A&A 665, A26 (2022)
https://doi.org/10.1051/0004-6361/202243587
© M. Mahlke et al. 2022

Astronomy & Astrophysics

Asteroid taxonomy from cluster analysis of spectrometry and albedo[⋆]

M. Mahlke[1], B. Carry[1], and P.-A. Mattei[2]

[1] Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, 06304 Nice Cedex 4, France
e-mail: `max.mahlke@oca.eu`

[2] Université Côte d'Azur, Inria, Maasai project-team, Laboratoire J.A. Dieudonné, UMR CNRS 7351, 06902 Sophia-Antipolis, France

Received 18 March 2022 / Accepted 25 May 2022

ABSTRACT

Context. The classification of the minor bodies of the Solar System based on observables has been continuously developed and iterated over the past 40 yr. While prior iterations followed either the availability of large observational campaigns or new instrumental capabilities opening new observational dimensions, we see the opportunity to improve primarily upon the established methodology.

Aims. We developed an iteration of the asteroid taxonomy which allows the classification of partial and complete observations (i.e. visible, near-infrared, and visible-near-infrared spectrometry) and which reintroduces the visual albedo into the classification observables. The resulting class assignments are given probabilistically, enabling the uncertainty of a classification to be quantified.

Methods. We built the taxonomy based on 2983 observations of 2125 individual asteroids, representing an almost tenfold increase of sample size compared with the previous taxonomy. The asteroid classes are identified in a lower-dimensional representation of the observations using a mixture of common factor analysers model.

Results. We identify 17 classes split into the three complexes `C`, `M`, and `S`, including the new `Z`-class for extremely-red objects in the main belt. The visual albedo information resolves the spectral degeneracy of the `X`-complex and establishes the `P`-class as part of the `C`-complex. We present a classification tool which computes probabilistic class assignments within this taxonomic scheme from asteroid observations, intrinsically accounting for degeneracies between classes based on the observed wavelength region. The taxonomic classifications of 6038 observations of 4526 individual asteroids are published.

Conclusions. The ability to classify partial observations and the reintroduction of the visual albedo into the classification provide a taxonomy which is well suited for the current and future datasets of asteroid observations, in particular provided by the *Gaia*, MITHNEOS, NEO Surveyor, and SPHEREx surveys.

Key words. minor planets, asteroids: general – methods: data analysis – techniques: spectroscopic

1. Introduction

The minor planets of the Solar System exhibit a wide range of surface compositions as outcomes of their diverse formation histories. Mineralogical insights into the main asteroid belt gained from observing the bodies' exteriors serve to constrain the dynamic evolution scenarios of our planetary environment (Morbidelli et al. 2015), to establish relationships in asteroid families (Masiero et al. 2015), and to identify the parent bodies of the members of the meteorite collection (Burbine et al. 2002; Granvik & Brown 2018). The conclusion of a static Solar System formation history (Gradie & Tedesco 1982) has since been discarded in favour of a dynamical version (Gomes et al. 2005; Morbidelli et al. 2005; Tsiganis et al. 2005) following the increasing resolution of the compositional distribution of asteroids in the main belt and in near-Earth orbits thanks to a growing number of minor bodies characterised by dedicated observational efforts (e.g. Bus & Binzel 2002b; Devogèle et al. 2019; Xu et al. 1995). Today, the majority of the mass in the main belt is thought to have been dynamically implanted during a later evolutionary stage of the Solar System (DeMeo & Carry 2014; Gradie & Tedesco 1982), including some of the largest members of the main belt (Bottke et al. 2006; De Sanctis et al. 2015; Vernazza et al. 2021; Vokrouhlický et al. 2016). Evidence of a dichotomous meteorite population further strengthens this interpretation of a large compositional variability among minor bodies as result of early-stage formation processes in the Solar System (Warren 2011).

To describe the compositional distribution, a classification scheme based on the observable features of asteroids is required. A common device used in the interpretation of observations is asteroid taxonomy. Taxonomic classification refers to the grouping of objects with shared characteristics (Candolle 1813). For asteroids, these characteristics are the observable surface properties, such as the absorption bands imprinted into their reflectance spectra or the surface albedos. The implicit assumption is that the observables are related to the minor planets' surface mineralogy (Gaffey & McCord 1979), though this is not a prerequisite for a practical taxonomy.

Schemes for the compositional classification of minor planets have been devised and iterated regularly since the 1970s (e.g. Bowell et al. 1978; Chapman et al. 1971; McCord & Chapman 1975). The initial division into carbonaceous `C`-types and silicaceous `S`-types was readily apparent in different observables, even for a small number of observed objects and limited observational detail. However, with an increasing number of smaller objects observed, the underlying continuum distribution between these complexes has been revealed (Bus & Binzel 2002a).

[⋆] The table of asteroid classifications and the templates of the defined taxonomic classes is only available at the CDS via anonymous ftp to `cdsarc.u-strasbg.fr` (`130.79.128.5`) or via `http://cdsarc.u-strasbg.fr/viz-bin/cat/J/A+A/665/A26`

The most commonly used taxonomies for minor bodies are the Tholen system (Tholen 1984) and the Bus-DeMeo system (Bus & Binzel 2002a; DeMeo et al. 2009). While the latter offers a feature-based classification which encompasses a wide range of the variability observed in spectral observations and has been adapted to visible and near-infrared (NIR) photometric observations (Carvano et al. 2010; DeMeo & Carry 2013; Popescu et al. 2018), the former has not been fully replaced, in part due to two advantages of the used asteroid observables: the visual albedo p_V and spectrophotometric observations down to ultraviolet (UV) wavelengths. Both features increase in particular the resolution of classes which only show faint features in the visible and NIR wavelength regimes.

In this work we aim to methodologically improve upon the existing taxonomic schemes for minor bodies with regard to three aspects. First, we introduce a method which enables the classification of complete and partial observations. This offers consistent class definitions across the visible-near-infrared (VisNIR) region. Second, the visual albedo p_V is reintroduced into the taxonomy observables, solving the degeneracy of the X-complex as a primary consequence. Third, asteroids are classified in a probabilistic model, yielding a vector of class probabilities rather than a definite class assignment, which enables taxonomic outliers and transitional populations to be identified.

In addition to the methodological advancement, we further aim to align the scheme of taxonomic classes with advancements in the understanding of asteroid surface compositions acquired over the last decade. Studies such as Rivkin (2012), Vernazza et al. (2014, 2015), and Shepard et al. (2015) have combined observational evidence for several asteroid and meteorite connections which show that the classes in the current schemes do not reflect mineralogical groups. While this is acceptable a priori as taxonomies are built on spectroscopic data alone, by taking into account the multi-observable studies we believe that a correction is acceptable and necessary.

In Sect. 2, we outline the collection of the observational data used in this study, as well as the methodological advancement of the clustering strategy with respect to previous taxonomies. In Sect. 3, we outline the clustering results and the strategy of identifying compositional classes. These classes are then discussed in detail in Sect. 4. In Sect. 5, we investigate degeneracies between the classes in this taxonomy and compare the classifications of asteroids in this study to those in the literature. The `classy` tool to classify asteroid observations in the framework of this taxonomy is presented. Finally, we draw conclusions and give an outlook in Sect. 6.

2. Method

In this section, we describe the compilation and preprocessing of the asteroid spectra and albedos for the cluster analysis, an overview of which is provided in Fig. 1. It is followed by a description of the issues that arise when working with partial observations (i.e. missing data). After motivating the split of the dataset into clustering and classification data, the section concludes with a description of our approach to the dimensionality reduction and clustering problem at hand.

2.1. Input data

2.1.1. Selecting the observables

The selection of asteroid observables to be included in a taxonomical system is a crucial decision in its design. A broad set of observables ensures its applicability to a large number of asteroids and high compositional resolution; however, it complicates the derivation of the classification scheme and limits the number of available observations as only the intersection in terms of observed asteroids can be considered when combining different datasets[1]. This first led (Tholen 1984) to apply the albedo only in a secondary classification step before the observable was completely dropped by Bus & Binzel (2002a).

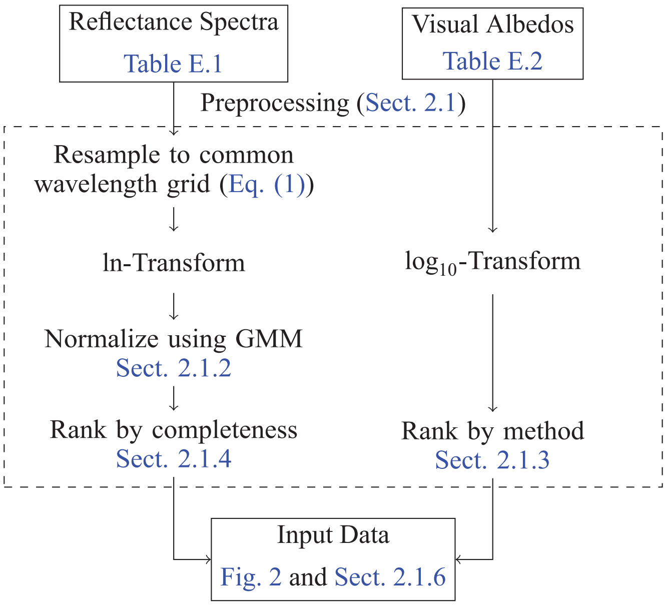


Fig. 1. Overview of preprocessing the input observations. The preprocessing steps encompassed in the dashed rectangle can be performed using the `classy` `python` package described in Sect. 5.

One of our main goals for this iteration of the taxonomy is the possibility to classify partial observations; we are a priori accepting gaps in the input data, and are thus not limiting the sample size when combining datasets and can use the union rather than the intersection of observations. Nevertheless, while we first considered a classification system based on spectrometric and photometric observations, and on visual albedos and phase curve coefficients, we found that including photometric observations and phase curve coefficients did not add to the compositional resolution of the resulting scheme as they are effectively low-resolution versions of the former (DeMeo & Carry 2013; Mahlke et al. 2021; Shevchenko et al. 2016). Therefore, we chose to build the taxonomy from asteroid VisNIR spectra and visual albedos.

2.1.2. Spectra

Spectrometric observations are the most compositionally informative asteroid features accessible via remote sensing. In preparing this work we focused both on building a large repository of asteroid spectra and on curating the data. In total, we acquired over 7500 spectra from online repositories, archived publications, and directly from the observers. The majority of spectra are unpublished spectra from the Small Main-belt Asteroid Spectroscopic Survey (SMASS) (Xu et al. 1995) and MIT-Hawaii Near-Earth Object Spectroscopic Survey (MITHNEOS)

[1] In machine learning literature, the observables used to identify groups in the input data are referred to as features, while the observations are referred to as samples. The input data is a matrix spanned by the features as columns and the samples as rows.

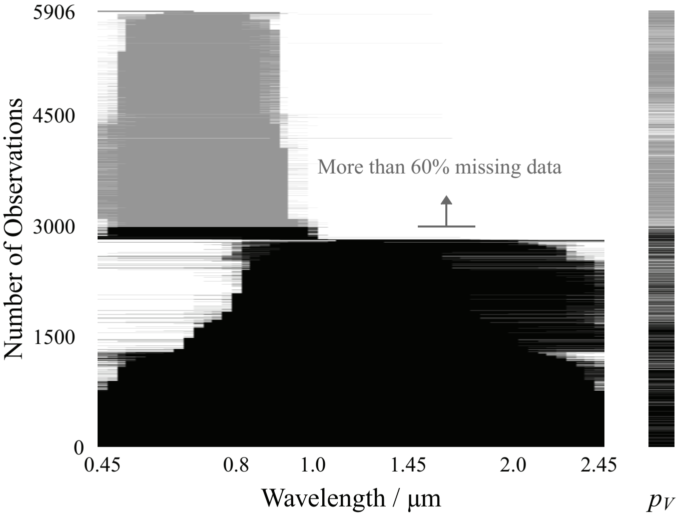


Fig. 2. Input data shown as a matrix. The columns represent the asteroid observables (i.e. the spectral wavelength bins and the visual albedo p_{V}) and each row represents one observation. The density of sampled wavelength bins is doubled in the visible compared to the near-infrared region. The cells are white if the corresponding value was not observed. The black cells indicate the samples used in the clustering analysis; the grey cells are samples that are classified but not used to build the taxonomy itself, due to the large degree of missing information in these spectra. 2983 observations are at least 40% complete and were used to train the clustering model. The matrix is sorted by increasing completeness of the asteroid spectra from top to bottom.

(Binzel et al. 2019; Marsset et al. 2022) surveys available online[2]. Literature sources of the spectra are given in Table E.1. After several iterations of visual inspection and rejection of low-quality data and duplicated observations, 6038 spectra of 4526 individual asteroids remained. About 50% of spectra cover the visible wavelength range only, while the sample of VisNIR spectra is about three times as large as in DeMeo et al. (2009) (see Fig. 2 in this paper).

The collection of spectra is heterogeneous in numerous aspects, including but not limited to their wavelength coverage, resolution, and sampling patterns. However, a consistent sampling pattern between all spectra is required for the following numerical analyses. We define this pattern in close resemblance to the one used by DeMeo et al. (2009), though we halve the sampling step size in the visible wavelength range as we find that possible superpositions of absorption features due to mafic minerals around 1 µm are better described by the finer sampling. The chosen sampling pattern is

$$\begin{aligned}\lambda_{\mathrm{S}} \in \{&0.45, 0.475, 0.50, \ldots, 1.0, 1.025,\\ &1.05, 1.10, 1.15 \ldots, 2.40, 2.45\}\ \mu\mathrm{m},\end{aligned} \tag{1}$$

totaling 53 wavelengths. In the following cluster analysis, each of these wavelengths represents one data dimension.

Before resampling the spectra, we apply a filter (Savitzky & Golay 1964) to smoothen features present in the spectra (e.g. telluric absorption features). The filter consists of applying least-squares fits of polynomials to a window of adjacent data points. The window size in units of data points and the degree of the polynomial dictate the amount of smoothing that is applied. We set these two parameters for each spectrum separately by visual inspection of the results. The smoothened spectra are then linearly interpolated and resampled to the pattern in Eq. (1). We then transform the spectra using the natural logarithm, which serves to approximate a zero mean and uniform standard deviation of the input spectra as they are generally normalised to unity at either 0.55 µm or 1.25 µm. This standardisation transform is generally beneficial to clustering and dimensionality reduction methods (Bouveyron et al. 2019).

The inclusion of missing data in the analysis poses a new challenge when it comes to normalising the spectral data. The common approach of multiplicatively setting the reflectance to unity at a shared wavelength is not possible as no single wavelength is shared among all spectra, as can be seen in Fig. 2. Furthermore, this approach would artificially decrease the variance in the wavelength chosen for normalisation and the neighbouring wavelength bins, causing the subsequent clustering analysis to effectively ignore the normalisation region.

Instead, we prepare the spectra in a way which benefits the following analysis most by employing a Gaussian mixture model (GMM). We assume that each spectrum can be written αy, where $\alpha \in \mathbb{R}$ is a normalisation constant that depends on the considered spectrum, and $y \in \mathbb{R}^{53}$ is a normalised spectrum. Further assuming that y follows a mixture of k log-normal distributions with diagonal covariances, all parameters of the models can be estimated from an incomplete data set via an expectation-maximisation algorithm (Dempster et al. 1977). This allows in particular to estimate the normalisation constants of all spectra, and to finally normalise them. By trial and error, we find that $k = 30$ mixture components result in a satisfying normalisation. As outlined in Sect. 2.2, the assumption of a normal distribution of the input samples in data space is also made in the clustering analysis.

Finally, we note that DeMeo et al. (2009) removed the slope component of the spectra to decrease the influence of space weathering on the taxonomy and to increase the depth of features present in the data. We cannot do this due to the missing data; however, we consider the presence of spectral-weathering effects in the taxonomy a beneficial rather than unfavourable aspect, as we further outline in Sect. 4.

2.1.3. Albedo

The visual albedos used in this study were compiled for the IMCCE's Solar system Open Database Network (SsODNet[3], Berthier et al., in prep.). The main contributors in this compilation are the Infrared Astronomical Satellite (IRAS) (Tedesco et al. 2002), the Wide-field Infrared Survey Explorer (WISE) (Masiero et al. 2011), AKARI (Usui et al. 2011), and Spitzer (Trilling et al. 2016). We use the SsODNet service to collect 4704 albedo measurements for the 3543 asteroids of which we have spectral observations using the `rocks`[4] `python`-interface (Berthier et al., in prep.).

When possible, we make use of several albedo measurements per asteroid when combining the input features (Sect. 2.1.4). To get the most accurate available value for each asteroid, we first compute the albedo based on the weighted averages of the asteroid's diameter and absolute magnitude provided by SsODNet following Harris & Lagerros (2002). These averages consist of the subjectively best available observations (Berthier et al., in prep.). In a second step, we compute the weighted mean of any albedo measurement available in the literature for the given asteroid and use this value as the second available albedo observation

[2] http://smass.mit.edu/

[3] https://ssp.imcce.fr/webservices/ssodnet/

[4] https://rocks.readthedocs.io

in the input data. Finally, the non-aggregated albedo observations are appended as additional available measurements. The literature sources we used to compile available albedo values and recompute updated ones from absolute magnitude and diameter are given in Table E.2

As for the spectra, we aim to have Gaussian distributions in the albedo data. Wright et al. (2016) shows that the distribution of albedos follows a double-Rayleigh distribution with a dark peak and a bright peak. To get a Gaussian distribution, we pass the $\log_{10}$-transform of the albedos to the clustering algorithm after limiting all albedo values to the interval [0.01, 1). We note that the albedo represents a single data dimension in the following analysis, compared to the 53 spectral data dimensions.

2.1.4. Merging of data samples

Previous taxonomies were derived based on photometry or spectrometry from a single dataset, for example the Eight Color Asteroid Survey (ECAS) (Zellner et al. 1985) for Tholen (1984) or the SMASS survey for Bus & Binzel (2002a). Individual observations of a single asteroid were combined in these datasets to give the single best-possible observation. In this work we do not combine the observations as they come from numerous different sources; for example, 549 of the 2125 asteroids have more than one observation in the input data.

When merging the asteroid spectra and albedo observations for each asteroid, we aim to create as many complete rows as possible. The array of albedo values is merged with the asteroid's spectral values in order of most complete observations. If there are more albedo observations than spectra, we discard the remaining values. We further set an upper limit of five spectra for each asteroid, removing any excess ones in order of the fewest observed wavelength bins. Many spectra of a single asteroid may cause artificial clusters or trends in the resulting taxonomy. This reduces the number of available spectra from 6038 to 5906. Figure 2 shows the final matrix of observations, colour-coded to differentiate partial and complete observations.

2.1.5. Clustering versus classification

When clustering the entire input dataset described above with the method outlined below, we note a population of clusters which contains mostly visible-only spectra. The intuitive explanation of these clusters is that when computing dimensionality reduction and projecting the spectra into a lower-dimensional space, they will be distributed over a smaller volume than the VisNIR spectra due to their large degree of missing information. This artificial accumulation of input data in the latent space disturbs the identification of real clusters in the data. We therefore set an upper limit of 60% of missing values per spectrum to be included into the clustering input data, which includes 2983 of the 5906 samples in the input data (see Fig. 2). The remaining 2923 are not used to derive the taxonomy; instead, they are classified in the resulting scheme and used to cross-validate the classification, as outlined in Sect. 5. In Fig. 3, we display the distribution of the 2125 individual asteroids in the data with which we derive the taxonomy in the sample over orbital classes and absolute magnitude.

2.1.6. Data availability

The dataset containing the input data samples, asteroid metadata, and the resulting classifications as outlined in the next sections is available at the Centre de Données astronomiques de Strasbourg (CDS).

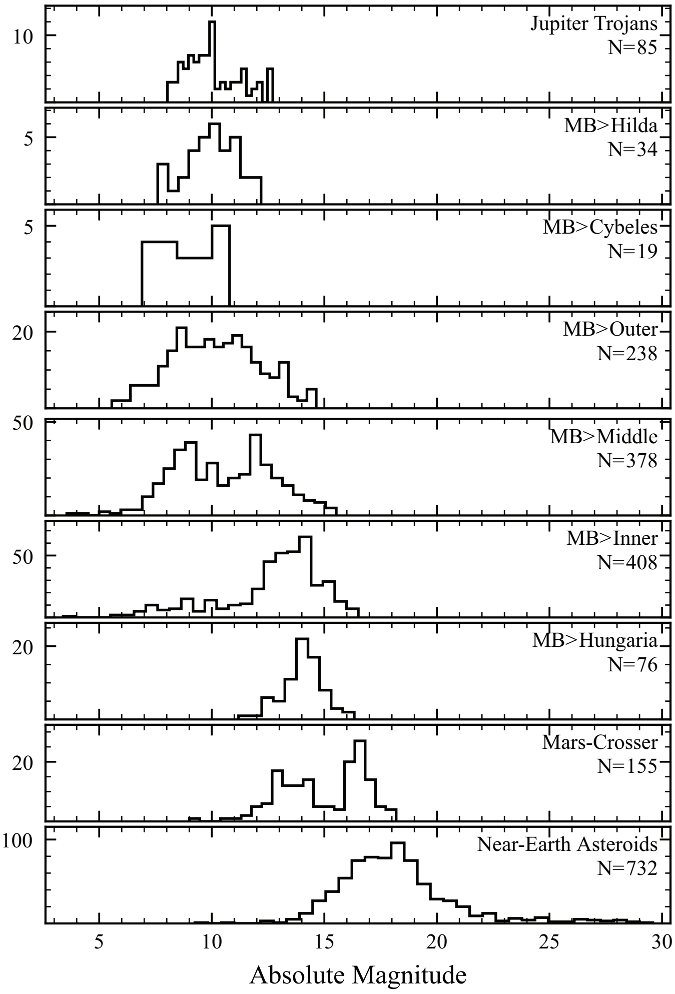


Fig. 3. Distribution of the 2125 individual asteroids used to build the taxonomy over orbital class and absolute magnitude. MB refers to the main belt. The number N of asteroids in the orbital population is given below each orbital class. The bin size of the histograms varies with N.

2.2. Dimensionality reduction and clustering

The derivation of a taxonomy falls into the realm of unsupervised machine learning. In the context of minor bodies, the approach consists of two steps: dimensionality reduction followed by clustering. Previous taxonomies have predominantly chosen principal component analysis (PCA) for the former and visual clustering for the latter. Given our goals for this new iteration of the current taxonomy as stated in Sect. 1, we need to evolve the established method, in particular to allow for the classification of partial observations. In the following we outline this method evolution, which arises naturally when challenging the PCA method with the requirements of our input data and the prior knowledge from previous taxonomies. The description of the resulting model is kept concise; the reader is referred to Tipping & Bishop (1999), Baek et al. (2010), and Montanari & Viroli (2010) for detailed explanations, and to Casey et al. (2019) for an example application of the same model but with a different treatment of missing data in the field of stellar physics.

2.2.1. Dimensionality reduction

The necessity of dimensionality reduction derives mainly from the spectrometric observations, where each bin of the sampling pattern represents one of the 54 data dimensions. Clustering in such high-dimensional space is not feasible as any model would be overparametrised given the limited sample size of the input data. Reducing the dimensionality of the observed data space is achieved by building linear combinations of the observed variables, referred to as latent variables[5], and projecting the input data into the space spanned by the latent variables, referred to as latent space.

We assume that we have N observations of a p-dimensional observable. The input data $\mathbf{Y}$ is thus of shape $N \times p$, denoted $\mathbf{Y}_{N\times p}$[6]. PCA can be described as eigendecomposition of the covariance matrix $\Sigma_{p\times p}$ of $\mathbf{Y}$,

$$\boldsymbol{W}^\intercal \Sigma \boldsymbol{W} = \mathbf{\Lambda}, \tag{2}$$

where $\mathbf{W}$ and $\mathbf{\Lambda}$ are the eigenvectors and eigenvalues of Σ (Pearson 1901). Expressing the p-dimensional Σ by the q-eigenvectors corresponding to the largest eigenvalues of Σ, where $q < p$, leads to dimensionality reduction while retaining the largest possible amount of variance within the data. The lower-dimensional representation $\mathbf{Z}_{N\times q}$ of the input data $\mathbf{Y}$ is given by the matrix product of $\mathbf{Y}$ with the matrix of the subset of eigenvectors $\mathbf{W}_{p\times q}$. In the following latent components are denoted $\mathbf{W}$ and latent scores $\mathbf{Z}$. The p elements of each latent component are referred to as latent loadings. They are the coefficients of the linear combination of dimensions of the input data. Latent components are constrained to unit L2-norm (i.e. the square root of the sum of the squared latent loadings is one). We note that the latent components and their loadings are determined from the data alone; no a priori information is used to influence the matrix.

PCA does not allow for missing data as it relies on the eigendecomposition of Σ. This limitation is overcome by reformulating PCA as a latent generative variable model. In essence, while computing the latent components and scores from the input data, we are making the assumption that there exists a Gaussian-distributed variable z in the latent space which causes the variance observed in the higher-dimensional data (i.e. that the resulting latent scores are normal distributed). A general model of this approach can be expressed as (Tipping & Bishop 1999)

$$\boldsymbol{Y} = \boldsymbol{f}(\boldsymbol{Z}, \boldsymbol{W}) + \boldsymbol{\epsilon}, \tag{3}$$

where f is a function of the latent scores $\mathbf{Z}$ and the latent components $\boldsymbol{W}$ and $\boldsymbol{\epsilon}_{p\times p}$ is a noise matrix independent of $\boldsymbol{Z}$. The reformulation of PCA in this model framework is referred to as probabilistic PCA (PPCA) (Tipping & Bishop 1999). The model parameters are fit using an expectation-maximisation algorithm (Dempster et al. 1977) and, when the input data is complete, gives the same solution as the conventional PCA.

PPCA assumes that the noise matrix $\boldsymbol{\epsilon}$ is isotropic (i.e. all data dimensions carry the same noise). This is not necessarily the case for our observations; the uncertainties between visible and NIR spectra may differ from one another and from that of the visual albedo. Factor analysis (FA) is another latent generative variable model analogous to PPCA except that the noise matrix $\boldsymbol{\epsilon}$ is assumed to be diagonal rather than isotropic (Rubin & Thayer 1982). The noise matrix is referred to as uniqueness as it captures the variance that is unique to each data dimension, effectively decoupling the measurement uncertainties from the data covariance.

In FA, the observations $\boldsymbol{Y}$ are modeled as

$$\boldsymbol{Y} = \mu + \boldsymbol{W}\boldsymbol{Z} + \boldsymbol{\epsilon}, \tag{4}$$

where $\mu_{p\times 1}$ is a p-dimensional vector containing the mean values of $\boldsymbol{Y}$ along the feature dimensions; $\boldsymbol{W}$ is the matrix of the latent components, as above; and $\boldsymbol{\epsilon}$ is a diagonal Gaussian noise matrix, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \Psi)$, where $\boldsymbol{\Psi}_{p\times p}$ is diagonal. The latent variables $\boldsymbol{Z}$ follow a normal distribution with zero mean and unit covariance, $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The model parameters can be determined by a maximum-likelihood approach, even in the case of missing data, under the assumption that the data is missing at random (i.e. its probability of being missing is independent of its value) (Little & Rubin 2019).

2.2.2. Clustering

Using the FA model given in Eq. (4), we assume that the distribution of the latent scores (i.e. the asteroid observations mapped into the latent space) follows a single Gaussian distribution. However, we know a priori from the previous taxonomic efforts that this is not the case; the `C`- and `S`-complexes form separate distributions, and endmember classes such as `A`, `K`, and `V` follow separate trends in the latent scores (see Fig. 2 in DeMeo et al. 2009). Instead of a single Gaussian, we therefore model the data as a mixture of g Gaussian distributions, an approach referred to as mixture of common factor analysers (MCFA, Baek et al. 2010) in the case where the model components are fit in the same latent space as is the case here. MCFA can be expressed as specialisation of the FA model in Eq. (4) using (Baek et al. 2010)

$$\begin{aligned} \mu_i &= \mathbf{A}\xi_i, \\ \mathbf{\Sigma_i} &= \mathbf{A}\mathbf{\Omega}_i\mathbf{A}^\intercal + \epsilon, \end{aligned} \tag{5}$$

where $i \in (1, \ldots, g)$, $\mathbf{A}$ is the common subspace of the mixture components (i.e. the matrix of latent components), ξ_i is the mean value of the ith mixture components in latent space, and $\mathbf{\Omega}_i$ is its variance. The noise matrix ϵ retains its definition as above, meaning that all mixture components share the same noise.

In MCFA, dimensionality reduction and clustering are achieved concurrently during the model training. Starting from an initial set of model parameters as outlined in Sect. 3.1.2, at each training epoch (i.e. the optimisation of the log-likelihood of the model against the entire input dataset), this model searches for the q-dimensional latent space and divides the input samples into g components, which gives the most likely projection of the input data assuming that it follows the mixture of g Gaussian distributions in the reduced space. The hyperparameters in the model are the number g of clusters and the number q of latent components.

2.3. Model implementation and availability

Implementations of the MCFA mixture-model approach are available in the `R` programming language[7] by Baek et al. (2010) and in the `python` language[8] by Casey et al. (2019). Nevertheless, we chose to write an alternative implementation in `python` as the implementation by the latter imputes the missing data via

[5] Latent can here be understood as a synonym for hidden or underlying as these variables are not directly observable.

[6] In the following we state the shape of the tensors in this manner when we first introduce them, and drop the notation afterwards.

[7] `https://github.com/suren-rathnayake/EMMIXmfa`

[8] `https://github.com/andycasey/mcfa`

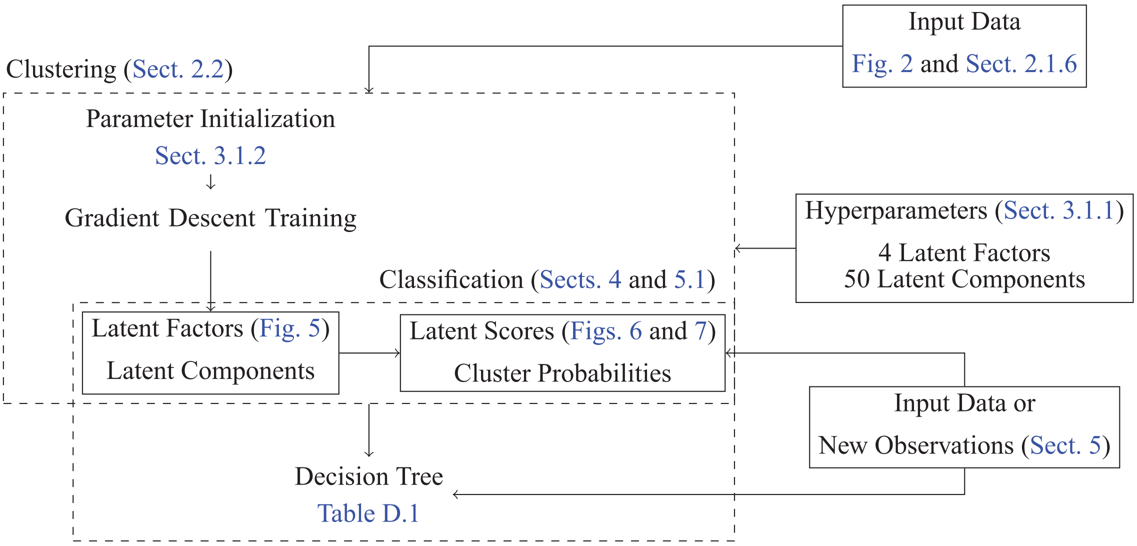


Fig. 4. Overview of the clustering and classification of the input observations. The MCFA model encompassed in the upper dashed rectangle can be computed using the `mcfa` `python` package. The classification of the input data or new observations in the lower dashed rectangle can be done using the `classy` `python` package described in Sect. 5.

mean imputation before training the model using an expectation-maximisation algorithm. Mean imputation is not appropriate for our dataset as we know that the spectra of different asteroid classes may appear entirely different in terms of absorption features and slope. Inserting the mean column value in each empty cell thus does not represent the missing data well. Instead, we use the `tensorflow` library (Abadi et al. 2015) to implement a stochastic gradient descent learning strategy which maximises the log-likelihood of the model given the observed data only, which is statistically sound under the missing-at-random assumption (Little & Rubin 2019), contrarily to using mean imputation. The stochastic gradient descent is of particular interest here as it estimates the model parameters based on batches of the input data, meaning that it can scale easily with an increasing number of observations. This MCFA implementation is independent of the taxonomy itself and may be applied in different studies. The implementation and documentation are available online[9].

3. Results

In this section, we present the results of fitting the MCFA model outlined in Sect. 2.3 to the input dataset described in Sect. 2.1. After depicting the latent space and the structure of the latent dimensions, we explain how the asteroid classes building this taxonomy are derived from the modelled Gaussian clusters. An overview of the clustering steps is given in Fig. 4.

3.1. Model fit

3.1.1. Parameters

We choose to cluster the asteroid observations in $q = 4$ latent dimensions using $g = 50$ Gaussian clusters. Both numbers are selected from a wide range of values after assessing the resulting model fits. Larger values retain and describe more variability in the data, and at the same time increase the number of free parameters in the model, hence a trade-off is made in both cases. The model fits obtained with four or five latent factors were comparable in terms of captured variability in the cluster, thus we opted for the smaller number of model parameters.

The large initial number of 50 clusters accounts for the model assumption of Gaussianity in the latent space. We have no reason to expect a Gaussian distribution of the asteroid classes; therefore, we model them as superpositions of one or more Gaussian clusters. The modelled clusters are later joined and mapped to build the asteroid classes using a many-to-many relationship and following a decision tree.

3.1.2. Initialisation and training

The latent loadings and cluster assignments of each observation have to be initialised at the start of the gradient descent algorithm to train the MCFA model. The initialisation dictates the global position in the Hamiltonian which is sampled by the training and thus has a significant impact on the final result.

A practical issue when reducing the dimensionality of asteroid data made up by different observables is the feature weighting. In our case the spectra contribute 53 data dimensions compared to the single dimension of the albedo. The summed variance in the former is much larger than the variance in the latter, resulting in a negligible contribution of the albedo to the latent space computation which does not reflect its actual information content. Tholen (1984) therefore chose not to include the albedo in the dimensionality reduction, using it in a subsequent manual clustering step instead. We employ an alternative strategy outlined below which allows us to account for the albedo while building the latent space.

We initialise the latent loadings using PPCA. This approach has two advantages. First, the latent loadings are set to the axes of largest variance in the data, ensuring a high resolution in the latent space, and second, PPCA is variant to feature scaling (i.e. data dimensions are weighted with respect to their variance when computing the dimensionality reduction). An effective way to increase the importance of the albedo information is hence to increase the variance of albedo values by some transformation prior to model training. We achieve this by means of the $\log_{10}$ transformation described in Sect. 2.1.3, which increases the variance in the albedo dimension by a factor of 6.8. During the

[9] https://github.com/maxmahlke/mcfa

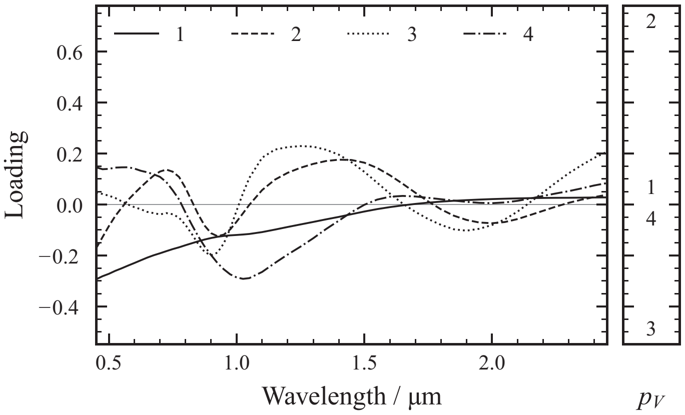


Fig. 5. Four latent components of the mixture of common factor analysers model trained on the input data. The left side gives the loading of the spectral data dimensions for each latent component, while the right side shows the loading corresponding to the albedo.

gradient-descent model training, we monitor the log-likelihood of the model given the data. As opposed to PPCA, MCFA is invariant to factor scaling, which leads to a decrease in the albedo loadings with each training step. Therefore, we do not train until the model has fully converged, instead stopping the training when a good balance between the weight of the albedo and of the spectra has been achieved. This subjective choice of training epochs is a concession we make to the challenge of combining different observables in the same model.

The latent cluster memberships are initialised by fitting a Gaussian mixture model with 50 components to the principal scores of the PPCA and assigning each sample to its most probable cluster. We train the MCFA model on the 2983 observations of 2125 individual asteroids as outlined in Sect. 2.1.4.

3.2. Latent space

During the model training the latent components matrix $\boldsymbol{W}$ is derived based on the covariance of the input observations. Each latent component contains one linear coefficient for each input data dimension (i.e. the latent loading). The absolute value of a loading indicates the degree to which the latent component responds to variance in the corresponding data dimension. Positive loadings lead to an increase in the latent scores z with increasing value in the data dimension, negative values to a decrease in z. The latent scores Z are essentially a vector product of the input data with the latent components. As such, both the spectra and the visual albedo of the observations influence the latent scores Z simultaneously.

The latent components resulting from the model training are depicted in Fig. 5, with the spectral loadings given on the left side and the latent loadings corresponding to the albedo dimension on the right side. We note that they are displayed separately only for visualisation purposes; for the clustering model itself, there is no principal distinction between the latent loadings corresponding to the spectra and that corresponding to the albedo.

The spectral loadings in Fig. 5 resemble different mineralogical features commonly present in asteroid spectra. The first component approximates a positive slope[10], with an inflection point around 1 µm. Components two and three resemble the spectra of pyroxene minerals due to their bands at 1 µm and 2 µm, though the band minima and depths differ between the components. The strongest distinction between these two components is the visible slope, which is positive for component two and negative for component three. Component three has a slight absorption feature at 0.7 µm. The fourth component depicts an olivine-like 1 µm band structure. The albedo contributes marginally to the first and fourth latent component, while component two has a large positive loading and component three a large negative loading to it.

The latent scores of the asteroid observations are shown in Figs. 6 and 7. The input data depicts a larger variance when projected along the first two components rather than along the last two due to the initialisation of the latent components with PPCA. It is clear that the featureless spectra will show little variance when projected along the pyroxene- and olivine-like axes z_2, z_3, and z_4.

Figures 6 and 7 additionally indicate the mean latent scores of all asteroids assigned to a given asteroid class, designated by the class letter and derived in the following sections. As an example for the interpretation of latent scores, we point out here that the degeneracy between classes E and S in the latent scores depicted in Fig. 6 is the result of the large loading of the albedo in the second component, which offsets the generally featureless E-types with respect to the feature-rich but darker S-types. This degeneracy is resolved in other latent components, as can be seen in Fig. 7.

3.3. Clusters

Concurrent with the dimensionality reduction, the input data is divided into 50 Gaussian clusters during the model training. The clusters are not constrained in their covariances, yielding a wide range of cluster shapes and orientations in the latent space. Illustrating the distribution of the clusters in the four-dimensional latent space is not practical due to their large number; instead, we show the distributions of input spectra and albedos over the clusters in Figs. 8 and A.1 respectively.

Most clusters occupy a narrow volume in the latent space and encompass Gaussian populations in previously recognised classes such as S and V. When building the asteroid classes from the clustering, we map the probability of any sample to belong to either of these narrow clusters one-to-one to the respective asteroid class. As an example, for all observations the probability of belonging to cluster 0 is added to the probability $p_{\mathtt{S}}$ of belonging to the S-types (see Fig. 8). Additional S-like clusters such as cluster 6 further add to $p_{\mathtt{S}}$; 33 clusters are mapped to a single asteroid class in this manner.

Other clusters either capture continuous trends between classes or the diffuse background population. An example of the former is cluster 22, containing spectra from both M- and P-types, and of the latter cluster 13, containing observations with varying spectral characteristics and albedos. For these clusters, we implement decision trees to separate the observations into mostly two or three distinct classes. These decision trees are described in Sect. 4 on a per class basis at the end of each class description. The probability of belonging to either of these clusters is split and added to the respective class probabilities following the decision trees. As an example, cluster 22 is resolved via the albedo. If no albedo is present in the observation, the cluster probability is added entirely to $p_{\mathtt{X}}$, otherwise it is split between $p_{\mathtt{M}}$ and $p_{\mathtt{P}}$ proportionally based on the albedo distribution of M- and P-types, derived in Sect. 3.4.2.

For clusters 13, 29, and 41, we note that they capture objects with high variability in their spectral and albedo features. These are either unique objects, such as the only O-types

[10] The latent loadings represent the variance of the ln-transformed spectra.

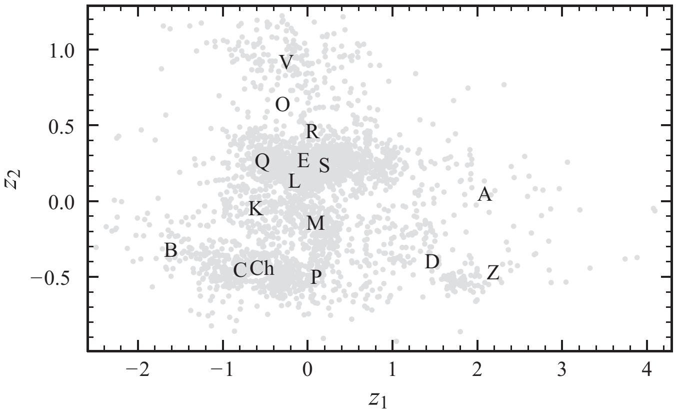


Fig. 6. Latent scores of the input data projected along the first two latent components (grey circles). The mean score of all asteroids assigned to a given class is indicated by the class letter. For better readability, the mean score of class C has been shifted by −0.1 in z_1.

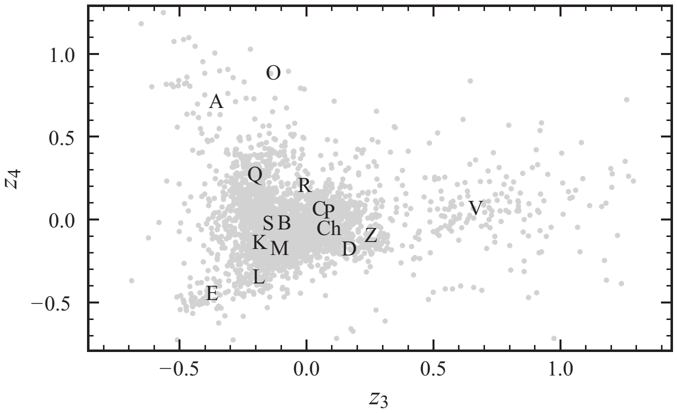


Fig. 7. As in Fig. 6, but giving the scores in the third and fourth latent components. For better readability, the mean score of class S has been shifted by −0.02 in z_3 and of classes C and P by 0.04 in z_4.

(3628) Boznemcova and (7472) Kumakiri in cluster 13, or spectra of questionable quality. We resolve these clusters with decision trees based on GMMs into different classes: cluster 13 into C, O, Q; cluster 41 into B and V; and cluster 29 into every class except for E, K, L, O, R, X, and Z (see Table D.1). Objects in either of the three clusters are flagged in the classification output as DIFFUSE and should undergo visual scrutiny.

3.4. Classes

3.4.1. Class continuity

When deriving the mapping of the Gaussian clusters to the asteroid classes, we strive to maximise the resemblance of the resulting taxonomy to the established system by Tholen (1984) and the Bus-DeMeo system. For any change in the classes, we weigh the evidence in the data to necessitate the change against the overall practicality of class continuity, opting for the latter when in doubt. Furthermore, we also take into account mineralogic and meteoritic interpretations established in the literature using observables outside this feature space, allowing us to derive classes which are more useful for communicating class properties within the community. These influences from outside the data-driven approach are stated in the description of the respective class in Sect. 4.

The main drivers for the evolution of the class scheme are twofold. The first is the fundamental difference between the probabilistic clustering employed here and the visual clustering used in previous schemes, affecting specifically classes that reflect continuous trends in the asteroid population. The second is the reintroduction of the albedo to the observables of the taxonomy.

The fundamental division of asteroids into feature-poor and feature-rich populations, the C- and S-complexes, is the baseline of our scheme, as it has been since the first taxonomic efforts by Chapman et al. (1975). A small population of asteroids with faint features occupies the space between these complexes in DeMeo et al. (2009), separated into the classes K, L, Xc, Xe, Xk, and T. Thanks to the taxonomic information provided by the albedo and targeted campaigns of these populations (e.g. Neeley et al. 2014; Ockert-Bell et al. 2010), this population has grown considerably, to the point that we recognise it as a third complex, which we dub the M-complex based on its most populous class.

Taxonomic constants such as the A- and V-types represent no challenge in identification. It is more difficult to prove the definition of the Q-types, which represent a continuous trend towards smaller slopes compared to the S-types and as such does not separate clearly in the latent space. In favour of class continuity, we still identify a population in the S-complex as Q-types. Subclasses such as Sa, Sq, and Sr are not identified, however, as we observe numerous clusters with varying slopes and mineralogies in the S-complex. Labelling each cluster with a secondary letter would increase the entropy of the taxonomic system, and would lead to more confusion than resolution. Furthermore, we note an overall large variability between observations of single asteroids which often exceeds the variability between these subclasses. Instead, we highlight the different mineralogical interpretations of these clusters in Sect. 4.5.1.

3.4.2. Resolving the X-complex

Solving the spectral degeneracy of the X-complex in the Bus-DeMeo scheme is the main motivation to reintroduce the albedo to the taxonomic system. We employ the system established in Tholen (1984); asteroids in the X-complex are differentiated based on their albedo values and are labelled P, M, and E in ascending order of albedo, while the letter X is retained for observations without albedo. However, instead of applying strict limits[11], we model the joint albedo distribution of all observations in clusters that we consider to be X-like based on their spectral appearance: clusters 17, 22, 35, 37, and 46. The employed model is a GMM with three components. In Fig. 9, we show the model fit to the albedo distribution of the X-complex, as well as the derived mean and standard deviations in p_V for classes E, M, and P. Any asteroid that falls into one of these clusters and has an albedo observation is assigned based on its probability in this model to the respective class. The subclassification indicating the presence of features in the spectra (e and k) is retained and discussed in the following subsection.

3.4.3. Feature flags

The Bus-DeMeo system recognises four classes which are based on the presence of distinct absorption features in addition to the overall shape of the spectra: (1) Ch, exhibiting a feature around 0.7 µm associated with possible surface hydration (e.g. Rivkin et al. 2015); (2) Xe, showing a narrow feature at 0.5 µm (Bus & Binzel 2002a); (3) Xk, depicting a faint broad feature between

[11] Tholen (1984) applied visual albedo separations of ~0.06 between P and M and ~0.28 for M and E.

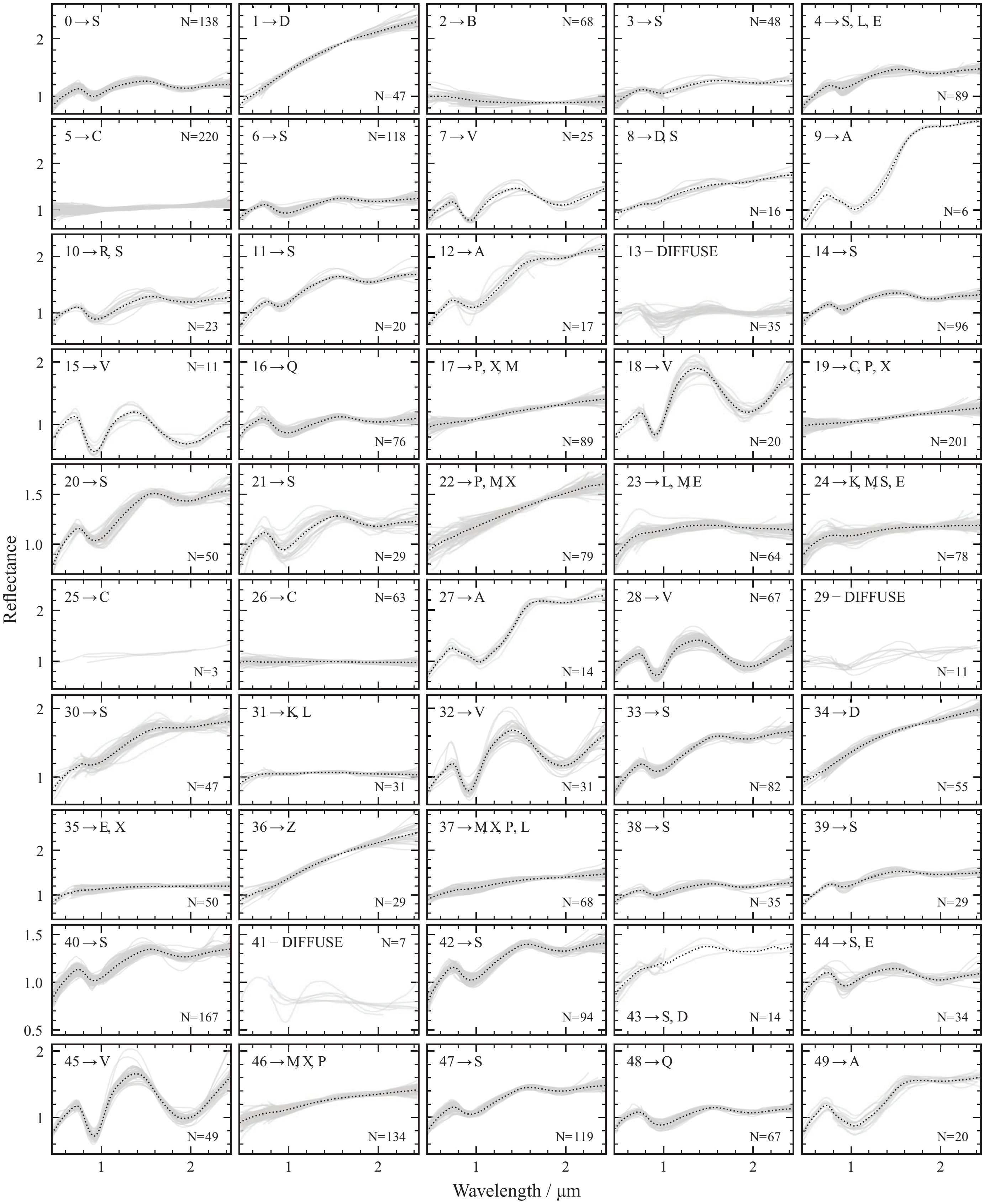


Fig. 8. Overview of asteroid spectra assigned to each cluster, including the number N of spectra and the asteroid classes to which the cluster contributes, excluding classes with fewer than three contributed observations except for cluster 25 which only has three members. The classes are sorted by the total number of observations the cluster contributed. The dotted line gives the mean value of the spectra per cluster except for diffuse clusters (defined in Sect. 3.3) and cluster 25. The mean spectra are normalised to unity at 0.55 µm. The y-axis limits change in each row.

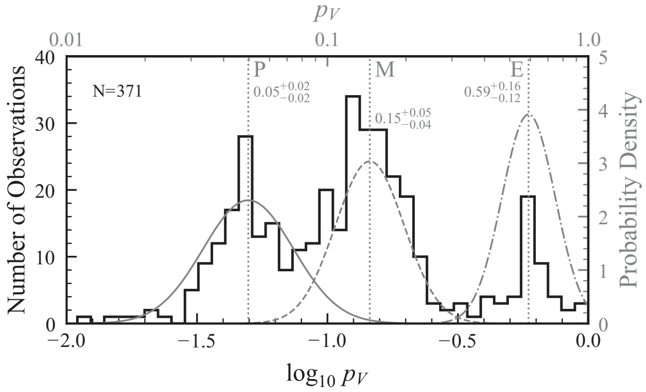


Fig. 9. Distribution of visual albedos in clusters associated with the X-complex. The spectral degeneracy of the X-complex is resolved by fitting a three-component Gaussian mixture model to its albedo distribution, consisting of N observations and shown in the histogram. The fitted components are given by the solid, dashed, and dash-dotted grey lines in terms of the probability distribution. The vertical dotted lines give the mean values of components, labelled by the established class designations P, M, and E in order of ascending albedo. The numbers below the class labels give the mean p_V and the upper and lower 1 σ limits per class. We note that these values slightly change later as other class members are added from clusters which are not assigned purely to the X-complex. The final albedo distributions are given in Table 3.

0.8–1.0 µm (Bus & Binzel 2002a); and (4) Xn, with a feature around 0.9 µm (Binzel et al. 2019). Example spectra carrying the e-, h-, and k-features are shown in Fig. 10.

The identification and flagging of these features by use of the secondary letters in the class designation is carried over in this scheme with a slight modification. First, we do not differentiate between the k- and the n-feature. Both are centred around 0.9 µm and after slope removal, we find no appreciable systematic difference between the features in a sample of spectra previously classified as Xk or Xn. We do not rule out that these features are imprinted by different surface mineralogies; however, we chose the evidence in the data over class continuity, we decided to drop the n-feature, and continued with only the k-feature.

Second, we do not reserve unique classes for observations depicting the e- or the k-feature. As discussed in Sect. 4.3, both features are prevalent in members of the X-complex showing a variety of spectral slopes and albedos. We judge these two properties to be more important when deriving classes than the presence of a single feature. Furthermore, we note that e and k are not mutually exclusive; for example, (2035) Stearns depicts both features as shown in Fig. 10. We thus decided to flag the presence of these features by appending the respective letter to the class designation without considering the resulting combinations such as Mk or Eek as proper classes.

On the other hand, the h-feature is treated consistently with the Bus-DeMeo system. It is exclusive to the members of the C-complex and displays a much narrower, continuous distribution than the other two features, as shown in Sect. 4.1. Any sample depicting the 0.7 µm band is assigned to the Ch-class, regardless of the subclass in the C-complex that the spectra falls in.

The features are identified in a semi-automated manner. For each feature we defined a wavelength interval around the band centre in which the spectral continuum is removed and the reflectance is fitted using a polynomial of fourth degree, following Fornasier et al. (2014). Both the interval and the expected band centre were defined heuristically using a training sample of visually identified features, and are given in Table B.1. Using the polynomial fit, we estimate the band depth with respect to the continuum, the band centre, and its signal-to-noise ratio. The last is given by the ratio of the band depth to the reflectance uncertainty, which is estimated using the residuals of the polynomial fit. The band is considered to be present if the band centre is within three standard deviations of the expected position derived from the training sample and the signal-to-noise ratio is higher than one.

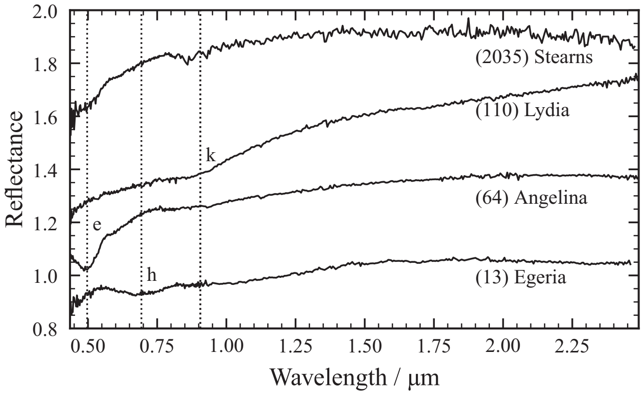


Fig. 10. Example spectra carrying the e-, h-, or k-feature which are recognised in this taxonomic system. The mean band centres derived from all visually identified features in the spectral observations is indicated by the vertical dotted lines (e: 0.50 µm, h: 0.69 µm, k: 0.91 µm). (2035) Stearns exhibits both the e- and the k-feature. Data from SMASS (http://smass.mit.edu).

The fitting procedure is run automatically to identify the h-feature in spectra classified as members of the C-complex (B, C, P, and the degenerate class X; see Sect. 4.1) and the e- and k-features for those belonging to the X-complex (E, M, P, and X). In practice, we find that relying on the automated band identification yields many false positives given the low threshold of one in the signal-to-noise ratio and the general uncertainty of the expected wavelength of the band centre. For example, Cloutis et al. (2018) give band centres between 0.6 µm–0.75 µm for the h-feature. Hence, we recommend a semi-automated approach where the bands are fitted automatically and the observer visually confirms the quality of the fit and the presence or absence of the band. The fitting and confirmation are handled by the classification tool presented in Sect. 5. In the 2983 spectra classified during the clustering, 13 (144, 135) carry the e-feature (h-feature, k-feature). For 392 spectra (361, 360), no conclusion could be made as the spectral region is missing.

The k-feature is particularly challenging to observe as it falls in the transition of visible and near-infrared spectra, which are acquired using different instruments. Merging the spectral parts is non-trivial and several subjective decisions have to be made, as outlined in Clark et al. (2009). The unknown offsets between visible and near-infrared can give rise to an artificial feature when joining the observations. In the case of the e-feature, Bus & Binzel (2002b) point out a systematic feature between 0.515 µm and 0.535 µm in the SMASS spectra, which are frequently used to complement acquired NIR-only spectra. Hence, we note here that the e-feature should only be considered present if its band centre is well below this wavelength range.

3.4.4. Class per asteroid

A total of 549 of 2125 asteroids in the input data have more than one sample in the input data. These observations may or may not

Table 1. Distribution of observations and asteroids over taxonomic classes and orbital populations.

			Fraction		Orbital Class								
Class	Samples	Asteroids	This work	DM09	NEA	MC	H	IMB	MMB	OMB	Cyb	Hilda	JT
A	57	32	1.5	1.6	2	3	2	7	10	8	–	–	–
B	68	45	2.1	1.1	15	4	1	12	5	8	–	–	–
C	299	221	10.4	7.3	69	8	2	89	72	79	2	2	5
Ch	144	107	5.0	4.8	9	2	–	20	47	26	2	–	1
D	119	82	3.9	4.3	6	1	–	1	4	5	5	16	44
E	65	46	2.2	–	7	4	27	4	3	1	–	–	–
K	59	42	2.0	4.3	21	2	–	5	2	12	–	–	–
L	76	58	2.7	5.9	20	4	3	4	22	3	–	–	2
M	252	142	6.7	–	29	7	2	17	47	28	–	2	10
O	4	2	0.1	0.3	–	–	–	–	1	1	–	–	–
P	195	135	6.4	–	14	6	1	11	26	36	12	12	17
Q	158	107	5.0	2.2	89	5	–	7	4	2	–	–	–
R	15	10	0.5	0.3	7	–	–	2	–	1	–	–	–
S	1188	898	42.3	53.8	404	101	35	140	172	45	–	1	–
V	206	142	6.7	4.6	28	2	–	104	4	4	–	–	–
X	50	33	1.6	8.6	20	8	2	1	–	2	–	–	–
Z	28	23	1.1	–	1	–	1	4	6	3	–	1	7
Σ	2983	2125	100	98.9	741	157	76	428	425	264	21	34	86

Notes. The second column gives the number of observations assigned to each class, while the third and all following columns refer to the number of individual asteroids assigned to the class. DM09 refers to DeMeo et al. (2009). The fractions in this column do not add up to 100%, due to the missing T-class in this scheme. The orbital classes use the following acronyms: NEA – near-Earth asteroids; MC – Mars-crosser; H – Hungaria; IMB – inner main belt; MMB – middle main belt; OMB – outer main belt; Cyb – Cybele; JT – Jovian trojans.

have been assigned to the same class, opening the possibility that asteroids have different classes assigned. We resolve these ambiguities by computing the sum of the class probabilities across all observations of the asteroid, weighted by the fraction of observed data dimensions. Observations with albedo values received an additional weight corresponding to 25 data dimensions, meaning that a visible-only spectrum including albedo has approximately as much weight as a VisNIR spectrum without albedo. If one of the e-, h-, or k-features is detected in any of the observations, the final class of the asteroid carries the respective suffix letter.

In Table 1, we report the total number of observations per taxonomic class, followed by the number of distinct asteroids in the class. The latter number only includes asteroids which were assigned to the class after the merging procedure outlined above in the case of multiple observations.

4. Discussion

In the following, we discuss the main properties of the 17 classes defined in this taxonomy in data and latent space, structured into three complexes: C, M, and S. We give our motivation for class scheme and point out where it aligns with or deviates from the existing classifications, in particular the taxonomy by Tholen (1984) and the Bus-DeMeo system (Bus & Binzel 2002a; DeMeo et al. 2009), which are the closest predecessors in terms of the observables. We further outline the decision tree used to derive the classes from the 50 clusters that were fit to the input observations in the previous section. An overview of this decision tree is given in Table D.1

A general overview of the class properties in data space is given in Figs. C.1 and C.2. Table 1 gives an overview of the number of samples and asteroids per taxonomical and orbital class. Tables 2 and 3 show an evolution of the taxonomic scheme and describe the classes defined in this taxonomy, including an overview of the spectra of class prototype asteroids, most of which are discussed in the text. The mean spectra and albedos for each class ('class templates') are available in the CDS repository. The X-class is not discussed separately as its members are covered by classes E, M, and P.

4.1. C-complex: B, C, Ch, P

The members of the C-complex are found throughout the main belt and dominate the regions past the 3:1 mean-motion resonance in terms of number and mass (DeMeo & Carry 2014; Vernazza et al. 2017). Their spectral appearance is generally feature-poor apart from the h-feature at 0.7 μm observed in about one-third of the population and associated with phyllosilicates present on the surface (Rivkin 2012). Instead, the diversity of the complex constituents is present in the slope and in the shape of the spectra, the former ranging from blue over neutral to red and the latter from overall linear to a concave appearance attributed to a carbonaceous surface composition including magnetite (Chapman et al. 1975; Cloutis et al. 1990b; Gaffey & McCord 1979).

Common meteorite linkages to the population of the C-complex involve carbonaceous chondrites such as CI, CK, CM, and CO with different degrees of thermal metamorphism or aqueous alteration (Clark et al. 2010; Cloutis et al. 2011; de León et al. 2012; Hiroi et al. 1996). However, the paucity of these meteorite groups among the falls even after bias-correction is difficult to reconcile with the abundance of the complex members in the main belt, leading Vernazza et al. (2015) to suggest interplanetary dust particles (IDPs) as analogues for the non-hydrated asteroids. Using a radiative transfer model, the spectral appearance of most C-complex asteroids is well matched using constituents of chondritic-porous IDPs. The open question on

Table 2. Evolution of taxonomic scheme from Tholen (1984) to Bus-DeMeo to this work.

Tholen		Bus-DeMeo		This work
B	→	B	→	B
F	↗			
G	→	Cg	↘	
	→	Cgh	↘	
C	→	C	→	C
	→	Ch	→	Ch
	→	Cb	↗	
D	→	D	→	D
			→	Z
P	...	Xc	...	P
M	...	Xk	...	M
X	...	X	...	X
E	...	Xe	...	E
	...	Xn	|	
T	→	T	|	
		K	→	K
		L	→	L
Q	→	Q	→	Q
		Sq		
	↗	Sr	↘	
S	→	S	→	S
	↘	Sa	↗	
		Sv		
O	→	O	→	O
R	→	R	→	R
A	→	A	→	A
V	→	V	→	V

Notes. Arrows are used to indicate the overall evolution of each class. The T-class is not present in this taxonomy and the feature characteristic of the Xn has been grouped into the k-feature. The evolution of the X-complex between the taxonomies is unclear as the visual albedo is not taken into account in the Bus-DeMeo system. No analogues for K and L were defined in Tholen (1984).

the surface composition is decisive for the behaviour of the asteroids under the influence of spectral weathering. Laboratory irradiation experiments (Lantz et al. 2017, 2018) and statistical approaches (Thomas et al. 2021) both show opposite trends for different initial surface compositions: while high-albedo material exhibits spectral reddening and surface darkening, low-albedo assemblages become bluer in slope and brighter.

Apart from the C-types, Tholen (1984) defined three smaller classes based on the albedo and UV distributions: B-types are ‘bright-C’ types with visual albedos around 10%, while F- and G-types are characterised by their behaviour in the UV wavelength region (the former flat, the latter showing strong absorption behaviour). The Bus-DeMeo system retained classes B and C and extended the taxonomy by addition of the Ch-class for hydrated C-type asteroids, as well as the classes Cb, Cg, and Cgh, which describe different slope behaviours in different wavelength regions. Neither system counts the members of P-class as members of C, but rather as member of the X-complex.

In this taxonomy, we divide the C-complex into four classes: B, C, Ch, P. The P-class is here defined for the first time in both albedo and spectral appearance, allowing us to move it from the X-complex and firmly establish it as part of the C-complex. Any object within the complex that exhibits the h-feature is classified as a Ch-type, even if it falls in B or P. The distribution of reflectance spectra and visual albedos for each class is shown in Fig. 11. The heterogeneous yet continuous distribution of the C-complex members in latent space is illustrated in Fig. 12. As change in slope and a broad feature around 1 µm–1.3 µm are the main differentiators, the complex members split best in the z_1 and z_4 latent dimensions[12]. We note that the apparent diagonal gaps between the C- and Ch-class members in the lower part of Fig. 12 are an artefact of the spectral normalisation (see Sect. 2.1.2) and are not of a physical nature, as shown by the large number of asteroids which have samples on either side of the gaps.

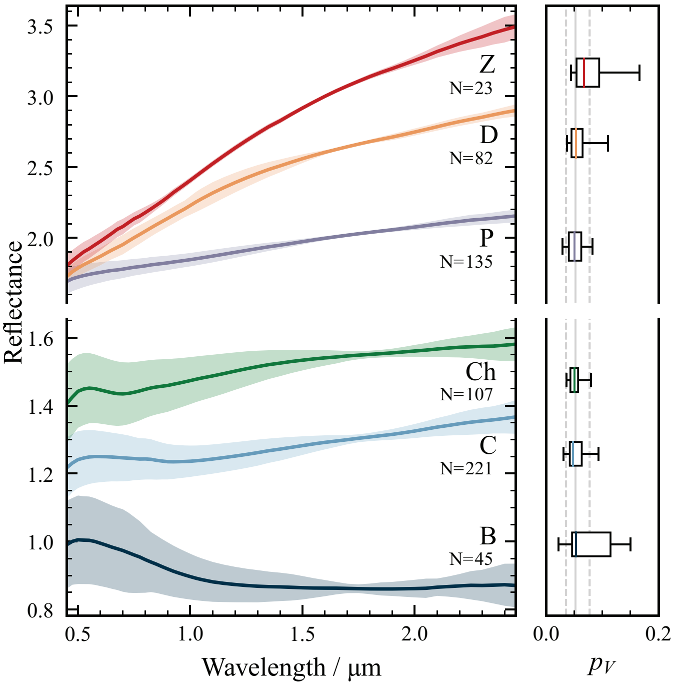


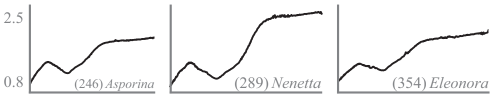
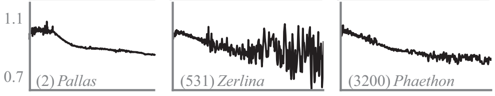
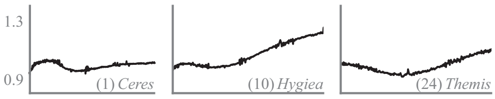
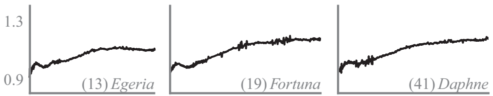
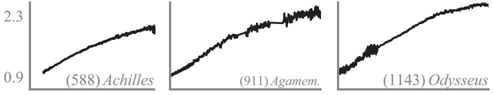
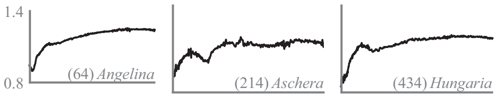
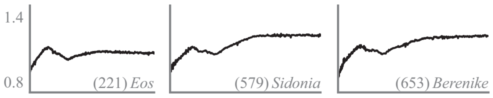
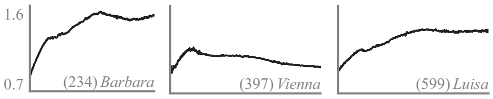
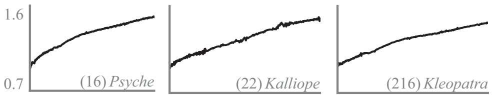
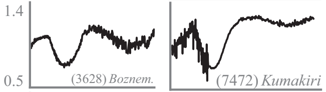
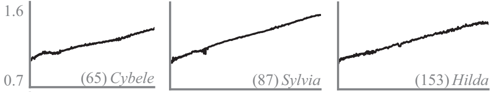
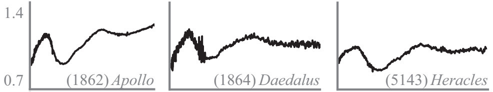
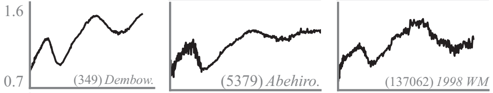
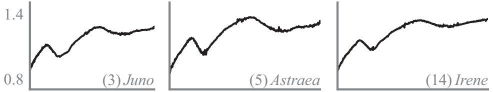
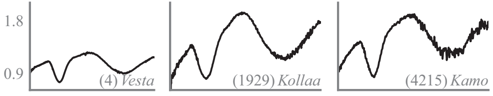
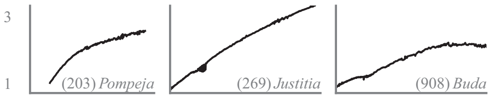
Fig. 11. Mean (solid line) and standard deviation (shaded area) of the reflectance spectra for each class and endmember of the C-complex on the left hand side. The spectra are shifted along the y-axis for comparability. The reflectance scale changes between B, C, Ch and P, D, Z. The number N of individual asteroids assigned to each class is given below the class letter. On the right side are given the median (solid line), the lower and upper quartiles (box), and the 5th and 95th percentiles of the distribution of visual albedos within the class. The vertical grey lines give the mean albedo (solid) and the upper and lower standard deviation (dashed) within the whole complex. These latter values are $0.05^{+0.03}_{-0.02}$ for the C-complex.

4.1.1. B-types

The B-class was first defined in Tholen (1984) based on their average albedo, which is higher in comparison to the other members of the C-complex. With the disappearance of the UV wavelength region from taxonomy, F-types are no longer distinguishable from B-types, and the distribution of generally high albedos of the latter has become a broad distribution

[12] We consider an absorption feature to be concave, while other works such as Lantz et al. (2018) define it as convex.

Table 3. Description of taxonomic classes defined in this work.

Class	Spectrum	Albedo	Prototypes
A	Broad and deep absorption feature at 1 µm, strong red slope in the near-infrared.	$0.25^{+0.09}_{-0.07}$	
B	Neutral to blue slope in the visible, blue slope in the near-infrared.	$0.06^{+0.05}_{-0.03}$	
C	Red visible slope with a possible broad feature around 1 µm and a red near-infrared slope. The spectrum might have an overall concave shape.	$0.05^{+0.02}_{-0.01}$	
Ch	Absorption feature at 0.7 µm. The near-infrared slope is red while the overall shape might be convex.	$0.05^{+0.02}_{-0.01}$	
D	Featureless with steep red slope with a possible convex shape longwards of 1.5 µm.	$0.06^{+0.03}_{-0.02}$	
E	Strong red slope in the visible with a feature around 0.9 µm of varying depth and a neutral near-infrared continuation.	$0.57^{+0.15}_{-0.12}$	
K	Strong red slope in the visible with a broad feature around 1 µm followed by a blue to neutral near-infrared slope.	$0.13^{+0.04}_{-0.03}$	
L	Variable appearance apart from a red visible slope. A small feature around 1 µm and a possible one at 2 µm. The near-infrared slope is blue or red.	$0.18^{+0.07}_{-0.05}$	
M	Linear red slope with possible faint features around 0.9 µm and 1.9 µm. Might show convex shape in the near-infrared.	$0.14^{+0.05}_{-0.04}$	
O	Broad, bowl-shaped 1 µm absorption feature and a weaker feature at 2 µm.	$0.26^{+0.02}_{-0.02}$	
P	Linear red slope and generally featureless. Less red than D-types.	$0.05^{+0.02}_{-0.01}$	
Q	Broad absorption at 1 µm and a shallow feature at 2 µm. An overall blue slope in the near-infrared.	$0.24^{+0.12}_{-0.08}$	
R	Strong feature at 1 µm and a feature at 2 µm. The latter feature is shallower than in V-types.	$0.30^{+0.05}_{-0.04}$	
S	Moderate features around 1 µm and 2 µm and a neutral to red near-infrared slope.	$0.24^{+0.10}_{-0.07}$	
V	Deep absorption features at 1 µm and 2 µm. The former is much narrower than the latter.	$0.29^{+0.11}_{-0.08}$	
Z	Extremely red slope, redder than the D-types. Featureless but may exhibit concave shape in the near-infrared.	$0.07^{+0.04}_{-0.03}$	

Notes. Listed are the spectral appearance, visual albedo distribution giving the mean value, the lower and upper standard deviation, and the spectral prototypes of the 17 classes defined in this taxonomy excluding the X-types.

with a standard deviation of around 10%, see Fig. 11. This distribution is further visible in the large variance in the z_2-scores of B-types (Fig. 12). Instead, the bright C (Tholen 1984) are best identified by another common interpretation of the class mnemonic, their blue slope longwards of ~0.7 µm, causing a readily apparent distinction from other classes specifically in the z_1 latent score. Nevertheless, the B-types do not separate entirely from the neighbouring C-types and form a diffuse but continuous branch of the complex, as shown in Fig. 12.

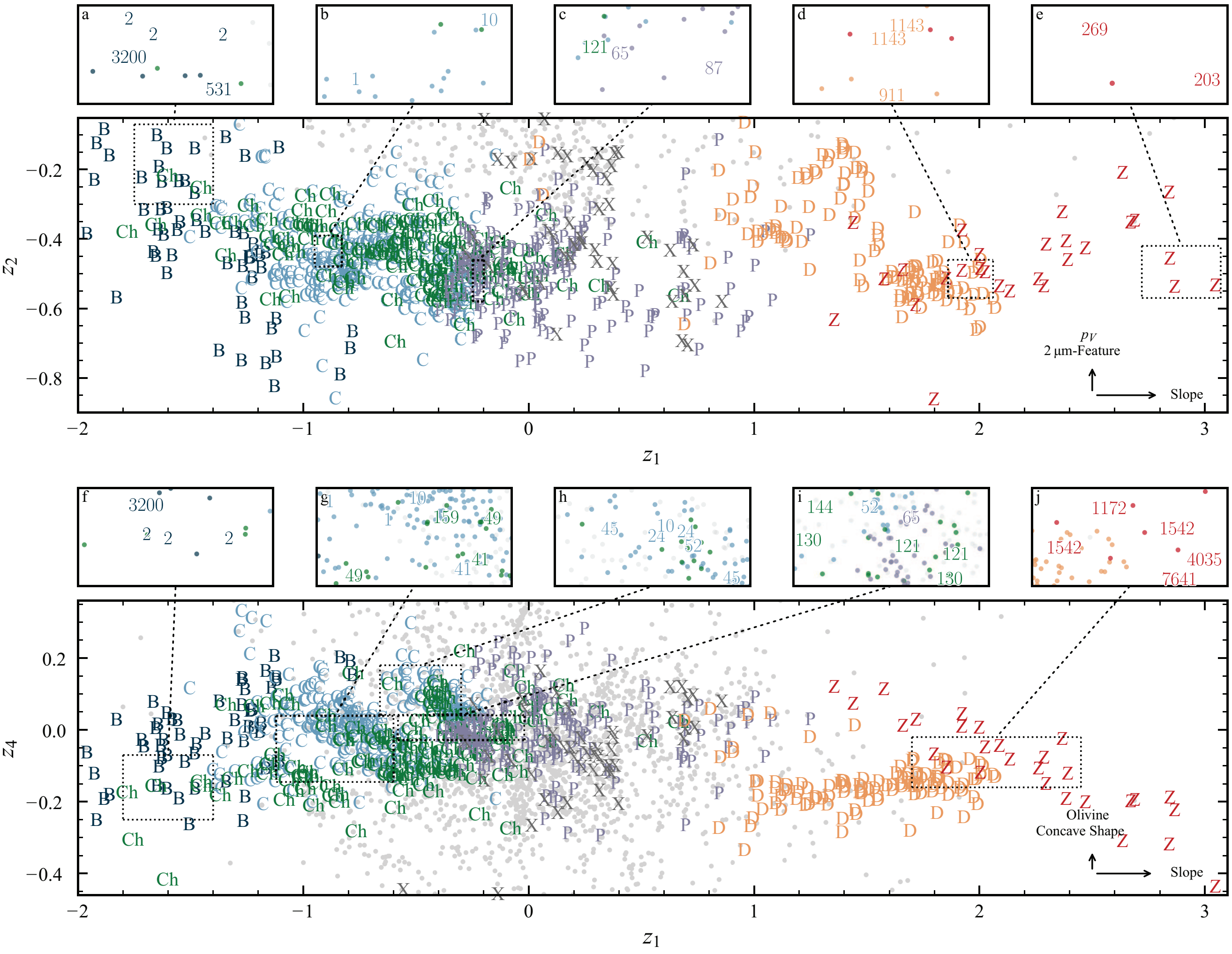


Fig. 12. Distribution of C-complex and its endmember classes D and Z in the first latent component vs the second (*top*) and the fourth (*bottom*) latent components. The samples assigned to each class are given with the respective class letter. The latent scores of all samples outside these classes are shown as grey circles. Some outliers in z_2 and z_4 are not shown for readability. The five subpanels above each panel show regions of interest where a selection of asteroids are highlighted by replacing the symbol with the respective asteroid's number. If more than one spectrum of the asteroid is in the input data, its number may appear several times.

The class variance in z_1–z_2 indicates that bluer B-types also tend to be brighter. As shown in subpanel a in Fig. 12, the archetype B-type (2) Pallas and near-Earth asteroid (3200) Phaethon are among the bluest and brightest class members. (531) Zerlina is further highlighted as a member of the Pallas collisional family, for which Alí-Lagoa et al. (2016) note a significantly higher average albedo compared to the remaining B-types. In z_4 B-types have higher scores than the other C-complex members, with the z_1 score due to the visible part of the fourth latent component resembling the B spectral region (compare Figs. 5 and 11).

A total of 45 asteroids (2.1%) are classified as B-types in this study. The B-class is made up of a single cluster (2) and is not subject to any decision tree. We note that the Themis-like B-types with a neutral-to-reddish slope in the NIR, as described in Clark et al. (2010) and de León et al. (2012), are C-types in this taxonomy, in agreement with their classification in the Bus-DeMeo system (see subpanel h in Fig. 12).

4.1.2. C-types

The carbonaceous C-types present spectra with a neutral to small red slope and are generally featureless except for a broad feature around 1.3 µm, which may give the spectrum an overall concave shape. In the upper part of Fig. 12, we observe a uniform distribution of the C-types in z_1–z_2 with the class variance aligned with the z_1 axis; z_2 is not a suitable projection for the C-types as they are featureless and present a narrow albedo distribution, as shown in Fig. 11. Instead, the concave feature shape is captured in z_4, hence in the lower part of Fig. 12 we observe a more structured clustering. The positive correlation of z_1 and z_4 scores among the C-types indicates that the spectra on average get more concave as they get redder. Nevertheless, the wide and continuous distribution around this general trend prevents us from defining analogues to the classes Cb, Cg, and Cgh in the Bus-DeMeo system as we aim to refrain from subjectively partitioning the latent space.

Both (1) Ceres and (10) Hygiea are members of the C-class (see subpanel b in Fig. 12). In subpanel h, we highlight (24) Themis, (45) Eugenia, and (52) Europa. All these asteroids are well matched by the models composed of IDP constituents as described in (Vernazza et al. 2015) and have on average higher z_4 scores than the Ch-class members of comparable slope.

A total of 221 asteroids (10.4%) are classified as C-types in this study. C-types are present in three different clusters (5, 19, 26, where the first two are the two largest of the 50 clusters in the model). Cluster 19 contains both prominent C-types such as (45) Eugenia and (52) Europa as well as prominent P-types such as (65) Cybele and (87) Sylvia, as shown in subpanel c in Fig. 12. The cluster resembles the Cb-class from the Bus-DeMeo system. We split this cluster into two components (C and P) using a GMM in z_1–z_4. While we generally aimed to keep the number of post-clustering decision trees to a minimum, we make the choice here to follow the mineralogical interpretation of the C-complex given in Vernazza et al. (2015) and Marsset et al. (2016), among others, and to increase class continuity for the objects in these clusters.

4.1.3. Ch-types

Unlike for the other feature flags outlined in Sect. 3.4.3, we reserve a unique class for the 0.7 µm h-feature, following the convention of the Bus-DeMeo system. We observe the continuous and narrow distribution of samples carrying this feature similar to the other classes in the C-complex. Furthermore, as above for the C-types, we recognise the mineralogical and meteoritic interpretation of the C-complex members in the literature (e.g. Cloutis et al. 2011; Marsset et al. 2016; Vernazza et al. 2015).

While degenerate with the distribution of C-types in z_1–z_2, the Ch-types generally have lower scores in z_4 than the C-types, corresponding to linear rather than concave spectra. In subpanels g and i of Fig. 12, we highlight asteroids (41) Daphne, (49) Pales, (121) Hermione (144) Vibilia, and (159) Aemilia, all of which are compatible with CM chondrite spectra following the interpretation in Vernazza et al. (2015). (130) Elektra is also linked to these objects based on the density measurements (Carry 2012; Hanuš et al. 2017; Yang et al. 2016).

The 0.7 µm h-feature has been observed in at least one observation of 107 asteroids (5.0%). Members of the Ch-class are found in clusters 2, 5, 17, 19, and 26. The assignment requires the identification of the 0.7 µm h-feature. Within the C-complex only, 20.4% of samples present the h-feature. The actual number is likely higher as 12.1% of samples in the C-complex are missing the visible wavelength range, for example a NIR-only spectrum of (41) Daphne indicated as C-type in subpanel g of Fig. 12.

4.1.4. P-types

The P-types have been absent from the taxonomic schemes since Bus & Binzel (2002a), and thus no definition of the VisNIR behaviour exists. As part of the X-complex, the 'pseudo-M' types (Gradie & Tedesco 1982) are spectrally degenerate to the E- and M-types in the visible wavelength range, specifically, the ECAS colours. In the NIR, P-types show a red linear slope (see Fig. 11). We find that the spectral degeneracy between P and M continues in the NIR, while E-types differentiate by showing overall neutral slopes. Classes P and M have to be distinguished by visual albedo observations, which is about 5% for P-types.

As the X-complex is dissolved in this taxonomy, we assign the class to the C-complex following the proximity to the other classes in Fig. 12. This assignment is also in line with the IDP interpretation (Marsset et al. 2016; Vernazza et al. 2015). In z_1–z_4 space we observe a high-density cluster of P-types immediately adjacent to C-types. These samples are spectrally similar to the Cb-class in the Bus-DeMeo system. Furthermore, there is a more diffuse population of P-types building a bridge between the C-complex and the D-class.

The P-class is part of the former X-complex and of the new C-complex. Observations assigned to the P-class are thus inspected for all three features. While 19.2% of samples in the P-class present the h-feature, we note that no sample carries the k-feature, which is most prominent in the M-class. Three samples assigned to P show the e-feature, yet they belong to asteroids which are later assigned to the M-class: (4660) Nereus and (5645) 1990 SP. The k-feature may thus be a reliable differentiator between the spectrally degenerate M and P. The distribution of these features is discussed further in Sect. 4.3.

A total of 135 asteroids (6.4%) are classified as P-types in this study. Class P is built primarily from clusters 17, 19, and 22, where cluster 19 entails the continuous transition to class C and the first and third M-types. As mentioned above, we used the prototypes (65) Cybele and (87) Sylvia to differentiate between the classes, though assigning both to the C-types would have been justified as well given the cluster trend depicted in Fig. 12 (see subpanel c).

4.2. Endmembers: D, Z

We refer to D and Z as endmembers, due to the visible gap between their members and the C-complex in the latent space in Fig. 12; however, some of the P-types form a bridge population between the two classes.

4.2.1. D-types

The defining property of dark D-type asteroids is their featureless and strongly red-sloped spectrum both in the visible and in the NIR (DeMeo et al. 2009; Tholen 1984). They are predominantly found beyond the outer main belt, especially among the Jupiter trojan population, where they dominate the region in terms of mass (DeMeo & Carry 2013, 2014).

D-types form a homogeneous population in spectral and in albedo space, as shown in Fig. 11. This homogeneity is mirrored in the latent scores z_1 and z_4 as well (see Fig. 12), where in subpanel d we show the positions of (911) Agamemnon and (1143) Odysseus. In the second latent score, the D-types appear to split into a blue and a red population. We attribute this again to the normalisation of the spectra, which can cause these spurious offsets. Comparing the samples in the clusters showed no significant difference in the observables, and (2246) Bowell and (2674) Pandarus are present in the two clusters. Nevertheless, this serves as an example that the normalisation algorithm we devised for the partial observations may require further improvement. Furthermore, all clusters in latent space have to be verified by comparing the members in the observed features.

A total of 82 asteroids (3.9%) are classified as D-types in this study. D-types appear predominantly in two clusters, the homogeneous main cluster 1 and a more diffuse cluster 34, which may contain interlopers of classes P and M. Furthermore, there are two small clusters containing both D- and S-types. Cluster 8 has 16 VisNIR spectra of D-types and strongly-sloped S-types, which are separated using a two-component GMM in z_2-z_4, where the feature-rich S-types have higher scores in z_2. Cluster 43 contains 14 spectra, which are mainly visible-only S-types but include five visible-only D-types, which we separate in the same way as in cluster 8.

4.2.2. Z-types

The clustering revealed a low-number diffuse cluster of featureless extremely red objects, showing larger slopes than the D-types. Figure 12 shows that in z_1 these objects form a continuum with the D-types; however, the classes show different

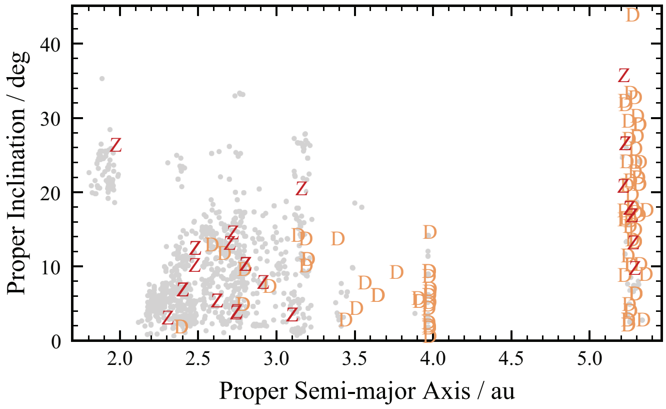


Fig. 13. Orbital distribution of `D`- and `Z`-types given by the respective class letters. The grey dots show the orbital elements of all other asteroids in the input data.

variances in the z_1–z_4 space: unlike `D`-types, `Z`-types show a clear trend towards a more convex shape with increasing slope. In addition, the classes show distinct orbital distributions, as illustrated in Fig. 13. While `D`-types are mostly situated among the Jupiter trojan population and the Hildas, these extremely red objects are largely scattered over the main belt. Three members of this population, (3283) Skorina, (15112) Arlenewolfe, and (17906) 1999 FG32, have previously been recognised in SDSS observations (e.g. Carvano et al. 2010) and described in a follow-up study by DeMeo et al. (2014), who further identified (908) Buda as a similar object.

The distinct distributions in latent and in orbital space prompt us to define a new class for this group of minor bodies. We propose the letter `Z`, which had previously been suggested by Mueller et al. (1992) for the extremely red Centaur (5145) Pholus. The 23 asteroids in the `Z`-class show overall low albedos, though we note the presence of outliers in Fig. 11.

The two reddest objects in this new class, (203) Pompeja and (269) Justitia, have been proposed as implanted trans-Neptunian objects by Hasegawa et al. (2021b). The authors suggest that complex organic material on the surface of these objects leads to the extremely red appearance. The prevalence of `Z`-types in the inner and middle main belt orbits of the objects could also indicate that a surface process such as spectral weathering is responsible.

A total of 23 asteroids (1.1%) are classified as `Z`-types in this study. They fall exclusively into cluster 36. Even so, we expect a certain number of `D`-type interlopers in this class as we observe an overlap in the latent space (see Fig. 12) and in subpanel (j), where we have highlighted the Trojan asteroids (1172) Aneas, (1542) Schalen, (4035) Thestor, and (7641) Cteatus, which spectrally match `D`-types.

4.3. `M`-complex: `K`, `L`, `M`

The `M`-complex comprises classes that fall in terms of spectra and albedo between the `C`- and the `S`-complex. Compositionally, it is the most diverse complex. For `C` and `S` the ensemble properties can be regarded as carbonaceous, primitive for the former and silicaceous, in part thermally metamorphosed for the latter (Cloutis et al. 1990a, 2011; Vernazza et al. 2014), while the likely mineralogical properties of any `M`-complex member cannot be given based solely on its complex membership. Meteorite analogues range from most carbonaceous chondrite clans in the meteorite collection to stony-iron and iron meteorites (Clark et al. 2009; Ockert-Bell et al. 2010; Sunshine et al. 2008; Eschrig et al. 2021; Shepard et al. 2010). Indeed, the only unifying property of these objects appears to be the spectral appearance with absent or generally faint features around 0.9 μm or 1.9 μm and an albedo around 15% with the exception of the endmember class `E`.

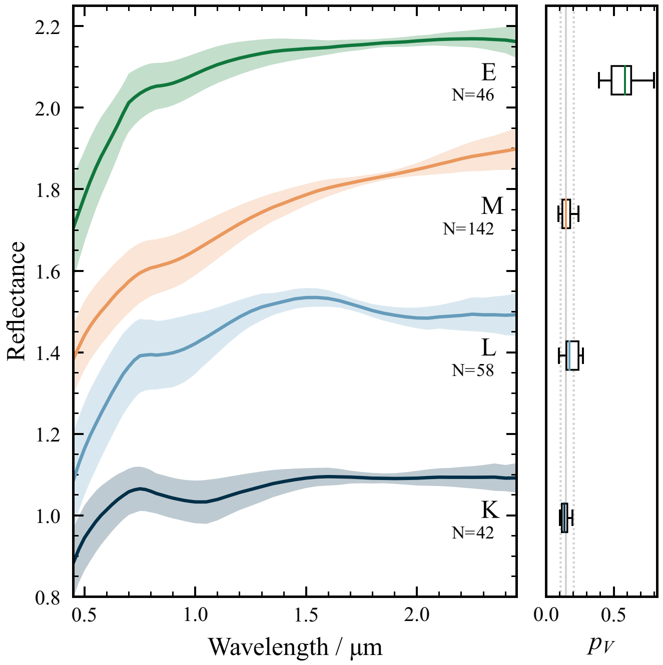


Fig. 14. As in Fig. 11, but for the data space properties of the `M`-complex. The `E`-class was excluded in the computation of the albedo distribution of the complex, indicated by the dotted linestyle of the upper and lower standard deviation. The albedo distribution of the `M`-complex excluding the `E`-types is $0.15^{+0.06}_{-0.04}$.

Devising the cluster-to-class decision tree proved challenging in this complex. In combination with the faint features, we observe slight variations in the slope in the NIR, and class degeneracies appear when the visible information is missing. Furthermore, we cannot rely as much on previously established terminology as this is a new complex in terms of taxonomic systems, replacing the `X`-complex as a third complex in previous taxonomic systems. Both the `K`- and the `L`-types are more recent than the Tholen (1984) taxonomy, which introduced the `X`-complex (Bell 1988; Bus & Binzel 2002a). The Bus-DeMeo system captures the diversity in the NIR in part in the form of the `X`- and `Xk` classes; however, no clear separation between the `X`- and the `C`-complex is achieved due to the lack of albedo information.

We split the complex into the three classes `K`, `L`, and `M`, shown in Fig. 14. Class `M`, in particular, contains a wide distribution of spectral appearances and likely mineralogical compositions. Nevertheless, we opt against a division of this class as no clear separation presents itself in this study, and we advocate for a division based on observables not included in this taxonomy. The `T`-class, which was tentatively introduced by Tholen (1984) and carried over in the Bus-DeMeo taxonomy, is dropped as prototypes (114) Kassandra and (308) Polyxo are well described by classes `P` and `M`.

4.3.1. `K`-types

Members of the `K`-class exhibit a red slope in the visible region with a 1 μm band associated with forsteritic olivine (Mothé-Diniz et al. 2008) and a neutral slope in the NIR. They have

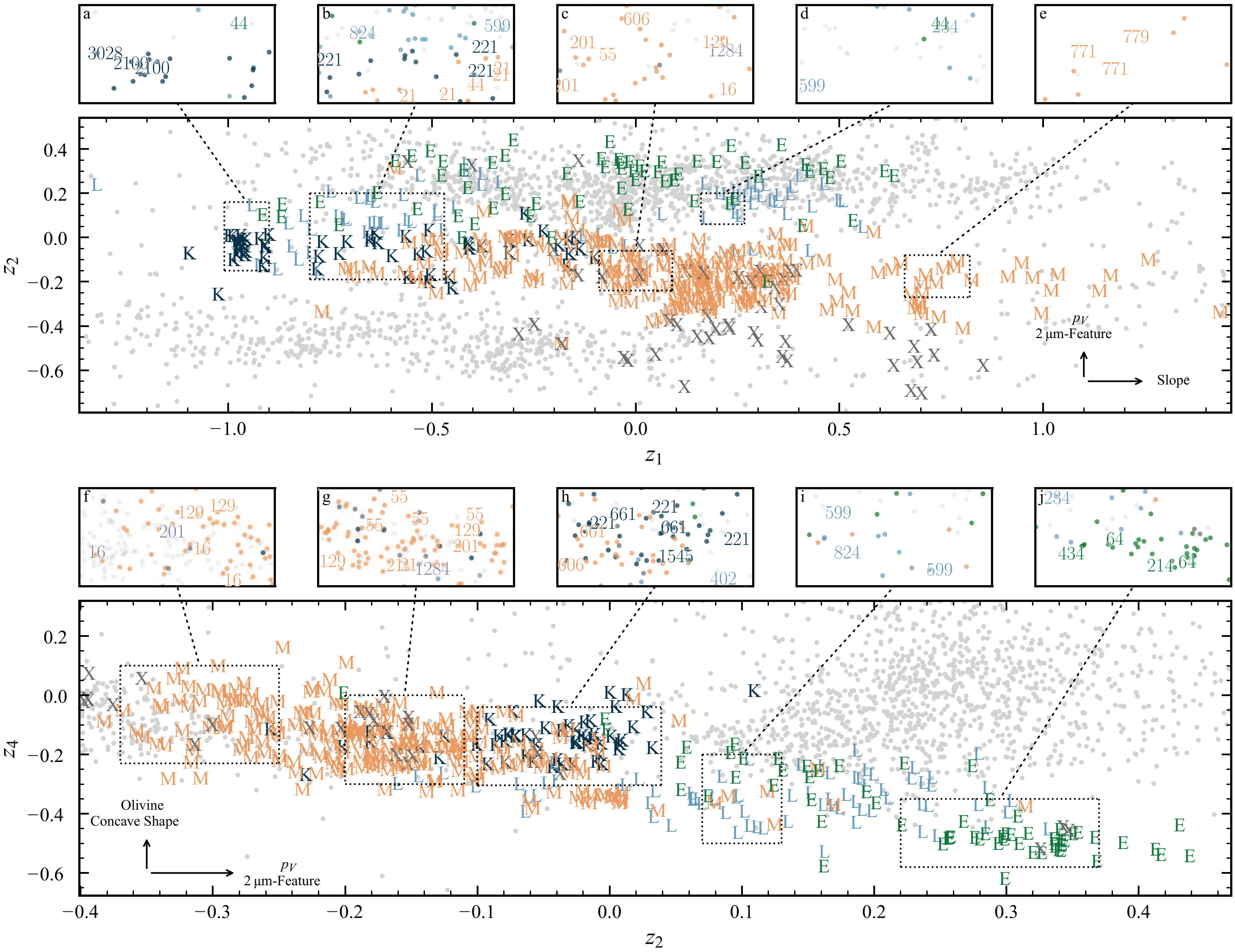


Fig. 15. As in Fig. 12, but for the member classes of the M-complex and its endmember class, the E-types.

low z_1 and high z_4 scores in comparison with the complex companion classes (see Figs. 14 and 15). Most K-types have visual albedos in the range 10%–15%, a narrow distribution which is comparable to the M-types and slightly lower than the L-types.

Dynamically, most main belt K-types are associated with the Eos family and depict on average a deeper 1 µm band than K-types outside the family based on the z_4 score (Clark et al. 2009), compare for example (402) Chloe and (1545) Thernoe to (221) Eos and (661) Cloelia in subpanel h in Fig. 15.

The class-averaged slope is neutral to slightly red in the NIR. However, some members, including the class archetype (221) Eos and (3028) Zhangguoxi, have a blue NIR slope, indicated by their low z_1 scores in subpanels a and b of Fig. 15. As the NIR spectrum is featureless above ~1 µm, this leads to a spectral degeneracy with the B-types, and the brighter part of the B-population requires the visible wavelength range information to be separated from the K-class. In subpanel a of Fig. 15, we see that (2100) Ra-Shalom is classified as a K-type, based on two NIR spectra. The only VisNIR sample of (2100) Ra-Shalom in this study is classified as a B-type. We note that (2100) Ra-Shalom is classified both as B and as K in the literature, based on its VisNIR spectrum (B: Binzel et al. 2019; de León et al. 2012 and K: Shepard et al. 2008a). The same degeneracy has been reported for B- and K-types in NIR spectra (Clark et al. 2009) and in the colour-space of the VISTA MOVIS survey (Popescu et al. 2018).

The distribution in z_1–z_2 shows a considerable overlap between M and K, with a slight gap between the populations around $z_1 = 0.3$. We considered whether the redder K-types may be Mk instead; however, among them are Eos family members such as (579) Sidonia and (653) Berenike, and thus we consider this slope variability to indicate K-types. The overlap is further resolved in z_3-z_4, where the K-class forms a denser population than the sparsely distributed Mk-types (not shown).

A total of 42 asteroids (2.0%) are classified as K-types in this study. K-types are found in two clusters, neither of which they populate entirely on their own. Cluster 24 is shared with M-types with neutral NIR slopes, while cluster 31 contains NIR-only observations of B-types as well as L-types. We resolve cluster 24 into K- and M-types using a two-component GMM fit to the cluster distribution in z_2-z_3, where K-types separate due to the large 1 µm band. Cluster 31 is only split into K and L members as the degeneracy with NIR observations of B cannot be resolved with the observables in this taxonomy. The cluster members are assigned based on their probability of belonging to cluster 23 (L) or 24 (K) in z_2-z_3.

4.3.2. L-types

L-type asteroids are associated with large abundances of spinel-bearing calcium–aluminium-rich inclusions due to a wide absorption feature around 2 µm (Sunshine et al. 2008). This composition would imply that the L-type parent bodies were among the first planetesimals to form in the accretion disk, making them of high interest for formation scenario studies (Devogèle et al. 2018). However, in addition to the 2 µm feature, L-types are spectrally heterogeneous in slope and shape of the visible and 1 µm region and in their albedo distribution, shown in Fig. 14. The diversity of L-types makes it difficult to reliably identify them in a taxonomy based on spectral features and opens up degeneracies with a handful of neighbouring classes, such as K, M, and S.

We find that many previously classified L-types cluster in dimensions z_2-z_4, where they branch off of the M-complex below the S-complex together with the E-types (see Fig. 15). The second latent component matches the spinel-associated 2 µm band best, giving L-types higher z_2 scores compared to the other classes in the complex, while compared to the S-types the 0.9 µm contribution to the z_4 score is missing.

In Fig. 15, we see that the L-types identified in z_2-z_4 exhibit a bimodality in terms of their slope in z_1, further shown in the spectral domain in Fig. C.1. This dichotomy is not caused by the normalisation of the spectra. Instead, we find that previously classified L-types with intermediate slope such as (606) Brangane are classified either as M or S as they lack the 2.0 µm feature (see subpanels c and h in Fig. 15). We regard the slope variability of the L-types classified here as intrinsic to the class, supported by (599) Luisa, which has both a blue and a red spectrum (see subpanels b, d, and i in Fig. 15).

Of particular interest among the L-types are the subgroup members referred to as Barbarians after (234) Barbara, which show anomalously high inversion angles in their negative polarisation branch (Cellino et al. 2014; Devogèle et al. 2018). We find that this group of asteroids has a large variance in latent space. In z_1–z_2, Barbarians such as (234) Barbara, (824) Anastasia, (599) Luisa, and (606) Brangane and (1284) Latvia (which are classified as M) are found in both the M- and S-complexes and at the transition region (see subpanels b–d and g–j) in Fig. 15). We also do not find a reliable clustering in the remaining latent scores. The spectral L-types do not include all Barbarians, among which we observe a diversity that is too large to derive a unique class in this taxonomy. Of the 16 Barbarians from Devogèle et al. (2018), 7 are L-types and 5 are M-types. An extension of the taxonomy observables with polarimetric observations is required to reliably identify Barbarians.

A total of 58 asteroids (2.7%) are classified as L-types in this study. L-types occur predominantly in clusters 4 and 23. As for the K-class, these two clusters are populated by members from other classes as well. For cluster 4, we split the L- and S-types based on a two-component GMM in z_3-z_4 trained on the distribution of the members of cluster 23 and cluster 40 in this space. For cluster 23, we split the L- and M-types based on a two-component GMM in z_1–z_4. A small fraction of L-types are also in cluster 37, which consists largely of M-types. The L-types are recovered using a two-component GMM in z_2-z_4.

4.3.3. M-types

The M-class is one of the oldest asteroid designations (Zellner & Gradie 1976). Originally introduced to describe asteroids representing presumably metallic cores of disrupted planetesimals (Bell et al. 1989; Gaffey & McCord 1979), dedicated observational efforts have revealed a variety of objects based on their densities (Carry 2012; Vernazza et al. 2021), hydration (Rivkin 1995, 2000), radar albedos (Shepard et al. 2010, 2015), and silicate spectral features (Clark et al. 2004; Fornasier et al. 2010; Neeley et al. 2014; Ockert-Bell et al. 2010).

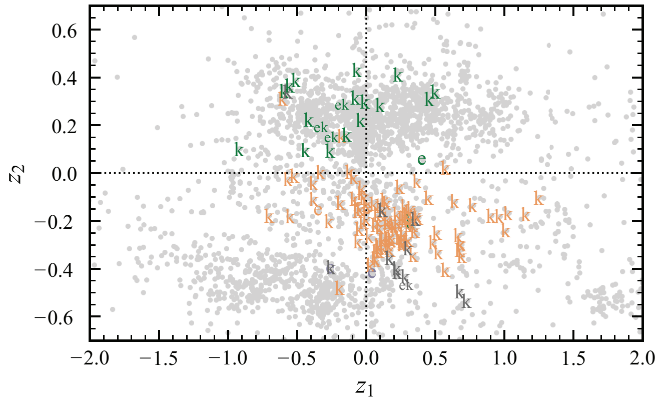


Fig. 16. Distribution of observations which carry the e- and k-feature in the first two latent scores, colour-coded by the class they are assigned to: green – E, orange – M, purple – P, grey – X. A smaller font size is used if the observation carries both e and k.

In spectral space, M-types asteroids are red with either linear or convex shapes, as shown in Fig. 14. The convex trend may even result in an overall blue slope longwards of 1.5 µm, as is the case for (21) Lutetia in four out of five observations in this study. M-types in the lower z_1 region around (21) Lutetia, highlighted in subpanels b and g in Fig. 15, closely resemble the Xc-class in the Bus-DeMeo system. At the other end of the class in z_1, asteroids like (771) Libera and (779) Nina are examples of red, linear slopes in the NIR, shown in subpanel e in Fig. 15. M-types have an albedo distribution of 10%–20%. We note that (55) Pandora has an albedo of 0.34, and one of its samples is classified as E, visible in the upper part of Fig. 15, around $(z_1, z_2) = (0.3, -0.2)$.

Silicate features at 0.9 µm or 1.9 µm are likely more common than an entirely featureless spectrum among M-types, with 40.9% of M-type samples exhibiting the k-feature. Of the samples, 30.2% lack the corresponding wavelength region observed. In Fig. 16, we display the first two latent scores of samples with the e- and k-feature. The latter feature is ubiquitous among M-types, and a concentration in latent space around (16) Psyche is visible. (55) Pandora, (129) Antigone, and (201) Penelope further show the k-feature in one or several samples, and are highlighted in subpanels c, f, and g in Fig. 15. The bands are linked to different pyroxenes (Hardersen et al. 2005), and the presence of the 1.9 µm band is accompanied by the 0.9 µm band, but not vice versa (Shepard et al. 2015).

The distribution of M-types in latent space and the results acquired in the studies cited above suggest that there are at least two populations of M-types, the chondritic population, of which (21) Lutetia may be the archetype, and the metallic population, of which (16) Psyche is the prototype (Vernazza et al. 2011; Viikinkoski et al. 2017). We see this as a reasonable division of the M-class to further dissolve the compositional degeneracy of the X-complex. However, this division cannot be done based on spectra alone. To not increase the entropy of the taxonomy in a false direction, we refrain here from dividing the M-class.

A total of 142 asteroids (6.7%) are classified as M-types in this study. The main clusters containing M-types are clusters 22, 37, and 46. Smaller contributors are clusters 17 and 35. All these clusters make up the X-complex in this taxonomy, and the spectra are split into E, M, and P as described in Sect. 3.4.2. Additional members of the M-class are found in clusters 23 and 24, which are spectrally close to L- and K-types.

4.4. Endmembers: E-types

E-type asteroids are linked to the enstatite achondrites (Gaffey et al. 1992). Their standout feature is a visual albedo generally above 50%, see Fig. 14. This unique property makes them easy to recognise in the reduced latent space, where they exhibit large absolute values in z_2 and z_3, with the former shown in Fig. 15.

Spectrally, E-types have a steep visible slope before flattening out in the NIR. In the case where the albedo observation is missing, E-types are degenerate with all classes of the M-complex. As an example, we observe samples of (44) Nysa located in subpanels a and d, around the K- and the L-types, correctly identified as an E-type, due to the albedo observation. However, the third sample in subpanel b lacks an associated albedo value and is classified as an M-type. As for L and M, we find a large intrinsic variability of the samples of individual asteroids in the E-class.

Most E-types in Tholen (1984) are classified as Xe in the Bus-DeMeo system due to the presence of the e-feature at 0.5 µm. In Fig. 16, we see that the e-feature is overall sparse compared to the k-feature. Thirteen samples in the M-complex exhibit the feature, while 65.4% of samples lack the corresponding wavelength region observed. Of the 13 samples, 4 are classified as E-type. Considering the relative sizes of the M- and E-class, the latter are hence more likely to exhibit the feature. We do not observe a clustering of the e-feature.

The bias towards E over M for e-feature presence may be of observational nature. As an abundance of metal on the surface of M-types may lead to a drop-off of the spectral reflectance in the UV, the 0.5 µm feature might not be observed as the reflectance does not increase again towards smaller wavelengths. The band is associated with the sulfide mineral oldhamite present in aubrites (Watters & Prinz 1979) or to titanium-bearing pyroxene (Shestopalov et al. 2010). The prototype for this feature is (64) Angelina, while the E-class archetype (434) Hungaria does not present it. The k-feature is present in 30.8% of E-type samples, while 36.9% of samples lack the corresponding wavelength region observed.

(214) Aschera highlights the benefit of resurrecting the visual albedo. Since its classification as E-type in Tholen & Barucci (1989), it has been classified as X, B, Cgh, and C in different works (de León et al. 2012; DeMeo et al. 2009; Lazzaro et al. 2004). With a visual albedo above 50%, (214) Aschera is here classified as Ek-type and concludes its spin through the proverbial ‘alphabet soup’. Observations of (64) Angelina, (214) Aschera, and (434) Hungaria are highlighted in subpanel j of Fig. 15.

A total of 46 asteroids (2.2%) are classified as E-types in this study. They are predominantly located in cluster 35, though other clusters of the M-complex may also contain single samples of E-types. These are identified and assigned to the E-class in a late branch of the decision tree using the albedo distributions of E, M, and P given in Fig. 9. E-types also appear in cluster 44 among the S-types, where they are identified based on a two-component GMM fitted to the albedo distribution of the cluster.

4.5. S-complex: S, Q

The S-complex is by far the largest complex in terms of individual asteroids, in this work and in previous taxonomies. This can be attributed to observational biases such as the numeric dominance of the S-types in the inner main belt and near-Earth space (Binzel et al. 2019; DeMeo & Carry 2013, 2014) and the high average albedo of more than 20%.

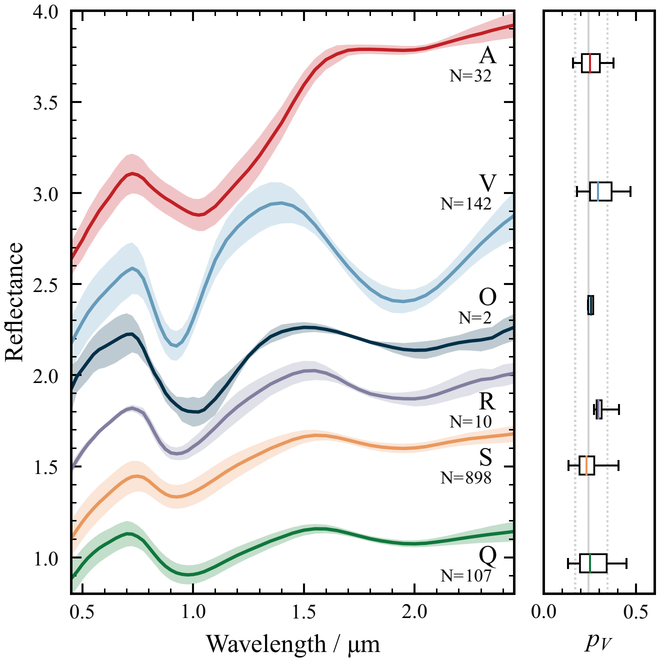


Fig. 17. As in Fig. 11, but for the data space properties of the S-complex. The albedo distribution of the S-complex is $0.24^{+0.10}_{-0.07}$.

The abundance of S-types makes their homogeneity both in spectra and albedos as shown in Fig. 17 even more remarkable. While trends in the slope and the silicate features at 0.9 µm, 1.0 µm, and 1.9 µm are observable, these are primarily continuous trends and well explained by variations in the mineral composition, in particular olivine and pyroxene, as well as trends of thermal alteration in ordinary chondrites (Eschrig et al. 2022; Vernazza et al. 2014),). S-types are one of two classes of asteroids that have an established meteorite analogue; they were linked to ordinary chondrites by the JAXA Hayabusa mission (Nakamura et al. 2011). This linkage in combination with the wealth of data on ordinary chondrites and S-types gives a solid understanding of the spectral weathering processes occurring on the surfaces of the minor bodies (Brunetto & Strazzulla 2005; Chrbolková et al. 2021; Thomas et al. 2012), which, unlike the C-complex members, shows a universal trend of surface darkening and spectral reddening with the surface age.

We divide the S-complex into two classes: S and Q. Including the endmember classes A, R, and V, we establish all classes defined in the Tholen (1984) system while extending it with the O-class. Compared to the Bus-DeMeo system, we reduce the taxonomy by subclasses of the S-class, as we explain in the following.

4.5.1. S-types

While class C has been split into subclasses since early taxonomic efforts (Gradie & Tedesco 1982; Tholen 1984), the S-class was not divided until Bus & Binzel (2002a) as the silicaceous surfaces are particularly subject to changes in slope and band structure induced by phase-angle effects (Sanchez et al. 2012) and space weathering (Strazzulla et al. 2005).

The Bus-DeMeo system accounts for these effects by subtracting the spectral slope before classification; however, as

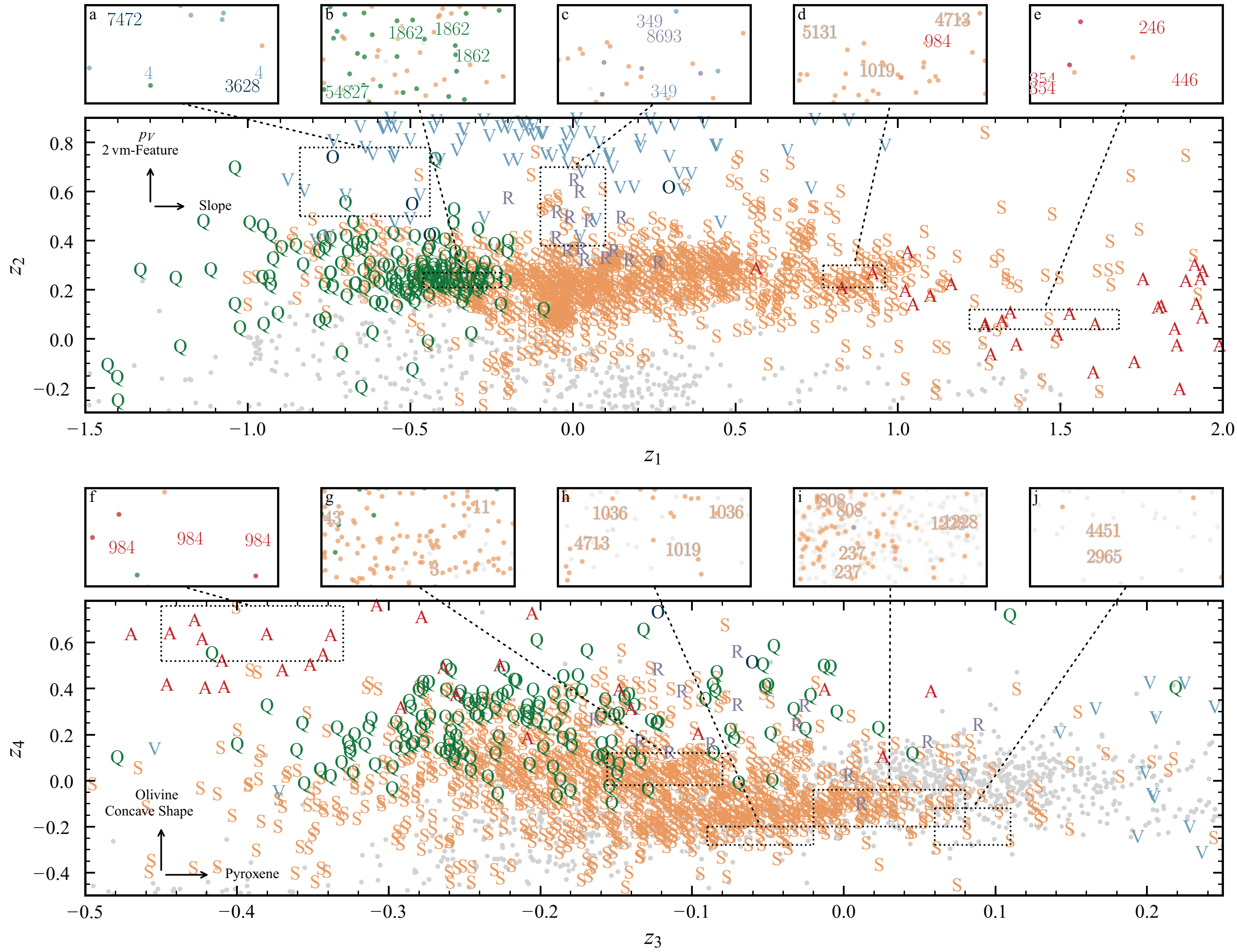


Fig. 18. As in Fig. 12, but for the member classes of the S-complex. For increased resolution of the S-class, the A- and V-class are only shown partially.

outlined in previous sections, the partial observations prevent us from applying this taxonomy. Instead, we rely on the interpretation of the latent components to serve as vectors in the compositional analysis of the S-types.

The second and third latent components both resemble pyroxene as this mineral dominates the S-class, in addition to the large contribution in terms of variance provided by the V-types. The first component resembles the slope, hence we can approximate the vector of space weathering within the S-complex with it (e.g. Brunetto et al. 2006). S-types denoted with the w-suffix for weathered in the Bus-DeMeo system exhibit higher z_1 scores than their class siblings with fresh surfaces. The degeneracy between a weathered S-type and an olivine-rich S-type (Sa in the Bus-DeMeo system), which is redder by mineralogy rather than by surface alteration, is resolved in the third and fourth latent component, which separates the pyroxene-olivine composition of objects.

As a practical example, in subpanel d in Fig. 18 we show the Bus-DeMeo Sa-types (984) Gretia and (5131) 1990 BG and the Sw-types (1019) Strackea and (4713) Steel. The subpanel h shows that both Sw-types have below average olivine components, indicating that the red surface is indeed due to weathering; also shown in this subpanel is the S-type (1036) Ganymed. (984) Gretia is classified as A-type in this study due to its high z_4 score (see subpanel f in Fig. 18).

The Bus-DeMeo system further recognises Sq-, Sr-, and Sv-types in addition to the regular S-types. The prototypes given in DeMeo et al. (2009) for these subclasses are highlighted respectively in subpanels g ((3) Juno, (11) Parthenope, (43) Ariadne), i ((237) Coelestina, (808) Merxia, (1228) Scabiosa), and j ((2965) Surikov, (4451) Grieve) in Fig. 18. The continuous distribution between the main S-complex and the subclasses confirms our decision to not subdivide the S-class.

A total of 898 asteroids (42.3%) are classified as S-types in this study. The class is made up of several clusters: 0, 3, 6, 11, 14, 20, 21, 30, 33, 38, 39, 40, 42, and 47. Clusters 4, 8, 10, 43, and 44 contain members from other classes, which we divide via GMMs, as described in the respective class descriptions (L, D, R, D, and E, in order of the clusters).

4.5.2. Q-types

Q-type asteroids are mostly found in the near-Earth asteroid population and resemble spectrally the ordinary chondrites in the meteorite collection (Binzel et al. 2004c). Compared to S-types, Q-types have a wider 1 μm band and a neutral to blue slope over

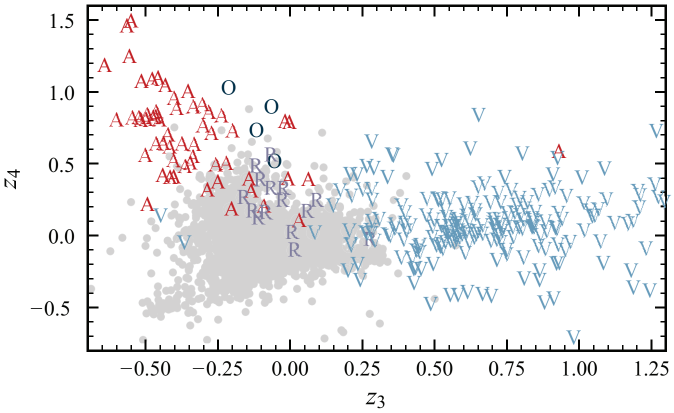


Fig. 19. Distribution of the S-complex endmember classes A, O, R, and V in the last two components of the latent space.

the whole spectral range (see Fig. 17). The albedo distribution is more extended towards higher albedo values, with values of 20%–35%, in agreement with space weathering models predicting darkening of silicaceous asteroids with increasing surface age (Brunetto et al. 2006).

In latent space, Q-types occupy the blue end of the S-complex in z_1. They are also distinguished from the less weathered S-types in the z_2-z_3 space based on their high z_3 scores due to the wide 1 µm band. The archetype (1862) Apollo and class member (54827) Kurpfalz are highlighted in subpanel b in Fig. 18.

A total of 107 asteroids (5.0%) are classified as Q-types in this study, 83.2% of which are near-Earth asteroids, which is considerably higher than the average of 34.4% over all asteroids in the input data. They populate clusters 16 and 48, as well as the diffuse cluster 13, further outlined in Sect. 4.6.3. We considered merging the Q-class into the S-class as it represents the overall continuity in the S-complex. However, as for the Z-class, the orbital distribution of the Q-types convinced us to keep this class.

4.6. Endmembers: A, O, R, V

The endmembers of the S-complex are the well-established classes A and V and the two classes that were initially built around single objects, O and R. Their distribution in the third and fourth latent scores is given in Fig. 19.

4.6.1. A-types

A-type asteroids are differentiated asteroids linked to brachinite achondrites (Burbine et al. 2002; Cruikshank & Hartmann 1984; DeMeo et al. 2019) and are easily recognised in spectral space by their strong red slope and deep olivine imprint at 1 µm (see Fig. 17). The albedo is within the complex average of about 20%–30%.

In latent space, the red colour of A-types leads to a high score in z_1, forming a diffuse branch off the S-type population. We highlight the prototypes (246) Asporina, (354) Eleonora, and (446) Aeternitas in subpanel e in Fig. 18. Further characteristic of A-types is a high z_4 score due to the high olivine content (see Fig. 19). Of all the classified asteroids, A-types have the highest z_1 and z_4 scores. We note that all three spectra of Mars-Crosser (1951) Lick are exceptionally red, even among A-types (Brunetto et al. 2007).

A total of 32 asteroids (1.5%) are classified as A-types in this study. They fall into clusters 9, 12, 27, and 49.

4.6.2. O-types

The class O was introduced in 1993 for supposedly ordinary-chondritic (3628) Boznemcova (Binzel et al. 1993). Its noteworthy characteristic is the wide round 1 µm feature as shown in Fig. 17, placing it between the known A-, Q-, and V-types. The albedo is close to the S-complex average at 25%.

None of the previously classified O-types, except for archetype (3628) Boznemcova remains as an O, and a comparison of these objects in spectral space showed little resemblance. While we assign with (7472) Kumakiri a second asteroid to the class, we find in this work that (3628) Boznemcova remains without a true spectral sibling. (7472) Kumakiri was previously classified as V (Solontoi et al. 2012); however, its spectral resemblance to (3628) Boznemcova has been pointed out by Burbine et al. (2011).

The unique appearance of the O-types can be seen by their position in the latent space shown in Figs. 18 and 19. The depth and shape of the 1 µm band in combination with the lack of overall slope place the O-types (3628) Boznemcova and (7472) Kumakiri between the classes Q and V in z_1–z_2 (see subpanel a in Fig. 18), while in z_3-z_4, they are closest to A-types.

Two asteroids (0.1%) are classified as O-types in this study. We debated whether keeping the O-class in the taxonomy is compatible with the overall approach of data-driven clustering. In the end, the unique feature and position of (3628) Boznemcova convinced us, although an argument against single-object classes can be made. The O-class was difficult to carve out from the clusters using the given method. It is derived from a three-component mixture model of the already diffuse cluster 13, which is split into C, O, and Q. Any assignment of the O-class by the classification tool should undergo visual scrutiny and direct comparison to the spectrum of (3628) Boznemcova.

4.6.3. R-types

The R-types are the second niche class of this taxonomy, built around (349) Dembowska. The unique nature of (349) Dembowska is recognised jointly with that of (4) Vesta in early works of taxonomy (Chapman et al. 1975; Zellner & Gradie 1976) and the R-class was introduced in Bowell et al. (1978). However, the A-class, which was split off the R-class in Veeder et al. (1983), has since been absorbed into most R-types. The continuity between A and R is visible in Fig. 19.

R-types show 1 µm and 2 µm features which are deeper than those in S-types. The width of the 1 µm is between the V- and the Q-types. They have albedos at the upper end of the S-complex distribution, around 28% (see Fig. 17). The spectral appearance is associated with low-iron ordinary chondrites (Zellner & Gradie 1976). We note that of the four samples of (349) Dembowska two are classified as R and another two as V (see subpanel c in Fig. 18, where we also give the position of R-class member (8693) Matsuki).

A total of 10 asteroids (0.5%) are classified as R-types in this study. The class is derived from cluster 10 in a two-component GMM fit in z_1–z_2, where objects with lower z_2 scores are assigned to the S-class.

4.6.4. V-types

(4) Vesta was the first asteroid to be observed spectrophotometrically (McCord et al. 1970) and the V-types have been an established and easily-recognizable class in all asteroid taxonomies since Tholen (1984). They are the second class, in addition to S, with an established meteoritic analogue, the HED

meteorites (e.g. Kelley et al. 2003). The class makes no exception here; its members are differentiated easily in both z_2 and z_3 due to the large contribution of pyroxene to the spectral appearance giving rise to the characteristic deep 1 µm and 2 µm features (see Figs. 17 to 19). The class archetype (4) Vesta is highlighted in subpanel a in Fig. 18.

The large class variance in z_2 and z_3 represents high variability in terms of band depth and position in the 0.9 µm and 2.0 µm features. However, we do not identify a subpopulation based on the band parameters, as was suggested by Binzel & Xu (1993).

A total of 142 asteroids (6.7%) are classified as V-types in this study. V-types populate clusters 7, 15, 18, 28, 32, and 45. V-types with a blue slope in the NIR further share the diffuse cluster 41 with the B-types.

5. Classification

In this section, we introduce the classification tool described in this work. We demonstrate the probabilistic classification results using asteroid observations with different wavelength regions covered. We further investigate degeneracies in the classification space. Finally, we compare the results obtained in this taxonomy to the previous systems.

5.1. Classification tool: Classy

To facilitate the classification of asteroid observations within the framework of this taxonomy, we provide the CLAssification of a Solar System bodY (classy[13]) tool written in Python. It is able to interactively smooth the input spectral observations prior to resampling them to the required wavelength grid, to automatically apply the necessary pre-processing steps outlined in Sect. 2 to both spectra and albedo, to identify features in the spectra as outlined in Sect. 3.4.3 (either fully automated or guided by the user), to execute the cluster-to-class decision tree, and to return the probabilistic classifications for each observation.

The classy tool provides a command-line interface written in Python and is available for Windows, MacOS, and Linux. The software is actively maintained and developed by the authors.

5.2. Class degeneracies

The probabilistic nature of the classifications in this taxonomy allow the degeneracies between classes to be quantified in certain wavelength regions and in albedo. One example is given in Sect. 4.1.1, where we point out the degeneracy of B and K in the case of a NIR-only observation.

We can quantify class degeneracies for three datasets in this work, with the aim of reflecting the most commonly available observation ranges of asteroid spectra: the 2983 spectra used to devise the clustering, the 2923 visible-only spectra shown in grey in Fig. 2 with 81.4% albedos observed, and the 2813 spectra from the clustering sample which have NIR information. For the last, we remove all observations of wavelengths below 0.8 µm and the albedo information present in the samples. We refer to these samples as the complete, the visible-only, and the NIR-only datasets; however, this wording is not entirely accurate as more than 50% of the samples in the complete sample are NIR-only spectra and the visible-only sample contains more than 80% of albedo observations.

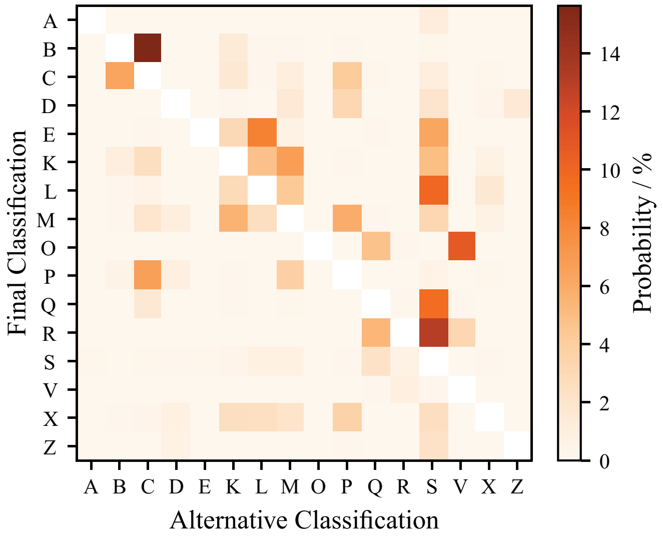


Fig. 20. Confusion matrix between the classes defined in this taxonomy in the visible-near-infrared and albedo input space. For each class in the taxonomic scheme, we give the average probability of its samples to be classified as any other class based on the complete dataset. The Ch-class is missing as it relies on the detection of the 0.7 µm h-feature and does not have an associated class probability. For better readability, the main matrix diagonal corresponding to the equal-class cases is left empty. These values are generally above 80% and lowest for K, L, M, and R.

5.2.1. Complete sample

To estimate the class degeneracy in the complete dataset we compute the average probability of belonging to any other class for all samples assigned to a given class. This comparison is given in Fig. 20. The Ch-class is missing as it relies on the detection of the h-feature, and as such does not have an associated class probability. Larger matrix element values indicate a higher degeneracy between the classes. A large sum per matrix row indicates that the class assignment is overall less certain.

Figure 20 shows the intuitive result that endmember classes such as A, V, and Z are assigned with a large probability. The largest degeneracies in pairs of classes are between B- and C-types and R- and S-types. Neither result is surprising as they overlap in latent space, and even in visual inspection these classes can be difficult to tell apart. The largest uncertainty overall for a single class (given by the sum per row in Fig. 20) is around 20 % for K, L, and M and also for the R-types. For the first three, we already pointed out in Sect. 4.3 the similarity in data space between these classes, hence this result is again expected.

5.2.2. Visible-only sample

The estimation of the class degeneracy is repeated for the visible-only dataset after classifying the samples therein using the classy tool. Figure 21 shows a result similar to that for the complete dataset, except that the overall values of uncertainty increase. Instead of 80%–99% certainty in the class assignment, we obtain values between 63%–91%. Except for this overall change in scale, we do not observe significant differences between the results for the visible-only and the complete datasets. K, L, and M are among the least-certain classes, while O-types have the largest uncertainty due to the missing 1 µm band. For classes from the M- and S-complex, we see an overall increasing probability to be classified as S-type.

[13] https://github.com/maxmahlke/classy

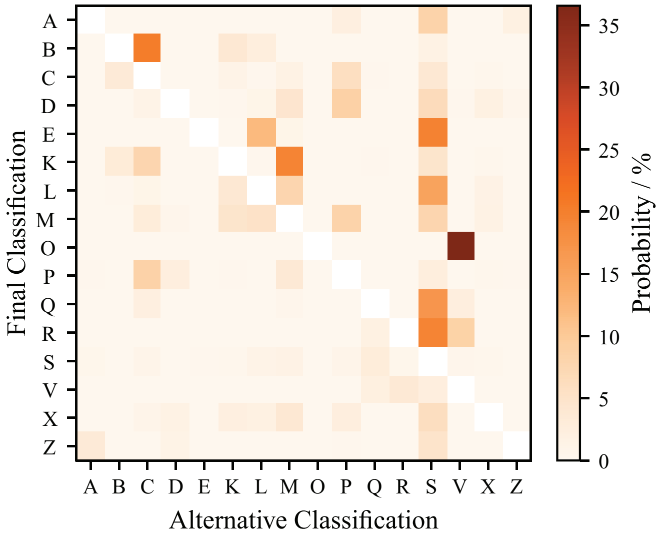


Fig. 21. As in Fig. 20, but using the dataset of 2923 visible-only spectra with 81.4% albedo observations. The colourbar scale is different to that in Fig. 20. The main matrix diagonal values are between 63%–91% and lowest for K, L, M, and O.

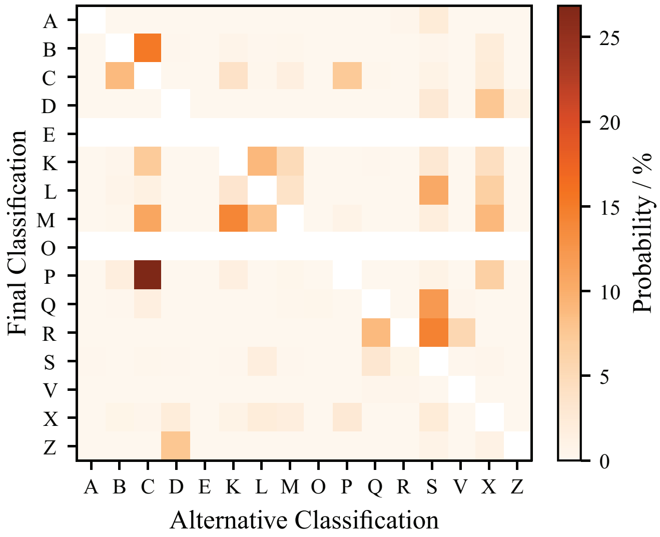


Fig. 22. As in Fig. 20, but using the dataset of 2813 NIR-only spectra without albedo information. The colourbar scale is different to that in Fig. 20. No observation in this sample is classified as E or O. The main matrix diagonal values are between 55%–99% and lowest for M and P.

5.2.3. NIR-only sample

The class degeneracy is next calculated for the NIR-only spectra that are part of the input observations used to train the MCFA model. We remove the albedo information present in 78.5% of the samples prior to classifying them. The confusion matrix is shown in Fig. 22. The overall scale of the uncertainty in the class assignment is between the results for the complete and the visible-only dataset, with the maximum average degeneracy between two classes just over 25% between P and C, likely due to both the missing albedo information and the truncation of the broad 1.3 µm feature in the C-types. We note that no sample is classified as E-type, due to the missing albedo information, and no sample is classified as O-type, as both (3628) Boznemcova and (7472) Kumakiri are classified as Q without the visible-wavelength information. The bowl-shaped 1 µm band of the O-types extends below the 0.8 µm limit we apply to this dataset, hence this misclassification is acceptable. The expected degeneracy between B and K in NIR-only data is not visible in Fig. 22 due to the presence of the 1 µm information. While visible-only spectra lead to uncertainty among the M- and S-complexes, in particular with respect to the S-types, this calculation shows that NIR-only spectra lead to greater confusion between the C- and the M-complexes.

We conclude that the class degeneracies in the complete, visible-only, and NIR-only samples follow an intuitive behaviour: the largest classes in terms of number of samples (S, C, and M) become more probable with decreasing observational data. This is in line with the established classification guideline that, when in doubt, assignment to small classes should only be done on the basis of convincing observational evidence.

5.2.4. Complete versus visible-only sample

Another way to investigate class degeneracies is the comparison of classifications resulting from samples with different wavelength regions observed. There are 267 asteroids present in both the complete and the visible-only datasets with a total of 328 observations. For these asteroids, we compare the resulting classifications based on the samples in both datasets, shown in Fig. 23. Each row gives for each class in the taxonomy the fraction of asteroids classified as any class based on the visible-only dataset. We note that the figure does not account for the different samples sizes: there are 2 samples classified as E in the intersection of the dataset and 140 classified as S. No samples classified as A, O, and X are present in both samples.

Figure 23 shows that Ch, S, and V are the most reliable when classified using visible-only data. Ch benefits from the binary classification which takes place once the h-feature is observed. The members of the M-complex show increasing degeneracy with the S-class with decreasing near-infrared coverage. The least-expected degeneracies are Z and C, as well as E and B, however, they are all based on a single sample.

Both results in Figs. 21 and 23 show that visible-only spectra in combination with the albedo place a strong constraint on the taxonomic class, as is well-established from the previous taxonomies which relied on the visible wavelength ranges exclusively. This highlights the strengths of the new method employed here: NIR-spectra are not strictly necessary to derive a classification as incomplete observations can be classified and the albedo as an accessible observable is accounted for.

We do not repeat this comparison for the complete and the NIR-only samples as the latter make up a significant fraction of the former, hence the agreement between the samples would be overestimated.

5.3. Comparison to previous taxonomies

Class continuity was one of the aspects which we considered when designing the scheme of classes in this taxonomy. We quantify this goal as above using a confusion matrix, except that we compare the classes assigned based on the complete dataset to the most-probable previous classification of the asteroid in the literature, retrieved for 2676 samples of 1852 individual asteroids from the SsODNet database. We convert the previous classifications done mostly in the Bus-DeMeo scheme to this scheme using the mapping given in Table 2.

Figure 24 shows an overall good agreement of the classes assigned in this work with the ones from the literature. Notable

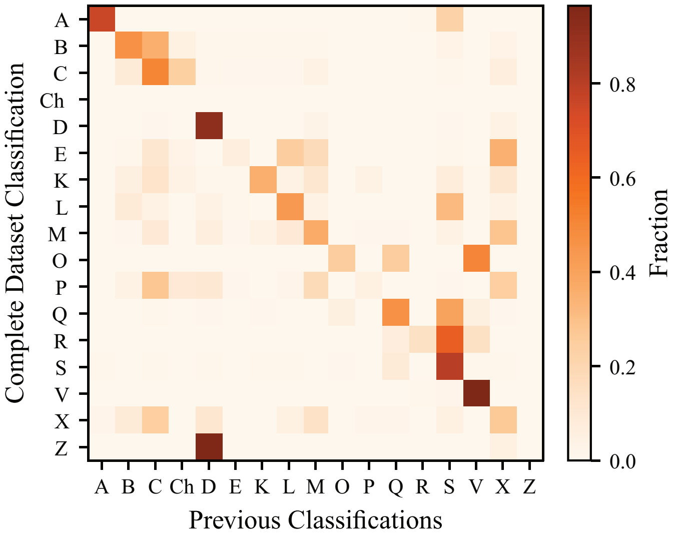


Fig. 23. Comparison of the classifications of 328 samples of 267 individual asteroids resulting from visible-only spectra with 81.4% observed albedos to the classifications of the same asteroids resulting from the complete sample classifications. The sample size iis different in each row: the intersection of asteroids present in both datasets gives 2 samples classified as E using complete samples as well as 2 samples classified as Z, while there are 140 samples entering the calculation in the row of S-types. No A-, O-, or X-types are present in both samples.

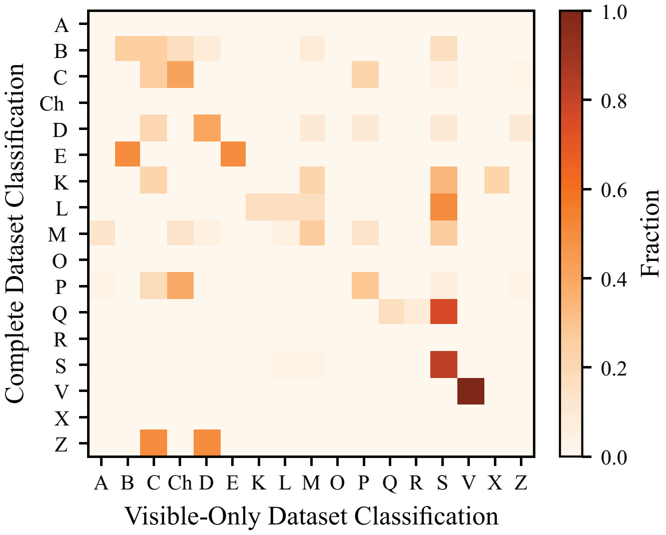


Fig. 24. Comparison of the classifications of 2676 samples of 1852 individual asteroids classified in the complete dataset to the classifications in the literature. The literature classifications were mapped into this taxonomy scheme following Table 2. The number of samples differs between the rows.

exceptions are the O-type, which has no legacy members apart from (3628) Boznemcova as pointed out in Sect. 4.6.2, and the new Z-class, which hosts almost exclusively previous D-types. Furthermore, the L loses members to the S as well as O to V.

6. Conclusion

The taxonomic scheme for minor body classification has been in development for close to 50 yr. During this time, numerous efforts to categorise the observational properties of asteroids have been driven forwards through dedicated observational campaigns and instrumental advancement. We focused on the methodology and statistical foundation, allowing us to increase the sample size by an order of magnitude compared to the previous taxonomy by DeMeo et al. (2009) and to reintroduce the albedo into the classifying observables as done in Tholen (1984).

The dimensionality reduction and clustering applied to 2983 spectra of 2125 asteroids revealed three main complexes: the well established C- and S-complexes and a restructured M-complex. While the S-complex is well understood in terms of mineralogy and meteoritic analogue material, both the C-complex and M-complex show a large degree of variability of so far unknown origin. We derive 17 classes from the three complexes, where the data-driven clustering is guided by the previous taxonomies and the goal of class continuity.

A classification tool named `classy` is available online and allows the user to classify asteroid observations covering the spectral VisNIR region and the visual albedo either completely or partially. The resulting array of class probabilities for each sample serves to estimate classification uncertainty and possible taxonomic trends.

We established a methodology for asteroid taxonomy which is well suited for the current and future datasets of asteroid observations. The ongoing MITHNEOS survey, the upcoming *Gaia* Data Release 3 (including visible spectra, Delbó et al. 2012), and the planned NEO Surveyor mission (Mainzer et al. 2015) and SPHEREx survey (Ivezić et al. 2022) will provide or continue to provide spectral and albedo observations of asteroids in different wavelengths, which are able to be classified within the framework of this taxonomy.

The dimensionality reduction and clustering are able to resolve more features and find more meaningful clusters when fed with more data. It may be worthwhile exploring how the model properties described in Sect. 3 change when fed with significantly more data. Nevertheless, during this work, we found that the latent space properties show little change whether we train with 500, 1000, or all samples in the dataset. Instead, we anticipate that a future taxonomy-revision will benefit more from an increased feature set. In particular the UV information offered by the *Gaia* data may solve degeneracies in the C- and M-complex. A further improvement should be the addition of polarimetric data, provided the amount of observations is comparable to the availability of the other features. The M-complex could benefit, and we consider that most work is left to be done in this complex. Extension of the spectral space into the 3 µm region is promising as well.

Acknowledgements. The authors thank the numerous members of the community who shared their spectral observations of minor bodies to support this work and the referee Pierre Vernazza for helpful comments during the review process. Rémi Flamary provided valuable support in the exploration of machine learning methods during the initial stage of the project. Unfortunately, the nature of asteroid B612 remains a mystery. This research has made use of IMCCE's SsODNet/Quaero VO tool. This research has made use of IMCCE's Miriade VO tool. This research has made use of the SVO Filter Profile Service supported from the Spanish MINECO through grant AYA2017-84089. All (or part) of the data utilised in this publication were obtained and made available by the MITHNEOS MIT-Hawaii Near-Earth Object Spectroscopic Survey. The IRTF is operated by the University of Hawaii under contract 80HQTR19D0030 with the National Aeronautics and Space Administration. The MIT component of this work is supported by NASA grant 80NSSC18K0849.

References

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, software available from `tensorflow.org`
Alí-Lagoa, V., & Delbó, M. 2017, A&A, 603, A55

Alí-Lagoa, V., de León, J., Licandro, J., et al. 2013, A&A, 554, A71
Alí-Lagoa, V., Licandro, J., Gil-Hutton, R., et al. 2016, A&A, 591, A14
Alí-Lagoa, V., Müller, T. G., Usui, F., & Hasegawa, S. 2018, A&A, 612, A85
Alvarez-Candal, A., Duffard, R., Lazzaro, D., & Michtchenko, T. 2006, A&A, 459, 969
Arredondo, A., Campins, H., Pinilla-Alonso, N., et al. 2021, Icarus, 358, 114210
Baek, J., McLachlan, G. J., & Flack, L. K. 2010, IEEE Trans. Pattern Anal. Mach. Intell., 32, 1298
Barucci, M. A., Perna, D., Popescu, M., et al. 2018, MNRAS, 476, 4481
Becker, T. M., Howell, E. S., Nolan, M. C., et al. 2015, Icarus, 248, 499
Bell, J. F. 1988, Meteoritics, 23, 256
Bell, J. F., Davis, D. R., Hartmann, W. K., & Gaffey, M. J. 1989, in Asteroids II, eds. R. P. Binzel, T. Gehrels, & M. S. Matthews, 921
Bendjoya, P., Cellino, A., Di Martino, M., & Saba, L. 2004, Icarus, 168, 374
Benner, L. 2002, Icarus, 158, 379
Berthier, J., Vachier, F., Marchis, F., Ďurech, J., & Carry, B. 2014, Icarus, 239, 118
Binzel, R. P. 2001, Icarus, 151, 139
Binzel, R. P., & Xu, S. 1993, Science, 260, 186
Binzel, R. P., Xu, S., Bus, S. J., et al. 1993, Science, 262, 1541
Binzel, R. P., Rivkin, A. S., Bus, S. J., Sunshine, J. M., & Burbine, T. H. 2001, Meteor. Planet. Sci. Suppl., 36, A20
Binzel, R. P., Birlan, M., Bus, S. J., et al. 2004a, Planet. Space Sci., 52, 291
Binzel, R. P., Perozzi, E., Rivkin, A. S., et al. 2004b, Meteor. Planet. Sci., 39, 351
Binzel, R. P., Rivkin, A. S., Stuart, J., et al. 2004c, Icarus, 170, 259
Binzel, R. P., Rivkin, A. S., Thomas, C. A., et al. 2009, Icarus, 200, 480
Binzel, R. P., DeMeo, F., Turtelboom, E., et al. 2019, Icarus, 324, 41
Birlan, M., Barucci, M. A., Vernazza, P., et al. 2004, New Astron., 9, 343
Birlan, M., Vernazza, P., Fulchignoni, M., et al. 2006, A&A, 454, 677
Birlan, M., Vernazza, P., & Nedelcu, D. A. 2007, A&A, 475, 747
Birlan, M., Nedelcu, D. A., Descamps, P., et al. 2011, MNRAS, 415, 587
Birlan, M., Nedelcu, D. A., Popescu, M., et al. 2014, MNRAS, 437, 176
Borisov, G., Christou, A., Bagnulo, S., et al. 2017, MNRAS, 466, 489
Borisov, G., Christou, A. A., Colas, F., et al. 2018, A&A, 618, A178
Bottke, W. F., Nesvorný, D., Grimm, R. E., Morbidelli, A., & O'Brien, D. P. 2006, Nature, 439, 821
Bouveyron, C., Celeux, G., Murphy, T. B., & Raftery, A. E. 2019, Model-Based Clustering and Classification for Data Science: With Applications in R, Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge University Press)
Bowell, E., Chapman, C. R., Gradie, J. C., Morrison, D., & Zellner, B. 1978, Icarus, 35, 313
Bowell, E., Muinonen, K., & Wasserman, L. H. 1994, in Asteroids, Comets, Meteors 1993, 160, eds. A. Milani, M. di Martino, & A. Cellino, 477
Brunetto, R., & Strazzulla, G. 2005, Icarus, 179, 265
Brunetto, R., Vernazza, P., Marchi, S., et al. 2006, Icarus, 184, 327
Brunetto, R., de León, J., & Licandro, J. 2007, A&A, 472, 653
Burbine, T. H. 2000, PhD thesis, Massachusetts Institute of Technology, Cambridge, USA
Burbine, T. H., McCoy, T. J., Meibom, A., Gladman, B., & Keil, K. 2002, Meteoritic Parent Bodies: Their Number and Identification, 653
Burbine, T. H., Buchanan, P. C., Dolkar, T., & Binzel, R. P. 2009, Meteor. Planet. Sci., 44, 1331
Burbine, T. H., Duffard, R., Buchanan, P. C., Cloutis, E. A., & Binzel, R. P. 2011, in 42nd Annual Lunar and Planetary Science Conference, Lunar and Planetary Science Conference, 2483
Bus, S. J. 1999, PhD thesis, Massachusetts Institute of Technology, Cambridge, USA
Bus, S. J., & Binzel, R. P. 2002a, Icarus, 158, 146
Bus, S. J., & Binzel, R. P. 2002b, Icarus, 158, 106
Candolle, A. P. D. 1813, Exposition des principes de la classification naturelle et de l'art de décrire et d'étudier les végétaux (Déterville)
Carry, B. 2012, Planet. Space Sci., 73, 98
Carvano, J. M., Hasselmann, P. H., Lazzaro, D., & Mothé-Diniz, T. 2010, A&A, 510, A43
Casey, A. R., Lattanzio, J. C., Aleti, A., et al. 2019, ApJ, 887, 73
Cellino, A., Bagnulo, S., Tanga, P., Novaković, B., & Delbó, M. 2014, MNRAS, 439, L75
Chapman, C. R., Johnson, T. V., & McCord, T. B. 1971, A Review of Spectrophotometric Studies of Asteroids, 267, ed. T. Gehrels (NASA), 51
Chapman, C. R., Morrison, D., & Zellner, B. 1975, Icarus, 25, 104
Chavez, C. F., Müller, T. G., Marshall, J. P., et al. 2021, MNRAS, 502, 4981
Chrbolková, K., Brunetto, R., Ďurech, J., et al. 2021, A&A, 654, A143
Clark, B. E., Veverka, J., Helfenstein, P., et al. 1999, Icarus, 140, 53
Clark, B. E., Bus, S. J., Rivkin, A. S., Shepard, M. K., & Shah, S. 2004, AJ, 128, 3070
Clark, B. E., Ockert-Bell, M. E., Cloutis, E. A., et al. 2009, Icarus, 202, 119
Clark, B. E., Ziffer, J., Nesvorný, D., et al. 2010, J. Geophys. Res. (Planets), 115, E06005
Cloutis, E. A., Gaffey, M. J., Smith, D. G. W., & Lambert, R. S. J. 1990a, J. Geophys. Res., 95, 8323
Cloutis, E. A., Gaffey, M. J., Smith, D. G. W., & Lambert, R. S. J. 1990b, J. Geophys. Res., 95, 281
Cloutis, E., Hudon, P., Hiroi, T., Gaffey, M., & Mann, P. 2011, Icarus, 216, 309
Cloutis, E. A., Izawa, M. R., & Beck, P. 2018, Reflectance Spectroscopy of Chondrites (Elsevier), 273
Cruikshank, D. P., & Hartmann, W. K. 1984, Science, 223, 281
de León, J., Licandro, J., Serra-Ricart, M., Pinilla-Alonso, N., & Campins, H. 2010, A&A, 517, A23
de León, J., Mothé-Diniz, T., Licandro, J., Pinilla-Alonso, N., & Campins, H. 2011, A&A, 530, A12
de León, J., Pinilla-Alonso, N., Campins, H., Licandro, J., & Marzo, G. 2012, Icarus, 218, 196
De Prá, M., Pinilla-Alonso, N., Carvano, J., et al. 2018, Icarus, 311, 35
De Sanctis, M. C., Ammannito, E., Migliorini, A., et al. 2011a, MNRAS, 412, 2318
De Sanctis, M. C., Migliorini, A., Luzia Jasmin, F., et al. 2011b, A&A, 533, A77
De Sanctis, M. C., Ammannito, E., Raponi, A., et al. 2015, Nature, 528, 241
Delbó, M., & Tanga, P. 2009, Planet. Space Sci., 57, 259
Delbó, M., Harris, A. W., Binzel, R. P., Pravec, P., & Davies, J. K. 2003, Icarus, 166, 116
Delbó, M., Gayon-Markt, J., Busso, G., et al. 2012, Planet. Space Sci., 73, 86
DeMeo, F. E., & Carry, B. 2013, Icarus, 226, 723
DeMeo, F. E., & Carry, B. 2014, Nature, 505, 629
DeMeo, F. E., Binzel, R. P., Slivan, S. M., & Bus, S. J. 2009, Icarus, 202, 160
DeMeo, F. E., Binzel, R. P., Carry, B., Polishook, D., & Moskovitz, N. A. 2014, Icarus, 229, 392
DeMeo, F. E., Polishook, D., Carry, B., et al. 2019, Icarus, 322, 13
Dempster, A. P., Laird, N. M., & Rubin, D. B. 1977, J. Roy. Stat. Soc. B (Methodological), 39, 1
Devogèle, M., Tanga, P., Cellino, A., et al. 2018, Icarus, 304, 31
Devogèle, M., Moskovitz, N., Thirouin, A., et al. 2019, AJ, 158, 196
Dong-fang, Z., Peng, L., Wei, Z., et al. 2016, Chinese Astron. Astrophys., 40, 555
Drummond, J., & Christou, J. 2008, Icarus, 197, 480
Drummond, J. D., Merline, W. J., Carry, B., et al. 2018, Icarus, 305, 174
Duffard, R., & Roig, F. 2009, Planet. Space Sci., 57, 229
Duffard, R., Lazzaro, D., Licandro, J., et al. 2004, Icarus, 171, 120
Emery, J., & Brown, R. 2003, Icarus, 164, 104
Emery, J. P., Burr, D. M., & Cruikshank, D. P. 2011, AJ, 141, 25
Eschrig, J., Bonal, L., Beck, P., & Prestgard, T. 2021, Icarus, 354, 114034
Eschrig, J., Bonal, L., Mahlke, M., et al. 2022, Icarus, 381, 115012
Fieber-Beyer, S. K. 2010, PhD thesis, University of North Dakota, Grand Forks, USA
Fieber-Beyer, S. K., & Gaffey, M. J. 2011, Icarus, 214, 645
Fieber-Beyer, S. K., & Gaffey, M. J. 2014, Icarus, 229, 99
Fieber-Beyer, S. K., & Gaffey, M. J. 2015, Icarus, 257, 113
Fieber-Beyer, S. K., Gaffey, M. J., Kelley, M. S., et al. 2011, Icarus, 213, 524
Fieber-Beyer, S. K., Gaffey, M. J., Hardersen, P. S., & Reddy, V. 2012, Icarus, 221, 593
Fornasier, S., Dotto, E., Hainaut, O., et al. 2007, Icarus, 190, 622
Fornasier, S., Clark, B., Dotto, E., et al. 2010, Icarus, 210, 655
Fornasier, S., Clark, B. E., Migliorini, A., & Ockert-Bell, M. 2011, NASA Planetary Data System, EAR
Fornasier, S., Lantz, C., Barucci, M., & Lazzarin, M. 2014, Icarus, 233, 163
Fornasier, S., Lantz, C., Perna, D., et al. 2016, Icarus, 269, 1
Fujiwara, A., Kawaguchi, J., Yeomans, D. K., et al. 2006, Science, 312, 1330
Gaffey, M. J., & McCord, T. B. 1979, in Asteroids, ed. T. Gehrels, & M. S. Matthews, 688
Gaffey, M. J., Reed, K. L., & Kelley, M. S. 1992, Icarus, 100, 95
Gartrelle, G. M., Hardersen, P. S., Izawa, M. R. M., & Nowinski, M. C. 2021, NASA Planetary Data System, 6
Gietzen, K. M., Lacy, C. H. S., Ostrowski, D. R., & Sears, D. W. G. 2012, Meteor. Planet. Sci., 47, 1789
Gomes, R., Levison, H. F., Tsiganis, K., & Morbidelli, A. 2005, Nature, 435, 466
Gradie, J., & Tedesco, E. 1982, Science, 216, 1405
Granvik, M., & Brown, P. 2018, Icarus, 311, 271
Grav, T., Mainzer, A. K., Bauer, J., et al. 2011, ApJ, 742, 40
Grav, T., Mainzer, A. K., Bauer, J., et al. 2012a, ApJ, 744, 197
Grav, T., Mainzer, A. K., Bauer, J. M., Masiero, J. R., & Nugent, C. R. 2012b, ApJ, 759, 49
Hanuš, J., Delbo', M., Ďurech, J., & Alí-Lagoa, V. 2015, Icarus, 256, 101

Hanuš, J., Delbo', M., Vokrouhlický, D., et al. 2016, A&A, 592, A34
Hanuš, J., Viikinkoski, M., Marchis, F., et al. 2017, A&A, 601, A114
Hanuš, J., Vokrouhlický, D., Delbo', M., et al. 2018, A&A, 620, A8
Hardersen, P., Gaffey, M., & Abell, P. 2005, Icarus, 175, 141
Hardersen, P. S., Cloutis, E. A., Reddy, V., Mothé-Diniz, T., & Emery, J. P. 2011, Meteor. Planet. Sci., 46, 1910
Hardersen, P. S., Reddy, V., Roberts, R., & Mainzer, A. 2014, Icarus, 242, 269
Hardersen, P. S., Reddy, V., & Roberts, R. 2015, ApJS, 221, 19
Hardersen, P. S., Reddy, V., Cloutis, E., et al. 2018, AJ, 156, 11
Harris, A. W., & Lagerros, J. S. V. 2002, Asteroids in the Thermal Infrared, 205
Hasegawa, S., Kuroda, D., Kitazato, K., et al. 2018, PASJ, 70, 114
Hasegawa, S., Kasuga, T., Usui, F., & Kuroda, D. 2021a, PASJ, 73, 240
Hasegawa, S., Marsset, M., DeMeo, F. E., et al. 2021b, ApJ, 916, L6
Helfenstein, P., Veverka, J., Thomas, P., et al. 1994, Icarus, 107, 37
Helfenstein, P., Veverka, J., Thomas, P., et al. 1996, Icarus, 120, 48
Herald, D., Frappa, E., Gault, D., et al. 2019, NASA Planetary Data System, 3
Hiroi, T., Zolensky, M. E., Pieters, C. M., & Lipschutz, M. E. 1996, Meteor. Planet. Sci., 31, 321
Huang, J., Ji, J., Ye, P., et al. 2013, Scientific Rep., 3, 3411
Hung, D., Hanuš, J., Masiero, J. R., & Tholen, D. J. 2022, Planet. Sci. J., 3, 56
Ieva, S., Dotto, E., Lazzaro, D., et al. 2018, MNRAS, 479, 2607
Ivezić, v., Ivezić, V., Moeyens, J., et al. 2022, Icarus, 371, 114696
Jasmim, F. L., Lazzaro, D., Carvano, J. M. F., Mothé-Diniz, T., & Hasselmann, P. H. 2013, A&A, 552, A85
Jiang, H., & Ji, J. 2021, AJ, 162, 40
Jorda, L., Lamy, P., Gaskell, R., et al. 2012, Icarus, 221, 1089
Kasuga, T., Usui, F., Ootsubo, T., Hasegawa, S., & Kuroda, D. 2013, AJ, 146, 1
Kasuga, T., Usui, F., Shirahata, M., et al. 2015, AJ, 149, 37
Keller, H. U., Barbieri, C., Koschny, D., et al. 2010, Science, 327, 190
Kelley, M. S., Vilas, F., Gaffey, M. J., & Abell, P. A. 2003, Icarus, 165, 215
Koren, S. C., Wright, E. L., & Mainzer, A. 2015, Icarus, 258, 82
Kuroda, D., Ishiguro, M., Takato, N., et al. 2014, PASJ, 66, 51
Landsman, Z. A., Campins, H., Pinilla-Alonso, N., Hanuš, J., & Lorenzi, V. 2015, Icarus, 252, 186
Lantz, C., Brunetto, R., Barucci, M., et al. 2017, Icarus, 285, 43
Lantz, C., Binzel, R., & DeMeo, F. 2018, Icarus, 302, 10
Lazzarin, M., Marchi, S., Barucci, M., Di Martino, M., & Barbieri, C. 2004, Icarus, 169, 373
Lazzarin, M., Marchi, S., Magrin, S., & Licandro, J. 2005, MNRAS, 359, 1575
Lazzaro, D., Angeli, C., Carvano, J., et al. 2004, Icarus, 172, 179
Lazzaro, D., Angeli, C. A., Carvano, J. M., et al. 2007, NASA Planetary Data System, EAR
Li, J.-Y., Le Corre, L., Schröder, S. E., et al. 2013, Icarus, 226, 1252
Li, J.-Y., Reddy, V., Nathues, A., et al. 2016, ApJ, 817, L22
Licandro, J., Alí-Lagoa, V., Tancredi, G., & Fernández, Y. 2016, A&A, 585, A9
Licandro, J., Popescu, M., de León, J., et al. 2018, A&A, 618, A170
Little, R., & Rubin, D. 2019, Wiley Series in Probability and Statistics
Lucas, M. P., Emery, J. P., Pinilla-Alonso, N., Lindsay, S. S., & Lorenzi, V. 2017, Icarus, 291, 268
Lucas, M. P., Emery, J. P., Hiroi, T., & McSween, H. Y. 2019, Meteor. Planet. Sci., 54, 157
Magri, C., Nolan, M. C., Ostro, S. J., & Giorgini, J. D. 2007, Icarus, 186, 126
Mahlke, M., Carry, B., & Denneau, L. 2021, Icarus, 354, 114094
Mainzer, A., Grav, T., Masiero, J., et al. 2011, ApJ, 741, 90
Mainzer, A., Grav, T., Masiero, J., et al. 2012, ApJ, 760, L12
Mainzer, A., Bauer, J., Cutri, R. M., et al. 2014a, ApJ, 792, 30
Mainzer, A., Bauer, J., Grav, T., et al. 2014b, ApJ, 784, 110
Mainzer, A., Grav, T., Bauer, J., et al. 2015, AJ, 149, 172
Marchi, S., Lazzarin, M., & Magrin, S. 2004, A&A, 420, L5
Marchi, S., Lazzarin, M., Paolicchi, P., & Magrin, S. 2005, Icarus, 175, 170
Marchis, F., Enriquez, J., Emery, J., et al. 2012, Icarus, 221, 1130
Marsset, M., Vernazza, P., Gourgeot, F., et al. 2014, A&A, 568, A7
Marsset, M., Vernazza, P., Birlan, M., et al. 2016, A&A, 586, A15
Marsset, M., DeMeo, F. E., Burt, B., et al. 2022, AJ, 163, 165
Masiero, J. R., Mainzer, A. K., Grav, T., et al. 2011, ApJ, 741, 68
Masiero, J. R., Mainzer, A. K., Grav, T., et al. 2012, ApJ, 759, L8
Masiero, J. R., Grav, T., Mainzer, A. K., et al. 2014, ApJ, 791, 121
Masiero, J. R., DeMeo, F. E., Kasuga, T., & Parker, A. H. 2015, Asteroid Family Physical Properties (University of Arizona Press)
Masiero, J. R., Nugent, C., Mainzer, A. K., et al. 2017, AJ, 154, 168
Masiero, J. R., Wright, E. L., & Mainzer, A. K. 2019, AJ, 158, 97
Masiero, J. R., Mainzer, A. K., Bauer, J. M., et al. 2020a, Planet. Sci. J., 1, 5
Masiero, J. R., Smith, P., Teodoro, L. D., et al. 2020b, Planet. Sci. J., 1, 9
Masiero, J. R., Mainzer, A. K., Bauer, J. M., et al. 2021, Planet. Sci. J., 2, 162
Matlovič, P., de Leon, J., Medeiros, H., et al. 2020, A&A, 643, A107
Matter, A., Delbo, M., Ligori, S., Crouzet, N., & Tanga, P. 2011, Icarus, 215, 47
Matter, A., Delbo, M., Carry, B., & Ligori, S. 2013, Icarus, 226, 419
McCord, T. B., & Chapman, C. R. 1975, ApJ, 195, 553
McCord, T. B., Adams, J. B., & Johnson, T. V. 1970, Science, 168, 1445
Migliorini, A., De Sanctis, M. C., Lazzaro, D., & Ammannito, E. 2017, MNRAS, 464, 1718
Migliorini, A., De Sanctis, M. C., Lazzaro, D., & Ammannito, E. 2018, MNRAS, 475, 353
Montanari, A., & Viroli, C. 2010, Stat. Model., 10, 441
Morbidelli, A., Levison, H. F., Tsiganis, K., & Gomes, R. 2005, Nature, 435, 462
Morbidelli, A., Walsh, K. J., O'Brien, D. P., Minton, D. A., & Bottke, W. F. 2015, The Dynamical Evolution of the Asteroid Belt (University of Arizona Press)
Moskovitz, N., Jedicke, R., & Willman, M. 2009, NASA Planetary Data System, EAR
Moskovitz, N. A., Willman, M., Burbine, T. H., Binzel, R. P., & Bus, S. J. 2010, Icarus, 208, 773
Moskovitz, N. A., Fatka, P., Farnocchia, D., et al. 2019, Icarus, 333, 165
Mothé-Diniz, T., Carvano, J., Bus, S., Duffard, R., & Burbine, T. 2008, Icarus, 195, 277
Mueller, B. E., Tholen, D. J., Hartmann, W. K., & Cruikshank, D. P. 1992, Icarus, 97, 150
Mueller, M., Delbó, M., Hora, J. L., et al. 2011, AJ, 141, 109
Müller, T. G., & Blommaert, J. A. D. L. 2004, A&A, 418, 347
Müller, T. G., Kiss, C., Scheirich, P., et al. 2014, A&A, 566, A22
Nakamura, T., Noguchi, T., Tanaka, M., et al. 2011, Science, 333, 1113
Nedelcu, D. A., Birlan, M., Vernazza, P., et al. 2007, A&A, 473, L33
Neeley, J., Clark, B., Ockert-Bell, M., et al. 2014, Icarus, 238, 37
Nugent, C. R., Mainzer, A., Masiero, J., et al. 2015, ApJ, 814, 117
Nugent, C. R., Mainzer, A., Bauer, J., et al. 2016, AJ, 152, 63
Ockert-Bell, M. E., Clark, B. E., Shepard, M. K., et al. 2008, Icarus, 195, 206
Ockert-Bell, M. E., Clark, B. E., Shepard, M. K., et al. 2010, Icarus, 210, 674
Ostrowski, D. R., Lacy, C. H., Gietzen, K. M., & Sears, D. W. 2011, Icarus, 212, 682
Oszkiewicz, D., Troianskyi, V., Föhring, D., et al. 2020, A&A, 643, A117
Pearson, K. 1901, London Edinburgh Dublin Philos. Mag. J. Sci., 2, 559
Perna, D., Barucci, M., Fulchignoni, M., et al. 2018, Planet. Space Sci., 157, 82
Pinilla-Alonso, N., de León, J., Walsh, K., et al. 2016, Icarus, 274, 231
Pinilla-Alonso, N., De Pra, M., de Leon, J., et al. 2021, NASA Planetary Data System, 8
Polishook, D., Moskovitz, N., Binzel, R. P., et al. 2014, Icarus, 233, 9
Popescu, M., Birlan, M., Binzel, R., et al. 2011, A&A, 535, A15
Popescu, M., Birlan, M., & Nedelcu, D. A. 2012, A&A, 544, A130
Popescu, M., Birlan, M., Nedelcu, D. A., Vaubaillon, J., & Cristescu, C. P. 2014, A&A, 572, A106
Popescu, M., Licandro, J., Carvano, J. M., et al. 2018, A&A, 617, A12
Popescu, M., Vaduvescu, O., de León, J., et al. 2019, A&A, 627, A124
Pravec, P., Harris, A. W., Kušnirák, P., Galád, A., & Hornoch, K. 2012, Icarus, 221, 365
Rayner, J. T., Toomey, D. W., Onaka, P. M., et al. 2003, PASP, 115, 362
Reddy, V. 2010, NASA Planetary Data System, EAR
Reddy, V., & Sanchez, J. A. 2016, NASA Planetary Data System, EAR
Reddy, V., & Sanchez, J. A. 2017, NASA Planetary Data System
Reddy, V., Carvano, J. M., Lazzaro, D., et al. 2011, Icarus, 216, 184
Reddy, V., Sanchez, J. A., Furfaro, R., et al. 2018, AJ, 155, 140
Rivkin, A. S. 1995, Icarus, 117, 90
Rivkin, A. S. 2000, Icarus, 145, 351
Rivkin, A. S. 2012, Icarus, 221, 744
Rivkin, A. S., Binzel, R. P., Sunshine, J., et al. 2004, Icarus, 172, 408
Rivkin, A. S., Thomas, C. A., Howell, E. S., & Emery, J. P. 2015, AJ, 150, 198
Rozitis, B., & Green, S. F. 2014, A&A, 568, A43
Rozitis, B., Duddy, S. R., Green, S. F., & Lowry, S. C. 2013, A&A, 555, A20
Rubin, D. B., & Thayer, D. T. 1982, Psychometrika, 47, 69
Russell, C. T., Raymond, C. A., Coradini, A., et al. 2012, Science, 336, 684
Russell, C. T., Raymond, C. A., Ammannito, E., et al. 2016, Science, 353, 1008
Ryan, E. L., & Woodward, C. E. 2010, AJ, 140, 933
Ryan, E. L., Mizuno, D. R., Shenoy, S. S., et al. 2015, A&A, 578, A42
Sanchez, J. A., Reddy, V., Nathues, A., et al. 2012, Icarus, 220, 36
Sanchez, J. A., Michelsen, R., Reddy, V., & Nathues, A. 2013, Icarus, 225, 131
Sanchez, J. A., Reddy, V., Kelley, M. S., et al. 2014, Icarus, 228, 288
Savitzky, A., & Golay, M. J. E. 1964, Anal. Chem., 36, 1627
Shepard, M. K., Clark, B. E., Nolan, M. C., et al. 2008a, Icarus, 193, 20
Shepard, M. K., Kressler, K. M., Clark, B. E., et al. 2008b, Icarus, 195, 220
Shepard, M. K., Clark, B. E., Ockert-Bell, M., et al. 2010, Icarus, 208, 221
Shepard, M. K., Taylor, P. A., Nolan, M. C., et al. 2015, Icarus, 245, 38

Shestopalov, D., Golubeva, L., McFadden, L., Fornasier, S., & Taran, M. 2010, Planet. Space Sci., 58, 1400

Shevchenko, V. G., Belskaya, I. N., Muinonen, K., et al. 2016, Planet. Space Sci., 123, 101

Sierks, H., Lamy, P., Barbieri, C., et al. 2011, Science, 334, 487

Solontoi, M. R., Hammergren, M., Gyuk, G., & Puckett, A. 2012, Icarus, 220, 577

Strazzulla, G., Dotto, E., Binzel, R., et al. 2005, Icarus, 174, 31

Sunshine, J. M., Bus, S. J., Corrigan, C. M., McCoy, T. J., & Burbine, T. H. 2007, Meteor. Planet. Sci., 42, 155

Sunshine, J. M., Connolly, H. C., McCoy, T. J., Bus, S. J., & La Croix, L. M. 2008, Science, 320, 514

Tatsumi, E., Domingue, D., Hirata, N., et al. 2018, Icarus, 311, 175

Tedesco, E. F., Noah, P. V., Noah, M., & Price, S. D. 2002, AJ, 123, 1056

Tholen, D. J. 1984, PhD thesis, University of Arizona, Tucson, USA

Tholen, D. J., & Barucci, M. A. 1989, in Asteroids II, eds. R. P. Binzel, T. Gehrels, & M. S. Matthews, 298

Thomas, P. 2000, Icarus, 145, 348

Thomas, P., Veverka, J., Simonelli, D., et al. 1994, Icarus, 107, 23

Thomas, P., Belton, M., Carcich, B., et al. 1996, Icarus, 120, 20

Thomas, P., Veverka, J., Bell, J., et al. 1999, Icarus, 140, 17

Thomas, C. A., Trilling, D. E., & Rivkin, A. S. 2012, Icarus, 219, 505

Thomas, C. A., Trilling, D. E., Rivkin, A. S., & Linder, T. 2021, AJ, 161, 99

Tipping, M. E., & Bishop, C. M. 1999, J. Roy. Stat. Soc. B (Stat. Methodol.), 61, 611

Trilling, D. E., Mueller, M., Hora, J. L., et al. 2010, AJ, 140, 770

Trilling, D. E., Mommert, M., Hora, J., et al. 2016, AJ, 152, 172

Tsiganis, K., Gomes, R., Morbidelli, A., & Levison, H. F. 2005, Nature, 435, 459

Usui, F., Kuroda, D., Müller, T. G., et al. 2011, PASJ, 63, 1117

Veeder, G., Matson, D., & Tedesco, E. 1983, Icarus, 55, 177

Vernazza, P., Mothé-Diniz, T., Barucci, M. A., et al. 2005, A&A, 436, 1113

Vernazza, P., Birlan, M., Rossi, A., et al. 2006, A&A, 460, 945

Vernazza, P., Lamy, P., Groussin, O., et al. 2011, Icarus, 216, 650

Vernazza, P., Zanda, B., Binzel, R. P., et al. 2014, ApJ, 791, 120

Vernazza, P., Marsset, M., Beck, P., et al. 2015, ApJ, 806, 204

Vernazza, P., Marsset, M., Beck, P., et al. 2016, AJ, 152, 54

Vernazza, P., Castillo-Rogez, J., Beck, P., et al. 2017, AJ, 153, 72

Vernazza, P., Ferrais, M., Jorda, L., et al. 2021, A&A, 654, A56

Veverka, J., Robinson, M., Thomas, P., et al. 2000, Science, 289, 2088

Viikinkoski, M., Hanuš, J., Kaasalainen, M., Marchis, F., & Ďurech, J. 2017, A&A, 607, A117

Vilas, F., Smith, B. A., McFadden, L. A., et al. 2006, NASA Planetary Data System, EAR

Vokrouhlický, D., Bottke, W. F., & Nesvorný, D. 2016, AJ, 152, 39

Warren, P. H. 2011, Earth Planet. Sci. Lett., 311, 93

Watters, T. R., & Prinz, M. 1979, Lunar Planet. Sci. Conf. Proc., 1, 1073

Willman, M., Jedicke, R., & Moskovitz, N. 2009, NASA Planetary Data System, EAR

Wong, I., Brown, M. E., & Emery, J. P. 2017, AJ, 154, 104

Wright, E. L., Mainzer, A., Masiero, J., Grav, T., & Bauer, J. 2016, AJ, 152, 79

Xu, S. 1994, PhD thesis, Massachusetts Institute of Technology, Cambridge, USA

Xu, S., Binzel, R. P., Burbine, T. H., & Bus, S. J. 1995, Icarus, 115, 1

Yang, B., & Jewitt, D. 2007, AJ, 134, 223

Yang, B., & Jewitt, D. 2011, AJ, 141, 95

Yang, B., Wahhaj, Z., Beauvalet, L., et al. 2016, ApJ, 820, L35

Yang, B., Hanuš, J., Brož, M., et al. 2020, A&A, 643, A38

Yu, L.-L., Ji, J., & Ip, W.-H. 2017, Res. Astron. Astrophys., 17, 070

Zellner, B., & Gradie, J. 1976, AJ, 81, 262

Zellner, B., Tholen, D., & Tedesco, E. 1985, Icarus, 61, 355

Appendix A: Distribution of albedos in cluster

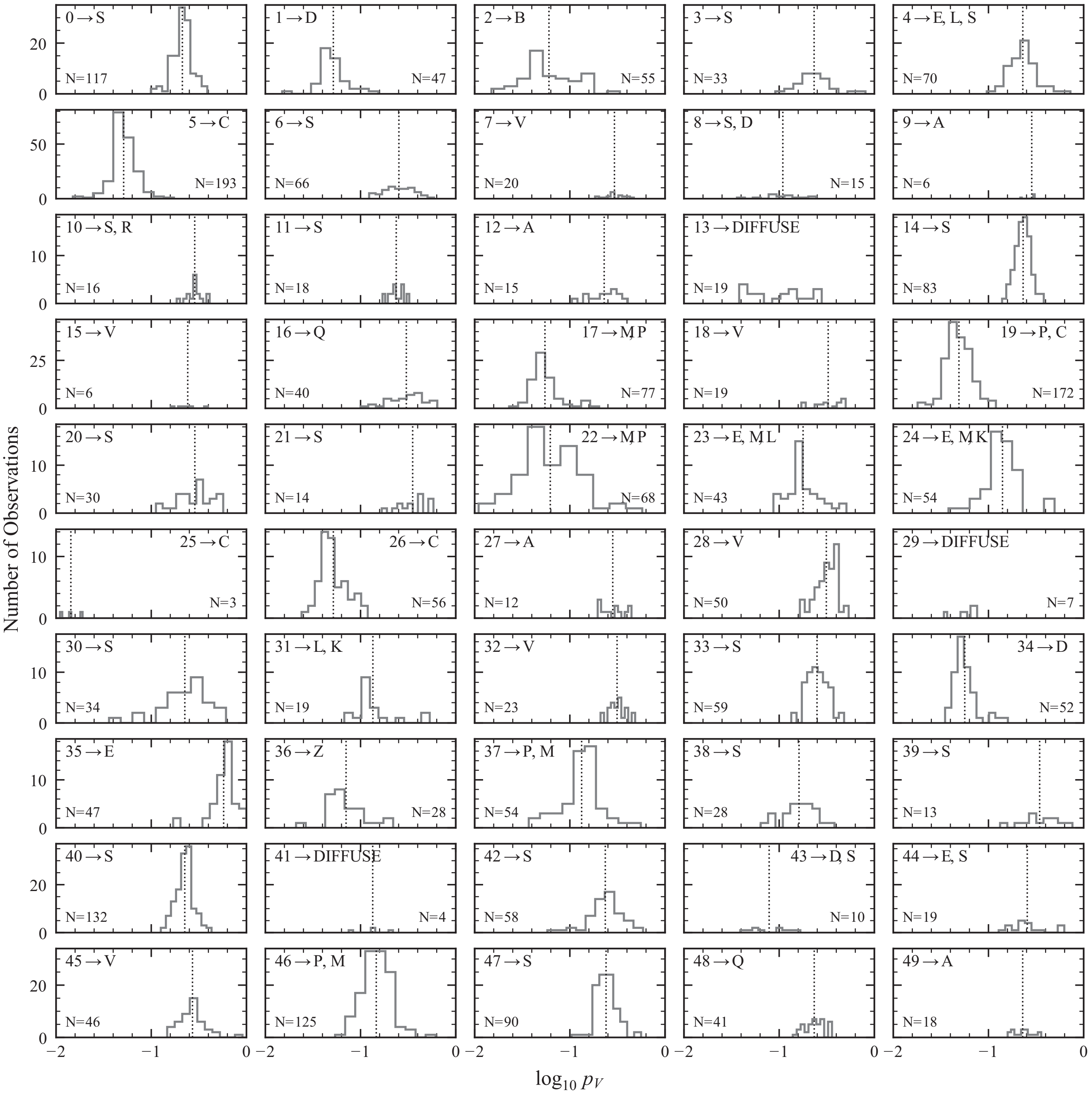


Fig. A.1. Overview of the albedo distribution per cluster, including the number N of albedos and the asteroid classes to which the cluster contributes , excluding classes with fewer than three contributed observations except for cluster 25 which has only three observations. The classes are sorted by the total number of observations the cluster contributed. The dotted line gives the mean value of the albedos per cluster except for diffuse clusters and cluster 25. The y-axis limit is different in each row.

Appendix B: Feature centres and windows

Table B.1. Listed are the mean band centres and the mean upper and lower band limits determined using the visually identified features in the input data.

Feature	Centre / µm	Lower Limit / µm	Upper Limit / µm
e	0.50 ± 0.01	0.450	0.539
h	0.69 ± 0.01	0.549	0.834
k	0.91 ± 0.02	0.758	1.060

Notes. These values are applied when using the automatic feature detection with the classy tool.

Appendix C: Distribution of spectra and albedos in classes

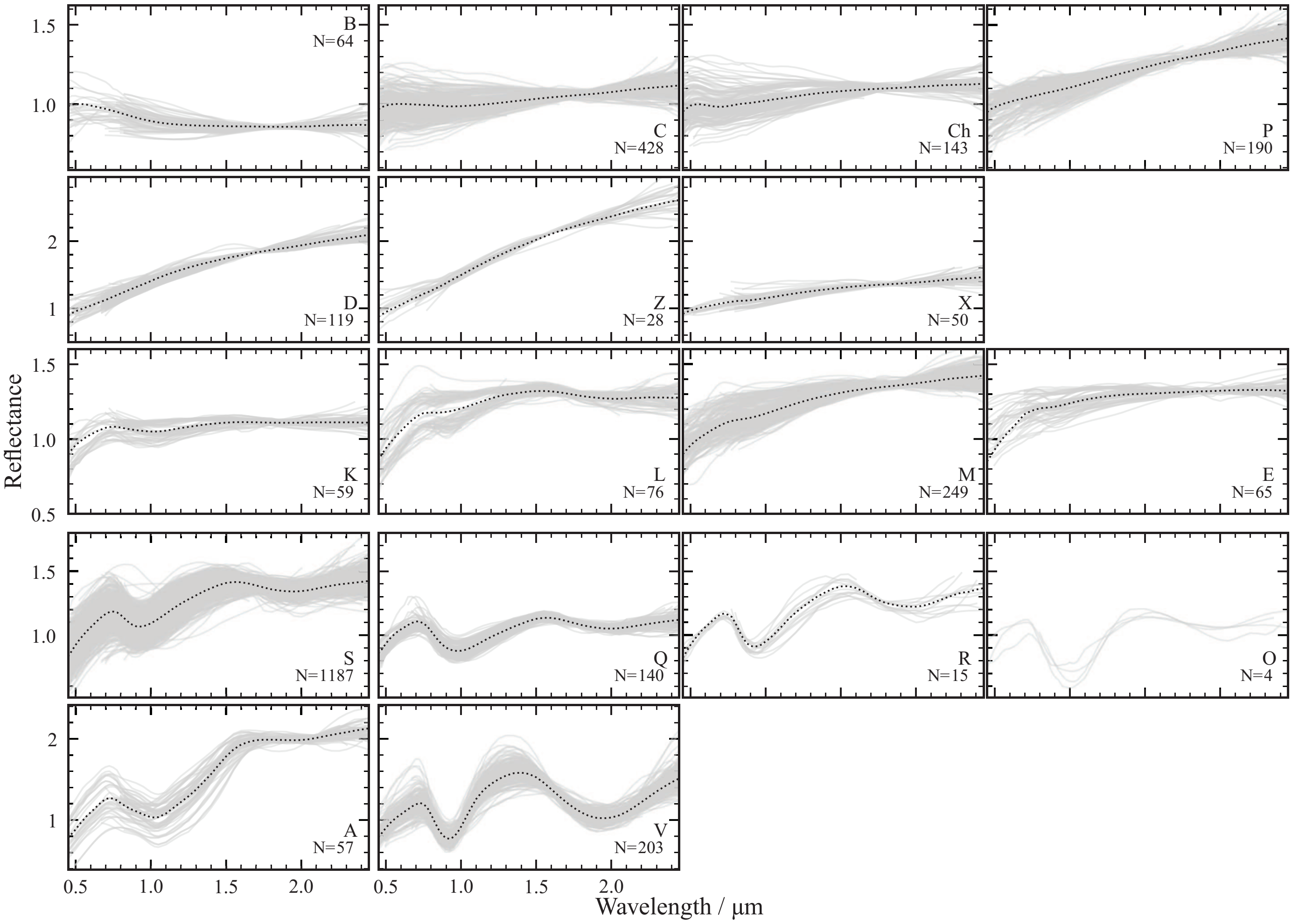


Fig. C.1. Distribution of spectral observations over the 17 classes assigned in this taxonomy. The number *N* of spectral observations assigned to the class is given under the respective letter. Spectra contributed by diffuse clusters are excluded.

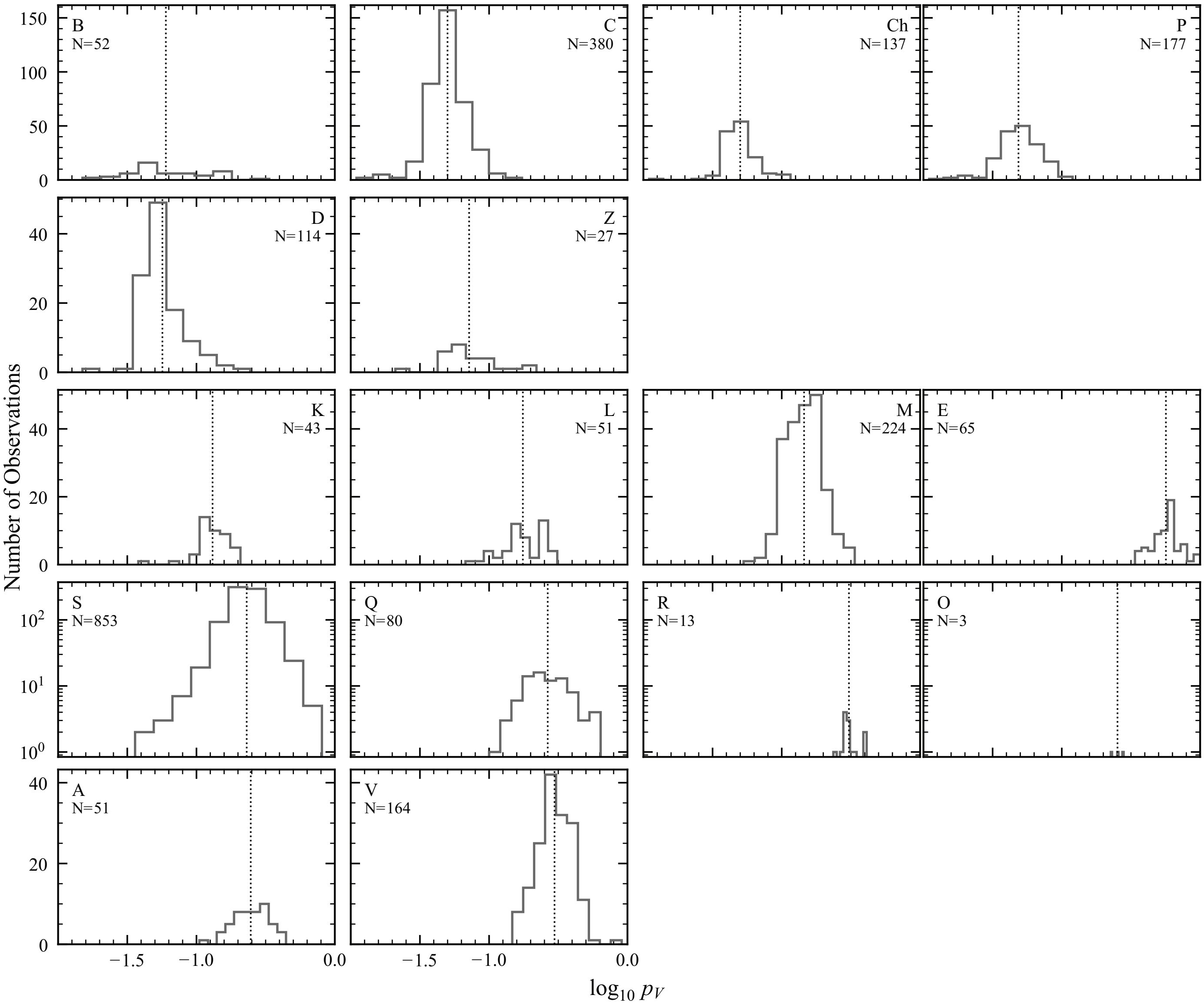


Fig. C.2. Distribution of albedo observations over the 17 classes assigned in this taxonomy excluding the X-class. The number N of albedo observations assigned to the class is given under the respective letter. Albedos contributed by diffuse clusters are excluded.

Appendix D: Cluster-to-class decision tree

Table D.1. Cluster-to-class decision tree.

Cluster		Class
0, 3, 6, 11, 14, 20, 21, 30, 33, 38, 39, 40, 42, 47	→	S
1, 34	→	D
2	→	B
5, 25, 26	→	C
7, 15, 18, 28, 32, 45	→	V
9, 12, 27, 49	→	A
16, 48	→	Q
36	→	Z
4	$P_{23}(z_3, z_4)/P_{40}(z_3, z_4)$	L, S
8, 43	GMM(z_2, z_4)	D, S
10	GMM(z_1, z_2)	R, S
13	GMM(z_2, z_4)	C, O, Q
17, 22, 35, 37, 46	$P_{\mathtt{E}}(p_V)/P_{\mathtt{M}}(p_V)/P_{\mathtt{P}}(p_V)$	E, M, P, X
19	GMM(z_1, z_4)	C, P
23	GMM(z_1, z_4)	L, M
24	GMM(z_2, z_3)	K, M
29	GMM(z_1, z_2)	A, B, C, D, M, P, S, Q, V
31	GMM(z_3, z_4)	K, L
37	GMM(z_2, z_4)	L, M
41	GMM(z_1, z_2)	B, V
44	$P_{\mathtt{E}}(p_V)/P_{\mathtt{M}}(p_V)$	E, S
Class is B, C, P, or X and h-feature is present		Ch

Notes. Overview of the computation of the asteroid-class probability for each observation based on its cluster probabilities. The *upper part* of the table contains clusters whose members are mapped to a single asteroid class. The *lower part* contains clusters where the resulting asteroid class probabilities depend on the criterion given in the *middle column*. GMM(z_x, z_y) means that the cluster probability is split based on a Gaussian mixture model with N components fit to all cluster members in z_x and z_y, where N is equal to the number of possible outcome classes (i.e. each mixture component represents one candidate class). $P_X(y)$ refers to the probability of belonging to the class or cluster X given the value of y. The last line gives the definition of the Ch-class, which is the last step of the classification.

Appendix E: References of spectra and visual albedos

Table E.1. Spectroscopic data references.

Alvarez-Candal et al. (2006); Arredondo et al. (2021); Barucci et al. (2018); Bendjoya et al. (2004); Binzel (2001); Binzel et al. (2001, 2004a,b,c, 2009); Birlan et al. (2004, 2006, 2007, 2011, 2014); Borisov et al. (2017, 2018); Burbine (2000); Burbine et al. (2009); Bus (1999); Bus & Binzel (2002a,b); Clark et al. (2004, 2009); de León et al. (2010, 2011); De Prá et al. (2018); De Sanctis et al. (2011a,b); Devogèle et al. (2018, 2019); Duffard & Roig (2009); Duffard et al. (2004); Emery & Brown (2003); Emery et al. (2011); Fieber-Beyer (2010); Fieber-Beyer & Gaffey (2011, 2014, 2015); Fieber-Beyer et al. (2011, 2012); Fornasier et al. (2007, 2011, 2014, 2016); Gartrelle et al. (2021); Gietzen et al. (2012); Hardersen et al. (2011, 2014, 2015, 2018); Hasegawa et al. (2018, 2021a); Ieva et al. (2018); Jasmim et al. (2013); Kasuga et al. (2013, 2015); Kuroda et al. (2014); Landsman et al. (2015); Lazzarin et al. (2004, 2005); Lazzaro et al. (2007); Licandro et al. (2018); Lucas et al. (2017, 2019); Marchi et al. (2004, 2005); Marsset et al. (2014, 2022); Matlovič et al. (2020); Migliorini et al. (2017, 2018); Moskovitz et al. (2009, 2010, 2019); Nedelcu et al. (2007); Neeley et al. (2014); Ockert-Bell et al. (2008, 2010); Ostrowski et al. (2011); Oszkiewicz et al. (2020); Perna et al. (2018); Pinilla-Alonso et al. (2016, 2021); Polishook et al. (2014); Popescu et al. (2011, 2012, 2014, 2019); Rayner et al. (2003); Reddy (2010); Reddy & Sanchez (2016, 2017); Reddy et al. (2011, 2018); Rivkin et al. (2004); Sanchez et al. (2013, 2014); Shepard et al. (2008a); Sunshine et al. (2007, 2008); Vernazza et al. (2005, 2006, 2014, 2016); Vilas et al. (2006); Willman et al. (2009); Wong et al. (2017); Xu (1994); Xu et al. (1995); Yang & Jewitt (2007, 2011); Yang et al. (2020)

Table E.2. Data references for albedos, diameters, and absolute magnitudes.

Alí-Lagoa & Delbó (2017); Alí-Lagoa et al. (2013); Alí-Lagoa et al. (2016); Alí-Lagoa et al. (2018); Becker et al. (2015); Benner (2002); Berthier et al. (2014); Bowell et al. (1994); Chavez et al. (2021); Clark et al. (1999); Delbó & Tanga (2009); Delbó et al. (2003); Dong-fang et al. (2016); Drummond & Christou (2008); Drummond et al. (2018); Fujiwara et al. (2006); Grav et al. (2011, 2012a,b); Hanuš et al. (2015, 2016, 2017, 2018); Helfenstein et al. (1994, 1996); Herald et al. (2019); Huang et al. (2013); Hung et al. (2022); Jiang & Ji (2021); Jorda et al. (2012); Keller et al. (2010); Koren et al. (2015); Li et al. (2013, 2016); Licandro et al. (2016); Magri et al. (2007); Mainzer et al. (2011, 2012, 2014a,b); Marchis et al. (2012); Masiero et al. (2011, 2012, 2014, 2017, 2019, 2020a,b, 2021); Matter et al. (2011, 2013); Mueller et al. (2011); Müller & Blommaert (2004); Müller et al. (2014); Nugent et al. (2015, 2016); Pravec et al. (2012); Rozitis & Green (2014); Rozitis et al. (2013); Russell et al. (2012, 2016); Ryan & Woodward (2010); Ryan et al. (2015); Shepard et al. (2008a,b); Sierks et al. (2011); Tatsumi et al. (2018); Thomas (2000); Thomas et al. (1994, 1996, 1999); Trilling et al. (2010, 2016); Usui et al. (2011); Vernazza et al. (2021); Veverka et al. (2000); Viikinkoski et al. (2017); Yu et al. (2017)