Latin American Journal of Sedimentology
and Basin Analysis

ISSN 1851-4979
Published by the Asociación Argentina de Sedimentología

ACCEPTED MANUSCRIPT

Predictive analysis of total organic carbon (TOC) in shale targets: example from the Lower Cretaceous of the Austral Basin (Patagonia, Argentina) using machine learning on outcrop data

Sebastián Richiano, *Federico Ares*

Received date: 5 March 2025
Accepted date: 24 April 2025
Available online: 6 June 2025

Handling Editor: José I. Cuitiño

Please cite this article as

Richiano, S., and Ares, F. Predictive analysis of total organic carbon (TOC) in shale targets: example from the Lower Cretaceous of the Austral Basin (Patagonia, Argentina) using machine learning on outcrop data (in press). *Latin American Journal of Sedimentology and Basin Analysis*.

This is a preliminary version of the manuscript accepted for publication in the Latin American Journal of Sedimentology and Basin Analysis. This version will be revised before final publication. Please note that some errors may be found during the final revision process. The same disclaimers as for printed and final versions apply for this early online version.

Predictive analysis of total organic carbon (TOC) in shale targets: example from the Lower Cretaceous of the Austral Basin (Patagonia, Argentina) using machine learning on outcrop data

Sebastián Richiano[1,2*], Federico Ares[1,2]

1 – Instituto Patagónico de Geología y Paleontología (CONICET-CENPAT), Boulevard Almirante Brown 2915, ZC: U9120ACD, Puerto Madryn, Chubut, Argentina.

2 – Universidad Nacional de la Patagonia San Juan Bosco, Boulevard Almirante Brown 3051, ZC: U9120ACD, Puerto Madryn, Chubut, Argentina.

* Corresponding author richiano@cenpat-conicet.gob.ar

ABSTRACT

The Río Mayer Formation (Lower Cretaceous) of the Austral Basin, Patagonia, is a key source rock for unconventional reservoirs. This study explores the potential of machine learning (ML) for predicting Total Organic Carbon (TOC) content using outcrop data, a novel approach compared to traditional subsurface data applications. Employing dimensional reduction techniques (PCA, T-SNE, UMAP), the analysis revealed clear clustering of high TOC values in feature space, supporting the feasibility of predictive modeling. Three ML models—Logistic Regression, Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN)—were tested using a feature set derived from ANOVA F-Score rankings. Dimensionality reduction improved model performance, with SVC achieving the most robust results. Despite limited labeled samples, predictions across models were consistent, identifying a promising region for high TOC. The study highlights the importance of integrating geological variables and XRD data in TOC modeling and emphasizes the need for expanded datasets and additional sedimentary sections to enhance regional interpretations.

Keywords: Unconventional reservoirs, Machine learning, TOC prediction, Austral Basin, Dimensional reduction

INTRODUCTION

In recent years, the use of machine learning methods to analyze, model, and predict various aspects of oil-bearing rocks has increased. These methods have been used to predict Total Organic Carbon (TOC; Handhal *et al*., 2020; Saporetti *et al*., 2022), distribution of facies associations (Tognoli *et al*., 2024), rock brittleness (Guo *et al*., 2022; Mustafa *et al*., 2022; Ore and Gao, 2023), hydrocarbon production predictions (Prochnow *et al*., 2022), and for reservoir characterization (Niu *et al*., 2022). These studies are based on subsurface data, using cores, cuttings, or petrophysical (wireline) data. Given the importance of analog outcrop data for petroleum system characterization (Busch *et al*., 2022), generating machine learning models from outcrop data is highly significant. However, scientific studies that integrate field and subsurface data using machine learning methods remain scarce (Milad *et al*., 2020). The TOC values represent the amount of organic carbon preserved in a rock sample, and are often used to estimate the type of hydrocarbon produced and/or retained, and it defines the possibility of that rock to be a source rock for hydrocarbons (more than 1%) (Passey *et al*., 1990; Handhal *et al*., 2020; Saporetti *et al*., 2022).

The Austral Basin is the southernmost oil-producing basin in Argentina (Fig. 1). Initially, oil production came from conventional reservoirs. However, in the last decade the basin has been intensely explored for unconventional reservoirs (e.g. Belotti *et al*., 2013; 2014). The Río Mayer Formation (=Palermo Aike Formation in subsurface) constitutes the main exploration target for unconventional reservoirs in the basin (Rojas *et al*., 2022; Melendo *et al*., 2023, and references therein). This unit is primarily composed of black shales, with thinly interbedded marls and sandstones (e.g., Richiano *et al*., 2012). Unconventional shale reservoirs must possess various characteristics, but foremost are high TOC, rock brittleness, significant stratigraphic thickness, and broad areal distribution. The analysis of TOC is critical in oil-exploration, and efforts have been made to measure it at lower costs and in less time-consuming ways (e.g., Handhal *et al*., 2020; Saporetti *et al*., 2022).

In this paper, machine learning methods are applied for the first time on the Río Mayer Formation, a shale target for non-conventional reservoirs. In addition, this work is one of the few available models based on outcrop data. In the next sections we describe the unit and the data previously published, then we describe the methods applied and run the database to finally model the TOC prediction. Taking the above into consideration, the objectives of this contribution are: (1) to use machine learning to model the distribution of sedimentological features (observed in the field), the mineralogical composition, and the TOC from selected samples; (2) to develop a workflow for predicting TOC values in samples where measurements are missing; and (3) to assess the accuracy of different mathematical models applied in this case study.

GEOLOGICAL SETTING

The Austral-Magallanes Basin (Jurassic to Cenozoic) is located in the southernmost part of Patagonia, Argentina (Cuitiño *et al.*, 2019) (Fig. 1). The basin was initiated by Late Jurassic extension associated with the El Quemado Complex (equivalent to the Tobífera Formation) syn-rift sequence (Féraud *et al.*, 1999; Pankhurst *et al.*, 2000). During subsequent transgression, the continental to shallow marine Springhill Formation (Tithonian to Berriasian) was deposited (Kraemer and Riccardi, 1997; Richiano *et al.*, 2016). During the Berriasian, the transgression continued, leading to the deposition of the Río Mayer Formation, marking the onset of post-rift (sag) conditions (Arbe, 2002). This unit mainly comprises black shales with fossiliferous levels indicating Berriasian-Albian deposition (i.e., Kraemer and Riccardi, 1997; Aguirre Urreta, 2002). The outcrops of the Río Mayer Formation are covered transitionally from north to south by the Piedra Clavada (=Kachaike), Lago Viedma and Cerro Toro formations during the Aptian/Albian (Richiano *et al.*, 2012; Cuitiño *et al.*, 2019).

At the Seccional Río Guanaco locality (Fig. 1) the Río Mayer Formation is *ca*. 400 m thick and was previously subdivided into three informal sections (Richiano *et al.*, 2012). The lower section is dominated by laminated black shales interbedded with marls, with abundant ammonites and belemnites, interpreted as deposited in an outer shelf setting (Richiano *et al.*, 2012). This section has the highest TOC content of the Río Mayer

Formation, ranging between 0.07 and 2.81% (Richiano, 2014). The middle section is 40 m thick and it is composed of intensely bioturbated dark marls and shales, characterized by trace fossils of the *Zoophycos* Ichnofacies (Richiano *et al.*, 2013; Richiano, 2015). The TOC in the middle part of the section is very low (< 0.58%; Richiano, 2014). The upper section is composed of massive black shales intercalated with very fine- to fine-grained sandstones, interpreted as an outer shelf with distal low-density turbidity current deposits (Richiano *et al.*, 2012). In this section, the *Zoophycos* Ichnofacies was also reported (Richiano *et al.*, 2013; Richiano, 2015). This section shows moderate TOC values at the base (0.5-2 %, average 1.12 %) and extremely low values towards the top. The frequent intercalation of sandstones in the uppermost part of the section is related to the distal influence of the deltaic deposition whose lithologic expression at the basin margins is the Piedra Clavada Formation to the north (Richiano *et al.*, 2012; 2015).

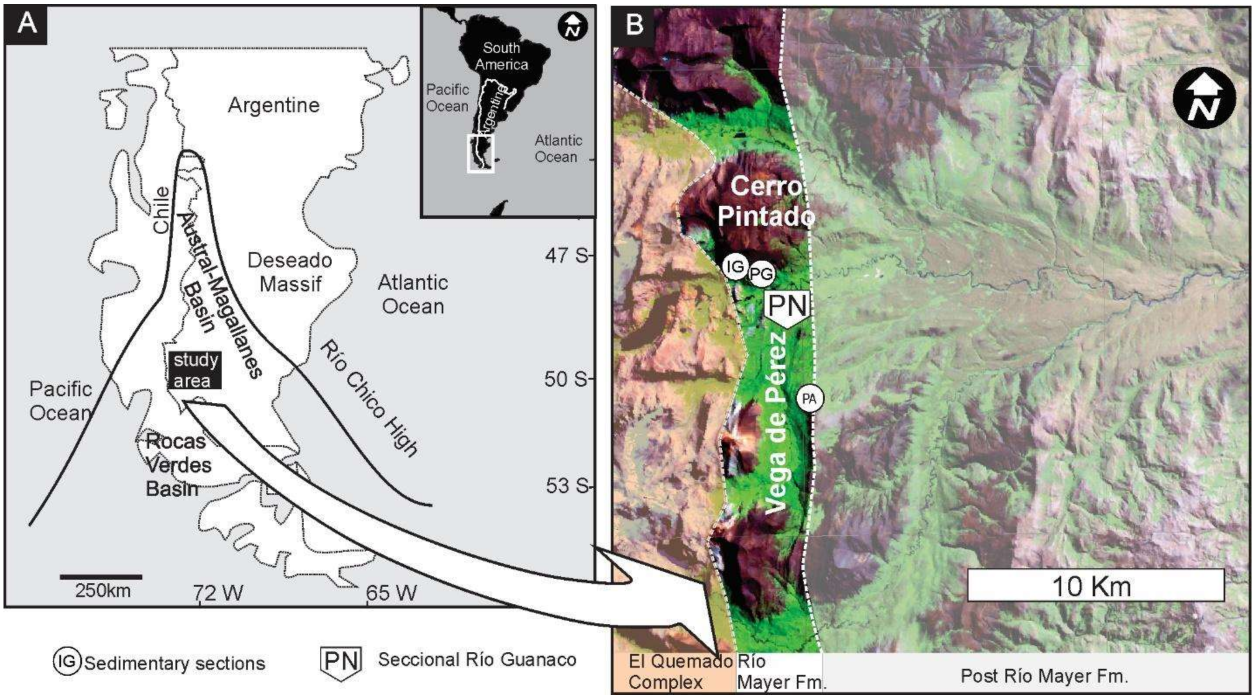

Figure 1. Location of the study area. **a)** General map of the Austral Basin in southern Patagonia. **b)** Position of the Seccional Río Guanaco (PN) of the Los Glaciares National Park, related stratigraphy and sedimentary sections used (full information at Richiano *et al.*, 2012, 2015, 2019). Modified from Richiano *et al.* (2019).

DATABASE AND METHODS

Sedimentary sections and samples analyzed

Three sedimentary sections of the Lower Cretaceous Río Mayer Formation were selected for sedimentological, mineralogical and geochemical analyses (Sections IG, PG, PA; Fig. 2). A total of 106 fine-grained rock samples from Río Mayer Formation were collected and analyzed (Fig. 2). The initial step involved converting the outcrop data into numerical values. In this sense, three "field parameters" were assigned to each sample. First, following the methodology used by Poiré *et al.* (2007), numerical values were assigned to sedimentary facies, wherein different values characterized the sedimentary texture and sedimentary structures (*i.e*., fabric). Secondly, different codes were applied to the sedimentary environments interpreted, using one (1) for outer shelf deposits and two (2) for outer shelf deposits influenced by deltaic environments. Finally, the last parameter is the bioturbation for which we use a binary discrimination between non-bioturbated (0) and bioturbated (1). The full sedimentary facies analysis, ichnology, and the compositional dataset used in this work are available in Richiano *et al.* (2012; 2013; 2015; 2019).

The X-ray diffraction (XRD) characteristics of the samples were conducted on an X-PANalytical model X´Pert PRO diffractometer located at the Centro de Investigaciones Geológicas (CONICET-UNLP, Argentina). The radiation source used was Cu/Ni, and the generation settings were set at 40 kV and 40 mA. For the whole-rock analysis, semi-quantification was obtained from the intensity of the main peak for each mineral (Schultz, 1964; Moore and Reynolds, 1997). Clay mineralogy was determined from diffraction patterns obtained using samples that were air-dried, ethylene glycol-solvated and heated to 550ºC for 2 h (Brown and Brindley, 1980).

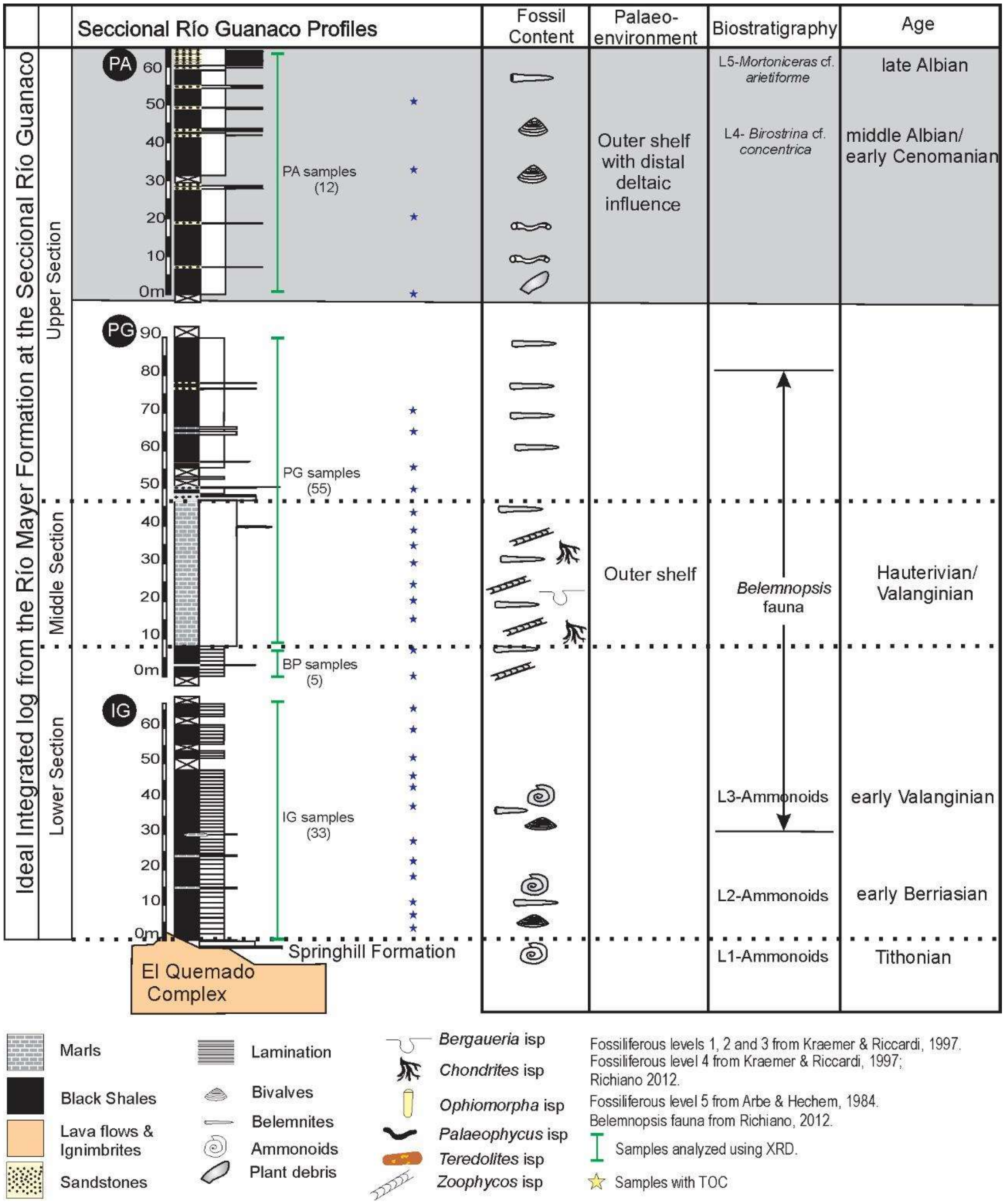

Figure 2. Sedimentary logs of the Río Mayer Formation at the Seccional Río Guanaco locality (profiles IG, PG, PA located in figure 1). XRD: x-ray diffraction analysis; TOC: Total Organic Carbon. Modified from Richiano *et al.* (2019).

Geochemical studies of the samples from the Río Mayer Formation include 17 samples analyzed for major, minor, trace elements, and rare earth elements (REE) by X-ray fluorescence spectrometry (XRF) and Inductively Coupled Plasma mass spectrometry (ICP-

MS) measurements performed by ACTLABS (Ontario, Canada). In addition, 28 samples were assessed to determine trace element composition at Centro de Investigaciones Geológicas laboratories (CONICET-UNLP, Argentina). These samples were treated by dissolving the silicates in acid, and analyzed using a Perkin-Elmer ICP-MS fitted with a Meinhardt concentric nebulizer. Finally, 29 TOC values were obtained from within five lateral meters from the collected outcrop profile. TOC was ascertained by Geolab Sur S.A. (Buenos Aires, Argentina).

The raw dataset consists of a total of 106 samples, including 103 samples with XRD analysis (103 whole rock and 101 of clay), 45 geochemical analyses and 29 samples of TOC (Table 1).

Data Analysis

Data analysis and modeling were performed using Python, primarily using commonly available libraries such as NumPy, pandas, matplotlib, seaborn, umap-learn, and scikit-learn. The dataset's features were grouped into two categories: 'geo' for geological data (3 features) and 'xrd' for X-ray diffraction data (10 features). Although X-ray fluorescence and ICP-MS data were included in the dataset, the sample size was too small for robust analysis. Missing XRD data were imputed using the population mean.

Exploratory Data Analysis (EDA) was conducted to investigate the dataset, applying dimensionality reduction techniques like PCA, T-SNE, and UMAP to project the feature space into two dimensions. Feature scaling was consistently applied using the StandardScaler class from scikit-learn. This process enabled the identification of high and low TOC areas, which are needed for TOC modeling. Visualizations of both labeled and unlabeled data (with and without TOC measurements) provide insight to the geological variability within and across sections.

Sample	Geological field data					X-Ray Diffraction										TOC
	Lithology		Bioturbation Index	Environment		Whole Rock						Clays				
	Sed. Fac.	NF		1	2	Qz	Pl	FK	Ca	Py	Arc	I	IS	Cl	K	
PA-15	Pm	12	0	0	1	6	3	1	1	1	3	32	19	49	0	
PA-14	Pm	12	0	0	1	6	3	1	3	1	3	35	16	49	0	0,09
PA-13	Pm	12	0	0	1	6	3	1	2	1	3	39	11	49	0	
PA-11	Pm	12	0	0	1	6	3	1	3	1	3	30	15	55	0	
PA-10	Pm	12	0	0	1	6	3	1	3	0	3	28	21	51	0	
PA-9	Pm	12	0	0	1	6	4	1	3	0	3	38	14	48	0	0,09
PA-8	Pm	12	0	0	1	6	3	1	2	1	3	28	20	51	0	
PA-6	Pm	12	0	0	1	6	3	1	2	1	3	29	30	41	0	0,09
PA-5	Pm	12	0	0	1	6	4	1	1	1	3	30	19	51	0	
PA-3	Pm	12	0	0	1	6	4	1	1	1	4	36	5	59	0	
PA-2	Pm	12	0	0	1	6	3	1	1	1	3	32	9	58	0	0,09
PA-1	Pm	12	0	0	1	6	3	1	1	1	3	30	24	46	0	
PG 64	Pl	10	0	1	0	6	2	1	2	0	2	41	24	35	0	
PG 60	Pl	10	0	1	0	6	3	1	4	0	3	34	21	45	0	1,48
PG 59	Pl	10	0	1	0	6	2	1	1	0	2	41	31	27	0	
PG 55	Pm	12	0	1	0	6	3	1	2	0	3	39	23	38	0	0,6
PG 54	Pm	12	0	1	0	6	3	1	2	0	3	45	14	41	0	
PG 52	Pm	12	0	1	0	6	2	2	2	0	3	26	34	40	0	
PG 51	Pm	12	0	1	0	6	2	1	3	0	3	22	34	44	0	
PG 50	Pm	12	0	1	0	6	1	1	4	0	3	24	40	36	0	1,81
PG 48	Pm	12	0	1	0	6	1	1	3	0	3	25	24	51	0	
PG 45	Pm	12	0	1	0	6	3	1	1	0	2	34	13	52	0	
PG 44	Gg	40	0	1	0	5	4	1	6	1	4	20	20	60	0	
PG 43	Sm	30	0	1	0	4	5	1	5	0	3	56	0	44	0	
PG 42	Pm	12	0	1	0	6	1	1	2	0	2	22	21	57	0	0,62
PG 41	Sl	33	0	1	0	5	5	1	2	1	4	66	2	32	0	
PG 40	Pm	12	0	1	0	6	3	1	1	1	4	61	11	28	0	
PG 39	Sl	33	0	1	0	4	5	1	3	1	4	85	0	15	0	
PG 38	Mb	13	1	1	0	6	3	1	1	1	4	22	0	78	0	
PG 37	Mb	13	1	1	0	6	4	1	1	0	4	71	0	29	0	
PG 36	Mb	13	1	1	0	6	3	1	2	1	3	37	12	51	0	
PG 35	Mb	13	1	1	0	6	3	1	3	0	3	38	23	39	0	0,58
PG 34	Mb	13	1	1	0	6	2	1	1	0	4	29	6	65	0	
PG 32	Mb	13	1	1	0	6	4	1	4	1	3	22	27	51	0	
PG 31	Mb	13	1	1	0	6	1	1	4	1	3	41	22	37	0	
PG 30	Mb	13	1	1	0	5	2	1	6	1	3	20	37	43	0	0,09
PG 29	Mb	13	1	1	0	6	3	1	5	1	3	61	17	23	0	
PG 28	Mb	13	1	1	0	6	1	1	5	0	2	36	27	35	2	
PG 27	Mb	13	1	1	0	6	3	1	4	0	3					
PG 26	Mb	13	1	1	0	6	1	1	5	0	3	33	30	36	0	
PG 25	Mb	13	1	1	0	6	1	1	4	1	2					
PG 24	Mb	13	1	1	0	6	3	1	4	0	3	66	18	16	0	0,17
PG 23	Mb	13	1	1	0	6	1	1	4	1	3	27	25	48	0	
PG 22	Mb	13	1	1	0	6	2	1	5	0	3	65	15	20	0	
PG 21	Mb	13	1	1	0	6	2	1	4	1	2	59	24	17	0	
PG 20	Mb	13	1	1	0	6	1	1	4	1	2	57	27	16	0	
PG 19	Mb	13	1	1	0	5	2	1	5	0	3	68	20	11	0	0,09
PG 18	Mb	13	1	1	0	6	1	1	5	1	2	43	47	9	1	
PG 17	Mb	13	1	1	0	5	2	1	5	0	2	85	12	3	1	
PG 16	Mb	13	1	1	0	6	2	1	4	1	2	64	23	13	0	
PG 15	Mb	13	1	1	0	5	2	1	5	0	3	61	20	11	8	0,07

PG 14	Mb	13	1	1	0	6	1	1	5	1	2	69	13	14	4	
PG 13	Mb	13	1	1	0	6	3	1	4	1	2	61	26	10	4	
PG 12	Mb	13	1	1	0	6	1	1	5	0	3	33	40	27	0	
PG 11	Mb	13	1	1	0	6	2	1	5	0	2	72	13	13	2	
PG 10	Mb	13	1	1	0	6	2	1	4	1	2	52	27	21	0	
PG 9	Mb	13	1	1	0	6	2	1	5	1	2	51	29	10	10	0,09
PG 8	Mb	13	1	1	0	6	1	1	4	1	3	62	23	15	0	
PG 7	Mb	13	1	1	0	6	1	1	5	1	3	37	34	19	10	
PG 6	Mb	13	1	1	0	6	2	1	4	1	3	50	26	18	6	
PG 5	Mb	13	1	1	0	6	2	1	5	0	2	71	19	11	0	0,13
PG 4	Mb	13	1	1	0	6	2	1	4	1	3	59	26	11	3	
PG 3	Mb	13	1	1	0	6	1	1	4	1	3	26	40	34	0	
PG 2	Mb	13	1	1	0	6	2	1	4	1	3	60	21	17	2	
PG 1	Mb	13	1	1	0	6	1	1	5	1	3	39	37	18	5	
BP 5	Pb	10	1	1	0	6	1	1	4	0	3	55	27	16	2	0,17
BP 4	Pb	10	1	1	0	6	2	1	3	0	3	40	32	28	0	
BP 3	Sm	30	1	1	0	6	5	1	5	0	3	6	17	76	0	
BP 2	Pb	10	1	1	0	6	3	1	4	0	3	26	29	41	4	
BP 1	Pb	10	1	1	0	6	1	1	4	0	2	41	27	23	9	0,31
🟩	🟩	🟩				🟩										
IG 44	Pl	10	0	1	0	6	2	1	1	0	3	37	40	20	3	1,49
IG 43	Pl	10	0	1	0	6	2	1	3	0	3	34	40	20	6	
IG 42	Pl	10	0	1	0	6	2	1	2	1	2	41	35	20	4	1,59
IG 41	Pl	10	0	1	0	5	1	1	5	0	2	48	34	12	6	
IG 40	Pl	10	0	1	0	6	1	1	4	0	2	34	44	15	7	2,44
IG 39	Pl	10	0	1	0	6	1	1	1	0	3	38	50	9	3	2,09
IG 37 - 38	Pl	10	0	1	0	6	2	1	5	0	3	46	37	16	0	1,88
IG 36	Pl	10	0	1	0	6	1	1	1	0	3	45	39	13	4	
IG 34 - 35	Pl	10	0	1	0	6	2	1	4	0	3	36	51	13	0	1,65
IG 32 - 33	Pl	10	0	1	0	6	2	1	1	1	3	36	43	21	0	
IG 30 - 31	Pl	10	0	1	0	6	2	1	3	0	2	35	44	18	3	
IG 29	Pl	10	0	1	0	3	1	1	6	0	2	52	31	12	6	
IG 27 - 28	Pl	10	0	1	0	6	2	1	2	0	3	40	39	17	4	1,56
IG 26	Pl	10	0	1	0	6	1	1	1	1	3	44	37	15	4	
IG -24-25	Pl	10	0	1	0	6	2	1	3	1	3	29	40	21	10	
IG 23	Mm	12	0	1	0	6	2	1	3	1	3	42	30	25	3	1,43
IG 22	Mm	12	0	1	0	6	2	1	3	0	3	41	37	20	2	
IG 20 - 21	Pl	10	0	1	0	6	1	1	3	0	3	31	42	24	4	
IG 18 - 19	Pl	10	0	1	0	6	1	1	3	0	3	50	32	17	2	
IG 17	Pl	10	0	1	0	6	1	1	5	0	3	34	42	20	4	2,81
IG 16	Mm	12	0	1	0	4	1	1	6	1	2	38	30	21	11	
IG14 - 15	Pl	10	0	1	0	6	1	1	1	0	3	46	36	16	2	
IG 12 - 13	Pl	10	0	1	0	6	2	1	3	0	3	47	28	25	0	1,52
IG 10 - 11	Pl	10	0	1	0	6	2	1	3	0	3	64	23	11	3	
IG 9	Pl	10	0	1	0	6	2	1	3	0	3	48	39	13	0	1,59
IG 8	Pl	10	0	1	0	6	2	1	3	0	3	59	26	15	0	
IG 7	Pl	10	0	1	0	6	1	1	3	0	3	46	32	15	7	
IG 6	Pl	10	0	1	0	6	1	1	3	0	3	57	28	15	0	
IG 5	Pl	10	0	1	0	6	1	1	1	0	3	46	33	21	0	
IG 4	Pl	10	0	1	0	6	1	1	1	1	3	60	19	21	0	0,07
IG 3	Cm	12	0	1	0	2	1	1	6	1	2	37	29	29	4	
IG 2	Cm	12	0	1	0	2	2	1	6	1	2	77	18	3	2	
IG 1	Cm	12	0	1	0	2	1	1	6	0	2	64	24	6	6	

Table 1. Data set used for modeling. References: Facies: P: Mudstone; M: Marl; S: Sandstone; G: Conglomerate/Sabulite; C: Carbonatic; m: massive; l: laminated; b: bioturbated; g: glauconitic. NF: numerical facies. Environment: 1- outer shelf, 2- outer shelf influenced by deltas; XRD: x-ray diffraction; Qz: Quartz, Pl:

Plagioclase; FK: K-feldespar; Ca: Calcite; Py: Pyrite; Arc: total clays; I: Illite; IS: Illite-Smectite; Cl: Chlorite; K: Kaolinite; TOC: Total Organic Carbon.

Data Preprocessing

Figure 3a illustrates the data preprocessing workflow. A threshold of >1 was applied to the measured TOC values to create a binary classification, with 1 indicating High TOC and 0 indicating Low TOC. This threshold resulted in a nearly balanced dataset for training the classification model, removing the necessity for additional techniques (e.g., precision/recall metrics, over- or under-sampling) to address dataset imbalance.

In order to address feature space dimensionality during TOC modeling (13 features for only 29 labeled samples), a feature selection process was implemented based on a feature importance metric. Several methods for computing feature importance were attempted (Random Forest, LinearSVC, Lasso and ANOVA). The quality of High/Low TOC separation in the feature space was evaluated using the silhouette coefficient (Rousseeuw, 1987). Feature selection was performed using an additive approach, prioritizing features in descending order of importance while monitoring changes in the silhouette score. The addition of features was halted upon observing a significant decline in the silhouette score.

Feature ranking by ANOVA F-Score and the computed silhouette are shown, as well as the impact of feature selection on sample spatial distribution in reduced dimension (T-SNE) in the results section.

TOC Modeling

The TOC modeling workflow is shown in figure 3b. A candidate model is fitted to a labeled dataset consisting of a set of selected features and their corresponding discretized High-TOC labels. In order to test the hypothesis (*i.e.* modeling of TOC is possible) we run three classification methods: Logistic Regression, Support Vector Classifier (SVC) and K-Nearest Neighbours Classifier (KNN). These models have different working principles, one is a parametric baseline model (Logistic regression), another is a state-of-the-art parametric model (SVC) and the last is a non- parametric method (k-NN). All of them are available in the scikit-learn library.

Optimal hyperparameters for each model were determined using a combination of Bayesian Hyperparameter Search and Leave-One-Out (LOO) cross-validation, with the mean accuracy of the left-out sample in each split serving as the performance metric. This method ensured the selection of hyperparameters that maximize predictive accuracy on unseen data.

Following hyperparameter selection, train/test accuracy plots were visually inspected to assess variance and bias, serving as indicators of potential overfitting or underfitting. The model was then retrained using the selected hyperparameters on all labeled samples and applied to predict the High-TOC content for both labeled and unlabeled samples.

TOC Prediction

The prediction workflow (Figure 3c) uses the full dataset, including both labeled and unlabeled samples, as input to the selected model and generates predictions for TOC values. These predictions can be either continuous (probability of High TOC) or discrete (High TOC probability > 0.5). Model predictions can be visualized in a reduced-dimensional space (*e.g.*, T-SNE) to identify regions in the feature space associated with a high probability of High TOC.

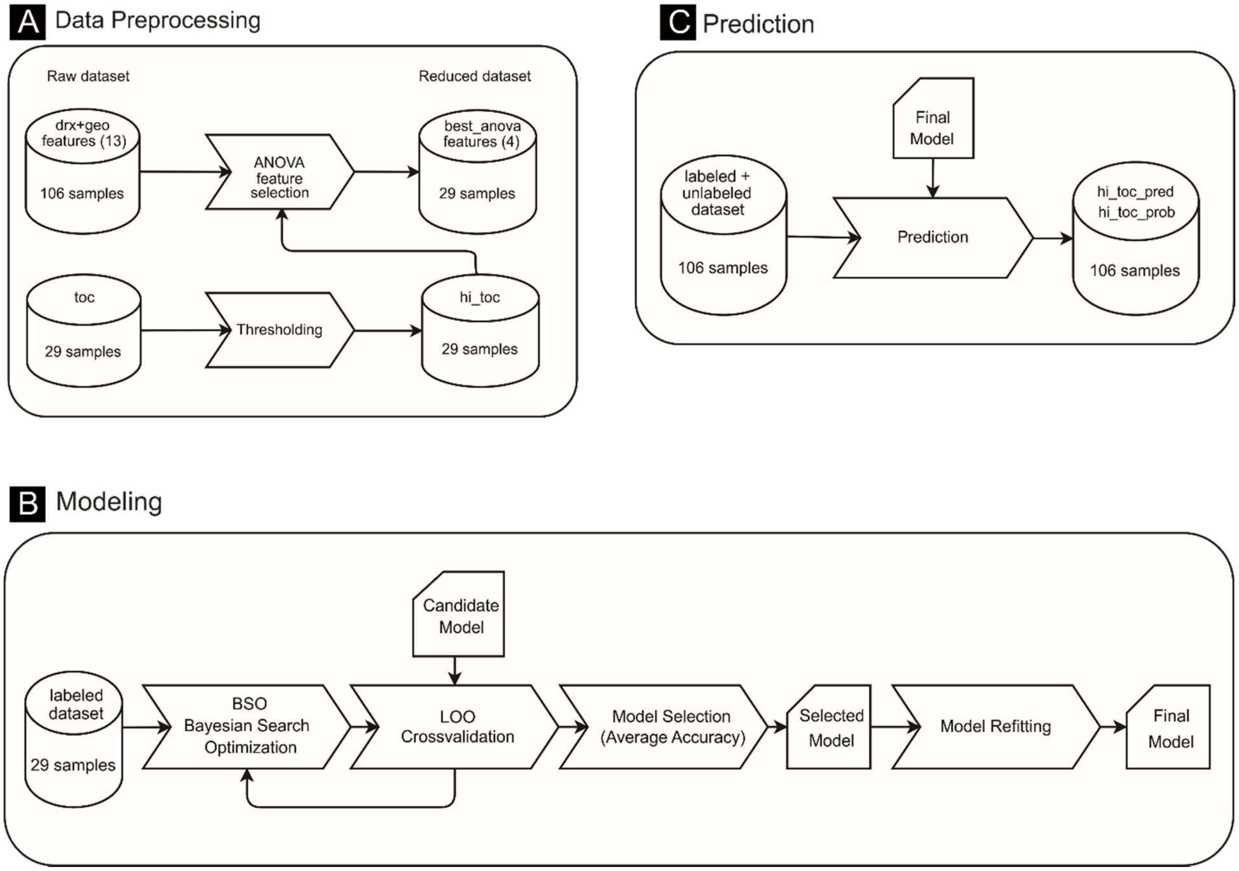

Figure 3. Modeling Workflow. a) Data Preprocessing Workflow. Continuous TOC values are thresholded to obtain a binary High TOC label. Using this categorical label, an additive feature selection process is performed by using ANOVA F-Score. Exploratory Data Analysis using dimensional reduction on feature space (all xrd+geo features). b) TOC Modeling Workflow. By using a labeled dataset, a candidate model is fitted. Model hyperparameters are adjusted by performing a Bayesian Search Optimization (BSO), and the model's generalization performance is estimated with Leave-One-Out (LOO) cross validation. The model with best average test accuracy is selected and refitted on the full dataset. c) High TOC Prediction Workflow. Using the trained model, High TOC is predicted for all samples (labeled and unlabeled), identifying potential regions of interest in feature space.

RESULTS

This work focuses on the mathematical modeling of the data needed to predict TOC contents. In this sense, sedimentological, ichnological, environmental and/or compositional data can be found in Richiano *et al.* (2012; 2013; 2015; 2019).

Exploratory Data Analysis (EDA)

Figure 4 presents the results of dimensionality reduction methods (PCA, T-SNE, UMAP). High TOC values cluster in a specific region corresponding to part of the PG section and nearly all labeled IG samples. In PCA space, the PG and IG sections significantly overlap, while PA is partially isolated with minor overlap with unlabeled PG samples. In T-SNE space, PA is distinctly isolated, and PG and IG show greater contrast. UMAP space reveals clear section contrasts: PA appears isolated in the lower left, while PG and IG show slight overlap in the lower right.

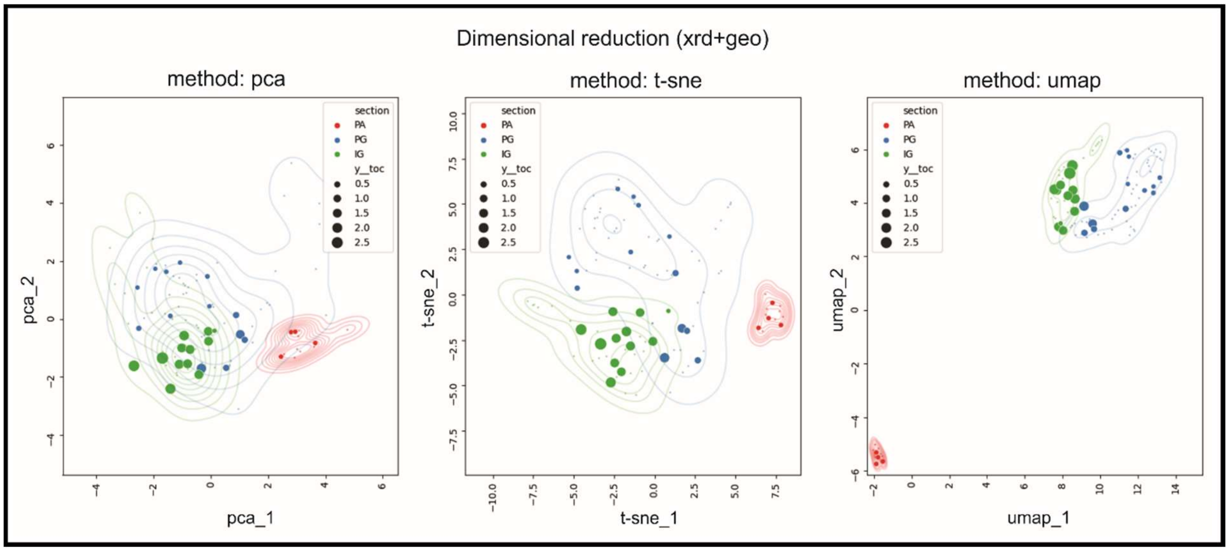

Figure 4. Exploratory Data Analysis via Dimensional Reduction. Left: Principal Component Analysis (PCA). Center: T-distributed Stochastic Neighbor Embedding (T-SNE). Right: Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP). TOC values are mapped in point sizes, while different colors are assigned to each section. High-TOC samples are consistently grouped in a region of space in all dimension reduction schemas. T-SNE provides the best results in visualizing both inter- and intra-section variability.

Feature ranking and selection

To reduce dimensionality before training the model, the most significant features were selected using ANOVA (Fig. 5a). Illite-Smectite abundance (xrd__arc_is) emerged as the most significant feature, followed by geological features: sedimentary facies (geo__num_facies), bioturbation (geo__bioturb), and environment (geo__amb). The Silhouette coefficient remained consistently high but dropped sharply after the fifth feature (Chlorite; xrd__arc_cl), so only the top four features were retained. Other ranking methods, though not shown, also identified the first three features as highly significant.

The spatial distribution of samples using all features versus the ANOVA-Silhouette selected features was compared in T-SNE reduced-dimensional space (Fig. 5b). In the feature selection scenario (right), high TOC samples clustered prominently in the upper left corner.

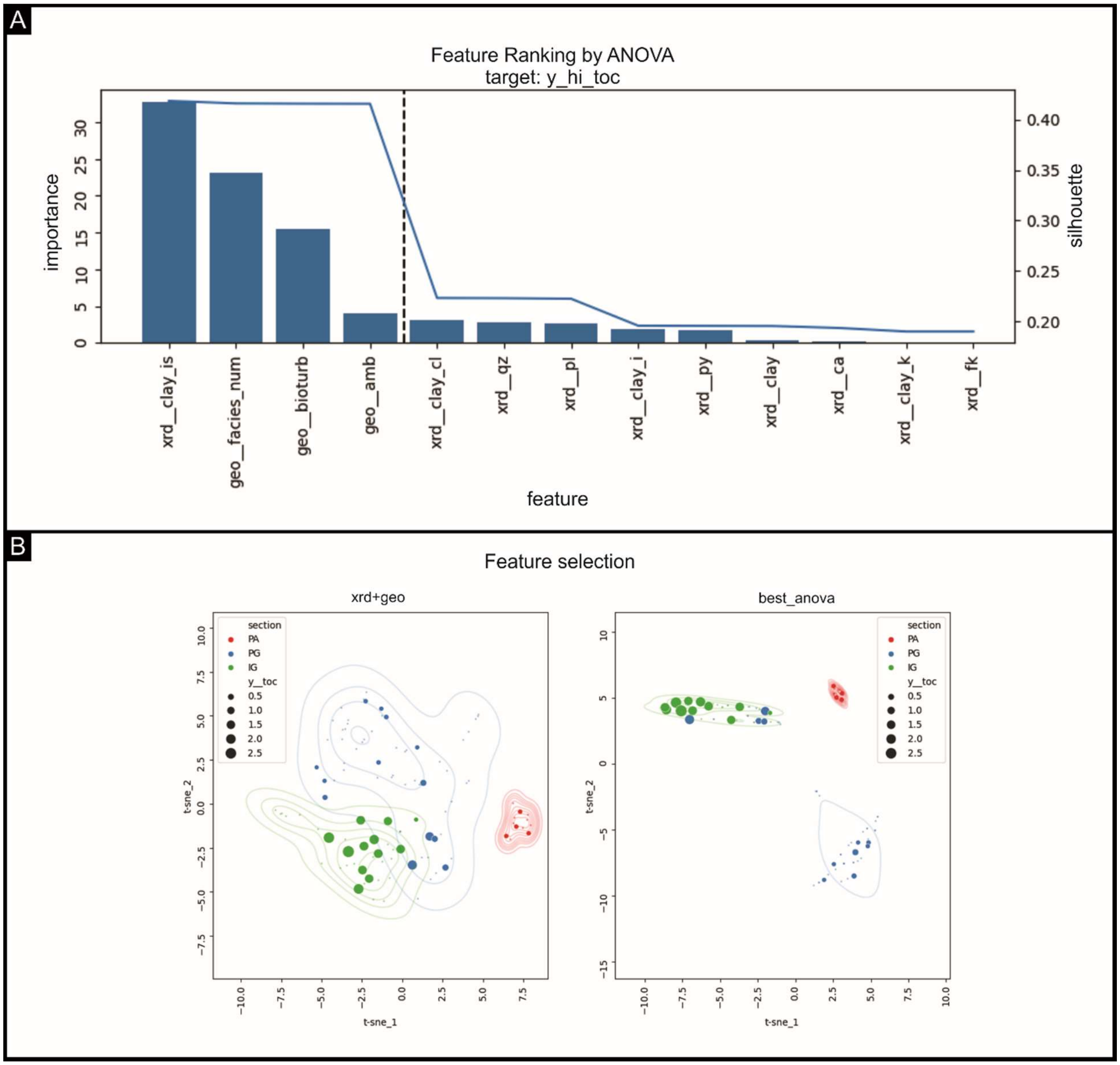

Figure 5. a) Feature importance (ANOVA F-score) and resulting Silhouette score after additive feature selection. The Silhouette score shows an abrupt decline after adding the fifth feature, indicating a potentially poor intra vs inter-cluster definition. **b)** Visualization of feature selection results in T-SNE reduced dimension. Left: Dimension reduction using all features (xrd+geo) as reference. Right: Dimension reduction of four most ANOVA F-Score relevant features (best_anova). Higher TOC values are considerably compacted in feature space, while being visibly isolated from samples with lower TOC content.

Modeling

Figures 6a and 6b show resulting mean accuracies of all trained models over train and test samples. Models trained on the full feature space (xrd+geo) show a greater tendency to overfit which is manifested as lower accuracies and a significant gap between train and test performance. While all models reported lower mean accuracies when trained in the full feature space, SVC appeared as the most robust model when working with higher dimensionality. When trained with a reduced feature set, the three proposed models showed improved performance. Logistic regression and SVC had a consistent test-train accuracy over 0.96, whereas KNN improved only slightly below this value and exhibited some overfitting behavior.

DISCUSSION

The analysis of the distribution of TOC content in shale targets is a crucial objective needed for the exploration and development of unconventional shale reservoirs. While detailed studies on the outcrops of the main source rock in southern Patagonia have been published in the past decade (*e.g.*, Richiano *et al.*, 2019 and references therein), this study represents the first application of mathematical modeling to test the potential of machine learning as a predictive method for TOC. Given the unbalanced dataset, only outcrop data and XRD were used, while geochemical rock studies were excluded except for TOC content.

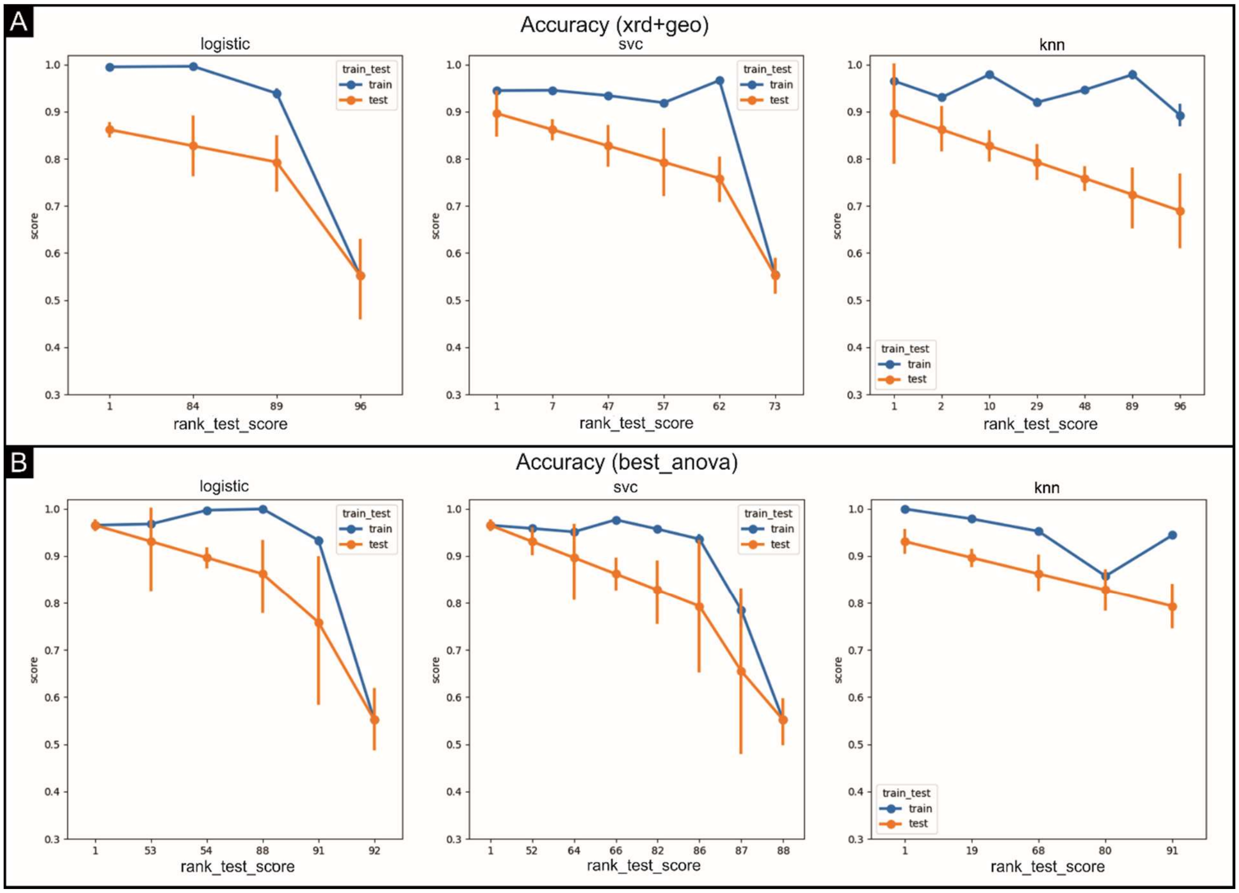

Figure 6. Train/test accuracy plots ranked by best average test accuracy. Left: Logistic Regression. Center: Support Vector Classifier (SVC). Right: K-Nearest Neighbors. **a)** Trained with all features (xrd+geo): All models exhibit moderate overfitting, as indicated by high training accuracy and low test accuracy. Logistic Regression shows the worst performance. **b)** Trained with ANOVA selected features (best_anova): There is a significant improvement in the Logistic and Support Vector Classifier with consistent train/test accuracy values over 0.96, while KNN shows a slight improvement. Error bars represent 95% Confidence interval.

During EDA all explored dimension reduction techniques proved to be useful to visualize distribution of labeled and unlabeled samples in feature space (Fig. 4). This is extremely useful to design a cost-effective analytical approach over remaining (TOC) unlabeled samples, to allow a uniform TOC sampling across feature space. Despite having only a limited number of labeled samples (approximately 30% of the dataset), spatial clustering of high TOC values was evident across all reduced-dimensionality scenarios.

A major difference was observed in the spatial discrimination of different sections. PCA shows a superposition of all three sections, particularly in the PG and IG sections. T-SNE clearly distinguished the PA section as a completely isolated cluster, while the PG and IG sections show greater contrast, with minor overlap in high TOC values within PG. UMAP provides the best sectional contrast: IG is isolated in the upper-left corner, while PG and IG occupy a broader area in the lower-left, with minimal overlap between their coverages. However, UMAP fails to display intra-section variability, especially after applying feature selection. Overall, T-SNE provides the best results in visualizing both inter- and intra-section variability, making it the preferred dimensionality reduction technique for visualizing feature selection and modeling results. This result is highly promising for successful modeling, as it demonstrates the existence of a nonlinear mapping in which the sections exhibit non-overlapping coverage.

Feature ranking by ANOVA F-Score was key in successfully identifying features that are strongly related to high TOC content. Silhouette index allowed to monitor the impact of additive feature selection on the spatial isolation of the high TOC samples helping to develop visual criteria to set the number of selected features (Fig. 5a). Although not included in this work, Random Forest and LinearSVC were also evaluated for feature importance, showing strong agreement on the top three features. However, their Silhouette performance was inferior. Figure 5b shows how reduced dimensionality from 13 to 4 did not mitigate the discrimination of the High TOC cluster and drastically improved its density.

With only 29 samples available for TOC analysis compared to 13 features, modeling and cross-validation are highly challenging. Dimensionality reduction through feature selection is essential for building a robust, generalizable model. Leave-One-Out cross-validation was used to estimate generalization accuracy but required extensive training iterations. Bayesian Search Optimization replaced exhaustive Grid Search, reducing training time.

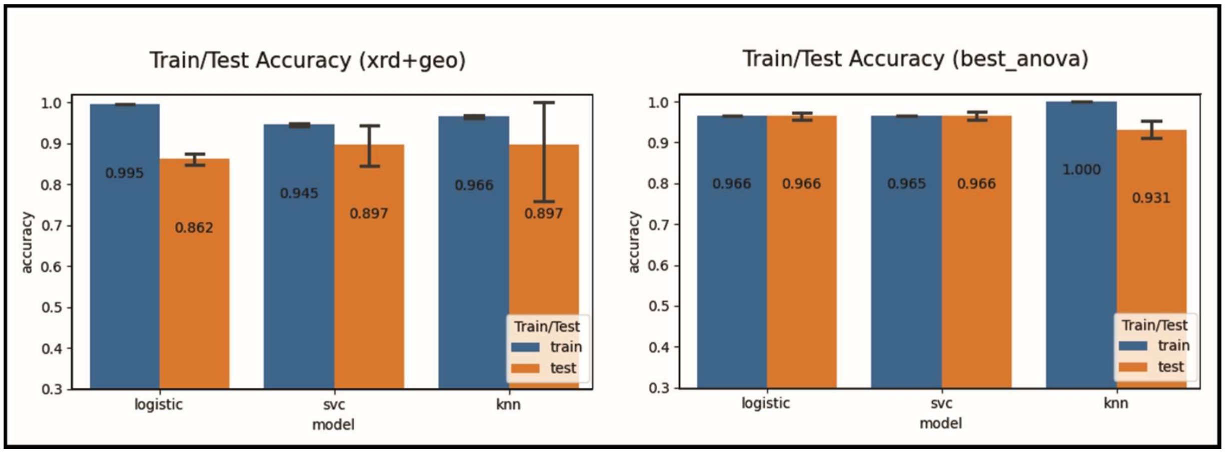

Figure 7. Best Train and Test accuracies. Left: models trained with full dataset (xrd+geo). Right: ANOVA selected features (best_anova). Error bars represent 95% Confidence intervals. Modeling performance in the test set is improved in the feature selection scenario.

Accuracy curves (Fig. 6a) for the full feature set (xrd+geo) reveal moderate overfitting, with high training accuracy and test accuracies below 0.9, indicating poor generalization, particularly for Logistic Regression. In the reduced dimension space (Fig. 6B), both Logistic Regression and Support Vector Classifier achieved consistent train/test accuracies above 0.96, significantly outperforming the full feature set. K-Nearest Neighbors remained overfitted due to the limited sample size. Figure 7 highlights substantial improvements in train/test accuracies across all models in the reduced dimension space.

The predictions run in the complete dataset (labeled and unlabeled; Fig. 8) show that all models have almost identical results, despite having completely different optimization objectives and internal structure. We restate here that these results were obtained by retraining the model on the entire labeled dataset. The green-highlighted area represents the region of interest with a high probability of predicting samples with High TOC values. This outcome supports the feasibility of modeling, as suggested during the EDA phase. The predicted values (low or high TOC) and the probability of the prediction using the SVC model on the best_anova dataset are shown in Figure 9.

Geological variables measured in the field significantly influence modeling and prediction (Fig. 4), particularly for the Río Mayer Formation. Among the XRD results, only the contribution of interstratified Illite-Smectite (IS) is notable. Interestingly, clay mineralogy of the unit is dominated by Illite (I) or Chlorite (Cl), both with IS as companion (Richiano *et al.*, 2015). Clearly, the combination of the IS content with one or more of the geological variables makes the difference for prediction of TOC content in this case study.

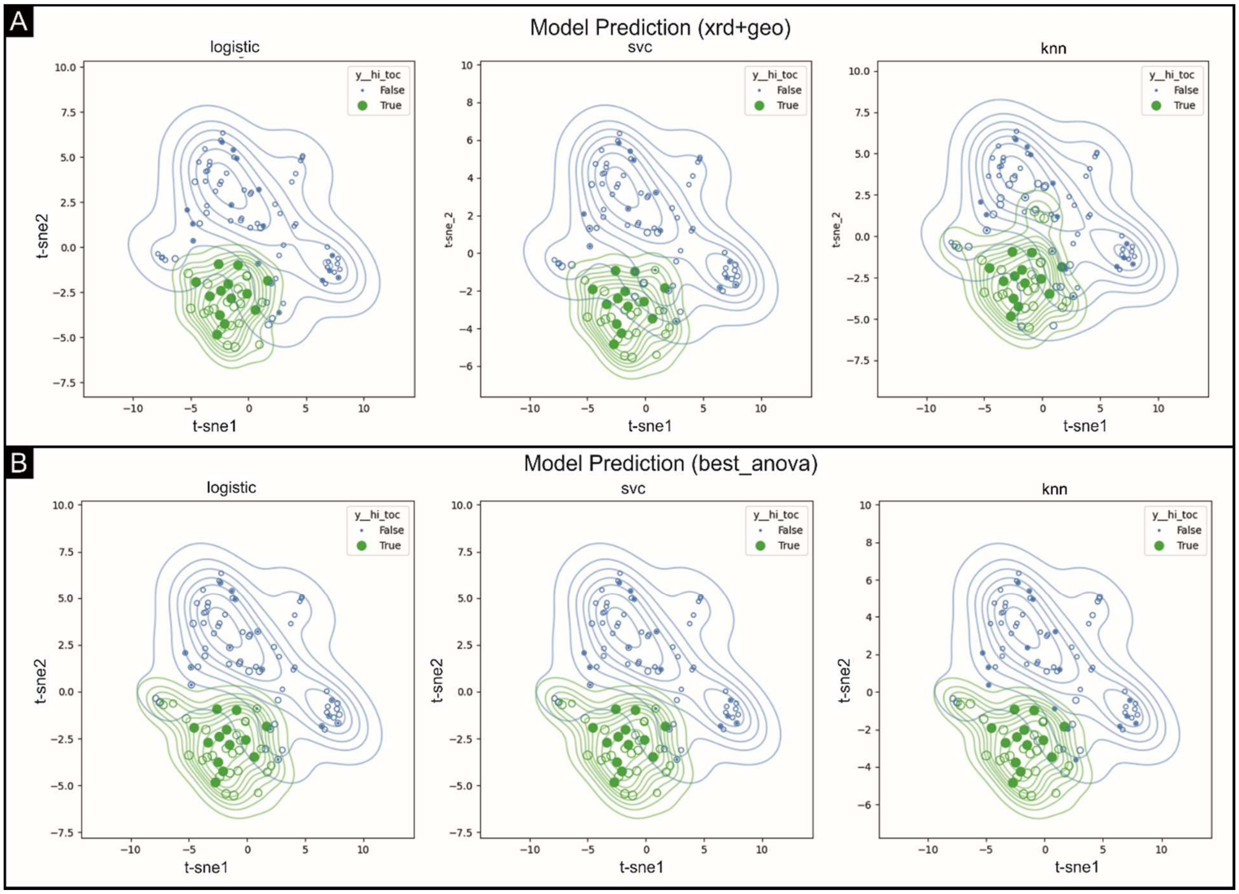

Figure 8. Predictions over T-SNE reduced dimension feature space for models trained with the full labeled dataset. Full circles represent labeled samples, while empty circles are predictions, with circle size representing the predicted probability. Green color represents a highly probable High-TOC value, while Blue circles have a low probability of high TOC. Left: Logistic Regression, Center: Support Vector Classifier, Right: K-Nearest Neighbors. **a)** Training with all features (xrd+geo). Predictions show slight variability between Logistic Regression and Support Vector Classifier models, while K-Nearest Neighbors provides a wider coverage. **b)** Training with reduced feature space (best_anova). Prediction is consistent across all models.

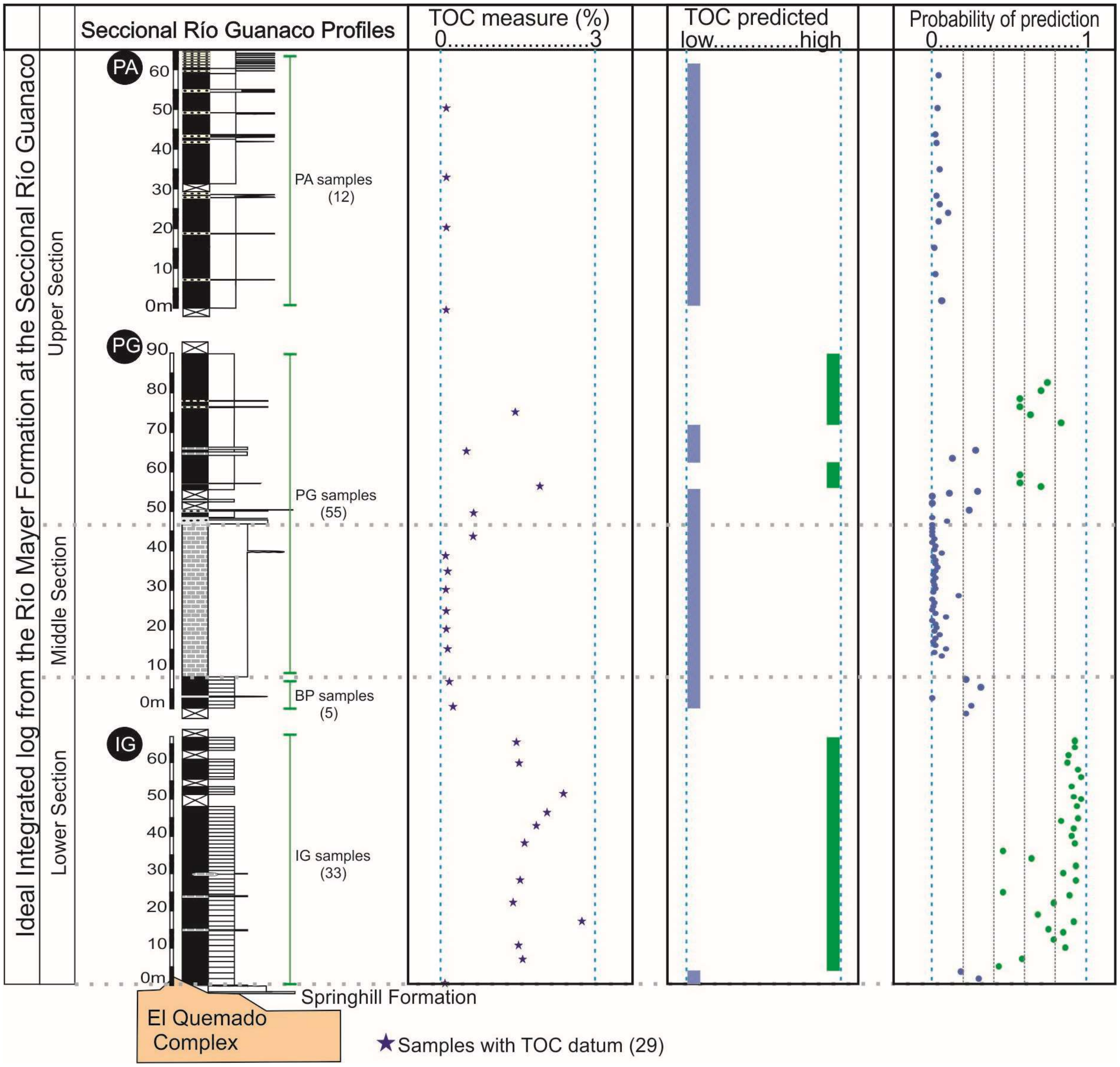

Figure 9. Visualization in the sedimentological profiles the TOC values measured in the Río Mayer Formation (n=29), the prediction of low vs high TOC and the probability of the prediction using the SVC model with the best_anova dataset (105 samples). The blue dots are considered by the model as low_TOC samples, while the green dots are interpreted as high_TOC.

Figures 8 and 9 clearly highlight the potential sweet spot within the Río Mayer Formation for unconventional targets. However, future modeling should incorporate additional factors, such as fracturing properties and areal distribution (*e.g.*, Niu *et al.*, 2022). A more advanced machine learning workflow, known as ALICE, was developed by

Chevron for unconventional plays (Prochnow *et al*., 2022). Tognolli *et al.* (2024) highlighted KNN as an effective method for classification and prediction, particularly for facies associations, but emphasized the need for additional studies to address algorithm limitations. For outcrop modeling of the Río Mayer Formation, expanding the TOC analysis database and incorporating more sedimentary sections will be crucial for enabling regional interpretations in future work.

CONCLUSIONS

The initial mathematical characterization of the main Lower Cretaceous back shale outcrops in southern Patagonia highlights several key findings. At the Exploratory Data Analysis (EDA) the dimensional reduction techniques revealed continuous regions in feature space with similar TOC values, showing the potential of the dataset to be modeled. The PCA, T-SNE, and UMAP methods progressively improved section differentiation, with UMAP delivering the best results, indicating distinct feature space fingerprints for different sections. T-SNE excelled in visualizing both inter- and intra-section variability. The ANOVA F-Score effectively ranked features associated with high TOC content, while the silhouette index identified the optimal number of features. For the Dimensionality Reduction the previous feature selection was critical for building robust, generalizable models, as evidenced by LOO cross-validation results. Logistic Regression was the most sensitive to high dimensionality. Finally, the Model Predictions across the dataset were consistent among models. While K-Nearest Neighbors struggled with generalization in reduced dimension space, its predictions aligned with other models, suggesting potential improvement with more samples. SVC emerged as the most robust method.

Acknowledgements. SR would like to acknowledge all the friends that helped to obtain data and discuss different topics about the Austral Basin in the last 18 years (Augusto Varela; Daniel Poiré; Damian Moyano-Paz; Leandro D'Elia; Andres Bilmes; Lucía Gómez-Peral; Abril Cereceda; Alejandro Gómez-Dacal; Ari Iglesias; Andrea Odino). To Prof. Murray Gingras for the final edits and comments of the manuscript. The manuscript was benefited by comments of Damián Jait, Ricardo Fabián Dominguez and an anonymous reviewer. This research was financially supported by grants by the CONICET [PIP 1016/10, PIP 0866, PUE-IPGP], the Universidad Nacional de La Plata [Project N/511] and Agencia Nacional de Promoción Científica y Tecnológica [PICT 2013-1298 and PICT 2022-0018]. The authors did not use any generative AI or AI-assisted technologies in the preparation of this work. All content was created, reviewed, and edited solely by the authors, who take full responsibility for the content of the published article. The code use for modeling is open-access at https://github.com/fares-ipgp/toc-paper-2024

REFERENCES

Aguirre Urreta, M.B. (2002). Invertebrados del cretácico inferior. In: Haller, M.J. (Ed.), *Geología y Recursos Naturales de Santa Cruz. Relatorio del XV Congreso Geológico Argentino*, 925 pp.

Arbe, H.A. (2002). Análisis estratigráfico del Cretácico de la Cuenca Austral. In: Haller, M.J. (Ed.), *Geología y Recursos Naturales de Santa Cruz. Relatorio del XV Congreso Geológico Argentino*, pp. 103–128.

Arbe, H.A. and Hechem, J. (1984). *Estratigrafía y facies de depósitos continentales, litorales y marinos del Cretácico superior, lago Argentino*. IX Congreso Geológico Argentino Actas 7:124-158.

Belotti, H., Pagan, F., Perez Mazas, A., Agüera, M., Rodriguez, J., Porras, J., Köhler, G., Weiner, G., Conforto, G., and Cagnolatti, M. (2013). *Geologic Interpretation and Assessment of Early Cretaceous Shale Oil and Gas Potential in Austral Basin, Santa Cruz, Argentina*, In: Unconventional Resources Technology Conference, Denver, Colorado, USA, August 2013.

Belotti, H., Rodriguez, J., Conforto, G., Pagan, F., Pérez Mazas, A., Agüera, M., Porras, J., Köhler, G., Cagnolatti, M., Weiner, G., Nigro, E., and Cangini, A. (2014). *La Formación Palermo Aike como Reservorio No Convencional en la Cuenca Austral, Provincia de Santa Cruz, Argentina*. IX Congreso de Exploración y Desarrollo, 1–17.

Brown, G., and Brindley, G.W. (1980). X-ray diffraction procedures for clay mineral identification. In: Brindley, G.W., Brown, G. (Eds.), *Crystal structures of Clay Minerals and Their X-ray Identification*. Mineralogical Society, London, pp. 305–359.

Busch, B., Spitzner, A., Adelmann, D., and Hilgers, C. (2022). The significance of outcrop analog data for reservoir quality assessment: A comparative case study of Lower Triassic Buntsandstein sandstones in the Upper Rhine Graben. *Marine and Petroleum Geology* 141, 105701. https://doi.org/10.1016/j.marpetgeo.2022.105701

Cuitiño, J., Varela, A., Ghiglione M., Richiano, S., and Poire, D.G. (2019). The Austral-Magallanes Basin (Southern Patagonia): A synthesis of its stratigraphy and evolution. *Latin American Journal of Sedimentology and Basin Analysis* 26, 155–166.

Féraud, G., Alric, V., Fornari, M., Bertrand, H., and Haller, M. (1999). 40Ar/39Ar dating of the Jurassic volcanic province of Patagonia: migrating magmatism related to Gondwana break-up and subduction. *Earth and Planetary Science Letter* 172, 83–98. https://doi.org/10.1016/S0012-821X(99)00190-9

Guo, J., Wang, S., Chang, L., Li, Y., Lyu, Q., and Fan, J. (2022). A Novel Method for Conventional Logging Prediction of Brittleness Index: A Case of Tight Sandstone in Western Ordos Basin. *Frontiers in Energy Research* 10.3389/fenrg.2022.819078.

Handhal, A., Al-Abadi, A., Chafeet, H., and Ismail, M. (2020). Prediction of total organic carbon at Rumaila oil field, Southern Iraq using conventional well logs and machine learning algorithms. *Marine and Petroleum Geology* 116, 104347. https://doi.org/10.1016/j.marpetgeo.2020.104347

Kraemer, P.E., and Riccardi, A.C. (1997). Estratigrafía de la región comprendida entre los lagos Argentino y Viedma (49º40' - 50º10' LS), Provincia de Santa Cruz. *Revista de la Asociación Geológica Argentina* 52, 333–360.

Melendo, F., Gargiulo, C., Garcia, N., and Jait, N. (2023). *Description and Analysis of the First Two Vertical Pilot Wells Completions to the Palermo Aike Shale: A Highly Promising Unconventional Reservoir in Austral Basin, Argentina*. Latin America Unconventional Resources Technology Conference *(LA URTeC)* https://doi.org/10.15530/urtec-2023-3969079

Milad, B., Slatt, R., and Fuge, Z. (2020). Lithology, stratigraphy, chemostratigraphy, and depositional environment of the Mississippian Sycamore rock in the SCOOP and STACK area, Oklahoma, USA: Field, lab, and machine learning studies on outcrops and subsurface wells. *Marine and Petroleum Geology* 115, 104278. https://doi.org/10.1016/j.marpetgeo.2020.104278

Moore, D.M., and Reynolds Jr., R.C. (1997). *X-ray Diffraction and the Identification and Analysis of Clay Minerals.* Oxford University Press, Oxford.

Mustafa, A., Tariq, Z., Abdulraheem, A., Mahmoud, M., Kalam, S., and Khan, R. (2022). Shale brittleness prediction using machine learning—A Middle East basin case study. *AAPG Bulletin*, 106 (11), 2275–2296. https://doi.org/10.1306/12162120181

Niu, D., Li, Y., Zhang, Y., Sun, P., Wu, H., Fu, H., and Wang, Z. (2022). Multi-scale classification and evaluation of shale reservoirs and 'sweet spot' prediction of the second and third members of the Qingshankou Formation in the Songliao Basin based on machine learning. *Journal of Petroleum Science and Engineering* 216, 110678. https://doi.org/10.1016/j.petrol.2022.110678

Ore, T. and Gao, D. (2023). Prediction of reservoir brittleness from geophysical logs using machine learning algorithms. *Computers & Geosciences* 171, 105266. https://doi.org/10.1016/j.cageo.2022.105266

Pankhurst, R.J., Riley, T.R., Fanning, C.M., and Kelley, S.P. (2000). Episodic silicic volcanism in Patagonia and Antarctic Peninsula: chronology of magmatism associated with the break-up of Gondwana. *Journal of Petrology* 41, 605–625.

Passey, Q., Creaney, S., Kulla, J., Moretti, F., and Stroud, J. (1990). A practical model for organic richness from porosity and resistivity logs. *AAPG Bulletin* 74, 1777–1794. https://doi.org/10.1306/0C9B25C9-1710-11D7-8645000102C1865D

Poiré, D., Richiano, S., Varela, A., and Sandoval, P.N. (2007). Descripción detallada de coronas, interpretación paleoambiental y logs numéricos de facies sedimentarias, Proyecto Lomitas. Informe confidencial preparado para Larriestra-Geotecnologías SA y Repsol-YPF SA. Parte I-II.

Prochnow, S., Raterman, N., Swenberg, M., Reddy, L., Smith, I., Romanyuk, M., and Fernadez, T. (2022). A subsurface machine learning approach at hydrocarbon production recovery & resource estimates for unconventional reservoir systems: Making subsurface predictions from multimensional data analysis. *Journal of Petroleum Science and Engineering* 215, 110598. https://doi.org/10.1016/j.petrol.2022.110598

Richiano, S. (2014). Lower cretaceous anoxic conditions in the Austral basin, south-western Gondwana, Patagonia Argentina. *Journal of South American Earth Sciences* 54, 37–46.

Richiano, S. (2015). Environmental factors affecting the development of the *Zoophycos* ichnofacies in the lower cretaceous Río Mayer Formation (Austral basin, Patagonia). *Palaeogeography Palaeoclimatology, Palaeoecology* https://doi.org/10.1016/j.palaeo.2015.03.029.

Richiano, S., Varela, A.N., Cereceda, A., and Poiré, D.G. (2012). Evolución paleoambiental de la Formación Río Mayer, cretácico inferior, cuenca austral, Patagonia Argentina. *Latin American Journal of Sedimentology and Basin Analysis* 19, 3–26.

Richiano, S., Poiré, D.G., and Varela, A.N. (2013). Icnología de la Formación Río Mayer, cretácico inferior, SO Gondwana, Patagonia, Argentina. *Ameghiniana* 50, 273–286.

Richiano, S., Varela, A.N., Gómez-Peral, L.E., Cereceda, A., and Poiré, D.G. (2015). Composition of the lower cretaceous black shales from the Austral basin (Patagonia, Argentina): implication for unconventional reservoirs in the southern Andes. *Marine and Petroleum Geology* 66, 764–790. http://dx.doi.org/10.1016/j.marpetgeo.2015.07.018

Richiano, S., Varela, A.N., and Poiré, D.G. (2016). Heterogeneous distribution of trace fossils across initial transgressive deposits in rift basins: an example from the Springhill Formation, Argentina. *Lethaia* 49, 524–539. https://doi.org/10.1111/let.12163

Richiano S., Gómez-Peral, L.E., Varela, A.N., Gómez Dacal, A., Cavarozzi, C., and Poiré, D.G. (2019). Geochemical characterization of black shales from the Río Mayer Formation (Early Cretaceous), Austral-Magallanes Basin, Argentina: Provenance response during Gondwana break-up. *Journal of South American Earth Sciences* 93, 67–83. https://doi.org/10.1016/j.jsames.2019.04.009

Rojas, C., Jait, D., Aimar, E., Rivero, M., Cevallos, M., Gargiulo, C., and Melendo, F. (2022). *Formación Palermo Aike, Caracterización y potencial productivo como reservorio no convencional, Cuenca Austral, Provincia De Santa Cruz, Argentina.* 11º Congreso de Exploración y Desarrollo de Hidrocarburos Exploración y Sistemas Petroleros, 73–96.

Rousseeuw, P.J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis”. *Computational and Applied Mathematics* 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Saporetti, C.M., Fonseca, D.L., Olivera, L.C., Pereira, E., and Goliatt, L. (2022). Hybrid machine learning models for estimating total organic carbon from mineral constituents in core samples of shale gas fields. *Marine and Petroleum Geology* 143, 105783. https://doi.org/10.1016/j.marpetgeo.2022.105783

Schultz, L.G. (1964). Quantitative interpretation of mineralogical composition from X-ray and chemical data for Pierra Shale. In: *U.S. Geological Survey Professional Paper*, vol. 391, pp. 1–31.

Tognoli, F. M. W., Spaniol, A., de Mello, M., and de Souza, L. (2024). A machine-learning based approach to predict facies associations and improve local and regional stratigraphic correlations. *Marine and Petroleum Geology* 160, 106636. https://doi.org/10.1016/j.marpetgeo.2023.106636