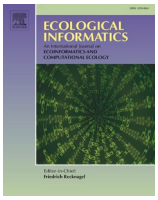
Ecological Informatics 94 (2026) 103669

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



AI-assisted multi-target classification for research-policy alignment in conservation science

Chris McCarthy [a,*], Cassandra Brooks [b,c], Troy Sternberg [d,e], Kyle Shaney [a,f], Buho Hoshino [g]

[a] American Association for the Advancement of Science (AAAS) Science & Technology Policy Fellow (STPF), Washington, DC 20005, USA
[b] Department of Environmental Studies, University of Colorado Boulder, Boulder, CO 80303, USA
[c] Institute of Arctic and Alpine Research, University of Colorado Boulder, Boulder, CO 80303, USA
[d] School of Geography, University of Oxford, Oxford OX2 6HY, UK
[e] CEI Centre for International Studies ISCTE – University Institute Lisbon, Avenida das Forças Armadas, 1649 Lisbon, Portugal
[f] Department of Biology, Health, and the Environment, The University of Texas at San Antonio, San Antonio, TX 78249, USA
[g] Lab of Environmental Remote Sensing, Department of Environmental Sciences, College of Agriculture, Food and Environment Sciences, Rakuno Gakuen University, Hokkaido 069-8501, Japan

ARTICLE INFO

Keywords:
Artificial intelligence
Automated classification
Conservation science
Evidence-based management
Multi-target learning
Natural language processing
Research coverage analysis
SciBERT

ABSTRACT

Scientific research underpins effective conservation policy, yet current approaches for assessing whether scientific outputs meaningfully support defined management objectives rely primarily on manual expert review. This limitation constrains scalability, is time intensive and introduces potential bias in identifying knowledge gaps. We present a framework combining AI-assisted multi-target classification with systematic coverage analysis for automated evaluation of research alignment with conservation objectives. We compare traditional machine learning (TF-IDF + logistic regression), a generic BERT baseline, and an enhanced SciBERT approach incorporating domain-specific adaptations including multi-target architecture, balanced loss functions, and target weighting optimized for conservation science. The framework classifies research topics and conservation objective alignment, two dimensions requiring comprehension of scientific content and policy implications. We demonstrate the approach using 295 expert-annotated peer-reviewed studies from the Ross Sea region Marine Protected Area in Antarctica. Our enhanced multi-target SciBERT model achieved 70.0% macro F1, outperforming TF-IDF (59.5%) and BERT (52.0%) baselines, with per-target improvements of 21% on research topics and 14.5% on conservation objectives. The framework achieved 78% agreement with expert annotations, with particularly strong performance on conservation objective alignment (87.7% F1, 94% agreement). The integrated system successfully identified and quantified descriptive patterns in research coverage across thematic and policy dimensions, enabling systematic assessment for research prioritization and automated coverage analysis. While demonstrated in the Antarctic context, the framework architecture is broadly transferable, though successful adaptation requires retraining with domain-specific expert annotations and fine-tuning to match local management frameworks.

1. Introduction

Conservation science increasingly grapples with the challenge of translating growing volumes of research into real-world outcomes. Despite widespread recognition that evidence-based decision-making is critical to addressing biodiversity loss, climate adaptation, and sustainability goals (Lemos et al., 2018; Sutherland et al., 2004), the link between scientific knowledge and conservation policy remains weak (Sabo et al., 2024). This disconnect, often termed the "knowledge–action" or "science–policy" gap, limits the impact of conservation investments, hampers adaptive management, and contributes to inefficiencies in research funding and planning (Cook et al., 2013; Cvitanovic et al., 2016; Toomey et al., 2017). While open-access publishing and data-sharing initiatives have expanded the availability of scientific knowledge (Piwowar et al., 2018), accessibility alone does not ensure that research outputs meaningfully support defined management

* Corresponding author.
E-mail addresses: cmccar27@jh.edu (C. McCarthy), cassandra.brooks@colorado.edu (C. Brooks), troy.sternberg@geog.ox.ac.uk (T. Sternberg), kjshaney@gmail.com (K. Shaney), aosier@rakuno.ac.jp (B. Hoshino).

https://doi.org/10.1016/j.ecoinf.2026.103669
Received 6 August 2025; Received in revised form 17 February 2026; Accepted 18 February 2026
Available online 19 February 2026

objectives. What is needed are scalable systems that can systematically evaluate whether scientific outputs align with conservation goals and provide automated tools for systematic research coverage assessment.

Marine Protected Areas (MPAs) offer an instructive lens for addressing this challenge. Although global MPA coverage has grown rapidly (UNEP-WCMC, IUCN, 2021), many remain "paper parks" lacking the scientific insight required to monitor performance or guide adaptive management (Gill et al., 2017; Jones et al., 2018; Pike et al., 2024). Ensuring that research activities align with conservation objectives and monitoring priorities is critical for MPA effectiveness, yet synthesizing whether existing studies support these priorities remains time-consuming, inconsistent, and largely manual. Traditional bibliometric approaches focus primarily on citation patterns and publication metrics rather than content alignment with specific conservation goals (Aria and Cuccurullo, 2017). Manual expert review, while thorough, cannot scale to assess the rapidly growing volume of scientific literature or provide consistent, unbiased evaluation across research domains and management contexts.

Recent advances in artificial intelligence, particularly natural language processing (NLP), offer promising solutions for automated scientific literature analysis. Transformer-based language models, such as BERT and its domain-specific variants, have demonstrated strong capabilities in understanding scientific text and performing complex classification tasks (Rogers et al., 2020). SciBERT, specifically pretrained on scientific literature, has shown superior performance in biomedical and scientific text classification compared to general-purpose models (Beltagy et al., 2019). However, most applications focus on single-target classification tasks, while comprehensive research assessment requires simultaneous evaluation across multiple dimensions including research topics, policy alignment, geographic scope, and methodological approaches (Chalkidis et al., 2020). Furthermore, robust, scalable automated approaches to identify research coverage gaps and inform strategic planning are still lacking (Westgate et al., 2015).

Multi-target learning approaches, where models simultaneously predict multiple related outcomes, have shown promise in various domains by leveraging shared representations and improving classification performance (Ruder, 2017; Zhang and Yang, 2017). In the context of scientific literature analysis, such approaches could enable comprehensive assessment of research alignment across multiple conservation dimensions simultaneously while reducing computational requirements compared to separate single-target models. The challenge extends beyond classification to systematic assessment of research coverage patterns and evaluation of how research portfolios align with existing conservation priorities. This requires frameworks that can systematically analyze research distribution across comprehensive literature corpora to assess coverage patterns across defined conservation priorities, ultimately enabling evidence-based research prioritization and resource allocation.

This study addresses these challenges by introducing an AI-assisted framework that combines enhanced multi-target SciBERT classification with research coverage assessment for evaluating research alignment with conservation objectives. The enhanced SciBERT architecture incorporates domain-specific adaptations including multi-target architecture, balanced loss functions, and target weighting optimized for conservation science. Through a unified model that leverages shared representations, the system classifies research topics and conservation objective alignment, two dimensions that require comprehension of scientific content and policy implications. While geographic dimensions (management zones and monitoring areas) were initially evaluated, these proved to be information extraction tasks rather than classification challenges requiring content understanding. The framework enables comprehensive evaluation across thematic and policy dimensions while revealing research coverage patterns and imbalances. The integrated framework enables systematic research assessment at scale, addressing the limitations of traditional manual review approaches.

The Ross Sea region Marine Protected Area (RSRMPA) provides an ideal demonstration case due to its well-defined conservation structure consisting of clearly articulated conservation objectives, structured management zones, and substantial research corpus that enables comprehensive validation of automated classification approaches. As the world's largest MPA, covering approximately 2 million km^2 (Marine Conservation Institute, 2024) and established in 2016 under the Commission for the Conservation of Antarctic Marine Living Resources (CCAMLR), the RSRMPA incorporates a zone-based management framework with three distinct management zones (General Protection Zone, Special Research Zone, and Krill Research Zone) and eleven specific conservation objectives outlined in CCAMLR Conservation Measure 91–05 (CCAMLR, 2016). This well-defined conservation structure, with explicit policy frameworks, geographic boundaries, and conservation targets, combined with the availability of expert-curated research datasets, provides the classification targets necessary to validate automated assessment approaches. These objectives include protecting ecological structure and function, maintaining reference areas for monitoring natural variability, and preserving important habitats for key species. We trained and validated our approach using a dataset of 295 expertly annotated studies compiled by Brooks & Ainley (Brooks and Ainley, 2022), achieved robust multi-target classification performance, and applied the framework to demonstrate systematic coverage assessment capabilities.

Our approach directly responds to the unique challenges that conservation domains present for automated text analysis. While the volume of relevant literature varies across conservation contexts, the need for expert annotation across multiple dimensions, including thematic content and policy alignment, makes comprehensive manual classification resource-intensive at any scale. Expert reviewers must possess deep domain knowledge to accurately assess research alignment with specific conservation objectives and thematic priorities. This assessment process typically requires considerable time and effort per paper when evaluating multiple classification dimensions. These demands create significant bottlenecks for grant program managers evaluating research portfolios, funding agencies assessing proposal alignment with priorities, conservation organizations conducting literature reviews, and policy makers requiring evidence synthesis for decision-making, a challenge increasingly recognized in recent literature (Gil-Clavel and Filatova, 2023; Kopperud et al., 2022). The resource intensity limits the scope and frequency of comprehensive research assessments, often resulting in ad hoc rather than structured evaluation of research coverage and gaps.

By combining advanced NLP techniques with research coverage assessment within a scalable AI framework, this study provides a transferable solution for evidence-based research assessment. The framework's ability to assess thematic content and policy alignment makes it particularly valuable for protected area management, where research must support diverse conservation objectives across defined spatial boundaries. While demonstrated using CCAMLR's conservation framework, the multi-target architecture applies broadly to diverse conservation contexts requiring simultaneous assessment across multiple dimensions, including terrestrial protected areas, marine conservation networks, and ecosystem-based management initiatives globally.

2. Methodology

2.1. Study area and dataset

This study focuses on the Ross Sea region Marine Protected Area (RSRMPA) in Antarctica, located between approximately 67°S to 78°S and 160°E to 160°W (spanning the International Date Line) (Fig. 1). Established in 2016 under CCAMLR Conservation Measure 91–05, the RSRMPA covers approximately 2 million km^2 of the Southern Ocean when including areas beneath the Ross Ice Shelf, making it the world's largest MPA (Marine Conservation Institute, 2024). The area is

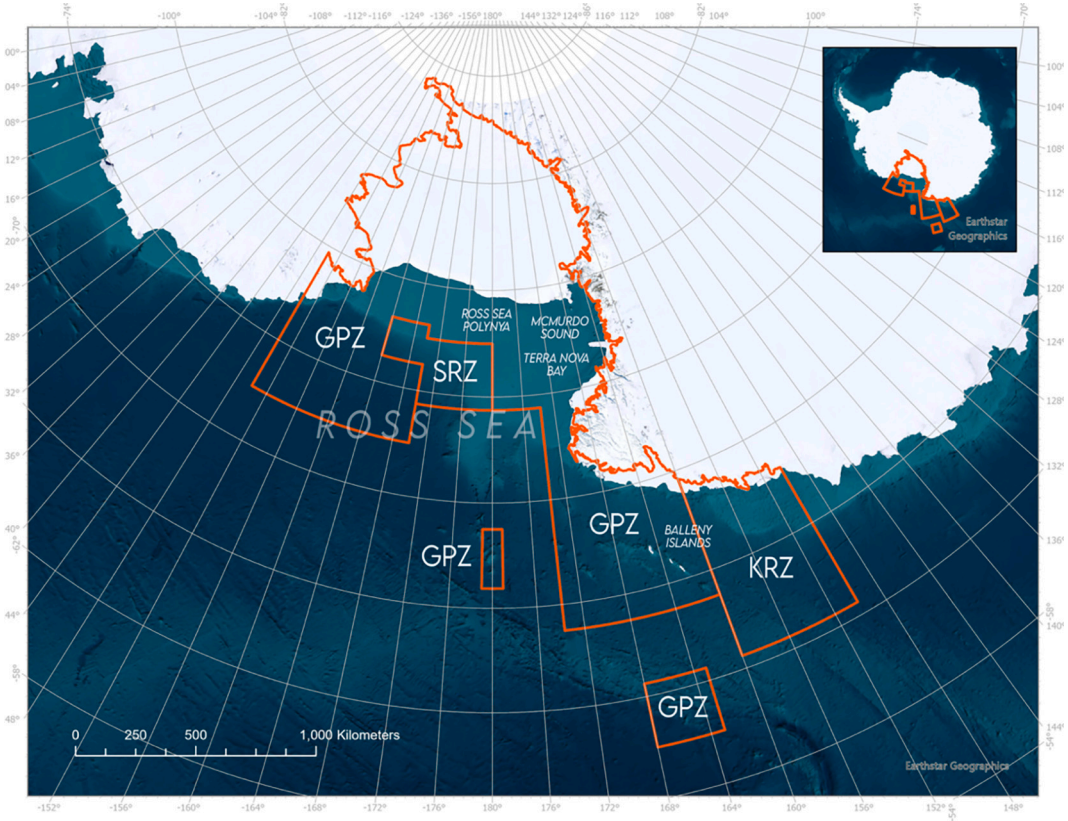


Fig. 1. Ross sea region marine protected area study region. Map showing the RSRMPA boundaries and management zones that comprise the conservation framework from which our research dataset was derived.

delineated into three primary management zones: the General Protection Zone (GPZ), Special Research Zone (SRZ), and Krill Research Zone (KRZ), each with specific regulations governing activities such as fishing and scientific research.

To ensure comprehensive coverage and policy relevance, we utilized the expert-curated dataset compiled by Brooks and Ainley (2022), consisting of 295 peer-reviewed articles published between 2010 and 2021. While 295 papers may seem modest by general ML standards, this represents a substantial portion of Ross Sea research literature and reflects the reality of specialized conservation domains where expert annotation is resource-intensive. This dataset reflects a systematic collection of research relevant to the RSRMPA and its conservation and monitoring priorities. Each paper was manually annotated by domain experts with comprehensive labels including research and monitoring topics (1–38 based on the RSRMPA Research and Monitoring Plan; (Dunn et al., 2017)), CCAMLR conservation objectives (I-XI; (CCAMLR, 2016)), management zones (GPZ, SRZ, KRZ), and monitoring areas. All included studies fall within CCAMLR Statistical Subareas 88.1 and 88.2 (the Ross Sea management unit). We initially evaluated all four annotation types but found management zones (94% single-class dominance) and monitoring areas (geographic entity recognition) to be information extraction rather than classification tasks. Therefore, our model addresses the two targets requiring content understanding, detailed in Table 1, with the complete research topics from the RSRMPA Research and Monitoring Plan detailed in Supplementary Table S1 and the full list of CCAMLR conservation objectives provided in Supplementary Table S2.

2.2. Multi-target SciBERT architecture

2.2.1. Foundation model

We compare two approaches to demonstrate the combined value of domain-specific pretraining and architectural innovations. The baseline model consists of standard BERT-base with simple multi-label classification heads using binary cross-entropy loss. The enhanced model employs SciBERT (Beltagy et al., 2019) with architectural adaptations including: (1) multi-target architecture with shared projection layer, (2) focal loss with class-specific weights, (3) label smoothing regularization, and (4) optimized target weighting. This comparison isolates the combined contribution of domain-specific pretraining (SciBERT's scientific

Table 1
Multi-target classification schema. Details of the two classification targets addressed by the enhanced multi-target SciBERT framework, both requiring content understanding. Class distribution metrics reveal extreme imbalance ratios (comparing most frequent to least frequent class within each target).

Target	Classes Total	Classes Present	Class Imbalance Pattern	Description	Examples
Research Topics	38	27	45:1 imbalance ratio, 59% of classes in <15 papers	Semantic classification of research focus areas from RSRMPA Research and Monitoring Plan	Bioregionalization and biodiversity mapping, Physical and biological habitat changes, Functional ecology processes, Krill population dynamics
CCAMLR Objectives	11	9	30:1 imbalance ratio, Dominant: "promote research"	Policy alignment classification based on conservation objectives in CCAMLR Conservation Measure 91–05	Conserve natural ecological structure, Promote research, Large scale ecosystem processes, Key top predator foraging distributions

corpus) and architectural enhancements designed for imbalanced multi-target classification, representing our complete methodological contribution rather than isolating individual components.

2.2.2. Multi-target learning framework

The Multi-Target SciBERT architecture consists of three main components: (1) the pre-trained SciBERT base model for feature extraction, (2) a shared projection layer (768 → 256 dimensions) for learning common representations across targets, and (3) target-specific classification heads for each prediction task. The shared projection layer enables the model to learn representations that benefit all classification targets while reducing computational complexity compared to separate single-target models. Target-specific classification heads consist of two-layer feedforward networks with ReLU (Rectified Linear Unit) activation and dropout regularization, allowing simultaneous prediction across both tasks while leveraging shared semantic understanding of research content. Fig. 2 illustrates the complete enhanced multi-target SciBERT architecture.

2.2.3. Classification targets and schema

Our framework addresses two classification targets that require understanding of scientific content and policy implications (Table 1). Research topics (38 total, 27 present in dataset) provide classification of research focus areas, including bioregionalization and biodiversity mapping, physical and biological habitat changes, functional ecology processes, and krill population dynamics. CCAMLR objectives (11 total, 9 present in dataset) provide policy alignment classification based on the conservation objectives outlined in CCAMLR Conservation Measure 91–05 and require inference of how research contributes to specific conservation goals.

2.2.4. Text processing

Input texts for each paper consisted of concatenated title, abstract, and keywords from the Brooks and Ainley (2022) dataset. While this constraint limits access to methodological details that may appear only in full text, this approach was necessitated by the available data, as the expert-curated dataset contained only these metadata fields rather than full text. Despite this constraint, these fields provide high information density for classification purposes: titles capture the primary research focus, abstracts summarize methods, findings, and study locations, while keywords explicitly identify key species, geographic areas, and research themes. This metadata-based approach also offers practical advantages for real-world deployment, as title, abstract, and keywords are readily available from research databases, grant proposals, and systematic review workflows, making the model immediately applicable for research assessment tasks without requiring access to full manuscripts.

The concatenated text was tokenized using SciBERT's tokenizer with a maximum sequence length of 512 tokens. We extended the tokenizer vocabulary with a small set of Antarctic-specific terms (e.g., "dissostichus-mawsoni", "euphausia-superba", "ross-sea-polynya") to reduce tokenization of domain-specific compound terms. For new paper classification, the same input format should ideally be maintained, which our implementation supports through automated extraction of these fields from PDF documents or structured metadata. This standard approach for document classification leverages SciBERT's pre-training on scientific literature while working within real-world data constraints.

2.3. Training strategy and optimization

2.3.1. Data splitting and cross-validation

The dataset was partitioned using iterative stratification to maintain label distribution across all targets, with 70% for training, 15% for validation, and 15% for testing. The final data splits consisted of 204 papers for training, 43 papers for validation, and 48 papers for testing.

Based on preliminary experiments with multiple configurations, we selected a configuration optimized for the two content understanding targets. This configuration uses differentiated target weights (topics: 3.0, objectives: 2.5) to balance the complexity difference between research topic classification and policy alignment, combined with gradient accumulation for effective batch size of 30, and label smoothing (0.08) for regularization.

2.3.2. Final model training

For the final model reported in this paper, we employed an enhanced training strategy to maximize data utilization. Following cross-validation for configuration selection, we combined the training and validation sets (247 papers total) and trained for 34 epochs using the

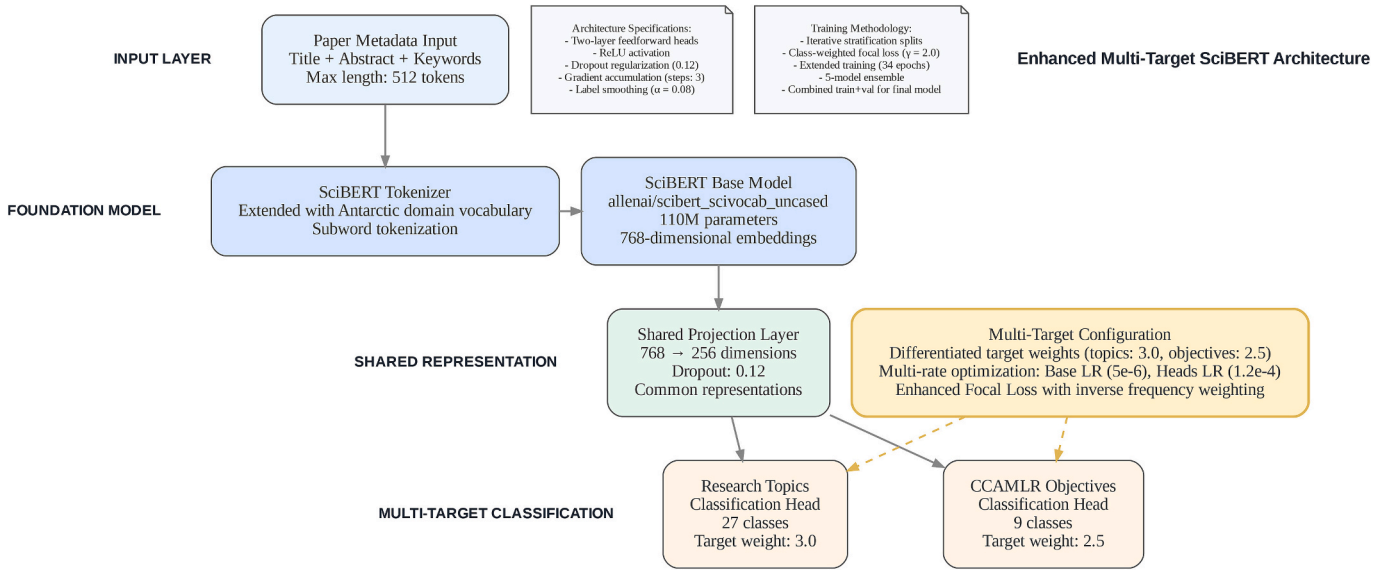


Fig. 2. Enhanced multi-target SciBERT architecture. Schematic diagram showing the workflow from paper metadata (title, abstract, keywords) through the SciBERT base model (110 M parameters) to multi-target predictions. The architecture employs a shared projection layer (768 → 256 dimensions) that learns common representations across both targets, followed by target-specific classification heads for research topics (27 classes) and CCAMLR objectives (9 classes). The configuration uses differentiated target weights (topics: 3.0, objectives: 2.5) with multi-rate optimization using differential learning rates for the base model (5e-6) and classification heads (1.2e-4).

optimized configuration. This approach maximizes the use of available annotated data while the held-out test set (48 papers) provides unbiased performance evaluation. To ensure robust performance estimates and enable statistical analysis, we trained an ensemble of five models with different random seeds (42, 123, 456, 789, 321). We report the performance of the ensemble using majority voting (3/5 agreement required for positive classification).

2.3.3. Multi-target loss function and class balancing

The framework employs Focal Loss (Lin et al., 2017) enhanced with class-specific weights and label smoothing to address severe label imbalance across targets. Class weights were calculated as the inverse frequency of each class in the training set. The multi-target loss combines weighted focal losses across both targets, with target-specific weights (w_target) applied to balance task complexity, label smoothing ($\alpha = 0.08$) for regularization, and gradient accumulation for stable training. This approach effectively handles the extreme class imbalance while maintaining focus on research topic and conservation objective classification. Multi-rate optimization employed different learning rates for the pre-trained base model (5e-6) versus classification heads (1.2e-4), determined through validation set tuning following standard practices for transformer fine-tuning, to prevent catastrophic forgetting while enabling task-specific adaptation.

2.3.4. Training implementation

All model development and training was conducted using Google Colab with NVIDIA A100 GPU acceleration, implemented using PyTorch 2.0.1 and Hugging Face Transformers 4.35.0. The final model was trained with the following key hyperparameters: batch size 10 with gradient accumulation steps of 3 (effective batch size 30), dropout rate 0.12, weight decay 0.012, and focal loss gamma 2.0. Decision threshold selection is described in Section 2.4.2. The complete model architecture contains 110,254,136 parameters, dominated by the pre-trained SciBERT base (110 M parameters) with a small number of additional parameters from the projection and classification heads.

2.3.5. TF-IDF baseline implementation

To provide a non-neural baseline comparison, we implemented a TF-IDF + Logistic Regression approach using scikit-learn 1.3.0. Text preprocessing included lowercase conversion and basic tokenization using the same concatenated title-abstract-keywords input as neural models. We configured TfidfVectorizer with max_features = 5000, ngram_range = (1,2) to capture unigram and bigram features, and sublinear term frequency scaling (sublinear_tf = True) to reduce the impact of term frequency differences.

For multi-label classification across both targets (themes and objectives), we employed MultiOutputClassifier wrapping LogisticRegression with balanced class weights.

(class_weight = 'balanced') to address class imbalance, L2 regularization (C = 1.0), and lbfgs solver (max_iter = 1000). The model was trained on the same 247-paper combined training and validation set and evaluated on the identical 48-paper held-out test set used for neural baselines, ensuring fair comparison. Classification thresholds were set to 0.5 for all targets. This traditional machine learning baseline represents what conservation organizations might deploy without deep learning infrastructure or domain-specific language models.

2.4. Model evaluation and validation framework

2.4.1. Performance metrics

Model performance was evaluated using macro F1 score as the primary metric for imbalanced multi-label data, complemented by micro F1 score, weighted F1 score, precision, recall, and Hamming loss (Zhang and Zhou, 2014). Additionally, we calculated Jaccard similarity (intersection over union) between predicted and true label sets to quantify expert agreement, providing a more intuitive measure of multi-label prediction quality. Final performance is reported on the held-out test set of 48 papers.

2.4.2. Decision thresholds

Classification thresholds were tuned on the validation split (43 papers) before any final training or test-set evaluation. A single shared threshold of 0.42, applied to both research topics and conservation objectives, was selected to maximize macro F1. A shared value was adopted because both targets use the same sigmoid output range and because per-target tuning on a 43-paper validation set risks overfitting to split-specific label noise. After threshold selection, the model was retrained on the combined training + validation set (247 papers) for 34 epochs. The fixed 0.42 threshold was then held constant across all five ensemble models during evaluation on the held-out test set (48 papers), ensuring no information leakage from test data into any model selection or calibration decision. For the TF-IDF baseline, a default threshold of 0.5 was applied to all targets.

2.4.3. Agreement with expert annotations

We conducted post-hoc expert validation using our test set of 48 papers containing both expert-assigned labels (from the Brooks & Ainley dataset) and model predictions. This approach treats the original dataset annotations as single-expert ground truth, comparing model predictions against expert classifications using Jaccard similarity for multi-label tasks.

Throughout the evaluation, overall scores are reported as weighted averages across the two classification targets, assigning 60% weight to research topics and 40% to conservation objectives. This weighting reflects three considerations: (1) research topics present substantially greater classification complexity (38 classes vs. 11, with 45:1 imbalance vs. 30:1), meaning that overall metrics dominated by the easier target would overstate practical system performance; (2) the same weighting was applied during training loss computation, so evaluation-time weighting mirrors the optimization objective; and (3) from a management perspective, thematic coverage assessment across 38 research priorities represents the primary use case driving this framework's development. This weighting scheme is applied consistently to all overall metrics reported in Tables 2 and 3, as well as the overall agreement scores in Section 3.3.

2.4.4. Statistical testing

To evaluate improvements over baseline approaches, we compared the enhanced multi-target SciBERT with a standard BERT model and a TF-IDF + Logistic Regression baseline, all trained on the same dataset. Statistical significance was assessed using paired *t*-tests comparing the five ensemble models from each approach. Effect sizes were calculated using Cohen's d to quantify the magnitude of improvements. Performance differences were assessed using the ensemble predictions with majority voting (3/5 agreement required for positive classification), with improvement percentages calculated for each target. All statistical tests used $\alpha = 0.05$ with Bonferroni correction for multiple comparisons across targets.

2.4.5. Research coverage analysis

Research coverage analysis quantifies the distribution of classified papers across themes and objectives to identify descriptive patterns in research coverage. We define research gaps normatively based on CCAMLR's policy framework: all 38 research topics in the RSRMPA Research and Monitoring Plan represent priorities that warrant scientific attention for effective MPA management. A gap exists when research activity (measured by paper count) is substantially lower for policy-designated priorities compared to others, descriptively indicating potential underinvestment in areas deemed important by management frameworks. The extreme class imbalances revealed in Table 1 directly indicate such gaps, with some themes appearing in fewer than 5 papers despite their designation as management priorities. This normative

Table 2
Enhanced multi-target SciBERT test set performance (5-model ensemble).

Target	Macro F1	Micro F1	Weighted F1	Jaccard	Hamming Loss	Precision	Recall
Research Topics	0.583	0.759	0.749	0.672	0.062	0.575	0.591
CCAMLR Objectives	0.877	0.932	0.933	0.941	0.034	0.883	0.871
Overall Weighted	0.700	–	–	0.780	–	–	–

Note: Dashes (−) indicate metrics where individual aggregation is not applicable. Overall weighted scores use the 60/40 target weighting described in Section 2.4.3. Jaccard similarity (0.780) is lower than Macro F1 (Topics: 0.583, Objectives: 0.877) because Jaccard uses intersection-over-union for multi-label sets, penalizing both false positives and false negatives simultaneously, while F1 separately considers precision and recall before harmonizing them.

Table 3
Performance comparison: enhanced SciBERT vs BERT baseline.

Target	TF-IDF + LR	BERT Baseline	Enhanced SciBERT	Improvement vs TF-IDF	*p*-value	Cohen's d
Research Topics	0.482	0.401	0.583	+21.0%	0.009**	3.01
CCAMLR Objectives	0.766	0.749	0.877	+14.5%	0.007**	3.24
Overall	0.595	0.520	0.700	+17.6%	<0.001***	5.20

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.
Note: TF-IDF + LR = TF-IDF vectorization with Logistic Regression (5000 features, unigrams + bigrams, balanced class weights). Overall scores use the 60/40 target weighting described in Section 2.4.3. Improvement percentages and statistical tests compare Enhanced SciBERT against TF-IDF baseline. All models evaluated on identical 48-paper test set.

approach reflects conservation practice where policy frameworks define what should be researched; our tool reveals what actually is being researched, with the delta indicating potential gaps for management consideration.

2.5. Code and data availability

The complete implementation of the Multi-Target SciBERT framework, including model training scripts, evaluation pipelines, ensemble training methodology, and classification analysis methodology, is available under an open-source license at https://github.com/mccarthy-conservation-ai/multitarget-scibert-ross-sea. The repository includes detailed documentation, example notebooks, and configuration files to enable reproduction of all results. Complete technical specifications and hyperparameter details are provided in Supplementary Table S3. The modular architecture separates domain-specific components (conservation objectives, Antarctic text preprocessing) from core methodology, facilitating adaptation to other research domains. Trained model weights for all ensemble models and preprocessing pipelines are provided to support immediate deployment. Access to the training dataset is freely available in the GitHub repository for testing and development purposes.

3. Results

3.1. Multi-target classification performance

The enhanced multi-target SciBERT framework achieved a weighted F1 score of 0.700 ± 0.021 on the held-out test set of 48 papers, representing a 35% improvement over the BERT baseline (0.520). Performance across the five ensemble models showed high consistency with F1 scores of 0.674, 0.680, 0.720, 0.714, and 0.713, demonstrating robust model behavior despite stochastic training variation.

Performance reflected the distinct characteristics of the two classification tasks (Table 2). Research topics classification (F1 = 0.583) required understanding complex scientific themes despite severe class imbalance, with papers averaging 2.56 topics each and 59% of topic classes appearing in fewer than 15 papers. CCAMLR objectives classification (F1 = 0.877) achieved higher performance due to clearer policy definitions and more balanced class distribution.

3.2. Comparison with baseline approaches

The enhanced multi-target SciBERT demonstrated statistically significant improvements over both traditional and neural baselines, as shown in Table 3. We compare three approaches representing different methodological paradigms: (1) TF-IDF + Logistic Regression representing traditional machine learning with domain-appropriate feature engineering, (2) BERT baseline representing generic neural language models without domain adaptation, and (3) Enhanced SciBERT representing domain-adapted transformers with architectural optimizations for imbalanced multi-label classification.

The TF-IDF baseline achieved overall performance of 0.595 (macro F1), demonstrating that conservation research classification is tractable with traditional machine learning methods. Notably, TF-IDF substantially outperformed the BERT baseline (0.595 vs 0.520, +14.4% improvement), highlighting that domain-appropriate feature engineering can match or exceed generic neural architectures without domain-specific pretraining.

Enhanced SciBERT achieved 0.700 overall macro F1, representing a 17.6% improvement over TF-IDF ($p < 0.001$, Cohen's d = 5.20), a substantial gain for practical deployment. This improvement is particularly meaningful because TF-IDF itself substantially outperformed generic BERT (0.595 vs 0.520, +14.4%), establishing it as a strong baseline that captures domain-specific keywords effectively. The per-target improvements are even more substantial and consistent: research topics improved 21.0% (0.583 vs 0.482, p = 0.009, Cohen's d = 3.01) and conservation objectives improved 14.5% (0.877 vs 0.766, p = 0.007, Cohen's d = 3.24). The substantial improvements on both classification targets demonstrate the value of combining domain-specific pretraining (SciBERT's scientific corpus) with architectural enhancements designed for imbalanced multi-label classification (focal loss, target weighting, multi-rate optimization).

We do not include keyword-based or rule-based baselines, as these approaches are impractical for multi-label classification across 49 conservation-specific classes (38 topics +11 objectives). Such baselines would require: (1) extensive expert curation of keyword lists for each class, (2) disambiguation rules for overlapping terminology (e.g., “ecosystem” appears in multiple topics), and (3) multi-label decision logic for papers addressing multiple themes simultaneously (papers average 2.56 topics and 2.3 objectives). More fundamentally, many classification targets are defined by conceptual rather than lexical patterns—distinguishing between CCAMLR objectives like “conserve ecological structure” versus “promote research” requires understanding

policy intent and research contribution, not keyword matching. The TF-IDF baseline provides a more meaningful comparison as it uses the same multi-label framework and captures semantic relationships through n-gram features.

The performance ranking (Enhanced SciBERT > TF-IDF > BERT) validates our methodological approach: domain-specific pretraining provides measurable benefits over generic models, while the substantial improvement over TF-IDF (which already captures domain keywords) demonstrates the value of semantic understanding and architectural optimizations for imbalanced multi-label classification. TF-IDF's competitive performance (0.595) establishes feasibility for resource-constrained organizations, while our enhanced approach provides substantial accuracy gains (17.6% improvement, 26% error reduction) and strong agreement with expert annotations (78% overall, 94% on objectives) that justify infrastructure investment for high-stakes conservation decision-making.

3.3. Expert agreement analysis

Expert validation demonstrated 78.0% overall agreement (Jaccard similarity) between model predictions and expert annotations across both classification targets. The five ensemble models showed consistent expert agreement with individual weighted Jaccard scores of 0.745, 0.804, 0.768, 0.754, and 0.752, indicating stable performance across different random initializations. Target-specific agreement rates revealed strong alignment with human judgment: CCAMLR objectives achieved 94.1% agreement while research topics showed 67.2% agreement. The higher agreement on conservation objectives reflects their well-defined nature in the CCAMLR framework, while moderate agreement on topics reflects the inherent ambiguity in scientific thematic classification and severe class imbalance.

3.4. Class imbalance analysis

Analysis of the Ross Sea research corpus reveals substantial class imbalance that reflects actual research priorities. Research topics showed severe imbalance with class frequencies ranging from 1 to 45 papers per topic (median: 8 papers). The most common topics included "Physical & biological habitat changes" (45 papers) and "Functional ecology processes" (41 papers), while specialized topics like "Toothfish spawning migrations" and "Balleny Islands endemic benthos" appeared in only 1–2 papers. This imbalance presents significant challenges for classification, particularly for underrepresented themes critical to conservation objectives.

CCAMLR objectives showed less severe but still notable imbalance, with "promote research" appearing most frequently while specific habitat protection objectives appeared less often. This distribution reflects the dual nature of the RSRMPA as both a conservation area and a scientific reference zone.

3.5. Model robustness

To ensure reliability, we trained five models with different random seeds (42, 123, 456, 789, 321), achieving consistent performance with a coefficient of variation of only 2.7% for overall F1 score. The variation across ensemble models (F1 range: 0.674–0.720, mean ± SD: 0.700 ± 0.021) provides empirical confidence intervals through bootstrap-style resampling with different random initializations. This 95% empirical confidence interval (approximately 0.658–0.742, assuming normality) demonstrates robust performance relatively insensitive to stochastic training factors, indicating that performance gains are reliable and reproducible rather than artifacts of fortunate initialization. Statistical significance testing comparing enhanced multi-target SciBERT with BERT baseline was performed using paired *t*-tests across these five ensemble models. All improvements showed statistical significance ($p < 0.05$) and large effect sizes (Cohen's d > 2.8), providing strong evidence for the robustness of the approach. The consistency across random initializations suggests the improvements derive from architectural and training strategy enhancements rather than fortunate random initialization.

3.6. Research distribution analysis

The classification results reveal descriptive patterns in research distribution across the RSRMPA research portfolio. Among the 27 research topics present in the dataset, 16 topics (59%) appear in fewer than 15 papers, descriptively indicating substantial underrepresentation. Descriptive gaps include evolutionary biology processes (4 papers), seamount benthic communities (3 papers), and Balleny Islands endemic benthos (1 paper). For conservation objectives, while "promote research" dominates with high frequency, several specific objectives receive notably limited attention. The class imbalance data shows a 30:1 ratio between the most and least represented objectives. Particularly underrepresented are objectives related to spatial protection (objective iii - "protection of representative areas"), monitoring baseline areas (objective iv), and specific habitat protection objectives. The implications of these coverage patterns for research prioritization are discussed in Section 4.1.

3.7. Practical deployment examples

To demonstrate real-world applicability, Table 4 presents classification results for three representative papers from recent Ross Sea research, showing how the framework provides comprehensive classification for research assessment. Researchers and program managers can test their own papers using the pre-trained models and classification tool available at https://github.com/mccarthy-conservation-ai/multitarget-scibert-ross-sea, which includes detailed documentation and example notebooks for immediate deployment.

These examples illustrate the framework's ability to provide multi-label classification with confidence scores that enable assessment of prediction reliability. The consistently high confidence for conservation objectives (0.795–0.971) indicates reliable policy alignment assessment, while varying confidence levels for research topics reflect the inherent classification challenges across different themes. The framework's ability to identify papers addressing multiple research priorities simultaneously supports evidence-based research portfolio assessment and systematic coverage assessment.

Table 4
Representative multi-target classification examples.

Paper	Research Topics (confidence)	CCAMLR Objectives (confidence)
Deep-sea skate nursery habitats; Finucci et al. (2024)	Functional ecology processes (0.522); Prey availability effects on predators (0.518); Physical & biological habitat changes (0.431)	Promote research (0.948); Conserve natural ecological structure (0.795)
Mesozooplankton distribution patterns; Minutoli et al. (2024)	Functional ecology processes (0.514); Physical & biological habitat changes (0.471)	Promote research (0.971); Conserve natural ecological structure (0.850)
Predator-prey dynamics in McMurdo Sound; Ainley et al. (2024)	Functional ecology processes (0.840); Prey availability effects on predators (0.719); Toothfish & predator distributions (0.429); Dependence on coastal habitats (0.383)	Conserve natural ecological structure (0.957); Promote research (0.950); Key top predator foraging distributions (0.853); Coastal/localized areas of ecosystem importance (0.657)

4. Discussion

4.1. Implications for evidence-based conservation science

This work demonstrates that automated multi-target classification can transform how conservation organizations evaluate research alignment with policy objectives. By achieving 78% agreement with expert annotations (noting that this reflects comparison with a single annotator rather than a consensus panel), the framework addresses a critical bottleneck in evidence-based conservation planning: the resource-intensive nature of manual research evaluation across multiple dimensions.

This 78% expert agreement should be interpreted in the context of multi-label classification challenges. Unlike single-label classification where only one correct answer exists, multi-label classification requires correctly identifying all applicable labels - a paper with three themes needs all three identified for perfect agreement. Given the severe class imbalance (45:1 ratio with 59% of classes appearing in fewer than 15 papers), achieving 67.2% agreement on research topics is particularly noteworthy, though direct comparison with human inter-annotator agreement is not possible from a single-annotator dataset and remains an important direction for future validation. The higher 94.1% agreement on conservation objectives reflects their clearer definition in the CCAMLR framework, objectives like “key top predator foraging distributions” or “coastal/localized areas of particular ecosystem importance” have explicit policy language that makes classification more straightforward than the nuanced thematic categories.

The 17.6% overall improvement of Enhanced SciBERT over the TF-IDF baseline (0.700 vs 0.595, $p < 0.001$, Cohen's d = 5.20) represents a substantial gain for practical deployment. This improvement is particularly meaningful because TF-IDF itself substantially outperformed generic BERT (0.595 vs 0.520, +14.4%), establishing it as a strong baseline that captures domain-specific keywords effectively. Beating this strong baseline by 17.6% demonstrates the value of domain-specific pretraining combined with architectural optimizations for imbalanced multi-label classification. In practical terms, this 17.6% improvement translates to a 26% reduction in classification errors. For a corpus of 1000 conservation papers, SciBERT would prevent approximately 105 misclassifications compared to TF-IDF, directly improving research prioritization decisions. The per-target improvements are even more substantial: research topics improved 21.0% (0.583 vs 0.482, $p = 0.009$) and conservation objectives improved 14.5% (0.877 vs 0.766, $p = 0.007$), with both targets showing consistent gains that validate the framework's robustness across different classification challenges.

For organizations evaluating automated classification approaches, the choice depends on operational context. TF-IDF provides competitive performance (0.595) with minimal infrastructure requirements, making it suitable for pilot deployments or resource-constrained settings. However, the 17.6% accuracy gain and strong agreement with expert annotations (78% overall, 94% on objectives) achieved by Enhanced SciBERT justify the additional infrastructure investment for organizations managing large research portfolios or making high-stakes funding decisions where classification accuracy directly impacts conservation outcomes.

The framework's practical significance extends beyond efficiency gains. Conservation programs often struggle to identify research that simultaneously addresses multiple objectives within their thematic scope. Our multi-target approach reveals these high-impact research opportunities that might be overlooked in traditional single-dimension assessments. For instance, the framework can instantly identify papers that combine underrepresented research topics (like evolutionary biology) with critical conservation objectives. These are connections that manual review might miss due to time constraints or reviewer expertise limitations.

Furthermore, the consistent performance across diverse research contexts (from novel species discoveries to functional ecology studies) validates the framework's role as a decision-support tool that complements expert judgment. This performance aligns with recent research showing that SciBERT-based models benefit from domain-specific adaptations (Gupta et al., 2022; Likhareva et al., 2024), validating our approach of combining multi-target optimization with conservation-specific strategies. This capability is particularly valuable for funding agencies and policy makers who must rapidly assess large research portfolios to ensure strategic alignment with conservation goals while maintaining evaluation consistency across reviewers and time periods.

4.2. Multi-target applications and deployment

The enhanced multi-target SciBERT framework's ability to classify research papers across thematic and policy dimensions provides immediate applications for systematic research portfolio evaluation across diverse stakeholder communities. While our implementation focuses on peer-reviewed research papers, the core multi-target architecture and methodology can be adapted for different document types through appropriate retraining on domain-specific corpora. Grant proposals, technical reports, and policy documents would require customized training datasets that reflect their distinct writing styles, structural elements, and evaluation criteria.

It's important to note that our initial evaluation revealed geographic dimensions (management zones and monitoring areas) to be information extraction tasks rather than classification challenges. Management zones showed 94% single-class dominance, while monitoring areas required only geographic entity recognition. This distinction informs deployment strategies: content understanding tasks like thematic classification benefit from the full SciBERT architecture, while geographic information can be extracted through simpler pattern matching techniques, allowing organizations to optimize computational resources.

The following applications demonstrate how the fundamental multi-target approach could be deployed across different organizational contexts and conservation scales, recognizing that each implementation would need domain-appropriate training data and potentially modified classification targets.

4.2.1. CCAMLR research portfolio assessment

CCAMLR and its Scientific Committee can use the framework to assess new research papers across research topics and conservation objectives. When processing incoming research proposals or publications, managers can quickly identify thematic gaps where certain research topics remain underrepresented and ensure policy alignment with CCAMLR objectives. The multi-target architecture enables identification of papers that address multiple conservation objectives through integrated thematic analysis. For example, a single paper might simultaneously address “Krill population dynamics” (topic) and “Large-scale ecosystem processes” (objective).

Our framework enables rapid thematic assessment reports for new literature, allowing research coordinators to guide future research priorities and helping program managers make evidence-based decisions about research needs and funding directions. This real-time capability addresses a critical need, as traditional manual assessment methods are often too slow for policy windows and adaptive management requirements (Atalay et al., 2025; Kaymaz Mühling, 2023). This may provide a useful tool for the five year reports and 10 year reviews required of the RSRMPA under Conservation Measure 91–05 (CCAMLR, 2016). Notably, the RSRMPA comes under its formal review in 2027, thus this tool is particularly timely to support this policy process, especially given that AI-powered species recognition and monitoring tools are increasingly seen as essential for uncovering ‘dark diversity’ and addressing gaps in conservation knowledge (Reynolds et al., 2025).

It is important to emphasize that this framework identifies thematic coverage patterns and policy-relevant topics but does not evaluate research quality, methodological rigor, or whether a policy issue has been adequately addressed. A high number of papers on a topic does not

necessarily indicate sufficient policy analysis, and conversely, a small number of high-quality studies may adequately address certain objectives. The framework serves as a screening and mapping tool to identify what topics are being researched, not to evaluate how well those topics are being addressed. This distinction is fundamental to appropriate deployment of automated classification in conservation decision-making.

4.2.2. Research funding portfolio optimization

Grant managers can leverage the framework's multi-target capability to evaluate incoming proposals for thematic coverage across research topics and conservation priorities within a single assessment. When reviewing funding applications, users can rapidly identify portfolio imbalances and ensure thematic diversity across research investments. This integrated analysis reveals funding applications that optimize multiple criteria simultaneously and can identify proposals that address underrepresented topics while aligning with conservation objectives. For instance, managers can instantly identify proposals that combine evolutionary biology research (underrepresented theme) with specific conservation objectives. The framework can help agencies generate portfolio diversity reports, make real-time funding allocation decisions, and justify strategic research investments with quantitative evidence rather than subjective assessments.

4.2.3. UN sustainable development goals assessment

International organizations can deploy the framework to simultaneously assess research thematic alignment with multiple United Nations Sustainable Development Goals (SDGs) targets and research methodologies in one integrated analysis (United Nations, 2015). When evaluating research contributions, users can quickly identify SDG coverage gaps and ensure alignment with priority development goals. The multi-target approach enables identification of research that contributes to multiple SDGs simultaneously, which is critical given that AI can enable the accomplishment of 134 SDG targets across all goals while potentially inhibiting 59 targets (Vinuesa et al., 2020). For example, using adapted classification targets, a marine conservation study can be automatically classified as addressing SDG 14.2 (sustainable management of marine ecosystems) and SDG 14.3 (ocean acidification), along with methodological approaches, all within a single integrated classification, enabling comprehensive impact assessment impossible with sequential single-target approaches. This integrated approach aligns with recent bibliometric analyses demonstrating the need for systematic methods to map research contributions across multiple SDGs, with studies showing that AI/ML techniques hold promise for SDG achievement but require regulatory oversight to ensure transparency and adherence to ethical standards (Meitei et al., 2023). The framework's ability to simultaneously assess multiple dimensions supports evidence-based approaches to SDG implementation, addressing the sociotechnical challenges of AI deployment while maintaining trust and transparency (Sachs et al., 2019; Visvizi, 2022). Program managers can generate comprehensive SDG alignment reports, track progress toward development targets, and guide strategic research investments toward underrepresented goals with evidence-based decision making.

4.2.4. Species conservation action planning

Conservation organizations and wildlife agencies can use the framework to automatically categorize incoming research by thematic content across conservation strategies and threat assessments in a single multi-target analysis. When evaluating new research, users can rapidly assess whether studies address priority conservation actions and identify thematic gaps in species research coverage. The integrated classification approach reveals research that simultaneously addresses multiple conservation needs. For example, in endangered species research, wildlife managers might discover that new papers simultaneously address habitat protection (conservation action) and disease threats (threat category). This capability aligns with growing efforts to leverage AI for conservation, where machine learning has shown potential to revolutionize how we process and synthesize conservation data (Reynolds et al., 2025). Our framework contributes to this movement by providing automated classification that can significantly reduce the time required for evidence synthesis, a critical bottleneck given that traditional systematic reviews are often too slow for policy windows (Cheng et al., 2018). The framework helps organizations prioritize research themes, justify conservation investments, and track research progress across conservation priorities with evidence-based assessments.

4.2.5. Corporate environmental research portfolio analysis

Private sector sustainability managers can use the framework's multi-target architecture to assess R&D investments across environmental themes and business applications in one integrated evaluation. When reviewing research portfolios, managers can quickly identify sustainability gaps and ensure alignment with corporate environmental goals. This simultaneous classification reveals projects that optimize multiple business criteria while addressing environmental priorities. For example, a sustainability project can be instantly classified as carbon reduction research (environmental theme) and supply chain application (business unit). The system enables companies to generate ESG compliance reports, make data-driven research investment decisions, and demonstrate environmental commitment to stakeholders with comprehensive portfolio analysis.

4.3. Limitations and future directions

Several limitations warrant consideration. The framework's performance depends on training annotation quality, requiring ongoing validation and potential retraining as conservation priorities evolve. While our 295-paper dataset represents substantial expert-annotated coverage for the specialized Ross Sea domain, successful application to other conservation contexts requires: (1) development of domain-specific classification taxonomies aligned with local management objectives, (2) expert annotation of comparable training datasets, (3) fine-tuning to capture domain-specific terminology and research patterns, and (4) validation against local expert judgment. The architectural approach is transferable; each application requires substantial domain-specific development rather than direct model deployment. Periodic model updates may be necessary to maintain optimal performance as scientific vocabulary and methodological approaches evolve. Most importantly, automated classification complements rather than replaces expert judgment, and integration with human expertise remains essential for comprehensive research assessment, aligning with emerging frameworks for human-AI collaboration in scientific domains (Wang et al., 2020).

Furthermore, because the training corpus was annotated by a single domain expert, our reported agreement levels reflect model-versus-one-annotator concordance rather than comparison with a community consensus. A doubly coded subset with independent annotators would be needed to benchmark the model against human inter-annotator agreement and to determine whether the observed 78% Jaccard similarity falls within or below the range of expert disagreement on these multi-label tasks.

Neural network training involves inherent randomness from weight initialization, data shuffling, and stochastic optimization processes that can produce different predictions across model instances trained with identical configurations and datasets. While our ensemble approach with five models and majority voting demonstrates consistent overall performance metrics, individual predictions may vary between training runs, particularly for borderline classifications near decision thresholds. This variability is most pronounced for papers with ambiguous thematic content, where different model instances might produce different classifications for the same research paper. Conservation managers should consider this prediction uncertainty when interpreting individual paper classifications, especially for critical decision-making scenarios, and

may benefit from expert review validation for high-stakes assessments.

The current implementation focuses on English-language peer-reviewed literature, potentially overlooking research published in other languages or formats. The framework's classification targets are inherently constrained by CCAMLR's conservation-focused taxonomy, which emphasizes biological and ecological research topics rather than geological or geophysical studies. While our model successfully identifies and processes geological papers (preventing systematic exclusion), such studies are necessarily classified using biological conservation themes due to the absence of geological categories in the CCAMLR framework. For example, geological studies of mud volcanoes or seismic processes receive classifications such as "Benthic community structure & function" or "Physical & biological habitat changes," which represent reasonable approximations within the conservation context but may not fully capture the geological research scope. This limitation reflects the framework's design for conservation research assessment rather than comprehensive scientific literature analysis, and adaptation to broader research domains would require expansion of classification taxonomies beyond CCAMLR's conservation-specific structure.

Several directions could enhance model performance and applicability. The reliance on title, abstract, and keywords rather than full text represents a meaningful constraint that may bias results toward papers with explicit topic statements in metadata while underrepresenting complex interdisciplinary studies where key themes emerge only in detailed methods or discussion sections. This limitation particularly affects classification of papers with nuanced or emergent research contributions not clearly signaled in abstracts. Incorporating full-text analysis could capture additional conceptual depth, though this must be balanced against practical accessibility advantages for real-world deployment. Training on larger, more diverse conservation corpora could improve generalization to other protected areas. Active learning approaches could prioritize the most informative papers for expert annotation, making dataset expansion more efficient. Integration with citation networks and research impact metrics could add another dimension to research assessment. Finally, developing confidence calibration techniques could provide more reliable uncertainty estimates for individual predictions, helping users identify when expert review is most needed.

Additional developments should consider multilingual capabilities, integration of grey literature and technical reports, ensemble methods and uncertainty quantification approaches to enhance prediction reliability, and expanded taxonomies for interdisciplinary research domains. While transformer-based models excel at content understanding and pattern recognition, they cannot assess research quality, methodological rigor, or actual conservation impact, fundamental limitations arising from their reliance on distributional patterns rather than genuine comprehension (Bender et al., 2021). These models effectively classify research content based on textual patterns but cannot evaluate whether research findings are scientifically sound, practically implementable, or cost-effective. However, emerging approaches in AI development, including retrieval-augmented generation, multi-modal learning, and improved reasoning capabilities, may eventually enable more sophisticated research assessment that goes beyond surface-level classification to evaluate methodological soundness and potential conservation impact. Our focus on the Ross Sea region Marine Protected Area provides proof-of-concept for the multi-target approach. The framework's demonstrated success in this domain establishes technical feasibility, though validation across diverse protected areas and conservation domains with varying management frameworks and research priorities would be necessary to support broader generalizability claims. The modular design enables continual learning capabilities where newly classified papers can be incorporated into the training dataset to create a self-improving system, potentially mitigating temporal degradation while reducing long-term annotation costs through active learning approaches that prioritize the most informative papers for expert review.

5. Conclusion

This paper presents an enhanced multi-target SciBERT framework that addresses the critical need for scalable, systematic assessment of research across thematic and policy dimensions. The framework achieved 70.0% macro F1, representing a 17.6% improvement over traditional machine learning approaches and demonstrating that domain-specific pretraining combined with architectural optimizations provides measurable benefits for conservation text classification. The system successfully identified and quantified descriptive patterns in research coverage across thematic and policy dimensions, with 59% of research topics appearing in fewer than 15 papers, providing quantitative foundations for evidence-based conservation planning through validated multi-target classification. Expert validation confirmed the framework's reliability with 78% agreement across both conservation dimensions and particularly strong performance on policy alignment (94% agreement), demonstrating automated classification with substantial agreement with single-expert annotations while enabling scalable deployment for systematic literature analysis. This transferable approach offers conservation organizations, funding agencies, and researchers a systematic tool to transform ad hoc research prioritization into strategic, evidence-based planning through automated coverage analysis methodology, ultimately strengthening the science-policy interface in conservation practice and complementing expert judgment in comprehensive research evaluation. As AI-assisted frameworks continue to mature, they have the potential to become standard tools integrated into conservation planning workflows, enabling evidence-based decision-making at unprecedented scale while maintaining essential human oversight.

CRediT authorship contribution statement

Chris McCarthy: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Cassandra Brooks:** Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Data curation, Conceptualization. **Troy Sternberg:** Writing – review & editing, Writing – original draft, Supervision, Formal analysis, Conceptualization. **Kyle Shaney:** Writing – review & editing, Writing – original draft, Investigation, Formal analysis, Conceptualization. **Buho Hoshino:** Writing – review & editing, Writing – original draft, Resources, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Dr. Andrew Titmus reports a relationship with National Science Foundation that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Dr. Andrew Titmus (Office of Polar Programs, National Science Foundation) for his valuable contributions to the conceptualization and development of this research. This work was supported by JSPS KAKENHI Project Number 25K03325 "Verification of abandoned oil well plugging effects for carbon offset credits by precise measurement of methane emissions".

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.

org/10.1016/j.ecoinf.2026.103669.

Data availability

The complete implementation of the Multi-Target SciBERT framework, including model training scripts, evaluation pipelines, TF-IDF baseline implementation, ensemble training methodology, and trained model weights, is available under an open-source license at https://github.com/mccarthy-conservation-ai/multitarget-scibert-ross-sea. The expert-annotated dataset of 295 Ross Sea region research papers used for training and evaluation is freely available in the same repository. All code and data necessary to reproduce the results reported in this paper are provided.

References

Ainley, D.G., Morandini, V., Salas, L., Nur, N., Rotella, J., Barton, K., Lyver, P.O., Goetz, K.T., Larue, M., Foster-Dyer, R., Parkinson, C.L., Arrigo, K.R., Van Dijken, G., Beltran, R.S., Kim, S., Brooks, C., Kooyman, G., Ponganis, P.J., Shanhun, F., Anderson, D.P., 2024. Response of indicator species to changes in food web and ocean dynamics of the Ross Sea, Antarctica. Antarctic Sci 36, 290–318. https://doi.org/10.1017/s0954102024000191.

Aria, M., Cuccurullo, C., 2017. Bibliometrix: an R-tool for comprehensive science mapping analysis. J. Inf. Secur. 11, 959–975. https://doi.org/10.1016/j.joi.2017.08.007.

Atalay, A., Perkumienė, D., Safaa, L., Škėma, M., Aleinikovas, M., 2025. Artificial intelligence technologies as smart solutions for sustainable protected areas management. Sustainability 17, 5006. https://doi.org/10.3390/su17115006.

Beltagy, I., Lo, K., Cohan, A., 2019. SciBERT: A Pretrained Language Model for Scientific Text. https://doi.org/10.48550/ARXIV.1903.10676.

Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., 2021. On the dangers of stochastic parrots: can language models be too big?. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Presented at the FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency. ACM, Virtual Event Canada, pp. 610–623. https://doi.org/10.1145/3442188.3445922.

Brooks, C.M., Ainley, D.G., 2022. A summary of United States research and monitoring in support of the Ross Sea region marine protected area. Diversity 14, 447. https://doi.org/10.3390/d14060447.

CCAMLR, 2016. Conservation measure 91-05: Ross Sea region marine protected area (conservation measure No. 91–05). In: Commission for the Conservation of Antarctic Marine Living Resources, Hobart, Australia.

Chalkidis, I., Fergadiotis, M., Kotitsas, S., Malakasiotis, P., Aletras, N., Androutsopoulos, I.. An Empirical Study on Large-Scale Multi-Label Text Classification Including Few and Zero-Shot Labels. https://doi.org/10.18653/v1/2020.emnlp-main.607.

Cheng, S.H., Augustin, C., Bethel, A., Gill, D., Anzaroot, S., Brun, J., DeWilde, B., Minnich, R.C., Garside, R., Masuda, Y.J., Miller, D.C., Wilkie, D., Wongbusarakum, S., McKinnon, M.C., 2018. Using machine learning to advance synthesis and use of conservation and environmental evidence. Conserv. Biol. 32, 762–764. https://doi.org/10.1111/cobi.13117.

Cook, C.N., Mascia, M.B., Schwartz, M.W., Possingham, H.P., Fuller, R.A., 2013. Achieving conservation science that bridges the knowledge–action boundary. Conserv. Biol. 27, 669–678. https://doi.org/10.1111/cobi.12050.

Cvitanovic, C., McDonald, J., Hobday, A.J., 2016. From science to action: principles for undertaking environmental research that enables knowledge exchange and evidence-based decision-making. J. Environ. Manag. 183, 864–874. https://doi.org/10.1016/j.jenvman.2016.09.038.

Dunn, A., Vacchi, M., Watters, G., 2017. The Ross Sea region Marine Protected Area Research and Monitoring Plan (CCAMLR Document No. SC-CAMLR-XXXVI/20). Commission for the Conservation of Antarctic Marine Living Resources (CCAMLR), Hobart, Australia.

Finucci, B., Chin, C., O'Neill, H.L., White, W.T., Pinkerton, M.H., 2024. First observation of a skate egg case nursery in the Ross Sea. J. Fish Bio. 104, 1645–1650. https://doi.org/10.1111/jfb.15688.

Gil-Clavel, S., Filatova, T., 2023. Using Natural Language Processing and Networks to Automate Structured Literature Reviews: An Application to Farmers Climate Change Adaptation. https://doi.org/10.48550/ARXIV.2306.09737.

Gill, D.A., Mascia, M.B., Ahmadia, G.N., Glew, L., Lester, S.E., Barnes, M., Craigie, I., Darling, E.S., Free, C.M., Geldmann, J., Holst, S., Jensen, O.P., White, A.T., Basurto, X., Coad, L., Gates, R.D., Guannel, G., Mumby, P.J., Thomas, H., Whitmee, S., Woodley, S., Fox, H.E., 2017. Capacity shortfalls hinder the performance of marine protected areas globally. Nature 543, 665–669. https://doi.org/10.1038/nature21708.

Gupta, T., Zaki, M., Krishnan, N.M.A., Mausam, 2022. MatSciBERT: a materials domain language model for text mining and information extraction. NPJ Comput. Mater. 8. https://doi.org/10.1038/s41524-022-00784-w.

Jones, K.R., Venter, O., Fuller, R.A., Allan, J.R., Maxwell, S.L., Negret, P.J., Watson, J.E.M., 2018. One-third of global protected land is under intense human pressure. Science 360, 788–791. https://doi.org/10.1126/science.aap9565.

Kaymaz Mühling, Ş.M., 2023. Utilizing artificial intelligence (AI) for the identification and management of marine protected areas (MPAs): a review. GEP 11, 118–132. https://doi.org/10.4236/gep.2023.119008.

Kopperud, B.T., Lidgard, S., Liow, L.H., 2022. Enhancing georeferenced biodiversity inventories: automated information extraction from literature records reveal the gaps. PeerJ 10, e13921. https://doi.org/10.7717/peerj.13921.

Lemos, M.C., Arnott, J.C., Ardoin, N.M., Baja, K., Bednarek, A.T., Dewulf, A., Fieseler, C., Goodrich, K.A., Jagannathan, K., Klenk, N., Mach, K.J., Meadow, A.M., Meyer, R., Moss, R., Nichols, L., Sjostrom, K.D., Stults, M., Turnhout, E., Vaughan, C., Wong-Parodi, G., Wyborn, C., 2018. To co-produce or not to co-produce. Nat. Sustainability 1, 722–724. https://doi.org/10.1038/s41893-018-0191-0.

Likhareva, D., Sankaran, H., Thiyagarajan, S., 2024. Empowering Interdisciplinary Research with BERT-Based Models: An Approach through SciBERT-CNN with Topic Modeling. https://doi.org/10.48550/ARXIV.2404.13078.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollar, P., 2017. Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV). Presented at the 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, Venice, pp. 2999–3007. https://doi.org/10.1109/ICCV.2017.324.

Marine Conservation Institute, 2024. MPAtlas – Marine Protected Areas Atlas.

Meitei, A.J., Rai, P., Rajkishan, S.S., 2023. Application of AI/ML techniques in achieving SDGs: a bibliometric study. Environ. Dev. Sustain. 27, 281–317. https://doi.org/10.1007/s10668-023-03935-1.

Minutoli, R., Bonanno, A., Guglielmo, L., Bergamasco, Alessandro, Grillo, M., Schiaparelli, S., Barra, M., Bergamasco, Andrea, Remirens, A., Genovese, S., Granata, A., 2024. Biodiversity and functioning of mesozooplankton in a changing Ross Sea. Deep Sea Research Part II. Topical Studies Oceanogra 217, 105401. https://doi.org/10.1016/j.dsr2.2024.105401.

Pike, E.P., MacCarthy, J.M.C., Hameed, S.O., Harasta, N., Grorud-Colvert, K., Sullivan-Stack, J., Claudet, J., Horta E Costa, B., Gonçalves, E.J., Villagomez, A., Morgan, L., 2024. Ocean protection quality is lagging behind quantity: applying a scientific framework to assess real marine protected area progress against the 30 by 30 target. Conserv. Lett. 17, e13020. https://doi.org/10.1111/conl.13020.

Piwowar, H., Priem, J., Larivière, V., Alperin, J.P., Matthias, L., Norlander, B., Farley, A., West, J., Haustein, S., 2018. The state of OA: a large-scale analysis of the prevalence and impact of open access articles. PeerJ 6, e4375. https://doi.org/10.7717/peerj.4375.

Reynolds, S.A., Beery, S., Burgess, N., Burgman, M., Butchart, S.H.M., Cooke, S.J., Coomes, D., Danielsen, F., Di Minin, E., Durán, A.P., Gassert, F., Hinsley, A., Jaffer, S., Jones, J.P.G., Li, B.V., Mac Aodha, O., Madhavapeddy, A., O'Donnell, S.A. L., Oxbury, W.M., Peck, L., Pettorelli, N., Rodríguez, J.P., Shuckburgh, E., Strassburg, B., Yamashita, H., Miao, Z., Sutherland, W.J., 2025. The potential for AI to revolutionize conservation: a horizon scan. Trends Ecol. Evol. 40, 191–207. https://doi.org/10.1016/j.tree.2024.11.013.

Rogers, A., Kovaleva, O., Rumshisky, A., 2020. A primer in BERTology: what we know about how BERT works. Trans. Assoc. Comp. Ling. 8, 842–866. https://doi.org/10.1162/tacl_a_00349.

Ruder, S., 2017. An Overview of Multi-task Learning in Deep Neural Networks. https://doi.org/10.48550/ARXIV.1706.05098.

Sabo, A.N., Berger-Tal, O., Blumstein, D.T., Greggor, A.L., Swaddle, J.P., 2024. Conservation practitioners' and researchers' needs for bridging the knowledge–action gap. Front. Conserv. Sci. 5, 1415127. https://doi.org/10.3389/fcosc.2024.1415127.

Sachs, J.D., Schmidt-Traub, G., Mazzucato, M., Messner, D., Nakicenovic, N., Rockström, J., 2019. Six transformations to achieve the sustainable development goals. Nat. Sustainability 2, 805–814. https://doi.org/10.1038/s41893-019-0352-9.

Sutherland, W.J., Pullin, A.S., Dolman, P.M., Knight, T.M., 2004. The need for evidence-based conservation. Trends Ecol. Evol. 19, 305–308. https://doi.org/10.1016/j.tree.2004.03.018.

Toomey, A.H., Knight, A.T., Barlow, J., 2017. Navigating the space between research and implementation in conservation. Conserv. Lett. 10, 619–625. https://doi.org/10.1111/conl.12315.

UNEP-WCMC, IUCN, 2021. Protected Planet Report 2020.

United Nations, 2015. Transforming Our World: The 2030 Agenda for Sustainable Development (Resolution No. A/RES/70/1). United Nations General Assembly.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S.D., Tegmark, M., Fuso Nerini, F., 2020. The role of artificial intelligence in achieving the sustainable development goals. Nat. Commun. 11. https://doi.org/10.1038/s41467-019-14108-y.

Visvizi, A., 2022. Artificial intelligence (AI) and sustainable development goals (SDGs): exploring the impact of AI on politics and society. Sustainability 14, 1730. https://doi.org/10.3390/su14031730.

Wang, D., Churchill, E., Maes, P., Fan, X., Shneiderman, B., Shi, Y., Wang, Q., 2020. From human-human collaboration to human-AI collaboration: designing AI systems that can work together with people. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. Presented at the CHI '20: CHI Conference on Human Factors in Computing Systems. ACM, Honolulu HI USA, pp. 1–6. https://doi.org/10.1145/3334480.3381069.

Westgate, M.J., Barton, P.S., Pierson, J.C., Lindenmayer, D.B., 2015. Text analysis tools for identification of emerging topics and research gaps in conservation science. Conserv. Biol. 29, 1606–1614. https://doi.org/10.1111/cobi.12605.

Zhang, Y., Yang, Q., 2017. A Survey on Multi-task Learning. https://doi.org/10.48550/ARXIV.1707.08114.

Zhang, M.-L., Zhou, Z.-H., 2014. A review on multi-label learning algorithms. IEEE Trans. Knowl. Data Eng. 26, 1819–1837. https://doi.org/10.1109/TKDE.2013.39.