Sedimentologika

Method article

Data-driven lithofacies prediction of unconsolidated sediments from wireline logs

Sebastian Schaller[1,2*], Patricio Becerra[3], Sarah Beraus[4,5], Marius W. Buechi[1,2], David Mair[1], Mehrdad Sardar Abadi[4], Bennet Schuster[1,2,6], Flavio S. Anselmetti[1,2]

[1] Institute of Geological Sciences, University of Bern, Baltzerstrasse 1+3, 3012 Bern, Switzerland
[2] Oeschger Centre for Climate Change Research, University of Bern, Hochschulstrasse 4, 3012 Bern, Switzerland
[3] Space Research and Planetary Sciences, Physics Institute, University of Bern , Sidlerstrasse 5, 3012 Bern, Switzerland
[4] LIAG Institute for Applied Geophysics, Stilleweg 2, 30655 Hannover, Germany
[5] Section Geology - Institute of Earth System Sciences, Leibniz University Hannover, Callinstraße 30, 30167 Hannover, Germany
[6] Institute of Earth and Environmental Sciences, University of Freiburg, Tennenbacher Str. 4, 79106 Freiburg, Germany

*corresponding author: Sebastian Schaller (sebastian.schaller@unibe.ch)

doi: 10.57035/journals/sdk.2025.e31.1819

Editors: Murray Gingras and Korhan Ayranci
Reviewers: Burak Tunc and one anonymous reviewer
Copyediting, layout and production: Romain Vaucher, Marine Prieur, and Seyedeh Mortazavi

Submitted: 12.01.2025
Accepted: 24.10.2025
Published: 12.11.2025

Abstract | Detailed knowledge of the lithostratigraphy of unconsolidated sediments is essential for many scientific and industrial applications. Although drill cores provide lithological and petrophysical information (e.g., lithofacies, permeability, consolidation), core recovery is time- and resource-intensive. In contrast, flush drillings are faster and less expensive but lack detailed geological context. Therefore, combining the advantages of both methods can significantly enhance subsurface investigations while reducing reliance on costly core drillings. This study explores the use of unsupervised machine learning to build data-driven stratigraphic models from standard petrophysical and geochemical wireline logging data. Specifically, it applies dimensionality reduction (UMAP) and hierarchical clustering to identify distinct lithofacies types. The method was tested on datasets from the former Rhine Glacier area in Germany, including one core-controlled well and a nearby flush-drilled borehole. The developed workflow predicted the lithofacies of the core-controlled well with ~76% accuracy, capturing both major stratigraphic units and finer internal features, directly linking them with the geology identified at the drilled site. This allows linking the reconstructed lithology of the core-controlled well with the flush-drilled well. The results demonstrate that unsupervised machine learning can significantly improve stratigraphic models of unconsolidated Quaternary sediments using wireline logs, with minimal dependence on core data.

Lay summary | Understanding the composition and layering of subsurface sediments is essential for a range of scientific and industrial applications, including groundwater management, engineering and construction, and resource assessment. Traditionally, this information is obtained through core drilling, which provides detailed physical samples and high-resolution geological insights, but is time-consuming and costly. Faster drilling methods, such as flush drilling, are more economical but break sediments into small fragments, making it difficult to reconstruct original layering and sediment properties. In this study, we applied unsupervised machine learning (a computational method that detects patterns in data without prior labeling) to standard downhole measurements of physical and chemical sediment properties. We validated the results against core samples. The approach was tested on two nearby boreholes in the former Rhine Glacier region of southern Germany: one fully cored and one drilled using the faster flush-drilling technique. Our analysis reproduced approximately 76% of the sedimentary layering. This enabled us to project the inferred stratigraphy onto the flush-drilled borehole, effectively creating an "artificial" drill core. These results demonstrate that machine learning can significantly improve subsurface geological interpretations while reducing reliance on costly core drilling.

Keywords: Unsupervised machine learning; Lithostratigraphy; Unconsolidated sediments; Wireline logging; Core drilling

1. Introduction

Understanding the stratigraphy of the terrestrial subsurface is essential for numerous scientific and industrial applications. However, due to its natural inaccessibility, it remains a primary challenge in geoscience. Applications include: 1) climate-change mitigation and paleoclimate studies (e.g., Drilling Overdeepened Alpine Valley Project; Anselmetti et al., 2022), 2) subsurface exploration and classification (e.g., potential and use of geothermal energy; Giardini et al., 2021), 3) natural hazard and risk assessments (e.g., earthquake vulnerability; Bergamo et al., 2023), and 4) infrastructure projects (e.g., railroad tunnels; Guntli et al., 2016). This challenge can be addressed either 1) indirectly via geophysical methods such as seismic or gravimetric surveys, or 2) directly via drilling. The two main drilling techniques are destructive flush drilling (e.g., Honer & Sherrell, 1977) – where the sediments/rocks are fragmented into small cuttings and flushed out by a drilling fluid (typically air or a mixture of water and additives termed "mud") – and core drilling (e.g., Marjoribanks, 2010), where the original stratigraphic order is preserved. Only core drilling provides direct, high-resolution access to in-situ geological information. In contrast, geophysical surveys return only indirect, low-resolution models, and flush drillings offer a limited and disturbed representation of the subsurface lithology.

Core drillings and its analysis require substantial economic and logistical commitments, despite their advantages. These costs significantly hinder project planning and complicate funding acquisition, often leading to the downsizing or cancellation of projects. As a result, valuable geological data often remains inaccessible. Although recent advances in automated image-based lithoclassification of sediment and rock cores (e.g., Di Martino et al., 2023; Lauper et al., 2021) have the potential to reduce some of these costs in the future, core drilling and analysis will remain a significant and critical investment. Limits in funding often lead to a lower number of drillings being realized, directly lowering the spatial resolution and quality of geological models – especially in complex sedimentary environments such as glacially overdeepened valleys (e.g., Buechi et al., 2024). Given these constraints, a method that combines the speed and cost-efficiency of flush drilling with the data quality of core drilling is highly desirable. While core drilling cannot be entirely replaced, combining flush drillings with wireline logging surveys – where a probe is inserted on a wire down the drill hole to create a continuous record of the properties of the surrounding geological formations (e.g., Mondol, 2015) – is an alternative . Wireline logs are traditionally used to evaluate carbonate or siliciclastic rocks with well-known petrophysical properties (i.e., electrofacies: Serra & Abbott, 1982; or acoustic properties: Anselmetti & Eberli, 1999), with well-established data-processing and interpretation (e.g., Ghosh, 2022; Lai et al., 2024). The combination of destructive and core drilling with wireline techniques (e.g., "Logging While Drilling"; Selley & Sonnenberg, 2023) is standard in energy exploration (e.g., hydrocarbon, geothermal).

Wireline logging techniques have proven effective in creating sophisticated stratigraphic models, however, they are rarely applied in Quaternary settings. However, the growing scientific and economic interest (e.g., groundwater, infrastructure) in these sediments may lead to broader application — particularly since many standard wireline logs record petrophysical and chemical parameters that can also be used for characterizing unconsolidated sediments because they provide information about: 1) lithological composition (e.g., magnetic susceptibility, photoelectric absorption, U-, Th-, and K-content), 2) depositional environment (e.g., clay/silt content: natural gamma radiation), or 3) geotechnical properties (e.g., water content and permeability: electric resistivity, formation porosity and compaction: porosity and bulk density).

Recent advances in data analysis have enabled the practical use of machine learning in geoscience (e.g., Dramsch, 2020; Lary et al., 2016), particularly for detecting and correlating wireline-data patterns with lithology using different variations of supervised and unsupervised approaches (e.g., Carrasquilla, 2023; Mukherjee & Sain, 2021; Popescu et al., 2021; Zekri et al., 2025), potentially reducing the need for manual interpretation. However, supervised approaches require substantial and representative training data and may perform poorly when used on data that differs too much from the original training data. In contrast, unsupervised methods (e.g., clustering algorithms) do not require a large amount of labeled data. They can be applied when data availability is limited due to logistical constraints (e.g., Abbas et al., 2023; Dixit et al., 2020; Hasan et al., 2023). In addition to their relatively fast implementation, unsupervised methods help to reduce user bias, allowing a focus on intrinsic data variability.

This study presents a new approach for predicting lithofacies from wireline-logging data using unsupervised machine learning. Combining dimensionality reduction and clustering techniques enables pattern recognition in high-dimensional feature space to identify complex relationships between multiple wireline logs. Where core data is available, predictions can be directly validated against the lithology. This method was tested and validated in the accompanying case study from the "Drilling Overdeepened Alpine Valleys" (DOVE) project (Anselmetti et al., 2022) in the wider Lake Constance area in Central Europe. The potential of the workflow was also evaluated in cross-hole applications. The study also evaluates the sensitivity of different log types for robust lithofacies classification in the encountered sediments.

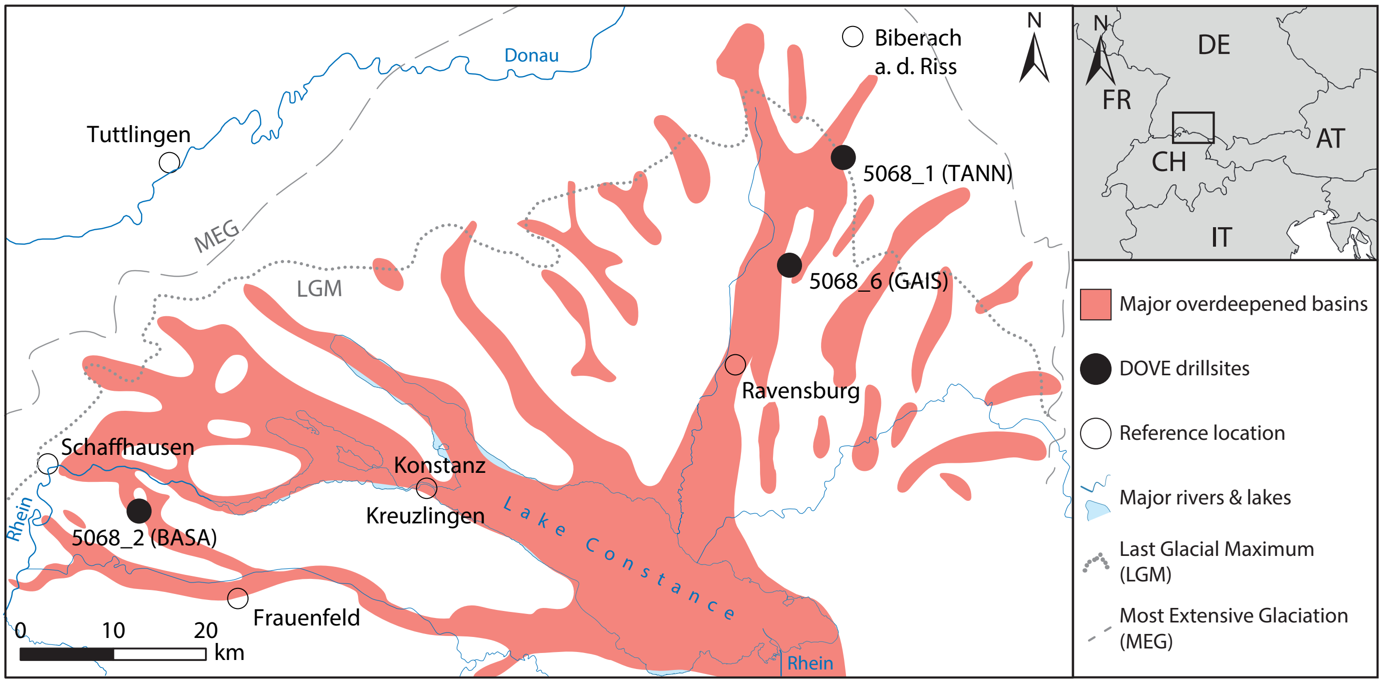


Figure 1 | Overview of the glacially overdeepened systems within the former Rhine Glacier lobe (modified from Ellwanger et al., 2011) with the location of three DOVE Phase I sites: 5068_1 (TANN), 5068_2 (BASA), and 5068_6 (GAIS).

2. Drill site and data selection

The modern Lake Constance area was heavily affected by multiple Quaternary glaciations, which formed a complex system of glacially overdeepened valleys and basins (Figure 1; e.g., Ellwanger et al., 2011; Preusser et al., 2011). These troughs–eventually refilled by unconsolidated Quaternary sediments (mainly sand, gravel, silt/clay, and diamicts) are key archives for paleolandscape and climate reconstructions (e.g., Buechi, 2016; Buechi et al., 2024; Dehnert et al., 2012). Therefore, the former Rhine Glacier lobe area became a primary target of DOVE Phase I (Anselmetti et al., 2022).

The dataset used in this study (Figure S1) includes a high-quality drill core and wireline logs from borehole 5068_1_C, as well as wireline logs from borehole 5068_1_A. The recovered core comprises ~156 m of unconsolidated Quaternary sediments, overlying ~10 m of Neogene sand and siltstone bedrock (for visual core examples, see Figure 2; Schuster et al., 2024). The Quaternary valley fill can be divided into two sections: an overdeepened unit (from ~156-40 m) and a non-overdeepened one (~40-0 m). The overdeepened sequence comprises, from bottom to top: 1) ~10 m of basal diamict overlying the bedrock, 2) a ~30 m thick clay- and silt-dominated bed, 3) a ~50 m thick sequence of interlayered sand and clay beds of varying thickness, 4) a ~10 m thick diamictic bed, and 5) topped by a sand-dominated section till a depth of ~40 m. This overdeepened strata is overlain by a ~40 m thick gravel layer. Schuster et al. (2024) provide a detailed sedimentological and stratigraphic interpretation of the core. A summary of the hierarchical lithofacies classification scheme used in this study is presented in Table S1.

The wireline data were acquired with a spacing of 10 cm but were not acquired above ~45 m (5068_1_A) and ~85 m (5068_1_C) due to the installation of steel casing. Nevertheless, the remaining open-hole sections provide sufficient vertical coverage. Table S3 summarizes the relevant wireline logs shown in Supplementary Figure S1. Further details on survey protocols, data quality, and preprocessing, as well as technical aspects of the drilling operations, are available in the DOVE Phase I Operational Report (DOVE-Phase 1 Scientific Team et al., 2023a) and the accompanying Data Report (DOVE-Phase 1 Scientific Team et al., 2023b).

This case study offers a unique opportunity to 1) directly validate the performance of the lithofacies-prediction workflow with the core-controlled logging data, 2) investigate the relationships between sedimentary and petrophysical properties, and 3) evaluate a cross-hole prediction approach for extending lithological models beyond core control. The high-quality dataset and complex setting make it an ideal case for developing and testing advanced wireline-based lithofacies analysis of Quaternary sediments.

3. Methods

This study presents a data-driven workflow (Figure 3) that links data clusters – defined by shared petrophysical and geochemical properties from wireline logs – with the corresponding drill core-based lithology. These data clusters are extracted from the measured wireline logs through the process of dimensionality reduction and hierarchical clustering. This direct link between the extracted data clusters and the actual lithology serves as a translation key for cross-correlation with similar wireline datasets

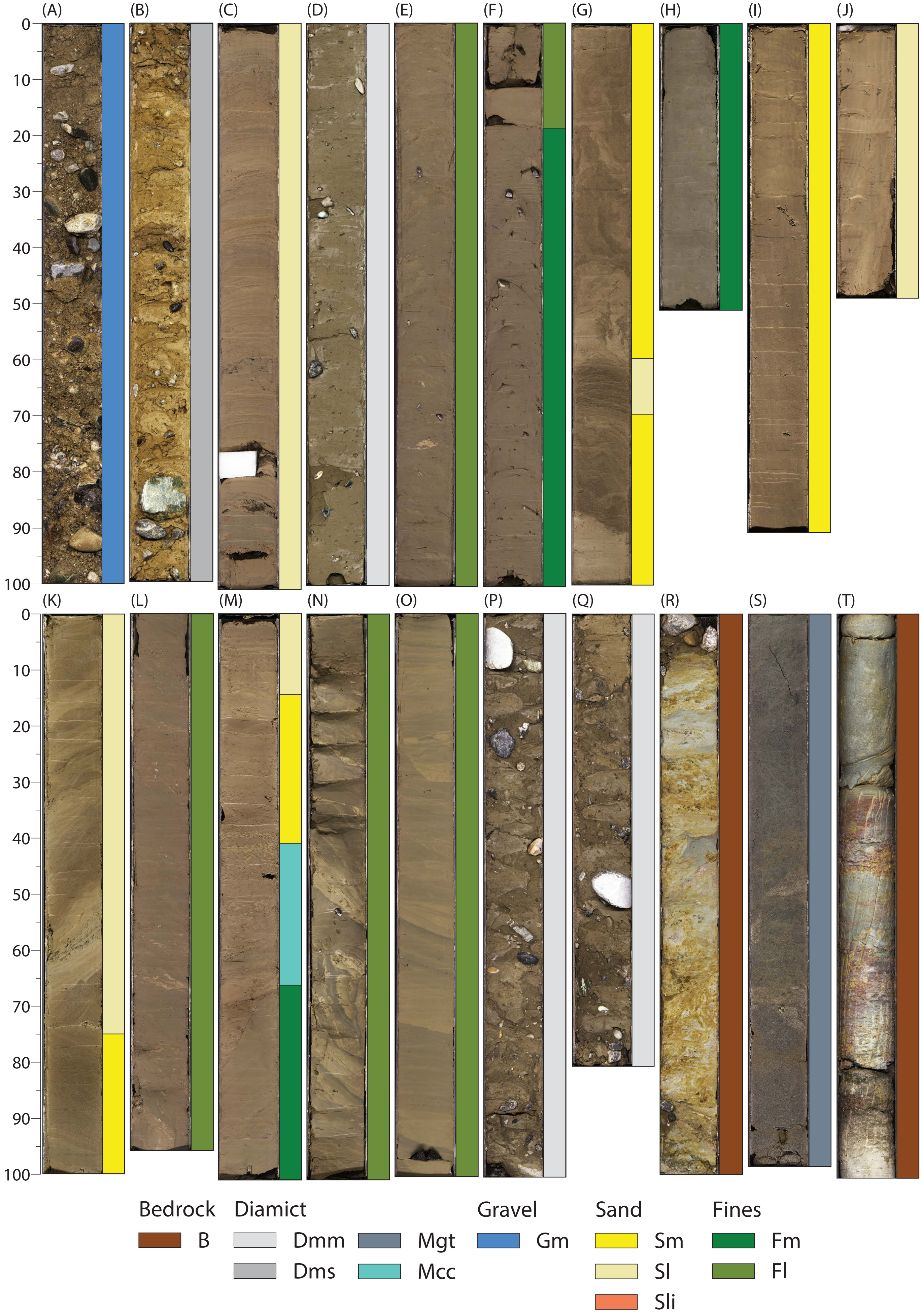


Figure 2 | Representative core examples of 5068_1_C with assigned lithotypes. An overview of the used lithotypes is provided in Table S1. The corresponding core section ID and composite depth of the used examples are given in Table S2. The vertical scale is in cm.

without direct lithological control. All data manipulation, processing, analysis, and visualization were performed using Python (v3.9.18) and Jupyter Notebooks (v6.4.5). The code of the workflow is based on the following key libraries: i) SciPy (v1.11.3; Virtanen et al., 2020), 2) Scikit-learn (v1.3.0; Pedregosa et al., 2011), and 3) UMAP (v0.5.5; McInnes et al., 2018). A complete list of the Python packages and libraries used is provided in the supplementary code.

3.1. Data acquisition, processing, and analysis

All details regarding the data acquisition are provided in DOVE-Phase 1 Scientific Team et al. (2023a, 2023b) and Schuster et al. (2024).

In the first step, the averaged borehole geometry was derived from caliper logs and corrected for shifts that led to unrealistic values, such as too-small or too-large values in cased sections with known diameters. This corrected geometry was then used to apply a caving correction (Krammer & Pohl, 1987; Schlumberger, 1972, and internal documents from the ICDP OSG) on wireline logs acquired without borehole-wall contact, including magnetic susceptibility (Susz), the combined spectral gamma log (SGR), and its three components (U, Th, and K). Sections measured in the cased hole and zones of intense caving (>15% of the nominal borehole diameter), indicating poor data quality, were removed. Additional optional filters were applied to individual logs to exclude implausible values (e.g., density <1.5 g/cm³). The corresponding supplementary code provides detailed information about the applied preprocessing, correction factors, and additional data filters. The original litholog, based on the initially defined DOVE-lithotypes, was condensed into 13 lithotypes (“full lithology”) by merging those with small-scale variations (e.g., laminated and cross-bedded sand or different types of laminated fines). These condensed lithotypes are further grouped into the following five "major lithologies" representing the dominant grain-size fraction: B: Bedrock; D: Diamict; G: Gravel; S: Sand; F: Fines. An overview of the hierarchy of the different lithotypes is given in Table S1. Based on this new litholog, a major lithology and full lithology label was assigned to each measurement point in the core-controlled wireline dataset, directly linking lithology with wireline logging data.

Prior to applying dimensionality reduction and cluster analysis to the wireline data (see optional data flow in Figure 3), standard methods for data investigation were employed, such as cross plots, depth plots, and density distributions. Additionally, principal component analysis (PCA) and covariance matrix analysis were employed to better understand the intrinsic structure of the data, such as degree of variance, lithological distribution, loadings, and relationships between individual logs.

3.2. Clustering

In preparation for dimension reduction and clustering, the 10 input logs (Table S3) were normalized into a distribution with a mean of 0 and a standard deviation of 1 by removing the mean and scaling to unit variance (using scikit-learn’s StandardScaler; see supplementary code for more details). Then, the original ten dimensions of the normalized input data (the selected wireline logs for clustering) were reduced to three dimensions using the "Uniform Manifold Approximation and Projection" method (UMAP; McInnes et al., 2018) The optimal UMAP hyperparameters (n_neighbors and min_dist) were qualitatively determined by iteratively plotting the 2D UMAP projection of the data set for n_neighbors (10 to 100, in steps of 10) and min_dist (0.1 to 0.9, in steps of 0.1) aiming to balance global structure and local detail while minimizing information loss (Coenen & Pearce, 2019). Hierarchical agglomerative clustering (HAC; Nielsen, 2016) was then applied to the reduced and projected dataset (from 10 to 3 dimensions) to extract data clusters using Ward linkage with Euclidean distance.

The optimal number of clusters was determined empirically by calculating and comparing three cluster-quality metrics that reflect how well the clusters are defined and separated: Silhouette-Score with optional Silhouette-Plots (Rousseeuw, 1987), Calinski-Harabasz-Score (Caliński & Harabasz, 1974), and Davies-Bouldin-Score (Davies & Bouldin, 1979) for cases with nclusters = 2-49. Ideally, the Silhouette and Calinski-Harabasz scores should be at a local maximum, and the Davies-Bouldin score should be at a local minimum for the same number of clusters. An initial quality and plausibility check of the clustering was done by interpreting/comparing 1) 3D-UMAP-scatter plots of the clusters and the assigned lithology, 2) hierarchical dendrograms, 3) depth plots of the clusters and the litholog, and 4) cross-checking with geological background knowledge such as the depth of bedrock contact.

3.3. Evaluation and depth-correlation

Prior to quantifying prediction quality, a dominant major lithology was assigned to each cluster based on the distribution of labeled data – the most frequent lithology within the cluster. The quality of the link between data-driven clusters and core-derived lithofacies classifications was then quantified by measuring the proportion of correctly matched lithologies – that is, the frequency with which the predicted cluster label aligned with the core-based lithology label.

In addition to this quantitative characterization, the plausibility of the assigned lithologies was further qualitatively evaluated by: 1) analyzing the intrinsic data variance of each cluster using PCA and creating covariance matrices, 2) examining the distribution of wireline-log values within each cluster, and 3) interpreting the structural relationships revealed in the dendrogram.

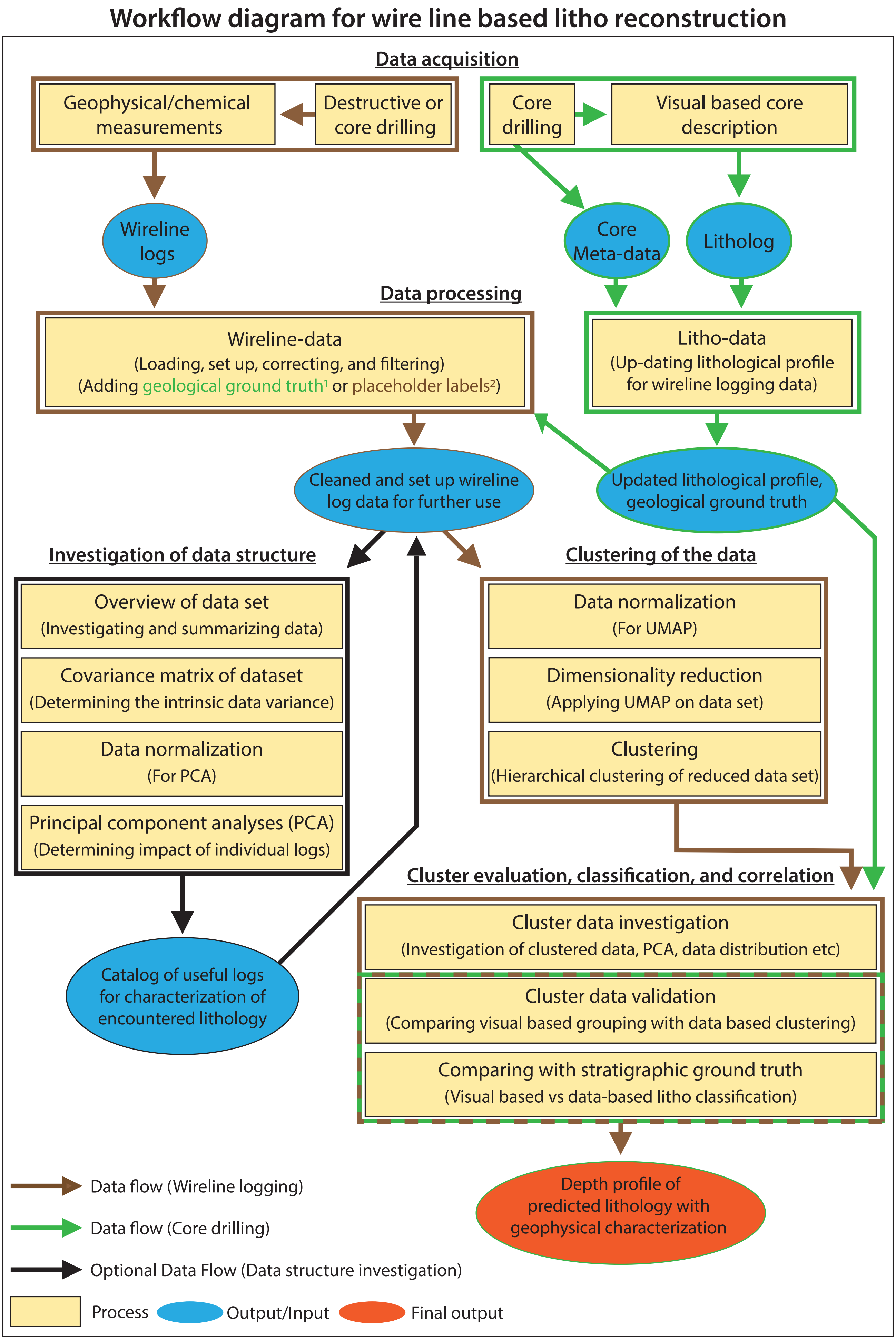

Workflow diagram for wire line based litho reconstruction
Data acquisition
Geophysical/chemical measurements
Destructive or core drilling
Core drilling
Visual based core description
Wireline logs
Core Meta-data
Litholog
Data processing
Wireline-data
(Loading, set up, correcting, and filtering)
(Adding geological ground truth¹ or placeholder labels²)
Litho-data
(Up-dating lithological profile for wireline logging data)
Cleaned and set up wireline log data for further use
Updated lithological profile, geological ground truth
Investigation of data structure
Overview of data set
(Investigating and summarizing data)
Covariance matrix of dataset
(Determining the intrinsic data variance)
Data normalization
(For PCA)
Principal component analyses (PCA)
(Determining impact of individual logs)
Catalog of useful logs for characterization of encountered lithology
Clustering of the data
Data normalization
(For UMAP)
Dimensionality reduction
(Applying UMAP on data set)
Clustering
(Hierarchical clustering of reduced data set)
Cluster evaluation, classification, and correlation
Cluster data investigation
(Investigation of clustered data, PCA, data distribution etc)
Cluster data validation
(Comparing visual based grouping with data based clustering)
Comparing with stratigraphic ground truth
(Visual based vs data-based litho classification)
Depth profile of predicted lithology with geophysical characterization
Data flow (Wireline logging)
Data flow (Core drilling)
Optional Data Flow (Data structure investigation)
Process
Output/Input
Final output

Figure 3 | Schematic diagram of the applied workflow; the standard workflow is represented by the combined green (core data) and brown (core-controlled wireline data[1]; without direct core control/lithodata[2]) data flows. The optional data flow represents the investigation of the intrinsic variance of the dataset.

Finally, comprehensive validation was performed by plotting the ground truth, predicted lithology, and validation results against depth. This projection enabled stratigraphic correlation and visualization of potential stratigraphic patterns in prediction accuracy and geological consistency.

3.4. Expansion of the core workflow to other wireline datasets

To extend the workflow to additional wireline datasets lacking core control such as from flush drillings, placeholder lithology labels (e.g., "no data") can be inserted (Figure 3). This enables unified preprocessing using the same pipeline as for core-controlled data and allows concatenation of datasets for joint clustering – this technique was applied to dataset 5068_1_A. The combined clustering allows for extrapolation of the dominant lithology from the core-controlled parts onto the uncontrolled parts of the combined wireline dataset and cross-correlation between the individual cluster-based depth profiles. It is also advisable to compare and examine the wireline datasets individually before combining them, as there may be shifts in the calibration of the logging tools between different surveys. In cases where the data come from the same geological system and are likely to represent identical lithological spectra, it is advisable to normalize the individual datasets prior to concatenation. After clustering, the combined dataset can be easily reseparated by their unique borehole-id and rearranged according to the original depth. This was done at the 5068_1 drill site (Figures S2 and S3). UMAP was run with a fixed random state to maintain reproducibility (see supplementary code for specific parameters).

4. Results

The results are based on 10 input wireline logs (Figure S1 and Table S3). The caliper log was used for data processing but was not included as direct input.

4.1. Data trends and impact of dimensionality reduction

The covariance data (Figures 4A and B) and the PCA results (Figures 4C and D; Tables 1 and 2; Tables S4 and S5) show strong similarities between the two wireline datasets (5068_1_A and C). First, both show a strong positive correlation between the deep and shallow resistivity (covariance ~1, very narrow angle between the corresponding loading vectors) and between the SGR, Th-, U-, and K-logs (covariance >0.5, loading vectors angles: ~20-30°) and a negative correlation between the resistivity logs and the porosity-, SGR-, Th-, U-, and K-logs (covariance: <-0.25, loading vector angles: 90-180°). Further, neither the loading plots nor the absolute values of the loading vectors show a strongly dominant log (Table 1). The two datasets differ mainly in the absolute value of the correlation, length, and direction of the loading vectors by preserving the primary trend. The range of variance ratio per principal component is narrow between the two datasets (e.g., var PC1: ~45-47.5%; var PC2: ~18-19.5%; Table 2). The projection of the first two principal components (PC1 and PC2, Figures 4C and D) contains ~63.5-67.5% of the data variance, and the 3-dimensional projection contains ~75-77% (Table 2).

The PCA biplots of PC1 and PC2 (Figures 4C and D) and the two-dimensional UMAP projection (Figures 5A-C) show a structured data distribution, both in general and within the assigned lithological data of 5068_1_C (Figures 4D and 5B). The distribution of the cleaned lithological ground truth data of 5068_1_C shows a distribution dominated by fines (~50%) with a recognizable component of diamictic sediments (~19%) and bedrock (~11%; Table 3). A more detailed visualization of the three-dimensional UMAP and PCA projection of the two individual wireline datasets of 5068_1_A and C is shown in Supplementary Figures S4 and S5.

4.2. Clustering results

The optimal number of clusters for the three-dimensional UMAP projection of the combined wireline datasets of 5068_1_A and 5068_1_C was determined to be 6 (CC0-CC5; Figures S6 and S7). The connection between individual clusters is illustrated in the dendrogram of the clustered combined dataset (5068_1_A and C), indicated by the vertical observation distance between connected clusters (Figure 6). After reseparating, the distribution between the individual clusters of the two datasets (5068_1_A and 5068_1_C) shows some variance in relative size but preserves the general trend in cluster size of the combined one (Table 4). Further, the wireline-log-data distribution of the two reseparated datasets matches (Figures S8 and S9), with slight variations in detailed distribution patterns visualized in the covariance and the PCA results (see Tables S6 and S7; Figures S10 and S11). A lithological characterization and classification of the clusters can be made by comparing the composition and distribution of the assigned core-based lithological data, if available (Table 3), with the individual data-based clusters (Table 4). A detailed lithological composition of each cluster is shown in Figure 7A for the core-controlled case

Dataset	Rshallow [Ohm m]	Rdeep [Ohm m]	SSPE [b/e]	DEN [g/ccm]	POR [%]	SGR [gAPI]	Th [ppm]	U [ppm]	K [%wt]	Susz [1E4SI]
5068_1_A	0.42	0.41	0.39	0.60	0.52	0.43	0.38	0.28	0.38	0.57
5068_1_C	0.44	0.43	0.46	0.64	0.38	0.47	0.39	0.35	0.38	0.47

Table 1 | Loading of the individual logs in the space spanned by PC1 and PC2 for each dataset.

Dataset	PC01	PC02	PC03	PC04	PC05	PC06	PC07	PC08	PC09	PC10
5068_1_A	47.64	19.48	10.04	5.65	5.39	4.95	3.86	2.94	0.05	0.02
5068_1_C	45.33	18.06	12.21	7.67	6.97	4.98	2.87	1.85	0.06	0.00

Table 2 | Explained variance in % of the principal component of each wireline dataset.

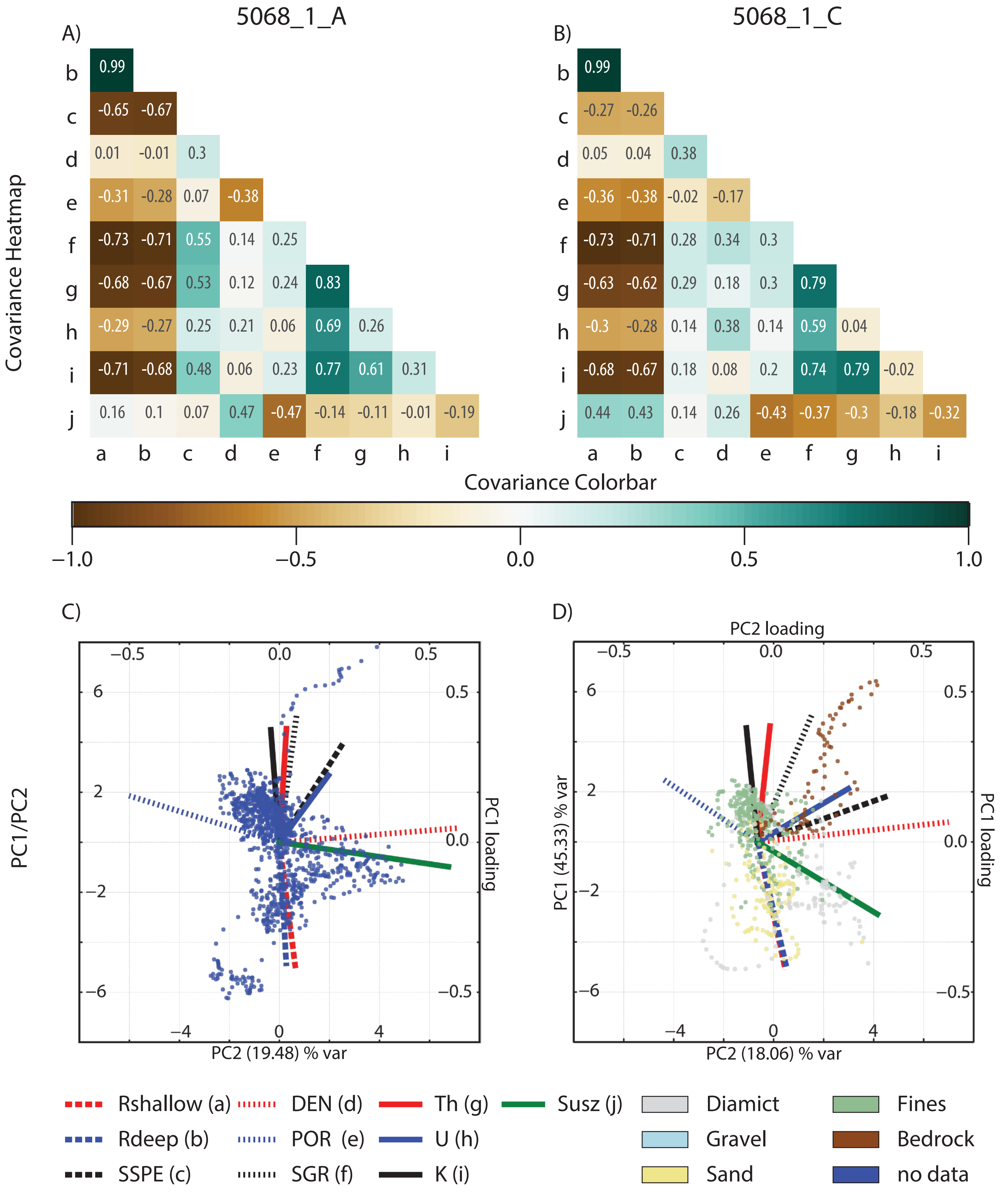


Figure 4 | Compilation and comparison of the two cleaned wireline datasets (from 5068_1_A and 5068_1_C) used for clustering. A and B): Heatmaps showing the variance between each log combination (a-j); 1.0-0.5: Strong positive correlation; 0.50-0.25 medium to weak positive correlation; 0.25- -0.25: weak to no correlation; -0.25 - -0.5: medium to weak negative correlation; -0.5 - -1.0: medium to strong negative correlation. C and D): Biplot of the first two principal components (primary x and y axis). D shows the visually-assigned major lithology from 5068_1_C. The integrated loading plots (secondary x and y axes scaled between ~0.7 and -0.7) indicate the relationship between individual logs (angle) and their contribution to the principal components (length).

of 5068_1_C. Integrating the litho-classification of the clusters with the dendrogram structure and the individual wireline log data distributions provides a more detailed characterization and understanding of the individual clustering and a direct link of log values with the sedimentological properties of stratigraphic units. The classification of value levels, such as high or low, is relative for each log and dataset due to the lack of cross-calibration between the datasets and lithological reference values.

The dendrogram reveals two principal branches converging at an observation distance of over 160. The first main branch consists of clusters CC0 and CC1, which converge at an observation distance of ~60. Both clusters display low resistivity (Rdeep and Rshallow), decreased density (DEN) and Susz, together with elevated SGR, Th, and K-values. The two clusters display slight differences in porosity, Th, and K values. CC0 shows a binomial distribution pattern in the resistivity logs. These clusters (CC0

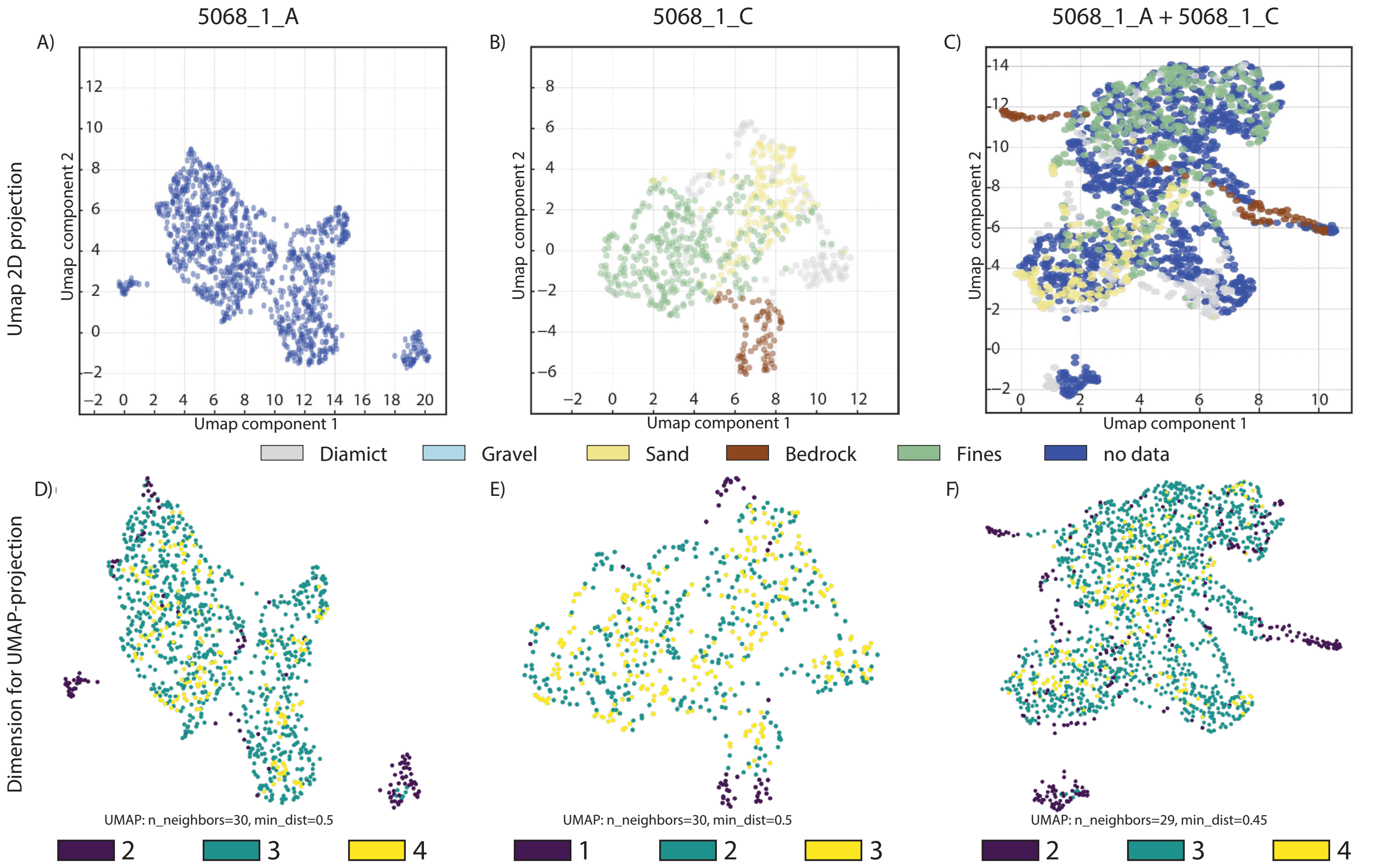


Figure 5 | UMAP results of the two individual and the combined wireline datasets. A-C): UMAP-projection from 10 to 2 dimensions. B-C): Points are color-coded with the visually assigned major lithology. D-F): The minimal number (color code is provided in the corresponding legend) of required dimensions (logs) to project each data point into a lower dimensional space while preserving as much individual variance between neighboring data points as possible. From left to right: 5068_1_A (A and D), 5068_1_C (B and E), and combined dataset of 5068_1_A and C (C and F).

Major Lithologies			Full Lithologies		
Lithotypes	**%**	**Data Points (n)**	**Lithotypes**	**%**	**Data Points (n)**
Bedrock (B)	10.7	74	B	10.7	74
Diamict (D)	18.8	130	Dmm	7.4	51
			Dms	2.4	17
			Dcm	0.0	0
			Mgt	7.8	54
			Mcc	1.1	8
Gravel (G)	0	0	Gm	0.0	0
Sand (S)	20.3	141	Sm	14.6	101
			Sl	5.8	40
			Sli	0.0	0
Fines (F)	50.2	348	Fm	13.1	91
			Fl	37.1	257

Table 3 | Distribution of geological ground-truth data (major and full lithology) of the cleaned dataset of 5068_1_C. Data in %, with numbers of corresponding data points (n; total = 693), color coding represents the used colors in the figures, for more information about the lithotypes see Table A2.

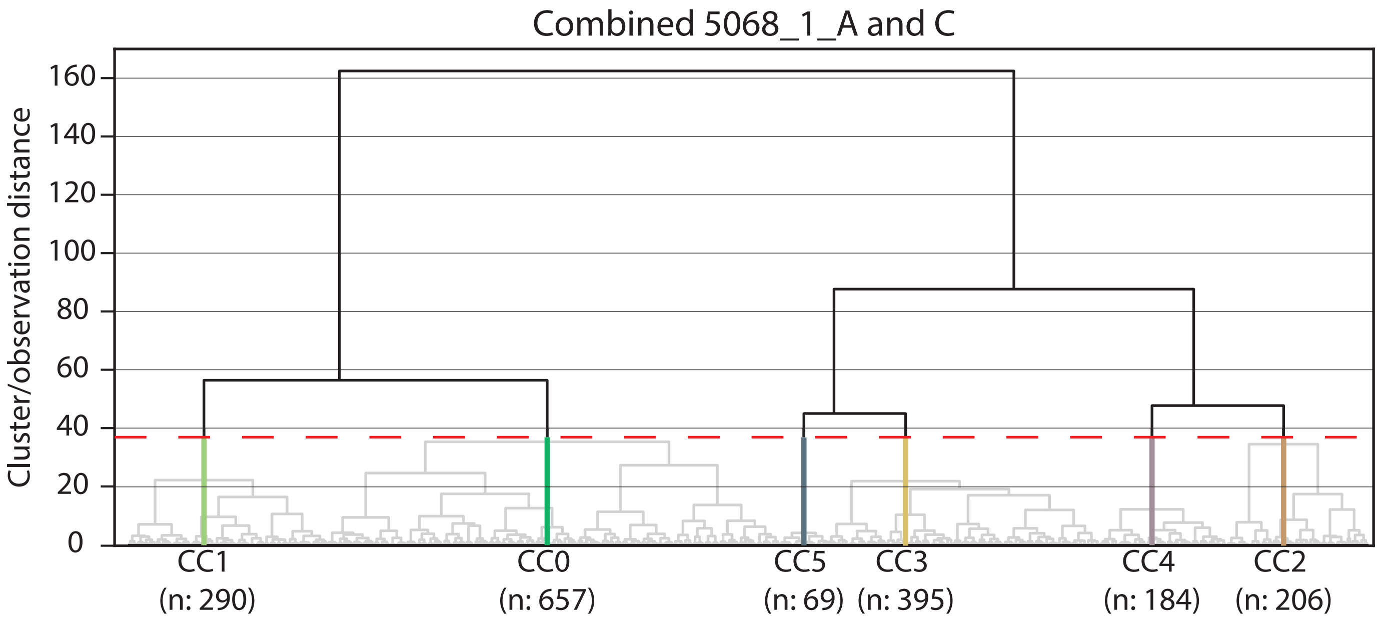


Figure 6 | Hierarchical cluster dendrograms of the combined clustering of 5068_1_A and C with 6 clusters (CC0-CC5). The y-axis displays the cluster/observation distance, a measure of the similarity between two neighboring clusters; the red dashed line indicates the approximate minimal cluster distance for the suggested number of clusters; semitransparent lines indicate the path of the merged clusters below the threshold. The x-axis displays the individual clusters and includes the total counts of data points belonging to each cluster. The colors of the cluster branches indicate the cluster colors used hereafter.

Clusters	5068_1_A + C (n =1801)		5068_1_A (n = 1108)		5068_1_C (n = 693)	
	%	Data points [n]	%	Data points [n]	%	Data points [n]
CC0	36.5	657	40.0	443	30.9	214
CC1	16.1	290	15.2	169	17.4	121
CC2	11.4	206	9.5	105	14.6	101
CC3	21.9	395	19.5	216	25.8	179
CC4	10.2	184	10.7	119	9.4	65
CC5	3.8	69	5.1	56	1.9	13

Table 4 | Data distribution of the individual clusters combined and after re-separation. Data distribution of the individual clusters of the combined data set (5068_1_A + C) and the two individual datasets (5068_1_A and 5068_1_C) after re-separation in % with the corresponding numbers of data points.

and CC1), representing nearly 45-50% of the data, coincide with fine-dominated lithologies in the litholog of 5068_1_C (Figure. 7A and 2N, 2O).

The second principal branch is formed in two steps: 1) the merging of four individual clusters, CC2–CC5, into pairs, CC5/CC3 and CC4/CC2, at observation distances of 45 and 50, respectively, and 2) the further merging these two pairs into one at an observation distance of ~90. There are notable divergences between CC2/CC4 and CC3/CC5, particularly in the distributions of density and partially in the resistivity logs. The other logs show overlapping distributions or varying patterns between these four clusters. Cluster CC3 shows a more widely distributed pattern than CC5, resulting in overlapping distributions with regard to density, porosity (POR), U-, and K-values. Further, CC3 shows elevated values for SGR, Th, Susz, and the short spectra photoelectric absorption (SSPE), as well as a reduction in resistivity compared to CC5. Cluster CC3 correlates with fine-rich sands (Figures 7A and 2G-K), whereas CC5 consists entirely of lithotype Mgt (Figure 7A), a subtype of diamict mainly composed of sand (Figure 2S and Table S1). Cluster CC2, dominantly consisting of Molasse bedrock (Figures. 7A and 2R-T), displays a broader distribution, particularly in the resistivity, density, SGR, Th-, U-, and K-logs, with some indications of several binomial distributions. This results in partial overlap with the distributions observed in CC4. Cluster CC4, linked to diamictic lithologies (Figures 7A and 2P-Q), also displays a broader distribution pattern, as evidenced by Susz, density, and resistivity logs, and exhibits elevated Susz and slightly diminished porosity, SGR, and Th-values relative to CC2.

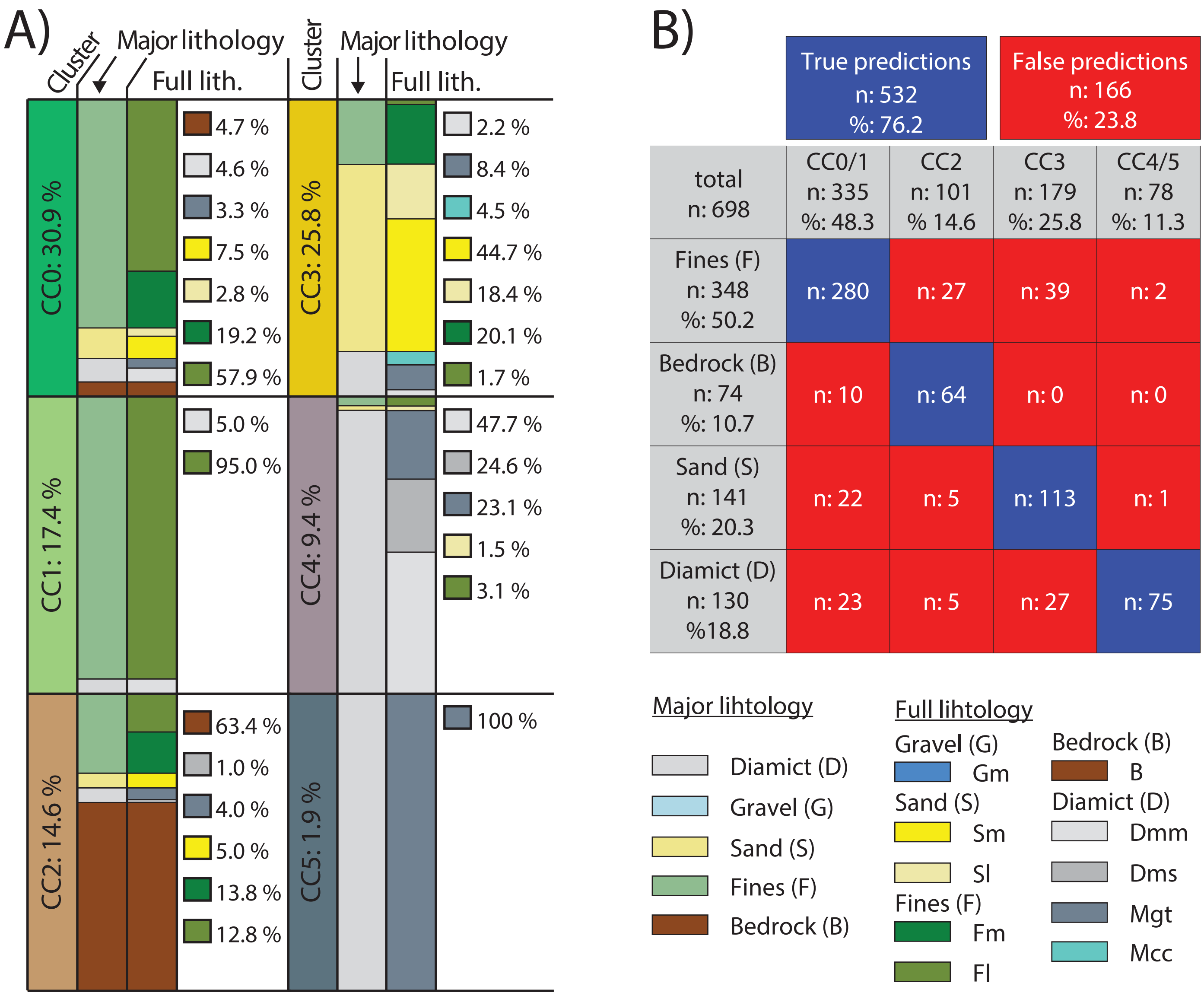
4.3. Comparison of lithological ground truth and cluster-based prediction

To quantify prediction quality, the six clusters were grouped based on their dominant major lithology (Figure 7A) an compared to the corresponding ground truth: 1) Fines (F) with CC0/CC1, 2) Bedrock (B) with CC2, 3) Sand (S) with CC3, and 4) Diamict (D) with CC4/CC5. The resulting confusion matrix of 5068_1_C (Figure 7B) shows an overall accuracy of 76.2% compared to the core-based ground truth. The prediction accuracy varies for the individual pairs: ~86% for bedrock, ~80% for the fines and sand, and ~57% for the diamicts (Table S8).

Overall, the stratified comparison between the lithological ground truth and the prediction for 5068_1_C shows strong agreement (Figure 8). The predicted cluster-based stratigraphy reveals major stratigraphic elements, which significantly contribute to the high overall prediction score, including 1) bedrock with its high prediction rate (CC2), 2) basal diamict bed that primarily contributes to the accurate predictions of this major lithology (CC4/CC5), 3) sand-dominated section with a solid prediction rate (CC3), and 4) the large-scale fine-dominated section, which was predicted with near-complete accuracy (CC0/CC1). Finer-scale structures, down to the m and sub-m scale, are also visible inside the major stratigraphic elements, such as the intercalated sand and fine-dominated layers. The prediction suggests that some features were not captured in the core-based lithological log, including a further split at the bedrock-basal diamictic contact (CC5) and some internal structures of the fine-dominated sections. False classifications are concentrated in areas of thinly interlayered sand and fines and sections of diamict. Furthermore, the reliability of the core-based lithological profile is based on the core quality, indicating areas of disturbed or missing data and interpolated sections. The overall strong prediction performance supports the use of the cluster-based profile of 5068_1_C as a reference in a cross-hole approach, enabling lithological extrapolation to boreholes such as 5068_1_A that lack direct core control.

4.4. Cross-borehole correlation

The combined clustering allows a correlation between the core-based lithological profile of 5068_1_C and the cluster-based stratigraphy of 5068_1_A. This is achieved by using the predicted cluster-based profile of 5068_1_C as a translation key between geological ground truth and



total n: 698	CC0/1 n: 335 %: 48.3	CC2 n: 101 % 14.6	CC3 n: 179 %: 25.8	CC4/5 n: 78 %: 11.3
Fines (F) n: 348 %: 50.2	n: 280	n: 27	n: 39	n: 2
Bedrock (B) n: 74 %: 10.7	n: 10	n: 64	n: 0	n: 0
Sand (S) n: 141 %: 20.3	n: 22	n: 5	n: 113	n: 1
Diamict (D) n: 130 %18.8	n: 23	n: 5	n: 27	n: 75

Figure 7 | Clustering and lithofacies prediction results. A) Cluster composition of 5068_1_C after combined clustering with 5068_1_A. The relative size of the clusters CC0–CC5 and their internal distribution of core-based individual major lithology and full lithology are given in %. B) Confusion matrix comparing each cluster's predicted dominant major lithology to the core-based labels of 5068_1_C. True predictions are shown in blue, and false ones in red. The legend provides the color codes for the used lithologies.

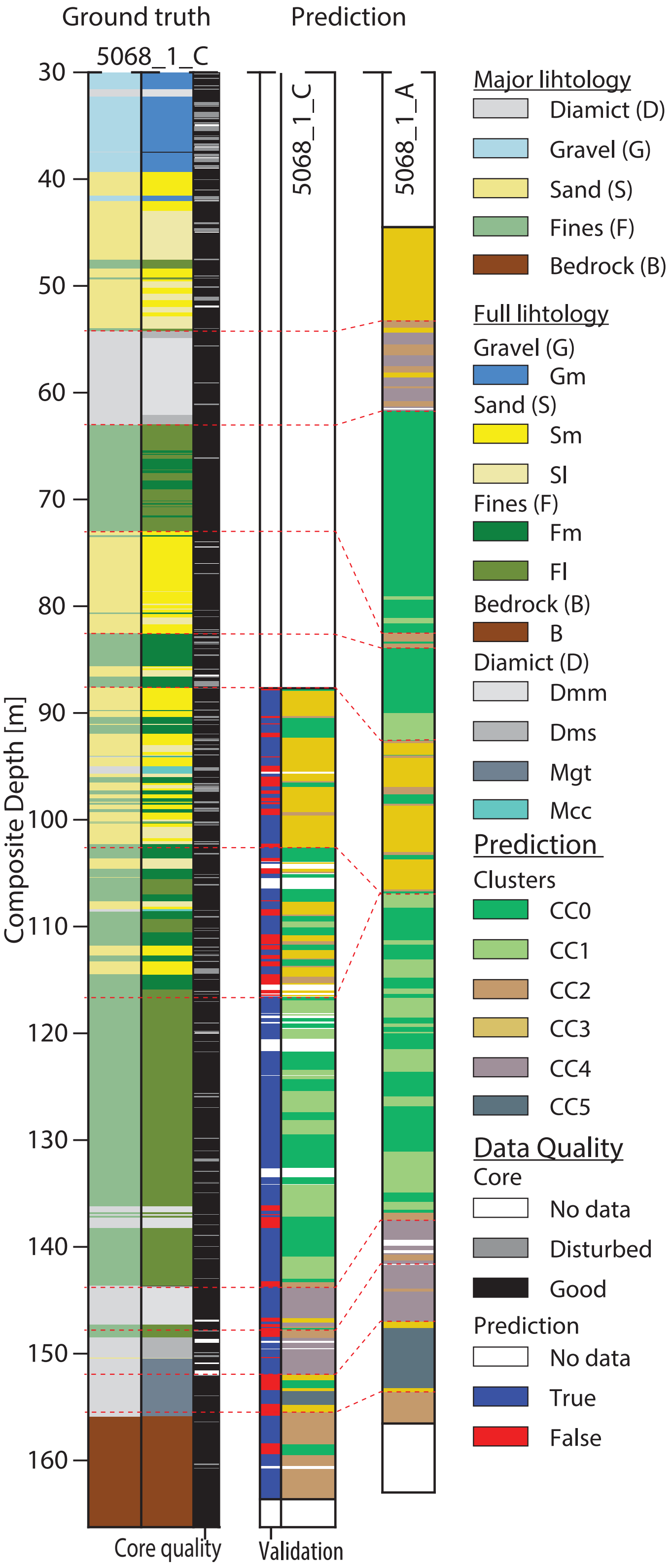


Figure 8 | Depth and cross-hole correlation of the lithology and cluster prediction. From left to right: 1) core-based Major and Full lithology for 5068_1_C with indicated core quality, 2) predicted lithology profiles for 5068_1_C with validation accuracy, and 3) predicted stratigraphy for 5068_1_A. Red dashed lines indicate the proposed stratigraphic correlation between lithology and cluster profile. The legend provides the color codes for the lithologies, clusters, and core- and prediction quality

data-based prediction (Figure 8). The cross-hole correlation between the predicted profiles of 5068_1_C and A reveals several stratigraphic trends: 1) thinning of the central sand-dominated section in 5068_1_A, located at ~73-116 m depth in 5068_1_C, 2) a near-horizontal correlation of the upper diamictic bed, situated at ~55-62 m depth in 5068_1_C, 3) increased thickness of the basal diamictic bed in 5068_1_A, and 4) more than a threefold increase in thickness of the upper fine-dominated section in 5068_1_A (~10 m vs. ~30 m).

5. Discussion

5.1. Importance of investigating the dataset structure

The importance of investigating and understanding the intrinsic data structures of the used datasets before confidentially combining them for dimensionality reduction and clustering is visualized by the rather surprising differences between the absolute values of the two individual wireline datasets (5068_1_A and 5068_1_C; Figure S2). These absolute differences are surprising since the two drill holes are only ~40 m apart, and a similar sedimentological composition can be assumed for both. Which can be considered as a typical heterogenic sedimentary composition for a glacially overdeepened setting (Figure 2 and Table 3; e.g., Buechi et al., 2018; 2024. Dehnert et al., 2012; Pomper et al., 2017). Both points, the individual internal heterogeneity and the common sedimentary composition of both boreholes, are supported by the matching patterns of the individual PCA and covariance analysis of the two separate datasets (Figure 4). Therefore, it is very likely that the reasons for the absolute wireline log values are due to shifts in the calibration of the logging tool before each measurement campaign. Further, based on the above points, the two datasets were individually normalized before being combined for clustering. This solution is further justified by the high similarity between the normalized wireline log values (Figure S3).

5.2. The case for dimensionality reduction

The combination of UMAP and hierarchical clustering provides a straightforward workflow for cluster analysis of time series data, relying on well-established and accepted methods in the data science community with a wide range of applications (UMAP: e.g., Becht et al., 2019; Cao et al., 2019; Hierarchical clustering: e.g., Belyadi & Haghighat, 2021). In the three wireline logging datasets examined, the two individual ones (A, C) and the combined one (A+C), UMAP introduces only a limited loss of information by reducing the 10-dimensional datasets to a 3-dimensional space (Figures 5D-F). This projection into a human-interpretable space allows an initial visual inspection and qualitative evaluation of the data distribution when combined with the corresponding lithological ground truth, visualized in two dimensions in Figures 5A-C. However, UMAP plots should be interpreted with caution. The distance between groups of data points and their shape and size may have no real meaning and highly depend on the selected hyperparameters, which should be chosen carefully for each dataset individually (e.g., n_neighbors and min_dist; McInnes et al., 2018; Coenen & Pearce, 2019).

Furthermore, these selected hyperparameters directly impact the suggested numbers of optimal clusters since the metrics use the UMAP projection as direct input. When comparing the explained/preserved variance of the UMAP reduction to that of PCA, UMAP outperforms PCA. For the three datasets, PCA explains between 75% and 77% of the variance (Table 2).

In contrast, UMAP preserves 100% of the variance for 5068_1_C and leads to some information loss in ~15% of the data points in the case of 5068_1_A and 13% of the data points in the combined dataset (5068_1_A and C). In all three cases, UMAP would maintain the complete variance if a reduction to four dimensions were applied (Figures 5D-F). However, reducing the data to four dimensions would hamper direct inspection of the visual data distribution. The superior performance of UMAP is likely due to its ability to detect nonlinear relationships in the datasets, unlike PCA, which is limited to linear combinations. Nevertheless, PCA and covariance analysis remain valuable tools for inspecting and understanding the data's internal structure since UMAP has no equivalent function to the PCA loading plots that would allow direct interpretations of the data. They are crucial to detect and exclude potential colinearities in the data (e.g., deep and shallow resistivity logs; Figure 4), and to reduce demand for computational resources, an aspect that gets critical in large datasets. This aspect was not further investigated in this study due to the relatively small size of the used dataset.

5.3. Cluster characterization and classification

Since there is no straightforward way to tell which of the input features were used for the UMAP dimension reduction, the proposed way is to compare the dendrogram (Figure 6) with the log-wise violin plots, which show the probability-density distributions for the individual clusters, to assume which logs most important for the projection (Figures S8 and S9). Some of these distributions indicate a more apparent separation of the data than others (Figure 9). For example, the SGR separates three groups of clusters (Figure 9A). In contrast, the U-log shows overlapping between all six clusters (Figure 9B). However, as UMAP indicates (Figure. 5F), no individual log leads to a clear separation of all clusters.

Furthermore, combining the information of the dendrogram (Figure 6) and the corresponding lithological composition (Figure 7A) helps to narrow down the dominant logs for cluster characterization. This combination directly links the petrophysical and chemical measurements and actual geology. Even without having a reference catalogue for typical wireline log values and uncertainties in the calibration, the plausibility of the individual cluster values can be evaluated for the assigned lithology. Additionally, individual PCA and covariance data analyses for each cluster may help define dominant features and similarities between clusters. They may help to visualize differences in the internal data distribution between separated clusters originating from combined datasets (see digital supplement). In cases of small cluster sizes (n ≤ 30; e.g., Fischer, 2011), the results should be treated with caution since they are most likely not statistically relevant anymore.

The re-divided clusters of both datasets show a substantial similarity, as indicated by the cluster data distribution, the PCA, and the covariance data. This similarity justifies the correlation of cluster-based profiles between the two boreholes. Consequently, the characterization and classification are primarily based on the 5068_1_C dataset to directly link the data-based clusters with the core-based geological ground truth. The assigned lithologies and the distributed logging data patterns support the proposed data-based linkage between the clusters. The two closely related clusters, CC0 and CC1, show reasonable log

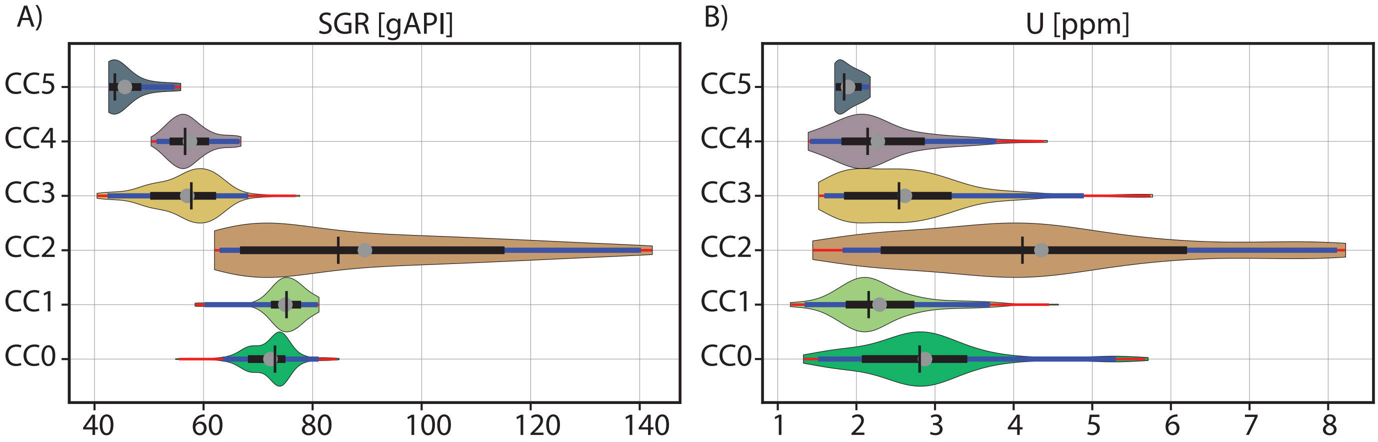


Figure 9 | Violin plot example of cluster data distribution of 5068_1_C. Visualizing a relatively clear separation (A) in the SGR and a rather unclear separation (B) in the U-log. Clusters are labeled on the y-axis and color-coded according to Table 4. Each violin plot includes the position of the mean (grey dot), the median (centered vertical black line), and the 1-, 2-, and 3-σ-ranges are displayed on the horizontal (1 σ: black, 2σ: black + red; 3σ: black + red + blue).

values for the assigned fine-dominated lithology, especially the high SGR log values. Their low resistivity and density suggest connected porosity, which aligns with the expected properties of these lithologies. Minor variations in sand or gravel content likely cause the small-scale log differences between the two clusters. While no clear multi-log trend explains the lower-level linkage between the remaining clusters, the assigned lithologies share some features supporting the clustering. CC3 and CC5, which are sand-dominated, have lower density and higher resistivity than CC2/CC4, indicating compaction and lower water content or permeability.

Further, the significant difference in the (SS)PE log of CC5 indicates a different lithological composition and likely a different origin. Although the lithological connection between the silt and sandstones of CC2 and the diamictic sediments of CC4 is surprising, it can be explained since both lithologies are mixtures of compacted or cemented sand and silt with some denser clasts or sections with lower porosity, yielding similar density logs. Additionally, the similar (SS)PE logs indicate a common sediment source, i.e., mostly reworked local bedrock, and the increased Susz-log of CC4 likely reflects the presence of crystalline clasts. Further, the moderate porosity with the wide range of the Gamma (SGR, U, Th, and K) logs matches the expected properties of a sand-siltstone mix with variable permeability, representing the local Molasses bedrock. UMAP likely projected CC0 and CC1 using a combination of resistivity, SGR, and density logs. However, due to the lack of a consistent trend among CC2-CC5 and some induced data blur by the projection into the three-dimensional space, identifying the key logs for UMAP projection is more challenging.

5.4. Quality of cluster-based prediction and stratigraphy

The lithofacies prediction for 5068_1_C achieved a solid overall accuracy of approximately 76% (Figure 7B), successfully capturing most of the lithologies identified through visual classification. This strong agreement is also displayed in the close visual match between the automated "cluster stratigraphy" and the core-based profile of 5068_1_C (Figure 8). The cluster-based stratigraphy reveals major stratigraphic packages: 1) bedrock, 2) basal diamict, 3) a central fines-dominated section, 4) a sand-dominated section, 5) interlayered sand- and fines-dominated layers. Except for the latter, these elements contribute significantly to the high overall prediction score (Figure 8). In addition, the model resolves internal structures within these packages, such as: 1) a fines-rich interval within the bedrock (likely claystone), 2) internal variation within the basal diamict, distinguishing between likely reworked bedrock (CC6) and overlying diamict (CC5), a feature undocumented in the visual core description, 3) variability within the fines-dominated section, and 4) the presence of interlayered sand and fines at dm-scale.

Most misclassifications are likely linked to the differing nature of the applied classification schemes. The core-based lithology is predominantly based on qualitative visual assessment of the dominant grain size and texture. In contrast, the clustering approach is strictly data-driven and sensitive to subtle variations in the petrophysical signal, particularly arising towards the center of the projected data point cloud (Figures 5, S4, and S5). These discrepancies explain confusion between sand and fines, or fines and bedrock.

In the case of diamicts, discrete classification is generally challenging due to their bi-modal grain-size distribution and textural variation. Moreover, classification often includes assumed depositional processes, criteria not entirely detectable by petrophysical data. This is reflected in the distribution of full-lithology prediction accuracy (Table S9). For example, 100% of Mcc and ~30% of Mgt are assigned to CC3, while ~25% of Mgt forms a separate cluster (CC6). These two clusters show a first-level connection in the dendrogram (Figure 6), and CC6 was lithologically described as mostly loose sand (Figure 2S).

Overall, the accuracy and quality of the predicted cluster-based stratigraphy depend on: 1) the quality and resolution of the wireline-log data, 2) the classification scheme and consistency of the lithological ground truth, and 3) the accuracy of the depth correlation between the core and log data. High-quality cores are critical for detecting small-scale features, determining the resolution of cluster composition, and minimizing uncertainty. The model's demonstrated ability to detect major stratigraphic packages and internal structures with robust accuracy underlines the high potential of this workflow for establishing a wireline-logging data-based stratigraphy for unconsolidated Quaternary sediments.

5.5. Cross-hole correlation of cluster-based and core-based stratigraphies

The strong agreement between the core-based and cluster-based stratigraphy of 5068_1_C supports its use as a reference for cross-hole correlation with 5068_1_A. Using the core-controlled stratigraphy as a translation key between clusters and lithology (Figure 7A), the detailed lithological interpretation of 5068_1_C can be extended to 5068_1_A, where only wireline logs are available, enabling a high-resolution lithostratigraphic prediction (Figure 8).

In addition, this approach enables direct correlation between the cluster-based stratigraphy of 5068_1_A and the core-based stratigraphy of 5068_1_C, including intervals where wireline data are missing due to casing or poor log quality. For example, the upper diamictic bed and underlying sand unit are present in the core of 5068_1_C but not covered by wireline data. While the diamictic bed is clearly reproduced in the cluster-based prediction for 5068_1_A, the sand unit is absent. This absence is likely real, as the lower sand bed is consistently predicted for both wells. The selected wireline logs, especially the SGR and, to some extent, its components (U, Th, K), proved valuable for detecting major trends in the encountered setting. The relatively small-scale stratigraphic variations highlight the complex, patchwork-like stratigraphy typical of glacially overdeepened deposits. These lithostratigraphic variations at the study site were also identified in the seismic cross-hole experiments done by Beraus et al. (2024, 2025). Log selection should, however, be adapted to the specific geological system and study objectives. For example, if integration with seismic survey data is planned, such as in core-to-seismic correlations, including the sonic log is recommended. The sonic log enables a robust connection between geological predictions, synthetic seismograms, vertical seismic profiles, and seismic data.

The study highlights the potential of the workflow for investigating subsurface geology in complex settings such as glacially overdeepened valleys. Its ability to resolve major and minor stratigraphic elements beyond the resolution of conventional geophysical surveys, such as seismic and gravimetry, while capturing fine-scale internal structure and linking sedimentological observations to petrophysical data, provides critical ground truth. By integrating these data, such models can serve as a geological equivalent to Building Information Models (BIM) commonly used in civil engineering.

5.6. Workflow stability and further developments

To evaluate the stability of the workflow, qualitative stability tests were performed, including: 1) varying the random state in the UMAP projection, 2) using four dimensions in the UMAP projection to preserve 100% of the intrinsic data variance (Figure 5F), and 3) testing different clustering linkage methods (Ward, Average, and Complete). Varying the random state resulted in minor fluctuations in prediction accuracy (~75–78%). Adding a fourth UMAP dimension improved prediction accuracy to ~80%, regardless of the selected random state. These variations are likely explained by numerical randomness during the nonlinear dimensionality reduction and clustering methods. Thus, a 2–3% reduction in prediction accuracy was considered an acceptable trade-off to preserve a visually interpretable 3D projection. Similarly, varying the linkage method had only a limited impact on clustering outcomes.

However, the variability observed under the tested conditions is negligible compared to the potential uncertainties introduced by: 1) the visually classified geological ground truth, 2) the overall input data quality, or 3) variations in hyperparameter selection. To address these sources of uncertainty, several improvements can be considered: 1) selecting a visual core classification scheme focusing on petrophysically detectable criteria to improve lithology log quality, 2) applying feature engineering techniques (e.g., Fourier or wavelet transformations) to reduce noise in the wireline signal, and 3) automating the testing of various combinations of workflow parameters (e.g., random states, UMAP dimensionality, linkage methods, and hyperparameter ranges) to identify the optimal settings. These points would require further development of the current code, ideally in combination with a graphical user interface. Additionally, the workflow could be expanded and tested with different types of sequential (geoscientific) data as input, such as X-ray fluorescence (XRF) or Multi-Sensor Core Logger (MSCL) data.

Beyond these practical improvements, the workflow could benefit from advances in geological data acquisition and interpretation. These include: 1) automated core description and lithofacies classification for improved ground-truthing, 2) integration of advanced logging techniques such as "Logging While Drilling", and 3) establishing a lithology–log value catalogue to support automated cluster–lithology correlation.

6. Conclusion

This study presents a robust, modular, and reproducible workflow for lithofacies prediction, combining unsupervised clustering with dimensionality reduction in a data-driven approach. The workflow is currently designed to analyze core and wireline log data and aims to minimize the subjectivity and operator bias of classical core descriptions. However, due to its modular and open design, the workflow is also adaptable to other forms of sequential (geoscientific) data.

The case study demonstrates the workflow's potential to correlate cored and uncored wells by identifying cluster-based signal patterns that correspond to geological ground truth, thereby achieving a prediction accuracy of ~76%. This approach enables detailed, data-driven lithostratigraphic models even when direct correlations between wireline logs and lithology are unavailable. As with all data-driven approaches, the output quality depends heavily on input quality, making input data quality optimization a critical aspect and a central point for further improvements. Despite these constraints, especially when integrated with geophysical survey data (e.g., seismic), the workflow can significantly contribute to the development of high-resolution geological models.

Acknowledgements

We are grateful for the continued support by the International Scientific Continental Drilling Program (ICDP) and the DOVE science team. A special thanks go to the technical staff of LIAG, who executed the downhole logging surveys and preprocessed the data. We thank the ICDP Operational Support Group, especially Simona Pierdominici and Jochem Kück, for providing support with the logging data corrections and handling, and the team of the Data Science Lab from the University of Bern for their technical and methodological support while developing the presented workflow. We acknowledge the strong support from the local inhabitants and the local authorities of the municipality of Ingoldingen in the district of Biberach during the drilling and logging campaigns. We thank the two anonymous reviewers and the executive editor, Murray Gingras, for their constructive comments and suggestions, significantly improving the original manuscript. This research has been supported by the ICDP, the Deutsche Forschungsgemeinschaft (DFG, grant nos. KR2073/3-1, BU 2467/1-2, GA749/5-1, BU 2467/3-1, BU 3894/2-1, BU 3894/3-1, and PR 957/6-1), Nagra, ENSI, LGRB, LFU, LIAG, BOKU Vienna, and University of Bern.

Author contribution

SS developed and implemented the presented workflow, processed and analyzed the data, and was the paper's main author with the support of all co-authors. DM and PB provided support during the development and implementation of the workflow. BS contributed to the initial analyses of the drill core and provided scientific input. MSA contributed to the (field) acquisition and initial preprocessing of the wireline logging data and provided, together with SB, scientific input regarding the wireline logging data. MWB and FSA provided significant scientific input to the paper as part of their role as PhD supervisors of SS. All authors approved the text and the figures.

Data availability

All data used and displayed in this study are publicly available according to the FAIR principle. The DOVE operational dataset (DOVE-Phase 1 Scientific Team et al., 2023b), containing all data concerning the drill core, is available on the ICDP DOVE project website: https://www.icdp-online.org/projects/by-continent/europe/dove-switzerland or can be accessed on the GFZ-library (https://doi.org/10.5880/ICDP.5068.001) together with the operational report (DOVE-Phase 1 Scientific Team et al., 2023a) and the explanatory remarks (DOVE-Phase 1 Scientific Team et al., 2023c). Other used data, the supplementary code, and the digital supplement are accessible under the following link: https://github.com/schallersebastian/Lithoprediction-with-UMAP

Conflict of interest

The authors declare that they have no known competing personal relationships or financial interests that could have appeared to influence the work reported in this paper.

References

Abbas, K. A., Gharavi, A., Hindi, N. A., Hassan, M., Alhosin, H. Y., Gholinezhad, J., Ghoochaninejad, H., Barati, H., Buick, J., Yousefi, P., Alasmar, R., & Al-Saegh, S. (2023). Unsupervised machine learning technique for classifying production zones in unconventional reservoirs. International Journal of Intelligent Networks, 4, 29–37. https://doi.org/10.1016/j.ijin.2022.11.007

Anselmetti, F., & Eberli, G. (1999). The Velocity-Deviation Log: A Tool to Predict Pore Type and Permeability Trends in Carbonate Drill Holes from Sonic and Porosity or Density Logs. Aapg Bulletin - AAPG BULL, 83, 450–466. https://doi.org/10.1306/00AA9BCE-1730-11D7-8645000102C1865D

Anselmetti, F. S., Bavec, M., Crouzet, C., Fiebig, M., Gabriel, G., Preusser, F., Ravazzi, C., & DOVE scientific team. (2022). Drilling Overdeepened Alpine Valleys (ICDP-DOVE): Quantifying the age, extent, and environmental impact of Alpine glaciations. Scientific Drilling, 31, 51–70. https://doi.org/10.5194/sd-31-51-2022

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. Nature Biotechnology, 37(1), 38–44. https://doi.org/10.1038/nbt.4314

Belyadi, H., & Haghighat, A. (2021). Chapter 4 - Unsupervised machine learning: Clustering algorithms. In H. Belyadi & A. Haghighat (Eds.), Machine Learning Guide for Oil and Gas Using Python (pp. 125–168). Gulf Professional Publishing. https://doi.org/10.1016/B978-0-12-821929-4.00002-0

Beraus, S., Burschil, T., Buness, H., Köhn, D., Bohlen, T., & Gabriel, G. (2024). A comprehensive crosshole seismic experiment in glacial sediments at the ICDP DOVE site in the Tannwald Basin. Scientific Drilling, 33(2), 237–248. https://doi.org/10.5194/sd-33-237-2024

Beraus, S., Köhn, D., Bohlen, T., Burschil, T., Schuster, B., Buness, H., & Gabriel, G. (2025). Seismic crosshole full-waveform inversion of high-frequency SV-waves for glacial sediment characterization. Geophysical Prospecting, 73(5), 1587–1605. https://doi.org/10.1111/1365-2478.70024

Bergamo, P., Fäh, D., Panzera, F., Cauzzi, C., Glueer, F., Perron, V., & Wiemer, S. (2023). A site amplification model for Switzerland based on site-condition indicators and incorporating local response as measured at seismic stations. Bulletin of Earthquake Engineering, 21(13), 5831–5865. https://doi.org/10.1007/s10518-023-01766-z

Buechi, M. W., Graf, H. R., Haldimann, P., Lowick, S. E., & Anselmetti, F. S. (2018). Multiple Quaternary erosion and infill cycles in overdeepened basins of the northern Alpine foreland. Swiss Journal of Geosciences, 111(1–2), 133–167. https://doi.org/10.1007/s00015-017-0289-9

Buechi, M. W., Landgraf, A., Madritsch, H., Mueller, D., Knipping, M., Nyffenegger, F., Preusser, F., Schaller, S., Schnellmann, M., & Deplazes, G. (2024). Terminal glacial overdeepenings: Patterns of erosion, infilling and new constraints on the glaciation history of Northern Switzerland. Quaternary Science Reviews, 344, 108970. https://doi.org/10.1016/j.quascirev.2024.108970

Burschil, T., Buness, H., Tanner, D. C., Wielandt-Schuster, U., Ellwanger, D., & Gabriel, G. (2018). High-resolution reflection seismics reveal the structure and the evolution of the Quaternary glacial Tannwald Basin. Near Surface Geophysics, 16(6), 593–610. https://doi.org/10.1002/nsg.12011

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics, 3(1), 1–27. https://doi.org/10.1080/03610927408827101

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., Trapnell, C., & Shendure, J. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature, 566(7745), 496–502. https://doi.org/10.1038/s41586-019-0969-x

Carrasquilla, A. (2023). Lithofacies prediction from conventional well logs using geological information, wavelet transform, and decision tree approach in a carbonate reservoir in southeastern

Brazil. Journal of South American Earth Sciences, 128, 104431. https://doi.org/10.1016/j.jsames.2023.104431

Coenen, A., & Pearce, A. (2019). Understanding UMAP. Understanding UMAP. https://pair-code.github.io/understanding-umap/

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2), 224–227. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.1979.4766909

Dehnert, A., Lowick, S. E., Preusser, F., Anselmetti, F. S., Drescher-Schneider, R., Graf, H. R., Heller, F., Horstmeyer, H., Kemna, H. A., Nowaczyk, N. R., Züger, A., & Furrer, H. (2012). Evolution of an overdeepened trough in the northern Alpine Foreland at Niederweningen, Switzerland. Quaternary Science Reviews, 34, 127–145. https://doi.org/10.1016/j.quascirev.2011.12.015

Di Martino, A., Carlini, G., Castellani, G., Remondini, D., & Amorosi, A. (2023). Sediment core analysis using artificial intelligence. Scientific Reports, 13(1), 20409. https://doi.org/10.1038/s41598-023-47546-2

Dixit, N., McColgan, P., & Kusler, K. (2020). Machine Learning-Based Probabilistic Lithofacies Prediction from Conventional Well Logs: A Case from the Umiat Oil Field of Alaska. Energies, 13(18), Article 18. https://doi.org/10.3390/en13184862

DOVE-Phase 1 Scientific Team, Anselmetti, F. S., Beraus, S., Buechi, M. W., Buness, H., Burschil, T., Fiebig, M., Firla, G., Gabriel, G., Gegg, L., Grelle, T., Heeschen, K., Kroe- mer, E., Lehne, C., Lüthgens, C., Neuhuber, S., Preusser, F., Schaller, S., Schmalfuss, C., Schuster, B., Tanner, D. C., Thomas, C., Tomonaga, Y., Wieland-Schuster, U., and Wonik, T. (2023a). Drilling Overdeepened Alpine Valleys (DOVE) – Operational Report of Phase 1, (ICDP Operational Report), GFZ German Research Centre for Geosciences, Potsdam, 70 pp., https://doi.org/10.48440/ICDP.5068.001

DOVE-Phase 1 Scientific Team, Anselmetti, F. S., Beraus, S., Buechi, M. W., Buness, H., Burschil, T., Fiebig, M., Firla, G., Gabriel, G., Gegg, L., Grelle, T., Heeschen, K., Kroe- mer, E., Lehne, C., Lüthgens, C., Neuhuber, S., Preusser, F., Schaller, S., Schmalfuss, C., Schuster, B., Tanner, D. C., Thomas, C., Tomonaga, Y., Wieland-Schuster, U., and Wonik, T. (2023b). Drilling Overdeepened Alpine Valleys (DOVE) – Operational Dataset of DOVE Phase 1, GFZ Data Services [data set], https://doi.org/10.5880/ICDP.5068.001

DOVE-Phase 1 Scientific Team, Anselmetti, F. S., Beraus, S., Buechi, M. W., Buness, H., Burschil, T., Fiebig, M., Firla, G., Gabriel, G., Gegg, L., Grelle, T., Heeschen, K., Kroe- mer, E., Lehne, C., Lüthgens, C., Neuhuber, S., Preusser, F., Schaller, S., Schmalfuss, C., Schuster, B., Tanner, D. C., Thomas, C., Tomonaga, Y., Wieland-Schuster, U., and Wonik, T. (2023c). Drilling Overdeepened Alpine Valleys (DOVE) – Explanatory remarks on the operational dataset, ICDP Operational Dataset – Explanatory Remarks, GFZ German Research Centre for Geosciences, Potsdam, 34 pp., https://doi.org/10.48440/ICDP.5068.002

Dramsch, J. S. (2020). Chapter One—70 years of machine learning in geoscience in review. In B. Moseley & L. Krischer (Eds.), Advances in Geophysics (Vol. 61, pp. 1–55). Elsevier. https://doi.org/10.1016/bs.agph.2020.08.002

Ellwanger, D., Wielandt-Schuster, U., Franz, M., & Simon, T. (2011). The Quaternary of the southwest German Alpine Foreland (Bodensee-Oberschwaben, Baden-Württemberg, Southwest Germany). E&G Quaternary Science Journal, 60(2/3), 306–328. https://doi.org/10.3285/eg.60.2-3.07

Fischer, H. (2011). A History of the Central Limit Theorem: From Classical to Modern Probability Theory. Springer. https://doi.org/10.1007/978-0-387-87857-7

Ghosh, S. (2022). A review of basic well log interpretation techniques in highly deviated wells. Journal of Petroleum Exploration and Production Technology, 12(7), 1889–1906. https://doi.org/10.1007/s13202-021-01437-2

Giardini, D., Guidati, G. (eds.), Amann, F., Driesner, T., Gischig, V., Guglielmetti, L., Hertrich, M., Holliger, K., Krause, R., Laloui, L., Lateltin, O., Lecampion, L., Löw, S., Maurer, H., Mazzotti, M., Meier, P., Moscariello, A., Saar, M.O., Spada, M., … Zappone, A. (2021). Swiss Potential for Geothermal Energy and CO2 Storage, Synthesis Report. ETH Zurich. https://doi.org/10.3929/ethz-b-000518184

Guntli, P., Keller, F., Lucchini, R., & Rust, S. (2016). Gotthard-Basistunnel: Geologie, Geotechnik, Hydrogeologie – zusammenfassender Schlussbericht (Vol. 7). Schweizerische Eidgenossenschaft, Bundesamt für Landestopografie (swisstopo). ISBN: 978-3-302-40105-8

Hasan, M. M. U., Hasan, T., Shahidi, R., James, L., Peters, D., & Gosine, R. (2023). Lithofacies Identification from Wire-Line Logs Using an Unsupervised Data Clustering Algorithm. Energies, 16(24), Article 24. https://doi.org/10.3390/en16248116

Honer, P. C., & Sherrell, F. W. (1977). The application of air-flush rotary percussion drilling techniques in site investigation. Quarterly Journal of Engineering Geology and Hydrogeology, 10(3), 207–220. https://doi.org/10.1144/GSL.QJEG.1977.010.03.04

Krammer, K., & Pohl, J. (1987). The susceptibility probe suslog 403-1. In KTB Report 87-2: Grundlagenforschung und Bohrlochgeophysik; Beiträge zur Tagung der Deutschen Geophysikalischen Gesellschaft in Clausthal-Zellerfeld (31.3. - 4.4.1987) (pp. 399–409). Projektleitung Kontinentales Tiefbohrprogramm der Bundesrepublik Deutschland im Niedersächsischen Landesamt für Bodenforschung. https://doi.org/10.48440/KTB.87-2_24

Lai, J., Su, Y., Xiao, L., Zhao, F., Bai, T., Li, Y., Li, H., Huang, Y., Wang, G., & Qin, Z. (2024). Application of geophysical well logs in solving geologic issues: Past, present and future prospect. Geoscience Frontiers, 15(3), 101779. https://doi.org/10.1016/j.gsf.2024.101779

Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. Geoscience Frontiers, 7(1), 3–10. https://doi.org/10.1016/j.gsf.2015.07.003

Lauper, B., Zimmerli, G. N., Jaeggi, D., Deplazes, G., Wohlwend, S., Rempfer, J., & Foubert, A. (2021). Quantification of Lithological Heterogeneity Within Opalinus Clay: Toward a Uniform Subfacies Classification Scheme Using a Novel Automated Core Image Recognition Tool. Frontiers in Earth Science, 9. https://doi.org/10.3389/feart.2021.645596

Marjoribanks, R. (2010). Diamond Drilling. In R. Marjoribanks (Ed.), Geological Methods in Mineral Exploration and Mining (pp. 99–136). Springer. https://doi.org/10.1007/978-3-540-74375-0_7

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software, 3(29), 861. https://doi.org/10.21105/joss.00861

Mondol, N. H. (2015). Well Logging: Principles, Applications and Uncertainties. In K. Bjørlykke (Ed.), Petroleum Geoscience: From Sedimentary Environments to Rock Physics (pp. 385–425).

Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-34132-8_16

Mukherjee, B., & Sain, K. (2021). Vertical lithological proxy using statistical and artificial intelligence approach: A case study from Krishna-Godavari Basin, offshore India. Marine Geophysical Research, 42(1), 3. https://doi.org/10.1007/s11001-020-09424-8

Nielsen, F. (2016). Hierarchical Clustering (pp. 195–211). https://doi.org/10.1007/978-3-319-21903-5_8

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

Pomper, J., Salcher, B. C., Eichkitz, C., Prasicek, G., Lang, A., Lindner, M., & Götz, J. (2017). The glacially overdeepened trough of the Salzach Valley, Austria: Bedrock geometry and sedimentary fill of a major Alpine subglacial basin. Geomorphology, 295, 147–158. https://doi.org/10.1016/j.geomorph.2017.07.009

Popescu, M., Head, R., Ferriday, T., Evans, K., Montero, J., Zhang, J., Jones, G., & Kaeng, G. C. (2021, November 23). Using Supervised Machine Learning Algorithms for Automated Lithology Prediction from Wireline Log Data. SPE Eastern Europe Subsurface Conference. https://doi.org/10.2118/208559-MS

Preusser, F., Graf, H. R., Keller, O., Krayss, E., & Schlüchter, C. (2011). Quaternary glaciation history of northern Switzerland. E&G Quaternary Science Journal, 60(2/3), 282–305. https://doi.org/10.3285/eg.60.2-3.06

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Schaller, S., Buechi, M. W., Schuster, B., & Anselmetti, F. S. (2023). Drilling into a deep buried valley (ICDP DOVE): A 252\,m long sediment succession from a glacial overdeepening in northwestern Switzerland. Scientific Drilling, 32, 27–42. https://doi.org/10.5194/sd-32-27-2023

Schlumberger. (1972). Log interpretation charts. Schlumberger-Doll Research Center: Schlumberger Technology Coporation.

Schuster, B., Gegg, L., Schaller, S., Buechi, M. W., Tanner, D. C., Wielandt-Schuster, U., Anselmetti, F. S., & Preusser, F. (2024). Shaped and filled by the Rhine Glacier: The overdeepened Tannwald Basin in southwestern Germany. Scientific Drilling, 33(2), 191–206. https://doi.org/10.5194/sd-33-191-2024

Selley, R. C., & Sonnenberg, S. A. (2023). 3—Methods of Exploration. In R. C. Selley & S. A. Sonnenberg (Eds.), Elements of Petroleum Geology (Fourth Edition) (Fourth Edition, pp. 43–166). Academic Press. https://doi.org/10.1016/B978-0-12-822316-1.00003-3

Serra, O., & Abbott, H. T. (1982). The Contribution of Logging Data to Sedimentology and Stratigraphy. Society of Petroleum Engineers Journal, 22(01), 117–131. https://doi.org/10.2118/9270-PA

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Zekri, H., Cohen, D., Rutherford, N., Folkes, C., & Thomas, M. (2025). Rapid analysis of drill core data for detection of geological boundaries. Journal of Geochemical Exploration, 269, 107634. https://doi.org/10.1016/j.gexplo.2024.107634

How to cite: Schaller, S., Becerra, P., Beraus, S., Buechi, M. W., Mair, D., Sardar Abadi, M., Schuster, B., & Anselmetti, F. S. (2025). Data-driven lithofacies prediction of unconsolidated sediments from wireline logs. Sedimentologika, 3(1), 1-18. https://doi.org/10.57035/journals/sdk.2025.e31.1819