Gaining Biological Insights through Supervised Data Visualization

Jake S. Rhodes[1†], Adrien Aumon[2,3†], Sacha Morin[3,4], Marc Girard[5], Catherine Larochelle[5,6,7], Boaz Lahav[5], Elsa Brunet-Ratnasingham[8], Wei Zhang[9,10], Adele Cutler[11], Anhong Zhou[9], Daniel E. Kaufmann[12,13,14], Stephanie Zandee[6,7], Alexandre Prat[5,6,7], Guy Wolf[2,3] and Kevin R. Moon[11*]

[1]Department of Statistics, Brigham Young University, West View Building, Provo, 84604, Utah, USA.
[2]Department of Mathematics and Statistics, Université de Montréal, 2900 Édouard-Montpetit Blvd, Montréal, H3T 1J4, QC, Canada.
[3]Mila - Quebec AI Institute, 6666 rue Saint-Urbain, Montréal, H2S 3H1, QC, Canada.
[4]Department of Computer Science and Operations Research, Université de Montréal, 2900 Édouard-Montpetit Blvd, Montréal, H3T 1J4, QC, Canada.
[5]Multiple Sclerosis Clinic, Division of Neurology, Centre Hospitalier de l'Université de Montréal (CHUM), 1560 rue Sherbrooke Est, Montréal, H2L 4M1, QC, Canada.
[6]Neuroimmunology Research Laboratory, Centre de Recherche du Centre Hospitalier de l'Université de Montréal (CRCHUM), 900 rue Saint-Denis, Montréal, H2X 0A9, QC, Canada.
[7]Department of Neurosciences, Université de Montréal, 2900 Édouard-Montpetit Blvd, Montréal, H3T 1J4, QC, Canada.
[8]Department of Medicine, University of California, 513 Parnassus ave, San Francisco, 94113, CA, USA.
[9]Department of Biological Engineering, Utah State University, 4105 Old Main Hill, Logan, 84322, Utah, USA.
[10]Department of Pathology, University of Utah, 15 N. Medical Drive East, Salt Lake City, 84112, Utah, USA.

[11]Department of Mathematics and Statistics, Utah State University, 3900 Old Main Hill, Logan, 84322, Utah, USA.
[12]Division of Infectious Diseases, Department of Medicine, Lausanne University Hospital and University of Lausanne, 46 rue du Bugnon, Lausanne, 1005, Switzerland.
[13]Centre de Recherche du Centre Hospitalier de l'Université de Montréal (CRCHUM), Université de Montréal, 900 Saint Denis St, Montréal, H2X 0A9, QC, Canada.
[14]Département de Médecine, Université de Montréal, 2900 Édouard-Montpetit Blvd, Montréal, H3T 1J4, QC, Canada.

*Corresponding author(s). E-mail(s): kevin.moon@usu.edu;
Contributing authors: rhodes@stat.byu.edu; adrien.aumon@umontreal.ca; sacha.morin@mila.quebec; m_girard@videotron.ca; catherine.larochelle.med@ssss.gouv.qc.ca; boblahav@gmail.com; elsa.brunetratnasingham@ucsf.edu; wei.zhang@hci.utah.edu; adele.cutler@usu.edu; anhong.zhou@usu.edu; daniel.kaufmann@umontreal.ca; stephanie.zandee.chum@ssss.gouv.qc.ca; a.prat@umontreal.ca; guy.wolf@umontreal.ca;
†These authors contributed equally to this work.

Abstract

Dimensionality reduction-based data visualization is pivotal in comprehending complex biological data. The most common methods, such as PHATE, t-SNE, and UMAP, are unsupervised and therefore reflect the dominant structure in the data, which may be independent of expert-provided labels. Here we introduce a supervised data visualization method called RF-PHATE, which integrates expert knowledge for further exploration of the data. RF-PHATE leverages random forests to capture intricate feature-label relationships. Extracting information from the forest, RF-PHATE generates low-dimensional visualizations that highlight relevant data relationships while disregarding extraneous features. This approach scales to large datasets and applies to classification and regression. We illustrate RF-PHATE's prowess through three case studies. In a multiple sclerosis study using longitudinal clinical and imaging data, RF-PHATE unveils a sub-group of patients with non-benign relapsing-remitting Multiple Sclerosis, demonstrating its aptitude for time-series data. In the context of Raman spectral data, RF-PHATE effectively showcases the impact of antioxidants on diesel exhaust-exposed lung cells, highlighting its proficiency in noisy environments. Furthermore,

RF-PHATE aligns established geometric structures with COVID-19 patient outcomes, enriching interpretability in a hierarchical manner. RF-PHATE bridges expert insights and visualizations, promising knowledge generation. Its adaptability, scalability, and noise tolerance underscore its potential for widespread adoption.

Keywords: Dimensionality reduction, data visualization, supervised learning, random forests, biological data analysis, high-dimensional data, feature-label relationships, multiple sclerosis sub-types, antioxidant effects, noisy data, COVID-19 outcomes, interpretability

Introduction

Data visualization via dimensionality reduction (DR) is an important component of exploratory data analysis in biology, especially with the recent explosion of high dimensional, high throughput biological data. Beyond data visualization, DR is also useful for downstream analysis such as classification, regression, or anomaly detection. Popular DR methods for visualization such as PHATE [1], t-SNE [2], UMAP [3], and others [4–6] are *unsupervised*. Thus these methods provide visualizations that preserve the dominant structure in the data. As such, these methods may focus on features that are not of interest to practitioners.

In many applications, we may have access to metadata or annotations, either manually given or inferred, that we want to incorporate to target the data. Often such annotations or labels are expert-derived and can provide valuable insights into relationships between variables and label information. The popular unsupervised DR-based visualization methods described previously do not use this auxiliary information during any step of the embedding process. Thus there is a need for *supervised* DR-based visualization methods that incorporate this information.

Many unsupervised DR methods begin by computing pairwise distances or dissimilarities between all points. Thus some of these methods have been extended to the supervised setting by incorporating data labels directly into the distance measure [7–11]. However, these approaches suffer from several limitations. First, these class-conditional dissimilarity functions are order non-preserving [12], which means a dissimilarity function, f, does not guarantee $f(d(x,y)) < f(d(x,z))$ when $d(x,y) < d(x,z)$ for any distance function d. Thus, these dissimilarity functions disrupt natural notions of distance, causing distortions in the relationships between data points that affect the visualization. Second, we will show that these methods tend to exaggerate the separation between observations of different classes, providing misleading visualizations and making downstream tasks unreliable. Third, the class-conditional dissimilarities are not defined for continuous labels nor extendable to unlabeled points.

We address these limitations with a new supervised DR method that we call random forest (RF)-PHATE. RF-PHATE uses a random forest to learn the relationships between the features and the supervised labels. The information is then extracted from the random forest and embedded into low dimensions, e.g., for visualization. The resulting embedding reflects the relationships between points when focusing on the features that are relevant to the supervised task while ignoring irrelevant features. RF-PHATE is also scalable to large datasets and can be applied to classification and regression problems.

We demonstrate the ability of RF-PHATE to meaningfully embed data in three case studies, including a longitudinal study of multiple sclerosis (MS) using clinical and imaging patient data, Raman spectral data collected from antioxidant-treated cancerous lung cells (A549), and COVID-19 patient plasma data. Fig. 1 provides a snapshot of each of these analyses. The ground truth of the multivariate longitudinal MS data is given in Fig. 1a followed by embeddings using t-SNE (Fig. 1b), UMAP (Fig. 1c), and RF-PHATE (Fig. 1d). Fig. 1ef shows the processing of the Raman spectral data from collection to baseline processing to finally the PHATE and RF-PHATE representations. Fig. 1h provides an overview of the data collection and processing for the COVID-19 clinical study. The patient grouping RF-PHATE representation is provided in Fig. 1i.

In each study, we show qualitatively and quantitatively that RF-PHATE provides clearer insights than competing methods. We show that RF-PHATE visualizations highlight a sub-group of patients with non-benign relapsing-remitting MS (RRMS), we demonstrate that antioxidants diminish the severity of diesel exhaust exposure to lung cells, and we show that RF-PHATE visualizations align known antibody profiles with clinically-established patient-outcome labels in COVID-19 cases. In particular, our experiments show that two significant sub-groups of RRMS patients exhibit characteristics in accordance with the traditional definition of BMS based on EDSS and disease duration. At the same time, other descriptors, such as disease worsening rate, annualized relapse rate, and treatment efficacy, distinguish our two RRMS sub-groups. A joint comparison with the SPMS group demonstrates that this distinction does not relate to RRMS patients being misclassified. Thus, our findings support the existence of benign and non-benign forms of RRMS defined as the combination of disease activity descriptors that better characterize the overall course of MS patients.

Results

RF-PHATE algorithm

RF-PHATE combines specialized RF proximities, known as RF geometry- and accuracy-preserving proximities (RF-GAP [14]), with the DR method PHATE to visualize the inherent structure of the features that are relevant to the supervised task while ignoring the irrelevant features. RF-PHATE has 5 main steps (Fig. 2):

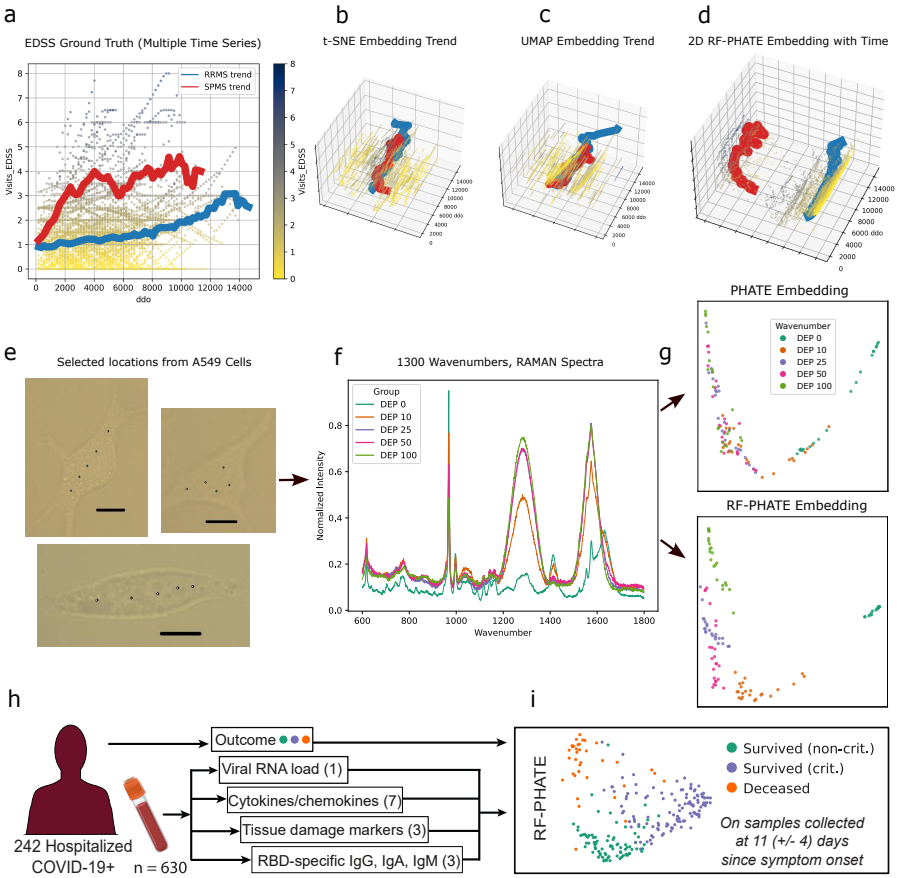

Fig. 1: **Overview of RF-PHATE and its ability to uncover latent, supervised structure**. **a**) Multiple MS patient measurements of their expanded disability status scale (EDSS) over time, along with trend averages for RRMS and secondary progressive MS (SPMS) groups. **bcd**) 2-dimensional embeddings of the MS data with the trend averages (t-SNE, UMAP, and RF-PHATE shown here). Only RF-PHATE follows the same trend as the ground truth time series. **e**) A549 cell images and locations from which Raman spectra were recorded. **f**) Baseline-corrected Raman spectral values from diesel exhaust particle exposed A549 cells across five exposure levels. **g**) PHATE, and RF-PHATE embeddings are generated from the Raman spectra. The RF-PHATE embeddings show group separation without disrupting the overall trend. **h**) COVID-19 clinical data flow resulting in interpretable RF-PHATE embeddings (**i**).

1. Train a random forest on the data to predict the labels (Methods and Fig. 2b).
2. Learn the local structure of the relevant variables using RF-GAP proximities [14] (Methods and Fig. 2c)
3. Learn the global structure of the relevant variables by diffusing the proximities (Methods and Fig. 2d)

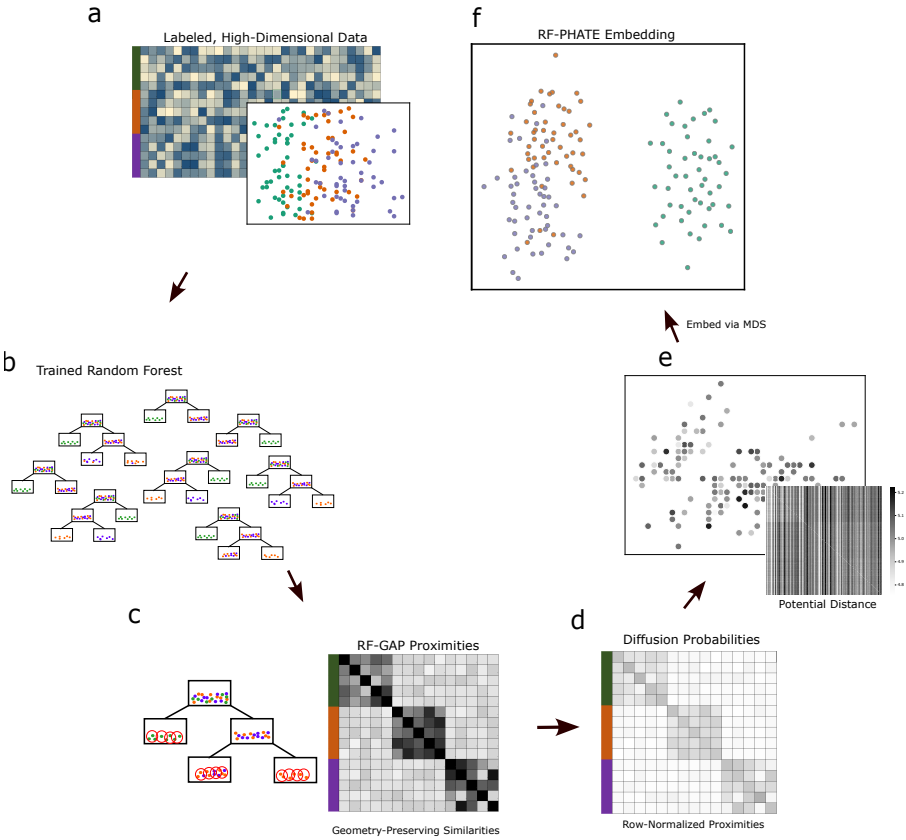

Fig. 2: A visual representation of the RF-PHATE algorithm. **a**) Noisy, high-dimensional data with labels is used to train a random forest [13] (**b**), from which local data similarities are computed using RF-GAP proximities [14] (**c**). The proximities are row-normalized, and damping [15] is applied to form the diffusion probabilities, which are stored in a Markov transition matrix (**d**). The global relationships are learned by diffusion, which is equivalent to simulating all possible random walks with time step t on the underlying Markov process. Mathematically, this is obtained by powering the transition matrix by t. A potential distance [1, 16] is then computed on the resulting diffused probabilities (**e**), which encodes both local and global relationships to which MDS [17] is applied to embed the data into low dimensions (**f**).

4. Extract the local and global structure via the potential distances (Methods and Fig. 2e)
5. Visualize the structure by embedding the potential distances using metric multidimensional scaling (MDS) into 2 or 3 dimensions (Fig. 2f)

RF-PHATE first trains a random forest to predict the provided labels. Random forests are powerful supervised learning tools that provide accurate predictions with little to no tuning [18]. Additionally, random forests are suited for both classification and regression problems and can be formed between labeled and unlabeled points once the RF is trained. Thus they naturally extend to out-of-sample points.

By using random forest geometry-preserving similarity measures, we reveal a data structure that naturally incorporates known auxiliary information while keeping the benefits of the powerful random forest model. Unlike other RF proximity measures, RF-GAP proximities theoretically and empirically capture the inherent structure of the features that are relevant to the labeled data structure [14]. RF-GAP proximities do not artificiality distort distances, but form an alternative measure relative to label information. Thus, the use of RF-GAP proximities overcomes the weaknesses of class-conditional dissimilarity measures. Thus, by using RF-GAP proximities as a local similarity measure, RF-PHATE is capable of learning a manifold structure that naturally incorporates label information while ignoring noise and features that are irrelevant to the supervised task.

The RF-GAP proximities are computed between all pairs of points. The proximities serve as an adaptive kernel, $\mathcal{P}$, capable of encoding the local data structure while retaining important information for the supervised task. From $\mathcal{P}$, we form a row-stochastic diffusion operator, P, by normalizing each row in $\mathcal{P}$. Each entry in P represents local, single-step transitions between points. We then learn the global structure of the data by powering P, simulating all possible t-step random walks.

The selection for t is often overlooked in diffusion processes. A good choice of t should denoise and learn global structure without oversmoothing. We follow the construct of PHATE [1] and use Von Neumann entropy to identify the appropriate number of steps where denoising transitions to oversmoothing. The powered diffusion operator, P^t, forms the basis for a potential distance that simultaneously retains local and global information which can be mapped to a lower-dimensional embedding. The potential distance, D_t, is computed by log-transforming entries of P^t and then computing the pairwise distances between the log-transformed columns. Other works have shown this diffusion potential is capable of mapping out noisy and complex data in biological and other contexts [1, 16, 19]. The potentially complex, nonlinear structural relationships are finally mapped to a 2- or 3-dimensional embedding, G, via metric MDS.

RF-PHATE finds meaningful structure, filtering out irrelevant features

Our goal with RF-PHATE is to devise a method that generates a low-dimensional embedding that accounts for variable importance relative to its auxiliary (i.e., label) information. As such, noise variables (features that are not related to the auxiliary information) can and should be discredited or ignored in the embedding construction. To qualitatively show that RF-PHATE is capable of extracting meaningful visualizations relative to the supervised problem, we generated a high-dimensional (60), artificial tree dataset with a known structure (Methods and Fig. 3a) and added 500 uniformly-distributed noise variables. The labels are discrete and correspond to individual branches.

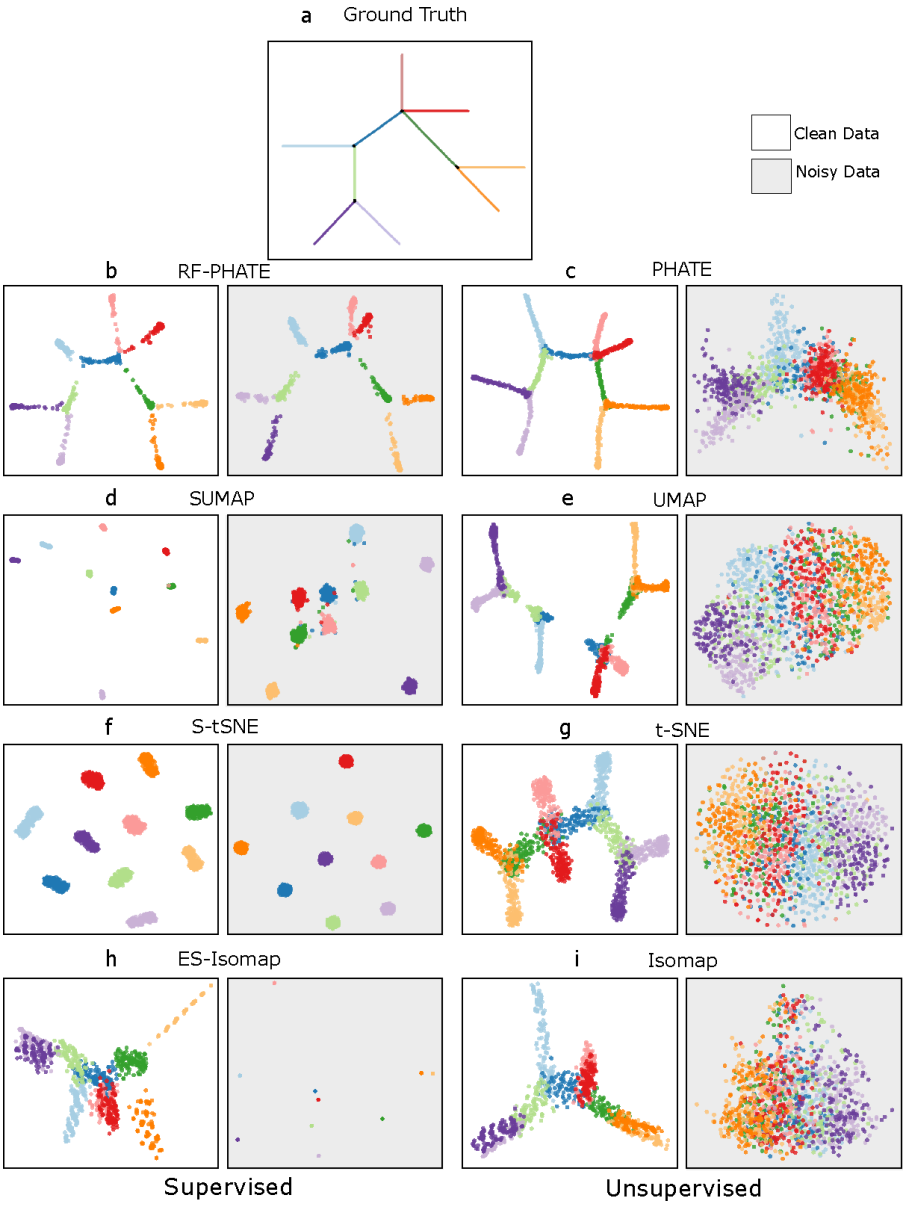

Fig. 3: RF-PHATE is robust to noisy, irrelevant features. We generated an artificial tree structure (**a**), (the details of the tree's construction can be found in Methods and [1]) and added 500 additional noise features (Methods). Without these noise variables (white background figures), RF-PHATE (**b**), PHATE (**c**), and to some degree t-SNE (**g**) accurately uncover the tree's structure, while all other supervised (**d, f, h**) and unsupervised methods (**e, i**) fail. This shows that RF-PHATE works as well as unsupervised PHATE when the features that dominate the structure are also relevant for the supervised task. However, in the noise case (gray background), none of the unsupervised nor supervised methods recover the tree structure except for RF-PHATE.

Without the extra noise variables, unsupervised PHATE (Fig. 3c) provides an accurate depiction of the true tree structure. Unsupervised t-SNE also captures the structure fairly well in the sense that all ten branches are visible and appropriately connected to each other (Fig. 3g). However, UMAP and Isomap both introduce distortions (Fig. 3e,i). After including the noise variables, all of the unsupervised methods fail to represent the branching structure. In contrast, RF-PHATE accurately depicts the branch structure with or without noise (Fig. 3b) while the other supervised methods tend to separate the branches into distinct clusters (Fig. 3d,f,h). This provides a qualitative demonstration of RF-PHATE's ability to represent the true structure of the data relative to the supervised task, in contrast with existing supervised and unsupervised DR methods.

Similar results were obtained on the Iris dataset (see Fig. S1). In particular, RF-PHATE shows the true structure of the data with or without the noise variables, while other supervised methods introduce distortions by exaggerating class separation. These trends are corroborated numerically in the next section.

Comparison of RF-PHATE to other methods

Here we quantitatively compare RF-PHATE with multiple unsupervised and supervised methods. Traditional performance metrics that evaluate an embedding's goodness-of-fit are not suitable measures in our context since our focus is on features that are important for the supervised task. Such traditional methods include those based on pairwise distance preservation [5, 17], methods based on Procrustes analysis [20], or denoised manifold affinity preservation scores (DeMAP) [1].

As an alternative, we provide three new quantitative metrics to score the quality of low-dimensional embeddings in a supervised context (see Methods for details). 1) Low-dimensional group separation (LGS) determines the extent to which classes are distinguishable in low dimensions. 2) Local structure preservation score (LSP) measures the structure of local variable importance preservation in the embeddings, and 3) global structure preservation (GSP) relays information about global variable importance in low dimensions, where variable importance is measured with respect to the supervised task. A good supervised visualization method should perform well at all three metrics if it separates class without distorting the local and global structure present in features that are relevant for the supervised task.

In a wide comparison consisting of 28 publicly available datasets (Methods, Table 7), we showed that RF-PHATE outperforms all unsupervised methods (10 total) using these metrics and generally outperforms existing supervised methods (14 total, including 5 RF methods). See Table 1. Thus based on these metrics, RF-PHATE outperforms other methods at learning an embedding that retains important information and structure relative to the supervised task in low dimensions.

Table 1: Compilation of results comparing embeddings across 24 DR methods, of which, 10 are unsupervised, 9 supervised, and 5 used RF-GAP proximities as an input. RF-PHATE outperformed all methods for both LGS and GSP and was bested only by S-tSNE for LSP. All scores were averaged across 28 publicly available datasets (Table 7) and 20 repetitions. Supervised methods (including RF methods) are marked with an asterisk.

Method	LGS	LSP	GSP	Avg. Rank
RF-PHATE*	**0.687** ± **0.01**	**0.780** ± **0.05**	**0.635** ± **0.05**	1.33
SUMAP*	**0.641** ± **0.04**	0.754 ± 0.06	**0.591** ± **0.12**	2.33
S-TSNE* [8]	0.596 ± 0.02	**0.790** ± **0.06**	0.485 ± 0.08	3.67
SNMF* [21]	0.428 ± 0.06	0.647 ± 0.07	0.449 ± 0.08	7.33
ES-LLE* [10]	0.374 ± 0.02	0.561 ± 0.04	0.381 ± 0.07	9.67
RF-UMAP*	0.327 ± 0.03	0.716 ± 0.08	0.569 ± 0.10	9.67
ES-ISOMAP* [9]	0.485 ± 0.00	0.685 ± 0.03	0.379 ± 0.03	10.33
RF-TSNE*	0.324 ± 0.02	0.752 ± 0.08	0.570 ± 0.09	10.33
NCA* [22]	0.320 ± 0.00	0.557 ± 0.03	0.411 ± 0.03	11.00
SPCA* [23]	0.241 ± 0.00	0.494 ± 0.03	0.427 ± 0.03	11.33
RF-ISOMAP*	0.206 ± 0.03	0.662 ± 0.08	0.585 ± 0.09	14.33
PLS-DA*	0.184 ± 0.00	0.531 ± 0.03	0.205 ± 0.03	16.67
T-SNE [2]	0.183 ± 0.02	0.581 ± 0.07	0.406 ± 0.09	15.33
KPCA	0.149 ± 0.00	0.444 ± 0.04	0.335 ± 0.04	18.67
PCA	0.147 ± 0.00	0.439 ± 0.04	0.296 ± 0.04	19.33
MDS [17]	0.137 ± 0.01	0.471 ± 0.11	0.331 ± 0.05	19.00
ISOMAP [5]	0.131 ± 0.00	0.536 ± 0.03	0.340 ± 0.03	17.67
UMAP [3]	0.117 ± 0.02	0.517 ± 0.07	0.366 ± 0.08	19.00
PHATE [1]	0.107 ± 0.01	0.530 ± 0.04	0.441 ± 0.06	17.67
KSPCA* [23]	0.183 ± 0.00	0.581 ± 0.03	0.267 ± 0.03	16.67
LAP. EIG. [6]	0.085 ± 0.00	0.495 ± 0.03	0.426 ± 0.03	14.67
RF-MDS* [18]	0.070 ± 0.02	0.516 ± 0.10	0.378 ± 0.11	15.67
DM [24]	0.032 ± 0.03	0.437 ± 0.06	0.281 ± 0.09	18.67
LLE [4]	-0.091 ± 0.01	0.544 ± 0.05	0.374 ± 0.06	21.33

RF-PHATE outperforms other methods in visualizing MS follow-up data in supervised settings

In this study, we extended RF-PHATE to time-series data and identified key trends underlying the course of MS. We pioneered the use of MS patient time-series data from disease onset by leveraging a unique MS database generated in the Centre de Recherche du Centre Hospitalier de l'Université de Montréal (CRCHUM), as described in Methods and Tables 4 and 5. Pairwise feature correlations can be found in Fig. S2. Diagnoses were retrospectively labeled by domain experts at the last follow-up as one of the two major forms of MS: RRMS or SPMS (secondary progressive MS). While these two forms are usually differentiated by a single progressive feature, namely the expanded disability status scale (EDSS [25, 26]), we explored other longitudinal descriptors such as functional system scores or MRI data to clinically and physiologically characterize RRMS and SPMS trajectories to potentially refine sub-types of RRMS and SPMS. Through RF-PHATE, we constructed a visualization of long-term MS courses that allows domain experts to easily retrieve known

differences between MS classes (domain knowledge retrieval) and detect potential between- and within-class differences in terms of additional longitudinal features (knowledge discovery).

We used existing labeling to project our 16-dimensional MS follow-up database into 3 dimensions using RF-PHATE and compared it with 23 other unsupervised and supervised DR algorithms using the same three metrics as before (Table 2), although they were moderately adapted to better suit time-series data (see Methods). We again showed that RF-PHATE outperforms other methods in embedding the original data while both separating the two classes and preserving the local and global structure with respect to the most important features of RRMS/SPMS classification.

Although supervised nonnegative matrix factorization (SNMF) [21] has a higher LGS score than RF-PHATE, its almost perfect score is symptomatic of overly separated classes. Indeed, SNMF fails to reach favorable performance in LSP, a lack of structure preservation corroborated by its corresponding plot (Fig. S4) in which trajectories of RRMS and SPMS patients are compressed within their respective clusters, making it difficult to extract any meaningful patterns. RF-PHATE and other competing methods showing relatively good LSP and GSP scores, such as supervised principal component analysis (SPCA [23]) or partial least squares-discriminant analysis (PLS-DA), are able to preserve the EDSS structure from the ground truth plot (Fig. S4), with RRMS and SPMS grouped motions starting from similar regions, and SPMS quickly diverging over time towards higher EDSS values. Since EDSS has been shown to be the most important feature in the RRMS/SPMS time-series classification (Table 8), this confirms the ability of these methods to embed patient trajectories while preserving the most important longitudinal structure of the original data. However, unlike RF-PHATE, SPCA and PLS-DA fail to demonstrate class separability and establish between- and within-class characterizations. Indeed, they fall far behind RF-PHATE in the LGS comparison. This is reflected visually in Fig. S4, as they do not highlight the well-separated clusters of RRMS and SPMS patient time series. As for the other embedding plots, they exhibit comparable weaknesses, if not more pronounced ones (Fig. S4). Thus, we have shown, both quantitatively and qualitatively, that RF-PHATE is the best overall method to embed our longitudinal MS database in low dimensions while separating known MS types and preserving the most important longitudinal structure from the original space. This shows that RF-PHATE is a strong approach to visualize and compare typical MS trajectories in a supervised context.

RF-PHATE provides insights into benign MS

Benign MS (BMS) is a sub-type of RRMS experiencing a mild course of disease that includes, for example, no severe relapses, complete recovery from relapses, or no progression outside of relapses. Given the heterogeneous nature of the disease, MS appears to encompass a sub-group of RRMS individuals who, even after numerous years, undergo minimal physical disability. While

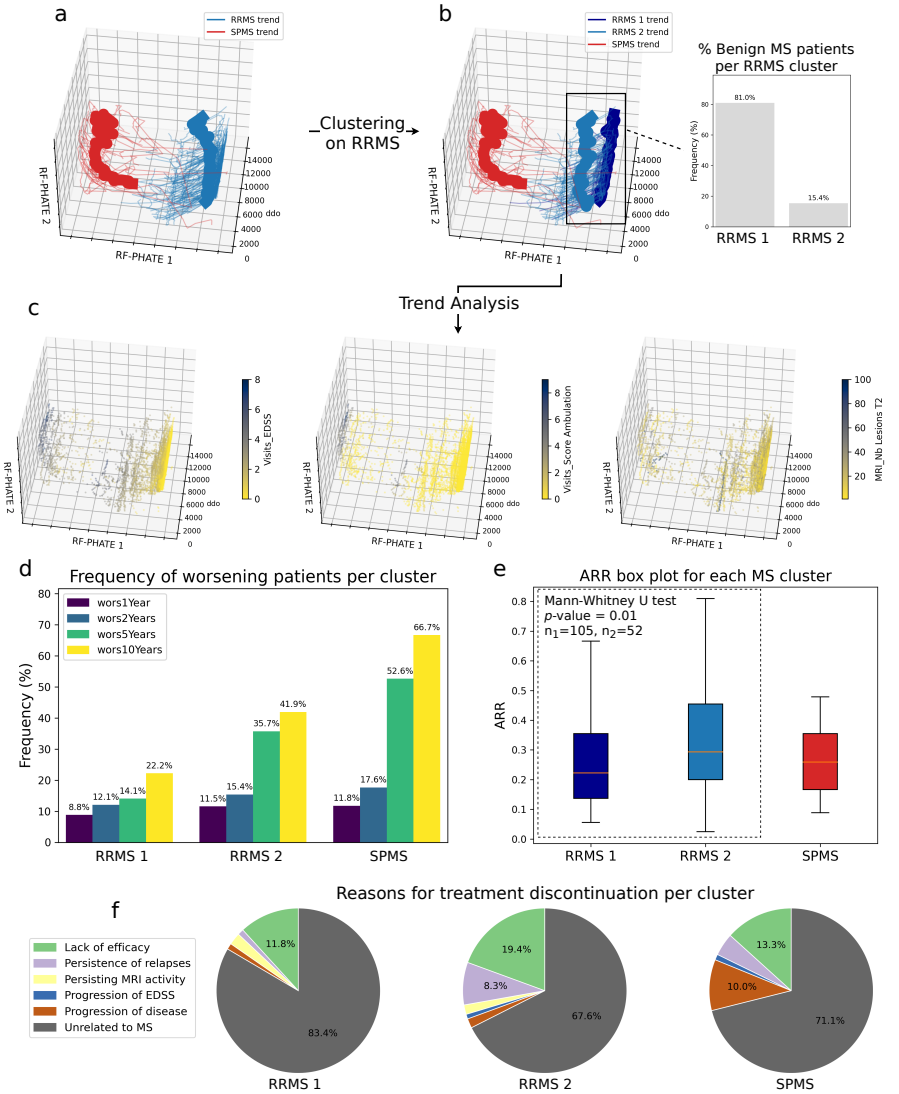

Fig. 4: **Overview of the exploratory analysis of MS follow-up data using RF-PHATE.** (**a**) Three-dimensional RF-PHATE embedding plot of MS patient trajectories colored by the two existing RRMS/SPMS classes. (**b**) Three-dimensional RF-PHATE embedding plot of MS patient trajectories colored by SPMS along with two sub-groups of RRMS detected using 2-medoids and dynamic time warping (DTW, see Methods). Bold lines show the global motion, computed as the time-wise average coordinates of the respective groups. RF-PHATE unveils two sub-groups of RRMS patients with very different profiles. Indeed, the light blue cluster exhibits high spatio-temporal variability and diverges from the dark blue cluster which consists of stable trajectories concentrated in the right portion of the plot. Moreover, thanks to RF-PHATE's ability to preserve the longitudinal structure of the data, the proximity observed between the RRMS 2 and SPMS clusters indicates similarities between these two groups of patient trajectories. (**c**) RF-PHATE plot of patient visits colored by 3 input features. (**d**) Frequency of patients worsening after $i = 1, 2, 5, 10$ years within each cluster. (**e**) Annualized relapse rate (ARR) distribution within each MS cluster. (**f**) Pie charts showing reasons for treatment discontinuation per cluster.

Table 2: Performance scores of 24 different embedding methods on MS time-series data using our three benchmark metrics LGS, LSP and GSP. Supervised embedding methods are marked by an asterisk. The numerical values are given as mean±standard deviation. We put the second-highest scores in bold, while the top scores are also underlined. Methods are sorted by their average rank across the three metrics (last column). RF-PHATE is the only embedding method ranked first (GSP) or second (LGS, LSP), making it the best overall method to embed our MS longitudinal database while separating RRMS/SPMS trends and preserving the local and global structure of the original data in terms of relevant features. While SNMF outperforms RF-PHATE in the LGS comparison, its almost perfect score is symptomatic of overly separated classes (Fig. S4). Moreover, although PLS-DA performs slightly better in the LSP comparison, it still does not effectively distinguish between the classes, as evidenced from its low LGS score and its embedding plot presented in Fig. S4.

Method	LGS	LSP	GSP	Avg. Rank
RF-PHATE*	**0.797 ± 0.01**	**0.756 ± 0.00**	**<u>0.934</u> ± <u>0.00</u>**	1.67
SPCA*	0.471 ± 0.00	0.700 ± 0.00	**0.903 ± 0.01**	3.00
SNMF*	**<u>0.999</u> ± <u>0.00</u>**	0.647 ± 0.00	0.901 ± 0.01	3.67
PLS-DA*	0.435 ± 0.00	**<u>0.779</u> ± <u>0.00</u>**	0.882 ± 0.00	4.33
RF-ISOMAP*	0.251 ± 0.02	0.664 ± 0.05	0.900 ± 0.05	7.00
KSPCA*	0.304 ± 0.00	0.677 ± 0.00	0.856 ± 0.00	7.67
RF-TSNE*	0.192 ± 0.01	0.676 ± 0.04	0.885 ± 0.05	8.67
MDS	0.299 ± 0.00	0.610 ± 0.04	0.858 ± 0.01	9.00
ES-ISOMAP*	0.476 ± 0.00	0.421 ± 0.00	0.809 ± 0.00	9.67
PCA	0.349 ± 0.00	0.526 ± 0.00	0.641 ± 0.00	10.00
RF-UMAP*	0.215 ± 0.02	0.620 ± 0.07	0.846 ± 0.05	10.33
ES-LLE*	0.341 ± 0.23	0.489 ± 0.18	0.616 ± 0.29	11.33
RF-MDS*	0.090 ± 0.00	0.605 ± 0.03	0.897 ± 0.02	11.33
NCA*	0.326 ± 0.00	0.453 ± 0.00	0.351 ± 0.04	12.67
KPCA	0.130 ± 0.00	0.515 ± 0.00	0.488 ± 0.00	15.00
ISOMAP	0.176 ± 0.00	0.241 ± 0.00	0.662 ± 0.00	16.00
SUMAP*	0.198 ± 0.04	0.370 ± 0.14	-0.104 ± 0.14	16.67
S-TSNE*	0.204 ± 0.04	0.262 ± 0.14	-0.210 ± 0.07	17.67
LLE	0.234 ± 0.00	-0.171 ± 0.00	-0.062 ± 0.00	17.67
LAP. EIG.	0.020 ± 0.00	0.356 ± 0.00	-0.129 ± 0.00	19.00
DM	-0.677 ± 0.00	0.206 ± 0.00	0.294 ± 0.00	20.33
UMAP	-0.068 ± 0.02	0.142 ± 0.10	-0.457 ± 0.05	22.00
PHATE	-0.029 ± 0.00	-0.221 ± 0.01	-0.342 ± 0.04	22.33
T-SNE	-0.075 ± 0.01	-0.115 ± 0.11	-0.573 ± 0.06	23.00

BMS is frequently observed in MS patients, there is a lack of consensus both in terms of defining this type of MS and acknowledging its existence [27, 28]. While the current standard definition of BMS is solely based on sustained EDSS scores below some threshold value [29], the practical significance of this entity in clinical practice remains uncertain [29]. This controversy arises from studies pointing out that symptoms such as fatigue, depression, cognitive dysfunction, and pain are not accounted for by the current definition of BMS despite their potentially great impact on patients' well-being and work ability [30–34]. Hutchinson [35] stated that truly benign MS is rare and strongly discouraged the clinical use of the traditional definition because it may lead clinicians to become complacent or overly-reassuring in their treatment plans. Yet, other studies found that basic EDSS-based definitions of BMS could be

sufficient to characterize patient well-being and work ability over time [36–38], despite variations in definitions across studies regarding EDSS score (0–4) and disease duration (5–15 years) [39].

In response to this controversy, several research groups have suggested alternative approaches to defining BMS. Recently, Morrow et al. [40] proposed an improved version of EDSS to consider cognitive decline during relapses. Ellenberger et al. [41] introduced a conservative definition (EDSS $\leq$ 1.0, absence from any disability, and the ability to work after 15 years of disease duration) to better reflect patients' well-being in their everyday life. From a biomarker standpoint, Correale et al. [42] reported the potential of novel quantitative MRI techniques and further assessment of spinal cord abnormalities in detecting BMS and distinguishing it accurately from other forms of the disease. In this direction, Golan et al. [43] found that computerized cognitive scores were significantly associated with quantified MRI. Additionally, Niiranen et al. [44] identified grey matter volume as a significant measure for distinguishing between EDSS-based BMS and non-BMS cases. To summarize, there is still no general agreement for the definition for BMS, but active research is being conducted toward an updated definition which covers a fuller spectrum of MS disease activity.

A powerful aspect of RF-PHATE is its ability to simultaneously consider other factors, such as the eight functional scores of EDSS and imaging data, to stratify the RRMS group into two sub-groups that could better define the distinction between benign and non-benign RRMS. Moreover, its capability to consider input features according to their importance to the RRMS/SPMS classification task means that the overall structure of the visualization is related to the long-term progression of the disease.

A detailed examination of the RF-PHATE embedding of the MS data uncovers a more severe and unstable sub-group of RRMS. The RF-PHATE plot (Fig. 4a) provides evidence for the presence of this sub-type with its capacity to distinguish between the RRMS and SPMS classes while maintaining the underlying structure. Indeed, our RRMS sub-typing (Fig. 4b) discovers 2 clusters of patient trajectories of large sizes (52 and 105 patients for the light blue and dark blue clusters, respectively), showing a significant spatial distinction. Visually, the light blue cluster exhibits RRMS trajectories characterized by high spatial variability, while the dark blue cluster gathers all of the stable RRMS trajectories. Since we have established in earlier experiments that RF-PHATE preserves the original structure, we are confident that this clear spatial differentiation directly relates to important differences within RRMS patient trajectories in terms of the input features.

Fig. 4c highlights three specific features that contribute to this differentiation (see Fig. S5 for the whole input feature set). The bottom-left part of the RRMS cluster shows a notable increase in values for EDSS (Fig. 4c, left) and nearly all other clinical indicators (Fig. S5). This suggests that the two detected RRMS sub-groups make an allusion to two time-series clusters associated with benign and non-benign RRMS trajectories that are explained

clinically. From a physiological perspective, the plots color-coded by MRI features do not clearly exhibit a distinct course toward a lower or higher number of brain lesions within each of the two RRMS clusters, except for a slight rise in the number of T2 lesions within the light blue cluster (Fig. 4c, right and Fig. S5). This implies that the overall clinical worsening of the light blue cluster in Fig. 4b is not necessarily linked to a progressive increase in the number of lesions over time, regardless of the brain region involved. However, this does not mean that there is no link between MRI data and clinical symptoms in the long run. Indeed, categories used in the McDonald criteria [45] do not seem refined enough to quantify meaningful radiological patterns. For instance, the McDonald Lesions T2 and McDonald Lesions Periventricular plots mostly consist of visits with scores 9+ and 3+, respectively, which fails to capture higher levels of quantification. Therefore, visually detecting physiological differences from our data is challenging, and further investigation is conducted in the next section.

At the same time, the light blue time-series cluster is similar to the neighboring SPMS cluster in terms of clinical and radiological features. This means that the light blue time-series cluster may also contain patients mislabelled as RRMS in place of SPMS. Indeed, SPMS patients could have been intentionally classified as RRMS to alleviate the financial and emotional burdens associated with being categorized as a more severe form of MS ("therapeutic mislabelling" [46]). It is also possible that SPMS was not diagnosed in RRMS patients due to the gradual nature of functional decline, which often goes unnoticed by both patients and doctors. This phenomenon, known as "silent progression" [47, 48], is observed in the RF-PHATE plots (Fig. 4c). While the second RRMS branch shares very similar features to the SPMS group, the latter generally displays high feature values earlier after disease onset (cf. Score Mental and Score Visual sub-plots of Fig. S5). Still, poor scores in Ambulation (Fig. 4c, center), Brainstem, and Cerebellar (Fig. S5) remain very specific to the SPMS cluster, which rather supports the existence of a distinct RRMS sub-group. Acknowledgment of this sub-group may have future implications on patient treatment, which is not currently optimized for this specific sub-group (see Fig. S6).

Quantitative investigation supports the existence of a more severe RRMS sub-group

Using the previous visual observations as a starting point, we examined other accessible patient data to provide a rationale for the distinct behaviors exhibited by the three clusters of interest, namely RRMS 1 (dark blue in Fig. 4), RRMS 2 (light blue), and SPMS (red).

We wanted to verify whether the two RRMS sub-clusters were indeed associated with benign and non-benign MS courses. To do so, we first used a basic definition of BMS that combines the EDSS score with disease duration. That is, BMS was identified in a patient if and only if their EDSS score was 3 or lower for a disease duration of 20 years or more, or if the EDSS score was 2 or

lower for a disease duration of 10 years or more. Consequently, out of the 105 patients in the dark blue cluster, 85 (81.0%) were assigned with BMS, whereas only 8 out of the 52 patients (15.4%) in the light blue cluster were assigned with BMS (Fig. 4b), which confirmed our previous statement. Another aspect of long-term clinical characterization regarding MS disease progression is the worsening of MS symptoms, as defined in Methods. Fig. 4d highlights a large increase in the frequency of worsening patients after 5 years since disease onset within the RRMS 2 (non-benign) cluster which is not apparent within the RRMS 1 (benign) cluster. Indeed, while the frequency of worsening patients within the RRMS 2 cluster remained below approximately 15% up to 2 years since disease onset, it escalated to 35.7% and 41.9% after 5 and 10 years, respectively. In contrast, RRMS 1 patients experienced a stable worsening over time, with only 22.2% of worsening patients after 10 years since disease onset.

However, since the definitions of BMS and worsening focus on EDSS and disease duration, the above results do not consider the radiological structure captured by our clustering approach and are thus limited to a clinical differentiation between the two RRMS trends. To also identify physiological differences between the two RRMS trends, we make use of the annualized relapse rate (ARR) of each patient. The main benefit of using the ARR is its ability to both clinically and physiologically quantify the relapsing condition of an RRMS patient, as relapses are defined as periods characterized by clinical deterioration and radiological activity [49]. Thus, the ARR is a way to complement our BMS/non-BMS differentiation from a physiological standpoint. Because of the oscillating trajectories depicted in the non-BMS cluster (Fig. 4b), we expect the non-BMS cluster to be associated with a higher ARR. We observed a mean ARR of 0.26 ± 0.16 within the dark blue cluster and a mean ARR of 0.37 ± 0.27 within the light blue cluster. Additionally, we performed a one-sided Mann-Whitney U test [50] ($\alpha = 0.05$) to determine whether the ARR distribution underlying the BMS cluster is stochastically less than the ARR distribution underlying the non-BMS cluster (Fig. 4e). The test yielded a p-value of 0.01, showing evidence of a difference in ARR between the two clusters. Likewise, Fig. 4f reveals that the non-benign RRMS cluster experiences a significantly higher rate of treatment discontinuation due to the persisting presence of relapses and MS symptoms.

One potential explanation for the existence of the two RRMS sub-clusters is that a substantial portion of the second RRMS cluster could have been misclassified as RRMS instead of SPMS. First, the potential for therapeutic mislabeling still exists, given that the process of verifying it would involve conducting interviews with all the neurologists responsible for diagnosing the patients, which can be a laborious task. Moreover, as of 2023, we have observed three patients from the RRMS 2 cluster transitioning into SPMS, and one patient has passed away. Therefore, it is possible that a few patients from the light blue cluster should have been categorized as SPMS. However, we believe that most of them are more appropriately associated with a distinct sub-group of RRMS, primarily due to important differences in terms of clinical worsening

rate (Fig. 4d) and some functional scores such as Ambulation (Fig. 4c, center) which are typically crucial factors in the transition to SPMS [51]. Physiologically, Fig. 4e also shows that the SPMS cluster has a lower ARR, indicating the ongoing presence of a relapsing condition within the RRMS 2 cluster. Thus, our findings support the existence of benign and non-benign forms of RRMS.

RF-PHATE has been an essential component of this investigation, and no other methods would have led us to explore these insights. Thanks to its ability to clearly distinguish between the RRMS and SPMS groups (Fig. S4), RF-PHATE enabled us to uncover the high variability within the RRMS cluster of trajectories, prompting us to apply a clustering algorithm to this specific cluster. Additionally, its capability to preserve the structure allowed us to swiftly identify similarities between the lower branch of the RRMS cluster and the SPMS cluster due to their spatial proximity.

Visualizing A549 antioxidant cell protection via Raman spectroscopy and RF-PHATE

Raman spectroscopy is a non-invasive technology used in various applications of cellular and other biological studies. Raman spectral readings are capable of detecting minute molecular variations and can be used to determine the health of infected or exposed cells [52–55]. These spectral readings often contain noise. Using RF-PHATE and relying on its ability to denoise data as a part of its embedding process, we visually validated the protective and restorative effects of the antioxidants cannabidiol (CBD) and resveratrol (RES) on the health of A549 [55] cells at various levels of exposure to diesel exhaust particles (DEPs), where the supervised labels are the exposure levels. The RF-PHATE embedding shows that the Raman spectra of cells express lower levels of change in the presence of antioxidants when compared to a control group (Fig. 5a,d). Protected cells (Fig. 5b,c,e,f) provide readings more similar to each other but different from the control group (DEP 0) no matter the level of DEP exposure (Fig. 5b,c). This suggests that the antioxidants provide some level of protection against DEP exposure but do not provide total protection. In contrast, PHATE is unable to detect this change as the cells with different exposure levels are difficult to distinguish with or without the antioxidants.

We numerically demonstrate this phenomenon by recording the k-NN scores of PHATE and RF-PHATE on the three spectral cell groups exposed to diesel exhaust particles. In each case, the k-NN 10-fold cross-validated accuracy means and standard deviations were recorded. The results can be found embedded in Fig. 5. When the cells are treated with CBD, or RES, the accuracy decreases when trained using the RF-PHATE embeddings, numerically showing that cells at all nonzero exposure levels are more similar to each other after treatment. In contrast, the PHATE embedding does not display this pattern.

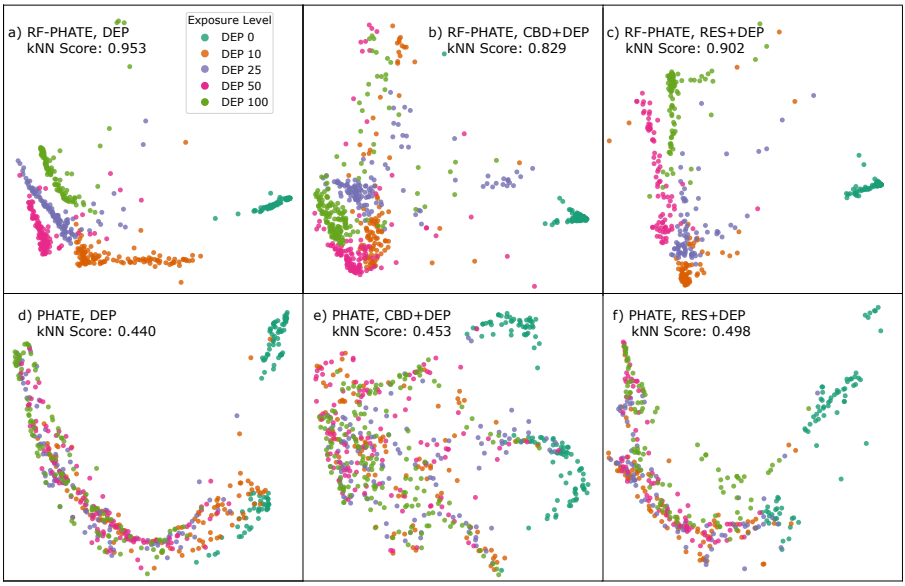

Fig. 5: Two-dimensional PHATE and RF-PHATE representations of the collected Raman spectra of A549 cells. Given the DEP exposure values as labels, RF-PHATE shows that the DEP exposure forms partially separable groups. Notably, when no antioxidants (RES or CBD) are present, the levels of DEP exposure are more distinguishable. When antioxidant protectants are added, the control group (no DEP) is distinguishable from the other groups, but exposed groups become less distinguishable from each other, suggesting some level of protection is given at all DEP levels. In contrast, the class separability does not significantly change in the PHATE embeddings (**d, e,** and **f**).

Hierarchical visualization of outcomes and immunovirological endotypes in COVID-19 patients

Coronavirus disease 2019 (COVID-19) induces a wide range of patient responses, spanning from asymptomatic cases to severe and fatal outcomes. Plasma viral RNA (vRNA) was quickly identified as a good predictor of COVID-19 patient fatality [56], but the range of patient responses to the disease prompted further investigations into additional immunological features to better understand disease progression. We identified [57] 4 patient endotypes using k-means on a set of cross-sectional measurements consisting of plasma vRNA, Spike Receptor Binding Domain (RBD)-specific antibody responses, and plasma levels of cytokines and tissue damage markers. Clusters were visualized using the PHATE embedding of the plasma samples (Fig. 6, a), revealing a global structure primarily driven by antibody and cytokine concentrations [57]. Statistical analyses further confirmed the association between k-means clusters and 3 clinically defined patient groups: non-critical survivors (clusters 3 and 4), mechanically ventilated survivors (i.e., critical survivors, cluster 2), and deceased patients (cluster 1).

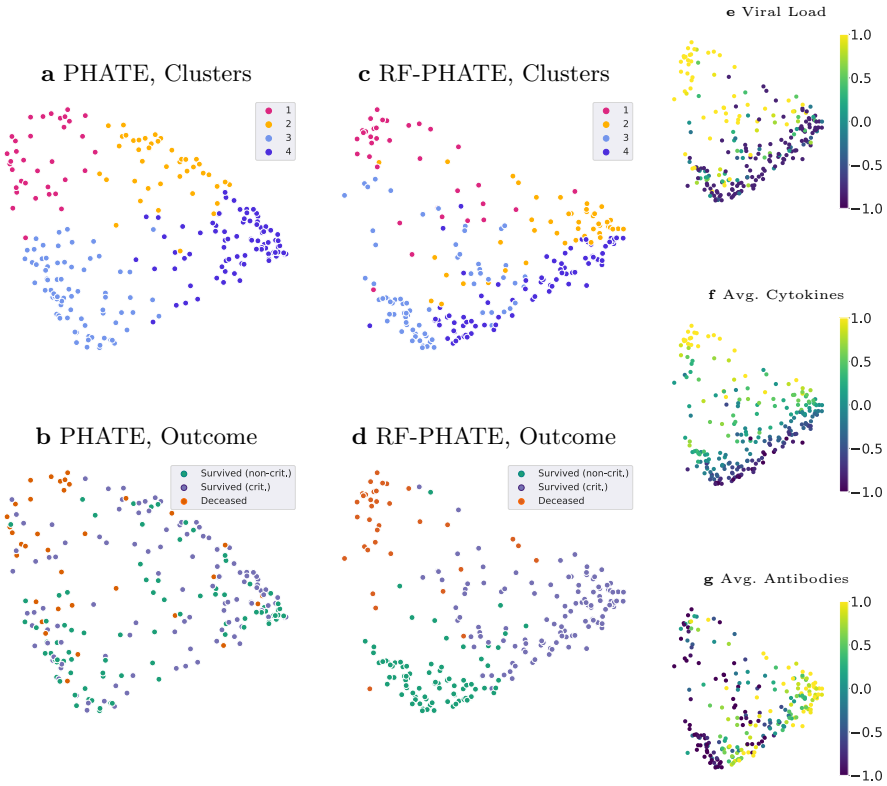

Fig. 6: PHATE and RF-PHATE embeddings of plasma samples from hospitalized COVID-19 patients at 11 days after symptom onset (DSO11) [57, 58]. (**a**), PHATE embedding and patient k-means clusters from [57]. The 4 clusters display distinct early immunological profiles and cluster 1 is particularly enriched in high-fatality cases. (**b**), PHATE embedding colored by outcome: survivors are divided into non-critical and critical based on the level of respiratory support they required. (**c**), RF-PHATE embedding using the 3-class outcome as the classification target. The original global geometry of the unsupervised PHATE embedding is still present and some clusters are preserved, illustrating how RF-PHATE factors in both the input data geometry and the supervised signal. (**d**), RF-PHATE embedding colored by the target outcome. As opposed to PHATE, RF-PHATE clearly separates the different outcomes, with both survivor groups located together at the bottom of the embedding. (**e, f, and g**), Average of different input variable categories as coloring for the RF-PHATE embedding. Input variables are standardized (mean 0, variance 1) and the color map is clipped to the range $[-1, 1]$ to enhance contrast. The embedding globally correlates with viral load and cytokine concentrations. The heterogeneity of the antibody response in the non-critical survivor group discussed in [57] can be visualized by observing the divide in the lower region of (**g**).

While the associations between clusters and patient outcomes can be measured numerically — e.g., cluster 1 (Fig. 6, a) is particularly fatal —, noise and class imbalance prevent proper visualization of the outcomes (Fig. 6, b). In contrast, RF-PHATE, supervised with the 3-class patient outcome, retains

much of the structure present in the original unsupervised setting. E.g., the k-means clusters (Fig. 6, c) are still largely separable and are continuously altered to align with patient outcomes (Fig. 6, d), with both survivor groups being located together at the bottom of the embedding. The clean separation between outcomes and the preservation of the original cluster regions demonstrate how RF-PHATE balances both the input data geometry and the guiding target. The overall RF-PHATE geometry separates patient outcomes and globally aligns with vRNA (Fig. 6, e) and average cytokine concentrations (Fig. 6, f). The embedding then exhibits a hierarchical structure: the 2 patient classes are locally organized according to the average antibody concentrations (Fig. 6, g), with a particular "gap" in the non-critical survivor cluster reflecting the heterogeneous antibody profiles of this group as discussed in [57].

Unsupervised embeddings of high dimensional data can identify biologically relevant clusters, such as types of immune response to acute SARS-CoV-2 infection. However, some stratification may be irrelevant in the setting of patient treatment - for example, the distinction between clusters 3 and 4. Supervised embeddings using RF-PHATE have the added benefit of identifying clinically relevant clusters based on an outcome deemed relevant while preserving the structure of the biological data. This type of stratification can be used for early identification of high-risk patients, allowing for treatment prioritization.

We also performed a numerical comparison on patient time series extracted from the COVID-19 data, shown in Table 3. As before, RF-PHATE outperforms all other methods on average across our three metrics.

Discussion

When working with high-dimensional biological or other complex data, a lower-dimensional visualization of the data can assist practitioners in better understanding the relationships between measured variables and a labeled response. However, the majority of existing DR approaches do not use extra, expert-provided auxiliary information to generate an embedding for this use case. Those that do use labels distort the global data structure by imposing class-driven constraints that artificially separate observations of different classes in the low-dimensional embedding. In this paper, we provided RF-PHATE as an alternative to these unsatisfactory supervised methods. RF-PHATE combines the powerful predictive power of random forests with diffusion-based informational geometry designed specifically for visualization. RF-PHATE is scaleable to small and large datasets, employs the flexibility of random forests, and works well for both continuous and categorical labels. RF-PHATE is robust to noise due to both the ability of random forests to determine feature importance and because of its diffusion dynamics.

We showed the ability of RF-PHATE to function with time-series data and produce interpretable insights in a longitudinal study of MS patients

Table 3: Performance scores of 24 different embedding methods on COVID-19 longitudinal data using our three benchmark metrics LGS, LSP and GSP. The scores are calculated by averaging the results from 20 repetitions. The mean and standard deviations are recorded. Methods are sorted by their average rank across the three metrics (last column). Supervised embedding methods are marked by an asterisk. Just like in our MS comparisons, RF-PHATE is still the best method overall.

Method	LGS	LSP	GSP	Avg. Rank
RF-PHATE*	**0.794 ± 0.01**	0.656 ± 0.01	**0.745 ± 0.01**	2.00
ES-ISOMAP*	0.715 ± 0.00	**0.757 ± 0.00**	**0.704 ± 0.00**	2.33
SNMF*	**0.999 ± 0.00**	0.631 ± 0.00	0.651 ± 0.02	2.67
SUMAP*	0.751 ± 0.06	**0.691 ± 0.11**	0.575 ± 0.05	3.67
SPCA*	0.313 ± 0.00	0.525 ± 0.00	0.593 ± 0.00	6.00
RF-UMAP*	0.294 ± 0.01	0.510 ± 0.09	0.586 ± 0.03	7.33
RF-TSNE*	0.283 ± 0.02	0.529 ± 0.06	0.557 ± 0.05	8.00
PLS-DA*	0.312 ± 0.00	0.457 ± 0.00	0.486 ± 0.00	8.67
RF-ISOMAP*	0.334 ± 0.02	0.349 ± 0.06	0.506 ± 0.11	9.33
PHATE	0.270 ± 0.00	0.596 ± 0.01	0.436 ± 0.00	9.33
S-TSNE*	0.600 ± 0.01	0.426 ± 0.03	0.135 ± 0.04	10.33
ISOMAP	0.276 ± 0.00	0.389 ± 0.00	0.421 ± 0.00	12.00
UMAP	0.233 ± 0.01	0.411 ± 0.08	0.255 ± 0.11	14.00
NCA*	0.286 ± 0.00	-0.150 ± 0.05	-0.179 ± 0.03	15.33
RF-MDS*	0.196 ± 0.00	0.289 ± 0.06	0.311 ± 0.03	15.33
LLE	-0.094 ± 0.00	0.439 ± 0.00	-0.336 ± 0.00	17.67
ESLLE*	-0.431 ± 0.01	0.250 ± 0.11	0.245 ± 0.02	17.67
PCA	0.257 ± 0.00	-0.296 ± 0.00	-0.214 ± 0.00	18.00
T-SNE	0.219 ± 0.00	-0.158 ± 0.02	-0.175 ± 0.01	18.00
MDS	0.238 ± 0.00	-0.327 ± 0.01	-0.197 ± 0.03	18.33
DM	-0.726 ± 0.00	0.136 ± 0.00	-0.150 ± 0.00	19.00
LAP. EIG.	0.234 ± 0.00	-0.539 ± 0.00	-0.382 ± 0.00	20.33
KPCA	0.186 ± 0.00	-0.511 ± 0.00	-0.389 ± 0.00	21.67
KSPCA*	0.175 ± 0.00	-0.814 ± 0.00	-0.729 ± 0.00	23.00

from which we uncovered a non-benign sub-group of RRMS patients. RF-PHATE demonstrated superior performance compared to existing DR methods in embedding MS follow-up data by effectively separating the two RRMS and SPMS groups while preserving the essential structure of the original space. In addition, RF-PHATE proved useful in practice by accomplishing two objectives: retrieving domain knowledge pertaining to RRMS and SPMS courses based on clinical and radiological features, and highlighting a non-benign RRMS sub-group that was not explicitly identified in our initial database. The non-benign RRMS sub-group does not primarily involve SPMS patients being misclassified. Instead, compared to SPMS, this RRMS sub-group exhibits a lower worsening rate ("silent progression") and a higher ARR showing an active relapsing-remitting condition. This conclusion strengthens the notion of a distinct RRMS sub-group with unique characteristics and specific needs.

In light of significant treatment failure (Fig. 4f) despite the use of similar medication on the non-benign RRMS cluster (Fig. S6), our future directions include increased attention to RRMS patients exhibiting a disease course similar to the non-benign RRMS cluster. This will enable them to receive

appropriate treatment as early as possible to improve their outcomes. Moreover, incorporating spinal cord MRI data could further improve the visual differentiation between our MS sub-groups, as spinal cord lesions are known to greatly affect clinical symptoms [59, 60]. Finally, another direction is the reclassification of RRMS patients into SPMS or another RRMS sub-type, with the aim of enhancing the efficiency of downstream prediction tasks related to the course of MS disease. Indeed, by specifically targeting three distinct profiles of MS courses, namely RRMS 1, RRMS 2, and SPMS, this approach would enable an early prognosis for MS patients regarding their forthcoming MS course into three degrees of severity.

We demonstrated RF-PHATE's ability to create meaningful visualizations in the noisy context of Raman spectroscopy. Here RF-PHATE displayed the protective effects of two types of antioxidants against diesel exhaust particles on A549 cells. In a COVID-19 case study, we showed that RF-PHATE embeddings aligned known geometric data structures with target labels to produce a visualization that both corroborates previous findings of covariates associated with patient outcomes and enriches these relationships in a hierarchical structure with increased interpretability. We anticipate that RF-PHATE is capable of producing valuable insights in a variety of fields and applications, both biological and otherwise. We expect to see new uses in private research and industry, with biomedical and clinical data, by revealing information in these fields via human-interpretable visualizations.

Methods

Here we provide additional details concerning our methods including background information on random forests, the RF-GAP proximities in manifold learning, descriptions and benefits of diffusion-based information geometry, approaches used to quantify embedding fit, and specific methodologies for data preparation and analysis.

Overview of supervised manifold learning

Visualizing high-dimensional data often serves as an early step in exploratory data analysis. Complex relationships between data points and features become more meaningful when practitioners can view relevant spatial or temporal representations of the data as a complement to summary statistics. Manifold learning provides a means to generate low-dimensional embeddings that can serve as tools for high-dimensional data visualization or data preprocessing.

Most manifold learning approaches do not use auxiliary information, such as labels, in the embedding-learning task. It is reasonable to ask if/how labels can be used in this traditionally unsupervised machine-learning task. Expert-provided labels should provide additional insights into the data structure relative to the label information. Providing an algorithm a target to seek (i.e., supervised learning) can assist the algorithm in learning a representation from features that relate to the supervised task. Supervised counterparts

for common manifold learning algorithms exist and are typically adapted at the algorithm's distance- or affinity-learning level. In some cases, distances are rescaled depending on class affiliation, as in the case of WeightedIso [61], a supervised version of Isomap [5]. Others additively increment distances between points of opposing classes, as is done in SLLE [11], a supervised version of local linear embedding (LLE) [4]. In the cases of supervised (S)-Isomap [62], enhanced supervised (ES)-Isomap [9], enhanced supervised (ES)-LLE [10], supervised Laplacian Eigenmaps (S-LapEig) [63], and supervised (S)-tSNE [8], dissimilarities are derived as class-conditional adaptations of a Gaussian kernel function.

However, manifold learning methods using these class-conditional dissimilarities suffer from a number of weaknesses: (1) The global structure relative to inter-class relations is disrupted. This is corroborated by the authors of [12, 64], who show these dissimilarity functions are order-non-preserving isomorphisms and should not be used for visualization tasks lest one draws misleading conclusions about the data structure. (2) The utility of downstream tasks is diminished. For example, class-conditional dissimilarities often induce perfect class separation which is inconsistent with the intrinsic dataset structure. Classification following DR in this manner can give unrealistically low error rates. (3) These dissimilarity measures do not provide a natural extension for continuous-valued labels and are thus incompatible with regression problems. (4) Dissimilarities between labeled data and unlabeled data (e.g., a training set vs. a test set) are not defined.

To counter each of these weaknesses, we designed RF-PHATE to use the random forest-based RF-GAP proximities [14] as a similarity-learning step instead of conditionally increasing distances between observations of different classes. RF-GAP similarities are computed in a manner that preserves random forest learning and thus provide a powerful local similarity representation for supervised manifold learning.

RF-GAP proximities serve as similarity measures that use data labels in accordance with the supervised model's learning and thus comprise a measure that overcomes the weaknesses of the class-conditional distances: (1) RF-GAP proximities form a local similarity measure that partially retains global information through the forest's recursive splitting process. Thus, the data geometry learned by the random forest captures hierarchical feature importance and does not induce artificial class separation but distinguishes between classes in a manner consistent with the intrinsic structure of the dataset relative to the labels. Thus instead of distorting the global structure, the proximities naturally retain observational relationships. (2) Rather than diminishing the integrity of downstream tasks, RF-GAP proximities retain random forest learning without overfitting to the training data. Thus, downstream tasks are positively influenced by this supervised information. (3) Class-conditional similarities are subject to discrete labels, but random forests can handle both discrete and continuous labels. RF-GAP proximities provide a way to perform supervised manifold learning with continuous labels. (4)

Class-conditional similarities are not extendable to an unlabeled test set, but a trained random forest can define similarities to out-of-sample observations, supporting semi-supervised learning and other tasks.

Similarity learning with random forests

Though dozens of variations have been introduced since the random forest's conception in 2001, the original random forest algorithm [13] is still used in a variety of applications today. Citing a few recent examples, random forests have been used to classify water salinity [65], groundwater nitrate pollution [66], to assess the risk of coronary heart disease [67], to predict cellular types and health using Raman Spectroscopic features [55], to classify oncosomatic gene variants using next-generation sequencing, outperforming recurrent neural networks [68], to predict compounds with anti-aging properties [69], and to predict the prognosis of diabetic patients [70].

In addition to prediction problems, random forests are also commonly used for feature selection or data preprocessing. Feature analysis for traditional ecological models was performed using random forests in [71]. Longitudinal predictors of health in exposome studies were identified using random forests in [72]. Feature selection has also been done via random forests in the context of surface-enhanced Raman scattering data [73], genome-wide associations using population structures and individual genetic variants [74], and for deep-learning preprocessing to prevent overfitting in gene-expression classification [75].

Random forest usage is universal because they are simple to use and they provide a broad array of benefits in addition to prediction. In particular, random forests perform well with little or no parameter tuning and can be used when the number of features exceeds the number of observations ($p > n$), which is often the case with biological data. They are compatible with both regression and classification problems and can be applied to both continuous and categorical features without feature engineering. This last benefit allows forest-based similarities to construct meaningful relationships with observations of mixed feature types which is often a factor in clinical studies [76]. In contrast, most manifold learning methods do not naturally handle mixed features.

In simple terms, random forests are an ensemble of randomized decision trees. Each decision tree is formed using a random bootstrap sample (sampled with replacement) of training data. In the literature, observations kept in the bootstrap sample are called in-bag observations, while those not included are called out-of-bag (OOB) observations. The OOB observations serve as a quasi-test set for model validation.

The trees are grown by a recursive splitting of the data across a randomized selection of features in a manner that maximizes purity within resulting nodes. The recursive process continues until a stopping criterion is met. The set of terminal nodes forms the decision space of the random forest. A new observation, x_0, is passed down each decision tree and a vote takes place within

its terminal node of residence. An overall, unweighted vote is counted across all trees in the forest, providing the predicted label for x_0. It is important to note that the collection of in-bag observations in the terminal nodes of x_0 exclusively determines its predicted label. Thus, the random forest can be viewed as a nearest-neighbor decision algorithm, where "neighbors" are in-bag observations in a shared terminal node [77].

The space of terminal nodes or "neighborhoods" can be used to define a similarity or proximity between observations. Since observations are organized according to splits across the features which are most useful for partitioning data, these similarities naturally incorporate feature importance. While two observations may be considered distant in a Euclidean space, they might be very similar with regard to important features for the supervised problem. Random forests are also robust to noise, making forest-derived proximities useful to process noisy, high-dimensional data.

Random forest proximities are typically constructed pair-wise according to terminal node residence. Random forest-based proximities were first defined as the proportion of trees in which a pair of observations shares a terminal node [13, 78]. However, this proximity definition tends to exaggerate class distinction as in-bag examples of opposing classes necessarily reside in different terminal nodes when trees are fully grown. As a fix, proximities are often constructed using only OOB pairs of points [79, 80]. These OOB proximities are still insufficient at capturing the random forest's learning because 1) they do not directly use training examples (in-bag observations), and 2) they fail to account for the number of observations in each terminal by which random forest voting is conducted [14].

To fully incorporate the random forest's learning and yet prevent overfitting to the training data, the proximities need to be constructed in a manner consistent with the random forest's modus operandi: predictions for new or OOB observations are made by majority vote or by averaging labels of in-bag observations within each terminal node. That is, instead of counting pair-wise interactions without regard to other points in a common terminal node, the proximities should be formed relative to the number of in-bag observations within the terminal nodes. Rather than constructing proximities between training points, the training points are used to construct proximities to OOB points. In this way, the proximities to a given OOB observation are consistent with the forest's predictive model. It was shown in [14] that thus defined RF-GAP proximities are capable of perfectly recreating the random forest OOB and test prediction when serving as weights in a weighted average (or weighted majority vote) of the corresponding data labels. It was also shown that multiple proximity-based applications are improved by using RF-GAP proximities. Thus, we use them as a supervised similarity measure for supervised manifold learning.

A few modifications of the RF-GAP proximities are required to use them in manifold learning. First, by default, the diagonal entries of the RF-GAP proximities are defined to be zero as an observation cannot be simultaneously

in- and out-of-bag for a given decision tree. To overcome this, we redefine the diagonal entries by, in essence, passing an identical, OOB point down each tree in which the original point is in-bag. Using this point to generate the diagonal proximities we maximize the similarity between an observation and itself while keeping these diagonal entries on a similar scale to the rest of the proximities. Additionally, RF-GAP proximities were designed to be capable of reconstructing the random forest predictions by using them as weights. Thus, they are not symmetric, although it was demonstrated in [14] that RF-GAP proximities are asymptotically symmetric as the number of trees increases. In applications requiring symmetric affinities, this can easily be fixed by symmetrizing the proximities. With these adjustments, we can use RF-GAP proximities as a supervised kernel to be used in manifold learning.

Diffusion-based information geometry

Early attempts at diffusion-based dimensionality reduction, such as Diffusion Maps [24] use a k-NN graph upon which a kernel function, such as a Gaussian kernel with a fixed bandwidth, is applied to represent observational similarities. The similarity matrix is row-normalized to form a stochastic matrix, P, known in this context as the diffusion operator. The diffusion operator represents the probabilities of all possible single-step transitions between observations. High affinities correspond to high transition probabilities. Through a powering process, P^t represents a random walk across the manifold structure where global relationships are learned. During this process, small probabilities are quickly reduced to zero, thus filtering or denoising the learned manifold. Finally, dimensionality reduction can be applied (e.g., eigendecomposition) to map to lower dimensions.

Although this diffusion process is intuitive and suitable for dimensionality reduction, some weaknesses need to be addressed to create meaningful visualizations. First, a fixed kernel bandwidth for all points is often not appropriate as the data may not be sampled uniformly. Second, choosing a good time scale t for the diffusion process is difficult and largely overlooked in classical diffusion maps. A small value of t can lead to insufficient denoising and an overemphasis on the local structure while a large value of t can lead to oversmoothing and an overemphasis on the global structure. Third, the eigendecomposition in Diffusion Maps often places the information learned in P^t into different and higher dimensions [1, 81], which is not amenable for visualization.

Instead of applying a kernel function to Euclidean distances, we encode local similarities using RF-GAP proximities. These locally-adaptive similarities contain local to long-range dependencies and overcome the weakness of a fixed-bandwidth kernel function. However, if the data has many disjoint clusters or clusters with limited connections, global relationships may not be properly reflected in the transition process. For example, the transition process may favor a single cluster with only transition inlets but no outlets. We counter this scenario with the introduction of a damping factor, $\beta \in (0, 1]$, inspired by Google's PageRank algorithm which is used to rate website importance [15].

The PageRank damping factor was inspired to overcome the problems of "traps" and "dead ends", that is, sets of websites from which all outward links point towards the same group of websites or towards a single external webpage. Analogously, we counter poor exit probabilities by redefining the diffusion operator P as $P_\beta = \beta P + \frac{(1-\beta)}{N}\mathbf{1}\mathbf{1}^T$, where $\mathbf{1}$ is a vector of ones of length N and β is typically set between 0.8 and 0.9. Here, the transition probabilities allow for random jumps creating small exit probabilities from isolated clusters or dead-ends. This has the effect of maintaining global structure in the presence of disjoint clusters.

To properly select t, we follow the inspiration of PHATE [1], which uses von Neumann Entropy (VNE) of the diffused operator to provide a good choice of t for visualization. The VNE is a soft proxy for the number of significant eigenvalues of P_β^t. Typically, t is chosen to be around the transition from rapid to slow decay in the VNE as this is considered to be a point in the diffusion process where noise has been eliminated and oversmoothing begins [1].

To overcome the weaknesses stemming from the eigendecomposition in standard Diffusion Maps, we apply an information distance [16] to the powered diffusion operator P_β^t to create a form of potential distance [1]. The potential distance is sensitive to differences in both the tails and the more dense regions of the diffused probabilities, resulting in a distance that preserves both local and global relationships. These distances are then embedded into low dimensions using metric MDS. This embeds the information in low dimensions for better visualization. The full RF-PHATE algorithm is given in Algorithm 1.

Algorithm 1 The RF-PHATE algorithm

Input: Training dataset $\mathcal{M} = \{(X, Y)\} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots (\mathbf{x}_N, y_N)\}$
Output: The RF-PHATE embedding, G

1: Train the random forest (RF) on $\mathcal{M}$
2: Generate RF-GAP [14] proximities ($\mathcal{P}$) from RF
3: Row-normalize $\mathcal{P}$ to form the initial diffusion operator P
4: Apply damping [15] to form $P_\beta = \frac{\beta}{N}P + \frac{(1-\beta)}{N}\mathbf{1}\mathbf{1}^T$
5: Perform t steps as selected using VNE to form P_β^t
6: Calculate the potential distance [1, 16] D_t from P_β^t by applying the log transform of P_β^t and then computing pairwise distances between columns
7: Form the embedding G by applying MDS to D_t

Embedding noisy data

Expert-generated data labels assist the learner in knowing what types of information to seek, but not how to perform the learning. In some cases, relationships between the inherent data structure and the data labels may not be trivial. In others, excessive noise in the data can prevent the learning of meaningful structures without the assistance of data labels. In Fig. 3, we showed

how RF-PHATE is able to ignore irrelevant (i.e., noise) features and capture the true structure of the artificial tree data, while other methods fail in the presence of noisy features.

Here we provide an additional example of a dataset with well-known structures to which we intentionally add noise features. With these examples, we demonstrate the ability of RF-PHATE to uncover a meaningful data structure relative to the labels in the presence of noise. This example presents a noised version of the simple and well-known Iris dataset [82]. In this example, we combined the original dataset (with features rescaled from 0 to 1) with an additional 500 uniformly sampled 0-1 noise features. The results are shown in Fig. S1. In this scenario, unsupervised approaches are incapable of determining a meaningful embedding. Each of the compared unsupervised methods generated a seemingly random cloud of points. Class-conditional supervised methods all tend to completely separate points of different classes, which is contrary to the ground-truth structure, in which there is a small amount of overlap between the Versicolor and Virginia species. In contrast, RF-PHATE preserves this structure even in the presence of noisy features.

Quantifying supervised embedding fit

Traditional tools for assessing DR quality are insufficient in a supervised setting and are not necessarily intended to provide a quantitative assessment of the embedding's useability for visualization. For example, MDS [17] seeks to minimize stress, the normalized squared differences between distances in the original space and the embedding space, but low stress does not necessarily correlate with the visual quality of the embedding [83]. To quantify the quality of supervised embeddings, one approach is to record the accuracy of a k-NN predictor trained on the embedding. However, this favors methods that emphasize class separation rather than structure preservation. Furthermore, perfect class separation can be attained by supervised manifold learning approaches using class-conditional distances (Fig. S1 and Fig. 3, for example).

To be useful in exploratory analysis, the goal of our method is not explicitly to separate classes but to uncover the intrinsic data structure relative to the supervised task. We thus use three different measures to quantify an embedding's goodness of fit. Low-dimensional Group Separation (LGS) assesses the extent to which the embedding is consistent with the classification problem; this method uses the silhouette score of the embedded points using data labels in lieu of its traditional use with cluster labels. LGS has also been adapted to time-series data as described in Methods. The Local Structure Preservation score (LSP) correlates the local feature importance of the supervised task with the feature importance for the embedding construction. This measures the extent to which the embedding can be recovered using the data features pertinent to the classification or regression problem at a local level. Global Structure Preservation (GSP) correlates the pairwise distances of each univariate feature with the pairwise distances of the embedding. The degree to

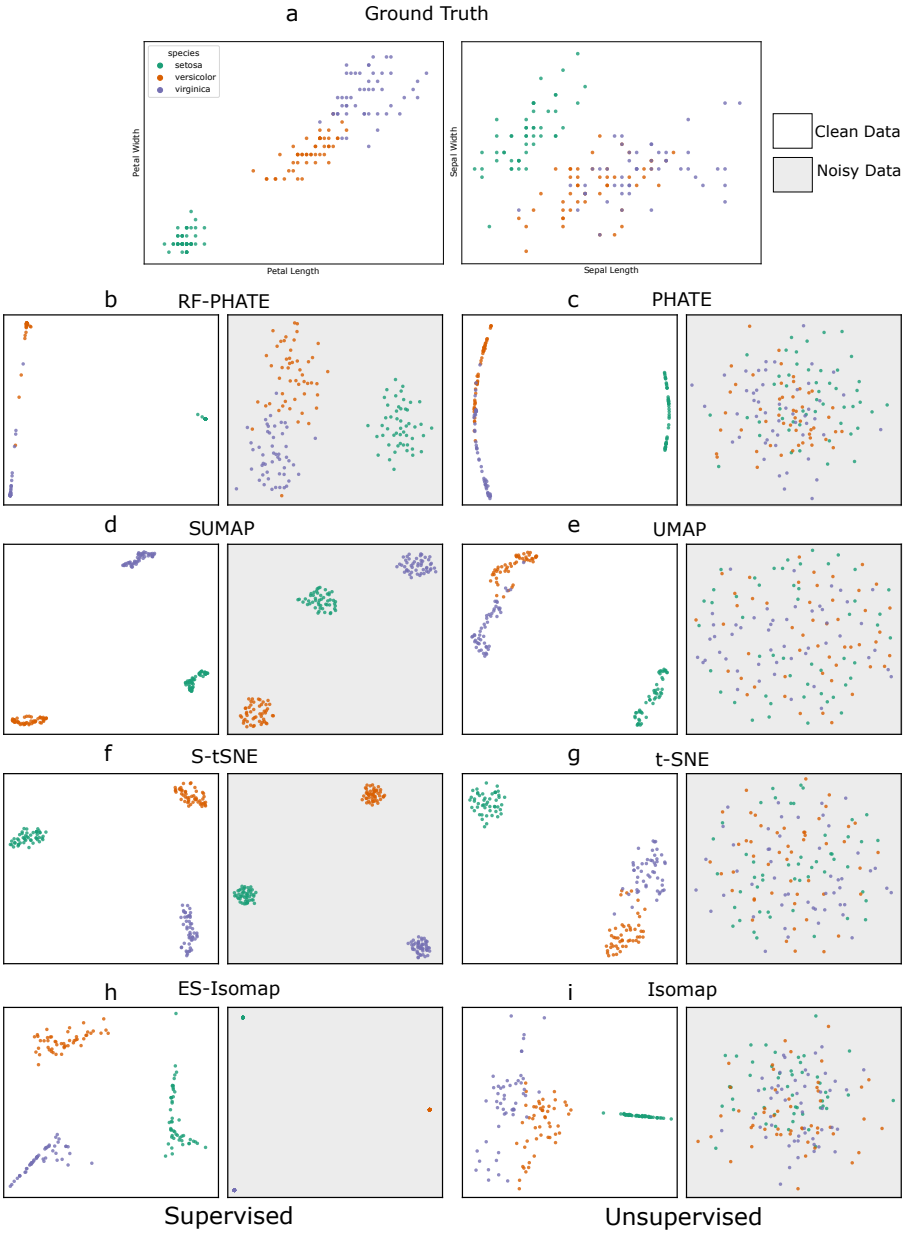

Fig. S1: Here we added 500 noise variables (each uniformly distributed from 0 to 1) to the well-known Iris dataset [82]. **a**, presents two of the pairwise variable plots, each of which demonstrates the separability of the Setosa class from the other two, while simultaneously showing that the other two classes have some overlap. **c, e, g, i**, unsupervised methods can generally reconstruct a meaningful representation in the absence of noise (white background) but are incapable of showing these relationships with noise (gray background). **d, f, h**, class-conditional supervised methods tend to completely separate the classes, which is not consistent with the known data structure. On the other hand, RF-PHATE, **b** preserves the inherent structure despite the noise.

which these correlations relate to the global feature importance of the original prediction problem determines the extent to which the global structure is maintained in low dimensions relative to the prediction problem. Further details on these metrics are provided below:

(a) **LGS**: the **L**ow-dimensional **G**roup **S**eparation score is defined as the overall silhouette score of the grouping in the embedding space determined by the label assignment $\mathbf{Y}$:

$$\text{LGS} = \frac{1}{N}\sum_{i=1}^{N} s(i),$$

where
- $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$ is the silhouette score of the ith embedded point,
- $a(i)$ is the average distance between embedded point i and all other embedded points belonging to the same class,
- $b(i)$ is the average distance between embedded point i and all other embedded points in the other class.

The overall silhouette score measures clustering quality [84] which is commonly used in the exploratory analysis of healthcare data [85–87]. It also has the convenience of ranging between -1 and 1, where values near 1 indicate that objects are very well matched to their own group and poorly matched to neighboring groups. Values near 0 indicate that objects are equally well matched to their own group and to neighboring groups, and values near -1 indicate that objects are poorly matched to their own group and well matched to neighboring clusters.

Instead of evaluating a clustering algorithm, we use the overall silhouette score to evaluate a DR method in separating the existing groups in the embedding space to provide the user with a visualization that clearly highlights class relationships. LGS does not have any tunable parameters.

(b) **LSP**: the **L**ocal **S**tructure **P**reservation score [88] is used to determine how important features are used in the embedding construction. In a supervised context, features important for classification or regression should dominate the low-dimensional local structure of the embedding. Thus, we produce two rankings of the input features with respect to (1) their importance to the classification or regression task and (2) their importance to the local embedding structure generation. A good embedding should preserve the respective feature rankings. LSP uses a trained prediction model (typically k-NN) to assess feature importance scores using randomized permutation importance, $\mathcal{C}$, for the original prediction task. A similar model is trained to regress onto the embedding and feature importance scores, $\mathcal{L}$, for this task are also assessed. LSP is calculated as the correlation between these two sets of feature importance scores, that is,

$$\text{LSP} = \rho(\mathcal{C}, \mathcal{L}).$$

This correlation score also ranges from -1 to 1, like LGS, allowing us to use them as a benchmark for a direct comparison between embedding algorithms.

(c) **GSP**: **G**lobal **S**tructure **P**reservation score is computed similarly to LSP. A model's feature importance scores, $\mathcal{C}$, is correlated with another set of feature importances $\mathcal{G}$ related to the global embedding structure generation.

Becht et al. [89] demonstrated that UMAP preserves global structure better than t-SNE by correlating pairwise distances in the original space with pairwise distances in the embedding space, where a high correlation is assumed to indicate good preservation of the original data structure from a macroscopic point-of-view. Here, we modify their approach to incorporate our hierarchical structure preservation with respect to feature importance for the prediction task. Instead of directly computing pairwise distances in the original, high-dimensional space, we compute pairwise distances within each univariate feature space and correlate each of the corresponding condensed pairwise distances $\mathrm{D}^j \in \mathbb{R}^{\frac{N(N-1)}{2}}$ with the condensed pairwise distances $\tilde{\mathrm{D}} \in \mathbb{R}^{\frac{N(N-1)}{2}}$ between the embedded space $\tilde{\mathbf{X}}$. Thus, the quantity $\mathcal{G}^j = \rho(\mathrm{D}^j, \tilde{\mathrm{D}}) \in [-1, 1]$ measures the importance of feature j in the global structure generation, and the set of global feature importances is defined as $\mathcal{G} = \left\{\mathcal{G}^j : j = 1, \ldots, K\right\}$. The GSP score of a DR method is then defined as the following Spearman's rank correlation coefficient:

$$\mathrm{GSP} = \rho(\mathcal{C}, \mathcal{G}) \in [-1, 1].$$

Artificial tree construction

The artificial tree data used in Fig. 3 follows the construction provided in the original PHATE paper [1]. The tree is crafted as follows: The initial branch has 100 linearly-spaced points spanning four dimensions, with the rest of the dimensions set to zero. The second branch's 100 points are constant in the first four dimensions, mirroring the first branch's endpoint. The subsequent four dimensions progress linearly, while others are set to zero. Similarly, the third branch progresses in dimensions 9–12 instead of 5–8. Other branches follow suit but with varying lengths, adding 40 points at endpoints and branch points, and zero-mean Gaussian noise (s.d. 7). This models gene expression advancement along branches. Additional noise dimensions total the data's dimensionality to 60. Prior to visualization, all features were normalized to fall between 0 and 1.

In the extremely noisy case (Fig. 3 with gray backgrounds), we added 500 variables simulated from a uniform (0, 1) distribution.

EDSS historical context

Accurately classifying the course of MS into sub-types is very important for communication, prognostication, design and recruitment of clinical trials, and

treatment decision-making [90]. To this end, Lublin et al. [90] proposed a baseline classification of MS into four major forms: Clinically Isolated Syndrome (CIS), Relapsing-Remitting MS (RRMS), Secondary-Progressive MS (SPMS) and Primary-Progressive MS (PPMS). Two main descriptors underlie these phenotypes: disease activity, as evidenced by relapses and/or new activity on magnetic resonance imaging (MRI), and progression of disability.

However, retrospective evaluations of physicians are inherently subjective and limited in terms of processing information. Additionally, the course of MS is widely recognized as highly heterogeneous and difficult to predict on an individual basis over time [91]. As a result, historical values of the expanded disability status scale (EDSS), computed as a nonlinear combination of eight functional scores [25, 92], remain the gold standard to characterize MS progression and relapses [26].

While other available historical data, such as the eight underlying functional scores or imaging features, are often not explicitly used by physicians in the MS diagnosis, our study is motivated by the assumption that there is an implicit interest in integrating these features to unveil more specific sub-types of MS. For instance, according to Lublin et al. [90], MRI data can help in determining active and inactive sub-types of RRMS and SPMS through contrast-enhancing lesions or new or unequivocally enlarging T2 lesions assessed at least annually with the aim of potentially switching to a high efficacy disease-modifying therapy and reduce the risk of relapse in case of high disease activity [93]. This also led researchers to classify MS patients using features that better reflect the pathophysiology of the disease, such as MRI lesions [94], underlying biological mechanisms, or the continuous aspect of disease progression [95]. Still, there is a lack of consensus regarding the clinical significance of MRI data as well as the clinically important changes that may result from changes in MRI data [96]. This means that, while additional features may be useful, EDSS remains the main differentiating factor between MS sub-types at this point.

MS data collection and description

The MS data was obtained from patients diagnosed with clinical and neuropathological MS diagnosis according to the revised 2010 McDonald's criteria [45]. Disability was scored by a qualified neurologist using the EDSS [25]. More specifically, our raw longitudinal database consists of two distinct MS follow-up data sets. The clinical set consists of 16,371 clinical visits with 636 distinct patients from which 9 features (Table 4) quantifying disability in MS were collected from a clinical examination. The radiological set consists of 6,327 radiological visits across 1,504 patients from which 6 radiological features related to the chronic inflammatory nature of the disease were recorded from MRI sessions. For a more detailed description of the available input features and their calculation, refer to Tables 4 and 5 as well as the guidelines of the reference papers [25, 92]. Note that both databases can be expressed as a

Table 4: Description of the 9 longitudinal input features recorded for each patient in the clinical database. Higher values indicate worse patient conditions.

Feature name	Range	Increment	Evaluated condition
EDSS	[0, 10]	1/2 unit	Nonlinear combination of the 8 following functional scores
Pyramidal Score	[0, 5]	unit	Deep tendon and cutaneous reflexes, limb and gait spasticity, limb strength,
Cerebellar Score	[0, 5]	unit	Tremor, coordination, ataxia, balance
Brainstem Score	[0, 5]	unit	Extraocular movements, nystagamus, trigeminal function, facial movements, hearing, dysarthria, dysphagia, other cranial nerve function
Sensory Score	[0, 6]	unit	Light touch and pain, vibration sense, position sense
Bladder/Bowel Score	[0, 6]	unit	Urinary hesitancy, retention, urgency, incontinence, bowel function
Visual Score	[0, 6]	unit	Visual acuity (Snellen Chart), disc (optic nerve) pallor, visual field
Mental Score	[0, 5]	unit	Depression, euphoria, mentation status, fatigue
Ambulation Score	[0, 9]	unit	Maximal walking distance (up to 500 meters), and type of assistance required

Table 5: Description of the 6 longitudinal input features recorded for each patient in the radiological database.

Feature name	Domain	Description
Nb Lesions T2	Any integer	Lesion count using the T2 imaging method [97]
Nb New or Enlarging Lesions	Any integer	New/enlarging lesion count with reference to a baseline scan
McDonald Lesions T2	0, 1–2, 3–8, 9+	T2 lesion count with categories according to the McDonald criteria
McDonald Lesions Infratentorial	0, 1+	McDonald lesion count located in the infratentorial brain region
McDonald Lesions Juxtacortical	0, 1+	McDonald lesion count located in the juxtacortical brain region
McDonald Lesions Periventricular	0, 1–2, 3+	McDonald lesion count located in the periventricular brain region

set of time series from each patient containing the history of their clinical and radiological visits.

At the end of the follow-up study, each patient was assigned one of the four typical MS types: Clinically Isolated Syndrome (CIS), Primary Progressive MS (PPMS), Relapsing-Remitting MS (RRMS), or Secondary Progressive MS (SPMS). According to the official classification provided by Lublin et al. [90], RRMS is a type of MS with relapses, defined as bouts of clinical deterioration (appearance of new symptoms or worsening) for a period of 24 hours or more—in the absence of an infection or fever. Relapses in RRMS are followed by a variable degree of recovery, without overall disease worsening outside of relapses. This form of MS can (but not always) progress to SPMS during which patients experience insidious accumulation of neurological disability outside of relapses. In that case, the SPMS type is assigned to the patient. Since CIS and

PPMS are less common forms of MS [90], we will be focusing on RRMS and SPMS throughout this study.

In addition, the collected data offers contextual information that can be valuable for correlating with our main longitudinal features and RRMS/SPMS classes. The worsening $\mathcal{W}_i$ of an MS patient after a period of i years is a clinically motivated binary (0/1) worsening score, defined from EDSS as

$$\mathcal{W}_i = \begin{cases} 1, & \text{if } \Delta^i_{EDSS} \geq 1 \quad \text{and } EDSS(0) \leq 5.5 \\ 1, & \text{if } \Delta^i_{EDSS} \geq 0.5 \text{ and } EDSS(0) > 5.5 \\ 0, & \text{otherwise} \end{cases}$$

where $\Delta^i_{EDSS} = EDSS(i) - EDSS(0)$ denotes the difference between the EDSS after i years and the EDSS at the beginning of the disease course. In the database, the worsening is given for $i = 1, 2, 5, 10$ years. The annualized relapse rate (ARR) is computed for each patient as the total number of relapses divided by the number of years since disease onset. In many clinical trials, it is a primary endpoint, because relapses are meaningful as they reflect how individuals experience fluctuations in symptoms that are central to relapsing MS.

Lastly, our data set contains typical demographic characteristics and some treatment details, including the duration of MS-specific medication usage, along with the reasons for discontinuation. Table 6 presents a summary of some patients' characteristics within the two RRMS and SPMS classes after data imputation and processing. As expected, the SPMS group exhibits a greater EDSS on average, as well as higher worsening rates. While the rate of relapses generally declines in SPMS patients, the similar ARR observed in Table 6 for both groups primarily results from the fact that SPMS patients inevitably experience RRMS before progressing to SPMS. This means some patients might have transitioned at a later stage or encountered high ARR during their RRMS phase, thereby contributing to an elevated ARR in the long run.

We also include the Spearman correlation matrix of the input features in Fig. S2.

MS data imputation and processing

We impute missing disability and MRI feature values using linear interpolation. For ordered categorical features, we first associate each category with an integer {0, 1, 2, ...} and round the imputed value to the nearest integer. To form a single database describing patients' follow-up in terms of both disability and MRI features, we linked clinical visits to MRI visits that happened no more than six months apart. Clinical visits without MRI data falling into that time period are simply removed. Afterward, we re-sample the combined dataset on a 6-month frequency and normalize the time between patients by setting the origin to the date of MS disease onset (ddo). In other words, the

Table 6: Summary characteristics of the RRMS and SPMS study cohorts

Feature	RRMS	SPMS
Number of		
patients	157	29
visits	4742	636
Male/Female ratio	1 : 4.6	1 : 2.2
Age at onset, mean ± SD	28.7 ± 9.0	31.0 ± 8.9
Disease duration, mean ± SD	23.0 ± 5.8	24.3 ± 6.7
EDSS (visits)		
Mean ± SD	1.3 ± 1.1	3.3 ± 1.7
Median	1.1	3.1
Worsening (% patients)		
1 year	9.6%	11.8%
2 years	13.1%	17.6%
5 years	20.7%	52.6%
10 years	28.2%	66.7%
McDonald Lesions (% visits)		
≥ 9	74.9%	69.8%
0 to 8	25.1%	30.2%
ARR (at last follow-up)		
Mean ± SD	0.30 ± 0.21	0.31 ± 0.29
Median	0.25	0.26

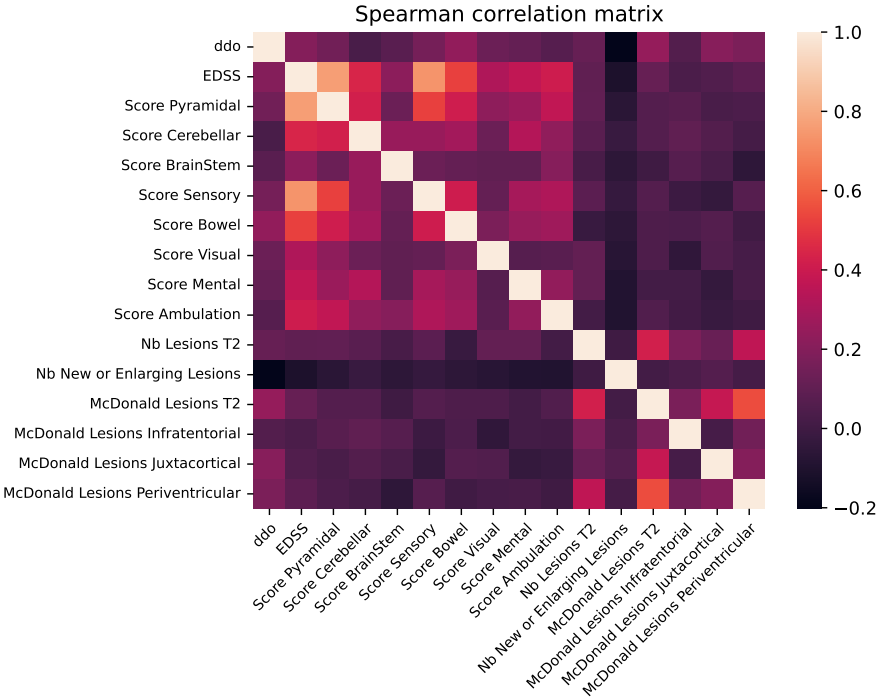

Fig. S2: Heatmap of the Spearman correlations between input features.

time (ddo) of a patient's visit corresponds to the number of days between the date of visit and the date of MS disease onset.

Since none of the DR methods presented in this work directly receive time-series objects as input, we also consider the normalized time component as an input feature to take the temporal structure of the input data into account, giving a total of $K = 16$ input features for each patient. Finally, we standardize features and remove patients without the available target of interest. That is, we removed the seven patients diagnosed with Clinically Isolated Syndrome (CIS) or Primary Progressive MS (PPMS), which are less common forms of MS. This pre-processing results in $M = 5{,}378$ visits, including both clinical and radiological features, spread across $N = 186$ patients. We refer to the combined and processed time-series data set as $\mathbf{X} = \{X_i\}_{i=1}^N$ such that $X_i \in \mathbb{R}^{T_i \times K}$ represents patient i's K-dimensional time series of length T_i (i.e., each row is a visit at some time point, and $\sum_{i=1}^N T_i = M$). We associate $\mathbf{X}$ with the binary targets $\mathbf{Y} \in \{0,1\}^N$ where 0 and 1 encode RRMS and SPMS labels, respectively. Conventionally since SPMS is a more advanced stage of MS which comes after RRMS, the positive class is set to SPMS.

MS time-series data formatting

For the purpose of visualization, we seek a d-dimensional representation of the K-dimensional time series in $\mathbf{X} = \{X_i\}_{i=1}^N$ such that $d << K$. Usually, this is performed by a DR algorithm, but most are built to transform static data, which is not the case here. To solve this, we build the flattened data matrix $\mathbf{X}_{flat} \in \mathbb{R}^{M \times K}$ which consists of the concatenation of all X_i. Each row of the matrix $\mathbf{X}_{flat}$ corresponds to a K-dimensional visit for any patient. Thus, $\mathbf{X}_{flat}$ can be seen as the phase-space trajectory for the time-varying system determined by the patient time series in $\mathbf{X}$. Its rows are also associated with $\mathbf{Y}_{flat} \in \{0,1\}^M$ which links each visit to the outcome of the corresponding patient. Then, $\mathbf{X}_{flat}$ is passed through a given DR algorithm, jointly with $\mathbf{Y}_{flat}$ when supervised, producing a d-dimensional representation of patient visits from which d-dimensional patient time series are re-constructed thanks to the original patients' identifiers and the timestamps of their visits.

Note that the above approach does not make explicit use of the longitudinal nature of the N observations. This is why we include time in the input feature set, and the longitudinal structure preservation of the original data set is taken into account in our evaluation framework. For the rest of our study, we refer to $\tilde{\mathbf{X}} = \{\tilde{X}_i\}_{i=1}^N$ ($\tilde{X}_i \in \mathbb{R}^{T_i \times d}$) as the resulting d-dimensional counterpart of the original patient time series $\mathbf{X} = \{X_i\}_{i=1}^N$ ($X_i \in \mathbb{R}^{T_i \times K}$).

Dynamic time warping

For time-series data, Euclidean distances are less meaningful and are thus replaced by dynamic time warping (DTW) [98, 99]. Given two time series Q and C, the DTW algorithm provides a dissimilarity measure $s_{DTW}(Q, C) \geq 0$ based on a non-linear matching of respective points, as opposed to the

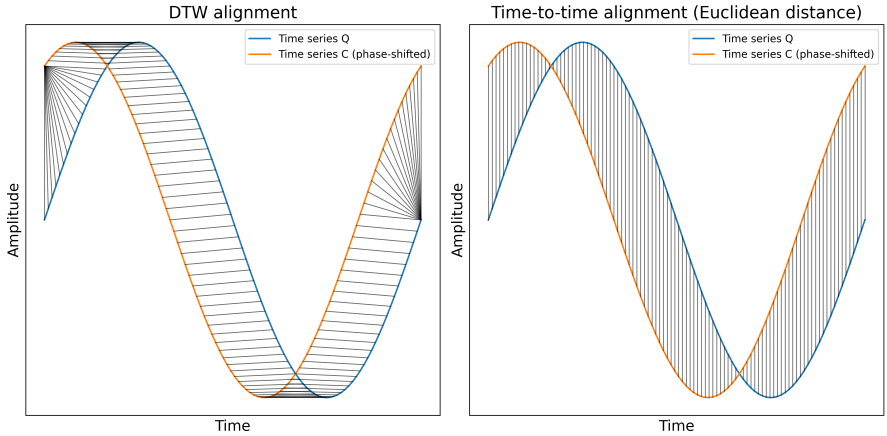

Fig. S3: Comparison of DTW (left) and Euclidean (right) alignments between two time series of similar shape but not aligned in the time axis. The Euclidean distance sums the time-to-time distances between matching points, while DTW allows time distortions. With its restricted alignment, the Euclidean distance misses the similar shape shared by those two time series.

Euclidean distance matching that calculates straight-line (geometric) distances (Fig. S3). More specifically, if n and m denote the lengths of Q and C, respectively, we write Q and C as K-dimensional ordered sequences,

$$Q = q_1, q_2, \ldots, q_i, \ldots, q_n, \qquad q_i \in \mathbb{R}^K,$$
$$C = c_1, c_2, \ldots, c_j, \ldots, c_m, \qquad c_j \in \mathbb{R}^K,$$

and construct an n-by-m matrix M such that each cell (denoted by their respective indices (i, j) for simplicity) contains the distance between the pair of points (q_i, c_j), usually the squared Euclidean distance: $d(q_i, c_j) = \|q_i - c_j\|_2^2$. A warping path W is a set of distinct cells of the constructed matrix M that defines a mapping between Q and C. That is, $W = w_1, w_2, \ldots, w_l, \ldots, w_{\mathrm{L}_W}$ with $w_l \in \{1, \ldots, n\} \times \{1, \ldots, m\}$. Then, we introduce the three following conditions to ensure that a warping path captures a "correct" sequence of (q_i, c_j) with respect to time, otherwise the time aspect of a time-series comparison would be completely lost:

(i) boundary conditions: $w_1 = (1, 1)$ and $w_{\mathrm{L}_W} = (n, m)$, i.e., the warping path starts and ends in diagonally opposite corner cells of the matrix M. This prevents the occurrence of time warping being limited to a small section of the sequences;

(ii) continuity: given $w_l = (a, b)$, then $w_{l-1} = (a', b')$, where $a - a' \leq 1$ and $b - b' \leq 1$. This restricts the allowable steps in the warping path to adjacent cells of M (including diagonally adjacent cells), as to avoid "time jumps";

(iii) monotonicity: given $w_l = (a, b)$, then $w_{l-1} = (a', b')$, where $a - a' \geq 0$ and $b - b' \geq 0$. This ensures that the warping path does not go backwards in time.

Finally, if $\mathbf{W}$ is the set of all possible warping paths subject to (i), (ii) and (iii), the DTW distance between Q and C is obtained from $W \in \mathbf{W}$ minimizing the warping cost $\sqrt{\sum_{l=1}^{\mathrm{L}_W} w_l}$:

$$s_{DTW}(Q, C) = \min_{W \in \mathbf{W}} \left\{ \sqrt{\sum_{l=1}^{\mathrm{L}_W} w_l} \right\}.$$

Under the above constraints (i), (ii) and (iii), DTW remains symmetric [100], that is, $s_{DTW}(Q, C) = s_{DTW}(C, Q)$, so DTW provides an intuitive proximity between time-series objects and can be used similarly to Euclidean distance matching in a wide range of applications.

Quantifying supervised embedding fit with DTW (MS Case)

The evaluation of the low-dimensional embeddings is inherently different for time-series data. Data points can no longer be viewed as static, independent observations. We want our embedding to capture the structure of the time series both locally and globally for meaningful exploratory analysis. We thus adapt our evaluation metrics LGS, LSP, and GSP to suit this purpose as follows.

LGS: In the context of our MS study, for the majority of cases where i and j represent different patients ($i \neq j$), we have $|T_i - T_j| >> 0$ and patients' time series are very likely to be recorded within different time (days from disease onset, (ddo)) periods. Moreover, physicians are rather interested in global long-term patterns of the course which is not restricted to time-to-time matching. This is why we make use of DTW which is more flexible than the traditional time-to-time Euclidean distance since it allows two time series that are similar but locally out of phase to align in a non-linear manner [101]. This variation of LGS still has the convenience of ranging between -1 and 1 for easy interpretation.

LSP: The low-dimensional embedding should maintain the local variable importance structure for features useful for RRMS/SPMS classification. In our supervised context, features important to classify patients' time series as RRMS/SPMS should dominate the low-dimensional local structure of the embedded time series to be clinically insightful from a physician's point of view. As in the non-time-series case, we produce two rankings of the input features with respect to their importance to the classification task and their importance to the local embedding structure generation. Clinically, important features can be thought of as the ones used by physicians in the retrospective RRMS and

SPMS classification. However, our data also includes features whose importance is not *a priori* known, forcing us to objectively assess feature importance using a data-driven approach. To illustrate, we know that EDSS has been primarily used by physicians to diagnose the transition from RRMS to SPMS, but the impact of each of the 8 functional scores is hazy.

We numerically provide feature importance scores $\mathcal{C}$ related to the RRMS/SPMS time-series classification by considering one feature at a time. First, we generate K data sets $\{\mathbf{X}^j\}_{j=1}^K$ from $\mathbf{X}$ such that $\mathbf{X}^j = \{X_i^j\}_{i=1}^N$ ($X_i^j \in \mathbb{R}^{T_i \times 1}$) corresponds to the set of univariate patient time series determined by the jth recorded feature. Then, for each $\mathbf{X}^j$ and the corresponding patient labels $\mathbf{Y}$, we apply a leave-one-out cross-validation (LOOCV) to a DTW k-NN binary classifier such that the outputs are positive class probabilities. Since the numbers of RRMS and SPMS patients are respectively 157 and 29, the predicted probabilities are proportionally weighted to the inverse of the distances from the query points to mitigate class imbalance that inclines towards RRMS labels [102]. Finally, the importance score of feature j is computed as the area under the receiver operating characteristic curve (AUROC) based on the predicted class probabilities over the test folds of the LOOCV and their corresponding true classes. Thus, if $\mathcal{C}^j$ denotes the resulting importance score of feature j, then the set of feature importances for the RRMS/SPMS classification task is defined as $\mathcal{C} = \left\{\mathcal{C}^j : j = 1, \ldots, K\right\}$. Feature importances $\mathcal{L}$ related to the local embedding structure generation are obtained similarly, except that we regress on the embedded time series, and each $\mathcal{L}^j$ is determined by summing the DTW distances between the predicted time series and the true embedded time series across the test folds. As previously described, the local structure preservation score is defined as the correlation between the two sets of importance vectors, $\mathcal{C}$ and $\mathcal{L}$.

GSP: We compute pairwise distances within each of the K univariate time-series data sets $\{\mathbf{X}^j\}_{j=1}^K$ using DTW, and correlate each of the K corresponding condensed pairwise distances $\mathrm{D}^j \in \mathbb{R}^{\frac{N(N-1)}{2}}$ with the condensed pairwise distances (also using DTW) $\tilde{\mathrm{D}} \in \mathbb{R}^{\frac{N(N-1)}{2}}$ between the embedded patient time series $\tilde{\mathbf{X}}$. The score is then determined as previously described.

Quantitative Experiment Details

Here we provide additional details about the quantitative comparisons of RF-PHATE to other methods using our three metrics, LGS, LSP, GSP, in Tables 1 and 2.

For LSP and GSP, we used an unweighted k-NN model, setting $k = 10$ to determine feature importance using a permutation approach. The metric used to assess the goodness of fit of the k-NN models was overall accuracy for the label classification task, and R^2 when regressing onto the embedding. 30 repetitions were used when extracting feature importance for stability. All other parameters were left as defaults. For these experiments, Spearman correlation was used for LSP and GSP, though Pearson's correlation coefficient

Table 7: This table provides a brief summary of all datasets used to quantify the embedding fit using LGS, LSP, and GSP in Table 1. For each dataset, observations with missing values were removed.

Data	Observations	Features	Number of Classes
Audiology	190	67	23
Balance Scale	625	4	3
Breast Cancer	699	16	2
Car	1728	6	4
Chess	3196	36	2
CRX	664	14	2
Diabetes	678	8	2
E. Coli	336	8	8
Flare1	323	10	3
Glass	214	9	6
Heart Disease	303	13	5
Heart Failure	299	10	2
Hepatitis	138	15	2
Hill Valley	606	100	2
Ionosphere	351	34	2
Iris	150	4	3
Lymphography	148	18	4
Optical Digits	3823	64	10
Parkinson's	195	22	2
Seeds	199	7	3
Segmentation	211	19	8
Tic-Tac-Toe	958	9	2
Titanic	712	7	2
Artificial Tree	1440	60	10
Waveform	5000	40	3
Wine	178	13	3
Zoo	101	17	100

can also be used. We compared the performance of RF-PHATE against 23 other DR methods (10 unsupervised, 9 supervised, and 4 random forest-based methods [103]).

We compared results across 28 publicly available datasets, with some details found in Table 7. No specific criteria were used to select the datasets; we tried to provide datasets with a variety of sizes and numbers of features. Across all datasets, the smallest number of observations was 98, and the largest was 5000. The number of classes ranged from 2 to 100. Due to the number of repetitions and number of embedding methods used we did not run this experiment with larger datasets. The number of features ranged from 4 to 100. Any observations with missing values were removed prior to embedding. The number of remaining observations is reflected in the table. All continuous features were normalized to range from 0 to 1. Categorical features were encoded as integers.

For the MS quantitative analysis, we set the hyper-parameter k of our k-NN classifier/regressor to $k = \lfloor \sqrt{N} \rfloor = 13$. The metric used was DTW rather than Euclidean. Otherwise, default parameters were used. Classification importances $\mathcal{C}$ are given in Table 8. The experiment was run 20 times and

Table 8: Feature importances computed as the AUROC from LOOCV test predictions using a DTW k-NN classifier on univariate time series determined by each feature (see Methods), in decreasing order (higher is better). As expected, EDSS is the best predictor in the RRMS/SPMS classification.

Feature name	AUROC
EDSS	0.918
Score Pyramidal	0.877
Score Ambulation	0.829
Score Cerebellar	0.717
Score Sensory	0.706
Score Bowel	0.617
Score Visual	0.606
Score Mental	0.590
Time (ddo)	0.569
Nb Lesions T2	0.568
McDonald Lesions Juxtacortical	0.543
Score Brainstem	0.541
McDonald Lesions T2	0.511
McDonald Lesions Infratentorial	0.501
Nb New or Enlarging Lesions	0.495
McDonald Lesions Periventricular	0.394

LGS, LSP, and GSP scores were each averaged across repetitions for each embedding. To qualitatively validate the quantitative comparison of the MS dataset, we also provide the 24 corresponding 3D plots of our MS database along with the ground truth univariate plot of historical EDSS in Fig. S4.

RRMS Cluster Details

Fig. 4a represents low-dimensional RF-PHATE patient trajectories with lines colored by their respective class labels. Given the high variability observed in patient courses within the RRMS group in Fig. 4a, we sought to determine more stable and typical groups of trajectories underlying RRMS by applying k-medoids clustering with $k = 2$ to the RRMS group. We used DTW as a proximity measure. Fig. 4b illustrates low-dimensional patient trajectories, with the RRMS patient trajectories colored based on the output of the clustering algorithm applied to the RF-PHATE embedding of the RRMS group of trajectories.

To facilitate a broader exploratory analysis that can be easily generalized, we chose $k = 2$ instead of higher values to maintain larger cluster sizes, resulting in 105 patients/3018 visits for the dark blue cluster, and 52 patients/1724 visits for the light blue cluster. To measure a level of spatial separation, we computed the silhouette score between these two clusters, which yielded a value of 0.59.

Raman data collection and processing

The A549 cells were purchased from ATCC (Manassas, VA, USA) and cultured in an F-12k medium containing 10% fetal bovine serum (Thermo Fisher

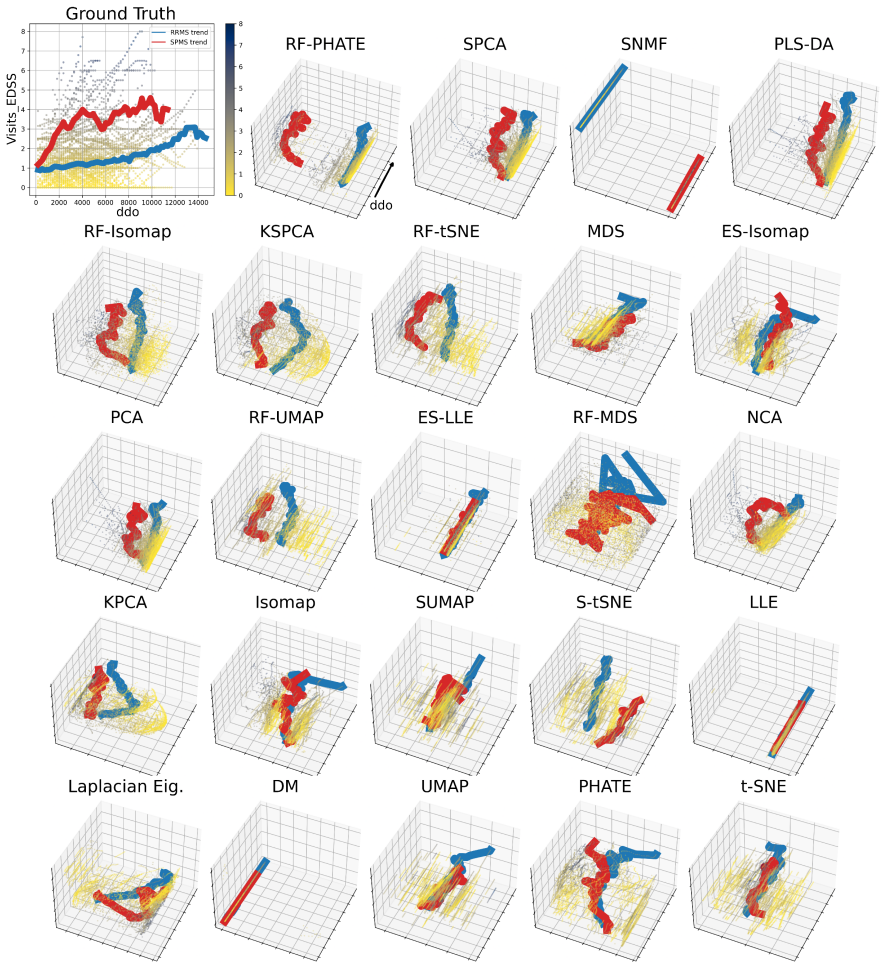

Fig. S4: Three-dimensional visualizations of the MS time-series database using RF-PHATE and 23 other embedding methods, sorted from left to right by their average rank in Table 2, along with the ground-truth EDSS patient trajectories (top-left plot). The three-dimensional plots have two axes for embedding dimensions and one for time, in days since MS onset (ddo). Each point represents a patient visit, colored by EDSS. Blue and red lines indicate the average trajectories of RRMS and SPMS groups, respectively. The plots align with our findings in Table 2. RF-PHATE remains the top choice for balancing class separability and structure preservation, while other embeddings perform progressively worse in both aspects as we move down the ranking.

Scientific, Waltham, MA, USA) at 37°C with 5% CO_2 in a humidified atmosphere. DEPs (4mg/ml) were prepared with a culture medium, vortexed for 10 seconds, and then sonicated for 20 minutes at room temperature. The A549 cells underwent pretreatment in a plain medium, a medium containing resveratrol (RES), or one containing Cannabidiol (CBD) at 10 µM for 24 hours.

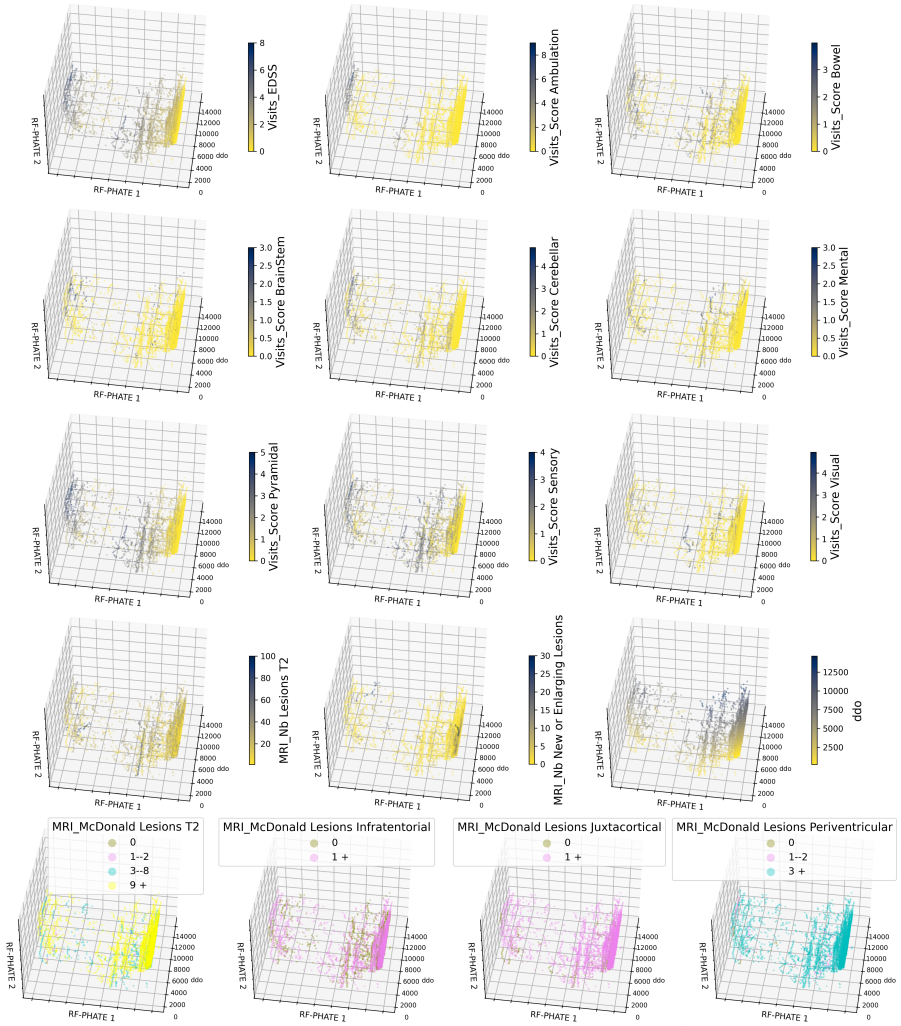

Fig. S5: Three-dimensional RF-PHATE plots of MS patient visits colored by each of the 16 input features. The embedded patient visits within the RRMS group (right cluster) exhibit lower values compared to those in the SPMS group (left cluster) across nearly all clinical indicators. This observation is in line with the well-established disparities between RRMS and SPMS in terms of clinical worsening (domain knowledge retrieval). Simultaneously, we observe a substantial number of visits in the bottom-left region of the RRMS group with scores similar to those in the SPMS group. Still, some aspects, such as high ambulation scores, remain very specific to the SPMS group. These observations suggest the existence of a more severe subgroup of RRMS patients that was not explicitly labelled in our database (knowledge discovery). As expected, the plots related to MRI features do not highlight distinct patterns because the MRI features included in our input feature set hold limited importance in the RRMS/SPMS classification (Table 8).

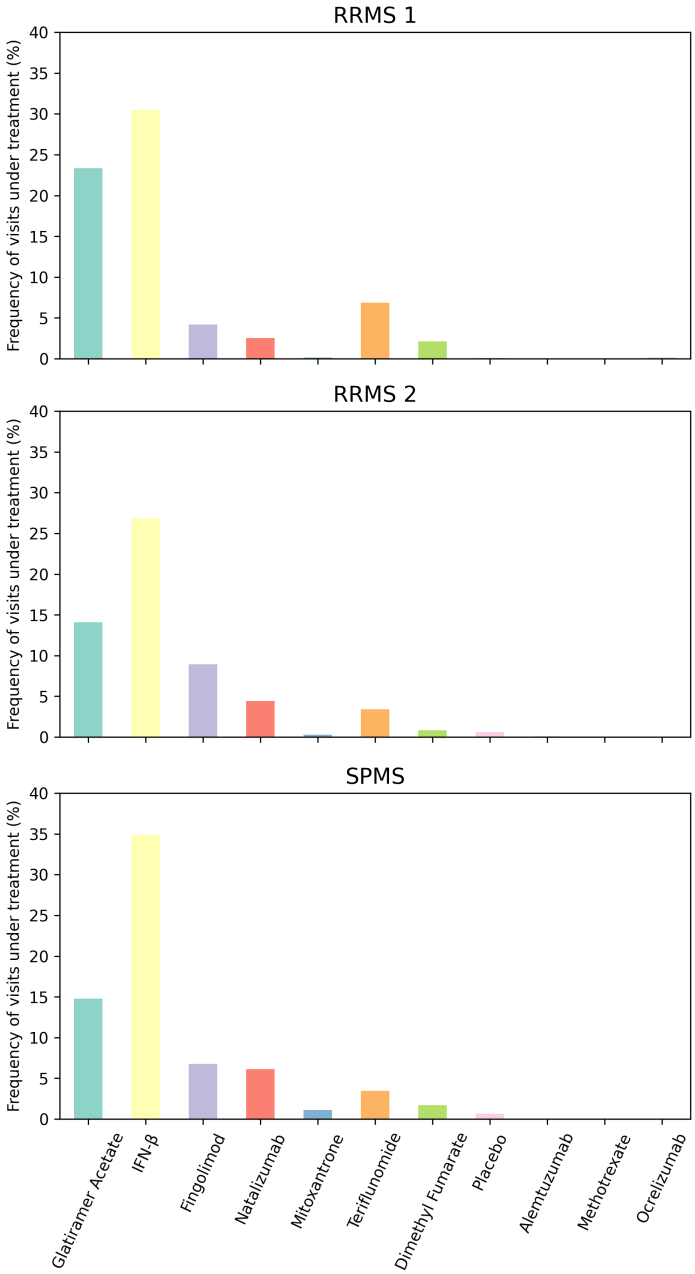

Fig. S6: Frequency of patient visits under treatment with various MS-specific medications within the RRMS 1 (top), RRMS 2 (middle), and SPMS (bottom) clusters. Surprisingly, patients within the three clusters have been prescribed similar medications despite their distinct MS trajectories (Fig. 4b). This suggests that an incorrect treatment strategy may have been employed, as RRMS 2 and SPMS patients might have benefited from targeted treatments earlier to improve their outcome.

Subsequently, cells were treated with a DEP solution at 0, 10, 25, 50, and 100 µg/ml for 24 hours.

The Raman spectra were measured by a Renishaw inVia Raman spectrometer (controlled by WiRE 3.4 software, Renishaw, UK) connected to a Leica microscope (Leica DMLM, Leica Microsystems, Buffalo Grove, IL, USA) equipped with a 785 nm near-infrared (IR) laser which was focused through a 63 ×NA = 0.90 water immersion objective (Leica Microsystems, USA). The standard calibration peak for the spectrometer with silicon mode at a static spectrum was 520.5 ± 0.1 cm^{-1}. A549 cells were cultured on MgF_2 (United Crystals Co., Port Washington, NY, USA) and imaged in Earle's balanced salt solution (EBSS). Raman spectra between 600 and 1800 cm^{-1} wavenumbers were recorded for 1 accumulation of 10-second laser exposure in static mode. Five points on each cell were randomly selected for measurements. At least 75 spectra (five points for each cell and fifteen cells) in each condition were collected for further analysis. The raw Raman spectra were baseline corrected by adaptive iteratively re-weighted penalized least squares. The Python library called BaselineRemoval (version 0.1.4) was used [104]. The spectrum was further normalized to [0,1].

Participants and samples – COVID-19 cohort

Plasma analytes were measured on prospectively enrolled COVID-19 individuals hospitalized between April 2020 and August 2021 with symptomatic infection with a positive SARS-CoV-2 nasopharyngeal swab (NSW) reverse-transcription polymerase chain reaction (RT-PCR) recruited into the Biobanque Québécoise de la COVID-19 (BQC19) [105]. Exclusion criteria were breakthrough or reinfection, plasma transfer therapy, or vaccination prior to infection. The study was approved by the respective IRBs (multicentric protocol: MP-02-2020-8929) and written informed consent was obtained from all participants or their legal guardians before enrollment. The research adhered to the standards indicated by the Declaration of Helsinki. COVID-19 hospitalized patients were stratified based on the severity of respiratory support at 11 days after symptom onset (DSO11): critical patients required mechanical ventilation (noninvasive ventilation, endotracheal intubation, extracorporeal membrane oxygenation), and non-critical patients required oxygen supplementation by nasal cannula or no supplemental oxygen. All-cause mortality was tracked up to 60 days after symptom onset (DSO60).

Quantification of plasma analytes – COVID-19 cohort

Plasma vRNA, cytokine, and RBD-specific antibody concentrations were measured as previously described [57] at DSO11 (+/- 4 days). Briefly, for SARS-CoV-2 viral RNA (vRNA), total RNA was extracted from 230 uL of plasma, and real-time PCR was performed using a N-specific primer of the SARS-CoV-2 genome. Purified RNA N transcripts (1328 bp) were quantified

by Nanodrop (Thermo), and RNA copy numbers were calculated using ENDMEMO. Cytokines were captured by magnetic beads array using a customized Human Luminex Discovery Assay (LXSAHM-26, R&D Systems) and acquired on the MagPix (Luminex). Only analytes significantly associated with a fatal outcome ($p<0.01$) in our previous work [58] were retained: TNFa, CXCL13, IL-6, IL-23, CXCL8, angiopoietin-2, RAGE, and Surfactant Protein D. RBD-specific IgG, IgM, and IgA were quantified using an in-house SARS-CoV-2 RBD ELISA assay, as described elsewhere [106]. Plasma from pre-pandemic uninfected donors was used as a negative control. The seropositivity threshold was established using the following formula: mean of all COVID-19 negative plasma + (3 standard deviations of the mean of all COVID-19 negative plasma).

COVID-19 longitudinal data details

In this section, we provide additional details on the patient time series extracted from COVID-19 data [57]. The initial database contains 604 plasma samples from 304 patients. We followed the same data imputation and formatting as in our MS case study (see Methods). Each patient time series consists of 14 immunological features recorded at different time points up to DSO30. We exclude five patients because one or more features are missing in their records. We also added the time component (in days since disease onset) to the existing feature set, resulting in 1,425 plasma samples from 299 patients associated with 15 features. The label of interest is the all-cause mortality (Yes/No) of any patient 60 days after enrolment. The positive group, which includes deceased patients, consists of 40 patients and 256 samples, while the negative group, comprising non-deceased patients, consists of 259 patients and 1,169 samples. The data processing is otherwise identical to the COVID-19 cross-sectional data described in Methods.

Data-processing - COVID-19 cohort

As in [57], we retain only plasma samples at DSO11 (+/- 4 days) and if a patient has multiple samples in the considered time frame, we include only the one closest to DSO11. The final input consists of 242 plasma samples.

We use the same data processing as in [57] to run RF-PHATE, PHATE, and k-means. All 14 input features are log-scaled and standardized (0 mean, variance 1). For RF-PHATE, we set the diffusion parameter t to 12 and otherwise use the default parameters of the random forest estimator of scikit-learn v1.1.1.

Supplementary information.

Acknowledgments.

Declarations

- Ethics approval. MS data was collected from MS patients with full ethical approval (BH07.001, Nagano 20.332-YP).
- Sacha Morin funding : Natural Sciences and Engineering Research Council of Canada (NSERC) PGS D Scholarship; Fonds de Recherche du Québec Nature et technologies (FRQNT) PhD Scholarship.
- Daniel E. Kaufmann funding : American Foundation for AIDS Research (amfAR) grant 110068-68-RGCV; Canada's COVID-19 Immunity Task Force (CITF), in collaboration with the Canadian Institutes of Health Research (CIHR) grant VR2-173203; CIHR grant # 178344.
- Elsa's funding : COVID-19 excellence scholarship from the Université de Montréal.

References

[1] Moon, K.R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D.B., Chen, W.S., Yim, K., Elzen, A.v.d., Hirn, M.J., Coifman, R.R., Ivanova, N.B., Wolf, G., Krishnaswamy, S.: Visualizing structure and transitions in high-dimensional biological data. Nat. Biotechnol. **37**(12), 1482–1492 (2019)

[2] van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)

[3] McInnes, L., Healy, J., Saul, N., Großberger, L.: Umap: Uniform manifold approximation and projection. Journal of Open Source Software **3**(29), 861 (2018). https://doi.org/10.21105/joss.00861

[4] Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000) https://arxiv.org/abs/https://www.science.org/doi/pdf/10.1126/science.290.5500.2323. https://doi.org/10.1126/science.290.5500.2323

[5] Tenenbaum, J.B., Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**(5500), 2319–2323 (2000)

[6] Belkin, M., Niyogi, P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. NIPS'01, pp. 585–591. MIT Press, Cambridge, MA, USA (2001)

[7] Cheng, J., Liu, H., Wang, F., Li, H., Zhu, C.: Silhouette analysis for human action recognition based on supervised temporal t-SNE and

incremental learning. IEEE Transactions on Image Processing **24**(10), 3203–3217 (2015). https://doi.org/10.1109/TIP.2015.2441634

[8] Hajderanj, L., Weheliye, I., Chen, D.: A new supervised T-SNE with dissimilarity measure for effective data visualization and classification. In: Proceedings of the 2019 8th International Conference on Software and Information Engineering. ICSIE '19, pp. 232–236. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3328833.3328853. https://doi.org/10.1145/3328833.3328853

[9] Ribeiro, B., Vieira, A., Carvalho das Neves, J.: Supervised isomap with dissimilarity measures in embedding learning. In: Ruiz-Shulcloper, J., Kropatsch, W.G. (eds.) Progress in Pattern Recognition, Image Analysis and Applications, pp. 389–396. Springer, Berlin, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85920-8_48

[10] Zhang, S.: Enhanced supervised locally linear embedding. Pattern Recognit. Lett **30**(13), 1208–1218 (2009)

[11] de Ridder, D., Kouropteva, O., Okun, O., Pietikäinen, M., Duin, R.P.W.: Supervised locally linear embedding. In: Kaynak, O., Alpaydin, E., Oja, E., Xu, L. (eds.) Artificial Neural Networks and Neural Information Processing — ICANN/ICONIP 2003, pp. 333–341. Springer, Berlin, Heidelberg (2003)

[12] Hajderanj, L., Chen, D., Weheliye, I.: The impact of supervised manifold learning on structure preserving and classification error: A theoretical study. IEEE Access **9**, 43909–43922 (2021). https://doi.org/10.1109/ACCESS.2021.3066259

[13] Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[14] Rhodes, J.S., Cutler, A., Moon, K.R.: Geometry- and accuracy-preserving random forest proximities. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–13 (2023). https://doi.org/10.1109/TPAMI.2023.3263774

[15] Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab (November 1999). Previous number = SIDL-WP-1999-0120. http://ilpubs.stanford.edu:8090/422/

[16] Duque, A.F., Wolf, G., Moon, K.R.: Visualizing high dimensional dynamical processes. In: 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing, Pittsburgh, PA, pp. 1–6 (2019). https://doi.org/10.1109/MLSP.2019.8918875. http://doi.org/10.

1109/MLSP.2019.8918875

[17] Kruskal, J.B., Wish, M.: Multidimensional Scaling. Multidimensional Scaling, vol. 11. Sage Publications, Newbury Park, California (1978)

[18] Cutler, A., Cutler, D.R., Stevens, J.R.: Random forests. In: Zhang, C., Ma, Y. (eds.) Ensemble Machine Learning: Methods and Applications, pp. 157–175. Springer, Boston, MA (2012). https://doi.org/10.1007/978-1-4419-9326-7_5. https://doi.org/10.1007/978-1-4419-9326-7_5

[19] Kuchroo, M., Huang, J., Wong, P., Grenier, J.-C., Shung, D., Tong, A., Lucas, C., Klein, J., Burkhardt, D.B., Gigante, S., Godavarthi, A., Rieck, B., Israelow, B., Simonov, M., Mao, T., Oh, J.E., Silva, J., Takahashi, T., Odio, C.D., Casanovas-Massana, A., Fournier, J., Obaid, A., Moore, A., Lu-Culligan, A., Nelson, A., Brito, A., Nunez, A., Martin, A., Wyllie, A.L., Watkins, A., Park, A., Venkataraman, A., Geng, B., Kalinich, C., Vogels, C.B.F., Harden, C., Todeasa, C., Jensen, C., Kim, D., McDonald, D., Shepard, D., Courchaine, E., White, E.B., Song, E., Silva, E., Kudo, E., DeIuliis, G., Wang, H., Rahming, H., Park, H.-J., Matos, I., Ott, I.M., Nouws, J., Valdez, J., Fauver, J., Lim, J., Rose, K.-A., Anastasio, K., Brower, K., Glick, L., Sharma, L., Sewanan, L., Knaggs, L., Minasyan, M., Batsu, M., Tokuyama, M., Muenker, M.C., Petrone, M., Kuang, M., Nakahata, M., Campbell, M., Linehan, M., Askenase, M.H., Smolgovsky, M., Grubaugh, N.D., Sonnert, N., Naushad, N., Vijayakumar, P., Lu, P., Earnest, R., Martinello, R., Herbst, R., Datta, R., Handoko, R., Bermejo, S., Lapidus, S., Prophet, S., Bickerton, S., Velazquez, S., Mohanty, S., Alpert, T., Rice, T., Schulz, W., Khoury-Hanold, W., Peng, X., Yang, Y., Cao, Y., Strong, Y., Farhadian, S., Dela Cruz, C.S., Ko, A.I., Hirn, M.J., Wilson, F.P., Hussin, J.G., Wolf, G., Iwasaki, A., Krishnaswamy, S., Team, Y.I.: Multiscale phate identifies multimodal signatures of covid-19. Nature Biotechnology **40**(5), 681–691 (2022). https://doi.org/10.1038/s41587-021-01186-x

[20] Goldberg, Y., Ritov, Y.: Local procrustes for manifold embedding: a measure of embedding quality and embedding algorithms. Machine Learning **77**(1), 1–25 (2009). https://doi.org/10.1007/s10994-009-5107-9

[21] Jia, Y., Kwong, S., Hou, J., et al.: Semi-supervised non-negative matrix factorization with dissimilarity and similarity regularization. IEEE Trans. Neural. Netw. Learn. Syst., 1–12 (2019)

[22] Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighbourhood components analysis. In: Adv. Neural. Inf. Process. Systs. NIPS'04, pp. 513–520. MIT Press, Cambridge, MA, USA (2004)

[23] Ghojogh, B., Crowley, M.: Unsupervised and Supervised Principal

Component Analysis: Tutorial (2022)

[24] Coifman, R.R., Lafon, S.: Diffusion maps. Appl. Comput. Harmon. Anal. **21**(1), 5–30 (2006). https://doi.org/10.1016/j.acha.2006.04.006. Special Issue: Diffusion Maps and Wavelets

[25] Kurtzke, J.F.: Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology **33**(11), 1444–1452 (1983)

[26] Bermel, R., Waldman, A., Mowry, E.M.: Outcome measures in multiple sclerosis. Mult Scler Int **2014**, 439375 (2014)

[27] Hawkins, S.: Truly benign multiple sclerosis is rare: let's stop fooling ourselves–no. Mult Scler **18**(1), 11–12 (2011)

[28] Amato, M.P., Portaccio, E.: Truly benign multiple sclerosis is rare: let's stop fooling ourselves–yes. Mult Scler **18**(1), 13–14 (2011)

[29] Reynders, T., D'haeseleer, M., De Keyser, J., Nagels, G., D'hooghe, M.B.: Definition, prevalence and predictive factors of benign multiple sclerosis. eNeurologicalSci **7**, 37–43 (2017)

[30] Meyer-Moock, S., Feng, Y.-S., Maeurer, M., Dippel, F.-W., Kohlmann, T.: Systematic literature review and validity evaluation of the expanded disability status scale (EDSS) and the multiple sclerosis functional composite (MSFC) in patients with multiple sclerosis. BMC Neurology **14**(1), 58 (2014)

[31] Paul, F.: Pathology and MRI: exploring cognitive impairment in MS. Acta Neurol Scand **134 Suppl 200**, 24–33 (2016)

[32] Penner, I.-K.: Evaluation of cognition and fatigue in multiple sclerosis: daily practice and future directions. Acta Neurol Scand **134 Suppl 200**, 19–23 (2016)

[33] Penner, I.-K., Paul, F.: Fatigue as a symptom or comorbidity of neurological diseases. Nat Rev Neurol **13**(11), 662–675 (2017)

[34] von Bismarck, O., Dankowski, T., Ambrosius, B., Hessler, N., Antony, G., Ziegler, A., Hoshi, M.-M., Aly, L., Luessi, F., Groppa, S., Klotz, L., Meuth, S.G., Tackenberg, B., Stoppe, M., Then Bergh, F., Tumani, H., Kümpfel, T., Stangel, M., Heesen, C., Wildemann, B., Paul, F., Bayas, A., Warnke, C., Weber, F., Linker, R.A., Ziemann, U., Zettl, U.K., Zipp, F., Wiendl, H., Hemmer, B., Gold, R., Salmen, A.: Treatment choices and neuropsychological symptoms of a large cohort of early MS. Neurol Neuroimmunol Neuroinflamm **5**(3), 446 (2018)

[35] Hutchinson, M.: Truly benign multiple sclerosis is rare: let's stop fooling ourselves–commentary. Mult Scler **18**(1), 15 (2011)

[36] McAlpine, D., Compson, N., *et al.*: Some aspects of the natural history of disseminated sclerosis. 1. the incidence, coarse, and prognosis. Quarterly Journal of Medicine **21**, 135–167 (1952)

[37] Confavreux, C., Compston, A.: The natural history of multiple sclerosis. In: McAlpine's Multiple Sclerosis, pp. 183–272. Elsevier, ??? (2006)

[38] Glad, S., Aarseth, J., Nyland, H., Riise, T., Myhr, K.-M.: Benign multiple sclerosis: a need for a consensus. Acta neurologica scandinavica **122**, 44–50 (2010)

[39] Ramsaransing, G.S.M., De Keyser, J.: Benign course in multiple sclerosis: a review. Acta Neurol Scand **113**(6), 359–369 (2006)

[40] Morrow, S.A., Conway, D., Fuchs, T., Wojcik, C., Unverdi, M., Yasin, F., Pol, J., Eckert, S., Hojnacki, D.H., Dwyer, M., Zivadinov, R., Weinstock-Guttman, B., Benedict, R.H.: Quantifying cognition and fatigue to enhance the sensitivity of the edss during relapses. Multiple Sclerosis Journal **27**(7), 1077–1087 (2021) https://arxiv.org/abs/https://doi.org/10.1177/1352458520973618. https://doi.org/10.1177/1352458520973618. PMID: 33259273

[41] Ellenberger, D., Flachenecker, P., Haas, J., Hellwig, K., Paul, F., Stahmann, A., Warnke, C., Zettl, U.K., Rommer, P.S.: Is benign ms really benign? what a meaningful classification beyond the edss must take into consideration. Multiple Sclerosis and Related Disorders **46**, 102485 (2020). https://doi.org/10.1016/j.msard.2020.102485

[42] Correale, J., Ysrraelit, M.C., Fiol, M.P.: Benign multiple sclerosis: Does it exist? Current Neurology and Neuroscience Reports **12**(5), 601–609 (2012)

[43] Golan, D., Doniger, G.M., Srinivasan, J., Sima, D.M., Zarif, M., Bumstead, B., Buhse, M., Van Hecke, W., Wilken, J., Gudesblatt, M.: The association between MRI brain volumes and computerized cognitive scores of people with multiple sclerosis. Brain and Cognition **145**, 105614 (2020)

[44] Niiranen, M., Koikkalainen, J., Lötjönen, J., Selander, T., Cajanus, A., Hartikainen, P., Simula, S., Vanninen, R., Remes, A.M.: Grey matter atrophy in patients with benign multiple sclerosis. Brain Behav **12**(7), 2679 (2022)

[45] Polman, C.H., Reingold, S.C., Banwell, B., Clanet, M., Cohen, J.A., Filippi, M., Fujihara, K., Havrdova, E., Hutchinson, M., Kappos, L., Lublin, F.D., Montalban, X., O'Connor, P., Sandberg-Wollheim, M., Thompson, A.J., Waubant, E., Weinshenker, B., Wolinsky, J.S.: Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Ann Neurol **69**(2), 292–302 (2011)

[46] Boissy, A.R., Ford, P.J.: A touch of MS: therapeutic mislabeling. Neurology **78**(24), 1981–1985 (2012)

[47] University of California, San Francisco MS-EPIC Team, Cree, B.A.C., Hollenbach, J.A., Bove, R., Kirkish, G., Sacco, S., Caverzasi, E., Bischof, A., Gundel, T., Zhu, A.H., Papinutto, N., Stern, W.A., Bevan, C., Romeo, A., Goodin, D.S., Gelfand, J.M., Graves, J., Green, A.J., Wilson, M.R., Zamvil, S.S., Zhao, C., Gomez, R., Ragan, N.R., Rush, G.Q., Barba, P., Santaniello, A., Baranzini, S.E., Oksenberg, J.R., Henry, R.G., Hauser, S.L.: Silent progression in disease activity-free relapsing multiple sclerosis. Ann Neurol **85**(5), 653–666 (2019)

[48] Cree, B.A.C., Mares, J., Hartung, H.-P.: Current therapeutic landscape in multiple sclerosis: an evolving treatment paradigm. Curr Opin Neurol **32**(3), 365–377 (2019)

[49] Avasarala, J.: Redefining acute relapses in multiple sclerosis: Implications for phase 3 clinical trials and treatment algorithms. Innov Clin Neurosci **14**(3-4), 38–40 (2017)

[50] Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics, 50–60 (1947)

[51] Cree, B.A.C., Arnold, D.L., Chataway, J., Chitnis, T., Fox, R.J., Pozo Ramajo, A., Murphy, N., Lassmann, H.: Secondary progressive multiple sclerosis: New insights. Neurology **97**(8), 378–388 (2021)

[52] Huang, H., Shi, H., Feng, S., Chen, W., Yu, Y., Lin, D., Chen, R.: Confocal raman spectroscopic analysis of the cytotoxic response to cisplatin in nasopharyngeal carcinoma cells. Analytical Methods **5**(1), 260–266 (2013). https://doi.org/10.1039/c2ay25684c. Cited by: 39

[53] Smith, R., Wright, K.L., Ashton, L.: Raman spectroscopy: an evolving technique for live cell studies. Analyst **141**, 3590–3600 (2016). https://doi.org/10.1039/C6AN00152A

[54] Zhang, W., Li, Q., Tang, M., Zhang, H., Sun, X., Zou, S., Jensen, J.L., Liou, T.G., Zhou, A.: A multi-scale approach to study biochemical and

biophysical aspects of resveratrol on diesel exhaust particle-human primary lung cell interaction. Scientific Reports **9**(1) (2019). https://doi.org/10.1038/s41598-019-54552-w. Cited by: 11; All Open Access, Gold Open Access, Green Open Access

[55] Zhang, W., Rhodes, J.S., Garg, A., Takemoto, J.Y., Qi, X., Harihar, S., Tom Chang, C.-W., Moon, K.R., Zhou, A.: Label-free discrimination and quantitative analysis of oxidative stress induced cytotoxicity and potential protection of antioxidants using raman micro-spectroscopy and machine learning. Anal. Chim. Acta **1128**, 221–230 (2020). https://doi.org/10.1016/j.aca.2020.06.074

[56] Fajnzylber, J., Regan, J., Coxen, K., Corry, H., Wong, C., Rosenthal, A., Worrall, D., Giguel, F., Piechocka-Trocha, A., Atyeo, C., et al.: Sars-cov-2 viral load is associated with increased disease severity and mortality. Nat Commun, 5493 (2020)

[57] Brunet-Ratnasingham, E., Morin, S., Randolph, H., Labrecque, M., Belair, J., Lima-Barbosa, R., Pagliuzza, A., Marchitto, L., Hultstrom, M., Niessl, J., et al.: Sustained ifn signaling is associated with delayed development of sars-cov-2-specific immunity. medRxiv, 2023–06 (2023)

[58] Brunet-Ratnasingham, E., Anand, S.P., Gantner, P., Dyachenko, A., Moquin-Beaudry, G., Brassard, N., Beaudoin-Bussières, G., Pagliuzza, A., Gasser, R., Benlarbi, M., *et al.*: Integrated immunovirological profiling validates plasma sars-cov-2 rna as an early predictor of covid-19 mortality. Science Advances **7**(48), 5629 (2021)

[59] Sechi, E., Messina, S., Keegan, B.M., Buciuc, M., Pittock, S.J., Kantarci, O.H., Weinshenker, B.G., Flanagan, E.P.: Critical spinal cord lesions associate with secondary progressive motor impairment in long-standing MS: A population-based case-control study. Mult Scler **27**(5), 667–673 (2020)

[60] Moccia, M., Ruggieri, S., Ianniello, A., Toosy, A., Pozzilli, C., Ciccarelli, O.: Advances in spinal cord imaging in multiple sclerosis. Ther Adv Neurol Disord **12**, 1756286419840593 (2019)

[61] Vlachos, M., Domeniconi, C., Gunopulos, D., Kollios, G., *et al.*: Non-linear dimensionality reduction techniques for classification and visualization. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '02, pp. 645–651. Association for Computing Machinery, New York, NY, USA (2002). https://doi.org/10.1145/775047.775143. https://doi.org/10.1145/775047.775143

[62] Geng, X., Zhan, D.-C., Zhou, Z.-H.: Supervised nonlinear dimensionality reduction for visualization and classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **35**(6), 1098–1107 (2005). https://doi.org/10.1109/TSMCB.2005.850151

[63] Jiang, Q., Jia, M.: Supervised laplacian eigenmaps for machinery fault classification. 2009 WRI World Congress on Computer Science and Information Engineering **7**, 116–120 (2009). https://doi.org/10.1109/CSIE.2009.765

[64] Balcan, M.-F., Blum, A., Srebro, N.: A theory of learning with similarity functions. Mach. Learn. **72**(1), 89–112 (2008). https://doi.org/10.1007/s10994-008-5059-5

[65] Ali Khan, M., Izhar Shah, M., Faisal Javed, M., Ijaz Khan, M., *et al.*: Application of random forest for modelling of surface water salinity. Ain Shams Engineering Journal **13**(4), 101635 (2022). https://doi.org/10.1016/j.asej.2021.11.004

[66] He, S., Wu, J., Wang, D., He, X.: Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest. Chemosphere **290**, 133388 (2022). https://doi.org/10.1016/j.chemosphere.2021.133388

[67] Wang, J., Rao, C., Goh, M., Xiao, X.: Risk assessment of coronary heart disease based on cloud-random forest. Artificial Intelligence Review **56**(1), 203–232 (2023). https://doi.org/10.1007/s10462-022-10170-z

[68] Pellegrino, E., Jacques, C., Beaufils, N., Nanni, I., Carlioz, A., Metellus, P., Ouafik, L.: Machine learning random forest for predicting oncosomatic variant ngs analysis. Scientific Reports **11**(1), 21820 (2021). https://doi.org/10.1038/s41598-021-01253-y

[69] Kapsiani, S., Howlin, B.J.: Random forest classification for predicting lifespan-extending chemical compounds. Scientific Reports **11**(1), 13812 (2021). https://doi.org/10.1038/s41598-021-93070-6

[70] Palimkar, P., Shaw, R.N., Ghosh, A.: Machine learning technique to prognosis diabetes disease: Random forest classifier approach. In: Bianchini, M., Piuri, V., Das, S., Shaw, R.N. (eds.) Advanced Computing and Intelligent Technologies, pp. 219–244. Springer, Singapore (2022)

[71] Simon, S.M., Glaum, P., Valdovinos, F.S.: Interpreting random forest analysis of ecological models to move from prediction to explanation. Scientific Reports **13**(1), 3881 (2023). https://doi.org/10.1038/s41598-023-30313-8

[72] Loef, B., Wong, A., Janssen, N.A.H., Strak, M., Hoekstra, J., Picavet, H.S.J., Boshuizen, H.C.H., Verschuren, W.M.M., Herber, G.-C.M.: Using random forest to identify longitudinal predictors of health in a 30-year cohort study. Scientific Reports **12**(1), 10372 (2022). https://doi.org/10.1038/s41598-022-14632-w

[73] Seifert, S.: Application of random forest based approaches to surface-enhanced raman scattering data. Scientific Reports **10**(1), 5436 (2020). https://doi.org/10.1038/s41598-020-62338-8

[74] Stephan, J., Stegle, O., Beyer, A.: A random forest approach to capture genetic effects in the presence of population structure. Nature Communications **6**(1), 7432 (2015). https://doi.org/10.1038/ncomms8432

[75] Kong, Y., Yu, T.: A deep neural network model using random forest to extract feature representation for gene expression data classification. Scientific Reports **8**(1), 16477 (2018). https://doi.org/10.1038/s41598-018-34833-6

[76] Gerasimiuk, M., Shung, D., Tong, A., Stanley, A., Schultz, M., Ngu, J., Laine, L., Wolf, G., Krishnaswamy, S.: Mural: An unsupervised random forest-based embedding for electronic health record data. In: 2021 IEEE International Conference on Big Data (Big Data), pp. 4694–4704 (2021). IEEE

[77] Lin, Y., Jeon, Y.: Random forests and adaptive nearest neighbors. J. Am. Stat. Assoc. **101**(474), 578–590 (2006)

[78] Breiman, L., Cutler, A.: Random Forests. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#prox. (Accessed on 4/10/2023)

[79] Hastie, T., Tibshirani, R., Friedman, J.: Random forests. In: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. edn., pp. 587–604. Springer, New York, NY (2009). https://doi.org/10.1007/978-0-387-84858-7_15

[80] Liaw, A., Wiener, M.: Classification and regression by randomForest. R News **2**(3), 18–22 (2002)

[81] Haghverdi, L., Buettner, M., Wolf, F.A., Buettner, F., Theis, F.J.: Diffusion pseudotime robustly reconstructs lineage branching. Nature Methods **13**(10), 845 (2016)

[82] Anderson, E.: The species problem in iris. Annals of the Missouri Botanical Garden **23**(3), 457–509 (1936). Accessed 2022-07-29

[83] Dexter, E., Rollwagen-Bollens, G., Bollens, S.M.: The trouble with stress: A flexible method for the evaluation of nonmetric multidimensional scaling. Limnology and Oceanography: Methods **16**(7), 434–443 (2018) https://arxiv.org/abs/https://aslopubs.onlinelibrary.wiley.com/doi/pdf/10.1002/lom3.10257. https://doi.org/10.1002/lom3.10257

[84] Shahapure, K.R., Nicholas, C.: Cluster quality analysis using silhouette score. In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), pp. 747–748 (2020). https://doi.org/10.1109/DSAA49011.2020.00096

[85] Lovmar, L., Ahlford, A., Jonsson, M., Syvänen, A.-C.: Silhouette scores for assessment of snp genotype clusters. BMC genomics **6**(1), 1–6 (2005)

[86] Ogbuabor, G., Ugwoke, F.: Clustering algorithm for a healthcare dataset using silhouette score value. Int. J. Comput. Sci. Inf. Technol **10**(2), 27–37 (2018)

[87] Pikoula, M., Quint, J.K., Nissen, F., Hemingway, H., Smeeth, L., Denaxas, S.: Identifying clinically important COPD sub-types using data-driven approaches in primary care population based electronic health records. BMC Medical Informatics and Decision Making **19**(1), 86 (2019)

[88] Rhodes, J.S., Cutler, A., Wolf, G., Moon, K.R.: Random forest-based diffusion information geometry for supervised visualization and data exploration. 2021 IEEE Statistical Signal Processing Workshop (SSP), 331–335 (2021). https://doi.org/10.1109/SSP49050.2021.9513749

[89] Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W., Ng, L.G., Ginhoux, F., Newell, E.W.: Dimensionality reduction for visualizing single-cell data using umap. Nature biotechnology **37**(1), 38–44 (2019)

[90] Lublin, F.D., Reingold, S.C., Cohen, J.A., Cutter, G.R., Sørensen, P.S., Thompson, A.J., Wolinsky, J.S., Balcer, L.J., Banwell, B., Barkhof, F., Bebo, B. Jr, Calabresi, P.A., Clanet, M., Comi, G., Fox, R.J., Freedman, M.S., Goodman, A.D., Inglese, M., Kappos, L., Kieseier, B.C., Lincoln, J.A., Lubetzki, C., Miller, A.E., Montalban, X., O'Connor, P.W., Petkau, J., Pozzilli, C., Rudick, R.A., Sormani, M.P., Stüve, O., Waubant, E., Polman, C.H.: Defining the clinical course of multiple sclerosis: the 2013 revisions. Neurology **83**(3), 278–286 (2014)

[91] Goldenberg, M.M.: Multiple sclerosis review. P T **37**(3), 175–184 (2012)

[92] Şen, S.: Neurostatus and EDSS calculation with cases. Noro Psikiyatr Ars **55**(Suppl 1), 80–83 (2018)

[93] Torkildsen, Ø., Myhr, K.-M., Bø, L.: Disease-modifying treatments for multiple sclerosis – a review of approved medications. European Journal of Neurology **23**(S1), 18–27 (2016) https://arxiv.org/abs/https://onlinelibrary.wiley.com/doi/pdf/10.1111/ene.12883. https://doi.org/10.1111/ene.12883

[94] Eshaghi, A., Young, A.L., Wijeratne, P.A., Prados, F., Arnold, D.L., Narayanan, S., Guttmann, C.R.G., Barkhof, F., Alexander, D.C., Thompson, A.J., Chard, D., Ciccarelli, O.: Identifying multiple sclerosis subtypes using unsupervised machine learning and MRI data. Nature Communications **12**(1), 2078 (2021)

[95] Kuhlmann, T., Moccia, M., Coetzee, T., Cohen, J., Correale, J., Graves, J., Marrie, R., Montalban, X., Yong, V.W., Thompson, E., Reich, D., Amato, M., Banwell, B., Barkhof, F., Chataway, J., Chitnis, T., Comi, G., Derfuss, T., Finlayson, M., Waubant, E.: Multiple sclerosis progression: time for a new mechanism-driven framework. The Lancet Neurology **22** (2022). https://doi.org/10.1016/S1474-4422(22)00289-7

[96] Cutter, G., Kappos, L.: Chapter 20 - clinical trials in multiple sclerosis. In: Goodin, D.S. (ed.) Multiple Sclerosis and Related Disorders. Handbook of Clinical Neurology, vol. 122, pp. 445–453. Elsevier, ??? (2014). https://doi.org/10.1016/B978-0-444-52001-2.00019-4. https://www.sciencedirect.com/science/article/pii/B9780444520012000194

[97] Trip, S.A., Miller, D.H.: Imaging in multiple sclerosis. Journal of Neurology, Neurosurgery & Psychiatry **76**(suppl 3), 11–18 (2005) https://arxiv.org/abs/https://jnnp.bmj.com/content/76/suppl_3/iii11.full.pdf. https://doi.org/10.1136/jnnp.2005.073213

[98] Berndt, D.J., Clifford, J.: Using dynamic time warping to find patterns in time series. In: KDD Workshop (1994). https://api.semanticscholar.org/CorpusID:929893

[99] Keogh, E., Ratanamahatana, C.A.: Exact indexing of dynamic time warping. Knowledge and Information Systems **7**(3), 358–386 (2005)

[100] Kruskal, J., Liberman, M.: The symmetric time-warping problem: From continuous to discrete. Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison (1983)

[101] Ratanamahatana, C., Keogh, E.: Everything you know about dynamic time warping is wrong. (2004)

[102] Dudani, S.A.: The distance-weighted k-nearest-neighbor rule. IEEE Transactions on Systems, Man, and Cybernetics **SMC-6**(4), 325–327 (1976). https://doi.org/10.1109/TSMC.1976.5408784

[103] Rhodes, J.S.: Supervised manifold learning via random forest geometry-preserving proximities. In: Fourteenth International Conference on Sampling Theory and Applications (2023). https://openreview.net/forum?id=t6E4dZjp-e

[104] Zhang, Z.-M., Chen, S., Liang, Y.-Z.: Baseline correction using adaptive iteratively reweighted penalized least squares. Analyst **135**(5), 1138–1146 (2010)

[105] Tremblay, K., Rousseau, S., Zawati, M., Auld, D., Chassé, M., Coderre, D., Falcone, E., Gauthier, N., Grandvaux, N., Gros-Louis, F., et al.: The biobanque québécoise de la covid-19 (bqc19)-a cohort to prospectively study the clinical and biological determinants of covid-19 clinical trajectories. PLoS One, 16–50245031 (2021)

[106] Prévost, J., Gasser, R., Beaudoin-Bussières, G., Richard, R. J abd Duerr, Laumaea, A., Anand, S., Goyette, G., Benlarbi, M., Ding, S., Medjahed, et al.: Cross-sectional evaluation of humoral responses against sars-cov-2 spike. Cell Rep Med, 1–7100126 (2020)