Universiteit Leiden

The Identification of Spectral Patterns in Low Resolution XP Spectra of Gaia DR3 Using Compression-based Pattern Mining.

Dean Kuurstra (s3343715)

Supervisors:
Dr. Saber Salehkaleybar & Dr. Anthony Brown

BACHELOR THESIS– DATA SCIENCE AND ARTIFICIAL INTELLIGENCE

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

August 24, 2026

Abstract

Gaia DR3 released 220 million XP spectra as coefficients of a basis of orthonormal Gauss-Hermite functions, which was chosen for compactness, rather than for physical meaning. This means that the largest homogeneous spectroscopic dataset is also one of the least interpretable. We apply sparse dictionary learning to $2.046 \cdot 10^6$ XP spectra to find an atom vocabulary made up of interpretable components that encode each spectrum. The dictionary size and sparsity are selected by minimum description length and class discovery. We find that MDL selects a dictionary consisting of 24 atoms, with a sparsity penalty $\lambda = 3.0$, which yields a median compression of $8.71\times$ at a reconstruction with an error equal to measurement noise, against a $13.6\times$ compression for an 8-component PCA at similar reconstruction accuracy. A 256-atom dictionary recovers 8 distinct classes in individual atoms with criteria set in Table 5, with one atom recovering 97% of all carbon stars, but does so at roughly $2.5\times$ the bit size. Approximately 42% of atoms are reproducible across random initializations beyond the similarity of the most similar atom in their own dictionary. Per-source codes are not well reproducible (median Jaccard at 0.46), and a large portion of the atoms correlate with apparent G magnitude, meaning some instrumental artifacts are left in the atoms. Spectral structures, specifically a set of pre-determined spectral lines, can be recovered from the atoms when they are converted to flux space.

Acknowledgements

I would like to express my most sincere gratitude to my supervisors for their assistance with this project, and for inspiring me to pursue this subject. I would like to thank Dr. Saber Salehkaleybar for his guidance and support, and for his adaptability to the extension of this project. I would like to thank Dr. Anthony Brown for his inspiring and groundbreaking work on the Gaia project, and for helping making this project possible in the first place. Thank you for guiding me through this important step in my academic career.

Contents

1 Introduction

Gaia Data Release 3 contains approximately 1.8 billion sources, of which around 220 million have accompanying low-resolution BP/RP spectra [Gaia Collaboration et al., 2023, De Angeli et al., 2023]. These sources were recorded on a single pair of instruments and reduced through a single calibration pipeline. The published sources were selected on the basis of apparent magnitude ($G < 17.65$) and number of CCD transits contributing to the mean spectra ($N > 15$) alone, without prioritization of any color, spectral type or members of any previous catalog [De Angeli et al., 2023], though the CCD transit count requirement does introduce sampling bias, as the amount of observations depends on the position of the object in the sky. There are no other spectroscopic datasets that are as large and as uniform as Gaia's, which provides a unique research opportunity. Gaia's selection function is comparatively quite simple, which entails structure in the data is more easily attributed to underlying patterns in the population rather than to the way the sample was selected, in comparison to other spectral surveys. There is no other survey that has such a large scale. In comparison, the largest ground based spectral survey is LAMOST, whose latest data release (DR12) contains approximately 12.6 million low-resolution spectra [CAS, 2024]. [1]

Although Gaia DR3 contains an incredible amount of spectroscopy data, its spectra contain a relatively low amount of information in comparison to other spectrographic surveys, which often have considerably higher resolving power, but much lower sample sizes. When combined, BP and RP spectra contain a total of 110 coefficients that are applied to standard basis functions to reconstruct a low-resolution spectrum. Together, these achieve an effective resolution of $R \sim 30 - 100$ for the BP spectrum, and one of $R \sim 70 - 100$ for the RP spectrum, though the lower wavelength side of the RP spectrum overlaps with measurements from the BP spectrum, mitigating the low resolutions to an extent [De Angeli et al., 2023]. Other spectral surveys generally achieve much higher resolutions. For instance, APOGEE has a resolution of approximately $R \sim 22500$ [Majewski et al., 2017], orders of magnitude above Gaia's resolution. In the resolution that Gaia has, most metal lines are unresolvable, and only the broad continuum shape and strongest lines and bands clearly survive, the rest of the features are largely transferred to the shape of the continuum, which is difficult for a human or classical technique to recover [Weiler et al., 2023].

The spectra provided in the Gaia data releases are not detailed, sampled spectra in the form of flux against wavelength, like most other spectral surveys. Instead, each source is described by 110 coefficients of an expansion onto 55 orthonormal Gauss-Hermite basis functions for each Photometer [De Angeli et al., 2023]. These were chosen because of their orthonormal nature, because they can be extended to provide higher detail by adding more bases, and because they decay to zero similarly to a real flux profile. For the published data, this set is then adapted to the data itself, specifically to the singular value decomposition of a set of L_2-normalized calibration spectra, so that most of the signal is concentrated in the low-order coefficients, possibly allowing for higher-order coefficient to be truncated as noise [De Angeli et al., 2023]. Nothing in the construction that was used directly relates a coefficient to a property of a star, although reconstructions on the basis of just the lower-order coefficients inevitably resemble a mean spectrum, because of the way they were determined. Additionally, as the bases are not localized, and rather extend over the entire wavelength space, they do not correspond to a specific spectral region, let alone to specific spectral lines. The consequence of this representation is that the largest homogeneous spectroscopic survey

[1] It also includes approximately 15.5 million medium-resolution spectra, but multiple passes of the same star are counted multiple times for this number.

becomes one of the least directly interpretable. What could partially alleviate this issue is if there was a vocabulary that spectra could be deconstructed into, a set of components that can be used to describe XP spectra as a combination of interpretable parts.

There are broadly two categories of previous analyses of the XP spectra. The first being parameter estimation, and the second being unsupervised representation learning. Neither of these categories find such a vocabulary.

The first of these categories, parameter estimation, includes data-driven models that are trained on external parameters, which now provide more accurate effective temperatures, surface gravities and metallicities for large swathes of the sources of Gaia DR3. Specifically, [Andrae et al., 2023] derives its parameters using an XGBoost algorithm, with APOGEE as training data, [Li et al., 2024] provides a neural network based regression model, also trained using APOGEE data, and [Zhang et al., 2023] trains a data driven forward model, which predicts a spectrum from parameters, and then fits sources to these, using LAMOST as training data. These papers establish that the XP spectra do carry physical information despite their low resolution. In particular, we use [Andrae et al., 2023] to validate our trained models that estimate atom importance in representing certain characteristics in Section 4.5. The published output of these methods, however, is a set of numbers for stellar parameters, meaning all information in the spectrum that does not correlate to the target label is discarded.

The second category is unsupervised representation learning, in which a variety of techniques have been employed. [Laroche and Speagle, 2025] developed a variational autoencoder that generates XP spectra without stellar labels. UMAP has been used to organize white dwarfs into distinct groups, resulting in the recovering of polluted white dwarfs [Kao et al., 2024]. Self Organizing Maps (SOMs) have also previously been applied to the same problem [Pérez-Couto et al., 2024]. These papers provide discoveries, but do not fill the specific gap we are looking for, namely that none of them produces a compositional description of any individual spectrum. A self-organizing map gives prototype spectra, but assigns each source to a single best matching unit, a manifold embedding yields unitless coordinates that are not additive components of a spectrum, and an autoencoder reconstructs a spectrum by passing a dense latent vector through a nonlinear decoder, so the spectrum is never expressed as a sum of identifiable parts.

This thesis will attempt to achieve this vocabulary through sparse dictionary learning. Rather than a small orthogonal basis in which every source uses every component, the aim is to create a large, overcomplete set of atoms, each of which being a set of coefficients that reconstruct an atom spectrum using the original basis functions (see Section 2.5.1), where each spectrum only uses a small subset of atoms [Mairal et al., 2010]. The expectation that a sparse representation of an input produces a set of inspectable components is not new, a previous application by [Olshausen and Field, 1996] produced localized, oriented and bandpass (localized in frequency, as well as space) basis functions resembling receptive fields in the primary visual cortex by imposing sparsity on a linear code for natural images. An important question this thesis asks is whether the same holds for Gaia's low resolution XP spectra.

There are other methods one may consider that fall short in this regard. Firstly, there is principal component analysis (PCA), which yields an orthogonal variance-ordered basis of which the components need not resemble any observed spectrum [Connolly et al., 1995]. Non-negative matrix factorization addresses this by making components additive and parts-based, but does not put any constraint on weights (other than them being non-negative) in standard form, meaning every source will still be composed of every component [Lee and Seung, 1999, Hoyer, 2004]. Archetypal analysis guarantees that each component resembles real data, as the archetypes themselves are mixtures of observed spectra, but represents each source as an interpolation between these entire spectra, rather than a sum of parts [Chan et al., 2003, Cutler and Breiman, 1994].

The solution sparse dictionaries aim for is different. Rather than that every source in a population decomposes into a set of components, it questions whether each given star is roughly composed of only a few identifiable components from a larger overarching set. We will use sparse dictionary learning as a compression model, where one of the objectives is a description length, balancing residual error against code cost, meaning the sparse code is less complex and thus, we hope, more interpretable, possibly revealing useful or notable patterns for future studies. This, to our knowledge, has not yet been applied to Gaia XP spectra.

1.1 Research Questions

RQ1. How compactly can Gaia XP spectra be represented as a sparse combination of learned atoms?

RQ2. Do the learned atoms and codes correspond to known stellar characteristics, when cross-matching against independent catalogs?

RQ3. Are atoms that correspond to stellar characteristics interpretable in their structure when transformed into flux space?

RQ4. Do the atoms or codes reveal clusters of sources that are not yet captured by Gaia or other external labels?

RQ5. Do poorly reconstructed spectra correspond to anomalous or rare objects?

1.2 Thesis Overview

This thesis contains the following sections. Section 1 contains an introduction to the topic and the motivation behind this thesis. Section 2 contains background relating to Gaia and sparse dictionary learning, which will be used in this thesis, and additionally contains some alternative methods which were not chosen for this thesis. Section 3 contains the methodology and setup for the results, and consists mostly of a justification of the chosen dictionary. Section 4 contains the experiments performed on the chosen dictionary and their results. Section **??** discusses future possibilities of extending this research and limitations of this thesis. Section 6 contains conclusions drawn from the experiment results.

2 Background

2.1 The Gaia Mission

Gaia is a space observatory launched by the European Space Agency in 2013, and operated until 2025, which is dedicated to the study of astrometry. Astrometry is a branch of astronomy that involves taking precise measurements of positions and movements of objects in the sky. Gaia is an especially important mission that has provided invaluable insight for astronomers because of the amount of data it has gathered [Gaia Collaboration et al., 2016]. In 2022, Gaia Data Release 3 (DR3) was published, including astrometric observations of over 1.8 billion sources, which mainly consist of stars, but also include solar system objects such as asteroids and comets, other extrasolar sources such as stellar remnants and planetary nebulae, and extragalactic sources such as active galactic nuclei [Gaia Collaboration et al., 2023]. Gaia DR3 not only expands the quantity of the observations, but also improves the data quality from previous missions and previous Gaia data releases [Gaia Collaboration et al., 2023]. Many objects were measured in unprecedented quality, and some objects that were measured by other missions also had data gathered on them. The primary aim of the Gaia mission is to study the formation and evolution of the Milky Way by detecting where objects and mass are situated throughout it [Gaia Collaboration et al., 2016].

2.2 Gaia XP Spectra

Gaia DR3 contains approximately 220 million low resolution XP spectra, which are generated by two of the instruments on board. These instruments are the Blue Photometer (BP) and Red Photometer (RP), which cover different parts of the electromagnetic spectrum. BP covers the wavelengths ranging from 330 to 680 nm and RP covers the wavelengths from 640 to 1050 nm. As Gaia scans the sky, the satellite measures the spectra for each source multiple times. For every scan, a new BP and RP spectrum is measured. Each of these epoch spectra are not directly comparable, as they are recorded at different positions on the CCDs, with varying instrument responses and with differing geometry across the field of view of the instrument. Therefore, processing occurs in stages. The first stage is internal calibration, where every epoch is transformed onto a common reference system so that observations taken under different conditions can be combined [Carrasco et al., 2021]. The calibrated spectra are then combined for each source into a single mean spectrum, and expressed in the previously mentioned basis functions [De Angeli et al., 2023]. External calibration happens on the user side, and it transforms the coefficient-space spectra into absolute flux and wavelength space [Montegriffo et al., 2023, Ruz-Mieres et al., 2025]. The mean spectral coefficients published in DR3, which are the ones used in this thesis, still have instrument response ingrained in them.

As described in Section 1, the published coefficients are expressed on a basis obtained by adapting 55 Gauss-Hermite functions per photometer according to the singular value decomposition of roughly $5 \cdot 10^4$ L_2-normalized calibration spectra. The purpose of this is to concentrate the majority of the information around the low order coefficients, so that the remainder could be dropped as noise if desired [De Angeli et al., 2023]. Since higher order coefficients are considerably more dominated by noise, a component dominated by them is likely an artifact of the representation rather than a real spectral feature. Truncation is not always safe though, as it may distort narrow spectral features in QSOs and strong emission-line sources [De Angeli et al., 2023]. As the Hermite functions were designed to fit typical stellar spectra, correlations between coefficients are low for these generic sources, whereas for other more rare objects or outliers, the basis may represent them somewhat poorly.

2.3 Stellar Labels for XP Sources

2.3.1 Apsis

Gaia DR3 provides its own stellar parameters through the Astrophysical Parameters Inference System (Apsis) [Creevey et al., 2023, Fouesneau et al., 2023]. These are deliberately derived under the restriction that the inference of astrophysical parameters is performed using *only* Gaia data of individual sources [Fouesneau et al., 2023]. For sources with XP spectra, the parameters are therefore in some way a function of the same coefficients this thesis will use to decompose stellar spectra. Their coverage is also uneven, as Apsis is not optimized to derive parameters for white dwarfs (WD), horizontal branch (HB) and asymptotic giant branch (AGB) stars [Fouesneau et al., 2023]. This is not necessarily a deficiency, but rather a scope decision, with the authors noting that more precise estimates are obtainable by cross-matching with external surveys, which is exactly what previously mentioned papers in Section 1 do [Andrae et al., 2023, Zhang et al., 2023, Li et al., 2024].

2.3.2 APOGEE

The Apache Point Observatory Galactic Evolution Experiment provides the independent reference labels which are used in this thesis. APOGEE has collected a half million high-resolution ($R \sim 22{,}500$), high signal-to-noise ratio (> 100), infrared spectra for 146,000 stars [Majewski et al., 2017]. Later data considerably increase these numbers, with DR17 containing $\sim 650{,}000$ stars [Abdurro'uf et al., 2022]. There are two important notes to make about this data. Firstly, the measurement APOGEE makes are independent of Gaia measurements, so if APOGEE labels align with an atom, that is evidence that cannot be explained by shared inputs. Secondly, APOGEE targets a color-selected sample dominated by red giants, so its coverage of a Gaia sample is small and not random. In our sample, 4,583 of $2.046 \cdot 10^6$ sources have corresponding APOGEE labels (see Section 3.1.2).

2.3.3 Data-driven Parameters from XP Spectra

To obtain labels which have reasonable coverage over the sample used, this thesis uses the catalog of [Andrae et al., 2023], who derived data-driven estimates of stellar metallicity for ~ 175 million stars with low-resolution XP spectra, along with T_{eff} and $\log g$. These were derived using the XGBoost algorithm, trained on APOGEE parameters, which means that the resulting parameters are tied to the APOGEE scale. Additionally, the model uses CatWISE magnitudes, along with parallax, which helps to constrain $\log g$ and [M/H] [Andrae et al., 2023]. Atoms corresponding to labels provided by [Andrae et al., 2023] are therefore neither fully independent, nor fully circular, but are compared to more accurate labels than if one were to simply use Gaia's Apsis labels.

2.3.4 Object Classification

Object type labels are taken from SIMBAD [Wenger et al., 2000], which provides classifications taken from published literature. Because SIMBAD records objects that have been subject to prior study, its catalog is non-random. Special star classes are more likely to be overrepresented, while regular field stars are under-represented, as there is a bias towards observing extraordinary objects, meaning they will be overrepresented in any sample that undergoes classification testing using SIMBAD. SIMBAD is therefore used for identifying which classes are over-represented among the sources that contain a specific atom, not for estimating how common a class is.

2.4 Compression & Pattern Mining

2.4.1 Minimum Description Length

The Minimum Description Length principle states that the best model for a dataset is the one that describes it with the shortest description. Data with regularities is data that can be compressed, because a pattern that recurs can be described once rather than describing each of its parts. You can code this by first encoding the model M and then the data given the model, and selecting the $M \in \mathcal{M}$ that minimizes the total $L(M) + L(D \mid M)$, where L is the description length in bits and D is the data. A pattern is only kept if it saves more bits than its implementation into the model costs, so the size of the pattern set follows from the encoding rather than from a custom fixed threshold, meaning that the trade-off between model and pattern complexity regulates itself [Galbrun, 2022].

This framework can be applied to this thesis. A code table is a dictionary, an itemset is a recurring pattern that individual records (in this case, sources) are built from, and the criterion for keeping a pattern is whether it shortens the description of the data. The difference is the data type. Krimp [Vreeken et al., 2011] and related algorithms operate on transactional data, like having items in a supermarket basket, which are either in the basket or not, drawn from a fixed set of items. XP spectra are real-valued vectors (consisting of 110 coefficients), and applying an itemset miner to them would require discretizing every coefficient into bins, which not only discards the metric structure that makes the orthonormal basis meaningful, but is also wildly inefficient. What is needed is a dictionary-based strategy that can apply to continuous data.

2.4.2 Sparse Dictionary Learning as a Compression Model

Sparse dictionary learning does precisely this. The dictionary plays the role of a code table, its atoms are the recurring patterns, and each source is described by the subset of atoms it activated together with their weights. The objective of Equation 3 balances the same two quantities as described before, the residual term measures how much of the source is unexplained, and the ℓ_1 term penalizes the length of the description. A pattern is only retained when the reduction it produces in the residual outweighs the cost of activating it, which is the criterion of Section 2.4.1.

2.5 Sparse Dictionary Learning

2.5.1 Model and Terminology

Sparse dictionary learning represents each data vector as a linear combination of a small number of elements drawn from a learned set. Let $b_s \in \mathbb{R}^N$ be the preprocessed coefficient vector of source s, where $N = 110$. The *dictionary* $D \in \mathbb{R}^{M \times N}$ consists of M *atoms*, each atom d_m being a row of D and therefore a vector in the same space as the data. The *code* $\beta_s \in \mathbb{R}^M$ gives the weight assigned to each atom for source s, so that

$$b_s \approx D^\top \beta_s = \sum_{m=1}^{M} \beta_{sm} d_m \tag{1}$$

The code is required to be sparse, which means most of its entries are exactly zero. We write $s_{\text{eff}} = \|\beta_s\|_0$ for the number of active atoms of a given source, and refer to this as the *sparsity* of the source.

2.5.2 The Optimization Problem

Given a fixed dictionary D, the code of a single source is obtained by solving a penalized least-squares problem,

$$\beta_s = \arg\min_{\beta \in \mathbb{R}^M} \tfrac{1}{2}\|b_s - D^\top \beta\|_2^2 + \lambda \|\beta\|_1 \tag{2}$$

in which the ℓ_1 penalty adds sparsity, shrinking coefficients to 0 [Tibshirani, 1996, Chen et al., 2006]. Learning the dictionary means minimizing the average of this variable over the sample. D requires a constraint for the error to not be trivially reduced by inflating atoms and shrinking codes. Therefore, one may constrain the rows of D, $d_1, \dots d_M$ to have an ℓ_2-norm less than or equal to one [Mairal et al., 2010]. If you take this constraint, and define a convex set $\mathcal{D}$, where $D \in \mathcal{D}$ if it satisfies this constraint, we can use Equation 2 to find the optimal dictionary rather than an optimal source representation.

$$\min_{D \in \mathcal{D}} \frac{1}{S} \sum_{s=1}^{S} \min_{\beta_s} \left(\tfrac{1}{2}\|b_s - D^\top \beta_s\|_2^2 + \lambda \|\beta_s\|_1 \right). \tag{3}$$

Where S is the total source count in the sample. The objective is not convex in D and β when combined, since the reconstruction $D^\top \beta$ multiplies two unknowns. However, it is convex in either of the two when the other is fixed, so algorithms in this family alternate between solving codes and updating the dictionary [Mairal et al., 2010]. This means the algorithms converge to a stationary point of the objective, and different initializations may give different dictionaries as solutions.

2.5.3 Overcompleteness

The dictionary need not form a basis in the usual sense, as the elements of the dictionary are not necessarily independent and the set can be overcomplete (meaning that it has more elements than the signal dimension). In this thesis, we will explore both an overcomplete and a non-overcomplete dictionary. If a dictionary is overcomplete, this automatically means the bases of the dictionary will not be independent, as no set composed of more elements than dimensions can be linearly independent. This is one of the clear distinctions between dictionary learning and orthogonal decompositions such as principal component analysis: the dictionary learning model does not require a basis to be orthogonal, which allows for more flexibility to adapt to the representation of the data [Mairal et al., 2010].

A larger, overcomplete M, in combination with a small s_{eff} allows for specialized atoms, each being used by a subset rather than the entire population and left inactive in the rest. This is contrary to a compact orthogonal basis, which uses the same global directions for every object. The sparsity penalty alone, without other constraints has been shown to be capable yielding inspectable components [Olshausen and Field, 1996, Olshausen and Field, 1997].

2.5.4 Sparse Coding Algorithms

Solving the sparse coding step for a given dictionary can be approached in multiple ways, and two will be addressed here.

OMP Orthogonal Matching Pursuit is a greedy method that selects atoms one at a time, at each step choosing the atom most correlated with the current residual, and re-estimating the weights of all selected atoms by using least squares. It addresses the ℓ_0-constrained problem, rather than Equation 2, and doesn't guarantee it will reach its optimum. It terminates after a fixed, specified number of atoms, which means the sparsity s_{target} is identical for each source [Pati et al., 1993]. This method was considered for the purposes of testing.

LARS-Lasso The LARS-Lasso solves Equation 2 by adding atoms into the code one at a time as the penalty is relaxed from a value large enough to zero every coefficient, down to a chosen λ. This penalty is fixed, while s_{eff} is left to be determined by the data, so that the spectra which are harder to represent using D receive longer atom codes [Efron et al., 2004]. This is the method used to learn the dictionaries recovered in Section 3.

2.5.5 Batch & Online Algorithms

Classical dictionary learning alternates code computation and dictionary updates over the entire dataset, which requires one sparse coding problem per source for each iteration. This is an extremely large task for the sample size we will be using.

Online algorithms instead draw a small mini-batch at each step, sparse-code only that batch, and update the dictionary immediately after, after which the batch is discarded. This is possible because the dictionary update does not require the data itself. Expanding the quadratic term of Equation 3 shows that it depends on all previously processed sources only through two accumulator matrices.

$$A \leftarrow \gamma_t A + \frac{1}{\eta} \sum_{s \in \text{batch}} \beta_s \beta_s^\top, \qquad B \leftarrow \gamma_t B + \frac{1}{\eta} \sum_{s \in \text{batch}} b_s \beta_s^\top. \tag{4}$$

where η is the batch size, and b_s and β_s are the preprocessed spectrum and the code of source s respectively, so that $A \in \mathbb{R}^{M \times M}$ and $B \in \mathbb{R}^{N \times M}$, where N is the size of the basis, in our case $N = 110$, and M is the chosen atom count of dictionary D. These two matrices are sufficient information for the dictionary update rather than using a compressed copy of the sample. When expanding the square in $\sum_s \frac{1}{2} \|b_s - D^\top \beta_s\|_2^2$, which aggregates all the information of Equation 3, A and B appear as sufficient metrics for the update. It should be noted that the algorithm uses earlier and worse dictionaries to compute the code stored in A and B, which means the function is not the empirical cost, but an upper bound on it, and the result here will be slightly worse, though [Mairal et al., 2010] shows that this replacement converges to the same value as the true cost. The algorithm minimizes this instead of the cost itself, which allows for each batch to be discarded once its contribution has been accumulated. This parallels how one might compute an average, by keeping the total count of number instances and their sum, but discarding each individual number [Mairal et al., 2010]. The dictionary is then updated by block coordinate descent, cycling over atoms, keeping all others fixed. Each atom is updated as follows.

$$d_m \leftarrow \frac{1}{A_{mm}} \left(B_{\cdot m} - D^\top A_{\cdot m} \right) + d_m, \qquad d_m \leftarrow \frac{d_m}{\max(1, \|d_m\|_2)} \tag{5}$$

Where $d_m \in \mathbb{R}^N$ is the mth atom, $A_{\cdot,m}$ and $B_{\cdot,m}$ is the mth columns of A and B. The second step is projection onto $\mathcal{D}$.

The first step requires no step size to be chosen. The factor $A_{mm} = \sum_s \beta_{sm}^2$ is the accumulated squared weight with which atom m has been used, so it is large for atoms that are commonly used and small for atoms that are rarely used. Dividing the correction by it means that each atom moves inversely proportionally to the evidence gathered about it. A heavily used atom is barely displaced, and a rarely used atom moves further. Each atom therefore automatically receives its own step size, rather than having it be tuned manually [Mairal et al., 2010]. Atoms that stop being used entirely across iterations are reinitialized from randomly drawn sources, as an atom with $A_{mm} \approx 0$ would otherwise never be updated again. The algorithm passes over the entire sample multiple times, which means that sources will be revisited in later iterations and re-coded against the improved dictionary. The factor $\gamma_t < 1$ discounts the accumulated statistics at each step, because codes computed earlier in training were obtained by using a worse dictionary, and without this factor the artifacts of this poor dictionary will be present forever. At $\gamma_t = 1$, there would be no discounting for past codes. Because A and B have a fixed size (they are formed by summation), the state carried by this algorithm scales as $\mathcal{O}(M^2 + MN)$, independent of S.

2.6 Alternative Decompositions

There are three other linear decompositions we covered earlier in this thesis, and each of these motivate the choice for sparse dictionary learning.

Principal component analysis. PCA finds an orthogonal basis ordered by explained variance, obtained from the eigendecomposition of the sample covariance. It is optimal in least-squares for any fixed number of components, but the representation is dense, meaning that every source has a weight for every component. Additionally, because the components must be orthogonal, they need not resemble any spectral structure, which makes the components difficult to interpret physically [Connolly et al., 1995].

Non-negative matrix factorization. NMF factorizes the data as $X \approx WH$, where both factors are constrained to be non-negative, which is physically natural for spectra, as flux is non-negative. The constraint yields additive components, rather than directions of highest variance [Lee and Seung, 1999]. However, in standard form, non-negativity is the only constraint, which means the weights remain dense, and like in PCA, each source will be described with all components. Sparse variants do exist [Hoyer, 2004], which changes the gap from dictionary learning from sparsity to overcompleteness.

Archetypal analysis. AA represents each source as a convex combination of archetype spectra, which are themselves a combination of observed data [Cutler and Breiman, 1994, Chan et al., 2003]. The components therefore always resemble real spectra, which is not guaranteed for dictionary learning. However, a source will be represented as an interpolation between entire spectra rather than a decomposition to parts.

Sparse dictionary learning deviates from all three approaches by combining the possibility of an overcomplete set with sparsity. The dictionary can contain many more atoms than the signal dimension, and each source may only activate a few of them, so atoms can specialize to subpopulations rather than being forced to describe the entire sample.

2.7 Interpretability of Components

Interpretability requires multiple conditions to be met, and the methods discussed in Section 2.6 provide different subsets of them. We define the following conditions:
Component Interpretability: Whether an individual atom can be inspected on its own and recognized as a spectral element.
Code Interpretability: Whether the description of a single object is short enough to read, rather than containing a weight on a large set of components.
Reconstruction Accuracy: Whether the representation actually reconstructs the data, so that what is being interpreted is a proper reconstruction of the spectrum.

Principal component analysis provides reconstruction accuracy, but does not satisfy either other condition. Non-negative matrix factorization has interpretable components, but leaves its code dense, dropping code interpretability. Manifold embeddings provide neither components, nor a reconstruction. A sparse dictionary can provide all three, as atoms are also in spectral space, each source displays only a small subset of them, and the reconstruction error is bounded by the objective.

Reconstruction accuracy can be directly measured, and code interpretability follows from the sparsity of the code, but component interpretability is harder to judge. Given an atom, almost any structure can be described as interpretable in some manner if one attempts to find this interpretability hard enough. Criteria must therefore be fixed in advance, rather than derived from the atoms that have already been learned.

We define the following metrics for component interpretability:

Plausibility. The atom, transformed into flux space[2], is smooth, rather than a set of constant, overwhelming oscillations. As mentioned in Section 2.2, high-order coefficients are dominated by noise, which will be further shown in Figure 2 in Section 3, so an atom composed only of these coefficients likely has the same property.

Stability. Whether a component can be identified in different random initializations. The objective is not jointly convex, so an atom that appears in one solution need not appear in another.

Distinctness. Whether an atom is sufficiently unique in a dictionary. If atoms are highly similar, the encoder's choice between them can be almost arbitrary, meaning the interpretability between both components is low.

Non-instrumentality. A component that is dominated by instrumental errors is not interpretable, as it does not correspond to astrophysical properties, but to instrumental ones, therefore not revealing any properties of the source it belongs to. Though this criterion is required for the interpretation of physical properties that the sources whose code contains this atom share, it may still be useful to retrieve atoms that contain isolated instrumental spectral components, as a means to improve the understanding of instrument behavior and data processing artifacts.

[2] The sampled space with wavelength on the x-axis and flux on the y-axis, where flux is the amount of incoming photons of a specific wavelength, per unit time and area.

Correspondence. The atom aligns with a spectral feature that is resolvable by Gaia, or the sources encoded by it concentrate around a specific region of parameter space, or consist largely of a known object class. If this criterion is missing for a subset of atoms, whereas other atoms have clear correspondence, this may point to — but certainly not guarantee — the fact that this subset contains an unknown class or cluster of objects that share an attribute.

2.8 Limitations - Instrument Errors

An important limitation with Gaia's XP spectra is the fact that they carry systematic errors that are not random with respect to source properties. Data quality varies with magnitude, as dimmer observed stars simply have less photons gathered for them, meaning that spectral errors are exacerbated, including for stars that are more distant. On the other hand, there is also evidence for imperfect calibration for brighter spectra with $G < 11$. The highest-quality spectra lie in the range $9 < G < 12$. Additionally, due to the fact that noise is correlated over long ranges within a spectrum, sampling in pseudo-wavelength may produce oscillations in the spectra [De Angeli et al., 2023]. Subsequent work has shown that flux errors depend on color, on magnitude and on reddening, and has derived corrections to account for it [Huang et al., 2024].

This has important implications for this thesis, as a compression model has no way to distinguish between structure that recurs because of stellar physics or because of observational systematics. It may, however, group astrophysical and instrumental structures differently. Systematic corrections to the XP fluxes by [Huang et al., 2024] were not applied. This is because the atoms and classes recovered in Section 4 are distinguished mostly by spectral morphology and shape rather than small calibration errors.

3 Methodology

3.1 Data Acquisition

3.1.1 Gaia Spectra

To start, we must first acquire the correct data, largely from the Gaia DR3 Catalog. The table that contains the most important spectral data for this thesis is `XP_CONTINUOUS_MEAN_SPECTRUM`, which contains the Hermite coefficients and the errors of the sources by `source_id` [Gaia Data Release 3, 2024].

3.1.2 Sampling

Gaia contains an extremely large amount of data, the spectra of the approximately 220 million sources that have had measurements of their spectra taken take up multiple terabytes of disk space. For that reason, Gaia partitions its sample using 3386 ranges of HEALPix level-8 indices [Gaia Data Release 3 Documentation, 2023]. In essence, HEALPix is a coordinate system for the location of stars in the night sky [Gorski et al., 2005]. For this thesis, a sample of approximately 1% of the total spectral sources is taken from Gaia DR3. Initially, the first approximately 500 thousand sources were taken as a subsample from the `XP_CONTINUOUS_MEAN_SPECTRUM` table of Gaia DR3, however, this is a biased sample as it contains only a specific section of the night sky. Stellar populations differ between the Milky Ways galactic disk and other parts of the night sky, which would overrepresent certain types of stars and underrepresent others. Therefore, the data that was used for this thesis will consist of a strided sample taken from Gaia DR3, containing a representative population from many differing sections of the night sky. Specifically, a sample of $2.046 \cdot 10^6$ sources was taken by taking 31 partitions from Gaia DR3s `XP_CONTINUOUS_MEAN_SPECTRUM` table. This sample accounts for approximately 1% of all XP spectral sources in Gaia DR3, which is large enough that the learned dictionary is not sample-limited while remaining manageable for parameter sweeps and learning.

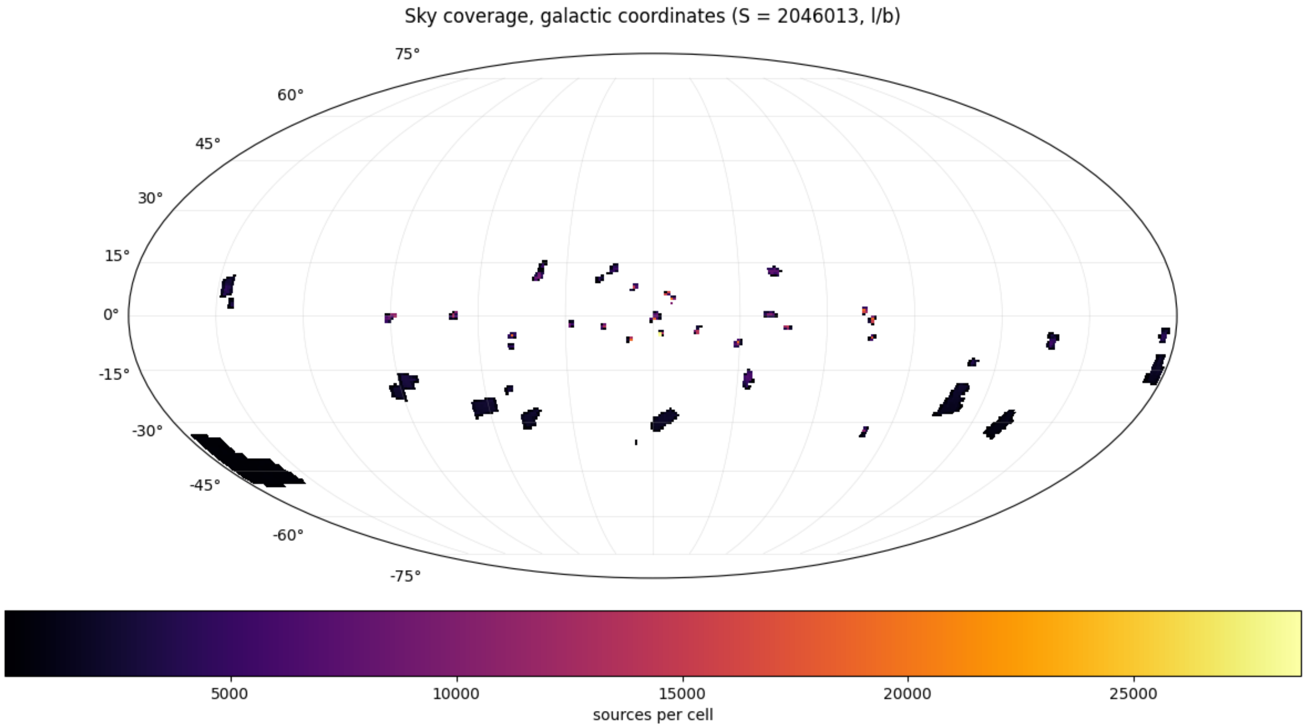

Figure 1: The sky coverage of the sample, showing that the partitions are distributed across both the galactic plane and the rest of the sky, distributed using population, rather than their position in the sky, which is why partitions in the galactic plane are smaller, as they contain more stars in a smaller sky patch. The sample is not precisely uniformly distributed, but since we are not attempting to predict class prevalence in a population, this should not be an issue.

3.1.3 Other Data & Sources

In addition to the Gaia spectra from `XP_CONTINUOUS_MEAN_SPECTRUM`, other parameters, characteristics and classes will be used in this thesis. Firstly, Gaia contains data about astrophysical parameters in the table `ASTROPHYSICAL_PARAMETERS` in the Astrophysical Parameters section of Gaia DR3. Specifically, the metallicity (`mh`), surface gravity (`logg`) and surface temperature (`teff`) labels will be used.

There are also some external datasets that will be used in this thesis, namely to provide more accurate data for characteristic prediction and training, and for sample and atom diagnostics. The first of these external datasets is APOGEE [Majewski et al., 2017], which will be used mainly as a verification set for training models for the earlier mentioned `mh`, `logg` and `teff` labels. However, this still leaves an issue, as the APOGEE dataset is quite small and has a low coverage of the sample. For the previously discussed sample of $2.046 \cdot 10^6$ sources, only 4,583 have APOGEE labels, giving a coverage of approximately 0.22%. To train prediction models on for determining atom importances, we require more data, in particular data that has a higher overlap with our sample than APOGEE. The dataset we will use to train these prediction models is obtained from [Andrae et al., 2023], who use a data driven approach to determine `mh`, `logg` and `teff` for about ~ 175 million stars that have accompanying low resolution XP spectra in Gaia DR3. These values were derived using the XGBoost algorithm and trained on the stellar parameters from APOGEE. This dataset improves the coverage of the accurately labeled stars from 0.22% to 77.2%. It is important to mention that many non-main sequence stars will not appear in these labels, and do not have any of these astrophysical parameters in Gaia DR3 either. The main focus for these objects will be using the atoms to determine their class, and attempting to discover atoms that are important in doing so, together with finding which atoms these classes are highly overrepresented in, rather than determining (and finding atom importance/clustering) for the previously mentioned astrophysical parameters.

Lastly, SIMBAD object classes are used for classification prediction models and for atom diagnostics, specifically to determine whether they are dominated by certain object classes. An important note for this is that only approximately 3.8% (76,766) of the stars in the sample have a corresponding label in SIMBAD. This is because SIMBAD only includes previously observed objects, which means that "extraordinary" objects are more likely to be represented, as there is a bias towards observing them. That means non-background star classes will be overrepresented in any sample that undergoes classification testing using SIMBAD [Wenger et al., 2000].

3.1.4 Sample Population

To provide a diagnostic for the source types present in the gathered sample, matching was performed to find each source in the Gaia sample in the SIMBAD catalog, by cross-identifying the Gaia `source_id` with the id given by SIMBAD. The following table contains a summary of the SIMBAD `otype` labels of all sources in the sample when matched to SIMBADs catalog [CDS, 2026]:

Type	Name	Count	Fraction
*	Star	41843	0.545072
EB*	Eclipsing binary	7164	0.093323
LP*	Long-period variable star	6801	0.088594
LP?	Long-period variable candidate	5393	0.070252
PM*	High proper-motion star	2901	0.037790
SB*	Spectroscopic binary	2016	0.026262
Em*	Emission-line star	1973	0.025701
RR*	RR Lyrae variable	1111	0.014473
RG*	Red giant branch star	1044	0.013600
WD*	White dwarf	943	0.012284
Mi*	Mira variable	643	0.008376
Y*?	Young stellar object candidate	494	0.006435
V*	Variable star	482	0.006279
AGN	Active galactic nucleus	462	0.006018
Pe*	Chemically peculiar star	386	0.005028
AB?	Asymptotic giant branch star cand.	290	0.003778
HB*	Horizontal branch star	277	0.003608
Pu*	Pulsating variable star	253	0.003296
RS*	RS Canum Venaticorum variable	219	0.002853
BY*	BY Draconis variable	217	0.002827
C*	Carbon star	197	0.002566
LM*	Low-mass star	185	0.002410
QSO	Quasar	152	0.001980
HS?	Hot subdwarf candidate	151	0.001967
**	Double or multiple star	126	0.001641
El*	Ellipsoidal variable	125	0.001628
cC*	Classical Cepheid variable	114	0.001485
dS*	Delta Scuti variable	81	0.001055
gD*	Gamma Doradus variable	63	0.000821
s*r	Red supergiant	63	0.000821
HS*	Hot subdwarf	45	0.000586
G	Galaxy	44	0.000573
AB*	Asymptotic giant branch star	34	0.000443
HB?	Horizontal branch star cand.	33	0.000430
EB?	Eclipsing binary candidate	32	0.000417
Ev*	Evolved star	23	0.000300
Be*	Be star	20	0.000261
Ro*	Rotationally variable star	20	0.000261
s*b	Blue supergiant	19	0.000248
WR*	Wolf-Rayet star	19	0.000248
WV*	Type II Cepheid (W Vir) variable	17	0.000221
a2*	Alpha2 CVn variable	13	0.000169
HV*	High-velocity star	13	0.000169
S*?	S star candidate	11	0.000143
BS*	Blue straggler	10	0.000130
RR?	RR Lyrae variable candidate	10	0.000130
Sy1	Seyfert 1 galaxy	10	0.000130
X	X-ray source	9	0.000117
NIR	Near-infrared source	9	0.000117
GiG	Galaxy towards group	9	0.000117
PN	Planetary nebula	9	0.000117
BLL	BL Lac object	8	0.000104
OH*	OH/IR star	8	0.000104
Sy2	Seyfert 2 galaxy	8	0.000104
pA?	Post-AGB star candidate	8	0.000104
EmG	Emission-line galaxy	7	0.000091
s?r	Red supergiant candidate	7	0.000091
blu	Blue object	6	0.000078
WD?	White dwarf candidate	6	0.000078
AG?	AGN candidate	6	0.000078
TT*	T Tauri star	6	0.000078
C*?	Carbon star candidate	6	0.000078
Bla	Blazar	6	0.000078
sg*	Evolved supergiant	6	0.000078
SX*	SX Phoenicis variable	5	0.000065
rG	Radio galaxy	5	0.000065
BS?	Blue straggler candidate	5	0.000065
S*	S star	5	0.000065
Or*	Variable star of Orion type	5	0.000065
Q?	Quasar candidate	5	0.000065
UV	UV-emission source	5	0.000065
RB?	Red giant branch star candidate	4	0.000052
HXB	High-mass X-ray binary	4	0.000052
CV*	Cataclysmic variable star	4	0.000052
Y*O	Young stellar object	4	0.000052
RV*	RV Tauri variable	4	0.000052
Ce*	Cepheid variable	4	0.000052
pA*	Post-AGB star	3	0.000039
Ir*	Irregular variable star	3	0.000039
BD*	Brown dwarf	3	0.000039
SB?	Spectroscopic binary candidate	3	0.000039
Er*	Eruptive variable star	3	0.000039
WV?	Type II Cepheid candidate	3	0.000039
GiP	Galaxy in a pair of galaxies	2	0.000026
bC*	Beta Cephei variable	2	0.000026
s?y	Yellow supergiant candidate	2	0.000026
FIR	Far-infrared source	2	0.000026
IR	Infrared source	2	0.000026
s*y	Yellow supergiant	2	0.000026
GiC	Galaxy towards cluster	1	0.000013
LIN	LINER-type active nucleus	1	0.000013
V*?	Variable star candidate	1	0.000013
GNe	Galactic nebula	1	0.000013
Psr	Pulsar	1	0.000013
TT?	T Tauri star candidate	1	0.000013
HII	HII region	1	0.000013
HX?	High-mass X-ray binary candidate	1	0.000013
Rad	Radio source	1	0.000013
C?G	Cluster of galaxies candidate	1	0.000013
RC*	R CrB-type variable star	1	0.000013

Table 1: Total count of each represented SIMBAD object type in our sample of 2,046,013 sources.

It is important to note that some of the objects listed in this table (and thus present in the Gaia sample used) are objects that are not the central focus of this thesis, as they are not stellar scale objects, which include stars, young stellar objects, brown dwarfs, degenerate stars and black holes. These non-stellar scale objects account for about 0.995% of all objects identified by SIMBAD. It is important to mention that special objects (not *) are likely highly overrepresented when compared to the actual Gaia sample, as these objects are logged more often than seemingly "uninteresting" background stars. The objects that do not fit in any of these categories are non-stellar scale intergalactic objects (AGN, QSO, G, Sy1, GiG, Sy2, EmG, AG?, Bla, rG, Q?, GiP, GiC, LIN, GNe, C?G), generic objects defined by their spectral output (X, NIR, blu, UV, FIR, IR, Rad) and emission nebulae (PN, HII), which predominantly have scales between stellar and galactic.

Object type	**Count**	**Total population fraction**	**SIMBAD population* fraction**
Non-SIMBAD classified	1969247	0.9625	N/A
Stellar-Scale	76002	0.03715	0.9900
Non-SS intergalactic	720	$3.519 \cdot 10^{-4}$	$9.379 \cdot 10^{-3}$
Generic	34	$1.662 \cdot 10^{-5}$	$4.429 \cdot 10^{-4}$
Emission nebulae	10	$4.888 \cdot 10^{-6}$	$1.303 \cdot 10^{-4}$

Table 2: Count of overarching object types in the sample, where Non-Stellar-Scale intergalactic objects consist of AGN, QSO, G, Sy1, GiG, Sy2, EmG, AG?, Bla, rG, Q?, GiP, GiC, LIN, GNe and C?G, generic objects consist of X, NIR, blu, UV, FIR, IR and Rad and emission nebulae consist of PN and HII, with full type labels given in Table 1.

3.2 Data Preprocessing

3.2.1 Coefficient Selection

The low resolution XP spectra require some preprocessing to be appropriate for the specific use cases and methods that were applied to them in this thesis. First of all, it is important to note that the data that will be used to represent these spectra is the 110 coefficients of each source that were obtained by Gaia's photometers. This method was chosen to avoid loss of any data that was provided by Gaia, though most higher-order coefficients are extremely noisy, having signal-to-noise ratios barely above 1 (See Figure 2).

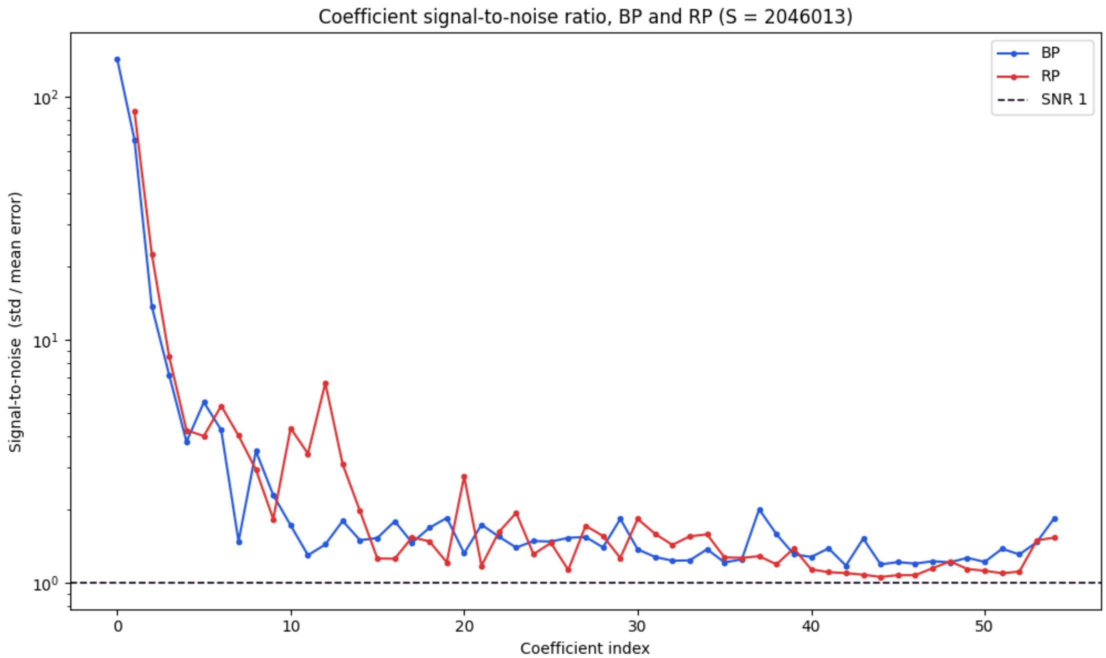

Figure 2: Signal-to-noise ratio of each BP and RP coefficient, using the mean error across the sample for each coefficient. A horizontal dashed line is shown for $SNR = 1$. The tail end of the graph shows that many higher-order coefficients have a signal-to-noise ratio barely exceeding 1, showing that information is indeed concentrated heavily around the lower-order coefficients [De Angeli et al., 2023].

3.2.2 Amplitude Normalization

Amplitude normalization is performed on each BP/RP spectrum that is used for dictionary learning. The reason for this normalization is that the characteristic we want to investigate is the spectral shape, and not normalizing the spectra would mean spectral amplitude would dominate the error optimization of dictionary learning algorithms. Additionally, were the spectral amplitudes not normalized, distance of the source would change the spectral amplitude.

Let c_{sn} be the raw coefficient n of source s, with $s = 1 \dots S$ ($S = 2046013$) and $n = 1 \dots 110$ (55 BP, 55 RP). The amplitude normalization is performed by dividing both the BP and RP coefficients by `rp_00` of the same source s. With this, `rp_00` is always equal to 1, meaning we can drop this coefficient.

$$x_{sn} = \frac{c_{sn}}{c_{sn_0}}, \qquad n_0 = 56, \tag{6}$$

where `rp_00` is the 56th coefficient. This also means that two identical sources at different distances will be normalized to the same spectral shape.

$$x'_{sn} = a x_{sn} = \frac{a c_{sn}}{a c_{sn_0}} = \frac{c_{sn}}{c_{sn_0}} = x_{sn}. \tag{7}$$

This method was specifically used because this is the least granular coefficient, and the one that is most representative of the entire spectrum, in the red section of the spectrum, which means that the size of this coefficients correlates directly with the overall amplitude of the spectrum, rather than specific structures located within it. Some tests were performed to see whether `rp_00` indeed makes a good candidate for a normalization variable. Another candidate that was considered is `bp_00`.

Metric	**bp_00**	**rp_00**
Normalization coefficient n_0		
Index	[0]	[55]
Min	-1.475	7.875
Negatives	290	0
$\lvert n_0 \rvert < 1$	960	0
$\lvert n_0 \rvert < 0.1$	123	0
Scaled $\lvert\beta\rvert_{\max}$		
Value	$1.429 \cdot 10^8$	1675
Target mass		
Top source	38.7%	0.0%
Top 100	98.3%	0.1%
Top 1000	99.9%	0.6%
Rows holding half of target mass	**2**	539,962

Table 3: Diagnostic statistics for `bp_00` and `rp_00` (out of 2,046,013 sources).

However, as seen Table 3, `bp_00` has many pathologies that disqualify it from being a good normalization factor. Some spectra have negative `bp_00`, many have coefficients where $bp_0 < 1$ or even $bp_0 < 0.1$. In this table, "target mass" refers to the amount of total coefficient value that is situated in a specific part of the sample, some stars have values of `bp_00` that are extremely close to 0 (the lowest being $bp_0 \approx 0.003$), which results in their target mass inflating to large portions of the total mass of the sample. Due to this large target mass, atoms tend to point in the direction of these "target massive" outliers (in this context, direction refers to the direction of an atom or vector when considering the 109-coefficient space, where the atoms tend to point along the directions with the greatest variance). Lastly, the `bp_00` coefficient is more directly tied to the color and temperature of a star rather than its luminosity, the latter of which being what we want to account for by normalizing in the first place. The `rp_00` for each source is stored for later use to retrieve the original spectra.

After this step was performed, the BP and RP coefficients are concatenated to create a set of 109 coefficients for each source, dropping `rp_00` due to it always being 1 after normalization. This step was performed to combine all of the stellar parameters into one representative set for each source that can be used by dictionary learning.

3.2.3 Per-coefficient Scaling

Squared error treats every coordinate identically, but the 110 coefficients do not carry comparable amounts of information. The low-order Hermite coefficients describe the broad continuum shape and are large, while the higher-order coefficients describe the finer structure and are small. Without any rescaling the objective is decided almost entirely by whichever coefficients are largest.

We considered two candidate scalers. The standard deviation of each coefficient across the sample, and the mean of Gaia's own published measurement uncertainty for each coefficient. A comparison is found in Table 4

Scaler	Min	Max	Max/Min	Max/Median
Standard deviation	$1.65 \cdot 10^{-4}$ (`bp_54`)	$2.64 \cdot 10^{-1}$ (`bp_00`)	1600.6	150.2
Measurement error	$8.95 \cdot 10^{-5}$ (`bp_54`)	$3.49 \cdot 10^{-3}$ (`bp_07`)	39.0	2.9

Table 4: Dynamic range of the two candidate per-coefficient scalers over 2,046,013 sources. The constant coefficient `rp_00`, divided out by the amplitude normalization, is excluded from both rows, as it has zero spread, though it retains a measurement uncertainty of $1.59 \cdot 10^{-3}$.

The standard deviation is largest at `bp_00`, and stretches a factor of 1,600 to equalize the coefficients. The measurement error on the other hand, peaks at `bp_07`, and varies far less. This makes it an important noise reference. While the spread of a coefficient across the sample mixes astrophysical variation with measurement noise, the published uncertainty measures the noise alone.

Scaling by the standard deviation gives every coefficient $\sigma = 1$, and therefore equal weight in the objective regardless of how well it was measured or how much information it contains. This counteracts Gaia's deliberate concentration of information around the lower order coefficients. They are bigger in magnitude, but they also contain much more information than higher-order coefficients [De Angeli et al., 2023]. Scaling by standard deviation would amplify coefficients whose variation is mostly noise, such as `rp_44`, which has a signal-to-noise ratio barely above 1, to have the same weight as `bp_00`, which has a signal to noise ratio of almost 200.

Scaling by the measurement error instead leaves every coefficient with a spread equal to this signal-to-noise ratio, meaning a coefficient that is measured at $SNR = x$ will have a spread x times larger than the spread of a coefficient at the noise floor. The residuals are now measured in units of mean measurement uncertainty, meaning the squared objective roughly approximates a χ^2 statistic, since correlation on coefficients is minimized. This means coefficients contribute proportionally to the information they contain. This is the scaling we will continue to use in this thesis.

3.3 Sparse Dictionary Learning

In this thesis, we will be using `scikitlearn`s function `MiniBatchDictionaryLearning`, as it provides a way to perform dictionary learning more efficiently on extremely large samples. Regular dictionary learning sparse-codes the entire matrix, updates dictionary D, and then repeats, which is excessively slow for a sample as large as $2.046 \cdot 10^6$ sources. This technique is something we discussed in Section 2.5.5.

The algorithm performs multiple passes over the given sample. Sources are revisited in later epochs and re-coded against the improved dictionary, with `max_iter` bounding the number of passes if none of the other stop conditions are met. These other conditions are, firstly, when the change in D between consecutive steps falls below `tol`, and secondly when an exponentially weighted average of the batch cost fails to improve for `max_no_improvement` consecutive steps. For this thesis, we use `max_iter = 20`, which has never resulted in early stopping, `tol = 0.01` and `max_no_improvement = None`.

A batch size of 2^{12} or 4096 was used, giving $\lceil 2.046 \cdot 10^6/2^{12} \rceil = 500$ dictionary updates for each pass over the entire sample. The batches are shuffled before formation so that the sources contributing to each update are drawn from every sampled region of the sky rather than a single (or two adjacent) HEALPix regions (Section 3.1.2).

3.3.1 Dictionary Size & Lambda Selection

We will perform a sweep to determine the optimal M and λ for our dictionary, given the criteria that will follow in Table 5. The sweep we use will span $\lambda \in [1.00, 100]$, using 10 logarithmically spaced values. These values were picked after attempting preliminary runs at 0.35 and 350. The former produced codes far denser than the measurement noise justifies, whereas the latter suppressed nearly all structure. We will be performing this sweep using $M \in \{32, 64, 128, 256, 512, 1024\}$, with each configuration fitted on a sample of $3 \cdot 10^5$ stars for each dictionary config in the sweep, using $2 \cdot 10^5$ for training and $1 \cdot 10^5$ for testing.

The following criteria will determine which dictionary will be used in this thesis.

Criterion	Description	Reason
Zero-fraction	The fraction of spectral codes that are completely zeroed out.	A significant fraction of empty codes means the dictionary is not representing the sample.
Dead atoms	The fraction of atoms that are unused in the test set.	Unused atoms means the data cannot support this many atoms. M is too large.
$\chi^2/dof \approx 1$	$\chi^2 = \sum_n \frac{r_n^2}{\sigma_n^2}$, then divide by $N-1$, the degrees of freedom (dof), as there are 109 data directions per source.	Shows whether the dictionary is fitting to the signal rather than to the noise, if $\chi^2/dof > 1$, we are missing information, if $\chi^2/dof < 1$, we are fitting to noise. Note that the σ used here is of the population mean, which does not affect this metric much, but does affect individual sources.
$\epsilon_{test} \approx \epsilon_{train}$	Whether a test set has the same error as the training set.	If the error for the train set is considerably lower than for the test set, the dictionary is overfitting for the train set.
Bit count	$L(M) + L(D \mid M)$, the total bit data to store the model and code, not including the residual.	See Section 2.4.1.
Median compression ratio	The median of the ratio between the bits the compressed spectrum takes to store and the bits the raw spectrum takes to store. This does not include the residual of a source.	This is a good number to diagnose how well the algorithm compresses a regular spectrum. The fact the residual is not included means this is a lossy compression, though median reconstruction accuracy is up to measurement uncertainty. The reason we do not store the residual is because storing it to any usefully small quantization would inflate the code dramatically. Additionally, the original measurements are also not a ground truth, and contain an error as well, so comparing these when storing the residual would not be a completely fair comparison. Without storing the residual, a dictionary with all zeroed-out atoms would give a perfect compression. We resolve this by selecting a dictionary at $\chi^2 \approx dof$.
Median/p75 MCCS	MCCS stands for mean-centered cosine similarity: The cosine similarity between the structures of each atom, with the mean direction of each atom removed.	A useful metric to determine how many atoms are essentially duplicates of each other. A low MCCS is not necessarily better, as PCA has a cosine similarity of 0 for all components, but does not have components which resembled observed spectra.
Refit stability	When optimally one-to-one matched between dictionary learning instances called with a different `random_state`, take the median MCCS between these atoms, and test whether it is a larger value than the median MCCS between the most similar atom to any other atom. Tested with 10 different seeds, where all combinations are averaged.	This metric determines whether the atoms have actual meaning, or whether they are simply one efficient deconstruction of many. It simultaneously shows whether there is a high redundancy of atoms, as MCCS between twin atoms in a single dictionary and MCCS between matched atoms in different generations both contribute.
Usage dynamic range	Min/max atom usage over all sources.	Determines how many generalist and specialist atoms there are. Dictionaries with low M tend to only have generalist atoms, which is bad for interpretability.
Distinct classes captured	A count of object types for which at least one atom has a specialization $\zeta > 0.05$. (See Equation 17)	Shows how many atoms there are that specialize in specific SIMBAD object types, shown in Table 1 [CDS, 2026].
Best atom label selectivity	Weighted interquartile range of the cluster in the label in comparison to the population IQR for labels T_{eff}, $\log g$ and $[M/H]$. (See Equation 18	Shows whether the dictionary has atoms that specialize for certain values of stellar properties.

Table 5: Criteria for successful dictionaries, split into five categories: General validity, reconstruction, compression, distinctness and specialization.

For these criteria, the median value will be taken across 10 different `random_state` values. A random seed will be picked to determine the dictionary used in this thesis, which will have its individual value reported as well. From the sweep, we determine that no dictionaries have a zero-fraction. Additionally, dead atoms only start appearing in dictionaries with a high λ, resulting in a very sparse dictionary. Specifically, dead atoms start appearing at $\lambda \geq 35.9$. Train and test error are also almost the same in every dictionary in the sweep, but the difference increases slightly with increasing dictionary size, maximizing at $\sim 2\%$.

The following shows whether the dictionary starts fitting to noise, by plotting χ^2/dof.

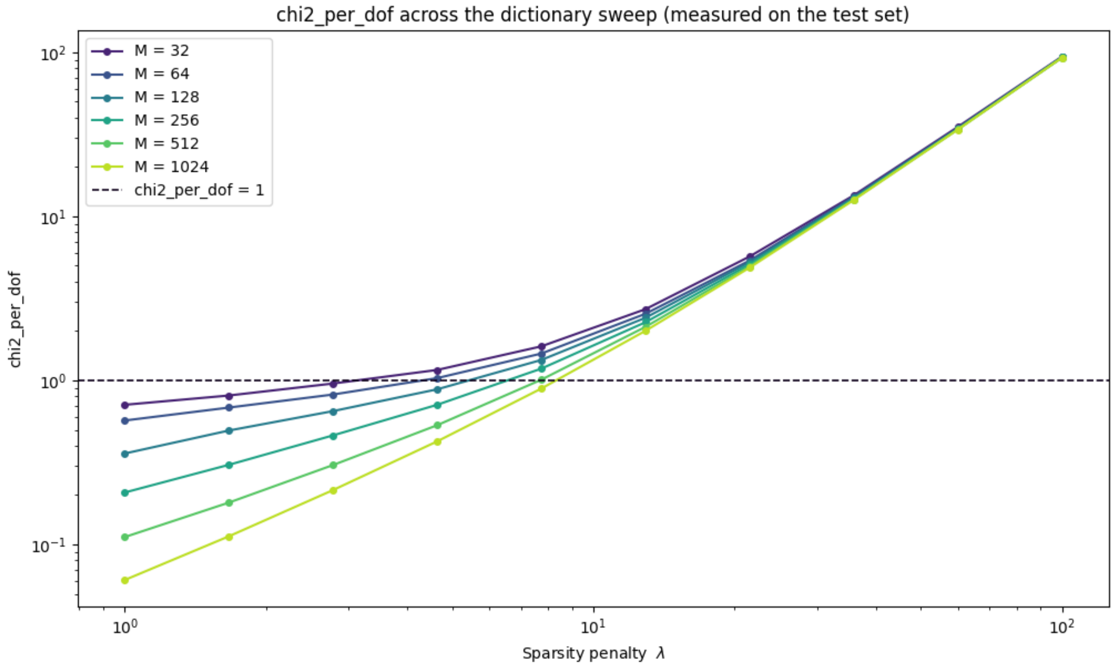

Figure 3: χ^2/dof, showing whether dictionaries fit to noise or to actual information. An optimal amount if given by $\chi^2 \approx 1$, where the residuals are approximately equal to the expected residuals if they were only variance, meaning no signal was missed and no noise was coded. The λ values that lie close to the $\chi^2 \approx 1$ boundary will be used for later sweeps (for instance, there will be no reasonable M for which $\lambda = 30$ will result in a reasonable χ^2).

The next criterion is one that comes from the minimum description length principle from Section 2.4.1, which is how many bits the code and model take up combined. We can graph this as well to find the minimum description length of the data. To do this, we need to know the bit size of the data and dictionary, which we can compute as follows:

$$L_s = L_{s,\text{support}} + L_{s,\text{weights}} + L_{s,\text{residual}}, \qquad L = L_{\text{dictionary}} + \sum_s L_s \tag{8}$$

Note how the second equation mirrors the equation of Section 2.4.1. We will now further express these bit lengths. Firstly, we will express the support of the code of each spectrum s. We encode the count of the atoms that are activated for a coded spectrum. Suppose an atom count M in the dictionary, this means that there are $M+1$ possible total active atoms for each coded spectrum (anywhere from 0 to M). Secondly, we need to encode which atoms are activated. There are $\binom{M}{s_{\text{eff}}}$ possible combinations of atoms, which means the information required is $\log_2 \binom{M}{s_{\text{eff}}}$. We can combine these two to get the bit length of the support:

$$L_{s,\text{support}} = \log_2(M+1) + \log_2 \binom{M}{s_{\text{eff}}} \tag{9}$$

Secondly, we need to express the weights of these activated atoms. An important step here is quantization, where we determine the precision at which each weight is stored. We name this quantization q. After this, we determine the minimum and maximum coefficient values over the entire dataset. By computing $2|\beta|_{max}$ we find the total amount that the coefficients vary, and upon dividing by q_β, we receive the total amount of possible values a coefficient can have. After multiplying by the atom count of the spectrum, this gives us Equation 10.

$$L_{s,\text{weights}} = s_{\text{eff}} \left(\log_2 \left(\frac{2|\beta|_{max}}{q_\beta} + 1 \right) \right) \tag{10}$$

Thirdly, now that we have encoded which atoms are active for each spectrum and by how much they are activated, we can encode the residual. Here, we assume the residual is Gaussian noise. For a Gaussian residual with unit variance, the following equation holds.

$$p(r) = \frac{1}{\sqrt{2\pi}} e^{-r^2/2} \tag{11}$$

And the information needed to describe a residual is its negative log probability, given by $-\log_2 p(r)$. We also must quantize this value to provide a finite-precision bit code. To do this, we ask what bin the residual falls into, with bin width q_r. We can therefore say the information required is $-\log_2(p(r)q_r)$, which we can rewrite as $-\log_2 p(r) - \log_2 q_r$, the latter term we can drop when comparing dictionaries, as it is constant, and we can also drop for our compression ratio, as we do not keep the residual for a lossless compression, rather compressing to measurement noise.

$$-\log_2 p(r) = -\log_2 \left(\frac{1}{\sqrt{2\pi}} e^{-r^2/2} \right) = -\log_2 \left(\frac{1}{\sqrt{2\pi}} \right) - \log_2(e^{-r^2/2}). \tag{12}$$

If we drop the first term, which is constant: $\frac{1}{2}\log_2(2\pi)$, we get $\frac{r^2}{2\ln 2}$. This is possible in our comparison, as it is the same for each dictionary. In our case, $\chi^2 = \sum_{n=1}^{N} r^2$, as r^2 is already divided by the measurement uncertainty σ_n due to our normalization.

$$L_{s,\text{residual}} = \frac{r^2}{2\ln 2}, \qquad L_{\text{residual}} = \frac{\chi^2}{2\ln 2} \tag{13}$$

Lastly, we need to encode the dictionary itself. We only need to do this once, rather than once for every source. A dictionary needs to encode each atom M, which all consist of a combination of the original basis functions. This means we need to store $M \cdot N$ numbers in the dictionary, each with another quantized coefficient. We will call this quantization q_b. Again, we must compute how many possible values the coefficient can take by computing $\frac{2|b|_{max}}{q_b}$. This gives us the bit length equation that follows.

$$L_{\text{dictionary}} = MN \left(\log_2 \left(\frac{2|b|_{max}}{q_b} + 1 \right) \right) \tag{14}$$

This means the total bit length of a dictionary with its code representation of the sources is as follows:

$$\begin{aligned} L = \sum_s \Bigg(\log_2(M+1) + \log_2 \binom{M}{s_{\text{eff}}} + s_{\text{eff}} \left(\log_2 \left(\frac{2|\beta|_{max}}{q_\beta} + 1 \right) \right) \Bigg) \\ + \frac{\chi^2}{2\ln 2} + MN \left(\log_2 \left(\frac{2|b|_{max}}{q_b} + 1 \right) \right) \end{aligned} \tag{15}$$

We use $q_b = 0.001$ and $q_\beta = 0.1$ as quantum sizes for b and β coefficients. These quantization steps were selected as the coarsest values for which the quantization increased the reconstruction error by no more than 0.1% above the baseline reconstruction error. Candidate quantization sizes were selected by direct reconstruction. We will additionally run a more localized sweep, where we focus on dictionaries where $\chi^2/dof \approx 1$, thus limiting $\lambda \in [2.00, 10.00]$.

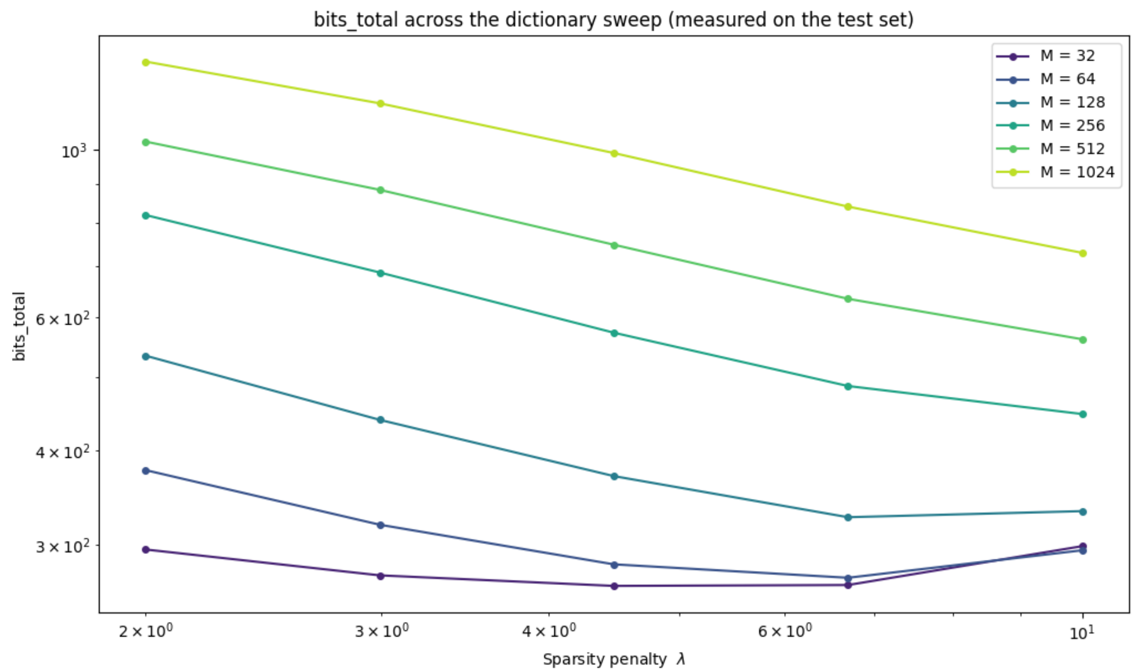

Figure 4: The y-axis shows the total bits **per source**, including the dictionary cost when divided by the sample size of $2.046 \cdot 10^6$, but not including the two residual constants. This figure shows the total bit count for models with $M \in \{32, 64, 128, 256, 512, 1024\}$. It shows a minimum around $\lambda = 5.0, M = 32$. However, since this is the smallest model, this lies on the edge of the sweep, meaning we will perform an additional sweep for lower M. The dictionaries where $M > 32$ show no bit count minimum in this figure, though they do have one that appears with higher λ. However, these dictionaries will have a $\chi^2/dof \gg 1$.

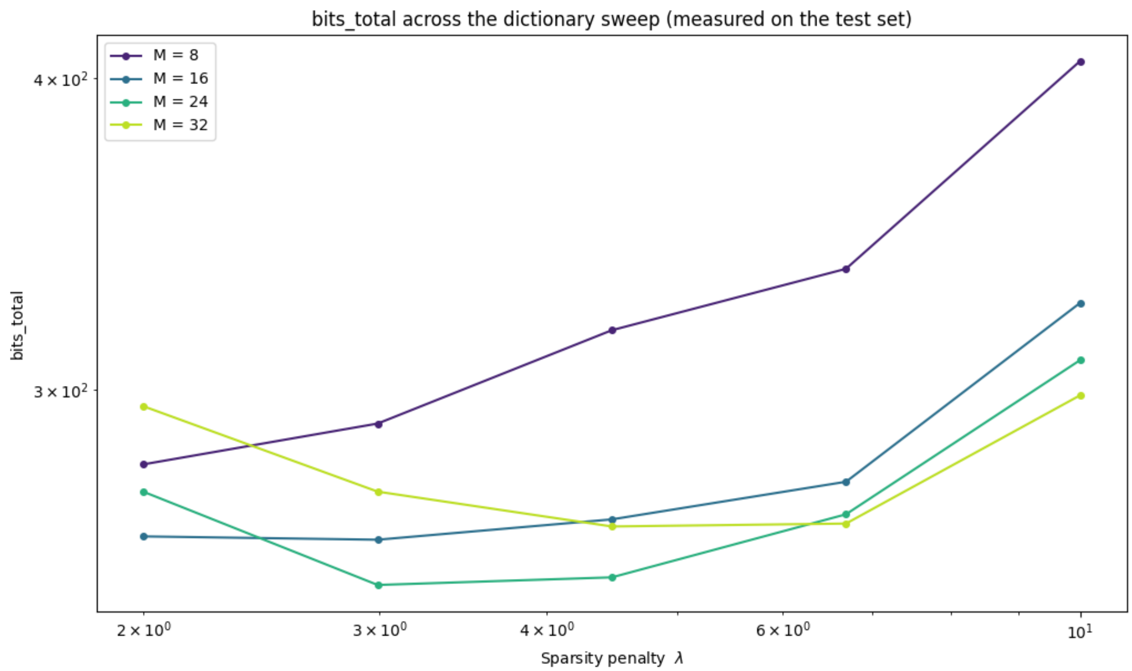

Figure 5: The y-axis shows the total bits **per source**, including the dictionary cost when divided by the sample size of $2.046 \cdot 10^6$, but not including the two residual constants. This figure shows the total bit count for models with $M \in \{8, 16, 24, 32\}$ and $\lambda \in [2.00, 10.00]$, showing that the minimum shown in Figure 4 was only the minimum because it was the minimum boundary of the sweep.

To provide more accurate data, we now use the entire dataset (80% train, 20% test) to train dictionaries and determine the proper dictionary parameters.

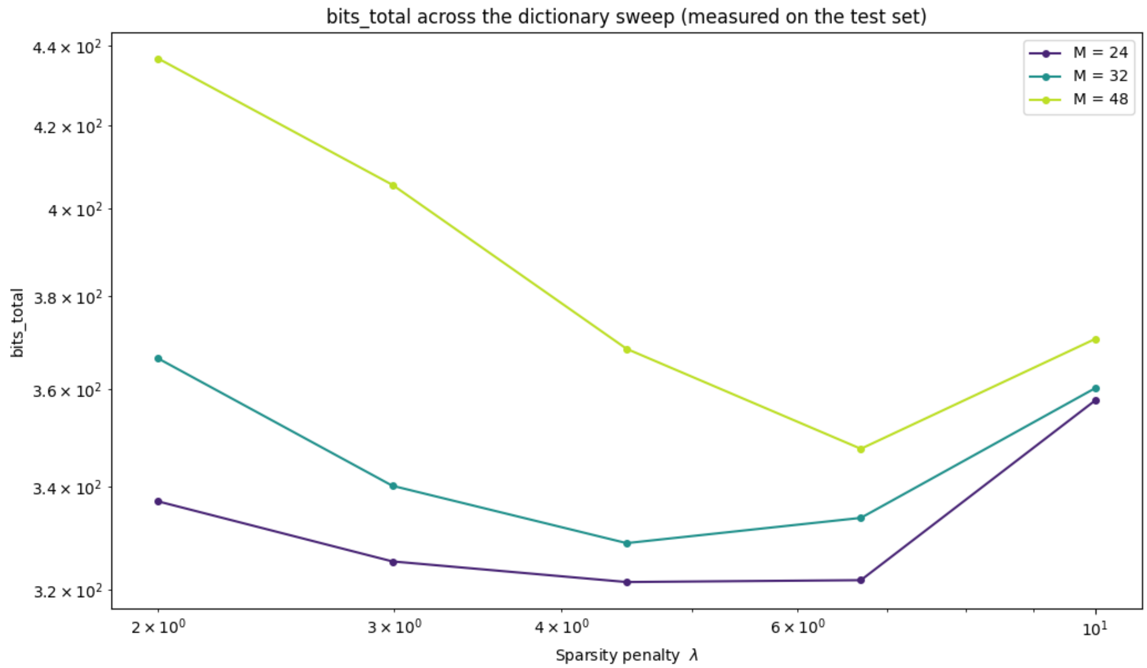

Figure 6: The y-axis shows the total bits **per source**, including the dictionary cost when divided by the sample size of $2.046 \cdot 10^6$, but not including the two residual constants.

The minimum bit count for a possible model lies around $M = 24, \lambda = 5.0$, with a further test showing that the bit size is essentially constant in $4 < \lambda < 7$. However, this model has a slightly elevated $\chi^2/dof > 1.0$. Instead of this model, we will pick $\lambda = 3.0$, which achieves a $\chi^2/dof = 1.00 \pm 0.01$ for a dictionary with $M = 24$ in the sweep, at the cost of a slightly higher bit size. From now on, we will call this dictionary the Compact Dictionary D_C. We expect that D_C will perform poorly on the specialization criteria, because of its small dictionary size. We will therefore add another dictionary with $M = 256$, $\lambda = 7.0$, which also achieves $\chi^2/dof = 1.00 \pm 0.01$. From now on, we will call this the High-Capacity Dictionary D_H.

Criterion	D_C	D_H
Zero-fraction	0.0	0.0
Dead atoms	0.0	0.0
$\chi^2/dof \approx 1$	1.0092 [1.0220]	0.9841 [0.9886]
$\epsilon_{test} \approx \epsilon_{train}$	$\epsilon_{test} \approx 1.05\epsilon_{train}$	$\epsilon_{test} \approx 1.00\epsilon_{train}$
Bit count	$182S$ [$185S$]	$461S$ [$450S$]
Median compression ratio	8.71× [8.73×]	3.64× [3.79×]
Median/p75 MCCS	0.510/0.740 [0.385/0.657]	0.567/0.753 [0.548/0.749]
Refit stability	0.418	0.427
Usage dynamic range	3.07×	23.9×
Distinct classes captured	4	8
Best atom label selectivity	0.533 (T_{eff}), 0.482 ($\log g$), 0.453 ($[M/H]$)	0.781 (T_{eff}), 0.705 ($\log g$), 0.633 ($[M/H]$)

Table 6: Filled in criteria for our two chosen dictionaries (D_C, the Compact Dictionary, $M = 24, \lambda = 3.0$ and D_H, the High-Capacity Dictionary, $M = 256, \lambda = 7.0$). When in the following format: x_1 [x_2], x_1 is the median for a 10-seed sample, and x_2 is the value for the random seed used in this thesis. An elaboration on these results is given in Section 4.2. Zero-fraction and dead atoms were 0.0 for all seeds. These metrics were found from the entire dictionary rather than a sample. These show that D_C is substantially better at compression tasks, but D_H has more specialist atoms, making them easier to interpret, unlike very generalist atoms.

3.4 Atom Interpretation & Experiment Setup

The following section will outline the setup for each experiment on the discovered dictionaries D_C and D_H.

Compression. This section will contain sparsity histograms for both dictionaries to explore any irregularities in these. The "default" result that is expected if no patterns are noticeable is a (skewed) normal distribution.

Additionally, this section will contain the relation between the sparseness of a code and certain source properties that were discussed earlier, namely T_{eff}, $[M/H]$ and $\log g$, in addition to the normalization coefficient `rp_00`.

This section will also contain our findings of the compressibility of Gaia spectra, by comparing the compression of our dictionaries to that of a principal component analysis.

Atom Stability. This section will cover the observed stability of atoms over multiple runs, reporting which atoms are unstable between runs by training the dictionary multiple times with different random states.

Another experiment performed in this section is a subsample stability test, which determines whether the atoms are stable when tested on different subsamples of our full Gaia sample.

Lastly, we determine the Jaccard index of the atom supports (when optimally matched) for different dictionary seeds, to find the code similarity between matched stars, and thus to determine whether code representation of stars is stable with different random initializations. We use a sample of 200 thousand stars for this test. The Jaccard index is defined in Equation 16.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{16}$$

Instrumental Error Correlation. This section will find whether the assignment of a subset of the atoms is highly correlated to the apparent magnitude of sources, which should not have any effect on source encoding if no instrument error was picked up, as the spectra were normalized by brightness using `rp_00`. To achieve this, we will calculate the Spearman rank correlation between the absolute activation of each atom and the apparent G magnitude of the sources for which that atom is active. The absolute activation is used because both positive and negative coefficients indicate that an atom contributes to the encoding of a source. A large absolute correlation coefficient indicates the strength of an atom's activation changes with apparent brightness. We set a limit of $|\rho| < 0.3$ for atoms that are not strongly correlated with apparent magnitude.

Atom Specialization. In this section we will determine which atoms contain an overrepresentation of a SIMBAD object type (see Table 1). The metrics we will use for this are: (1) Enrichment factor, o/e, and (2) Poisson significance score $z = \frac{o-e}{\sqrt{e}}$, where o is the observed count of an object type, and e is the expected count of an object type.

As previously stated, SIMBAD cross-matching recovered 76,766 stars, or approximately 3.8% coverage, and 99% of matches are stellar-scale objects. Note that SIMBAD records consist of objects that have already been studied, meaning rare classes are overrepresented relative to their true frequency. This section can determine which classes often contain a particular atom in their code, but not how common each class is. To perform an informative analysis on the overrepresentation of SIMBAD classes in our atoms, we must set some criteria as to not see extreme overrepresentations of tiny object types. For instance, if an object type contains 2 sources, having 1 of the sources contain a certain atom would be an extreme overrepresentation of this object type in the atom. Therefore, we set a limit `min_type_count = 50`, meaning that a class must contain at least 50 cross-matched instances in the entire sample. When referencing Table 1, we can see that the dominant portion of categories does not meet this criterion. Specifically, 30 classes meet this requirement.

To derive the significant z, we calculate the statistical test count, $k = M \cdot C$, where C is the amount of detectable classes (30), as each class is tested against every atom. A Bonferroni correction at $\alpha = 0.05$ gives a per-test false positive rate of α/k. This gives us a per-test threshold of $z > 4.0$ for D_C, and $z > 4.5$ for D_H. All reported overrepresentations will exceed this value.

Additionally, to derive specialization, we will determine whether atoms have a coverage higher than 50% of specific object types, and determine the fraction that this class represent in all of the atoms activations. To determine whether an atom consistently captures an object type, and predominantly this object type, we will use the following metric for the specialization of atom m in object type t:

$$\zeta_{t,m} = t_m \frac{N_d}{N_t} \tag{17}$$

Where t_m is the count of the object type which contains the atom d_m in its code, N_t is the total count of that object type in the sample, and N_d is the total count of objects cross-matched with SIMBAD that contain atom d. A fully specialized atom m_s would have $\zeta_{t,m_s} = 1$ for a single object type t. Note that this will not identify an atom which is capable of identifying a subclass, as it will have a high atom fraction t_m/N_d, but a low coverage t_m/N_t.

Stellar Parameter Correspondence. This section contains the degree to which atoms cluster around each of the previously discussed labels (T_{eff}, $[M/H]$ and $\log g$), measured by the ratio of the activation-weighted standard deviation of the label across the activated sources, divided by the standard deviation of the label across the entire population, and by the selectivity.

$$\text{sel} = 1 - \frac{\text{IQR}_{\text{active}}}{\text{IQR}_{\text{pop}}} \tag{18}$$

Note that this does not capture negative activations, though $> 99\%$ of all activations are positive. It will also display Hertzsprung-Russel diagrams for all (D_C) or interesting (D_H) atoms for illustration.

Additionally, this section will train an HGB model to predict stellar labels and SIMBAD classes through each atom, finding which atom is of highest importance in these predictions. For T_{eff}, $\log g$ and $[M/H]$, we first perform hyperparameter tuning of the HGB model using the Python package `optuna`. After this, we train the model to predict a specified label from the sparse codes or raw spectral coefficients, using the values provided by [Andrae et al., 2023] as training labels, and the values from [Abdurro'uf et al., 2022] as test labels. We then perform an importance analysis for each feature (atom/raw coefficient) j by shuffling j across sources and re-scoring the fitted model, for 10 different, randomly shuffled instances. We determine the importance of j as follows: $\text{imp(j)} = \text{R}^2 - \text{R}^2_{\text{shuffled}}$, where R^2 is the coefficient of determination. We compare the findings of this experiment to the results using the original coefficient dataset in place of the sparse coding.

Atoms in Flux Space. In the rest of the experiments, we will determine the metrics Stability, Distinctness, Non-instrumentality from Section 2.7. Plausibility and Correspondence remain to be tested, though.

Firstly, we can determine the Plausibility of an atom by determining its signal-to-noise ratio, which is as follows. The SNR of each coefficient is defined by $\text{SNR}_\text{n} = \frac{\sigma_\text{n}}{\epsilon_\text{n}}$. Since $\|d\|_2 = 1$, we can say coefficient n contributes d_{mn}^2 length to d as a vector. We use this as a weight to sum all signal-to-noise ratios of each coefficient.

Correspondence is significantly harder to determine a metric for, as it is very hard to define what is a "spectral feature" and what is simply noise, in particular for atoms. The metric we define for correspondence is therefore somewhat subjective, and is mostly meant as a proof that there is correspondence between atom and spectral features, rather than determining which spectral features are exactly present in each atom. We therefore define a diagnostic set of spectral lines which we will attempt to detect in the atoms. We define the line windows we will be measuring in Table 7. Learned atoms will be returned from coefficient to flux space, by first un-normalizing the coefficients by their error, and then passing them through `GaiaXPy` [Ruz-Mieres et al., 2025], which yields flux against wavelength. The amplitude normalization is not reversed, as an atom describes a spectral shape rather than an absolute flux. Most atoms are dominated by a broad continuum, with narrow deviations. To find these deviations, a continuum estimate is obtained with

Spectral line	Wavelength window [nm]
Hα	654–658
Hβ	484–488
Hγ	432–436
Ca II K	392–395
Ca II H	396–399
Na I D	588–590
Mg I b	516–519

Table 7: Spectral line windows used in the analysis.

a Savitzky-Golay filter of width w and polynomial order 3, after which we compute the residual by subtracting the found continuum from the total spectrum, which retains only the structure narrower than the filter. At ~ 2 nm sampling of a calibrated XP spectrum, $w = 41$ corresponds to approximately 80 nm, several times the width of a resolution element at $R \sim 70$. With this, real spectral features will be preserved in the residual while slope and curvature are absorbed into the continuum.

Detection thresholds are set locally in the spectrum, rather than globally. This is due to the fact that the Hermite basis produces oscillations whose amplitude grows toward shorter wavelength, so a single threshold across the spectrum would detect all blue oscillations, while dropping real spectral features in the red section of the spectrum. A local scale is therefore estimated as a rolling median absolute deviation,

$$\sigma(\lambda) = 1.4826 \cdot \mathrm{med}_{\mathrm{W}(\lambda)} |\mathrm{r} - \mathrm{med}_{\mathrm{W}(\lambda)} \mathrm{r}| \tag{19}$$

over a window $W = 101$ samples, wider than the continuum filter so that a feature does not inflate its own noise floor. The factor 1.4826 converts the median absolute deviation to the equivalent standard deviation of a Gaussian. Peaks are then identified in r and $-r$, corresponding to features that are higher and lower than the continuum, and retained where prominence exceeds $3\sigma(\lambda)$. Each detection is recorded together with its wavelength, amplitude, significance and sign, and is measured against the set of spectral features in Table 7. These windows were fixed in advance before analysis.

A very important note is that none of the atoms contain a `rp_00`, which means they will have an unnatural shape, with a steep slope around the BP/RP separation. To alleviate this, we need to recover the optimal `rp_00` of each atom, even though it is not included. `rp_00` is not taken into account when learning, but we can recover the approximate value it would have had if the atoms were structured as they are now, because we know that this coefficient will have to be 1 for every star. Therefore, we assign the value that would recover this 1 to the `rp_00` coefficient for each atom, to retain a natural shape. We find this value for each atom by minimizing the square error of `rp_00` approximating 1 over the sample. This change is only made for displaying atoms and for the locating of features, as this steep slope will count as a feature without this added coefficient.

Anomalous Spectra. This section will cover the sources that have the highest χ^2, and cross-match them against SIMBAD to find the fraction of these spectral anomalies that have been previously observed, and concluding at which ratio these objects have been previously observed compared to the average object in the sample.

4 Results

4.1 Compression

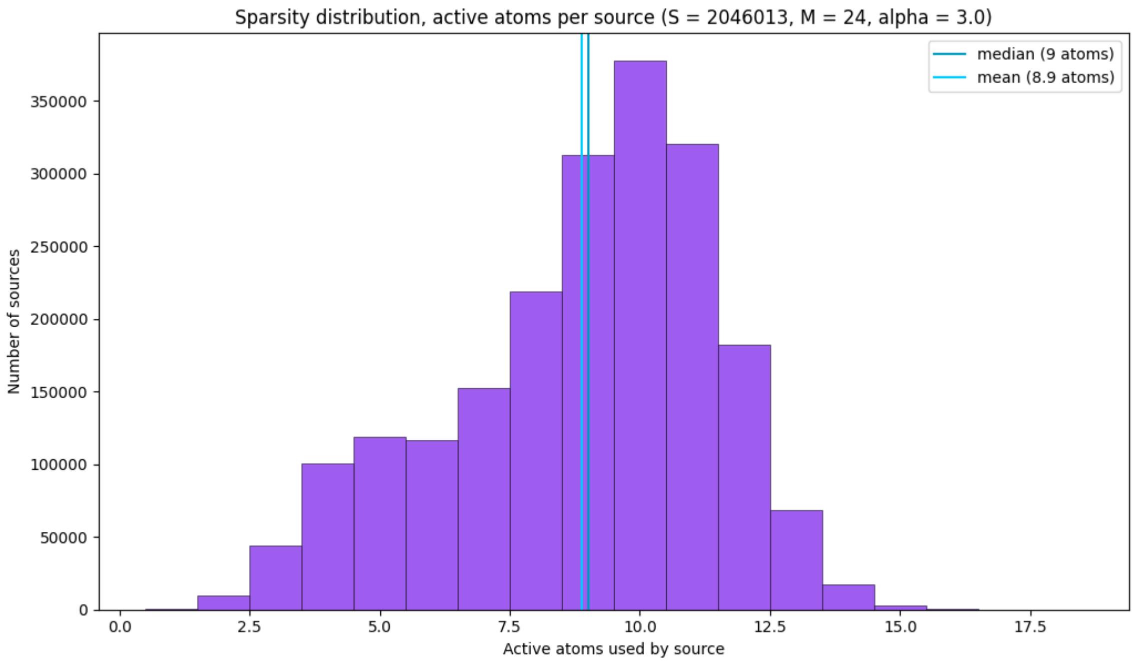

Figure 7: A sparsity distribution over all sources of D_C, showing that the mean and median s_{eff} lie around 9 atoms, meaning each source holds on average around 37% of all atoms. This points to the fact that D_C is a good representation from an MDL perspective, but likely consists entirely of generalist atoms.

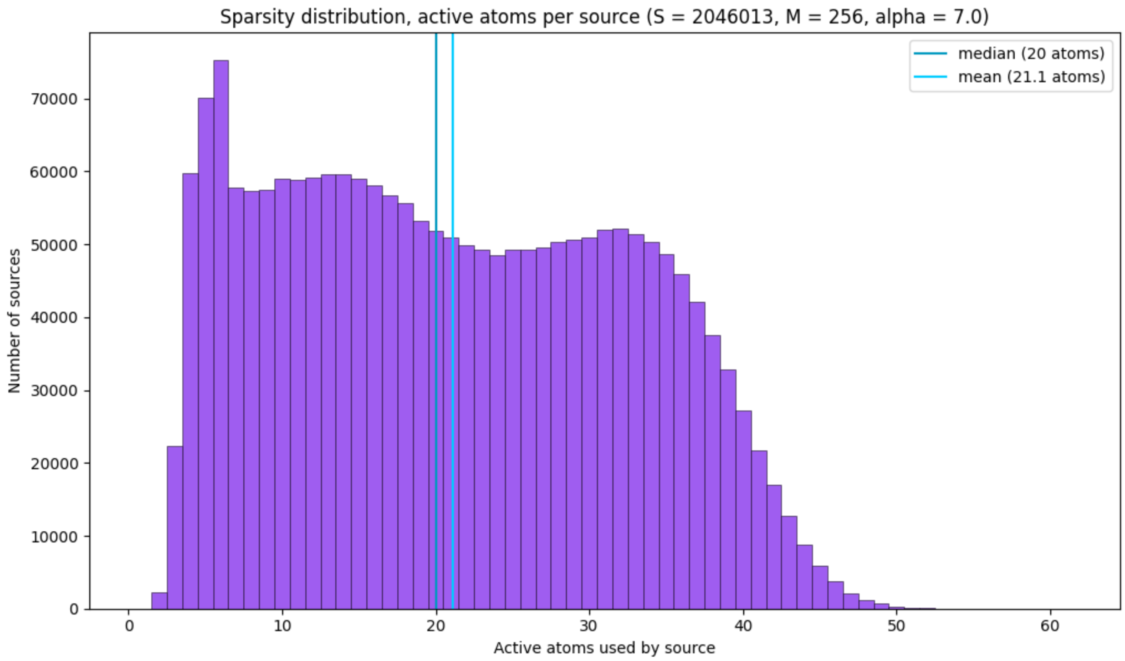

Figure 8: A sparsity distribution over all sources of D_H, showing that the mean and median s_{eff} lie around 20 atoms, meaning each source holds on average around 7.8% of all atoms.

Figure 8 shows that D_H is a better representation when looking for specialist atoms. An additional observation is that the distribution of s_{eff} over sources when coded by D_H is clearly bimodal, and arguably trimodal, with peaks at $s_{\text{eff}} = 6$, $s_{\text{eff}} = 13$ and $s_{\text{eff}} = 32$. This peak at $\sim s_{\text{eff}} = 6$ also survives in the sparsity distribution of D_C (though showing at $s_{\text{eff}} = 5$ here), showing there is likely a subpopulation whose spectra are very easily encoded. The fact that this bimodality is much more visible in the sparsity distribution of D_H points to the fact that it preserved some structure that D_C was unable to capture as clearly.

We can test the claim that the sparsely encoded stars of the bimodality correspond to a certain class directly by viewing the properties of the sources which have $s_{\text{eff}} \approx 6$. We conclude the following from a sample survey of D_H: sources with $s_{\text{eff}} \in [4, 6]$ consist largely of cool stars, with long period variable stars and LPV candidates (at 2.6×, z=59 and 3.1×, z=72 respectively) in this data. Emission line stars are also overrepresented at this sparsity (2.1×, z=23). Eclipsing binaries are strongly overrepresented at above-average s_{eff} (3.4×, z=80), which is to be expected for unresolved pairs whose observed spectrum contains the added light of two stars that may have distinct characteristics, therefore requiring two sets of atoms that code for different regions. Note that these metrics come from the SIMBAD-matched population in the sparsity bin, not the bin in its entirety.

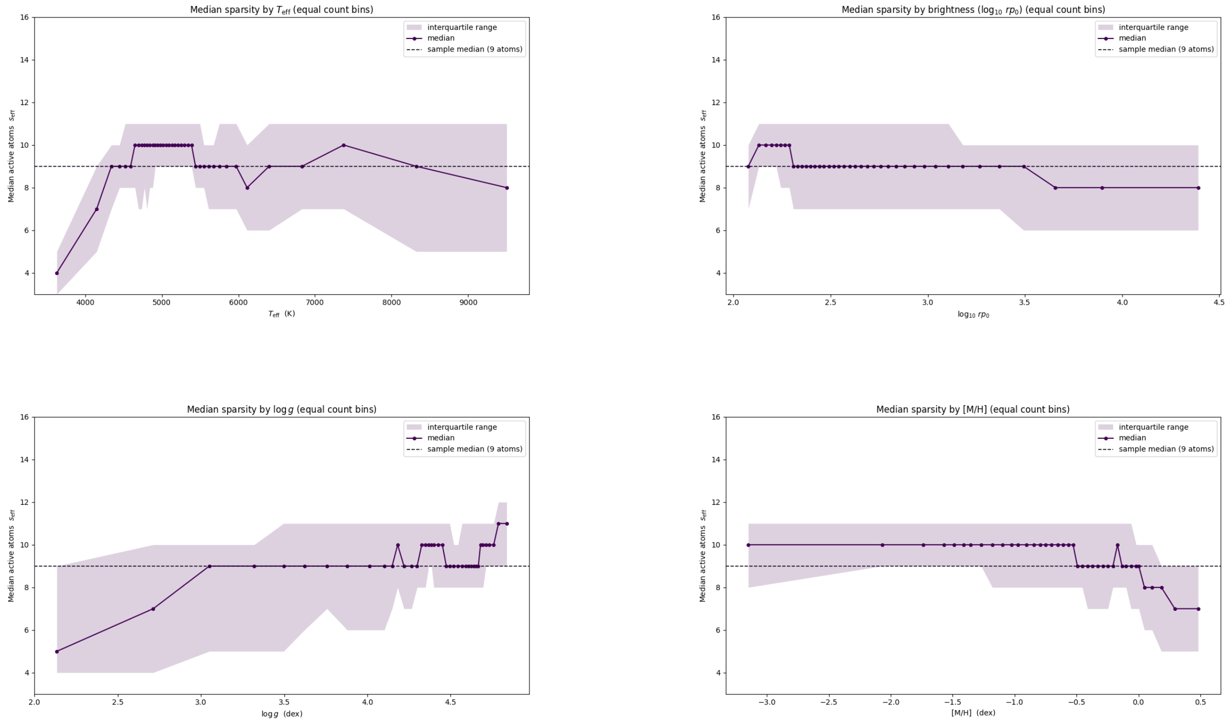

Figure 9: The relationship between sparsity and $T_{\rm eff}$, $[M/H]$, $\log g$ and `rp_00` of dictionary D_C. Previous analyses of a dictionary of $M = 32, \lambda = 3.2$ show a more clear resemblance to the D_H figure, meaning that the shape of the relation is likely the similar no matter the dictionary size.

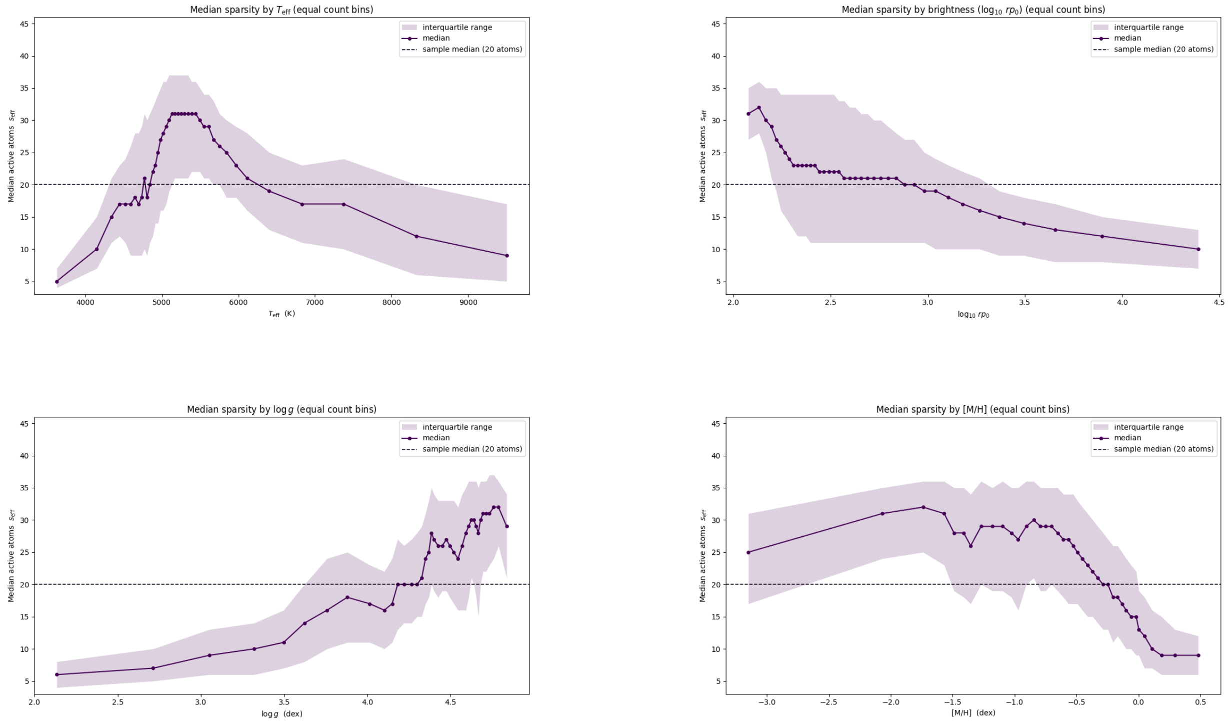

Figure 10: The relationship between sparsity and T_{eff}, $[M/H]$, $\log g$ and `rp_00` of dictionary D_H. There appears to be a clear relationship between sparsity and these four properties, which is persistent in both dictionaries, pointing to the fact that stars with certain values of these properties may have spectra that are inherently harder to compress (here it is important to note that `rp_00` is not necessarily a stellar property, but also dependent on distance).

Lastly, we perform a principal component analysis on the sample to determine the effectiveness of compression of sparse dictionary learning. We find that a PCA of 8 components achieves a $\chi^2/dof \approx 1$. The median compression ratio we find is 13.6× (once again leaving out the residual), which is considerably more than both dictionaries.

4.1.1 Conclusion

Both dictionaries produce sparse codes that are well short of saturation, with maximum s_{eff} of 18 out of 24 atoms for D_C, and 61 out of 256 atoms for D_H. A population of 5-6 active atoms appears in both distributions, and its persistence across an over 10x increase in dictionary size indicates that there is a population of simple spectra rather than an artifact of representation. The distributions differ in shape strongly, D_C is largely unimodal, while D_H is bimodal or even trimodal, with a clear second mode near $s_{\text{eff}} = 33$. This second mode is likely present due to sources drawing from specialist atoms that are possessed in much greater numbers by D_H, also being supported by the fact that the first peak does not grow nearly as much as the second. This provides evidence to the claim that additional dictionary capacity is used on specializing atoms, rather than wasted, independent of Section 4.4.

There appears to be a clear relationship between the four source properties that were examined, and code sparseness, which persists in shape and amplitude through both dictionaries, showing that sources that have certain discovered properties are likely inherently harder to compress. There is a relationship between `rp_00` and sparsity, which points to an instrumental effect or to a shortcoming of the normalization. The effect appears to show a decrease in sparsity for redder stars, which is also visible in the temperature relation, and in the fact that cooler stars are overrepresented in low sparsity.

We conclude that sparse dictionary learning is not as effective at compression as principal component analysis, as it achieves a compression ratio of 8.71 at measurement noise precision ($\chi^2 \approx 1$), while PCA achieves a compression ratio of 13.6.

4.2 Atom Stability

We first determine the stability of each dictionary by computing 10 random seeds with their metrics. This was previously used in Section 3, as it was a requirement to choose a dictionary.

Criterion	D_C	D_H
$\chi^2/dof \approx 1$		
Used	1.0220	0.9886
Q25	0.9979	0.9824
Median	1.0092	0.9841
Q75	1.0316	0.9869
Bit count		
Used	$185S$	$450S$
Q25	$175S$	$455S$
Median	$182S$	$461S$
Q75	$185S$	$469S$
Median compression ratio		
Used	8.73×	3.79×
Q25	8.36×	3.61×
Median	8.71×	3.64×
Q75	9.19×	3.71×
Median/p75 MCCS		
Used	0.385/0.657	0.548/0.749
Q25	0.445/0.694	0.551/0.742
Median	0.510/0.740	0.567/0.753
Q75	0.540/0.805	0.579/0.764
Refit stability		
Frac. matched	0.4176	0.4266
Matched median/p25 MCCS	0.856/0.692	0.857/0.714

Table 8: Dictionary criteria for D_C and D_H, reporting the detailed results of sample of 10 random initializations.

We perform another test with the same random state, but using a randomly sampled subsample consisting of 50% of the complete sample. For this test we log the same metrics as the ones measured in Table 8.

Criterion	D_C	D_H
$\chi^2/dof \approx 1$		
Used	1.0220	0.9886
Q25	1.0245	0.9861
Median	1.0357	0.9886
Q75	1.0476	0.9964
Bit count		
Used	$185S$	$450S$
Q25	$158S$	$420S$
Median	$166S$	$429S$
Q75	$172S$	$452S$
Median compression ratio		
Used	8.73×	3.78×
Q25	9.42×	3.62×
Median	9.58×	3.78×
Q75	10.4×	3.94×
Median/p75 MCCS		
Used	0.385/0.657	0.548/0.749
Q25	0.357/0.661	0.516/0.726
Median	0.414/0.692	0.557/0.749
Q75	0.442/0.700	0.571/0.759
Refit stability		
Frac. matched	0.2120	0.4131
Matched median/p25 MCCS	0.720/0.543	0.849/0.709

Table 9: Dictionary criteria for both dictionaries, compared to metrics computed from a set of 10 dictionaries learned from a randomly selected 50% subsample.

Computing a Jaccard index sweep for 10 random initializations of dictionaries with D_C parameters yields a median Jaccard index of 0.462, with the upper quartile at 0.571, revealing that code representations do not necessarily match between sources if matching atom pairs are created. This number is similar to the earlier fraction of atoms that were matched to other dictionaries at higher similarity than its twin in its own dictionary. This means that individual stellar code representations will — in most cases — have a low code interpretability, as these codes will not match well between runs in most cases. For D_H, which has a larger variety of specialized atoms, and thus also a larger room for different codes, we find an expectedly lower Jaccard index of 0.273, with an upper quartile at 0.333.

4.2.1 Conclusion

We can conclude from Table 5 that the chosen dictionary reconstructions do not contain entirely stable and unique atoms in most cases, though there is a subset of atoms that persists. Additionally, earlier observations show that atom structure of a dictionary is often similar with different random initializations and different parameters. The dictionaries compress at similar effectiveness at different random initializations. Under 50% subsampling, 43% of D_H atoms match, but only 21% of D_C atoms, a very significant decrease. This may mean that a dictionary of 24 atoms might take on a very different structure at a lower sample size. The Jaccard index experiment reveals that codes of the same sources are not very consistent between runs when atoms are optimally matched, and shows that larger dictionaries have less consistent codes.

4.3 Instrumental Error Correlation

We find that the spearman rank correlation between atom activation and source apparent G magnitude is strong ($|\rho| > 0.3$) for 109/256 atoms in D_H, and for 2/24 atoms in D_C (the ρ values for these atoms are displayed in Section A of the appendix, and their flux structure is displayed in Figures 18 and 19), with a maximum absolute correlation of $\rho \approx 0.5$ for both dictionaries. This indicates that the activation of a substantial fraction of the learned atoms is directly related to the apparent magnitude of a source. Since the input spectra are normalized by `rp_00`, this dependence is not expected to arise strongly simply from the overall flux level of the source. Instead, it suggests that the learned representation is sensitive to properties of the data that vary with apparent magnitude, which may include instrumental noise effects. It can also be seen that these G magnitude correlating atoms have a range of different ρ, which points to the fact that instrumental and astrophysical structures are not well separated.

We additionally determine whether this observed relationship can be explained by the correlation between apparent magnitude and source color, by repeating the analysis within 16 $BP - RP$ bins. This follow-up gives similar results, with the correlation remaining present for the same atoms, only exhibiting slight changes in either direction for both dictionaries, showing that the observed dependence is not removed when correcting for color.

This alone is not enough to establish that instrumental noise is the cause of this relationship, since other source properties and observational effects or data quality artifacts may also covary with both apparent magnitude and atom activation. We therefore conclude that this result is evidence that the learned atoms of both dictionaries are associated with the apparent magnitude of sources, which is consistent with the presence of instrumental errors or noise-related information in the code.

4.4 Atom Specialization

4.4.1 Compact Dictionary D_C

We determine the overrepresented types of each atom in D_C and D_H. We observe that the dominant class of all except 2 D_C atoms, when covering all activations of each atom, is ordinary field stars ($*$). This is an expected outcome, as $M = 24, \operatorname{med} s_{\text{eff}} = 9$ is a considerable compression from 110 down to 9 median coefficients, with no overcompleteness. The two atoms that are not majority background stars are majority long period variable stars, which may be considered specialist atoms. These specializations will be discussed later in this section.

In Figure 11, we observe that the two atoms with the lowest mean (14, 7) have a distinctly different distribution of sparsity than all other atoms, and we observe that 5 out of the following 7 atoms in this order (6, 20, 17, 4, 3) have a similar distribution to the other atoms, but contain a strong peak around $4 \leq s_{\text{eff}} \leq 6$, which we also observe in Figure 7. The two atoms with a different distribution are also the two atoms with the highest significance score, and both of the atoms has an overrepresentation of long period variable stars, providing more evidence that long period variable stars are more easily compressible by sparse dictionary learning. Atoms 8, 9, 12 and 10 show the highest overrepresentations of eclipsing binaries, and three of these four atoms are also in the top 25% atoms with highest s_{eff}, which is more evidence for the fact that eclipsing binaries are inherently harder to compress than other stars.

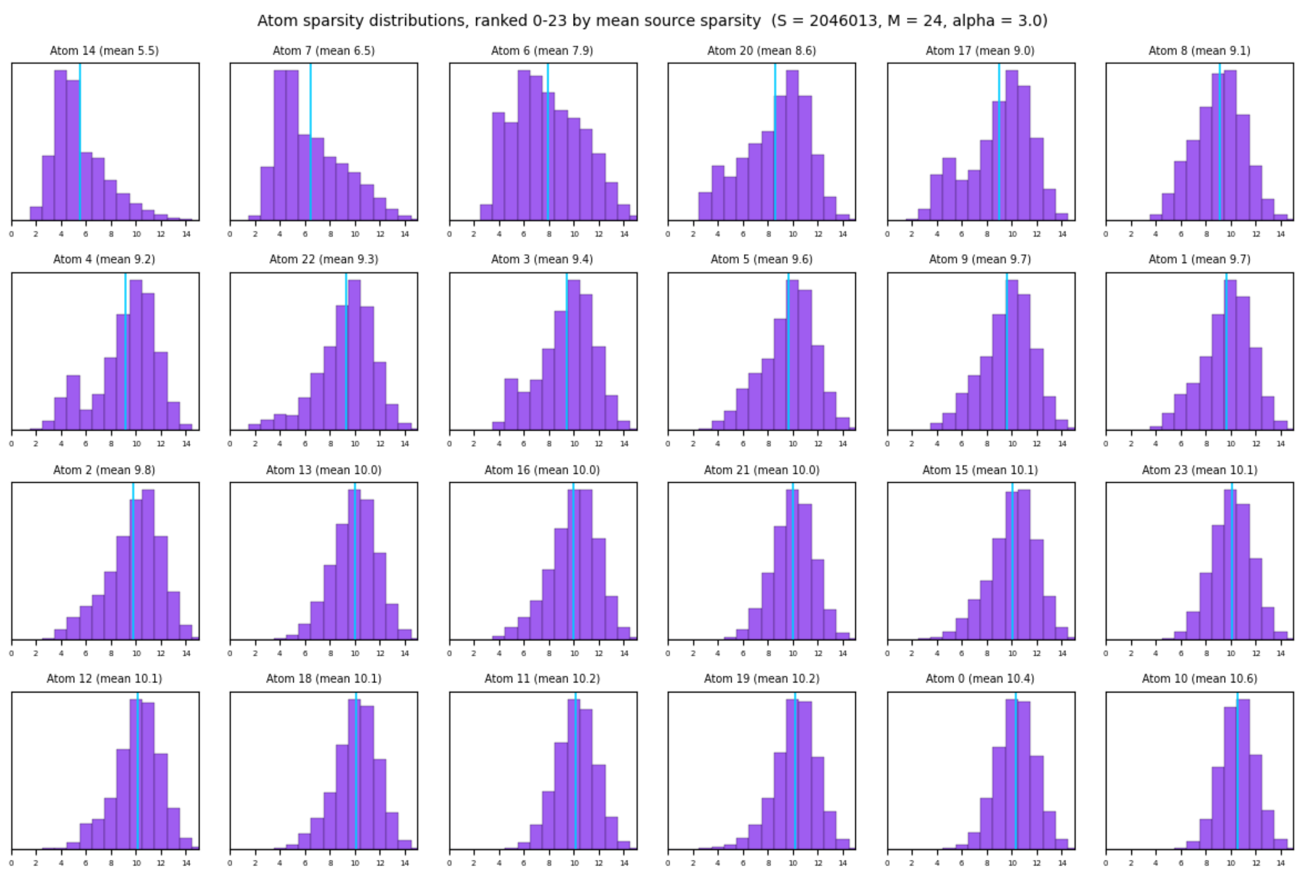

Figure 11: The sparsity distribution of sources that activate each atom, sorted by mean sparsity of sources activating the atom.

We observe that the top two specializations of any atom ($\zeta_{*,3} = 0.54, \zeta_{*,17} = 0.43$) are both ordinary background star specializations, which are the only cases of this happening in this thesis. These atoms also have high coverages of many other classes, such as high proper motion stars, spectroscopic binaries, red giants and white dwarfs, showing that these atoms, though counting as "specialized" through our metric ζ, are actually very generalist atoms, as it is "specialized" in many different classes. This is a flaw in our statistic, but not one that will display itself further.

Atoms 7 and 14 display extremely high enrichment in long period variable stars, with candidates and confirmed long period variables having $z \geq 79$ in both atoms. These atoms are also the atoms that have the lowest mean s_{eff} among its sources, once again confirming that long period variables tend to be easily compressible.

Additionally, we observe that some atoms have a coverage of 50% of at least one class at only top 5% activations of said atom. The specializations of these atoms in these top 5% activations are $\zeta > 0.1$ in D_C, the atoms being atom 4, 7, 20 and 6, with specializations $\zeta_{\texttt{LP?},4} = 0.37$, $\zeta_{\texttt{LP*},4} = 0.24$, $\zeta_{\texttt{LP?},7} = 0.29$, $\zeta_{\texttt{C*},20} = 0.11$ and $\zeta_{\texttt{WD*},6} = 0.10$. One of these atoms shows a very high atom fraction in D_C when taken at high activations only, that being atom d_4, with over 83% of all sources that contain this atom at top 5% activation being long period variables or candidates.

When observing the top 1% of absolute activations of each atom in D_C, there is a much more clear split between each object type, meaning most background stars do not have a very high activation of any atom. Long period variables, Mira variables and eclipsing binaries once again often appearing classes. Contrary to the initial table, though, there is a subset of atoms that has an extreme overrepresentation of white dwarfs, often coinciding with spectra of AGNs or QSOs. Perhaps the single most interesting finding in this table is d_{20}, which has an extreme overrepresentation of carbon stars, containing almost half of all SIMBAD confirmed carbon stars in the entire sample in its top 1% activations. Expanding to top 5% of activations only yields 120 carbon stars, meaning not all carbon stars are highly activated by this atom. Additionally, other activations of this atom include white dwarf and Mira variables, which is interesting, as carbon stars and white dwarfs often exhibit opposite spectral features, with carbon stars having strong features and white dwarfs having largely featureless XP spectra. We therefore decided to investigate further by determining whether either white dwarfs or carbon stars activate this atom negatively rather than positively, finding that white dwarfs strongly negatively activate atom 20, whereas carbon stars strongly positively activate it. This is another good candidate for a specialized atom. To further compound on this evidence, we find that this atom persists through dictionary changes, as a dictionary with $M = 32, \lambda = 3.2$ also displays an atom that shares all previously discussed properties, meaning this atom is a consistent finding in dictionaries with varying initializations and parameters.

4.4.2 High-Capacity Dictionary D_H

In D_H, we observe that the dominant class of each atom varies, in particular displaying long period variables, eclipsing binaries and white dwarfs as the dominant class for a small subset of atoms. We observe that a very large subset of atoms has eclipsing binaries as their most statistically overrepresented class, which is to be expected, as these objects are also highly overrepresented at higher sparsities, meaning they contain more atoms overall. This is a similar pattern as seen in D_C. Another observation we make is atom 191, which is the 14th least activating atom in the dictionary, containing white dwarfs as its dominant fraction. This atom does contain other stars, however, not all white dwarfs are caught by this atom, so the specialization of this atom will be low.

Inspecting the top 1% activations of each atom in D_H reveals some atoms with an extremely high hit rate and an extremely high statistical over-representation. For instance, the atom 144 has a hit rate of 86%, with a carbon star enrichment of $158.0\times$, $z = 167$. This atom contains 93% of all carbon stars at top 5% activation, and 97% for all activations. Unlike in D_C, there is no significant white dwarf negative counterpart to these carbon stars, showing the atom is likely more specialized in this dictionary. Additionally, there are atoms that have a hit rate of over 90%, specifically d_{36}, d_{89} and d_{93} at 98%, having extreme overrepresentations of Mira variables, and d_{71}, d_{229} and d_{98} at 94-96%, with extreme overrepresentations for long period variables. A final observation that can be made is that most of the atoms have very high activations from white dwarfs. This dictionary displays highly specialized atoms for long period variables again, with the fraction of SIMBAD objects in d_{229} being $> 80\%$ long period variables. Another finding is that atoms 65, 40 and 88 have almost complete coverage for hot subdwarf candidates, and high coverage of AGNs, QSOs and white dwarfs.

The tables of overrepresented object types and atom specializations (ζ) are located in Section B of the appendix.

4.4.3 Conclusion

The Compact Dictionary D_C contains few atoms that are entirely specialized over all absolute activations, however, it does show that there are distinguishable patterns in atom activation, where long period variables and eclipsing binaries are again shown to be split in their atom activations. There are some clear exceptions to the mostly generalist atoms, though, with atoms 7 and 14 consisting for around 50% of long period variables each. Additionally, we observe that in top 1% of absolute activations, atoms are much more concentrated around particular classes, finding that almost half of all carbon stars activate atom 20 positively within the top 1% of activations, whereas negative activations of this atom are exhibited often by white dwarfs.

The High-Capacity Dictionary D_H contains strongly specialized atoms, with again showing multiple atoms that have an atom fraction of over 50% for long period variables, it also still contains an atom where carbon stars are extremely overrepresented, though it contains 97% of all carbon stars this time.

Overall, there are no clear-cut class separation atoms, but some atoms for both dictionaries show highly specialized populations using them.

4.5 Stellar Parameter Correlation

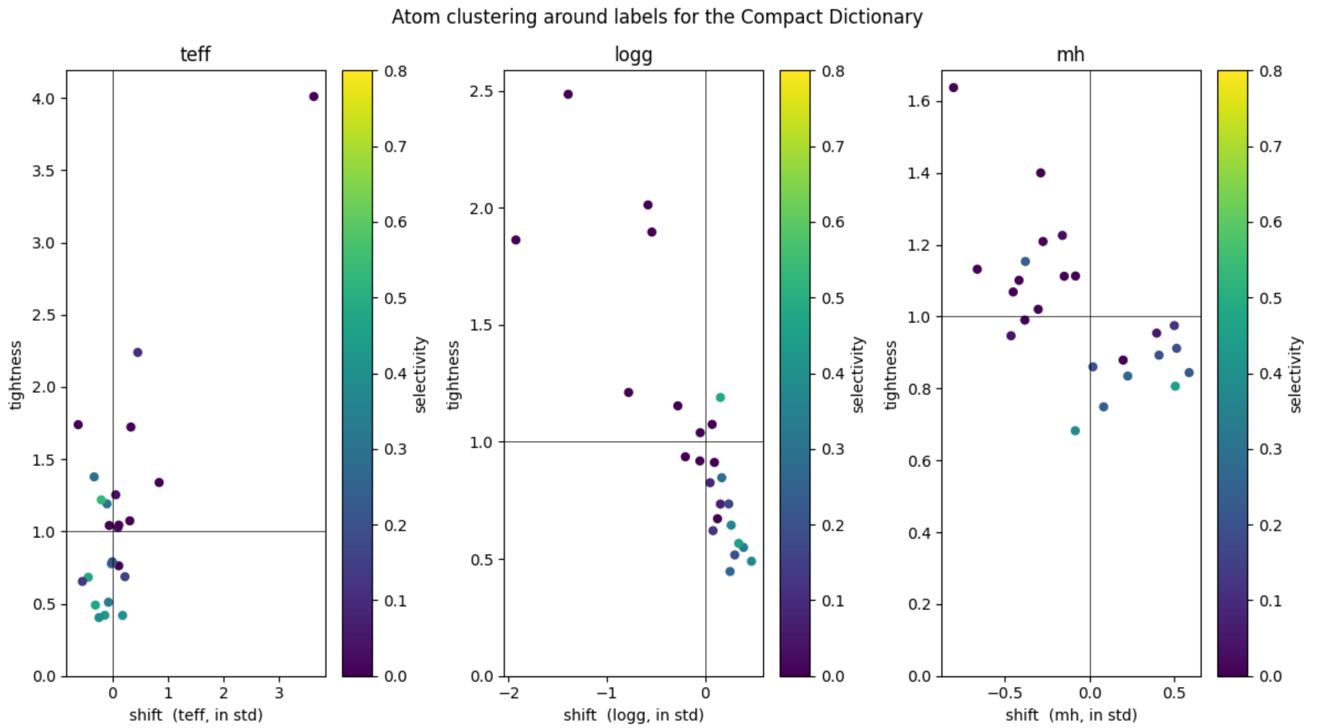

Figure 12: Clustering of atoms around source properties for D_C, colored by selectivity $1 - \frac{\text{IQR}_{\text{active}}}{\text{IQR}_{\text{pop}}}$.

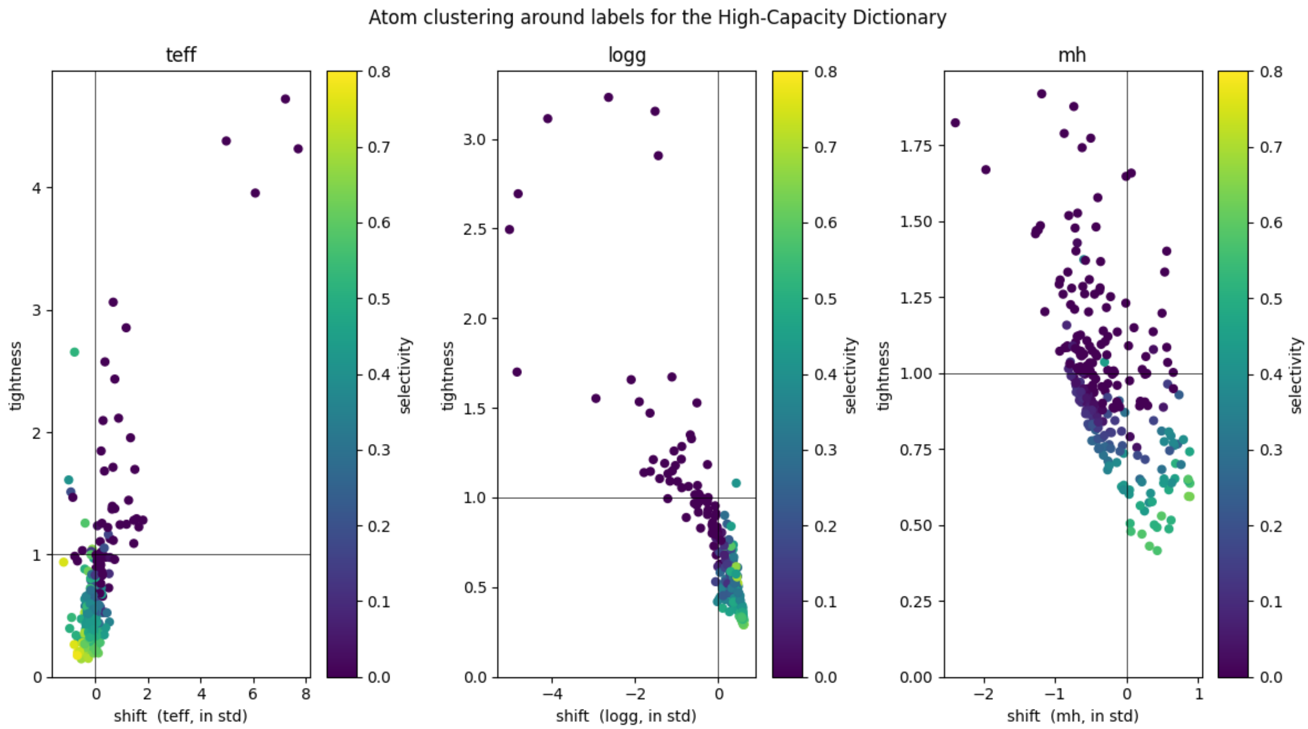

Figure 13: Clustering of atoms around source properties for D_H, colored by selectivity $1 - \frac{\mathrm{IQR}_{\mathrm{active}}}{\mathrm{IQR}_{\mathrm{pop}}}$.

Figure 12 and Figure 13 show that most atoms are not highly clustered around a specific label, especially for D_C, which contains more generalist atoms. For D_H, some T_{eff} atoms are highly clustered around specific temperatures, but this is to be expected, as T_{eff} is the most easily found property of the three that are plotted here. We find that the best selectivities for each dictionary are 0.533 (T_{eff}), 0.482 ($\log g$), 0.453 ($[M/H]$) for D_C and 0.781 (T_{eff}), 0.705 ($\log g$), 0.633 ($[M/H]$) for D_H. We again see that temperature is more easily recovered in Figure 16, where most atoms in D_H are highly clustered around a specific temperature, therefore often showing vertical clusters. For D_C, in Figure 14, we observe that most atoms are quite generalist, though some cluster somewhat tightly on a certain location in the HR diagram. Atom d_{20} can be clearly seen exhibiting its previously discovered characteristic, which is the fact that it is triggered negatively by the white dwarfs at the bottom left of the HR diagram.

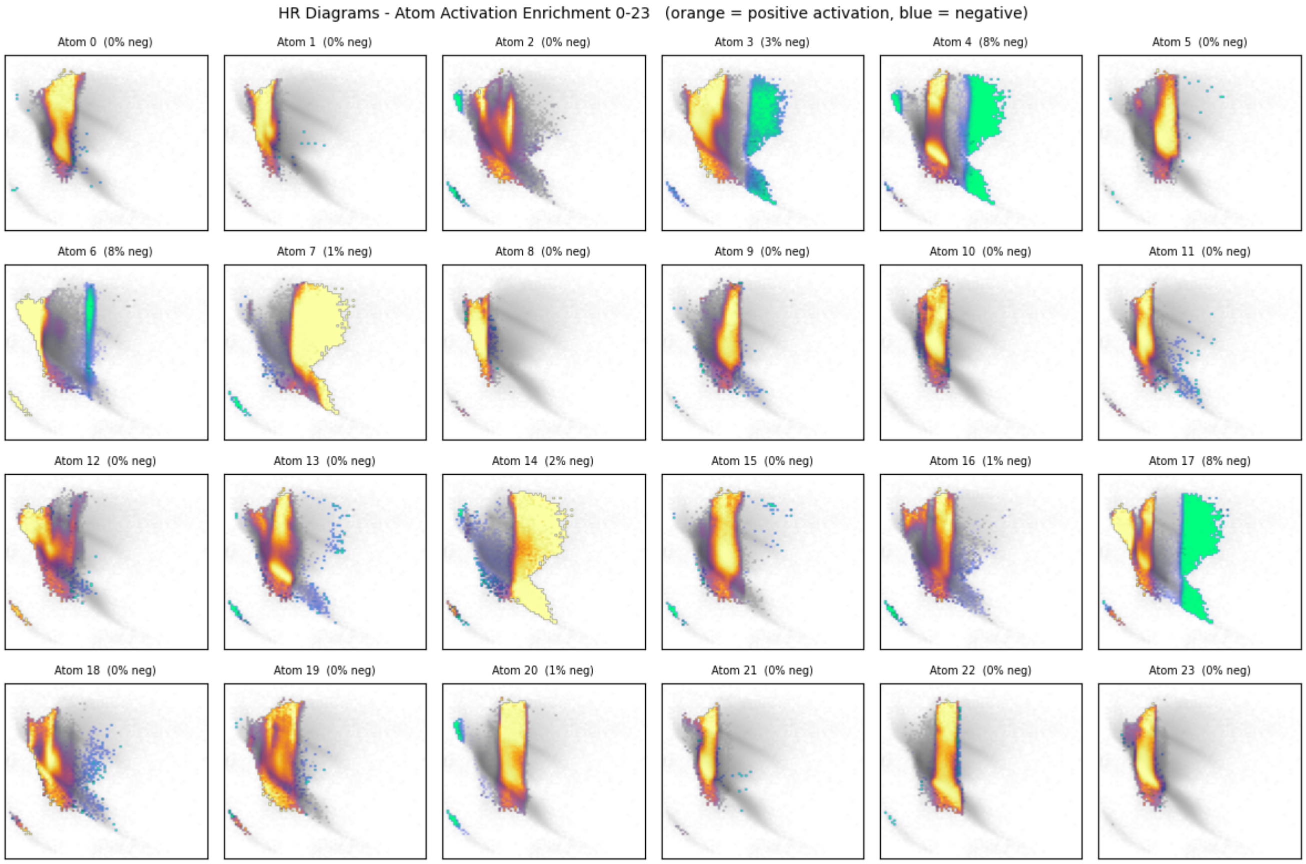

Figure 14: Hertzsprung-Russel diagrams with plotted atom enrichment of D_C. Absolute G magnitude on the y-axes, BP-RP magnitude on the x-axes. Light gray background contains the entire sample population. Heatmap is the fraction of the atoms fired on, low activations are dark grey, negative activations shown in blue/green.

Dust between Gaia and a source absorbs and scatters the light of the source, meaning it is dimmed and reddened, as light of shorter wavelengths is affected more than those of longer wavelengths. On a Hertzsprung-Russel diagram, that entails that a star is displaced down and to the right of its true position, along a direction set by the extinction law. Figure 14 shows a diagram of the observation, without extinction taken into account. In the following Figure 15, we correct the magnitude and color of objects that are dimmed and reddened using Gaia's GSPPhot `ag` and `ebpminrp` values. This will narrow the diagram, as Figure 14 has its main sequence and giants broadened by extinction.

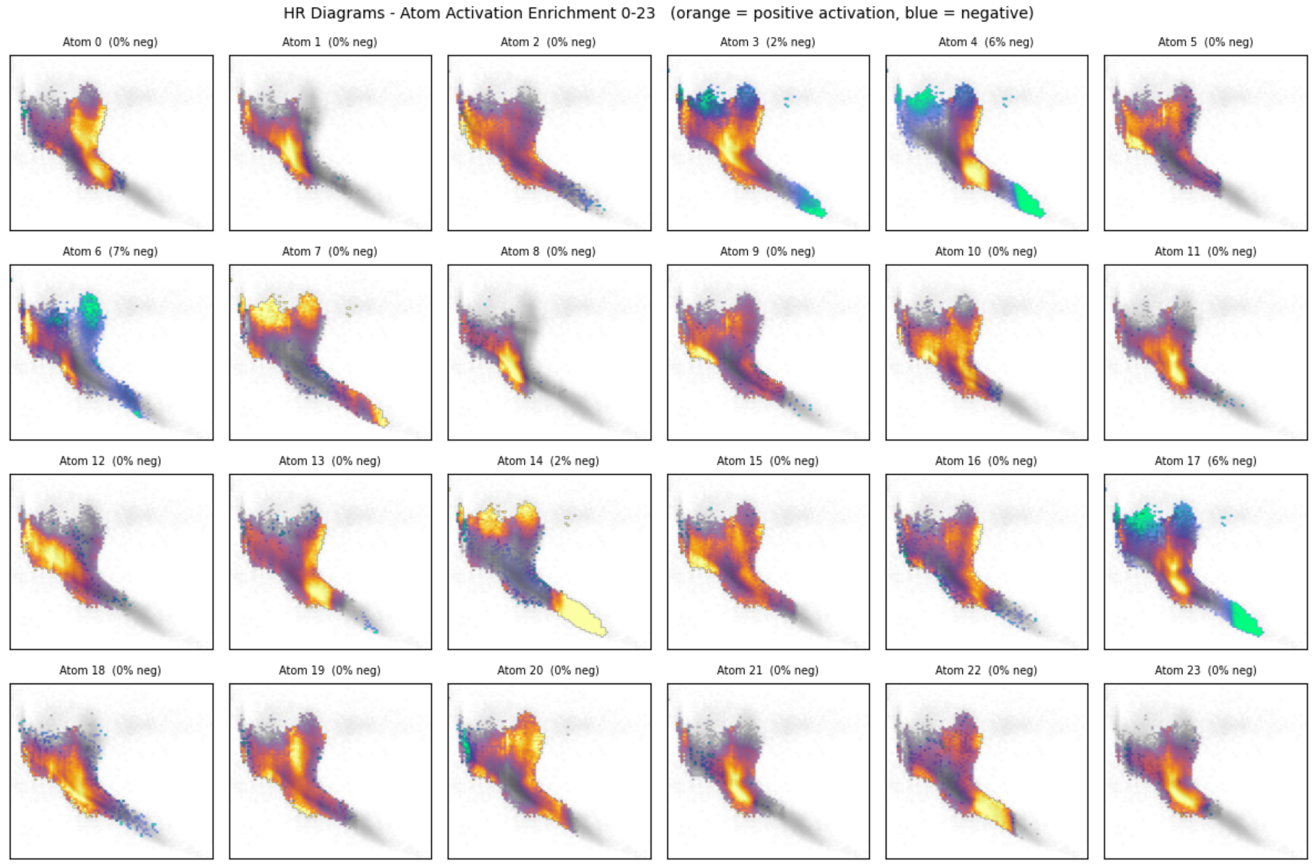

Figure 15: Hertzsprung-Russel diagrams with plotted atom enrichment of D_C. Absolute G magnitude with G-band extinction subtracted on the y-axes, BP-RP magnitude, with reddening subtracted, on the x-axes. Light gray background contains the entire sample population. Heatmap is the fraction of the atoms fired on, low activations are dark grey, negative activations shown in blue/green. Considerable distortion of color concentrated atoms is noticeable in the HR-diagram with extinction factored in, meaning that this is likely a very significant factor in the indecisive atoms of the dictionaries.

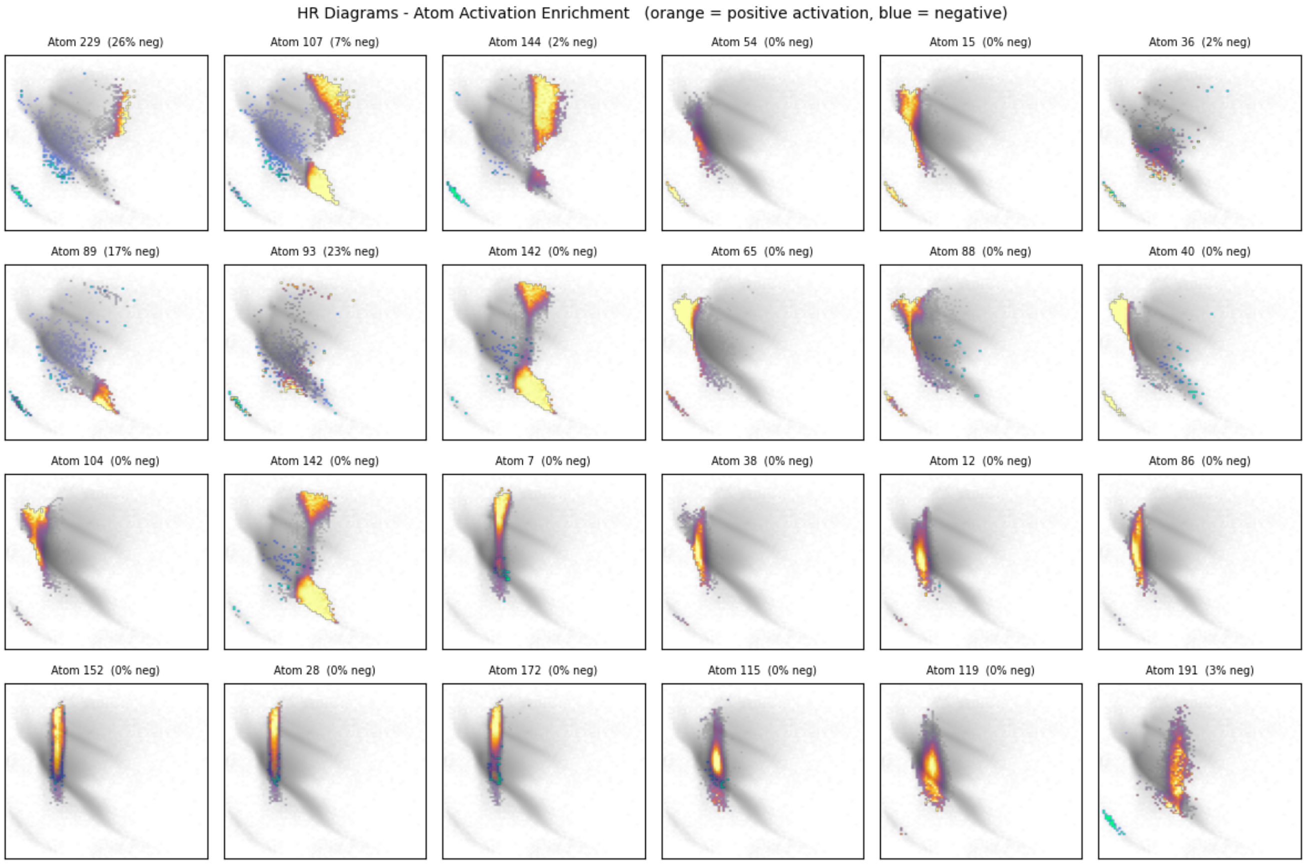

Figure 16: Hertzsprung-Russel diagrams with plotted atom enrichment of D_H. Absolute G magnitude on the y-axes, BP-RP magnitude on the x-axes. Light gray background contains the entire sample population. Heatmap is the fraction of the atoms fired on, low activations are dark grey, negative activations shown in blue/green. The atoms shown in this figure will be listed in this section.

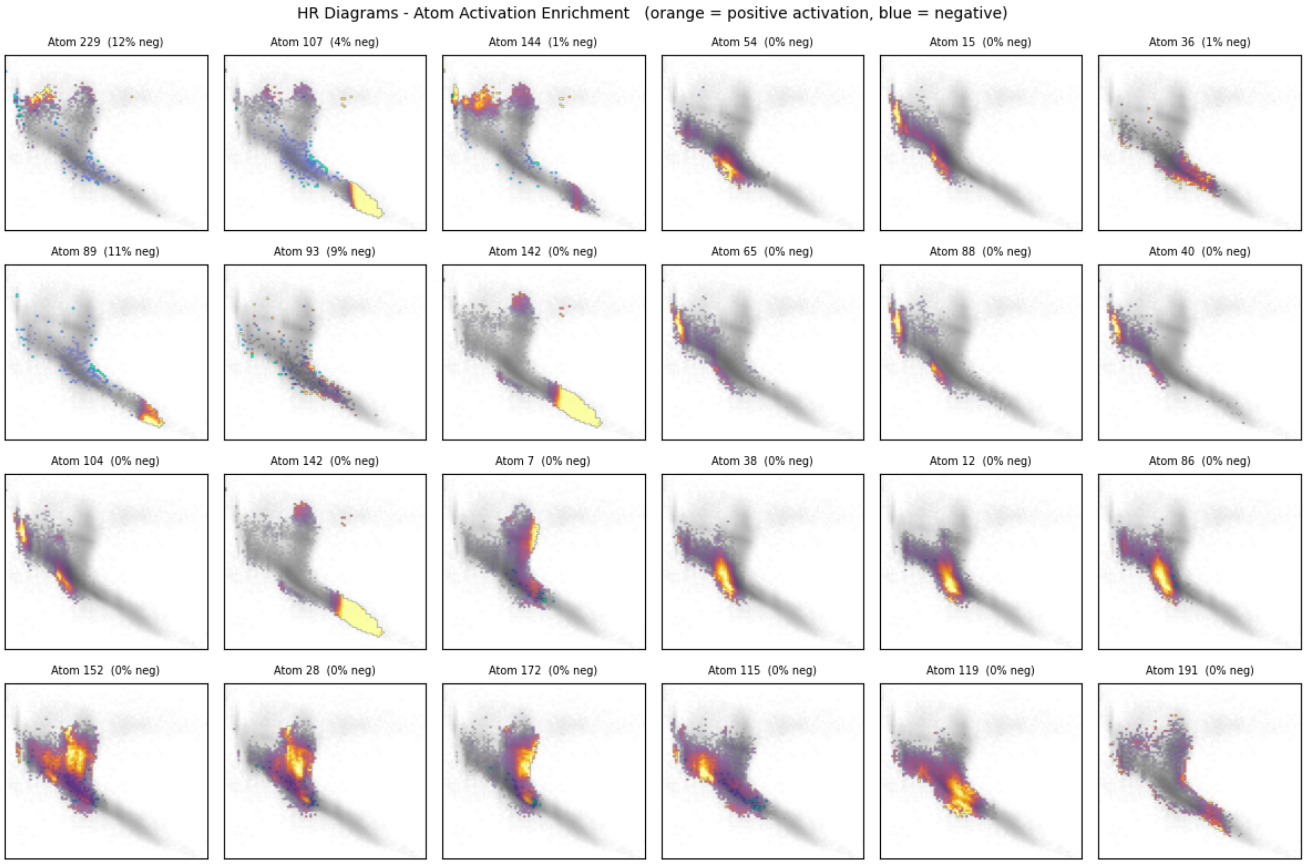

Figure 17: Hertzsprung-Russel diagrams with plotted atom enrichment of D_H. Absolute G magnitude with G-band extinction subtracted on the y-axes, BP-RP magnitude, with reddening subtracted, on the x-axes. Light gray background contains the entire sample population. Heatmap is the fraction of the atoms fired on, low activations are dark grey, negative activations shown in blue/green. Once again, factoring in extinction causes significant distortion in the shape of the atom clusters. The atoms shown in this figure will be listed in this section.

Figure 16 and 17 contain the following atoms:
d_{229}: An LPV specialist atom. The atom with the highest overrepresentation of any atom in D_H, with an enrichment of `LP?` at $6.1\times, z = 137$ and `LP*` at $4.3\times, z = 98$, with its matched sources consisting for over 80% of long period variables and candidates. Additionally, we see many of the stars activated by this atom are high extinction stars that move from the red to the blue side of the HR diagram when extinction is taken into account. Also the atom with the highest two specializations: $\zeta_{\texttt{LP?}} = 0.348, \zeta_{\texttt{LP*}} = 0.213$. The atom has a very low mean sparsity of sources it encodes, approximately $s_{\text{eff}} = 5$.
d_{107}: An LPV specialist atom. The atom with the third highest overrepresentation of any atom in D_H, with enrichments for `LP*` and `LP?` at $4.1\times$, with $z = 96$ and $z = 87$ respectively. Also the atom with the third and fourth highest specializations: $\zeta_{\texttt{LP?}} = 0.173, \zeta_{\texttt{LP*}} = 0.210$. The atom has a very low mean sparsity of sources it encodes, approximately $s_{\text{eff}} = 5$.
d_{144}: The atom identified earlier to have almost every carbon star in its top 1% activations, with

97% of carbon stars having it contained in their sparse code.
d_{54} and d_{15}: Atoms that have extremely high overrepresentations of white dwarfs, active galactic nuclei and quasars in their top 1% activations. d_{54} also has the fourth highest z_{max} of any atom in D_H.
d_{36}, d_{89} and d_{93}: Atoms that have an extremely high hit rate at top 1% activations, meaning most of the sources they encode have been measured and are stored in SIMBAD. They each have overrepresentations of Mira variables.
d_{142}: The only atom that specialized in emission line stars, containing approximately 60% of them, being the atom with the 5th highest ζ_{max} of all atoms in D_H. This atom is also the highest importance atom for T_{eff}.
d_{65}, d_{88} and d_{40}: Atoms with very high coverage of white dwarfs, active galactic nuclei, quasars and hot subdwarf candidates, the latter category of which contains atom 40 and 65 in its code at all times, and contains atom 88 in its code approximately 97% of the time.
d_{104} and d_7: The atoms with the highest importance for predicting $[M/H]$ and $\log g$ respectively.
d_{38}, d_{12} and d_{86}: The atoms with the highest correlation to G-magnitude, meaning these may contain instrumental artifacts in some way.
d_{51}: The atom with the highest enrichment of all atoms in D_H at top 1% activation, having an enrichment of 556× for galaxies, containing 37 out of 44 galaxies in its top 1% activations.
d_{152}, d_{28} and d_{172}: Atoms that largely consist of background stars.
d_{115} and d_{119}: The two atoms with the highest statistical overrepresentations of eclipsing binaries. Atoms that overrepresent these objects are very common, but typically do not have them highly overrepresented.
d_{191}: The only atom that has white dwarfs as the most dominant fraction when queried across all activations.

After training HGB models to predict labels using features and coefficients, we determine the following R^2 on the test set:

Metric	D_C atoms	Coefficients	D_H atoms
$[M/H]$	0.714	0.826	0.724
T_{eff}	0.915	0.941	0.916
$\log g$	0.756	0.885	0.847

Table 10: R^2 of HGB models trained on the data from [Andrae et al., 2023] and tested on [Abdurro'uf et al., 2022].

We observe that R^2 values are always lower for dictionaries than for raw coefficients, which is an expected observation, as dictionaries will always reduce the data stored in them, and never create new data. The score discrepancy measures what was discarded by the sparsity constraint. D_H achieves consistently higher R^2 than D_C, pointing to the fact that D_C is too far compressed to resolve these labels as accurately as D_H, which has a median sparsity over double that of D_C. Additionally, using the tables in Section C of the appendix, we observe that no single atom dominates label prediction, with the highest importance being 0.47. D_C appears to achieve a fairly concentrated prediction, relying on a small set of atoms to make an accurate prediction. We also observe that one raw coefficient is very good at predicting both T_{eff} and $\log g$.

4.5.1 Conclusion

From the HR diagrams, we can conclude that D_C does indeed have atoms that are much more spread across different stellar populations, whereas D_H more readily contains quite narrow clusters, although many less specialist atoms such as 152, 28 and 172 are much more spread out across the HR diagram.

From Table 10, we can conclude that compression costs least where information is most robust, as the decrease in R^2 of T_{eff} (2.8%) is far smaller than the decrease in R^2 for $\log g$ (14.6%) and $[M/H]$ (13.6%), neither of which are as readily apparent as T_{eff}. This is likely because temperature is encoded into the rough continuum shape, whereas metallicity is blended into line depths and other fine structure, which sparse coding likely discards first.

The fact this is the case is also backed up by the coefficient importances, as $\text{imp}_{\texttt{rp03}} = 0.691$ for $\log g$, and $\text{imp}_{\texttt{bp01}} = 0.710$ for T_{eff}, whereas the highest importance raw coefficient for $[M/H]$ is $\text{imp}_{\texttt{bp08}} = 0.118$, a somewhat higher-order coefficient, with a much bigger spread across other raw coefficients. Both dictionaries dilute each label in comparison to the raw coefficients, with the exception of D_C concentrating $[M/H]$, which is arguably an improvement for interpretability, though not by much. D_H dilutes the label importances at a consistently higher rate, likely due to highly cosine-similar atoms taking away parts of the prediction.

4.6 Atoms in Flux Space

Firstly, we determine that the atom signal-to-noise ratio of any atom in both dictionaries does not fall below 34. Additionally, the mean measured atom signal-to-noise ratios are 137 and 139 for D_C and D_H respectively. This is largely due to the representation of `bp_00` in each atom, as it is a coefficient with a very large signal-to-noise ratio.

Secondly, we look at the correspondence of a small set of selected features from Table 7. We find 325 and 4750 peaks/troughs above 3σ for D_C and D_H respectively, with all atoms having at least one feature. Many of these are false positives, but 38 and 368 of these fall in a known spectral line for D_C and D_H respectively. The entire spectrum spans 684 nm, with wavelength windows covering 23 nm, meaning about 3.4% of the spectrum is covered. The rate of observed features in these windows exceeds the expected rate by a factor of $3.4\times, z = 8.1$ and $2.3\times, z = 16.2$. This is evidence that some atoms encode identifiable spectral features rather than continuum shape alone.

Spectral line	(D_C)	(D_H)
Hα	12	114
Hβ	4	36
Hγ	6	49
Ca II K	5	75
Ca II H	4	60
Na I D	1	2
Mg I b	6	32

Table 11: Spectral line counts, detected in atoms of both dictionaries.

The following figures are representations of atoms in both dictionaries in flux space, which

Atom	Line	Type	SNR
12	Hα	absorption	12.157062
3	Ca II K	absorption	6.730269
17	Hα	absorption	5.214912
5	Hα	absorption	5.002140
23	Mg I b	absorption	4.354033
2	Ca II H	absorption	4.307493
8	Hα	absorption	4.218751
10	Hα	absorption	3.916823
18	Hα	absorption	3.898866
1	Ca II H	absorption	3.768701
8	Ca II K	absorption	3.709094
6	Hα	absorption	3.576402
10	Ca II K	absorption	3.414504
18	Ca II K	absorption	3.313488
12	Hβ	absorption	3.207874

Table 12: Most significant spectral features of atoms in D_C identified by signal-to-noise ratio.

Atom	Line	Type	SNR
112	Hα	absorption	8.927103
40	Hα	absorption	8.520469
99	Hα	absorption	8.466310
17	Hα	absorption	8.310777
88	Hα	emission	7.469739
56	Hα	absorption	7.232686
90	Ca II H	absorption	6.691134
61	Hα	absorption	6.504998
19	Hα	absorption	6.307916
69	Hα	absorption	6.147154
245	Hα	absorption	5.816514
70	Hα	emission	5.353859
194	Ca II K	absorption	5.209357
176	Hα	absorption	4.983453
30	Ca II K	absorption	4.862756

Table 13: Most significant spectral features of atoms in D_H identified by signal-to-noise ratio.

reveals some of their characteristics. For instance, in Figure 18, a clear example of spectral lines can be seen in atom 12.

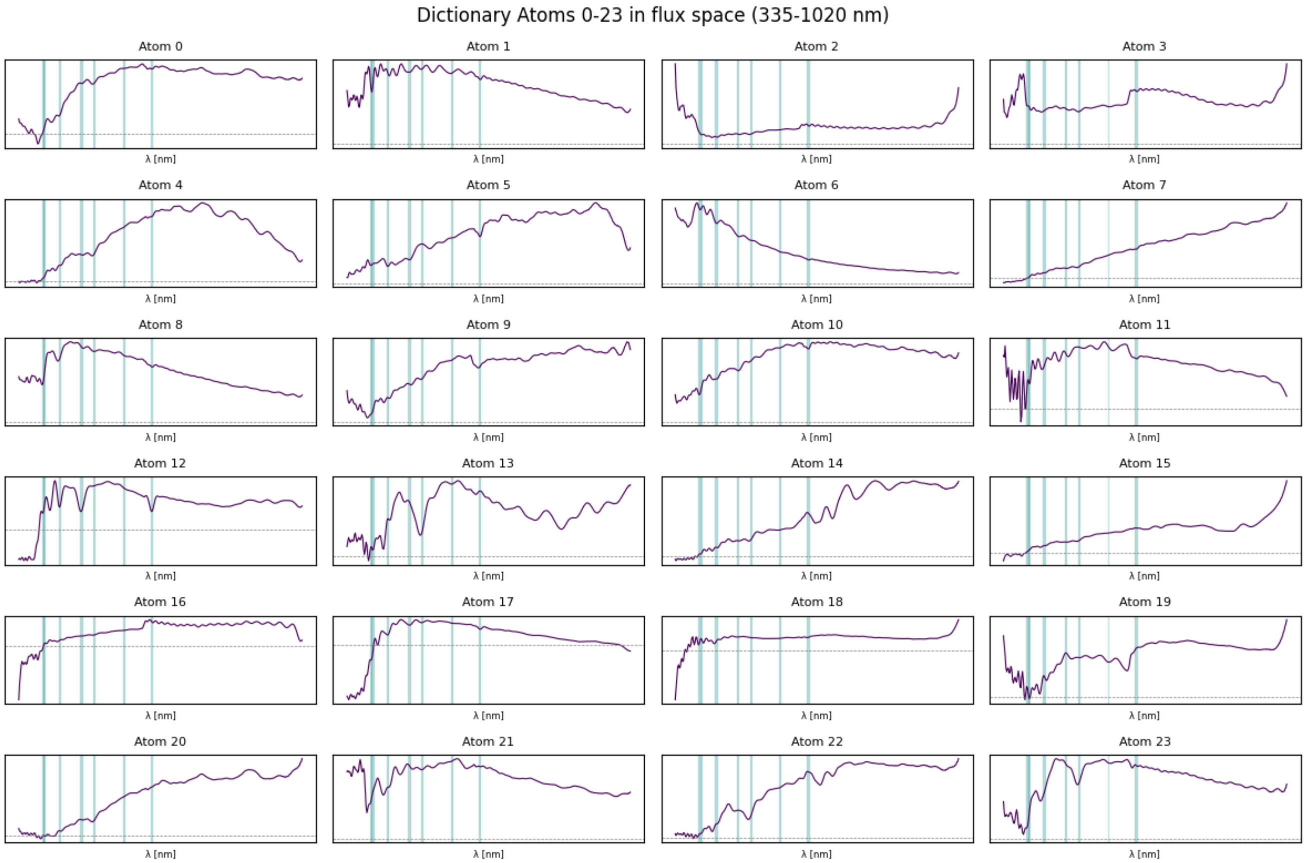

Figure 18: The dictionary atoms of D_C, with the spectral features from Table 7 highlighted in green.

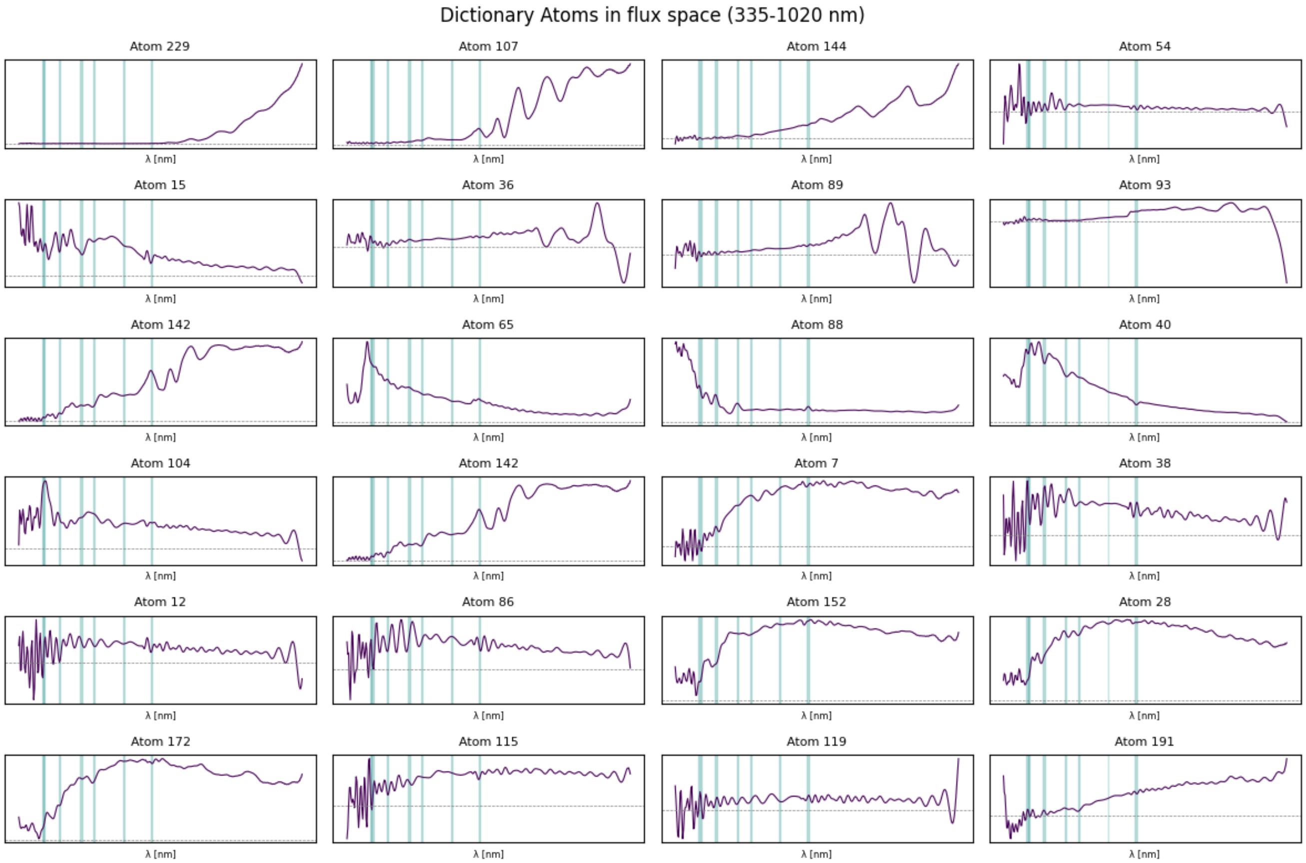

Figure 19: The dictionary atoms of D_H, with the spectral features from Table 7 highlighted in green. This figure contains the same atoms as are displayed in Figure 17 and 16. d_{229} and d_{107} show to be very red, which is to be expected for LPV specialized spectra. d_{65}, d_{88} and d_{40} show very blue spectra, as they specialize in blue objects such as white dwarfs, AGNs and QSOs. d_{142}, which is the only atom that specializes in emission line stars, has a clear emission line in the $H\alpha$ window. d_{38}, d_{12} and d_{86} have high correlation with G-magnitude, meaning the shapes of their spectra possibly correspond to instrument noise of some kind.

4.7 Anomalous Spectra

We rank the 0.1% of spectra with the highest error using χ^2, finding that 1,577/2,046 (D_C) and 1,429/2,046 (D_H) spectra are matched with known SIMBAD objects, an approximately 20-fold increase for both dictionaries from 3.8% to 77% and 70% coverage for D_C and D_H respectively. Poorly represented spectra are not simply noisy outliers, but overwhelmingly known objects of specific classes, namely white dwarfs, active galactic nuclei, quasars and Mira variables, for both D_C and D_H, meaning this poor compressibility is not related to the individual dictionaries themselves, with regular background stars being highly underrepresented ($z < -27$) for both dictionaries. Interestingly, each of these 4 classes has somewhat specialized atoms that correspond to them, but they are still poorly reconstructed. Additionally, several rare classes have very high overrepresentations, though we limit the tables to listing only otypes that have more than

5 representatives in the worst 0.1% errors, so that singular outliers don't give their entire classes extremely high overrepresentations.

Otype	**Count**	**Frac.**	**Pop. Frac.**	**Enrichment**	z
WD*	695	0.44071	0.01228	35.9x	154
AGN	299	0.18960	0.00602	31.5x	94
Mi*	285	0.18072	0.00838	21.6x	75
QSO	99	0.06278	0.00198	31.7x	54
Sy1	7	0.00444	0.00013	34.1x	15
WR*	8	0.00507	0.00025	20.5x	12
PN	5	0.00317	0.00012	27.0x	11
C*	24	0.01522	0.00257	5.9x	10
PM*	22	0.01395	0.03779	0.37x	-5
Em*	6	0.00381	0.02570	0.15x	-5
LP*	60	0.03805	0.08859	0.43x	-7
LP?	22	0.01395	0.07025	0.20x	-8
*	6	0.00381	0.54507	0.0070x	-29

Table 14: SIMBAD object-type distribution and enrichment of the top 0.1% highest χ^2 for D_C, excluding types that have less than 5 representatives in this set.

Otype	**Count**	**Frac.**	**Pop. Frac.**	**Enrichment**	z
WD*	707	0.49475	0.01228	40.3x	165
AGN	341	0.23863	0.00602	39.7x	113
QSO	113	0.07908	0.00198	39.9x	65
Mi*	131	0.09167	0.00838	10.9x	34
Sy1	8	0.00560	0.00013	43.0x	18
WR*	8	0.00560	0.00025	22.6x	13
PN	5	0.00350	0.00012	29.8x	12
LM*	9	0.00630	0.00241	2.6x	3
HS?	5	0.00350	0.00197	1.8x	1
PM*	23	0.01610	0.03779	0.43x	-4
Em*	9	0.00630	0.02570	0.25x	-5
LP*	23	0.01610	0.08859	0.18x	-9
EB*	6	0.00420	0.09332	0.045x	-11
*	13	0.00910	0.54507	0.017x	-27

Table 15: SIMBAD object-type distribution and enrichment of the top 0.1% highest χ^2 for D_H, excluding types that have less than 5 representatives in this set.

There are also approximately 500 more sources for each dictionary (which may overlap) that are part of this top 0.1% χ^2 population that are not recorded in SIMBAD. For these sources, we can use Gaia's Discrete Source Classifier to investigate. Gaia logs a probability it gives a source to be a white dwarf or a quasar (among some other object types).

Firstly, to determine the accuracy of these labels, we compare whether SIMBADs `QSO` and `AGN` labels correspond to a high p_{quasar}. We observe that the fraction of these SIMBAD matched objects where $p_{\text{quasar}} > 0.9$ is 0.961 and the fraction for which $p_{\text{quasar}} > 0.5$ is 0.986. However, not all $p_{\text{quasar}} > 0.5$ sources are SIMBAD classified quasars or AGNs. These sources (among the entire sample) consist of 457 AGNs, 149 quasars, 49 white dwarfs, 22 RR-Lyrae variables, 9 Seyfert 1 galaxies and 8 ordinary stars. For white dwarfs, DSC is much less accurate. For SIMBAD classified white dwarfs, $\mathrm{med}(\mathrm{p_{wd}}) = 0.9998$, however the fraction for which $p_{\mathrm{wd}} > 0.5$ is only 0.593, meaning that DSC often misclassifies white dwarfs as other objects.

The unmatched stars have the following DSC labels:

Class	(D_C)	(D_H)
Star	338	454
Quasar	125	146
White dwarf	4	11
Galaxy	2	4
Binary star	0	2

Table 16: Distribution of the number of sources that are not classified by SIMBAD over different Gaia DSC classes in the worst 0.1% errors.

The DSC classifications show a considerable difference in population. Though quasars are still highly represented, white dwarfs are not. However, as DSC often mistakes white dwarfs for regular stars, the star portion of this group may still hold a considerable amount of white dwarfs. It cannot be confidently concluded that the unmatched stars have a considerably different population distribution to matched stars.

When looking at the worst 25 unmatched sources, one source stands out. The χ^2 of the sources gradually increases, until reaching Gaia DR3 4658446877901307904, which has a catastrophically bad reconstruction. The dictionary appears to be completely incapable of representing this singular object. A possible cause for this is that this object retains very high values for coefficients, even in higher, more noisy coefficients. This object is also unique in the entire sample for having a sum of raw coefficients that is negative.

4.7.1 Conclusion

Poor compressibility recovers known rare classes at 20x base rate, consistently across both dictionaries, with white dwarfs, AGN/QSO and Mira variables dominating this population. Roughly 500 sources for each dictionary have no SIMBAD matched entry. The unmatched population does not appear to differ systematically from the matched one, as DSC classifies a similar fraction of quasars, with as a remainder ordinary stars, which is consistent with the observation that DSC often misclassifies white dwarfs as ordinary stars. The existence of a SIMBAD entry for a star therefore cannot be said to show a different population, rather than reflecting observational history, as the stars in this group may be misclassified white dwarfs, meaning that the unmatched and matched populations are very similar.

Lastly, Gaia DR3 4658446877901307904 may be an object of interest, although it may be an outlier because of faulty coefficients.

5 Further Research & Limitations

This thesis has three main limitations. Amplitude normalization in this thesis is one, as normalizing by `rp_00` does rely on the color of a spectrum to an extent, though it is the least biased coefficient. A better amplitude normalization could provide a less biased dictionary. The second is the fact that the sample is highly concentrated around the galactic disk. Though the sample was uniformly gathered from Gaia's database, higher populations in the galactic disk result in many chunks being located here. Stars around the galactic disk are more often heavily extincted than stars outside it, which is a confounding factor. Thirdly, extinction was not removed before dictionary learning, which means observed spectra may be reddened, so an atom encoding a broad continuum will conflate temperature with extinction. This is shown in resulting atoms, often containing many stars that are much bluer than they appear due to their extinction. A clear example of this is D_Hs d_{229}. Extinction causes clear clusters in the HR diagram to become much more smudged across different populations, and accounting for it may improve results drastically. Learning dictionaries on spectra where extinction is taken into account is left to future work.

Another possible improvement to make is using per-source coefficient errors rather than sample mean errors to train. This matters little when finding the optimal χ^2 of an entire sample, but could matter considerably while training and when determining anomalous spectra, though no significant signs of contamination of the latter are showing. The previously provided corrections by [Huang et al., 2024] may improve accuracy as well.

Using a larger sample size could provide more accurate data, and future Gaia data releases will provide even more XP spectra to be used. Future research may expand upon this subject using the entire catalog of XP spectra, as this is not within the scope of this thesis. There are many future experiments that can be done in the context of pattern mining using Gaia XP spectra that did not fit in the scope of this thesis. In the context of dictionary learning, one possibility we suggest is to use the sparse codes of sources to find source clusters, which may show certain properties more accurately than singular atoms. To do this, one would likely have to set all zeroed out values for each to a different value to "penalize" this in relation to an atom that *is* used, but with a small activation. Additional research is also necessary to properly identify Gaia DR3 4658446877901307904. Attempting to determine whether the atoms that have a high correlation to G-magnitude have a meaningful structure that can be used to learn more about Gaia's instrumental artifacts is also a suggested extension of this work.

6 Conclusion

After having performed the experiments of Section 4, we find that Gaia XP spectra can be represented as a sparse combination of learned atoms with a median compression of $8.71\times$ (interquartile range of 0.83), determined by finding the minimum median bit length for each source when represented by a dictionary with $\chi^2/dof \approx 1$, and dropping the residual in the bit representation. The compression of a PCA with similar reconstruction accuracy is approximately $13.6\times$. This entails that sparse dictionary learning is an inferior compressor for the sample, though it still contains considerably more information than a dense code of 8 directions, which is what PCA provides. The highest SDL compression is achieved with a dictionary that has $M = 24, \lambda = 3.0$. Thus, an overcomplete dictionary does not satisfy the MDL constraint. While the MDL constraint results in a dictionary that has $M = 24, \lambda = 3.0$, class discovery is more effective at higher atom count M, as atoms need not be as generalist for these dictionary sizes. A dictionary with $M = 256, \lambda = 7.0$ has consistently better atom label selectivity and captures more distinct classes, as seen in Table 6.

Atoms have some level of correspondence to known source properties, and are not spread generally throughout the population, but are not highly specialized. For instance, certain object types are recovered at very high rates in certain atoms, but these atoms do not exclusively contain this type. Dictionaries with lower M show much higher generalization when compared to dictionaries with higher M. For SIMBAD classes, we determine that atoms cannot exclusively capture a single class, but are shown to have a definite level of specialization, specifically for long period variable stars, white dwarfs and carbon stars, with higher specializations at higher atom count M.

The atom interpretability criteria we define in Section 2.7 were Plausibility, Stability, Distinctness, Non-instrumentality and Correspondence. For Plausibility, we find that all atoms consistently have a $SNR \gg 1$, mostly because they all include coefficients with a high SNR at high magnitude. For Stability, in Section 4.2, we find that the code for the same source is not very stable, with approximately 46% of codes matching when using optimally matched atoms, meaning the atoms can change depending on the random initialization. This means that a certain atom type, even if in both dictionaries, does not need to represent the same sources. However, it is clear that many specialized atoms persist at different random initializations, and even different M, meaning most "non-matching" atoms may be a sort of "filler" atom, while specialized atoms do still exist. For Distinctness, we determine that atoms are only moderately separated, with a median MCCS of 0.51 for D_C and one of 0.57 for D_H, with these values rising to 0.74 and 0.75 respectively at the upper quartile. This means that the encoder may readily pick between atoms that are not very specialized, without losing much accuracy. This finding shows that, though there are distinct, stable and specialized atoms, a lot of atoms are essentially "filler" atoms, which can be highly similar. This means that for many sources, the code may change considerably without significantly affecting reconstruction accuracy. For Non-instrumentality, we determine in Section 4.3 that many of the atoms have a correlation with apparent G magnitude, which is evidence for the fact that the atoms contain some instrumental artifacts, as this points to the fact that atoms do not exclusively obtain physical source properties, but also depend on distance of the source. For Correspondence, there is evidence that atoms are somewhat interpretable in flux space. Specifically, the spectral lines in Table 7 are overrepresented at $3.4\times, z = 8.1$ for D_C and $2.3\times, z = 16.2$ for D_H, which is evidence for this claim. Larger, continuum structure can also be easily identified in atoms. It is reasonable to hypothesize there are other spectral structures that show up in flux space as well.

The atoms are likely not capable of revealing clusters of sources, clustered by physical properties of the source, that are not yet captured by Gaia or external labels, as they are not entirely effective at distinguishing between classes or properties of sources that *are* captured by external labels. There is still a possibility is that atoms that are highly correlated to instrumental properties may reveal clustering that is previously unknown, though not revealing properties of the sources themselves, though this would require further research.

The codes of sources that are poorly reconstructed do not necessarily correspond to anomalous objects, but correspond mostly to white dwarfs and active galactic nuclei/quasars, as shown in Section 4.7. There is no conclusive evidence that the SIMBAD-matched population differs considerably from the population that was not matched.

Disclosure of AI Usage

Claude by Anthropic, a large language model, was used to aid in the preparation of this thesis. Specifically, it was used for the formatting of LaTeX, especially tables, Python assistance, and as an independent proof reader to ensure that no errors in writing were missed. The author takes full responsibility for all derivations and results of this thesis, as they are all the authors original work, unless cited.

References

[Abdurro'uf et al., 2022] Abdurro'uf, Accetta, K., Aerts, C., Silva Aguirre, V., Ahumada, R., Ajgaonkar, N., Filiz Ak, N., Alam, S., Allende Prieto, C., Almeida, A., Anders, F., Anderson, S. F., Andrews, B. H., Anguiano, B., Aquino-Ortíz, E., Aragón-Salamanca, A., Argudo-Fernández, M., Ata, M., Aubert, M., Avila-Reese, V., Badenes, C., Barbá, R. H., Barger, K., Barrera-Ballesteros, J. K., Beaton, R. L., Beers, T. C., Belfiore, F., Bender, C. F., Bernardi, M., Bershady, M. A., Beutler, F., Bidin, C. M., Bird, J. C., Bizyaev, D., Blanc, G. A., Blanton, M. R., Boardman, N. F., Bolton, A. S., Boquien, M., Borissova, J., Bovy, J., Brandt, W. N., Brown, J., Brownstein, J. R., Brusa, M., Buchner, J., Bundy, K., Burchett, J. N., Bureau, M., Burgasser, A., Cabang, T. K., Campbell, S., Cappellari, M., Carlberg, J. K., Wanderley, F. C., Carrera, R., Cash, J., Chen, Y.-P., Chen, W.-H., Cherinka, B., Chiappini, C., Choi, P. D., Chojnowski, S. D., Chung, H., Clerc, N., Cohen, R. E., Comerford, J. M., Comparat, J., da Costa, L., Covey, K., Crane, J. D., Cruz-Gonzalez, I., Culhane, C., Cunha, K., Dai , Y. S. Damke, G., Darling, J., Davidson Jr., J. W., Davies, R., Dawson, K., De Lee, N., Diamond-Stanic, A. M., Cano-Díaz, M., Sánchez, H. D., Donor, J., Duckworth, C., Dwelly, T., Eisenstein, D. J., Elsworth, Y. P., Emsellem, E., Eracleous, M., Escoffier, S., Fan, X., Farr, E., Feng, S., Fernández-Trincado, J. G., Feuillet, D., Filipp, A., Fillingham, S. P., Frinchaboy, P. M., Fromenteau, S., Galbany, L., García, R. A., García-Hernández, D. A., Ge, J., Geisler, D., Gelfand, J., Géron, T., Gibson, B. J., Goddy, J., Godoy-Rivera, D., Grabowski, K., Green, P. J., Greener, M., Grier, C. J., Griffith, E., Guo, H., Guy, J., Hadjara, M., Harding, P., Hasselquist, S., Hayes, C. R., Hearty, F., Hernández, J., Hill, L., Hogg, D. W., Holtzman, J. A., Horta, D., Hsieh, B.-C., Hsu, C.-H., Hsu, Y.-H., Huber, D., Huertas-Company, M., Hutchinson, B., Hwang, H. S., Ibarra-Medel, H. J., Chitham, J. I., Ilha, G. S., Imig, J., Jaekle, W., Jayasinghe, T., Ji, X., Johnson, J. A., Jones, A., Jönsson, H., Katkov, I., Khalatyan, D. A., Kinemuchi, K., Kisku, S., Knapen, J. H., Kneib, J.-P., Kollmeier, J. A., Kong, M., Kounkel, M., Kreckel, K., Krishnarao, D., Lacerna, I., Lane, R. R., Langgin, R., Lavender, R., Law, D. R., Lazarz, D., Leung, H. W., Leung, H.-H., Lewis, H. M., Li, C., Li, R., Lian, J., Liang, F.-H., Lin , L. Lin, Y.-T., Lin, S., Lintott, C., Long, D., Longa-Peña, P., López-Cobá, C., Lu, S., Lundgren, B. F., Luo, Y., Mackereth, J. T., de la Macorra, A., Mahadevan, S., Majewski, S. R., Manchado, A., Mandeville, T., Maraston, C., Margalef-Bentabol, B., Masseron, T., Masters, K. L., Mathur, S., McDermid, R. M., Mckay, M., Merloni, A., Merrifield, M., Meszaros, S., Miglio, A., Di Mille, F., Minniti, D., Minsley, R., Monachesi, A., Moon, J., Mosser, B., Mulchaey, J., Muna, D., Muñoz, R. R., Myers, A. D., Myers, N., Nadathur, S., Nair, P., Nandra, K., Neumann, J., Newman, J. A., Nidever, D. L., Nikakhtar, F., Nitschelm, C., O'Connell, J. E., Garma-Oehmichen, L., Luan Souza de Oliveira, G., Olney, R., Oravetz, D., Ortigoza-Urdaneta, M., Osorio, Y., Otter, J., Pace, Z. J., Padilla, N., Pan, K., Pan, H.-A., Parikh, T., Parker, J., Peirani, S., Peña Ramírez, K., Penny, S., Percival, W. J., Perez-Fournon, I., Pinsonneault, M., Poidevin, F., Poovelil, V. J., Price-Whelan, A. M., Bárbara de Andrade Queiroz, A., Raddick, M. J., Ray, A., Rembold, S. B., Riddle, N., Riffel, R. A., Riffel, R., Rix, H.-W., Robin, A. C., Rodríguez-Puebla, A., Roman-Lopes, A., Román-Zúñiga, C., Rose, B., Ross, A. J., Rossi, G., Rubin, K. H. R., Salvato, M., Sánchez, S. F., Sánchez-Gallego, J. R., Sanderson, R., Santana Rojas, F. A., Sarceno, E., Sarmiento, R., Sayres, C., Sazonova, E., Schaefer, A. L., Schiavon, R., Schlegel, D. J., Schneider, D. P., Schultheis, M., Schwope, A., Serenelli, A., Serna, J., Shao, Z., Shapiro, G., Sharma, A., Shen, Y., Shetrone, M., Shu, Y., Simon, J. D., Skrutskie, M. F., Smethurst, R., Smith, V., Sobeck, J., Spoo, T., Sprague, D., Stark, D. V., Stassun, K. G., Steinmetz, M., Stello, D., Stone-Martinez, A., Storchi-Bergmann, T., Stringfellow, G. S., Stutz, A., Su, Y.-C., Taghizadeh-Popp, M., Talbot, M. S., Tayar, J., Telles, E., Teske, J., Thakar, A., Theissen, C., Tkachenko, A., Thomas, D., Tojeiro, R., Hernandez Toledo, H., Troup, N. W., Trump, J. R., Trussler, J., Turner, J., Tuttle, S., Unda-Sanzana, E., Vázquez-Mata, J. A., Valentini, M., Valenzuela, O., Vargas-González, J., Vargas-Magaña, M., Alfaro, P. V., Villanova, S., Vincenzo, F., Wake, D., Warfield, J. T., Washington, J. D., Weaver, B. A., Weijmans, A.-M., Weinberg, D. H., Weiss, A., Westfall, K. B., Wild, V., Wilde, M. C., Wilson, J. C., Wilson, R. F., Wilson, M., Wolf, J., Wood-Vasey, W. M., Yan , R. Zamora, O., Zasowski, G., Zhang, K., Zhao, C., Zheng, Z., Zheng, Z., and Zhu, K. (2022). The seventeenth data release of the sloan digital sky surveys: Complete release of manga, mastar, and apogee-2 data. *The Astrophysical Journal Supplement Series*, 259(2):35.

[Andrae et al., 2023] Andrae, R., Rix, H.-W., and Chandra, V. (2023). Robust data-driven metallicities for 175 million stars from gaia xp spectra. *The Astrophysical Journal Supplement Series*, 267(1):8.

[Carrasco et al., 2021] Carrasco, J. M., Weiler, M., Jordi, C., Fabricius, C., De Angeli, F., Evans, D. W., van Leeuwen, F., Riello, M., and Montegriffo, P. (2021). Internal calibration of ¡i¿gaia¡/i¿ bp/rp low-resolution spectra. *Astronomy amp; Astrophysics*, 652:A86.

[CAS, 2024] CAS (2024). Lamost dr12. https://www.lamost.org/dr12/. Accessed July 2026.

[CDS, 2026] CDS (2026). Simbad object type labels. https://simbad.cds.unistra.fr/guide/otypes.labels.txt. Accessed July 2026.

[Chan et al., 2003] Chan, B. H. P., Mitchell, D. A., and Cram, L. E. (2003). Archetypal analysis of galaxy spectra. *Monthly Notices of the Royal Astronomical Society*, 338(3):790–795.

[Chen et al., 2006] Chen, S., Donoho, D., and Saunders, M. (2006). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61.

[Connolly et al., 1995] Connolly, A. J., Szalay, A. S., Bershady, M. A., Kinney, A. L., and Calzetti, D. (1995). Spectral classification of galaxies: an orthogonal approach. *The Astronomical Journal*, 110:1071.

[Creevey et al., 2023] Creevey, O. L., Sordo, R., Pailler, F., Frémat, Y., Heiter, U., Thévenin, F., Andrae, R., Fouesneau, M., Lobel, A., Bailer-Jones, C. A. L., Garabato, D., Bellas-Velidis, I., Brugaletta, E., Lorca, A., Ordenovic, C., Palicio, P. A., Sarro, L. M., Delchambre, L., Drimmel, R., Rybizki, J., Torralba Elipe, G., Korn, A. J., Recio-Blanco, A., Schultheis, M. S., De Angeli, F., Montegriffo, P., Abreu Aramburu, A., Accart, S., Álvarez, M. A., Bakker, J., Brouillet, N., Burlacu, A., Carballo, R., Casamiquela, L., Chiavassa, A., Contursi, G., Cooper, W. J., Dafonte, C., Dapergolas, A., de Laverny, P., Dharmawardena, T. E., Edvardsson, B., Le Fustec, Y., García-Lario, P., García-Torres, M., Gomez, A., González-Santamaría, I., Hatzidimitriou, D., Jean-Antoine Piccolo, A., Kontiza, M., Kordopatis, G., Lanzafame, A. C., Lebreton, Y., Licata, E. L., Lindstrøm, H. E. P., Livanou, E., Magdaleno Romeo, A., Manteiga, M., Marocco, F., Marshall, D. J., Mary, N., Nicolas, C., Pallas-Quintela, L., Panem, C., Pichon, B., Poggio, E., Riclet, F., Robin, C., Santoveña, R., Silvelo, A., Slezak, I., Smart, R. L., Soubiran, C., Süveges, M., Ulla, A., Utrilla, E., Vallenari, A., Zhao, H., Zorec, J., Barrado, D., Bijaoui, A., Bouret, J.-C., Blomme, R., Brott, I., Cassisi, S., Kochukhov, O., Martayan, C., Shulyak, D., and Silvester, J. (2023). ¡i¿gaia¡/i¿data release 3: Astrophysical parameters inference system (apsis). i. methods and content overview. *Astronomy amp; Astrophysics*, 674:A26.

[Cutler and Breiman, 1994] Cutler, A. and Breiman, L. (1994). Archetypal analysis. *Technometrics*, 36(4):338–347.

[De Angeli et al., 2023] De Angeli, F., Weiler, M., Montegriffo, P., Evans, D. W., Riello, M., Andrae, R., Carrasco, J. M., Busso, G., Burgess, P. W., Cacciari, C., Davidson, M., Harrison, D. L., Hodgkin, S. T., Jordi, C., Osborne, P. J., Pancino, E., Altavilla, G., Barstow, M. A., Bailer-Jones, C. A. L., Bellazzini, M., Brown, A. G. A., Castellani, M., Cowell, S., Delchambre, L., De Luise, F., Diener, C., Fabricius, C., Fouesneau, M., Frémat, Y., Gilmore, G., Giuffrida, G., Hambly, N. C., Hidalgo, S., Holland, G., Kostrzewa-Rutkowska, Z., van Leeuwen, F., Lobel, A., Marinoni, S., Miller, N., Pagani, C., Palaversa, L., Piersimoni, A. M., Pulone, L., Ragaini, S., Rainer, M., Richards, P. J., Rixon, G. T., Ruz-Mieres, D., Sanna, N., Sarro, L. M., Rowell, N., Sordo, R., Walton, N. A., and Yoldas, A. (2023). ¡i¿gaia¡/i¿data release 3: Processing and validation of bp/rp low-resolution spectral data. *Astronomy amp; Astrophysics*, 674:A2.

[Efron et al., 2004] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2).

[Fouesneau et al., 2023] Fouesneau, M., Frémat, Y., Andrae, R., Korn, A. J., Soubiran, C., Kordopatis, G., Vallenari, A., Heiter, U., Creevey, O. L., Sarro, L. M., de Laverny, P., Lanzafame, A. C., Lobel, A., Sordo, R., Rybizki, J., Slezak, I., Álvarez, M. A., Drimmel, R., Garabato, D., Delchambre, L., Bailer-Jones, C. A. L., Hatzidimitriou, D., Lorca, A., Le Fustec, Y., Pailler, F., Mary, N., Robin, C., Utrilla, E., Abreu Aramburu, A., Bakker, J., Bellas-Velidis, I., Bijaoui, A., Blomme, R., Bouret, J.-C., Brouillet, N., Brugaletta, E., Burlacu, A., Carballo, R., Casamiquela, L., Chaoul, L., Chiavassa, A., Contursi, G., Cooper, W. J., Dafonte, C., Demouchy, C., Dharmawardena, T. E., García-Lario, P., García-Torres, M., Gomez, A., González-Santamaría, I., Jean-Antoine Piccolo, A., Kontizas, M., Lebreton, Y., Licata, E. L., Lindstrøm, H. E. P., Livanou, E., Magdaleno Romeo, A., Manteiga, M., Marocco, F., Martayan, C., Marshall, D. J., Nicolas, C., Ordenovic, C., Palicio, P. A., Pallas-Quintela, L., Pichon, B., Poggio, E., Recio-Blanco, A., Riclet, F., Santoveña, R., Schultheis, M. S., Segol, M., Silvelo, A., Smart, R. L., Süveges, M., Thévenin, F., Torralba Elipe, G., Ulla, A., van Dillen, E., Zhao, H., and Zorec, J. (2023). ¡i¿gaia¡/i¿data release 3: Apsis. ii. stellar parameters. *Astronomy amp; Astrophysics*, 674:A28.

[Gaia Collaboration et al., 2016] Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., Brown, A. G. A., Vallenari, A., Babusiaux, C., Bailer-Jones, C. A. L., Bastian, U., Biermann, M., Evans, D. W., Eyer, L., Jansen, F., Jordi, C., Klioner, S. A., Lammers, U., Lindegren, L., Luri, X., Mignard, F., Milligan, D. J., Panem, C., Poinsignon, V., Pourbaix, D., Randich, S., Sarri, G., Sartoretti, P., Siddiqui, H. I., Soubiran, C., Valette, V., van Leeuwen, F., Walton, N. A., Aerts, C., Arenou, F., Cropper, M., Drimmel, R., Høg, E., Katz, D., Lattanzi, M. G., O'Mullane, W., Grebel, E. K., Holland, A. D., Huc, C., Passot, X., Bramante, L., Cacciari, C., Castañeda, J., Chaoul, L., Cheek, N., De Angeli, F., Fabricius, C., Guerra, R., Hernández, J., Jean-Antoine-Piccolo, A., Masana, E., Messineo, R., Mowlavi, N., Nienartowicz, K., Ordóñez-Blanco, D., Panuzzo, P., Portell, J., Richards, P. J., Riello, M., Seabroke, G. M., Tanga, P., Thévenin, F., Torra, J., Els, S. G., Gracia-Abril, G., Comoretto, G., Garcia-Reinaldos, M., Lock, T., Mercier, E., Altmann, M., Andrae, R., Astraatmadja, T. L., Bellas-Velidis, I., Benson, K., Berthier, J., Blomme, R., Busso, G., Carry, B., Cellino, A., Clementini, G., Cowell, S., Creevey, O., Cuypers, J., Davidson, M., De Ridder, J., de Torres, A., Delchambre, L., Dell'Oro, A., Ducourant, C., Frémat, Y., García-Torres, M., Gosset, E., Halbwachs, J.-L., Hambly, N. C., Harrison, D. L., Hauser, M., Hestroffer, D., Hodgkin, S. T., Huckle, H. E., Hutton, A., Jasniewicz, G., Jordan, S., Kontizas, M., Korn, A. J., Lanzafame, A. C., Manteiga, M., Moitinho, A., Muinonen, K., Osinde, J., Pancino, E., Pauwels, T., Petit, J.-M., Recio-Blanco, A., Robin, A. C., Sarro, L. M., Siopis, C., Smith, M., Smith, K. W., Sozzetti, A., Thuillot, W., van Reeven, W., Viala, Y., Abbas, U., Abreu Aramburu, A., Accart, S., Aguado, J. J., Allan, P. M., Allasia, W., Altavilla, G., Álvarez, M. A., Alves, J., Anderson, R. I., Andrei, A. H., Anglada Varela, E., Antiche, E., Antoja, T., Antón, S., Arcay, B., Atzei, A., Ayache, L., Bach, N., Baker, S. G., Balaguer-Núñez, L., Barache, C., Barata, C., Barbier, A., Barblan, F., Baroni, M., Barrado y Navascués, D., Barros, M., Barstow, M. A., Becciani, U., Bellazzini, M., Bellei, G., Bello García, A., Belokurov, V., Bendjoya, P., Berihuete, A., Bianchi, L., Bienaymé, O., Billebaud, F., Blagorodnova, N., Blanco-Cuaresma, S., Boch, T., Bombrun, A., Borrachero, R., Bouquillon, S., Bourda, G., Bouy, H., Bragaglia, A., Breddels, M. A., Brouillet, N., Brüsemeister, T., Bucciarelli, B., Budnik, F., Burgess, P., Burgon, R., Burlacu, A., Busonero, D., Buzzi, R., Caffau, E., Cambras, J., Campbell, H., Cancelliere, R., Cantat-Gaudin, T., Carlucci, T., Carrasco, J. M., Castellani, M., Charlot, P., Charnas, J., Charvet, P., Chassat, F., Chiavassa, A., Clotet, M., Cocozza, G., Collins, R. S., Collins, P., Costigan, G., Crifo, F., Cross, N. J. G., Crosta, M., Crowley, C., Dafonte, C., Damerdji, Y., Dapergolas, A., David, P., David, M., De Cat, P., de Felice, F., de Laverny, P., De Luise, F., De March, R., de Martino, D., de Souza, R., Debosscher, J., del Pozo, E., Delbo, M., Delgado, A., Delgado, H. E., di Marco, F., Di Matteo, P., Diakite, S., Distefano, E., Dolding, C., Dos Anjos, S., Drazinos, P., Durán, J., Dzigan, Y., Ecale, E., Edvardsson, B., Enke, H., Erdmann, M., Escolar, D., Espina, M., Evans, N. W., Eynard Bontemps, G., Fabre, C., Fabrizio, M., Faigler, S., Falcão, A. J., Farràs Casas, M., Faye, F., Federici, L., Fedorets, G., Fernández-Hernández, J., Fernique, P., Fienga, A., Figueras, F., Filippi, F., Findeisen, K., Fonti, A., Fouesneau, M., Fraile, E., Fraser, M., Fuchs, J., Furnell, R., Gai, M., Galleti, S., Galluccio, L., Garabato, D., García-Sedano, F., Garé, P., Garofalo, A., Garralda, N., Gavras, P., Gerssen, J., Geyer, R., Gilmore, G., Girona, S., Giuffrida, G., Gomes, M., González-Marcos, A., González-Núñez, J., González-Vidal, J. J., Granvik, M., Guerrier, A., Guillout, P., Guiraud, J., Gúrpide, A., Gutiérrez-Sánchez, R., Guy, L. P., Haigron, R., Hatzidimitriou, D., Haywood, M., Heiter, U., Helmi, A., Hobbs, D., Hofmann, W., Holl, B., Holland, G., Hunt, J. A. S., Hypki, A., Icardi, V., Irwin, M., Jevardat de Fombelle, G., Jofré, P., Jonker, P. G., Jorissen, A., Julbe, F., Karampelas, A., Kochoska, A., Kohley, R., Kolenberg, K., Kontizas, E., Koposov, S. E., Kordopatis, G., Koubsky, P., Kowalczyk, A., Krone-Martins, A., Kudryashova, M., Kull, I., Bachchan, R. K., Lacoste-Seris, F., Lanza, A. F., Lavigne, J.-B., Le Poncin-Lafitte, C., Lebreton, Y., Lebzelter, T., Leccia, S., Leclerc, N., Lecoeur-Taibi, I., Lemaitre, V., Lenhardt, H., Leroux, F., Liao, S., Licata, E., Lindstrøm, H. E. P., Lister, T. A., Livanou, E., Lobel, A., Löffler, W., López, M., Lopez-Lozano, A., Lorenz, D., Loureiro, T., MacDonald, I., Magalhães Fernandes, T., Managau, S., Mann, R. G., Mantelet, G., Marchal, O., Marchant, J. M., Marconi, M., Marie, J., Marinoni, S., Marrese, P. M., Marschalkó, G., Marshall, D. J., Martín-Fleitas, J. M., Martino, M., Mary, N., Matijevič, G., Mazeh, T., McMillan, P. J., Messina, S., Mestre, A., Michalik, D., Millar, N. R., Miranda, B. M. H., Molina, D., Molinaro, R., Molinaro, M., Molnár, L., Moniez, M., Montegriffo, P., Monteiro, D., Mor, R., Mora, A., Morbidelli, R., Morel, T., Morgenthaler, S., Morley, T., Morris, D., Mulone, A. F., Muraveva, T., Musella, I., Narbonne, J., Nelemans, G., Nicastro, L., Noval, L., Ordénovic, C., Ordieres-Meré, J., Osborne, P., Pagani, C., Pagano, I., Pailler, F., Palacin, H., Palaversa, L., Parsons, P., Paulsen, T., Pecoraro, M., Pedrosa, R., Pentikäinen, H., Pereira, J., Pichon, B., Piersimoni, A. M., Pineau, F.-X., Plachy, E., Plum, G., Poujoulet, E., Prša, A., Pulone, L., Ragaini, S., Rago, S., Rambaux, N., Ramos-Lerate, M., Ranalli, P., Rauw, G., Read, A., Regibo, S., Renk, F., Reylé, C., Ribeiro, R. A., Rimoldini, L., Ripepi, V., Riva, A., Rixon, G., Roelens, M., Romero-Gómez, M., Rowell, N., Royer, F., Rudolph, A., Ruiz-Dern, L., Sadowski, G., Sagristà Sellés, T., Sahlmann, J., Salgado, J., Salguero, E., Sarasso, M., Savietto, H., Schnorhk, A., Schultheis, M., Sciacca, E., Segol, M., Segovia, J. C., Segransan, D., Serpell, E., Shih, I.-C., Smareglia, R., Smart, R. L., Smith, C., Solano, E., Solitro, F., Sordo, R., Soria Nieto, S., Souchay, J., Spagna, A., Spoto, F., Stampa, U., Steele, I. A., Steidelmüller, H., Stephenson, C. A., Stoev, H., Suess, F. F., Süveges, M., Surdej, J., Szabados, L., Szegedi-Elek, E., Tapiador, D., Taris, F., Tauran, G., Taylor, M. B., Teixeira, R., Terrett, D., Tingley, B., Trager, S. C., Turon, C., Ulla, A., Utrilla, E., Valentini, G., van Elteren, A., Van Hemelryck, E., van Leeuwen, M., Varadi, M., Vecchiato, A., Veljanoski, J., Via, T., Vicente, D., Vogt, S., Voss, H., Votruba, V., Voutsinas, S., Walmsley, G., Weiler, M., Weingrill, K., Werner, D., Wevers, T., Whitehead, G., Wyrzykowski, , Yoldas, A., Žerjal, M., Zucker, S., Zurbach, C., Zwitter, T., Alecu, A., Allen, M., Allende Prieto, C., Amorim, A., Anglada-Escudé, G., Arsenijevic, V., Azaz, S., Balm, P., Beck, M., Bernstein, H.-H., Bigot, L., Bijaoui, A., Blasco, C., Bonfigli, M., Bono, G., Boudreault, S., Bressan, A., Brown, S., Brunet, P.-M., Bunclark, P., Buonanno, R., Butkevich, A. G., Carret, C., Carrion, C., Chemin, L., Chéreau, F., Corcione, L., Darmigny, E., de Boer, K. S., de Teodoro, P., de Zeeuw, P. T., Delle Luche, C., Domingues, C. D., Dubath, P., Fodor, F., Frézouls, B., Fries, A., Fustes, D., Fyfe, D., Gallardo, E., Gallegos, J., Gardiol, D., Gebran, M., Gomboc, A., Gómez, A., Grux, E., Gueguen, A., Heyrovsky, A., Hoar, J., Iannicola, G., Isasi Parache, Y., Janotto, A.-M., Joliet, E., Jonckheere, A., Keil, R., Kim, D.-W., Klagyivik, P., Klar, J., Knude, J., Kochukhov, O., Kolka, I., Kos, J., Kutka, A., Lainey, V., LeBouquin, D., Liu, C., Loreggia, D., Makarov, V. V., Marseille, M. G., Martayan, C., Martinez-Rubi, O., Massart, B., Meynadier, F., Mignot, S., Munari, U., Nguyen, A.-T., Nordlander, T., Ocvirk, P., O'Flaherty, K. S., Olias Sanz, A., Ortiz, P., Osorio, J., Oszkiewicz, D., Ouzounis, A., Palmer, M., Park, P., Pasquato, E., Peltzer, C., Peralta, J., Péturaud, F., Pieniluoma, T., Pigozzi, E., Poels, J., Prat, G., Prod'homme, T., Raison, F., Rebordao, J. M., Risquez, D., Rocca-Volmerange, B., Rosen, S., Ruiz-Fuertes, M. I., Russo, F., Sembay, S., Serraller Vizcaino, I., Short, A., Siebert, A., Silva, H., Sinachopoulos, D., Slezak, E., Soffel, M., Sosnowska, D., Straižys, V., ter Linden, M., Terrell, D., Theil, S., Tiede, C., Troisi, L., Tsalmantza, P., Tur, D., Vaccari, M., Vachier, F., Valles, P., Van Hamme, W., Veltz, L., Virtanen, J., Wallut, J.-M., Wichmann, R., Wilkinson, M. I., Ziaeepour, H., and Zschocke, S. (2016). The gaia mission. *Astronomy amp; Astrophysics*, 595:A1.

[Gaia Collaboration et al., 2023] Gaia Collaboration, Vallenari, A., Brown, A. G. A., Prusti, T., de Bruijne, J. H. J., Arenou, F., Babusiaux, C., Biermann, M., Creevey, O. L., Ducourant, C., Evans, D. W., Eyer, L., Guerra, R., Hutton, A., Jordi, C., Klioner, S. A., Lammers, U. L., Lindegren, L., Luri, X., Mignard, F., Panem, C., Pourbaix, D., Randich, S., Sartoretti, P., Soubiran, C., Tanga, P., Walton, N. A., Bailer-Jones, C. A. L., Bastian, U., Drimmel, R., Jansen, F., Katz, D., Lattanzi, M. G., van Leeuwen, F., Bakker, J., Cacciari, C., Castañeda, J., De Angeli, F., Fabricius, C., Fouesneau, M., Frémat, Y., Galluccio, L., Guerrier, A., Heiter, U., Masana, E., Messineo, R., Mowlavi, N., Nicolas, C., Nienartowicz, K., Pailler, F., Panuzzo, P., Riclet, F., Roux, W., Seabroke, G. M., Sordo, R., Thévenin, F., Gracia-Abril, G., Portell, J., Teyssier, D., Altmann, M., Andrae, R., Audard, M., Bellas-Velidis, I., Benson, K., Berthier, J., Blomme, R., Burgess, P. W., Busonero, D., Busso, G., Cánovas, H., Carry, B., Cellino, A., Cheek, N., Clementini, G., Damerdji, Y., Davidson, M., de Teodoro, P., Nuñez Campos, M., Delchambre, L., Dell'Oro, A., Esquej, P., Fernández-Hernández, J., Fraile, E., Garabato, D., García-Lario, P., Gosset, E., Haigron, R., Halbwachs, J.-L., Hambly, N. C., Harrison, D. L., Hernández, J., Hestroffer, D., Hodgkin, S. T., Holl, B., Janßen, K., Jevardat de Fombelle, G., Jordan, S., Krone-Martins, A., Lanzafame, A. C., Löffler, W., Marchal, O., Marrese, P. M., Moitinho, A., Muinonen, K., Osborne, P., Pancino, E., Pauwels, T., Recio-Blanco, A., Reylé, C., Riello, M., Rimoldini, L., Roegiers, T., Rybizki, J., Sarro, L. M., Siopis, C., Smith, M., Sozzetti, A., Utrilla, E., van Leeuwen, M., Abbas, U., Ábrahám, P., Abreu Aramburu, A., Aerts, C., Aguado, J. J., Ajaj, M., Aldea-Montero, F., Altavilla, G., Álvarez, M. A., Alves, J., Anders, F., Anderson, R. I., Anglada Varela, E., Antoja, T., Baines, D., Baker, S. G., Balaguer-Núñez, L., Balbinot, E., Balog, Z., Barache, C., Barbato, D., Barros, M., Barstow, M. A., Bartolomé, S., Bassilana, J.-L., Bauchet, N., Becciani, U., Bellazzini, M., Berihuete, A., Bernet, M., Bertone, S., Bianchi, L., Binnenfeld, A., Blanco-Cuaresma, S., Blazere, A., Boch, T., Bombrun, A., Bossini, D., Bouquillon, S., Bragaglia, A., Bramante, L., Breedt, E., Bressan, A., Brouillet, N., Brugaletta, E., Bucciarelli, B., Burlacu, A., Butkevich, A. G., Buzzi, R., Caffau, E., Cancelliere, R., Cantat-Gaudin, T., Carballo, R., Carlucci, T., Carnerero, M. I., Carrasco, J. M., Casamiquela, L., Castellani, M., Castro-Ginard, A., Chaoul, L., Charlot, P., Chemin, L., Chiaramida, V., Chiavassa, A., Chornay, N., Comoretto, G., Contursi, G., Cooper, W. J., Cornez, T., Cowell, S., Crifo, F., Cropper, M., Crosta, M., Crowley, C., Dafonte, C., Dapergolas, A., David, M., David, P., de Laverny, P., De Luise, F., De March, R., De Ridder, J., de Souza, R., de Torres, A., del Peloso, E. F., del Pozo, E., Delbo, M., Delgado, A., Delisle, J.-B., Demouchy, C., Dharmawardena, T. E., Di Matteo, P., Diakite, S., Diener, C., Distefano, E., Dolding, C., Edvardsson, B., Enke, H., Fabre, C., Fabrizio, M., Faigler, S., Fedorets, G., Fernique, P., Fienga, A., Figueras, F., Fournier, Y., Fouron, C., Fragkoudi, F., Gai, M., Garcia-Gutierrez, A., Garcia-Reinaldos, M., García-Torres, M., Garofalo, A., Gavel, A., Gavras, P., Gerlach, E., Geyer, R., Giacobbe, P., Gilmore, G., Girona, S., Giuffrida, G., Gomel, R., Gomez, A., González-Núñez, J., González-Santamaría, I., González-Vidal, J. J., Granvik, M., Guillout, P., Guiraud, J., Gutiérrez-Sánchez, R., Guy, L. P., Hatzidimitriou, D., Hauser, M., Haywood, M., Helmer, A., Helmi, A., Sarmiento, M. H., Hidalgo, S. L., Hilger, T., Hładczuk, N., Hobbs, D., Holland, G., Huckle, H. E., Jardine, K., Jasniewicz, G., Jean-Antoine Piccolo, A., Jiménez-Arranz, , Jorissen, A., Juaristi Campillo, J., Julbe, F., Karbevska, L., Kervella, P., Khanna, S., Kontizas, M., Kordopatis, G., Korn,

A. J., Kóspál, , Kostrzewa-Rutkowska, Z., Kruszyńska, K., Kun, M., Laizeau, P., Lambert, S., Lanza, A. F., Lasne, Y., Le Campion, J.-F., Lebreton, Y., Lebzelter, T., Leccia, S., Leclerc, N., Lecoeur-Taibi, I., Liao, S., Licata, E. L., Lindstrøm, H. E. P., Lister, T. A., Livanou, E., Lobel, A., Lorca, A., Loup, C., Madrero Pardo, P., Magdaleno Romeo, A., Managau, S., Mann, R. G., Manteiga, M., Marchant, J. M., Marconi, M., Marcos, J., Marcos Santos, M. M. S., Marín Pina, D., Marinoni, S., Marocco, F., Marshall, D. J., Martin Polo, L., Martín-Fleitas, J. M., Marton, G., Mary, N., Masip, A., Massari, D., Mastrobuono-Battisti, A., Mazeh, T., McMillan, P. J., Messina, S., Michalik, D., Millar, N. R., Mints, A., Molina, D., Molinaro, R., Molnár, L., Monari, G., Monguió, M., Montegriffo, P., Montero, A., Mor, R., Mora, A., Morbidelli, R., Morel, T., Morris, D., Muraveva, T., Murphy, C. P., Musella, I., Nagy, Z., Noval, L., Ocaña, F., Ogden, A., Ordenovic, C., Osinde, J. O., Pagani, C., Pagano, I., Palaversa, L., Palicio, P. A., Pallas-Quintela, L., Panahi, A., Payne-Wardenaar, S., Peñalosa Esteller, X., Penttilä, A., Pichon, B., Piersimoni, A. M., Pineau, F.-X., Plachy, E., Plum, G., Poggio, E., Prša, A., Pulone, L., Racero, E., Ragaini, S., Rainer, M., Raiteri, C. M., Rambaux, N., Ramos, P., Ramos-Lerate, M., Re Fiorentin, P., Regibo, S., Richards, P. J., Rios Diaz, C., Ripepi, V., Riva, A., Rix, H.-W., Rixon, G., Robichon, N., Robin, A. C., Robin, C., Roelens, M., Rogues, H. R. O., Rohrbasser, L., Romero-Gómez, M., Rowell, N., Royer, F., Ruz Mieres, D., Rybicki, K. A., Sadowski, G., Sáez Núñez, A., Sagristà Sellés, A., Sahlmann, J., Salguero, E., Samaras, N., Sanchez Gimenez, V., Sanna, N., Santoveña, R., Sarasso, M., Schultheis, M., Sciacca, E., Segol, M., Segovia, J. C., Ségransan, D., Semeux, D., Shahaf, S., Siddiqui, H. I., Siebert, A., Siltala, L., Silvelo, A., Slezak, E., Slezak, I., Smart, R. L., Snaith, O. N., Solano, E., Solitro, F., Souami, D., Souchay, J., Spagna, A., Spina, L., Spoto, F., Steele, I. A., Steidelmüller, H., Stephenson, C. A., Süveges, M., Surdej, J., Szabados, L., Szegedi-Elek, E., Taris, F., Taylor, M. B., Teixeira, R., Tolomei, L., Tonello, N., Torra, F., Torra, J., Torralba Elipe, G., Trabucchi, M., Tsounis, A. T., Turon, C., Ulla, A., Unger, N., Vaillant, M. V., van Dillen, E., van Reeven, W., Vanel, O., Vecchiato, A., Viala, Y., Vicente, D., Voutsinas, S., Weiler, M., Wevers, T., Wyrzykowski, , Yoldas, A., Yvard, P., Zhao, H., Zorec, J., Zucker, S., and Zwitter, T. (2023). Gaia data release 3: Summary of the content and survey properties. *Astronomy amp; Astrophysics*, 674:A1.

[Gaia Data Release 3, 2024] Gaia Data Release 3 (2024). Gaia/gdr3/spectroscopy/xp_continuous_mean_spectru https://cdn.gea.esac.esa.int/?prefix=Gaia/gdr3/Spectroscopy/xp_continuous_mean_spectrum/. Accessed July 2026.

[Gaia Data Release 3 Documentation, 2023] Gaia Data Release 3 Documentation (2023). 13.2.4 data partitioning. Gaia Data Release 3 Documentation, release 1.3. Accessed July 2026.

[Galbrun, 2022] Galbrun, E. (2022). The minimum description length principle for pattern mining: a survey. *Data Mining and Knowledge Discovery*, 36(5):1679–1727.

[Gorski et al., 2005] Gorski, K. M., Hivon, E., Banday, A. J., Wandelt, B. D., Hansen, F. K., Reinecke, M., and Bartelmann, M. (2005). Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2):759–771.

[Hoyer, 2004] Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints.

[Huang et al., 2024] Huang, B., Yuan, H., Xiang, M., Huang, Y., Xiao, K., Xu, S., Zhang, R., Yang, L., Niu, Z., and Gu, H. (2024). A comprehensive correction of the gaia dr3 xp spectra. *The Astrophysical Journal Supplement Series*, 271(1):13.

[Kao et al., 2024] Kao, M. L., Hawkins, K., Rogers, L. K., Bonsor, A., Dunlap, B. H., Sanders, J. L., Montgomery, M. H., and Winget, D. E. (2024). Hunting for polluted white dwarfs and other treasures with gaia xp spectra and unsupervised machine learning.

[Laroche and Speagle, 2025] Laroche, A. and Speagle, J. S. (2025). Closing the stellar labels gap: Stellar label independent evidence for [/m] information in gaia bp/rp spectra. *The Astrophysical Journal*, 979(1):5.

[Lee and Seung, 1999] Lee, D. and Seung, H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–91.

[Li et al., 2024] Li, J., Wong, K. W. K., Hogg, D. W., Rix, H.-W., and Chandra, V. (2024). Aspgap: Augmented stellar parameters and abundances for 37 million red giant branch stars from gaia xp low-resolution spectra. *The Astrophysical Journal Supplement Series*, 272(1):2.

[Mairal et al., 2010] Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2010). Online learning for matrix factorization and sparse coding.

[Majewski et al., 2017] Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., Prieto, C. A., Barkhouser, R., Bizyaev, D., Blank, B., Brunner, S., Burton, A., Carrera, R., Chojnowski, S. D., Cunha, K., Epstein, C., Fitzgerald, G., Pérez, A. E. G., Hearty, F. R., Henderson, C., Holtzman, J. A., Johnson, J. A., Lam, C. R., Lawler, J. E., Maseman, P., Mészáros, S., Nelson, M., Nguyen, D. C., Nidever, D. L., Pinsonneault, M., Shetrone, M., Smee, S., Smith, V. V., Stolberg, T., Skrutskie, M. F., Walker, E., Wilson, J. C., Zasowski, G., Anders, F., Basu, S., Beland, S., Blanton, M. R., Bovy, J., Brownstein, J. R., Carlberg, J., Chaplin, W., Chiappini, C., Eisenstein, D. J., Elsworth, Y., Feuillet, D., Fleming, S. W., Galbraith-Frew, J., García, R. A., García-Hernández, D. A., Gillespie, B. A., Girardi, L., Gunn, J. E., Hasselquist, S., Hayden, M. R., Hekker, S., Ivans, I., Kinemuchi, K., Klaene, M., Mahadevan, S., Mathur, S., Mosser, B., Muna, D., Munn, J. A., Nichol, R. C., O'Connell, R. W., Parejko, J. K., Robin, A. C., Rocha-Pinto, H., Schultheis, M., Serenelli, A. M., Shane, N., Aguirre, V. S., Sobeck, J. S., Thompson, B., Troup, N. W., Weinberg, D. H., and Zamora, O. (2017). The apache point observatory galactic evolution experiment (apogee). *The Astronomical Journal*, 154(3):94.

[Montegriffo et al., 2023] Montegriffo, P., De Angeli, F., Andrae, R., Riello, M., Pancino, E., Sanna, N., Bellazzini, M., Evans, D. W., Carrasco, J. M., Sordo, R., Busso, G., Cacciari, C., Jordi, C., van Leeuwen, F., Vallenari, A., Altavilla, G., Barstow, M. A., Brown, A. G. A., Burgess, P. W., Castellani, M., Cowell, S., Davidson, M., De Luise, F., Delchambre, L., Diener, C., Fabricius, C., Frémat, Y., Fouesneau, M., Gilmore, G., Giuffrida, G., Hambly, N. C., Harrison, D. L., Hidalgo, S., Hodgkin, S. T., Holland, G., Marinoni, S., Osborne, P. J., Pagani, C., Palaversa, L., Piersimoni, A. M., Pulone, L., Ragaini, S., Rainer, M., Richards, P. J., Rowell, N., Ruz-Mieres, D., Sarro, L. M., Walton, N. A., and Yoldas, A. (2023). ¡i¿gaia¡/i¿data release 3: External calibration of¡i¿bp/rp¡/i¿low-resolution spectroscopic data. *Astronomy amp; Astrophysics*, 674:A3.

[Olshausen and Field, 1996] Olshausen, B. and Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–9.

[Olshausen and Field, 1997] Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311–3325.

[Pati et al., 1993] Pati, Y., Rezaiifar, R., and Krishnaprasad, P. (1993). Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, pages 40–44 vol.1.

[Pérez-Couto et al., 2024] Pérez-Couto, X., Pallas-Quintela, L., Manteiga, M., Villaver, E., and Dafonte, C. (2024). Identifying new high-confidence polluted white dwarf candidates using gaia xp spectra and self-organizing maps. *The Astrophysical Journal*, 977(1):31.

[Ruz-Mieres et al., 2025] Ruz-Mieres, D., zuzannakr, and Angeli, F. D. (2025). gaia-dpci/gaiaxpy: Gaiaxpy v2.1.4.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

[Vreeken et al., 2011] Vreeken, J., Leeuwen, M., and Siebes, A. (2011). Krimp: Mining itemsets that compress. *Data Min. Knowl. Discov.*, 23:169–214.

[Weiler et al., 2023] Weiler, M., Carrasco, J. M., Fabricius, C., and Jordi, C. (2023). Analysing spectral lines in gaia low-resolution spectra. *Astronomy amp; Astrophysics*, 671:A52.

[Wenger et al., 2000] Wenger, M., Ochsenbein, F., Egret, D., Dubois, P., Bonnarel, F., Borde, S., Genova, F., Jasniewicz, G., Laloë, S., Lesteven, S., and Monier, R. (2000). The simbad astronomical database: The cds reference database for astronomical objects. *Astronomy and Astrophysics Supplement Series*, 143(1):9–22.

[Zhang et al., 2023] Zhang, X., Green, G. M., and Rix, H.-W. (2023). Parameters of 220 million stars from gaia bp/rp spectra. *Monthly Notices of the Royal Astronomical Society*, 524(2):1855–1884.

A Atoms with high correlation with G ($\rho > 0.3$)

Atom	ρ_g
18	0.458224
2	0.346503

Table 17: Atoms with high correlation with G ($\rho > 0.3$) in D_C.

Atom	ρ_g
38	0.500679
12	0.484630
86	0.475780
123	0.451486
149	0.450542
81	0.449086
52	0.448065
197	0.445908
18	0.440602
47	0.438554
125	0.438337
255	0.437747
110	0.437608
39	0.437352
179	0.435869
8	0.435841
53	0.435633
168	0.432387
171	0.430278
205	0.429576
215	0.429153
204	0.428207
116	0.423955
82	0.423293
9	0.422352
243	0.422346
75	0.422294
50	0.421256
196	0.420663
184	0.420581
151	0.418883
100	0.416295
92	0.413285
145	0.412514
155	0.412058
225	0.411651
217	0.411288
94	0.411153
108	-0.409178
188	0.408364
78	0.406909
13	0.406544
233	0.406237
232	0.405133
251	0.402185
247	0.401716
72	0.399798
216	0.399182
241	0.398675
54	0.398459
84	0.398273
29	0.397965
121	0.397868
249	0.397711
48	0.397173
45	0.396846
126	0.396496
67	0.393427
62	0.391601
246	0.391555
202	0.391084
189	0.389971
79	0.389397
169	0.388248
25	0.387568
74	0.386860
111	0.385696
244	0.381033
174	0.379192
143	0.378555
60	0.377736
235	0.376506
112	-0.373295
213	0.372872
175	0.372594
146	0.370766
87	0.370740
150	0.369349
43	0.367613
209	0.365114
239	0.364268
210	0.363289
83	0.363208
26	0.361497
119	0.361484
165	0.360231
238	0.358210
254	0.356337
157	0.355834
95	0.350337
49	0.346535
220	0.342300
32	0.341803
120	0.339190
4	0.339184
24	0.338855
135	0.337951
34	0.337842
85	0.334373
107	-0.333646
178	0.328482
228	-0.327442
161	0.325912
89	-0.320160
172	-0.316509
20	0.315838
221	-0.307590
177	-0.304024
129	0.303312

Table 18: Atoms with high correlation with G ($|\rho_g| > 0.3$) in D_H.

B Per-atom SIMBAD cross-match statistics

B.1 Compact Dictionary D_C

Table 19: SIMBAD object types of the stars each atom in D_C activates on, ranked by Poisson significance of the largest overrepresented class.

Atom	N_{match}	Hit rate	Dominant	Frac.	Enrich.	$z_{\max}$	Over-represented	Types
7	20670	0.047	LP?	0.256	3.646	100.8	LP?:3.6x(z101), LP*:2.9x(z80), Mi*:3.7x(z35)	LP?:5295, LP*:5266, *:3991
14	20604	0.059	LP*	0.254	2.863	79.6	LP*:2.9x(z80), LP?:3.1x(z79), Em*:2.6x(z38)	LP*:5227, LP?:4471, *:4004
18	18519	0.018	*	0.566	1.039	53.1	EB*:2.3x(z53), RR*:2.6x(z26), WD*:2.1x(z16)	*:10486, EB*:3934, RR*:690
4	36405	0.046	*	0.439	0.805	49.8	LP?:2.0x(z50), LP*:1.8x(z46), Mi*:2.1x(z19)	*:15966, LP*:5838, LP?:5075
9	10021	0.017	*	0.464	0.850	48.5	EB*:2.6x(z48), RR*:2.4x(z17), El*:3.3x(z9)	*:4645, EB*:2417, LP*:938
12	26761	0.030	*	0.640	1.174	42.0	EB*:1.8x(z42), WD*:1.8x(z15), Pu*:2.5x(z14)	*:17121, EB*:4597, SB*:681
2	22647	0.031	*	0.390	0.715	40.7	LP?:2.0x(z41), WD*:2.5x(z26), EB*:1.5x(z23)	*:8822, LP?:3215, EB*:3179
10	19954	0.023	*	0.642	1.179	38.0	EB*:1.9x(z38), RG*:2.5x(z24), Pe*:2.3x(z13)	*:12818, EB*:3502, RG*:668
5	11207	0.021	*	0.460	0.844	36.9	EB*:2.1x(z37), RR*:2.1x(z14), Mi*:2.2x(z12)	*:5158, EB*:2238, LP*:1319
15	22539	0.028	*	0.436	0.800	35.8	LP*:1.8x(z36), Mi*:2.8x(z25), LP?:1.5x(z20)	*:9837, LP*:3597, EB*:2971
6	26741	0.074	*	0.550	1.009	34.0	RR*:2.7x(z34), WD*:2.7x(z31), AGN:2.9x(z23)	*:14709, EB*:2615, LP?:1577
11	22989	0.025	*	0.672	1.233	27.5	EB*:1.6x(z27), RG*:1.7x(z13), Pe*:1.6x(z7)	*:15454, EB*:3419, SB*:694
19	21822	0.023	*	0.599	1.099	24.1	EB*:1.5x(z24), RG*:2.0x(z17), Pe*:2.1x(z11)	*:13073, EB*:3126, LP*:912
0	23596	0.029	*	0.685	1.257	24.1	RG*:2.3x(z24), SB*:1.5x(z11), Pe*:1.9x(z10)	*:16167, EB*:2377, LP*:1134
22	14960	0.024	*	0.523	0.959	24.0	EB*:1.6x(z24), BY*:3.7x(z18), Em*:1.8x(z16)	*:7820, EB*:2293, LP*:1394
16	21131	0.034	*	0.566	1.038	22.2	WD*:2.4x(z22), RR*:2.0x(z18), HS?:2.5x(z9)	*:11960, EB*:2345, LP*:1082
20	13898	0.025	*	0.509	0.934	21.1	WD*:2.6x(z21), C*:4.5x(z21), AB?:3.5x(z18)	*:7073, LP*:1768, EB*:1626
1	27020	0.033	*	0.699	1.282	18.5	EB*:1.4x(z19), RR*:1.9x(z19), Pe*:1.6x(z7)	*:18878, EB*:3453, SB*:888
23	27737	0.026	*	0.714	1.311	15.4	EB*:1.3x(z15), SB*:1.4x(z9), BY*:2.1x(z9)	*:19816, EB*:3370, PM*:1131
13	32357	0.036	*	0.541	0.993	15.1	LP?:1.3x(z15), Mi*:1.8x(z13), RR*:1.5x(z11)	*:17516, LP*:3356, LP?:2992
3	45567	0.042	*	0.703	1.290	12.4	SB*:1.4x(z12), RG*:1.3x(z9), PM*:1.1x(z5)	*:32029, EB*:3358, PM*:1932
21	26010	0.029	*	0.732	1.342	11.0	RG*:1.6x(z11), SB*:1.4x(z10), HB*:1.3x(z3)	*:19032, EB*:2565, PM*:1066
17	42357	0.046	*	0.650	1.193	11.0	PM*:1.3x(z11), Pu*:1.5x(z6), SB*:1.2x(z5)	*:27552, EB*:2941, PM*:2039
8	25309	0.040	*	0.747	1.370	9.8	SB*:1.4x(z10), gD*:3.0x(z9), Pu*:1.9x(z8)	*:18901, EB*:2435, PM*:1076

Table 20: SIMBAD object types of the stars that make up the top 1% of activations of each atom in D_C, ranked by Poisson significance of the largest overrepresented class.

Atom	N_{match}	Hit rate	Dominant	Frac.	Enrich.	$z_{\max}$	Over-represented	Types
20	269	0.049	C*	0.338	131.823	108.7	C*:131.8x(z109), WD*:26.6x(z47), Mi*:11.5x(z16)	C*:91, WD*:88, Mi*:26
4	5155	0.657	LP?	0.459	6.530	105.2	LP?:6.5x(z105), LP*:4.7x(z78), Mi*:12.6x(z76)	LP?:2365, LP*:2130, Mi*:543
12	949	0.107	*	0.415	0.762	100.9	WD*:30.5x(z101), QSO:10.1x(z12), AGN:4.2x(z8)	*:394, WD*:356, EB*:72
7	3041	0.690	LP?	0.498	7.087	89.0	LP?:7.1x(z89), Mi*:13.2x(z62), LP*:4.2x(z53)	LP?:1514, LP*:1143, Mi*:337
15	4170	0.509	LP?	0.425	6.052	86.5	LP?:6.1x(z86), Mi*:14.2x(z78), LP*:4.1x(z60)	LP?:1773, LP*:1517, Mi*:497
13	3739	0.417	LP?	0.426	6.061	82.0	LP?:6.1x(z82), LP*:4.7x(z68), Mi*:10.0x(z51)	LP?:1592, LP*:1561, Mi*:314
6	649	0.180	EB*	0.277	2.972	75.0	HS?:67.4x(z75), WD*:20.7x(z56), EB*:3.0x(z15)	EB*:180, WD*:165, *:134
18	626	0.061	WD*	0.305	24.838	66.1	WD*:24.8x(z66), AGN:27.9x(z52), QSO:30.7x(z33)	WD*:191, AGN:105, *:77
5	473	0.090	Mi*	0.264	31.551	60.8	Mi*:31.6x(z61), C*:28.0x(z30), AGN:8.8x(z13)	Mi*:125, LP?:81, EB*:56
11	225	0.024	WD*	0.413	33.648	54.3	WD*:33.6x(z54), AGN:11.8x(z13), EB*:2.6x(z7)	WD*:93, EB*:54, *:31
14	1778	0.507	PM*	0.215	5.700	38.5	PM*:5.7x(z39), Em*:4.8x(z26), Mi*:7.6x(z25)	PM*:383, LP?:374, LP*:294
2	396	0.053	LP?	0.197	2.804	35.0	Mi*:20.2x(z35), RR*:7.7x(z16), LP?:2.8x(z10)	LP?:78, Mi*:67, RR*:44
1	317	0.039	RR*	0.243	16.784	33.8	RR*:16.8x(z34), WD*:11.3x(z20), Pe*:5.6x(z6)	RR*:77, *:72, EB*:61
17	2328	0.251	*	0.631	1.158	31.5	HS?:15.7x(z32), WD*:5.8x(z26), Pu*:7.2x(z17)	*:1470, EB*:252, WD*:165
3	357	0.033	*	0.218	0.401	20.6	RR*:10.1x(z21), WD*:6.4x(z11), EB*:2.3x(z7)	*:78, EB*:76, RR*:52
21	289	0.032	*	0.671	1.232	19.2	AGN:15.5x(z19), QSO:12.2x(z8), WD*:2.8x(z3)	*:194, AGN:27, EB*:19
19	330	0.035	*	0.448	0.823	18.5	AGN:14.1x(z18), WD*:7.4x(z13), **:12.9x(z9)	*:148, EB*:41, WD*:30
10	150	0.018	EB*	0.547	5.858	18.2	EB*:5.9x(z18), Pe*:6.6x(z5), SB*:1.5x(z1)	EB*:82, *:46, SB*:6
22	192	0.030	*	0.458	0.841	15.5	PM*:6.8x(z15), Mi*:13.1x(z15), LM*:17.3x(z11)	*:88, PM*:49, Mi*:21
9	168	0.028	*	0.220	0.404	15.0	QSO:27.1x(z15), AGN:13.8x(z13), Mi*:10.7x(z11)	*:37, EB*:34, Em*:22
16	144	0.023	LP?	0.299	4.251	12.2	WD*:10.2x(z12), LP?:4.3x(z10), RR*:7.2x(z9)	LP?:43, EB*:19, WD*:18
23	180	0.017	*	0.511	0.938	10.0	PM*:4.9x(z10), SB*:2.1x(z2), EB*:1.5x(z2)	*:92, PM*:33, EB*:26
0	1464	0.183	*	0.891	1.634	6.7	SB*:2.1x(z7), RG*:1.0x(z0)	*:1304, SB*:80, LP*:23
8	1111	0.174	*	0.840	1.541	6.6	SB*:2.2x(z7), PM*:1.8x(z5)	*:933, PM*:76, SB*:65

Table 21: Atoms in D_C holding a majority of a single SIMBAD class, ranked by specialization ζ.

Atom	Class	N_{atom}	N_{total}	Coverage	Atom frac.	Spec.
3	*	32029	41843	0.765	0.703	0.538
17	*	27552	41843	0.658	0.650	0.428
7	LP?	5295	5393	0.982	0.256	0.252
7	LP*	5266	6801	0.774	0.255	0.197
14	LP*	5227	6801	0.769	0.254	0.195
14	LP?	4471	5393	0.829	0.217	0.180
4	LP*	5838	6801	0.858	0.160	0.138
4	LP?	5075	5393	0.941	0.139	0.131
18	EB*	3934	7164	0.549	0.212	0.117
12	EB*	4597	7164	0.642	0.172	0.110
2	LP?	3215	5393	0.596	0.142	0.085
15	LP*	3597	6801	0.529	0.160	0.084
13	LP?	2992	5393	0.555	0.092	0.051
14	Em*	1402	1973	0.711	0.068	0.048
6	RR*	1056	1111	0.950	0.039	0.038
17	PM*	2039	2901	0.703	0.048	0.034
6	WD*	895	943	0.949	0.033	0.032
7	Mi*	637	643	0.991	0.031	0.031
3	SB*	1625	2016	0.806	0.036	0.029
3	PM*	1932	2901	0.666	0.042	0.028
7	Em*	1059	1973	0.537	0.051	0.027
2	WD*	709	943	0.752	0.031	0.024
18	RR*	690	1111	0.621	0.037	0.023
0	RG*	752	1044	0.720	0.032	0.023
4	PM*	1557	2901	0.537	0.043	0.023
10	RG*	668	1044	0.640	0.033	0.021
15	Mi*	535	643	0.832	0.024	0.020
17	SB*	1282	2016	0.636	0.030	0.019
16	WD*	618	943	0.655	0.029	0.019
1	RR*	757	1111	0.681	0.028	0.019
14	Mi*	498	643	0.774	0.024	0.019
6	AGN	459	462	0.994	0.017	0.017
4	Mi*	630	643	0.980	0.017	0.017
16	RR*	616	1111	0.554	0.029	0.016
2	Mi*	485	643	0.754	0.021	0.016
19	RG*	591	1044	0.566	0.027	0.015
3	RG*	833	1044	0.798	0.018	0.015
12	WD*	603	943	0.639	0.023	0.014
13	RR*	696	1111	0.626	0.022	0.013
18	WD*	472	943	0.501	0.025	0.013
11	RG*	539	1044	0.516	0.023	0.012

Atom	Class	N_{atom}	N_{total}	Coverage	Atom frac.	Spec.
21	RG*	560	1044	0.536	0.022	0.012
3	RR*	760	1111	0.684	0.017	0.011
14	AGN	327	462	0.708	0.016	0.011
13	Mi*	480	643	0.747	0.015	0.011
12	RR*	562	1111	0.506	0.021	0.011
4	RG*	622	1044	0.596	0.017	0.010
23	RG*	535	1044	0.512	0.019	0.010
14	AB?	241	290	0.831	0.012	0.010
20	C*	160	197	0.812	0.012	0.009
6	Pu*	247	253	0.976	0.009	0.009
17	WD*	585	943	0.620	0.014	0.009
20	AB?	183	290	0.631	0.013	0.008
18	AGN	265	462	0.574	0.014	0.008
7	C*	182	197	0.924	0.009	0.008
22	BY*	157	217	0.724	0.010	0.008
17	RR*	596	1111	0.536	0.014	0.008
6	Mi*	357	643	0.555	0.013	0.007
17	RG*	566	1044	0.542	0.013	0.007
12	Pu*	221	253	0.874	0.008	0.007
10	Pe*	233	386	0.604	0.012	0.007
15	C*	174	197	0.883	0.008	0.007
18	Pe*	216	386	0.560	0.012	0.007
19	AGN	254	462	0.550	0.012	0.006
7	AB?	194	290	0.669	0.009	0.006
19	Pe*	226	386	0.585	0.010	0.006
0	Pe*	228	386	0.591	0.010	0.006
6	HS?	151	151	1.000	0.006	0.006
6	QSO	149	152	0.980	0.006	0.005
20	HS?	107	151	0.709	0.008	0.005
14	LM*	144	185	0.778	0.007	0.005
4	Y*?	307	494	0.621	0.008	0.005
18	Pu*	149	253	0.589	0.008	0.005
1	Pe*	220	386	0.570	0.008	0.005
4	AGN	278	462	0.602	0.008	0.005
10	BY*	140	217	0.645	0.007	0.005
3	Pe*	281	386	0.728	0.006	0.004
10	RS*	140	219	0.639	0.007	0.004
17	Pu*	215	253	0.850	0.005	0.004
23	BY*	161	217	0.742	0.006	0.004
3	V*	305	482	0.633	0.007	0.004
0	HB*	164	277	0.592	0.007	0.004
8	Pu*	160	253	0.632	0.006	0.004
23	Pe*	200	386	0.518	0.007	0.004

Atom	Class	N_{atom}	N_{total}	Coverage	Atom frac.	Spec.
14	QSO	107	152	0.704	0.005	0.004
17	Y*?	276	494	0.559	0.007	0.004
19	BY*	131	217	0.604	0.006	0.004
13	BY*	157	217	0.724	0.005	0.004
12	RS*	143	219	0.653	0.005	0.003
17	HB*	196	277	0.708	0.005	0.003
16	HS?	102	151	0.675	0.005	0.003
4	LM*	145	185	0.784	0.004	0.003
19	RS*	120	219	0.548	0.005	0.003
0	BY*	124	217	0.571	0.005	0.003
7	LM*	104	185	0.562	0.005	0.003
3	Y*?	250	494	0.506	0.005	0.003
18	QSO	88	152	0.579	0.005	0.003
12	dS*	77	81	0.951	0.003	0.003
20	s*r	48	63	0.762	0.003	0.003
4	HB*	159	277	0.574	0.004	0.003
3	HB*	177	277	0.639	0.004	0.002
6	dS*	73	81	0.901	0.003	0.002
13	C*	124	197	0.629	0.004	0.002
8	gD*	62	63	0.984	0.002	0.002
18	dS*	60	81	0.741	0.003	0.002
2	HS?	90	151	0.596	0.004	0.002
6	gD*	63	63	1.000	0.002	0.002
23	RS*	118	219	0.539	0.004	0.002
1	RS*	116	219	0.530	0.004	0.002
12	gD*	61	63	0.968	0.002	0.002
13	RS*	125	219	0.571	0.004	0.002
7	HS?	79	151	0.523	0.004	0.002
6	**	82	126	0.651	0.003	0.002
17	C*	129	197	0.655	0.003	0.002
3	RS*	141	219	0.644	0.003	0.002
17	**	103	126	0.817	0.002	0.002
14	s*r	50	63	0.794	0.002	0.002
17	HS?	110	151	0.728	0.003	0.002
2	QSO	79	152	0.520	0.003	0.002
3	Pu*	144	253	0.569	0.003	0.002
7	El*	65	125	0.520	0.003	0.002
8	dS*	57	81	0.704	0.002	0.002
3	**	94	126	0.746	0.002	0.002
4	QSO	92	152	0.605	0.003	0.002
17	gD*	62	63	0.984	0.001	0.001
1	dS*	56	81	0.691	0.002	0.001
1	cC*	65	114	0.570	0.002	0.001

Atom	Class	N_{atom}	N_{total}	Coverage	Atom frac.	Spec.
18	gD*	39	63	0.619	0.002	0.001
17	cC*	79	114	0.693	0.002	0.001
2	cC*	57	114	0.500	0.003	0.001
13	cC*	68	114	0.596	0.002	0.001
8	**	63	126	0.500	0.002	0.001
3	gD*	56	63	0.889	0.001	0.001
3	cC*	75	114	0.658	0.002	0.001
3	dS*	59	81	0.728	0.001	0.001
4	s*r	40	63	0.635	0.001	0.001
1	gD*	34	63	0.540	0.001	0.001
13	dS*	41	81	0.506	0.001	0.001

B.2 High-Capacity Dictionary D_H

Table 22: SIMBAD object types of the stars each atom in D_H activates on, ranked by Poisson significance of the largest excess.

Atom	N_{match}	Hit rate	Dominant	Frac.	Enrich.	$z_{\max}$	Over-represented	Types
229	10143	0.267	LP?	0.430	6.1	136.8	LP?:6.1x(z137), LP*:4.3x(z98), Mi*:7.1x(z56)	LP?:4365, LP*:3833, Mi*:604
36	1821	0.091	LP*	0.259	2.9	100.1	Mi*:26.6x(z100), AGN:9.8x(z29), LP*:2.9x(z24)	LP*:472, Mi*:406, LP?:358
107	11092	0.189	LP*	0.359	4.1	95.7	LP*:4.1x(z96), LP?:4.1x(z87), PM*:2.1x(z23)	LP*:3982, LP?:3220, *:1125
54	2472	0.017	*	0.364	0.7	92.7	WD*:17.8x(z93), AGN:8.9x(z31), RR*:3.9x(z17)	*:901, WD*:541, EB*:487
55	4212	0.119	LP?	0.430	6.1	88.1	LP?:6.1x(z88), LP*:3.9x(z56), Mi*:3.5x(z15)	LP?:1812, LP*:1457, *:287
93	3181	0.262	LP?	0.340	4.8	87.4	Mi*:17.9x(z87), LP?:4.8x(z57), C*:17.6x(z48)	LP?:1081, LP*:799, Mi*:478
142	6074	0.087	*	0.266	0.5	80.7	Em*:7.5x(z81), PM*:4.5x(z54), LP*:2.0x(z23)	*:1613, Em*:1165, LP*:1083
40	8232	0.187	*	0.522	1.0	77.0	WD*:8.7x(z77), AGN:8.9x(z55), HS?:9.3x(z34)	*:4299, WD*:875, EB*:806
89	7046	0.462	LP*	0.355	4.0	75.5	LP?:4.4x(z75), LP*:4.0x(z75), Mi*:9.8x(z67)	LP*:2499, LP?:2174, Mi*:576
71	4546	0.084	LP?	0.363	5.2	74.6	LP?:5.2x(z75), Mi*:11.0x(z62), LP*:3.6x(z52)	LP?:1652, LP*:1456, Mi*:418
98	2724	0.173	LP?	0.345	4.9	72.9	Mi*:16.3x(z73), LP?:4.9x(z54), LP*:3.0x(z32)	LP?:939, LP*:731, Mi*:371
191	696	0.015	WD*	0.302	24.6	68.9	WD*:24.6x(z69), Em*:3.8x(z12), El*:5.3x(z5)	WD*:210, *:123, LP*:95
70	3593	0.100	LP*	0.321	3.6	66.5	WD*:11.0x(z67), HS?:19.8x(z50), LP*:3.6x(z47)	LP*:1152, *:762, WD*:486
82	2898	0.016	*	0.375	0.7	63.6	RR*:10.8x(z64), EB*:3.1x(z35), cC*:10.2x(z19)	*:1087, EB*:841, RR*:454
31	2658	0.038	*	0.346	0.6	62.9	Em*:8.6x(z63), LP*:2.0x(z16), PM*:2.3x(z13)	*:920, Em*:588, LP*:476
144	3544	0.070	LP*	0.285	3.2	60.3	C*:21.0x(z60), LP*:3.2x(z39), Mi*:8.0x(z38)	LP*:1010, LP?:646, *:477
226	2897	0.063	LP?	0.367	5.2	60.2	LP?:5.2x(z60), LP*:4.0x(z47), RG*:2.4x(z9)	LP?:1062, LP*:1014, *:273
252	3154	0.069	LP?	0.343	4.9	57.8	LP?:4.9x(z58), Mi*:10.6x(z49), LP*:3.5x(z42)	LP?:1082, LP*:979, Mi*:279
163	2810	0.031	*	0.449	0.8	57.6	RR*:10.0x(z58), EB*:2.3x(z21), AGN:3.2x(z9)	*:1261, EB*:610, RR*:408
15	7782	0.058	*	0.646	1.2	56.9	WD*:6.8x(z57), Pu*:4.9x(z20), AGN:3.7x(z19)	*:5029, EB*:1013, WD*:652
115	2809	0.015	*	0.449	0.8	56.3	EB*:4.5x(z56), RS*:3.7x(z8), V*:2.4x(z6)	*:1261, EB*:1173, RR*:55
210	4995	0.026	*	0.532	1.0	56.0	RR*:7.6x(z56), AGN:3.4x(z13), EB*:1.5x(z10)	*:2659, EB*:688, RR*:548
119	2533	0.011	*	0.448	0.8	54.4	EB*:4.5x(z54), RR*:2.3x(z8), RG*:2.0x(z6)	*:1134, EB*:1073, RR*:83
88	6848	0.091	*	0.463	0.8	54.0	RR*:6.4x(z54), AGN:9.2x(z52), WD*:5.8x(z44)	*:3168, EB*:878, RR*:637
64	2378	0.046	LP?	0.359	5.1	53.1	LP?:5.1x(z53), LP*:2.6x(z24), Mi*:5.9x(z22)	LP?:854, LP*:554, *:308
116	2791	0.014	*	0.431	0.8	53.1	RR*:9.4x(z53), EB*:3.0x(z32), AGN:3.5x(z10)	*:1204, EB*:779, RR*:378
175	2549	0.011	*	0.464	0.9	52.7	EB*:4.4x(z53), RR*:2.6x(z10), V*:2.3x(z5)	*:1183, EB*:1051, RR*:95
208	2149	0.029	*	0.315	0.6	52.3	WD*:11.2x(z52), RR*:6.5x(z31), EB*:2.3x(z18)	*:676, EB*:459, WD*:295
67	3291	0.015	*	0.420	0.8	52.2	EB*:4.0x(z52), RR*:3.9x(z20), RG*:3.2x(z15)	*:1381, EB*:1222, RR*:185
60	2903	0.012	*	0.467	0.9	51.9	EB*:4.2x(z52), RR*:3.0x(z13), V*:2.6x(z7)	*:1357, EB*:1125, RR*:126
42	2556	0.015	*	0.443	0.8	49.9	EB*:4.2x(z50), RR*:4.1x(z19), RG*:2.3x(z8)	*:1132, EB*:1009, RR*:151
25	2510	0.010	*	0.519	1.0	49.6	EB*:4.2x(z50), BY*:1.8x(z2), RS*:1.7x(z2)	*:1302, EB*:993, RG*:26
104	8348	0.059	*	0.602	1.1	49.5	RR*:5.5x(z50), AGN:3.9x(z21), Pu*:4.1x(z16)	*:5028, EB*:947, RR*:665
239	2615	0.012	*	0.474	0.9	49.4	EB*:4.2x(z49), RR*:2.7x(z11), El*:3.1x(z4)	*:1239, EB*:1015, RR*:103
227	2357	0.022	*	0.535	1.0	48.8	C*:20.8x(z49), s*r:13.4x(z17), LP*:2.1x(z15)	*:1261, LP*:430, EB*:195
251	2836	0.012	*	0.490	0.9	48.5	EB*:4.0x(z49), RR*:2.2x(z8), cC*:2.6x(z3)	*:1389, EB*:1054, RR*:90
133	4081	0.017	*	0.508	0.9	47.6	EB*:3.4x(z48), RG*:2.8x(z14), Pe*:3.0x(z9)	*:2074, EB*:1310, RG*:157

Atom	N_{match}	Hit rate	Dominant	Frac.	Enrich.	$z_{\max}$	Over-represented	Types
202	2462	0.011	*	0.452	0.8	47.1	EB*:4.1x(z47), RR*:3.7x(z16), cC*:4.1x(z6)	*:1114, EB*:944, RR*:131
65	9965	0.151	*	0.613	1.1	46.9	RR*:4.9x(z47), HS?:7.7x(z30), AGN:4.6x(z28)	*:6108, EB*:847, RR*:708
166	4607	0.088	*	0.198	0.4	46.7	PM*:4.5x(z47), Em*:4.7x(z40), Mi*:6.6x(z35)	*:913, PM*:790, LP*:746
235	2557	0.011	*	0.497	0.9	46.3	EB*:4.0x(z46), RR*:2.5x(z9), El*:4.8x(z8)	*:1270, EB*:954, RR*:91
220	2187	0.013	*	0.481	0.9	46.1	EB*:4.2x(z46), El*:5.1x(z8), RR*:2.1x(z6)	*:1052, EB*:863, RR*:66
17	2449	0.022	*	0.483	0.9	45.3	EB*:4.0x(z45), RS*:6.4x(z14), RR*:3.3x(z14)	*:1184, EB*:914, RR*:116
216	2480	0.010	*	0.501	0.9	44.9	EB*:3.9x(z45), RR*:2.5x(z9), V*:2.1x(z4)	*:1242, EB*:914, RR*:90
135	2109	0.013	*	0.473	0.9	44.5	EB*:4.2x(z44), El*:5.0x(z7), RR*:1.5x(z3)	*:997, EB*:821, LP*:85
134	2187	0.022	LP*	0.370	4.2	44.2	LP*:4.2x(z44), AB?:4.6x(z10), Em*:2.0x(z8)	LP*:809, *:705, EB*:303
85	2414	0.011	*	0.522	1.0	43.5	EB*:3.9x(z43), V*:2.2x(z5), RS*:2.8x(z5)	*:1259, EB*:878, RR*:47
129	2157	0.014	*	0.470	0.9	42.5	EB*:4.0x(z42), El*:4.3x(z6), Pe*:2.5x(z5)	*:1014, EB*:804, RR*:57
174	2682	0.011	*	0.483	0.9	42.4	EB*:3.7x(z42), RR*:3.7x(z17), RG*:2.2x(z7)	*:1295, EB*:921, RR*:144
111	2743	0.015	*	0.444	0.8	42.1	WD*:8.3x(z42), EB*:3.1x(z33), AGN:5.1x(z17)	*:1218, EB*:782, WD*:278
179	3068	0.015	*	0.473	0.9	41.9	EB*:3.5x(z42), RR*:3.9x(z20), AGN:2.6x(z7)	*:1451, EB*:995, RR*:175
243	2989	0.011	*	0.529	1.0	41.7	EB*:3.5x(z42), RR*:2.5x(z10), cC*:2.7x(z4)	*:1582, EB*:975, RR*:107
121	2917	0.012	*	0.542	1.0	41.3	EB*:3.5x(z41), RG*:2.7x(z11), RS*:4.2x(z9)	*:1580, EB*:954, RG*:106
50	2351	0.011	*	0.487	0.9	40.7	EB*:3.7x(z41), AGN:3.5x(z9), WD*:2.1x(z6)	*:1144, EB*:822, WD*:62
87	6740	0.035	*	0.655	1.2	40.4	RR*:5.1x(z40), EB*:1.5x(z13), Pu*:3.2x(z10)	*:4416, EB*:952, RR*:497
143	2249	0.010	*	0.531	1.0	40.2	EB*:3.8x(z40), El*:3.3x(z4), V*:2.1x(z4)	*:1194, EB*:792, RR*:38
215	3017	0.016	*	0.471	0.9	40.2	RR*:7.1x(z40), EB*:3.0x(z34), cC*:7.6x(z14)	*:1420, EB*:850, RR*:309
73	2660	0.013	*	0.449	0.8	40.0	EB*:3.5x(z40), RR*:4.6x(z22), RG*:3.7x(z16)	*:1195, EB*:878, RR*:177
94	2532	0.012	*	0.545	1.0	39.9	EB*:3.6x(z40), El*:3.2x(z4), HB*:1.8x(z2)	*:1381, EB*:850, RG*:47
0	3126	0.043	*	0.355	0.7	39.6	AB?:12.5x(z40), LP*:3.1x(z36), Em*:3.4x(z22)	*:1110, LP*:869, Em*:275
213	2341	0.010	*	0.532	1.0	39.4	EB*:3.7x(z39), RR*:2.1x(z6), BY*:2.6x(z4)	*:1246, EB*:801, RR*:70
100	2950	0.011	*	0.558	1.0	39.3	EB*:3.4x(z39), cC*:3.2x(z5), BY*:2.3x(z4)	*:1645, EB*:927, RG*:63
246	2530	0.011	*	0.517	0.9	39.0	EB*:3.5x(z39), RR*:3.8x(z17), cC*:4.0x(z6)	*:1307, EB*:836, RR*:138
249	2874	0.013	*	0.495	0.9	38.9	EB*:3.4x(z39), RR*:3.3x(z15), WD*:3.1x(z12)	*:1423, EB*:906, RR*:137
242	4863	0.050	*	0.615	1.1	38.7	WD*:6.0x(z39), AGN:6.8x(z32), Pu*:4.5x(z14)	*:2992, EB*:502, WD*:359
190	3596	0.018	*	0.491	0.9	38.4	EB*:3.1x(z38), RG*:3.1x(z15), Pe*:3.7x(z11)	*:1767, EB*:1039, RG*:153
244	2375	0.010	*	0.500	0.9	38.3	EB*:3.6x(z38), RR*:4.4x(z20), RG*:1.6x(z3)	*:1187, EB*:792, RR*:152
78	2422	0.010	*	0.530	1.0	38.3	EB*:3.5x(z38), RR*:2.5x(z9), cC*:3.1x(z4)	*:1283, EB*:802, RR*:89
84	3654	0.016	*	0.559	1.0	38.3	EB*:3.1x(z38), RG*:2.6x(z11), V*:2.1x(z5)	*:2044, EB*:1048, RG*:127
199	1618	0.009	*	0.448	0.8	38.2	EB*:4.1x(z38), WD*:3.0x(z9), RR*:2.6x(z8)	*:725, EB*:621, RR*:60
238	2481	0.012	*	0.531	1.0	38.2	EB*:3.5x(z38), BY*:9.3x(z22), RS*:5.2x(z11)	*:1318, EB*:813, BY*:65
204	2039	0.011	*	0.453	0.8	38.2	EB*:3.8x(z38), WD*:3.8x(z14), RR*:3.3x(z12)	*:924, EB*:717, RR*:96
224	3950	0.021	*	0.525	1.0	38.1	EB*:3.0x(z38), BY*:4.4x(z11), RS*:4.1x(z10)	*:2073, EB*:1101, SB*:120
62	2185	0.009	*	0.549	1.0	37.5	EB*:3.6x(z38), El*:4.2x(z6), V*:1.9x(z3)	*:1200, EB*:740, RR*:46
49	2591	0.011	*	0.569	1.0	37.4	EB*:3.4x(z37), RG*:3.3x(z14), cC*:2.1x(z2)	*:1473, EB*:824, RG*:117
114	3069	0.020	*	0.431	0.8	37.1	LP*:3.3x(z37), EB*:1.9x(z16), El*:2.8x(z4)	*:1324, LP*:884, EB*:557

Atom	N_{match}	Hit rate	Dominant	Frac.	Enrich.	$z_{\max}$	Over-represented	Types
145	2671	0.010	*	0.575	1.1	37.1	EB*:3.3x(z37), RG*:2.1x(z7), HB*:2.6x(z5)	*:1535, EB*:835, RG*:78
236	3162	0.017	*	0.485	0.9	36.8	EB*:3.1x(z37), RR*:6.0x(z34), RS*:3.9x(z9)	*:1532, EB*:928, RR*:273
74	2714	0.010	*	0.557	1.0	36.7	EB*:3.3x(z37), RG*:1.9x(z5), V*:2.1x(z4)	*:1513, EB*:837, RG*:69
26	3533	0.016	*	0.548	1.0	36.5	EB*:3.0x(z36), RR*:3.7x(z19), V*:2.0x(z5)	*:1935, EB*:992, RR*:188
56	4608	0.024	*	0.608	1.1	36.5	EB*:2.8x(z36), RS*:3.7x(z10), dS*:4.7x(z8)	*:2803, EB*:1186, SB*:135
254	2481	0.011	*	0.551	1.0	36.4	EB*:3.4x(z36), BY*:3.6x(z7), V*:2.6x(z6)	*:1366, EB*:786, LP*:49
205	2353	0.013	*	0.449	0.8	36.4	EB*:3.5x(z36), RR*:4.3x(z19), AGN:5.4x(z16)	*:1057, EB*:759, RR*:147
68	5138	0.040	*	0.526	1.0	36.3	RR*:5.2x(z36), cC*:3.4x(z7), Pu*:2.5x(z6)	*:2705, EB*:554, LP?:421
203	2484	0.070	LP*	0.268	3.0	36.0	Mi*:8.9x(z36), AGN:9.0x(z31), LP*:3.0x(z30)	LP*:665, LP?:415, *:264
209	3635	0.015	*	0.510	0.9	35.9	EB*:2.9x(z36), RG*:3.7x(z19), RR*:2.9x(z13)	*:1855, EB*:1000, RG*:183
189	2386	0.009	*	0.541	1.0	35.6	EB*:3.4x(z36), RR*:3.4x(z14), RG*:1.5x(z3)	*:1290, EB*:754, RR*:119
232	3352	0.014	*	0.541	1.0	35.6	EB*:3.0x(z36), RR*:2.5x(z10), V*:2.7x(z8)	*:1813, EB*:942, RR*:120
184	2340	0.012	*	0.451	0.8	35.4	EB*:3.4x(z35), RR*:5.6x(z26), cC*:6.9x(z11)	*:1056, EB*:742, RR*:188
162	4843	0.024	*	0.527	1.0	35.0	EB*:2.6x(z35), BY*:5.7x(z17), Pe*:3.2x(z11)	*:2551, EB*:1195, RR*:146
156	2659	0.042	LP*	0.290	3.3	35.0	LP*:3.3x(z35), Em*:3.4x(z20), AB?:7.0x(z19)	LP*:772, *:608, LP?:433
29	3229	0.013	*	0.554	1.0	34.9	EB*:3.0x(z35), RG*:2.9x(z13), Pe*:3.6x(z11)	*:1789, EB*:908, RG*:129
57	1839	0.011	*	0.463	0.8	34.5	EB*:3.6x(z34), RR*:4.5x(z18), RG*:2.5x(z7)	*:851, EB*:623, RR*:121
126	2340	0.010	*	0.535	1.0	34.4	EB*:3.3x(z34), RR*:2.3x(z7), cC*:4.3x(z6)	*:1253, EB*:727, RR*:77
241	2853	0.011	*	0.557	1.0	34.3	EB*:3.1x(z34), RG*:1.9x(z5), cC*:2.6x(z3)	*:1588, EB*:826, RG*:73
13	2254	0.010	*	0.495	0.9	34.2	EB*:3.4x(z34), RR*:3.9x(z17), WD*:2.5x(z8)	*:1115, EB*:706, RR*:127
155	2226	0.009	*	0.541	1.0	34.1	EB*:3.4x(z34), RR*:3.1x(z12), WD*:2.3x(z7)	*:1204, EB*:699, RR*:101
255	2343	0.010	*	0.540	1.0	34.0	EB*:3.3x(z34), RR*:3.4x(z14), cC*:6.3x(z10)	*:1265, EB*:721, RR*:116
99	1946	0.016	*	0.550	1.0	33.8	EB*:3.5x(z34), RS*:3.4x(z6), El*:3.8x(z5)	*:1071, EB*:637, SB*:38
24	1818	0.011	*	0.492	0.9	33.4	EB*:3.6x(z33), WD*:3.1x(z10), cC*:3.7x(z4)	*:894, EB*:605, WD*:70
117	1691	0.021	*	0.406	0.7	33.1	Em*:6.0x(z33), AB?:8.0x(z18), RG*:2.7x(z8)	*:686, Em*:262, LP*:172
196	3502	0.013	*	0.613	1.1	33.0	EB*:2.8x(z33), RR*:1.5x(z4), RS*:1.9x(z3)	*:2147, EB*:923, RR*:77
123	2380	0.010	*	0.531	1.0	32.6	EB*:3.2x(z33), RR*:3.1x(z12), WD*:2.7x(z9)	*:1263, EB*:708, RR*:106
8	2784	0.011	*	0.499	0.9	31.9	EB*:3.0x(z32), RR*:5.2x(z26), cC*:7.5x(z13)	*:1388, EB*:774, RR*:208
150	2494	0.010	*	0.603	1.1	31.0	EB*:3.0x(z31), RS*:2.1x(z3), V*:1.7x(z3)	*:1505, EB*:705, RR*:50
6	2091	0.013	*	0.545	1.0	30.9	BY*:13.7x(z31), EB*:2.9x(z27), RS*:4.9x(z9)	*:1140, EB*:575, BY*:81
120	1432	0.011	*	0.490	0.9	30.8	EB*:3.7x(z31), El*:5.6x(z7), RR*:1.8x(z4)	*:701, EB*:490, LP*:92
225	2411	0.010	*	0.606	1.1	30.8	EB*:3.1x(z31), RG*:1.4x(z2), cC*:2.0x(z2)	*:1460, EB*:687, RG*:45
90	1260	0.016	EB*	0.285	3.1	30.3	AGN:12.0x(z30), RR*:7.5x(z28), QSO:16.8x(z25)	EB*:359, *:334, RR*:136
4	3095	0.014	*	0.556	1.0	30.3	EB*:2.8x(z30), Pe*:3.5x(z10), RG*:2.2x(z8)	*:1722, EB*:803, RG*:94
181	2455	0.023	*	0.464	0.9	30.1	LP*:3.0x(z30), LM*:4.9x(z9), AB?:2.4x(z4)	*:1138, LP*:661, EB*:228
247	2687	0.011	*	0.593	1.1	30.0	EB*:2.9x(z30), RR*:1.9x(z6), cC*:3.0x(z4)	*:1593, EB*:726, PM*:86
132	2062	0.024	*	0.476	0.9	30.0	LP*:3.2x(z30), s*r:14.8x(z18), AB?:5.6x(z13)	*:981, LP*:588, EB*:116
146	2255	0.013	*	0.553	1.0	30.0	EB*:3.1x(z30), RR*:2.0x(z6), BY*:3.0x(z5)	*:1248, EB*:645, RR*:66
9	2230	0.010	*	0.535	1.0	29.9	EB*:3.1x(z30), cC*:8.5x(z14), RR*:3.0x(z12)	*:1193, EB*:640, RR*:98

Atom	N_{match}	Hit rate	Dominant	Frac.	Enrich.	z_{max}	Over-represented	Types
39	2742	0.013	*	0.528	1.0	29.9	RR*:5.7x(z30), EB*:2.9x(z30), WD*:2.3x(z7)	*:1449, EB*:730, RR*:228
66	1481	0.011	*	0.402	0.7	29.8	EB*:3.5x(z30), RR*:3.4x(z11), LP*:1.6x(z7)	*:596, EB*:488, LP*:207
48	3668	0.014	*	0.603	1.1	29.6	EB*:2.6x(z30), cC*:5.9x(z11), RG*:1.9x(z7)	*:2211, EB*:890, RG*:96
72	3769	0.014	*	0.591	1.1	29.1	EB*:2.6x(z29), BY*:3.8x(z9), V*:2.2x(z6)	*:2227, EB*:897, PM*:185
170	821	0.011	EB*	0.403	4.3	29.1	EB*:4.3x(z29), RR*:6.6x(z19), QSO:4.3x(z4)	EB*:331, *:289, RR*:78
2	4773	0.026	*	0.572	1.0	29.1	RG*:4.6x(z29), EB*:2.2x(z26), Pe*:4.3x(z16)	*:2730, EB*:997, RG*:299
177	3341	0.029	*	0.559	1.0	28.7	LP*:2.7x(z29), AB?:4.6x(z13), SB*:1.8x(z7)	*:1866, LP*:790, EB*:193
164	3691	0.014	*	0.640	1.2	28.5	EB*:2.5x(z29), V*:2.1x(z5), HB*:2.2x(z4)	*:2362, EB*:874, SB*:80
16	5150	0.025	*	0.642	1.2	28.4	BY*:8.4x(z28), EB*:1.6x(z14), RG*:2.3x(z11)	*:3306, EB*:793, SB*:199
77	1450	0.025	EB*	0.162	1.7	28.4	Mi*:9.1x(z28), WD*:6.6x(z24), RR*:5.1x(z19)	EB*:235, LP?:219, LP*:215
63	3296	0.030	*	0.491	0.9	28.1	LP*:2.6x(z28), s*r:10.4x(z15), AB?:2.9x(z7)	*:1618, LP*:772, EB*:330
187	3616	0.050	*	0.407	0.7	28.0	LP*:2.6x(z28), AB?:6.7x(z21), BY*:6.8x(z19)	*:1470, LP*:822, EB*:309
127	1202	0.011	*	0.377	0.7	28.0	RR*:7.7x(z28), EB*:3.2x(z24), Pe*:3.3x(z6)	*:453, EB*:363, RR*:134
171	2444	0.011	*	0.548	1.0	27.9	EB*:2.8x(z28), RR*:3.9x(z17), cC*:8.0x(z13)	*:1339, EB*:650, RR*:137
141	1393	0.027	LP?	0.268	3.8	27.9	LP?:3.8x(z28), Y*?:5.9x(z15), LP*:2.2x(z13)	LP?:374, *:294, LP*:273
43	2970	0.013	*	0.570	1.0	27.9	EB*:2.7x(z28), RR*:2.7x(z11), AGN:2.8x(z8)	*:1693, EB*:741, RR*:116
83	2953	0.015	*	0.583	1.1	27.9	EB*:2.7x(z28), BY*:7.9x(z20), V*:2.9x(z8)	*:1721, EB*:738, BY*:66
106	3722	0.017	*	0.605	1.1	27.8	EB*:2.5x(z28), RR*:3.0x(z15), RS*:2.4x(z5)	*:2253, EB*:866, RR*:164
157	3799	0.016	*	0.570	1.0	27.7	EB*:2.5x(z28), RR*:2.6x(z12), BY*:4.1x(z10)	*:2167, EB*:876, RR*:142
131	593	0.014	LP?	0.202	2.9	27.7	WD*:11.3x(z28), Em*:5.2x(z16), LP?:2.9x(z12)	LP?:120, *:108, WD*:82
112	10918	0.122	*	0.757	1.4	27.5	Pu*:5.6x(z28), RR*:2.9x(z24), gD*:6.8x(z17)	*:8269, EB*:915, RR*:459
69	1263	0.017	*	0.541	1.0	27.2	EB*:3.5x(z27), RS*:3.3x(z4), RR*:1.6x(z3)	*:683, EB*:413, RR*:30
53	2590	0.012	*	0.530	1.0	27.0	EB*:2.7x(z27), RR*:4.1x(z19), cC*:2.9x(z4)	*:1372, EB*:662, RR*:155
151	2030	0.011	*	0.478	0.9	27.0	EB*:3.0x(z27), RR*:5.5x(z25), WD*:4.7x(z18)	*:971, EB*:561, RR*:163
233	3210	0.012	*	0.632	1.2	26.7	EB*:2.5x(z27), RG*:2.8x(z12), HB*:3.0x(z7)	*:2030, EB*:762, RG*:121
5	2040	0.012	*	0.557	1.0	26.6	EB*:2.9x(z27), RR*:2.4x(z7), QSO:3.5x(z5)	*:1137, EB*:558, RR*:70
79	3131	0.011	*	0.658	1.2	26.6	EB*:2.6x(z27), V*:2.1x(z5), RS*:2.6x(z5)	*:2059, EB*:747, RG*:53
47	3573	0.014	*	0.610	1.1	26.5	EB*:2.5x(z26), RS*:2.1x(z3), cC*:2.3x(z3)	*:2178, EB*:817, PM*:161
125	3943	0.017	*	0.593	1.1	26.4	EB*:2.4x(z26), RR*:2.7x(z13), Pe*:2.1x(z5)	*:2340, EB*:874, RR*:156
130	1160	0.017	*	0.309	0.6	26.1	AGN:10.9x(z26), QSO:13.9x(z20), Em*:4.6x(z19)	*:359, EB*:195, Em*:136
148	4905	0.019	*	0.650	1.2	25.4	EB*:2.2x(z25), RG*:2.4x(z12), HB*:3.0x(z8)	*:3190, EB*:1001, RG*:162
197	2754	0.012	*	0.592	1.1	25.1	EB*:2.6x(z25), RR*:2.7x(z11), cC*:4.2x(z6)	*:1631, EB*:660, RR*:109
122	1851	0.019	*	0.516	0.9	25.1	EB*:2.9x(z25), Pe*:2.5x(z4), El*:3.0x(z3)	*:956, EB*:503, LP*:89
153	1816	0.020	*	0.405	0.7	25.1	Em*:4.7x(z25), EB*:1.9x(z12), LM*:4.6x(z7)	*:736, EB*:326, Em*:218
200	1080	0.010	*	0.474	0.9	24.9	EB*:3.5x(z25), LM*:4.2x(z5), RR*:1.9x(z3)	*:512, EB*:351, LP*:52
61	4163	0.035	*	0.630	1.2	24.7	EB*:2.3x(z25), RR*:2.9x(z15), Pu*:4.8x(z14)	*:2621, EB*:875, RR*:176
152	6682	0.034	*	0.727	1.3	24.7	RG*:3.6x(z25), BY*:3.5x(z11), HB*:3.1x(z10)	*:4861, EB*:592, RG*:326
28	9691	0.049	*	0.771	1.4	24.6	RG*:3.1x(z25), HB*:3.0x(z12), SB*:1.5x(z8)	*:7473, EB*:639, RG*:414
140	3216	0.025	*	0.626	1.1	24.5	RR*:4.6x(z25), **:4.0x(z7), cC*:4.0x(z7)	*:2012, EB*:336, RR*:214

Atom	N_{match}	Hit rate	Dominant	Frac.	Enrich.	z_{max}	Over-represented	Types
81	3721	0.016	*	0.658	1.2	24.5	EB*:2.3x(z25), cC*:3.6x(z6), dS*:2.0x(z2)	*:2447, EB*:804, PM*:99
20	2106	0.011	*	0.557	1.0	24.3	EB*:2.7x(z24), RR*:2.1x(z6), El*:3.5x(z5)	*:1174, EB*:537, LP*:133
178	1508	0.012	*	0.507	0.9	24.2	EB*:3.0x(z24), BY*:3.5x(z5), El*:3.7x(z4)	*:764, EB*:428, LP*:117
52	3187	0.015	*	0.598	1.1	23.9	EB*:2.4x(z24), RR*:4.2x(z22), cC*:5.7x(z10)	*:1907, EB*:710, RR*:195
45	2657	0.011	*	0.641	1.2	23.6	EB*:2.5x(z24), RG*:2.4x(z8), BY*:2.4x(z4)	*:1704, EB*:620, RG*:85
248	2347	0.014	*	0.572	1.0	23.4	EB*:2.6x(z23), BY*:3.8x(z7), RS*:3.3x(z6)	*:1343, EB*:566, LP*:129
250	1151	0.013	*	0.367	0.7	23.4	RR*:6.7x(z23), Em*:4.5x(z19), EB*:2.8x(z19)	*:422, EB*:301, Em*:132
12	2417	0.010	*	0.651	1.2	22.9	EB*:2.5x(z23), cC*:2.8x(z3), WD*:1.5x(z3)	*:1574, EB*:569, WD*:46
32	3619	0.015	*	0.623	1.1	22.8	EB*:2.2x(z23), V*:2.7x(z8), BY*:3.4x(z8)	*:2254, EB*:757, PM*:158
212	2215	0.017	*	0.520	1.0	22.6	EB*:2.6x(z23), BY*:7.0x(z15), Pe*:4.8x(z13)	*:1151, EB*:531, RR*:59
21	911	0.015	*	0.472	0.9	22.2	Em*:5.6x(z22), RG*:3.5x(z9), AB?:5.2x(z8)	*:430, Em*:131, EB*:78
222	5394	0.019	*	0.648	1.2	22.0	EB*:2.0x(z22), RG*:2.4x(z12), RS*:2.7x(z7)	*:3496, EB*:998, RG*:179
80	1278	0.013	*	0.548	1.0	21.5	EB*:3.0x(z21), BY*:4.2x(z6), El*:4.8x(z5)	*:700, EB*:354, SB*:41
37	1999	0.014	*	0.449	0.8	21.4	LP*:2.6x(z21), EB*:2.0x(z14), Pe*:3.4x(z8)	*:898, LP*:462, EB*:381
186	1578	0.021	*	0.374	0.7	21.2	EB*:2.8x(z21), AGN:6.6x(z17), QSO:6.1x(z9)	*:590, EB*:405, LP?:81
41	2126	0.020	*	0.498	0.9	21.0	EB*:2.5x(z21), BY*:7.2x(z15), RR*:2.3x(z7)	*:1058, EB*:494, Em*:93
76	1807	0.021	*	0.391	0.7	20.9	Em*:4.1x(z21), LP*:2.4x(z17), AB?:6.2x(z13)	*:707, LP*:378, EB*:255
165	4714	0.028	*	0.639	1.2	20.9	EB*:2.0x(z21), Pu*:5.5x(z18), RR*:3.1x(z18)	*:3014, EB*:878, RR*:213
51	1473	0.026	LP?	0.214	3.0	20.8	LP?:3.0x(z21), Mi*:5.9x(z17), WD*:4.8x(z16)	LP?:315, LP*:303, *:262
188	10538	0.059	*	0.733	1.3	20.8	RR*:2.7x(z21), Pu*:2.5x(z9), SB*:1.5x(z8)	*:7725, EB*:1001, RR*:409
217	4047	0.018	*	0.561	1.0	20.7	RR*:3.7x(z21), EB*:2.0x(z19), cC*:6.2x(z13)	*:2271, EB*:745, PM*:254
207	1030	0.016	*	0.417	0.8	20.5	Em*:5.0x(z21), El*:9.5x(z11), EB*:2.1x(z11)	*:429, EB*:201, Em*:132
195	904	0.009	*	0.403	0.7	20.3	EB*:3.2x(z20), LP*:2.2x(z10), El*:7.5x(z8)	*:364, EB*:271, LP*:174
136	5116	0.029	*	0.666	1.2	20.2	EB*:1.9x(z20), RS*:3.6x(z10), dS*:4.6x(z8)	*:3407, EB*:919, SB*:152
44	924	0.016	*	0.396	0.7	20.1	AB?:11.7x(z20), RG*:4.9x(z14), Em*:3.8x(z14)	*:366, LP?:111, Em*:91
218	8796	0.033	*	0.689	1.3	19.9	PM*:2.1x(z20), SB*:1.8x(z12), BY*:2.2x(z6)	*:6064, EB*:964, PM*:696
19	1457	0.018	*	0.406	0.7	19.9	EB*:2.7x(z20), LP*:2.1x(z12), AB?:4.2x(z7)	*:591, EB*:368, LP*:268
105	855	0.020	*	0.199	0.4	19.7	Mi*:8.4x(z20), WD*:6.7x(z18), AGN:8.0x(z16)	*:170, LP?:158, Em*:96
211	4495	0.023	*	0.647	1.2	19.7	RG*:3.5x(z20), EB*:1.8x(z16), Pe*:2.8x(z9)	*:2909, EB*:750, RG*:215
219	1414	0.014	*	0.354	0.6	19.6	LP*:2.8x(z20), EB*:2.1x(z13), s*r:11.2x(z11)	*:500, LP*:345, EB*:281
22	967	0.011	*	0.444	0.8	19.3	EB*:3.0x(z19), cC*:4.9x(z5), El*:4.4x(z4)	*:429, EB*:274, Em*:43
183	729	0.009	*	0.460	0.8	19.2	EB*:3.3x(z19), LM*:6.8x(z8), WD*:2.0x(z3)	*:335, EB*:226, LP*:53
201	4315	0.034	*	0.582	1.1	19.1	BY*:6.5x(z19), LM*:4.5x(z11), LP*:1.4x(z9)	*:2511, LP*:552, EB*:388
160	1658	0.014	*	0.463	0.8	18.8	EB*:2.5x(z19), RG*:3.3x(z11), RR*:3.3x(z11)	*:767, EB*:388, LP*:94
214	7389	0.029	*	0.642	1.2	18.5	RG*:2.8x(z19), EB*:1.6x(z17), Pe*:3.5x(z15)	*:4742, EB*:1132, RG*:286
118	3096	0.014	*	0.690	1.3	18.2	EB*:2.1x(z18), RG*:1.8x(z5), El*:2.8x(z4)	*:2136, EB*:599, RG*:74
172	9560	0.050	*	0.814	1.5	17.4	RG*:2.5x(z17), SB*:1.8x(z13), HB*:2.4x(z8)	*:7783, SB*:458, EB*:416
167	5059	0.024	*	0.697	1.3	17.1	RG*:3.1x(z17), EB*:1.4x(z8), Pe*:2.4x(z7)	*:3526, EB*:653, RG*:211
221	9053	0.044	*	0.805	1.5	17.1	PM*:1.9x(z17), SB*:1.7x(z10)	*:7287, PM*:658, EB*:399

Atom	N_{match}	Hit rate	Dominant	Frac.	Enrich.	z_{max}	Over-represented	Types
38	4919	0.025	*	0.688	1.3	17.0	EB*:1.8x(z17), AGN:2.5x(z8), Pu*:2.6x(z6)	*:3385, EB*:823, SB*:140
158	5552	0.025	*	0.650	1.2	17.0	RG*:3.0x(z17), EB*:1.6x(z14), Pe*:3.3x(z12)	*:3610, EB*:846, RG*:223
185	6618	0.030	*	0.689	1.3	17.0	EB*:1.7x(z17), RG*:2.0x(z10), Pe*:2.6x(z9)	*:4558, EB*:1039, SB*:214
147	1205	0.016	*	0.528	1.0	16.6	EB*:2.6x(z17), HB*:4.1x(z7), RG*:2.3x(z5)	*:636, EB*:288, SB*:40
27	8652	0.037	*	0.798	1.5	16.4	PM*:1.9x(z16), SB*:1.5x(z7), **:1.5x(z2)	*:6903, PM*:624, EB*:506
33	6944	0.026	*	0.680	1.2	16.4	EB*:1.6x(z16), RG*:2.3x(z12), Pe*:3.0x(z12)	*:4723, EB*:1066, SB*:231
1	6018	0.025	*	0.772	1.4	16.3	PM*:2.1x(z16), SB*:1.4x(z5), cC*:1.8x(z2)	*:4646, PM*:473, EB*:404
137	1304	0.020	*	0.383	0.7	16.1	Em*:3.8x(z16), AB?:7.3x(z14), Y*?:4.5x(z10)	*:500, LP*:148, EB*:145
59	2097	0.022	*	0.422	0.8	15.7	Em*:3.1x(z16), PM*:2.1x(z10), AB?:3.9x(z8)	*:885, EB*:287, LP*:276
18	6573	0.032	*	0.714	1.3	15.7	RR*:2.6x(z16), PM*:1.5x(z8), SB*:1.3x(z4)	*:4692, EB*:700, PM*:371
192	5090	0.026	*	0.668	1.2	15.6	EB*:1.7x(z16), RG*:2.8x(z15), Pe*:3.0x(z10)	*:3401, EB*:816, RG*:192
124	2684	0.019	*	0.514	0.9	15.6	EB*:2.0x(z16), LP*:1.8x(z13), El*:3.0x(z4)	*:1380, EB*:498, LP*:437
234	1100	0.014	*	0.467	0.9	15.6	Em*:3.9x(z16), AB?:5.1x(z8), RG*:2.9x(z7)	*:514, Em*:111, EB*:104
169	4676	0.019	*	0.677	1.2	15.5	PM*:2.2x(z16), EB*:1.5x(z9), RG*:1.6x(z5)	*:3165, EB*:633, PM*:383
23	1276	0.016	*	0.292	0.5	15.5	WD*:4.9x(z15), EB*:2.0x(z11), LP*:1.7x(z7)	*:373, EB*:237, LP*:189
7	7216	0.038	*	0.711	1.3	15.4	RG*:2.6x(z15), SB*:1.5x(z7), HB*:1.5x(z3)	*:5131, LP*:546, EB*:512
182	1412	0.014	*	0.485	0.9	15.1	LP*:2.4x(z15), EB*:1.4x(z4), s*r:4.3x(z4)	*:685, LP*:294, EB*:179
176	3238	0.022	*	0.598	1.1	15.0	EB*:1.9x(z15), BY*:5.4x(z13), Pe*:3.7x(z11)	*:1937, EB*:563, SB*:129
58	1300	0.015	*	0.551	1.0	14.9	AB?:7.7x(z15), HB*:4.9x(z8), SB*:2.4x(z8)	*:716, LP*:144, EB*:114
194	656	0.008	*	0.399	0.7	14.9	EB*:2.9x(z15), El*:5.6x(z5), Em*:1.7x(z3)	*:262, EB*:178, LP?:61
101	11247	0.095	*	0.800	1.5	14.7	gD*:5.9x(z15), Pu*:3.3x(z14), SB*:1.5x(z8)	*:8995, EB*:774, SB*:441
35	7082	0.042	*	0.801	1.5	14.6	PM*:1.9x(z15), SB*:1.7x(z10), **:1.3x(z1)	*:5674, PM*:507, SB*:325
14	3829	0.028	*	0.644	1.2	14.6	EB*:1.8x(z15), RR*:2.1x(z8), PM*:1.3x(z4)	*:2464, EB*:633, PM*:190
34	4603	0.018	*	0.686	1.3	14.2	EB*:1.7x(z14), RG*:2.3x(z10), Pe*:2.9x(z9)	*:3158, EB*:724, RG*:144
180	10184	0.066	*	0.751	1.4	14.1	Pu*:3.4x(z14), gD*:5.4x(z13), RR*:2.0x(z12)	*:7647, EB*:898, SB*:361
230	6864	0.038	*	0.797	1.5	14.0	PM*:1.9x(z14), SB*:1.6x(z8)	*:5469, PM*:485, EB*:387
173	8290	0.035	*	0.708	1.3	14.0	RG*:2.3x(z14), Pe*:2.8x(z12), PM*:1.6x(z11)	*:5872, EB*:700, PM*:510
30	13154	0.107	*	0.791	1.5	13.9	Pu*:3.1x(z14), gD*:4.9x(z13), SB*:1.5x(z10)	*:10404, PM*:636, EB*:606
159	9812	0.047	*	0.785	1.4	13.9	RG*:2.2x(z14), SB*:1.8x(z12), HB*:2.3x(z8)	*:7699, EB*:599, SB*:451
103	3193	0.019	*	0.674	1.2	13.7	EB*:1.8x(z14), **:4.6x(z8), V*:1.6x(z3)	*:2152, EB*:535, PM*:99
86	4755	0.019	*	0.707	1.3	13.4	EB*:1.6x(z13), SB*:1.3x(z3), dS*:2.0x(z2)	*:3363, EB*:726, SB*:157
149	2976	0.014	*	0.701	1.3	13.2	EB*:1.8x(z13), cC*:3.4x(z5), RR*:1.5x(z3)	*:2087, EB*:497, PM*:75
245	1535	0.025	*	0.403	0.7	13.1	Em*:3.1x(z13), EB*:2.0x(z12), Y*?:4.0x(z10)	*:618, EB*:286, LP?:135
46	9986	0.047	*	0.799	1.5	13.1	PM*:1.7x(z13), SB*:1.7x(z11), **:2.0x(z4)	*:7976, PM*:632, EB*:556
223	7043	0.028	*	0.765	1.4	13.1	RG*:2.3x(z13), Pe*:2.3x(z8), SB*:1.4x(z6)	*:5386, EB*:609, SB*:260
110	4697	0.023	*	0.677	1.2	12.4	EB*:1.6x(z12), RR*:2.1x(z9), Pu*:3.1x(z8)	*:3180, EB*:698, PM*:195
138	612	0.013	*	0.428	0.8	12.4	RG*:5.3x(z12), El*:9.0x(z8), Em*:2.5x(z6)	*:262, EB*:95, RG*:44
108	11660	0.073	*	0.831	1.5	11.9	SB*:1.7x(z12), PM*:1.3x(z7), **:1.8x(z3)	*:9694, PM*:594, SB*:514
228	5442	0.033	*	0.738	1.4	11.7	SB*:2.0x(z12), RG*:2.3x(z11), Y*?:2.0x(z6)	*:4015, EB*:343, SB*:283

Atom	N_{match}	Hit rate	Dominant	Frac.	Enrich.	$z_{\max}$	Over-represented	Types
161	7393	0.034	*	0.750	1.4	11.5	EB*:1.4x(z12), Pu*:1.9x(z4), SB*:1.2x(z3)	*:5547, EB*:992, SB*:237
102	10066	0.047	*	0.813	1.5	11.4	PM*:1.6x(z11), SB*:1.6x(z10), **:1.2x(z1)	*:8181, PM*:602, EB*:522
10	5669	0.027	*	0.759	1.4	11.3	PM*:1.8x(z11), SB*:1.3x(z3), V*:1.5x(z3)	*:4300, EB*:444, PM*:380
11	8596	0.038	*	0.819	1.5	11.1	RG*:2.0x(z11), SB*:1.6x(z9), HB*:2.6x(z9)	*:7036, EB*:398, SB*:365
168	3966	0.017	*	0.732	1.3	11.0	EB*:1.6x(z11), cC*:4.2x(z8), RR*:1.4x(z3)	*:2902, EB*:582, PM*:113
240	11844	0.073	*	0.812	1.5	11.0	gD*:4.5x(z11), Pu*:2.5x(z9), SB*:1.4x(z8)	*:9616, EB*:737, SB*:451
206	5964	0.040	*	0.745	1.4	11.0	RG*:2.2x(z11), BY*:3.6x(z11), SB*:1.6x(z7)	*:4442, EB*:379, PM*:297
154	9900	0.060	*	0.813	1.5	10.9	SB*:1.7x(z11), Pu*:1.5x(z3), PM*:1.1x(z2)	*:8052, EB*:600, SB*:436
96	2495	0.018	*	0.639	1.2	10.9	BY*:5.1x(z11), EB*:1.6x(z10), Pe*:3.2x(z8)	*:1594, EB*:382, SB*:106
91	9800	0.038	*	0.779	1.4	10.0	SB*:1.6x(z10), PM*:1.5x(z9), Pe*:1.1x(z1)	*:7635, EB*:735, PM*:542
97	5428	0.032	*	0.790	1.4	8.5	SB*:1.7x(z8), PM*:1.4x(z5), Pe*:1.2x(z1)	*:4286, EB*:336, PM*:279
95	5236	0.023	*	0.726	1.3	8.2	EB*:1.4x(z8), RG*:1.9x(z7), BY*:2.3x(z5)	*:3802, EB*:669, SB*:165
193	6302	0.035	*	0.722	1.3	7.7	EB*:1.3x(z8), Pe*:2.1x(z6), cC*:2.7x(z5)	*:4548, EB*:775, SB*:221
139	6402	0.023	*	0.757	1.4	7.7	PM*:1.5x(z8), SB*:1.4x(z5), BY*:1.6x(z3)	*:4849, EB*:590, PM*:361
109	6706	0.033	*	0.738	1.4	7.3	SB*:1.6x(z7), Pe*:2.0x(z6), EB*:1.2x(z5)	*:4947, EB*:753, SB*:273
231	7713	0.029	*	0.785	1.4	6.9	HB*:2.3x(z7), SB*:1.4x(z5), RG*:1.5x(z5)	*:6057, EB*:608, SB*:276
92	7105	0.030	*	0.770	1.4	6.8	SB*:1.5x(z7), Pu*:1.6x(z3), EB*:1.1x(z3)	*:5471, EB*:734, SB*:279
128	6696	0.040	*	0.782	1.4	6.6	RG*:1.7x(z7), SB*:1.4x(z5), PM*:1.1x(z2)	*:5234, EB*:387, PM*:289
253	1442	0.017	*	0.561	1.0	5.9	SB*:2.0x(z6), RG*:2.1x(z5), EB*:1.4x(z5)	*:809, EB*:190, LP*:158
198	5801	0.021	*	0.739	1.4	5.7	EB*:1.2x(z6), RG*:1.6x(z6), HB*:2.1x(z5)	*:4289, EB*:675, PM*:179
75	7997	0.036	*	0.757	1.4	5.0	RR*:1.5x(z5), dS*:2.4x(z4), EB*:1.1x(z4)	*:6054, EB*:848, PM*:310
113	3254	0.023	*	0.708	1.3	4.7	LM*:2.7x(z5), PM*:1.3x(z3), RR*:1.2x(z2)	*:2304, EB*:304, LP*:163
237	4025	0.027	*	0.680	1.2	3.6	LM*:2.2x(z4), RG*:1.4x(z3), Y*?:1.5x(z2)	*:2735, LP*:375, EB*:329
3	3338	0.022	*	0.780	1.4			*:2602, EB*:246, PM*:120

Table 23: SIMBAD object types of the stars that make up the top 1% of activations of each atom in D_H, ranked by Poisson significance of the largest overrepresented class.

Atom	N_{match}	Hit rate	Dominant	Frac.	Enrich.	$z_{\max}$	Over-represented	Types
144	439	0.864	C*	0.405	158.0	166.6	C*:158.0x(z167), Mi*:29.4x(z54), LP*:2.5x(z9)	C*:178, Mi*:108, LP*:98
54	520	0.360	WD*	0.763	62.2	154.6	WD*:62.2x(z155), AGN:25.2x(z43), QSO:27.2x(z27)	WD*:397, AGN:79, QSO:28
15	662	0.494	WD*	0.642	52.3	146.2	WD*:52.3x(z146), AGN:20.1x(z38), QSO:18.3x(z20)	WD*:425, AGN:80, *:71
36	196	0.975	Mi*	0.668	79.8	101.0	Mi*:79.8x(z101), AGN:15.3x(z15), LP*:1.6x(z3)	Mi*:131, LP*:28, AGN:18
89	149	0.980	Mi*	0.732	87.3	96.5	Mi*:87.3x(z96), LP*:2.1x(z4)	Mi*:109, LP*:28, AB*:4
227	63	0.059	C*	0.587	228.9	91.6	C*:228.9x(z92), LP*:2.0x(z2)	C*:37, LP*:11, *:7
40	182	0.413	WD*	0.703	57.3	84.1	WD*:57.3x(z84), HS?:61.5x(z36)	WD*:128, HS?:22, EB*:12
166	187	0.356	Mi*	0.561	67.0	82.6	Mi*:67.0x(z83), LM*:15.5x(z10), LP*:1.9x(z4)	Mi*:105, LP*:31, PM*:14

Atom	N_{match}	Hit rate	Dominant	Frac.	Enrich.	$z_{\max}$	Over-represented	Types
111	204	0.109	WD*	0.642	52.3	81.2	WD*:52.3x(z81), AGN:33.4x(z36), QSO:27.2x(z17)	WD*:131, AGN:41, QSO:11
210	331	0.173	RR*	0.535	36.9	78.7	RR*:36.9x(z79), AGN:35.1x(z48), QSO:33.6x(z26)	RR*:177, AGN:70, WD*:23
104	451	0.321	AGN	0.279	46.4	74.8	AGN:46.4x(z75), QSO:54.9x(z51), RR*:16.1x(z39)	AGN:126, RR*:105, WD*:70
88	267	0.353	*	0.255	0.5	65.5	HS?:91.4x(z66), WD*:17.1x(z29), EB*:2.0x(z5)	*:68, WD*:56, EB*:51
93	118	0.975	Mi*	0.559	66.8	65.4	Mi*:66.8x(z65), LP?:3.6x(z8), LP*:1.6x(z2)	Mi*:66, LP*:30, LP*:17
70	165	0.458	WD*	0.394	32.1	59.1	HS?:104.8x(z59), WD*:32.1x(z44)	WD*:65, HS?:34, *:24
71	510	0.943	LP?	0.549	7.8	57.0	Mi*:28.6x(z57), LP?:7.8x(z41), LP*:2.0x(z7)	LP?:280, Mi*:122, LP*:91
151	100	0.053	WD*	0.610	49.7	53.9	WD*:49.7x(z54), AGN:18.3x(z13), QSO:30.3x(z13)	WD*:61, AGN:11, RR*:8
123	68	0.029	WD*	0.706	57.5	51.6	WD*:57.5x(z52), EB*:1.1x(z0)	WD*:48, EB*:7, AGN:4
217	109	0.048	AGN	0.376	62.5	49.8	AGN:62.5x(z50), QSO:55.6x(z25), WD*:14.2x(z15)	AGN:41, WD*:19, QSO:12
0	96	0.133	*	0.365	0.7	49.6	s*r:177.7x(z50), AB?:33.1x(z19), LP*:3.1x(z6)	*:35, LP*:26, s*r:14
249	76	0.034	WD*	0.632	51.4	48.7	WD*:51.4x(z49), AGN:19.7x(z13), EB*:1.1x(z0)	WD*:48, AGN:9, EB*:8
204	80	0.041	WD*	0.600	48.8	47.4	WD*:48.8x(z47), AGN:39.5x(z27), QSO:31.6x(z12)	WD*:48, AGN:19, QSO:5
155	56	0.023	WD*	0.661	53.8	43.8	WD*:53.8x(z44)	WD*:37, QSO:4, EB*:4
87	88	0.046	RR*	0.568	39.3	43.2	RR*:39.3x(z43), WD*:9.3x(z9), EB*:1.8x(z2)	RR*:50, EB*:15, WD*:10
205	114	0.062	WD*	0.456	37.1	42.8	WD*:37.1x(z43), QSO:70.9x(z33), AGN:39.4x(z32)	WD*:52, AGN:27, QSO:16
229	363	0.955	LP?	0.642	9.1	41.1	LP?:9.1x(z41), Mi*:20.1x(z33), LP*:1.8x(z4)	LP?:233, Mi*:61, LP*:57
13	68	0.030	WD*	0.544	44.3	39.6	WD*:44.3x(z40), AGN:19.5x(z12), EB*:1.3x(z1)	WD*:37, AGN:8, EB*:8
199	58	0.033	WD*	0.586	47.7	39.4	WD*:47.7x(z39), EB*:2.2x(z3)	WD*:34, EB*:12, *:7
184	79	0.041	WD*	0.494	40.2	38.6	WD*:40.2x(z39), AGN:40.0x(z27), RR*:4.4x(z4)	WD*:39, AGN:19, EB*:9
98	149	0.943	LP?	0.530	7.5	38.3	Mi*:35.3x(z38), LP?:7.5x(z21)	LP?:79, Mi*:44, LP*:13
165	96	0.058	WD*	0.438	35.6	37.6	WD*:35.6x(z38), AGN:24.2x(z18), QSO:26.3x(z11)	WD*:42, AGN:14, *:10
39	64	0.029	WD*	0.531	43.2	37.5	WD*:43.2x(z37), AGN:20.8x(z12), RR*:6.5x(z5)	WD*:34, EB*:10, AGN:8
82	134	0.074	WD*	0.366	29.8	36.9	WD*:29.8x(z37), AGN:36.0x(z31), QSO:56.5x(z29)	WD*:49, AGN:29, RR*:26
173	176	0.075	*	0.409	0.8	36.3	Pe*:39.5x(z36), RG*:16.3x(z24), SB*:1.3x(z1)	*:72, RG*:39, Pe*:35
38	81	0.041	WD*	0.383	31.2	33.7	AGN:49.2x(z34), WD*:31.2x(z30), QSO:56.1x(z22)	WD*:31, AGN:24, QSO:9
130	37	0.055	AGN	0.432	71.9	33.4	AGN:71.9x(z33), QSO:81.9x(z22)	AGN:16, WR*:6, QSO:6
8	64	0.026	WD*	0.469	38.2	32.9	WD*:38.2x(z33), AGN:26.0x(z15)	WD*:30, AGN:10, *:10
214	105	0.041	RG*	0.362	26.6	30.6	RG*:26.6x(z31), WD*:3.9x(z3), EB*:1.0x(z0)	RG*:38, *:38, EB*:10
28	193	0.097	*	0.503	0.9	30.5	RG*:19.8x(z30), Pe*:22.7x(z21)	*:97, RG*:52, Pe*:22
50	65	0.029	WD*	0.431	35.1	30.4	WD*:35.1x(z30), AGN:43.5x(z27), QSO:46.6x(z16)	WD*:28, AGN:17, *:6
149	45	0.021	AGN	0.356	59.1	30.2	AGN:59.1x(z30), WD*:23.5x(z17)	AGN:16, WD*:13, QSO:4
116	96	0.048	WD*	0.354	28.8	30.2	WD*:28.8x(z30), AGN:27.7x(z20), QSO:31.6x(z13)	WD*:34, RR*:17, AGN:16
142	88	0.127	PM*	0.659	17.4	30.0	PM*:17.4x(z30), Em*:3.5x(z4)	PM*:58, *:17, Em*:8
126	42	0.019	WD*	0.524	42.6	29.9	WD*:42.6x(z30)	WD*:22, *:11, AGN:4
9	59	0.025	WD*	0.441	35.9	29.7	WD*:35.9x(z30), AGN:25.3x(z15), EB*:1.3x(z1)	WD*:26, *:10, AGN:9
191	11	0.023	WD*	1.000	81.4	29.6	WD*:81.4x(z30)	WD*:11
107	542	0.925	LP*	0.459	5.2	29.0	LP*:5.2x(z29), LM*:20.7x(z22), LP?:4.6x(z22)	LP*:249, LP?:175, PM*:36
43	79	0.034	AGN	0.253	42.1	28.3	AGN:42.1x(z28), WD*:14.4x(z13), QSO:32.0x(z12)	AGN:20, *:16, WD*:14

Atom	N_{match}	Hit rate	Dominant	Frac.	Enrich.	z_{max}	Over-represented	Types
179	61	0.029	WD*	0.393	32.0	26.9	WD*:32.0x(z27), QSO:49.7x(z17), AGN:27.2x(z16)	WD*:24, AGN:10, EB*:9
208	61	0.082	RR*	0.426	29.5	26.7	RR*:29.5x(z27)	RR*:26, *:24, EB*:4
203	38	0.107	AGN	0.342	56.8	26.7	AGN:56.8x(z27), QSO:66.5x(z18)	AGN:13, QSO:5, Em*:4
24	36	0.021	WD*	0.500	40.7	26.4	WD*:40.7x(z26), EB*:1.5x(z1)	WD*:18, *:5, EB*:5
86	26	0.010	WD*	0.577	47.0	26.0	WD*:47.0x(z26)	WD*:15, *:4, AGN:3
244	38	0.016	WD*	0.474	38.6	25.7	WD*:38.6x(z26), EB*:1.7x(z1)	WD*:18, *:6, EB*:6
163	99	0.109	*	0.263	0.5	25.1	AGN:33.6x(z25), RR*:12.6x(z14), QSO:30.6x(z13)	*:26, AGN:20, RR*:18
53	30	0.013	WD*	0.500	40.7	24.1	WD*:40.7x(z24)	WD*:15, *:4, EB*:3
64	50	0.097	LP?	0.540	7.7	24.1	Mi*:38.2x(z24), LP?:7.7x(z13)	LP?:27, Mi*:16, LP*:4
100	39	0.015	WD*	0.436	35.5	23.9	WD*:35.5x(z24), EB*:3.0x(z4)	WD*:17, EB*:11, *:7
197	57	0.024	WD*	0.333	27.1	23.3	AGN:40.8x(z23), WD*:27.1x(z22), QSO:53.2x(z18)	WD*:19, AGN:14, *:8
251	51	0.021	WD*	0.373	30.3	23.2	WD*:30.3x(z23), EB*:3.6x(z6)	WD*:19, EB*:17, *:8
255	46	0.020	WD*	0.391	31.9	23.2	WD*:31.9x(z23), AGN:28.9x(z15), EB*:2.3x(z3)	WD*:18, EB*:10, AGN:8
202	46	0.020	WD*	0.391	31.9	23.2	WD*:31.9x(z23), AGN:28.9x(z15), EB*:2.6x(z3)	WD*:18, EB*:11, AGN:8
78	46	0.018	WD*	0.391	31.9	23.2	WD*:31.9x(z23), AGN:28.9x(z15), EB*:2.8x(z4)	WD*:18, EB*:12, *:5
206	127	0.085	*	0.480	0.9	22.9	PM*:11.5x(z23)	*:61, PM*:55, SB*:4
81	59	0.025	WD*	0.339	27.6	22.6	WD*:27.6x(z23), AGN:19.7x(z11), EB*:3.6x(z6)	WD*:20, EB*:20, AGN:7
156	25	0.040	*	0.440	0.8	22.5	AB?:74.1x(z22), LP*:2.3x(z2)	*:11, AB?:7, LP*:5
158	71	0.032	*	0.338	0.6	21.7	RR*:22.4x(z22), EB*:2.0x(z2)	*:24, RR*:23, EB*:13
188	76	0.042	RR*	0.263	18.2	21.5	AGN:32.8x(z22), RR*:18.2x(z18), WD*:11.8x(z10)	RR*:20, AGN:15, EB*:14
232	39	0.016	WD*	0.359	29.2	19.5	WD*:29.2x(z20), EB*:1.6x(z1)	WD*:14, *:9, EB*:6
106	54	0.024	RR*	0.333	23.0	19.5	RR*:23.0x(z19), EB*:3.4x(z5)	RR*:18, EB*:17, *:10
136	219	0.126	*	0.689	1.3	19.5	RS*:25.6x(z19), Em*:3.2x(z5), V*:3.6x(z3)	*:151, Em*:18, RS*:16
171	45	0.019	WD*	0.333	27.1	19.4	WD*:27.1x(z19), AGN:33.2x(z17), EB*:1.9x(z2)	WD*:15, AGN:9, EB*:8
12	30	0.013	WD*	0.400	32.6	19.2	WD*:32.6x(z19), AGN:33.2x(z14)	WD*:12, AGN:6, *:4
215	60	0.032	WD*	0.283	23.1	18.9	WD*:23.1x(z19), AGN:30.5x(z18), RR*:10.4x(z9)	WD*:17, EB*:13, AGN:11
225	31	0.012	WD*	0.387	31.5	18.8	WD*:31.5x(z19), EB*:2.8x(z3)	WD*:12, *:9, EB*:8
55	22	0.062	Mi*	0.364	43.4	18.2	Mi*:43.4x(z18), LP?:4.5x(z4)	Mi*:8, LP?:7, LP*:4
216	40	0.017	WD*	0.325	26.5	17.8	WD*:26.5x(z18), EB*:2.7x(z3)	WD*:13, EB*:10, *:5
246	35	0.015	WD*	0.343	27.9	17.6	WD*:27.9x(z18), EB*:2.8x(z3)	WD*:12, *:10, EB*:9
247	26	0.010	EB*	0.308	3.3	17.3	AGN:44.7x(z17), EB*:3.3x(z4)	EB*:8, AGN:7, *:3
145	32	0.012	WD*	0.344	28.0	16.9	WD*:28.0x(z17), EB*:2.7x(z3)	WD*:11, *:11, EB*:8
56	98	0.051	EB*	0.612	6.6	16.8	EB*:6.6x(z17)	EB*:60, *:32, RS*:2
52	45	0.021	EB*	0.267	2.9	16.8	AGN:33.2x(z17), WD*:14.5x(z10), RR*:7.7x(z5)	EB*:12, AGN:9, WD*:8
183	18	0.023	WD*	0.444	36.2	16.5	WD*:36.2x(z17)	WD*:8, RR*:3, *:3
103	108	0.064	*	0.833	1.5	16.2	**:39.5x(z16)	*:90, **:7, EB*:5
180	62	0.040	*	0.403	0.7	15.9	RR*:17.8x(z16), EB*:2.1x(z3)	*:25, RR*:16, EB*:12
148	42	0.017	EB*	0.833	8.9	15.7	EB*:8.9x(z16)	EB*:35, *:4, HB?:1
161	104	0.047	*	0.404	0.7	15.6	AGN:20.8x(z16), EB*:3.7x(z8), RR*:3.3x(z3)	*:42, EB*:36, AGN:13

Atom	N_{match}	Hit rate	Dominant	Frac.	Enrich.	$z_{\max}$	Over-represented	Types
243	33	0.013	EB*	0.394	4.2	15.1	WD*:24.7x(z15), EB*:4.2x(z6)	EB*:13, WD*:10, *:6
213	35	0.015	*	0.314	0.6	14.8	AGN:33.2x(z15), EB*:2.8x(z3)	*:11, EB*:9, AGN:7
29	41	0.016	EB*	0.317	3.4	14.8	WD*:21.8x(z15), EB*:3.4x(z5)	EB*:13, *:11, WD*:11
60	28	0.012	EB*	0.393	4.2	14.8	WD*:26.2x(z15), EB*:4.2x(z5)	EB*:11, WD*:9, *:5
174	42	0.017	WD*	0.262	21.3	14.6	WD*:21.3x(z15), AGN:19.8x(z9), EB*:2.0x(z2)	WD*:11, *:8, EB*:8
47	35	0.014	WD*	0.286	23.3	14.6	WD*:23.3x(z15), AGN:23.7x(z10), EB*:3.1x(z4)	WD*:10, EB*:10, *:6
16	99	0.049	*	0.485	0.9	14.6	BY*:28.6x(z15), EB*:3.2x(z7)	*:48, EB*:30, BY*:8
175	50	0.022	EB*	0.380	4.1	14.5	WD*:19.5x(z15), EB*:4.1x(z7)	EB*:19, WD*:12, *:8
110	27	0.013	AGN	0.222	36.9	14.5	AGN:36.9x(z14), RR*:15.4x(z9), EB*:2.4x(z2)	AGN:6, EB*:6, RR*:6
75	39	0.018	EB*	0.282	3.0	14.0	AGN:29.8x(z14), RR*:12.4x(z9), EB*:3.0x(z4)	EB*:11, AGN:7, *:7
233	25	0.009	WD*	0.320	26.0	13.9	WD*:26.0x(z14), EB*:3.4x(z4)	WD*:8, *:8, EB*:8
18	40	0.019	RR*	0.275	19.0	13.7	RR*:19.0x(z14), AGN:24.9x(z12), WD*:16.3x(z11)	RR*:11, EB*:8, WD*:8
241	48	0.019	*	0.354	0.6	13.6	WD*:18.7x(z14), AGN:24.2x(z12), EB*:1.8x(z2)	*:17, WD*:11, EB*:8
125	42	0.018	*	0.286	0.5	13.4	AGN:27.7x(z13), WD*:15.5x(z10), EB*:1.8x(z2)	*:12, WD*:8, EB*:7
41	42	0.040	RR*	0.262	18.1	13.3	RR*:18.1x(z13), EB*:2.0x(z2)	RR*:11, EB*:8, *:7
84	39	0.017	EB*	0.744	8.0	13.3	EB*:8.0x(z13)	EB*:29, *:6, WD*:4
252	41	0.089	LP?	0.610	8.7	13.0	LP?:8.7x(z13), LP*:1.4x(z1)	LP?:25, LP*:5, Mi*:4
187	43	0.059	PM*	0.419	11.1	12.8	PM*:11.1x(z13), Em*:5.4x(z5)	PM*:18, *:12, Em*:6
119	37	0.016	EB*	0.405	4.3	12.7	WD*:19.8x(z13), EB*:4.3x(z6)	EB*:15, WD*:9, *:4
62	30	0.012	*	0.300	0.6	12.6	WD*:21.7x(z13), EB*:2.9x(z3)	*:9, EB*:8, WD*:8
115	43	0.024	EB*	0.674	7.2	12.5	EB*:7.2x(z12)	EB*:29, *:11, RR*:2
114	40	0.026	EB*	0.675	7.2	12.0	EB*:7.2x(z12)	EB*:27, *:10, LP*:2
235	34	0.015	EB*	0.353	3.8	11.7	WD*:19.2x(z12), EB*:3.8x(z5)	EB*:12, WD*:8, *:7
67	35	0.016	EB*	0.457	4.9	11.5	WD*:18.6x(z12), EB*:4.9x(z7)	EB*:16, WD*:8, *:6
83	42	0.021	WR*	0.190	769.6	11.4	AGN:23.7x(z11), EB*:1.5x(z1)	WR*:8, *:7, EB*:6
143	36	0.016	EB*	0.417	4.5	11.4	WD*:18.1x(z11), EB*:4.5x(z6)	EB*:15, *:9, WD*:8
74	28	0.010	*	0.321	0.6	11.3	WD*:20.4x(z11), EB*:3.4x(z4)	*:9, EB*:9, WD*:7
201	30	0.024	*	0.533	1.0	11.1	PM*:11.5x(z11)	*:16, PM*:13, LM*:1
133	34	0.014	EB*	0.647	6.9	10.6	EB*:6.9x(z11)	EB*:22, *:12
219	49	0.049	LP?	0.469	6.7	10.5	LP?:6.7x(z11), WD*:8.3x(z6)	LP?:23, WD*:5, LP*:4
73	36	0.018	EB*	0.444	4.8	10.4	RR*:15.4x(z10), EB*:4.8x(z7)	EB*:16, RR*:8, *:6
31	26	0.037	PM*	0.346	9.2	10.2	Em*:13.5x(z10), PM*:9.2x(z8)	PM*:9, Em*:9, *:8
48	34	0.013	*	0.324	0.6	10.2	WD*:16.8x(z10), EB*:1.6x(z1)	*:11, WD*:7, EB*:5
7	93	0.049	LP*	0.398	4.5	10.0	LP*:4.5x(z10), RG*:7.9x(z8)	LP*:37, *:36, RG*:10
135	29	0.017	EB*	0.655	7.0	9.9	EB*:7.0x(z10)	EB*:19, *:9, PN:1
25	46	0.018	EB*	0.348	3.7	9.9	WD*:14.2x(z10), EB*:3.7x(z6)	EB*:16, *:14, WD*:8
218	117	0.044	*	0.402	0.7	9.9	BY*:18.1x(z10), EB*:2.9x(z6), PM*:3.4x(z5)	*:47, EB*:32, PM*:15
2	30	0.016	EB*	0.633	6.8	9.7	EB*:6.8x(z10)	EB*:19, *:7, V*:2
35	113	0.067	*	0.619	1.1	9.5	PM*:5.6x(z10), SB*:3.4x(z4), Em*:1.7x(z1)	*:70, PM*:24, SB*:10

Atom	N_{match}	Hit rate	Dominant	Frac.	Enrich.	$z_{\max}$	Over-represented	Types
189	29	0.011	EB*	0.345	3.7	9.5	WD*:16.8x(z9), EB*:3.7x(z4)	EB*:10, WD*:6, *:5
19	14	0.018	EB*	0.857	9.2	9.4	EB*:9.2x(z9)	EB*:12, *:1, RR*:1
128	65	0.039	*	0.631	1.2	9.3	PM*:6.9x(z9)	*:41, PM*:17, RR*:4
160	17	0.014	EB*	0.765	8.2	9.1	EB*:8.2x(z9)	EB*:13, *:2, Em*:1
222	25	0.009	EB*	0.640	6.9	8.9	EB*:6.9x(z9)	EB*:16, *:8, BY*:1
129	13	0.008	EB*	0.846	9.1	8.9	EB*:9.1x(z9)	EB*:11, *:2
51	116	0.201	G	0.319	556.5	8.7	RR*:7.7x(z9), WD*:7.0x(z7)	G:37, RR*:13, WD*:10
162	63	0.031	EB*	0.429	4.6	8.7	EB*:4.6x(z9), RR*:8.8x(z7)	EB*:27, *:14, RR*:8
186	24	0.031	WD*	0.208	17.0	8.7	WD*:17.0x(z9)	WD*:5, Em*:3, QSO:2
152	65	0.033	*	0.554	1.0	8.6	RG*:10.2x(z9), V*:12.3x(z7), EB*:1.3x(z1)	*:36, RG*:9, EB*:8
254	21	0.009	EB*	0.667	7.1	8.6	EB*:7.1x(z9)	EB*:14, *:4, Em*:1
92	21	0.009	EB*	0.381	4.1	8.5	RR*:16.5x(z9), EB*:4.1x(z4)	EB*:8, RR*:5, WD*:3
26	25	0.011	EB*	0.360	3.9	8.5	WD*:16.3x(z8), EB*:3.9x(z4)	EB*:9, *:6, WD*:5
65	51	0.077	EB*	0.353	3.8	8.5	RR*:10.8x(z8), EB*:3.8x(z6)	EB*:18, *:12, RR*:8
30	820	0.666	*	0.779	1.4	8.3	**:8.2x(z8), PM*:2.2x(z6), gD*:7.4x(z5)	*:639, PM*:67, SB*:36
220	28	0.017	EB*	0.571	6.1	8.3	EB*:6.1x(z8)	EB*:16, *:9, RR*:2
211	15	0.008	EB*	0.733	7.9	8.1	EB*:7.9x(z8)	EB*:11, *:2, El*:1
94	31	0.014	EB*	0.484	5.2	7.5	WD*:13.1x(z7), EB*:5.2x(z7)	EB*:15, *:6, WD*:5
85	17	0.008	EB*	0.647	6.9	7.5	EB*:6.9x(z7)	EB*:11, *:6
6	15	0.010	EB*	0.667	7.1	7.3	EB*:7.1x(z7)	EB*:10, *:2, PM*:1
57	15	0.009	EB*	0.667	7.1	7.3	EB*:7.1x(z7)	EB*:10, *:4, RR*:1
248	13	0.008	EB*	0.692	7.4	7.1	EB*:7.4x(z7)	EB*:9, *:2, WR*:1
22	31	0.034	Em*	0.226	8.8	6.9	Em*:8.8x(z7)	Em*:7, AGN:3, *:3
239	33	0.015	EB*	0.455	4.9	6.8	EB*:4.9x(z7)	EB*:15, *:8, WD*:4
157	26	0.011	EB*	0.500	5.4	6.8	EB*:5.4x(z7)	EB*:13, *:8, RR*:3
238	26	0.012	EB*	0.500	5.4	6.8	EB*:5.4x(z7)	EB*:13, *:12, WD*:1
224	41	0.022	*	0.488	0.9	6.7	EB*:4.4x(z7)	*:20, EB*:17, RR*:3
80	14	0.015	EB*	0.643	6.9	6.7	EB*:6.9x(z7)	EB*:9, *:3, PN:1
122	23	0.023	EB*	0.522	5.6	6.7	EB*:5.6x(z7)	EB*:12, *:9, LP*:1
193	95	0.052	*	0.632	1.2	6.5	RR*:6.5x(z7), SB*:2.0x(z2), EB*:1.2x(z1)	*:60, EB*:11, RR*:9
4	21	0.010	EB*	0.524	5.6	6.5	EB*:5.6x(z6)	EB*:11, *:10
190	28	0.014	EB*	0.464	5.0	6.4	EB*:5.0x(z6)	EB*:13, *:12, RR*:1
178	18	0.014	EB*	0.556	6.0	6.4	EB*:6.0x(z6)	EB*:10, *:8
32	32	0.013	*	0.469	0.9	6.4	EB*:4.7x(z6)	*:15, EB*:14, PN:1
192	36	0.019	EB*	0.417	4.5	6.4	EB*:4.5x(z6)	EB*:15, *:15, El*:2
207	10	0.016	EB*	0.700	7.5	6.3	EB*:7.5x(z6)	EB*:7, *:2, El*:1
150	29	0.012	EB*	0.448	4.8	6.3	EB*:4.8x(z6)	EB*:13, *:11, WD*:2
108	438	0.273	*	0.838	1.5	6.2	**:8.3x(z6), SB*:2.0x(z3), PM*:1.6x(z2)	*:367, PM*:26, SB*:23
46	32	0.015	*	0.375	0.7	6.2	PM*:6.6x(z6), EB*:1.7x(z1)	*:12, PM*:8, EB*:5

Atom	N_{match}	Hit rate	Dominant	Frac.	Enrich.	$z_{\max}$	Over-represented	Types
153	13	0.014	EB*	0.615	6.6	6.2	EB*:6.6x(z6)	EB*:8, *:2, Em*:1
131	11	0.026	LP?	0.545	7.8	5.9	LP?:7.8x(z6)	LP?:6, EB*:3, PN:1
127	17	0.016	EB*	0.529	5.7	5.9	EB*:5.7x(z6)	EB*:9, RR*:3, WD*:2
76	9	0.011	EB*	0.667	7.1	5.6	EB*:7.1x(z6)	EB*:6, RR*:2, *:1
185	56	0.025	*	0.464	0.9	5.6	EB*:3.4x(z6)	*:26, EB*:18, RG*:4
146	26	0.015	*	0.462	0.8	5.5	EB*:4.5x(z6)	*:12, EB*:11, RR*:2
121	27	0.011	EB*	0.407	4.4	5.3	EB*:4.4x(z5)	EB*:11, *:9, WD*:3
167	16	0.008	EB*	0.500	5.4	5.3	EB*:5.4x(z5)	EB*:8, *:7, RG*:1
212	13	0.010	EB*	0.538	5.8	5.3	EB*:5.8x(z5)	EB*:7, RR*:4, cC*:1
134	10	0.010	EB*	0.600	6.4	5.2	EB*:6.4x(z5)	EB*:6, *:4
33	53	0.020	*	0.453	0.8	5.0	RG*:6.9x(z5), EB*:3.2x(z5)	*:24, EB*:16, RG*:5
42	25	0.015	*	0.440	0.8	5.0	EB*:4.3x(z5)	*:11, EB*:10, RR*:4
177	63	0.055	*	0.746	1.4	4.9	SB*:4.8x(z5)	*:47, SB*:8, LP*:4
34	16	0.006	*	0.438	0.8	4.5	EB*:4.7x(z5)	*:7, EB*:7, WR*:1
176	38	0.026	*	0.421	0.8	4.5	EB*:3.4x(z4)	*:16, EB*:12, Y*?:3
196	28	0.011	*	0.429	0.8	4.0	EB*:3.4x(z4)	*:12, EB*:9, RR*:3
112	205	0.230	*	0.863	1.6	3.9	Pe*:4.9x(z4)	*:177, EB*:9, Pe*:5
159	289	0.138	*	0.872	1.6	3.9	HB*:4.8x(z4), RG*:2.3x(z3), SB*:1.8x(z2)	*:252, SB*:14, RG*:9
228	85	0.051	*	0.824	1.5	3.9	SB*:3.6x(z4)	*:70, SB*:8, Em*:2
154	167	0.101	*	0.856	1.6	3.9	V*:4.8x(z4), SB*:1.4x(z1)	*:143, EB*:9, SB*:6
221	230	0.112	*	0.813	1.5	3.8	PM*:2.3x(z4), SB*:2.2x(z3), Em*:1.2x(z0)	*:187, PM*:20, SB*:13
164	68	0.026	*	0.471	0.9	3.8	EB*:2.5x(z4)	*:32, EB*:16, AGN:4
198	27	0.010	*	0.519	1.0	3.5	EB*:3.2x(z3)	*:14, EB*:8, RR*:2
120	13	0.010	EB*	0.385	4.1	3.4	EB*:4.1x(z3)	EB*:5, *:5, WD*:2
124	13	0.009	EB*	0.385	4.1	3.4	EB*:4.1x(z3)	EB*:5, *:5, Mi*:1
245	13	0.021	EB*	0.385	4.1	3.4	EB*:4.1x(z3)	EB*:5, *:3, Sy1:1
109	135	0.067	*	0.756	1.4	3.4	SB*:2.8x(z3), EB*:1.3x(z1)	*:102, EB*:17, SB*:10
168	18	0.008	EB*	0.333	3.6	3.3	EB*:3.6x(z3)	EB*:6, *:5, WD*:4
5	24	0.014	*	0.500	0.9	3.2	EB*:3.1x(z3)	*:12, EB*:7, WD*:2
49	15	0.006	*	0.600	1.1	3.0	EB*:3.6x(z3)	*:9, EB*:5, PN:1
209	31	0.013	*	0.484	0.9	3.0	EB*:2.8x(z3)	*:15, EB*:8, AGN:4
72	22	0.008	*	0.364	0.7	2.8	EB*:2.9x(z3)	*:8, EB*:6, AGN:4
118	17	0.008	*	0.588	1.1	2.7	EB*:3.2x(z3)	*:10, EB*:5, pA*:1
236	17	0.009	*	0.529	1.0	2.7	EB*:3.2x(z3)	*:9, EB*:5, RG*:1
11	452	0.199	*	0.931	1.7	2.6	HB*:3.1x(z3), SB*:1.5x(z2)	*:421, SB*:18, HB*:5
17	29	0.026	*	0.724	1.3	2.6	EB*:2.6x(z3)	*:21, EB*:7, V*:1
63	24	0.022	*	0.333	0.6	2.5	EB*:2.7x(z3)	*:8, EB*:6, PM*:3
223	39	0.015	*	0.718	1.3	2.3	EB*:2.2x(z2)	*:28, EB*:8, SB*:1
101	95	0.080	*	0.821	1.5	1.7	EB*:1.6x(z2)	*:78, EB*:14, V*:2

Atom	N_{match}	Hit rate	Dominant	Frac.	Enrich.	z_{max}	Over-represented	Types
140	40	0.032	*	0.550	1.0	1.7	EB*:1.9x(z2)	*:22, EB*:7, **:2
172	734	0.387	*	0.931	1.7	1.5	SB*:1.3x(z2)	*:683, SB*:26, RG*:8
99	80	0.067	*	0.838	1.5	1.3	EB*:1.5x(z1)	*:67, EB*:11, PN:1
242	47	0.048	*	0.574	1.1	1.2	EB*:1.6x(z1)	*:27, EB*:7, PM*:3
1	10	0.004	*	0.900	1.7			*:9, EB*:1
3	7	0.005	*	0.714	1.3			*:5, WD*:1, AGN:1
10	10	0.005	*	0.800	1.5			*:8, EB*:2
14	10	0.007	*	0.400	0.7			*:4, EB*:3, RR*:1
20	7	0.004	*	0.429	0.8			*:3, EB*:2, WD*:2
21	3	0.005	*	0.667	1.2			*:2, LP*:1
23	18	0.022	G	0.278	484.6			G:5, EB*:3, PN:2
27	5	0.002	*	0.600	1.1			*:3, WD*:1, EB*:1
37	3	0.002	EB*	0.667	7.1			EB*:2, *:1
44	2	0.003	RG*	0.500	36.8			RG*:1, HB*:1
45	13	0.005	*	0.692	1.3			*:9, EB*:2, QSO:1
58	13	0.015	*	0.846	1.6			*:11, SB*:1, Em*:1
59	4	0.004	*	0.500	0.9			*:2, EB*:1, LP?:1
61	39	0.033	*	0.897	1.6			*:35, EB*:2, PN:1
66	5	0.004	*	0.800	1.5			*:4, RR*:1
68	12	0.009	*	0.500	0.9			*:6, EB*:3, RR*:3
69	11	0.015	*	0.545	1.0			*:6, EB*:4, PN:1
77	16	0.027	LP?	0.250	3.6			LP?:4, RR*:3, GiG:2
79	11	0.004	EB*	0.273	2.9			EB*:3, *:3, AGN:2
90	11	0.014	EB*	0.364	3.9			EB*:4, *:3, LP?:2
91	14	0.005	*	0.643	1.2			*:9, EB*:4, QSO:1
95	24	0.010	*	0.833	1.5			*:20, EB*:2, WD*:1
96	7	0.005	EB*	0.429	4.6			EB*:3, *:2, SB*:1
97	32	0.019	*	0.906	1.7			*:29, Em*:2, SB*:1
102	18	0.008	*	0.778	1.4			*:14, EB*:4
105	7	0.017	Em*	0.286	11.1			Em*:2, LP?:2, Mi*:1
113	9	0.006	*	0.556	1.0			*:5, EB*:2, WR*:1
117	7	0.009	*	0.429	0.8			*:3, EB*:2, LP?:1
132	21	0.025	*	0.762	1.4			*:16, SB*:4, AB?:1
137	8	0.013	*	0.625	1.1			*:5, EB*:2, LP?:1
138	2	0.004	*	1.000	1.8			*:2
139	8	0.003	*	0.625	1.1			*:5, EB*:3
141	13	0.025	*	0.538	1.0			*:7, EB*:4, LP?:2
147	10	0.013	*	0.900	1.7			*:9, EB*:1
169	16	0.006	*	0.750	1.4			*:12, EB*:3, RR*:1

Atom	N_{match}	Hit rate	Dominant	Frac.	Enrich.	$z_{\max}$	Over-represented	Types
170	7	0.009	WD*	0.286	23.3			WD*:2, RR*:2, EB*:1
181	20	0.019	*	0.750	1.4			*:15, PM*:2, QSO:1
182	1	0.001	LP?	1.000	14.2			LP?:1
194	4	0.005	LP?	0.500	7.1			LP?:2, Em*:1, *:1
195	3	0.003	El*	0.333	204.7			El*:1, EB*:1, RR*:1
200	6	0.006	EB*	0.500	5.4			EB*:3, WD*:1, LP?:1
226	1	0.002	EB*	1.000	10.7			EB*:1
230	82	0.045	*	0.890	1.6			*:73, EB*:4, PM*:3
231	12	0.005	*	0.500	0.9			*:6, cC*:3, EB*:3
234	2	0.003	*	1.000	1.8			*:2
237	6	0.004	*	1.000	1.8			*:6
240	77	0.048	*	0.883	1.6			*:68, EB*:5, Pe*:2
250	6	0.007	Em*	0.333	13.0			Em*:2, EB*:2, *:2
253	14	0.016	*	0.643	1.2			*:9, El*:1, HB*:1

Table 24: Atoms in D_H holding a majority of a single SIMBAD class, ranked by specialization ζ.

Atom	Class	N_{atom}	N_{total}	Coverage	Atom frac.	Spec.
229	LP?	4365	5393	0.809	0.430	0.348
229	LP*	3833	6801	0.564	0.378	0.213
107	LP*	3982	6801	0.586	0.359	0.210
107	LP?	3220	5393	0.597	0.290	0.173
36	Mi*	406	643	0.631	0.223	0.141
54	WD*	541	943	0.574	0.219	0.126
142	Em*	1165	1973	0.590	0.192	0.113
93	Mi*	478	643	0.743	0.150	0.112
40	WD*	875	943	0.928	0.106	0.099
98	Mi*	371	643	0.577	0.136	0.079
89	Mi*	576	643	0.896	0.082	0.073
70	WD*	486	943	0.515	0.135	0.070
71	Mi*	418	643	0.650	0.092	0.060
15	WD*	652	943	0.691	0.084	0.058
229	Mi*	604	643	0.939	0.060	0.056
88	RR*	637	1111	0.573	0.093	0.053
144	C*	191	197	0.970	0.054	0.052
40	AGN	439	462	0.950	0.053	0.051
104	RR*	665	1111	0.599	0.080	0.048
65	RR*	708	1111	0.637	0.071	0.045
88	AGN	378	462	0.818	0.055	0.045
88	WD*	488	943	0.517	0.071	0.037
70	HS?	140	151	0.927	0.039	0.036
227	C*	126	197	0.640	0.053	0.034
93	C*	144	197	0.731	0.045	0.033
70	C*	131	197	0.665	0.036	0.024
0	AB?	148	290	0.510	0.047	0.024
88	HS?	147	151	0.974	0.021	0.021
40	HS?	151	151	1.000	0.018	0.018
65	AGN	273	462	0.591	0.027	0.016
40	QSO	142	152	0.934	0.017	0.016
65	HS?	151	151	1.000	0.015	0.015
112	Pu*	201	253	0.794	0.018	0.015
16	BY*	123	217	0.567	0.024	0.014
88	QSO	117	152	0.770	0.017	0.013
65	Pu*	176	253	0.696	0.018	0.012
40	Pu*	156	253	0.617	0.019	0.012
166	LM*	93	185	0.503	0.020	0.010
89	C*	114	197	0.579	0.016	0.009
229	C*	117	197	0.594	0.012	0.007
0	s*r	36	63	0.571	0.012	0.007

Atom	Class	N_{atom}	N_{total}	Coverage	Atom frac.	Spec.
30	Pu*	135	253	0.534	0.010	0.005
112	gD*	61	63	0.968	0.006	0.005
107	LM*	103	185	0.557	0.009	0.005
70	s*r	33	63	0.524	0.009	0.005
15	gD*	47	63	0.746	0.006	0.005
65	gD*	53	63	0.841	0.005	0.004
101	gD*	54	63	0.857	0.005	0.004
65	QSO	76	152	0.500	0.008	0.004
30	gD*	53	63	0.841	0.004	0.003
180	gD*	45	63	0.714	0.004	0.003
30	**	67	126	0.532	0.005	0.003
240	gD*	44	63	0.698	0.004	0.003
104	gD*	36	63	0.571	0.004	0.002
40	gD*	35	63	0.556	0.004	0.002
112	dS*	43	81	0.531	0.004	0.002

C Feature Importances (Label Prediction)

Atom	Importance	Std.
1	0.332151	0.007198
17	0.176136	0.003224
6	0.172336	0.003407
13	0.152408	0.004180
0	0.140128	0.004309
12	0.137989	0.004401
5	0.112480	0.003810
20	0.105708	0.001922
18	0.080809	0.001877
23	0.051297	0.000866
3	0.039189	0.001271
21	0.038001	0.000907
8	0.037757	0.001037
14	0.032926	0.000889
15	0.030868	0.001361
19	0.028337	0.000695
10	0.025074	0.001239
11	0.021089	0.001043
4	0.013651	0.000407
9	0.011718	0.000778

Table 25: Top 20 D_C atoms for $[M/H]$ using HGB permutation importance.

Coefficient	Importance	Std.
`bp_08`	0.117630	0.002750
`bp_01`	0.089495	0.002242
`bp_05`	0.085685	0.002085
`bp_16`	0.083862	0.002717
`rp_04`	0.081180	0.001312
`rp_02`	0.072944	0.000970
`bp_02`	0.066996	0.002051
`bp_04`	0.065952	0.001823
`bp_27`	0.043150	0.001386
`bp_12`	0.041054	0.001146
`rp_11`	0.037130	0.000907
`bp_03`	0.035122	0.000702
`rp_14`	0.035083	0.001314
`rp_03`	0.034259	0.000610
`bp_18`	0.031766	0.000572
`rp_06`	0.024380	0.000447
`rp_09`	0.019955	0.000990
`rp_05`	0.019824	0.000596
`bp_20`	0.014935	0.000419
`rp_10`	0.014403	0.000543

Table 26: Top 20 coefficients for $[M/H]$ using HGB permutation importance.

Atom	Importance	Std.
104	0.111152	0.002944
35	0.039710	0.001472
214	0.039004	0.002307
7	0.037484	0.001967
173	0.029387	0.001149
230	0.028212	0.001267
144	0.025324	0.000428
152	0.023876	0.000684
172	0.020402	0.000985
44	0.018776	0.000897
227	0.018260	0.000852
97	0.017518	0.000888
233	0.016546	0.000743
166	0.016487	0.000503
108	0.015366	0.000503
192	0.014428	0.000622
228	0.013738	0.000668
109	0.013237	0.000577
67	0.012942	0.000880
185	0.012549	0.000569

Table 27: Top 20 D_H atoms for $[M/H]$ using HGB permutation importance.

Atom	Importance	Std.
1	0.223693	0.004790
14	0.146207	0.003502
22	0.144071	0.002328
10	0.125408	0.002479
8	0.093381	0.001850
23	0.086287	0.001298
12	0.085514	0.001899
5	0.070540	0.001295
0	0.051261	0.000994
17	0.041110	0.000511
6	0.030197	0.000488
20	0.022723	0.000504
21	0.015730	0.000404
19	0.012325	0.000311
4	0.011575	0.000295
11	0.010715	0.000176
9	0.007861	0.000202
2	0.007588	0.000271
7	0.007014	0.000183
13	0.006869	0.000256

Table 28: Top 20 D_C atoms for T_{eff} using HGB permutation importance.

Coefficient	Importance	Std.
`bp_01`	0.710030	0.008942
`bp_00`	0.152244	0.001405
`bp_05`	0.100661	0.000870
`bp_02`	0.046777	0.000756
`bp_08`	0.037430	0.000607
`bp_04`	0.024891	0.000256
`rp_02`	0.021367	0.000384
`rp_06`	0.020208	0.000392
`bp_06`	0.016848	0.000699
`bp_03`	0.013145	0.000329
`rp_05`	0.010794	0.000239
`rp_07`	0.010470	0.000294
`rp_11`	0.005183	0.000231
`rp_01`	0.004887	0.000092
`rp_04`	0.004461	0.000275
`rp_03`	0.003718	0.000087
`rp_08`	0.003640	0.000190
`bp_16`	0.002818	0.000133
`bp_07`	0.002418	0.000146
`rp_14`	0.002304	0.000149

Table 29: Top 20 coefficients for T_{eff} using HGB permutation importance.

Atom	Importance	Std.
142	0.186289	0.003207
107	0.059977	0.002580
56	0.059158	0.001648
201	0.051725	0.001434
112	0.043187	0.001433
17	0.036019	0.001355
240	0.025381	0.001120
2	0.025162	0.000893
99	0.016838	0.001009
192	0.014083	0.000446
109	0.013401	0.000481
154	0.013400	0.000393
31	0.012613	0.000583
27	0.010931	0.000690
176	0.010372	0.000483
187	0.009531	0.000412
19	0.008728	0.000369
193	0.008115	0.000206
122	0.007717	0.000303
92	0.007224	0.000301

Table 30: Top 20 D_H atoms for T_{eff} using HGB permutation importance.

D Code Availability

The code used in this thesis is available at: https://github.com/dean-UL/Bachelor_Thesis.

Atom	Importance	Std.
7	0.469790	0.003680
0	0.371673	0.004854
20	0.244978	0.005334
13	0.060734	0.002268
22	0.055181	0.001896
17	0.041433	0.000997
4	0.041029	0.000484
14	0.040866	0.000887
8	0.024996	0.000651
10	0.024400	0.001019
23	0.022506	0.000777
5	0.020063	0.000859
3	0.018952	0.000881
12	0.015627	0.000414
19	0.015503	0.000779
6	0.013436	0.000571
15	0.010997	0.000631
1	0.009044	0.000554
9	0.007789	0.000341
2	0.005360	0.000660

Table 31: Top 20 D_C atoms for $\log g$ using HGB permutation importance.

Coefficient	Importance	Std.
`rp_03`	0.691009	0.006383
`bp_00`	0.119871	0.002072
`bp_03`	0.106587	0.002177
`bp_01`	0.078336	0.001636
`rp_04`	0.077928	0.001418
`bp_04`	0.066999	0.001041
`bp_05`	0.066204	0.001782
`bp_06`	0.028254	0.001250
`rp_01`	0.024926	0.000599
`rp_02`	0.022762	0.000842
`bp_02`	0.017444	0.000724
`bp_08`	0.010422	0.000466
`bp_16`	0.007999	0.000286
`bp_21`	0.005549	0.000320
`rp_05`	0.005543	0.000299
`rp_07`	0.003976	0.000197
`bp_27`	0.003677	0.000164
`rp_09`	0.003571	0.000224
`rp_06`	0.003076	0.000239
`rp_12`	0.002238	0.000119

Table 32: Top 20 coefficients for $\log g$ using HGB permutation importance.

Atom	Importance	Std.
7	0.097881	0.001755
177	0.070644	0.001642
0	0.069884	0.002252
228	0.063455	0.001540
58	0.061654	0.001782
44	0.059677	0.001332
142	0.058259	0.001466
229	0.053383	0.001661
172	0.048548	0.000850
156	0.047269	0.001471
132	0.046853	0.002103
234	0.035562	0.001611
226	0.035335	0.001452
152	0.022495	0.000407
182	0.022386	0.000726
31	0.021377	0.000771
144	0.019372	0.001042
159	0.018881	0.000586
107	0.017426	0.000727
181	0.017379	0.001178

Table 33: Top 20 D_H atoms for $\log g$ using HGB permutation importance.