THE ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES, 284:67 (36pp), 2026 June https://doi.org/10.3847/1538-4365/ae6068

HETDEX Public Data Release 1: Source Catalog 2 and Data Cubes from $\sim$90 deg^2 of Integral-field Optical Spectroscopy

Erin Mentuch Cooper[1], Karl Gebhardt[1], Dustin Davis[1], Chenxu Liu (刘辰旭)[2], Barbara G. Castanheira[3], Owen Chase[1,4], Óscar A. Chávez Ortiz[1], Robin Ciardullo[5,6], Olivia Curtis[5,6], Delaney A. Dunne[7], Neal J. Evans, II[8], Daniel J. Farrow[9,10], Maximilian Fabricius[11,12], Steven L. Finkelstein[1,13], Caryl Gronwall[5,6], Nathaniel J. Hamme[5], Gary J. Hill[1,14], Lindsay R. House[15,16], Matt J. Jarvis[17,18], Donghui Jeong[5,6,19], Andreas Kelz[20], Eiichiro Komatsu[21,22,23], Mahan Mirza Khanlari[1], Hasti Khoraminezhad[24], Wolfram Kollatschny[25], Maja Lujan Niemeyer[21,22], Hanshin Lee[14], Phillip MacQueen[14], Deeshani Mitra[24], Shiro Mukae[1,26], Masami Ouchi[23,27,28,29], Jennifer Poppe[1], Meredith C. Powell[20], Mahdi Qezlou[1], Shun Saito[23,24], Donald P. Schneider[5,6], Laurel Weiss[1], Lutz Wisotzki[20], and Gregory R. Zeimann[30]

[1] Department of Astronomy, The University of Texas at Austin, 2515 Speedway Boulevard, Austin, TX 78712, USA; erin.hetdex@gmail.com
[2] South-Western Institute for Astronomy Research, Key Laboratory of Survey Science of Yunnan Province, Yunnan University, Kunming, Yunnan 650500, People's Republic of China
[3] Department of Physics and Astronomy, Baylor University, Waco, TX 76798-7316, USA
[4] Weinberg Institute, The University of Texas at Austin, 2515 Speedway Boulevard, Austin, TX 78712, USA
[5] Department of Astronomy & Astrophysics, The Pennsylvania State University, University Park, PA 16802, USA
[6] Institute for Gravitation and the Cosmos, The Pennsylvania State University, University Park, PA 16802, USA
[7] California Institute of Technology, 1200 E. California Boulevard, Pasadena, CA 91125, USA
[8] Department of Astronomy, The University of Texas at Austin, 2515 Speedway, Stop C1400, Austin, TX 78712-1205, USA
[9] E. A. Milne Centre for Astrophysics University of Hull, Cottingham Road, Hull, HU6 7RX, UK
[10] Centre of Excellence for Data Science, Artificial Intelligence & Modelling (DAIM), University of Hull, Cottingham Road, Hull, HU6 7RX, UK
[11] Max Planck Institute for Extraterrestrial Physics, Giessenbachstr. 1, 85748 Garching, Germany
[12] Universitäts-Sternwarte München, Fakultät für Physik, Ludwig-Maximilians-Universität München, Scheinerstrasse 1, 81679, München, Germany
[13] Cosmic Frontier Center, The University of Texas at Austin, Austin, TX, USA
[14] McDonald Observatory, The University of Texas at Austin, 2515 Speedway Boulevard, Austin, TX 78712, USA
[15] Data Science Institute, The University of Chicago, 5460 S University Avenue, Chicago, IL 60615, USA
[16] NSF-Simons AI Institute for the Sky (SkAI), 172 E. Chestnut Street, Chicago, IL 60611, USA
[17] Astrophysics, Department of Physics, University of Oxford, Keble Road, Oxford, OX1 3RH, UK
[18] Department of Physics and Astronomy, University of the Western Cape, Robert Sobukwe Road, 7535 Bellville, Cape Town, South Africa
[19] School of Physics, Korea Institute for Advanced Study, Seoul 02455, Republic of Korea
[20] Leibniz-Institut für Astrophysik Potsdam (AIP), An der Sternwarte 16, 14482 Potsdam, Germany
[21] Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, 85741 Garching, Germany
[22] Ludwig-Maximilians-Universität München, Schellingstr. 4, 80799, München, Germany
[23] Kavli Institute for the Physics and Mathematics of the Universe (WPI), The University of Tokyo Institutes for Advanced Study (UTIAS), The University of Tokyo, Chiba 277-8583, Japan
[24] Institute for Multi-messenger Astrophysics and Cosmology, Department of Physics, Missouri University of Science and Technology, 1315 N. Pine Street, Rolla, MO 65409, USA
[25] Institut für Astrophysik, Universität Göttingen, Friedrich-Hund-Platz 1, 37077, Göttingen, Germany
[26] MIRAI Technology Institute, Shiseido Co., Ltd., 1-2-11, Takashima, Nishi-ku, Yokohama, Kanagawa, 222-0011, Japan
[27] National Astronomical Observatory of Japan, 2-21-1 Osawa, Mitaka, Tokyo 181-8588, Japan
[28] Institute for Cosmic Ray Research, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8582, Japan
[29] Department of Astronomical Science, SOKENDAI (The Graduate University for Advanced Studies), Osawa 2-21-1, Mitaka, Tokyo, 181-8588, Japan
[30] Hobby-Eberly Telescope, University of Texas at Austin, Austin, TX 78712, USA

Received 2025 November 21; revised 2026 April 10; accepted 2026 April 11; published 2026 June 3

Abstract

The Hobby–Eberly Telescope Dark Energy Experiment (HETDEX) is a wide-field, integral-field spectroscopic survey designed to map the large-scale distribution of Lyα-emitting galaxies (LAEs) at $1.88 < z < 3.52$ and constrain dark energy at cosmic noon. Using the 10 m Hobby–Eberly Telescope and the Visible Integral-Field Replicable Unit (IFU) Spectrograph, HETDEX obtains >35,000 spectra per exposure over 3500–5500 Å at $R \sim 800$ with $\sim 1''\!.8$ image quality, enabling an untargeted census of emission-line galaxies across 540 deg^2. We present HETDEX Public Data Release 1 (PDR1), comprising 431,713 IFU observations covering 86.67 deg^2 of noncontiguous sky in the Spring ($13^{\rm h}$, $+51°$) and Fall ($1^{\rm h}\!.5$, $0°$) fields, along with legacy regions (Cosmic Evolution Survey, Great Observations Origins Deep Survey North, North Ecliptic Pole, SA22). PDR1 includes the HETDEX Public Source Catalog 2 (HPSC2), an expanded and reprocessed version of E. Mentuch Cooper et al. (2023) incorporating four additional years of data, improved quality control, and new machine learning classifiers. HPSC2 contains 426,654 LAEs, 491,411 [O II] emitters, 19,457 low-z galaxies, 18,303 active galactic nuclei, and 150,608 stars, providing coordinates, redshifts or stellar velocities, and 1D spectra for each source. Because the data cubes use local sky subtraction optimized for faint emission-line detection, they are not suited for absolute surface-brightness measurements or very extended nearby galaxies. Appendix materials include the full detection catalog, the 1.6 million–candidate LAE sample, and raw detection databases. All products are publicly accessible through the HETDEX data portal (https://hetdex.org/data-results/), including access to a public JUPYTERLAB. HPSC2 is also publicly available via Zenodo (doi:10.5281/zenodo.19581262).

Unified Astronomy Thesaurus concepts: Astronomy databases (83); Galaxy spectroscopy (2171); Lyman-alpha galaxies (978); Catalogs (205); Redshift surveys (1378); Emission line galaxies (459)

1. Introduction

Once considered a simple celestial sphere, the sky is now understood in full three dimensions thanks to wide-area spectroscopic surveys. Over the last two decades, surveys such as the Sloan Digital Sky Survey (SDSS; D. G. York et al. 2000), the SDSS-III Baryon Oscillation Spectroscopic Survey (BOSS; K. S. Dawson et al. 2013), the SDSS-IV extended BOSS (eBOSS; K. S. Dawson et al. 2016), and the Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration et al. 2016; B. Abareshi et al. 2022) have revealed the 3D, large-scale structure of the Universe. Galaxy spectra encode the distance measurements that allow us to map the 3D positions of millions of galaxies and quantify the evolving cosmological state of the Universe.

Astronomical spectra provide both precise distance measurements and detailed diagnostics of a source's astrophysical nature. Galaxy properties derived from spectra—including star formation rate, stellar mass, and metallicity—statistically reveal how galaxies evolve as a population. Spectra from stars reveal their temperature, surface gravity, metallicity, and atmospheric structure, just to name some of the vast range of content encapsulated in a source's spectrum.

Most spectroscopic surveys such as SDSS, eBOSS, and DESI, select their targets based on multiwavelength photometric imaging. Targets are chosen to maximize detection and ensure enough signal at a range of wavelengths such that multiple spectroscopic features can secure a spectroscopic redshift with high confidence. This approach generally requires long exposures, often many hours. These surveys position hundreds of fibers—or thousands in DESI's case—and obtain spectra of previously detected objects over a large field of view (FOV) in order to maximize the area surveyed. They are optimized for brighter galaxies and only obtain a single spectrum per source target.

The Hobby–Eberly Telescope Dark Energy Experiment (HETDEX; K. Gebhardt et al. 2021) takes a different approach by using an array of up to 78 integral field units (IFUs) to perform an untargeted search of the sky. Designed to detect strong line emission from $1.9 < z < 3.5$ Lyα-emitting galaxies (LAEs), HETDEX trades continuum sensitivity for sky coverage. The strong Lyα line emission allows for detection over a wide range of stellar masses (e.g., A. E. Shapley et al. 2003; E. M. Hu & L. L. Cowie 2006) and redshifts for objects generally too faint for detection in broadband images (A. Hagen et al. 2016; G. A. Oyarzún et al. 2017; S. Santos et al. 2020). See M. Ouchi et al. (2020) and references therein for a thorough review.

HETDEX is performed with the Visible IFU Replicable Unit Spectrograph (VIRUS; G. J. Hill et al. 2021), on the 10 m Hobby–Eberly Telescope (HET; L. W. Ramsey et al. 1998; G. J. Hill et al. 2021). VIRUS obtains ≈ 35,000 spectra simultaneously, each covering the wavelength range $3470\,\text{Å} \leqslant \lambda \leqslant 5540\,\text{Å}$ with a corresponding spectral resolving power $750 \lesssim R \lesssim 950$. HETDEX uses LAEs as a (biased) tracer of dark matter density; by measuring their clustering, HETDEX characterizes the Universe's dark energy density and tests for potential evolution (M. Shoji et al. 2009). HETDEX aims to measure the Hubble parameter, $H(z)$, and the angular diameter distance, $D_A(z)$, to better than 1% accuracy in the redshift range $1.9 < z < 3.5$. To achieve this precision, approximately 1 million LAE positions must be mapped over 540 deg^2 of sky, or 10.9 Gpc^3 in the targeted redshift range.

HETDEX is unlike traditional LAE narrowband surveys that identify LAE candidates via a narrowband minus broadband color excess (e.g., L. L. Cowie & E. M. Hu 1998; J. E. Rhoads et al. 2000; C. Gronwall et al. 2007; M. Ouchi et al. 2008; N. Benitez et al. 2014; A. Konno et al. 2016; D. Sobral et al. 2018; D. Spinoso et al. 2020; Y. Ono et al. 2021). These narrowband LAE surveys can effectively cover large areas at a slice of redshift space, but lack line-of-sight volume information; although, multiple narrowband filters can increase the survey volume and/or redshift resolution (M. Eriksen et al. 2019; S. Bonoli et al. 2021; K.-S. Lee et al. 2024). With an IFU, not only is the redshift coverage much wider and more continuous, but the effective bandpass for emission-line detection is generally 10–50 times narrower, enabling deeper LAE detections in shorter exposures and with higher-redshift precision. Searches for LAEs have proved successful at fainter sensitivities (C. van Breukelen et al. 2005; J. J. Adams et al. 2011; R. Bacon et al. 2015; T. Urrutia et al. 2019) but not over large volumes. HETDEX trades line-flux sensitivity for volume.

With only two to three LAEs detected in a typical VIRUS IFU observation, they make up a minuscule fraction of the total spectral data obtained by HETDEX. For example, a typical LAE's signal is spread across 3–15 fibers, depending on observational quality, and the spectral signature is typically only 4–8 Å or 3–5 pixels, in the spectral dimension. An IFU consists of 1344 fiber spectra with spectral coverage from 3500–5500 Å; thus, LAEs are <0.01% of the signal in an HETDEX IFU data cube, with HETDEX IFUs observing far more empty sky than LAEs. The remainder of the IFU fiber coverage is filled with hundreds of thousands of other astronomical sources, along with statistical signatures of large-scale unresolved Lyα emission (M. Lujan Niemeyer et al. 2026) and absorption (M. Mirza Khanlari et al. 2025; L. H. Weiss et al. 2025).

In the context of the number of spectra collected, HETDEX is currently the largest spectroscopic sky survey ever conducted. The main survey was completed in 2024 July; its database consists of over 600 million observed fiber spectra of the night sky obtained from 2500 dark sky observing hours. After being interpolated and cleaned for various quality issues, this dataset comprises 6778 observations and 431,713 HETDEX IFU data cubes. Figure 1 shows the distribution of HETDEX sources as a function of R.A. and cosmological comoving distance, collapsed in decl. The diagram illustrates

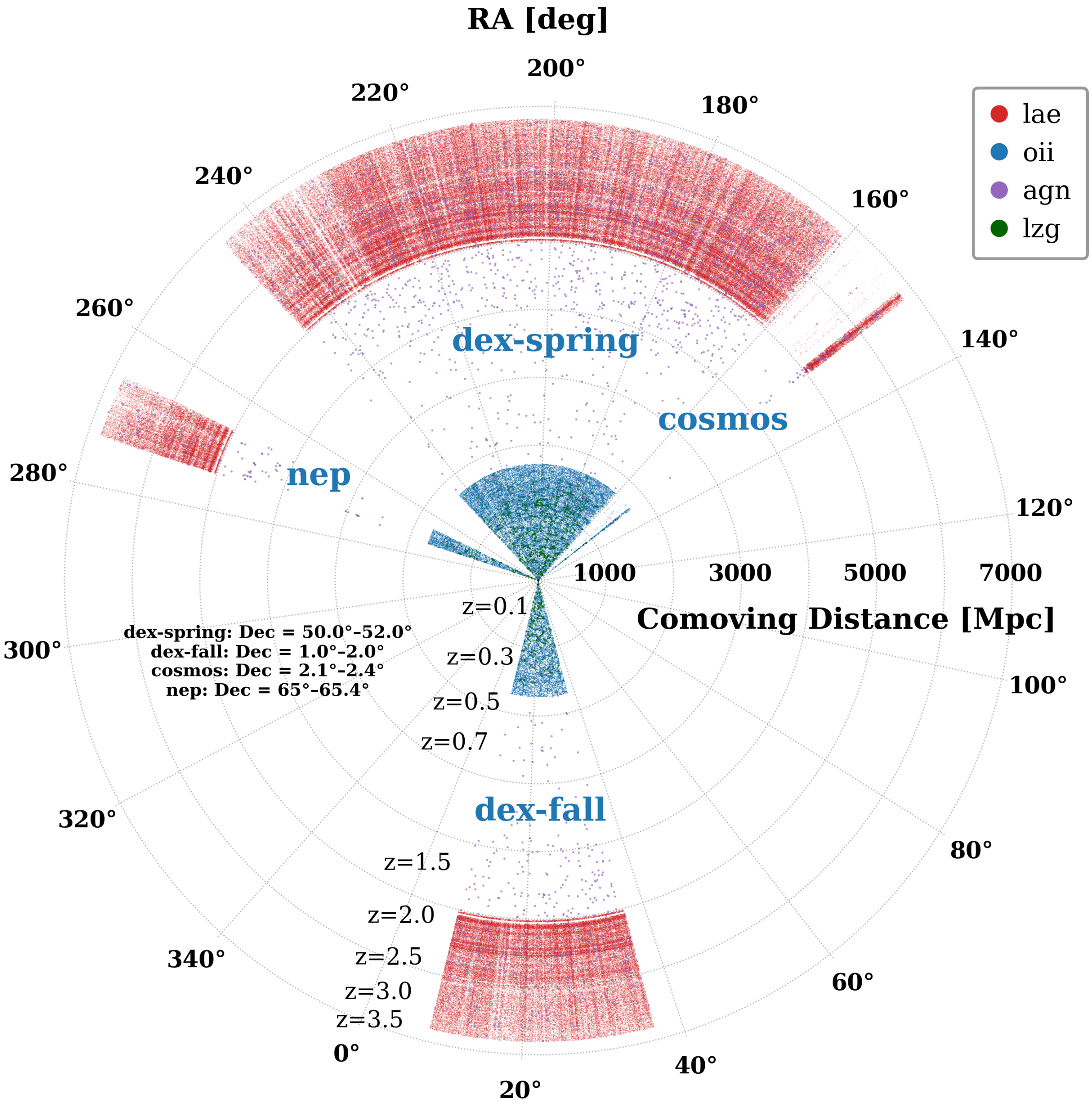


Figure 1. Projected distribution of Lyα emitters ("lae," red), active galactic nuclei ("agn," purple), [O II] emitters ("oii," blue), and low-redshift, non-emission-line galaxies ("lzg," green) in comoving space, collapsed in decl. Radial labels indicate comoving distance in Mpc, with corresponding redshift ticks annotated along the wedge. Clustering is evident at low redshift among [O II] emitters and passive low-z galaxies, while the distribution of the higher-redshift LAEs appears relatively smooth. The variation in LAE number density reflects the effect of coverage variation, wavelength-dependent flux sensitivity, and an intrinsic decline toward higher redshift. Only AGNs are found in the region between $0.5 < z < 1.9$, identified from their bright continuum emission and broad spectral lines in C. Liu et al. (2025).

the survey's ability to map large-scale structure across several Gpc, with high-redshift Lyα emitters (red) dominating the volume at $z \gtrsim 2$, while [O II] emitters (blue), active galactic nuclei (AGNs, purple), and luminous low-redshift galaxies (green) trace the foreground.

We present the first full public data release (PDR1) from the HETDEX main survey, providing access to over 430,000 science-quality IFU data cubes across the survey footprint. The paper describes the PDR1 data model, IFU data products, and methods for accessing and manipulating data cubes, with example tutorial notebooks. This paper also serves as a reference for the second HETDEX public source catalog (HPSC2) for HETDEX, which is an expansion of the early HETDEX Public Source Catalog 1 (HPSC1) described in E. Mentuch Cooper et al. (2023). Future data releases may consist of updated mask modeling, revised flux calibration, and/or expanded data from post main survey operations.

The outline of this paper is as follows. Section 2 describes the HETDEX observations, the generation of the IFU data cubes, and details about the data masking. Section 3 describes the catalog accompanying this release, HETDEX Public Source Catalog 2. Section 4 outlines the data model for the public data release and describes the format of the index and data cube files. In Section 4.4, a description of access options is provided, including a list of tutorial Jupyter notebooks, although access is not limited to the Python computing language, as the data cubes follow standard FITS file format (D. C. Wells et al. 1981; W. D. Pence et al. 2010). Section 5 describes the limitations of this data release.

HETDEX public data products are found at the data mount hosted by the Texas Advanced Computing Center (TACC) at The University of Texas at Austin at https://web.corral.tacc.utexas.edu/hetdex/HETDEX/pdr/. Additional information and data hosting mirrors are provided at https://hetdex.org/data-results/. Example access using Python can be found at https://github.com/HETDEX/dexcube. Notebooks within this repository demonstrate remote access for users who wish to download data cubes of interest and provide tutorials showing how to access source catalogs, extract spectra, and generate narrowband images directly from the data cubes.

All positions reported in this paper are in the International Celestial Reference System (ICRS; epoch J2000). We adopt the flat Lambda cold dark matter cosmology with $H_0 = 67.7$ km s^{-1} Mpc^{-1} and $\Omega_{\rm m,0} = 0.31$ measured by Planck Collaboration et al. (2020). All magnitudes are expressed in the AB system (J. B. Oke & J. E. Gunn 1983). We assume a rest-frame vacuum wavelength of $\lambda = 1215.67$ Å for Lyα and rest-frame air wavelength of $\lambda = 3727.8$ Å for the [O II] doublet, integrated to our instrumental resolution. Observed wavelengths expressed in this paper and associated data products are as measured in air. Redshifts are calculated from rest-frame wavelengths with conversions from air to vacuum (E. W. Greisen et al. 2006) and include a radial-velocity correction to the solar system barycenter (Astropy Collaboration et al. 2013). All spectral data are provided in air in the wavelength range from 3470–5540 Å in steps of 2 Å.

2. Observations

2.1. Survey Overview

The data cubes and catalogs released in PDR1 consist of the full main HETDEX survey (K. Gebhardt et al. 2021), drawn from the fifth internal data release (HDR5) from the HETDEX Collaboration. It contains 6778 observations obtained starting in 2017 January, when the VIRUS IFU assembly contained just 16 operational IFUs, and concluding on 2014 July 31 with the full 78 IFUs installed within the VIRUS array in 2021 July.

The HETDEX footprint consists of two primary fields that allow for full-year surveying as shown in Figure 2. The HETDEX-Spring Field, labeled as the `dex-spring` field throughout this paper and in the associated catalog, covers 390 deg^2 of high-decl. ($\delta \sim 51°$) sky, while the HETDEX-Fall Field, labeled as `dex-fall` in the catalog, covers 150 deg^2 along the celestial equator (see K. Gebhardt et al. 2021 for full details on field selection). The project does not need complete coverage within this sky area to accomplish its scientific goals, as discussed by C.-T. Chiang et al. (2013); a fill factor of 0.22 (1/4.6), which optimizes the number of IFUs given the area of the focal plane of the HET, is sufficient.

In addition to the two main survey fields, this HETDEX data release also includes coverage from multiple legacy fields, as presented in Figure 3. The most extensive coverage comes from collaborative observations with the Texas Euclid Survey for Lyα (Ó. A. Chávez Ortiz et al. 2023) of the north ecliptic pole (NEP). Nearly full field coverage of the central 1 deg^2 of the Cosmic Evolution Survey survey (COSMOS; N. Scoville et al. 2007) is included, as well as sparse coverage of the SA22 (C. C. Steidel et al. 1998) and GOODS-N (M. Dickinson et al. 2002) fields. Field coordinates, coverage area, and the number of IFU observations contained in each field are summarized in Table 1.

HETDEX uses VIRUS (G. J. Hill et al. 2021), the fiber-fed, multispectrograph instrument of the upgraded Hobby–Eberly Telescope (L. W. Ramsey et al. 1998; G. J. Hill et al. 2021). Each IFU connects to a pair of VIRUS spectrographs with 448 $1\farcs5$ diameter fibers positioned on a rectangular array with fiber center separations of $2\farcs5$. In a single HETDEX observation, three exposures, each typically lasting 6–7 minutes (the exposure times range from 3.6–12 minutes, depending on observing conditions), are captured; the telescope is dithered in a triangular pattern to ensure complete fill factor for each of the $51'' \times 51''$ IFU fields (see G. J. Hill et al. 2021; K. Gebhardt et al. 2021). The bottom-right panel in Figure 2 provides an example IFU fiber layout for this three-dither pattern. In a single IFU observation, 1344 fiber spectra are collected, providing full sky coverage of the IFU. At full completion of the VIRUS instrument, its 78 IFUs cover approximately 21.7% (a factor of 4.6) of the HET's $18'$-diameter FOV. The exact layout and numbering scheme for the array of IFUs is shown in Figure 3 in K. Gebhardt et al. (2021).

HETDEX data processing is described in detail in K. Gebhardt et al. (2021). Briefly, reduction begins with CCD-level calibrations, including bias subtraction, pixel flats, and twilight sky flats. These are followed by fiber extraction, wavelength calibration, and sky subtraction to produce a calibrated spectrum for each fiber in the array.

Sky subtraction is performed using a local sky model constructed at the amplifier level from fibers within each dithered exposure. This local sky subtraction is optimized for the detection of low signal-to-noise emission lines, which are the primary targets of HETDEX.

In addition to this approach, the internal HETDEX data model also supports a full-frame sky estimate constructed using all IFUs across the VIRUS focal plane within a given observation. While this full-frame model can better capture large-scale sky structure, it introduces low-level systematics in the calibrated spectra that can degrade the flux calibration on small spatial scales. For this reason, the public data products are calibrated using the local sky model. As a consequence, flux measurements near bright or extended sources, or on spatial scales comparable to the IFU FOV, may be affected by oversubtraction of extended emission.

Astrometric calibrations are achieved by measuring the centroid of each field star using the signal between 4400 Å and 5200 Å and comparing their IFU positions to the stars' equatorial coordinates in SDSS (D. G. York et al. 2000) and Gaia (Gaia Collaboration et al. 2018) catalogs. This process typically results in global solutions that are accurate to $\sim 0\farcs2$.

Absolute flux calibrations are achieved using $g < 24$ SDSS field stars as in situ flux standards. Their ugriz colors (N. Padmanabhan et al. 2008), Gaia parallaxes (Gaia Collaboration et al. 2018), and foreground reddening estimates (E. F. Schlafly & D. P. Finkbeiner 2011) are used to determine their most likely spectral energy distribution in a grid of model spectra (A. J. Cenarro et al. 2007; J. Falcón-Barroso et al. 2011). The final system throughput curve is derived from the most likely flux distribution of ∼20 stars, and is generally accurate to ∼5% (K. Gebhardt et al. 2021). Uncertainties in the calibrated fiber spectra are derived by propagating the photon

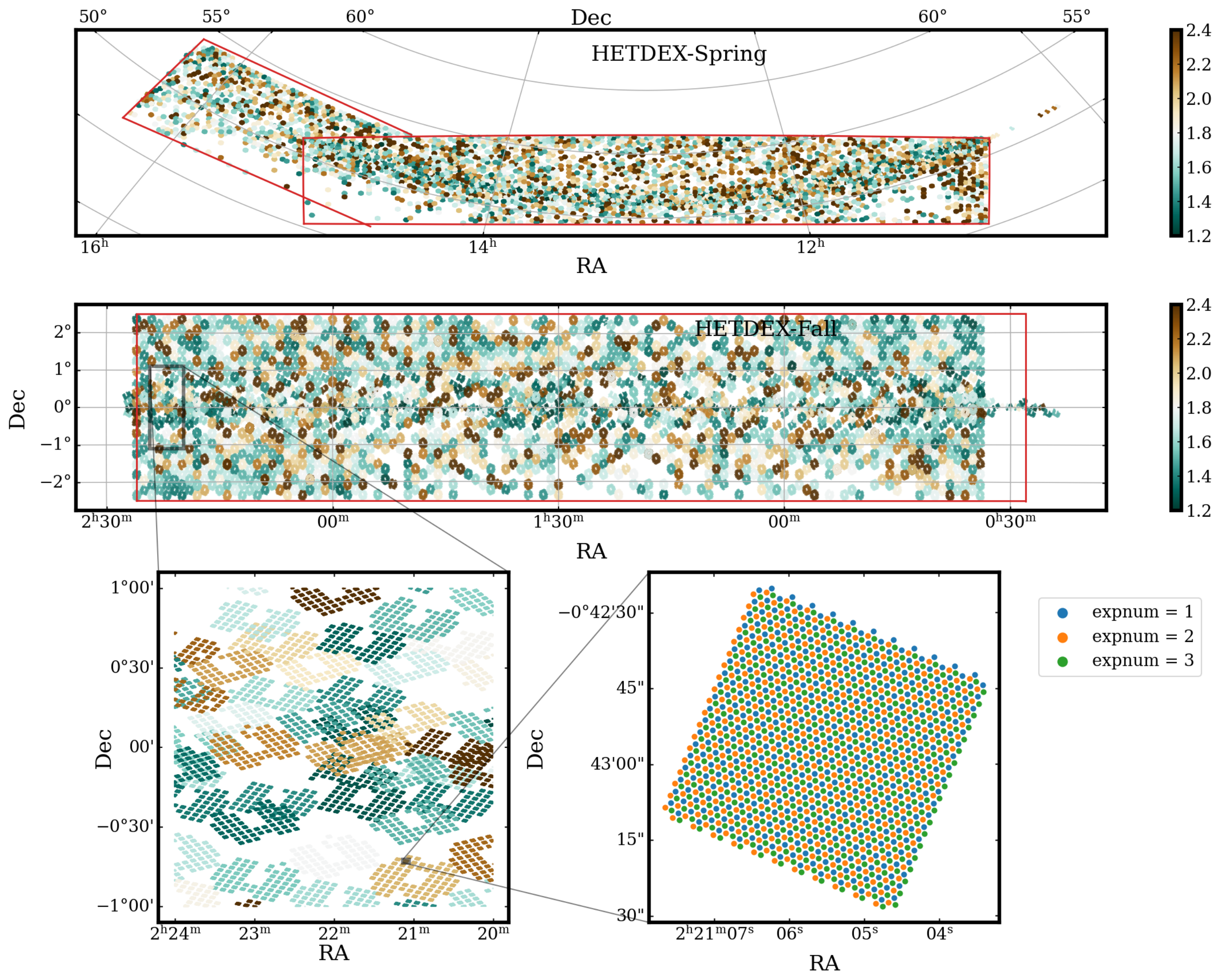


Figure 2. Outline of the two main HETDEX science fields (in red) and the footprint of IFU data cubes in this data release. Colors indicate the image quality of individual VIRUS IFU pointings, quantified as the FWHM of the point-spread function (PSF) in arcseconds. Fields are (1) the high decl. Spring field (top), which is centered at ($13^{\rm h}.5$, +51°) and covers ~390 deg^2 of the sky, and (2) the equatorial Fall field (middle), which is centered at ($1^{\rm h}.5$, 0°), and covers ~150 deg^2 of the sky. Each IFU has a field of view (FOV) of 51″ × 51″, which means that the full VIRUS IFU array has a 0.22 fill factor in the HET's 22′-diameter FOV. The expanded inset on the bottom-left presents a zoomed-in area on a typical 2 deg^2 region in the HETDEX-Fall field. The expanded inset on the bottom right shows a zoom in on a singular IFU indicating the array of fiber spectra, each 1″.5 in diameter. Each color represents one of the three dithered exposures. A single HETDEX observation can contain up to 78 IFU observations, consisting of 936 CCD exposures and 104,832 fiber spectra. An individual IFU three-dither observation of 1344 (448 × 3) fibers is used to create a single HETDEX data cube.

noise through the reduction steps and tracking the uncertainties that come with each step.

Each dither in an HETDEX observation is flux calibrated independently, as there may be small differences in their relative throughput due to variations in the observing conditions. For inclusion in the HETDEX survey, we require that a nominal throughput (assuming a 360 s exposure time) exceeds 08, and that the relative throughput of each dithered exposure cannot differ by more than a factor of 3. The most common reason for rejection by this criterion is a significant drop in transparency during the third dithered exposure when clouds drifted into the FOV.

The data are organized into a database of flux-calibrated, 1D fiber spectra each with their own corresponding sky coordinate. This forms the fundamental data product of the internal HETDEX data model. The internal data model stores not only the calibrated fiber data but also a number of intermediate data products including the raw and calibrated CCD images from each dithered observation. Internally, these intermediate data are used for quality control and artifact identification. Cutouts of these products are used in machine learning–based classification and vetting of detections (L. R. House et al. 2023, 2024; S. Mukae et al. 2026).

Internally, HETDEX data are stored in HDF5 files (The HDF Group 2024), with the primary data model constructed around a database of calibrated fiber spectra. Direct use of the fiber data is nontrivial: because the HET and is not equipped with an atmospheric dispersion corrector, each fiber samples a slightly different sky position as a function of wavelength. This atmospheric differential refraction (ADR) is corrected in

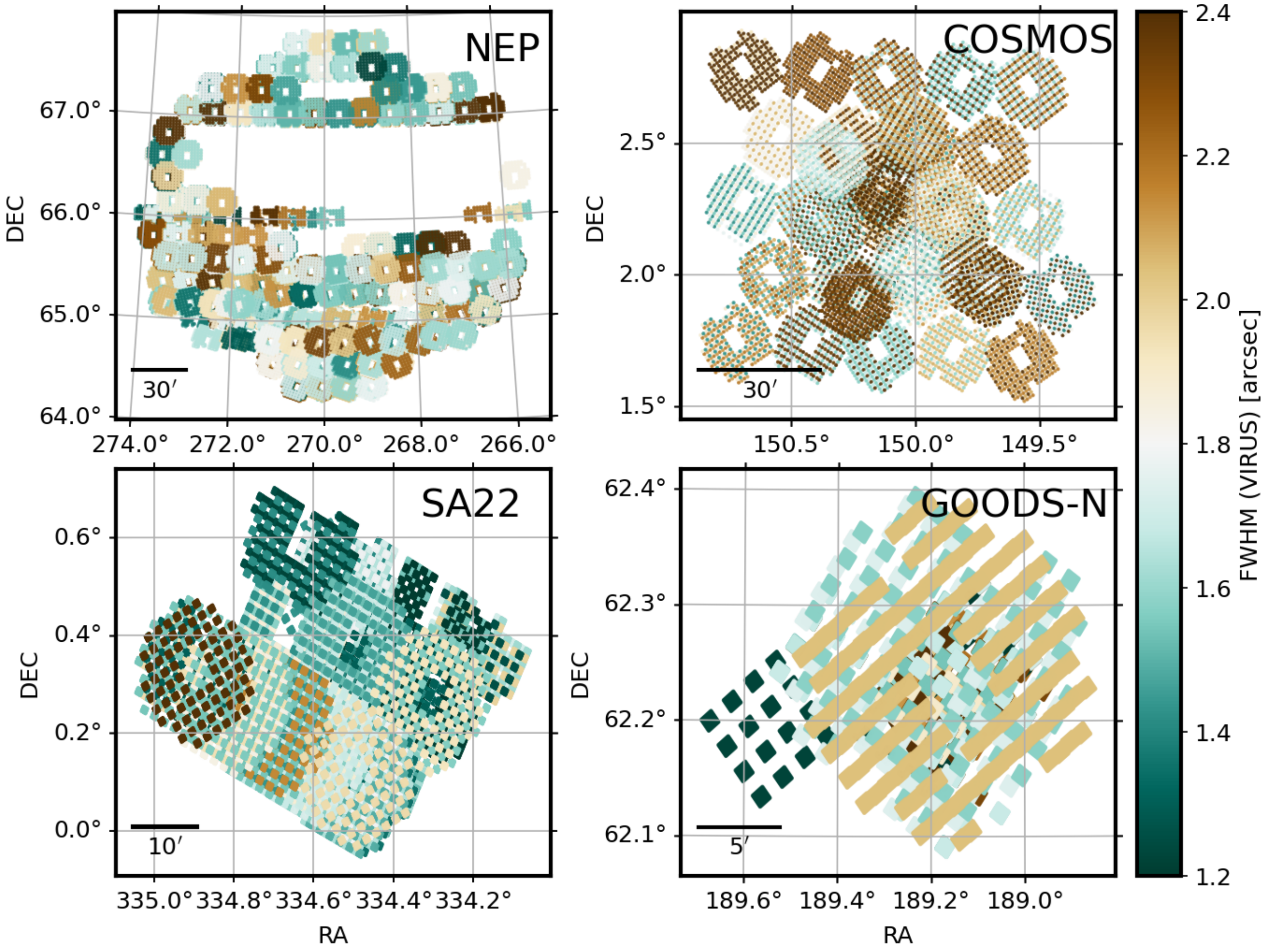


Figure 3. Coverage of the four legacy fields included in this release: NEP, COSMOS, SA22, and GOODS-N. Colors indicate the image quality of individual VIRUS IFU pointings (FWHM in arcseconds). The bar in each panel indicates the image scale. As in Figure 2, the IFU layout creates gaps in each observation that are most notable in the higher-resolution panels on the bottom row.

Table 1
Catalog Release Survey Statistics

Field ID	Center (deg, deg)	*N* (IFU)	Area (deg^2)	*N*(Source)	*N*(LAE)	*N*(OII)	*N*(AGN)	*N*(LzG)	*N*(Star)
dex-spring	(195°.00°, +51.00°)	244,176	49.00	640,603	253,458	287,017	11,093	10,807	77,522
dex-fall	(22.50°, +0.00°)	136,829	27.46	305,915	114,419	140,133	5,531	6,586	38,886
nep	(270.00°, +66.00°)	34,269	6.88	100,872	25,649	45,035	1,232	1,381	27,359
cosmos	(150.12°, +2.21°)	11,271	2.26	24,662	7,794	11,521	431	498	4,390
ssa22	(336.50°, +0.00°)	4,393	0.88	10,905	3,975	4,036	146	93	1,930
goods-n	(189.18°, +62.24°)	775	0.16	2,172	1,025	877	42	27	195
Total	⋯	431,713	86.64	1,085,129	406,320	488,619	18,475	19,392	150,282

Note. Listed is the number count of IFU observations observed and the count included after observation quality inclusion criteria. Reported area coverage includes repeat observations and therefore does not represent unique sky coverage. Field centers are given in equatorial coordinates (ICRS; epoch J2000).

the internal pipeline with dedicated software (HETDEX-API[31]), which is also required for subsequent interpolation and spectral extraction.

To simplify data access and avoid imposing these software dependencies, the public release is distributed as interpolated IFU data cubes rather than the native fiber database. This format also allows users to download only data cubes of interest, each of modest size (< 20 MB), and analyze the data using standard astronomical software tools.

In contrast to the internal release, the public release excludes the raw and intermediate data products. This decision is motivated primarily by the large data volume and complexity

---

[31] https://github.com/HETDEX/hetdex_api

Table 2
FITS HDU Structure for HETDEX IFU Data Cubes

HDU Index	EXTNAME	Description	Data Shape
0	PRIMARY	Primary HDU with no data	None
1	DATA	Flux data cube in units of 10^{-17} erg s^{-1} cm^{-2} per 2Å spectral bins	(1036, 104, 104)
...	...	Flux is sampled from 3470 Å to 5540 Å in steps of 2 Å	...
2	ERROR	1σ uncertainty estimates for each spaxel	(1036, 104, 104)
3	MASK	Bitmask cube indicating quality flags per spaxel	(1036, 104, 104)

of the internal data model: the full internal dataset, including raw CCD frames, calibrated products, and intermediate processing files, exceeds ∼350 TB, while the processed internal data products alone comprise ∼150 TB. By comparison, the public release of data cubes and catalogs requires <10 TB of storage. At present, we do not plan a full public release of the intermediate data products. However, we are happy to provide access to these data upon request for users interested in reproducing specific aspects of the internal processing or investigating instrumental effects in greater detail. The publicly released data cubes and catalogs are sufficient for the majority of scientific investigations with the HETDEX survey.

2.2. Data Cube Generation

The database of calibrated fibers is interpolated using custom software from HETDEX-API[32] into a set of IFU data cubes in a format that is similar to other IFU survey data from Calar Alto Legacy Integral Field Area (CALIFA; S. F. Sánchez et al. 2012), Mapping Nearby Galaxies at APO (MaNGA; K. Bundy et al. 2015; N. Drory et al. 2015; D. R. Law et al. 2015), and MUSE (T. Urrutia et al. 2019). As VIRUS is composed of an array of many IFUs with 1′ gaps between them, the natural unit of data cube storage is that of a single IFU. A given HETDEX exposure is thus composed of 16–78 IFU data cubes depending on how many units were active on the instrument at the time of observation. Each IFU is stored in an individual FITS (D. C. Wells et al. 1981) file with the format described in Table 2. The spectral sampling of each cube is set to 2 Å, the same as the spectral sampling of the internal fiber spectra. Pixel spatial sampling is set to 0.″5. With a median image quality of FWHM = 1.″8, this is sufficient to sample the point-spread function (PSF) of HETDEX observations while keeping storage requirements manageable.

The IFU data cubes cover the same spectral range and resolution as the internally calibrated fibers, sampling a linear wavelength grid from 3470 to 5540 Å in steps of 2 Å. For each wavelength slice, the apparent sky position of every fiber is adjusted to account for ADR using the astrometric solution from the internal pipeline (K. Gebhardt et al. 2021). The resulting irregularly spaced fiber positions are then resampled onto a 52″ × 52″ square grid with 0.″5 spatial sampling. Interpolation is performed with the `griddata` routine from SCIPY (P. Virtanen et al. 2020), which linearly interpolates the flux and associated uncertainties of all contributing fibers to the output coordinate grid. The WCS of each cube follows standard FITS conventions, with the CRPIXn values specifying 1-indexed reference pixel coordinates at pixel centers rather than pixel edges. The data cubes are oriented according to the position angle of the telescope and VIRUS focal plane to minimize any unnecessary data storage space. Thus, the projected orientation is not constrained to the conventional "north up, east left" display; users should rely on the WCS headers when visualizing the data.

The resulting data cube spaxels are per-bin flux values sampled in 2 Å wide spectral bins in units of 10^{-17} erg s^{-1} cm^{-2} per 2 Å bin. Flux values are stored in Header Data Unit 1 (HDU1), and their associated flux uncertainties are stored in HDU2. Spaxels that fall outside the region with valid fibers or within fully masked regions are assigned `NaN`. To prevent spurious interpolation across masked fibers, a Boolean coverage map is propagated by resampling the good-pixel indicator with nearest-neighbor interpolation and blanking any output spaxels where this indicator is false.

In addition, a bitmask array, described in detail in Section 2.3, is interpolated from the internal data model to the data cube wavelength slices following the same methodology as the flux interpolation except that, for the bitmask, the nearest-neighbor method is used for interpolation.

2.3. Masking Model

The calibrated fiber spectra in the internal HETDEX data model include a single mask for problems automatically identified by the reduction pipeline. This includes detector artifacts, poor-quality fibers yielding unusable spectra, transient cosmic rays, and various calibration failures. Such cases are flagged through a fill value of 0.0 in the calibrated fiber error arrays. While the fiber spectrum itself may still report a flux value, it is considered invalid when flagged in the error array.

In practice, given the complexity and sheer volume of HETDEX data, a wide range of previously undiscovered issues must also be masked from the data. Over the years, these have been identified through an extensive manual effort. In particular, data reports generated by the Emission Line eXplorer (ELIXER) tool developed by D. Davis et al. (2023a) have proved invaluable in diagnosing artifact features that are clear in the view of the raw CCD science frames but may not be obvious in the interpolated fiber spectrum. This manual inspection led to the development of a number of automated procedures to remove spurious artifacts from our source catalogs and fiber spectral data.

For each flux-calibrated spectrum, these masks are stored as bitmask values for every spectral element in the array. In the public data model, these bitmasks are interpolated into HDU3 of the data cube FITS files as listed in Table 2. The interpolation is performed such that the nearest bitmask value is stored in the cube spaxel.

Table 3 provides a reference for the bits used to flag various issues in the data. The bitmask format offers the compact representation for a large array of flagging options. A user can

---

[32] https://github.com/HETDEX/hetdex_api/blob/master/hetdex_tools/create_dex_cube.py

Table 3
Masking Description

Hex Value	Binary	Integer Value	Name	Description
0x00000000	0	0	Good	No flag
0x00000001	1	1	MAIN	Value flagged in reduction (calfibe= = 0.0)
0x00000002	10	2	FTF	Average fiber-to-fiber in spectrum is > 0.5
0x00000004	100	4	CHI2FIB	chi2fib > 150
0x00000008	1000	8	BADPIX	On a bad pixel region
0x00000010	10000	16	BADAMP	On a bad amplifier
0x00000020	100000	32	LARGEGAL	Located within a large galaxy mask
0x00000040	1000000	64	METEOR	On a known meteor track
0x00000080	10000000	128	BADSHOT	In bad shot list
0x00000100	100000000	256	THROUGHPUT	Relative response at 4540 Å < 0.08
0x00000200	1000000000	512	BADFIB	On a known bad fiber
0x00000400	10000000000	1024	SAT	On a known satellite track
0x00000800	100000000000	2048	BADCAL	Masking due to sky and calibration issues
0x00001000	1000000000000	4096	PIXMASK	Masking spectrum==0 in native spectrum counts
0x00002000	10000000000000	8192	BADDET	$5 \times 5 \times 5$ pixel mask where a detection has been flagged

select which flags they wish to include, or simply use the default setting where any nonzero value in the mask is flagged. In this section we provide additional detail on each bitmask flag, in the order they are encoded into the model.

2.3.1. MAIN (Pipeline Reduction Flags)

The most basic mask in the data model is inherited from the HETDEX data reduction pipeline (K. Gebhardt et al. 2021). Any fiber with a calibrated fiber error array value of zero (calfibe= = 0.0 in HETDEX's internal data model) is flagged, indicating that the pipeline determined the data to be unusable. This flag may occur for cosmic rays, readout glitches, or fibers where the calibration is poor, causing a specific spectral region of the fiber to be flagged. These defects are directly marked in the error arrays and carried through to the bitmask as the `MAIN` flag. It is strongly recommended to always use this mask.

2.3.2. FTF (Fiber-to-fiber Outliers)

This flag identifies spectra where the average deviation between neighboring fibers is anomalously high. Such cases indicate strong flat-fielding residuals or improperly calibrated regions, which can mimic emission lines when interpolated into data cubes. Masking these features ensures a more uniform background. It is recommended to always use this mask.

2.3.3. CHI2FIB (High χ^2 Fibers)

Spectral array elements in a fiber spectrum where the fiber profile is not fitted well by the fiber profile model are flagged by the CHI2FIB bitmask flag (indicated by $\chi^2_{\rm fib} > 5$). This cut identifies fibers whose calibration may result in spurious detections if left unmasked. Pixel elements in the spectral direction are masked, but the rest of the fiber spectrum remains unflagged. It is recommended to always use this mask.

2.3.4. BADPIX (Bad Pixel Regions)

This bit marks all fibers intersecting detector regions with persistent defects, dust spots, or charge traps that cannot be corrected by flat-fielding alone. As described in K. Gebhardt et al. (2021) and further in E. Mentuch Cooper et al. (2023), several detectors have significant features, including large dust spots, many charge traps, and a "pox" contamination where the quantum efficiency of individual pixels can be suppressed by 10%–40%. Issues such as these were especially prevalent for data taken during the first 2 yr of observations. While the flat-field calibrations identify many of the worst features automatically, many low-count defects remain in the data and can produce false-positive line detections. It is recommended to always use this mask.

2.3.5. BADAMP (Amplifier Variability)

Every VIRUS IFU is composed of 448 fibers, which are divided into two spectrograph channels. Each channel has a 2064 × 2064 CCD detector, which is read out by two amplifiers binned 2× in the spectral direction. Thus, each IFU consists of four amplifier ("amp" for short) channels, labeled "RU," "RL," "LL," and "LU," each with 112 fiber spectra that are individually stored in separate 1032 × 1032 raw data arrays and processed independently within the HETDEX reduction pipeline. Our standard three-dithered observation set generates 936 FITS files (three dithered observations consisting of 312 amplifier readouts for the 78 IFUs). Although the IFU spectrographs are designed to be identical, the "R" in VIRUS stands for "replicable" after all; in practice, there are a number of issues that alter the consistency and quality of the fibers on any given amplifier. Some examples include low photon counts, electronic interference, scattered light, and poor calibration due to bright stars and/or galaxies.

The PDR1 dataset contains 5,180,556 amplifier images. Over time, a combination of statistics that are calculated either during the reduction process or in post-reduction provide a good assessment on whether to reject the data produced by an amplifier's readout. Please see Table 2 and related discussion in E. Mentuch Cooper et al. (2023) for more details on criteria used to flag bad amplifiers. Roughly 92% of the FITS files are of the necessary science quality to detect the faint signature of emission-line galaxies. Those that are not are flagged by the BADAMP bitmask.

A conservative approach is taken when assigning automated criteria to identify poor-quality amplifiers. This decision is based on our desire to mitigate false positives in the LAE sample. It is likely that a small fraction of flagged amplifiers are of science quality for brighter sources. If a user plans to visually inspect their data, then this flag can be ignored.

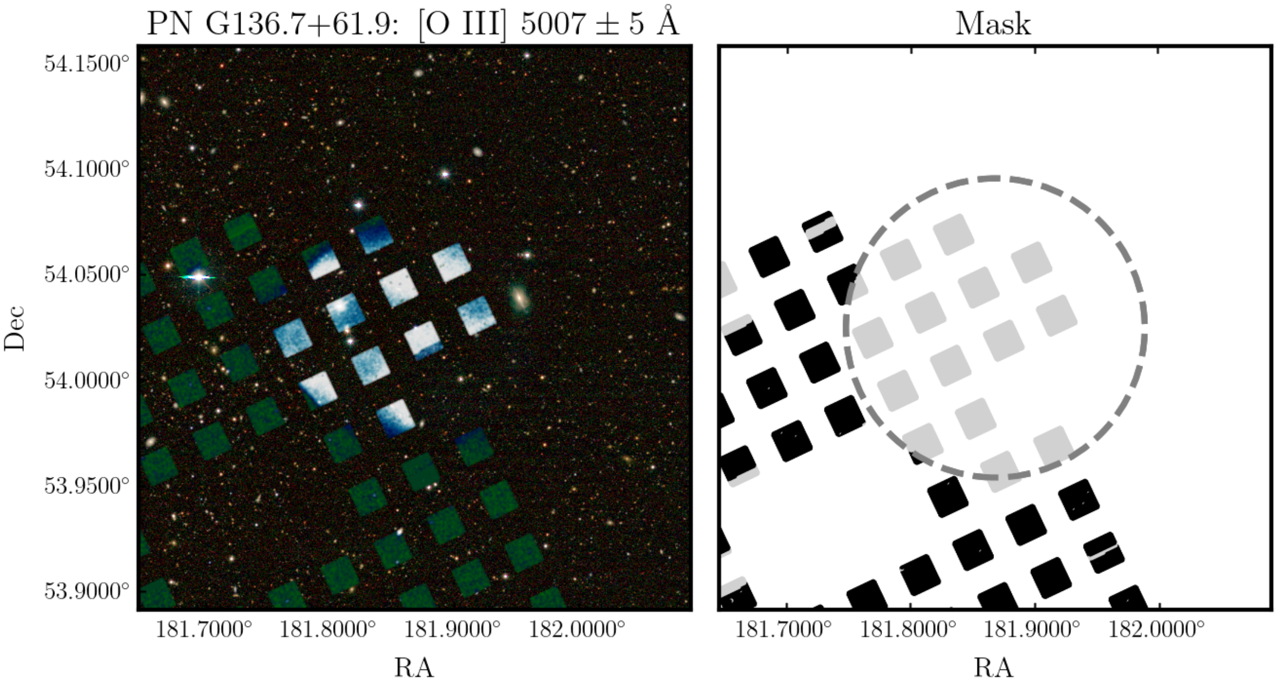


Figure 4. This example demonstrates how large extended sources, such as the planetary nebulae PN G136.7+61.9 displayed here, are identified and masked to prevent false emission-line detections in the catalog and data cubes. Left panel: collapsed data cube highlighting the [O III] 5007 ± 5 Å emission-line flux over an 8′ field centered at the planetary nebula coordinates (R.A.: 181°. 8685, decl.: 54°. 0248). The RGB background image is constructed from Legacy Survey data (A. Dey et al. 2019). Right panel: visualization of the mask for the same region, where gray regions indicate masked HETDEX data, and black indicates nonmasked data. The gray dashed ellipse represents the sky region masked from the HETDEX fiber data when the `LARGEGAL` bitmask is set (as described in Section 2.3.6), PNe are treated similar to large galaxies for masking purposes.

Table 4
Planetary Nebulae in HETDEX PDR1

PNG ID	Coordinates (R.A., Decl.)	shotid	r (arcmin)
PNG 085.3+52.3	(230.°444, 52.°368)	20210510014	4.08
PNG 085.3+52.3	(230.°444, 52.°368)	20220131013	4.08
PNG 144.8+65.8	(179.°437, 48.°938)	20240314023	3.00
PNG 144.8+65.8	(179.°437, 48.°938)	20240403008	3.00
PNG 136.6+61.9	(181.°868, 54.°025)	20210309019	2.83
PNG 136.6+61.9	(181.°868, 54.°025)	20210401013	2.83

2.3.6. LARGEGAL (Large Galaxy and Planetary Nebula Masks)

We mask bright galaxies whose optical diameters D_{25} (measured at the 25 mag arcsec^{-2} isophote) exceed 1′. Their locations and shapes are provided by the Third Reference Catalog of Bright Galaxies (RC 3; G. de Vaucouleurs et al. 1991) and the Uppsala General Catalog of Galaxies (UGC; P. Nilson 1973). The Spring field contains 644 such galaxies; the fall field, 447. All detections that fall within 1.5 × the D_{25} scale of a bright galaxy's elliptical isophote are removed. This factor was determined by examining the HETDEX spectra at different scalings and ensuring that all detections related to the bright galaxy were encompassed in the aperture mask. These galaxy masks are consistently applied to the HETDEX Source Catalogs and the PDR1 data cubes.

Multiple planetary nebulae (PNe) were identified in the dataset due to an excess of bright emission lines found in the emission-line detection search. An example of a resolved PN is shown in Figure 4. Spanning several arcminutes in diameter, the PN G136.7+61.9 is seen in emission at 5007 Å across several VIRUS IFUs. We mask these in a similar manner as the large galaxy mask. The masked regions are listed in Table 4, with their central coordinate information and angular sizes provided by the HASH PN catalog[33] from Q. A. Parker et al. (2016). Note that all three PNe extend over multiple HETDEX observations due to their large angular extent. Flagged detections are removed from our catalog and masked as circular regions similar to the large galaxy masking.

Users interested in studying these masked galaxies or PNe can ignore this flag. Similarly, there are background sources that can be seen in the outskirts of these galaxies that are masked conservatively at 1.5 × the isophotal sizes. With visual inspection, this mask can be ignored, but the calibration may be off in the vicinity of bright sources. Users ignoring this flag should proceed with caution.

2.3.7. METEOR (Meteors)

Meteors crossing the focal plane generate spatially extended line emission unassociated with fixed astronomical sources. These tracks are identified in two stages. Then for each shot, exhibiting many of these detections, we manually fit the locations with a linear fit to create a meteor mask.

The first search method comes from the ELIXER classifying software tool. Strong line emission appearing in just a single dithered HETDEX observation, notably associated with Mg, Al, Ca, and Fe, is flagged as a meteor candidate. The second search method is performed using AI software from RAIC Labs,[34] which uses Rapid Automatic Image Categorization (RAIC) software to quickly generate a labeled dataset that can

---

[33] https://hashpn.space

[34] https://raiclabs.com/

be trained to classify test data. Streaks in emission-line maps generated by meteor detections are easily identified using this software. We discuss this method in greater detail later in Section 3.7.1. We use the meteor labels generated by both ELIXER and RAIC to identify meteor detections in the HETDEX catalog.

For any observation with more than 10 such detections, we visually inspect the data to confirm the presence of a meteor, then create a simple linear mask by fitting to the positions of the flagged detections. This mask extends 12″ above and below the linear fit; while a narrower mask could suffice in some cases, this conservative width ensures that the brightest events are fully removed. The linear mask is consistently applied to both the line emission and continuum-emission raw catalogs and is masked in the PDR1 data cubes at all wavelengths under the METEOR bitmask.

In total, 96 meteors have been detected so far in HETDEX PDR1 observations, with the streaks often spanning several IFU data cubes. A list of the affected observations and their linear fits can be found in the HETDEX-API GitHub repository,[35] but this information is not needed for PDR1 users, as meteors are easily identified using the METEOR bitmask. The bitmask is activated at all wavelengths in the fiber spectra for regions that contain meteors. We do note that in most cases, meteor emission affects only a limited region of our spectra. However, the compromised wavelengths are not consistent and are dependent on a number of properties of the meteor itself (e.g., chemical composition, size, altitude) that is beyond the scope of this work.

While only a small amount of data is flagged by the meteor masking (∼0.04%), the affect on line-emission detections is significant. A single meteor can create several hundred emission-line candidates. Without the bitmask, a meteor streak without any associated continuum emission appears as very bright and strongly clustered LAE candidates. It is, thus, very important to remove these features for HETDEX cosmological analysis.

Users are recommended to always apply the METEOR bitmask unless they are either (1) studying the meteor spectra or (2) manually masking the affected emission lines.

2.3.8. BADSHOT

Entire exposures (“shots”) may be deemed unsuitable for analysis due to poor observing conditions, failed guiding, or instrumental issues. A list of such exposures is maintained internally, and fibers from these shots are flagged. This bitmask is not relevant for the public data release, as these observations are not included in the public data products. The bitmask flag remains for consistency with the internal data model structure.

2.3.9. THROUGHPUT

Fibers with an effective throughput response at 4540 Å below 0.08 (8%) are flagged within the internal HETDEX data model. PDR1 data has all low throughput data removed prior to creation, so this is not applicable to the public data products. It is retained for consistency.

2.3.10. BADFIB (Known Bad Fibers)

Fiber spectra are deemed unfit for scientific use for multiple reasons. Some may have too low a throughput, while others fail fiber profile fitting. Additionally, a small number of fibers create a surplus of spurious detections at multiple wavelengths; these are poorly handled by our noise model. These fibers are masked from the catalogs and are masked within the data cubes with the BADFIB flag. Less than 0.2% of fibers are flagged overall, but these are important to remove, as some can create multiple emission-line artifacts that can introduce both spatial clustering between HETDEX tiled observations and line-of-sight source clustering in the spectral dimension. It is recommended to always use this mask.

2.3.11. SAT (Satellites)

Satellites trails are a significant source of continuum contamination, with approximately one-third of each HETDEX observation being affected by a satellite in one of its dithers. Though emission-line contamination caused by satellites is less common, it can still occur. Masking these regions and accounting for the lost survey volume is important, as these streaks and the increase in continuum they produce results in the loss and/or misidentification of LAE candidates in the areas of the sky they traverse.

Satellites are identified using classifications generated by RAIC Labs AI classification platform (see Section 3.7.1 for more details). In spectrally collapsed images at the locations of HETDEX continuum detections, satellite contamination appears as a streak. When more than five satellite detections are found, a linear model is fit to the positions of all point-source satellite detections in an HETDEX observation. All detections and fibers within ±6″ of the fitted line are masked.

A satellite will only contaminate a single dithered observation, but we mask data in all dithers, since HETDEX detections are based on the three-dither combined data. In PDR1, we mask a total of 546 satellite streaks. In some cases, some HETDEX observations contain two to three satellite streaks. A list is maintained in the HETDEX-API github repository[36], but users can mask out the streaks using the SAT bitmask option.

Several fainter satellites, beyond those currently masked, are still evident when visually inspecting spectrally collapsed IFU data cubes. These objects are missed, as they do not contain many continuum detections related to the streaks, and so are not found with the current detection algorithms.

The satellite mask is conservative, often masking more regions than is needed, so a user can consider not using this flag, provided that they include visual inspection.

2.3.12. BADCAL (Calibration Residual Issues)

This bit identifies wavelength regions where strong sky lines, calibration errors, or other systematic issues result in unreliable flux calibration. These wavelength-dependent masks reduce spurious emission-line detections in these contaminated regions. There are three spectral windows that are flagged with this bitmask. The latter two windows are not always flagged and are observation and amplifier dependent.

---

[35] https://github.com/HETDEX/hetdex_api/blob/master/known_issues/hdr3/meteor.txt

[36] https://github.com/HETDEX/hetdex_api/blob/master/known_issues/hdr3/satellite_tracks.txt

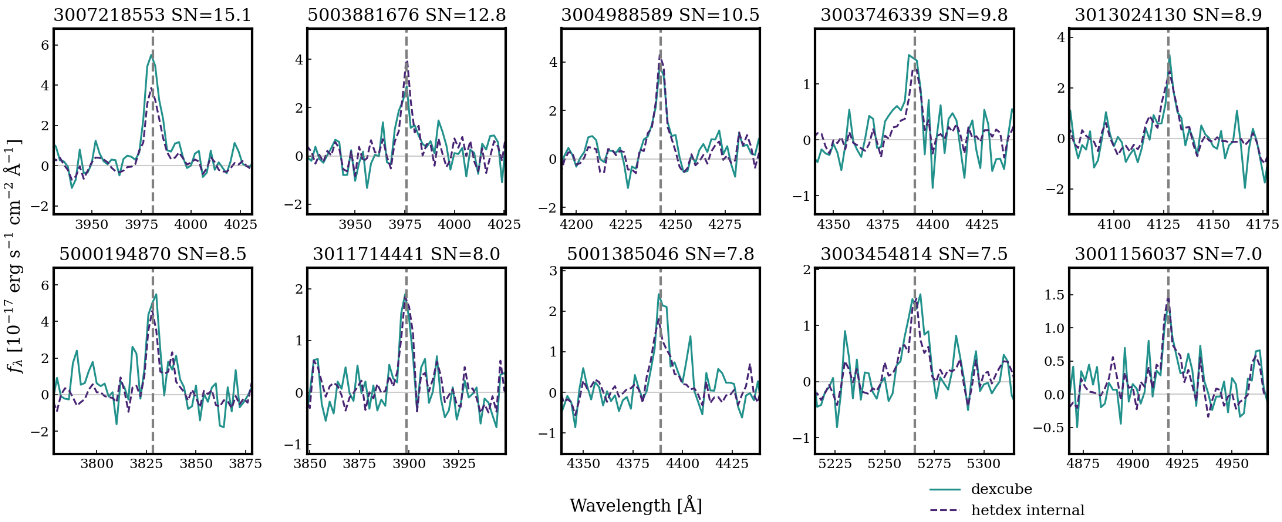


Figure 5. Comparison between aperture-corrected, extracted spectra of LAEs (with their emission wavelength indicated by the gray dashed vertical line) from the internal data model (shown with the purple dashed line) and the public data model (shown with the blue solid line). The internal spectra come from summing the flux in PSF-weighted aperture, while the dexcube spectral extractions are created from circular apertures with a radius of $3\farcs5$. Interpolation and differences in aperture definitions account for minor difference between the two, with a moderate 5%–10% difference in the blue region of the spectrum.

1. 3534−3556 Å: this sky line is difficult to fully remove and is masked in every observation under the BADCAL bitmask.
2. 5194−5197 Å and 5200−5205 Å: Although these two sky regions are often corrected for through flat-fielding, a sharp spectral feature often persists and can create false emission lines. The bitmask is set for this region for the shotid/amplifiers listed in the HETDEX-API github repository.[37]
3. 5456−5466 Å: This sky line can be difficult to fully remove. The bitmask is set for this region for the amplifier/shotid combinations also contained on `hetdex_api` repository.[38]

This bitmask can be ignored for analysis performed on bright objects like stars, nearby galaxies, most AGNs, meteors, and satellites.

2.3.13. PIXMASK (Spectrum==0)

Spectral regions in which the raw spectral data are set to zero are also flagged. These wavelength-dependent pixel masks ensure that no spurious flux is interpreted from missing data. The bitmask is expanded ±1 pixel in the spectral dimension to minimize low signal-to-noise ratio (S/N) false-positive detections. It is recommended to always use this wavelength-dependent bitmask.

2.3.14. BADDET

In Section 3.7, we describe targeted machine learning methods used to identify and flag catalog detections classified as artifacts. The majority of these artifacts are already masked in the HETDEX data cubes through the procedures described above; however, roughly 10% (65,078 detections) remain unaccounted for.

To mitigate these remaining artifacts, we provide a cubic bitmask for detections flagged by machine learning methods or other manual inspection efforts. For these sources, we apply a $5 \times 5 \times 5$ pix mask (corresponding to $2\farcs5 \times 2\farcs5 \times 10$ Å in spatial and spectral dimensions) centered on the detection coordinates and emission-line wavelength (`RA_det`, `DEC_det`, `wave`).

These objects do not appear in the source catalog but can be found in the raw detection database (included in this release and described in Appendix A). We recommend applying this mask when performing independent emission-line searches. However, if extracting spectra at known object positions and performing visual inspection, this mask can be ignored.

2.4. Spectral Calibration Comparisons

The internal HETDEX data model is shown to be well matched to SDSS spectra (K. Gebhardt et al. 2021; E. Mentuch Cooper et al. 2023). In this section, we perform two tests to ensure that the extracted spectra from HETDEX data cubes are consistent with the internal model and with SDSS external spectral data.

2.4.1. Comparison with Internal Pipeline Spectra

HETDEX pipeline spectra are composed of spectral extractions using several fibers and adding their flux values based on PSF-weighted extraction (K. Horne 1986; K. Gebhardt et al. 2021). For point-source objects, this will be well approximated by a circular aperture extraction. In Figure 5, we provide examples of LAE spectra, focusing on a 100 Å window surrounding their emission-line wavelength. These panels display examples for sources with decreasing in S/N and compare spectra generated by the HETDEX pipeline in dashed purple to extracted circular ($r = 3\farcs5$) aperture spectra generated by the data cubes. We see that for these emission-line sources, the spectra between the public and internal model are well matched.

---

[37] https://github.com/HETDEX/hetdex_api/blob/master/known_issues/hdr3/flag5200.txt

[38] https://github.com/HETDEX/hetdex_api/blob/master/known_issues/hdr3/flag5460.txt

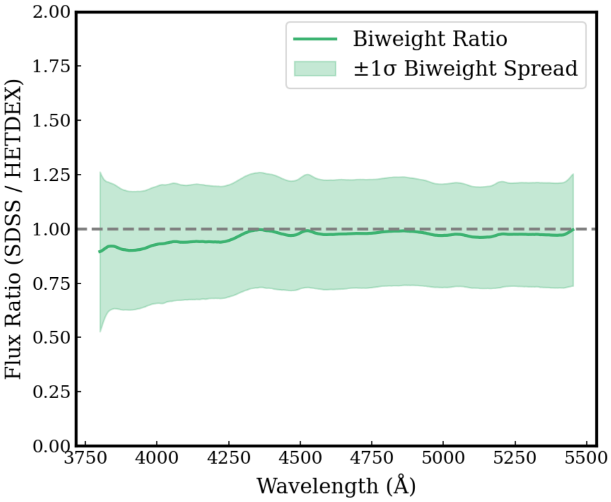


Figure 6. Biweight average flux ratio spectrum comparing SDSS and HETDEX observations of 8076 AGNs. Each spectrum was Gaussian-smoothed (σ = 5 pixels) before computing the flux ratio SDSS/HETDEX at each wavelength. The solid green line shows the biweight location of the flux ratio across all matched sources, while the shaded region indicates the $\pm 1\sigma$ biweight spread after sigma-clipping outliers at each wavelength bin. The dashed line at ratio = 1 indicates perfect flux agreement. Overall, the SDSS and HETDEX spectra show good agreement.

2.4.2. Comparisons with SDSS

We consider spectral comparisons between a sample of 8076 AGNs that are found in both HETDEX and SDSS in Figure 6. Spectral extractions are performed on the HETDEX IFU data cubes with 3.″5 circular apertures. The median SDSS/HETDEX flux ratio is consistent with unity with a scatter in the agreement of under 20%. A modest offset of <10% is evident in the blue region on the spectrum.

3. Catalog

HETDEX Public Source Catalog 1 (HPSC1; E. Mentuch Cooper et al. 2023) describes the generation of the first public source catalog taken from data that is a subset of this release (early data spanning from 2017 January 1 to 2020 June 26; internally referred to as HDR2.1). Included in PDR1 is an update to this catalog, titled HETDEX Public Source Catalog 2, which contains four additional years of data and incorporates several updates to masking, source selection and false-positive mitigation. Hereafter, we refer to this catalog as HPSC2.

Details of generating the catalog are described in HPSC1, but brief descriptions are included here for ease and clarity. Source detection, data quality selections, redshift assignment, and the derivation of resolved [O II] fluxes for low-redshift galaxies are applied consistently in both catalogs. Given the considerable updates to data quality assessment and detection masking (see Section 2.3), we outline the differences between HPSC1 and HPSC2 throughout this section.

As issues have been found in HETDEX data and the masking model has developed, we aimed to maintain consistent masking of both HETDEX detections and calibrated fiber data in our data model. However, we have also investigated and found success using multiple means of classifying and cleaning catalog sources through citizen science, also known as participatory science, and through multiple machine learning/artificial intelligence (ML/AI) classification approaches. We discuss these issues in Section 3.7.

Upon the identification of systematic issues through these methods, such as a pixel defect in the detector, a faulty fiber, or a meteor streak, we can reliably mask the problem at both the fiber spectrum (and interpolated data cube) and detection level. For transient features that could not be masked through an automated method, we mask at the detection level through the BADDET mask (described in Section 2.3.14). These include horizontal readout artifacts caused by by cosmic rays, calibration problems due to scattered light, and residuals produced sky lines detected by the different classification methods.

Before getting into the details, we provide a few HETDEX catalog concepts to highlight:

1. HETDEX source detection (see Section 3.1) is performed under a point-source assumption. Fluxes and extracted spectra are reported under this assumption unless otherwise stated. Detections that are within 3″ of each other and 3 Å in spectral separation are reduced to a single detection in the HETDEX pipeline. For low-redshift, resolved galaxies, we supplement the pipeline fluxes and provide an aperture flux measurement for the [O II] emission line (see Section 3.5).
2. `detectid`: this is a unique integer identifier from the raw detection pipeline. Nine digits long, the first digit signals the internal release database of the detection. The third digit indicates whether the detection is a line-emission detection (third digit = 0), or a continuum detection (third digit = 9). For example, the detection 5090045587 is from the internal HDR5 data release, and the “9” in 509 indicates that it is a continuum detection.
3. `source_id`: a single astronomical source can comprise many `detectids` if it is spatially resolved or if it has multiple emission lines. These detections are grouped together (as described in Section 3.3) to make one source, identified by the integer `source_id`. If the same astronomical source is observed in multiple observations, it is listed multiple times in the catalog and will have a different `source_id`. It is up to the user to combine the data from repeat observations into a single source.
4. `z_hetdex` and `source_type`: We assign source classifications and redshifts using three methods: continuum template fitting with DIAGNOSE (M. H. Debski et al. 2025; M. Debski & G. Zeimann 2024), probabilistic emission-line identification via ELIXER (D. Davis et al. 2023a), and secure AGN redshifts from dedicated AGN catalogs (C. Liu et al. 2022, 2025). See Section 3.4 for a full description. This approach achieves an overall redshift accuracy of ∼96% compared to external spectroscopic catalogs. HETDEX sources are grouped into five `source_types`: “lae,” “agn,” “oii,” “lzg,” and “star” (see Section 3.4.1 for definitions).

3.1. Object Detection

Two independent, but complementary, object-detection search techniques are performed as part of the main HETDEX reduction pipeline: one to identify emission lines, the other to detect continuum sources. In the fifth internal data release for

HETDEX (HDR5), a search was performed across more than $>6 \times 10^8$ flux-calibrated fibers. The search method details can be found in K. Gebhardt et al. (2021). We briefly summarize the procedures here.

During this process, no imaging preselection is used; the HETDEX data itself provides object detection. Both the emission-line and continuum detection algorithms are designed to identify point sources and account for the variable image quality, or PSF of each independent HETDEX three-dither observation.

The raw detections are stored in HDF5 files. We include these files in the PDR1 data release and describe their content and access in Appendix A.

3.1.1. Emission-line Detection

Line emission is identified using a two-stage Gaussian fitting procedure. An initial grid search steps through spatial ($0\farcs5$) and spectral (8 Å) dimensions, fitting a fixed-width Gaussian ($\sigma = 1.7$ Å) after subtracting the local continuum. Candidate emission lines with an S/N >4 and a reduced chi-square (χ^2) <3 are retained and refit at higher ($0\farcs15$) spatial resolution, this time allowing σ to vary. The raster location that maximizes S/N defines the source centroid, and the best-fit amplitude yields the line flux. Duplicate detections within $3''$ and 3 Å are merged by selecting the highest-S/N measurement.

Final line catalog membership is determined by additional quality cuts: a detection must satisfy S/N, line width, and χ^2 combinations that help to identify real galaxies from artifacts. Spurious features from negative continua, poor fiber fits, or detector artifacts are removed through these cuts but also through a number of automated methods described in Section 3.2.

The raw emission-line database consists of 31,961,716 detections and reaches ~50% completeness at $11 \times 10^{-17}\,\mathrm{erg\,s^{-1}\,cm^{-2}}$, although this number varies with wavelength, image quality, and observing conditions.

3.1.2. Continuum Detection

Continuum sources are searched for independently by measuring fiber counts in two 200 Å wide spectral regions, one in the blue (3700–3900 Å) and one in the red (5100–5300 Å). If either region exceeds 50 counts per 2 Å pixel, the fiber is stored as a continuum detection. This translates to roughly $g_{\mathrm{HETDEX}} \sim 22.5$, where g_{HETDEX} is a pseudomagnitude derived from the HETDEX source spectrum by convolving it with the SDSS g-band transmission curve and integrating the resulting flux. A PSF model fit is then used to centroid the source with a 15 × 15 grid raster in $0\farcs1$ steps, and a point-source spectrum is extracted at the best-fit position following the methodology of K. Horne (1986). As with line detections, quality checks remove sources located on low-quality detectors or those flagged as artifacts or transients. The raw continuum catalog contains 682,067 detections, with sensitivity set by photon statistics and typically reaching $g_{\mathrm{HETDEX}} \sim 22.5$ under typical HETDEX observing conditions. Continuum detections fainter than $g_{\mathrm{HETDEX}} = 22$ are not included in the final catalog, as they do not provide enough sensitivity to determine a redshift.

3.1.3. AGN Catalog Selection

The HETDEX AGN catalog (C. Liu et al. 2022, 2025), derived from the same internal HDR5 release, is incorporated into our combined source catalog. While it shares the same raw detection databases, their selection criteria differ. The AGN catalog results from targeted searches for broad emission lines, whereas the main catalog, as discussed in Section 3.2, excludes lines with large Gaussian widths ($\sigma > 14$ Å) to mitigate artifacts that often appear as broad features. The AGN classification method also identifies candidates through the presence of common spectral line pairs across the full emission-line database, enabling the identification of both Type I and Type II AGNs. Roughly a quarter of the AGN sample overlaps with our line and continuum catalogs, but the majority represent broad-line AGNs that fail the Gaussian line-fit cuts due to their large line widths and high χ^2 values. These sources are essential for capturing the diversity of AGNs and are retained after visual verification to mitigate contamination from artifacts.

In the combined source catalog, AGN detections are grouped with any associated line or continuum emission according to the detection grouping method (Section 3.3), with redshift assignments taken from the AGN catalog classifications of C. Liu et al. (2025).

We note that the HSPC2 source catalog applies stricter data quality cuts than the C. Liu et al. (2025) AGN catalog. The former removes frames with poor observing conditions or failed amplifiers, and eliminates objects that happen to lie on regions masked due to satellite trails, meteor emission, and the like. The AGN catalog retains some of these sources through visual inspection. In addition, detection grouping in HPSC2 is done only by observation, so the same AGN may have multiple entries in HPSC2, while C. Liu et al. (2025) opted to merge all observations of the same AGN into a single `agnid`.

3.2. Quality Control and Selection of Raw Detections

The internal HETDEX pipeline produced 32,643,783 raw detections for HDR3/4/5 line and continuum emission all combined. These files are included in this release and are described in Appendix A. This raw database is processed to a quality-controlled catalog (internal version "5.0.2") of 4,627,812 detections. The outputs from the three object-detection methods (continuum, line, and AGN) are down-selected from the raw detection database through two primary steps: (1) artifact mitigation via the application of the mask model, and (2) filtering based on detection parameters.

1. *Detection masking.* Internal masking of HETDEX data is done at a detection level. For each issue described in Section 2.3, the relevant detection parameters are searched to determine if the source candidate should be flagged. For example, for meteors and satellites, a combination of observation name (e.g., `shotid`) and sky coordinates are used to decide whether to exclude the detection. In contrast, for calibration, amplifier, or fiber issues, the relevant instrument info is queried for an affected time range.
2. *Detection parameter selections.* For the line search, the following line-parameter ranges are required for catalog

inclusion:

$$\begin{cases} 3510\ \text{Å} \leqslant\ \lambda \leqslant 5496\ \text{Å} \\ \text{S/N} \geqslant 4.8 \\ \chi^2 \leqslant 2.5 \\ \texttt{continuum} > -1 \\ \chi^2_{\text{fib}} < 4.5 \\ 1.7\ \ \text{Å} \leqslant\ \texttt{sigma} \leqslant 14\ \text{Å} \\ \texttt{apcor} \geqslant 0.45 \end{cases} . \tag{1}$$

Here, S/N and χ^2 are the signal-to-noise ratio and reduced quality of fit of the emission line, respectively, and `sigma` is the line width measured from the Gaussian line fit. `continuum` measures the local continuum around the HETDEX emission-line detection. It is derived from a linear fit to two 50 Å spectral windows placed 3× sigma on either side of the detected line. χ^2_{fib} is the quality of fit of the fiber profile of the highest weighted fiber for the detection. High values of χ^2_{fib} are associated with detection artifacts. We only include detections with an aperture correction factor, `apcor`, >0.45. This factor is the fraction of the 3.″5 radius circular aperture with fiber coverage.

For continuum sources, the only exclusion is based on `apcor` and g_{HETDEX}, the magnitude derived from the detection spectrum by folding the detected spectrum through the SDSS-g filter's bandpass and summing the flux.

$$\begin{cases} g_{\text{HETDEX}} \leqslant 22 \\ \texttt{apcor} \geqslant 0.45 \end{cases} \tag{2}$$

3.3. Detection Grouping

Both the line- and continuum-emission pipelines are optimized for point-source detection. For LAEs and many faint [O II] emitters, this assumption is generally valid, although neither LAEs nor [O II] emitters are truly pointlike. As shown in E. Mentuch Cooper et al. (2026), more than 50% of LAEs with S/N > 6 are better described by a two-component surface-brightness model consisting of a compact core and an extended exponential halo. Roughly 25% of HETDEX sources, such as $z < 0.5$ galaxies, AGNs, and stars, are composed of multiple detections, arising either from extended emission or from multiple emission lines. In addition, bright sources can be independently detected in both the line- and continuum-emission searches.

To robustly associate detections and mitigate confusion from extended low-z [O II] emission, we apply a 3D friends-of-friends (FOF) clustering algorithm. Using spatial and spectral linking lengths of 6″ and 8 Å, we combine emission features into coherent source groups as described in E. Mentuch Cooper et al. (2023). This procedure ensures multiline galaxies are properly grouped, although in some blended cases, background sources may be merged into a foreground object. This will cause some background sources in gravitational lenses to be merged with the foreground object; although, these can be discovered through independent searches (I. H. Laseter et al. 2022) within the expanded catalog described in Appendix B.

For each source, we designate a single representative detection, identified by the value in the `selected_det` column of the expanded catalog Detection Information Table (Appendix B). In the main catalog table, only this detection is printed for each source (including its exact detection position at `RA_det`, `DEC_det`); all other detection entries are removed. Non-LAEs are assigned the brightest g_{HETDEX} member, while LAEs are assigned the highest-S/N Lyα detection. This procedure produces a quality-controlled source catalog in which multiple detections are linked through the unique source identifier `source_id`. In summary, `source_id` provides a common identifier for all detections associated with the same astrophysical object, enabling the association of extended emission features, multiple emission lines, and the assignment of a consistent redshift and source classification.

3.4. Source Classification and Redshift Assignment

Each HETDEX source group is assigned a classification and redshift using one of three methods depending on brightness and spectral features. For sources with $g < 22$, we apply DIAGNOSE (M. H. Debski et al. 2025; M. Debski & G. Zeimann 2024), a PCA-based template-fitting algorithm developed for the Hobby–Eberly Telescope VIRUS Parallel Survey (HETVIPS; G. R. Zeimann et al. 2024). DIAGNOSE assigns classifications such as `star`, `galaxy`, or `qso` for bright sources, along with a redshift derived from spectral template matching. It recovers ∼97% of classifications reliably at $g_{\text{HETDEX}} < 22$. However, for fainter objects, confusion between Lyα and low-z lines demands that we rely on alternative methods.

Objects with little to no continuum emission are classified using ELIXER (D. Davis et al. 2023a). This program is based on the initial Bayesian work presented in A. S. Leung et al. (2017), incorporating equivalent widths, broadband imaging, and luminosity priors to infer line identity and redshift. Each detection is assigned a likelihood `P(Lyα)` of being Lyα, with a threshold of 0.4 chosen to balance recovery of true LAEs (96%) and [O II] contamination in the LAE candidate sample (<3%). For high-redshift sources, ELIXER assigns the redshift of the most central detection, considering clustering of lines and multiline associations.

Redshifts and classifications for AGNs are adopted from the HETDEX AGN catalog (C. Liu et al. 2022, 2025), which provides redshift assignments for single broad-line detections (generally assuming the line is Lyα), for emission-line pairs indicative of AGN activity, and for broad-line sources identified through crossmatching with SDSS DR14.

A logical sequence is implemented to assign each source a best redshift and object classification. Priority is given to AGN matches, followed by DIAGNOSE redshifts for brighter objects ($g_{\text{HETDEX}} < 22$), and ELIXER results otherwise. A histogram of spectroscopic redshifts is presented in Figure 7.

We assess the accuracy of HETDEX redshifts through a comparison to two independent spectroscopic samples. The first is the COSMOS field spectroscopic redshift compilation created by A. A. Khostovan et al. (2026); this sample is skewed toward brighter sources, and, when restricted to objects with high-quality redshifts (`flag` ⩾ 3), the cross-matched sample (with a matching radius of 1.″0) consists of 395 LAEs, 303 AGNs, 347 low-redshift galaxies, and 4767 [O II] emitting sources. We note that since COSMOS was used by HETDEX as a science verification field, the region was

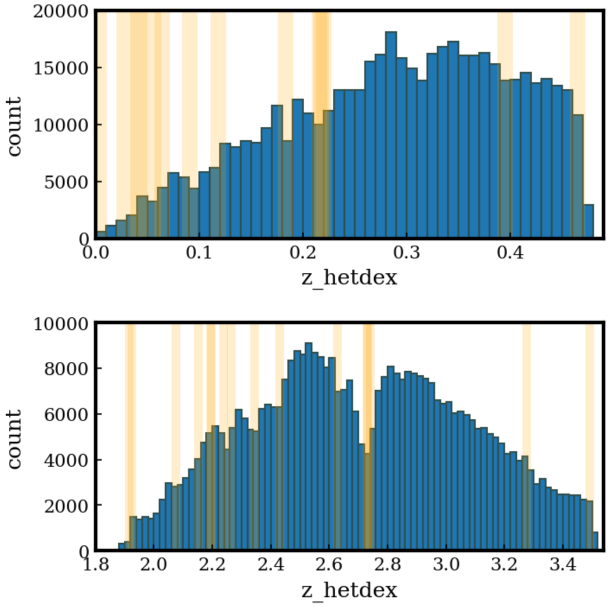


Figure 7. The redshift distribution of the low-z ($\Delta z_{\mathrm{bin}} = 0.01$) and high-$z$ ($\Delta z_{\mathrm{bin}} = 0.02$) galaxy samples in the top and bottom panels, respectively. The low-z sample is a combination of [O II] emitters and LZGs; the high-z dataset includes LAEs and AGNs. The brightest sky lines are marked by light-yellow vertical bars; these lines suppress the number counts in both distributions.

visited multiple times in the course of the survey. The data acquired during each visit were reduced and analyzed independent of the other visits; as a result, a single COSMOS source can potentially have different HETDEX redshifts and object classifications. The numbers listed above therefore represent the total number of redshift comparisons, not the number of unique sources on the sky.

The second sample comes from objects observed by the DESI–HETDEX survey (M. Landriau et al. 2025). This program concentrated on sources with HETDEX classifications that straddled the decision boundary between LAE and [O II] galaxy, and is composed of 851 LAEs, 67 AGNs, and 139 [O II] galaxies, all with high-quality (`VI_QUALITY`>3) redshifts. Thus, our redshift comparison involves over 6800 measurements in both the local and high-redshift Universe. Note that both DESI and HETDEX measure the LAE redshifts assuming that it is a single Gaussian emission-line profile, even though it is not uncommon to observe a double-peak feature (M. Landriau et al. 2025). Using the mock LAE spectral profile in H. Khoraminezhad et al. (2025), we have checked that the impact of this assumption is given by $\Delta z/(1+z) \sim 0.0005$ for DESI and $\Delta z/(1+z) \sim 0.0001$ for HETDEX, where HETDEX has a smaller bias due to its slightly poorer spectral resolution than DESI.

Figure 8 shows the direct comparison between z_{HETDEX} and z_{spec}, as well as the distribution of normalized residuals, defined as $(z_{\mathrm{spec}} - z_{\mathrm{HETDEX}})/(1 + z_{\mathrm{spec}})$. The ensemble statistics demonstrate that HETDEX redshifts are unbiased, with an average offset of 0.0026 and a scatter of $\sigma = 0.22$. The normalized median absolute deviation is consistent with zero bias (NMAD = 0.0003), and the catastrophic outlier fraction (defined by $|\Delta z|/(1+z) > 0.2$) is 5.9%.

As noted above, the DESI–HETDEX program specifically targeted faint emission-line galaxies near the classification boundary between Lyα and [O II] and, therefore, consists primarily of objects with $g > 23$. By design, there are no false positives or neighbor confusion. In contrast, the heterogeneous collection of object classes in the COSMOS spectroscopic compilation is biased toward brighter sources. Despite these differences, the catastrophic outlier fractions are nearly identical in the two samples: 5.97% for COSMOS and 6.03% for DESI–HETDEX.

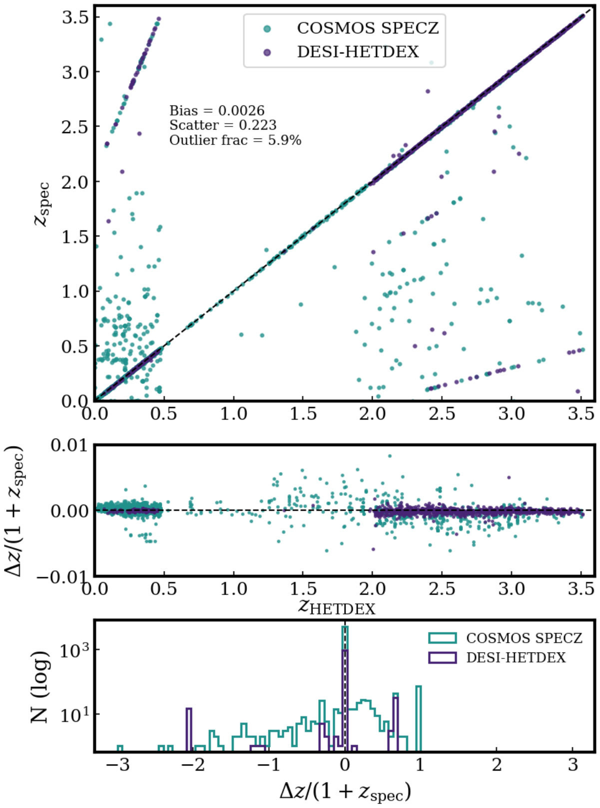


Figure 8. Comparison of HETDEX redshifts with external spectroscopic measurements in the COSMOS legacy field compiled by A. A. Khostovan et al. (2026) and visual spectroscopic confirmation from DESI–HETDEX (M. Landriau et al. 2025). Top panel: direct comparison between z_{HETDEX} and z_{spec}, with the dashed line showing the one-to-one relation. Middle panel: normalized residuals, defined as $(z_{\mathrm{spec}} - z_{\mathrm{HETDEX}})/(1 + z_{\mathrm{spec}})$, as a function of z_{HETDEX}. Bottom panel: distribution of residuals on a logarithmic scale, highlighting the extended tails. Across the full sample, we measure a small bias (0.0026), scatter ($\sigma = 0.22$), and an outlier fraction of 5.9%.

In the COSMOS comparison, catastrophic outliers can arise from incorrect counterpart associations, including occasional HETDEX false positives or nearby on-sky neighbors within the matching radius. In contrast, the DESI–HETDEX sample is constructed from emission-line detections identified independently by both surveys, so catastrophic failures primarily result from incorrect emission-line identification. The most common case is confusion between Lyα and [O II], which produces the two distinct outlier tracks visible in Figure 8. Overall, these results demonstrate that HETDEX redshifts are robust across the survey footprint, with only a modest tail of failures dominated by low-S/N detections or occasional misclassifications.

3.4.1. Source Classification

HETDEX sources are broken into five `source_types`. Source counts by field are provided in Table 1, and the magnitude distributions for each source type are shown in the histograms in Figure 9.

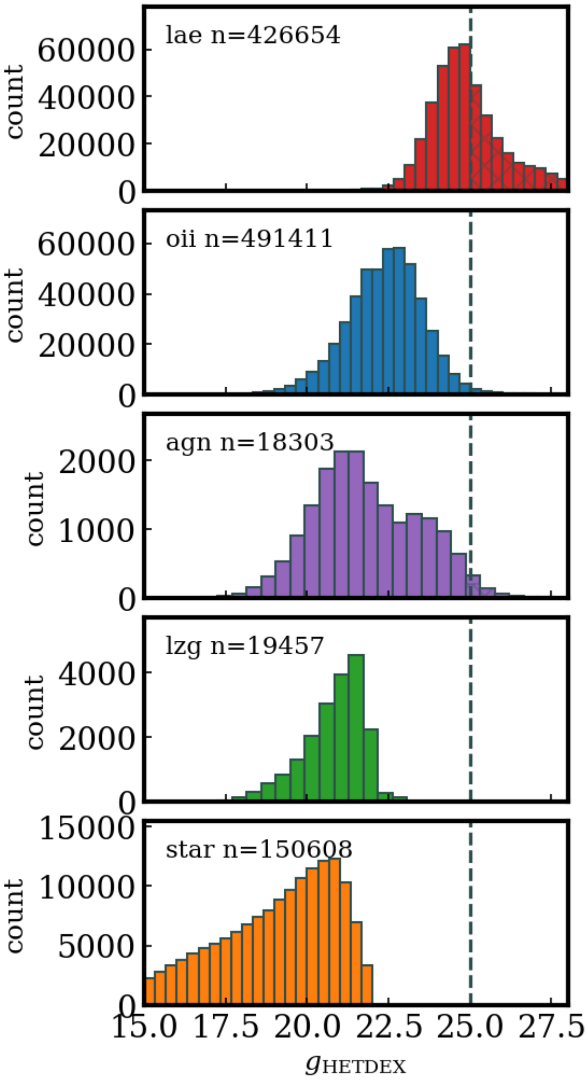


Figure 9. Histogram of $g_{\rm HETDEX}$ magnitudes for each source type as measured by summing the 1D extracted spectra, weighted by the SDSS g-band response curve. If multiple detections exist for the source, the brightest detection is used. The vertical dashed line at $g_{\rm HETDEX} = 25$ represents the HETDEX average sensitivity limit.

1. `star`—Galactic stellar sources; classified by DIAGNOSE and assigned `z_hetdex` = 0.
2. `agn`—Active galactic nuclei; identified via the HETDEX AGN catalog or DIAGNOSE; redshift range: 0 < `z_hetdex` < 4.6.
3. `lae`—Lyα emitters classified by ELIXER; redshift range: 1.88 < `z_hetdex` < 3.52.
4. `oii`—Low-redshift galaxies with detected [O II] emission (λ3727 Å), classified via DIAGNOSE or ELIXER; redshift range: 0 < `z_hetdex` < 0.48.
5. `lzg`—Low-redshift galaxies without detected emission lines classified by DIAGNOSE using continuum template fits; redshift range: 0 < `z_hetdex` < 0.5.

3.5. *Spatially Resolved [O II]*

The HETDEX emission-line detection algorithm produces PSF-weighted line-flux values by assuming the object being measured is a point source. However, many nearby [O II] line emitters are extended, resulting in underestimated measurements of the objects' fluxes. In addition, the emission-line detection pipeline contains an upper limit on the continuum value, so very bright emission-line galaxies are completely missing from the emission-line database, although they are found in the HETDEX continuum catalog and are assigned a redshift. Without an associated emission line, they are initially classified as a low-redshift galaxy (an "lzg"). Figure 10 illustrates this issue. The two leftmost sources are only found in the continuum search method with no reported [O II] line flux; in the third galaxy, the HETDEX-report line flux severely underestimates the total [O II] emission from the galaxy.

For these resolved galaxies, we supplement the HETDEX line-flux measurements with estimates of galaxies' total line emission. A major strength of the wide-IFU (dithered) coverage provided by HETDEX is that the observations automatically produce an emission-line map of resolved galaxies. However, due to the grid-like distribution of IFUs in the HET's focal plane, many of these systems have incomplete coverage, as their light extends beyond their IFU's limits. We deal with this issue by using shape information from ancillary imaging observations either provided by HSC or DECam Legacy Survey (DECALs; A. Dey et al. 2019).

Isophotal elliptical apertures based on broadband imaging are included in HETDEX's ELIXER classification tool (D. Davis et al. 2023a). The program applies SOURCE EXTRACTION AND PHOTOMETRY (SEP; K. Barbary 2016) to all available broadband imaging at the location of each HETDEX detection. We use the ELIXER catalog `selected=True` option (indicating this was the imaging counterpart of choice) and preferentially choose r-band measurements over g-band measurements to define each galaxy's elliptical aperture. In general, the r-band imaging has a fainter limiting magnitude, and better image quality.

Elliptical parameters for each low-z galaxy are provided in HPSC2 (columns described in Table 5) under the columns `major`, `minor`, and `theta`. Additionally, the aperture center and the measured continuum aperture magnitude are in the expanded Detection Information Table in Appendix B in the columns `ra_aper`, `dec_aper`, `mag_aper`, and `mag_aper_err`.

For every galaxy (both "lzg" and "oii") at `z_hetdex` < 0.5, we first construct a sky-subtracted narrowband image centered on the observed [O II] wavelength (given by `z_hetdex`). The cube is collapsed spectrally over a ± 15 Å window to form the line-flux map. To remove continuum emission, we create two additional 50 Å wide collapsed images placed ± 10 Å outside the line-map limits. The average of these two continuum images is subtracted from the narrowband image, producing a continuum-subtracted [O II] emission-line-flux map.

Because many resolved galaxies extend beyond the HETDEX IFU footprint, only part of the elliptical aperture defined from broadband imaging is actually covered by IFU data. After generating the continuum-subtracted map, we therefore measure the [O II] flux within the portion of the aperture that has fiber coverage, compute the fraction of the aperture that is sampled, and scale the measured flux by the inverse of this coverage fraction. This yields an aperture-corrected estimate of each galaxy's total [O II] flux assuming the flux distribution is smooth and evenly distributed within the aperture. In practice, emission-line regions can exhibit radial structure and clumpy morphology, so this correction

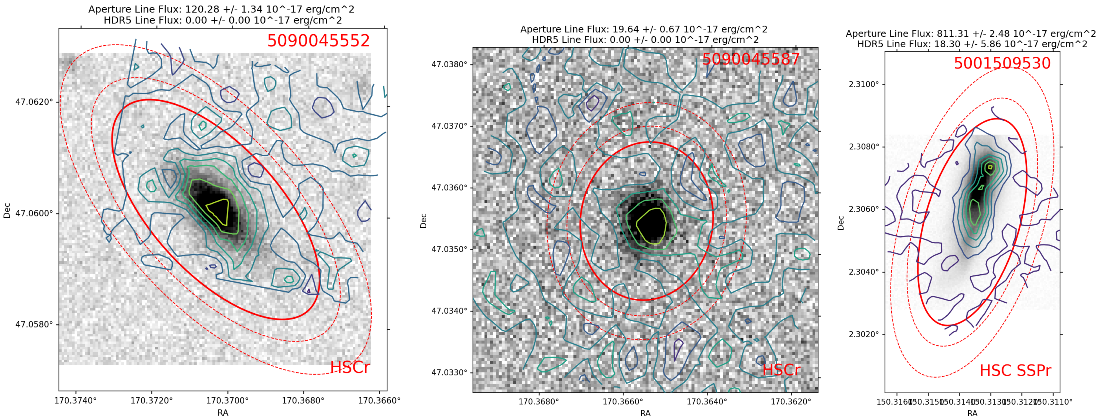


Figure 10. Examples of [O II] line-flux maps (shown in contours) that are used to measure continuum-subtracted [O II] fluxes for spatially resolved low-redshift galaxies. Elliptical apertures are defined using imaging data from HSC or DECALs as measured in the ELIXER catalog (D. Davis et al. 2023a). In total, 32,031 galaxies are found to have [O II] line emission that was not initially measured by the HETDEX reduction pipeline.

should be regarded as an approximate estimate of the total flux. The aperture-corrected fluxes and their uncertainties are provided in the catalog under the columns `flux_aper` and `flux_aper_err`, respectively.

A significant number of low-z galaxies exhibit a positive [O II] line-flux value within the resolved galaxy aperture, despite an absence of an [O II] flux detection in the initial pipeline database. While the redshift of these systems remains unchanged, their classification is revised from "lzg" to "oii." Using this method, 32,031 sources in the HPSC2 catalog have been reclassified as [O II] emitters from an initial classification as a low-z passive galaxy.

3.6. [O II] and Lyα Line Fluxes and Luminosities

For the two main emission-line datasets, we include columns for their continuum-subtracted, extinction-corrected line fluxes (`flux_oii, flux_lya`) and corresponding luminosities (`logL_oii, logL_lya`), along with their respective uncertainties marked by the `_err` subscript. This information is only relevant to sources with [O II] and Lyα emission lines and pertains exclusively to objects categorized with `source_type` as "oii," "lae," or "agn." Only the AGN with Lyα-emission within the HETDEX spectral range will have reported flux values. For accessing other emission-line-flux data, users are directed to Appendix B.

The native source detection pipeline as described in K. Gebhardt et al. (2021) is designed to deliver continuum-subtracted line fluxes under the assumption that the source is unresolved. However, many HETDEX sources are resolved. Moreover, an additional shortcoming in the HETDEX detection procedure is that an upper bound on continuum counts is set during the search for emission lines. As a result, the line flux emitted at the central location of a bright [O II] emitter or AGN is not reported.

As outlined in Section 3.5, the issue is addressed for [O II] emitters by fitting spatially resolved line fluxes for each system with `z_hetdex` < 0.5. When the aperture flux, denoted as `flux_aper`, exceeds the pipeline-reported `flux`, this spatially integrated flux measurement value is assigned to `flux_oii`, and the `flag_aper` is set to 1. Conversely, if the HETDEX pipeline's estimate of the [O II] surpasses the aperture flux, the pipeline flux is preserved in `flux_oii`, and `flag_aper` is set to 0. If no [O II] flux is associated with the source, the `flag_aper` is assigned a value of -1, and the source is not to be classified as an "oii."

For LAEs, it is generally acceptable to assume that their morphology resembles that of a point source, especially considering the relatively poor image quality delivered by the HET. Research focused on quantifying surface-brightness profiles and determining isophotal radii suggests that for a subset of LAEs with an $S/N > 6$ (E. Mentuch Cooper et al. 2026), an exponential model with a scale length of $\sim$10 kpc is a more accurate representation of LAE morphology. This improved model increases the reported Lyα flux by a factor of $\sim$1.3 over the point-source calculation as reported in E. Mentuch Cooper et al. (2026). However, we have chosen not to implement any corrections to adjust for this discrepancy. We also have opted to not provide spatially resolved flux values for extended Lyα emitters; these measurements can be found in E. Mentuch Cooper et al. (2026).

As mentioned above, some AGNs are often bright enough so that their continuum counts are above the maximum allowed for an emission-line search. In these objects, the Lyα line flux is measured by subtracting the surrounding continuum and fitting a multicomponent Gaussian model to the emission line (C. Liu et al. 2025). These line-flux values are reported in `flux_lya` in HPSC2. For AGNs whose Lyα emission lies outside the spectral range of the VIRUS spectrographs, no flux is reported. Readers that desire more advanced line-flux values are referred to the AGN catalogs created from HETDEX data release HDR4 (C. Liu et al. 2025).

3.6.1. Extinction Correction due to Dust Attenuation

All flux values, unless designated with an `_obs` subscript have been extinction corrected. The Python software package DUSTMAPS (G. Green 2018) provides local Milky Way Dust reddening values for each source's coordinates as measured by D. J. Schlegel et al. (1998). The software returns the locally measured color excess value, $E(B - V)$, based on a source coordinate. We assume the ratio of V-band extinction, A_V, to color excess, $E(B - V)$, to be $R_V = 3.1$ and apply a factor of 2.742 to measure the local V-band extinction as $A_V = 2.742 \times E(B - V)$ according to the re-calibration using SDSS stars of the D. J. Schlegel et al. (1998) maps by E. F. Schlafly & D. P. Finkbeiner (2011). However, we note that large variations of this re-calibration are seen in the Milky

Table 5
Source Observation Table Column Descriptions

Name	Description
source_name	HETDEX IAU designation (e.g., `HETDEX J123449.19+511733.7`).
source_id	HETDEX Source Identifier.
shotid	Integer observation ID: `int(date+obsid)`.
ifuslot	String identifier of IFU location in focal plane.
RA	R.A. for source (ICRS deg).
DEC	Decl. for source (ICRS deg).
RA_det	Representative detectid R.A. (ICRS J2000 deg)
DEC_det	Representative detectid decl. (ICRS J2000 deg)
gmag	SDSS g magnitude measured in the HETDEX spectrum.
Av	Applied extinction correction in the V band.
z_hetdex	HETDEX spectroscopic redshift.
z_hetdex_src	Source/method for HETDEX spectroscopic redshift.
z_hetdex_conf	Confidence (0–1) in HETDEX spectroscopic redshift source.
source_type	One of `star`, `lae`, `agn`, `lzg`, `oii`.
detectid	Detection ID of representative detection (where `selected_det` = `True` in the *Detection Info Table*).
field	Survey field: `dex-fall`, `dex-spring`, `cosmos`, `goods-n`, `nep`, or `ssa22`.
flux	Extinction-corrected line flux in 10^{-17} erg s^{-1} cm^{-2} at `wave`. −999.0 for continuum sources but is correctly assigned in `flux_lya` or `flux_oii`.
flux_err	Markov Chain Monte Carlo (MCMC) error in extinction-corrected line flux.
flux_aper	Extinction-corrected [O II] line flux in 10^{-17} erg s^{-1} cm^{-2} measured in the resolved galaxy aperture.
flux_aper_err	Uncertainty in `flux_aper`.
flag_aper	Aperture-flux usage flag for `lum_oii`: 1 = use `flux_aper`; 0 = use PSF line flux `flux`.
major	Major axis (arcseconds) of the resolved [O II] aperture ellipse (from imaging).
minor	Minor axis (arcseconds) of the resolved [O II] aperture ellipse (from imaging).
theta	Position angle (degrees) of the aperture ellipse.
logL_lya	$\log_{10}$ Lyα luminosity erg s^{-1} from extinction-corrected `flux`.
logL_lya_err	Uncertainty in `logL_lya`.
logL_oii	$\log_{10}$ [O II] luminosity erg s^{-1} from `flux` if `flag_aper=0` or from `flux_aper` if `flag_aper=1`.
logL_oii_err	Uncertainty in `logL_oii`.
flux_lya	Lyα flux in 10^{-17} erg s^{-1} cm^{-2} from extinction-corrected `flux`.
flux_lya_err	Uncertainty in `flux_lya`.
flux_oii	[O II] flux in 10^{-17} erg s^{-1} cm^{-2} from `flux` if `flag_aper=0` or `flux_aper` if `flag_aper=1`.
flux_oii_err	Uncertainty in `flux_oii`.
sn	Signal-to-noise ratio for line emission. Note that continuum sources (`det_type`==“cont”) will have `sn` = −999.0.
det_type	Detection type of the main detectid for the source: “line” or “cont.”
apcor	Aperture correction applied to the spectrum at 4500 Å.
p_conf	LAE/faint [O II] emitter confidence score from RF classifier. 1 = high confidence, 0 = low confidence.
p_cnn	LAE/faint [O II] emitter confidence score based on CNN classifier. 1 = high confidence, 0 = low confidence.

Note. Bad values are −999.0 for floating-point columns and `n/a` for string columns. This table is available in its entirety in machine-readable form at https://hetdex.org/data-results/ and on Zenodo.

Way interstellar medium (X. Zhang & G. M. Green 2025; D. Lee et al. 2025). A_V values range from 0.01–1.44 with a median value of 0.04. Extinction corrections of line fluxes are applied at the central wavelength of each cataloged emission line according to the $R_V = 3.1$ extinction curve of E. L. Fitzpatrick (1999), implemented using the open-source Python software EXTINCTION.[39]

3.7. ML/AI-assisted Classification

ML and AI tools provide efficient methods for data classification, and the HETDEX collaboration has explored several such approaches. Unsupervised learning has proven valuable for identifying supernovae (J. Vinkó et al. 2023) and AGNs (V. Tardugno Poleo et al. 2023) within the rich HETDEX dataset, while also offering interactive methods for artifact removal and sample cleaning.

For HPSC2, ML/AI methods have been most successful in identifying and removing artifacts from the raw detection database and improving the low-S/N emission-line sample. In this section, we describe three approaches that have proven particularly effective and that are incorporated into HPSC2 and the PDR1 data cube mask model.

3.7.1. Rapid Automatic Image Categorization

Every emission-line and continuum detection in the HETDEX catalog is passed through a classification workflow using the RAIC Labs.[40]

---

[39] https://github.com/kbarbary/extinction

[40] The https://raiclabs.com classification platform (RAIC), developed as a scalable visual AI system for analyzing large image datasets using semisupervised learning and human-in-the-loop interaction. This system provides a user-friendly interface for semiautomated labeling of astronomical detections, enabling rapid and scalable classification of large survey datasets without requiring fully labeled training sets.

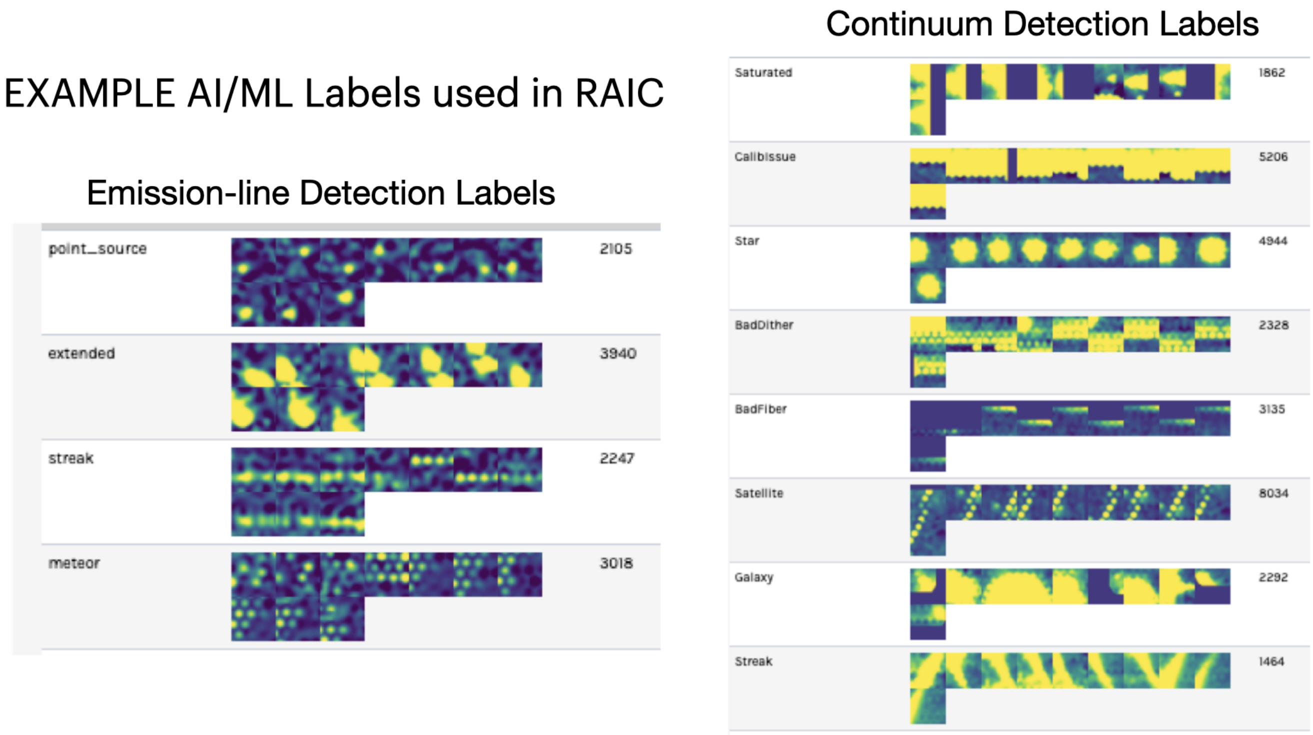


Figure 11. RAIC Labs classification label sets used for HETDEX detections. Panel (a) shows the classes applied to emission-line cutouts, while panel (b) shows the classes for continuum cutouts.

For each emission-line detection, we generate a narrowband cutout, 20″ on a side, with the wavelength dimension collapsed over $\pm 2\sigma$ around the fitted central wavelength of the line, where σ is taken from the Gaussian profile fit produced by the HETDEX pipeline. No continuum subtraction is performed. For continuum detections, cutouts of the same size are produced by collapsing the spectrum over the rest-frame wavelength range 4000–5000 Å. In both cases, the detection of interest is centered within the cutout. These narrowband cutouts are uploaded to the RAIC Labs classification platform, which projects the image set into a low-dimensional feature space where unsupervised learning naturally groups detections with similar morphology. The tool provides an interactive GUI in which users can visually explore this feature space and interactively choose selections to assign labels to groups of sources simultaneously. This approach enables efficient labeling of large numbers of detections with consistent morphological classifications.

The resulting label sets are shown in Figure 11. For emission-line detections, the labeled sets are point source, extended object, streak, and meteor. For continuum detections, the categories include saturated object, calibration issue, star, bad dither, bad fiber, satellite, galaxy, streak, low counts, and faint star. Each category is built up to include at least 1000 example labels.

Once a sufficient number of labels has been collected via the RAIC Labs interactive GUI, the system trains its internal classification model to apply those labels to previously unlabeled detections. Each source in the HETDEX catalog (emission-line or continuum) is passed through the relevant classification model, and is assigned a likelihood score, set between 0.0 and 1.0, for each label for each detection. This process means that any HETDEX detection can be assigned multiple labels. Those that are above the threshold classification score of 0.5 are assigned that label.

The labels for "meteor" and "satellite" are used to help identify contaminants in the catalog and create the masking for each category, respectively (see Sections 2.3.7 and 2.3.11.) From the continuum labels, we flag the labels "calibissue," "baddither," and "badfiber" from the catalog when their label score is >0.3. From the emission-line labels, we use the flag and remove any detections with the "streak" label and a score >0.35.

From our internal quality-controlled catalog 113,954 detections are flagged by RAIC. We do not include these in any public catalogs. While many of these artifacts are found through our other automated flagging procedures, 41,002 are only found using the RAICLabs AI classification platform; this demonstrates the success of this software as a tool for efficiently labeling and classifying HETDEX detections. These detections are masked in the data cubes under the BADDET bitmask described in Section 2.3.14.

3.7.2. Citizen Science + Unsupervised Learning

Another method to reduce false positives due to noise and artifacts is to combine large-scale citizen science vetting from the NASA Zooniverse project *Dark Energy Explorers*[41] (DEE) with unsupervised learning on the detection spectra (L. R. House et al. 2023). DEE presents compact "mini" views of each detection produced by HETDEX custom software ELIXER and collects a minimum of 10 binary classifications per source. This result is a per-object DEE probability, p_{DEE}, that a candidate is real, as opposed to an artifact. Since its launch in

---

[41] https://www.zooniverse.org/projects/erinmc/dark-energy-explorers

2021 February, DEE has amassed over 4 million classifications from over ∼22,000 volunteers.

As described in L. R. House et al. (2023), we then map spectra into a low-dimensional space using t-distributed stochastic neighbor embedding (t-SNE), operating on a ± 50 Å window centered on the emission-line wavelength. We found that the labels alone lead to too many real sources being assigned low p_{DEE} scores; however, the combination of an object's p_{DEE} and its location in t-SNE space—particularly in regions where the majority of detections also have low p_{DEE} values—was very successful at labeling artifacts with $\mathrm{S/N} > 5.5$.

As of this writing, the final set of HDR5 detections with $\mathrm{S/N} > 5.5$ is still being evaluated, but we have flagged all candidates that satisfy the exclusion conditions described in L. R. House et al. (2023) and removed them from the public catalog. These objects are marked `flag_best==0` in our internal catalog but are not retained in the public release. In total, 28,841 detections are flagged and removed. Many of these are also identified by the masking procedure described in Section 2.3. For those missed by our primary masking methods, the remaining sources are masked in the data cubes with the cubic BADDET mask.

3.7.3. CNN Classifier: `p_cnn`

The largest challenge for HETDEX and its untargeted emission-line search is validating the low-S/N emission-line detections found by the HETDEX reduction pipeline. Standard quality flags and conservative S/N thresholds (>5.5) are effective at reducing contamination but also exclude a large fraction of genuine LAEs, limiting the source density needed for cosmological analyses.

Extending the threshold down to the pipeline limit ($\mathrm{S/N} = 4.8$) requires an additional layer of classification to suppress residual noise and artifacts. To address this issue, S. Mukae et al. (2026) developed a convolutional neural network (CNN) tailored for the HETDEX 2D spectra. The model was trained on ∼4300 emission-line candidates from the COSMOS field, with validation from HETDEX repeat observations, ancillary spectroscopy, and multiple narrowband LAE surveys and citizen science classifications. By incorporating both the spatial and spectral structure of detections, the CNN achieves ∼90% accuracy overall and maintains strong performance (86%) in the low-S/N regime. Visualization methods confirm that the network identifies smooth, centrally concentrated features as true LAEs, while rejecting irregular or noisy patterns.

Applied to the full catalog, the classifier extends the usable S/N range, suppresses spurious sky-line-induced redshift spikes in the blue region of the spectrum, and recovers the target LAE density required for cosmological measurements. This approach demonstrates that domain-specific deep learning methods can enhance the reliability of blind spectroscopic surveys and improve the scientific yield of HETDEX.

The CNN score, `p_cnn`, is provided in HPSC2 and the supplemental Detection Information Table in Appendix B. It is measured for the full LAE sample of ∼1.6 million LAE candidates as well as a subset of ∼400,000 faint [O II] emitters with $g_{\mathrm{HETDEX}} > 22$, which are found to be well described by this CNN classifying approach.

3.7.4. Random Forest Classifier: `p_conf`

We train a single, compact random forest (RF) classifier to distinguish high-confidence LAEs from low-confidence detections in the COSMOS Legacy field. This field benefits from multiple HETDEX observations, which allow for confirmation of genuine LAEs, and provides a set of high-confidence LAEs suitable for use in supervised machine learning methods.

Training is restricted to COSMOS LAE emission-line detections and excludes any objects previously classified as AGNs. It is further limited to faint objects with $g_{\mathrm{HETDEX}} > 22$. Classification labels are binary, with 1 denoting a high-confidence LAE candidate and 0 denoting a low-confidence LAE candidate. It is important to note that verifying that an HETDEX detection is real is much more straightforward than proving the candidate is noise, unless the object exhibits characteristics of an obvious artifact. Because HETDEX is dominated by low-S/N emission lines, uncertainties in the noise model lead to detections with questionable validity.

High-confidence classifications originate from several sources. HETDEX observed some regions of the COSMOS Legacy field multiple times; any LAE detected in a second observation beyond its original detection is considered a validated source. We also include objects confirmed through positional and redshift and/or spectral matches to other spectroscopic programs (via the compilation by A. A. Khostovan et al. 2026), or deep narrowband surveys. Included in the latter category are SILVERRUSH ($z = 2.2$ and 3.3; S. Kikuta et al. 2023), ODIN ($z = 2.4$ and 3.1; N. M. Firestone et al. 2024), and the SC4K narrow- and medium-band surveys ($z \sim 2.2$–3.4; D. Sobral et al. 2018). To increase sample size and better match the distribution of HETDEX LAEs, we also include sources classified as "Real" in the citizen science project DEE (L. R. House et al. 2023), with a confidence score of $p_{\mathrm{DEE}} \geqslant 0.8$. In total, 2935 sources are assigned real. The 4582 low-confidence sources come exclusively from DEE with $p_{\mathrm{DEE}} \leqslant 0.1$.

We explore a number of catalog columns as possible feature vectors for our LAE training set. Ultimately, we find that a compact feature set performs well, relying only on parameters measured directly from the detection pipeline:

$$\boldsymbol{x} = \{\texttt{wave},\ \texttt{sn},\ \texttt{chi2},\ \texttt{linewidth},\ \texttt{continuum}\},$$

where `wave` is the wavelength of the line, `sn` is the pipeline's signal-to-noise, `chi2` is the reduced χ^2 value of the Gaussian line-fits to the possible feature, `linewidth` is the standard deviation for the line width Gaussian, and `continuum` is the pipeline's fitted value for the continuum underlying the emission line.

We use SCIKIT-LEARN's `RandomForestClassifier` (version 1.5.2) with standard settings (F. Pedregosa et al. 2011), including `n_estimators=100`, `criterion='gini'`, and `max_features='sqrt'`. The classifier's output is a probability-like score, `p_conf`, assigned to each detection in our catalog. As with the CNN classifier, `p_conf` is computed for both the full LAE sample and for a subset of fainter [O II] emitters. Our Random Forest model achieves an accuracy of 0.89 and an area under the receiver operating characteristic curve (AUC) of 0.95 when trained on COSMOS repeat detections and ancillary spectroscopic matches. It reliably separates high-confidence LAEs from low-confidence detections, with most predictive power supplied by the `sn` (0.303) and `continuum` (0.286) features. The remaining features—`wave`, `linewidth`,

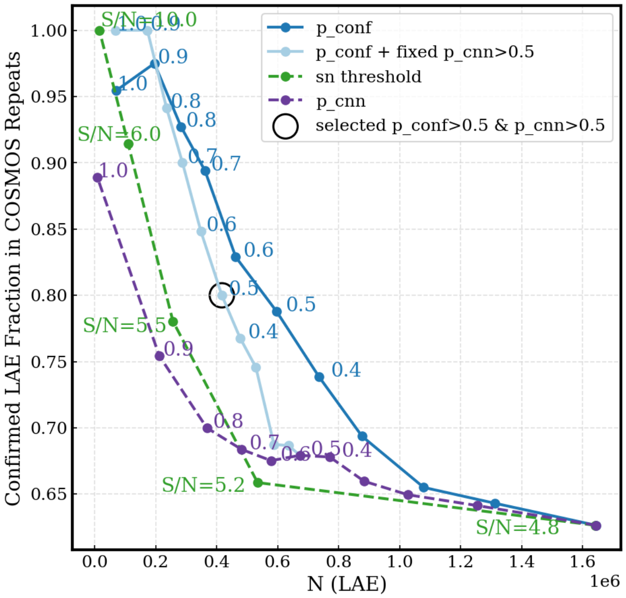


Figure 12. The *y*-axis represents the fraction of LAEs that are confirmed from multiple methods (repeat HETDEX observations, ancillary spectroscopic, and narrowband confirmation). The *x*-axis indicates total LAE count from the full LAE catalog based on the plotted thresholds. We consider different confirmation rates for various LAE selection cuts based on an S/N threshold cut (green curve), and threshold cuts for our ML/AI classifiers based on a convolution neural network (CNN) (`p_cnn`; see Section 3.7.3) and a random forest (RF) classifier (`p_conf`; see Section 3.7.4). This figure visualizes the trade-off between purity (confirmation rate) and yield (total LAEs) for RF-only, RF+CNN, S/N-only, and CNN-only operating strategies. The supplemental table in Appendix B contains the full sample, while the main catalog, HPSC2, contains a limited more robust combined sample selection of `p_conf` > 0.5 and `p_cnn` > 0.5 indicated in the black open circle selected.

and `chi2`—provide smaller but meaningful contributions, with relative importances of 0.189, 0.115, and 0.107, respectively.

3.8. LAE Confirmation and Best LAE Selection

To determine the optimal thresholds that maximize the LAE sample size and minimize false-positive contamination, we take a systematic evaluation approach that is demonstrated in Figure 12. This figure shows how the confirmation rate of a subset of LAEs from the COSMOS Legacy field varies with a number of different selection criteria thresholds. As stated above, it is difficult to prove with certainty that a low-S/N emission-line detection is not real unless it is visually verified as an artifact: when the completeness fraction is below 50%, it can take a prohibitively large number of observations to say with >90% confidence that a candidate flagged by the HETDEX pipeline is a false detection. Conversely, it is relatively straightforward to prove an object is real. If the object is found in more than one HETDEX observation, or it has already been cataloged by an independent spectroscopic or narrowband survey, then there is no question that the object exists. Thus, instead of plotting the false-positive rate on the *y*-axis, we plot the confirmation rate, which can be considered an upper limit on the inverse of the false-positive rate.

The key to measuring the confirmation rate is to select a sample in which confirmation is possible. While the entire COSMOS field has a large number of spectroscopic redshift sources, provided by the compilation by A. A. Khostovan et al. (2026), and narrowband sources that overlap with the HETDEX LAE population, the majority of HETDEX objects will not be found in ancillary studies. Conventional spectroscopic surveys, which require a broadband detection to identify candidates for follow-up redshift measurements, generally target brighter galaxies (with consequently larger stellar masses) than the average HETDEX LAE. Moreover, while the physical properties of the LAEs identified in deep narrowband surveys, such as SILVERRUSH, ODIN, and SC4K, are similar to those of HETDEX LAEs, these programs target very narrow redshift slices of the Universe. The parameter space explored by HETDEX is quite unlike any other LAE survey except for the small coverage (∼20′) provided by the MUSE COSMOS sample (i.e., T. Urrutia et al. 2019). Operating with longer exposure times and reaching deeper flux sensitivity, this program is too limited in size (and redshift overlap) to provide useful comparison.

The most effective method to confirm an HETDEX LAE is by using HETDEX itself. Unfortunately, due to the large variation in observing conditions, the flux sensitivity of any given observation can be vastly better than another. One repeat observation alone is generally not enough to confirm. Most HETDEX LAEs straddle the 50% completeness limit in a typical HETDEX observation and in that regime, it takes at least five independent observations to excluded the reality of a detection with 95% confidence. By design, we limit the COSMOS confirmation sample to those observations that contain at least eight overlap apertures. This results in a small set of 414 LAEs that have the potential to be confirmed by repeat HETDEX observations as well as ancillary spectroscopic and narrowband surveys. Future dedicated HETDEX validation observations at the HET/VIRUS (obtained by PI Laurel Weiss in 2025) will expand this statistical baseline.

For any threshold on a classifier score, we define the confirmed LAE fraction as

$$f_{\mathrm{conf}}\,(\mathrm{threshold}) = \frac{N_{\mathrm{confirmed}}\,(>\mathrm{threshold})}{N_{\mathrm{total}}\,(>\mathrm{threshold})}, \tag{3}$$

where the denominator is the number of LAEs in the confirmation sample whose classifier scores exceed the threshold (either a selection-cut based on S/N, `p_cnn`, or `p_conf` or a combination of these), and the numerator is the subset of those LAEs that are confirmed by repeat HETDEX observation or by ancillary COSMOS catalogs in that same sample.

In Figure 12, the solid dark-blue curve shows the confirmation fraction for different `p_conf` thresholds, described in Section 3.7.4, as a function of the total number of LAEs included in the full HETDEX catalog sample above the indicated threshold; this is annotated with text for clarity. As expected, for all metrics, we see a trend where higher thresholds correspond to an increased confirmed LAE fraction (and consequently less contamination from false positives), indicating that the use of more stringent criteria results in a more reliable selection of LAEs. However, this comes at the cost of decreasing the overall sample size, as fewer candidates meet the stricter selection criteria. For example, a threshold of `p_conf` = 0.5 increases the confirmation fraction from 62% to 80% but reduces the sample of LAE candidates from 1.6 million candidates to 600,000 LAE candidates. We therefore adopt `p_conf` = 0.5 as a practical compromise between confirmation fraction and sample size, as higher thresholds yield only modest additional improvements in confirmation while rapidly reducing the number of LAE candidates available for statistical studies.

Variation based on thresholds for the CNN classifier, `p_cnn`, described in Section 3.7.3, is shown in the dashed purple line. Confirmation of the sample increases as the threshold for `p_cnn` is raised, but the sample size decreases significantly. As the curve plateaus at `p_cnn` ∼ 0.3–0.7, the confirmation rate does not indicate a single obvious threshold from this metric alone. In S. Mukae et al. (2026), additional criteria are explored, including the recovery of known LAEs in the DESI–HETDEX validation sample (M. Landriau et al. 2025) and ancillary COSMOS catalogs. These analyses show that adopting a CNN threshold of `p_cnn` = 0.5 provides a practical balance between completeness and contamination, recovering the majority of confirmed LAEs while maintaining a catalog size close to the survey's target LAE number density. Because the CNN score uncertainties are typically ∼0.2, sources with scores in the range ∼0.3–0.7 are intrinsically ambiguous, making a threshold near 0.5 a natural dividing point between high- and low-confidence classifications.

To compare against alternative selection criteria, we also compute (i) a combined RF + fixed value CNN sequence requiring a fixed `p_cnn` = 0.5 alongside varying `p_conf` (light blue), (ii) pure S/N cuts at $S/N \in \{4.8, 5.2, 5.5, 6.0, 10.0\}$, and (iii) pure CNN cuts at thresholds from 0 to 1.

It is crucial to strike a balance between maximizing the confirmation rate and preserving a sufficiently large sample for statistical analysis. In this release, we adopt thresholds of `p_conf` ⩾ 0.5 and `p_cnn` ⩾ 0.5 to provide a robust sample while maintaining a sufficiently large catalog for statistical analyses. The choice of `p_conf` = 0.5 reflects a balance between confirmation fraction and sample size, as higher thresholds yield only modest improvements in confirmation while rapidly reducing the number of LAE candidates. The CNN threshold of `p_cnn` = 0.5 is motivated by the analysis of S. Mukae et al. (2026), which shows that this value provides a practical balance between completeness and contamination while preserving the survey's target LAE number density. These values reduce the LAE sample from 1,632,604 to 426,654, and the [O II] sample from 559,544 to 491,411.

Future work will involve exploring alternative selection methods and thresholds to optimize these criteria. For users who are interested in exploring a less stringent selection criteria, we provide the full suite of LAE candidates in the Detection Information Table in Appendix B. Candidates at lower `p_conf` and `p_cnn` values are included. This flexibility allows future users to tailor their analyses based on specific scientific goals.

4. Public Data Model and Access

HETDEX public data products are found at the data mount hosted by the Texas Advanced Computing Center (TACC) at The University of Texas at Austin at https://web.corral.tacc.utexas.edu/hetdex/HETDEX/pdr/. Additional information and data hosting mirrors are provided at https://hetdex.org/data-results/.

A schematic of the PDR1 data model is shown in Figure 13. The primary data products are

1. *IFU metadata table.* The `ifu-index` file (FITS/HDF5/ASCII) provides the master manifest with astrometric, observational, and quality assurance information for every IFU.
2. *IFU data cubes.* 431,713 FITS files, each corresponding to a single IFU observation (`dex_cube_<SHOTID>_<IFUSLOT>.fits`), with dimensions 1036 × 104 × 104 (wavelength × spatial × spatial). Each cube covers 3470–5540 Å at 2 Å sampling, with 0.″5 spatial resolution, and contains `DATA`, `ERROR`, and `MASK` extensions. The WCS header information must be used for orientation as the cubes are on a grid that is aligned with the original telescope rotation angle.
3. *Source catalogs.* .fits/.dat files containing spectra and properties for all detected sources, described in detail in Section 3 and Appendix B.
4. *Raw detection databases.* A record of the raw emission-line and continuum detection search output. A detailed description is given in Appendix A.

The data release is 9.8 TB in size with the data cubes making up the bulk of storage size at 7.4 TB. For the typical user, downloading the full data release is not practical, so we provide example notebooks on how to query the database based on coordinates or observing quality and download only the data cubes of interest. We first describe the data products that make up the PDR1 data release and then provide instructions on how to access the data.

4.1. IFU Index Table

The file `ifu-index.fits` (also available in .h5/.txt format) is a master look-up table for each IFU data cube observation made by HETDEX.

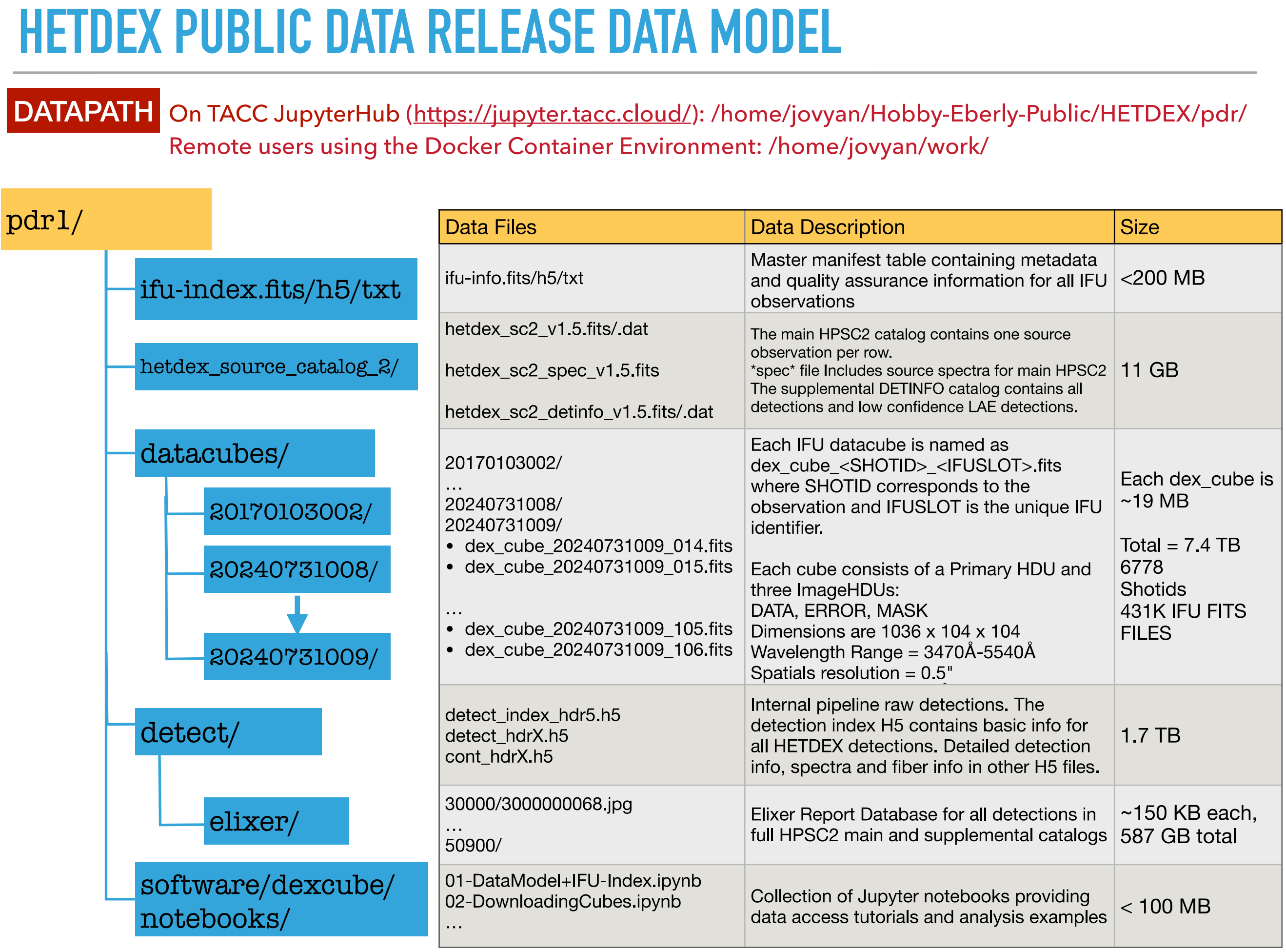
Table 6 describes the columns from the IFU index table. Each row in the table represents a single IFU observation, with the parent observation labeled as `shotid`. An HETDEX observation will have between 14 and 78 IFU observations, identified by column `ifuslot`. The `shotid/ifuslot` combination provides the unique identifier of an HETDEX IFU data cube.

The central sky coordinate of each IFU is listed in Table 6. Users can search for data cubes that overlap within an arcminute of this coordinate and then check the WCS of the data cube for exact coverage. An example query is given in Section 4.5.1. Observational quality assessment values for seeing (FWHM), and transparency (`response_4540`) are tabulated to allow for querying of desired shots. We include the `fwhm_virus` value, which is produced by the HETDEX pipeline based on VIRUS observations.

For each IFU, the fraction of the data cube flagged due to bad amplifiers, bad fibers, meteors, and satellites is reported in the aptly named columns. Of the full set of IFU data cubes, 91.6% of the data remain unflagged, meaning that 8.4% of the data cubes are flagged at all wavelengths. Breaking this down by source of contamination, 7.8% of the total data cube volume are flagged due to bad amplifiers, 0.2% due to bad fibers, 0.12% due to satellites, and 0.04% due to meteors. Additional wavelength-dependent masking can occur; this mask is contained in the BITMASK HDU.

4.2. Catalog Data Products

The creation of the primary catalog delivered in this release is the HETDEX Public Source Catalog 2 (HPSC2), which was described in Section 3. Here we describe its data model.



On TACC JupyterHub (https://jupyter.tacc.cloud/): /home/jovyan/Hobby-Eberly-Public/HETDEX/pdr/
Remote users using the Docker Container Environment: /home/jovyan/work/

Data Files	Data Description	Size
ifu-info.fits/h5/txt	Master manifest table containing metadata and quality assurance information for all IFU observations	<200 MB
hetdex_sc2_v1.5.fits/.dat hetdex_sc2_spec_v1.5.fits hetdex_sc2_detinfo_v1.5.fits/.dat	The main HPSC2 catalog contains one source observation per row. *spec* file Includes source spectra for main HPSC2 The supplemental DETINFO catalog contains all detections and low confidence LAE detections.	11 GB
20170103002/ … 20240731008/ 20240731009/ • dex_cube_20240731009_014.fits • dex_cube_20240731009_015.fits … • dex_cube_20240731009_105.fits • dex_cube_20240731009_106.fits	Each IFU datacube is named as dex_cube_<SHOTID>_<IFUSLOT>.fits where SHOTID corresponds to the observation and IFUSLOT is the unique IFU identifier. Each cube consists of a Primary HDU and three ImageHDUs: DATA, ERROR, MASK Dimensions are 1036 x 104 x 104 Wavelength Range = 3470Å-5540Å Spatials resolution = 0.5"	Each dex_cube is ~19 MB Total = 7.4 TB 6778 Shotids 431K IFU FITS FILES
detect_index_hdr5.h5 detect_hdrX.h5 cont_hdrX.h5	Internal pipeline raw detections. The detection index H5 contains basic info for all HETDEX detections. Detailed detection info, spectra and fiber info in other H5 files.	1.7 TB
30000/3000000068.jpg … 50900/	Elixer Report Database for all detections in full HPSC2 main and supplemental catalogs	~150 KB each, 587 GB total
01-DataModel+IFU-Index.ipynb 02-DownloadingCubes.ipynb …	Collection of Jupyter notebooks providing data access tutorials and analysis examples	< 100 MB

Figure 13. Structure of the HETDEX Public Data Release 1 (PDR1) IFU data cubes. Each cube is stored in an FITS file named `dex_cube_<SHOTID>_<IFUSLOT>.fits`, where `SHOTID` identifies the observation and `IFUSLOT` the IFU. Cubes have dimensions of 1036 × 104 × 104 (wavelength × spatial × spatial), with a spectral resolution of 2 Å over the range 3470–5540 Å and a spatial sampling of $0\farcs5$. Each file contains a primary HDU and three image HDUs (`DATA`, `ERROR`, `MASK`), with a typical size of 19 MB, totaling ∼7.4 TB for 431,713 IFU data cubes across 6778 observations.

HPSC2 provides one row per source observation and summarizes the key physical and observational properties for each object. HPSC2 includes the source position, redshift, physical size when relevant, and the dust- and aperture-corrected [O II] or Lyα fluxes and luminosities. It contains all necessary information for most scientific analyses. To support more specialized use cases, we also provide the full Detection Information Table, described in Appendix B, which lists every individual HETDEX line emission and continuum detection that passes the quality checks and line-parameter criteria in Sections 2.3 and 3.2. This expanded table contains per-detection measurements and metadata for all HETDEX observations of a source and can be used to trace each HPSC2 entry back to its constituent detections. While the detailed Detection Information Table enables advanced analyses, HPSC2 provides a streamlined and user-friendly summary appropriate for most users.

An HETDEX source, identified by `source_id`, is a collection of all detections at the same on-sky position combined through the detection grouping method described in Section 3.3. If the source is observed more than once, its `source_name` will be the same, but the `source_id` will be different, as will the reported catalog measurements. For each `source_id` row, we report a single representative detection identifier, `detectid`, which may be matched to the Detection Information Table for each source observation in the `detectid` column; this column corresponds to the detection member with the brightest (i.e., smallest) g_{HETDEX} value for all sources that are not LAEs. For LAEs, the highest-S/N Lyα line detection is the selected representative `detectid`. A user may search the Detection Information Table for this representative detectid by the selecting the column `selected_det==True`.

For sources identified as `lae` or `oii`, we provide the objects' Lyα or [O II] line-flux in the columns `flux_lya` and `flux_oii`, along with the corresponding errors on these quantities. As discussed in Section 3.5, for each low-z galaxy, an aperture [O II] line flux is measured: `flux_aper` at `z_hetdex`. This flux is assigned as the source's `flux_oii` if it is a positive value and the major axis of the galaxy (as measured on broadband imaging) is $>2''$; otherwise, the measured line flux comes from the native pipeline `flux` value obtained from the line fit to the extracted spectrum of the brightest `detectid` in the source group. Line fluxes and

Table 6
IFU Metadata Schema

Column Name	Description	Data Type	Units
shotid	Shot ID integer representing a shot: DATEOBS	int64	⋯
ifuslot	IFU slot number (e.g., 046, 063, etc.)	str	⋯
ra_cen	R.A. of IFU center (ICRS J2000)	float32	deg
dec_cen	Decl. of IFU center (ICRS J2000)	float32	deg
flag	Fraction of IFU with usable data (1 = fully usable, 0 = fully flagged)	float32	⋯
flag_badamp	Fraction of IFU unaffected by bad amplifier(s)	float32	⋯
flag_badfib	Fraction of IFU unaffected by bad fiber(s)	float32	⋯
flag_meteor	Fraction of IFU unaffected by meteor trail contamination	float32	⋯
flag_satellite	Fraction of IFU unaffected by satellite trail contamination	float32	⋯
flag_shot	Fraction of IFU unaffected by shot-level flagging	float32	⋯
flag_throughput	Fraction of IFU unaffected by low throughput	float32	⋯
field	Field identifier (e.g., “dex-spring”)	str	⋯
objid	Object ID (string identifier)	str	⋯
date	Date of the observation (YYYYMMDD)	int32	⋯
obsid	Observation index	int32	⋯
ra_shot	R.A. of shot center (ICRS J2000)	float64	deg
dec_shot	Decl. of shot center (ICRS J2000)	float64	deg
pa[a]	Position angle of focal plane	float64	deg
n_ifu	Number of active IFUs in the shot	int32	⋯
fwhm_virus	Seeing FWHM from VIRUS	float32	arcsec
fwhm_virus_err	Error in seeing measurement	float32	arcsec
response_4540	Effective normalized system throughput at 4540 Å in 360 s	float32	⋯
ambtemp	Ambient temperature	float32	°C
datevobs	Observation date string	str	⋯
dewpoint	Dew point	float32	°C
exptime	Exposure time	float32	s
humidity	Relative humidity	float32	%
mjd	Modified Julian Date	float32	d
nstars_fit_fwhm	Number of stars used for seeing estimate	int32	⋯
obsind	Observation index (unused)	int32	⋯
pressure	Barometric pressure	float32	Torr
structaz	Telescope azimuth	float32	deg
time	Start time of exposure	str	UTC
trajcdec	TRAJDEC from telescope header	float32	deg
trajcpa	TRAJPA (position angle) from telescope header	float32	deg
trajcra	TRAJRA from telescope header	float32	deg

Notes.
[a] The `pa` column is the HET tracker rotation angle. The position angle east of north on the celestial sphere is given by $\theta = 360^\circ - (90^\circ + \mathtt{pa} + 1\overset{\circ}{.}55)$.
This table is available in its entirety in machine-readable form at https://hetdex.org/data-results/ and on Zenodo.

associated errors are converted to intrinsic [O II] and Lyα line luminosities (denoted `logL_lya` and `logL_oii`) using our best measured redshift, `z_hetdex`, and the cosmological model defined by Planck Collaboration et al. (2020).

The following files are included in this release:

1. HETDEX Public Source Catalog 2 (HPSC2, columns described in Table 5): `hetdex_sc2_vX.X.dat/.fits`. This table consists of one row per source observation. For each source observation, it provides the source’s J2000 equatorial coordinates, and redshift (`z_hetdex`). Every source is classified into one of the following source_type options: `lae`, `oii`, `agn`, `lzg`, or `star` as described in Section 3.4. For sources with either Lyα or [O II] line emission, the table provides the optimal measurement for the dust-corrected, aperture-corrected flux, and luminosity in `flux_lya`, `flux_oii`, `logL_lya`, and `logL_oii`.
2. HPSC2 spectra in FITS format: `hetdex_sc2_spec_vX.X.fits`. Each row in HPSC2 provides the corresponding 1D extracted spectra in an FITS file format consisting of four Header Data Units (HDUs). Multiple HDUs are included as listed in Table 7. The primary HDU is empty. HDU1:INFO contains a copy of the main HPSC2 catalog. At the same row index for each source in this table, HDU2:SPEC and HDU3:SPEC_ERR contain the aperture-corrected, 1D PSF-weighted extinction-corrected spectra and their associated uncertainties in 10^{-17} erg s^{-1} cm^{-2} Å^{-1}, computed according to the procedure outlined in Section 3.6. The final HDU4:WAVELENGTH is a 1036 array corresponding to the spectral dimension in angstroms. All spectra have the same spectral range from 3470–5540 Å in steps of 2 Å.
3. Supplemental Detection Information Table (columns described in Table 11): `hetdex_sc2_detinfo_vX.X.dat/.fits`. This table contains specific information for every quality-controlled line emission and continuum detection. Every emission-line detection row contains all parameter information, including the emission line’s observed wavelength, its fitted parameters, and measured flux. If the observed wavelength corresponds to a commonly found spectral species[42] at redshift `z_hetdex`, the species is indicated in the column `line_id`. There are also several columns related to imaging counterpart matches, redshift assignments, and emission-line classification as found by ELIXER (D. Davis et al. 2023a). Also included are a number of columns containing details about the specific observation, the instrument, and the detection grouping parameters. A full column description is provided in Table 11. Detailed information concerning this catalog is provided in Appendix B.

4.3. ELiXer Reports

An essential diagnostic for HETDEX detections is provided by the ELIXER reports that are produced for every detection in the quality-controlled HETDEX catalog. A full description of an ELIXER detection report is provided in the Appendix of D. Davis et al. (2023a). These reports provide a summary of ancillary imaging data at the location of the detection. Small cutouts of the amplifier array containing the four highest weighted fibers that contribute to the PSF-extraction of the

---

[42] http://classic.sdss.org/dr6/algorithms/linestable.html

Table 7
Format of the HETDEX Public Source Catalog 2 (HPSC2, Columns Described in Table 5) Spectra FITS File

HDU Name	Type	Dimensions	Description
0:PRIMARY	PrimaryHDU	...	...
1:INFO	BinTableHDU	7367R × 27C	Source information for each catalog source, one row per source observation.
2:SPEC	ImageHDU	(1036 × 1,107,763)	Extinction- and aperture-corrected, PSF-weighted 1D spectrum at the source's representative `detectid` at position `RA_det`, `DEC_det` in units of 10^{-17} erg s^{-1} cm^{-2} Å^{-1}.
3:SPEC_ERR	ImageHDU	(1036 × 1,107,763)	Uncertainty in SPEC.
4:WAVELENGTH	ImageHDU	(1036,)	Wavelength array from 3470–5540 Å in 2 Å bins.

Note. This file contains the full HPSC2 table (Table 5) in HDU1 (also available in a simple .dat ASCII format). The spectrum provided is a 1D PSF-extracted spectrum at the representative `detectid` location. Different spectral extractions can be obtained from HETDEX data cubes directly.

detection. Also included is a full plot of the HETDEX PSF-extracted spectrum, associated error with an expanded view of the spectral region of the detection. These reports are used to visually validate HETDEX detections. We include in this release 3,296,101 ELIXER reports for all detections in full supplemental HETDEX catalog described in Appendix B. This includes every detection in the main HPSC2 catalog.

They can be accessed individually through the HETDEX data mount at https://web.corral.tacc.utexas.edu/hetdex/HETDEX/pdr/pdr1/detect/elixer/. The files are organized based on the first five digits of their detectid into directories to allow for easier navigation. For example, `detectid`=4022202460 is located at https://web.corral.tacc.utexas.edu/hetdex/HETDEX/pdr/pdr1/detect/elixer/40222/4022202460.jpg.

4.4. Access Options

There are two primary ways a user can access HETDEX public data: (1) use the IFU index table to determine dexcubes of interest and then download these cubes locally for personal analysis, or (2) utilize a public JupyterHub access point. In this section we outline these two points of access and present some basic data access examples including making line-flux maps at a spectral region of interest and doing 1D spectral extractions.

4.4.1. Remote Access

For users without direct HETDEX JupyterHub access, data cubes may be downloaded manually. First, you will need to download the IFU index file from https://web.corral.tacc.utexas.edu/hetdex/HETDEX/pdr/pdr1/ifu-index.fits. Then files can be downloaded through a terminal using `wget`:

```
$ wget --cut-dirs=4 -nH -x -i {url}
```

where url = http://web.corral.tacc.utexas.edu/hetdex/HETDEX/pdr/pdr1/datacubes/20190405020/dex_cube_20190405020_034.fits, for example. The options here ensure the same file structure as is located on the host directory system so that only minor edits of the base path are require in the tutorial notebooks.

The script can be easily adapted to download many cubes by looping over lists of `shotid` and `ifuslot` values. For examples of downloading data cubes, including batch parallel downloading options, see https://github.com/HETDEX/dexcube/blob/main/notebooks/02-DownloadingCubes.ipynb and https://github.com/HETDEX/dexcube/blob/main/notebooks/12-BatchDownloads-ForRemoteUsers.ipynb.

4.4.2. JupyterHub Access

A public JupyterHub is available for all public users through the Texas Advanced Computing Center at https://jupyter.tacc.cloud/. The JupyterHub provides access to the full data release and is the most efficient way to perform bulk actions on the survey data. For example, one can perform searches on a catalog of coordinates and extract spectra, create narrowband like images and run other analysis packages. We also include an interactive widget, CubeWidget, to explore HETDEX IFU cubes. An example of CubeWidget is shown in Figure 14.

A TACC account is required for access, but users do not need to be associated with a project allocation. More details on JupyterHub access can be found at https://hetdex.org/data-results/. No persistent storage is offered to public users. Tutorial notebooks can be run and edited, but when the user's server is shut down, the server will reboot back to the original container. It is strongly recommended you should download your work and any data products that you produce.

4.5. Examples

The best approach for a user to become familiar with the data model is to run the notebooks found on the `dexcube` repository.[43] These tutorials introduce the data model and provide simple examples of querying the survey and downloading data cubes of interest for remote users. Users will generally want to do two things with HETDEX data cubes: (1) extract spectra at a list of catalog coordinates and (2) create line-flux maps at specific wavelength and coordinate locations. In this section we demonstrate, using Python, how to query, access and interpolate HETDEX data cubes.

4.5.1. Coordinate Query Example

The following Python snippet demonstrates how to query the HETDEX survey for data cubes near a given sky coordinate. This example searches for IFU observations within $37''$ of a Lyα nebula at $z = 2.53$.

---

[43] https://github.com/HETDEX/dexcube

```python
import os.path as op
from astropy.table import Table
from astropy.coordinates import SkyCoord
import astropy.units as u

# Load IFU index
pdr_dir = '/pathto/pdr1/'
ifu_data = Table.read(op.join(pdr_dir,
    'ifu-index.fits'))

# Create SkyCoord array of IFU centers
ifu_coords =
    SkyCoord(
        ra=ifu_data['ra_cen']*u.deg,
        dec=ifu_data['dec_cen']*u.deg
    )

# Target coordinate (a LAB at z=2.53)
coord =
    SkyCoord(
        ra=228.78581*u.deg,
        dec=51.268036*u.deg
    )

# Select IFUs within 37 arcsec
sel = coord.separation(ifu_coords) < 37*u.arcsec

# Print matching datacube path(s)
for row in ifu_data[sel]:
    shotid  = row['shotid']
    ifuslot = row['ifuslot']
    path =
        op.join(
            pdr_dir,
            'datacubes', str(shotid),
            f'dex_cube_{shotid}_{ifuslot}.fits')
    print(path)
```

4.5.2. Spectral Extractions

Spectra can be extracted from HETDEX data cubes by applying a circular aperture at a given sky position and summing the flux across spatial pixels within the aperture. This process is the most direct way to obtain 1D spectra from the 3D cubes, and it can be adapted depending on the science case. A simple example is shown in Figure 15. Additional examples are provided on the `dexcube` repository.[44]

For more robust use cases, we provide a dedicated spectral extraction script, `get_spectra`.[45] This tool performs proper aperture corrections and allows for a variety of masking options. For example, in Figure 15, we removed the `BADCAL` mask flag so that the AGN's Lyα-emission line at $\lambda = 3563$ Å could be recovered, even though it overlaps with a region that is typically masked. Users are encouraged to adapt the example code to their own needs and apply masking strategies appropriate for their targets.

Finally, while spectra can be extracted directly from the full data cube, efficiency can often be improved by slicing the cube along the spectral axis to limit the wavelength range of interest. This reduces memory load and speeds up the extraction when only a narrow spectral region is required.

```python
from astropy.io import fits
from astropy.wcs import WCS
from astropy.coordinates import SkyCoord
import astropy.units as u
import numpy as np, matplotlib.pyplot as plt

# Open datacube
hdul = fits.open("dex_cube_20181118020_036.fits")
flux, hdr = hdul["DATA"].data, hdul["DATA"].header
wcs, coord = WCS(hdr), SkyCoord(150.23189*u.deg,
↪   2.363963*u.deg)

# Pixel center + radius in pixels
x,y = wcs.celestial.world_to_pixel(coord)
r_pix = 2.0 / wcs.proj_plane_pixel_scales()[0].t⌋
↪   o(u.arcsec).value

# Circular aperture mask
yy,xx = np.indices(flux.shape[1:])
mask = (xx-x)**2 + (yy-y)**2 < r_pix**2

# Extract spectrum
spec = np.nansum(flux[:,mask],axis=1)
wave = hdr['CRVAL3'] + (np.arange(flux.shape[0])
↪   - hdr['CRPIX3'] + 1) * hdr['CDELT3']

plt.plot(wave*1.e10, spec);
↪   plt.xlabel("Wavelength (Å)")
```

4.5.3. Collapsing Cubes to Make Line-flux Maps

To visualize emission at a known wavelength, data cubes can be collapsed along the spectral axis across a small wavelength window around the targeted line. This process produces a pseudonarrowband "line-flux" image. A similar procedure can be performed on an adjacent/wide-pass set of wavelengths to build a "continuum" image, which can be scaled and subtracted to remove underlying continuum. During this interpolate process, masked pixels (bitmasks or zeros) should be set to "NaN" before summation so they don't bias the collapse.

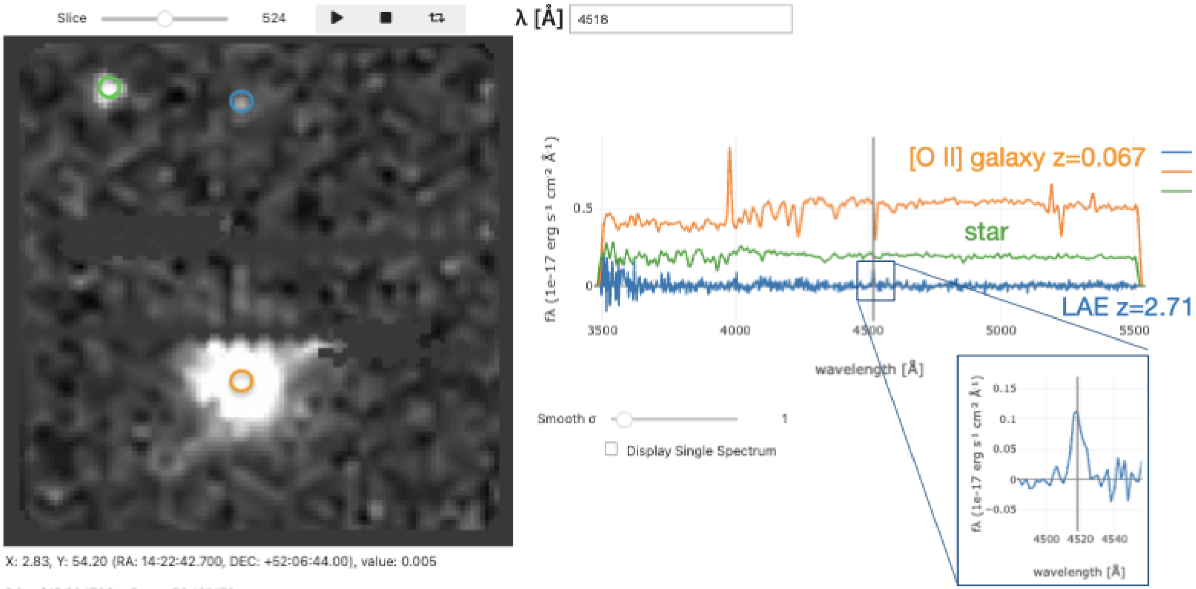
Below is an example of collapsing in a $\pm\Delta\lambda$ window around Lyα at a known observed wavelength. In this case, we collapse $\pm$ 10 Å at $\lambda = 4295.9$ Å. No continuum subtraction is done here, but an example is available online.[46]

---

[44] https://github.com/HETDEX/dexcube/blob/main/notebooks/08-ExtractingSpectra.ipynb

[45] https://github.com/HETDEX/dexcube/blob/main/notebooks/dexcube/get_spec.py

[46] https://github.com/HETDEX/dexcube/blob/main/notebooks/07-CollapsingCubes.ipynb



X: 2.83, Y: 54.20 (RA: 14:22:42.700, DEC: +52:06:44.00), value: 0.005

RA = 215.684736° **Dec** = 52.108173°

Figure 14. Example HETDEX IFU data cube visualization using the `CubeWidget` tool for `shotid = 20190405020`, `ifuslot = 034`. The left panel shows a 2 Å wide spectral region at $\lambda = 4518$ Å, overlaid with circular apertures marking three sources. The right panel displays the extracted spectra: HETDEX_J142245.70+520636.2, an LAE at $z = 2.717$ (blue; detectid 3002870541, S/N=8.7); HETDEX_J142244.63+520702.8, an [O II] galaxy at $z = 0.067$ (orange; detectid 3002870503, S/N=64.5); and HETDEX_J142244.36+520629.4, a foreground star (green; detectid 3090069113). A zoomed-in cutout panel highlights the Lyα-emission-line feature at $\lambda = 4517.5$ Å. There is a noteable masked region in this wavelength slice, caused by a known artifact that occurs at 4500 Å due to CCD readout present in many HETDEX data cubes. Despite this issue, the LAE is clearly detected in both the slice and its spectrum.

```
from astropy.io import fits
from astropy.wcs import WCS
import numpy as np

hdul = fits.open("dex_cube_20180618017_024.fits")

F, M = hdul["DATA"].data, hdul["MASK"].data

# mask data cube
F[(M>0) | (F==0)] = np.nan

wave = hdr['CRVAL3'] + (np.arange(flux.shape[0]) -
↪   hdr['CRPIX3'] + 1) * hdr['CDELT3']

lam0, dlam = 4295.9, 10.0
sel = np.abs(wave - lam0) <= dlam

#collapse along spectral dim
img_nb = np.nansum(F[sel], axis=0)
```

In Figure 16, we extend this simple example by collapsing the data cube into narrowband and continuum images and then visualizing their spatial distribution across multiple epochs for the same changing-look AGN shown in Figure 15. For each observation covering the AGN, we build (i) a line-flux map by collapsing the cube within ±50 Å of Lyα at 3563 Å and (ii) a continuum map from 3800–5200 Å. Square cutouts centered on the AGN are extracted with consistent physical size, but the orientation here is in the IFU plane, so they are not matched in WCS. Lyα narrowband images are overlaid as contours on top of the gray-scale continuum maps made from the HETDEX data cubes. This approach highlights the combined spatial and spectral advantage offered by HETDEX observations.

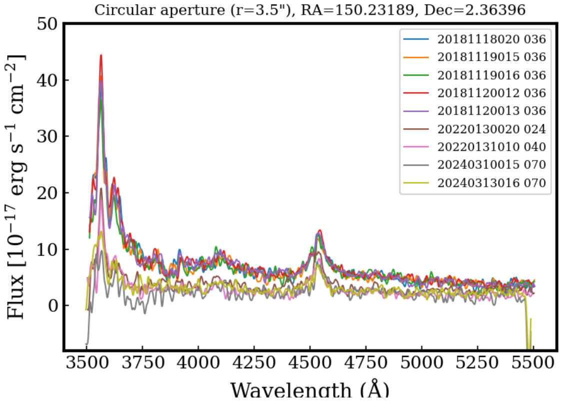


Figure 15. Example aperture-extracted spectra of an AGN in COSMOS at R.A. = 150.23189°, decl. = 2.36396° using a 3.″5 circular aperture. The source is observed multiple times across HETDEX, showing clear evidence of variability. In early observations from 2018, strong Lyα emission is present, but the feature diminishes in later data from 2022 and 2024. This "changing-look" AGN demonstrates the time-domain potential of repeated spectroscopic coverage within the survey.

5. Caveats: How You Can Use These Data and How You Should Not

The optimal use of HETDEX data cubes is to retrieve data related to sources that are in the public source catalog provided with this release (Section 3) or extract spectra or data cube cutouts on your own list of known sources. The HSPC2 spectral FITS file only provides simple 1D, PSF-weighted spectra for the representative `detectid` of the source. Many of the objects, however, are resolved and the HETDEX data cubes provide the spatial information needed to explore their physical properties. For example, E. Mentuch Cooper et al. (2026) reported surface-brightness profile measurements and

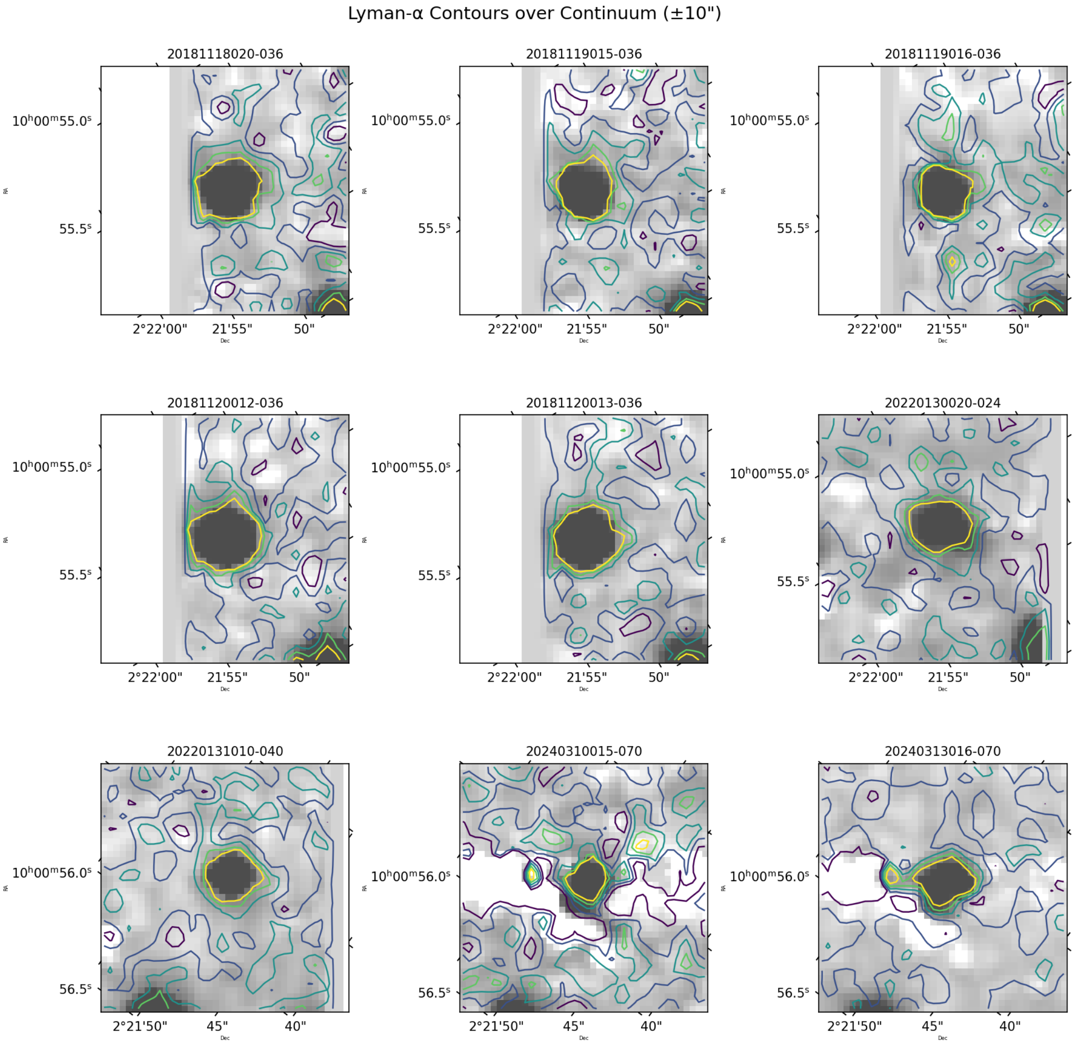


Figure 16. Collapsing the HETDEX data cubes into pseudonarrowband (Lyα, λ = 3563 Å ± 50 Å) and continuum (3800–5200 Å) images for the example changing-look AGN in COSMOS shown in Figure 15. Each panel represents a $20'' \times 20''$ cutout, with gray-scale continuum flux calculated from the HETDEX data cubes and Lyα contours overlaid. These maps are not projected to a common WCS orientation but are in the orientation native to the data cubes. These maps demonstrate both the temporal and spatial information available from repeated IFU coverage in HETDEX.

isophotal sizes for a sample of high-S/N LAEs revealing that over 50% of this population has extended emission.

Extracting spectra at the positions of known objects is also an ideal way to access HETDEX data. Objects with $g < 22$ are sufficiently bright for radial-velocity measurements of Milky Way stars ($\sigma_v < 30\,\mathrm{km\,s^{-1}}$; K. Hawkins et al. 2021), or redshift estimates based on a galaxy's absorption lines (see, for example, E. Mentuch Cooper et al. 2023; G. R. Zeimann et al. 2024; M. H. Debski et al. 2025).

5.1. Beware of Early Data

During its first year (2017 January 1 to 2018 January 1), VIRUS operated with fewer than 20 IFUs. The data quality was poor, with about half of the amplifiers flagged as bad, and detectors were plagued by pox and charge traps. Over time, IFU units were replaced, and new generations of detectors proved to be more stable and clean of artifacts. As a result, early catalog sources and IFU data cubes are of subpar quality. Early observations can be identified using the `shotid` variable. Values beginning with 2017 correspond to early data and should be used with caution. The `shotid` appears in the naming structure of each data cube and is included in all source catalogs. We do not recommend the use of LAE samples from this era without manual inspection. Users should be cautious; these data are included mainly for follow-up studies of external samples.

5.2. Flux Sensitivity and Luminosity Functions

HETDEX observations are very heterogeneous. The sensitivity of each spaxel varies considerably depending on wavelength, detector/amplifier, and observing conditions. The flux sensitivity of HETDEX is quantified through extensive source injection simulations calibrated against the survey noise properties (K. Gebhardt et al. 2021). In these simulations, artificial emission lines are added directly into the fiber spectra across the full wavelength range of VIRUS, and recovery rates are measured using the same detection algorithms applied to the real data.

The survey reaches ∼50% completeness at a line flux of 1.1×10^{-16} erg s^{-1} cm^{-2}, depending on wavelength and seeing conditions, based on simulations. Sensitivity decreases toward the blue end of the spectrograph, but remains relatively uniform redward of $\lambda \simeq 4700$ Å.

We do not include completeness corrections, as they need to be appropriately adapted and tested with the interpolated data cube model presented in PDR1. As a result, PDR1 is not suited for direct luminosity function or number density measurements of faint line emitters without additional modeling. While PDR1 provides the essential raw data products to do so, the completeness models needed to translate catalog detections into population statistics are beyond the scope of the release.

5.3. Source Detection

Users may attempt their own source detection algorithms on the released data cubes; however, this should be done with caution. The primary targets of HETDEX, LAEs, are extremely faint emission-line sources that typically occupy only a couple spatial and spectral resolution elements in the VIRUS IFU data. As a result, noise fluctuations and instrumental artifacts can occasionally mimic emission-line detections.

We have made a concerted effort to provide systematic masking of nonastronomical signals through the mask model described in Section 2.3. In addition, some detections were removed during catalog construction using automated machine learning classifiers and manual vetting procedures that operate at the detection level rather than directly within the cube masks. To improve reproducibility for users performing independent analyses, we introduce an additional bitmask flag (see Section 2.3.14), BADDET, which masks the locations of detections that were rejected by these procedures but were not captured by the previous noise model. This mask is applied as a $5 \times 5 \times 5$ pixel cube (corresponding to $2\farcs5 \times 2\farcs5 \times 10$ Å) centered on the emission-line detection and is included in the MASK array HDU of the data cube FITS files.

Despite these masking efforts, some artifacts may still remain and may be difficult to identify using the data cubes alone. Intermediate data processing products, particularly the 2D spectral images used in S. Mukae et al. (2026), provide additional diagnostic power for distinguishing real emission lines from artifacts. Users developing independent detection algorithms should therefore consider using the provided mask extensions and, where possible, auxiliary data products to minimize contamination from nonastronomical features.

Source detection is performed on the internal fiber database using a PSF-extracted grid search, with noise estimated by propagating fiber uncertainties into PSF-weighted spectral extractions. All detections are assumed to be point sources. Due to the interpolated nature of the public IFU data cubes, one-to-one recovery is not expected, even with masking applied.

To quantify agreement between independent detections on the public data cubes and the released catalog, we re-measured the S/N for 1000 faint emission-line sources drawn from the expanded source catalog (see Appendix B), including those below the primary selection threshold (`p_conf` < 0.5 or `p_cnn` < 0.5). Using the same PSF-weighted Moffat extraction ($\beta = 3.0$, FWHM from the IFU index), and S/N estimator as the HETDEX pipeline (K. Gebhardt et al. 2021), we find a median recovered S/N of 0.75 ± 0.23 relative to catalog values.

Agreement improves at higher significance: sources with $\mathrm{S/N} \gtrsim 6$ approach unity with low scatter, while dispersion increases near the detection threshold ($\mathrm{S/N} \sim 5$). The ~25% deficit arises from correlated noise introduced by spatial interpolation onto the $0\farcs5$ grid, unlike the uncorrelated noise in native fiber spectra. For high-confidence detections (`p_conf` > 0.5 and `p_cnn` > 0.5), 95% of sources with $\mathrm{S/N} \geqslant 6$ are recovered at $\mathrm{S/N} > 4$, decreasing to 81% for $\mathrm{S/N} \geqslant 5.5$ and 5% near $\mathrm{S/N} \sim 5$.

5.4. Local Sky Subtraction and Extended Flux Calibration Limitations

Flux calibration in HETDEX is performed independently within each IFU. While this approach works well for detecting the low-S/N emission lines that are the primary targets of HETDEX, it can lead to inaccurate flux measurements near bright or extended sources such as large galaxies, bright stars, or the occasional planetary nebula that would benefit from larger-scale modeling of the sky.

To address this, we also generate a full-focal-plane sky estimate using the entire VIRUS array for each observation. In practice, however, the current implementation of the full-frame sky model introduces low-level spectral features that degrade the default flux calibration. For this reason, fiber spectra calibrated with the full-frame sky model are not included in this release, although this may be revisited in future data products. Despite these limitations, M. Lujan Niemeyer et al. (2026) successfully used full-frame sky-subtracted data to perform Lyα intensity mapping by statistically removing contaminating components.

Because of these issues, we caution users against performing intensity-mapping analyses (e.g., M. Lujan Niemeyer et al. 2026) or absolute sky-brightness measurements (L. H. Weiss et al. 2025) directly from the public PDR1 data cubes. The local sky subtraction provided here is optimized for source detection but is not suitable for analyses requiring a globally uniform sky model. Users studying very extended sources should also note that the local sky subtraction can suppress low surface-brightness emission on spatial scales comparable to the IFU footprint.

5.5. Stacking

One of the biggest strengths of HETDEX is its potential for studying the properties of faint targets via spectral stacking. By combining thousands to hundreds of thousands of spectra, the faint spectral continuum signal from LAEs is boosted revealing details about its stellar population (D. Davis et al. 2023b) and signatures of neutral hydrogen in the circumgalactic medium (L. H. Weiss et al. 2024; M. Mirza Khanlari et al. 2025).

When stacking HETDEX spectra, we account for faint sky residuals by constructing a correction spectrum from stacks of 200 randomly selected empty-sky apertures per field (D. Davis et al. 2023b). This residual stack, typically <1% of the sky level in individual spectra, is subtracted from source stacks to remove small calibration offsets that only become significant when combining large numbers of spectra. This measured offset, which is applied in published HETDEX stacking analyses, such as D. Davis et al. (2023b), L. H. Weiss et al. (2024), and M. Mirza Khanlari et al. (2025), is not included in the HETDEX data cubes.

6. Summary

HETDEX is a medium-wide area, IFU spectroscopic survey that covers the wavelength range 3470–5540 Å at a resolving power of $750 < R < 950$. Main survey operations are now complete with 86.67 deg^2 of noncontiguous sky coverage mapped across $\sim$540 deg^2.

This data release provides access to all science-quality IFU data cubes from the main HETDEX survey. It consists of 431,713 data cubes, reduced from >600 million fiber spectra obtained by the VIRUS instrument on the HET. The data cubes are offered at a spatial resolution of $0\farcs5$, and a spectral resolution of 2Å. Because the data cubes use a local sky-subtraction procedure optimized for detecting faint emission-line sources, they are not designed for analyses requiring absolute surface-brightness measurements (e.g., intensity mapping) or for studies of very extended nearby galaxies.

This paper describes the public release of HETDEX Public Source Catalog 2 (HPSC2), which contains 426,654 LAEs, 491,411 [O II]-emitting galaxies, 150,608 stars, 19,457 low-z continuum selected galaxies, and 18,303 AGNs. By utilizing a three-pronged classification approach, we provide robust spectroscopic redshifts and classifications for the entire catalog. When compared to spectroscopic redshifts from external catalogs, 94.1% of the sources are within $|\Delta z| < 0.02$.

Data access and details about the catalog can be found at http://hetdex.org. The link to the data at TACC is at https://web.corral.tacc.utexas.edu/hetdex/HETDEX/pdr/pdr1/. A copy of the HETDEX Public Source Catalog 2 is also publicly available via Zenodo (doi:10.5281/zenodo.19581262), ensuring long-term access and reproducibility.

Papers making use of the HETDEX dataset should include citations for the HET (L. W. Ramsey et al. 1998; G. J. Hill et al. 2021), the VIRUS instrument (G. J. Hill et al. 2021), and the HETDEX survey (K. Gebhardt et al. 2021), and acknowledgments as specified at https://hetdex.org/papers/.

Acknowledgments

HETDEX is led by the University of Texas at Austin McDonald Observatory and Department of Astronomy with participation from the Ludwig-Maximilians-Universität München, Max-Planck-Institut für Extraterrestrische Physik (MPE), Leibniz-Institut für Astrophysik Potsdam (AIP), Texas A&M University, The Pennsylvania State University, Institut für Astrophysik Göttingen, The University of Oxford, Max-Planck-Institut für Astrophysik (MPA), The University of Tokyo, and Missouri University of Science and Technology.

Observations for HETDEX were obtained with the Hobby–Eberly Telescope (HET), which is a joint project of the University of Texas at Austin, the Pennsylvania State University, Ludwig-Maximilians-Universität München, and Georg-August-Universität Göttingen. The HET is named in honor of its principal benefactors, William P. Hobby and Robert E. Eberly.

The Visible Integral-field Replicable Unit Spectrograph (VIRUS) was used for HETDEX observations. VIRUS is a joint project of the University of Texas at Austin, Leibniz-Institut für Astrophysik Potsdam (AIP), Texas A&M University (TAMU), Max-Planck-Institut für Extraterrestrische Physik (MPE), Ludwig-Maximilians-Universität Muenchen, Pennsylvania State University, Institut für Astrophysik Göttingen, University of Oxford, and the Max-Planck-Institut für Astrophysik (MPA). In addition to Institutional support, VIRUS was partially funded by the National Science Foundation, the State of Texas, and generous support from private individuals and foundations.

The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing high-performance computing, visualization, and storage resources that have contributed to the research results reported within this paper (http://www.tacc.utexas.edu).

We acknowledge RAIC Labs (https://raiclabs.com) for providing access to the RAIC platform, which enabled the efficient classification and identification of artifacts, meteor tracks, and satellite streaks within the HETDEX dataset.

Dark Energy Explorers is recognized as an official NASA Citizen Science partner. This publication utilizes data generated through the Zooniverse.org platform, the development of which is supported by generous funding, including a Global Impact Award from Google and a grant from the Alfred P. Sloan Foundation.

The Institute for Gravitation and the Cosmos is supported by the Eberly College of Science and the Office of the Senior Vice President for Research at the Pennsylvania State University. The Kavli IPMU is supported by World Premier International Research Center Initiative (WPI), MEXT, Japan.

This work makes use of the Sloan Digital Sky Survey IV, with funding provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS website is www.sdss.org.

In addition to Institutional support, HETDEX is funded by the National Science Foundation (grant AST-0926815), the State of Texas, the US Air Force (AFRL FA9451-04-2-0355), and generous support from private individuals and foundations. K.G. acknowledges support from NSF-2008793. S.S. and H.K. acknowledge support from the National Science Foundation under grants NSF-2219212 and NSF-2511145.

Facility: HET

Software: ASTROPY (Astropy Collaboration et al. 2018), PYTHON (G. Van Rossum & F. L. Drake 2009), NUMPY (C. R. Harris et al. 2020), SCIPY (P. Virtanen et al. 2020), SCIKIT-LEARN (F. Pedregosa et al. 2011), HETDEX-API (https://github.com/HETDEX/hetdex_api), ELIXER (https://github.com/HETDEX/elixer; D. Davis et al. 2023a), DIAGNOSE (https://github.com/grzeimann/Diagnose; M. Debski & G. Zeimann 2024), PHOTUTILS (L. Bradley et al. 2021), DUSTMAPS (G. Green 2018), EXTINCTION (https://github.com/kbarbary/extinction).

Appendix A
Raw Emission-line and Continuum Detections

In addition to the IFU data cubes, this data release includes the raw databases of all emission-line and continuum detections from the HETDEX internal data pipeline. Emission lines capture discrete spectral features such as Lyα and [O II], while continuum detections identify bright broadband sources such as galaxies and stars. Both are distributed as HDF5 files under `pdr1/detect/`. Each file contains three main tables: `Detections`, `Spectra`, and `Fibers`. Continuum catalogs use the same structure, though many line-specific columns are not applicable and are filled with 0.0 values.

The raw detection HDF5 files provide a structured record of every emission-line or continuum detection found by the HETDEX pipeline, along with its associated spectra and

Table 8
Detection Catalog Files in PDR1

File	Description	Date Range	Number of Detections
`cont_hdr3.h5`	Continuum detections	2017-01-01 to 2021-08-31	297,877
`cont_hdr4.h5`	Continuum detections	2021-09-01 to 2023-08-31	276,465
`cont_hdr5.h5`	Continuum detections	2023-09-01 to 2024-07-31	107,725
`detect_hdr3.h5`	Emission-line detections	2017-01-01 to 2021-08-31	13,843,051
`detect_hdr4.h5`	Emission-line detections	2021-09-01 to 2023-08-31	14,180,087
`detect_hdr5.h5`	Emission-line detections	2023-09-01 to 2024-07-31	3,938,578
`detect_index_hdr5.h5`	Index for HDR5 detections	2023-09-01 to 2024-07-31	32,643,783
`elixer_hdr345_cluster_cat.h5`	Filtered subset of clustered detections	2017-01-01 to 2024-07-31	4,710,195

fiber-level extractions. Each file is organized into three linked groups: *Detections*, *Spectra*, and *Fibers*.

The *Detections* group contains one row per candidate detection, including positional information (R.A., decl., IFU coordinates), best-fit line measurements (central wavelength, flux, line width, continuum), and quality metrics (S/N, χ^2 values, noise estimates). Instrumental identifiers (e.g., `shotid`, `specid`, `ifuslot`) and fiber metadata (brightest fiber ID, aperture correction factors) are also stored here.

The *Spectra* group links to each detection via `detectid` and records the 1D extracted spectrum. Both aperture-corrected and uncorrected versions are provided, with associated errors. Additional arrays include the wavelength grid, raw counts, PSF-weighted counts, and the wavelength-dependent aperture correction. No dust-extinction correction is applied to the raw database spectral data.

The *Fibers* group stores the fiber-level information for each fiber contained in a single detection. This dataset includes fiber coordinates (on-sky and within the IFU), CCD positions, weights, flags, time stamps, and instrument identifiers. Each fiber entry is linked back to its parent detection, allowing for reconstruction of the full fiber ensemble contributing to a given source.

Together, these three groups allow users to trace each detection from the raw fiber spectra, through PSF-weighted extraction, to the final line and continuum measurements.

Two additional tables are contained in this directory. The Detection Index table provides a look-up reference so that the full collection of raw detection databases may be queried. The column names for the Detection Index table are given in Table 9. Querying may be done using healpix with a resolution of $N_{\rm side} = 2^{15}$.

A.1. Detection Methods

Emission-line detections are identified through a two-stage grid search of the data cubes. An initial coarse search is performed in $0\farcs5$ spatial steps and 8 Å spectral steps using a Gaussian line profile with the instrumental resolution ($\sigma = 1.7$ Å). Continuum emission is subtracted locally, and candidate lines with $\mathrm{S/N} > 4$ and $\chi^2 < 3$ are retained. A refined raster with $0\farcs15$ spatial steps and an unconstrained σ then optimizes the fit, with duplicates within $3''$ and 3 Å merged. Continuum detections are identified by searching for fibers with more than 50 counts in either a blue (3700–3900 Å) or red (5100–5300 Å) window (corresponding to $g \sim 22.5$). Candidate positions are refined with a PSF raster fit, and spectra are extracted at the best-fit location.

Table 9
Schema of the DetectIndex Table (file: detect_index_hdr5.h5)

Column	Type	Notes
`detectid`	Int64	Unique detection identifier. Indexed (CSI)
`shotid`	Int64	Observation identifier. Indexed (CSI)
`ra`	Float32	R.A. (ICRS J2000 deg). Indexed (CSI)
`dec`	Float32	Decl. (ICRS J2000 deg)
`wave`	Float32	Central wavelength (Å)
`sn`	Float32	Signal-to-noise ratio of the detection.
`healpix`	Int64	HEALPix index of detection location. Indexed (CSI)
`det_type`	String(4)	Detection type (`line` or `cont`)
`survey`	String(4)	Survey flag (e.g., HDR2, HDR3, HDR4, HDR5)
`fiber_id`	String(38)	Unique fiber identifier string

Note. The table contains 32,643,783 rows.

Each row in the catalogs is assigned a unique integer `detectid`. A single astronomical source may have multiple detection entries if it is spatially extended, has multiple emission lines, or is bright enough to be detected in via its line and continuum emission. Additionally, the same emission line observed in separate exposures will receive independent `detectid` identifiers. Users should therefore treat the detection catalogs as a superset of measurements from which higher-level source catalogs can be constructed.

A.2. Access Example

The catalogs are stored in HDF5 format and can be accessed directly with PYTABLES (F. Alted et al. 2002). Each internal data release has its own set of detections for the date range, as outlined in Table 8. To streamline querying, a master index file `detect_index_hdr5.h5` can first be queried either by coordinates, detectid, or other variables (see Table 9) to determine the detection's survey and detection type (`line` versus `cont`). This query indicates which survey HDF5 file to open (e.g., `detect/detect_{survey}.h5` for line detections, or `cont/cont_{survey}.h5` for continuum detections). The matching rows in the `/Detections` and `/Spectra` tables are then read to obtain sky coordinates, `shotid/ifuslot`, the central wavelength for line detections, and the pipeline 1D spectrum and errors.

```python
import tables as tb
import os.path as op
import numpy as np
from astropy.coordinates import SkyCoord

# --- Inputs ---
detectid_obj = 3002870541   # <- example detectid (int)

pdr_dir = '/home/jovyan/Hobby-Eberly-Public/HETDEX/internal/pdr1/'

# 1) Open detection index to access pipeline detection index for the full survey
DI = tb.open_file(op.join(pdr_dir, 'detect/detect_index_hdr5.h5'), 'r')

det_file_info = DI.root.DetectIndex.read_where(f'detectid == {detectid_obj}')[0]
# PyTables string columns are byte strings; we use \texttt{.decode()} when needed.
survey = det_file_info['survey'].decode()
det_type = det_file_info['det_type'].decode()

# 2) Choose file family based on detection type
if det_type == 'line':
    det_file_type = 'detect'
else:
    det_file_type = 'cont'

# 3) Spectra and IFU/shotid information are accessed from H5 files for
#    each survey/detection-type combination

det_file = tb.open_file(op.join(pdr_dir, f'{det_file_type}/{det_file_type}_{survey}.h5'), 'r')

# Pull basic detection info
det_info = det_file.root.Detections.read_where(f'detectid == {detectid_obj}')[0]
coord = SkyCoord(ra=det_info['ra'], dec=det_info['dec'], unit='deg')
shotid = det_info['shotid']
ifuslot = det_info['ifuslot'].decode()

cw = None
if det_type == 'line':
    cw = det_info['wave']   # central wavelength [Å] for line detections

# 4) Load the pipeline 1D spectrum row for this detection
spec_pipeline = det_file.root.Spectra.read_where(f'detectid == {detectid_obj}')[0]

# 5) Fix for calibration issue in pipeline spectra

corr_file = np.loadtxt(op.join(pdr_dir, 'detect/wdcor.txt'))
spectrum = spec_pipeline['spec1d']/corr_file[:,1]

# 6) If desired convert to a flux density
spectrum /= 2.0

# (Optional) clean-up
# DI.close()
# det_file.close()
```

For emission-line detections, `cw` stores the fitted line center in angstroms. The pipeline spectrum (column `spec1d`, more information in Table 10) contains PSF-weighted flux in units of 10^{-17} erg s^{-1} cm^{-2} in 2 Å bins (users should divide by 2 to obtain conventional flux density units of 10^{-17} erg s^{-1} cm^{-2} Å^{-1}, as is provided in the 1D spectra provided in HPSC2). A calibration error is present in the raw H5 detection spectra. Output spectra should be normalized by the spectrum "/pdr1/detect/wdcor.txt" as shown in the example above.

Continuum catalogs (`cont_hdr`*) have the same structure, although many line-specific columns (e.g., `wave`, `flux`, `sn`) are zero.

Table 10
Column Descriptions in the Raw Detections HDF5 Table Groups

Table	Column	Description
Detections	`detectid`	Unique integer detection ID
	`inputid`	Input extraction ID (`DATEvOBS_inputname`)
	`shotid`	Shot identifier (`YYYYMMDDnnn`)
	`date`	Observation date (YYYYMMDD)
	`obsid`	Observation number on that date
	`ra`, `dec`	R.A. and decl. (ICRS J2000 deg)
	`wave`, `wave_err`	Line central wavelength (Å) and uncertainty
	`flux`, `flux_err`	Line flux 10^{-17} erg s^{-1} cm^{-2} and uncertainty
	`linewidth`, `linewidth_err`	Gaussian σ (Å) and uncertainty
	`continuum`, `continuum_err`	Continuum flux density 10^{-17} erg s^{-1} cm^{-2} $Å^{-1}$ and uncertainty
	`sn`, `sn_err`	Emission-line signal-to-noise ratio and uncertainty
	`chi2`, `chi2_err`	χ^2 of line fit and uncertainty
	`fiber_id`	Brightest fiber identifier
	`multiframe`	`specid/ifuslot/ifuid/amp` identifier
	`fibnum`	Fiber number
	`expnum`	Dither exposure number
	`x_raw`, `y_raw`	CCD coordinates of detection
	`x_ifu`, `y_ifu`	IFU position (arcseconds)
	`specid`, `ifuslot`, `ifuid`, `amp`	Instrument identifiers
	`weight`	Weight of brightest fiber
	`apcor`	Aperture correction factor
	`sn_cen`	S/N summed across 7 pixels at line center
	`flux_noise_1sigma`	1σ noise estimate on flux
	`sn_3fib`, `sn_3fib_cen`	S/N for three-fiber extractions
Spectra	`detectid`	Unique integer detection ID (links to Detections)
	`wave1d`	Wavelength array (Å)
	`spec1d`, `spec1d_err`	Flux-calibrated, PSF-weighted, aperture-corrected spectrum and error in 10^{-17} erg s^{-1} cm^{-2} in 2 Å bins
	`counts1d`, `counts_err`	Counts and uncertainty
	`apsum_counts`, `apsum_counts_err`	PSF-weighted counts and uncertainty
	`spec1d_nc`, `spec1d_nc_err`	Spectrum without aperture correction and uncertainty
	`apcor`	Applied aperture correction (wavelength dependent)
Fibers	`detectid`	Unique integer detection ID (links to Detections)
	`ra`, `dec`	Fiber R.A., decl. (ICRS J2000 deg)
	`fiber_id`	Fiber identifier string
	`x_ifu`, `y_ifu`	Fiber position in IFU (arcseconds)
	`multiframe`	`specid/ifuslot/ifuid/amp` identifier
	`fibnum`	Fiber number
	`expnum`	Dither exposure number
	`distance`	Distance to requested R.A., decl. (arcseconds)
	`wavein`	Input wavelength for extraction
	`timestamp`	Exposure time stamp
	`date`, `obsid`	Observation date and number
	`flag`	Fiber flag
	`weight`	Weight applied in PSF sum
	`ADC`	Five-element array of ADC values
	`specid`, `ifuslot`, `ifuid`, `amp`	Instrument identifiers
	`x_raw`, `y_raw`	CCD coordinates of fiber

Appendix B
HPSC2 Supplemental Detection Information Table

This appendix describes the supplemental Detection Information Table, noted by the `hetdex_sc2_detinfo_vX.X.fits/.dat` file name, which contains information for every emission-line and continuum detection found by the HETDEX pipeline (see Section 3.1). As described in Section 3.3, an HETDEX source can be composed of a collection of line-emission and continuum-emission detections. The *HETDEX Public Source Catalog 2* (HPSC2), outlined in Table 5, provides a simplified version of the Detection Information Table with one row per source observation; it provides basic information about a source such as coordinates, redshift, $g_{\rm HETDEX}$ magnitude, and the [O II]/Lyα line flux and luminosity, where applicable. The Detection Information Table presented in this appendix is expanded to provide additional information for every detection in a source. Many columns are the same to those in HPSC2, e.g., `source_id`, `source_name`, `RA`, `DEC`, and `z_hetdex`. Additional information is provided regarding line fit parameter

Table 11
Detection Information Table Column Descriptions

Name	Description
source_id	HETDEX Source Identifier
source_name	HETDEX IAU designation
RA	source_id R.A. (ICRS J2000 deg)
DEC	source_id decl. (ICRS J2000 deg)
z_hetdex	HETDEX spectroscopic redshift
z_hetdex_src	HETDEX spectroscopic redshift source
z_hetdex_conf	0 to 1 confidence HETDEX spectroscopic redshift source. Not well calibrated
source_type	options are `star`, `lae`, `agn`, `lzg`, `oii`, and `none`
detectid	emission-line or continuum detection ID
selected_det	best detectid for Lyα flux or [O II] line flux
det_type	detection type: “line” or “cont”
shotid	integer represent observation ID: int(date+obsid)
ifuslot	string identifier of IFU location in focal plane
line_id	line identification at observed wavelength (wave) assuming redshift of `z_hetdex`
p_conf	LAE/faint O II emitter confidence score from RF classifier. 1 = high confidence, 0 = low confidence
p_cnn	LAE/faint O II emitter confidence score based on CNN classifier. 1 = high confidence, 0 = low confidence
RA_det	detectid R.A. (ICRS J2000 deg)
DEC_det	detectid decl. (ICRS J2000 deg)
src_separation	separation in arcseconds between the detectid (RA_det, DEC_det) and the source_id center (R.A., decl.)
n_members	number of detections in the source group. Note that some may be missing from the final catalog due to selections
gmag	SDSS-g magnitude measured in HETDEX spectrum
Av	applied extinction correction in the V band
ebv	applied selective extinction
wave	central wavelength of line emission (Å)
wave_err	MCMC error in fitted central wavelength (Å)
flux	extinction-corrected line flux 10^{-17} erg s^{-1} cm^{-2}
flux_err	MCMC error in extinction-corrected line flux
flux_obs	observed line flux 10^{-17} erg s^{-1} cm^{-2}
flux_obs_err	MCMC error in observed line flux
flux_aper	extinction-corrected, [O II] line flux measured in elliptical galaxy aperture in 10^{-17} erg s^{-1} cm^{-2}
flux_aper_err	error in flux_aper
flux_aper_obs	[O II] line flux measured in elliptical galaxy aperture in 10^{-17} erg s^{-1} cm^{-2}
flux_aper_obs_err	error in flag_aper_obs
flag_aper	1 = aperture line flux used for `lum_oii`, 0 = PSF line flux used from "flux" column
flux_lya	Lyα flux in 10^{-17} erg s^{-1} cm^{-2} from dust-corrected `flux`
flux_lya_err	uncertainty in `flux_lya`
flux_oii	[O II] flux in 10^{-17} erg s^{-1} cm^{-2} from `flux` if `flag_aper=0` or `flux_aper` if `flag_aper=1`
flux_oii_err	uncertainty in `flux_oii`
logL_lya	$\log_{10}$ Lyα luminosity (erg s^{-1}) from extinction-corrected `flux`
logL_lya_err	uncertainty in `logL_lya`
logL_oii	$\log_{10}$ [O II] luminosity (erg s^{-1}) from `flux` if `flag_aper=0` or from `flux_aper` if `flag_aper=1`
logL_oii_err	uncertainty in `logL_oii`
sigma	sigma linewidth in Gaussian line fit (Å)
sigma_err	MCMC error in sigma linewidth (Å)
continuum	local fitted extinction-corrected continuum in 10^{-17} erg s^{-1} cm^{-2} Å^{-1}
continuum_err	MCMC error in continuum in 10^{-17} erg s^{-1} cm^{-2} Å^{-1}
continuum_obs	local fitted observed continuum in 10^{-17} erg s^{-1} cm^{-2} Å^{-1}
continuum_obs_err	MCMC error in continuum in 10^{-17} erg s^{-1} cm^{-2} Å^{-1}
sn	signal-to-noise ratio for line emission
sn_err	MCMC error in signal-to-noise
chi2	reduced χ^2 quality of line fit
chi2_err	MCMC uncertainty in reduced χ^2
flux_noise_1sigma_obs	observed 1σ flux sensitivity in 10^{-17} erg s^{-1} cm^{-2}
flux_noise_1sigma	extinction-corrected 1σ flux sensitivity in 10^{-17} erg s^{-1} cm^{-2}
apcor	aperture correction applied to spectrum at 4500 Å
counterpart_mag	selected closest counterpart magnitude from source extraction on image data
counterpart_mag_err	uncertainty in counterpart_mag
counterpart_dist	distance to closest counterpart
counterpart_catalog_name	image catalog source of counterpart
counterpart_filter_name	image filter of counterpart
plya_classification	ELIXER confidence line in Lyα ranges 0 to 1 (1 = high probility line is Lyα)

Table 11
(Continued)

Name	Description
z_elixer	ELIXER best redshift
best_pz	confidence in best_z
z_diagnose	best-fit redshift from `Diagnose`
cls_diagnose	best classification from `Diagnose`. Options are "STAR," "GALAXY," "QSO," and "UNKNOWN"
stellartype	`Diagnose` spectral type classification for stars
agn_flag	−1 not an AGN, 0 broad-line source but not confirmed AGN, 1 confident AGN
wave_group_id	id for 3D friend-of-friends (FOF) clustering at common R.A., decl., wave
wave_group_a	semimajor axis in arcseconds from 3D FOF clustering
wave_group_b	semiminor axis in arcseconds from 3D FOF clustering
wave_group_pa	positional angle from 3D FOF clustering
wave_group_ra	mean R.A. from 3D FOF clustering (ICRS J2000 deg)
wave_group_dec	mean decl. from 3D FOF clustering (ICRS J2000 deg)
wave_group_wave	mean wavelength from 3D FOF clustering
fwhm	measured seeing of the observation in arcseconds
throughput	relative spectral response at 4540 assuming a 360 s nominal exposure
field	field ID: cosmos, goods-n, dex-fall, dex-spring, nep, ssa22
date	date
obsid	observation number
multiframe	string identifier for the ifuslot/specid/ifuid/amp combination
fiber_id	string identifier for the highest weight fiber
weight	flux weight of the highest weight fiber
x_raw	x-value on the CCD of the detection (ds9 x-value)
y_raw	y-value on the CCD of the detection (ds9 y-value)
x_ifu	x-position in the ifu in arcseconds
y_ifu	y-position in the ifu in arcseconds
ra_aper	R.A. of aperture center of imaging counterpart (ICRS J2000 deg)
dec_aper	decl. of aperture center of imaging counterpart (ICRS J2000 deg)
catalog_name_aper	imaging source for measuring [O II] resolved apertures
filter_name_aper	filter of imaging used for measuring [O II] resolved apertures
dist_aper	distance between aperture center and detectid position in arcseconds
mag_aper	photometric magnitude in aperture in imaging source
mag_aper_err	photometric magnitude error in aperture in imaging source
major	major axis of aperture ellipse of resolved [O II] galaxy defined by imaging
minor	minor axis of aperture ellipse of resolved [O II] galaxy defined by imaging
theta	angle in aperture ellipse

Note. Bad values are −999.0 for floating-point columns and `n/a` for string columns.This table is available in its entirety in machine-readable form at https://hetdex.org/data-results/ and on Zenodo.

information, including the specific position of the detection (`RA_det`, `Dec_det`) and wavelength (`wave`) for the detection, the detection's line width, (σ: `sigma`), continuum-subtracted line flux, and the local continuum measurement. Each observed wavelength is checked for a rest-frame match to a common line species at `z_hetdex`. Specifically, we consider C III] λ1909, C IVλ1550, Hβ, Hδ, Hγ, He II, Lyα, [O II] λ3727, and [O III] $\lambda\lambda$4959, 5007.[47] If a match is found, it is listed in `line_id`. Not all detections have a `line_id`, as some HETDEX detections can result from discontinuities in a spectrum or calibration issues. We attempt to mitigate these by excluding high line width sources that are not selected as the main detection (ie., `selected_det==`True) of a source. Other information as described in the text is also provided, including (1) detection group information from 3D and 2D FOF detection clustering, (2) ELIXER imaging counterpart information, (3) specific observation parameters such as image quality (e.g., `fwhm`), `shotid`, `date`, `obsid`, `field`), and (4) specific information related to the highest weight fiber in the spectral extraction (such as `multiframe`, `fiber_id`, `weight` and others). The detectid whose spectrum is included in HPSC2 for the source (typically the brightest magnitude detection) is identified by `selected_det==True`. The description of all parameters is provided in Table 11.

ORCID iDs

Erin Mentuch Cooper https://orcid.org/0000-0002-2307-0146
Karl Gebhardt https://orcid.org/0000-0002-8433-8185
Dustin Davis https://orcid.org/0000-0002-8925-9769
Chenxu Liu (刘辰旭) https://orcid.org/0000-0001-5561-2010
Barbara G. Castanheira https://orcid.org/0000-0001-7010-7637
Owen Chase https://orcid.org/0000-0002-0304-5701
Óscar A. Chávez Ortiz https://orcid.org/0000-0003-2332-5505
Robin Ciardullo https://orcid.org/0000-0002-1328-0211
Olivia Curtis https://orcid.org/0000-0002-0212-4563
Delaney A. Dunne https://orcid.org/0000-0002-5223-8315
Neal J. Evans, II https://orcid.org/0000-0001-5175-1777
Daniel J. Farrow https://orcid.org/0000-0003-2575-0652
Maximilian Fabricius https://orcid.org/0000-0002-7025-6058
Steven L. Finkelstein https://orcid.org/0000-0001-8519-1130

---

[47] http://classic.sdss.org/dr6/algorithms/linestable.html

Caryl Gronwall https://orcid.org/0000-0001-6842-2371
Nathaniel J. Hamme https://orcid.org/0009-0003-7103-9076
Gary J. Hill https://orcid.org/0000-0001-6717-7685
Lindsay R. House https://orcid.org/0000-0002-1496-6514
Matt J. Jarvis https://orcid.org/0000-0001-7039-9078
Donghui Jeong https://orcid.org/0000-0002-8434-979X
Andreas Kelz https://orcid.org/0000-0002-2196-4699
Eiichiro Komatsu https://orcid.org/0000-0002-0136-2404
Mahan Mirza Khanlari https://orcid.org/0009-0003-1893-9526
Hasti Khoraminezhad https://orcid.org/0000-0001-5610-4405
Wolfram Kollatschny https://orcid.org/0000-0002-0417-1494
Maja Lujan Niemeyer https://orcid.org/0000-0002-6907-8370
Hanshin Lee https://orcid.org/0000-0002-3559-5310
Phillip MacQueen https://orcid.org/0009-0006-7054-0100
Deeshani Mitra https://orcid.org/0000-0002-1350-019X
Shiro Mukae https://orcid.org/0000-0003-3823-8279
Masami Ouchi https://orcid.org/0000-0002-1049-6658
Meredith C. Powell https://orcid.org/0000-0003-2284-8603
Mahdi Qezlou https://orcid.org/0000-0001-7066-1240
Shun Saito https://orcid.org/0000-0002-6186-5476
Donald P. Schneider https://orcid.org/0000-0001-7240-7449
Laurel Weiss https://orcid.org/0000-0002-4974-1243
Lutz Wisotzki https://orcid.org/0000-0003-2977-423X
Gregory R. Zeimann https://orcid.org/0000-0003-2307-0629

References

Abareshi, B., Aguilar, J., Ahlen, S., et al. 2022, AJ, 164, 207
Adams, J. J., Blanc, G. A., Hill, G. J., et al. 2011, ApJS, 192, 5
Alted, F., Vilata, I., Prater, S., et al. 2002, PyTables: Hierarchical Datasets in Python, v3.9.2, https://www.pytables.org/
Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123
Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33
Bacon, R., Brinchmann, J., Richard, J., et al. 2015, A&A, 575, A75
Barbary, K. 2016, JOSS, 1, 58
Benitez, N., Dupke, R., Moles, M., et al. 2014, arXiv:1403.5237
Bonoli, S., Marín-Franch, A., Varela, J., et al. 2021, A&A, 653, A31
Bradley, L., Sipőcz, B., Robitaille, T., et al. 2021, Python, v1.3.0, Zenodo, doi:10.5281/zenodo.5796924
Bundy, K., Bershady, M. A., Law, D. R., et al. 2015, ApJ, 798, 7
Cenarro, A. J., Peletier, R. F., Sánchez-Blázquez, P., et al. 2007, MNRAS, 374, 664
Chávez Ortiz, Ó. A., Finkelstein, S. L., Davis, D., et al. 2023, ApJ, 952, 110
Chiang, C.-T., Wullstein, P., Jeong, D., et al. 2013, JCAP, 2013, 030
Cowie, L. L., & Hu, E. M. 1998, AJ, 115, 1319
Davis, D., Gebhardt, K., Cooper, E. M., et al. 2023a, ApJ, 946, 86
Davis, D., Gebhardt, K., Cooper, E. M., et al. 2023b, ApJ, 954, 209
Dawson, K. S., Kneib, J.-P., Percival, W. J., et al. 2016, AJ, 151, 44
Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, AJ, 145, 10
Debski, M., & Zeimann, G. 2024, Diagnose: Spectral classification code, Astrophysics Source Code Library, ascl:2411.020
Debski, M. H., Zeimann, G. R., Hill, G. J., et al. 2025, ApJ, 978, 101
de Vaucouleurs, G., de Vaucouleurs, A., Corwin, H. G. J., et al. 1991, Third Reference Catalogue of Bright Galaxies (Springer)
DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016, arXiv:1611.00036
Dey, A., Schlegel, D. J., Lang, D., et al. 2019, AJ, 157, 168
Dickinson, M., Giavalisco, M. & GOODS Team 2003, in The Great Observations Origins Deep Survey, ed. R. Bender & A. Renzini (Springer-Verlag), 324
Drory, N., MacDonald, N., Bershady, M. A., et al. 2015, AJ, 149, 77
Eriksen, M., Alarcon, A., Gaztanaga, E., et al. 2019, MNRAS, 484, 4200
Falcón-Barroso, J., Sánchez-Blázquez, P., Vazdekis, A., et al. 2011, A&A, 532, A95
Firestone, N. M., Gawiser, E., Ramakrishnan, V., et al. 2024, ApJ, 974, 217
Fitzpatrick, E. L. 1999, PASP, 111, 63
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, A&A, 616, A1
Gebhardt, K., Mentuch Cooper, E., Ciardullo, R., et al. 2021, ApJ, 923, 217
Green, G. 2018, JOSS, 3, 695
Greisen, E. W., Calabretta, M. R., Valdes, F. G., & Allen, S. L. 2006, A&A, 446, 747
Gronwall, C., Ciardullo, R., Hickey, T., et al. 2007, ApJ, 667, 79
Hagen, A., Zeimann, G. R., Behrens, C., et al. 2016, ApJ, 817, 79
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Natur, 585, 357
Hawkins, K., Zeimann, G., Sneden, C., et al. 2021, ApJ, 911, 108
Hill, G. J., Lee, H., MacQueen, P. J., et al. 2021, AJ, 162, 298
Horne, K. 1986, PASP, 98, 609
House, L. R., Gebhardt, K., Finkelstein, K., et al. 2024, ApJ, 975, 172
House, L. R., Gebhardt, K., Finkelstein, K., et al. 2023, ApJ, 950, 82
Hu, E. M., & Cowie, L. L. 2006, Natur, 440, 1145
Khoraminezhad, H., Saito, S., Gronke, M., & Byrohl, C. 2025, OJAp, 8, 51254
Khostovan, A. A., Kartaltepe, J. S., Salvato, M., et al. 2026, ApJS, 282, 6
Kikuta, S., Ouchi, M., Shibuya, T., et al. 2023, ApJS, 268, 24
Konno, A., Ouchi, M., Nakajima, K., et al. 2016, ApJ, 823, 20
Landriau, M., Mentuch Cooper, E., Davis, D., et al. 2025, ApJ, 995, 220
Laseter, I. H., Finkelstein, S. L., Bagley, M. J., et al. 2022, ApJ, 940, 9
Law, D. R., Yan, R., Bershady, M. A., et al. 2015, AJ, 150, 19
Lee, D., Hensley, B. S., Chang, T.-C., & Doré, O. 2025, ApJ, 994, 61
Lee, K.-S., Gawiser, E., Park, C., et al. 2024, ApJ, 962, 36
Leung, A. S., Acquaviva, V., Gawiser, E., et al. 2017, ApJ, 843, 130
Liu, C., Gebhardt, K., Cooper, E. M., et al. 2022, ApJS, 261, 24
Liu, C., Gebhardt, K., Mentuch Cooper, E., et al. 2025, ApJS, 276, 72
Lujan Niemeyer, M., Komatsu, E., Bernal, J. L., et al. 2026, ApJ, 999, 177
Mentuch Cooper, E., Gebhardt, K., Davis, D., et al. 2023, ApJ, 943, 177
Mentuch Cooper, E., Gebhardt, K., Davis, D., et al. 2026, ApJ, 1000, 38
Mirza Khanlari, M., Gebhardt, K., Weiss, L. H., et al. 2025, ApJ, 989, 169
Mukae, S., Mentuch Cooper, E., Gebhardt, K., et al. 2026, AJ, 171, 318
Nilson, P. 1973, Uppsala General Catalogue of Galaxies (Astronomiska Observatorium)
Oke, J. B., & Gunn, J. E. 1983, ApJ, 266, 713
Ono, Y., Itoh, R., Shibuya, T., et al. 2021, ApJ, 911, 78
Ouchi, M., Ono, Y., & Shibuya, T. 2020, ARA&A, 58, 617
Ouchi, M., Shimasaku, K., Akiyama, M., et al. 2008, ApJS, 176, 301
Oyarzún, G. A., Blanc, G. A., González, V., Mateo, M., & Bailey, J. I. I. 2017, ApJ, 843, 133
Padmanabhan, N., Schlegel, D. J., Finkbeiner, D. P., et al. 2008, ApJ, 674, 1217
Parker, Q. A., Bojičić, I. S., & Frew, D. J. 2016, JPhCS, 728, 032008
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825
Pence, W. D., Chiappetti, L., Page, C. G., Shaw, R. A., & Stobie, E. 2010, A&A, 524, A42
Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020, A&A, 641, A1
Ramsey, L. W., Adams, M. T., Barnes, T. G., et al. 1998, SPIE, 3352, 34
Rhoads, J. E., Malhotra, S., Dey, A., et al. 2000, ApJL, 545, L85
Sánchez, S. F., Kennicutt, R. C., Gil de Paz, A., et al. 2012, A&A, 538, A8
Santos, S., Sobral, D., Matthee, J., et al. 2020, MNRAS, 493, 141
Schlafly, E. F., & Finkbeiner, D. P. 2011, ApJ, 737, 103
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Scoville, N., Aussel, H., Brusa, M., et al. 2007, ApJS, 172, 1
Shapley, A. E., Steidel, C. C., Pettini, M., & Adelberger, K. L. 2003, ApJ, 588, 65
Shoji, M., Jeong, D., & Komatsu, E. 2009, ApJ, 693, 1404
Sobral, D., Santos, S., Matthee, J., et al. 2018, MNRAS, 476, 4725
Spinoso, D., Orsi, A., López-Sanjuan, C., et al. 2020, A&A, 643, A149
Steidel, C. C., Adelberger, K. L., Dickinson, M., et al. 1998, ApJ, 492, 428
Tardugno Poleo, V., Finkelstein, S. L., Leung, G., et al. 2023, AJ, 165, 153
The HDF Group 2024, Hierarchical Data Format, v5, GitHub, https://github.com/HDFGroup/hdf5
Urrutia, T., Wisotzki, L., Kerutt, J., et al. 2019, A&A, 624, A141
van Breukelen, C., Jarvis, M. J., & Venemans, B. P. 2005, MNRAS, 359, 895
Van Rossum, G., & Drake, F. L. 2009, Python 3 Reference Manual (CreateSpace)
Vinkó, J., Thomas, B. P., Wheeler, J. C., et al. 2023, ApJ, 946, 31
Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, NatMe, 17, 261
Weiss, L. H., Davis, D., Gebhardt, K., et al. 2024, ApJ, 962, 102
Weiss, L. H., Gebhardt, K., Davis, D., et al. 2025, ApJ, 983, 72
Wells, D. C., Greisen, E. W., & Harten, R. H. 1981, A&AS, 44, 363
York, D. G., Adelman, J., Anderson, J. E. J., et al. 2000, AJ, 120, 1579
Zeimann, G. R., Debski, M. H., Schneider, D. P., et al. 2024, ApJ, 966, 14
Zhang, X., & Green, G. M. 2025, Sci, 387, 1209