A curated bioacoustics dataset of European bat vocalizations

A. Leonie Baier[1*], Panagiotis Georgiakakis[2], Karl-Heinz Frommolt[3], Signe M.M. Brinkløv[4], Simeon Q. Smeele[4], Bob Planqué[5], Willem-Pier Vellinga[5], Burooj Ghani[1], Dan Stowell[1,6†], Vincent J. Kalkman[1†]

[1*]Naturalis Biodiversity Centre, Darwinweg 2, Leiden, 2333 CR, The Netherlands.

[2]Natural History Museum of Crete, University of Crete, Knossos Avenue, Heraklion, 71409, Greece.

[3]Museum für Naturkunde Berlin, Invalidenstraße 43, Berlin, 10115, Germany.

[4]Department of Ecoscience, Aarhus University, C.F. Møllers Allé 3, Aarhus, 8000, Denmark.

[5*]Xeno-Canto Foundation, Laan van Nieuw Oost-Indië 177, The Hague, 2593BN, The Netherlands.

[6*]Department of Cognitive Science and Artificial Intelligence, Tilburg University, Warandelaan 2, Tilburg, 5037 AB, The Netherlands.

*Corresponding author(s). E-mail(s): leonie.baier@naturalis.nl;
†These authors contributed equally to this work.

Abstract

Acoustic surveys represent the primary method for monitoring bat populations and assessing their habitat use, yet reference recordings and associated metadata remain fragmented across repositories, recording programs, and private collections. We present ChirosetEurope, a curated dataset of European bat vocalizations (echolocation and social calls) assembled from multiple open repositories and expert-curated collections. The dataset integrates recordings from diverse geographic regions, habitat types, recording devices, and species groups, together with harmonized taxonomic labels, standardized metadata, and group-aware train/validation/test partitions optimized for machine-learning applications. To respect the licensing terms under which recordings were originally published, it is distributed as a metadata resource providing provenance information and persistent links to the original recordings, rather than redistributing all audio files. ChirosetEurope is designed to support a broad range of applications, including bioacoustic research, biodiversity monitoring, species identification, ecological analyses, and machine-learning development. By consolidating previously dispersed resources into a unified and versioned framework, the dataset provides a community resource for research on European bat bioacoustics and automated acoustic analysis.

Keywords: Bats, Bioacoustics, Biodiversity Monitoring, Acoustic Ecology

1 Background & Summary

Bats are among the most important yet least visible components of European ecosystems[1,2]. As nocturnal animals with largely hidden lifestyles, the ways in which bats shape ecosystems—including insect population regulation, nutrient transport, and trophic regulation—often remain unnoticed[3,4]. Nevertheless, bats are of major conservation and regulatory relevance. All European bat species are protected under Annex IV of the EU Habitats Directive, reflecting both their ecological importance and the long-term population declines experienced by many species during the twentieth century. Shifts in their populations function as sensitive bioindicators of broader ecosystem transitions[5–7].

In response to these declines, bat monitoring schemes emerged across many European countries long before the establishment of modern EU biodiversity legislation. While direct field methods remain central components of bat monitoring programs, including surveys and captures at roosts, hibernation sites and other key habitats, acoustic monitoring has long served as an important complementary approach because many bat species are difficult to detect and identify directly. The history of bat acoustics reflects broader advances in wildlife monitoring, progressing from heterodyne bat detectors to frequency-division and time-expansion systems, and later to digital full-spectrum recording and passive acoustic monitoring systems capable of collecting large quantities of ultrasonic recordings over extended periods using accessible, cost-effective hardware. In parallel, methods for analysing acoustic data have evolved from expert listening and visual interpretation of spectrograms to increasingly automated classification approaches, including machine-learning methods[8,9].

Bat acoustics present substantial challenges for automated species recognition. Bat vocalizations relevant to acoustic monitoring can be broadly grouped into search-phase echolocation calls, feeding buzzes, and social vocalizations. Importantly, all European bats navigate and forage using high-frequency echolocation, typically above 15 kHz, necessitating specialized recording equipment. Search-phase echolocation calls are emitted frequently during flight and contain species-specific information, yet their structure also reflects adaptation to particular foraging niches[10,11]. Bats further adjust these signals to habitat structure, flight conditions, and behavioural tasks[12,13], causing calls from different species to overlap while those of a single species vary across contexts and thereby complicating acoustic identification. Feeding buzzes, by contrast, are emitted specifically during the final phase of prey capture and are characterized by a rapid succession of short, low-amplitude, and often lower-frequency pulses [14]. Social vocalizations are used primarily for communication [15]. Because they serve functions such as distress signalling, maternal direction, mating, and song, these calls typically occupy lower frequency ranges, undergo less atmospheric attenuation, and can travel further. However, they are restricted to specific behavioural and social contexts and therefore occur less frequently in the wild. Because feeding buzzes and social vocalizations are context-specific, search-phase echolocation calls constitute the dominant signal type in passive acoustic monitoring. This contrasts with most other taxa, including birds, for which communication signals are generally used in automated species identification. However, the environmental plasticity of search-phase calls produces substantial acoustic overlap among species and high variability within species [10–13]. Consequently, reliable identification from recordings remains difficult even for experienced bat workers.

The increasing scale of passive acoustic monitoring has motivated growing interest in automated classification approaches based on deep learning [16]. Automated bat call classifiers, including commercial software tools, have been used in ecological monitoring workflows for more than a decade. Yet classification systems are often developed and evaluated on isolated, localized datasets, limiting the generalizability of models across different geographic regions and recording conditions. Because the performance and ecological validity of any machine learning system depends fundamentally on the quantity, quality, and representativeness of its training data [17], there is a critical need for a unified, large-scale reference resource that reflects this real-world variability [18,19].

At the same time, recent developments in open science and biodiversity data sharing have encouraged the publication of bat sound recordings through public repositories and collaborative monitoring initiatives. Citizen scientists, bat workers, research groups, and increasingly also governmental agencies are contributing recordings and associated metadata to openly accessible collections, creating new opportunities for large-scale aggregation and reuse of acoustic biodiversity data [20]. As part of the present work, we also encouraged and supported contributors in mobilizing existing recordings into public repositories suitable for long-term reuse.

Here we present ChirosetEurope, a curated dataset of European bat vocalizations spanning search-phase echolocation calls, feeding buzzes, and social calls. The dataset was assembled from multiple open and expert-validated sources and combines recordings from diverse geographic regions, habitat types, recording devices, and species groups. Within the constraints of currently available data, ChirosetEurope aims to maximize taxonomic coverage, geographic diversity, and annotation quality. It is intended as an adaptable reference resource for comparative bioacoustics, acoustic monitoring, and automated acoustic analysis. For machine-learning applications, its predefined group-aware partitions support reproducible model comparisons using consistent training and evaluation data. One such application is reported by Ghani et al. [21], who used the dataset to develop and evaluate an automated classifier for European bat vocalizations. That study focuses on model development and performance, whereas the present work describes the dataset structure, source aggregation, curation workflow, metadata standards, provenance, and technical validation. We additionally provide a starter notebook for flexible dataset reconstruction and preprocessing, together with comprehensive Usage Notes addressing dataset heterogeneity, imbalance, labelling certainty, and potential data-leakage risks.

2 Methods

2.1 Input data

Sources. Bat vocalization recordings and associated metadata were compiled from publicly accessible sources and a privately curated collection. Recordings were included only where reproducible user access could be provided, either through persistent recording-level links to the original source or through redistribution within the accompanying archive under terms permitting such sharing. All inputs were ingested using custom Python scripts and standardized to a unified data structure.

Publicly accessible recordings were obtained from *xeno-canto*[23], *iNaturalist*[24], *Observation.org*[25] (the latter accessed via the Global Biodiversity Information Facility, *GBIF*[26]), and a historical collection of European bat recordings archived at the *Tierstimmenarchiv*[27] of the Museum für Naturkunde Berlin. In addition, a privately curated collection provided by Maike Schuchmann, focusing on horseshoe bats (Rhinolophidae), was incorporated and is made publicly available as part of the present dataset.

The public repositories and the Rhinolophidae collection consist of native digital files, whereas the *Tierstimmenarchiv* collection contains historical recordings that were subsequently digitized. This archive was assembled by Reinald Skiba (1932—2013), author of a comprehensive description of European bat echolocation calls[28]. The original collection comprises analogue recordings captured using heterodyne, frequency-division, and time-expansion bat detectors and stored on magnetic tape with a Sony WM-D6C cassette recorder. Only the time-expansion recordings were included in ChirosetEurope as 'Skiba collection' (see Section 2.2), comprising 1,595 high-quality recordings from 27 European species. These particular recordings were acquired between 1997 and 2007 using a Pettersson D980 bat detector operating at a 350 kHz sampling rate with 1:10 time expansion and an original acquisition resolution of 8 bits. Following the archive's donation to the Museum für Naturkunde, the analogue tapes were digitized at 96 kHz/24 bit using the original playback equipment and an RME ADI 96 Pro analogue-to-digital converter. Although the archive was digitized at 96 kHz/24 bit to preserve the analogue recordings at high fidelity, this process cannot recover information beyond the original detector's 8-bit acquisition, which therefore remains the limiting factor for dynamic range. In addition, small tape-speed fluctuations (approximately 1%) introduce minor variations in both the frequency and time domains.

Taxonomic scope. The dataset targeted European bat species commonly encountered in acoustic monitoring and bioacoustic research applications (Table 1.) Repository-specific synonym substitutions were required for some taxa because different repositories indexed species under alternative generic assignments or historical nomenclature.

To resolve taxonomic discrepancies among source repositories, all species names were standardized to a unified reference taxonomy[29]. Taxonomic harmonization included the reassignment of Western Palearctic *Eptesicus* species to *Cnephaeus* following recent phylogenetic evidence,[30] the harmonization of records labelled as *Myotis oxygnathus* to *Myotis blythii*, and of records labelled as *Pipistrellus creticus* to *Pipistrellus hanaki*. In contrast, recordings labelled as species groups, unresolved compound taxa, or slash-delimited identifications (e.g., *M. brandtii/mystacinus*) were excluded because they could not be assigned unambiguously to a single species.

Table 1 Taxonomic coverage of ChirosetEurope, including the number of recordings, total audio duration, and geographic coverage for the 35 bat species represented in v1.0.

Genus	Species	n recordings	Duration (min)	n countries
Barbastella	*barbastellus*	284	46	14
Cnephaeus	*nilssonii*	177	35	12
Cnephaeus	*serotinus*	455	111	17
Hypsugo	*savii*	659	60	14
Miniopterus	*schreibersii*	251	41	6
Myotis	*alcathoe*	8	< 1	5
Myotis	*bechsteinii*	10	1	4
Myotis	*blythii*	7	< 1	3
Myotis	*brandtii*	53	12	5
Myotis	*capaccinii*	166	20	2
Myotis	*dasycneme*	220	48	10
Myotis	*daubentonii*	425	109	15
Myotis	*davidii*	16	< 1	1
Myotis	*emarginatus*	38	8	5
Myotis	*myotis*	73	24	7
Myotis	*mystacinus*	76	24	8
Myotis	*nattereri*	198	45	12
Nyctalus	*lasiopterus*	10	3	5
Nyctalus	*leisleri*	332	72	17
Nyctalus	*noctula*	871	202	19
Pipistrellus	*hanaki*	170	14	1
Pipistrellus	*kuhlii*	803	125	13
Pipistrellus	*maderensis*	44	16	2
Pipistrellus	*nathusii*	796	238	16
Pipistrellus	*pipistrellus*	1602	296	19
Pipistrellus	*pygmaeus*	508	86	20
Plecotus	*auritus*	258	28	13
Plecotus	*austriacus*	14	< 1	4
Rhinolophus	*blasii*	306	61	2
Rhinolophus	*euryale*	299	18	4
Rhinolophus	*ferrumequinum*	351	27	14
Rhinolophus	*hipposideros*	412	26	11
Rhinolophus	*mehelyi*	404	20	1
Tadarida	*teniotis*	974	102	8
Vespertilio	*murinus*	247	57	7

Data acquisition. For repository sourcing, data extraction and preprocessing were implemented in Python using a modular repository-specific download framework. The workflow used `pandas`, `librosa`, `pydub`, `requests`[31–34], and repository-specific API clients for automated querying, downloading, metadata aggregation, and preprocessing. Queries and local file directories were structured directly around the standardized scientific names defined in our unified reference taxonomy, applying the necessary synonym substitutions dynamically at the point of ingestion.

A shared base downloader class (a master code template that defines standard download and error-handling steps) was implemented to standardize acquisition across repositories. The framework maintained versioned metadata tables for each source repository and supported incremental dataset updates by tracking previously downloaded recording identifiers. For each downloaded recording, the workflow extracted or generated repository-specific recording identifiers, species names, licensing information, recording duration, local file paths, and content-derived audio hashes (unique digital fingerprints calculated directly from the audio data to identify duplicate files, even if they have different filenames). Metadata tables were stored as CSV files and versioned sequentially across dataset releases.

Recordings from *xeno-canto*[23] were retrieved through the xeno-canto API using the Python package `xenocanto-api`[35]. Species-level searches were restricted to `taxonomic group: bats`, `geographic region: Europe`, and recordings not under active identification review. For each target species, all recordings returned by the API were downloaded. Metadata fields were extracted from *xeno-canto* batch metadata tables and merged into the unified dataset schema. Metadata retained from *xeno-canto* included recording identifiers, scientific names, licensing information, source file names, and repository metadata associated with each upload. Downloaded recordings were organized into species-specific directories using standardized scientific names.

Recordings from *iNaturalist*[24] were retrieved using the iNaturalist API through the Python package `pyinaturalist`[36]. For each target species, species-level taxon identifiers were first resolved programmatically. Observation retrieval was restricted to observations containing associated sound files, "research-grade" observations only (see section 6.1), and observations assigned to the geographic region "Europe" using the *iNaturalist* place identifier '97391'. Several taxa required synonym remapping because *iNaturalist* indexed some species under alternative generic assignments. Repository-specific synonym substitutions were therefore applied during query generation to maximize retrieval completeness. For each downloaded sound file, the workflow retained observation identifiers, sound identifiers, scientific names, common names, observer usernames, geographic coordinates, country and locality information, observation URLs, recording timestamps, licensing information, audio duration, local storage paths, and content-derived audio hashes.

Observation.org[25] records obtained through *GBIF*[26] were processed from a Darwin Core Archive (DwC-A)[37] occurrence download[38] using custom Python scripts. The workflow parsed the `occurrence.txt` and `multimedia.txt` tables included in the DwC-A release. Media records were filtered to retain only entries with media type 'Sound'. Occurrence records were filtered to retain only taxa matching the predefined target species list. Following filtering, occurrence and multimedia tables were merged using an inner join on `gbifID`, thereby retaining only occurrence records with associated downloadable audio media. Metadata fields retained from the DwC-A archive included occurrence identifiers, scientific names, vernacular names, observer information, geographic coordinates, country information, event dates, licensing metadata, media download links, and occurrence URLs.

The Rhinolophidae and Skiba collections were processed using custom Python-based ingestion workflows to harmonize externally curated metadata and audio archives with the unified dataset structure. Metadata spreadsheets were matched against locally stored audio files using source file names. Files lacking accessible audio recordings or corresponding metadata entries were excluded during import. The Skiba collection additionally provided metadata describing recording methods, recording equipment, localities, and recording remarks. Time-expansion recordings were retained with explicit metadata annotations describing the associated time-expansion factors.

2.2 Data curation

Audio preprocessing. Audio recordings were subject to a series of curation steps to ensure consistency and suitability for bioacoustic analysis. Operations on audio files were performed solely for metadata derivation and curation decisions; no processed audio products are part of the released dataset.

Bat echolocation calls frequently contain substantial ultrasonic energy extending far beyond the human audible range. Conventional audio sampling rates such as 44.1 kHz discard biologically relevant spectral information, rendering files unsuitable for high-bandwidth bioacoustic analysis. Consequently, native digital recordings with original sampling rates below 96 kHz were excluded from the dataset. This threshold was employed as a conservative criterion for identifying recordings that likely retained their original ultrasonic content (rather than as an assertion that 96 kHz sampling fully captures the frequency range of all bat species). Recordings below this threshold typically represent heterodyned, frequency-divided, or otherwise destructively downsampled detector outputs for which the original ultrasonic frequency content cannot be reliably reconstructed. For recordings obtained from public repositories, metadata fields were additionally screened for indications of time expansion (e.g., ‘TE‘ or ‘expan*‘), with candidate files manually verified. Importantly, the Skiba collection contains files that were digitized at 96 kHz, but this sampling rate reflects the time-expanded audio signal rather than the original ultrasonic acquisition bandwidth. Compressed formats such as MP3 or M4A were also excluded because they typically contain transformed detector outputs (e.g. heterodyne or frequency-division recordings) rather than native full-spectrum audio and do not preserve the original ultrasonic signal. Files that could not be decoded reliably were also removed from the final manifest.

For metadata derivation and quality control purposes, audio files underwent the following internal signal processing steps:

Audio duration was calculated directly from decoded waveform data using the original audio stream. Durations were stored with millisecond resolution in a standardized MM:SS.mmm format. Files that could not be decoded or processed successfully were excluded from downstream curation.

Audio hashes were generated from decoded mono waveform data rather than from original binary file contents. Recordings were loaded using `librosa`[32] with original sampling frequencies preserved (sr=None) and converted to mono prior to hash generation. MD5 hashes were then calculated directly from waveform sample arrays. This procedure enabled stable duplicate detection independent of differences in metadata encoding, file containers, or source-specific formatting.

Metadata harmonization. Following acquisition, metadata originating from all repositories and collections were merged into a unified manifest structure. Metadata fields describing geographic coordinates, recording dates, licensing information, and locality descriptions were normalized into consistent formats across repositories. Processing was organized as a sequence of versioned metadata manifests. Each curation stage generated a new intermediate manifest file to preserve provenance and enable reproducible reconstruction of the full workflow. Although not implemented as a Darwin Core schema, many metadata fields correspond directly to Darwin Core terms.

Unique identifier assignment. To ensure unambiguous trackability and long-term reproducibility across dataset versions, each curated recording was provisioned with a standardized, structured alphanumeric identifier (`chiroset_id`) formatted as `CSE-[SpeciesCode]-[Source]-[RecordID]`. For example, `CSE-MYOBEC-XC-805730` represents a *Myotis bechsteinii* recording from Xeno-Canto (`XC`) with the native repository ID `805730`. Species codes are generated using the first three letters of the genus and species names. The record identifier is programmatically parsed from native metadata fields and zero-padded to a minimum of five digits where applicable, or assigned via a sequential counter fallback.

In some source collections, multiple distinct audio clips share the same source metadata from which the CSE identifier is derived. To ensure that each recording receives a unique identifier, newly generated IDs are checked against the waveform hash registry created during audio preprocessing. If the same structural identifier is generated for a non-identical audio clip, a numeric suffix is appended, e.g. `CSE-MYOEMA-SKIB-00112_3` (indicating the fourth distinct clip matching that specific Skiba record). Conversely, exact duplicate entries sharing the same content hash receive identical structural IDs to facilitate downstream deduplication.

Conservation status integration. To provide conservation context for downstream bioacoustic modeling, the unified manifest was supplemented with conservation status metadata from the International Union for Conservation of Nature (IUCN) Red List of Threatened Species (v4 API)[39]. Although legacy IUCN metadata were natively present in a subset of source records (e.g., observations sourced via GBIF), these historical entries were programmatically overwritten to ensure methodological uniformity across the entire dataset. For each unique combination of species and recording country, the IUCN API was queried to retrieve the current standardized European Regional assessment profile. This focused geographic scope provides a uniform, continental-scale threat status for every record. To facilitate analyses incorporating legal protection status, species were annotated under the EU Habitats Directive (Council Directive 92/43/EEC): all European bats are listed in Annex IV (strict protection), with Annex II status indicating species requiring designation of special areas of conservation[40].

In future releases of ChirosetEurope, IUCN status will be systematically updated to reflect the current conservation classification at the time of each respective version's release. However, because IUCN Red List assessments as well as Habitats Directive Annex classifications may change over time, end-users retain the final responsibility for verifying that the archival statuses embedded at the time of a dataset release remain aligned with current official listings at the time of active data usage.

License filtering. License metadata obtained from source repositories were normalized into standardized Creative Commons[41] categories. Records with unrecognized or incompatible licensing metadata were excluded from the curated dataset manifest. The dataset distinguishes between licenses that permit commercial reuse and modification of the underlying audio, and licenses that impose additional restrictions, such as limitations on commercial use or the creation of derivative works. Licenses permitting commercial reuse include CC0, CC-BY, CC-BY-SA, and Public Domain Mark (PDM). Restrictive licenses include CC-BY-NC, CC-BY-NC-SA, CC-BY-ND, and CC-BY-NC-ND. To ensure compliance with these source-specific licensing conditions, ChirosetEurope is distributed as a standardized metadata registry providing provenance information and links to the respective audio sources rather than redistributing all audio files. The accompanying Python notebook allows users to filter recordings according to licensing requirements prior to download or analysis.

Duplicate detection. Exact duplicate recordings, including files deposited in multiple source repositories or submitted more than once with different species labels, were identified using content-derived audio hashes generated from decoded mono waveform data, supplemented by duplicate filename detection. For each detected duplicate set, the associated source metadata and species labels were used to determine retention priority. Duplicate files with the same focal-species label were reduced to a single retained record. When identical audio had been associated with different focal-species labels—for example, where a recording containing multiple bats had been submitted separately under each species—the retained record was selected according to the species' representation in the curated dataset. The record assigned to the less-represented species was retained, rather than selecting a duplicate at random, to preserve information for taxonomically under-represented species and improve overall species balance. This procedure does not resolve the species identity of all sounds present in a recording; it determines which of multiple identically encoded copies is retained as the dataset record.

Minimum data thresholds. Species represented by fewer than five retained recordings after filtering and deduplication were excluded from the dataset to ensure sufficient sample sizes for statistical analysis and robust model training. Species excluded under this criterion included *Myotis crypticus*, *M. escalerai*, *M. punicus*, *Plecotus kolombatovici*, and *P. macrobullaris*. Several additional European bat species listed under the EUROBATS Agreement were not represented in the source collections (Supplementary Table S2), notably several island endemics and geographically restricted taxa. We highlight these species as priority targets for future recording efforts and dataset contributions. Future releases of ChirosetEurope will incorporate currently excluded species as additional labelled recordings become available.

Dataset splitting. To support future machine-learning development and ensure reproducible model evaluation, the curated manifest includes pre-assigned 'training', 'validation', and 'test' `split` tags. Splits were assigned at the level of metadata-derived grouping keys (`group_key`), designed to represent biologically or observationally coherent recording sessions. A recording session was defined spatio-temporally as recordings made at the same location on the same biological night. Location was defined by rounding geographic coordinates to three decimal places (approximately 100 m precision), and biological night was defined by assigning recordings made before 12:00 to the preceding night. When geographic coordinates were unavailable, sessions were instead defined using biological recording night, recordist identity, and country. This grouping design minimises data leakage from repeated recordings of the same recording session appearing across dataset partitions.

Dataset splitting was ordinarily conducted at the `group_key` level: all recordings in a group were assigned to a single split, irrespective of their species labels. Thus, recordings of different species originating from the same inferred recording session were kept together, reducing leakage of site- and time-specific acoustic conditions.

The Rhinolophidae collection was treated differently because its recordings were made under controlled, anechoic-chamber conditions at a shared location by the same recordist. A spatio-temporal grouping rule would therefore place most or all recordings in a single group and would not provide a useful basis for splitting the collection. Instead, recordings with reliable filename-derived individual identifiers were grouped by animal, preventing recordings from the same known individual from appearing in different splits. Pooled recordings and recordings without a reliable individual identifier were grouped using the ordinary spatio-temporal procedure. Because the acoustic recording environment was shared across this collection, split assignment cannot eliminate overlap in recording conditions; the individual-based grouping specifically addresses repeated-animal leakage. The dataset also contains recordings of *Rhinolophus* species from other source collections (currently all except *R. mehelyi*), providing examples recorded under different acoustic conditions.

Two limited exceptions to the grouping rule were made to ensure adequate species representation across splits. First, selected 'training' and 'test' assignments were swapped for *Myotis alcathoe*, *Pipistrellus maderensis*, and *Plecotus austriacus* to increase representation in the training set. This reassignment affected eight recording groups that also contained recordings belonging to other classes (i.e., different species). It therefore introduced limited, documented cross-split overlap between recordings from the same inferred recording session, but did not create cross-split overlap within any species. A second exception was made for *Myotis davidii*, which was represented by only two recording groups: the larger group (14 recordings) remained in 'training', while the two recordings in the smaller group were assigned individually to 'validation' and 'test' to permit representation of the species in both evaluation subsets.

Groups were allocated to the 'training', 'validation', or 'test' split using a reproducible MD5 hash of the text-based `group_key`. This procedure yields stable and fully reproducible splits that remain unchanged across dataset versions, even when new recordings are added, provided that the defined grouping identifiers are unchanged.

The resulting split proportions were approximately 60%, 20%, and 20% for the 'training', 'validation', and 'test' subsets, respectively (Supplementary Table S1).

By embedding these deterministic split assignments directly within the metadata manifest, ChirosetEurope provides a standardized baseline partition scheme that minimizes information leakage, while still allowing users the freedom to define alternative custom splitting strategies based on the provided spatio-temporal metadata.

3 Data Records

The dataset is hosted on Zenodo at https://doi.org/10.5281/zenodo.20773227[42]. ChirosetEurope consists of a harmonized metadata resource derived from multiple biodiversity and bioacoustic repositories as well as curated collections. The release does not redistribute audio files that are already available online; instead, it provides a standardized metadata registry and persistent access pathways, paired with a companion Python notebook for automated data retrieval, filtering, and preprocessing.

3.1 Dataset structure

The Zenodo archive contains three top-level files organized into three functional components:

(1) Unified metadata registry. All recordings are indexed in `ChirosetEurope_manifest_v1.csv`. This file contains one row per dataset record and provides standardized taxonomic labels, source information, provenance data, recording-level access information, and the recommended training, validation, or test assignment for machine-learning applications.

(2) Recording access and provenance. Audio files from contributing collections for which redistribution permission was obtained are provided in `audio.zip` and referenced in the metadata using relative file paths. Recordings that remain available through public repositories are represented in the manifest by persistent identifiers and recording-level source URLs, allowing users to retrieve them under the applicable licensing terms. Sources for which neither persistent recording-level access nor permission for redistribution could be established were not included in this release.

(3) User workflow notebook. A working starter notebook is provided as `ChirosetEurope_starter.ipynb`. It demonstrates how to query and filter the metadata registry, select recordings according to criteria such as licensing conditions and predefined partitions, retrieve audio files from their respective locations, and apply configurable preprocessing operations. These include channel conversion, resampling, and segmentation into fixed-length excerpts suitable for automated acoustic analysis and other bioacoustic applications. Users seeking additional practical guidance on applying deep-learning methods to animal-sound data are referred to Ghani et al.[43].

3.2 Metadata fields

Metadata are unified across sources and comprise the fields detailed in Table 2.

Table 2 Core metadata fields included in the unified ChirosetEurope manifest.

Field Name	Source Context	Description
`chiroset_id`	All	Unique identifier prefixed by species, source, and record ID.
`source_id`	All	Unique recording or observation ID within source repository.
`genus`	All	Genus component of the resolved taxon.
`species`	All	Species component of the resolved taxon.
`subspecies`	Where available	Subspecies taxonomic rank modifier if determined.
`group`	Where available	Assigned species grouping or acoustic classification clade.
`latin`	All	Standardized binomial species name.
`common_name`	Where available	Normalized common English species name.
`IUCN_status`	All	IUCN Red List status in respective country of recording.
`IUCN_scope`	All	Geographic scope of the IUCN assessment.
`annex_ii`	All	Annex II inclusion status under the EU Habitats Directive.
`annex_iv`	All	Annex IV inclusion status under the EU Habitats Directive.
`recordist`	Where available	Identity of the recordist(s) or original data contributor(s).
`time`	Where available	Time of day reference for temporal activity profiling.
`date`	Where available	Calendar date reference of the recording session.
`when_uploaded`	Public repos	Repository submission timestamp.
`country`	Where available	Standardized country name of the recording location.
`location`	Where available	Descriptive name or textual tag of the observation site.
`latitude`	Where available	Latitude coordinate of the observation location.
`longitude`	Where available	Longitude coordinate of the observation location.
`altitude`	Where available	Elevation height above sea level in meters.
`remarks`	Where available	Miscellaneous field notes or metadata annotations.
`rec_method`	Where available	Recording methodology (e.g., 'from hand', 'field recording').
`voc_type`	Where available	Vocalization type (e.g., social call, echolocation).
`sex`	Where available	Biological sex of the recorded individual.
`life_stage`	Where available	Ontogenetic stage of the subject (e.g., adult, juvenile).
`background_species`	Where available	Secondary species label.
`animal_seen`	Where available	Boolean indicator stating if subject was visually confirmed.
`playback_used`	Where available	Boolean indicator stating use of acoustic lures.
`temperature`	Where available	Ambient temperature recorded during the session.
`ring_number`	Curated sets	Banding ring number assigned to the subject.
`automated_rec`	Where available	Boolean indicator stating whether recording was unattended.
`device`	Where available	Hardware model of the bat detector or recording unit used.
`mic`	Where available	Specific microphone or transducer equipment profile used.
`sampling_rate`	All	Native audio sampling rate in Hertz (Hz).
`time_exp_factor`	All	Time expansion factor coefficient applied to the recording.
`length`	All	Duration of the audio recording in seconds.
`observation_url`	Public repos	Persistent web link to original occurrence record page.
`audio_url`	Public repos	Persistent link to the online audio file stream.
`sound_id`	Where available	Secondary sound or identifier if applicable.
`file_name`	All	Audio file name of the original raw recording.
`license`	Where available	Normalized Creative Commons license identifier string.
`audio_quality`	Where available	Subjective audio quality classification rating.
`verification`	Where available	Method of species label verification.
`source`	All	Originating repository identifier code (e.g., `xc`, `inat`).
`file_path`	All	Local archive filesystem path structure.
`retrieval_path`	All	Designated file acquisition path.
`audio_hash`	All	Content-based MD5 waveform hash.
`group_key`	All	Relational key mapping for managing grouped events.
`split`	All	ML partition designation (`train`, `val`, `test`).

3.3 Data versioning

This release corresponds to `ChirosetEurope v1.0`. The dataset is intended as a living resource that can evolve through the incorporation of additional recordings, metadata corrections, and improved annotations. To ensure reproducibility, each release will remain permanently archived and versioned while also preserving backward compatibility of the core schema. Split assignments for recordings already included in a dataset release will not be modified in subsequent versions. This design ensures that published analyses remain reproducible and that downstream modelling or statistical evaluations can be directly interpreted in the context of a specific dataset version.

4 Data Overview

ChirosetEurope comprises 11,517 bat vocalisation recordings representing 35 species across 32 European countries. Regional representativeness and taxonomic completeness relative to national bat faunas are summarized in Figure 1.

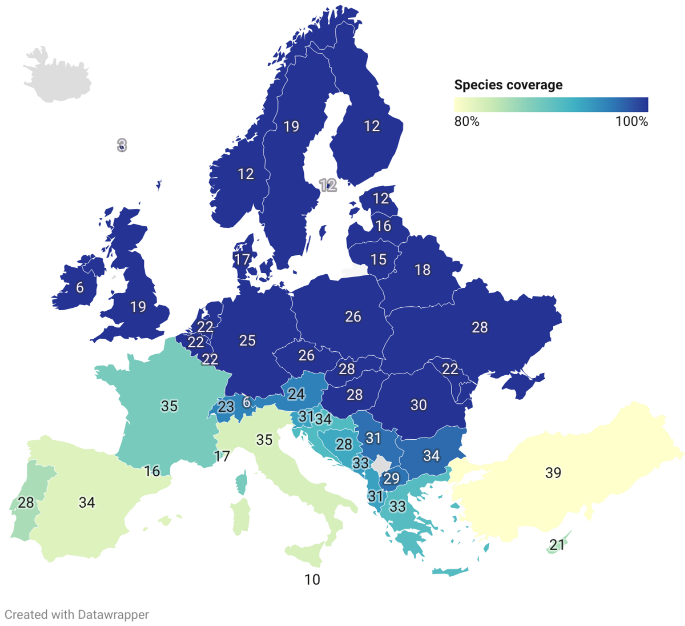

Fig. 1 Regional representativeness and taxonomic completeness of ChirosetEurope across national faunas. Shading indicates the percentage of locally occurring bat species represented within the dataset relative to national checklists; no inventory was available for Kosovo. Numeric labels within each country show its native bat species count [22]. While the dataset demonstrates high overall coverage across Europe (> 75%), an inverse trend is visible in high-biodiversity regions: countries with the highest number of native species tend to have lower representation. Notably, the dataset includes *Pipistrellus hanaki* and *Pipistrellus maderensis*, both represented in Europe only on islands. It currently lacks the island-endemic species *Plecotus sardus*, *Nyctalus azoreum*, and *Plecotus teneriffae*, slightly reducing the national coverage scores for Italy, Portugal, and Spain, respectively. Similarly, the species count for Turkey reflects its entire transcontinental inventory, including Anatolian and Middle Eastern species that do not occur in its European territory.

5 Technical Validation

Technical validation examined the integrity of the dataset throughout the curation workflow, from record attrition and provenance preservation to construction of the final metadata registry and group-aware partitions (Figure 2). Results from a separate classifier study provided further evidence that the dataset can support machine-learning-based bioacoustic analysis.

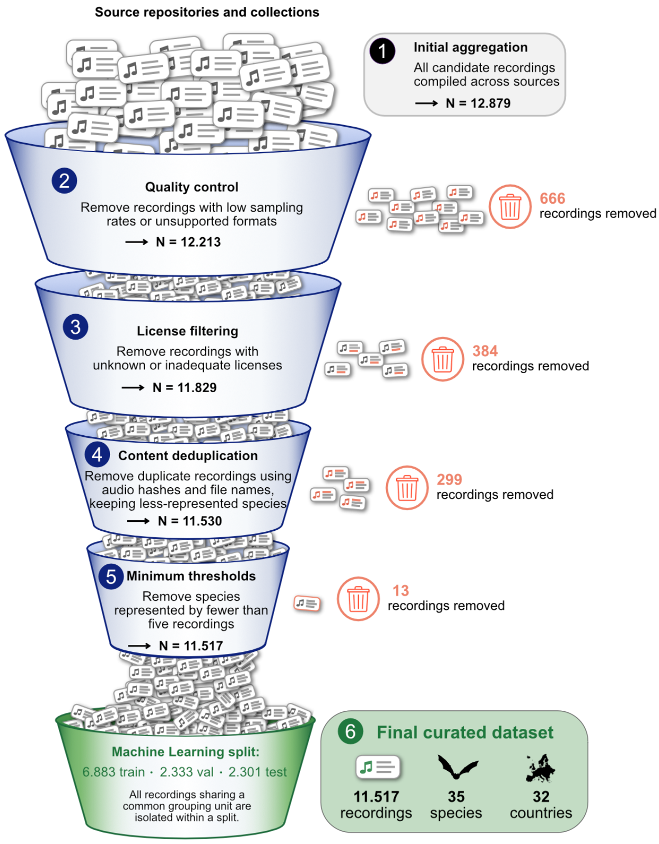

Fig. 2 Sequential curation workflow used to construct ChirosetEurope. Candidate recordings from all contributing repositories and collections were progressively filtered according to acoustic quality, recording format, licensing compatibility, duplicate content, and minimum species representation criteria. The number of recordings retained after each filtering stage is shown within each stage. Orange annotations indicate the number of recordings excluded by each filter. The final metadata registry preserves complete provenance and source traceability for every recording and forms the basis of the group-aware train/validation/test partitions distributed with the dataset.

The initial aggregation stage yielded 12,879 candidate recordings from all contributing repositories and collections. During acoustic and format quality control, 666 recordings were excluded because they did not meet the sampling-rate requirements (96 kHz or higher) or were stored in formats that prevented reliable recovery of the original signal. A further 384 recordings were removed due to incompatible licensing conditions, and 299 duplicate records were identified and eliminated based on exact waveform MD5 hashes and filename matching. Finally, 13 recordings were excluded because their species was represented by fewer than five remaining files. The final validated registry comprises 11,517 recordings representing 35 species across 32 countries. Provenance information, source identifiers, and links to the original recordings were retained for 100% of records, ensuring complete traceability to the source material. To validate the ML partitions, we verified that all recordings sharing a common metadata-derived grouping unit were assigned to the same subset. The resulting partition structure (6,883/2,333/2,301) approximates the intended 60%/20%/20% training, validation, and test split while preventing information leakage between subsets based on the available metadata. This grouping remains valid even when users apply additional filters to the registry, such as restricting analyses to recordings with specific licensing conditions.

As complementary application-level validation, a separate study trained a supervised classifier using the distributed group-aware partitions and recording-level species labels. In a closed-set evaluation covering all 35 species and an independently constructed noise class, the classifier achieved a species-level F1-score of 0.836, mean average precision of 0.694, and AUROC of 0.990 on the held-out test partition. These results provide evidence that the curated recordings and weak species labels can support automated taxonomic classification. Full methodological details, genus-level results, and further analyses are reported in the classification study by Ghani et al. [21].

6 Usage Notes

6.1 Dataset heterogeneity

This release represents an initial integrated version of a multi-source bat acoustic dataset compiled from heterogeneous biodiversity and bioacoustic repositories as well as curated collections. Users should be aware that recordings originate from heterogeneous recording devices and field conditions, which may introduce variability in amplitude scaling, frequency response, and background noise characteristics.

ChirosetEurope aggregates recordings containing search-phase echolocation, feeding buzzes, and social calls. While these acoustic sub-types are not uniformly annotated across all source files, the metadata fields `voc_type` or, in some cases, `remarks` frequently indicate which specific signals are present. Users wishing to train models exclusively on specific vocal types can programmatically parse these fields to filter for relevant keywords prior to feature extraction.

6.2 Audio preprocessing flexibility

The dataset preserves the original temporal structure of all recordings and does not include segmentation or trimming of audio files into fixed-length excerpts. This design choice maintains compatibility with a wide range of downstream machine-learning approaches—which may require different input window lengths depending on model architecture—while preserving the complete acoustic context necessary for traditional bioacoustic analysis. Recordings can be retrieved in their original form, allowing flexible preprocessing tailored to specific analytical needs, with configurable segmentation code provided in the starter notebook at `dataset/notebooks/ChirosetEurope_starter.ipynb`.

6.3 Data imbalance

As is typical for large-scale bioacoustic compilations, the dataset is highly imbalanced. It exhibits a heavily skewed taxonomic distribution, with a small number of common species accounting for the majority of records (Table 1), as well as a non-uniform distribution of file lengths that reflects the standardized recording durations of the underlying source platforms (Table 1). Downstream machine-learning architectures should implement mitigation strategies—such as stratified batch sampling, focal loss functions, or data augmentation—to prevent classification biases toward overrepresented taxa.

6.4 Labelling certainty

Because ChirosetEurope aggregates data from diverse platforms, the verification rigour behind species-level labels varies across recordings and may result in differing levels of expected label noise. Users should consider these source-specific validation profiles when constructing training, validation, or benchmarking subsets. The unified metadata manifest includes a dedicated verification field describing the label verification procedure associated with each recording.

Xeno-canto: Species identifications on *xeno-canto* rely primarily on the uploader's expertise. However, identifications can be challenged by other users through the platform's community review process. While recordings with explicitly flagged identification disputes were excluded during data retrieval, the remaining species identifications should not be considered definitive. Note that xeno-canto's A–E quality ratings refer exclusively to recording quality and should not be interpreted as indicators of label accuracy or verification confidence.

iNaturalist: To maximize label reliability, data ingestion from *iNaturalist* was strictly restricted to 'Research Grade' observations. On this platform, observations only attain Research Grade status when a community consensus is reached—specifically, when more than 2/3 of identifying users agree on a species-level taxonomy and the community taxon matches the observer's assignment.

Observation.org: Sourced via *GBIF*, records from *Observation.org* rely on a distinct validation pipeline combining automated scoring filters and manual verification by designated regional taxonomic experts who review submitted audio evidence before observations are tagged as validated.

Rhinolophidae collection: This subset features absolute labelling certainty. It contains recordings of known individuals captured and recorded under controlled laboratory conditions, ensuring highly reliable species assignments.

Skiba collection: Sourced from the Animal Sound Archive ('Tierstimmenarchiv') of the Museum für Naturkunde Berlin, this collection provides an expert-identified dataset of Central European bat acoustics. It focuses primarily on directed handheld or active field recordings and serves as a reference database for bioacoustic, ethological, and automated pattern recognition research.

Where available, the unified metadata manifest preserves original repository quality flags and validation metadata. Together with the `verification` field, these attributes enable users to filter recordings according to their desired balance between dataset size and expected label noise.

6.5 Licensing compliance

The dataset aggregates recordings originating from sources with differing licensing conditions. To ensure compliance with source-specific requirements, the release is distributed as a harmonized metadata resource containing provenance information, licensing metadata, and access information for each recording. Because ChirosetEurope does not redistribute the underlying audio recordings, long-term availability of individual files depends on the originating repositories. Consequently, recordings may subsequently become unavailable or inaccessible if they are removed or their access conditions change. Users should consult the original repository records or accompanying archive documentation for the applicable license terms before retrieving, reusing, or redistributing audio data. The accompanying Python notebook provides functionality for filtering recordings according to licensing constraints prior to download or analysis. Because the ML partitions are deterministically assigned at the group level, users restricting the dataset to permissive (commercial-friendly) licenses can maintain the original train/validation/test split assignments, ensuring leak-free modelling even if the precise 60/20/20 proportions shift slightly.

6.6 Data leakage risks

Because the dataset aggregates recordings from repositories that have been publicly available for varying periods of time, some recordings or subsets thereof may previously have been used for training, validation, or evaluation of existing bat acoustic classification systems. Consequently, users aiming to perform rigorous model evaluation or comparative studies should assess the potential for data leakage between their training data and the recordings contained in this release.

Where prior use of source repositories is known or suspected, additional filtering or source-aware partitioning strategies may be appropriate to ensure independent evaluation. To facilitate such analyses, the metadata include a `source` field identifying the repository or collection from which each recording was obtained. However, users should note that prior use of recordings is not necessarily restricted to their repository of origin. In some cases, recordings may have been incorporated into training datasets before being deposited in the repositories represented here, meaning that repository-level filtering should be regarded as a useful first step rather than a guarantee of independence.

6.7 Future development

ChirosetEurope is designed as an evolving resource built upon systematic license filtering, deduplication, and taxonomic harmonization. The dataset reconstruction and preprocessing workflows provided with this initial release are intended as reference implementations that users can adapt to specific analytical requirements, with future updates expected to continuously improve both data quality and annotation structure. In particular, two areas are prioritised for further development: (i) validation and addition of new recordings across contributing sources to continually expand dataset volume and coverage, and (ii) the enrichment of annotation granularity, specifically the transition from recording-level (weak) labels toward temporally and spectrally localized (strong) annotations. At present, *xeno-canto*[23] is the only contributing platform that provides infrastructure for detailed annotations; this discrepancy across repositories precludes the uniform integration of strong labelling in the current dataset release. Resolving this limitation in future versions will provide the precise temporal labels required for both advanced acoustic behaviour research and automated call detection algorithms. In parallel, future releases will seek to align the ChirosetEurope metadata schema with emerging community standards for bat acoustic metadata as these mature, improving interoperability with other datasets and analysis tools.

7 Data Availability

All data have been deposited in the following Zenodo repository[42]: https://doi.org/10.5281/zenodo.20773227.

8 Code Availability

The accompanying Zenodo release (DOI: 10.5281/zenodo.20773227)[42] contains a starter Jupyter notebook (`ChirosetEurope_starter.ipynb`) intended for dataset users. This notebook demonstrates how to query the metadata manifest, retrieve recordings from their original repositories or archive locations, and apply standard preprocessing steps for downstream bioacoustic and machine-learning applications.

The custom code used to compile, curate, and validate the ChirosetEurope metadata registry is available through GitHub at https://github.com/anlebai/ChirosetEurope and archived as a versioned release on Zenodo at https://doi.org/10.5281/zenodo.21992417[44]. The code is provided as a Jupyter notebook (`Creating_ChirosetEurope.ipynb`) which also includes optional analysis routines for exploratory dataset inspection, such as summary statistics and visualisations of class imbalance, and supports reproduction of the figures and tables presented in this manuscript. All code is released under the MIT License.

Author contributions

A. Leonie Baier: Conceptualization, Methodology, Software, Validation, Formal analysis, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Panagiotis Georgiakakis:** Data Curation, Writing - Review & Editing. **Karl-Heinz Frommolt:** Data Curation, Writing - Review & Editing, Funding acquisition. **Simeon Q. Smeele:** Data Curation, Writing - Review & Editing. **Signe M.M. Brinkløv:** Data Curation, Writing - Review & Editing. **Bob Planqué:** Resources, Data Curation, Writing - Review & Editing. **Willem-Pier Vellinga:** Resources, Data Curation, Writing - Review & Editing. **Burooj Ghani:** Conceptualization, Writing - Review & Editing. **Dan Stowell:** Conceptualization, Writing - Review & Editing, Project administration, Funding acquisition. **Vincent J. Kalkman:** Conceptualization, Writing - Review & Editing, Project administration, Funding acquisition.

Funding

This work was supported by funding received via the Horizon Europe projects TETTRIs (grant no. 101081903, supporting work by BG, DS, VJK) and MAMBO (grant no. 101060639, supporting work by ALB, PG, BG, DS, VJK). The views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The digitization of the Skiba Collection was supported by a grant from the Pawel von Rammingen Foundation.

References

[1] Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., ... Narme, S. Biodiversity loss and its impact on humanity. *Nature* **486**(7401), 59–67 (2012). https://doi.org/10.1038/nature11148

[2] Locke, H., Rockström, J., Adams, R., Bäckstrand, K., Bennett, E., Biggs, R., Boers, N., de Vries, W., Echo-Hawk, A., Gifford, L., Gordon, L. J., Kotzé, L. J., Krug, R. M., Liverman, D., Loriani, S., Luke, O., Miller, C., Obura, D., O'Farrell, P., ... Rocha, J. C. Nature Positive: halting and reversing biodiversity loss toward restoring Earth system stability. *Front. Sci.* **4**, 1609998 (2026). https://doi.org/10.3389/fsci.2026.1609998

[3] Ghanem, S. J., & Voigt, C. C. (2012). Increasing awareness of ecosystem services provided by bats. In *Advances in the Study of Behavior* (Vol. 44, pp. 279–302). Elsevier. https://doi.org/10.1016/B978-0-12-394288-3.00007-1

[4] Sánchez-Soto, M. F., Gaona, O., Mercado-Juárez, R., Yanez-Montalvo, A., de León-Lorenzana, A., Borja-Martínez, G., ... Falcón, L. I. Ecological-based insights into bat populations in the Yucatán Peninsula under a One Health

approach: coexistence or biophobia. *Front. Conserv. Sci.* **5** (2025). https://doi.org/10.3389/fcosc.2024.1488378

[5] Jones, G., Jacobs, D. S., Kunz, T. H., Willig, M. R. Racey, P. A. Carpe noctem: the importance of bats as bioindicators. *Endanger. Species Res.* **8**(1-2), 93–115 (2009). https://doi.org/10.3354/esr00182

[6] Russo, D., Salinas-Ramos, V. B., Cistrone, L., Smeraldo, S., Bosso, L. Ancillotto, L. Do We Need to Use Bats as Bioindicators? *Biology* **10**(8), 693 (2021). https://doi.org/10.3390/biology10080693

[7] Siya, A., Matovu, B., Nalukenge, L., Mutebi, M., Nalikka, B., Castle, K., Dewey, T., Williams, K. M., Wickenkamp, N. R., Harris, E. K., Rwego, I. B., Sande, E., Masembe, C., Kading, R. C., & Kityo, R. M. (2025). Bat Cave Vulnerability to Anthropogenic Factors: Status and Priorities for Conservation Within the Mount Elgon Region, Uganda. *Life*, *15*(12), 1940. https://doi.org/10.3390/life15121940

[8] Frick, W. F. (2013). Acoustic monitoring of bats, considerations of options for long-term monitoring. *Therya*, *4*(1), 69–78. https://doi.org/10.12933/therya-13-109

[9] Roemer, C., Haquart, A., López-Baucells, A., & Besnard, A. (2025). Current frontiers in the passive acoustic monitoring of bats. *Methods in Ecology and Evolution*, *16*, 2534–2544. https://doi.org/10.1111/2041-210X.70157

[10] Murray, K. L., Britzke, E. R. and Robbins, L. W. (2001). Variation in search-phase calls of bats. *Journal of Mammalogy*, 82(3):728–737. https://doi.org/10.1644/1545-1542(2001)082⟨0728:VISPCO⟩2.0.CO;2

[11] Siemers, B. M. and Schnitzler, H.-U. (2004). Echolocation signals reflect niche differentiation in five sympatric congeneric bat species. *Nature*, 429(6992):657–661. https://doi.org/10.1038/nature02547

[12] Obrist, M. K. (1995). Flexible bat echolocation: The influence of individual, habitat and conspecifics on sonar signal design. *Behavioral Ecology and Sociobiology*, 36:207–219. https://doi.org/10.1007/BF00177798

[13] Denzinger, A. and Schnitzler, H.-U. (2013). Bat guilds, a concept to classify the highly diverse foraging and echolocation behaviors of microchiropteran bats. *Frontiers in Physiology*, 4:164. https://doi.org/10.3389/fphys.2013.00164.

[14] Ratcliffe, J. M., Elemans, C. P. H., Jakobsen, L. and Surlykke, A. (2013). How the bat got its buzz. *Biology Letters*, 9(2):20121031. https://doi.org/10.1098/rsbl.2012.1031

[15] Middleton, N., Froud, A., & French, K. (2022). *Social Calls of the Bats of Britain and Ireland: Expanded and Revised Second Edition*. Pelagic Publishing Ltd.

[16] Stowell, D. (2022). Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10:e13152. https://doi.org/10.7717/peerj.13152.

[17] Brinkløv, S.M.M., Macaulay, J., Bergler, C., Tougaard, J., Beedholm, K., Elmeros, M. and Madsen, P.T. (2023). Open-source workflow approaches to passive acoustic monitoring of bats. *Methods in Ecology and Evolution*, 14, 1747–1763. https://doi.org/10.1111/2041-210X.14131.

[18] Milner-Gulland, E. J. A Critical Year for Nature: Now is the time to accelerate action on the Global Biodiversity Framework. Preprint at https://doi.org/10.32942/X2MQ1T (2026).

[19] Gonzalez, A., et al. From data to decisions: Toward a Biodiversity Monitoring Standards Framework. *UU Research Portal* (2026).

[20] Murdoch Research Portal. Building support for 'non-charismatic' species conservation through citizen science: bats and bioacoustics. *Systematic Review (2014-2025)* (2025).

[21] Ghani B., Eversteijn W., van Hirtum M., Cañas J.S., Kalkman V.J., Stowell D., Baier A.L. ChiroEcho: extending automated bat vocalisation classification beyond the learned taxonomy. *Computer Vision – ECCV 2026 Workshops* (2026) Accepted for publication. Preprint: https://arxiv.org/abs/2608.18191

[22] UNEP/EUROBATS. Agreement on the Conservation of Populations of European Bats (2026). *Protected Bat Species*. Retrieved from https://www.eurobats.org/about_eurobats/protected_bat_species (Accessed: July 9, 2026).

[23] Vellinga, W.-P. Planqué, R. The Xeno-canto Collection and its Relation to Sound Recognition and Classification. *CLEF (Working Notes)* **1391** (2015).

[24] iNaturalist. iNaturalist Open Data Initiative. https://www.inaturalist.org (2026).

[25] Observation International and local partners. Observation.org: Nature observations worldwide. https://observation.org (2025).

[26] Telenius, A. Biodiversity information goes public: GBIF at your service. *Nordic Journal of Botany* **29**(3), 378–381 (2011). https://doi.org/10.1111/j.1756-1051.2011.01167.x

[27] The Animal Sound Archive (Tierstimmenarchiv) at the Museum für Naturkunde Berlin. Tierstimmenarchiv Bioacoustic Collection. https://www.tierstimmenarchiv.de/ (2025).

[28] Skiba, R. *Europäische Fledermäuse. Kennzeichen, Echoortung und Detektoranwendung* 2nd edn (Westarp Wissenschaften, 2009).

[29] Simmons, N.B. and Cirranello, A.L. (2018). *Bat Species of the World: A Taxonomic and Geographic Database*. Available at: `https://batnames.org` (accessed 16 June 2026).

[30] Cláudio, V. C., Novaes, R. L. M., Gardner, A. L., Nogueira, M. R., Wilson, D. E., Maldonado, J. E., Oliveira, J. A. & Moratelli, R. Taxonomic re-evaluation of New World *Eptesicus* and *Histiotus* (Chiroptera: Vespertilionidae), with the description of a new genus. *Zoologia (Curitiba)* **40**, e22029 (2023). https://doi.org/10.1590/S1984-4689.v40.e22029

[31] McKinney, W., et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* **445**, 51–56 (2010).

[32] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E. & Nieto, O. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference* **18**, 25–30 (2015).

[33] Robert, J., et al. Pydub: Manipulate audio with a simple and easy high level interface. Software available at https://github.com/jiaaro/pydub (2021).

[34] Reitz, K., et al. Requests: HTTP for Humans. Software available at https://github.com/psf/requests (2026).

[35] Ghani, B. xenocanto-api: A Python library for querying and downloading recordings from the Xeno-canto API, version 0.1.1. *Zenodo* https://doi.org/10.5281/zenodo.10468000 (2025).

[36] Cook, J. pyinaturalist: Python client for the iNaturalist API, version 0.20.2. *PyPI* https://pypi.org/project/pyinaturalist/ (2025).

[37] Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Doering, M., Giovanni, R., Robertson, T. Vieglais, D. Darwin Core: an evolving community-developed biodiversity data standard. *PLOS ONE* **7**(1), e29715 (2012). https://doi.org/10.1371/journal.pone.0029715

[38] GBIF.org (2026). GBIF Occurrence Download. Downloaded 17 June 2026. doi:10.15468/dl.g26vt7.

[39] IUCN. (2026). *The IUCN Red List of Threatened Species API*. Version 2026-1. Retrieved from https://api.iucnredlist.org/ (Accessed: June 19, 2026).

[40] European Environment Agency. (2026). *EUNIS, the European Nature Information System*. Retrieved from https://eunis.eea.europa.eu/ (Accessed: June 20, 2026).

[41] Creative Commons. About the Licenses. https://creativecommons.org/share-your-work/cclicenses/ (2026).

[42] Baier A.L., Georgiakakis P., Frommolt K.H., Brinkløv S.M.M., Smeele S.Q., Planqué B., Vellinga W.P.., Ghani B, Stowell D., Kalkman V.J. ChirosetEurope: A curated bioacoustics dataset of European bat vocalizations [Data set]. Zenodo; 2026. Available from: https://doi.org/10.5281/zenodo.20773226

[43] Ghani B., Baier A.L., Kalkman V.J., Stowell D. Twelve quick tips for applying deep learning to animal sounds. *PLOS Computational Biology* **22**(8), e1014604 (2026). https://doi.org/10.1371/journal.pcbi.1014604

[44] Baier A.L. Creating ChirosetEurope [Software]. Zenodo; 2026. Available from: https://doi.org/10.5281/zenodo.21992417

[45] S. Barova and A. Streit (eds.), *Action Plan for the Conservation of All Bat Species in the European Union, 2018–2024*, Technical report, European Commission and UNEP/EUROBATS, October 2018. Main report cited together with the accompanying complementary document. Available at: https://environment.ec.europa.eu/topics/nature-and-biodiversity/habitats-directive_en (accessed July 9 2026).

Supplementary Material

Table S1 Composition of ChirosetEurope: For the 35 represented species, the table reports recording counts by split, total audio duration, and the number of recording countries.

Species	**Train**	**Val**	**Test**	**Total**	**Duration**	**Countries**
Barbastella barbastellus	170	77	37	284	00:46:29	14
Cnephaeus nilssonii	79	54	44	177	00:34:44	12
Cnephaeus serotinus	253	123	79	455	01:51:08	17
Hypsugo savii	399	100	160	659	01:00:08	14
Miniopterus schreibersii	112	109	30	251	00:40:56	6
Myotis alcathoe	5	2	1	8	00:00:45	5
Myotis bechsteinii	6	1	3	10	00:01:05	4
Myotis blythii	3	3	1	7	00:00:28	3
Myotis brandtii	38	12	3	53	00:12:11	5
Myotis capaccinii	152	12	2	166	00:19:41	2
Myotis dasycneme	118	63	39	220	00:48:03	10
Myotis daubentonii	254	81	90	425	01:49:22	15
Myotis davidii	14	1	1	16	00:00:30	1
Myotis emarginatus	35	1	2	38	00:08:20	5
Myotis myotis	66	5	2	73	00:23:35	7
Myotis mystacinus	67	8	1	76	00:23:30	8
Myotis nattereri	119	46	33	198	00:44:48	12
Nyctalus lasiopterus	7	1	2	10	00:03:20	5
Nyctalus leisleri	235	52	45	332	01:12:24	17
Nyctalus noctula	518	186	167	871	03:21:40	19
Pipistrellus hanaki	76	35	59	170	00:13:34	1
Pipistrellus kuhlii	462	77	264	803	02:05:05	13
Pipistrellus maderensis	22	5	17	44	00:16:27	2
Pipistrellus nathusii	426	154	216	796	03:58:11	16
Pipistrellus pipistrellus	1048	344	210	1602	04:56:23	19
Pipistrellus pygmaeus	297	114	97	508	01:26:27	20
Plecotus auritus	158	40	60	258	00:27:58	13
Plecotus austriacus	8	1	5	14	00:00:28	4
Rhinolophus blasii	135	124	47	306	01:01:27	2
Rhinolophus euryale	210	62	27	299	00:17:30	4
Rhinolophus ferrumequinum	206	109	36	351	00:26:36	14
Rhinolophus hipposideros	283	61	68	412	00:26:17	11
Rhinolophus mehelyi	269	36	99	404	00:20:02	1
Tadarida teniotis	497	190	287	974	01:42:04	8
Vespertilio murinus	136	44	67	247	00:57:11	7
Total	**6,883**	**2,333**	**2,301**	**11,517**	**32:58:47**	**32**

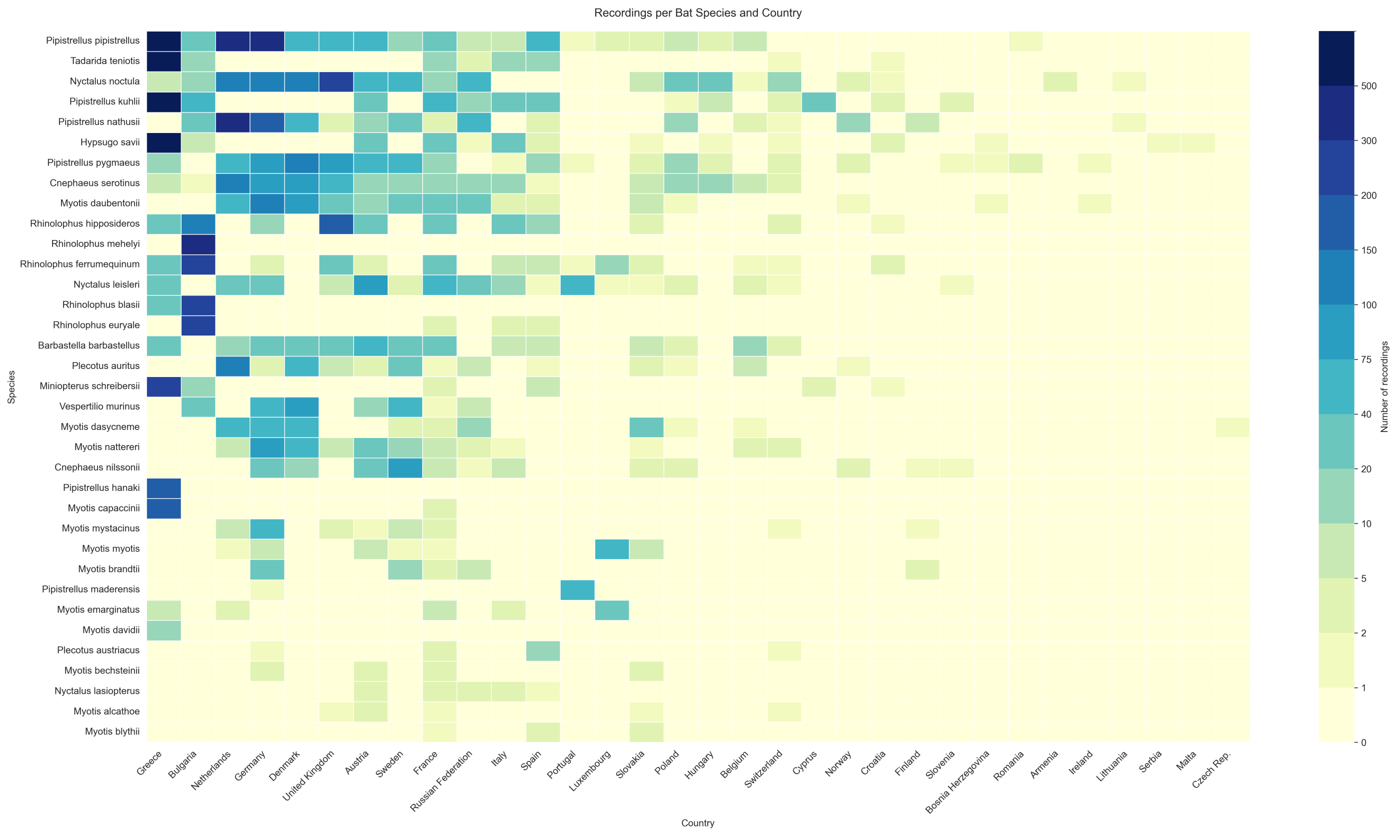

Fig. S1 Geographic and taxonomic distribution of dataset recordings. Heatmap depicting the number of audio recordings per bat species across sampling countries in the ChirosetEurope dataset. Both species (vertical axis) and countries (horizontal axis) are ordered by total recording volume (descending). Color intensity reflects binned recording counts on a non-linear scale to highlight sampling coverage across data-sparse and data-dense regions.

Table S2 The 55 bat taxa listed under the EUROBATS Agreement on the Conservation of Populations of European Bats[22], presented using current nomenclature, and their representation in ChirosetEurope ('CSE'). 'No.' indicates the number of EUROBATS Range States ($n = 60$) in which a species is listed as present. Species marked with * have no documented occurrence within the territory covered by the latest EU Action Plan for Bats (2018–2024)[45]. Data current as of July 2026.

Scientific name	CSE	No.	Distribution within EUROBATS area
Barbastella barbastellus	✓	45	Widespread across Europe and adjacent western Asia
*Barbastella caspica**		4	Caucasus and Caspian region
*Cnephaeus anatolicus**		5	Eastern Mediterranean
Cnephaeus isabellinus		6	Iberia and North Africa
Cnephaeus nilssonii	✓	30	Northern, central and eastern Europe
Cnephaeus ognevi		6	Caucasus and Middle East
Cnephaeus serotinus	✓	53	Widespread across Europe, North Africa and western Asia
Hypsugo savii	✓	40	Southern and central Europe, North Africa and western Asia
*Miniopterus pallidus**		5	Caucasus, Turkey and western Asia
Miniopterus schreibersii	✓	28	Southern Europe, the Balkans and the Mediterranean region
Myotis alcathoe	✓	31	Central, southern and eastern Europe
Myotis bechsteinii	✓	31	Western, central and southeastern Europe
Myotis blythii/oxygnathus	✓	40	Southern and central Europe, North Africa and western Asia
Myotis brandtii	✓	37	Northern, central and eastern Europe
Myotis capaccinii	✓	24	Mediterranean region and Middle East
Myotis crypticus		8	Southwestern Europe
Myotis dasycneme	✓	24	Northern, central and eastern Europe
Myotis daubentonii	✓	42	Widespread across Europe and adjacent western Asia
Myotis davidii	✓	11	Eastern Europe, the Caucasus and western Asia
Myotis emarginatus	✓	45	Southern and central Europe, North Africa and western Asia
Myotis escalerai		3	France and the Iberian Peninsula
Myotis hoveli		4	Eastern Mediterranean
Myotis myotis	✓	35	Western, central and southeastern Europe
Myotis mystacinus	✓	43	Widespread across Europe and adjacent western Asia
Myotis nattereri	✓	33	Western, central and northern Europe
Myotis punicus		7	Western Mediterranean and North Africa
*Myotis schaubi**		2	Armenia and Iran
*Myotis tschuliensis**		5	Caucasus and Turkey
Nyctalus azoreum		1	Azores (Portugal)
Nyctalus lasiopterus	✓	27	Southern, central and eastern Europe
Nyctalus leisleri	✓	47	Widespread across Europe and adjacent regions
Nyctalus noctula	✓	48	Widespread across Europe and western Asia
*Otonycteris hemprichii**		11	North Africa and Middle East
Pipistrellus hanaki/creticus	✓	1	Greece
Pipistrellus kuhlii	✓	44	Southern and central Europe, North Africa and western Asia
Pipistrellus maderensis	✓	2	Madeira and the Canary Islands
Pipistrellus nathusii	✓	43	Widespread across Europe and western Asia
Pipistrellus pipistrellus	✓	53	Widespread across the EUROBATS area
Pipistrellus pygmaeus	✓	44	Widespread across Europe and western Asia
Plecotus auritus	✓	47	Widespread across Europe and adjacent western Asia
Plecotus austriacus	✓	33	Western, central and southern Europe
Plecotus gaisleri		6	Central Mediterranean and North Africa
Plecotus kolombatovici		8	Adriatic, Balkan and eastern Mediterranean region
Plecotus macrobullaris		22	Mountain regions of southern Europe and western Asia
Plecotus sardus		1	Sardinia (Italy)
Plecotus teneriffae		1	Canary Islands (Spain)
Rhinolophus blasii	✓	20	Balkans, eastern Mediterranean and western Asia
Rhinolophus euryale	✓	32	Southern Europe, North Africa and western Asia
Rhinolophus ferrumequinum	✓	45	Widespread across southern Europe and western Asia
Rhinolophus hipposideros	✓	50	Widespread across Europe, North Africa and western Asia
Rhinolophus mehelyi	✓	27	Mediterranean region, the Balkans and western Asia
Rousettus aegyptiacus		9	Eastern Mediterranean and Middle East
Tadarida teniotis	✓	36	Mediterranean region and western Asia
*Taphozous nudiventris**		10	North Africa and Middle East
Vespertilio murinus	✓	40	Central, northern and eastern Europe