RAS Techniques and Instruments

RASTI **4,** 1–10 (2025) https://doi.org/10.1093/rasti/rzaf044
Advance Access publication 2025 October 3

Finding rare classes in large data sets: the case of polluted white dwarfs from *Gaia* XP spectra

Xander Byrne [1]★ Amy Bonsor [1] Laura K. Rogers [1,2] and Mariona Badenas-Agusti [1]

[1] *Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*
[2] *NOIRLab, 950 N Cherry Ave, Tucson, AZ 85719, USA*

Accepted 2025 September 28. Received 2025 September 15; in original form 2025 June 18

ABSTRACT

The *Gaia* mission's third data release recorded low-resolution spectra for about 100 000 white dwarf candidates. A small subset of these spectra show evidence of characteristic broad Ca II absorption features, implying the accretion of rocky material by the so-called polluted white dwarfs–important probes of the composition of exoplanetary material. Several supervised and unsupervised data-intensive methods have recently been applied to identify polluted white dwarfs from the *Gaia* spectra. We present a comparison of these methods, along with the first application of *t*-distributed stochastic neighbour embedding (*t*SNE) to this data set. We find that *t*SNE outperforms the similar technique Uniform Manifold Approximation and Projection, isolating over 50 per cent more high-confidence polluted candidates, including 39 new candidates which are not selected by any other method investigated and which have not been observed at higher resolution. Supervised methods benefit greatly from data labels provided by earlier works, selecting many known polluted white dwarfs which are missed by unsupervised methods. Our work provides a useful case study in the selection of members of rare classes from a large, sporadically labelled data set, with applications across astronomy.

Key words: Data Methods – White Dwarfs – Machine Learning – Extrasolar Planets.

1 INTRODUCTION

White dwarfs (WDs) are remnant stellar cores, the final form of all main-sequence (MS) stars with zero-age masses below 9–12 $M_\odot$ (L. G. Althaus et al. 2010, 2021; Lauffer, A. D. Romero & S. O. Kepler 2018). WDs constitute a highly uniform population, and are therefore ideal astrophysical laboratories with which to probe a wide range of astrophysical phenomena, including the bulk composition of exoplanetary material.

The presence of metal 'pollution' in the spectra of WDs reveals the composition of rocky bodies in the system. Following a star's post-MS evolution, exoplanetary bodies can be scattered onto star-grazing orbits (J. H. Debes & S. Sigurdsson 2002; A. Bonsor, A. J. Mustill & M. C. Wyatt 2011; S. F. N. Frewen & B. M. S. Hansen 2014; A. J. Mustill et al. 2018; R. F. Maldonado et al. 2020a, b), tidally disrupted (M. Jura 2003; D. Veras et al. 2014), and then accreted onto the WD (M. G. Brouwers, A. Bonsor & U. Malamud 2022). This has been shown to be the dominant mechanism behind the appearance of metal features in the spectra of so-called polluted WDs (e.g. J. Farihi et al. 2010; D. Veras 2016). By fitting WD atmosphere models to the polluted WD spectra (e.g. D. Koester et al. 2005, 2011; P. Dufour et al. 2007; M. A. Hollands et al. 2017; M. A. Hollands, B. T. Gänsicke & D. Koester 2018; S. Blouin et al. 2018; M. Badenas-Agusti et al. 2024), one can post-mortem constrain the bulk composition, and thus geology, of the accreted parent body (e.g. S. Xu & A. Bonsor 2021).

The *Gaia* mission (Gaia Collaboration 2016) has revolutionized the study of WDs. The mission's second data release (DR2) identified over 262 000 high-confidence WD candidates, an increase by almost an order of magnitude on the then-state-of-the-art (Gaia Collaboration 2018; N. P. Gentile Fusillo et al. 2019). This increased further to over 359 000 thanks to the improved astrometry and photometry of *Gaia*'s Early Data Release 3 (EDR3; N. P. Gentile Fusillo et al. 2021; Gaia Collaboration 2023). In addition to astrometric and photometric measurements, the full *Gaia* DR3 also provides low-resolution ($R \sim 70$) spectra for 219 million sources (J. M. Carrasco et al. 2021; Gaia Collaboration 2023), including $\approx 100\,000$ WD candidates (N. P. Gentile Fusillo et al. 2021). These spectra – called 'BP/RP spectra'; collectively 'XP spectra' – are stored not as fluxes in a sequence of wavelength bins, but as a pair of 55-coefficient Hermite polynomials. They can therefore be represented by a 110D coefficient vector $\mathbf{x} \in \mathbb{R}^{110}$. The Hermite polynomials approximating the spectrum can be reconstructed using the `GaiaXPy` Python package.[1]

While an XP spectrum is not sufficient to classify a WD as polluted or not (let alone fit elemental abundances) they indicate which WDs from this large data set to observe with targeted spectroscopic campaigns at higher resolution. The high surface gravities of WDs ($\sim 10^8$ g cm^{-2}) cause significant collisional broadening, allowing metal pollution to be detected in even low-resolution spectra, usually from the strong Ca H and K lines at 3933 and 3968 Å.

★ E-mail: xbyrne@ast.cam.ac.uk

[1] https://gaia-dpci.github.io/GaiaXPy-website/

Several techniques have recently been applied to the large *Gaia* XP sample to identify polluted WD candidates:

(i) **Uniform manifold approximation and projection (UMAP; McInnes, J. Healy & J. Melville 2018)**. UMAP is an example of a dimensionality reduction technique, wherein a data set of high-dimensional points is projected into a 2D map, while approximately preserving the original structure of the data set. In the case of UMAP, this is achieved by finding a 2D submanifold of the high-dimensional data space that is as close as possible to the data points. M. L. Kao et al. (2024) apply UMAP to a sample of 96 134 XP spectra, identifying a distinct group of 465 WDs of which 90 were known *a priori* to be polluted. The remaining 375 are subject to an ongoing high-resolution spectroscopic campaign; they report that 99 per cent of those observed at the time of publication showed multiple metal lines.

(ii) **Self-organizing maps (SOMs; T. Kohonen 1982)**. SOMs employ neural networks to fit a flexible grid of neurons to the data, 'assigning' each data point to a particular neuron together with similar data points. This partitions the data set, such that within each subset the data are similar to each other. X. Pérez-Couto et al. (2024) use a sequence of two SOMs: one to remove contaminants such as quasars and galaxies, and another to partition the data set into spectral classes. They find two populations of 249 and 218 polluted WD candidates, the latter of which was not identified by UMAP.

(iii) **Gradient tree boosting (J. H. Friedman 2001)**. Gradient tree boosting trains a sequence of decision trees to regress the error on the prediction from the tree before it. O. Vincent et al. (2024) apply gradient tree boosting to classify 101 783 WD candidates into one of six primary WD spectral types: DA (characterized by hydrogen features in the spectrum), DB (neutral helium features), DC (no features), DO (ionized helium), DQ (carbon), and DZ (metals).[2] They classify 1272 WDs as DZs.

(iv) **Random forests (L. Breiman 2001)**. A random forest consists of a large ensemble of decision trees, which are trained to classify data. E. M. García-Zamora et al. (2025) apply this method to classify 78 920 *Gaia* WDs within 500 pc, to (i) differentiate between DA and non-DA WDs; (ii) classify non-DA WDs into DB, DC, DQ, and DZ WDs. They classify 785 WDs as DZ.

The first two of these methods are *unsupervised*; the latter two are *supervised*. Supervised methods rely on a labelled training set, which in this case means WDs in the sample that have a known spectral classification, perhaps from existing higher resolution observations. Although data labels constitute highly relevant information, any biases or imbalances present in the training data can propagate through to the resulting model's predictions. Unsupervised methods do not require this ground-truth information, simply processing all of the data at face value. They therefore evade label bias while enabling the serendipitous discovery of anomalies (e.g. D. Giles & L. Walkowicz 2019; S. Webb et al. 2020).

t-distributed Stochastic Neighbour Embedding (*t*SNE; L. Van der Maaten & G. Hinton 2008) is another example of an unsupervised method. It is a dimensionality reduction technique, similar to UMAP, which attempts to project the data set into 2D by preserving as well as possible the 'similarity' between pairs of data points (see Section 2.1). A key empirical difference between *t*SNE and UMAP is that UMAP prioritizes the global structure of a data set more than *t*SNE (L. McInnes et al. 2018; S. Fotopoulou 2024). The embeddings generated by *t*SNE may therefore better reveal the similarities between similar data points, at the cost of presenting a less realistic picture of the data set at large.

C. L. Steinhardt et al. (2020) use *t*SNE to identify quiescent galaxies from UltraVISTA photometry (H. J. McCracken et al. 2012). K. Hawkins et al. (2021) likewise use *t*SNE to identify 416 metal-poor stars from a sample of 14 000 stars in the Hobby-Eberly Telescope Dark Energy Experiment (HETDEX; K. Gebhardt et al. 2021) survey. More recently, X. Byrne et al. (2024) demonstrate the use of *t*SNE to explore the intermediate-resolution Dark Energy Spectroscopic Instrument's Early Data Release (DESI EDR; DESI Collaboration 2023). Among their findings, cataclysmic variables (CVs) are identified from the catalogue at human-level recall in seconds.

This work compares the ability of various techniques, including *t*SNE, to identify polluted WDs from *Gaia* XP spectra. The isolation of rare classes from a large data set with sparse labels is a common problem in Astronomy, so our findings are expected to apply more broadly. Section 2 outlines details on *t*SNE and characterizes the data set. Section 3 demonstrates *t*SNE's ability to select polluted WDs, as well as other patterns that appear in the resulting embedding. Section 4 compares *t*SNE's capability in this task to that of other methods, and suggests use cases for the various techniques in attaining other insights from large data sets. Section 5 concludes our work.

2 METHODS

2.1 *t*SNE

*t*SNE is a technique for mapping a data set of high-dimensional points $\mathbf{x} \in \mathbb{R}^D$ into lower dimensional points $\mathbf{y} \in \mathbb{R}^d$, while preserving local structure probabilistically. The lower dimensionality $d < D$ is usually either 2 or 3; we use $d = 2$ throughout. This dimensionality reduction is achieved by defining two *similarity* functions – one for the original D-dimensional space, another for d dimensions – and then minimizing the difference between the two similarity distributions with respect to the projected (d-dimensional) points.

The similarity between two points $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^D$ is defined by

$$p_{ij} = \frac{1}{2N}\left(p_{i|j} + p_{j|i}\right), \tag{1}$$

where

$$p_{i|j} = \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|/2\sigma^2\right)}{\sum_{k\neq j}\exp\left(-\|\mathbf{x}_k - \mathbf{x}_j\|/2\sigma^2\right)}. \tag{2}$$

The parameter σ is internally optimized such that this similarity distribution achieves a particular user-defined value of the *perplexity* $\mathrm{Perp}(p_i)$:

$$\log_2 \mathrm{Perp}(p_i) \equiv -\sum_j p_{j|i}\log_2 p_{j|i}; \tag{3}$$

the right-hand side being recognizable as the Shannon entropy of the similarity distribution in bits (L. Van der Maaten & G. Hinton 2008). In 2D, the similarity is instead defined according to a Cauchy distribution:

$$q_{ij} = \frac{\left(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2\right)^{-1}}{\sum_{k\neq j}\left(1 + \|\mathbf{y}_k - \mathbf{y}_j\|^2\right)^{-1}}. \tag{4}$$

The use of a Cauchy distribution – which has broader tails than a normal distribution – addresses the 'crowding problem', a phenomenon permitting high-dimensional points to have many more close neighbours than 2D points (L. Van der Maaten & G. Hinton 2008).

[2] This classification scheme was first outlined in E. M. Sion et al. (1983); the 'D' stands for 'degenerate'.

Ideally, the pairwise distances $\|\mathbf{y}_i - \mathbf{y}_j\|$ in the embedding should be as close as possible to the corresponding $\|\mathbf{x}_i - \mathbf{x}_j\|$ in the original data space. *t*SNE therefore proceeds by probabilistically minimizing the difference between the two similarity distributions, as quantified by the Kullback–Leibler divergence (S. Kullback & R. A. Leibler 1951):

$$\mathcal{KL}(p||q) \equiv \sum_{i,j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right). \quad (5)$$

The embedding is constructed by minimizing this divergence with respect to the 2D points $\{\mathbf{y}_i\}$, for instance by gradient descent.

For our purposes, each vector $\mathbf{x}_i$ is a 110D vector, each of whose components is an XP spectral coefficient. To account for sources being of different apparent magnitude – a parameter irrelevant to their classification – these coefficients are first divided by the *G*-band flux; we found that this normalization enabled easier isolation of polluted WD candidates in the embedding, compared to L2 or unit-Gaussian normalization. Our results are not sensitive to perplexity values between about 30 and 120; we henceforth use a perplexity of 50. We use the implementation of *t*SNE provided in the SCIKIT-LEARN Python package (F. Pedregosa et al. 2011) throughout, and use their default values for hyperparameters other than the perplexity.

2.2 Sample selection

The sample of 359 073 high-confidence WD candidates created by N. P. Gentile Fusillo et al. (2021) was selected based on *Gaia* EDR3 photometry, as well as a series of quality filters to exclude sources with e.g. unreliable astrometry. Of these, 107 797 sources had XP spectra released in DR3, as obtained from the Gaia@AIP service.[3]

Some of these sources inevitably contain low-signal-to-noise data, and further selection criteria are needed. There seems to be little consensus as to how best to achieve this: previous works have performed various combinations of cuts on the distance, astrometric noise, number of *Gaia* observations, or the *probability of being a white dwarf* ($P_{\rm WD}$; N. P. Gentile Fusillo et al. 2015). Comparing these criteria, we suggest that the criteria of X. Pérez-Couto et al. (2024) are the most inclusive and well-justified. These criteria are:

(i) (`phot_bp_n_obs` > 10) and (`phot_rp_n_obs` > 15), as recommended by R. Andrae et al. (2023);
(ii) `visibility_periods_used` > 10, as recommended by L. Lindegren et al. (2018).

We note that X. Pérez-Couto et al. (2024) also list a cut on the `phot_bp_rp_excess_factor_corrected`, however this same cut is already made in assembling the supersample of N. P. Gentile Fusillo et al. (2021, their equation 21), and is therefore not necessary. Following these cuts, we obtain a sample of size 107 164.

2.3 Evaluation of polluted WD selection

Each method for identifying polluted WDs will be imperfect: some polluted WDs will inevitably escape selection, and some other objects will be erroneously selected. In evaluating the different methods, it is thus crucial to know which sources are genuinely polluted WDs.

Unambiguously identifying a polluted WD usually requires much higher resolution spectroscopy than the *Gaia* XP spectra, with an R of at least 1000 and ideally $R > 20\,000$ to detect weaker pollution lines. Such observations are however time- and resource-intensive and have therefore only been conducted for a small fraction of the data set. For most of the sources, their pollution status is unknown. For those that *have* been observed, they have either been confirmed as polluted WDs, or as WDs of different, non-polluted spectral type.

We categorize each source as either 'known polluted', 'known non-polluted', and 'unknown' according to their classifications in three data sets, each of which collates higher resolution observations. These are:

(i) Montreal White Dwarf Data base (MWDD; P. Dufour et al. 2017);
(ii) *Gaia*-SDSS spectroscopic sample (N. P. Gentile Fusillo et al. 2021);
(iii) Planetary Enriched White Dwarf Data base (PEWDD; J. T. Williams et al. 2024).

If a source is assigned a 'polluted' spectral class (DZ, DAZ etc.) in any of these data sets, we categorize it as 'known polluted'. If assigned a non-polluted spectral class in any data set (e.g. DA), it is categorized as 'known non-polluted'. Otherwise, its pollution status is categorized 'unknown'.

For WDs with weak metal lines, these features may be fundamentally undetectable in the low-resolution XP spectra. Furthermore, some WDs have only been identified as polluted in the ultraviolet ($\lambda < 300$ nm; e.g. D. Koester, B. T. Gänsicke & J. Farihi 2014; L. B. Ould Rouis et al. 2024), which is outside the XP spectral window. As such, even an idealized method would not be able to identify all 'known polluted' WDs.

When evaluating the performance of *supervised* methods, it is important to note that many 'known polluted' WDs likely appear in their training sets. Supervised methods are much more likely to correctly classify data present in their training sets, so a large proportion of 'known polluted' WDs will inevitably be identified by these methods. However, it is unlikely that supervised methods would perform equally well on unseen data: as such, a lower proportion of the 'unknown' WDs selected by these methods would be expected to be true polluted WDs, than the proportion of correctly classified *known* polluted WDs. A second reason for this is that WDs with an 'unknown' pollution status are generally fainter (being less likely to have been observed at higher resolution). Their XP spectra are thus on average lower signal-to-noise than those used in the training set, which may make the supervised methods (inevitably trained on higher signal-to-noise data) less likely to identify polluted WDs that have not been observed at higher resolution.

3 RESULTS

3.1 Identification of polluted WDs by *t*SNE

The *t*SNE embedding of the sample (Fig. 1) features a large 'N'-shaped structure, surrounded by several smaller 'islands'. Two of these islands show a high purity of known polluted WDs, suggesting that many of the 'unknown' WDs in these islands might be polluted WDs that have not yet been observed at higher resolution.

Which sources belong to these islands can be decided objectively using clustering algorithms, such as the well-known Density-Based Spatial Clustering of Applications with Noise (DBSCAN; M. Ester et al. 1996) algorithm. We apply the `scikit-learn` implementation of this algorithm to the *t*SNE embedding, with

[3] https://gaia.aip.de/

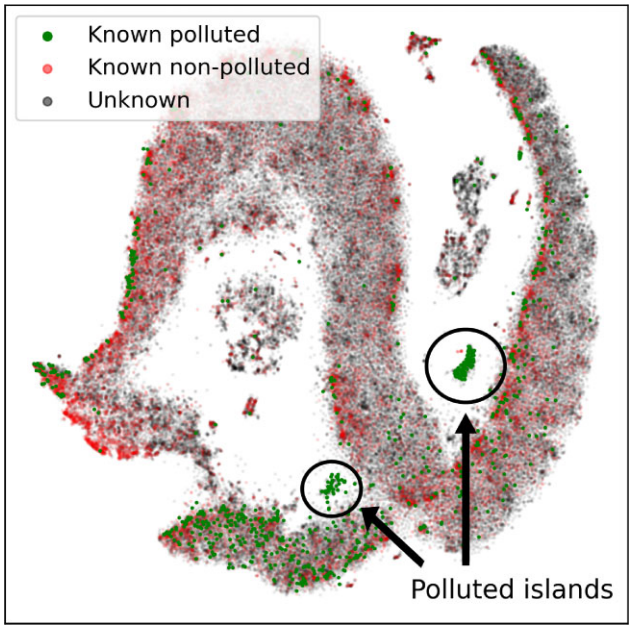

Figure 1. *t*SNE embedding of normalized *Gaia* WD candidate XP spectra. The embedding's primary feature is an 'N'-shaped sequence, with a number of islands scattered around it. Two highlighted islands contain a large number of known polluted WDs, and few known non-polluted WDs.

parameters[4] `eps` = 2 and `min_samples` = 30, identifying twelve clusters, of which two correspond to the polluted islands (Fig. 2a). We refer to these as the 'cool' and the 'warm' islands, owing to their respective temperature distributions (see Section 4.2).

The UMAP embedding constructed in M. L. Kao et al. (2024) also contained a feature populated mostly by many known polluted WDs (see their figs 2 and 3), but it is slightly blurred with other features, making membership less certain for some sources. However, the *t*SNE embedding contains sufficiently distinct clusters that algorithms such as DBSCAN can evaluate cluster membership objectively. The reasons for this difference in clustering behaviour between the two algorithms are discussed in Section 4.3.

With the members of each island definitively ascertained, we find that the 'cool' island contains 478 sources, of which 97 are known polluted WDs and 11 are known non-polluted WDs. The 'warm' island contains 229 sources, of which 50 are known polluted WDs and 2 are known non-polluted WDs. The remaining 547 'unknown' objects (across both islands) do not appear in any of the spectroscopic data bases mentioned in Section 2.3, but given that these objects have been clustered together based on similar features in their XP spectra, it is likely that a large proportion are as – yet-unidentified polluted WDs. 39 of these sources are not selected by any of the other methods compared later in Section 3.3, and are selected only by *t*SNE. In the co-added XP spectrum of these 'unknown' sources (Fig. 3), there is a clear Ca II feature, suggesting that a significant fraction of these candidates are indeed polluted WDs.

Of the 13 'known non-polluted' objects seemingly erroneously located on the two polluted islands, most may in fact be polluted after all. Eight are classified as DC or DA based on spectra whose wavelength coverage does not include the diagnostic Ca II features at 3933 and 3968 Å. Two others rely on very low-quality (S/N $\lesssim 2$) Sloan Digital Sky Survey (SDSS) spectra classified tentatively as DCs by the data-driven pipeline of O. Vincent, P. Bergeron & P. Dufour (2023). As such, ten of the 13 putative false positives might not be false positives at all, though the remaining three objects appear to be genuine DCs that have been erroneously selected here.

Running UMAP on the sample (Fig. 2b) gives a qualitatively similar embedding to that of M. L. Kao et al. (2024, cf. their fig. 2).[5] The WDs from our 'cool' island appear in a similar location to those selected in their work; indeed 435 sources are common to both. However, the WDs from our 'warm' island are deeply entrenched in the diagonal feature of the UMAP embedding, and could therefore not have been isolated using UMAP.

3.2 Other features of the *t*SNE embedding

Other than polluted WDs, the *t*SNE embedding also locates other types of source, in specific locations in the embedding. The spectral classifications in this subsection are from the *Gaia*-SDSS spectroscopic sample (N. P. Gentile Fusillo et al. 2021).

The spectra of DA WDs are dominated by Balmer features due to atmospheric hydrogen. Plotting their $BP - RP$ colours in the embedding (Fig. 4), we see that these DAs populate the 'N'-shaped sequence, in decreasing temperature order from left to right. No doubt the tilt of the blackbody continuum is reflected in the XP coefficients, and hence in the embedding. X. Byrne et al. (2024) identify a similar trend in their application of this method to DESI EDR WD spectra, in which the primary feature of their embedding is a sequence of DAs arranged in temperature order.

The spectral classes DB, DQ, and DC correspond respectively to He I features, C_2 Swan bands, and no features at all. These three classes occupy the large lower island of the embedding (Fig. 5), where there is a sequence from DBs, through DCs, to DQs. There are also large numbers of DCs and DQs along the cooler end of the 'N'-shaped DA sequence.

Some aspects of the embedding mirror the principles of WD spectral evolution. Consider a DA: it begins its life very warm, before gradually cooling over time; such a WD could be thought of as 'travelling' along the 'N'-shaped DA sequence. Once the temperature cools below about 5500 K, hydrogen is no longer excited and the Balmer features gradually fade into the featureless spectrum of a DC; indeed DCs occupy the 'cool end' of this sequence. A weaker trend is seen in the helium-dominated sequence, wherein DBs transition into DQs (if carbon-rich and low-mass) or DCs (e.g. A. Bédard 2024).

Finally, we show in Fig. 6 the locations of MS stars (erroneously included in the sample), as well as WD+MS binaries and CVs. MS stars occupy mostly the spur on the left of the embedding. WD+MS binaries are found in several locations in the embedding: an island in the top right; an island below left of centre; an island above right of centre. CVs are found primarily in two places: just below the WD+MS island in the top right; on a tight island at the very bottom of the embedding. The reasons for the identification of these spectral classes is beyond the scope of this work; suffice to say that the XP spectra of these types of object are sufficiently distinctive that they can be roughly isolated using *t*SNE.

[4]The two parameters `eps` and `min_samples` define the clustering. Briefly, each member of a cluster either has at least `min_samples` points within a distance `eps`, or is at most `eps` away from such a point.

[5]The same hyperparameters were used as in the cited work: `n_neighbours` = 25 and `min_dist` = 0.05.

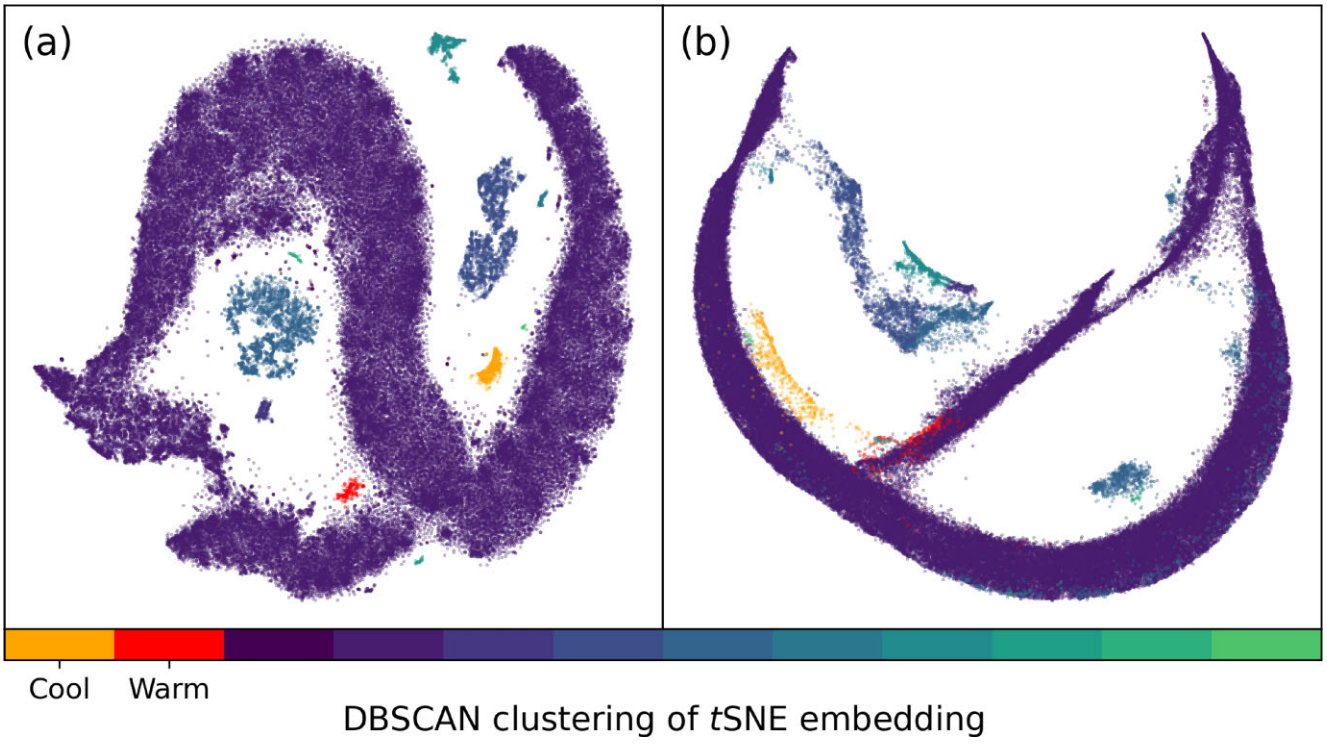

Figure 2. (a) DBSCAN clustering applied to the *t*SNE embedding. Different clusters in the embedding are colour-coded with different colours; the two polluted islands ('cool' and 'warm') are highlighted. (b) UMAP embedding of the sample, using the colour-coding from (a). The embedding is similar to that of M. L. Kao et al. (2024) (see their fig. 2), with the WDs from our 'cool' island corresponding roughly to those found in their work. However, the WDs from our 'warm' island are buried within the diagonal feature of the UMAP embedding.

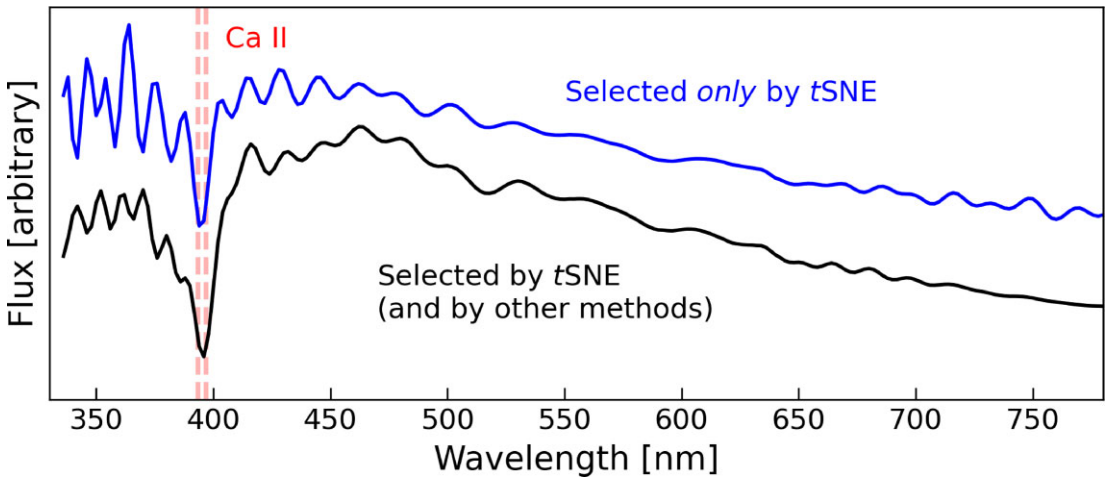

Figure 3. Co-added *Gaia* XP spectra of polluted WD candidates selected by *t*SNE that have not been observed at higher resolution ('unknown'). The upper blue line stacks the 39 candidates which are selected only by *t*SNE. The lower black line stacks all 547 'unknown' candidates selected by *t*SNE, most of which are also selected by at least one other method (see Section 3.3). The two spectra are offset arbitrarily for visualization purposes. The clear Ca II feature just below 400 nm suggests that many of these candidates are genuine polluted WDs.

3.3 Comparison of methods

Here we compare the ability of five data-driven methods, including *t*SNE, in identifying polluted WDs from the *Gaia* XP spectra.

E. M. García-Zamora et al. (2025) apply random forests, listing the classifications of 78 920 WDs within 500 pc, classifying 785 as polluted (DZ). O. Vincent et al. (2024) use gradient tree boosting to classify a larger sample of 100 886 WD candidates, classifying 1272 DZs. M. L. Kao et al. (2024) obtain 465 polluted WD candidates from a sample of 96 134 using UMAP; although this sample is not publicly available, it was acquired by personal communication. We were able to approximately reproduce the SOM-based selection of polluted WDs by X. Pérez-Couto et al. (2024) identifying 451 WDs in two neurons containing a high number of known DZs (the original work reports 467). Finally, using *t*SNE, we identify 707 polluted WD candidates across the two islands identified above (see Section 3.1).

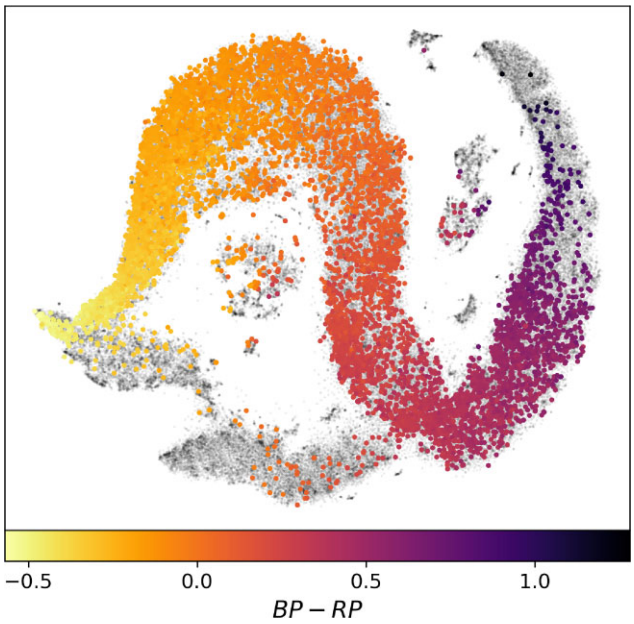

Figure 4. Embedding and $BP - RP$ colour of spectroscopically confirmed DA WDs. These WDs populate the 'N'-shaped sequence, with a temperature trend: the hottest DAs appear at the left-hand end; the coolest at the right.

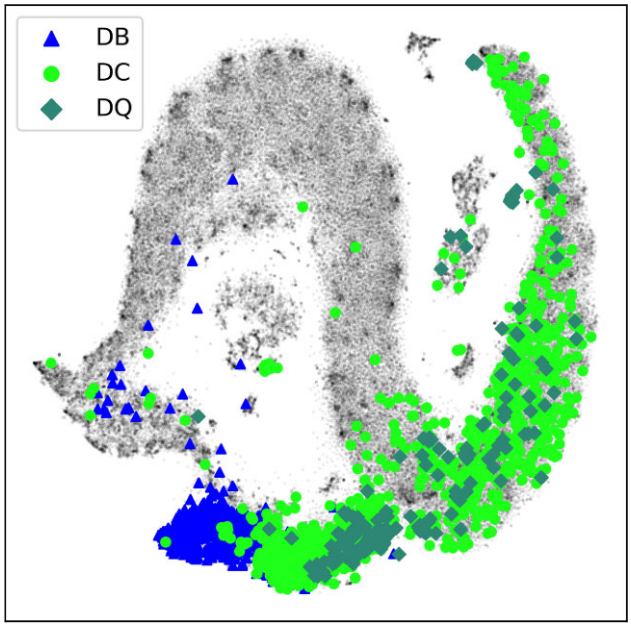

Figure 5. Locations of spectral classes DB, DC, and DQ in the *t*SNE embedding. The lower island is dominated by these spectral classes, as is the 'cooler' end of the 'N'-shaped DA sequence. A smaller number of these WDs are also scattered elsewhere in the embedding.

We list the numbers of 'known polluted', 'known non-polluted', and 'unknown' sources (see Section 2.3) in each selection in Table 1. An ideal method would identify a high proportion of known polluted WDs (effectively the true positive rate) and a low proportion of known non-polluted WDs (false positive rate), and ideally a large number of sources overall. The performance of the supervised methods is likely exaggerated, as many of the 'known (non)-polluted' WDs are present in their respective training sets (see Section 2.3).

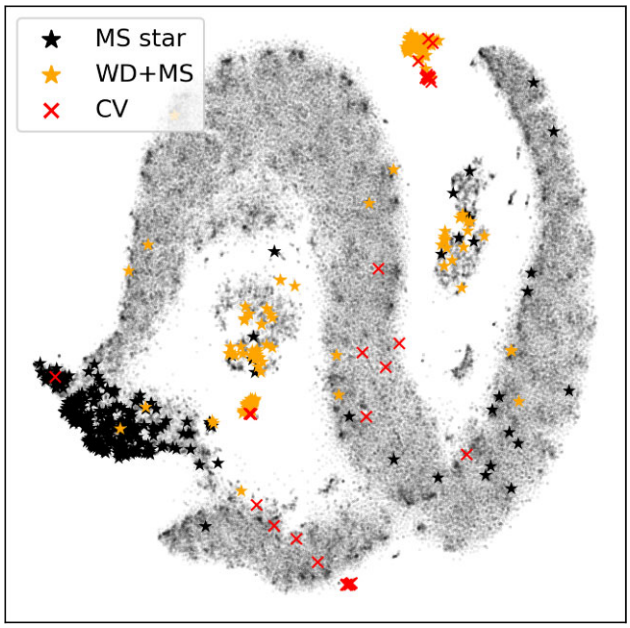

Figure 6. Locations of MS stars, WD+MS binaries, and CVs in the *t*SNE embedding. MS stars are found in a trail to the left of the embedding; WD+MS binaries are found in several regions of the embedding; CVs are mostly found in an island to the top right and bottom centre.

Random forests are by far the most successful in terms of true positive rate (53 per cent) and false positive rate (0.3 per cent). Furthermore, E. M. García-Zamora et al. (2025) quote an impressive precision (fraction of selected objects which are selected correctly) of 95 per cent for DZs. As such, a large proportion of the 364 'unknown' sources selected by random forests are expected to be genuine polluted WDs, though perhaps not 95 per cent: 'known' objects are naturally biased towards objects which are nearer, have higher quality data, and are thus more likely to be classified correctly.

Gradient tree boosting, another supervised method, outperforms the three unsupervised methods (SOMs, UMAP, *t*SNE) in identifying polluted WD candidates. On a similar sample size ($\approx$ 100 000), this method correctly recovers 350 known polluted WDs, while maintaining a low false positive rate (3.0 per cent). Again, however, many of the known polluted WDs are present in the training set used by O. Vincent et al. (2024), so a smaller proportion of the unknown sources might turn out to be genuine polluted WDs.

Among the unsupervised methods, UMAP and *t*SNE have similar true positive rates (21 per cent), though *t*SNE has a slightly better false positive rate (1.8 per cent versus 2.6 per cent). Additionally, *t*SNE proposes a larger number of polluted candidates than UMAP (707 versus 465); follow-up high-resolution spectroscopic campaigns would thus be expected to return over 50 per cent more confirmed polluted WDs from the 547 unknown candidates selected by *t*SNE than the 357 selected by UMAP. Finally, we note that SOMs show a lower true positive (11 per cent) and higher false positive rate (3.1 per cent) than any of the other methods. However, we show that it can be a useful complement to other methods in Section 3.4.

3.4 Overlap between polluted candidate selections

It might be expected that each method would 'learn' to identify metal-polluted XP spectra in similar ways, and hence that the candidates selected by the different methods would significantly overlap. However, a significant number of candidates are selected

Table 1. Pollution status of WDs classified as polluted based on their *Gaia* XP spectra, according to five methods: random forests, gradient tree boosting, SOMs, UMAP, and *t*SNE.

	RF[a]	GTB[a]	SOM[b]	UMAP	*t*SNE
Known polluted	419	350	51	96	147
	(53%)	(28%)	(11%)	(21%)	(21%)
Known non-poll.	2	38	14	12	13
	(0.3%)	(3.0%)	(3.1%)	(2.6%)	(1.8%)
Unknown	364	884	386	357	547
	(46%)	(69%)	(86%)	(77%)	(77%)
Total	785	1 272	451	465	707

Notes. [a]Supervised method: performance likely overestimated (see Section 2.3).
[b]Based on an approximate reproduction of X. Pérez-Couto et al. (2024).

only by one or two of the five methods explored here, as shown in Fig 7. Only 126 candidates are selected by all five. Even among the known polluted WDs, every method misses a significant number which were selected by at least one other method, and conversely each method selects some known polluted WDs which are not selected by any other method.

These discrepancies suggest that different methods pick up on different features in the data to identify polluted WDs. To obtain a large polluted candidate list, evidently the best strategy would be to apply all five methods to the data set and take the union of the methods' selections.

4 DISCUSSION

The previous section demonstrates the use of *t*SNE in selecting candidate polluted WDs, and compares with several other data-driven methods in achieving this goal. We begin this section by justifying the use of *Gaia* XP spectra in identifying polluted candidates. We then attempt to explain why *t*SNE isolates two islands of polluted WD candidates, rather than just one. We then compare the *t*SNE and UMAP embeddings and discuss their differing behaviour. Finally, we provide recommendations as to which of the methods compared here are best suited to different problems.

4.1 Trustworthiness of *Gaia* XP spectra in revealing pollution

Given the very low resolution of the XP spectra ($R \sim 70$), it is reasonable to question the validity of methods based on these spectra to select polluted WD candidates. The alternative is to suppose that the methods presented here merely select fortuitously aligned noise, rather than genuine features. We briefly justify here that these spectra are indeed useful in identifying polluted WDs.

The ultimate justification is empirical: the vast majority of sources identified as polluted WDs based on XP spectra do indeed turn out to be so, when observed at higher resolution. A preliminary success rate of 99 per cent is reported on spectroscopic follow-up survey of the sample identified by UMAP (M. L. Kao et al. 2024). We report a similarly high preliminary success rate for the *t*SNE-identified polluted candidates by cross-matching to DESI DR1 (DESI Collaboration 2025) and briefly visually inspecting the higher resolution spectra therein.

4.2 Why two groups of polluted WDs?

The *t*SNE embedding contains two islands populated by a large portion of known polluted WDs (Fig. 1). One might instinctively suspect that these two islands represent two distinct populations, but they turn out to correspond to a continuum, bisected in the *t*SNE embedding by the 'N'-shaped DA sequence. If *t*SNE is run on *only* the spectra on these two islands, the result is a single cluster (Fig. 8a), implying that there is no natural split between the two populations. Only when some spectra from the rest of the data set are added to the sample do the polluted islands split into two (Figs 8b–d).

This is likely a symptom of dimensionality reduction's inability to faithfully represent the structure of the 110D data set in 2D. Perhaps the polluted WDs in some sense 'bridge' over the DA sequence in the high-dimensional data space, and when projecting into 2D the optimal solution (in terms of similarity distributions; see Section 2.1) is to split the polluted WDs into two groups.

The estimated temperature distributions of the two polluted islands (Fig. 9) corroborate this interpretation. While the 'warm' sources are on average warmer than the 'cool' sources (hence our choice of group names), they do not clearly form two distinct distributions. Additionally, the temperatures of the 'warm' sources are similar to those of the DAs near to it in the embedding; likewise the 'cool' sources. *t*SNE empirically appears to have prioritized the co-location of sources of similar temperatures nearby, over the co-location of all polluted WDs, which are therefore split into two groups.

4.3 Comparing *t*SNE and UMAP

Of the two sets of polluted WDs identified by *t*SNE, only the 'cool' island corresponds to those identified using UMAP (see also Fig. 2 M. L. Kao et al. 2024). The 'warm' island is embedded by UMAP among non-polluted WDs (Fig. 2b). The two methods have similar true and false positive rates (Table 1), but *t*SNE selects an overall larger number of candidates, making it a preferable method for this use case.

We suspect that *t*SNE's advantage over UMAP in this case is its deprioritization of global structure. UMAP is known to preserve global data set structure better than *t*SNE (L. McInnes et al. 2018; S. Fotopoulou 2024), but for the purpose of selecting members of a rare class from a large data set, this is not of particular importance. Of greater importance here is the preservation of *local* structure, so that as many members as possible of the rare class can be co-identified. In prioritizing global structure, UMAP appears to sacrifice somewhat the co-identification of polluted WDs. Another manifestation of this is the 'N'-shaped DA sequence of the *t*SNE embedding, corresponding to the 'U'-shaped sequence in the UMAP embedding. As shown in Fig. 4 (as well as fig. 2 of M. L. Kao et al. 2024), these correspond to temperature sequences, with the hottest DAs at one end and the coolest at the other. In the UMAP embedding, this sequence is stretched out as much as possible, to preserve the distance in data space between DAs of very different temperatures. By contrast, in the *t*SNE embedding, this sequence is coiled up into an 'N' shape, as this method places less importance on

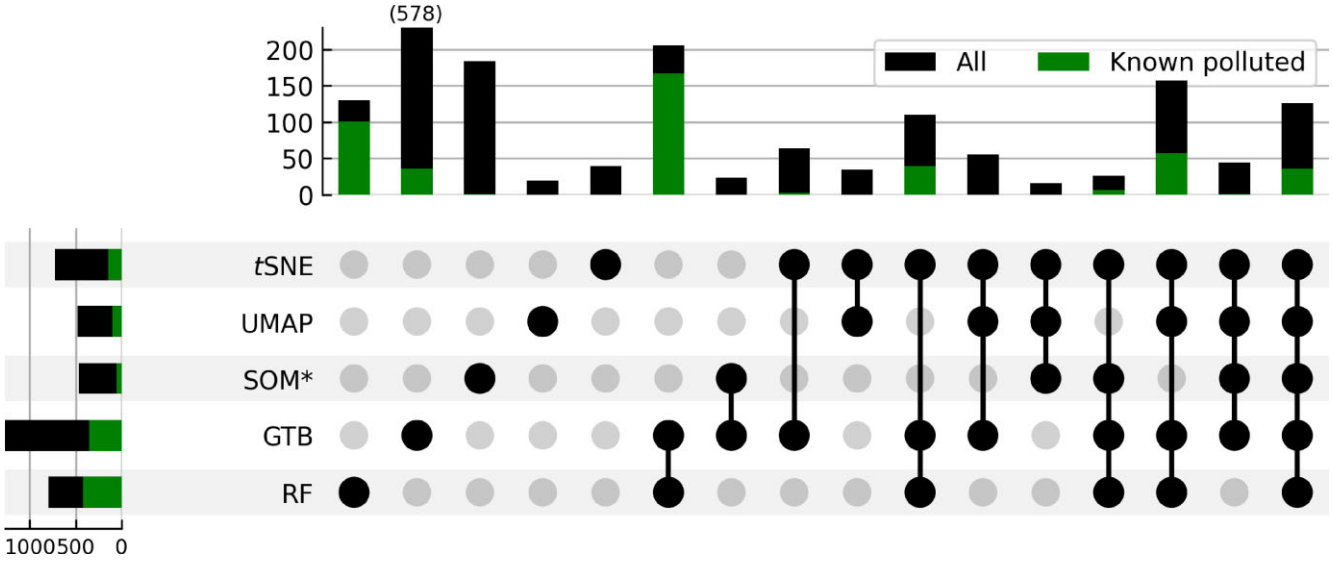

Figure 7. UpSet plot (A. Lex et al. 2014) showing the sources selected by various combinations of methods. These plots show the same information as a five-set Venn diagram, but in a less cluttered manner. The lower left panel shows the total number of sources identified as polluted by each method, as well as the subset of those that are known polluted WDs (these data are also given in Table 1). The rest of the plot shows the numbers of candidates selected by only one method (first five columns), and by various combinations of methods (shown by the connectors). For example, 130 sources are selected only by random forests; 64 are selected by *t*SNE and gradient tree boosting and none of the other three methods; 126 are selected by all five methods (rightmost column). For visualization purposes, only subsets with at least 15 sources are shown, and the second column (candidates selected only by gradient tree boosting) is clipped. A large number of candidates are selected only by one or two of the methods, even among the known polluted WDs. 39 unknown candidates are selected by *t*SNE and no other method.

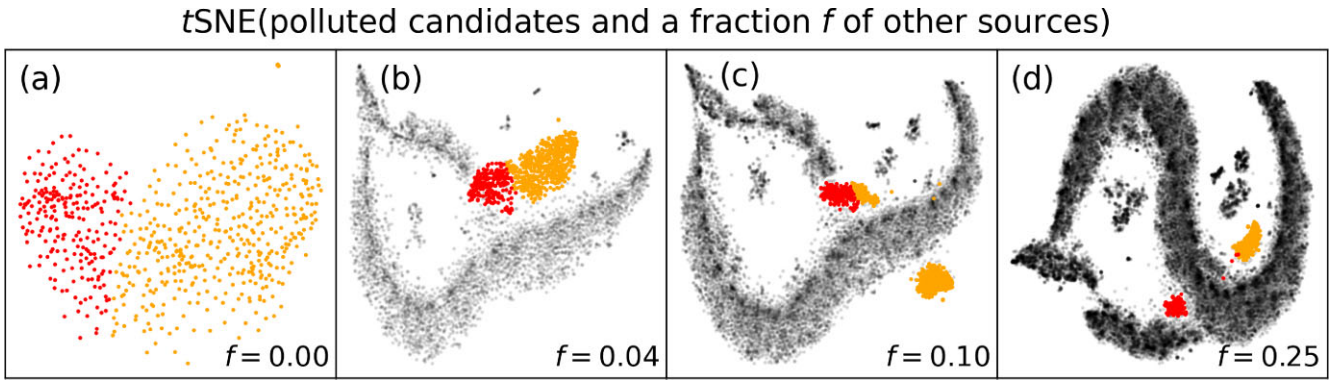

Figure 8. *t*SNE embeddings of (a) just the polluted candidates, (b–d) the polluted candidates as well as a fraction f of the rest of the sample. When just the polluted candidates are embedded, they do not form two distinct clusters. As more other objects are included, the island gradually splits into two.

distancing DAs of different temperatures. As such, the cool end of the DA sequence (upper right of Fig. 4) wraps around quite close to the region occupied by intermediate-temperature DAs, even though their spectra are quite different. Of course, global structure preservation is desirable for other use cases, such as visualizing trends across the whole data set; for such tasks *t*SNE is known to be less appropriate than UMAP (L. McInnes et al. 2018).

*t*SNE is much slower than UMAP for this data set of $\approx 100\,000$ objects, taking of order 5–10 min compared to $\sim$ 1 min for UMAP; we corroborate this finding of M. L. Kao et al. (2024). *t*SNE has a time complexity of $\mathcal{O}(N \log N)$, so its use may become impractical for data sets a few orders of magnitude larger.[6]

[6]Standard *t*SNE has a complexity of $\mathcal{O}(N^2)$, but the commonly-used Barnes-Hut algorithm (J. Barnes & P. Hut 1986) accelerates this to $\mathcal{O}(N \log N)$ (L. Van Der Maaten 2013).

4.4 Recommendations on method choice

Selecting members of a rare class from a large data set is a common task in Astronomy. In this subsection we make recommendations on which of the methods mentioned above are suited to different use cases, based on insights from the search for polluted WDs from *Gaia* XP spectra.

Where a significant number of high-quality training labels are available, supervised methods outperform unsupervised methods. These labels constitute highly relevant information that supervised methods can exploit to select members of rare classes at high true positive and low false negative rates, as shown in Table 1. Care must be taken to ensure that the labels are accurate, as inaccurate training labels can have a highly detrimental effect on model performance. Furthermore, for selecting members of very rare classes, class imbalance may become problematic, though there exist several mitigation strategies such as data augmentation and resampling (see

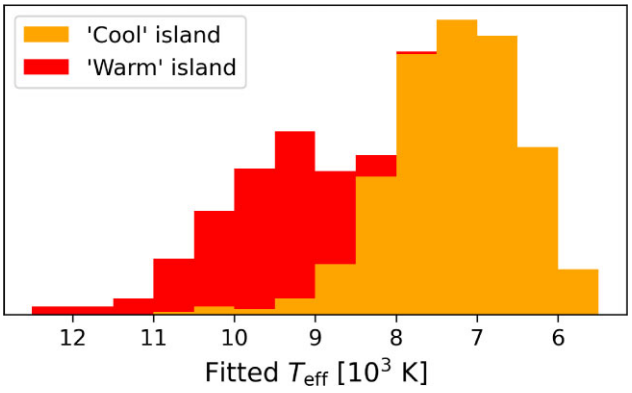

Figure 9. Stacked histogram of estimated effective temperatures of sources selected by *t*SNE on the 'cool' and 'warm' islands. The temperatures here are of the H-atmosphere WD atmosphere model that best fit the sources' de-reddened photometry (N. P. Gentile Fusillo et al. 2021); the temperature axis is inverted to follow the embeddings in Fig. 8. Although the temperature distributions are systematically different, there is a significant overlap.

also O. Vincent, P. Bergeron & P. Dufour 2025). In this case, the high performance of random forests and gradient tree boosting show that the benefits of training labels outweigh the obstacle of class imbalance, even without these mitigation strategies.

Among unsupervised methods, *t*SNE outperforms UMAP and SOMs in the selection of rare classes. As discussed in Section 4.3, UMAP can sacrifice the co-identification of members of rare classes in favour of preserving the global data set structure. SOMs – and indeed any neural-network-based methodology – are highly flexible, but require the tuning of a large number of hyperparameters: numbers of neurons, hidden layer sizes, learning rate, etc. We also attempted to apply two further neural-network-based methods–contrastive learning (T. Chen et al. 2020) and disentangled representation learning (e.g. X. Wang et al. 2024) – but found there to be far more hyperparameters to tune, compared to *t*SNE and UMAP.[7]

The methods examined here show complementary behaviour, each identifying candidates that were missed by the other methods (Fig. 7). This suggests that a good strategy would be to apply a suite of methods and select the union of the candidates selected by each method. However, for tasks with a different profile of class imbalance and label availability, some methods may contribute more candidates than others. For example, a data set with a very low number of training labels would favour the use of unsupervised methods such as *t*SNE.

5 CONCLUSIONS

We have compared five different methods in their ability to select polluted WDs from low-resolution *Gaia* XP spectra. Among these methods we present the use of *t*SNE, which we find to outperform other unsupervised methods, identifying a population of warmer polluted candidates that was missed by UMAP, a similar technique. 39 candidates selected using *t*SNE are not selected by any other method and lack spectroscopic observations at higher resolution. The availability of high-quality training labels is found to confer a significant advantage to supervised methods such as random forests and gradient tree boosting, especially for the highest signal-to-noise data. We frame this work as a case study in the common task of selecting members of a rare class from a large, sparsely labelled data set. As Astronomy surges into the 'Big Data' era, the ability to identify such interesting classes of object is crucial to a broad range of astronomical problems.

[7] *t*SNE has only one main hyperparameter: the perplexity (see Section 2.1). UMAP has two: `n_neighbours` and `min_dist` (see e.g. M. L. Kao et al. (2024), their section 3.1). While both methods do have further hyperparameters (such as the number of iterations to perform), the results are generally less sensitive to these.

ACKNOWLEDGEMENTS

We thank Malia Kao and Xabier Pérez-Couto for detailed discussions and suggestions regarding their methodologies. We thank Sarah Kane for advice on the acquisition of the *Gaia* XP spectra, and Keith Hawkins for advice on normalizing them.

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. We have also made use of the Python package GAIAXPY, developed and maintained by members of the Gaia DPAC, and in particular, Coordination Unit 5 (CU5), and the Data Processing Centre located at the Institute of Astronomy, Cambridge, UK (DPCI).

We are grateful to the Leibniz-Institute for Astrophysics Potsdam (AIP) for hosting the GAIA@AIP service, which greatly streamlined the collection of the *Gaia* data used in this work.

This research has made use of the VizieR catalogue access tool, CDS, Strasbourg, France (DOI: 10.26093/cds/vizier). The original description of the VizieR service was published in F. Ochsenbein, P. Bauer & J. Marcout (2000).

In addition to Python packages referenced in the text, we also acknowledge the use of ASTROPY (Astropy Collaboration 2013, 2018, 2022), BOKEH (Bokeh Development Team 2018), MATPLOTLIB (J. D. Hunter 2007), NUMPY (C. R. Harris et al. 2020), PANDAS (W. McKinney 2010; The pandas development team 2020), and UPSETPLOT[8].

LKR is supported by the international Gemini Observatory, a program of NSF NOIRLab, which is managed by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the U.S. National Science Foundation, on behalf of the Gemini partnership of Argentina, Brazil, Canada, Chile, the Republic of Korea, and the United States of America.

Finally, we gratefully acknowledge two anonymous reviewers, whose numerous comments and suggestions greatly improved this work.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY

The classifications of E. M. García-Zamora et al. (2025)[9] and O. Vincent et al. (2024)[10] are publicly available on the VizieR platform (F. Ochsenbein et al. 2000). The list of polluted WD candidates selected in M. L. Kao et al. (2024) is not publicly available, but the

[8] https://github.com/jnothman/UpSetPlot
[9] https://cdsarc.cds.unistra.fr/viz-bin/cat/J/A+A/699/A3
[10] https://cdsarc.cds.unistra.fr/viz-bin/cat/J/A+A/682/A5

embeddings are available as a VizieR catalogue.[11] All other data used in this work, including the *Gaia* XP spectra, are publicly available.

Scripts used to download and process data are available at https://github.com/xbyrne/wd_xp_methods.

REFERENCES

Althaus L. G., Córsico A. H., Isern J., García-Berro E., 2010, A&AR, 18, 471
Althaus L. G. et al., 2021, A&A, 646, A30
Andrae R. et al., 2023, A&A, 674, A27
Astropy Collaboration, 2013, A&A, 558, A33
Astropy Collaboration, 2018, AJ, 156, 123
Astropy Collaboration, 2022, ApJ, 935, 167
Badenas-Agusti M., Viaña J., Vanderburg A., Blouin S., Dufour P., Xu S., Sha L., 2024, MNRAS, 529, 1688
Barnes J., Hut P., 1986, Nature, 324, 446
Bédard A., 2024, Ap&SS, 369, 43
Blouin S., Dufour P., Allard N. F., 2018, ApJ, 863, 184
Bokeh Development Team, 2018, Bokeh: Python Library for Interactive Visualization
Bonsor A., Mustill A. J., Wyatt M. C., 2011, MNRAS, 414, 930
Breiman L., 2001, Mach. Learn., 45, 5
Brouwers M. G., Bonsor A., Malamud U., 2022, MNRAS, 509, 2404
Byrne X., Bonsor A., Rogers L. K., Manser C. J., 2024, MNRAS, 535, 2246
Carrasco J. M. et al., 2021, A&A, 652, A86
Chen T., Kornblith S., Norouzi M., Hinton G., 2020, in International Conference on Machine Learning. PMLR, p. 1597
Debes J. H., Sigurdsson S., 2002, ApJ, 572, 556
DESI Collaboration, 2023, preprint (arXiv:2306.06308)
DESI Collaboration, 2025, preprint (arXiv:2503.14745)
Dufour P. et al., 2007, ApJ, 663, 1291
Dufour P., Blouin S., Coutu S., Fortin-Archambault M., Thibeault C., Bergeron P., Fontaine G., 2017, in Tremblay P. E., Gaensicke B., Marsh T., eds, ASP Conf. Ser. Vol. 509, 20th European White Dwarf Workshop. Astron. Soc. Pac., San Francisco, p. 3
Ester M., Kriegel H.-P., Sander J., Xu X. 1996, Proc. Int. Conf. Knowledge Discovery and Data Mining, Vol. 96, A Density Based Algorithm for Discover Clusters in Large Spatial Datasets with Noise. p. 226
Farihi J., Barstow M. A., Redfield S., Dufour P., Hambly N. C., 2010, MNRAS, 404, 2123
Fotopoulou S., 2024, Astron. Comput., 48, 100851
Frewen S. F. N., Hansen B. M. S., 2014, MNRAS, 439, 2442
Friedman J. H., 2001, Ann. Stat., 29, 1189
Gaia Collaboration, 2016, A&A, 595, A1
Gaia Collaboration, 2018, A&A, 616, A1
Gaia Collaboration, 2023, A&A, 674, A1
García-Zamora E. M., Torres S., Rebassa-Mansergas A., Ferrer-Burjachs A., 2025, A&A, 699, A3
Gebhardt K. et al., 2021, ApJ, 923, 217
Gentile Fusillo N. P., Gänsicke B. T., Greiss S., 2015, MNRAS, 448, 2260
Gentile Fusillo N. P. et al., 2019, MNRAS, 482, 4570
Gentile Fusillo N. P. et al., 2021, MNRAS, 508, 3877
Giles D., Walkowicz L., 2019, MNRAS, 484, 834
Harris C. R. et al., 2020, Nature, 585, 357
Hawkins K. et al., 2021, ApJ, 911, 108
Hollands M. A., Koester D., Alekseev V., Herbert E. L., Gänsicke B. T., 2017, MNRAS, 467, 4970
Hollands M. A., Gänsicke B. T., Koester D., 2018, MNRAS, 477, 93
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Jura M., 2003, ApJ, 584, L91
Kao M. L., Hawkins K., Rogers L. K., Bonsor A., Dunlap B. H., Sanders J. L., Montgomery M. H., Winget D. E., 2024, ApJ, 970, 181
Koester D., Rollenhagen K., Napiwotzki R., Voss B., Christlieb N., Homeier D., Reimers D., 2005, A&A, 432, 1025
Koester D., Girven J., Gänsicke B. T., Dufour P., 2011, A&A, 530, A114
Koester D., Gänsicke B. T., Farihi J., 2014, A&A, 566, A34
Kohonen T., 1982, Biol. Cybern., 43, 59
Kullback S., Leibler R. A., 1951, Ann. Math. Stat., 22, 79
Lauffer G. R., Romero A. D., Kepler S. O., 2018, MNRAS, 480, 1547
Lex A., Gehlenborg N., Strobelt H., Vuillemot R., Pfister H., 2014, IEEE Trans. Visual. Comput. Graph., 20, 1983
Lindegren L. et al., 2018, A&A, 616, A2
Maldonado R. F., Villaver E., Mustill A. J., Chavez M., Bertone E., 2020a, MNRAS, 497, 4091
Maldonado R. F., Villaver E., Mustill A. J., Chavez M., Bertone E., 2020b, MNRAS, 499, 1854
McCracken H. J. et al., 2012, A&A, 544, A156
McInnes L., Healy J., Melville J., 2018, preprint (arXiv:1802.03426)
McKinney W., 2010, in van der Walt S., Millman J., eds, Proc. 9th Python in Science Conference. p. 56
Mustill A. J., Villaver E., Veras D., Gänsicke B. T., Bonsor A., 2018, MNRAS, 476, 3939
Ochsenbein F., Bauer P., Marcout J., 2000, A&AS, 143, 23
Ould Rouis L. B. et al., 2024, ApJ, 976, 156
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Pérez-Couto X., Pallas-Quintela L., Manteiga M., Villaver E., Dafonte C., 2024, ApJ, 977, 31
Sion E. M., Greenstein J. L., Landstreet J. D., Liebert J., Shipman H. L., Wegner G. A., 1983, ApJ, 269, 253
Steinhardt C. L., Weaver J. R., Maxfield J., Davidzon I., Faisst A. L., Masters D., Schemel M., Toft S., 2020, ApJ, 891, 136
The pandas development team, 2020, pandas-dev/pandas: Pandas. Pandas!
Van Der Maaten L., 2013, preprint (arXiv:1301.3342)
Van der Maaten L., Hinton G., 2008, J. Mach. Learn. Res., 9, 2579
Veras D., 2016, Roy. Soc. Open Sci., 3, 150571
Veras D., Leinhardt Z. M., Bonsor A., Gänsicke B. T., 2014, MNRAS, 445, 2244
Vincent O., Bergeron P., Dufour P., 2023, MNRAS, 521, 760
Vincent O., Barstow M. A., Jordan S., Mander C., Bergeron P., Dufour P., 2024, A&A, 682, A5
Vincent O., Bergeron P., Dufour P., 2025, MNRAS, 538, 2233
Wang X. et al., 2024, IEEE Trans. Pattern Anal. Mach. Intell.
Webb S. et al., 2020, MNRAS, 498, 3077
Williams J. T., Gänsicke B. T., Swan A., O'Brien M. W., Izquierdo P., Cutolo A. M., Cunningham T., 2024, A&A, 691, A352
Xu S., Bonsor A., 2021, Elements, 17, 241

[11] https://cdsarc.cds.unistra.fr/viz-bin/cat/J/ApJ/970/181

This paper has been typeset from a TEX/LATEX file prepared by the author.