: Data Preprocessing

The THz-TDS system provides a time-domain waveform that profiles the interactions between THz radiation and measured tablets. The reference waveform $S_{ref}(t)$ and the tablet waveform $S_{tablet}(t)$, where $t$ is the time index, refer to the measured waveform without and with a tablet, respectively. After obtaining reference waveform $S_{ref}(t)$ and tablet waveform $S_{tablet}(t)$, the fast Fourier transform (FFT) was applied to convert time-domain waveforms to the frequency-domain reference spectrum $S_{ref}(\omega)$ and tablet spectrum $S_{tablet}(\omega)$, where $\omega$ is the angular frequency. To further extract the material absorption spectrum ($\alpha(\omega)$) of the tablet, the thickness of the tablet is an important parameter for the derivation. The thickness of the tablet is measured by the electric ruler, which provides a 0.1 mm resolution. Compared to the thickness of the tablet (i.e., 3 mm), this resolution of 0.1 mm introduces approximately 3% inaccuracy in the measurement. The material absorption spectra of the tablets are derived from the Fresnel equations and the THz wave propagation model based on homogeneous and planar materials. Furthermore, according to this derivation, the thickness measurement error of 3% will only introduce approximately 3% error in material absorption spectra. Since the particle size of the tablet powder is vastly smaller than the wavelength of the THz waves (Supplementary Note 16) Additionally, the thickness of the tablet is significantly larger than the effective wavelength of the THz signal in the 0.2 - 1.75 THz frequency range such that the Fabry-Pérot effect [17] does not introduce drastic damping to the measured spectrum.

Under this condition, we do not include the scattering influence and the Fabry-Pérot effect in this model. The Fresnel equations describe the transmission and the reflection of a THz wave traveling through an interface, which are based on the complex refractive index of the material, $\tilde{n}(\omega) = n(\omega) - j\kappa(\omega)$, where $n(\omega)$ is the real refractive index and $\kappa(\omega)$ represents the extinction coefficient which is proportional to the material absorption coefficient, $\kappa(\omega) = [\alpha(\omega) \cdot \lambda] / 4\pi$, where $\lambda$ is the wavelength. The material absorption coefficients describe the power loss when the THz waves travel through the material. Considering a typical incident THz wave propagates a material with thickness $d$ (Supplementary Fig. 11), the Fresnel equation [17] at an interface can be represented as

$$t_{ab}(\omega) = \frac{2\tilde{n}_a(\omega)}{\tilde{n}_a(\omega) + \tilde{n}_b(\omega)};$$

$$r_{ab}(\omega) = \frac{\tilde{n}_a(\omega) - \tilde{n}_b(\omega)}{\tilde{n}_a(\omega) + \tilde{n}_b(\omega)},$$

where $t_{ab}(\omega)$ is the transmission coefficient of a THz wave from the region $a$ (which is air in our experiment) to region $b$ (which is the tablet in our experiment) and $r_{ab}(\omega)$ is the reflection coefficient of a THz wave at the $a$-$b$ interface. Additionally, $\tilde{n}_a(\omega)$ and $\tilde{n}_b(\omega)$ represent the complex refractive
index of region $a$ and $b$, respectively. The THz wave propagation constant, $p_b(\omega, d)$, is governed by
\[
p_b(\omega, d) = \exp \left[ -j \tilde{n}_b(\omega) \omega d \right],
\]
where $c$ is the speed of light. Consequently, the equation for the received THz tablet signal in the frequency domain can be represented as
\[
S_{\text{tablet}}(\omega) = S_{\text{ref}}(\omega) \cdot t_{ab} \cdot t_{ba} \cdot p_b(\omega, d) \quad (\text{The Fresnel loss is included in $t_{ab}$ and $t_{ba}$}.)
\]
\[
= S_{\text{ref}}(\omega) \cdot t_{ab} \cdot t_{ba} \cdot \exp \left[ -j \tilde{n}_b(\omega) \omega d \right]
\]
\[
= S_{\text{ref}}(\omega) \cdot t_{ab} \cdot t_{ba} \cdot \exp \left\{ -j [n_b(\omega) - j \kappa_b(\omega)] \omega d \right\}
\]
\[
= S_{\text{ref}}(\omega) \cdot t_{ab} \cdot t_{ba} \cdot \exp \left[ \kappa_b(\omega) \omega d \right] \cdot \exp \left[ -j n_b(\omega) \omega d \right].
\]
The reference signal normalizes the tablet signal, and its amplitude can be obtained from equation (8)
\[
\frac{|S_{\text{tablet}}(\omega)|}{S_{\text{ref}}(\omega)} = |t_{ab} \cdot t_{ba} \cdot \exp \left[ \kappa_b(\omega) \omega d \right]| \triangleq |H(\omega)|. \quad (8)
\]
The extinction coefficient can be derived by taking the natural logarithm operation of equation (8)
\[
\ln(|H(\omega)|) = \frac{\kappa_b(\omega) \omega d}{c} + \ln(|t_{ab} \cdot t_{ba}|),
\]
\[
\kappa_b(\omega) = \frac{c [\ln(|H(\omega)|) - \ln(|t_{ab} \cdot t_{ba}|)]}{\omega d}.
\]
Therefore, the material absorption coefficient can be represented as
\[
\alpha'(\omega) = \frac{4 \pi \kappa(\omega)}{\lambda} = \frac{2 \cdot \ln(|H(\omega)|) - \ln(|t_{ab} \cdot t_{ba}|)}{d}. \quad (9)
\]
Based on the derivation, to acquire the material absorption spectrum, we have first to calculate the transmission coefficients, $t_{ab}$ and $t_{ba}$. However, the interface power loss is relatively small since the thickness of the sample is greatly larger than the THz wavelength, and the tablet surface is smooth and flat compared to the THz wavelength. This relatively small interface power loss results in the high transmission coefficients of $t_{ab}$ and $t_{ba}$. Thus, we can neglected the second term of the right-hand side in equation (9) to form the equation (10) since the multiplication of $t_{ab}$ and $t_{ba}$ approaches to 1.
\[
\alpha'(\omega) = \frac{2 \cdot \ln(|H(\omega)|)}{d} = \frac{2 \cdot \ln \left( \frac{|S_{\text{tablet}}(\omega)|}{S_{\text{ref}}(\omega)} \right)}{d}. \quad (10)
\]
To further match the notation in blind source separation fields, the equation (10) can be expressed
Figure 11. Illustration of THz wave traveling through a material. A THz wave transmission and reflection pathways through a planar, homogeneous material with a thickness $d$. The THz signals shown in this figure are represented in time-domain.

In ordinary frequency, $f$, by the equation of $\omega = 2\pi f$:

$$\alpha^\prime(\omega) = \alpha^\prime(2\pi f) \triangleq \alpha(f).$$

(11)

Supplementary Note 5: Mild algorithmic requirement

The aforementioned problem (i.e., recovering $s_1, \ldots, s_q$ from their mixtures $x_1, \ldots, x_\ell$) has been studied in the machine learning literature; for example, in the successive projection algorithm (SPA) [47], the problem is solved under one additional assumption called separability assumption, i.e.,

$$s_1, \ldots, s_q \in \{x_1, \ldots, x_\ell\},$$

meaning that, for each material, there is a pure observation solely composed of such material. Under the separability assumption, one can show that $\text{conv}\{x_1, \ldots, x_\ell\} = \text{conv}\{s_1, \ldots, s_q\}$, and hence identifying the signatures is nothing but seeking the vertices of the observable geometry $\text{conv}\{x_1, \ldots, x_\ell\}$, which is polynomial-time solvable. Remarkably, by following the definition of data purity $\gamma \in (0, 1]$, defined in [27], the separability assumption holds if, and only if, $\gamma = 1$ (the
highest data purity). So, the separability assumption is quite restricted, and such assumption is often violated in practical hyperspectral applications [27].

In fact, the separability assumption may not hold in our THz application, as a tablet typically involves highly mixed chemicals, implying a low data purity. In recent mathematics literature, the LJE theory has been proven to work well even with very low data purity; specifically, the signatures can be perfectly recovered under a much milder assumption $\gamma > \frac{1}{\sqrt{q-1}}$ [22]. Also, as aforementioned, the LJE criterion is theoretically and experimentally proven to be robust against the ill-conditioning of the signatures, hence quite suitable for our THz applications. Therefore, in this paper, the convex LJE criterion is applied to the transmitted (penetrating-type) THz signals for the first time, to be detailed later.

We remark that there is other non-convex criterion that can also perform HU under the mild requirement of $\gamma > \frac{1}{\sqrt{q-1}}$ [27, 48], but such criterion cannot achieve the theoretical bound $\frac{1}{\sqrt{q-1}}$ in practice due to its non-convexity [22]. By contrast, the LJE criterion has the same theoretical bound, and in the meanwhile, it is experimentally shown to be able to achieve the bound due to its convexity nature [22].

**Supplementary Note 6: Implementation of THz HYPERION**

To restore the chemical compositions, the equation (12) should be solved for every tablet.

$$\min_{r_i \in \mathbb{R}^3} \|x_i - Ar_i\|_1$$

s.t. $r_i \succeq 0$, $1^T r_i = 1$. 

Since the equation (12) is convex, it can be solved by the solver CVX [37] as the following pseudo-code:

```cvx
cvx_begin
    variable r(3) nonnegative;
    minimize( norm(x - A * r, 1) );
    subject to
    sum(r) == 1;
    cvx_end
```
Supplementary Note 7: How the LJE problem (13) can be solved in practice?

One of the core in HYPERION is the LJE problem:

\[
(F^*, c^*) = \arg \max_{F \in S^{p-1}_{++}, c \in \mathbb{R}^{q-1}} \log \det(F) \\
\text{s.t.} \quad \|F b_i\| \leq h_i - b_i^T c, \quad \forall i = 1, \ldots, H,
\]

where \(S^{p-1}_{++}\) is the positive semidefinite (PSD) cone, the halfspace parameters \(\{(b_1, h_1), \ldots, (b_H, h_H)\}\) come from the \(\mathcal{H}\)-polytope representation of the data \(X\) and the optimal argument \((F^*, c^*)\) gives the desired maximum-volume ellipsoid inscribed in the penetrating-type THz data.

We remark that there is an algorithm [42] that can be adapted to solve equation (13) exactly, but it is mainly designed to handle the case with million-scale constraints \(H\) induced by NASA’s benchmark hyperspectral data [49]. In typical THz applications, such as measuring the chemicals in a given tablet, we just have a few samples, and hence we do not need to use such a sophisticated method. Instead, the general-purpose convex software CVX [37] is effective and fast when the problem size is not too large, and hence is suitable to be adopted to solve equation (13).

To adopt CVX, we need to implement the constraints and the objective function of equation (13). The constraints of equation (13) are the well-known Lorentz cone constraints \(\|F b_i\| \leq h_i - b_i^T c\), theoretically equivalent to require the LJE to lie within the penetrating-type data convex hull \(\mathcal{H}(PT\Sigma) \equiv \mathcal{H}(\text{conv}\{x_1, \ldots, x_I\})\). As for the objective function, for PSD matrix \(F\), we observe that

\[
\log \det(F) \propto \det(F) \propto (\det(F))^{\frac{1}{q-1}}.
\]

So, the objective function can be implemented by the command \(\text{det}_\text{rootn}(\cdot)\) supported by the CVX package. Solving the LJE problem by the general-purpose convex software CVX normally requires some prior knowledge to transform the problem into the supporting functions. To this end, we provide the pseudo-code for the better understanding. The pseudo-code using CVX is as
following:

```python
cvx_begin
    variable F(q - 1, q - 1) symmetric;
    variable c(q - 1);
    maximize( det_rootn(F) );
    subject to
        for i = 1 : H
            norm( F * b(i, :) - 2 ) + b(i, :)’ * c <= h(i);
        end
    cvx_end
```

It shows that CVX can handle a non-standard form of convex programming.

**Supplementary Note 8: Proof of Lemma 1.**

Following the definition of precondition operator $C$, we have

$$
C(X) \triangleq (F^*)^\dagger (PT\Sigma - c^*1_\ell^T)
$$

$$
= (F^*)^\dagger PT\Sigma - (F^*)^\dagger c^*1_\ell^T
$$

$$
= (F^*)^\dagger PT\Sigma - (F^*)^\dagger c^*1_q^T T\Sigma
$$

$$
= (F^*)^\dagger [s_1, \ldots, s_q] T\Sigma - (F^*)^\dagger c^*1_q^T T\Sigma
$$

$$
= \left( (F^*)^\dagger [(s_1, \ldots, s_q) - c^*1_q^T] \right) T\Sigma
$$

$$
= \tilde{S}T\Sigma,
$$

where we have used the observation that each column of the matrix $T\Sigma$ actually provides a set of convex combination coefficients [44], implying the equation "$1_q^T T\Sigma = 1_\ell^T$".

**Supplementary Note 9: Algorithm details for solving the HYPERION**

As there are four variables in equation (14),

$$
\min_{\tilde{S}, T, \Sigma, U} \left\| C(X) - \tilde{S}T\Sigma \right\|_F^2 + \lambda\phi(\tilde{S}) + I_{\text{unitary}}(U)
$$

s.t. $T \geq 0$, $\Sigma \geq 0$, $1_q^T T\Sigma = 1_\ell^T$,
solving it directly using the alternating optimization may not be effective. Considering that what we are really interested in is the THz signatures \( \tilde{S} \), we do not need to separately optimize all the variables, allowing us to introduce the trick of changes of variables to reduce the number of block variables. Specifically, by letting \( \tilde{T} = T \Sigma \), equation (14) can be reformulated as

\[
\min_{\tilde{S}, \tilde{T}, U} \left\| C(X) - \tilde{S} \tilde{T} \right\|_F^2 + \lambda \phi(\tilde{S}) \\
\text{s.t.} \quad \tilde{T} \geq 0, \quad 1_q^T \tilde{T} = 1_T^T, \quad U^T U = I_{q-1}.
\]

After reducing the number of block variables, it is natural to solve equation (15) using alternating optimization that alternatively updates \( \tilde{S}, \tilde{T}, U \), as detailed next.

The subproblem for solving \( \tilde{S} \) (with \( \tilde{T}, U \) fixed) can be explicitly written as

\[
\tilde{S}^* = \arg \min_{\tilde{S}} \left\| C(X) - \tilde{S} \tilde{T} \right\|_F^2 + \lambda \left\| \tilde{S} - \alpha U^T S_0 \right\|_F^2.
\]

By defining \( \tilde{m} = \text{vec}(\tilde{S}) \), \( v = [\text{vec}(C(X))^T, \sqrt{\lambda}(\text{vec}(\alpha U^T S_0))^T]^T \) and \( \tilde{P} = [\tilde{T} \otimes I_{q-1}, \sqrt{\lambda}I_{q(q-1)}]^T \), equation (16) can be simplified as \( \text{vec}(\tilde{S}^*) = \arg \min_{\tilde{m}} \| \tilde{P} \tilde{m} - v \|_2^2 \), which closed-form solution can be derived as

\[
\tilde{m}^* = (\tilde{P}^T \tilde{P})^{-1}(\tilde{P}^T v) = ([\tilde{T} \tilde{T}^T + \lambda I_q]^{-1} \otimes I_{q-1}) \text{vec}(C(X) \tilde{T}^T + \lambda \alpha U^T S_0) = \text{vec} \left( I_{q-1} [C(X) \tilde{T}^T + \lambda \alpha U^T S_0] [\tilde{T} \tilde{T}^T + \lambda I_q]^{-T} \right),
\]

where \( \otimes \) and \( \text{vec}(\cdot) \) denote Kronecker product and vectorization operator, respectively. Therefore, we have \( \tilde{S}^* = [C(X) \tilde{T}^T + \lambda \alpha U^T S_0] [\tilde{T} \tilde{T}^T + \lambda I_q]^{-T} \).

Next, note that the subproblem for solving \( \tilde{T} \) (with \( \tilde{S}, U \) fixed) is

\[
\tilde{T}^* = \arg \min_{\tilde{T} \geq 0, \ 1_q^T \tilde{T} = 1_T^T} \left\| C(X) - \tilde{S} \tilde{T} \right\|_F^2,
\]

which can be easily solved as the well-known fully-constrained least-squares problem [50, 51]. Finally, the subproblem for solving \( U \) (with \( \tilde{S}, \tilde{T} \) fixed) is a non-convex optimization problem

\[
U^* = \arg \min_{U^T U = I_{q-1}} \left\| \tilde{S} - \alpha U^T S_0 \right\|_F^2,
\]

but can be elegantly solved using the following inequalities for square matrices \( A, B \), i.e., trace\((A) \leq \|A\|_* = \|A\|_{S^1} \) and \( \|AB\|_{S^1} \leq \|A\|_{S^p} \|B\|_{S^q} \), where \( \frac{1}{p} + \frac{1}{q} = 1 (1 \leq p, q \leq \infty) \), \( \| \cdot \|_* \) denotes nuclear norm, and \( \| \cdot \|_{S^p} \) denotes Schatten p-norm. Since \( U \) is unitary, the objective function of equation
Figure 12. THz-TDS system noise distribution. The noise distribution in our THz-TDS system follows the Gaussian distribution. The noise standard deviation is 0.48% of the mean of the peak amplitude. Based on this information, we can infer the noise level in different average number configuration.

(17) can be simplified as \( \text{trace}(U \tilde{S} S_0^T) \), whose upper bound is derivable using the inequalities as \( \text{trace}(U \tilde{S} S_0^T) \leq \| U \tilde{S} S_0^T \|_{S^1} \leq \| U \Sigma \|_{S^1} \| \tilde{S} S_0^T \|_{S^1} = \| \tilde{S} S_0^T \|_{S^1} \), and is achievable by \( U^* = D_1 D_2^T \), where \( D_1, D_2 \) are obtained from the singular value decomposition of \( \tilde{S} S_0^T = D_2 \Sigma D_1^T \).

The algorithm, termed HYperspectral Penetrating-type Ellipsoidal ReconstructION (HYPERION), has been completed. Remarkably, HYPERION does not use any information about the pattern of resonant peaks and is designed under a fully blind setting.

Supplementary Note 10: Noise Modeling of the System

The noise in our THz-TDS system corresponds to additive white Gaussian noise (AWGN), whose standard deviation in the single time-domain trace is approximately 0.48% of the peak-to-peak amplitude as shown in Supplementary Fig. 12. According to the law of large numbers, the standard deviation of AWGN is linearly proportional to \( \sqrt{n} \), where \( n \) is the number of time traces. In this regard, we applied the AWGN to our dataset to simulate the different levels of noise conditions. We also present this noise amplitude in terms of SNR by adopting the equation (18) from the speech signal processing field.

\[
\text{SNR} = 10 \log_{10} \left( \frac{P(S) - P(N)}{P(N)} \right),
\]

where \( P(S) \) and \( P(N) \) are the power of signal and noise, respectively.
Supplementary Note 11: Asynchronized Optical Sampling (ASOPS) THz-TDS System

The conventional THz time-domain spectroscopy (THz-TDS) system is essential for many THz applications, such as remote sensing, material characterization, imaging, and defect inspection. As THz-TDS systems are capable of sampling ultrafast THz signals in the time domain, THz-TDS systems can provide more information, such as time delay, spectral phase, and spectral amplitude [52]. A conventional THz-TDS system utilizes a mechanical delay-line stage to introduce a time delay between two split femtosecond laser beams: the so-called pump beam and probe beam [53]. The pump beam is coupled to the THz photoconductive antenna emitter to generate the THz radiations. The probe beam is fed to the THz photoconductive antenna detector to gate received THz signals [52, 54]. With the accumulated time delay between two beams, a THz-TDS system can profile the THz time-domain signal, where the effective sampling frequency is typically tens of THz, which is the reciprocal of the time delay [53]. However, the sampling time and the quality of the time-domain signal are severely limited by the mechanical delay line stage condition, such as maximum moving velocity, position resolution, and operational stability. To address the limitation, the asynchronized optical sampling (ASOPS) THz-TDS system is developed to increase the sampling frequency. Additionally, the ASOPS THz-TDS system is more stable due to the elimination of the mechanical delay line stage interference [55]. By replacing the mechanical delay line stage with the two asynchronized femtosecond lasers, the time delay between the probe beam and the pump beam can be accumulated in a short time. More specifically, the sampling time of a single time trace, \( t_{\text{sample}} \), can be decreased down to within millisecond timescale. Additionally, \( t_{\text{sample}} \) is related to the repetition rate of the lasers as

\[
\frac{1}{f_{\text{rep}}} - \frac{1}{f_{\text{rep}} + \Delta f} \approx \frac{1}{\Delta f},
\]

where \( f_{\text{rep}} \) and \( \Delta f \) are the repetition rate of the pump beam laser and repetition rate difference between the two femtosecond lasers. In our ASOPS THz-TDS system, the repetition rate of the pump beam laser is 100 MHz. To achieve the high time-domain sampling resolution and the fast sampling time simultaneously, the repetition rate difference is set as 10 Hz, delivering 5 fs sampling resolution and 100 ms sampling time. However, under this configuration, the size of the full single-time trace is considerably large, requiring high data transfer rate and large storage space. To prevent those issues, we have only taken 100 ps segment of the single time trace and set the sampling rate of the DAQ card as 20 MHz. Accordingly, our ASOPS THz-TDS system provides 0.01 THz resolution in the spectrum and 5 fs time-domain sampling resolution.
Supplementary Note 12: Frequency Band Selection

The detectable bands of the spectra are between 0.2 THz and 1.75 THz due to physical and hardware limitations [56, 57]; the THz spectral signal below 0.2 THz suffers from the physical and instrumental limitations, such as photoconductive material carrier lifetime, bandwidth, and sensing area of the THz photoconductive antennas; the THz spectral signal above 1.75 THz decays to the system noise level due to the absorption of the tablets. Consequently, we have only extracted THz bands between 0.2 THz and 1.75 THz as our valid spectra for the further unmixing experiments.

Supplementary Note 13: Nonnegative Matrix Factorization (NMF)

The nonnegative matrix factorization (NMF) is one of the commonly used methods to recover material signatures in the blind sense. To address the non-convexity of NMF, the result of NMF is taken from one of the 10 trials under different initialization based on the best root-mean-squared error (RMSE). The RMSE, \( D \), is defined as:

\[
D = \frac{\|X - WH\|_F}{\sqrt{nm}},
\]

where \( X \in \mathbb{R}^{n \times m} \), \( W \in \mathbb{R}^{n \times k} \), and \( H \in \mathbb{R}^{k \times m} \) are the target, signatures matrix, and abundance matrix, respectively. In the experiment, \( k \) equals to 3 and 5 in the ternary and quinary cases, respectively. Additionally, we adopt the naive NMF in equation (19) since validity of the commonly-used regularizers, such as \( L^2 \) norm or \( L^1 \) norm, have not yet been verified with reasonable physical meaning. The alternating least square method is used to solve this NMF optimization problem:

\[
\begin{align*}
\min_{W, H} \quad & \|X - WH\|_F \\
\text{s.t.} \quad & W \geq 0, \ H \geq 0.
\end{align*}
\]

Supplementary Note 14: Linear Mixing Model

Since the sum-to-one criterion derived in the “Methods” section is based on the assumption where interface power loss is relatively small compared to material absorption. To validate the assumption in practical cases, we have measured the spectra of three mixture tablets with different mixing approaches as shown in Supplementary Fig. 14. The three tablets are composed of the same amount of substances. As shown in Supplementary Fig. 15, the deviation among three measured spectra by different mixing approaches is only less than 2 dB. Based on the experiment, it concludes that the spectrum is not severely affected by the mixing approaches.

In addition to the mixing approaches, the dynamic range of the THz-TDS system affects the accuracy of linear mixing model on the material absorption spectra since the largest detectable
absorption coefficients $\alpha_{\text{max}}(f)$ are determined by the THz-TDS system dynamic range [58]. More specifically, the performance of HYPERION will be decreased when the material absorption coefficients are larger than the $\alpha_{\text{max}}(f)$. To this sense, it is important to compare the $\alpha_{\text{max}}(f)$ of our ASOPS THz-TDS system to the pure substance absorption spectra. As shown in Supplementary Fig. 13, in the range between 1.3 THz and 1.5 THz, the material absorption coefficients of D-lactose monohydrate and D-glucose exceed the $\alpha_{\text{max}}(f)$. Except this region, the material absorption spectra are less than the $\alpha_{\text{max}}(f)$. Although the measured absorption coefficients are not accurate in this region, HYPERION can still deliver the acceptable unmixing performance due to use of the material absorption information from all measured frequency bands.

Supplementary Note 15: Composition Estimation Comparison of HYPERION, NMF, HMFA, nICA, and SPA

We compared the material composition estimation by the material absorption spectra unmixed from different methods, including HYPERION, NMF, HMFA, nICA, and SPA. The configuration of the test set is same as in the “Application” section. As shown in Supplementary Table. 1, HYPERION delivers superior performance of estimating material composition since HYPERION unmixs the accurate material absorption spectra. HMFA and nICA have the much inferior performance on material estimation since the unmixed material absorption spectra are much suppressed in magnitude. The unmixed material spectra by the quinary dataset without pure substances are
used to restore the chemical compositions of the 15 tablets in the test set by the convex optimization problem in equation (20).

$$\min_{r_i \in \mathbb{R}^3} \| x_i - A r_i \|_1 $$

(20)

$$\text{s.t. } r_i \geq 0, \quad 1^T r_i = 1, $$

(21)

**Supplementary Note 16: Grain Size of Pure Substances and the Scattering Effect**

Here, we will investigate the effect of the scattering on the unmixed spectra results. Since the effect of scattering is proportional to the grain size of the tablet [59], an optical microscopy with 1.25 numerical aperture (NA) is utilized to measure the grain sizes of different tablets. In the measurement, all chemicals follow the same grounding approach (see “Method” for the details) and are placed on the standard microscopy slide. For each chemical, the grain size is determined by taking the average of the grain size of 7 randomly selected grains. As shown in Supplementary Table. 2, the grain sizes of all chemicals fall in 1 µm range. Thus, the scattering effect in different chemicals falls in the same scale. To this sense, we can model the measured material absorption as the addition of true material absorption and the scattering effect as in equation (22).

$$u_{msr}(f) = u_{abs}(f) + u_{sct}(f),$$

(22)

where $u_{msr}$, $u_{abs}$, and $u_{sct}$ are the measured material absorption, material absorption and the scattering effect, respectively. Since the grain sizes of different chemicals fall in the same range, $u_{sct}(f)$ is considerably identical for different chemicals. In HYPERION, this constant shift $u_{sct}(f)$ will be calibrated in the affine fitting step and added to the unmixed spectrum in the transformation between the preconditioned space and original space. Thus, the scattering effect will not affect the unmixing performance of HYPERION when the grain sizes of different chemical fall in the same scale.
Table 1: The composition estimation comparison between HYPERION, HMFA, NMF, SPA and nICA.

| Index (ground truth) | HYPERION | HMFA     | NMF      | SPA      | nICA     |
|---------------------|----------|----------|----------|----------|----------|
| 1 (0.00/0.00/1.00)  | 0.09/0.00/0.91 | 1.00/0.00/0.00 | 0.03/0.00/0.97 | 0.00/0.00/1.00 | 0.00/0.00/1.00 |
| 2 (0.25/0.00/0.75)  | 0.35/0.00/0.65 | 1.00/0.00/0.00 | 0.39/0.00/0.61 | 0.47/0.00/0.53 | 0.00/0.00/1.00 |
| 3 (0.00/0.25/0.75)  | 0.11/0.13/0.75 | 1.00/0.00/0.00 | 0.19/0.19/0.62 | 0.11/0.00/0.89 | 0.00/0.00/1.00 |
| 4 (0.50/0.00/0.50)  | 0.52/0.00/0.48 | 1.00/0.00/0.00 | 0.60/0.00/0.40 | 1.00/0.00/0.00 | 0.00/0.00/1.00 |
| 5 (0.25/0.25/0.50)  | 0.43/0.14/0.43 | 1.00/0.00/0.00 | 0.61/0.12/0.27 | 1.00/0.00/0.00 | 0.00/0.00/1.00 |
| 6 (0.00/0.50/0.50)  | 0.11/0.45/0.44 | 1.00/0.00/0.00 | 0.18/0.32/0.50 | 0.09/0.39/0.52 | 0.00/0.00/1.00 |
| 7 (0.75/0.00/0.25)  | 0.61/0.03/0.36 | 1.00/0.00/0.00 | 0.70/0.01/0.30 | 1.00/0.00/0.00 | 0.00/0.00/1.00 |
| 8 (0.50/0.25/0.25)  | 0.55/0.14/0.30 | 1.00/0.00/0.00 | 0.67/0.19/0.14 | 1.00/0.00/0.00 | 0.00/0.00/1.00 |
| 9 (0.25/0.50/0.25)  | 0.29/0.42/0.29 | 1.00/0.00/0.00 | 0.39/0.33/0.28 | 0.67/0.00/0.33 | 0.00/0.00/1.00 |
| 10 (0.00/0.75/0.25) | 0.05/0.52/0.43 | 1.00/0.00/0.00 | 0.06/0.39/0.54 | 0.00/0.65/0.35 | 0.00/0.00/1.00 |
| 11 (1.00/0.00/0.00) | 0.75/0.03/0.22 | 1.00/0.00/0.00 | 0.94/0.01/0.05 | 1.00/0.00/0.00 | 0.00/0.00/1.00 |
| 12 (0.75/0.25/0.00) | 0.63/0.17/0.20 | 1.00/0.00/0.00 | 0.85/0.15/0.00 | 1.00/0.00/0.00 | 0.00/0.00/1.00 |
| 13 (0.50/0.50/0.00) | 0.46/0.49/0.06 | 1.00/0.00/0.00 | 0.61/0.36/0.03 | 0.83/0.17/0.00 | 0.00/0.00/1.00 |
| 14 (0.25/0.75/0.00) | 0.26/0.61/0.12 | 1.00/0.00/0.00 | 0.32/0.47/0.20 | 0.26/0.74/0.00 | 0.00/0.00/1.00 |
| 15 (0.00/1.00/0.00) | 0.04/0.99/0.06 | 0.59/0.41/0.00 | 0.07/0.69/0.24 | 0.00/1.00/0.00 | 0.00/0.00/1.00 |

RMSE 0.11 0.54 0.15 0.25 0.57
Table 2: The grain size of five pure substances. The measurement is performed by a optical microscopy with numerical aperture (NA) of 1.25.

| Pure substances        | Sampled Grain Size (μm) |
|------------------------|-------------------------|
|                        | #1 | #2 | #3 | #4 | #5 | #6 | #7 | Average |
| D-glucose              | 1.02 | 1.14 | 1.02 | 1.02 | 1.14 | 1.02 | 1.02 | 1.05 |
| D-lactose monohydrate  | 1.36 | 1.14 | 1.03 | 1.14 | 0.91 | 0.92 | 0.92 | 1.06 |
| L-tyrosine             | 1.02 | 1.14 | 1.14 | 1.03 | 0.91 | 0.91 | 0.80 | 0.99 |
| L-histidine            | 0.91 | 0.91 | 1.02 | 1.02 | 1.02 | 1.02 | 1.02 | 0.99 |
| D-sucrose              | 1.04 | 1.02 | 1.26 | 1.02 | 1.02 | 1.26 | 1.02 | 1.09 |

Figure 14. Three different permutation approaches. The 1\textsuperscript{st} way is to mix chemicals randomly, the 2\textsuperscript{nd} way is to mix chemicals in order, and the 3\textsuperscript{rd} way is to mix chemicals in a staggered manner.

Figure 15. The THz frequency-domain spectrum of different permutation approaches. The amplitude deviation among three different permutation approaches (uniform mix, stack mix, and staggered mix) is less than 2 dB within 0.2 - 1.0 THz.
Table 3: The quantitative comparison between the HYPERION and NMF on five substances (HYPERION/NMF) under different noise levels. Performances of HYPERION and NMF are evaluated by the root mean square error (RMSE). Superior values are labeled in bold and are highlighted with different colors (HYPERION: orange; NMF: blue).

| Noise level | D-lactose monohydrate | D-glucose | L-tyrosine | L-histidine | D-sucrose | Average |
|-------------|------------------------|-----------|------------|-------------|-----------|---------|
| $\sigma V_{\rho} = 0.001\%$ | 3.6140/3.4503 | 2.6160/4.6785 | 4.9644/5.1320 | 2.9567/3.4479 | 3.1040/4.7188 | 3.5034/4.3420 |
| (75.16 dB)   |           |           |            |             |           |         |
| $\sigma V_{\rho} = 0.002\%$ | 3.6403/3.3226 | 2.6683/2.6344 | 4.8826/4.9983 | 2.8614/3.4611 | 3.1237/3.8781 | 3.5203/3.7412 |
| (69.15 dB)   |           |           |            |             |           |         |
| $\sigma V_{\rho} = 0.003\%$ | 3.5377/2.9873 | 2.6230/4.1950 | 4.8368/5.1681 | 2.8525/3.6541 | 3.0069/4.8048 | 3.4632/2.2349 |
| (65.57 dB)   |           |           |            |             |           |         |
| $\sigma V_{\rho} = 0.004\%$ | 3.6261/3.1662 | 2.5114/3.4721 | 4.8981/5.3325 | 2.8963/3.4692 | 3.0972/4.2565 | 3.5043/4.0167 |
| (63.08 dB)   |           |           |            |             |           |         |
| $\sigma V_{\rho} = 0.005\%$ | 3.6292/3.7409 | 2.6276/3.9129 | 4.8068/5.2014 | 2.8535/3.5072 | 3.1220/4.8088 | 3.4949/4.3106 |
| (61.15 dB)   |           |           |            |             |           |         |
| $\sigma V_{\rho} = 0.006\%$ | 3.5050/3.2655 | 2.4541/4.3019 | 4.8946/5.4083 | 2.8899/3.7053 | 3.2840/4.7271 | 3.4264/4.3473 |
| (59.60 dB)   |           |           |            |             |           |         |
| $\sigma V_{\rho} = 0.007\%$ | 3.5730/3.6872 | 2.8368/4.1089 | 4.8503/5.1466 | 2.9333/3.2788 | 3.2998/4.9054 | 3.5730/4.1864 |
| (58.19 dB)   |           |           |            |             |           |         |
| $\sigma V_{\rho} = 0.008\%$ | 3.5089/3.0500 | 2.5347/4.3118 | 4.9483/5.2554 | 2.9330/3.3217 | 3.1104/4.4470 | 3.5217/4.1550 |
| (57.02 dB)   |           |           |            |             |           |         |
| $\sigma V_{\rho} = 0.009\%$ | 3.6000/3.4783 | 2.7313/3.6033 | 4.9100/5.2351 | 2.8537/3.5643 | 3.0564/4.4112 | 3.5217/4.1199 |
| (56.02 dB)   |           |           |            |             |           |         |
| $\sigma V_{\rho} = 0.01\%$ | 3.5416/3.1702 | 2.5252/4.9713 | 4.9761/5.2580 | 2.8540/3.4808 | 3.1656/4.9449 | 3.5170/4.4491 |
| (55.05 dB)   |           |           |            |             |           |         |
| $\sigma V_{\rho} = 0.02\%$ | 3.7670/4.8878 | 2.7651/2.6345 | 4.8250/4.8321 | 2.8209/3.4781 | 3.5658/3.8478 | 3.6274/4.0270 |
| (49.07 dB)   |           |           |            |             |           |         |
| $\sigma V_{\rho} = 0.03\%$ | 3.6676/5.4747 | 2.8837/2.8082 | 5.0670/4.5282 | 3.0939/4.4415 | 2.9091/4.7917 | 3.6188/4.4950 |
| (45.48 dB)   |           |           |            |             |           |         |
| $\sigma V_{\rho} = 0.04\%$ | 3.7316/3.7302 | 2.8532/2.6471 | 4.9270/4.8576 | 2.8964/4.1455 | 3.3033/4.2762 | 3.6304/4.0271 |
| (43.04 dB)   |           |           |            |             |           |         |
| $\sigma V_{\rho} = 0.05\%$ | 3.7342/3.7029 | 2.2775/2.3344 | 5.1655/5.2176 | 3.0811/4.0381 | 3.8650/4.3404 | 3.7454/4.0381 |
| (41.11 dB)   |           |           |            |             |           |         |
| $\sigma V_{\rho} = 0.06\%$ | 4.1557/6.0125 | 2.8107/3.5731 | 5.3080/5.0936 | 3.2992/3.3370 | 3.3463/4.9151 | 3.8840/4.6937 |
| (39.63 dB)   |           |           |            |             |           |         |
| $\sigma V_{\rho} = 0.07\%$ | 3.4130/4.5678 | 3.9707/3.3905 | 4.8356/4.7439 | 3.9909/5.7263 | 4.0345/6.1963 | 4.0742/5.0209 |
| (38.21 dB)   |           |           |            |             |           |         |
| $\sigma V_{\rho} = 0.08\%$ | 3.3973/5.9899 | 3.4688/3.6627 | 4.5378/4.8091 | 3.8142/4.1332 | 4.1016/6.1756 | 3.9618/5.0526 |
| (37.04 dB)   |           |           |            |             |           |         |
| $\sigma V_{\rho} = 0.09\%$ | 4.2028/5.4271 | 2.5348/3.3437 | 5.3325/5.1830 | 3.1667/3.5642 | 4.3521/4.5720 | 4.0906/4.4999 |
| (36.05 dB)   |           |           |            |             |           |         |
| $\sigma V_{\rho} = 0.1\%$  | 3.6703/5.9964 | 2.5039/3.6946 | 5.7183/6.0404 | 3.4767/5.0297 | 3.1753/6.1030 | 3.8829/5.3667 |
| (35.07 dB)   |           |           |            |             |           |         |
| Noise-free   | 3.6111/3.4958 | 2.5919/4.5175 | 4.8647/5.1673 | 2.8487/3.4869 | 3.0766/4.6753 | 3.4930/4.3208 |
Table 4: The quantitative comparison between the HYPERION and HMFA on five substances (HYPERION/HMFA) under different noise levels. Performances of HYPERION and HMFA are evaluated by the root mean square error (RMSE). Superior values are labeled in bold and are highlighted with different colors (HYPERION: orange; HMFA: blue).

| Noise Level | D-lactose monohydrate | D-glucose | L-tyrosine | L-histidine | D-sucrose | Average |
|-------------|------------------------|-----------|------------|-------------|-----------|---------|
| Noise-free | 3.6149/9.0792          | 2.6160/8.7788 | 4.3644/6.5146 | 2.8567/7.4226 | 3.1040/4.1796 | 3.5034/7.4095 |
| σV_p = 0.001% (75.16 dB) | 3.6403/9.0864          | 2.6683/8.7858 | 4.8626/6.5221 | 2.8614/7.4311 | 3.1237/4.1848 | 3.5203/7.4165 |
| σV_p = 0.002% (69.15 dB) | 3.5377/9.0318          | 2.6230/8.7249 | 4.3368/6.4634 | 2.8525/7.3845 | 3.0063/4.1701 | 3.4632/7.3634 |
| σV_p = 0.003% (65.57 dB) | 3.6261/9.5155          | 2.5114/6.9090 | 4.8816/6.1629 | 2.8963/7.5512 | 3.0927/3.9291 | 3.5043/7.4881 |
| σV_p = 0.004% (63.08 dB) | 3.6292/9.1243          | 2.6276/8.8220 | 4.8068/6.5587 | 2.8535/7.4707 | 3.1220/4.1957 | 3.4949/7.4500 |
| σV_p = 0.005% (61.15 dB) | 3.5050/9.1524          | 2.4641/8.8438 | 4.8946/6.0773 | 2.8898/7.4934 | 2.8460/4.1976 | 3.4284/7.4701 |
| σV_p = 0.006% (59.66 dB) | 3.5730/9.4469          | 2.9368/8.8466 | 4.5503/6.1008 | 2.9333/7.4862 | 3.2999/3.9245 | 3.5730/7.4304 |
| σV_p = 0.007% (58.19 dB) | 3.5809/9.5430          | 2.5347/9.2611 | 4.9483/6.0801 | 2.9330/7.4908 | 3.1194/4.2811 | 3.5217/7.8351 |
| σV_p = 0.008% (57.02 dB) | 3.6000/8.9617          | 2.7313/8.6499 | 4.9100/6.3905 | 2.8537/7.2869 | 3.0564/4.1693 | 3.5217/7.5456 |
| σV_p = 0.009% (56.02 dB) | 3.5416/9.5706          | 2.5252/8.9867 | 4.9761/6.2499 | 2.8540/7.6356 | 3.1656/3.9316 | 3.5170/7.5527 |
| σV_p = 0.01% (55.05 dB) | 3.7670/8.9409          | 2.7851/8.2717 | 4.8250/5.5812 | 2.8209/6.8551 | 3.5683/3.9119 | 3.6274/6.9542 |
| σV_p = 0.02% (49.07 dB) | 3.6678/9.5448          | 2.8337/9.2946 | 5.0670/7.0108 | 3.0933/7.9226 | 2.9091/4.3685 | 3.6188/7.8670 |
| σV_p = 0.03% (45.48 dB) | 3.7316/8.6366          | 2.5532/7.9801 | 4.9270/5.3014 | 2.5864/6.5176 | 3.3933/3.8310 | 3.5304/6.6863 |
| σV_p = 0.04% (43.04 dB) | 3.7342/9.0481          | 2.2775/8.4761 | 5.1555/5.7343 | 3.0811/7.0827 | 3.8650/3.8647 | 3.7454/7.0950 |
| σV_p = 0.05% (41.11 dB) | 4.1657/9.6244          | 2.8197/9.1878 | 5.3080/6.8676 | 3.2992/7.5950 | 3.3463/3.9511 | 3.8840/7.7137 |
| σV_p = 0.06% (39.63 dB) | 3.4130/10.9347         | 3.9707/10.4539 | 4.8356/7.7910 | 3.9900/9.2055 | 4.0345/4.2419 | 4.0742/8.8576 |
| σV_p = 0.07% (38.21 dB) | 3.3973/12.5002         | 3.4689/8.4531 | 4.5378/7.8309 | 3.8142/9.2486 | 4.1016/4.1776 | 3.9518/8.8546 |
| σV_p = 0.08% (37.04 dB) | 4.2028/11.8970         | 2.5349/11.4039 | 5.3239/8.7368 | 3.1667/10.1914 | 4.3521/4.4985 | 4.0906/9.7162 |
| σV_p = 0.09% (36.05 dB) | 3.6703/10.6357         | 2.5039/10.1022 | 5.7183/7.4071 | 3.4767/8.8139 | 3.1753/4.0896 | 3.8628/8.5375 |
| σV_p = 0.1% (35.07 dB) | 3.6111/9.0780          | 2.5919/8.7767 | 4.8647/6.5122 | 2.8487/7.4204 | 3.0766/4.1793 | 3.4930/7.4078 |
Table 5: The quantitative comparison between the HYPERION and nICA on five substances (HYPERION/nICA) under different noise levels. Performances of HYPERION and nICA are evaluated by the root mean square error (RMSE). Superior values are labeled in bold and are highlighted with different colors (HYPERION: orange; nICA: blue).

|                        | D-lactose monohydrate | D-glucose  | L-tyrosine | L-histidine | D-sucrose | Average |
|------------------------|------------------------|------------|------------|-------------|-----------|---------|
| $\sigma_{V_p} = 0.001\%$ (75.16 dB) | 3.614/12.3305          | 2.616/10.1294 | 4.864/4.82883 | 2.856/8.8289 | 3.104/11.1584 | 3.503/10.2544 |
| $\sigma_{V_p} = 0.002\%$ (69.15 dB) | 3.640/12.3452          | 2.668/10.1298 | 4.862/4.82758 | 2.861/8.8293 | 3.123/11.1610 | 3.520/10.2570 |
| $\sigma_{V_p} = 0.003\%$ (65.57 dB) | 3.537/12.3238          | 2.623/10.1287 | 4.836/4.82968 | 2.852/8.8498 | 3.006/11.1387 | 3.463/10.2537 |
| $\sigma_{V_p} = 0.004\%$ (63.06 dB) | 3.626/12.3427          | 2.511/10.1329 | 4.898/4.82858 | 2.896/8.8111 | 3.092/11.1725 | 3.504/10.2580 |
| $\sigma_{V_p} = 0.005\%$ (61.15 dB) | 3.629/12.3265          | 2.627/10.1258 | 4.806/4.82837 | 2.863/8.8154 | 3.122/11.1739 | 3.494/10.2534 |
| $\sigma_{V_p} = 0.006\%$ (59.60 dB) | 3.505/12.3139          | 2.454/10.1440 | 4.894/4.83074 | 2.999/8.8623 | 2.846/11.1218 | 3.428/10.2546 |
| $\sigma_{V_p} = 0.007\%$ (58.19 dB) | 3.573/12.3433          | 2.836/10.1234 | 4.850/4.82807 | 2.933/8.8388 | 3.298/11.1558 | 3.573/10.2566 |
| $\sigma_{V_p} = 0.008\%$ (57.02 dB) | 3.583/12.3223          | 2.534/10.1438 | 4.948/4.83024 | 2.933/8.8416 | 3.110/11.1379 | 3.521/10.2566 |
| $\sigma_{V_p} = 0.009\%$ (56.02 dB) | 3.620/12.3359          | 2.731/10.1415 | 4.910/4.83015 | 2.853/8.8548 | 3.056/11.1241 | 3.521/10.2574 |
| $\sigma_{V_p} = 0.01\%$ (55.05 dB)  | 3.541/12.3296          | 2.525/10.1487 | 4.976/4.83163 | 2.854/8.8161 | 3.165/11.1538 | 3.517/10.2597 |
| $\sigma_{V_p} = 0.02\%$ (49.07 dB)  | 3.767/12.3552          | 2.786/10.1010 | 4.829/4.82462 | 2.820/8.8121 | 3.586/11.2376 | 3.627/10.2627 |
| $\sigma_{V_p} = 0.03\%$ (45.48 dB)  | 3.667/12.3446          | 2.883/10.1619 | 5.067/4.8322 | 3.093/8.8898 | 2.909/11.1037 | 3.618/10.2703 |
| $\sigma_{V_p} = 0.04\%$ (43.04 dB)  | 3.731/12.2789          | 2.653/10.1392 | 4.927/4.82763 | 2.586/8.7841 | 3.303/11.2213 | 3.530/10.2484 |
| $\sigma_{V_p} = 0.05\%$ (41.11 dB)  | 3.734/12.2762          | 2.277/10.0697 | 5.155/4.8446 | 3.081/8.7891 | 3.850/11.3440 | 3.745/10.2573 |
| $\sigma_{V_p} = 0.06\%$ (39.63 dB)  | 4.166/12.3103          | 2.810/10.2029 | 5.308/4.8201 | 3.299/8.9418 | 3.346/11.0313 | 3.884/10.2751 |
| $\sigma_{V_p} = 0.07\%$ (38.21 dB)  | 3.413/12.2656          | 3.970/10.2234 | 4.835/4.8408 | 3.990/8.9401 | 4.035/11.1737 | 4.074/10.2728 |
| $\sigma_{V_p} = 0.08\%$ (37.04 dB)  | 3.397/12.2939          | 3.469/10.2199 | 4.537/4.82679 | 3.814/8.9403 | 4.101/11.1738 | 3.951/10.2833 |
| $\sigma_{V_p} = 0.09\%$ (35.65 dB)  | 4.208/12.2938          | 2.534/10.1510 | 5.532/4.8377 | 3.166/8.8246 | 4.352/11.2604 | 4.096/10.2781 |
| $\sigma_{V_p} = 0.1\%$ (35.07 dB)   | 3.670/12.3379          | 2.003/10.2379 | 5.718/4.82616 | 3.476/8.9012 | 3.175/11.1288 | 3.862/10.2798 |
| Noise-free             | 3.611/12.3288          | 2.591/10.1299 | 4.864/8.2909 | 2.848/8.8275 | 3.076/11.1562 | 3.493/10.2542 |
Table 6: The quantitative comparison between the HYPERION and SPA on five substances (HYPERION/SPA) under different noise levels. Performances of HYPERION and SPA are evaluated by the root mean square error (RMSE). Superior values are labeled in bold and are highlighted with different colors (HYPERION: orange; SPA: blue).

|                  | D-lactose monohydrate | D-glucose | L-tyrosine | L-histidine | D-sucrose | Average |
|------------------|------------------------|-----------|------------|-------------|-----------|---------|
| σv_p = 0.001%   | 3.614/5.2706           | 2.616/2.8631 | 4.864/4.2611 | 2.856/4.0573 | 3.104/3.0517 | 3.503/4.2292 |
| (75.16 dB)      |                        |           |            |             |           |         |
| σv_p = 0.002%   | 3.640/5.2613           | 2.688/2.8599 | 4.862/5.2641 | 2.861/4.0528 | 3.123/3.0802 | 3.520/4.2304 |
| (69.15 dB)      |                        |           |            |             |           |         |
| σv_p = 0.003%   | 3.537/5.2784           | 2.623/2.8636 | 4.838/5.2481 | 2.852/4.0454 | 3.006/3.0094 | 3.463/4.2196 |
| (65.57 dB)      |                        |           |            |             |           |         |
| σv_p = 0.004%   | 3.626/5.2706           | 2.511/2.8692 | 4.898/5.2478 | 2.896/4.0426 | 3.092/3.0614 | 3.504/4.2299 |
| (63.08 dB)      |                        |           |            |             |           |         |
| σv_p = 0.005%   | 3.629/5.2641           | 2.627/2.8678 | 4.808/5.2487 | 2.863/4.0426 | 3.122/3.0523 | 3.485/4.2224 |
| (61.15 dB)      |                        |           |            |             |           |         |
| σv_p = 0.006%   | 3.505/5.3009           | 2.451/2.8985 | 4.894/5.2152 | 2.899/4.0630 | 2.846/2.9213 | 3.428/4.2129 |
| (59.60 dB)      |                        |           |            |             |           |         |
| σv_p = 0.007%   | 3.573/5.2487           | 2.836/2.8025 | 4.850/5.3039 | 2.933/4.0311 | 3.298/3.1155 | 3.573/4.2427 |
| (56.19 dB)      |                        |           |            |             |           |         |
| σv_p = 0.008%   | 3.580/5.4537           | 2.534/3.8927 | 4.943/5.4157 | 2.933/3.4633 | 3.110/4.4109 | 3.521/4.3938 |
| (57.02 dB)      |                        |           |            |             |           |         |
| σv_p = 0.009%   | 3.600/5.2669           | 2.731/2.8662 | 4.910/5.2682 | 2.853/4.0548 | 3.056/3.0524 | 3.521/4.2301 |
| (56.02 dB)      |                        |           |            |             |           |         |
| σv_p = 0.01%    | 3.541/4.2048           | 2.525/3.9579 | 4.976/6.0753 | 2.854/3.2431 | 3.165/3.2821 | 3.517/4.2789 |
| (55.05 dB)      |                        |           |            |             |           |         |
| σv_p = 0.02%    | 3.767/5.2760           | 2.786/2.9482 | 4.828/5.3367 | 2.820/4.0886 | 3.565/3.2473 | 3.627/4.2958 |
| (49.07 dB)      |                        |           |            |             |           |         |
| σv_p = 0.03%    | 3.667/4.5014           | 2.883/3.9625 | 5.067/5.4005 | 3.093/3.4466 | 2.909/4.6888 | 3.618/4.4499 |
| (45.48 dB)      |                        |           |            |             |           |         |
| σv_p = 0.04%    | 3.731/4.6153           | 2.653/4.5767 | 4.927/5.2415 | 2.584/3.0794 | 3.303/3.2805 | 3.530/4.2418 |
| (43.04 dB)      |                        |           |            |             |           |         |
| σv_p = 0.05%    | 3.734/4.2384           | 2.277/3.9538 | 5.155/5.9286 | 3.081/3.0119 | 3.865/3.2723 | 3.745/4.2077 |
| (41.11 dB)      |                        |           |            |             |           |         |
| σv_p = 0.06%    | 4.155/5.1791           | 2.810/3.3936 | 5.308/6.2274 | 3.299/3.6143 | 3.483/3.8846 | 3.884/4.4386 |
| (39.63 dB)      |                        |           |            |             |           |         |
| σv_p = 0.07%    | 3.413/5.3799           | 3.970/4.2422 | 4.835/5.6287 | 3.990/3.3939 | 4.034/3.0721 | 4.072/4.5515 |
| (38.21 dB)      |                        |           |            |             |           |         |
| σv_p = 0.08%    | 3.397/5.3310           | 3.469/4.5682 | 4.537/5.3959 | 3.814/3.8825 | 4.101/3.3322 | 3.951/4.5734 |
| (37.04 dB)      |                        |           |            |             |           |         |
| σv_p = 0.09%    | 4.202/6.1861           | 2.534/4.4227 | 5.532/5.7359 | 3.166/4.9885 | 4.352/3.4175 | 4.096/5.0456 |
| (36.05 dB)      |                        |           |            |             |           |         |
| σv_p = 0.1%     | 3.670/4.8825           | 2.603/3.9650 | 5.719/5.7829 | 3.476/3.6058 | 3.175/3.3515 | 3.863/4.4099 |
| (35.07 dB)      |                        |           |            |             |           |         |
| Noise-free      | 3.611/5.2688           | 2.591/2.8649 | 4.864/5.2569 | 2.848/4.0571 | 3.076/3.0426 | 3.493/4.2266 |
Table 7: Tablet thickness in the quinary dataset.

| Tablet        | Glucose | Lactose | Tyrosine | Histidine | Sucrose |
|---------------|---------|---------|----------|-----------|---------|
| Thickness (mm)| 2.77    | 2.85    | 3.27     | 3.18      | 2.85    |
| Tablet        | Glucose + Lactose | Glucose + Tyrosine | Glucose + Histidine | Glucose + Sucrose | Lactose + Tyrosine |
| Thickness (mm)| 2.85    | 3.04    | 3.40     | 2.98      | 3.04    |
| Tablet        | Lactose + Histidine | Lactose + Sucrose | Tyrosine + Histidine | Tyrosine + Sucrose | Histidine + Sucrose |
| Thickness (mm)| 3.13    | 2.89    | 3.29     | 3.17      | 3.00    |
Table 8: The convergence time of HYPERION, HMFA, NMF, SPA, and nICA.

| Algorithm   | HYPERION | HMFA | NMF  | SPA  | nICA |
|-------------|----------|------|------|------|------|
| Convergence time (s) | 2.52  | 0.23 | 0.18 | 0.01 | 1.07 |
References

1. Koenig, S. et al. Wireless sub-THz communication system with high data rate. Nat. Photonics 7, 977–981 (2013).
2. Curran, P. J. Review Article Remote sensing methodologies and geography. Int. J. Remote Sens. 8, 1255–1275 (1987).
3. Akiyama, K. et al. First M87 event horizon telescope results. IV. Imaging the central supermassive black hole. Astrophys. J. 875, L4 (2019).
4. Gorelick, N. et al. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27 (2017).
5. Faccio, D. & Velten, A. A trillion frames per second: the techniques and applications of light-in-flight photography. Rep. Prog. Phys. 81, 105901 (2018).
6. Globus, T. et al. THz-spectroscopy of biological molecules. J. Biol. Phys. 29, 89–100 (2003).
7. Markelz, A. G., Knab, J. R., Chen, J. Y. & He, Y. Protein dynamical transition in terahertz dielectric response. Chem. Phys. Lett. 442, 413–417 (2007).
8. Hung, Y.-C., Lin, C.-H., Wang, F.-Y. & Yang, S.-H. Penetrating Terahertz Hyperspectral Unmixing via Löwner-John Ellipsoid (THz HU-LJE): An Unsupervised Algorithm in 2020 45th International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz) (2020), 1–2.
9. George, P. A. et al. Ultrafast optical-pump terahertz-probe spectroscopy of the carrier relaxation and recombination dynamics in epitaxial graphene. Nano Lett. 8, 4248–4251 (2008).
10. Kawase, K., Ogawa, Y., Watanabe, Y. & Inoue, H. Non-destructive terahertz imaging of illicit drugs using spectral fingerprints. Opt. Express 11, 2549–2554 (2003).
11. Federici, J. F. et al. THz imaging and sensing for security applications—explosives, weapons and drugs. Semicond. Sci. Technol. 20, S266 (2005).
12. Huber, A. J., Keilmann, F., Wittborn, J., Aizpurua, J. & Hillenbrand, R. Terahertz near-field nanoscopy of mobile carriers in single semiconductor nanodevices. Nano Lett. 8, 3766–3770 (2008).
13. Cocker, T. L. et al. An ultrafast terahertz scanning tunnelling microscope. Nat. Photonics 7, 620–625 (2013).
14. Schmuttenmaer, C. A. Exploring dynamics in the far-infrared with terahertz spectroscopy. Chem. Rev. 104, 1759–1780 (2004).
15. Cherkasova, O., Nazarov, M., Konnikova, M. & Shkurinov, A. THz spectroscopy of bound water in glucose: Direct measurements from crystalline to dissolved state. *J. Infrared Millim. Terahertz Waves* **41**, 1057–1068 (2020).

16. Upadhya, P., Shen, Y., Davies, A. & Linfield, E. Terahertz time-domain spectroscopy of glucose and uric acid. *J. Biol. Phys.* **29**, 117–121 (2003).

17. Dorney, T. D., Baraniuk, R. G. & Mittleman, D. M. Material parameter estimation with terahertz time-domain spectroscopy. *J. Opt. Soc. Am. A* **18**, 1562–1571 (2001).

18. Zhong, H., Redo-Sanchez, A. & Zhang, X.-C. Identification and classification of chemicals using terahertz reflective spectroscopic focal-plane imaging system. *Opt. Express* **14**, 9130–9141 (2006).

19. Li, X., Hou, D., Huang, P., Cai, J. & Zhang, G. Component spectra extraction from terahertz measurements of unknown mixtures. *Appl. Opt.* **54**, 8925–8934 (2015).

20. Ma, Y. *et al.* THz spectral data analysis and components unmixing based on non-negative matrix factorization methods. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **177**, 49–57 (2017).

21. Balci, S. *et al.* Independent component analysis applications on THz sensing and imaging in *Image Sensing Technologies: Materials, Devices, Systems, and Applications III* **9854** (2016), 98540K.

22. Lin, C.-H., Wu, R., Ma, W.-K., Chi, C.-Y. & Wang, Y. Maximum Volume Inscribed Ellipsoid: A New Simplex-Structured Matrix Factorization Framework via Facet Enumeration and Convex Optimization. *SIAM J. Imaging Sci.* **11**, 1651–1679 (2018).

23. Wei, Q., Bioucas-Dias, J., Dobigeon, N. & Tourneret, J.-Y. Hyperspectral and multispectral image fusion based on a sparse representation. *IEEE Trans. Geosci. Remote Sens.* **53**, 3658–3668 (2015).

24. Loncan, L. *et al.* Hyperspectral pansharpening: A review. *IEEE Geosci. Remote Sens.* **3**, 27–46 (2015).

25. Nowak, K. *et al.* Selected aspects of terahertz spectroscopy in pharmaceutical sciences. *Acta Pol. Pharm.* **72**, 851–866 (2015).

26. Gamble, J. F. *et al.* Investigation into the degree of variability in the solid-state properties of common pharmaceutical excipients—anhydrous lactose. *AAPS PharmSciTech* **11**, 1552–1557 (2010).

27. Lin, C.-H., Ma, W.-K., Li, W.-C., Chi, C.-Y. & Ambikapathi, A. Identifiability of the Simplex Volume Minimization Criterion for Blind Hyperspectral Unmixing: The No Pure-Pixel Case. *IEEE Trans. Geosci. Remote Sens.* **53**, 5530–5546 (2015).
28. Heshmat, B. et al. Terahertz scattering and water absorption for porosimetry. Opt. Express 25, 27370–27385 (2017).
29. Lin, C.-H. & Bioucas-Dias, J. M. Nonnegative blind source separation for ill-conditioned mixtures via John ellipsoid. IEEE Trans. Neural Netw. Learn. Syst. 32, 2209–2223 (2020).
30. Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P. & Plemmons, R. J. Algorithms and applications for approximate nonnegative matrix factorization. Comput. Stat. Data Anal. 52, 155–173 (2007).
31. Plumbley, M. D. Algorithms for nonnegative independent component analysis. IEEE Trans. Neural Netw. 14, 534–543 (2003).
32. Kriesten, E. et al. Identification of unknown pure component spectra by indirect hard modeling. Chemometr. Intell. Lab. Syst. 93, 108–119 (2008).
33. Arora, S. et al. A practical algorithm for topic modeling with provable guarantees in International conference on machine learning (2013), 280–288.
34. Lin, C.-J. On the convergence of multiplicative update algorithms for nonnegative matrix factorization. IEEE Trans. Neural Netw. 18, 1589–1596 (2007).
35. Albright, R., Cox, J., Duling, D., Langville, A. N. & Meyer, C. Algorithms, initializations, and convergence for the nonnegative matrix factorization tech. rep. (Tech. rep. 919. NCSU Technical Report Math 81706. http://meyer. math. ncsu . . ., 2006).
36. Ma, W.-K. et al. A signal processing perspective on hyperspectral unmixing: Insights from remote sensing. IEEE Signal Process. Mag. 31, 67–81 (2013).
37. Grant, M. & Boyd, S. CVX: Matlab Software for Disciplined Convex Programming, version 1.21 http://cvxr.com/cvx/. 2011.
38. Sun, Q. et al. Recent advances in terahertz technology for biomedical applications. Quant. Imaging. Med. Surg. 7, 345 (2017).
39. Seco-Martorell, C. et al. Goya’s artwork imaging with Terahertz waves. Opt. Express 21, 17800–17805 (2013).
40. Lawrence, X. Y. et al. Applications of process analytical technology to crystallization processes. Adv. Drug Deliv. Rev. 56, 349–369 (2004).
41. Duvillaret, L., Garet, F. & Coutaz, J.-L. Highly precise determination of optical constants and sample thickness in terahertz time-domain spectroscopy. Appl. Opt. 38, 409–415 (1999).
42. Lin, C.-H. & Bioucas-Dias, J. M. Nonnegative blind source separation for ill-conditioned mixtures via John ellipsoid. IEEE Trans. Neural Netw. Learn. Syst. (2020).
43. Lin, C.-H., Chi, C.-Y., Wang, Y.-H. & Chan, T.-H. A fast hyperplane-based minimum-volume enclosing simplex algorithm for blind hyperspectral unmixing. *IEEE Trans. Signal Process.* **64**, 1946–1961 (2016).

44. Chi, C.-Y., Li, W.-C. & Lin, C.-H. *Convex Optimization for Signal Processing and Communications: From Fundamentals to Applications* (CRC Press, Boca Raton, FL, 2017).

45. Barber, C. B., Dobkin, D. P. & Huhdanpaa, H. The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.* **22**, 469–483 (1996).

46. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

47. Arora, S. *et al.* A practical algorithm for topic modeling with provable guarantees. *International Conference on Machine Learning* (2013).

48. Craig, M. D. Minimum-volume transforms for remotely sensed data. *IEEE Trans. Geosci. Remote Sens.* **32**, 542–552 (1994).

49. Green, R. O. *et al.* Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS). *Remote Sens. Environ.* **65**, 227–248 (1998).

50. Bioucas-Dias, J. M. & Figueiredo, M. A. T. *Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing* in 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS) (2010), 1–4.

51. Heinz, D. & Chang, C.-I. Fully constrained least squares linear mixture analysis for material quantification in hyperspectral imagery. *IEEE Trans. Geosci. Remote Sens.* **39**, 529–545 (2001).

52. Jepsen, P. U., Cooke, D. G. & Koch, M. Terahertz spectroscopy and imaging—Modern techniques and applications. *Laser Photonics Rev.* **5**, 124–166 (2011).

53. Neu, J. & Schmuttenmaer, C. A. Tutorial: An introduction to terahertz time domain spectroscopy (THz-TDS). *J. Appl. Phys.* **124**, 231101 (2018).

54. Burford, N. M. & El-Shenawee, M. O. Review of terahertz photoconductive antenna technology. *Opt. Eng.* **56**, 010901 (2017).

55. Janke, C., Först, M., Nagel, M., Kurz, H. & Bartels, A. Asynchronous optical sampling for high-speed characterization of integrated resonant terahertz sensors. *Opt. Lett.* **30**, 1405–1407 (2005).

56. Tani, M., Herrmann, M. & Sakai, K. Generation and detection of terahertz pulsed radiation with photoconductive antennas and its application to imaging. *Meas. Sci. Technol.* **13**, 1739 (2002).
57. Preu, S., Döhler, G., Malzer, S., Wang, L. & Gossard, A. Tunable, continuous-wave terahertz photomixer sources and applications. *J. Appl. Phys.* **109**, 4 (2011).

58. Jepsen, P. U. & Fischer, B. M. Dynamic range in terahertz time-domain transmission and reflection spectroscopy. *Optics letters* **30**, 29–31 (2005).

59. Bandyopadhyay, A. *et al.* Effects of scattering on THz spectra of granular solids. *International Journal of Infrared and Millimeter Waves* **28**, 969–978 (2007).