



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS TRINDADE
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA E ENGENHARIA DE
MATERIAIS

Henrique de Medeiros Back

Aplicações de aprendizado de máquina para identificar e abordar lacunas e vieses na pesquisa sobre o destino do Lixo Plástico Marinho

Florianópolis

2024

Henrique de Medeiros Back

Aplicações de aprendizado de máquina para identificar e abordar lacunas e vieses na pesquisa sobre o destino do Lixo Plástico Marinho

Tese submetida ao Programa de Pós-Graduação em Ciência e Engenharia de Materiais da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de Doutor.

Orientador: Prof. Orestes Estevam Alarcon, Dr.
Coorientadora: Prof^a Patrizia Ziveri, Dra.

Florianópolis

2024

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC. Dados inseridos pelo próprio autor.

Back, Henrique de Medeiros

Aplicações de aprendizado de máquina para identificar e abordar lacunas e vieses na pesquisa sobre o destino do Lixo Plástico Marinho / Henrique de Medeiros Back ; orientador, Orestes Estevam Alarcon, coorientadora, Patrizia Ziveri, 2024.
152 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Ciência e Engenharia de Materiais, Florianópolis, 2024.

Inclui referências.

1. Ciência e Engenharia de Materiais. 2. Poluição Marinha. 3. Poluição por plásticos. 4. Aprendizado de Máquina. 5. Inteligência Artificial. I. Alarcon, Orestes Estevam. II. Ziveri, Patrizia. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência e Engenharia de Materiais. IV. Título.

Henrique de Medeiros Back

Aplicações de aprendizado de máquina para identificar e abordar lacunas e vieses na pesquisa sobre o destino do Lixo Plástico Marinho

O presente trabalho em nível de Doutorado foi avaliado e aprovado, em 14 de maio de 2024, pela banca examinadora composta pelos seguintes membros:

Prof. Alexander Turra, Dr.

Instituto Oceanográfico - Universidade Federal de São Paulo

Prof. Carlos Eduardo Peres Teixeira, Dr.

Instituto de Ciências do Mar - Universidade Federal do Ceará

Prof. Pedro Veras Guimarães, Dr.

Departamento de Eng. Mecânica - Universidade Federal de Santa Catarina

Prof. Sergio Rossi, Dr.

Università del Salento

Certificamos que esta é a versão original e final do trabalho de conclusão que foi julgado adequado para obtenção do título de Doutor em Materiais.



Documento assinado digitalmente

Aloisio Nelmo Klein

Data: 12/07/2024 16:27:02-0300

CPF: ***.316.770-**

Verifique as assinaturas em <https://v.ufsc.br>

Coordenação do Programa de Pós-Graduação



Documento assinado digitalmente

ORESTES ESTEVAM ALARCON

Data: 07/07/2024 10:58:58-0300

CPF: ***.644.417-**

Verifique as assinaturas em <https://v.ufsc.br>

Prof. Orestes Estevam Alarcon, Dr.

Orientador

PATRIZIA ZIVERI -
DNI X7352385K

Digitally signed by PATRIZIA
ZIVERI - DNI X7352385K

Date: 2024.07.05 15:54:33 +02'00'

Profa. Patrizia Ziveri, Dra.

Coorientadora

Florianópolis, 2024

AGRADECIMENTOS

Agradeço, em primeiro lugar, ao meu orientador, Orestes Estevam Alarcon, por me guiar nessa caminhada.

Agradeço à minha coorientadora, Patrizia Ziveri, por todo o incentivo e por acreditar no meu trabalho.

Agradeço, imensamente, à minha família por todo o apoio, sem o qual esta tese não seria possível.

Agradeço à Georgia, minha esposa, que esteve ao meu lado nos momentos mais difíceis e celebrando cada vitória.

Agradeço aos meus amigos e coautores, Daphiny, Edson, Camila e Mika pela dedicação e pela orientação. Este trabalho deve muito à sua colaboração.

Agradeço aos meus colegas e amigos na UFSC, que contribuíram para minha evolução como pesquisador e profissional e pela colaboração nos tantos projetos em que estivemos envolvidos juntos.

Agradeço à equipe do Veleiro ECO por todo o aprendizado, pelo companheirismo e por ampliar meus horizontes.

Agradeço aos colegas e amigos da Universidade Autônoma de Barcelona por me receberem tão bem e pelas conversas empolgantes e perspicazes.

Agradeço aos amigos que tive a sorte de conhecer antes e durante o doutoramento.

Agradeço a todos que contribuíram direta ou indiretamente e que, por ventura, não foram mencionados aqui.

RESUMO

Embora tenha havido investigação significativa sobre o Lixo Plástico Marinho (MPL), especialmente na última década, ainda existem lacunas de conhecimento sobre as suas fontes, destinos e impactos. No que diz respeito ao balanço de massa de plásticos no oceano, recursos e infraestruturas limitados, bem como as dificuldades de amostragem e identificação, dificultam a produção de dados de boa qualidade sobre a degradação, distribuição e tempo de residência dos plásticos em diferentes compartimentos marinhos. No Sul Global, isto pode ser ainda mais preocupante devido aos fundos de investigação mais escassos. Esta tese contribui para uma melhor compreensão do MPL ao utilizar métodos de Machine Learning para melhorar o uso de dados e conhecimentos já disponíveis, otimizando esforços com menos infraestrutura e demandas de financiamento. A tese é composta por três artigos científicos que tratam o tema por diferentes abordagens. O Artigo 1 fornece uma visão histórica do campo de pesquisa, traçando tendências desde a década de 1970 até hoje, particularmente a mudança de foco da escala macro para a micro. A utilização de métodos de Processamento de Linguagem Natural (NLP) permitiu levar em conta mais de 9 mil publicações na área, oferecendo informações valiosas sobre o impacto das políticas e da conscientização pública nos esforços de pesquisa - e vice-versa. Também mostra a distribuição regional de estudos pelo mundo e tópicos de pesquisa existentes. Por último, demonstra que os investigadores dependem cada vez mais de uma parcela menor das publicações, o que significa esforços de investigação menos eficientes. O Artigo 2 centra-se na degradação ambiental dos plásticos e destaca a importância das colaborações entre investigadores na área do MPL e da ciência dos materiais. Ele usa métodos de NLP semelhantes para avaliar a frequência com que propriedades relevantes de polímeros estão presentes em diferentes tópicos no campo do MPL. Ele contém uma longa introdução que visa educar pesquisadores de diferentes disciplinas em conceitos básicos da ciência dos polímeros e destaca os principais componentes dos estudos de degradação que parecem estar faltando na pesquisa de MPL. Ao mesmo tempo, aponta o impacto que esta área pode ter na ciência dos polímeros, preenchendo lacunas anteriormente ignoradas por ela. Finalmente, o Artigo 3 define um método para treinar e avaliar algoritmos de Aprendizado de Máquina (ML) para a classificação de microplásticos a partir dos seus espectros de FTIR. Isto é crucial em muitas

investigações, uma vez que a identificação adequada dá pistas sobre as fontes, destinos e impactos dos plásticos. Os modelos de classificação permitem a identificação rápida em massa e o método robusto garante confiança na atribuição de classes, melhorias significativas em comparação com a inspeção manual ou as tradicionais buscas por similaridade em bases de dados. Em última análise, demonstra que os modelos de ML podem ajudar os investigadores a construir bases de conhecimento mais abrangentes sobre a distribuição de plásticos no oceano.

Keywords: poluição marinha; lixo marinho; microplásticos; interdisciplinaridade; aprendizado de máquina.

ABSTRACT

While there has been significant research on MPL, particularly in the last decade, there are still knowledge gaps on its sources, fates and impacts. In what concerns the mass balance of plastics in the ocean, limited funds and infrastructure, as well as difficulties in sampling and identification hinder the production of good-quality data on plastics degradation, distribution and residence time at different marine compartments. In the Global South, this can be even more concerning due to more scarce research funds. This thesis contributes to a better understanding of MPL by using Machine Learning methods to improve the use of data and knowledge that is already available, optimising efforts with less infrastructure and funding demands. It consists of three scientific articles that address the issue through different approaches. Article 1 provides a historical overview of the research field tracing trends from the 1970s to today, particularly the change in focus from the macro to the micro scale. Using Natural Language Processing (NLP) methods allowed taking into account over 9 thousand publications in the field, offering valuable insights into the impact of policy and public awareness on research efforts - and vice versa. It also shows the world's regional distribution of studies and topics of research. Finally, it shows that researchers are increasingly reliant on fewer publications, meaning less efficient research efforts. Article 2 focuses on the environmental degradation of plastics and highlights the importance of collaborations between researchers in the field of MPL and materials science. It uses similar NLP methods to assess how frequently relevant polymer properties are present in different topics within the MPL field. It contains a lengthy introduction that aims to educate researchers from different disciplines in basic polymer science concepts and highlights key components of degradation studies that seem to be missing in MPL research. At the same time, it points out the impact this field can have on polymer science, filling gaps previously overlooked by it. Finally, Article 3 defines a method to train and evaluate machine learning algorithms for the classification of microplastics from their FTIR spectra. This is crucial in many investigations as proper identification gives hints to plastics sources, fates and impacts. The classification models allow for the rapid identification in bulk and the robust method ensures confidence in the assignment of classes, significant improvements compared to manual inspection or traditional database

searches. Ultimately, it demonstrates that ML models can help researchers build more comprehensive knowledge bases on plastics distribution in the ocean.

Keywords: marine pollution; marine litter; microplastics; interdisciplinarity; machine learning.

RESUMO EXPANDIDO

Esta tese explora as intersecções entre o campo de estudos sobre Lixo Plástico Marinho, a Ciência dos Materiais e o Aprendizado de Máquina (ML, do inglês *Machine Learning*), visando aprofundar a nossa compreensão do destino e dos efeitos do Lixo Plástico Marinho no ambiente e na sociedade. Ao elucidar essas conexões, esta pesquisa pretende contribuir para o debate em torno das alternativas ao modelo atual de produção e o descarte desenfreado de plásticos, que abrange toda cadeia produtiva, desde a extração de matéria-prima fóssil até o consumo e descarte excessivos.

Os plásticos emergiram como poluentes emblemáticos no século XXI, paralelamente ao desenvolvimento global, à industrialização e ao consumo. Embora os plásticos só tenham surgido há algumas décadas, o seu impacto na sociedade foi profundo, remodelando os hábitos de consumo e promovendo uma cultura de uso único e descarte. O advento dos polímeros sintéticos marcou uma revolução tecnológica, oferecendo uma vasta gama de materiais com propriedades diversas. Essa versatilidade permitiu a integração de produtos plásticos em todas as facetas da atividade humana, desde a lavoura até a mesa, do transporte à construção e tantos outros setores. No entanto, essa rápida adoção de plásticos levou à proliferação de produtos descartáveis, perpetuando uma economia linear de utilização e descarte. A produção global de plástico vem se acelerando desde a década de cinquenta, atingindo 400 milhões de toneladas em 2022, com o setor de embalagens como principal impulsionador. Consequentemente, a geração de resíduos plásticos aumentou, com uma parcela significativa chegando ao meio ambiente. Os plásticos não só contaminam os oceanos, mas também permeiam o ar e a terra, estando presentes em matrizes bióticas e abióticas impondo desafios complexos à gestão de resíduos e à proteção ambiental.

O lixo plástico marinho (MPL, do inglês *Marine Plastic Litter*) é um problema premente, com uma estimativa de 4,8 a 12,7 milhões de toneladas entrando anualmente nos oceanos e o fluxo pode possivelmente triplicar até 2040. A acumulação de plásticos no ambiente está correlacionada com a sua produção, mas embora possam persistir indefinidamente no ambiente, existe um grande desalinhamento entre o volume de acumulação observado e o fluxo esperado, sugerindo lacunas na nossa compreensão do seu destino e distribuição. Igualmente, os impactos da poluição ainda não são amplamente compreendidos e dificilmente quantificados. No oceano, o macrolixo apresenta riscos em razão da ingestão e emaranhamento, enquanto o microlixo levanta preocupações sobre a toxicidade e a bioacumulação em organismos marinhos. Já a exposição humana aos microplásticos, através do ar e dos alimentos, levanta questões quanto à saúde. Isso porque a presença desses poluentes em tecidos de muitos órgãos e até no sistema cardiovascular geram preocupação, com estudos já demonstrando seus efeitos.

O início das pesquisas sobre MPL data de meados dos anos 1970, mas expandiu-se rapidamente na última década impulsionada pelo crescente reconhecimento da ubiquidade da poluição e dos seus impactos ambientais e à saúde humana. Apesar disso, ainda existem grandes lacunas de dados e conhecimento acerca do MPL, o que dificulta o entendimento das suas fontes, destinos e impactos. Elas estão relacionadas com limitações de recursos e infraestrutura e com a própria dificuldade de obtenção de dados e monitoramento ambientais de qualidade, tanto sob o aspecto da amostragem quanto da identificação do lixo plástico marinho.

A colaboração interdisciplinar é crucial para abordar as complexidades do MPL. As ciências ambientais constituem o núcleo do campo de pesquisa, mas as contribuições da farmacologia, toxicologia, ciências sociais e economia, por exemplo, são vitais. A ciência dos materiais, com a sua experiência em engenharia de polímeros, pode oferecer insights sobre o comportamento e o destino dos plásticos no meio ambiente. Por outro lado, as ciências da computação, particularmente o Aprendizado de Máquina, pode ser utilizada como método capaz de aprimorar o uso de dados e acelerar descobertas.

O Aprendizado de Máquina é uma subárea da Inteligência Artificial (IA), mas é frequentemente tratado como seu sinônimo, devido à importância fundamental no avanço deste campo. Como uma analogia ao aprendizado humano pela experiência, modelos de ML são capazes de fazer distinções e previsões ao analisar eventos passados. A IA está se tornando cada vez mais presente na ciência e no cotidiano, especialmente após o surgimento de IA conversacionais avançadas, como o GPT (*Generative Pre-trained Transformer*) e o BERT (*Bidirectional Encoder Representations from Transformers*). Essas tecnologias estão revolucionando a forma como interagimos com a informação, oferecendo novas possibilidades para a pesquisa científica e transformando a maneira como realizamos tarefas diárias.

Considerando a abrangência desta tese, ela foi estruturada em três objetivos específicos principais. Cada um deles é abordado de forma independente em cada um dos três capítulos empíricos (artigos científicos). Embora sejam estudos individuais, existe um elemento comum que os integra: a hipótese é que avanços notáveis no conhecimento sobre lixo marinho e poluição plástica são proporcionados pela aplicação da AI na literatura científica disponível.

Objetivo principal	Entrega	Objetivos específicos	Subobjetivos
Avançar nossa compreensão da poluição marinha por plásticos, empregando uma abordagem interdisciplinar que integra as ciências ambientais e de materiais e o aprendizado de máquina	Artigo 1	Fornecer uma análise da literatura sobre Lixo Plástico Marinho para compreender:	os desenvolvimentos e tendências emergentes no campo
			a distribuição global de estudos
			como o campo está avançando em termos do impacto das publicações nas citações ao longo do tempo.
	Artigo 2	Fomentar colaborações entre a ciência dos materiais e a pesquisa sobre lixo plástico marinho	identificar as principais propriedades dos polímeros que influenciam sua degradação ambiental
			avaliar sua ocorrência em estudos de MPL
			educar os pesquisadores do MPL sobre importantes fatores de degradação
	Artigo 3	Melhorar os métodos de classificação de microplásticos por meio de uma metodologia robusta de Aprendizado de Máquina	desenvolver um pipeline estatisticamente robusto para treinar e testar modelos de ML
			avaliar o desempenho de diferentes modelos e escolher o melhor

O primeiro artigo (Capítulo 3 - MAPPING THE LANDSCAPE OF MARINE PLASTIC LITTER RESEARCH WITH LARGE LANGUAGE MODELS) tem por objetivo fornecer um panorama abrangente do estado da arte sobre o lixo plástico marinho e entender as causas e implicações do crescimento abrupto da literatura científica no tema. O artigo fornece uma visão histórica do campo de pesquisa, classificando publicações em tópicos de interesse e traçando tendências no número de publicações e citações para cada um desde a década de 1970 até os tópicos emergentes de hoje. Igualmente, levanta dados sobre a distribuição geográfica dos estudos e aponta lacunas de pesquisa nas diversas bacias oceânicas. Foram utilizadas técnicas de Processamento de Linguagem Natural (NLP) para identificar tópicos e extrair as informações a partir dos resumos dos artigos de maneira automática. Notavelmente, este método permitiu ter em conta mais de nove mil publicações, o que não seria factível sem o uso desta metodologia específica. O artigo oferece análises valiosas sobre o impacto das políticas públicas e da sensibilização do público nos esforços de investigação, e vice-versa. Os dados coletados fornecem uma visão geral importante dos avanços na investigação do lixo marinho e de eventos importantes de alto nível internacional que estimularam o campo de pesquisa e o financiamento. O artigo mostra ainda que as publicações mais antigas na área, datando dos anos 70 até aproximadamente o começo da década de 2010, tinham um foco marcado no estudo do macro lixo e dos impactos relacionados a ele, tais como, o emaranhamento de focas e tartarugas em redes de pesca fantasma. Contudo, o crescimento acelerado no número de publicações, que se observa a partir de meados da última década, é impulsionado sobretudo por publicações em temas como microplástico, microrganismos e impactos toxicológicos. Isso mostra uma transformação do campo de estudo, muito clara, do foco macro para o microscópico. Pode-se dizer que o reconhecimento da existência de partículas nessa escala como contaminantes ambientais fez soar um alerta para a comunidade científica e os tomadores de decisão que ampliaram os esforços de pesquisa nesta direção. Por consequência, isso fez surgir novas linhas de pesquisa antes não exploradas. Aqui, a metáfora da Ilha do Conhecimento se aplica muito bem: à medida que a ilha cresce, sua fronteira também se expande. Contudo, o artigo traz dois contrapontos à ideia do crescimento no número de publicações como manifestação simples de um maior entendimento do problema do lixo marinho. O primeiro trata da heterogeneidade geográfica do campo de pesquisa e o segundo da difícil assimilação de novidades pelos pesquisadores. O Processamento de Linguagem Natural (NLP) permitiu categorizar a literatura de acordo com o local de estudo de cada publicação. Uma ferramenta chamada Reconhecimento de Entidades Nomeadas identificou menções à localização dos estudos presentes nos resumos das publicações. Com isso, foi possível obter a posição geográfica exata do local mencionado. Os resultados mostram que a pesquisa não é feita de maneira homogênea pelo planeta. Ao contrário, há um forte viés em direção ao Norte Global, entendido como o conjunto dos países *desenvolvidos*, que também apresentam maior relação do investimento em pesquisa frente ao tamanho da sua economia. Portanto, o problema do lixo marinho permanece ainda mais incompreendido no Sul Global, onde seus impactos podem ser ainda maiores, visto a maior dependência dessas economias à exploração da natureza, seja pelo extrativismo da pesca ou pelo turismo, atividades que sofrem

diretamente com os impactos da poluição por plásticos. Mais ainda, a pesquisa na área encontra-se concentrada em poucos tópicos de investigação. Nessas regiões onde o número absoluto de trabalhos é pequeno, isso implica em muitas lacunas de pesquisa. É verdade que muito do que se entende sobre as causas e consequências da poluição marinha por plásticos independe da localização geográfica e, por isso, pode ser generalizado. Entretanto, há também particularidades no que tange às fontes e à concentração dos diversos tipos de plásticos, e aos impactos em organismos e ecossistemas específicos que requerem pesquisa e monitoramento locais.

Por outro lado, o artigo aponta desafios que parecem emergir no campo de pesquisa em decorrência do crescimento acelerado no número de publicações. Apesar do aumento na produção científica, os pesquisadores estão cada vez mais dependentes de um número reduzido de publicações. O aumento da desigualdade na citação de artigos ao longo dos anos, evidenciado pelo coeficiente de Gini, sugere uma crescente dificuldade na disseminação eficaz do conhecimento, o que pode limitar a diversidade de ideias e abordagens no campo.

O segundo artigo (Capítulo 4 - INTERSECTIONS BETWEEN MATERIALS SCIENCE AND MARINE PLASTICS TO ADDRESS ENVIRONMENTAL DEGRADATION DRIVERS: A MACHINE LEARNING APPROACH) centra-se na degradação ambiental dos plásticos, e tem por objetivo construir pontes entre a ciência dos materiais e a pesquisa sobre lixo plástico marinho. A investigação sobre o destino dos plásticos no ambiente, especificamente a sua dispersão e degradação, é um dos pilares para entender os processos de remoção do plástico da superfície do oceano. Ela está na fronteira do conhecimento na área, devido à incapacidade dos modelos atuais em aproximar o balanço de massa. O trabalho apresentado neste artigo é um esforço no sentido de entender o destino dos plásticos a partir de fundamentos da Ciência dos Materiais. Concretamente, buscou-se identificar qual o descompasso entre esta ciência e a pesquisa em MPL e elucidar conceitos importantes para que ambas possam colaborar efetivamente. Diante da vasta literatura sobre a degradação de polímeros, buscou-se avaliar como ela está sendo empregada na mais recente área de MPL. O artigo identifica as principais propriedades dos polímeros que influenciam sua degradação ambiental e avalia sua ocorrência em estudos de MPL.

O artigo contém uma longa introdução sobre os mecanismos de degradação e como as propriedades dos polímeros influenciam sua taxa de fragmentação. Pretende, portanto, informar pesquisadores de diferentes disciplinas acerca de conceitos básicos da Ciência dos Polímeros. Argumenta-se que uma compreensão mais aprofundada dos *drivers* de degradação dos plásticos a partir da ciência dos polímeros deve ajudar os estudos fundamentais de degradação realizados pela comunidade de pesquisa do Lixo Plástico Marinho.

Este artigo utiliza métodos de NLP semelhantes ao do primeiro artigo para classificar resumos em tópicos, filtrar publicações relevantes com intervenção manual mínima e identificar os tópicos e artigos mais voltados à degradação. A partir disso, pôde-se avaliar a frequência com que *drivers* de degradação aparecem na literatura de MPL, com foco nos tópicos relacionados à degradação. Os resultados mostram que os *drivers* ambientais, como raios ultravioleta e temperatura, são mais recorrentes do que as propriedades dos polímeros, como cristalinidade e peso molecular. Ainda assim, os fatores abióticos, como luz UV, são pouco abordados, se comparados aos fatores bióticos, como microrganismos. O tópico biodegradação, portanto, aparece como um tema de crescente interesse nos estudos de degradação dos plásticos.

Os resultados apontam uma grande oportunidade destes estudos contribuírem para a ciência dos materiais em si. Isto porque os *drivers* biológicos da degradação não tiveram o mesmo destaque dos *drivers* físicos e até mesmo químicos ao longo da história da Engenharia de Polímeros, uma vez que ela está tradicionalmente menos ocupada com o fim da vida útil dos produtos plásticos. Similarmente, o artigo destaca a importância de cientistas de materiais no estudo acerca do destino dos plásticos marinhos.

O terceiro artigo (Capítulo 5 - TRAINING AND EVALUATING MACHINE LEARNING ALGORITHMS FOR OCEAN MICROPLASTICS CLASSIFICATION THROUGH VIBRATIONAL SPECTROSCOPY) objetiva melhorar os métodos de classificação de plásticos marinhos a fim de facilitar a obtenção de dados de qualidade sobre a composição e concentração desses poluentes. Os modelos de aprendizado de máquina utilizados têm o potencial de ajudar os pesquisadores a construir bases de conhecimento mais amplas sobre a distribuição de plásticos no oceano e melhorar as estimativas do balanço de massa.

A caracterização adequada desses contaminantes pode fornecer informações substanciais sobre as entradas e o transporte nos oceanos, as taxas de degradação e fragmentação, a interação com a biota, as consequências de sua presença em habitats naturais e, finalmente, apoiar a avaliação e gestão de riscos. Múltiplos parâmetros são relevantes para essas análises, como tamanho, massa, local de amostragem e DNA da plastisfera. Em particular, a determinação da composição química dos plásticos amostrados tem sido usada para supor de onde eles se originaram, qual uso humano eles poderiam ter tido e, até mesmo, estimar sua idade.

A diversidade de materiais é interessante do ponto de vista das aplicações, mas complexifica o estudo dos impactos ambientais e dificulta a identificação de partículas de plástico em meio a substratos ambientais. O método apresentado neste trabalho permite uma caracterização mais confiável mesmo de plásticos bastante degradados.

Este artigo define um método para treinar e avaliar algoritmos de aprendizado de máquina para classificação de microplásticos a partir de seus espectros de FTIR (da sigla em inglês, Espectroscopia no Infravermelho por Transformada de Fourier). Os resultados demonstram que algoritmos de ML podem ter performance significativamente melhor do que as tradicionais buscas por similaridade em bases de dados. Tais modelos permitem a rápida identificação e o método robusto garante confiança na atribuição das classes. Apesar de ainda não poder ser usado para classificar qualquer partícula possível, já que o modelo só reconhece pouco mais de uma dúzia de plásticos, é provável que funcione na maioria dos casos, uma vez que a base de dados de treino reflete o que é normalmente encontrado no ambiente marinho. Não obstante, uma pontuação para a probabilidade de erro do modelo permite a identificação dessas instâncias discrepantes (*outliers*) que podem ser caracterizadas manualmente.

Mais amplamente, esta tese representa um esforço para integrar a Inteligência Artificial ao estudo do lixo plástico marinho, explorando o seu potencial para melhorar a nossa compreensão de um dos desafios sociais e ambientais mais prementes do nosso tempo. Ela está centrada na grande inconsistência no balanço de massa dos plásticos no oceano, que aponta a existência de lacunas de conhecimento significativas.

A produção de dados para investigar e monitorar o estado da poluição em nível regional e global requer financiamento e infraestrutura robustos, que impõem

limitações, e demanda alternativas inteligentes que possam melhorar a utilização dos dados e informações já disponíveis. Do ponto de vista do Sul Global, onde a limitação de recursos para a pesquisa é ainda mais determinante, melhorar métodos de análise de dados, apontar lacunas de conhecimento no nível local e facilitar a produção de dados de alta qualidade sem a necessidade de infraestrutura adicional, representa avanços importantes.

Ao abordar lacunas e vieses nos dados, esta tese contribui para o desenvolvimento de estratégias de mitigação do lixo plástico nos ambientes marinhos. A pesquisa interdisciplinar, envolvendo pesquisadores de diversas regiões e disciplinas, é essencial para preencher essas lacunas e avançar o nosso conhecimento sobre o lixo plástico marinho no intuito de entender suas implicações e encontrar saídas dessa crise. Enquanto os métodos de IA e ML continuam a surgir e a desenvolver-se, a sua aplicação numa série de campos científicos irá crescer. Esta tese mostra apenas algumas maneiras pelas quais eles podem ser úteis para orientar a tomada de decisões mais informadas e acelerar a descoberta científica, analisando grandes quantidades de dados e revelando padrões latentes.

Palavras-chave: poluição marinha; microplásticos; interdisciplinaridade; aprendizado de máquina.

LISTA DE FIGURAS

2 INTRODUÇÃO

- Figura 1: Um artigo publicado no Times Mazine, de 1955, ilustra o zeitgeist em relação aos plásticos à época. 22
- Figura 2: Tendências de crescimento de 1750 a 2019 de múltiplos indicadores socioeconômicos e ambientais, incluindo a produção de plástico 24
- Figura 3: Modelo conceitual de distribuição 3D de plásticos no oceano dado por mudanças na densidade por meio da fragmentação e pela probabilidade $P(x)$ de que ele tenha sido pulverizado na atmosfera, consumido pela biota ou encalhado na costa. d é a densidade da partícula e d_{sea} é a densidade da água do mar. 33
- Figura 4: Exemplo de modelo de classificação K-nn treinado em dados gerados aleatoriamente para 3 clusters. As cores de fundo indicam o limite de decisão do modelo para previsões futuras. 39

3 MAPPING THE LANDSCAPE OF MARINE PLASTIC LITTER RESEARCH WITH LARGE LANGUAGE MODELS

- Figure 1: Topic Frequencies: Changes in the frequency of publications on MPL from 2008 to 2022. Previous years are omitted as there were only a few publications and no apparent trend. A. The number of publications in each topic found by this study. B. Total number of publications over time. 52
- Figure 2 - Distribution of studies across ocean basins: Papers were classified into basins (as defined by the NOAA's World Ocean Database) using an automatic Named Entity Recognition algorithm freely available at spaCy (spacy.io). Colours indicate the total number of papers in each basin according to the colourbar at the bottom of the figure. The spacing of bins in the colourbar is proportional to the number of papers and represents exact values found in the data. Values on the colourbar are to the left of their respective bin. 53
- Figure 3: Distribution of publications per topic and basin following the classifications made with ML algorithms described in the methods. A log scale is used to better visualise differences in values, given their wide range. It is calculated by applying the natural logarithm of $(1+N)$ to avoid errors when calculating log for $N=0$. 55
- Figure 4: Citations distribution: a) Gini coefficients per topic per year. Colours represent topics and lines are "loess" trendlines for each topic. b) The cumulative share of citations by topic versus the cumulative share of publications. Each paper had citations averaged by their age. Values were min-max normalised to highlight the similarity in the behaviour of distinct topics. 56

4 INTERSECTIONS BETWEEN MATERIALS SCIENCE AND MARINE PLASTICS TO ADDRESS ENVIRONMENTAL DEGRADATION DRIVERS: A MACHINE LEARNING APPROACH

- Figure 1: Workflow of the methods for vocabulary construction and text mining. 72
- Figure 2: (a) The materials science tetrahedron; (b) particle size reduction as caused by degradation and fragmentation of plastics. 75
- Figure 3: Example section of a polyethylene macromolecule with possible deviations from the ideal structure and corresponding dissociation energy of some typical bonds. Values are related to the bonds made between carbon and carbon

or hydrogen or oxygen. Source of values: Huheey, Keiter and Keiter, 1993.14 77

Figure 4: Linkages between degradation drivers and processes given with high, moderate and low relevance. Details on these linkages are expressed in the previous sections. 83

Figure 5: Topic trends in marine plastics research – some topics were omitted to facilitate visualisation. 84

Figure 6: Sum of normalized recurrence of degradation drivers in the marine plastics literature by topic considering topic size (number of publications in each topic, set as size of bubbles) and topic impact factor (total number of citations for all publications in a given topic divided by the number of publications in that same topic, set as colour). 86

Figure 7: Normalized recurrence of drivers by topic (colour) for top 10 topics. 87

5 TRAINING AND EVALUATING MACHINE LEARNING ALGORITHMS FOR OCEAN MICROPLASTICS CLASSIFICATION THROUGH VIBRATIONAL SPECTROSCOPY

Figure 1: Graphical representation of the pipeline used with Grid Search for assessing multiple classifiers. 109

Figure 2: Graphical representation of the pipeline used in the Monte Carlo Cross-Validation. After 1700 repetitions the results are averaged. The process was repeated on the seven most promising models. 114

Fig. 3. Graphical representation of the full pipeline methodology. 114

Figure 4: Best performance found within the gridsearch for each evaluated algorithm, namely Random Forests (RF), Support Vector Machine Classifier (SVC), Logistic Regression (LR), K-Nearest Neighbors (kNN), Decision Trees (DT), Gaussian Naive Bayes (GB), considering the best combination of hyperparameters and pre-processing techniques - the baseline correction, Standard scaler, Principal Component analysis (PCA) and (Naive) oversampling. The log-loss shown here is the mean of the five results (cross-validation). The correspondent standard deviation is also presented. 117

Figure 5: Cross-entropy log-loss for the seven most promising models: Logistic Regression with baseline correction; SVC with linear kernel, baseline correction, and oversampling; SVC with rbf kernel, baseline correction, and oversampling; kNN with baseline correction and PCA; RF with baseline correction, standard scaler, and oversampling; Sigmoid NN with baseline correction, standard scaler, and oversampling; Sigmoid NN with baseline correction and standard scaler. 119

Figure 6: Confusion Matrix for the results of MCCV for the final hypothesis. Each row represents the true class, meanwhile each column is the class predicted proportion, on average, by baseline_over_SVC_linear_100.0. Ideally, the matrix should be an identity matrix (perfect match). 121

Figure 7: Sensitivity for every class in the dataset according to results published by Kedzierski et al. (2019) using a k-Nearest Neighbour algorithm, Support Vector Machine Classifier (SVC) in the methodology proposed by Kedzierski et al. (2019), SVC in the Monte Carlo Cross-Validation and SVC in the final test. 123

Figure 8: Detailed score for the MCCV applied to the Jung dataset considering five classes: High-density Poly(ethylene) (HDPE), Low-density Poly(ethylene) (LDPE), a mixture of PE and PP, Poly(propylene) (PP) and Unknown. 125

SUMÁRIO

1 OBJETIVOS E ESTRUTURA

2 INTRODUÇÃO

- 2.1 Plásticos: de solução à poluição
 - 2.1.1 Acumulação no meio ambiente
 - 2.1.2 Composição do Lixo Plástico Marinho
 - 2.1.3 Distribuição do Lixo Plástico Marinho
 - 2.1.4 Relevância da pesquisa em Lixo Plástico Marinho
- 2.2 Aprendizado de Máquina
 - 2.2.1 Classificação
 - 2.2.2 Análise de Aglomerados
 - 2.2.3 Processamento de Linguagem Natural

3 MAPPING THE LANDSCAPE OF MARINE PLASTIC LITTER RESEARCH WITH LARGE LANGUAGE MODELS

- 3.1 Introduction
- 3.2 A Topic Model of the Literature
- 3.3 Topic Trends
- 3.4 Regional distribution of studies
- 3.5 Focal Points and Research Gaps Across Basins
- 3.6 Greedy publications
- 3.7 Discussion
- 3.8 Methods
- Code Availability
- Notes
- References

4 INTERSECTIONS BETWEEN MATERIALS SCIENCE AND MARINE PLASTICS TO ADDRESS ENVIRONMENTAL DEGRADATION DRIVERS: A MACHINE LEARNING APPROACH

- Environmental significance
- Introduction
- Methods
- Degradation impact drivers
- Material characteristics
 - Composition
 - Microstructure
 - Properties
- Weathering factors
 - Ultraviolet radiation
 - Temperature
 - Chemicals
 - Microbes
- Results & discussion
- Conclusions

Notes and references

Appendix A

Degradation Mechanisms

Topic Modeling and Text Mining

Results

Supplementary References

5 TRAINING AND EVALUATING MACHINE LEARNING ALGORITHMS FOR OCEAN MICROPLASTICS CLASSIFICATION THROUGH VIBRATIONAL SPECTROSCOPY

5.1 Introduction

5.2 Materials and methods

5.2.1 The dataset

5.2.2 Dataset pre-processing pipeline

5.2.3 The holdout strategy

5.2.4 The classification algorithms

5.2.5 Gridsearch

5.2.6 Monte Carlo Cross-validation

5.2.7 The Pipeline

5.2.8 Comparability and External Validation

5.2.9 Application of the classifier

5.3 Results/discussion

5.4 Conclusion

Appendix A. Programming and results details

Appendix B. Log-loss: maximum likelihood estimation

Appendix B References

Appendix C. Using the Support Vector Machine Classifier

References

6 SÍNTESE

7 PERSPECTIVAS FUTURAS

REFERÊNCIAS

1 OBJETIVOS E ESTRUTURA

O objetivo geral desta tese de doutorado é avançar na nossa compreensão da poluição marinha por plásticos, empregando uma abordagem interdisciplinar que integra ciências ambientais e de materiais e aprendizado de máquina.

Durante o período em que esta tese foi conduzida, procurou-se evoluir com o campo e os métodos mais modernos. Notavelmente, os métodos de Processamento de Linguagem Natural evoluíram rapidamente com o advento dos Grandes Modelos de Linguagem e foram incorporados nas análises feitas nesta tese.

Esta tese de doutorado é composta por 3 artigos científicos que são resultados da investigação de doutoramento, cada um deles abordando um objetivo específico. Ela contém, ainda, uma introdução aos temas de investigação e um capítulo de síntese que integra os resultados e as perspectivas futuras.

Optou-se por manter a edição dos artigos no formato adequado para cada jornal. Portanto, cada capítulo poderá ter estrutura e organização distintas. Por exemplo, o artigo 1 tem a seção de metodologia ao final, após resultados e discussão, conforme a estrutura dos artigos publicados em jornais da Nature Portfolio (para os quais o artigo foi preparado e submetido). As referências de cada artigo estão listadas ao final de cada um, enquanto as referências dos outros capítulos estão ao final da tese. De maneira semelhante, a numeração de figuras e tabelas é específica para cada artigo.

Objetivos específicos:

- fornecer uma análise da literatura sobre lixo plástico marinho para compreender: a) os desenvolvimentos e tendências emergentes no campo, bem como; b) a distribuição global dos estudos, particularmente quais são as regiões menos estudadas e, finalmente; c) se o campo continua a avançar relativamente ao impacto das publicações nas citações ao longo do tempo. - objetivo abordado no primeiro artigo (Capítulo 3 - MAPPING THE LANDSCAPE OF MARINE PLASTIC LITTER RESEARCH WITH LARGE LANGUAGE MODELS)
- construir pontes entre a ciência dos materiais e a pesquisa sobre lixo plástico marinho, identificando as principais propriedades dos polímeros que

influenciam sua degradação ambiental e avaliando sua ocorrência em estudos de MPL - objetivo abordado no segundo artigo (Capítulo 4 - INTERSECTIONS BETWEEN MATERIALS SCIENCE AND MARINE PLASTICS TO ADDRESS ENVIRONMENTAL DEGRADATION DRIVERS: A MACHINE LEARNING APPROACH)

- aprimorar os métodos de classificação de microplásticos por meio de uma metodologia robusta de aprendizado de máquina - objetivo abordado no terceiro artigo (Capítulo 5 - TRAINING AND EVALUATING MACHINE LEARNING ALGORITHMS FOR OCEAN MICROPLASTICS CLASSIFICATION THROUGH VIBRATIONAL SPECTROSCOPY),

2 INTRODUÇÃO

O lixo plástico marinho (MPL, do inglês Marine Plastic Litter) é um campo de estudo que emergiu em meados dos anos 1970, mas que ganhou importância na esfera pública e no meio acadêmico. Isso se deve à recente descoberta dos microplásticos e de sua pervasividade, mas sobretudo da exposição humana a esses contaminantes. Mesmo com a conscientização e os esforços para conter o descarte inadequado, a emissão para o ambiente continua a crescer (OECD, 2022). Apesar das incertezas quanto ao grau de exposição necessário para que se observem impactos significativos à saúde humana e ao meio ambiente, grande preocupação reside no fato de os plásticos serem contaminantes persistentes, que permanecem no ambiente indefinidamente. Ainda, a fragmentação de produtos plásticos ao longo do tempo produz partículas muito pequenas, ditas micro ou nanoplásticos, que têm efeitos ainda mais adversos. Portanto, baseando-se no princípio da precaução, devemos, como sociedade, evitar a dispersão destes polímeros sintéticos no ambiente.

Em razão da contaminação ambiental por plásticos ser um fato já consumado, surge a necessidade de compreender o destino destes contaminantes na natureza. Esta tese está inserida no âmbito da busca por uma melhor compreensão da escala da contaminação e dos seus efeitos adversos na saúde ambiental e humana. Em particular, faz parte dos esforços da Universidade Federal de Santa Catarina e do grupo de pesquisas do Veleiro ECO em contribuir para o campo de estudos. Ela complementa outras frentes de atuação do ECO, como a mitigação do excesso de gás carbônico na atmosfera, a produção de matéria prima renovável e o desenvolvimento sustentável de comunidades costeiras com o cultivo de algas, as quais o autor também contribuiu paralelamente a este trabalho.

Muito se sabe sobre a degradação dos plásticos por ação de intemperismos, que causam desgaste, perda das propriedades de engenharia e necessidade de reposição. Entretanto, isso não causa a eliminação completa do polímero (conversão em pequenas moléculas, como CO₂ e metano) e de outras substâncias associadas, como os aditivos. A persistência de fragmentos degradados no ambiente requer estudos acerca dos compartimentos ambientais de maior acúmulo, do tempo de residência nestes compartimentos e a direção de potenciais fluxos entre eles. Ainda, é interessante entender em que condição estes fragmentos se encontram. Por

exemplo, qual a distribuição de tamanho das partículas, qual a proporção entre as concentrações de diferentes tipos de polímero ou tipos de geometria (fibras, filmes, fragmentos, etc). Tudo isso tem o potencial de influenciar o tipo e a magnitude de potenciais impactos associados aos plásticos no ambiente.

Por algum tempo imaginava-se que os plásticos eventualmente acabavam na superfície dos oceanos, sobretudo nos giros subtropicais e subpolares. Entretanto, percebeu-se que há um déficit de plásticos no oceano (CÓZAR et al., 2014). Isto é, dado o fluxo de entrada, as concentrações esperadas são muito maiores do que o observado. Por isso, hipotetizou-se vários mecanismos de remoção, como a ingestão pela biota ou a precipitação para o fundo numa espécie de smog marinho. A isso, devem ser adicionadas também a incerteza e a imprecisão nas medidas, sobretudo por conta de limitações metodológicas na detecção de partículas muito pequenas ou muito grandes. Esta tese tem por objetivo contribuir para o entendimento do destino dos plásticos no oceano identificando lacunas e vieses de dados e conhecimento na pesquisa sobre o tema. Ela baseia-se na hipótese de que seria possível avançar o conhecimento na pesquisa sobre lixo plástico marinho com o uso de Inteligência Artificial, a partir de informações que já estejam disponíveis na literatura. Este esforço, não pretende suplantar análises ambientais e estudos experimentais, mas sim incluir uma nova dimensão à pesquisa, favorecendo o uso eficiente dos recursos empregados para produção científica. Isto pode aliviar a pressão por mais dados ambientais e é particularmente relevante no contexto da pesquisa no Sul Global que, como veremos adiante, sofre ainda mais com limites no financiamento da pesquisa.

MPL é um campo de pesquisa bastante interdisciplinar, que agrega disciplinas tão diversas quanto física e sociologia ou farmacologia e materiais. Defendo o argumento desta tese de que a Inteligência Artificial poderia ajudar a integrar o campo, amalgamar conhecimentos separados e, com isso, oferecer novas perspectivas para o enfrentamento deste grande desafio. De fato, já observa-se nos mais recentes relatórios acerca do combate à poluição por plásticos o reconhecimento da necessidade de se utilizar métodos de IA (“Science-Policy Interface for Plastic Pollution”, 2023; SPRING et al., 2023). Entretanto, eles ainda são vagos quanto às maneiras em que esses métodos podem ser empregados. Apesar da utilização da IA crescer no âmbito das pesquisas acadêmicas em geral, é

fato que seu uso é preponderante nas áreas já relacionadas às Ciências da Computação (VAN NOORDEN; PERKEL, 2023).

Portanto, esta tese visa também promover a inclusão da IA no vocabulário do campo da poluição marinha por plásticos e favorecer a sua disseminação. Além disso, os métodos desenvolvidos nesta tese podem ser replicados em outros contextos, inclusive em outras áreas de estudo.

2.1 PLÁSTICOS: DE SOLUÇÃO À POLUIÇÃO

Os avanços na ciência, na tecnologia, na indústria, na produção e no consumo, desencadeados por um processo múltiplo de desenvolvimento, globalização e ocidentalização, degradam implacavelmente a biosfera e ameaçam os sistemas vivos em todo o mundo. Uma faceta dessa ameaça é o fluxo contínuo de substâncias nocivas, resíduos das atividades humanas, para os habitats naturais. Durante a primeira e a segunda décadas do novo milênio, os plásticos tornaram-se, provavelmente, o mais emblemático desses resíduos, juntamente com os gases de efeito estufa. A sua ubiquidade no ambiente reflete aquela na vida cotidiana das pessoas. Isto é, os humanos estão enchendo suas vidas de plástico e, com isso, também os ambientes naturais onde eles não têm utilidade e só podem causar prejuízos. Esse é um exemplo muito claro e didático da nossa indissociabilidade da natureza, mas, acima de tudo, do nosso impacto no meio ambiente.

Embora possa ser paralisante enxergar essa questão como uma verdade inescapável do nosso tempo, os plásticos têm sido uma presença certa nas nossas vidas apenas há algumas décadas. O primeiro plástico totalmente sintético, uma resina termofixa de fenol formaldeído, chamada baquelite, só foi inventado no início do século XX (NATIONAL HISTORIC CHEMICAL LANDMARK, 1993). Mas foi só durante a Segunda Guerra Mundial que a produção de plástico explodiu, desencadeada pela necessidade de substituir materiais como aço, alumínio e borracha na indústria militar, e impulsionada pela indústria química, que crescia rapidamente no período da Guerra (CRAWFORD; QUINN, 2017). Já no início da década de 1950, uma miríade de outros plásticos derivados do petróleo já havia sido inventada. Combinados entre si e com outras substâncias, as características desses materiais podiam ser as mais diversas. Essa versatilidade permitiu a inserção do plástico em todos os nichos de atividade humana. Ao mesmo tempo, um enorme complexo industrial químico desenvolvido durante a guerra poderia, agora,

concentrar-se em aplicações civis, voltando-se, em grande parte, para produtos agro químicos e plásticos (SMITH, 1988). E isso gerou implicações profundas para os seres humanos e para o meio ambiente.

Em contrapartida, há muito tempo os humanos sabem utilizar polímeros produzidos naturalmente. Assim, eles desempenharam funções essenciais à nossa sobrevivência como espécie. Alguns deles acompanharam o ser humano ao longo de sua evolução e foram fundamentais para o desenvolvimento de ferramentas, abrigo e vestimentas. Couro, madeira, lã e outras fibras vegetais, como algodão e cânhamo são polímeros utilizados pelo homem há milênios e seu uso continua difundido até hoje (LAVAN; SWIFT; PUTZEYS, 2008; YADAV et al., 2015). A utilização destes materiais permitiu-nos desenvolvimentos técnicos desde a pré-história e deu-nos grandes vantagens evolutivas.

O desenvolvimento dos plásticos foi um enorme salto tecnológico, proporcionando ao ser humano um catálogo praticamente infinito de materiais com diferentes características. Eles foram rapidamente incorporados em produtos de todos os setores, desde a agricultura e sistemas de abastecimento de alimentos até indústrias de transformação, transporte e construção (GEYER, 2020). Além disso, as tecnologias modernas romperam com as formas tradicionais de obtenção e produção de produtos poliméricos ao serem capazes de sintetizá-los a partir de matérias-primas minerais, particularmente do petróleo.

É fato que o desenvolvimento e a produção de polímeros sintéticos provocaram uma revolução tecnológica. Como resultado de tantas mudanças, uma revolução cultural também pôde ocorrer. Marcadamente, a *cultura do descarte*. Alguns plásticos chegaram ao mercado a preços muito atrativos. Aliada a processos de fabricação eficientes, a produção desses materiais tornou-se tão barata que eles podiam ser usados uma vez e jogados fora, substituídos por outro produto similar de acordo com a conveniência. Foi o início da produção em massa de produtos descartáveis. A adoção desse tipo de consumo pela sociedade foi rápida e moldou muitos hábitos culturais a partir de então.

A Figura 1, extraída de um artigo publicado na Times Magazine em 1955, ilustra bem o *zeitgeist* à época, que permitiu o surgimento de novos mercados e a expansão de outros globalmente. Por exemplo, os alimentos poderiam ser transportados por mais tempo e mantidos frescos em embalagens plásticas de baixo valor ou servidos em um restaurante para serem consumidos em outro lugar sem a

necessidade de devolução dos recipientes. Além disso, a lógica do descarte tomou conta dos hábitos de consumo em geral. Roupas, acessórios, eletrodomésticos, eletrônicos e até carros têm vida útil cada vez menor (SLADE, 2007). Assim, também cresceu a nossa dependência desses produtos e materiais.

Figura 1: Um artigo publicado no Times Magazine, de 1955, ilustra o *zeitgeist* em relação aos plásticos à época



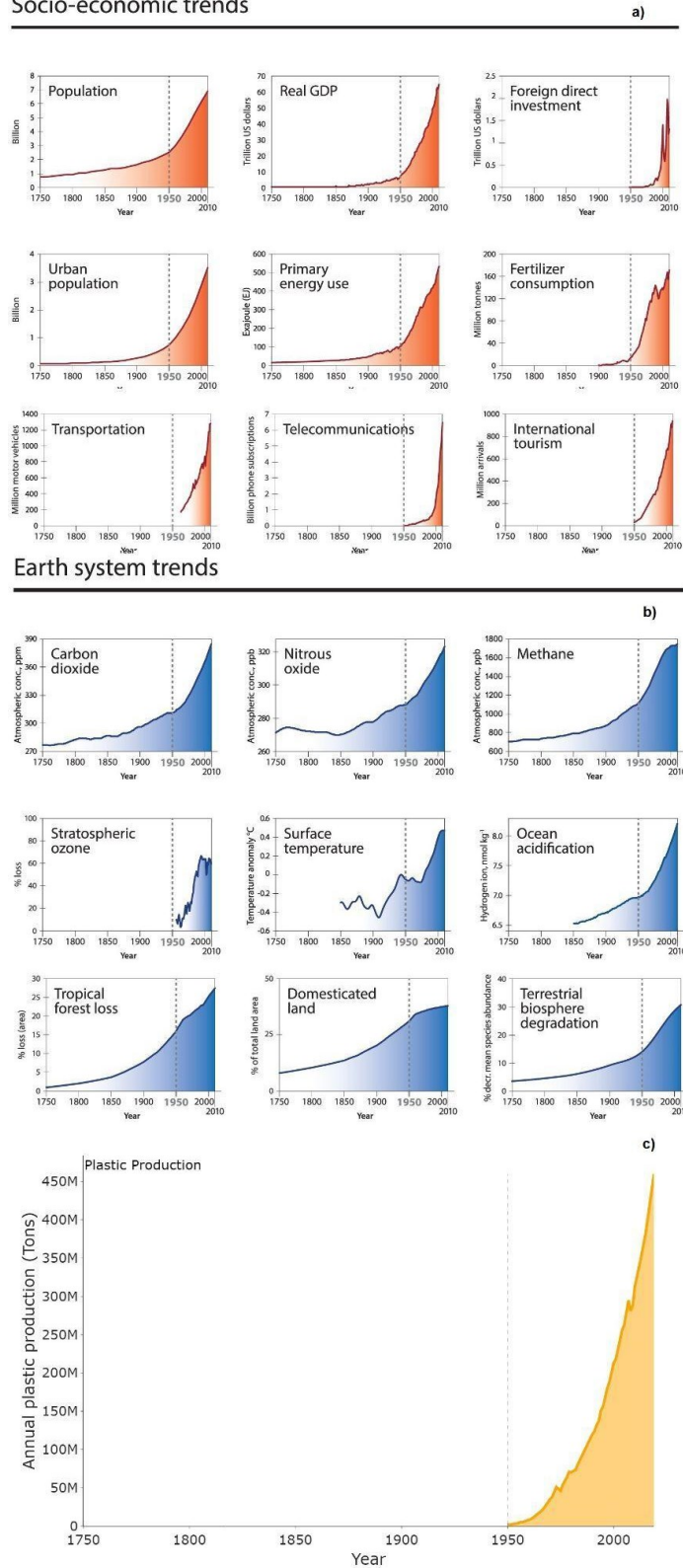
Fonte: Peter Stackpole—*The LIFE Picture Collection/Getty Images*
(COSGROVE, 2014)

A produção global de plásticos já era de 2 milhões de toneladas por ano em 1950, mas esse número é insignificante em comparação as 400 milhões de

toneladas produzidas em 2022 (PLASTICS EUROPE, 2023). Essa tendência é semelhante a de muitas outras atividades humanas no mesmo período, que foi chamado de “A Grande Aceleração” do Antropoceno (STEFFEN et al., 2015). A Figura 2 mostra as tendências de crescimento de múltiplos indicadores. Percebe-se que a produção de plástico segue uma curva semelhante em comparação principalmente, a outros indicadores socioeconômicos. Na indústria do plástico, o setor de embalagens tem sido o principal motor da produção (GEYER; JAMBECK; LAW, 2017), da mesma forma, é o principal contribuinte para a geração de resíduos. Estima-se que a maioria dos produtos plásticos desse setor tendem a sair do ciclo de utilização no mesmo ano da sua produção (GEYER; JAMBECK; LAW, 2017). Comparativamente, os plásticos utilizados na construção civil demoram em média 20 anos. Os resíduos plásticos gerados em 2019 chegaram a 353 milhões de toneladas ($\frac{3}{4}$ da produção nesse ano), e quase dois terços provêm de produtos plásticos com vida útil inferior a cinco anos, sendo 40% provenientes de embalagens, 12% de bens de consumo e 11 % proveniente de vestuário e têxteis. Globalmente, metade dos resíduos é gerada nos países da OCDE (OECD, 2022). Isso mostra como a geração de resíduos está fortemente ligada às taxas de consumo, as quais esses países também lideram. No entanto, países da OCDE representam apenas 11% das emissões de macropásticos para o ambiente, uma vez que as infraestruturas de gestão de resíduos estão mais bem estabelecidas.

Por outro lado, o comércio de resíduos para países de menor renda per capita estimula a utilização insustentável nos países de origem, ao mesmo tempo em que exporta os custos da eliminação de resíduos e da contaminação ambiental.

Figura 2: Tendências de crescimento de 1750 a 2019 de múltiplos indicadores socioeconômicos e ambientais, incluindo a produção de plástico
Socio-economic trends



Fonte: Adaptada pelo autor. Figuras a e b de Steffen *et al.*, 2015 (STEFFEN *et al.*, 2015). Figura c elaborada pelo autor com dados compilados por *Our World In Data* (RITCHIE; SAMBORSKA; ROSER, 2023)

2.1.1 Acumulação no meio ambiente

Durante muito tempo, a contaminação ambiental por resíduos plásticos foi tratada como um problema individual e de educação, considerando o lixo jogado nas ruas como o principal causador da poluição (PATHAK; MARK, 2021). Da mesma forma, a reciclagem tem sido promovida, particularmente pela indústria do plástico, como uma panaceia para os problemas relacionados ao uso excessivo de plásticos, transferindo a responsabilidade pela geração e má gestão de resíduos para indivíduos e governos (PERKINS, 2023).

Hoje, há uma consciência crescente de que a questão é complexa e sistêmica. Embora os indivíduos sejam parte do problema, os sistemas de abastecimento tornam quase impossível consumir sem gerar resíduos plásticos. Os sistemas de gestão de resíduos sólidos urbanos estão sobrecarregados pelo volume de resíduos plásticos gerados nas cidades, especialmente, nos países de menor renda, onde as infraestruturas são mais deficientes. Sem uma gestão adequada, os plásticos provavelmente serão emitidos para o meio ambiente.

A essas emissões diretas podemos acrescentar também a perda de pellets plásticos (grânulos de resina polimérica que servem de matéria-prima à manufatura de produtos de plástico) a durante o transporte para indústrias de transformação e a emissão de micropartículas presentes em certos produtos de limpeza ou cosméticos, por exemplo (AUTA; EMENIKE; FAUZIAH, 2017). Todas estas fontes são facilmente identificadas e poderiam ser mitigadas se fossem implementadas infraestruturas e estratégias adequadas.

No entanto, existem alguns tipos de emissões indiretas, resultantes do desgaste e erosão dos polímeros em uso, que podem ser mais difíceis de evitar. Alguns exemplos são fragmentos de tintas e pneus e fibras de tecidos sintéticos (GESAMP, 2016; WEINSTEIN et al., 2019). Por exemplo, estima-se que uma única lavagem de roupa pode emitir até 700 mil microfibras ao meio ambiente (NAPPER; THOMPSON, 2016). Nesses casos, uma vez que não é possível evitar completamente o desprendimento de partículas na fonte, a mitigação exigiria materiais alternativos que não prejudicasse o ambiente e as pessoas e/ou que não fossem persistentes, bem como melhores tratamentos de águas residuais.

Pode-se dizer que os plásticos já contaminam todos os lugares do planeta. Onde quer que procuremos por plásticos, eles são encontrados. Os oceanos são os

maiores reservatórios, mas sua distribuição ali é heterogênea. As áreas costeiras, os giros subtropicais (LEBRETON et al., 2018) e o fundo do mar são notavelmente mais afetados (KANE; CLARE, 2019). Estima-se que a maior parte do lixo plástico marinho tem origem em terra (LEBRETON et al., 2017).

Como tudo, os plásticos tendem a mover-se para altitudes mais baixas a fim de reduzir a sua energia potencial gravitacional. Essa tendência se concretiza devido à facilidade com que os plásticos são transportados por agentes naturais. Sua baixa densidade e tamanho, portanto, seu baixo peso, fazem com que sejam facilmente carregados por ventos e chuvas (SU et al. 2022), acabando por encontrar um caminho para os rios e para o mar. Estima-se que entre $\frac{1}{4}$ e $\frac{1}{6}$ de todos os plásticos marinhos chegam ao oceano através dos rios, sendo os 20 rios mais poluentes responsáveis por $\frac{2}{3}$ das emissões fluviais (LEBRETON et al., 2017).

Apesar da sua tendência para se acumularem no mar, os plásticos também podem ser encontrados no ar e na terra, tanto em ambientes de água doce como no solo. De certa forma, mesmo os plásticos que são descartados e manejados adequadamente acabam no solo, enterrados em aterros sanitários. Embora os contenham, essas instalações apresentam elevadas concentrações de plástico e são reservatórios que podem apresentar algum risco a longo prazo dada a persistência desses materiais nessas condições. Outras fontes de plásticos terrestres são a deposição atmosférica e as práticas agrícolas, como a cobertura do solo com filmes plásticos e o uso de lodo ou água de estações de tratamento de efluentes para fertilização e irrigação (HASSAN et al., 2023). No entanto, se considerarmos os sedimentos costeiros (por exemplo, areia de praia) como reservatórios terrestres, as suas origens são muito mais diversas e os volumes de acumulação muito maiores.

A presença de plástico na atmosfera não surpreende, visto que materiais particulados suspensos são comuns no ar. Em geral, partículas muito pequenas e/ou partículas com alta razão de aspecto (como pelos ou fibras), mesmo que mais densas que o ar, permanecem suspensas em movimento browniano. É o caso das partículas naturais, como o pólen ou a fuligem (PRATA, 2018). Os plásticos no ar podem ter origem no desprendimento de fibras de tecidos sintéticos de roupas, móveis, etc. ou em materiais de construção, como espumas de isolamento termoacústico. Além disso, o desgaste dos pneus e da pintura nas estradas pode gerar partículas muito pequenas que permanecem em suspensão (DRIS; GASPERI;

TASSIN, 2018). Essas partículas são facilmente transportadas na atmosfera e podem atingir regiões remotas, impactando ecossistemas e pessoas, mesmo que haja pouca atividade humana localmente. No entanto, pensa-se que mesmo essas partículas provavelmente acabarão no oceano (BOUCHER; FRIOT, 2017).

A emissão de plásticos do oceano para a atmosfera também já foi documentada (ALLEN et al., 2020). Plásticos muito pequenos, especialmente fibras, são expelidos por bolhas de ar que estouram ao atingir a superfície da água e permanecem em suspensão através dos mesmos mecanismos discutidos acima. Essas partículas podem ficar presas em ciclos de ejeção e deposição oceânica, efetivamente “saltando” da água, razão pela qual esse processo é chamado de “efeito gafanhoto” (JURADO; DACHS, 2008)

É preciso considerar, além das fontes do plástico ambiental, a sua persistência no meio ambiente. Os plásticos se acumulam porque apresentam taxas de mineralização muito lentas. Ou seja, não podem ser rapidamente transformados em dióxido de carbono, água e sal, como a maior parte da matéria orgânica. Isso se deve, em grande parte, as suas longas cadeias moleculares, que são dificilmente assimiladas por microrganismos que poderiam mineralizá-los completamente. Apesar disso, os plásticos sofrem rápida degradação por agente bióticos e abióticos que comprometem as suas propriedades mecânicas, levando à fragmentação. Os plásticos ambientais tendem a se fragmentar em partículas cada vez menores, um único saco plástico pode dar origem a milhões de partículas plásticas, chamadas micro e nanoplásticos (BRANDON; GOLDSTEIN; OHMAN, 2016).

A definição exata do tamanho dos microplásticos ainda é uma questão em debate. Na ciência, o prefixo micro é usado para indicar a milionésima parte de qualquer unidade de medida. No caso do comprimento, o micrômetro é igual a 10^{-6} m. Contudo, estritamente falando, o radical “micro” em grego significa apenas pequeno. Portanto, os microplásticos poderiam ter um tamanho da ordem de alguns milímetros ou mais. De fato, a Organização Internacional de Normalização (ISO) define microplásticos como qualquer partícula polimérica de origem sintética com tamanho inferior a 5 mm (ISO/TR 21960, 2020). Essa definição também foi a primeira dada por Thompson et al. (THOMPSON et al., 2004). A sua adoção está ligada a razões metodológicas relacionadas aos processos de amostragem e de identificação e foi amplamente utilizado antes mesmo da publicação da norma ISO.

Pelas mesmas razões, os nanoplásticos já foram definidos como partículas menores que 300 micrômetros. Abaixo disso, a amostragem seria improvável usando métodos de filtração de água que normalmente usam malhas de 330 micrômetros. Por outro lado, adotar o termo, conforme usado na engenharia, é preferível em alguns casos. Assim, os nanoplásticos seriam partículas poliméricas de origem sintética com tamanho entre 1 e 100 nm (em alternativa, também foi proposto o limite superior de 1000 nm). Por conveniência, é comum agrupar os nanoplásticos na categoria de microplásticos, exceto quando o tamanho e sua relação com as propriedades desses contaminantes é o foco do estudo (LEHNER et al. 2019; MATERIC et al., 2020). Por exemplo, partículas de 30 micrômetros são muitas vezes ditas microplásticos, apesar de terem tamanho inferior ao limite estabelecido. Isso porque os métodos de amostragem e detecção, nesses casos, é o mesmo.

Apesar de sofrerem degradação e fragmentação contínua até a escala nano, tem sido sugerido que os plásticos nunca se degradam completamente, permanecendo no oceano indefinidamente (WAYMAN; NIEMANN, 2021). Contudo, as taxas de acumulação esperadas, em resposta ao aumento da produção e descarte desses materiais, não foram observadas (LAW et al., 2010; THOMPSON et al., 2004). De fato, Law et al. não observaram nenhuma tendência de acumulação entre 1986 e 2008 nas regiões de maior acúmulo do Oceano Atlântico Norte e do Mar do Caribe. Isso sugere que há uma rápida remoção de plásticos da superfície e/ou há graves limitações na detecção de plásticos marinhos.

Se por um lado espera-se que os plásticos permaneçam indefinidamente nos oceanos, por outro, não somos capazes de perceber seu acúmulo. Para entender o tamanho das lacunas de conhecimento que impedem esse balanço de massa, é preciso olhar para os números. Os primeiros valores de uma emissão anual global de plástico para os oceanos foram fornecidos por Jambeck et al. (JAMBECK et al., 2015), que estimou uma emissão anual entre 4,8 e 12,7 milhões de toneladas. Esses valores foram obtidos por meio da análise dos dados sobre os volumes de produção e descarte de plástico e incorporam números sobre a má gestão dos resíduos sólidos urbanos em 192 países. Esse resultado tem sido amplamente aceito nos meios científicos, apesar das suas limitações. O estudo não leva em conta outras formas de emissões, decorrentes da erosão e do desgaste de materiais poliméricos ou de microplásticos primários provenientes de cosméticos,

por exemplo. Também considera apenas as emissões provenientes de atividades humanas que ocorrem até 50 km da costa, abrangendo cerca de 2 bilhões de pessoas (cerca de $\frac{1}{4}$ da população mundial).

Em contrapartida, outros estudos se propuseram a calcular a quantidade total de plástico presente nos oceanos. Essas estimativas contribuem para a compreensão do destino dos plásticos, diferente dos levantamentos das concentrações locais que têm maior valor nas avaliações de risco ambiental. Ainda assim, estão sujeitos a uma maior incerteza devido à dinâmica dos oceanos, às limitações metodológicas e à falta de conjuntos de dados abrangentes em ambientes marinhos. Eriksen et al., Cózar et al. e van Sebille et al., (CÓZAR et al., 2014; ERIKSEN et al., 2014; SEBILLE et al., 2015) realizaram extrapolações de dados sobre concentrações locais na superfície do oceano usando modelos oceanográficos de distribuição de partículas na superfície. Eriksen et al., por exemplo, estimou no máximo 250 mil toneladas de microplásticos flutuando na superfície do oceano, enquanto que Cózar et al. estimou apenas 35 mil toneladas. Apesar da diferença nas estimativas, ambos os valores estão bem abaixo do que seria esperado considerando o aporte anual proposto por Jambeck et al. (na casa de milhões de toneladas anuais). O mesmo vale para os valores intermediários encontrados por van Sebille et al. (de 93 a 235 mil toneladas) que alega ter utilizado o conjunto de dados mais abrangente até o momento.

2.1.2 Composição do Lixo Plástico Marinho

O Lixo Plástico Marinho é composto por uma ampla gama de plásticos, comumente classificados de acordo com seu tamanho, forma, cor e tipo de polímero. Sua abundância varia espacial e temporalmente dependendo dessas características. O tipo de polímero é uma característica chave, pois fornece informações sobre os produtos originais e hábitos de consumo aos quais esses plásticos estão relacionados, além de informações sobre seus impactos e tempos de residência.

No ambiente marinho, a distribuição dos tipos de polímeros reflete de perto a proporção daquilo que é produzido por tipo, mas, ainda mais, a proporção daquilo que é descartado. Os plásticos *commodities*, como o polietileno (PE), o polipropileno (PP) e o poliestireno (PS), representam mais de 50% do consumo de plásticos na Europa, por exemplo, e são abundantes em embalagens, sacos e artigos descartáveis (“Plastics - the Facts 2021”, 2021). Eles também são os tipos de resina

mais prevalentes na superfície do oceano (DE HAAN; SANCHEZ-VIDAL; CANALS, 2019; ERNI-CASSOLA et al., 2019). A presença no oceano de outros plásticos *commodities* com alta demanda pode ter algumas particularidades. O tereftalato de polietileno (PET) é encontrado principalmente em sedimentos já o cloreto de polivinila (PVC) é raramente encontrado. No caso do PVC, é provável que haja uma subestimação devido à má identificação, mas também se espera uma baixa taxa de emissão, uma vez que é consumido principalmente na indústria da construção, imobilizado em edifícios durante décadas (FERNÁNDEZ-GONZÁLEZ et al., 2022).

O MPL varia desde pedaços de alguns metros de comprimento até nanoplásticos, desde *pellets* de plástico até fragmentos de tinta, tecido ou sacos, incluindo garrafas, tampas, bitucas de cigarro e talheres descartáveis. A finalidade original do produto pode ser identificada através da observação direta das formas dos itens, mas não pode ser deduzida quando os plásticos estão demasiado fragmentados. Nesses casos, a melhor alternativa é caracterizar a composição química da partícula. A espectroscopia vibracional, seja através da espectroscopia Raman ou no Infravermelho, são técnicas acessíveis que fornecem boas medidas da identidade de uma partícula (JUNG et al., 2018). No entanto, os espectros geralmente diferem significativamente dos plásticos disponíveis no mercado em razão da degradação e absorção/dessorção de produtos químicos, incluindo aditivos. Isso dificulta a determinação do polímero principal naquela partícula, mas, principalmente, o seu uso original. Isso porque as composições dos produtos plásticos são adaptadas a utilizações específicas e, embora produtos diferentes possam partilhar o tipo de polímero, teriam aditivos diferentes para controlar as suas propriedades.

Isso nos leva a uma distinção importante a ser feita entre polímeros e plásticos. Os polímeros pertencem a uma classe de materiais variados que compartilham uma estrutura química semelhante. Todos formam longas cadeias a partir da repetição de uma unidade monomérica. Podem ocorrer naturalmente, como proteínas ou celulose, ou ser sintetizados artificialmente, como polietileno ou poliéster. O termo tem origem no grego antigo *polus*, que significa muitos, e *meros*, que significa partes. Plástico, por outro lado, é um termo comumente usado para descrever polímeros sintéticos. Tem origem no termo técnico termoplástico, referindo-se a materiais poliméricos que podem ser modelados e moldados (possuem plasticidade) sob a aplicação de calor. Além desse tipo, os polímeros

também podem ser classificados em elastômeros e termofixos (CALLISTER, 2019). Assim, o uso do termo plástico para se referir a todos os tipos de polímeros sintéticos pode ser ambíguo. No entanto, esse parece ser o termo mais comum na pesquisa ambiental e do lixo marinho, bem como na mídia em geral, por esse motivo foi também adotado aqui. Para evitar ambiguidade, o autor preferiu então utilizar o termo *polímero* para se referir às cadeias moleculares ou ao principal componente dos plásticos. Já *plástico* é utilizado para se referir ao material como um todo, que contém, além do polímero, outros aditivos. É utilizado como sinônimo de materiais poliméricos ou materiais plásticos.

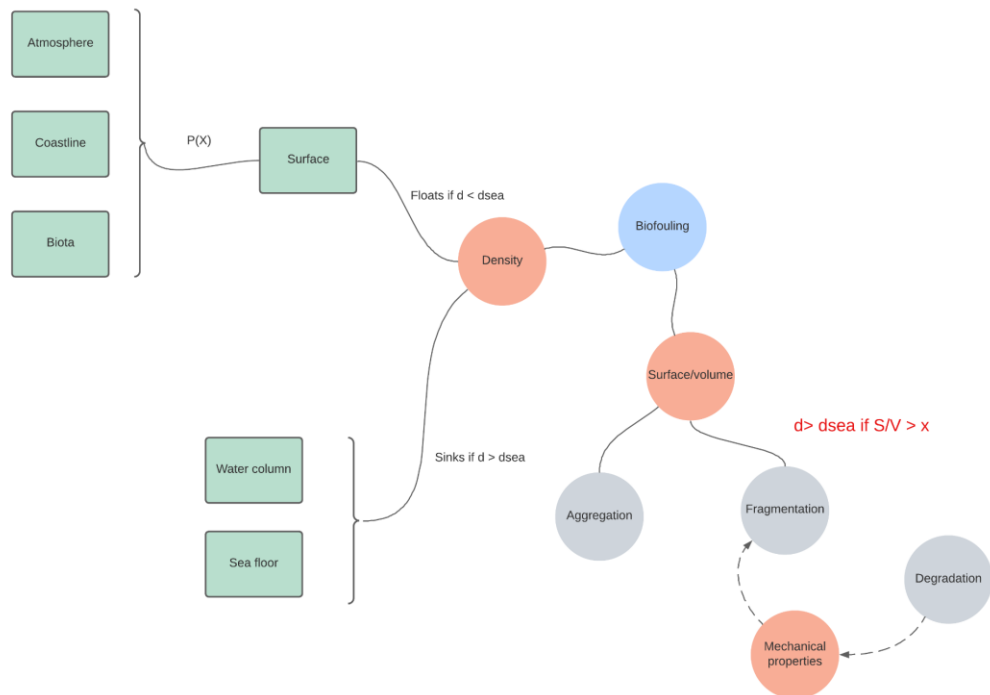
2.1.3 Distribuição do Lixo Plástico Marinho

A maior parte do plástico emitido para os oceanos flutua na superfície. Sua dispersão, portanto, ocorre através das correntes marítimas. Emitidos no litoral, passam a ser transportados por correntes de transporte locais e depois por correntes de maior importância regional (LEBRETON; EGGER; SLAT, 2019). Dado um tempo suficiente, esse material flutuante é transportado para as grandes zonas de convergência das correntes oceânicas: os giros subtropicais ou subpolares. Uma vez nos giros, há baixa probabilidade desses detritos escaparem, portanto, esses são locais de alto acúmulo (LEBRETON et al., 2018). Os giros são frequentemente tratados como “ilhas de lixo” pela mídia. Esse nome é impreciso, dando a ideia de que realmente existe uma ilha, quando, na verdade, está mais próxima de algo como uma “sopa”. Apesar disso, alguns estudos já apontaram que é necessário muito tempo para que esses detritos sejam de fato capturados pelos giros (LEBRETON; EGGER; SLAT, 2019). Nesse caso, permaneceriam próximos à costa, provavelmente em ciclos de encalhe e reflutuação.

Por outro lado, uma porção não negligenciável do lixo plástico marinho tem uma densidade superior a da água salgada. Esses detritos afundam rapidamente, acumulando-se no fundo do mar junto com muitos outros materiais que precipitam (KOELMANS et al., 2017). Tanto é que os plásticos já são utilizados como indicadores estratigráficos em análises de sedimentos (ZALASIEWICZ et al., 2016). Independentemente da sua densidade original, os plásticos têm a tendência de afundar com o tempo. Isso se deve à fixação e desenvolvimento de organismos na superfície das partículas plásticas, criando um biofilme que altera a densidade geral da partícula (KOOI et al., 2017). Esse efeito é especialmente importante para

partículas menores, que têm uma maior relação superfície/volume, de modo que o biofilme na superfície é relativamente maior. A Figura 3 mostra um modelo conceitual de distribuição 3D do plástico no oceano por mudanças na densidade por meio da fragmentação e pela probabilidade $P(x)$ de que ele tenha sido emitido para a atmosfera, consumido pela biota ou esteja encalhado na costa. No entanto, o biofilme pode desagregar-se em profundidades mais elevadas, por não estar adaptado às condições destas profundidades, particularmente à falta de luz solar, e as partículas podem voltar à superfície.

Figura 3: Modelo conceitual de distribuição 3D de plásticos no oceano dado por mudanças na densidade por meio da fragmentação e pela probabilidade $P(x)$ de que ele tenha sido pulverizado na atmosfera, consumido pela biota ou encalhado na costa. d é a densidade da partícula e d_{sea} é a densidade da água do mar.



Fonte: preparado pelo autor

2.1.4 Relevância da pesquisa em Lixo Plástico Marinho

Os plásticos são considerados contaminantes ambientais emergentes (SAUVÉ; DESROSIERS, 2014), embora a sua presença no ambiente tenha sido relatada já na década de 1960, foi após um artigo seminal de Thompson et al. (THOMPSON et al., 2004) em 2004 que o assunto começou a atrair ampla atenção e o campo de estudo se expandiu. Apesar de estudos anteriores terem se

concentrado principalmente no macrolixo, em 2004, Thompson et al. propuseram que esses detritos maiores se fragmentassem em pequenas partículas, que chamaram de microplásticos. Posteriormente, a ocorrência desses contaminantes foi amplamente relatada em todo o mundo e em matrizes ambientais, o que levantou preocupações sobre seus impactos em diferentes ecossistemas, bem como nos seres humanos.

O MPL pode impactar organismos e ecossistemas de muitas maneiras diferentes e a probabilidade e a força dos seus impactos estão relacionadas com as características das partículas, particularmente a composição química e a morfologia. O macrolixo, com tamanho superior a alguns centímetros, tende a causar maiores impactos devido à ingestão e emaranhamento em animais marinhos. As redes de pesca abandonadas, perdidas ou descartadas são notórias por causarem a morte de focas, tartarugas, golfinhos e outros animais marinhos (GILMAN et al., 2021). Outros detritos, como sacos, tampas de garrafas, recipientes de alimentos, etc. podem ser confundidos com alimentos e causar obstrução do trato digestivo, levando à fome, fato frequentemente observado em aves marinhas. Porém, mesmo que o animal não morra, a dificuldade na alimentação causa prejuízos, reduzindo sua taxa de crescimento e fertilidade (GALL & THOMPSON, 2015).

O microlixo, ao contrário, pode gerar impactos mais fortemente associados à toxicidade. A lixiviação de substâncias presentes nos microplásticos é motivo de preocupação e já foi observada em alguns animais, como os moluscos filtradores (CAPOLUPO et al., 2020). Embora os polímeros em si sejam considerados seguros, sabe-se que os microplásticos estão associados a uma série de contaminantes, alguns deles com efeitos mais conhecidos. Isso se deve à mistura intencional de algumas substâncias no processo de produção de materiais poliméricos, conhecidas como aditivos, e à presença de monômeros residuais do processo de polimerização no produto final. Além disso, os microplásticos concentram contaminantes que ocorrem na água, servindo como vetores dessas substâncias na cadeia trófica. Essa absorção acontece devido à característica hidrofóbica de ambos e à estrutura porosa dos fragmentos degradados de polímeros ambientais (COSTIGAN et al., 2022).

A concentração de plásticos num determinado ecossistema é um parâmetro fundamental para avaliar os seus impactos ambientais. Entretanto, na maioria das situações, a contaminação ocorre em razão de uma matriz complexa de

contaminantes, incluindo muitos tipos de plásticos e outras substâncias. Isso representa um desafio para a investigação dos efeitos ecotoxicológicos, que serão diferentes para cada interação material-organismo. Para ajudar a determinar os danos causados aos organismos e ecossistemas pelos plásticos, eles são classificados por tipo de polímero. Como já vimos, essa classificação fornece pistas sobre a sua origem e destino, mas também sobre os seus impactos no ambiente. Apesar disso, a complexidade na avaliação dos impactos aumenta devido à necessidade de considerar outros fatores de influência ecotoxicológica como: tamanho das partículas, estágio de degradação e composição do biofilme. Por exemplo, tamanhos menores de partículas, especialmente quando a relação entre área superficial e volume é maior (por exemplo, nanoplásticos ou fibras), aumentam a disponibilidade para consumo na cadeia trófica e a reatividade com organismos (AVIO; GORBI; REGOLI, 2017).

A exposição humana a esses contaminantes também já foi demonstrada. Inicialmente indiretamente através da detecção no ar e nos alimentos. Mais recentemente, foi confirmada sua presença no estômago, nos pulmões e até na corrente sanguínea (LESLIE et al., 2020; WRIGHT; KELLY, 2017). Os plásticos podem conter uma série de substâncias com potencial de ruptura metabólica (estrogênios, antiandrogênicos, adipogênicos, etc.) (“Plastics, EDCs & Health”, 2020). Os microplásticos ambientais também estão relacionados a substâncias não misturadas intencionalmente com eles na produção, pois são capazes de absorver contaminantes do meio ambiente (HARTMANN et al., 2017). Porém, ainda existem incertezas sobre os reais riscos relacionados à exposição aos microplásticos, principalmente no que diz respeito às concentrações dessas substâncias. Estudos laboratoriais sobre efeitos ecotoxicológicos em animais são realizados em concentrações de microplásticos superiores às encontradas no meio ambiente (UNDERWOOD; CHAPMAN; BROWNE, 2017). No entanto, há uma acumulação esperada nas próximas décadas que poderá apoiar este tipo de evidência.

O MPL é uma questão ampla e complexa que requer esforços científicos de uma gama de disciplinas. Pode-se argumentar que o campo está centrado na Ciência Ambiental – que por si só já é bastante extensa. Ainda assim, contribuições das áreas de Farmacologia e Toxicologia, Ciência da Computação, Ciências Sociais, Economia, Química e Bioquímica são importantes e frequentes. A Ciência dos Materiais, em particular, tem imenso conhecimento acumulado sobre engenharia e

comportamento de plásticos, por isso, pode ser um parceiro importante na compreensão do destino e dos impactos do MPL. Embora as contribuições dessa área sejam cada vez mais comuns, os números totais ainda são comparáveis aos da Física e da Astronomia, por exemplo. Esta tese pretende encontrar as intersecções entre MPL e a Ciência dos Materiais, particularmente no que diz respeito ao destino dos plásticos no ambiente, abrindo novos caminhos para colaborações.

2.2 APRENDIZADO DE MÁQUINA

O aprendizado de máquina (ML, do inglês *Machine Learning*) é considerado um subcampo da Inteligência Artificial (IA), que é um grande campo dedicado ao desenvolvimento de máquinas inteligentes que possam ser capazes de raciocínio e interações semelhantes aos humanos. O ML é indiscutivelmente a área que mais impulsiona a IA e, muitas vezes, é tratado como sinônimo dela. Numa analogia à aprendizagem humana pela experiência, os computadores são capazes de fazer distinções e previsões observando eventos passados. Por exemplo, pode ser ensinado a distinguir maçãs, laranjas e melões se forem apresentadas imagens dessas frutas. Embora o termo esteja na moda nos noticiários e em muitos campos científicos, o campo de Aprendizado de Máquina existe pelo menos desde a década de 1980 (LANGLEY, 2011). No entanto, as altas demandas computacionais desse tipo de algoritmo impediam que fossem implementados até para funções simples, pois sua performance era superada por outros métodos estatísticos mais convencionais.

Algoritmos de aprendizado de máquina são capazes de identificar padrões e relações em um conjunto de dados e então usar esse aprendizado para fazer previsões sobre dados novos. Nesse sentido, não diferem muito das análises estatísticas mais convencionais. Na verdade, há um debate contínuo sobre se são a mesma coisa. Em consonância com o Prof. Michale I. Jordan e outras figuras eminentes na área, esta tese entende que o aprendizado de máquina e a estatística podem ser colocados sob a égide do campo emergente da Ciência de Dados, que envolve o conhecimento e as ferramentas que usamos para entender os dados (JORDAN, 2019). No entanto, o aprendizado de máquina está mais interessado nas previsões ao invés de descrever e interpretar observações, como a estatística. Os métodos de aprendizado de máquina geralmente são mais adequados para conjuntos de dados maiores e mais complexos, que podem conter centenas de

variáveis. Isso também limita a nossa capacidade de explicar esses modelos e utilizá-los de modo eficaz para fazer inferências sobre os mecanismos por trás das relações observadas.

Ainda assim, há um componente importante no ML que geralmente está ausente em outros tipos de análise estatística, que são as divisões de treino e teste. Na estatística, os modelos são usados para descrever as observações, o que pode ser feito ajustando uma curva aos dados. Como os modelos de ML são utilizados para fazer previsões, é importante que sejam generalizáveis, ou seja, descrevam com precisão dados não vistos, não apenas a amostra analisada. Para evitar sobreajuste (e até sub ajuste) da curva ou modelo, as abordagens de aprendizado de máquina usam um subconjunto do conjunto de dados para criar um modelo estatístico (treino) que é então testado em relação ao restante do conjunto de dados que não foi selecionado (teste). Isso pode ser feito várias vezes, com subconjuntos diferentes a cada vez, para que o modelo e seu desempenho sejam equilibrados pela variabilidade no conjunto de dados. Essa técnica é chamada de validação cruzada e é predominante no aprendizado de máquina.

Hoje, o aprendizado de máquina é prevalente em áreas como visão computacional e processamento de linguagem natural, incluindo geração de imagens e texto, que tem ganhado manchetes nos últimos dois anos. Mas, além dessas aplicações chamativas, o aprendizado de máquina está agora difundido em todas as áreas onde há dados disponíveis. Por exemplo, é usado no transporte para analisar padrões de tráfego históricos e em tempo real, como no Google Maps (LAU, 2020) ou para guiar um aspirador robô dentro das casas (MOON et al., 2022). É utilizado na área da saúde, para encontrar genes associados a determinadas doenças ou compostos ativos eficazes no tratamento dessas doenças (GAN et al., 2023), ou mesmo para auxiliar no diagnóstico através de imagens ou amostras de sangue (SUTTON et al., 2020). A previsão do tempo também fornece um ótimo exemplo de aplicações de ML de ponta. Embora no século passado as previsões meteorológicas tenham evoluído por meio de uma melhor compreensão da física que rege o clima e os eventos atmosféricos, hoje os modelos de ML podem superar esses modelos tradicionais, aprendendo com os padrões climáticos do passado. Eles podem fornecer previsões mais precisas em uma fração do tempo (LAM et al., 2023).

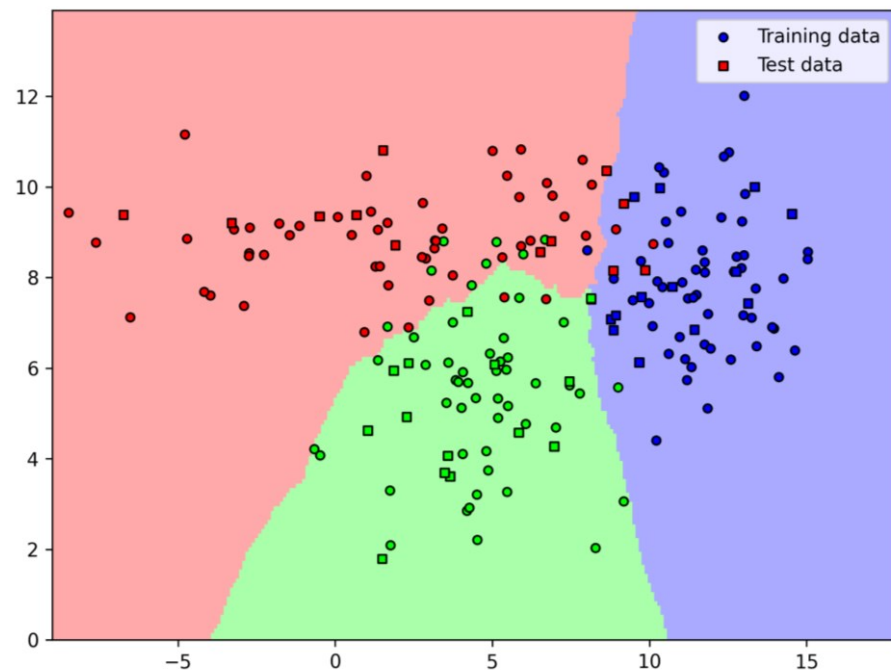
2.2.1 Classificação

O aprendizado de máquina pode ser subdividido em 3 tipos principais de abordagens que dizem respeito à forma como o aprendizado é realizado: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço. As tarefas de classificação enquadram-se na primeira categoria, pois requerem informações sobre a variável alvo para aprendizagem (GÉRON, 2019). No exemplo anterior de maçãs, laranjas e melões, esses são os tipos de frutas. Para melhor ilustrar, podemos dizer que o conjunto de dados contém informações sobre peso, cor, espessura da casca, esfericidade e tipo de fruto. Um algoritmo de aprendizado de máquina é fornecido, digamos, para 80% das observações e é treinado nessas características das amostras de frutas. Em seguida, tenta prever o tipo de fruta dos 20% restantes das amostras e comparar as previsões com o valor real para avaliar o seu desempenho.

A Figura 4 ilustra um modelo de ML treinado em dados gerados aleatoriamente. As cores de fundo na figura definem o limite de decisão calculado para prever dados futuros. Pode-se observar que alguns pontos estão fora do limite da classe correta, mesmo para dados de treinamento. Existe uma clara sobreposição nos agrupamentos de pontos de diferentes classes, o que cria essas regiões dúbias, onde a probabilidade de as amostras pertencerem a cada classe é menor (mais incerteza). Um modelo de ML como este poderia fornecer valores de probabilidade para cada atribuição de amostra a fim de evitar que amostras com alto grau de incerteza na atribuição de classe sejam classificadas erroneamente, quando falsos positivos são importantes, por exemplo.

Algoritmos de classificação são utilizados quando o objetivo é atribuir uma classe conhecida a cada amostra desconhecida de um novo conjunto de dados, com base em exemplos anteriores dessas classes. Alguns algoritmos comumente usados são: Regressão Logística, K-vizinhos mais próximos, Árvores de Decisão, Florestas Aleatórias e Máquinas de Vetores de Suporte.

Figura 4: Exemplo de modelo de classificação K-nn treinado em dados gerados aleatoriamente para 3 clusters. As cores de fundo indicam o limite de decisão do modelo para previsões futuras.



Fonte: elaborado pelo autor

De maneira semelhante, quando a variável alvo é contínua, em vez de categórica, tem-se uma análise de regressão. Nesse caso, poderíamos prever, por exemplo, o preço de casas com base no seu tamanho e endereço.

2.2.2 Análise de Aglomerados

A análise de aglomerados é análoga à classificação. No entanto, é uma técnica de aprendizagem não supervisionada, pois não possui uma classe alvo definida (não explicitada pelo conjunto de dados original). O objetivo dessa análise é agrupar observações semelhantes, em vez de atribuir-lhes uma classe (GÉRON, 2019). Assim, no nosso exemplo de frutas, o agrupamento poderia ser feito usando um conjunto de dados de imagens de frutas que não estavam rotuladas. A vantagem desse método é que ele não requer um conjunto de dados de treinamento, dispensando a trabalhosa tarefa de atribuir manualmente classes a um conjunto de imagens a serem utilizadas para treinamento. Por outro lado, não teremos base factual para comparar nossos modelos. Em outras palavras, não podemos testar a precisão dos nossos modelos na distinção entre tipos de frutas.

Os modelos de clustering, como são chamados em inglês, podem ser benéficos por não precisarem enquadrar os dados em categorias predefinidas. Por exemplo, poderíamos identificar, no nosso conjunto de dados, variedades de maçã ou mesmo maçãs com defeitos, se houvesse diferenças significativas e consistentes entre os grupos de maçãs. Esse não seria o caso se os algoritmos se limitassem a distinguir maçãs, laranjas e melões. Isso foi particularmente útil nesta tese para revelar padrões na literatura, como veremos nos capítulos 3 e 4.

Existe ainda um terceiro tipo de aprendizado de máquina: o aprendizado por reforço, onde um agente aprende como tomar decisões interagindo com seu ambiente. O agente aprende a atingir uma meta em um ambiente incerto e potencialmente complexo, realizando ações e recebendo recompensas ou penalidades em troca. Essa abordagem é, por vezes, considerada um tipo de aprendizagem não supervisionada, porque não requer dados de entrada. Ainda assim, acumula dados por tentativa e erro. Essa abordagem não foi utilizada na tese, portanto, não será discutida em mais detalhes.

2.2.3 Processamento de Linguagem Natural

O processamento de linguagem natural (NLP, do inglês *Natural Language Processing*) é um subcampo da ciência de dados que se concentra na interação entre computadores e humanos usando linguagem natural. Particularmente, visa tornar a linguagem natural (texto e fala) “compreensível” para os computadores. Isso pode ser feito de diferentes maneiras. Por exemplo, algoritmos podem ser instruídos explicitamente sobre os padrões de sentenças. No entanto, a linguagem natural tem uma estrutura mais orgânica e menos fixa que desafia essa abordagem. O aprendizado de máquina, por outro lado, pode ser usado para reconhecer padrões na linguagem, às vezes sem qualquer conhecimento prévio de como a linguagem está estruturada (VASWANI et al., 2023). Na verdade, o ML contribuiu para o rápido avanço do campo nos últimos anos, em especial com os Grandes Modelos de Linguagem (LLM, da sigla em inglês *Large Language Models*). Esses avanços resultaram também no surgimento de habilidades não previstas para esses modelos, como a resolução de problemas matemáticos e, até certo ponto, o raciocínio e a compreensão da física do “mundo real” (WEI et al., 2022). Além disso, os LLMs também podem derivar “significado” do contexto, ou seja, são capazes de distinguir

homônimos. Por exemplo, os LLMs podem dizer se a palavra manga se refere à parte da camisa ou à fruta dependendo do contexto em que é utilizada.

O desenvolvimento de modelos que reconhecem padrões em linguagem natural foi a base de modelos de linguagem generativa, que podem produzir textos, imagens ou vídeos a partir de entradas de texto. Existem hoje muitos aplicativos disponíveis para o público em geral, como ChatGPT, Google Bard, Midjourney, Dall-e e SORA. Esses LLMs proporcionaram grandes melhorias em relação aos chatbots anteriores (VICTOR, 2016) e mudaram a percepção geral negativa que as pessoas tinham em relação à “IA conversacional”. O ChatGPT, por exemplo, obteve impressionantes 15,5 milhões de visitas na primeira semana após o lançamento, mas o número agora é próximo de 2 bilhões por mês. Na ciência, o ChatGPT, assim como outras IAs generativas, já estão sendo usadas por pesquisadores de diferentes disciplinas para ajudá-los a escrever códigos, manuscritos e debater ideias, entre outras coisas (VAN NOORDEN; PERKEL, 2023). Embora essas ferramentas possam beneficiar profissionais de todos os setores, ainda é preciso aumentar a consciencialização sobre as suas potenciais implicações na sociedade, incluindo questões relacionadas a preconceitos, desinformação, privacidade e autoria (TYLER et al., 2023; “Why Nature will not allow the use of generative AI in images and video”, 2023).

Além da IA conversacional, o NLP possui outras aplicações que podem ser úteis, principalmente para pesquisadores. Como muitas outras atividades humanas, a produção científica explodiu durante a Grande Aceleração (BORNMANN; HAUNSCHILD; MUTZ, 2021). Embora isso se deva em parte às ramificações e ao surgimento de novos tópicos de pesquisa, a literatura especializada também cresceu. Repositórios web, como Scopus e OpenAlex, transformaram enormemente a forma como as publicações científicas são acessadas pelo público, facilitando em grande parte sua distribuição e acesso. Essa larga disponibilidade tem um custo: analisar *corpora* cada vez maiores e de rápido crescimento pode ser tedioso e caro, o que por vezes torna-se inviável. Os pesquisadores geralmente contornam essa limitação reduzindo o escopo de suas análises. Em consequência, estão confinados a campos cada vez mais especializados. No entanto, os problemas sociais contemporâneos, como o MPL, requerem uma visão mais holística e uma integração interdisciplinar, razão pela qual essa abordagem pode ser insuficiente. Alternativamente, o NLP pode ser usado para extrair automaticamente insights

valiosos a partir de texto. Isso acelera o processo, tornando possível analisar milhões de frases de uma só vez, mesmo em tempo real, se necessário. Ferramentas que possam extrair dados e informações dessas fontes de forma eficiente podem ser úteis para o avanço dos campos científicos e para orientar decisões políticas com a ciência mais atualizada e relevante (TYLER et al., 2023).

3 MAPPING THE LANDSCAPE OF MARINE PLASTIC LITTER RESEARCH WITH LARGE LANGUAGE MODELS

Submetido em Abril de 2024 em Nature Sustainability.

Henrique de Medeiros Back^{1,2}, Michael Grelaud¹, Orestes Estevam Alarcon², Patrizia Ziveri^{1,3}

¹ Institute of Environmental Science and Technology (ICTA-UAB), Universitat Autònoma de Barcelona (UAB), Barcelona, Spain

² Federal University of Santa Catarina (UFSC), Florianópolis, Brazil

³ Catalan Institution for Research and Advanced Studies, ICREA, Barcelona, Spain

ABSTRACT

Plastics are among the most studied pollutants in environmental science, particularly during the last decade. The literature on the issue has grown and diversified rapidly, which contributes to the accumulated knowledge, but poses difficulties for researchers, policymakers and funding agencies in synthesising the relevant information. Here we utilise state-of-the-art Machine Learning tools to swiftly analyse topic trends on marine plastic litter. We show the recent surge in publications was driven by an increasing focus on impact studies and a shift to the microscale, accompanied by funding and policy strategies. We also found large discrepancies in the number of studies between oceans, with the Global South greatly overlooked. Increasing citation inequalities on all topics indicate difficulties in knowledge dissemination that progressively hinder efforts to understand marine plastic litter and reduce funding outcomes. Finally, we advocate for the use of Artificial Intelligence to mitigate the effects of information overload in science.

3.1 INTRODUCTION

Marine Plastic Litter (MPL) covers all plastic waste that enters the coastal and marine environment, either discarded or abandoned by sea-going vessels or more commonly originating from land-based sources as a result of poor waste management processes¹. Evidence of widespread contamination and subsequent ecological, social, and economic impacts² have brought MPL to the attention of scholars, environmental regulation agencies^{3,4}, policymakers⁵ and the general population⁶, seeking to better understand the problem, its implications and potential solutions⁷. As a result, the scientific literature on MPL has been rapidly expanding, amassing evidence and revealing previously unknown threats. Sources of plastic litter are incredibly diverse⁸, emerging from activities such as transportation⁸, clothing⁹, agriculture¹⁰, packaging¹¹ and health¹². Likewise, its impacts range from the subatomic to the ecological and global level and have been linked to oxidative stress, inflammation, obstruction of the digestive system and the spread of pathogens and invasive species, just to name a few^{13–16}. Enough evidence now exists to target plastic production models, consumption patterns and waste disposal practices^{17,18}. However, effective dissemination of critical knowledge is needed for strategic implementations grounded on the latest and most comprehensive scientific evidence^{19–21}.

Systematic reviews collectively evaluate multiple findings to produce syntheses useful for interested parties. However, researchers often have to narrow the focus of their reviews due to the difficulties in assessing many hundreds of publications in large and diverse fields, such as MPL. While these reviews provide a close-up into particular aspects of a field and may still combine meta-analyses, and expert opinions, they are limited in their ability to provide a broader perspective on the field as a whole. Comprehensive assessments of MPL, such as those produced by GESAMP²² and UNEP^{7,23–26}, may provide some perspective but are still restricted in how many papers they can account for, and rely on the subjective experience of experts. The constant and rapid growth of the literature can exacerbate these constraints, with late recognition and response to shifting trends in the community's priorities and emerging new topics of research. Furthermore, this surge in new publications in other fields has been shown to slow down scientific progress²⁷ as overloaded scholars concentrate their attention on the already well-established literature, which encourages authors to align their research within the same

frameworks, reducing the probability of the emergence of novel ideas and adding disproportionately to the citation of canon publications. The overflow of newly available knowledge might lead to slower progress in the field as a significant portion of scientific and funding efforts result in knowledge that is not transferred to strategic audiences.

The challenges of accumulating publications affect researchers from all scientific fields^{27,28}, as well as policymakers who rely on scientific knowledge to make informed decisions²⁹. There is a growing need for approaches to automate information extraction (IE) from texts that can be used to disseminate instrumental knowledge for research, policy and sustainability purposes^{20,30}. Despite the recognition that Artificial Intelligence (AI) should be employed, reports are still unclear on how AI can effectively be used. The medical sciences provide examples of evidence-based decision-making in applications such as Clinical Decision Support Systems^{31,32} and Drug Discovery and Development^{33,34}. Recent advances in Natural Language Processing (NLP) driven by progress in Machine Learning (ML)^{35,36}, and computational linguistics³⁷ more broadly, have enabled the development of new tools that can significantly aid in the extraction and synthesis of information from textual databases. They are capable of recognising patterns in language without being explicitly programmed with the rules of grammar and syntax, and they can incorporate semantic relations between words, such as those captured by text embeddings - also known as distributed representations^{38,39}. These tools can be used effectively for a variety of tasks, such as identifying and comparing the content of texts (topic modelling) and recognizing entities, such as names of places, people, or organisations, from texts (named entity recognition). They provide experts and non-experts with ways of synthesising the scientific literature and visualising trends based on the relevant information that is dispersed throughout texts. This helps us to organise and classify knowledge to extract actionable insights, which we can then incorporate into new research and policymaking.

In light of the rapid expansion of the scientific literature on MPL and the challenges imposed by its large and diverse corpus, this study employs cutting-edge ML and NLP tools to offer a comprehensive perspective on past and current trends in MPL research. The proposed tools can be instrumental when combined with detailed systematic reviews on specific topics based on a more limited number of articles. Here, we categorised over 9,000 peer-reviewed publications into 38 research topics

based on the hierarchical, density-based clustering (HDBSCAN) of abstracts' embeddings. We visualise topic trends over the past 15 years and observe an increasing diversification of topics, as well as a general shift from studying macro to microorganisms and from ingestion and entanglement to toxicity. This reflects a growing concern about the environmental threats to our oceans imposed by the widespread contamination of micro and nanoplastics, particularly their consequences on the building blocks of marine ecosystems²² and biogeochemical cycles⁴⁰. Additionally, this shift was favoured by the development of analytical capabilities that allowed the detection of smaller debris in natural environments and organisms.

In parallel, Named-Entity Recognition was used to classify publications into ocean basins by automatically identifying mentions of locations in texts. This analysis revealed a large imbalance in areas surveyed for plastics across the world and that it correlates with the gross expenditure in research and development. Moreover, the less studied regions compound a less diverse literature in terms of topics, presenting potential inequalities in mitigation and adaptation strategies.

Finally, we analyse the development of the field through citation behaviours. We observe that a few papers published in recent years have accumulated citations rapidly, which, along with the diversification of topics, indicates the field is rapidly evolving. However, our results also show high citation inequalities on every topic, and that these increase with the rate of publications. This means a large proportion of the new research is not being incorporated, either because it does not contain novel or relevant ideas or simply because readers are cognitively overloaded. In any case, the trend is likely to intensify in the future with more and more publications in the MPL field, posing challenges to research communication, priority setting and resource allocation. We highlight the importance of mapping the landscape of the research field and the integral part Machine Learning can have in providing researchers, funding agencies and policymakers with resources and tools to address marine pollution and other key societal challenges.

3.2 A TOPIC MODEL OF THE LITERATURE

A topic model was created based on texts, for all abstracts of the MPL literature retrieved from SCOPUS. We used BERTopic⁴¹, a software that leverages BERT embeddings and unsupervised machine learning to cluster similar abstracts based on the content of their text. This enables the identification of research topics

and the assignment of publications into these detected categories. Topics are understood here as dense areas in the representation space of abstracts, that is, areas with similar embeddings. Since topics are direct outcomes of the density-based clustering, they may have been determined with more or less broad content. This has an effect on topic sizes, which can vary widely. Nonetheless, the definition of the model's parameter was made to favour the recognition of a larger scale structure and avoid too many small clusters, which would hinder interpretation.

To enable the interpretation of topics, BERTopic calculates representative words and abstracts by comparing word frequencies inside and outside each cluster and identifying the documents which are closer to the cluster's centroid, respectively. A label is then provided using a Text Generation tool (see methods for details). A more in-depth look at topics can be found in SI 1, which contains a brief description of each topic. Table 1 shows the identified topics with their respective labels. Topic -1 (representing 16.84% of the total articles) contains all abstracts that were not densely clustered with any topic and didn't form clusters themselves. These unclustered documents lack clear patterns and may indicate open questions where more research is needed. Topic 30 is clearly out of the scope of this analysis and will not be further discussed. Other topics (2, 12, 13, 16, 25, 32) focusing on terrestrial and freshwater systems were kept due to potential overlap with marine debris, especially regarding transition zones and emitting sources.

The Unsupervised Learning method applied here produces topics of publications that contrast with traditional classifications often prescribed within academic disciplines. Alternative classifications serve different purposes. For instance, to account for the complexity and interdisciplinarity in MPL, it is frequently organized into three main categories of "sources, fate and impacts"²². Our methods, on the other hand, allow a more flexible approach to categorisation as topics are found from the corpus and not pre-fixed. This is particularly useful for large and diverse corpora, such as the MPL literature, allowing the production of a classification model that fits well with the available data, and not necessarily with pre-existing frameworks. However, this poses other challenges as there can be many models of classification. Particularly, variations in the dataset or in the definition of model parameters can produce different outcomes. Since there is no "ground truth" to evaluate the performance of each possible model (documents are not previously associated with categories), the aim is to find a model that can be useful for

classification. The topic model used here was chosen among a list of models with high confidence in the assignment of documents into topics and was preferred over others because it contained a fair number of topics*. This is crucial because models with too many classes can be difficult to interpret and models with too few classes add little understanding about the literature. Importantly, the model can be audited and reproduced using the publicly available code (see Methods for further details).

Table 1: List of topics and their corresponding size (paper count and percentage relative to the total number of publications considered by this study)

Topic	Topics	Count	% of the literature
-1	Outliers	1542	16,84
0	Plastic Pollution and Marine Debris	1230	13,43
1	Microbial Communities on Marine Plastic Debris	811	8,86
2	Microplastic Pollution in Freshwater Environments	797	8,70
3	Adverse Effects of Microplastic Exposure on Marine Organisms	758	8,28
4	Microplastic Contamination in Commercial Fish Species	579	6,32
5	Plastic Ingestion by Marine Species	363	3,96
6	Floating Plastic Debris and Ocean Transport	329	3,59
7	Oceanic Microplastic Pollution	273	2,98
8	Microplastics and Human Health Impacts	202	2,21
9	Plastic Debris on Coastal Beaches	186	2,03
10	Detection of Marine Debris Using Spectral Images	172	1,88
11	Fishing Gear as a Source of Marine Plastic Pollution	148	1,62
12	Microplastic Pollution in Soil and Its Environmental Impacts	145	1,58
13	Microplastic Research in Aquatic and Terrestrial Environments	134	1,46
14	Transport and Settling of Microplastic Particles in Aquatic Environments	107	1,17
15	Impacts of Marine Debris on Sea Turtles and Seals	103	1,12
16	Microplastics Removal in Wastewater Treatment Plants	102	1,11
17	Adsorption of Microplastics for Contaminant Removal	95	1,04
18	Environmental Impacts and Research on Microplastics	93	1,02
19	Environmental Impact of Face Masks and PPE in the COVID-19 Pandemic	82	0,90
20	Spatial and Temporal Variability of Microplastic Contamination in Coastal Sediments	82	0,90
21	Microplastic Pollution in Coral Reef Ecosystems	82	0,90
22	Transoceanic Rafting of Species on Marine Debris	73	0,80
23	Biological Effects of Pollution and Human Disturbances	70	0,76

	on Marine Ecosystems		
24	Tire Wear Particles (TWP) as a Source of Microplastic Pollution from Road Dust	63	0,69
25	Distribution and Impact of Microplastics in Freshwater River Systems	57	0,62
26	Marine Pollution by Plastic Debris and its Chemical Contaminants	53	0,58
27	Microplastic Abundance in Seabed Sediments	49	0,54
28	Microplastic Ingestion in Fish Species	41	0,45
29	Microplastic Pollution in Marine Environments	41	0,45
30	Dual Parametric Optical Fiber Sensor for Temperature and Pressure Measurement in Seawater	37	0,40
31	Marine Litter Pollution and Plastic Accumulation in the Ocean	35	0,38
32	Microplastic Pollution in Freshwater Environments	35	0,38
33	Ocean Health and Pollution	35	0,38
34	Impact of Mangrove Forests on Sediment Microplastic Abundance	32	0,35
35	Adsorption and Speciation of Heavy Metals on EPS Microplastics	32	0,35
36	Phthalate Contamination in Marine Organisms and Human Health Risks	32	0,35
37	Plastic Litter Pollution in Coastal Areas of the Philippines	28	0,31
38	Microplastic Contamination in Sea Salt Products	28	0,31

Source: prepared by the author

3.3 TOPIC TRENDS

To investigate the prevalence of certain topics in the literature, and how they change over time, we visualise the number of publications per year in each topic. Figure 1 shows the increasing frequency of publications in the literature as a whole (Fig 1a) and for specific topics (Fig 1b) for the past 15 years. While the literature addressing MPL extends back to the 1970s (data preceding 2008 were excluded from Figure 1 for enhanced visualization), publications were sparse. Early publications were often classified into the topics: 0: Plastic Pollution and marine Debris, 5: Plastic Ingestion by Marine Species, 6: Floating Plastic Debris and Ocean Transport, 9: Plastic Debris on Coastal Beaches, 11: Fishing Gear as a Source of Marine Plastic Pollution, 15: Impacts of Marine Debris on Sea Turtles and Seals. A closer look at their representative words through the years (refer to the interactive figure `topics_over_time` available in the SI), also shows that early research often

mentioned birds and hatchlings for topic 5 and entanglement for topic 15. This observation reveals the early interests of researchers investigating MPL, perhaps indicative of the most evident aspects of the problem at the time.

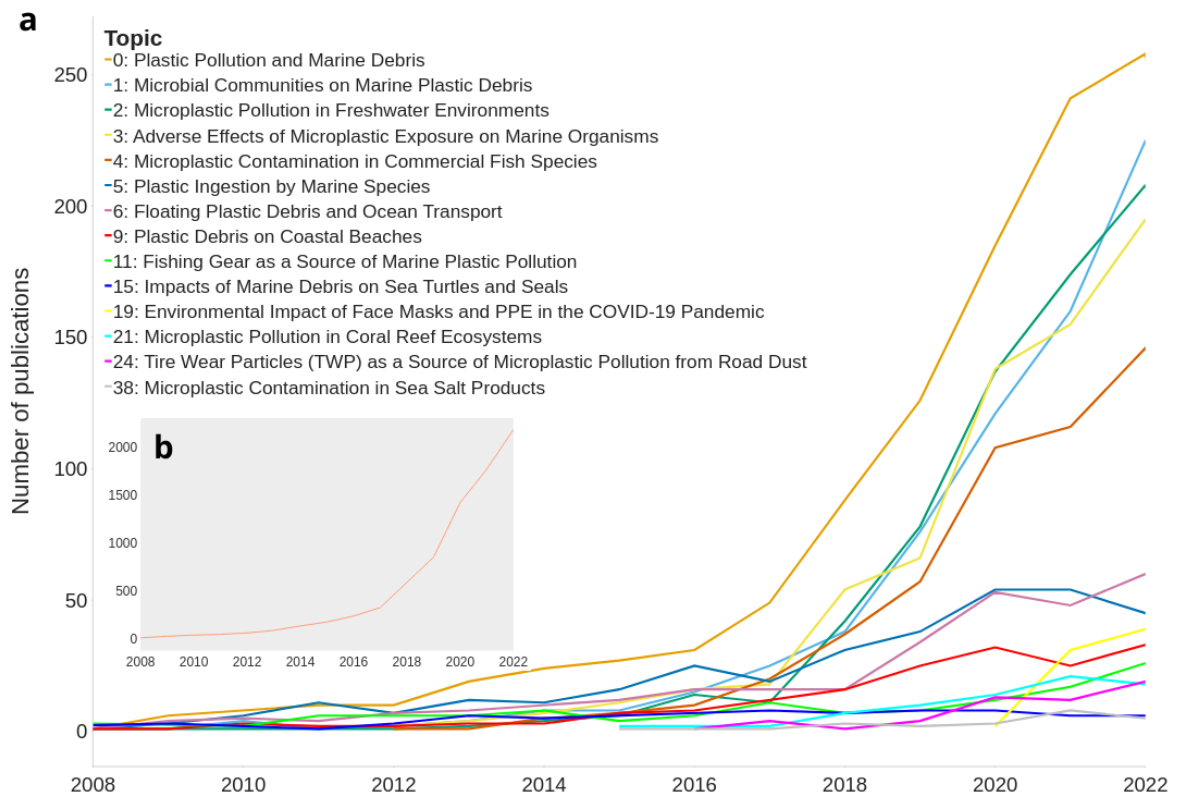
In the early 2010s, these early topics continue to dominate, though we begin to observe an increasing diversification of topics with consistent occurrence. However, it is not until 2017 that the field gains substantial momentum, with a sharp increase in the number of publications. The general trend (Fig 1a) is driven by the rapid growth of the top 5 topics, which were virtually absent from the literature before 2013: 1: Microbial Communities on Marine Plastic Debris, 2: Microplastic Pollution in Freshwater Environments, 3: Adverse Effects of Microplastic Exposure on Marine Organisms and 4: Microplastic Contamination in Commercial Fish Species. Based solely on these topic labels, we can swiftly infer that there was a shift in interest towards microplastics, toxicity and microbial communities from those of earlier topics which focused on debris, entanglement and marine animals, such as birds, seals and turtles. Moreover, from the interactive figure (topics_over_time) in the SI, we see topic 1 presented degradation as a representative word in 2015 and 2018, just before the explosion in publications on that topic. This indicates some interrelation between studies on degradation and microbial assemblages and the potential for concomitant investigations on both topics.

This recent period of intense scientific production and the diversification of topics in the field of MPL follows the increasing consideration of plastic pollution at high-level international discussions. It was raised as an emerging issue in the UNEP YearBook of 2011, the Global Partnership on Marine Litter was launched in Rio de Janeiro in 2012, and the United Nations Environment Assembly (UNEA) adopted a series of resolutions and decisions in 2014, 2016 and 2017⁷. Other landmarks, such as the UN's 2030 Agenda and the Ocean Decade framework, have also likely contributed to raising awareness of the problem. We cannot address changes in funding volume from our dataset and information on the matter is not easily available, especially at the multi-national level. However, the surge in publications suggests that funding and resources were allocated to MPL research around this period of intense societal debate on MPL.

As any field of science grows, its boundaries also grow, opening up new questions and paths of investigation⁴². The need for a more thorough understanding of MPL, required to define and implement effective policies, seems to be driving

recent diversification of research, and expanding the boundaries of the field. Continued evolution and funding enables new topics to emerge from previous knowledge on older ones. We can point out the emergence of topic 19: Environmental Impact of Face Masks and PPE in the COVID-19 Pandemic emerged in 2020, topic 21: Microplastic Pollution in Coral Reef Ecosystems and 38: Microplastic Contamination in Sea Salt Products emerged in 2015. and topic 24: Tire Wear Particles (TWP) as a Source of Microplastic Pollution from Road Dust emerged in 2016. The latter can be traced to a publication in that year by Dris et al.⁴³ on the atmospheric fall-out of synthetic fibres. This indicates that much of the content subsequently present in that topic, first appeared in this publication, suggesting it was a precursor to this topic of research. This also illustrates how BERTopic identifies topics and assigns documents to them.

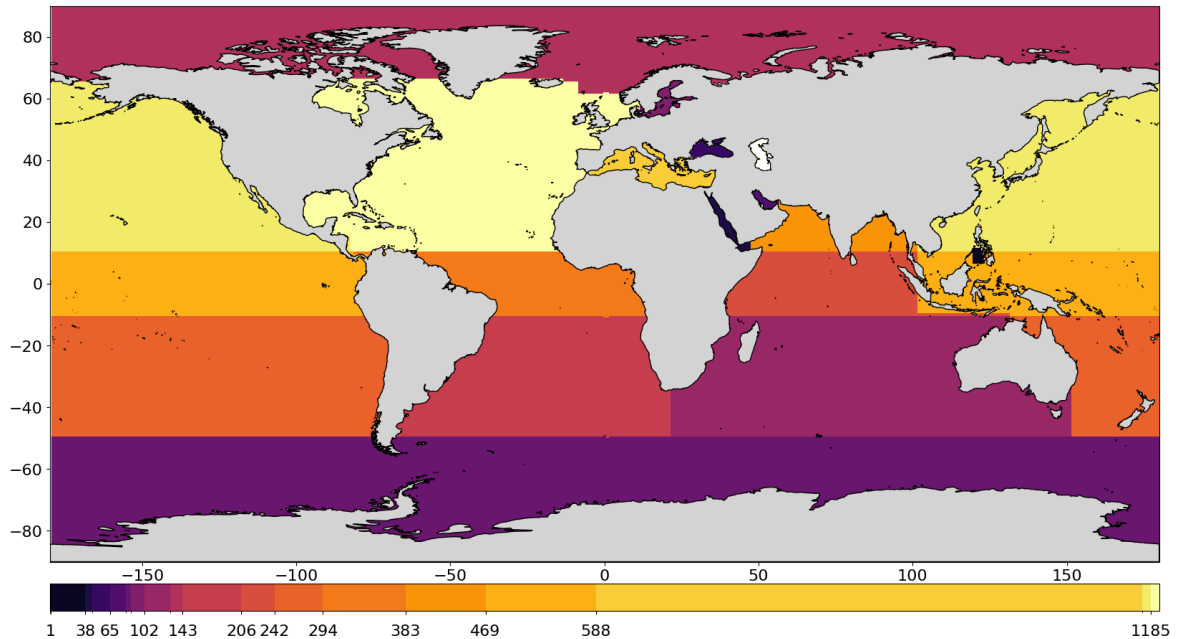
Figure 1: Topic Frequencies Changes in the frequency of publications on MPL from 2008 to 2022. Previous years are omitted as there were only a few publications and no apparent trend. A. The number of publications in each topic found by this study. B. Total number of publications over time.



Source: prepared by the author

3.4 REGIONAL DISTRIBUTION OF STUDIES

Figure 2 - Distribution of studies across ocean basins Papers were classified into basins (as defined by the NOAA's World Ocean Database) using an automatic Named Entity Recognition algorithm freely available at spaCy (spacy.io). Colours indicate the total number of papers in each basin according to the colourbar at the bottom of the figure. The spacing of bins in the colourbar is proportional to the number of papers and represents exact values found in the data. Values on the colourbar are to the left of their respective bin.



Source: prepared by the author

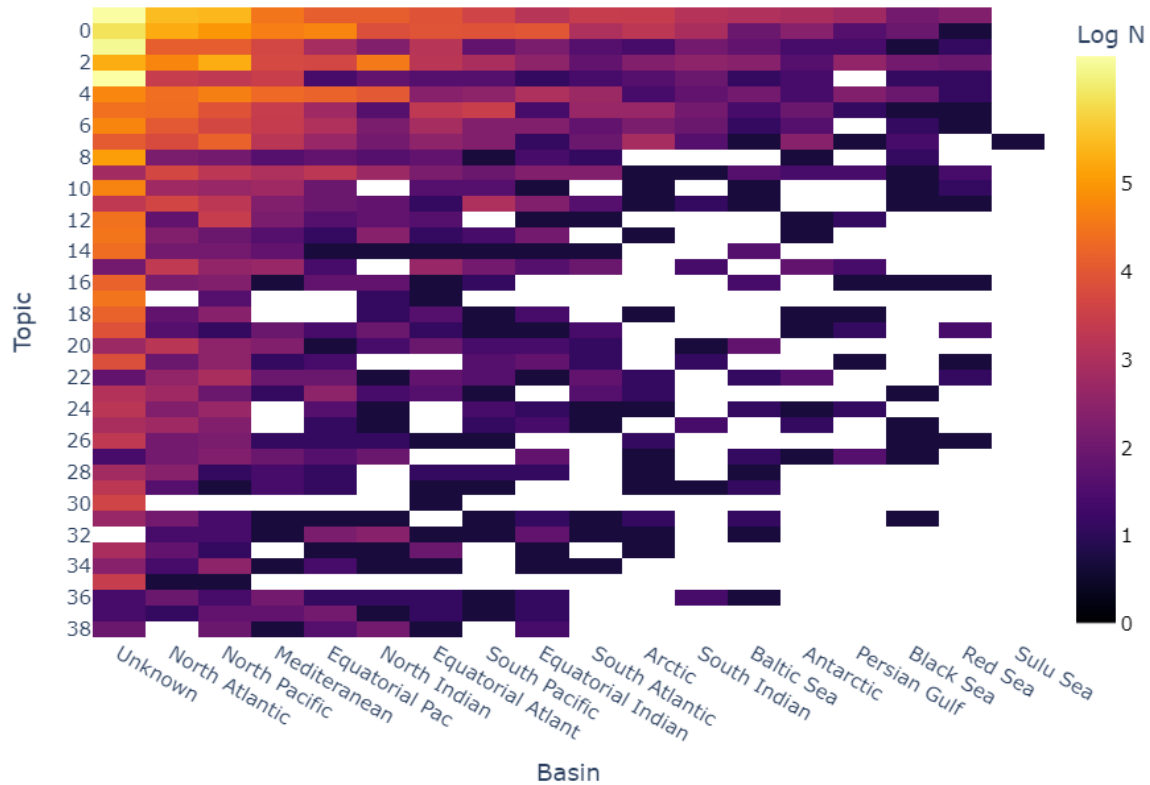
To assess the global distribution of MPL research, we classified publications into basins using the automatic recognition of place names in abstracts. The results, presented in Figure 2, reveal an uneven distribution, with a clear bias towards the Global North. The North Atlantic, North Pacific, and Mediterranean Sea dominate the literature. While this distribution aligns with the higher concentrations of MPL found in these regions⁴⁴, it also reflects the greater availability of research resources in these areas. Notably, we observe a strong correlation between the number of published MPL studies and the Gross Expenditure on Research and Development (GERD) as a percentage of GDP⁴⁵. This suggests that funding is a major factor driving MPL research. Although the GERD indicator, as presented by the UNESCO Institute for Statistics (UIS), does not distinguish between general and MPL research investments, regions with high GERDs tend to produce more research in this particular field. These are namely: North America and Western Europe, and East Asia and the Pacific.

After the Global North, the North Indian and Equatorial Pacific present a relatively high number of publications. These regions are hotspots for marine litter accumulation due to the discharge of plastic litter into the oceans^{46,47}. This is reflected in the number of publications, which appears higher than expected from trends in research spending. Given the global relevance of these regions, studies conducted there are likely to be supported by funding from elsewhere.

The map also reveals blind spots in the evidence base for MPL in the world's oceans. Smaller basins, such as the Persian Gulf, the Red Sea, the Black Sea, and the Sulu Sea together account for just 150 publications (less than 2%). This will in part be related to the smaller size of these basins. However, it is worth noting that the Mediterranean Sea, a small basin accounting for 0.8% of the world's ocean area, concentrates 6.42% of publications. This further corroborates the observed correlation of government spending on research and development (GERD) in the Global North, but also highlights a significant data gap in these regions.

3.5 FOCAL POINTS AND RESEARCH GAPS ACROSS BASINS

Figure 3: Distribution of publications per topic and basin following the classifications made with ML algorithms described in the methods. A log scale is used to better visualise differences in values, given their wide range. It is calculated by applying the natural logarithm of (1+N) to avoid errors when calculating log for N=0.



Source: prepared by the author

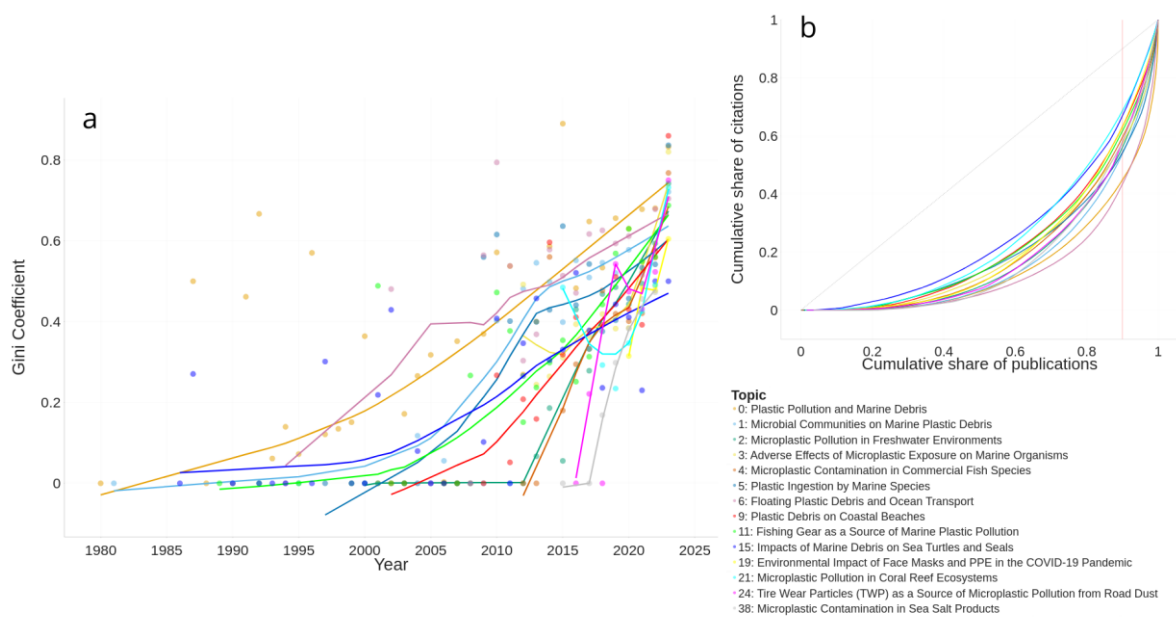
Figure 3, shown on a logarithmic scale (see Figure log_heatmap in the SI for absolute values) compares the two previous classifications and highlights the concentration of MPL research on just a few key topics and basins. It demonstrates that research across basins tends to be focused on dominant topics. Indeed, the earliest topics in MPL, (topics 11 and 15) tend to be more evenly distributed across basins, though our analysis found that there are still gaps. For example, topic 11: Fishing gear... is absent in the Antarctic and the Persian Gulf and topic 15: Impacts of Marine Debris... is not found in the North Indian, the Arctic, or the Baltic, Black and Red Seas. Overall, current knowledge on topics appears disproportionately biased by data from a few regions, risking generalisations about certain aspects of MPL that may not apply to lesser-studied locations.

However, a few topics stand out from this general trend: topics 17 and 35 have publication gaps across most basins (refer to the interactive Figure log_heatmap in the SI, for more details). In these cases, publications are

coincidentally concentrated in the “Unknown” location class, because the method was not able to identify locations in their abstracts. This suggests that these topics are rarely field-based studies and explains, at least in part, the gaps in the heatmap for the four most studied basins.

3.6 GREEDY PUBLICATIONS

Figure 4: Citations distribution: a) Gini coefficients per topic per year. Colours represent topics and lines are “loess” trendlines for each topic. b) The cumulative share of citations by topic versus the cumulative share of publications. Each paper had citations averaged by their age. Values were min-max normalised to highlight the similarity in the behaviour of distinct topics.



Source: prepared by the author

To assess how the research is being incorporated, we look at citation patterns. We observe a high citation inequality in the MPL literature with highly cited papers accounting for 30% to 82% of topic citations. We define highly cited publications as the top 10% of papers by citation count. The overall Gini coefficient at 0.66, is higher than that observed in larger fields of science, like Artificial Intelligence and Applied Physics, with up to 100,000 publications per year²⁷. In MPL, Gini coefficients on individual topics range from 0.48 to 0.83 (0.62 on average), with no correlation to topic size, that is, topics with just a few publications can be as unequal as topics with several hundred publications. This can be seen in Figure 4b which shows the similarity of the citation accumulation curve for individual topics.

While the field is not trapped in ‘old’ literature, knowledge dissemination has become less efficient. Figure 4a, shows that Gini coefficients tend to increase over

the years. The growing literature tends to cite relatively new papers, which underscores the evolving nature of the field and the researchers' urge to stay relevant. However, citations get progressively more concentrated in a smaller number of publications. This suggests that when there is too much new information available, researchers are more likely to cite the same publications. If a paper is cited more often, it is more likely to be cited again, creating a positive feedback loop. This leads to researchers becoming even more reliant on a few well-known publications. This mechanism prevents other publications on the same topic from being read and cited, effectively hindering knowledge dissemination.

3.7 DISCUSSION

Our topic model is based on coherent research themes rather than disciplines which are broad subject categories that often lack clear boundaries. This approach incorporates the dynamic and interdisciplinary nature of the research with a more granular classification. Our literature analysis highlights the diversification of topics of research since 2016, and reveals the growing specialisation in the field from the macro- toward the micro-scale.

Disparities in knowledge production within the field can create inequalities in mitigation and adaptation strategies to MPL. We found the North Indian and Equatorial Pacific, are still understudied compared to the North Atlantic, the North Pacific and the Mediterranean, despite the prevalence of riverine discharge of marine litter⁴⁶. The Global South, more broadly, was found to be characterised by more limited research and narrower scope. In these regions, vital economic sectors such as fishing and tourism are particularly vulnerable to the impacts of marine pollution. Addressing these regional knowledge gaps can help decision-makers elaborate more effective actions. In light of insufficient regional funding for research, international cooperation can be instrumental in developing the evidence base through the sharing of infrastructure and expertise. Finally, ensuring there is a common understanding of the major threats and how to address them can also enable consensus in international agreements, such as the international legally binding instrument on plastic pollution called for at the fifth UNEA⁴⁸.

Increased scientific production has certainly resulted in advancements in the field, with the volume of literature increasing and diversifying topics. We demonstrate that the most recent publications rapidly accumulate citations indicating the novelty of

contemporary research. This is expected in new fields, but the pattern is apparent in older topics as well. However, we highlight that despite the growth in the number of sources, researchers increasingly rely on fewer highly cited publications. We argue that the community's research efforts that resulted in a deluge of new publications are unfortunately causing an information overload that hinders the dissemination and reincorporation of new research into the literature. This trend will likely worsen in the future as the rate of publication continues to grow. This is particularly concerning given the need for an improved understanding of MPL, especially in the Global South, to support mitigation and prevention strategies and actions.

Curtailling the scientific output in favour of less but more innovative publications can help dampen the effects of information overload. Yet, this can be difficult to implement within current academic structures that favour productivity and would likely have detrimental impacts on early career researchers. Instead, we demonstrate with this study, how AI tools can be used to overcome human capacity constraints in digesting vast amounts of information and defining research gaps. While they have their limitations and don't supersede human labour entirely, they can augment our abilities, allowing us to process significantly more data, similar to traditional computing methods. While their value is evident in various contexts, they stand out when analysing large corpora.

By rapidly assessing the current state and trends in the MPL literature, our results enable policy and future research that target key areas and knowledge gaps. These AI methods can be replicated in different contexts and can have considerable impact in other fields. Increasing the granularity of Machine Learning models to identify more specific emerging issues or areas lacking, and improving interpretability, can widen the scope of their applications. Other methods based on Question Answering, Text Generation, Information Extraction, etc. can further improve how we access information in the future, though factors such as referencing and dealing with contradictory information will pose some reliability issues. Similar to the revolution enabled by traditional computing, we argue AI will likely become common practice in research and decision-making and will complement existing approaches, shaping the future of policy and research.

3.8 METHODS

The methods described hereafter were assembled to allow easy reproducibility by reducing subjectivity in every step, especially in the definition of the topic model and the interpretation of topics. Therefore, the only manual entry made by the researchers was done at the beginning, at the definition of the terms in the search query (TITLE-ABS-KEY (("marine debris" OR "marine plastic debris" OR "marine plastic pollution" OR "plastic pollution" OR "marine plastic" OR "plastic debris" OR microplastic OR nanoplastic OR microfiber OR "plastic pellet") AND (marine OR coast* OR sea OR ocean* OR *shore OR *pelagic OR benth* OR sediment))). The query was based on the terms suggested by Ivar do Sul⁴⁹, and chosen to be reasonably broad. We used the Scopus API to programmatically extract all the data from the database, which was useful since the results exceeded the download limits of the provider's website. To avoid errors, entries without abstracts or duplicates (based on DOI) were removed.

The classification of publications was made in terms of the topic of research. For that, we used the BERTopic software⁴¹ (version: 0:15:0 running on Windows 11), which applies a sequence of Machine Learning tasks to cluster documents into topics, and to produce interpretable representations of them. In summary, it is an unsupervised clustering of documents based on semantically meaningful vector representations of texts called embeddings^{38,39}. The software allows some modularity, so we can choose and define which algorithms are more suited to our application. For the embedding model, we opted for SPECTER⁵⁰ which is based on the Transformers architecture, also used by the more recent and conspicuous text or image generation tools. SPECTER was specifically trained on scientific publications, including their inter-relations based on shared citations, so it is expected to show enhanced embedding capacity for this type of text. The dimensional reduction and clustering steps were performed with the standard algorithms for BERTopic which are UMAP⁵¹ and HDBSCAN⁵², respectively.

The unsupervised nature of the modelling means we have no prior knowledge about the classes, or topics, that we want to classify publications into. This has many advantages, such as not needing to label a training set – as would be the case in supervised learning –, which is onerous, and being able to discover new classes. On the other hand, the definition of classes is left to the algorithm and can present some challenges to their interpretation. However, as in any classification task

and especially in machine learning, the goal is to find a model that is useful for the task at hand. In our case, this means having distinctive clusters that can be interpreted. To find the most suited topic model for our data, we created many models by randomly varying some hyperparameters for the UMAP and HDBSCAN algorithms, in what is known in machine learning as Randomized Search. Since our data is not labelled, it is not possible to validate the performance of the models with a test set. Instead, we compared them by calculating a cost function based on the proportion of publications that were classified with a less than 5% probability into the most likely topic. Essentially, the cost function measures the uncertainty or reliability of the cluster assignments based on probability values. However, there is no measure of the interpretability of topics, which is intrinsically subjective. The main factor that hinders the analysis of topics is the number of topics, where having too few results in over-generalization and too many results in over-specificity, and neither give much new information on our data. Therefore, we chose the model that provided the lowest cost while keeping a reasonable amount of topics, allowing both diversity and interpretability.

Finally, topic labels and short descriptions were automatically generated using a text-generation software: ChatGPT. It was given a list of the three most representative documents for each topic, which are nearest to the cluster centroids, and the 10 most relevant keywords and their associated c-TF-IDF score (refer to Topics Catalogue in the SI for definition). Both are provided by BERTopic for the user's interpretation but are limited in their capacity to convey meaning. While reducing the researchers' subjectivity, ChatGPT also facilitated the interpretation of topics by generating comprehensible labels and descriptions. They are fully shown in the Supplementary Material (Topic Catalogue.docx), which serves as an index for our analysis.

The second classification of papers was made into ocean basins according to basins described by the World Ocean Database (WOD)⁵³ from the American National Oceanic and Atmospheric Administration (NOAA). Considering the large number of publications that needed to be classified, we performed a Named Entity Recognition task, employing a pre-trained Natural Language Model to get information about locations from abstracts. The model automatically tags named entities in texts, including GPE (Geopolitical Entities) and LOC (all other location types) which we then use to classify the publications. Each GPE and LOC was attributed to a basin

based on its geographical coordinates, programmatically extracted from the GeoNames⁵⁴ API, and its distance to the nearest ocean calculated using the Haversine formula.

CODE AVAILABILITY

All data and code used for running experiments, model fitting, and plotting is available on a GitHub repository at https://github.com/henrique-back/mpl_topics.git

NOTES

* When the number of topics approaches 1 or the number of documents, the likelihood of documents being assigned to topics with high confidence increases, but these models can be difficult to interpret and provide little insight into the content of abstracts.

REFERENCES

1. Jambeck, J. R. *et al.* Plastic waste inputs from land into the ocean. *Science* **347**, 768–771 (2015).
2. Beaumont, N. J. *et al.* Global ecological, social and economic impacts of marine plastic. *Mar. Pollut. Bull.* **142**, 189–195 (2019).
3. Veiga, J. *et al.* Marine Litter in Europe - An integrated assessment from source to sea. *European Environment Information and Observation Network*
<https://www.eionet.europa.eu/etcs/etc-icm/products/etc-icm-reports/etc-icm-report-5-2022>.
4. *Trash Free Waters International Implementation Guide*.
https://www.epa.gov/sites/default/files/2021-01/documents/tfw_howtguide_20201228_finalprint_508_final.pdf (2020).
5. Bruch, C. *et al.* *Marine Litter Legislation: A Toolkit for Policymakers*.
<http://www.unep.org/resources/report/marine-litter-legislation-toolkit-policymakers> (2016).
6. van Oosterhout, L. *et al.* Public Perceptions of Marine Plastic Litter: A Comparative Study Across European Countries and Seas. *Front. Mar. Sci.* **8**, (2022).

7. Jacqueline, M. *From Pollution to Solution: A Global Assessment of Marine Litter and Plastic Pollution*. <http://www.unep.org/resources/pollution-solution-global-assessment-marine-litter-and-plastic-pollution> (2021).
8. Rochman, C. M. Microplastics research—from sink to source. *Science* **360**, 28–29 (2018).
9. Singh, R. P., Mishra, S. & Das, A. P. Synthetic microfibers: Pollution toxicity and remediation. *Chemosphere* **257**, 127199 (2020).
10. Steinmetz, Z. *et al.* Plastic mulching in agriculture. Trading short-term agronomic benefits for long-term soil degradation? *Sci. Total Environ.* **550**, 690–705 (2016).
11. Jadhav, E. B., Sankhla, M. S., Bhat, R. A. & Bhagat, D. S. Microplastics from food packaging: An overview of human consumption, health threats, and alternative solutions. *Environ. Nanotechnol. Monit. Manag.* **16**, 100608 (2021).
12. Dharmaraj, S. *et al.* The COVID-19 pandemic face mask waste: A blooming threat to the marine environment. *Chemosphere* **272**, 129601 (2021).
13. Rochman, C. M. *et al.* The ecological impacts of marine debris: unraveling the demonstrated evidence from what is perceived on JSTOR. <https://www-jstor-org.are.uab.cat/stable/24703091?seq=5> (2016).
14. Savoca, M. S., Wohlfeil, M. E., Ebeler, S. E. & Nevitt, G. A. Marine plastic debris emits a keystone infochemical for olfactory foraging seabirds. *Sci. Adv.* **2**, e1600395 (2016).
15. Tekman, M. B. *et al.* Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research and Meeresforschung. LITTERBASE. Online Portal for Marine Litter. <https://litterbase.awi.de/> (2020).
16. Soares, M. O. *et al.* Marine debris provide long-distance pathways for spreading invasive corals. *Sci. Total Environ.* **900**, 165637 (2023).
17. Borrelle, S. B. *et al.* Why we need an international agreement on marine plastic

- pollution. *Proc. Natl. Acad. Sci.* **114**, 9994–9997 (2017).
18. *Intergovernmental Negotiating Committee on Plastic Pollution*.
<http://www.unep.org/inc-plastic-pollution>.
 19. Letter to the Intergovernmental Negotiation Committee from Independent Scientists,
11th November 2023.
 20. *Science-Policy Interface for Plastic Pollution*. <https://www.grida.no/publications/1007>
(2023).
 21. Spring, M. *et al. Policy Brief: Creating a Strong Interface between Science, Policy
and Society to Tackle Global Plastic Pollution*. (2023).
 22. GESAMP. *Sources, Fate and Effects of Microplastics in the Marine Environment
(Part 2)*. [http://www.gesamp.org/publications/microplastics-in-the-marine-environment-
part-2](http://www.gesamp.org/publications/microplastics-in-the-marine-environment-part-2) (2016).
 23. Jeftic, L., Hägerhäll, B. & Aniansson, B. *Marine Litter, an Analytical Overview*.
<http://www.unep.org/resources/report/marine-litter-analytical-overview> (2005).
 24. Environment Program, U. N. *Marine Litter: A Global Challenge*.
<http://www.unep.org/resources/report/marine-litter-global-challenge> (2009).
 25. Notten, P. *Addressing Marine Plastics: A Systemic Approach – Recommendations for
Actions*. [http://www.unep.org/resources/report/addressing-marine-plastics-systemic-
approach-recommendations-actions](http://www.unep.org/resources/report/addressing-marine-plastics-systemic-approach-recommendations-actions) (2020).
 26. Environment, U. N. Policy Options to Eliminate Additional Marine Plastic Litter.
UNEP - UN Environment Programme [http://www.unep.org/resources/publication/policy-
options-eliminate-additional-marine-plastic-litter](http://www.unep.org/resources/publication/policy-options-eliminate-additional-marine-plastic-litter) (2021).
 27. Chu, J. S. G. & Evans, J. A. Slowed canonical progress in large fields of science.
Proc. Natl. Acad. Sci. **118**, e2021636118 (2021).
 28. Park, M., Leahey, E. & Funk, R. J. Papers and patents are becoming less disruptive

- over time. *Nature* **613**, 138–144 (2023).
29. Ziveri, P., Grelaud, M. & Pato, J. Actions of cities and regions in the Mediterranean Sea area to fight sea pollution.
 30. *Assessment of Options for Strengthening the Science-Policy Interface at the International Level for the Sound Management of Chemicals and Waste*.
<https://wedocs.unep.org/bitstream/handle/20.500.11822/33808/OSSP.pdf?sequence=1&isAllowed=y> (2020).
 31. Sutton, R. T. *et al.* An overview of clinical decision support systems: benefits, risks, and strategies for success. *Npj Digit. Med.* **3**, 1–10 (2020).
 32. Esteva, A. *et al.* A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
 33. Gan, X. *et al.* Network medicine framework reveals generic herb-symptom effectiveness of traditional Chinese medicine. *Sci. Adv.* **9**, eadh0215 (2023).
 34. Wang, K. *et al.* NERO: a biomedical named-entity (recognition) ontology with a large, annotated corpus reveals meaningful associations through text embedding. *Npj Syst. Biol. Appl.* **7**, 1–8 (2021).
 35. Vaswani, A. *et al.* Attention Is All You Need. Preprint at <http://arxiv.org/abs/1706.03762> (2023).
 36. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. Preprint at <http://arxiv.org/abs/1301.3781> (2013).
 37. Church, K. & Liberman, M. The Future of Computational Linguistics: On Beyond Alchemy. *Front. Artif. Intell.* **4**, (2021).
 38. Bengio, Y., Ducharme, R., Vincent, P. & Janvin, C. A neural probabilistic language model. *J. Mach. Learn. Res.* **3**, 1137–1155 (2003).
 39. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Preprint at

- <https://doi.org/10.48550/arXiv.1810.04805> (2019).
40. Wang, J., Peng, C., Li, H., Zhang, P. & Liu, X. The impact of microplastic-microbe interactions on animal health and biogeochemical cycles: A mini-review. *Sci. Total Environ.* **773**, 145697 (2021).
 41. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. Preprint at <https://doi.org/10.48550/arXiv.2203.05794> (2022).
 42. Gleiser, M. *The Island of Knowledge: The Limits of Science and the Search for Meaning*. (Hachette UK, 2014).
 43. Dris, R., Gasperi, J., Saad, M., Mirande, C. & Tassin, B. Synthetic fibers in atmospheric fallout: A source of microplastics in the environment? *Mar. Pollut. Bull.* **104**, 290–293 (2016).
 44. Isobe, A., Iwasaki, S., Uchida, K. & Tokai, T. Abundance of non-conservative microplastics in the upper ocean from 1957 to 2066. *Nat. Commun.* **10**, 417 (2019).
 45. Unesco Institute for Statistics. Research and experimental development (full dataset). (2023).
 46. Meijer, L. J. J., van Emmerik, T., van der Ent, R., Schmidt, C. & Lebreton, L. More than 1000 rivers account for 80% of global riverine plastic emissions into the ocean. *Sci. Adv.* **7**, eaaz5803 (2021).
 47. Omeyer, L. C. M. *et al.* Priorities to inform research on marine plastic pollution in Southeast Asia. *Sci. Total Environ.* **841**, 156704 (2022).
 48. END PLASTIC POLLUTION - TOWARDS AN INTERNATIONAL LEGALLY BINDING INSTRUMENT - English.pdf.
 49. Ivar do Sul, J. A., Tagg, A. S. & Labrenz, M. Exploring the common denominator between microplastics and microbiology: a scientometric approach. *Scientometrics* **117**, 2145–2157 (2018).

50. Cohan, A., Feldman, S., Beltagy, I., Downey, D. & Weld, D. S. SPECTER: Document-level Representation Learning using Citation-informed Transformers. Preprint at <https://doi.org/10.48550/arXiv.2004.07180> (2020).
51. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at <https://doi.org/10.48550/arXiv.1802.03426> (2020).
52. McInnes, L., Healy, J. & Astels, S. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* **2**, 205 (2017).
53. World Ocean Database. *National Centers for Environmental Information (NCEI)* <https://www.ncei.noaa.gov/products/world-ocean-database> (2020).
54. GeoNames. <https://www.geonames.org/>.

ACKNOWLEDGEMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

This work contributes to the ICTA-UAB “María de Maeztu” Programme for Units of Excellence funded by the Spanish Ministry of Science, Innovation and Universities (CEX2019-000940-M) and was funded by the i-plastic project (JPI-Oceans—Grant PCI2020-112059) and the Generalitat de Catalunya (MERS, 2021 SGR 00640).

AUTHOR INFORMATION

Authors and affiliations

Institut de Ciència i Tecnologia Ambientals, Universitat Autònoma de Barcelona, Barcelona, Catalonia, Spain

Henrique de M. Back, Michael Grelaud, Patrizia Ziveri

Departamento de Engenharia Mecânica, Universidade Federal de Santa Catarina, Florianópolis, SC, Brazil

Henrique de M. Back, Orestes E. Alarcon

Author contributions

All authors collaboratively contributed to the conception and design of the study. H.M.B, M.G and P.Z. collaboratively contributed to the analysis and interpretation of the data. H.M.B. created the software used in this study and drafted the manuscript. M.G. and P.Z. revised the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Additional information can be found in the Supplementary Information folder on article page of the publisher's website.

4 INTERSECTIONS BETWEEN MATERIALS SCIENCE AND MARINE PLASTICS TO ADDRESS ENVIRONMENTAL DEGRADATION DRIVERS: A MACHINE LEARNING APPROACH

Publicado em Outubro de 2023 em Environmental Science: Advances

Henrique de Medeiros Back^a, Daphiny Pottmaier^b, Camilla Kneubl Andreusi^a, and Orestes Estevam Alarcon^a

a Universidade Federal de Santa Catarina, Brazil

b Nottingham Trent University, UK

ABSTRACT

Plastics are an integral part of the material structure in modern societies. However, their widespread contamination in the environment raises concerns regarding the sustainable use of these materials. Plastic pollution research has accelerated rapidly in the past 20 years and developed as a broad and multidisciplinary field, due to its complex nature. Research on the fate of plastics in the environment, specifically their dispersal and degradation, is one of the pillars in the field. In this study, we used machine learning and text mining tools to bridge the marine plastics community to the existing knowledge from polymer science regarding degradation drivers. Topic modelling enabled visualising hot topic trends in marine plastics research and filtering for relevant publications with minimal expert intervention. The recurrence of drivers was verified in the literature, indicating particular areas of focus on the marine degradation of plastics. The results show weathering conditions to be more recurrent than polymer properties and the latter to be rarely discussed in depth. Also, biodegradation is found to be a hot topic in the field, while degradation caused by abiotic factors is less addressed. This may be necessary as polymer engineering is traditionally less concerned about the end-of-life of plastic products. Finally, we argue that not only a deep understanding of plastics from polymer science should aid fundamental degradation studies by the marine plastics community, but also that the latter has the opportunity to largely contribute to the former by filling the gaps it has left.

ENVIRONMENTAL SIGNIFICANCE

Plastics are ubiquitous in everyday life and the environment. Recently, the academic community has put tremendous effort into trying to describe its causes and understand its consequences. Notably, the fates of plastic particles are of particular interest. This has led to a rapidly growing repository of new publications that even experts struggle to be up to date. Machine learning can be a powerful tool to help synthesize and filter through these large amounts of unstructured textual data dispersed in scientific publications. This paper uses text mining to semi-automatically filter the relevant literature, explore trends in the field and verify possible mismatches between polymer science and the environmental plastics community. Our results show essential features of degradation not receiving due attention in the more recent marine plastics research. Topic trends also reveal hot topics in the field and the impact of the COVID-19 pandemic. These allow a broad view of how the field is evolving and of the relative importance of particular topics while showing the usefulness of machine learning for the apprehension of rapidly growing scientific pieces of literature.

INTRODUCTION

Plastic production has grown exponentially since the end of World War II and they are now used in all sectors of industry and day-to-day applications, replacing a multitude of other materials, especially natural ones. Since then, they have become so important to modern life that the current period of human history has been described as the Plastic Age, in relation to the Bronze and Iron Ages.^{1–3} The accumulation of plastic waste in nature, especially in the oceans, has been a topic of great public and scientific interest in recent years. Even if a more thorough understanding of impacts that range from entanglement, ingestion, and toxicity is still developing, the scale of this issue demands precautionary action.

Plastic pollution research typically focuses on three areas: sources, fates, and impacts.⁴ The fates of plastic particles in different environments are influenced by their behaviours, such as horizontal and vertical transport, weathering and fragmentation, particle aggregation, biofouling, and ingestion.^{5,6} New findings in these areas help identify accumulation hotspots, removal processes, regional/global mass budgets, degradation/fragmentation rates, and affected areas/species. Knowledge of particle concentrations and characteristics is crucial in understanding

how plastic pollution affects ecosystems. The issue of plastic pollution is also being studied from the perspective of social and political sciences, ocean physics, and chemistry. Particularly regarding the behaviour and fate of plastics in the environment, chemical and materials engineering can contribute with knowledge of polymer structures, their properties and mechanisms that lead to their degradation and change in characteristics.

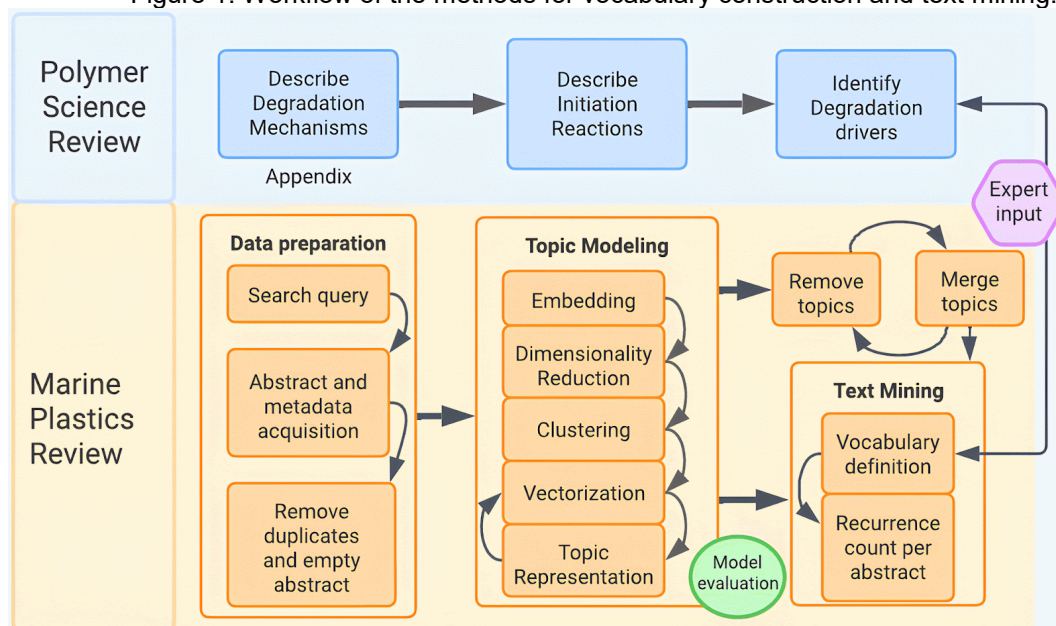
Reviews provide mechanisms to synthesise the literature and generate new insights, but traditional systematic reviews are time-consuming and subjective. In a fast-growing scientific field, by the time the review is published, it is already outdated. A new era in research calls for new methods of analysis and artificial intelligence (AI) may be our way to adapt. The use of AI and machine learning (ML) has grown in plastic pollution research and other environmental fields as tools to analyze large amounts of data, create predictive models, and reduce human error. Natural language processing (NLP) allows the transformation of unstructured text into data that can feed ML models. These versatile tools can help researchers navigate and analyze the growing flood of information.

This paper aims to visualize trends in plastic pollution research and, specifically, how often particular terms associated with the degradation of plastics are being considered. We used text mining and BERTopic, a topic modeling software, for this purpose. We show the focus given by the community on the plastics fate problem and point out the aspects of degradation that are being left out and should be considered.

METHODS

A workflow of the methods is shown in Fig. 1. To assess how polymer properties are related to materials degradation in the environment, we draw from the polymer science literature and briefly explain some concepts in the Degradation impact drivers section (more detailed explanations can be found in Appendix A). In the Results section we present a table summarizing the linkages between drivers and degradation processes as having high, moderate or low relevance. This is done similarly to how the IPCC's Sixth Assessment Report presents the relevance of key climatic impact drivers. 'High relevance' indicates drivers that are most prominent for their direct connection to degradation processes, while 'lower relevance' indicates weaker linkages or driving behaviours.

Figure 1: Workflow of the methods for vocabulary construction and text mining.



Source: Author

To understand how this knowledge is being incorporated by the marine plastics community, we counted the recurrence of key terms in the abstracts of this literature. The construction of a vocabulary for text mining was made concurrently with the identification of drivers, to find compatible terms between both literature studies. Only abstracts were used because they are readily available as metadata on SCOPUS and are summaries of the full text. They are more commonly used, but full texts may be employed, especially if there is a smaller literature available.

Papers were sourced from Scopus using the following search query, adapted from Ivar do Sul, 2018:7 “TITLE-ABS-KEY ((“marine debris” OR “marine plastic debris” OR “marine plastic pollution” OR “plastic pollution” OR “marine plastic” OR “plastic debris” OR microplastic OR nanoplastic OR microfiber OR “plastic pellet”) AND (marine OR coastal OR sea OR ocean OR beach)). The search was performed on February 24, 2022, and resulted in 8319 publications. Then, all abstracts were retrieved alongside the corresponding year of publication and DOI. We included a step to remove duplicates based on DOI identifiers as well as instances without abstracts.

For topic modelling, documents were embedded using SPECTER, a pre-trained BERT language model that is trained on scientific text.⁸ This was preferred over the standard MiniLM used by BERTopic, as it should be more capable of capturing semantic relationships between words in this kind of text. Indeed, initial

trials with both models showed more favourable results with SPECTER. Dimensionality reduction and clustering steps were defined using the standard BERTopic framework. It consists of performing these operations using UMAP (uniform manifold approximation and projection) and HDBSCAN (hierarchical density-based spatial clustering of applications with noise), respectively. The seed of randomness was fixed at “42” (this choice is arbitrary) to allow the reproducibility of results since the modelling is a stochastic process. It's important to note that the method utilises an unsupervised learning approach, which means we have no a priori knowledge of the topics present in our dataset. For further details refer to Appendix A.

A topic model was generated using BERTopic over the embedded text. It attributes each document to a topic based on the calculated probabilities of each document belonging to each topic. However, if none of these is higher than a minimum threshold, the model assigns them to an outlier topic. An initial analysis of this topic's representation indicated that many documents relevant to our study were set to be outliers. To avoid that, the threshold was set to 0.01 – after also having checked the topic probability distribution for several documents and tried using 0.02 and 0.005, which, respectively, left thousands and dozens of outliers. This was made with the assumption that the literature was well represented by all generated topics and that most documents would belong to at least one of them.

To evaluate the topic modelling, we chose the silhouette score as it is intuitive and simple to implement. It is a measure of the similarity between a document and its own topic in relation to its difference from other topics. Scores can range from -1 to 1 , where 1 means topics are cohesive and well separated. However, it should be kept in mind when evaluating the model that all evaluation metrics are only indicators of a model's performance and have better use in benchmarking different algorithms. Their evaluation and utility are ultimately dependent on the user's subjectivity.

To be able to identify what topics are about, we created topic representations using vectorization and c-TF-IDF (see Appendix A for further information). During

vectorization, the `stop_word` parameter was set to “English” to exclude words that add no meaning to the text (i.e., and, the, in, etc.) and could hinder topic differentiation. Another parameter, “`ngram_range`” was set as (1, 2) to include compound words like “plastic waste” instead of looking just at “plastic” and “waste”.

By analysing the hierarchy of topics, we were able to remove topics unrelated to the scope of the study and merge topics that were closely related. This step required a more in-depth look into topics by inspecting topic word scores, that is, how much each word is associated with a certain topic, and representative docs, which are the abstracts that better represent a topic (all of this can be seen in detail in the `intersections.ipynb` file in the ESI†). This step (merging and removing) was conducted recursively two other times as topic representations are updated and may better reveal the meaning of certain topics.

To visualize the development of the field, the number of publications was plotted per year and by topic. We used the “topics over time” function on BERTopic to calculate the frequency of papers present in each topic over the studied period. Topic representations are dynamic which allows us to see how topics evolve as well (interactive plot available in the `intersections.ipynb` file in the ESI†).

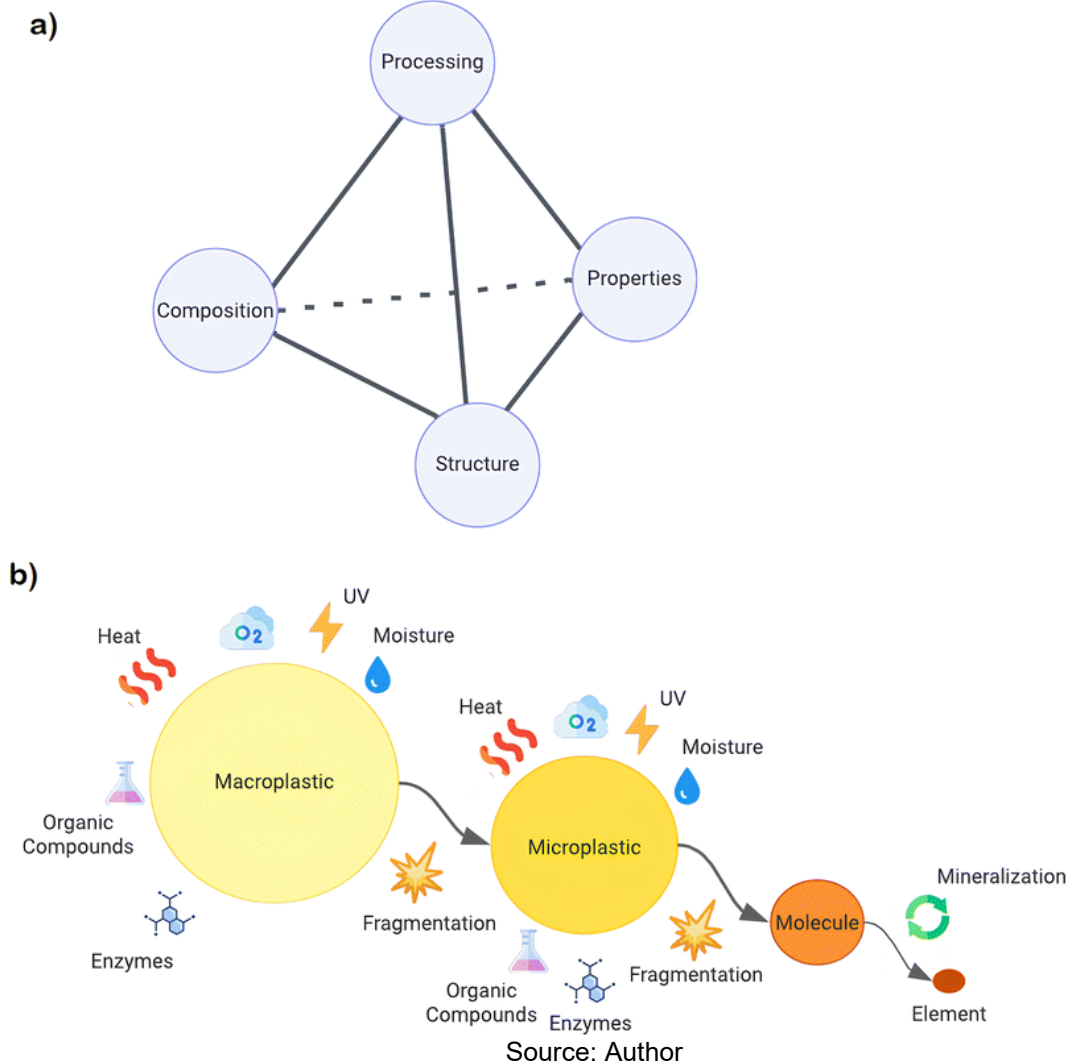
After the modelling and screening of papers, keyword recurrence was verified in the corpus of each topic. For this, we defined a function in Python and applied it to our corpus using the constructed vocabulary. The output recurrence is the number of abstracts that contain at least one mention of a particular word, which is not the same as word occurrence which would count every mention of that word. Details can be found in the `text_mining.ipynb` file in the ESI.†

DEGRADATION IMPACT DRIVERS

Degradation is any form of loss of function or property impaired by some kind of agent as a result of physicochemical modifications in the material. In other terms, degradation imparts changes in composition and/or structure that affect properties. This relationship between a material's composition, microstructure, processing, and properties is widely accepted in materials science. A conventional way to represent this is shown in Fig. 2a. On a molecular level,‡ any degradation of polymers starts with the breaking of a covalent bond in the macromolecule, which generates reactive species, called free radicals, responsible for propagating the degradation reaction. This energy can be provided through heat, light, chemical potential or mechanical

stress (Fig. 2b). The magnitude of the energy needed to break the bond depends on the material properties. Therefore, we define degradation drivers as material properties and environmental factors that interact to cause the weathering of plastics.

Figure 2: (a) The materials science tetrahedron; (b) particle size reduction as caused by degradation and fragmentation of plastics.



MATERIAL CHARACTERISTICS

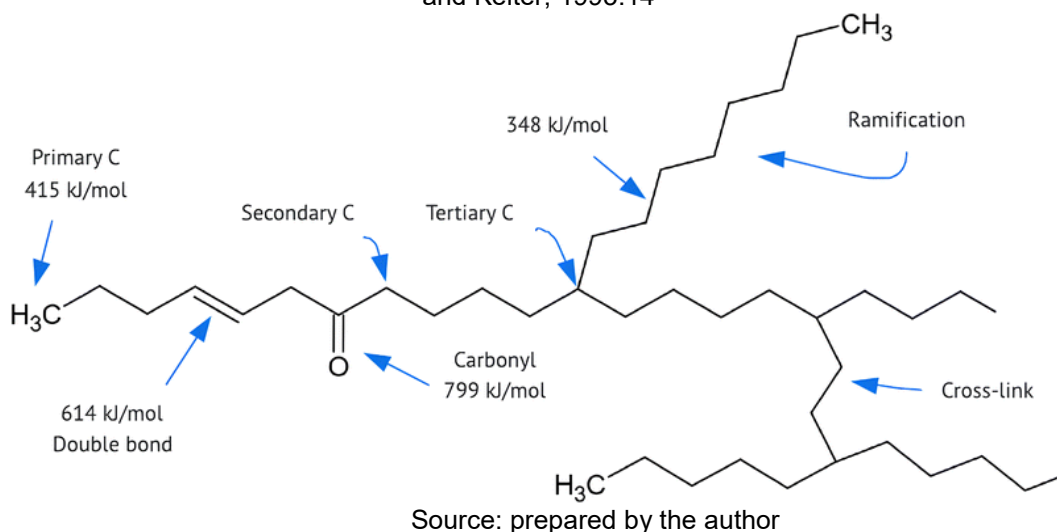
Composition

The terms “plastic” and “polymer” are often used interchangeably in the environmental literature, which is not accurate. Plastics are materials that belong to the class of polymers, which has a unifying factor in their chemical structure: all polymers form long chains from the repetition of a monomeric unit. They can occur naturally, like proteins or cellulose, or be synthesized artificially. The latter are commonly referred to as “plastics”, due to the plasticity that many of them have –

especially under the application of heat, a characteristic that differentiates them from the majority of natural polymers. Considering the varied mixture of substances (additives) in plastics, the term “polymer” can also be used when referring to the main resin that makes up a plastic product/particle or simply to its macromolecular chains.

These macromolecules consist mostly of C–C (carbon–carbon), C–H (carbon–hydrogen) and C–R (carbon–radical) bonds where R can be a halogen, a functional group, such as an aromatic ring, etc.^{10,11} The energy of bonds depends on their nature and that of the surrounding bonds in the molecule. C–C bonds can also be of different types, or hybridised states, as shown in Fig. 3. Primary carbons will have higher C–H bond energy and will be less prone to bond dissociation. Tertiary carbons are more susceptible to bond breaking in the C–H bond than the others because that carbon is already strongly bonded to other carbons.¹² The same effect is observed for C–H bonds surrounded by other chemical species, such as C[double bond, length as m-dash]C (polystyrene), C–O–C (like in polyesters), and C–N–C (like in polyamides) that have stronger bond energies.¹³ Fig. 3 illustrates a section of a polymer chain with different carbons and other structures mentioned further in the text and provides typical bond energy for several species. The actual values can only be stated in relation to the surrounding chemical structure. Softwares like Gaussian and Gamess can be used to predict actual bond energies in complex molecules.¹⁵ Generally, strong bonds will reduce the energy of surrounding bonds.

Figure 3: Example section of a polyethylene macromolecule with possible deviations from the ideal structure and corresponding dissociation energy of some typical bonds. Values are related to the bonds made between carbon and carbon or hydrogen or oxygen. Source of values: Huheey, Keiter and Keiter, 1993.¹⁴



Microstructure

When studying polymer degradation, another point that can be broadly addressed is the strength of intermolecular forces, as its increase will impart greater stability towards degradation.¹⁶ Stronger intermolecular forces can be imparted by strong polar groups or by closely packed macromolecules. The first is directly related to chemical composition (already addressed), while the second is related to the physical microstructure of the polymeric material.

Polymers' macromolecules have a high molecular weight (MW), and for that, they will behave very differently compared to the substances containing just the monomers. Increasing MW increases the intermolecular forces between the polymer chains, which is related to the cohesive energy responsible for the mechanical strength and chemical stability of polymers.⁹ MW is always in some range and this distribution is also a factor of importance for properties.

Polymers can also be amorphous or semi-crystalline. The degree of crystallinity, or the ratio between crystalline and amorphous phases in the material, affects the strength of intermolecular forces because crystals have a more compact microstructure. The closer these molecules are, the more they will interact with each other. This increases mechanical strength, melting temperature and stability towards oxidation and many other degrading agents.^{17,18} This includes stability to biodegradation, as we will see further in the text. Engineers control this ratio to produce materials with the desired properties, by controlling processing conditions, MW and composition, especially bond rigidity and side group sizes.

A similar effect is the result of cross-linking, which is the formation of covalent bonds between polymer chains that can happen during polymerization or as a result of degradation.¹⁹ It reduces molecular mobility, because it ties one molecule to the other, restricting their conformation, thus increasing bulk stiffness and elevating the glass transition temperature (T_g) and melting temperature (T_m). Contrarily, branching, or ramifications, which are secondary polymer chains attached to a primary backbone, weaken intermolecular forces because they reduce densification and impair crystallization. Additionally, they weaken intramolecular forces through the presence of tertiary carbons.²⁰

Some other considerations can be made regarding physical arrangements and the microstructure, such as the form of crystals or the degree and length of

ramifications, but they will not be addressed here for brevity. In any case, they will also influence chain rigidity or crystallinity.²¹

Properties

Plastic properties are the macroscopic characteristics of the material that determine its behaviour during use or after end-of-life. Degradability, or even compostability, for instance, are some of these properties. As stated before and shown in Fig. 2a, properties are conditioned to the composition and microstructure. However, they are more convenient to measure. A good example is the glass transition temperature, a typical property of amorphous or semicrystalline materials that is commonly measured for polymers. It indicates the mobility capacity of polymeric chains and it's correlated to several other properties, such as molar mass, degree of crosslinking and crystallinity.^{21,22} While it is not involved in the degradation process, it may be a relevant measure as it reflects composition and structure factors that are.

Likewise, density can also be of interest. Determining whether a particle will float or sink away from the surface of a waterbody and away from sun exposure influences the intensity of degradation. Mechanical strength, on the other hand, is important for fragmentation and when reduced will tend to increase the surface-area-to-volume ratio, but it is not directly related to degradation itself. However, it is a useful measure of the degree to which a plastic particle has been degraded already and can thus, appear in the literature as such.

WEATHERING FACTORS

Ultraviolet radiation

Photons of ultraviolet radiation can initiate a photochemical reaction that causes degradation. Once started, it will propagate due to the presence of free radicals. It requires a chemical species to be in an excited electronic state, which happens when an unbound electron increases its energy by absorbing a photon, mostly in the ultraviolet (295–380 nm) range.²³ The mere incidence of light, however, does not necessarily cause excitation. It must be absorbed. And since energy is quantized, a chemical bond will only absorb photons at specific frequencies. The chemical group responsible for light absorption in the UV-Vis region

is called a chromophore. It can give colour to a material since only part of the incident radiation will be reflected but is also responsible for photodegradation. There are two types of chromophores: intrinsic (present in the macromolecule) and extrinsic (additives or contaminants). Photostabilizers are additives used in plastics production as protection from photodegradation because they contain chromophore groups that preferentially absorb UV.

Most polymers have no intrinsic chromophore groups but can degrade under UV because of impurities or additives. PE, for instance, can undergo photochemical degradation, because in real-world conditions there will necessarily be contaminants or previous degradation that allow it. Similarly, PS does not absorb at any frequencies in the solar spectrum, but irradiation can generate an excimer (association of two adjacent aromatic rings). This excited state is responsible for initiating the photodegradation process even in energy bands not absorbed by chromophores. In most polymers, these defects will be present in the form of C[double bond, length as m-dash]C conjugated or aromatic double bonds or even as carbonyl groups (C[double bond, length as m-dash]O).

Photodegradation will cause the appearance of carbonyl groups (rigid double bonds) and crosslinking, both reducing the mobility of polymeric chains in the plastic material. However, since the yield of chain scission depends on chain mobility, the process will slow down with further degradation²⁴ Additionally, photodegradation only occurs on the surface of the material up to a few micrometres, depending on the energy of the photon and the material.²⁵

Temperature

Due to their relatively weak covalent bonds, the maximum service temperature of polymers is considered low if compared to other materials, like most metals and ceramics, ranging typically from 50 °C to 200 °C.²⁸ The effect of temperature on degradation cannot be neglected though, as it influences many other chemical processes, such as migration of low molecular weight additives, hydrolysis and photo-oxidation.^{26,27,29} Due also to their thermal insulation capacity, polymers will suffer thermal degradation differently depending on how long they are exposed to higher temperatures.³⁰

Polymers with the same types of covalent bonds will have different thermal decomposition temperatures due to molecular arrangements. For instance, PE and

PP only have C–C and C–H bonds, however, PE will only contain low-energy tertiary carbons in its ramifications, while PP has one of such carbons at each monomeric repetition. This makes PP less stable and more prone to oxidation.²⁶

Once thermal degradation is initiated it is highly favoured by free radicals and other compounds typically found in polymers, such as residual monomers and catalysts, as well as defects. These may originate from the production process itself, either during polymerization or from heating or chemical contamination during manufacturing. Additionally, chain defects caused by previous degradation can also reduce thermal stability.³¹ In contrast to photodegradation, thermal degradation occurs in the bulk of the material, not just at the surface.

Chemicals

Polymer degradation can also occur by reactions with chemicals. It can happen at the surface of the material due to contact with substances in the environment, or in the bulk, caused by impurities or by the diffusion of liquids, such as water. Unlike UV or temperature, chemicals can cause depolymerization, where the polymer is converted back into monomers, which is good for recycling but is a source of toxic compounds in the environment.

The reactivity depends on the affinity of both polymer and chemical. Therefore, it should be analysed on a case-by-case basis. However, shared characteristics of substances can be indicators of reactivity. For instance, it will be higher for organic compounds than for inorganic salts, alkaline or weak acid solutions. Additionally, more hydrophobic materials, such as polyolefins, will tend to react more intensely with oils and other hydrophobic compounds. Persistent organic pollutants are some of these compounds that co-occur with plastics in marine environments and can favour their degradation. Similarly, polymers containing polar groups such as polyamides, polyesters and polycarbonates, have higher interactions with water that can cause hydrolysis.

Microbes

The general process of biodegradation involves the excretion of enzymes by microorganisms that cause hydrolysis of the polymer chain into intermediate compounds that can then be assimilated by these organisms as a source of

carbon.³² Several polymers have been reported to suffer biodegradation, mostly from bacteria or fungi,³³ but polymers containing ester, amide or carbonate bonds in the backbone are more prone to it, because of their susceptibility to hydrolysis in these bonds.

To produce biodegradation, microbes must have access to water and the polymer chain. This is generally related to how compact a certain microstructure is and how easy it is to rotate macromolecules around molecular bonds. Strong bonds, such as C[double bond, length as m-dash]C and C[double bond, length as m-dash]O, have the tendency to reduce the energy of adjacent bonds, which makes rotating around them easier.³⁴ These carbonyl bonds also increase hydrophilicity, which facilitates microbe adhesion and water availability. Additionally, bulky side groups, ramifications and cross-linking all limit the molecules' conformability, thus hindering biodegradation.³⁴ The degree of crystallinity also has a role, as the crystalline structure is more compact and less free to move. The molecular weight is of particular importance, as microbes cannot assimilate long chains. Even polymers considered biodegradable, such as PLA and PHB may have to suffer previous chain scission in order to be metabolized by microbes.³⁵

Aside from polymer properties, environmental conditions also have an influence on biodegradation. Temperature is the first factor to be considered. Increases in temperature increase the rate of biodegradation, until an optimum value above which biological activity is reduced. Generally, thermophilic composting (up to 70 °C) in institutional facilities produces faster biodegradation than mesophilic home compost (around 35 °C).³⁶ Moisture, pH, oxygen and UV radiation are also decisive factors for microbial growth and will influence the rate of biodegradation.³⁷

RESULTS & DISCUSSION

Fig. 4 shows the linkages between key degradation drivers and degradation processes and compartments summarized from the previous sections where we can easily identify drivers with high relevance for one or more degradation processes. These linkages will be considered when we evaluate the recurrence of drivers in the marine plastics literature. But first, we classify the publications in this literature by performing our topic modelling with BERTopic as described in the Methods section. A total of 7828 abstracts from the preprocessed dataset based on our search query on SCOPUS were included. Initially, the modelling resulted in 89 topics and around

3000 documents were assigned to an outlier topic. After applying a new threshold, only 41 documents were classified as outliers.

Figure 4: Linkages between degradation drivers and processes given with high, moderate and low relevance. Details on these linkages are expressed in the previous sections.

Weathering Factor	Degradation process			
	Photo	Chemical	Thermal	Bio
UV				
Temperature				
Moisture				
pH				
Oxygene				
Organic Compounds				
Microorganisms/enzymes				

Polymer properties				
Bond energy				
Linearity				
Functional groups				
Cristallinity				
Molecular weight				
Chain mobility				
Glass transition temperature				
Defects				
Chromophores		-	-	-
Carbonyl group				
Free radicals				
Residual Monomers				
Catalysts				
Additives				

Source: prepared by the author

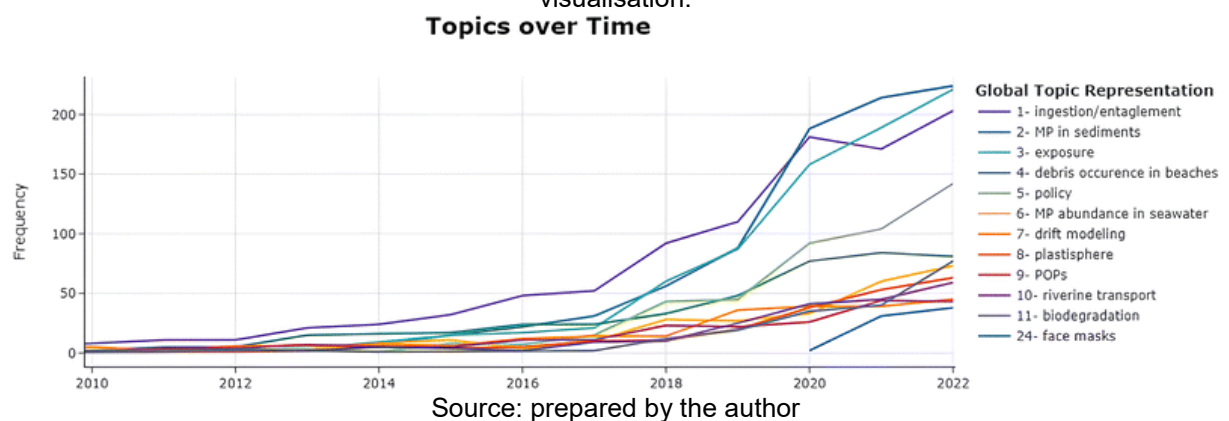
It is important to note that the topics found can differ using other hyperparameters, models or even slightly different datasets.³⁸ In any case, the modelling gives us enough information about topics contained in our dataset, that we can use to understand it. Despite being able to distinguish between very specific topics, the great number of topics given by the model hindered a better comprehension of what is going on in the literature. So, topics were merged by similarity as described in the Methods section. Additionally, we removed a total of 6 topics consisting of 242 abstracts (3%) that were outside the scope of our analysis in

topics such as “facies_sandstone_deposits_mass transport_mounds”, “sensor_sensitivity_salinity_microfiber_high sensitivity”, “composites_nanofibers_microfibers_melt_collagen” and “soil_soils_microplastics soil_agricultural_terrestrial”. Finally, the literature could be represented by 41 different topics.

The calculation of the silhouette score for the evaluation of topic cohesion resulted in a value of 0.26. With values ranging from -1 to 1 , a score of 0.26 is reasonable. While we cannot be certain if the topic assignment of documents is correct and we could be leaving out papers that should be considered, the silhouette score, accompanied by subsequent topic explorations, is sufficient for enough confidence in the results. Moreover, the method drastically reduces individual subjectivity which is concerning when screening the literature.

In Fig. 5 we show the frequency of new publications per topic per year since 2010. The period was only chosen to facilitate the visualisation, as frequencies remained relatively constant before that. Similarly, not every topic was selected to facilitate visualization. The internal build of the algorithm allows the user to visualise not only the frequency of documents per topic per year but also how each topic representation changes over time. We have found that they don't change significantly, but sometimes a word may appear in a topic's representation, indicating that it has gained relevance in that topic.

Figure 5: Topic trends in marine plastics research – some topics were omitted to facilitate visualisation.



From Fig. 5, we observe that “topic 1 – ingestion/entanglement” and “topic 4 – debris occurrence in beaches” are the most frequent until 2015, and then “topic 2 – MP in sediments” starts to grow faster and from 2017 onwards “topic 3 – exposure” has a sharp increase as well. Overall, these are the most common topics, but this result may be biased to indicate how broadly those topics were defined by the model. So we should be careful with that interpretation. However, in a general sense, the results are something that could be expected.

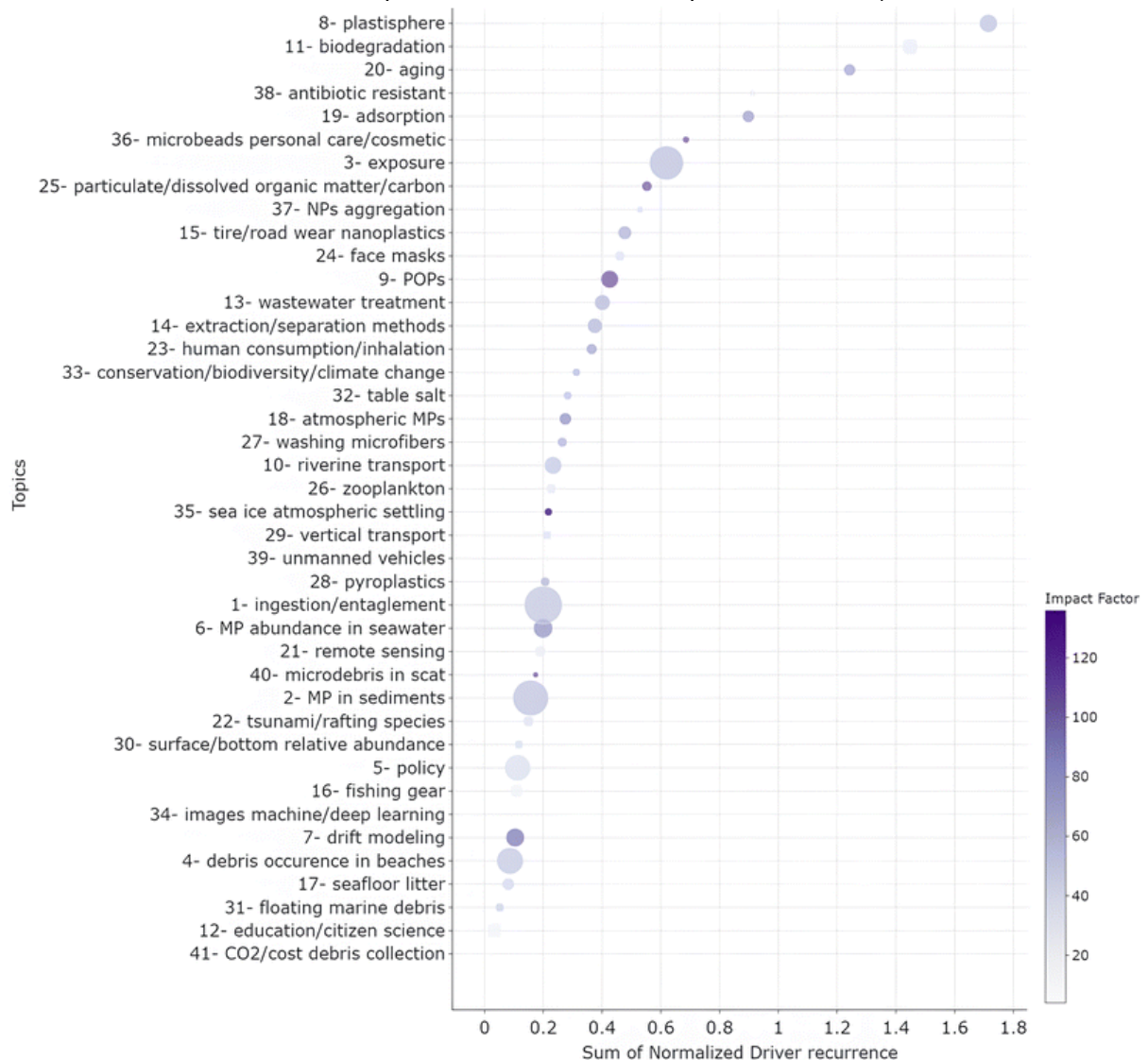
It may be more insightful to look at growth trends, instead. For instance, the number of publications per year in “topic 4 – occurrence in beaches” has had little change since 2020, while “topic 5 – policy” grew much faster. This may indicate a change of interest inside the research community towards new legislation instead of accounting for the occurrence of plastics. We could also say that “topic 11 – biodegradation” was a hot topic in 2022, because of the sharper increase in publications in that year. Particularly in the case of Topic 5, we can speculate the reasons for this growth as being related to key international agreements being drafted and signed, such as the plastic waste amendment to the basel convention on hazardous wastes and the UNEA resolution passed in 2022. These clearly indicate an overall interest in policy that can have an effect on scientific publications as well.

In general, all of the major topics inside the field of plastic debris in marine environments have experienced changes in their trends between 2020 and 2022. This may be explained, at least partially, by the COVID-19 pandemic. Due to lockdowns, difficulties in accessing public places and laboratories may have influenced how researchers conducted their new and ongoing projects, but also, funding and personal interests could have been shifted from some areas to others.^{39,40} One example of this is the emergence, in 2020, of “topic 24 – face masks”.

For some topics, you could argue that they are no longer frequent because they are saturated, with a few studies having a high citation impact and less interest in conducting new research on that topic. In Fig. 6, a look into the impact factor by topic can help validate this assumption. This may be true for topics 7 and 9 (“drift modelling” and “persistent organic pollutants”, respectively). Both have a similar slow growth curve in Fig. 4 and a higher impact factor. Additionally, documents associated with “topic 11 – biodegradation” seemed to have a smaller impact than those associated with “topic 20 – ageing” which tends to be more about physicochemical

degradation. Again, this could be because the latter is already saturated and looking at this topic's trend we see it has a small and steady growth over the years compared to a more abrupt growth in Topic 11, as mentioned previously. It could be helpful to compare “Citations over time” with “topics over time”, but unfortunately, the former is not provided as metadata by SCOPUS.

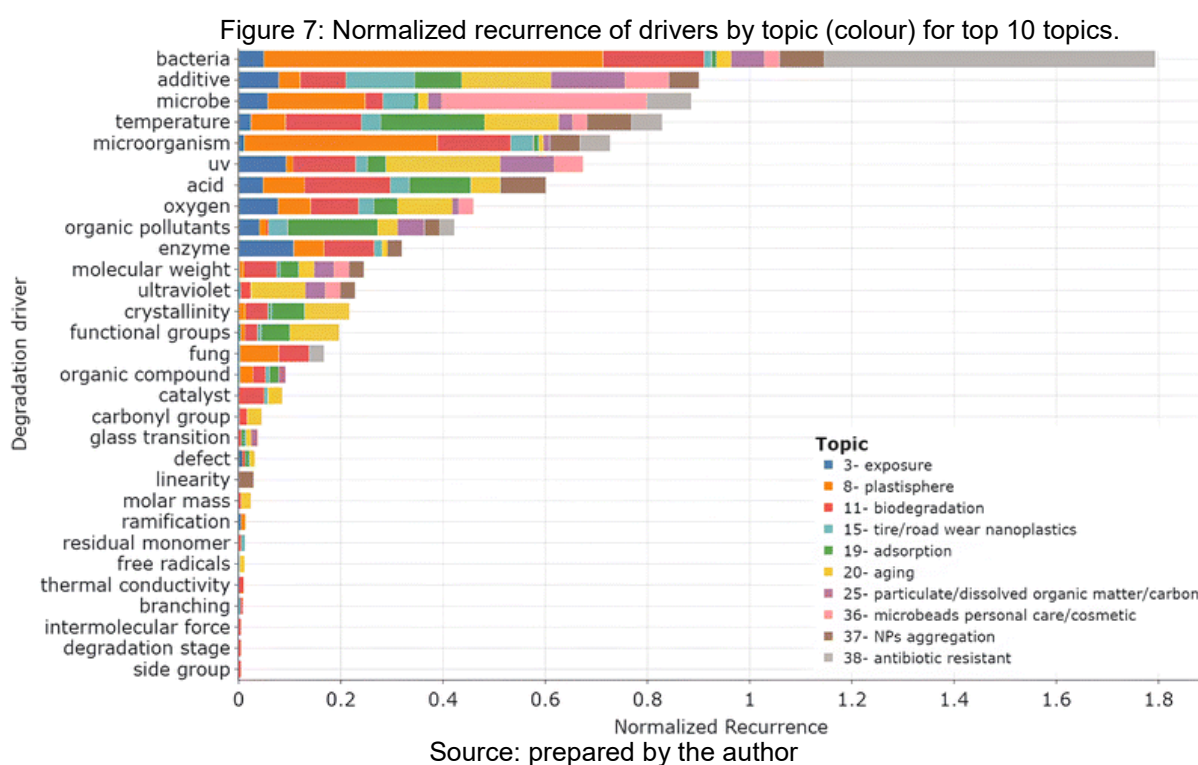
Figure 6: Sum of normalized recurrence of degradation drivers in the marine plastics literature by topic considering topic size (number of publications in each topic, set as size of bubbles) and topic impact factor (total number of citations for all publications in a given topic divided by the number of publications in that same topic, set as colour).



Aside from topics' size and impact factor, Fig. 6 also shows the sum of normalised driver recurrence per topic. This variable accounts for the sum of recurrences of every degradation driver for each topic. This means that when the algorithm finds a word from our drivers vocabulary in an abstract it will count 1, no

matter how many times that word appears in that abstract. The normalization, or division by the size of the topic, allows a better comparison between small and large topics. Interestingly, topics “8 – plastisphere”, “11 – biodegradation” and “20 – aging” have the highest recurrence of degradation drivers, which is expected and helps validate our topic modelling. Regarding Topic 8, it may not be specifically about degradation, but it certainly overlaps with it.

To gain a better understanding of which drivers are being addressed the most by each topic, we created a stacked bar graph for the normalised driver recurrence by word by topic for the top ten topics in Fig. 6, which is shown in Fig. 7.



Considering the findings shown in Fig. 4, each weathering factor can have a different impact depending on the type of degradation. By comparing them with Fig. 7, a possible interpretation is that biodegradation is being more thoroughly addressed than other modes of degradation, followed by photodegradation. This is corroborated by the fact that topics “8 – plastisphere” and “11 – biodegradation” have higher normalized sums of degradation drivers, but are also more frequent topics overall (size of bubbles in Fig. 6).

We can observe, from Fig. 7, that weathering factors are much more recurrent than material properties in topics associated with degradation except for the term “additive”, which seems to be an outlier. It is relatively well mentioned in most of

the selected topics, but it can be noted that its recurrence is greater in topics “15 – tire/road wear nanoplastics”, “20 – ageing” and “25 – particulate/dissolved organic matter/carbon”, indicating the relevance of this property in these topics. However, due to the limitations of this analysis, it is unclear if the term “additive” is being mentioned as a potential driver of degradation or not. The same reasoning can be applied when looking at weathering factors such as oxygen, which can be present in the context of degradation or not. Mentions by topic may provide some context in the use of words. For instance, oxygen must be present in abiotic degradation to produce oxidation and formation of carbonyl bonds, so mentions in “topic 20 – ageing” are likely to be used in that context. In the context of “topic 11 – biodegradation”, oxygen may be a parameter for biofouling and microorganism development. In any case, oxygen will be relevant for degradation even if indirectly.

It is pertinent to point out that “topic 20 – ageing” has a small recurrence of words like bacteria or microbe, while the recurrences for temperature, UV, and oxygen are higher. This is in good agreement with the label given to this topic which is based on the most important words for that topic (refer to Methods and Appendix for more details). Also, for material properties like crystallinity and functional groups, the recurrences are greater than in other topics. By this, we can infer that topic 20 deals more with drivers related to materials and physicochemical degradation than the other topics, especially topics 8 and 11 which are more biologically focused. The fact that our topic model was able to differentiate these topics and the finding that driver recurrences vary significantly between them show an apparent compartmentalization of studies. Although it's not possible to assert the researchers' background on any of these publications, we see that biologically focused researchers are much more likely to limit their analysis to environmental factors and dismiss material characteristics, contrarily to studies on physicochemical degradation. This result clearly points to a lack of fundamental concepts of polymer science in the environmental science community, given it is clear that both material and environment interact to produce degradation and should be equally assessed.

“Topic 38 – antibiotic resistance” contains more mentions of words like bacteria and temperature, and little to no mentions of other drivers, including enzymes. This suggests it may be a false interpretation to consider it a relevant degradation topic and that these drivers are mentioned in other contexts. The same may be true for “topic 19 – adsorption”, where “additive” and “organic pollutants” may

be mentioned not as degradation drivers but as substances that are adsorbed by plastics.

In the case of “topic 37 – NPs aggregation”, there seems to be an even recurrence of physical and biological weathering factors, but also material properties. Although it's hard to say for certain what they are referring to, this visualisation helps getting an understanding of particular focuses on this field. From Fig. 4, we also know that molecular weight and crystallinity are material properties that have a high relevance for degradation, especially in biodegradation. The importance given to them by studies on marine plastics, relative to other material properties reflects this. Interestingly, crystallinity does not appear in biodegradation studies as often as molecular weight and that could be because the latter would appear as a product of degradation as well – so it would be more recurrent, than crystallinity. Also, the reduction of the molecular weight or the formation of carbonyl groups are the main ways to verify degradation, so it was expected to see these terms being frequently mentioned. Interestingly, this is not the case for the latter.

The term “bond energy”, which is the main property to be considered in the degradation of polymers, doesn't appear in any of the topics. Although it could be addressed indirectly by other material properties, this doesn't seem to be the case. The recurrences of “free radical” and “chromophore” follow the same tendency. Other important properties such as chain mobility, which can be addressed indirectly by linearity, ramifications or the glass transition temperature are rarely present. It is true that crystallinity also plays a role in chain mobility. However, it may often be addressed as a factor of increased density and reduced intermolecular space, which hinders degradation, especially by biological drivers. Nevertheless, this highlights the fact that a more in-depth focus on polymer properties and degradation mechanisms is still lacking.

CONCLUSIONS

ML and NLP tools boosted the process of literature screening and topic understanding. Nevertheless, they can't be used automatically without human intervention. Removing and merging topics as well as constructing a vocabulary to search demands expert knowledge. While there is a lot more they can accomplish and future developments promise greater innovations, artificial intelligence does not supersede human intelligence. The usefulness of these tools demonstrates that they in fact expand our capacity to grasp and build knowledge.

The finding that biological degradation is more prevalent in the marine plastics literature, might be because ecologists are commonly looking at microorganisms since they are a major component of the marine environment. In spite of that, there is a lack of previous understanding of this kind of degradation. Because polymer scientists are mostly focused on designing materials for applications, they study degradation mainly under processing and use conditions. Since biological degradation is minimal under these conditions, it only appears as a topic of great interest in the marine plastics community, which is concerned about the consequences of the end-of-life of plastic products.

We advocate for the complementarity of polymer science and marine plastics research and praise the efforts of publishers to bring different disciplines together, such as in this very journal. Our results indicate that there is a big opportunity for researchers working in environmental plastics to contribute to polymer science, considering the gap of biodegradation in polymer science literature, as a study on one may turn out to be influential on the other. Yet, for it to have a greater impact across fields, it must be aware of the vocabulary and significance of at least the most important polymer properties. In this work, we intended to give an overview of degradation processes and polymer properties that many marine scientists may not be fully aware of. Although not exhaustive, this basic dictionary aims to enrich the vocabulary and facilitate the connection between terms that may seem unrelated at first for researchers not trained as polymer scientists. We understand that without a thorough consideration of the polymer properties that arise from their composition and structure, the behaviour of plastic particles cannot be fully understood. Therefore, their fate and accumulation characteristics can be inappropriately defined.

NOTES AND REFERENCES

‡ Plastic degradation can occur without molecular changes, as in the case of additive desorption or water adsorption and contribute to loss of mechanical resistance, thus increasing the fragmentation rate. However, these were not discussed.

- 1 L. Simon-Sánchez, M. Grelaud, C. Lorenz, J. Garcia-Orellana, A. Vianello, F. Liu, J. Vollertsen and P. Ziveri, *Environ Sci Technol*, 2022, **56**, 16780–16788.

- 2 J. Zalasiewicz, C. N. Waters, J. A. Ivar do Sul, P. L. Corcoran, A. D. Barnosky, A. Cearreta, M. Edgeworth, A. Gałuszka, C. Jeandel, R. Leinfelder, J. R. McNeill, W. Steffen, C. Summerhayes, M. Wapreisch, M. Williams, A. P. Wolfe and Y. Yonan, *Anthropocene*, 2016, **13**, 4–17.
- 3 R. C. Thompson, S. H. Swan, C. J. Moore and F. S. vom Saal, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2009, **364**, 1973–1976.
- 4 GESAMP (Kershaw P.J., Turra A. and Galgani F. editors), Rep. Stud. GESAMP No. 99, 2019, 130p. GESAMP, Guidelines for the Monitoring and Assessment of Plastic Litter in the Ocean, 2019 .
- 5 K. L. Law, S. Morét-Ferguson, N. A. Maximenko, G. Proskurowski, E. E. Peacock, J. Hafner and C. M. Reddy, *Science* (1979), DOI:10.1126/science.1192321.
- 6 A. Cózar, M. Sanz-Martín, E. Martí, J. I. González-Gordillo, B. Ubeda, J. Á.gálvez, X. Irigoien and C. M. Duarte, *PLoS One*, DOI:10.1371/journal.pone.0121762.
- 7 J. A. Ivar do Sul, A. S. Tagg and M. Labrenz, *Scientometrics*, 2018, **117**, 2145–2157.
- 8 A. Cohan, S. Feldman, I. Beltagy, D. Downey, D. S. Weld and † Allen, arXiv:2004.07180, 2020.
- 9 S. S. Stivala and L. Reich, *Polymer Engineering and Science*, 1980, 20, No. 10: 654–661 *Structure vs Stability in Polymer Degradation*, .
- 10 P. R. Sundararajan, *Physical Properties of Polymers Handbook*, ed. J. E. Mark, Springer, Springer New York, NY, 2nd edn, 2007, Chain Structures, pp. 3-24 .
- 11 J. K. Jayasuriya Chandima Kumudinie and Premachandra, *Physical Properties of Polymers Handbook*, ed. J. E. Mark, Springer New York, New York, NY, 2007, pp. 25–40.
- 12 L. S. Reich and S. Stivala, *Journal of Pharmaceutical Sciences*, 1970, vol. 59, 1, pp. 131-132.

- 13 Y.-R. Luo, *Comprehensive Handbook of Chemical Bond Energies*, CRC Press, 2007.
- 14 J. E. Huheey, E. A. Keiter and R. L. Keiter, *Inorganic Chemistry: Principles of Structure and Reactivity*, Pearson Education, 4th edn., 1997.
- 15 J. B. Huang, G. S. Zeng, X. S. Li, X. C. Cheng and H. Tong, *IOP Conference Series: Earth and Environmental Science*, 2018, vol. 167.
- 16 A. A. Askadskii, M. Popova, T. Matseevich and E. Kurskaya, *Adv Mat Res*, 2014, **864–867**, 751–754.
- 17 W. L. Hawkins, W. Matreyek and F. H. Winslow, *Journal of Polymer Science*, 1959, **41**, 1–11.
- 18 G. Ayoub, F. Zaïri, C. Frédérix, J.M. Gloaguen, M. Naït-Abdelaziz, R. Seguela, J.M. Lefebvre, *International Journal of Plasticity*, 2011, Volume 27, Issue 4, Pages 492-511, ISSN 0749-6419 D. A. S. Ravens, *Polymer (Guildf)*, 1960, 1, 375–383.
- 19 N. P. Cheremisinoff, *Condensed Encyclopedia of Polymer Engineering Terms*, 2001, 39–81.
- 20 X. Zhu, Y. Zhou and D. Yan, *J Polym Sci B Polym Phys*, 2011, **49**, 1277–1286.
- 21 Timothy P. Lodge and Paul C. Hiemenz, in *Polymer Chemistry*, 3rd edn., 2020.
- 22 James E Mark, *Physical Properties of Polymers Handbook*, Springer, Springer New York, NY, 2nd edn, 2007.
- 23 D. Feldman, *Polymer Weathering: Photo-Oxidation*, 2002, vol. 10.
- 24 V. V. Krongauz, *Thermochim Acta*, 2010, **503–504**, 70–84.
- 25 S. Girois, L. Audouin, J. Verdu, P. Delprat and G. Marot, *Polym Degrad Stab*, 1996, **51**, 125–132.

- 26 P. Gijsman, *e-Polymers*, 2013, vol. 8,
1<https://doi.org/10.1515/epoly.2008.8.1.727> .
- 27 Samuel Leo madorsky, *Thermal Degradation of Organic Polymers*, John Wiley & Sons, Inc, 1st edn., 1964.
- 28 V. V Kholodovych and W. J. Welsh, in *Physical Properties of Polymers Handbook*, ed. E. M. James, Springer, Springer New York, NY, 2nd edn, 2007, *Thermal-Oxidative Stability and Degradation of Polymers*, pp. 927-938.
- 29 J. Izdebska, *Printing on Polymers: Fundamentals and Applications*, 2016, 353–370.
- 30 P. Gijsman, J. Hennekens and J. Vincent, *Polym Degrad Stab*, 1993, **42**, 95–105.
- 31 J. Vohlidal, *Chemistry Teacher International*, 2021, 3, 213–220.
- 32 S. J. Joshi, S. Kant Bhatia, K. University, S. Korea, W.-M. Wu, D. B. Levin, N. Mohanan, Z. Montazer and P. K. Sharma, , DOI:10.3389/fmicb.2020.580709.
- 33 J. Yuan, J. Ma, Y. Sun, T. Zhou, Y. Zhao and F. Yu, *Science of the Total Environment*, 2020, **715**, 136968.
- 34 T. Kijchavengkul and R. Auras, *Polym Int*, 2008, **57**, 793–804.
- 35 G. Kale, T. Kijchavengkul, R. Auras, M. Rubino, S. E. Selke and S. P. Singh, *Macromol Biosci*, 2007, 7, 255–277.
- 36 S. Toze and J. Sidhu, *Biosolids: Human Health Impacts*, .
- 37 A.-U.-R. Bacha, I. Nabi, M. Zaheer, W. Jin and L. Yang, *Science of The Total Environment*, 2023, **858**, 160108.
- 38 R. Egger and J. Yu, *Frontiers in Sociology*, , DOI:10.3389/fsoc.2022.886498.
- 39 E. Callaway, H. Ledford, G. Viglione, T. Watson and A. Witze, *Nature*, 2020, **588**, 550–552.
- 40 E. Gibney, *Nature*, 2020, **582**, 173–174.

APPENDIX A

Degradation changes the chemical structure of plastic particles and is expected to alter their behaviour in their environment. So, in order to predict future scenarios of plastic pollution and evaluate their impacts, we must understand how these changes occur and their effects on the physical chemistry of plastic particles (e.g., molecular weight, crystallinity and density). Being so diverse in composition and properties, each type of plastic can undergo different degradation mechanisms¹. In this Appendix, we provide: a brief overview of the mechanisms and chemical reactions involved in plastics' degradation; a short description of Topic Modeling and Text Mining and additional discussions on the results that support the conclusions stated in the main text.

Degradation Mechanisms

The reactions happening during degradation can be one of three main types: scission, elimination or depolymerization. Modes of degradation are strongly affected by the chemical structure and the presence of defects in the polymer chain, but also by the conditions of degradation^{1,2}. It should be noted that degradation reactions occur mainly at the surface of particles. Given that for a volume of material, the smaller the particles, the more important the ratio between surface area and volume, we should expect higher rates of degradation for smaller particles.

Scission happens when the local energy overcomes the bond energy, thus breaking it. Well illustrated by polyolefins, such as PE and PP, that are polymerized by the addition of monomers, scission can also be seen as the reverse reaction of polymerization. Breakage will happen randomly along the chain, forming two macroradicals (the two ends of the broken chain). They can either recombine or diffuse in the material. If they recombine, they will likely increase the mean molar mass by crosslinking, but it will reduce mechanical strength, nonetheless³. Otherwise, it will cause a sharp decline in the mean molar mass of the material. The result of chain scission is governed by the rate of diffusion of the macromolecules in the material, thus, by their temperature, degree of crystallinity and molar mass of the molecule. For instance, when scission occurs in a tertiary carbon of a lateral group, the resulting macromolecule is very small and is likely to diffuse in the material, propagating the reaction. Additionally, the presence of oxygen will determine the

result of the reaction. Those macroradicals are reactive species that will tend to react with O₂ molecules. This oxidation, however, is an auto-catalytic reaction, appropriately called auto-oxidation. That is because the hydroperoxides that are the products of these reactions continue to be reactive. Thus, the reaction will propagate until two macroradicals recombine or until they react with H₂O forming alcohol⁴.

Elimination, or substituent reactions, is a mode of degradation where the radical attached to the polymer backbone, otherwise called substituent, is removed by the breakage of the bond between the radical and the main chain. This is then followed by the break of the adjacent C-H bond and the formation of a double bond between both carbons. Since there is no breaking of bonds in the backbone, this kind of reaction doesn't lead to decreased molar mass⁵. However, a severe change in physicochemical properties is observed. It is most well represented by the degradation of PVC, which rapidly forms HCl. HCl is a highly toxic compound and also catalyses a chain scission reaction that eventually reduces molar mass^{6,7}. Macroscopically, the most evident effect is the strong formation of a reddish colouration. It also catalyzes the hydrolysis of PET, which may be detrimental in plastic recycling plants if these polymers are processed together.

The difference between the energies of the bonds in the backbone and in the radicals will largely determine if degradation will occur by elimination or by scission or depolymerization. Elimination generally ensues at energies smaller than those needed for backbone bond break. So, it will probably happen first.

Depolymerization is the process of turning a polymer into a monomer or mixture of monomers, which are the main products of this type of reaction⁸. Defects in the polymer chain, such as ether bonds formed during polymerization in the presence of O₂, and other weaker bonds, such as C-C adjacent to C=C double bonds, favour it. Therefore, it will start at these "weak links", where there will be a reduction of bond energy. It is disfavored by the presence of O₂, as it will produce oxidation of macroradicals. It will reduce the mean molar mass, but the effect is initially negligible. It is unlikely to happen in the environment by action of abiotic factors alone, as it tends to occur at high temperatures. For instance, depolymerization of PS only happens above 250°C. However, the presence of some reagents allows depolymerization at lower temperatures, as is the case for PET which can undergo depolymerization at around 70°C in the presence of ammonia or

strong acids ⁹. Equally, specific enzymes excreted by microorganisms can produce some degree of depolymerization at ambient temperature.

The presence of some enzymes and moisture causes the hydrolysis of polymer molecules. It consists of the reaction of a water molecule with the polymer chain that can cause depolymerization and chain scission. Hydrolysis can be accelerated in the presence of acids, alkalis, specific enzymes or higher temperatures. It is a common mechanism of digestion in organisms. The reactions of water with starch, protein and fat catalyzed by specific enzymes are responsible to transform food into small molecules that can be absorbed by cells in human intestines.

Besides these chemical reactions that occur during degradation, it is useful to understand what starts these reactions. In marine conditions, plastics are subject to a number of weathering factors. Among them, some will initiate degradation reactions, like UV, temperature and moisture. After initiation, reactions can propagate because of the formation of free radicals and be facilitated by primary degradation.

In this section, we briefly describe a few initiation mechanisms that the literature reports as the most likely to occur in marine environments. However, there are others, such as ozonolysis and mechanical degradation under high stress, that may not be relevant in the ocean. In the next section of this appendix, we discuss how weathering factors vary in intensity between environmental compartments.

Topic Modeling and Text Mining

The statistical analysis of documents can be traced at least to the 1960s¹⁰. However, methods and data changed significantly with the advent of the internet, the creation of large repositories of texts and new statistical techniques, more prominently, those related to machine learning.

Natural Language Processing (NLP) is an interdisciplinary field that associates linguistics and computer science. Naturally, it developed greatly with the advance of computer science, but even more from the more recent use of neural networks. NLP has found applications in sentiment analysis, text translation, large language models and smart assistants, text-to-image generators, etc. A fundamental operation in this field is to transform words into vectors, called word embedding¹¹, or even sentence embedding as in SBERT, which is able to derive semantically meaningful sentence embeddings. Neural networks have shown great results in the

generation of contextual word-sentence vectors, which allow the meaning of texts to be encoded¹². Topic Analysis allows clustering or classifying these encoded documents into distinct classes based on their vector representations¹³. Sample size should be taken into account when discussing results, as they provide an indication of the generalizability of the findings¹⁴.

More generally, Text Mining uses embedded words as data in a range of algorithms to discover information present in text or even to derive new information hidden in multiple sources. Topic Modelling is an unsupervised approach that differs from other rule-based text mining approaches that use regular expressions or dictionary-based keyword-searching techniques. Text mining is most commonly found in Information and Communication Technologies (ICT), probably due to the familiarity researchers in this field have with these tools. Yet, the frequency of its use in Environmental Science is also significant as it was shown to be useful in document classification and summarization, topic modelling and in identifying research trends. Authors have preferred Latent Dirichlet Allocation (LDA) for text mining applications, more commonly using only paper abstracts¹⁴. However, while LDA considers a document as a bag-of-words, a collection of vectors representing each word, to reveal latent topics in a set of documents, it disregards semantic relationships among words. To overcome this limitation, Bidirectional Encoder Representations from Transformers (BERT), a natural language processing technique developed by Google and published in 2018 was developed¹³. Another drawback of LDA is that it demands the number of topics to be set by the user previously, which was not known in our case.

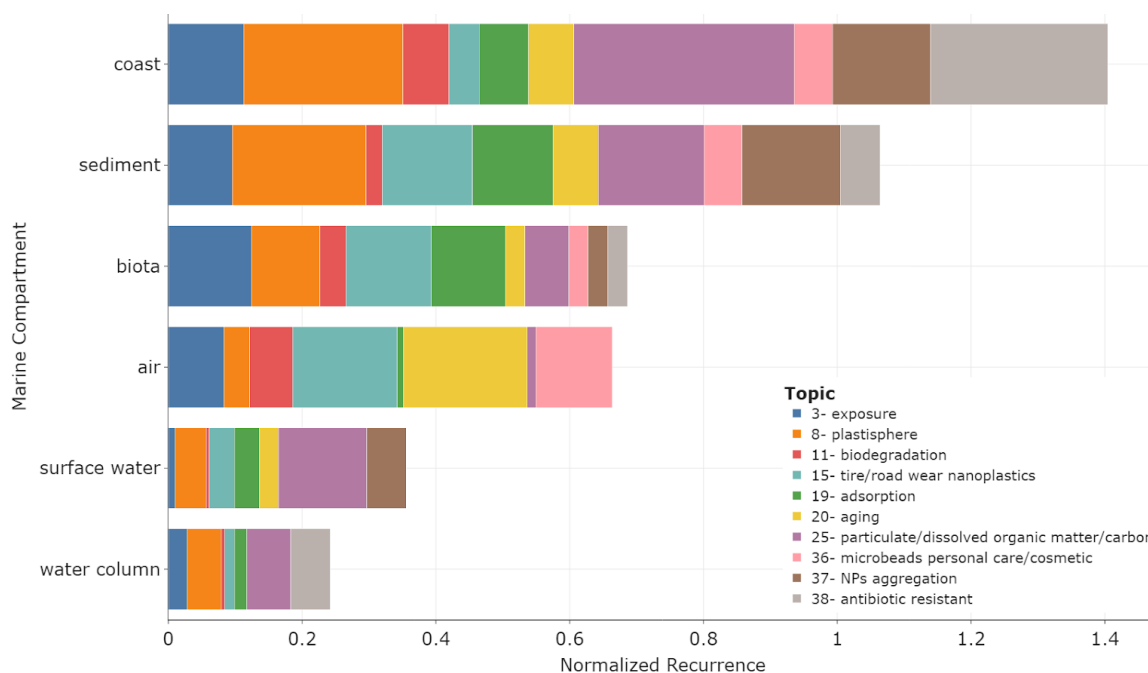
A limitation of this method resides in the length of the text data used. BERT models have a maximum input length of 512 tokens. If a word encoder is used, simple texts should contain around 512 words (more complex words may require up to 3-4 tokens to encode). If a sentence encoder is used, as was the case in this study, longer texts can be used^{13,15,16}. Considering this, the model is only capable of handling between 128 and 512 words. When the model reaches its maximum number of tokens, it is truncated and leaves out the remaining tokens. There are workarounds to this, such as chunking and summarization^{17,18}. The user must be aware of this when replicating our methods, but we don't expect this to have an impact on our results, as most abstracts never reach maximum tokens.

A few variations of BERT have been developed afterwards and can be used with satisfying results. Among them, BERTopic was chosen for this study for its ease of use and the vast supporting information available. It is a state-of-the-art algorithm for topic analysis that combines clustering techniques with TF-IDF to generate topic representations. The latter is a short term for “term frequency-inverse document frequency”, a statistical measure that evaluates the importance of words for a document in a collection of documents. The reasoning behind it is that the more common a word is across all documents, the lesser its importance for the current document. It works firstly by creating document embeddings using a pre-trained language model (usually MiniLM). Secondly, the dimensionality of the embeddings is reduced and clusters are defined (usually by HDBSCAN). Finally, TF-IDF is used to extract topic representations from each topic. Each of these steps is independent which allows flexible topic modelling, the choice of different tools for each step, and dynamic topic modelling over a time period. Nonetheless, BERTopic is an out-of-the-box tool, optimised for use without the need to fine-tune parameters.

Results

Each weathering factor has a different intensity depending on the marine compartment. To assess which compartments are being more frequently addressed, their recurrences were also checked and results are presented in Figure S1. For instance, the recurrence of the term **temperature** in topics 8, 11 and 20, suggests a relationship with the study of plastics that are exposed in the atmosphere or on the coastline. That is, compartments where the temperature factor exposure is higher. Therefore the term **organic pollutants** in topics such as exposure, adsorption, aging and dissolved organic matter suggest a relationship with the study of plastics that are exposed in the atmosphere, biota, and seabed.

Figure S1 - Normalized recurrence of terms associated with marine compartments given by topic



Source: Elaborated by the author

Considering mentions of the 6 marine compartments evaluated, the coast and sediment are the ones that stand out. This could be related to the accessibility of these compartments for research but could also be biased by the selected topics. The seawater compartment is considerably more difficult to access than the coastal region, for example. In the case of the air, this term was preferred over atmosphere as the latter revealed little to no mentions. However, results should be regarded with care, since plastics could be only exposed to the air rather than suspended in the air/atmosphere. Nevertheless, the greater recurrence of this word in the topics aging and tire/road wear nanoplastics is in good accordance with what would be expected for these two topics

When evaluating the recurrence of each topic, results demonstrate topics are more or less evenly studied across all 6 marine compartments. This indicates the interest of researchers in understanding particular topics across diverse compartments. We can point out some deviations from this pattern in topics such as NPs aggregation, antibiotic resistance and biodegradation. The latter, specifically, is much more present outside of the water, having low recurrences for compartments “surface water” and “water column”. This seems to be in accordance with the understanding that biodegradation occurs at even lower rates in water than in humid soil. Overall, the two compartments mentioned are less recurrent in the abstracts under these topics, which may indicate a lack of attention to degradation studies

under these conditions. However, here again the methodology can only offer an overview of the researchers interest considering mentions to specific terms. We reiterate that the construction of the vocabulary was made iteratively to optimize for the best terms used to refer to each compartment.

Supplementary References

- 1 F. W. Billmeyer, *Textbook of Polymer Science*, Wiley-Interscience: John Wiley & Sons, New York, 3rd edn., 1990.
- 2 H. R. Allcock and F. W. Lampe, *Contemporary Polymer Chemistry*, Prentice-Hall, Inc., Prentice-Hall, 3rd edn., 2003.
- 3 I. H. Craig, J. R. White, A. V. Shyichuk and I. Syrotynska, *Polym Eng Sci*, 2005, **45**, 579–587.
- 4 Jan. F. Rabek, *Polymer Photodegradation*, 1995, 1–23.
- 5 V. Kholodovych and W. J. Welsh, *Physical Properties of Polymers Handbook*, 2007, 927–938.
- 6 W. H. Starnes, *Prog Polym Sci*, 2002, **27**, 2133–2170.
- 7 J. Yu, L. Sun, C. Ma, Y. Qiao and H. Yao, *Waste Management*, 2016, **48**, 300–314.
- 8 V. Gold, Ed., *The IUPAC Compendium of Chemical Terminology*, International Union of Pure and Applied Chemistry (IUPAC), Research Triangle Park, NC, 2019.
- 9 A. B. Raheem, Z. Z. Noor, A. Hassan, M. K. Abd Hamid, S. A. Samsudin and A. H. Sabeen, *J Clean Prod*, 2019, **225**, 1052–1064.
- 10 F. Mosteller and D. L. Wallace,
<http://dx.doi.org/10.1080/01621459.1963.10500849>, 2012, **58**, 275–309.
- 11 T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, .
- 12 J. Devlin, M.-W. Chang, K. Lee, K. T. Google and A. I. Language, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, .

- 13 R. Egger and J. Yu, *Frontiers in Sociology*, , DOI:10.3389/fsoc.2022.886498.
- 14 K. Thakur and V. Kumar, <https://doi.org/10.1080/13614533.2021.1918190>, 2021, **28**, 279–302.
- 15 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, .
- 16 Q. Xie, Z. Dai, E. Hovy, M.-T. Luong and Q. V. Le, .
- 17 G. Erkan and D. R. Radev, *Journal Of Artificial Intelligence Research*, 2011, **22**, 457–479.
- 18 K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2014, 1724–1734.

5 TRAINING AND EVALUATING MACHINE LEARNING ALGORITHMS FOR OCEAN MICROPLASTICS CLASSIFICATION THROUGH VIBRATIONAL SPECTROSCOPY

Publicado em Janeiro de 2022 em Chemosphere

Henrique de Medeiros Back, Edson Cilos Vargas Junior, Orestes Estevam Alarcon, Daphiny Pottmaier

Universidade Federal de Santa Catarina, Brazil

ABSTRACT

Microplastics are contaminants of emerging concern - not only environmental, but also to human health. Characterizing them is of fundamental importance to evaluate their potential impacts and target specific actions aiming to reduce potential harming effects. This study extends the exploration of machine learning classification algorithms applied to FTIR spectra of microplastics collected at sea. A comparison of successful classification models was made in order to evaluate prediction performance for 13 classes of polymers. A rigorous methodology was applied using a pipeline scheme to avoid bias in the training and selection phases. The application of an oversampling technique also contributed by compensating unbalanceness in the dataset. The log-loss was used as the minimization function target and to assess performance. In our analysis, Support Vector Machine Classifier provides a good relationship between simplicity and performance, for a fast and useful automatic characterization of microplastics.

Keywords: Microplastics, Marine pollution, Chemical identification, Vibrational spectroscopy, FTIR, Machine learning

5.1 INTRODUCTION

Plastic debris is today found in virtually every habitat on earth (Free et al., 2014; Lebreton et al., 2018; Saito et al., 2018; Wang et al., 2019; Zhang et al., 2020). Yet, major scientific concern has been given to plastics at sea, where they are almost omnipresent and when found in large quantities and/or concentrations their negative impacts may be serious (Barnes, 2002; Chae and An, 2017). Originated mainly on

land as a result of large consumption of disposable items and poor waste management, plastics compose most marine debris (Galgani et al., 2015).

According to the International Organization for Standardization (ISO), microplastics are any solid plastic particle insoluble in water with any dimension from 1 μm to 1000 μm (1 mm) and large microplastics between 1 mm and 5 mm. (ISO/TR 21960:2020). They can originate from the degradation and fragmentation of larger plastic debris when exposed to environmental conditions or can be directly emitted in their form as, for example, microbeads, pellets, or textile fiber (Fendall and Sewell, 2009; GESAMP, 2019; Hidalgo-Ruz et al., 2012; Thompson et al., 2009). Their wide occurrence and physical characteristics, such as density and chemical composition, have contributed to consider these as emerging environmental contaminants (Sauvé and Desrosiers, 2014).

An adequate characterization of these contaminants can provide substantial information on the inputs and transport in the oceans, rates of degradation and fragmentation, interaction with biota, consequences of their presence in natural habitats, and, finally, risk assessment and management (Rocha-Santos and Duarte, 2015). Multiple parameters are relevant for these analyses, such as size, mass, sampling site, and DNA from the plastisphere (Zettler et al., 2013). Determination of the chemical composition of sampled plastics has been used to suppose where they originated from, what human use they might have had, and even estimate their age (Song et al., 2015; Turner and Holmes, 2011).

While diversity is wanted in the production and use phases of polymer products, it adds complexity to the identification process. For macroplastics, characterization may still be done by recycling codes in the product or by physical characterization. Yet, for microplastics, these procedures have been shown to result in high rates of false positives when not followed by any chemical analysis (GESAMP, 2019; Hidalgo-Ruz et al., 2012). Thus, a chemical characterization of microplastics is fundamental to more adequately assess the sources, fate, and impacts of microplastics.

Vibrational Spectroscopy, namely Raman and Fourier Transform Infrared Spectroscopy (FTIR), are established techniques to assess chemical composition (Käppler et al., 2016). Both analyses provide information on specific chemical bonds and functional groups, albeit by different methods (Kuptsov and Zhizhin, 1998). They allow differentiation between synthetic and natural polymers, identification of polymer

type, and degree of weathering, hence their wide utilization in microplastic identification methods (Andrady, 2017).

FTIR analysis can be done visually by an expert, but it is often more convenient to use a peak matching dedicated software with the available databases (Li et al., 2006). They perform what is called a library search, going through all spectra in the database and showing those with better ranking. This strategy allows a faster polymer identification, but since sampled microplastics are somehow dirty and weathered, peaks absent in the same non-weathered polymers can be seen in microplastics. (Melissa R. Jung et al., 2018; Xu et al., 2019). Therefore, their chemical signature (spectrum) can be significantly different from virgin polymers that commonly constitute databases (GESAMP, 2019; Primpke et al., 2018). Procedural factors can also influence the quality of measurements and impact identification accuracy and time. The size of identifiable microplastics, for example, is limited by the FTIR setup and should not be less than 500 μm (unless the spectrometer is coupled with a microscope) (Käppler et al., 2016; Shim et al., 2017). Weathered microplastics could be fragile and shatter while being manipulated, which also hinders spectrum acquisition (Shim et al., 2017).

A few alternatives have been proposed and tested to surpass these limitations. Pre-processing of the raw data, the usage of specifically designed databases, and dual database searches are some of the approaches that have been demonstrated to improve match results and allow library searches to be used with good confidence (Primpke et al., 2018, 2020; Renner et al., 2019).

Most notably, μ -FTIR and μ -Raman techniques are being used to address common limitations on microplastics analysis. Coupling a spectrometer and a microscope, allows the characterization of smaller particles and with an additional Focal Plane Array (FPA) detector, several particles can be characterized at once (Cabernard et al., 2018; Brandt et al., 2020). In this sense, these techniques can provide a more accurate picture of the evaluated microplastic environment. Yet, this kind of equipment may be less available mainly due to cost issues.

Improving methodologies for conventional vibrational spectroscopy could provide better alternatives to marine pollution researchers in places where financial support is insufficient, mainly the global south. Having said that, the proposed methodology could be used to evaluate the performance of classification algorithms using data from different spectroscopic techniques, especially FTIR imaging where

data volumes are significantly larger (~10⁶ spectra per image) and spectra quality tends to be lower (Primpke et al., 2017).

Sampling microplastics at sea produces hundreds or even thousands of individual particles, which by manual methodologies would take far too long to be characterized. Increased automation of the characterization process of microplastic samples with higher confidence in the results could quicken information acquisition on this emerging contaminant, filling multiple knowledge gaps and significantly advancing understanding on the field. Indeed, that is the main objective of the study presented in this paper. A machine learning pipeline was proposed for the selection of the best among a few machine learning algorithms to classify microplastics spectra, then discuss the main findings.

ML algorithms have recently been proposed as faster and more accurate methods to analyze spectra from marine microplastics and have been successfully used in chemometrics (and many other fields) for more than a decade (Conroy et al., 2005). Most efforts have been applied to imaging techniques, probably due to high computational demands associated with higher data volumes - thus, the necessity to improve efficiency (Hufnagl et al., 2019). Nevertheless, conventional spectroscopic techniques could benefit from these improvements.

In machine learning, classification algorithms are a category of prediction models used to attribute unidentified samples to a given class based on a set of variables. These algorithms have been applied by Hufnagl (Random Forest Classifier) and Kedzierski (K-nearest neighbors) to spectral data of microplastics and resulted in expressive classification accuracies (Hufnagl et al., 2019; Kedzierski et al., 2019). However, a comparison of different algorithms hasn't been made. Given the already demonstrated potential of these techniques, further investigation could lead to even better results. Since the learning process highly depends on the database available for training, comparable results should ideally derive from the same dataset. In the present study, data previously published by Kiedzierski were used for this purpose (Kedzierski et al., 2019).

Methodologies focusing on the selection of relevant attributes on microplastics spectra have been proposed by Renner (2017), Hufnagl et al., 2019, and da Silva et al., 2020. Indeed, it is the approach experts take when visually interpreting a spectrum. In this paper, a dimensionality reduction (DR) technique called Principal Component Analysis (PCA) was applied to extract the most relevant

features in the whole dataset. Some machine learning models lose performance when the data is represented in a high dimension space, falling into the so-called curse of dimensionality (Trunk, 1979). Feature selection or extraction is used to reduce the number of variables that describe a certain set of instances (samples) while retaining most of the information. This can, sometimes, reduce the time of implementation without significant loss of information and, when the curse of dimensionality is observed, even improve prediction performance.

The main disadvantage of the machine learning approach to this classification task is that it is limited to analyzing classes of polymers that are well represented in the training database. Since there is an enormous variety of plastics, setting up rather complete databases would be laborious. However, from all the known range of polymers, there are only a few commonly found as marine microplastics, making the setup of representative databases a lot more feasible. The method described in this paper goes further in the process of setting up this type of database by using an artificial oversampling technique to mitigate the effects of imbalanced datasets in the learning process, which can produce models that are biased towards the majority class.

5.2 MATERIALS AND METHODS

This study extends previous efforts by combining several machine learning techniques and comparing algorithms, to find, among them, the best method to automatically classify microplastic spectra. It differs from other approaches by proposing a pipeline methodology to train, evaluate and select classification models. The methodology described next was rigorously defined to justify choosing one model over all others. Once the model is chosen, it can be trained and called to make a prediction. A researcher interested in reproducing the experiment must follow the entire methodology, whereas a potential user of the proposed predictive tool can skip to section 2.9 for a short explanation on how it can be used.

All programming was made in a Core i5-7200U with 16 GB ram, using only 1 core of the processor to avoid issues related to parallel programming. We used Python programming language and some of its data analysis and machine learning libraries, namely: scikit-learn (Pedregosa et al., 2011), scipy (Virtanen et al., 2020), and imblearn (Lemaître et al., 2017). All programming is available in a supplementary file (Appendix D) so that the methodology described in this paper can be audited,

reproduced and even improved by peers. It is also available online at GitHub (https://github.com/EdsonCilos/mp_classification).

5.2.1 The dataset

The data used to train the algorithms were previously published by Kedzierski et al. (2019), and were generated from the Attenuated Total Reflection FTIR spectroscopy of samples collected during expeditions in the Mediterranean Sea. Spectra were recorded in absorbance mode in the range of 4000 to 600 cm^{-1} with 4 cm^{-1} resolution and 16 scans. The labeled dataset was constructed using the raw data consisting of 9581 spectra previously identified and assigned to 17 different classes.

5.2.2 Dataset pre-processing pipeline

Typically in the spectra evaluation by an expert, the analysis of some frequencies is not considered due to the presence of noise or just because the knowledge in the field prescribes that samples can be distinguished by some specific peaks. Our perspective, however, stands in the Machine Learning background, assigning to the algorithm the task to find the most successful pattern matching. As pointed by Vapnik: “In a wide philosophical sense, predictive models do not necessarily connect prediction of an event with an understanding of the law that governs the event; they are just looking for a function that explains the data best.” (Vapnik, 2006).

In other words, our goal is to look for a good explanation of the available data, and therefore, no feature selection based on expert knowledge was done, nor seemed to be necessary. As a consequence, all recorded wavelengths were used for the analysis. Additionally, a major drawback of a priori feature selection is that it may be suboptimal and should be revised when new classes are introduced, while PCA is embedded in the pipeline.

Since the original dataset was highly imbalanced, the first step in pre-processing consisted of renaming the classes that were underrepresented in the

dataset, assigning them to a generic class called “unknown”. Samples from the 4 least represented classes, namely, Polyurethane, Animal fiber-like, Poly(vinyl chloride), and acrylic (PMMA) were moved to the unknown class, effectively removing 4 classes. The latter having only 3 samples, while the others only 1 sample, making it impossible to train a model using the methodology adopted in this paper. The dataset then consisted of 958 samples from 14 different classes (13 microplastic classes + unknown).

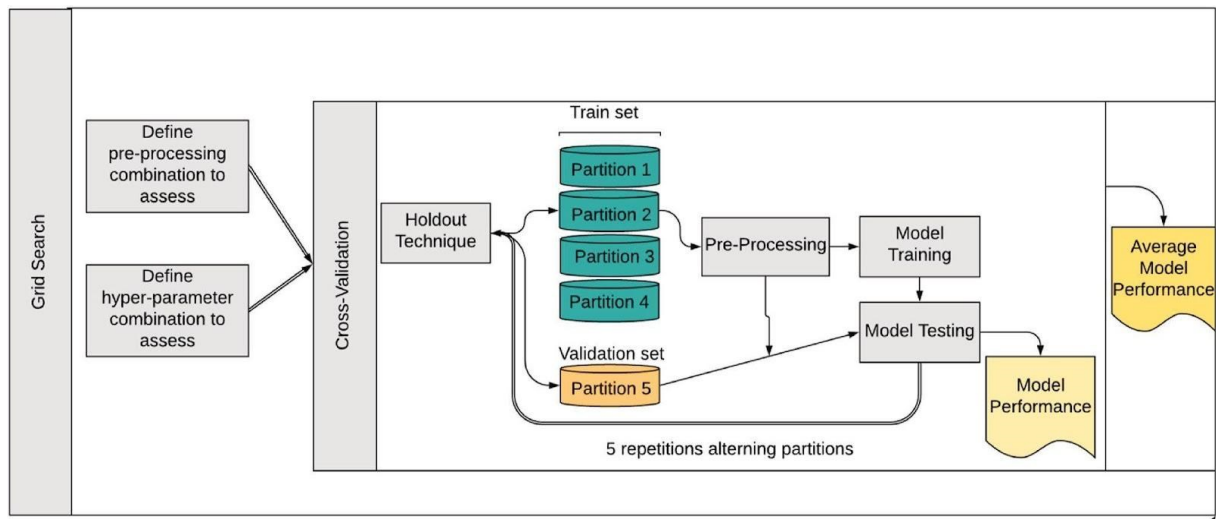
It is important to note that apart from relabeling underrepresented classes and correcting the baseline, all preprocessing steps must be done after the train/test split, to avoid the so-called “snooping bias/data leakage” (see “Results and Discussion” for further details). It is a common mistake when coding machine learning algorithms. It happens when some information that typically would be available only in subsequent steps (like in a “production environment”), is introduced in the learning process.

The remaining pre-processing steps were assembled in a pipeline, which we will call, for short, a Pre-pipeline (Later on in Section 2.6 we shall discuss the full Pipeline of the methodology, which includes the Pre-pipeline). All steps were tested in the “turn on/turn off” configuration, to evaluate which of those was able to improve the final model's performance. It consisted of the application of the following techniques in the raw data (in this order):

- A. Baseline correction using Asymmetric Least Squares (Eilers and Boelens, 2005). Parameters were: λ (2nd derivative constraint) = $1e5$; p (weighting deviations) = 0.05; itermax (number of iterations to perform) = 10
- B. Standard scaler (Z-transform), since the features are arranged at different scales, which could affect the model performance (Géron, 2017);
- C. Principal component analysis (PCA), as a method to verify if higher dimensions could impact the algorithm's performance (Jolliffe, 2002). PCA is an unsupervised learning technique that creates a new linear space with orthogonal variables, called principal components (PCs), which are the directions of most variance in the dataset. If the original features are uncorrelated, the new space contains the same number of dimensions as the original space, but, generally, a much smaller number of variables is sufficient to describe the data without much loss of information. In our analysis, the first n PCs containing 99% of the explained variance were used.

D. Oversampling to mitigate the effects of imbalance in the dataset. Despite having assembled underrepresented classes, the dataset was still highly imbalanced, as the 2 most populated classes contained 45% of all samples in the dataset. This can cause distortions in the learning process, as the algorithm would be more likely to assign a new sample to these classes (Prati et al., 2009). This was done using the Imbalanced Learning API inside a Pipeline scheme (Fig. 1, Fig. 3). It generates new instances by randomly sampling the under-represented classes, effectively copying existing spectra (Lemaître et al., 2017).

Figure 1: Graphical representation of the pipeline used with Grid Search for assessing multiple classifiers.



Source: (Author)

The four pre-processing techniques described (Baseline correction, Standard Scaler, PCA, and Oversampling) were subjected to analysis to evaluate if including them would improve the performances of the classifiers. Several combinations were tried for each classifier (e.g. kNN without PCA, with Standard Scaler, and with oversampling). The best combination was forwarded to another trial to examine if a frequency smoothing strategy could improve the performance likely by suppressing features that arise from non-ideal instrument and sample conditions (Renner et al., 2019; Zimmermann and Kohler, 2013). Ergo, a Savitzky-Golay filter was included before the standard scaler step, in which an evaluation of 115 combinations of windows and degrees was performed. This step was done separately in order to speed the evaluation, considering it already took 4 days.

5.2.3 The holdout strategy

Initially, the dataset was split into training and test sets (holdout strategy), stratifying proportionally to the number of classes, to make sure every class is present both in the train and test sets. The test set ratio choice was 25%. To allow reproducibility, we fixed the seed, as well as all random states, equal to 0. It is worth mentioning that the authors used the holdout strategy to be able to evaluate, under mathematical assumptions, an “unbiased”² estimate of the generalization error. The holdout strategy allows the trained algorithm to be tested on an “external” dataset, completely unknown to the algorithm. This can be done using mathematical tools like Hoeffding's inequality (Hoeffding, 1963), Vapnik-Chervonenkis theory (Vapnik, 2000) or usual statistical inference techniques. Thereby, a unique and final evaluation in the test set is capable of providing a confidence interval for the proposed algorithm, as we shall see in the Results/Discussion section.

5.2.4 The classification algorithms

According to the “No Free Lunch Theorem” (Wolpert, 1996): there is no a priori reason to prefer one machine learning model over another without making any assumption about the data. Therefore, in practice, we must select some models and test them on the problem that we aim to solve. For this reason, we chose some of the most popular and relevant models in machine learning: Decision Tree (DT) (Moore, 1987; Murphy, 2012, p. 544), Gaussian Naive Bayes (GNB) (Murphy, 2012, p. 82), k-Nearest Neighbour (kNN) (Nordhausen, 2009, p. 14), Random Forest (RF) (Nordhausen, 2009, p. 587), Logistic Regression (LR) (Murphy, 2012, p. 14; Nordhausen, 2009, p. 119), Classifier (SVC) (Murphy, 2012, p. 496; Nordhausen, 2009, p. 417), and Neural Networks (Goodfellow et al., 2016).

Each model has general characteristics that may be more suited to one application than the other (Classifier comparison,). A comparison of 5 of the models proposed here is presented in Table 1, adapted from Kotsiantis (2007). For instance, GNB has a high speed of learning and classification but is highly affected by interdependent attributes. SVC, on the other hand, has a high speed of classification and is very tolerant to irrelevant attributes, but is slow to learn and falls short in explainability (Kotsiantis, 2007).

Table 1. A comparison of models' characteristics (Adapted from Kotsiantis, 2007).

	Decision Trees	Neural Networks	Naive Bayes	kNN	SVC
Accuracy in general	**	***	*	**	****
Speed of learning	***	*	****	****	*
Speed of classification	****	****	****	*	****
Tolerance to missing values	***	*	****	*	**
Tolerance to irrelevant attributes	***	*	**	**	****
Tolerance to redundant attributes	**	**	*	**	***
Tolerance to highly interdepende nt attributes	**	***	*	*	***
Dealing with discrete/binar y/continuous attributes	****	***	***	***	**
Tolerance to noise	**	**	***	*	**
Dealing with danger of overfitting	**	*	***	***	**
Attempts for incremental learning	**	***	****	****	**
Explanation ability	****	*	****	**	*
Model parameter handling	***	*	****	***	*

Source:(Author)

The algorithms have been configured to return probabilities (provided by Sklearn; sklearn. calibration) rather than just a deterministic output (sklearn.calibration.CalibratedClassifierCV, 2021). This seems to be suitable for the current task, allowing the researcher to be more confident when the probability score

is higher. In this context, the cross-entropy (or log-loss) of a multinoulli distribution was chosen as an objective function to be minimized, to make the probability distribution of the model as close as possible to the empirical distribution (see Appendix B for further details) (Goodfellow et al., 2016).

5.2.5 Gridsearch

Every machine learning model has a set of parameters that must be predefined by the user and are not learned during training. These are called “hyperparameters”. Grid Search is an efficient tool to find the best combination of hyperparameters, given a predefined grid.

All combinations of hyperparameters can be found in the project's file (Appendix D) (refer to `param_grid.py`). Just to mention a few, Logistic Regression, for example, was tested using diverse penalty schemes (l1, l2 and Elastic Net with different l1 ratios ranging from 0.1, 0.2, ..., 0.9), different regularization parameter C ($C = 10^j$ for $j = -2, -1, \dots, 3$) and several solvers depending on the penalty (newton-cg, sag, saga, lbfgs and/or liblinear).

In a cross-validation we split the data in k folds (or partitions), for each fold we train the model in the remaining data (all data except the fold) and evaluate the model in the fold; in the end the k results are averaged producing a score. Such a strategy was used to evaluate the grid search, in a 5-fold cross-validation scheme (Fig. 1). We fixed the seed that generated the partition in the cross-validation, therefore a reproduction of this experiment is likely to achieve the same result.

Throughout this section whenever we refer to a “training set” we are referring to a new training set created by cross-validation or another validation mechanism. Otherwise, when referring to the initial training set, we will refer to it as the “original training set”.

Cross-validation was performed within Grid Search to avoid overfitting on the validation set - because these hyperparameters can be tuned optimally for a specific validation set. That is, for a certain combination of hyperparameters, 5 different training and test sets are used to train and test the model's performance. The final

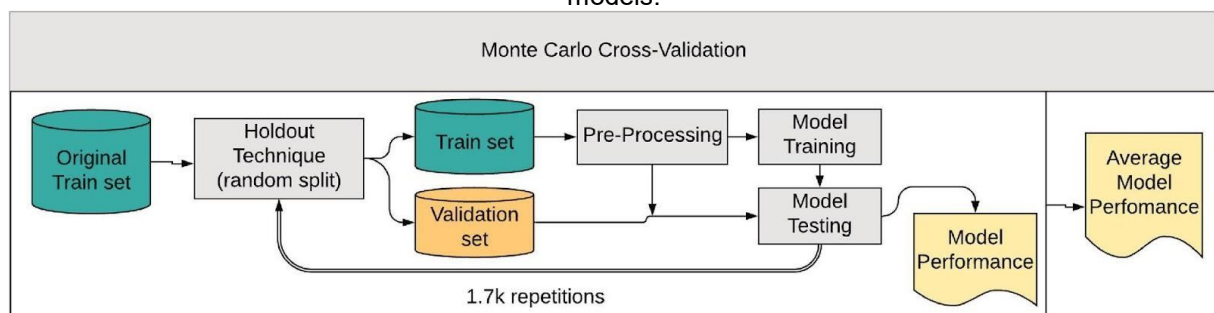
result for that specific combination is the average of these tests (standard deviations were also computed).

5.2.6 Monte Carlo Cross-validation

After the GridSearch, we looked for the best scores. Decision Tree e Gaussian Naive Bayes classifiers performed very poorly. Among the remaining models, we selected the most promising configuration in each case: the top 2 Neural Networks (with less or more neurons), top 2 (linear kernel and 'rbf' kernel), the best Random Forest, the best kNN and the best Logistic Regression. These models were subjected to subsequent analysis considering a Monte Carlo Cross-Validation (MCCV) (briefly described next) (Xu and Liang, 2001). The remaining classifiers showed a considerably higher log-loss and/or its parameters were unlikely to improve the performance; therefore, they were not forwarded to the next step. The selected model will be referred to as “the final hypothesis”.

An MCCV consists in randomly splitting the data in train and validation sets, several times. It's a kind of holdout strategy with many trials, also known as “repeated learning-test”. This technique allows us to draw a “monte carlo picture” and check a model's performance in a histogram, allowing us to compute some statistics like mean and standard variation. In our case, the MCCV consisted of 1700 trials (see Fig. 2). Since the grid search was done with a fixed seed, it can produce models that perform well in rare cases. By introducing randomness through MCCV, we can make a second filter in the models, in an attempt to avoid “rare best performing models” (Burman, 1989).

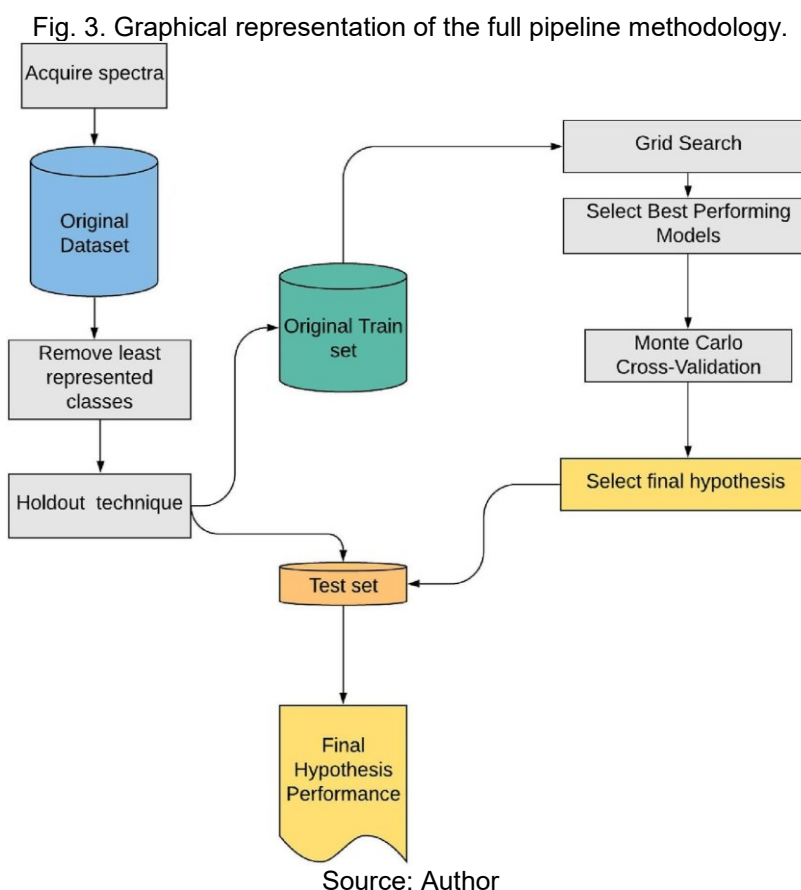
Figure 2: Graphical representation of the pipeline used in the Monte Carlo Cross-Validation. After 1700 repetitions the results are averaged. The process was repeated on the seven most promising models.



Source: Author

5.2.7 The Pipeline

The application of the tools described until now was done by means of a pipeline methodology. In total, 878 models under 5 combinations of pre-processing techniques were subjected to the same workflow. Thus, using a pipeline was useful to automate the sequence of operations that we proposed (Fig. 3).



One important step to mention is the pre-processing of the data, which happens after the train/test split, something that is commonly overlooked, but that avoids the snooping bias. Notice, as well, that all steps concerning the selection of the best model are done using the “original training set”, not the “original dataset”. Keeping part of the original dataset aside, we can use it in the end to test the selected model on a dataset completely new to the algorithm.

We suggest this pipeline to be used to evaluate and compare the performances of different models in the task of microplastics classification with ATR-FTIR spectroscopy. It can be equally used with different datasets, containing, for example, other polymer classes.

5.2.8 Comparability and External Validation

In order to be able to compare our methodology with previous works, we applied, to the final hypothesis, the training methodology described by Kedzierski et al. (2019) (which consisted essentially of a MCCV with 1000 slightly different splits) with three differences: Firstly, we used a stratified split. This is important because our dataset is highly unbalanced with some classes underrepresented, therefore it is possible that in certain splits some classes remain only in the training set (more probable) or only in the test set (less probable), which can, in any case, introduce bias in our model. Secondly, Kedzierski et al. assign all unclear predictions (less than 3 votes in kNN) to the “unknown” class, which the present methodology does not. Finally, and most importantly, it seems that the authors included a standardization pre-process in the entire dataset, which introduces a “snooping” bias in the model. This means, for example, that the model “already knows” the mean and standard deviation of the features of the validation set before even being trained. On the other hand, our pre-processing methodology is built inside a pipeline that uses, for example, the standardization learned in the training set and applies it in the validation set, which avoids such bias.

A further verification was included to prove that the approach proposed here is useable outside the Kedzierski dataset. To do so, a different environmental dataset was used to train and evaluate the final hypothesis, which was done by MCCV, as described in section 2.6. There was no tuning of hyperparameters at this step. The dataset contained 800 FTIR spectra of polymers ingested by turtles and was published in a previous study by M.R Jung et al. (2018). There were initially 9 classes, including a differentiation between high and low density poly(ethylene). As previously, the less representative classes were suppressed and spectra moved to the “unknown” class, which resulted in 5 classes.

5.2.9 Application of the classifier

If the reader is particularly interested in using the proposed final hypothesis (classifier), they can do so by running the code “final_classifier.py” in a python console, which is located in the root folder of the project repository.

It should be stressed that the data provided for classification should not have been pre-processed at all, as the appropriate pre-processing will be performed by the algorithm considering the results presented in the next section of this paper.

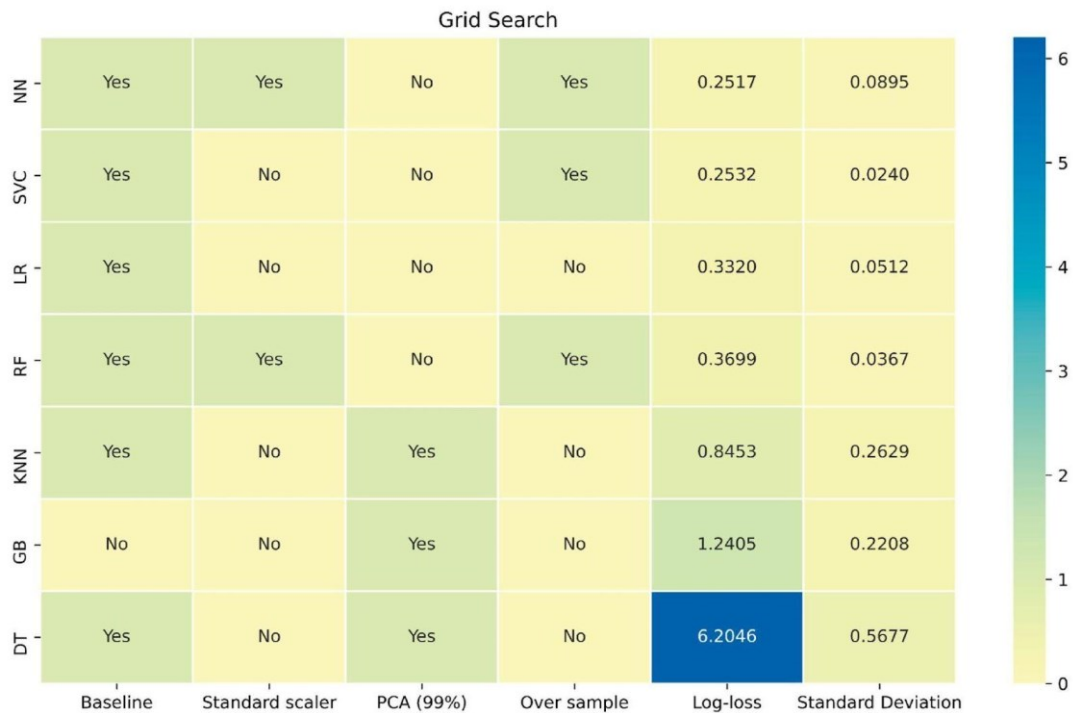
Additionally, the spectral data should have been collected at the same frequency range described in section 2.1.

Given the limitation of the classifier to recognize only the classes it was trained on, the classification result will be given as a probability of that sample belonging to the assigned class. If the user is then unsure of the assignment, it can be visually confirmed. An example of this application and output is given in Appendix C.

5.3 RESULTS/DISCUSSION

Using the methodology described in the previous section, 28,096 models were evaluated in the 5-fold cross-validation. Results (log-loss and standard deviation) for the best combination of hyperparameters and pre-processing techniques for each different model are presented in Fig. 4.

Figure 4: Best performance found within the gridsearch for each evaluated algorithm, namely Random Forests (RF), Support Vector Machine Classifier (SVC), Logistic Regression (LR), K-Nearest Neighbors (kNN), Decision Trees (DT), Gaussian Naive Bayes (GB), considering the best combination of hyperparameters and pre-processing techniques - the baseline correction, Standard scaler, Principal Component analysis (PCA) and (Naive) oversampling. The log-loss shown here is the mean of the five results (cross-validation). The correspondent standard deviation is also presented.



Source: Author

For each simulation, a more detailed table can be found in Appendix D (paths can be found in Appendix A). Analyzing the GridSearch, the seven most promising

models, based on the proposed metrics, were: LR with baseline correction; Sigmoid NN with baseline correction, standard scaler, and oversampling; Sigmoid NN with baseline correction and standard scaler; SVC with linear kernel, baseline correction, and oversampling; SVC with rbf kernel, baseline correction, and oversampling; RF with baseline correction, standard scaler, and oversampling; kNN with baseline correction and PCA. The reader can refer to the “gridsearch” subfolder in the project repository for further details.

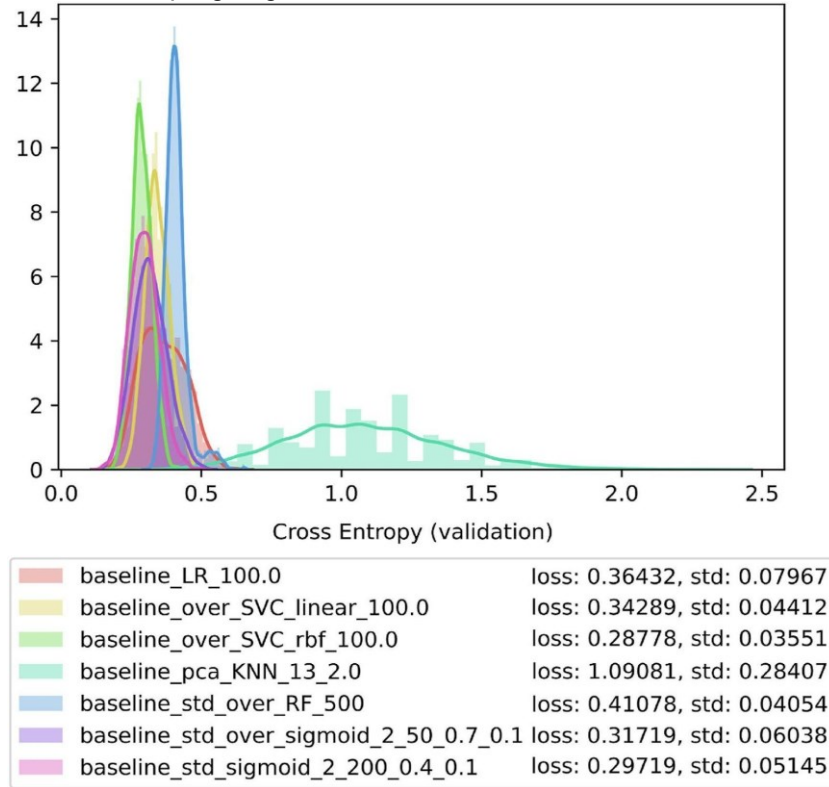
The pre-processing combination that improved the performance was different for each model. For instance, SVC did not benefit from PCA, while kNN, DT, and GB did.

When comparing several models with and without PCA, most of them present little scoring difference when PCA is applied. Meanwhile, we verified significant differences in the learning time. For example, baseline + oversample took around 5 h to run, whereas baseline + pca + oversample took about 17 min. It should be noticed, however, that this is not a detailed picture of which algorithm could benefited from PCA, but an experimental evidence to affirm that future analyses, mainly when they involve the comparison of several algorithms, can be done with PCA (99%) without significant loss of information but with a saving time bonus. All grid search and MCCV (next) simulations have their time computed and stored in the file “time.csv” in the “results” project's folder.

Previous results suggested the use of peak smoothing techniques over the spectra (da Silva et al., 2020; Zimmermann et al., 2013). In our case, however, a Savitzky-Golay filter did not improve the performance of any model. For this reason, the procedure was not included in the project. Such a phenomenon probably happens because smoothing peaks favours experts’ visual inspection, while in a pattern recognition this is likely to be irrelevant.

The next step was to subject the most promising models to an MCCV in order to validate the consistency of their performances when randomness is allowed. The resulting histogram of each model is presented in Fig. 5.

Figure 5: Cross-entropy log-loss for the seven most promising models: Logistic Regression with baseline correction; SVC with linear kernel, baseline correction, and oversampling; SVC with rbf kernel, baseline correction, and oversampling; kNN with baseline correction and PCA; RF with baseline correction, standard scaler, and oversampling; Sigmoid NN with baseline correction, standard scaler, and oversampling; Sigmoid NN with baseline correction and standard scaler.



Source: Author

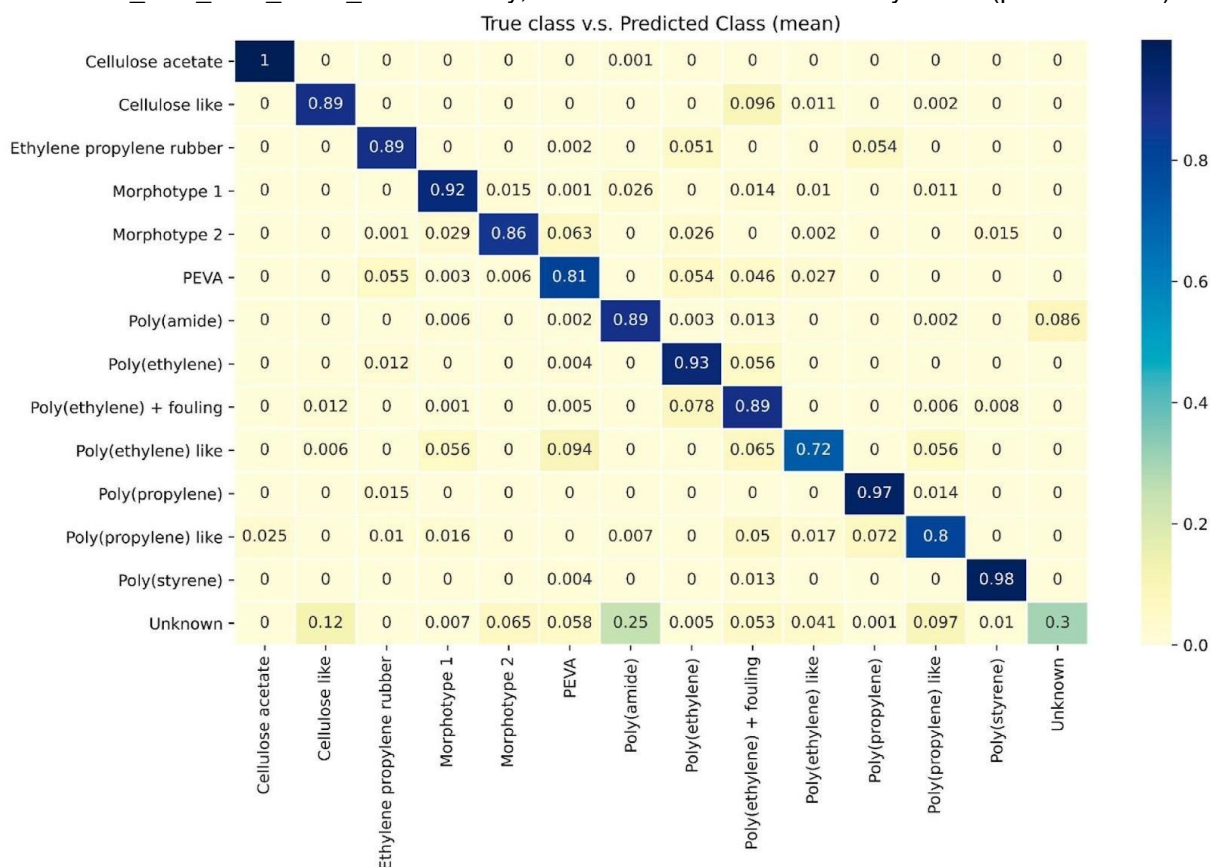
Overall, the seven models had a decrease in performance (increased log-loss) but it could be said that they are, in fact, robust against randomness. Even though the results indicate that kNN and RF are the worst models, only by looking at these results it is not possible to choose the best model. To extend the analysis, the authors examined their accuracies and the results were very similar between models, so no further conclusion could be made. Considering the methodology proposed by (Renner et al., 2019), which was also used by (Kedzierski et al., 2019), we evaluated, for each model, how many classes were “working well” (sensitivity greater than 75%) and equal conclusions were possible. Therefore, since none of these metrics provided clear boundaries for a final decision, we shall take into account Occam's Razor, which states that “the simplest model that fits the data is also the most plausible” (Abu-Mostafa et al., 2012, p. 167). Appropriately, SVC with a linear kernel was chosen as the final hypothesis. The reader can refer to the “results/graphics” subfolder in the project repository for further graphs in this matter.

More complex, non-linear models, such as neural networks, can be tuned to fit the data well. Nonetheless, if the data were collected under different circumstances, or other pre-processing techniques or parameters were changed, this would cause minor changes in the final performance of every model. The neural networks, for instance, would probably still perform well, but there would be many more hyper-parameters to tune again in order to obtain optimal results. Linear SVC has only one such parameter (C) that could be readily tuned in case new reference samples or a whole new training set was provided. Having considered this, SVC is expected to perform efficiently with low tuning and prediction times.

Additionally, regarding the similarity in the performance of distinct models under various criteria, it is the authors' opinion that the "ideal theoretical structure" of the data is indeed linearly separable. By "ideal theoretical structure" of the data, we mean: no error in data labeling, materials without wear, and all micro-plastic polymers well represented, thus we conjecture that under FTIR the microplastics exhibits an almost linear pattern, with linearity suffering small violations due to the presence of these "non-ideal elements". Another important remark is that our model improved performance when combined with oversample technique, agreeing with the fact that in an imbalanced problem, a SVC classifier can produce suboptimal models that are biased toward the majority class (He, H., Ma, Y., 2013, p. 83). Lastly, Linear SVC can be implemented using the LIBLINEAR library, which is capable of handling large data sets (Fan et al., 2008). Therefore, we expect that the model could be implemented in a massive dataset, in which we expect similar results.

Since the methodology describes supervised learning procedures, it can be expected to observe where the algorithm makes mistakes when tested on a validation set. This is possible by looking at the confusion matrix shown in Fig. 6, generated by averaging results obtained for every split in the MCCV for the selected model.

Figure 6: Confusion Matrix for the results of MCCV for the final hypothesis. Each row represents the true class, meanwhile each column is the class predicted proportion, on average, by baseline_over_SVC_linear_100.0. Ideally, the matrix should be an identity matrix (perfect match).



Source: Author

A confusion matrix compares the true class in the validation set with the class predicted by the algorithm for the same sample (Ballabio et al., 2018). Considering this, it can be observed, for instance, that samples from the “Cellulose Acetate” class were predicted to belong to their true class every time, whereas for the “Cellulose Like” class, 9.6% of all predictions were incorrectly assigned to the “Poly(ethylene) + fouling” class and 1.1% were incorrectly assigned to the “Poly(ethylene) like” class. The same kind of error was observed by Kedzierski et al. (2019) using a kNN classifier. Since the algorithm is trained to distinguish between classes by identifying patterns in the provided dataset, which only contains information on spectral features and polymer classes, this could indicate similarities between spectra of both classes. As pertinently pointed out by the referred authors, spectral bands associated with aging and biofouling of microplastics could be the cause, with a preponderance of the latter. A visual examination of spectra from the “Poly(ethylene) + fouling” and “Cellulose Like” classes was done to check for specific bands related to these changes in their molecular fingerprint. However, it was not possible to visually

establish any specific relations that could be causing the algorithm to mistake both classes, as spectra had too many overlapping peaks. We emphasize that the results are promising, nonetheless.

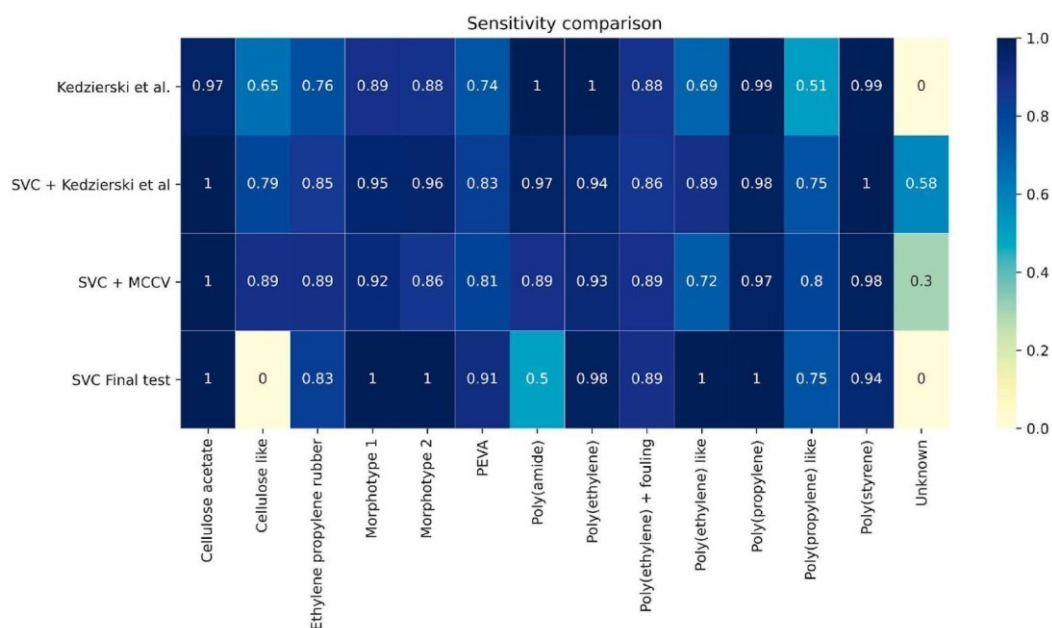
It may be clearer to depict faults made by the algorithm by observing the “Ethylene Propylene Rubber” (EPR) class, where the algorithm wrongly predicted the “Poly(ethylene)” (PE) and “Poly(propylene)” (PP) classes. This is comprehensible since EPR is made of the same monomers as PE and PP and share their chemical characteristics, hence, their spectral features. For classes such as “Poly(ethylene) like”, “Poly(ethylene) + fouling” and “Poly(propylene) like”, the main error made by the algorithm was to assign samples truly belonging to these classes to their regular counterparts (i.e. “Poly(ethylene)” and “Poly(propylene)”).

In the case of “Poly(propylene) like”, which had one of the worst results, most incorrect predictions (7,2%) were made to the “Poly(propylene) class. Since they were originally assigned to these classes by an expert due to their spectral similarities, this could be expected - and, in fact, was also observed by Kedzierski et al. (2019). However, despite having similar spectra, the algorithm was able to differentiate them with considerable accuracy, showing better class results than the k-nearest neighbour (k-NN) method originally proposed by the referred authors. Further research on the quality of the data can respond more assertively if errors of the algorithm are indeed mistakes or indicative of a human error in labelling the data. The class “Morphotype 2” was mistaken mostly with “Poly(amide)”, however, it returned excellent results meaning this grouping is probably very concise and should contain samples from remarkably similar polymers, or even, mostly samples from a single polymer type. It is also unlikely that this polymer is one of the others already in this database.

Concerning the oversample methodology, adding new samples in the distribution of the training set can “impose non-uniform error costs, causing the learner to be biased in favour of predicting the rare class” (He and Ma, 2013, p. 37). As we can see in Fig. 6, the proposed model does not suffer from this problem, since the less representative classes are not “stealing” sensitivity of the more representative ones, like (PE) or (PP). This indicates, within the results shown, that the model properly handled the unbalanced dataset problem.

After these considerations, the sensitivity calculated from results shown in Fig. 6 is presented in Fig. 7, along with results from a previous paper and the final test results.

Figure 7: Sensitivity for every class in the dataset according to results published by Kedzierski et al. (2019) using a k-Nearest Neighbour algorithm, Support Vector Machine Classifier (SVC) in the methodology proposed by Kedzierski et al. (2019), SVC in the Monte Carlo Cross-Validation and SVC in the final test.



Source: Author

Assuming that the dataset is statistically representative, the Final test results are what could be approximately expected of the performance of the algorithm in “real” conditions, that is, with the application of the deployed model to non-labeled or unknown spectra. Contrarily to the results given by the MCCV and the methodology adapted from Kedzierski et al. (2019) (refer to section 2.7. for a comparison between methodologies), which used the entire dataset for training, having test samples used for training even if in different splits/simulations, the test set in this case was completely new to the algorithm, giving an “unbiased” prediction.

All methodologies had comparable sensitivities, with those from Kedzierski et al. (2019) being lower for some classes (Fig. 7). Nonetheless, the referred author had shown that conventional machine learning algorithms could be powerful tools for classifying microplastics spectra. Interestingly, classes that did not perform as well with the kNN algorithm, like “PEVA”, “Poly(ethylene) like” and Poly(propylene) like” also had worse sensitivities with SVC in the MCCV, but presented a slight improvement. In general, this comparison indicates that Support Vector Machine

Classifiers offer more appropriate classification boundaries for this specific task with even better scalability, since kNN algorithms are highly computationally demanding with large datasets (Gutierrez et al., 2016).

The “Cellulose like” class presented no sensitivity, however. Among 240 samples, 4 of them belong to this class, therefore this undesirable event is plausible to happen and should not be interpreted as the actual performance for that class. In this regard, we expect that more samples in this class could improve the model performance.

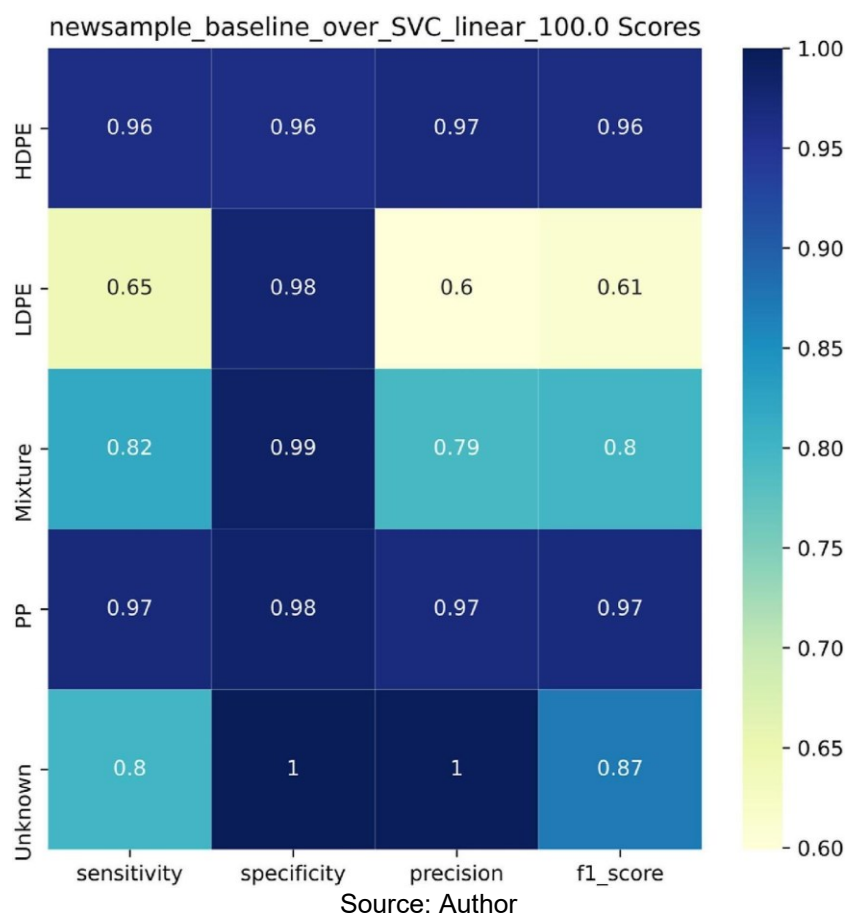
Assuming the data is large enough, independent and identically distributed (i.i.d), the overall accuracy (non class-specific) of the model (which can be given with a 95% probability) is approximately $91.25\% \pm 3.6\%$ (Mitchell, 1997, p. 132, formula 5.1). The condition “large enough data” ensures that the sample mean will be approximately normal, due to the Central Limit Theorem (Mitchell, 1997, p. 142). A more conservative confidence interval without such assumption can be computed by using Hoeffding's Inequality (Hoeffding, 1963), which can establish the following bound for the generalization accuracy (Abu-Mostafa et al., 2012, p. 40):

$$|A_{out}(g) - A_{in}(g)| = |E_{out}(g) - E_{in}(g)| \leq \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}},$$

with probability greater than where is the number of final models to be evaluated in the test set, is the generalization accuracy of the model , is the in-sample accuracy, similarly the generalization error (the ratio of misclassified labels) and the in-sample error. With and = 0.05 and assuming the data is i. i.d, the confidence interval is $91.25\% \pm 8.8\%$ with 95% of probability.

In regard to the model tested on the Jung dataset, a detailed score is presented in Fig. 8. The lower sensitivities for “LDPE” and “Mixture” are attributed to errors similar to those found in the confusion matrix in Fig. 6.

Figure 8: Detailed score for the MCCV applied to the Jung dataset considering five classes: High-density Poly(ethylene) (HDPE), Low-density Poly(ethylene) (LDPE), a mixture of PE and PP, Poly(propylene) (PP) and Unknown.



The model's cross-entropy log loss was 0.24554 (with a standard deviation of 0.0662) and the accuracy was 94% (with a standard deviation of 0.01209). This result is in accordance with those from the Kedzierski dataset in the MCCV (shown in Fig. 5), although their comparison is not straightforward since the number of classes is different. The excellent performance of the Linear SVC classifier on a different dataset is another evidence of the linear separability of MP spectra.

5.4 CONCLUSION

The present study was able to demonstrate the performance of different machine learning classification algorithms to the classification of ocean microplastics using previously identified samples' ATR-FTIR spectra as input data. For this purpose, we proposed log-loss to measure models' multi-class probabilities, an approach that has not been used previously in machine learning methods for microplastics characterization. We presented a machine learning combination not yet proposed for classification of microplastics, choosing linear SVC as the final classifier after thoroughly evaluating multiple conventional classification algorithms. The

rigorous pipeline methodology described in this paper is essential to avoid introducing bias in the model training and evaluation, which supports the selection of the algorithm and substantiates the performance obtained.

The best pre-processing of the raw spectra was also evaluated for each classifier. We identified that for the selected model, only a baseline correction and a naive oversampling was more effective.

Linear SVC is well suited for scaling up. After being trained, the classifier can be directly deployed and applied to classify unknown spectra. The user doesn't have to upload the database or train the SVC. Given a spectra, the proposed model would return a probability score for each class, rather than simply a deterministic output. This procedure seems to be more realistic with practice, since the researcher may have greater or lesser confidence in the resulting model's evaluation, depending on the probabilities returned.

This study attests for the use of this methodology applied to ATR-FTIR data. In many cases, other vibrational spectroscopic techniques may have been used, namely Raman and microspectroscopy. Despite the new challenges these techniques impose, such as particle morphology, moisture, and blank spectra, in the case of an FPA detector, due to the similarities between them, similar results could be expected. This hypothesis could be verified by applying the same evaluation methodology described in this paper.

Machine learning models work within the learned database, which means that the database must be representative of the required task. Thus, it should be applied with caution to spectra collected under different circumstances. Beyond that, the major limitation of the learning approach is that the final model cannot predict samples for a class it does not know, or doesn't know well, which could restrict the full deployment of this procedure, seeing that polymers can be of many kinds.

The disproportionate amount of samples in some classes in the dataset was addressed in this study, but not solved. Regardless of the method adopted, "artificially balancing the training distribution may help with the effects of class imbalance, but does not remove the underlying problem" (He and Ma, 2013). In other words, the most appropriate solution to the problem needs to be solved through some specialized algorithm, which does not seem to be the case, since the statistical support will be missing anyway, or through the acquisition of new data, which seems to be the critical point to solve the issue of the unbalanced dataset and even improve

the results in this and future works. Data sharing is necessary for the improvement of machine learning algorithms as it increases the size and diversity of training databases. Nonetheless, the authors came across some obstacles, such as: different file formats, huge files with lots of redundant metadata, missing class labels on identified spectra and data that were already pre-processed. We recommend the publication of raw spectra, as the prior pre-processing may introduce bias. The JCAMP-DX file format is a standard defined by the International Union of Pure and Applied Chemistry (IUPAC) and is considered the optimal format for sharing spectroscopy data. Also, providing the polymer class labels of every spectra should reduce redundant work and speed the collective learning process.

Since identification is a critical step in the study of these contaminants, improving confidence and speeding up the process are crucial to the advancement of the area. Despite not yet being able to stand alone as a method to automatically classify every conceivable sample, given its limitations, this study presents robust statistics to support the utilization of machine learning methods to the problem of automatic classification of microplastics.

¹ Comparing with Kedzierski et al. (2019) dataset, we deleted 12 spectra, since they were not available as raw data. Spectras related to Cellulose Acetate are also unavailable in the raw data, but we requested to Kedzierski, which gently provided us the data.

² Quotes were used here because, despite every effort to minimize bias, machine learning models cannot really be unbiased.

³ As long as this massive dataset adequately represents the distribution of MP classes in a way that the imbalance is less deleterious. Even in a scenario with little data, Linear SVC without oversample is still highly competitive.

APPENDIX A. PROGRAMMING AND RESULTS DETAILS

All programming is available in the project's repository at GitHub: https://github.com/edsonjunior14/mp_classification.

Below, the reader will find a table indicating the location of specific parts of the methodology and results.

The reader can also collaborate on the project via GitHub.

DESCRIPTION	FILE NAME	FOLDER
ORIGINAL DATA	<i>D4_REBUILD</i>	<i>DATA</i>
X, Y, TRAIN/TEST SPLIT	MULTIPLE (FILE NAMES ARE DESCRIPTIVE)	<i>DATA</i>
LOG-LOSS AND ACCURACY FOR EACH MODEL IN EVERY SPLIT IN MCCV	<i>TOTAL_SCORE.CSV</i>	<i>RESULTS / MCCV/ "MODEL NAME"</i>
SENSITIVITY, SPECIFICITY, PRECISION, F1_SCORE FOR EVERY SPLIT IN MCCV	<i>DETAILED_SCORE.CSV</i>	<i>RESULTS / MCCV/ "MODEL NAME"</i>
CLASS PROBABILITIES FOR EVERY SPLIT IN MCCV	<i>PROBABILITY.CSV</i>	<i>RESULTS / MCCV/ "MODEL NAME"</i>
DETAILED RESULTS FOR THE FINAL HYPOTHESIS ON THE EXTERNAL ENVIRONMENTAL DATASET (THE JUNG DATASET)	<i>TOTAL_SCORE.CSV;</i> <i>DETAILED_SCORE.CSV;</i> <i>PROBABILITY.CSV</i>	<i>RESULTS /</i> <i>MCCV_NEWSAMPLE /</i> <i>BASELINE_SVC_LINEAR_1</i> <i>00.00</i>
DETAILED SCORE OF SVC WITH METHODOLOGY PROPOSED BY KEDZIERSKI ET AL. 2019	<i>FINAL_MODEL_MCCV_ALL_DATA_DETAILED_SCORE.PNG</i>	<i>RESULTS</i>
DETAILED SCORE OF THE FINAL TEST	<i>FINAL_TEST_DETAILED_SCORE.PNG</i>	<i>RESULTS</i>
DETAILED SCORE OF SVC IN MCCV	<i>SVC_LINEAR_0.002_DETAILED_SCORE.PNG</i>	<i>RESULTS</i>

OVERALL LOG-LOSS AND ACCURACY OF THE FINAL TEST	<i>FINAL_TEST_TOTAL_SCORE.CSV</i>	<i>RESULTS</i>
OVERALL LOG-LOSS AND ACCURACY OF SVC WITH METHODOLOGY PROPOSED BY KEDZIERSKI ET AL. 2019	<i>FINAL_MODEL_MCCV_ALL_DATA_TOTAL_SCORE.PNG</i>	<i>RESULTS</i>
DETAILED GRID SEARCH RESULTS FOR EVERY COMBINATION OF PRE-PROCESSING	<i>"PRE-PROCESSING COMBINATION".CSV</i>	<i>RESULTS / GRID_SEARCH</i>
CODE FOR GENERATING GRAPHS AND TABLES	<i>GRAPHICS.PY</i>	<i>MP_CLASSIFICATION</i>
CODE FOR THE GRID SEARCH STEP	<i>GRID_SEARCH.PY</i>	<i>MP_CLASSIFICATION</i>
CODE FOR THE MCCV STEP	<i>MCCV.PY</i>	<i>MP_CLASSIFICATION</i>
CODE FOR DEFINITION OF THE PIPELINE, FINAL TESTING, AND FINAL HYPOTHESIS	<i>MODEL_EVALUATION.PY</i>	<i>MP_CLASSIFICATION</i>
PLOT AN UNIDENTIFIED SPECTRUM OVER THE AVERAGE SPECTRUM OF THE PREDICTED CLASS	<i>SPECTRAL_PLOT.PY</i>	<i>MP_CLASSIFICATION</i>
CODE FOR TREATING THE JUNG DATASET, APPLICATION OF MCCV, AND EVALUATION	<i>EVALUATE_NEW-SAMPLE.PY</i>	<i>MP_CLASSIFICATION</i>

APPENDIX B. LOG-LOSS: MAXIMUM LIKELIHOOD ESTIMATION

Let n be the sample size, K the number of classes and y_k representing the label of sample k . Furthermore, consider $p_{k,j}$ if $j = y_k$, and zero otherwise, that is, $p_{k,j}$ when the label of sample k is j , for $j \in \{1, \dots, K\}$. Additionally, let $\hat{y}_k$ where $\hat{y}_k$ is the label predicted by a classifier parametrized by a parameter θ (possibly multivariate). In short, $p_{k,j}$ is the probability of our classifier predict the sample k as a member of class j .

Therefore, given a classifier that provides class probabilities, assuming that samples are i.i.d, we can model the probability of achieving the correct label as the likelihood of a multinoulli distribution (categorical distribution) (Bishop, 2006):

Since we are interested on maximizing this likelihood, it's equivalent to minimize its cross-entropy loss (Friedman et al., 2001). Even though we explicitly wrote the cross-entropy loss function for the likelihood of a categorical distribution, it should be noticed that the term "cross-entropy" has a broader meaning:

"Any loss consisting of a negative log-likelihood is a cross-entropy between the empirical distribution defined by the training set and the probability distribution defined by model" (Goodfellow, 2016).

In summary, by minimizing the cross-entropy we attempt to make the probability distribution of the model as close as possible to the empirical distribution.

Appendix B References

BISHOP, C.M., 2006. Pattern recognition and machine learning. New York: Springer-Verlag.

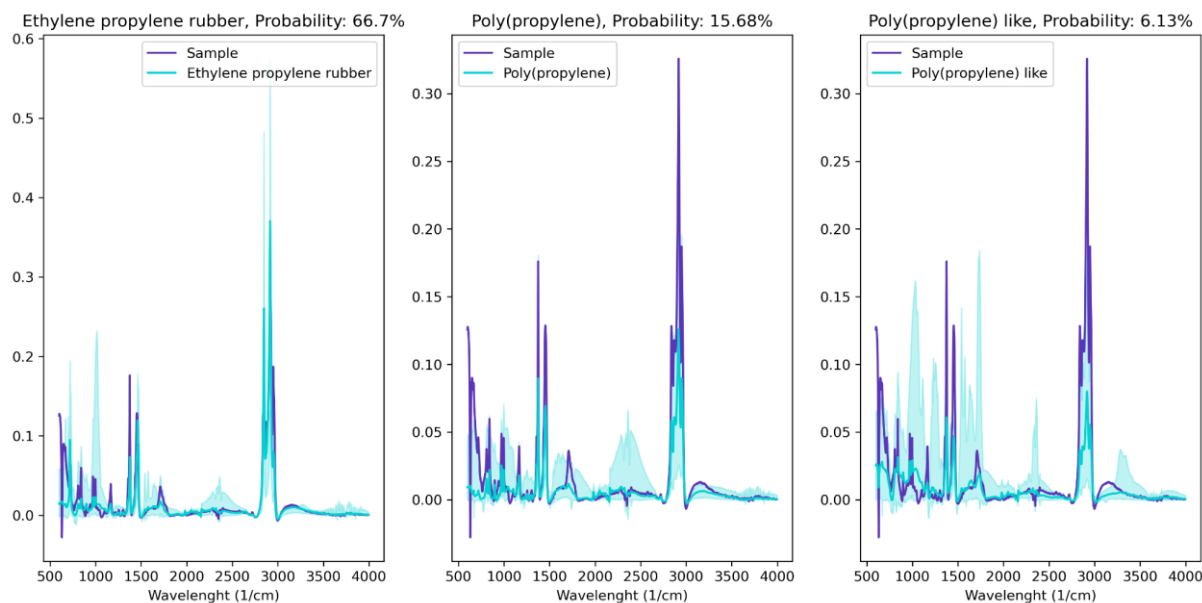
GOODFELLOW, I., BENGIO, Y., COURVILLE, A., 2016. Deep Learning. MIT Press, page 130.

HASTIE, T. TIBSHIRANI, R., FRIEDMAN, J., 2009. The elements of statistical learning. New York: Springer-Verlag.

APPENDIX C. USING THE SUPPORT VECTOR MACHINE CLASSIFIER

After having been selected and evaluated, the SVC model was trained on the entire dataset, and saved. A script (available in the project's repository) was run to use it to classify unknown MP samples. The authors used samples from the Kedzierski et al. (2019) dataset that had not been classified to show how the results of the model are presented. Figure 8 illustrates with an example the comparison of a spectrum and probable polymer classes.

Figure 9 - Each graph shows a comparison between the unknown sample spectrum and the predicted class showing the probability calculated by the model for the sample to belong to that class. The background curve area in light blue displays the full range of known samples in the predicted class, while darker blue shows the average spectrum of the class and the purple curve shows the spectrum of the unknown sample.



Source: prepared by the author

The user of the model can display a comparison between the sample and the predicted classes. This allows the user to be more confident in the results when the assigned class probability is lower. If used by an expert, possible errors are rapidly corrected by a visual examination. This is not expected to be necessary in most cases, however, as the model has shown to perform very well for most classes.

REFERENCES

- Abu-Mostafa, Y.S., Magdon-Ismael, M., Lin, H.-T., 2012. Learning from Data.
- Andrady, A.L., 2017. The plastic in microplastics: a review. *Mar. Pollut. Bull.* 119, 12–22. <https://doi.org/10.1016/J.MARPOLBUL.2017.01.082>.
- Ballabio, D., Grisoni, F., Todeschini, R., 2018. Multivariate comparison of classification performance measures. *Chemometr. Intell. Lab. Syst.* 174, 33–44.
- Barnes, D.K.A., 2002. Biodiversity: invasions by marine life on plastic debris. *Nature* 416. <https://doi.org/10.1038/416808a>.
- Brandt, J., Bittrich, L., Fischer, F., et al., 2020. High-Throughput analyses of microplastic samples using fourier transform infrared and Raman spectrometry. *Appl. Spectrosc.* 74 (9), 1185–1197. <https://doi.org/10.1177/0003702820932926>.
- Burman, P., 1989. A comparative study of ordinary cross-validation, v-fold crossvalidation and the repeated learning-testing methods. *Biometrika* 76. <https://doi.org/10.1093/biomet/76.3.503>.
- Cabernard, L., Roscher, L., Lorenz, C., Gerdts, G., Primpke, S., 2018. Comparison of Raman and fourier transform infrared spectroscopy for the quantification of microplastics in the aquatic environment. *Environ. Sci. Technol.* 52 (22), 13279–13288. <https://doi.org/10.1021/acs.est.8b03438>, 2018.
- Chae, Y., An, Y.J., 2017. Effects of micro- and nanoplastics on aquatic ecosystems: current research trends and perspectives. *Mar. Pollut. Bull.* 124 <https://doi.org/10.1016/j.marpolbul.2017.01.070>.
- [WWW Document], n.d. Classifier comparison https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html.
- Conroy, J., Ryder, A.G., Leger, M.N., Hennessey, K., Madden, M.G., 2005. Qualitative and quantitative analysis of chlorinated solvents using Raman

- spectroscopy and machine learning. *Proc. SPIE 5826, Opto-Ireland 2005: Optical Sensing and Spectroscopy*. <https://doi.org/10.1117/12.605056>.
- da Silva, V.H., Murphy, F., Amigo, J.M., Stedmon, C., Strand, J., 2020. Classification and quantification of microplastics (<100 μm) using a focal plane array-fourier transform infrared imaging system and machine learning. *Anal. Chem.* 92 (20), 13724–13733.
- Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C., 2008. LIBLINEAR: a library for large linear classification. *J. Mach. Learn. Res.* 9, 1871–1874.
- Fendall, L.S., Sewell, M.A., 2009. Contributing to marine pollution by washing your face: microplastics in facial cleansers. *Mar. Pollut. Bull.* 58 <https://doi.org/10.1016/j.marpolbul.2009.04.025>.
- Free, C.M., Jensen, O.P., Mason, S.A., Eriksen, M., Williamson, N.J., 2014. High-levels of microplastic pollution in a large , remote , mountain lake. *Mar. Pollut. Bull.* 85, 156–163. <https://doi.org/10.1016/j.marpolbul.2014.06.001>.
- Galgani, F., Hanke, G., Maes, T., 2015. Global distribution, composition and abundance of marine litter. *Marine Anthropogenic Litter*. https://doi.org/10.1007/978-3-319-16510-3_2.
- Géron, A., 2017. *Hands-on Machine Learning with Scikit-Learn and TensorFlow*. O'Reilly Media.
- GESAMP, 2019. *Guidelines for the Monitoring and Assessment of Plastic Litter in the Ocean*.
- Goodfellow, I., Bengio, Y., Courville, A., 2016. *Deep Learning*. MIT Press.
- Gutierrez, P., Lastra, M., Bacardit, J., Benitez, J., Herrera, F., 2016. GPU-SME-kNN: scalable and memory efficient kNN and lazy learning using GPUs. *Inf. Sci.* 373, 165–182, 2016.
- He, H., Ma, Y., 2013. *Imbalanced Learning: Foundations, Algorithms, and Applications*. <https://doi.org/10.1002/9781118646106>.

- Hidalgo-Ruz, V., Gutow, L., Thompson, R.C., Thiel, M., 2012. Microplastics in the marine environment: a review of the methods used for identification and quantification. *Environ. Sci. Technol.* 46, 3060–3075. <https://doi.org/10.1021/es2031505>.
- Hoeffding, W., 1963. Probability inequalities for sums of bounded random variables. *J. Am. Stat. Assoc.* 58 (301), 13–30. <https://doi.org/10.1080/01621459.1963.10500830>.
- Hufnagl, B., Steiner, D., Renner, E., Loder, M.G.J., Laforsch, C., Lohninger, H., 2019. A methodology for the fast identification and monitoring of microplastics in environmental samples using random decision forest classifiers. *Analytical Methods* 11, 2277–2285. <https://doi.org/10.1039/C9AY00252A>.
- Jolliffe, I.T., 2002. *Principal Component Analysis*. Springer.
- Jung, M.R., Balazs, G.H., Work, T.M., Todd Jones, T., Orski, S.V., Viviana Rodriguez, C., Beers, K.L., Brignac, K.C., David Hyrenbach, K., Jensen, B.A., Lynch, J.M., 2018. Polymer identification of plastic debris ingested by pelagic-phase sea turtles in the central pacific. *Environ. Sci. Technol.* 52 (20), 11535–11544. <https://doi.org/10.1021/acs.est.8b03118>.
- Jung, M.R., Horgen, F.D., Orski, S.v., Rodriguez, C.V., Beers, K.L., Balazs, G.H., Jones, T.T., Work, T.M., Brignac, K.C., Royer, S.J., Hyrenbach, K.D., Jensen, B.A., Lynch, J.M., 2018. Validation of ATR FT-IR to identify polymers of plastic marine debris, including those ingested by marine organisms. *Mar. Pollut. Bull.* 127, 704–716. <https://doi.org/10.1016/j.marpolbul.2017.12.061>.
- Kappler, A., Fischer, D., Oberbeckmann, S., Schernewski, G., Labrenz, M., Eichhorn, K.-J., Voit, B., 2016. Analysis of environmental microplastics by vibrational microspectroscopy: FTIR, Raman or both? *Anal. Bioanal. Chem.* 408, 8377–8391. <https://doi.org/10.1007/s00216-016-9956-3>.
- Kedzierski, M., Falcou-Préfol, M., Kerros, M.E., Henry, M., Pedrotti, M.L., Bruzaud, S., 2019. A machine learning algorithm for high throughput identification of

FTIR spectra. Chemosphere 234, 242–251.
<https://doi.org/10.1016/j.chemosphere.2019.05.113>.

Kotsiantis, S.B., 2007. Supervised machine learning: a review of classification techniques. *Informatica* 31 (3), 249–268.

Kuptsov, a. h., Zhizhin, g. n., 1998. Handbook of Fourier Transform Raman and Infrared Spectra of Polymers. Bioseparation.

Lebreton, L., Slat, B., Ferrari, F., Aitken, J., Marthouse, R., Hajbane, S., 2018. Evidence that the great pacific garbage patch is rapidly accumulating plastic. *Sci. Rep.* 1–15. <https://doi.org/10.1038/s41598-018-22939-w>.

Lemaître, G., Nogueira, F., Aridas, C.K., 2017. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* 18.

Li, J., Hibbert, D.B., Fuller, S., Vaughn, G., 2006. A comparative study of point-to-point algorithms for matching spectra. *Chemometr. Intell. Lab. Syst.* 82 <https://doi.org/10.1016/j.chemolab.2005.05.015>.

Mitchell, T.M., 1997. Machine Learning. McGraw-Hill Science.

Moore, D.H., 1987. Classification and Regression Trees, by Leo Breiman, Jerome H.Friedman, Richard A. Olshen, and Charles J. Stone, 1984. Brooks/Cole Publishing, Monterey, p. 358. <https://doi.org/10.1002/cyto.990080516>. Cytometry 8.

Murphy, K.P., 2012. Machine Learning: a Probabilistic Perspective (Adaptive Computation and Machine Learning Series). Mit Press.

Nordhausen, K., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., vol. 77. https://doi.org/10.1111/j.1751-5823.2009.00095_18.x by Trevor Hastie, Robert Tibshirani, Jerome Friedman. International Statistical Review.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J.,

- Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12.
- Prati, R.C., Batista, G.E.A.P.A., Monard, M.C., 2009. Data mining with unbalanced class distributions: concepts and methods. In: *Proceedings of the 4th Indian International Conference on Artificial Intelligence. IICAI, 2009.*
- Primpke, S., Lorenz, C., Gerdts, G., 2017. An automated approach for microplastics analysis using focal plane array (FPA) FTIR microscopy and image analysis. *Analytical Methods* 1499–1511. <https://doi.org/10.1039/c6ay02476a>.
- Primpke, S., Wirth, M., Lorenz, C., Gerdts, G., 2018. Reference database design for the automated analysis of microplastic samples based on Fourier transform infrared (FTIR) spectroscopy. *Anal. Bioanal. Chem.* 5131–5141.
- Primpke, S., Cross, R.K., Mintenig, S.M., Simon, M., Vianello, A., Gerdts, G., Vollertsen, J., 2020. Toward the systematic identification of microplastics in the environment: evaluation of a new independent software tool (siMPle) for spectroscopic analysis. *Appl. Spectrosc.* 74 (9), 1127–1138.
- Renner, G., Nellessen, A., Schwieters, A., Wenzel, M., Schmidt, T.C., Schram, J., 2019. Data preprocessing & evaluation used in the microplastics identification process : a critical review & practical guide. *Trends Anal. Chem.* 111, 229–238. <https://doi.org/10.1016/j.trac.2018.12.004>.
- Rocha-Santos, T., Duarte, A.C., 2015. A critical overview of the analytical approaches to the occurrence, the fate and the behavior of microplastics in the environment. *Trac. Trends Anal. Chem.* 65, 47–53. <https://doi.org/10.1016/j.trac.2014.10.011>.
- Saito, H., Fletcher, R., Yogi, T., Kayo, M., Miyagi, S., Ogido, M., Fujikura, K., 2018. Human footprint in the abyss : 30 year records of deep-sea plastic debris. *Sanae Chiba. Mar. Pol.* 96, 204–212. <https://doi.org/10.1016/j.marpol.2018.03.022>.
- Sauvé, S., Desrosiers, M., 2014. A review of what is an emerging contaminant. *Chem. Cent. J.* 8, 1–7. <https://doi.org/10.1186/1752-153X-8-15>.

- Shim, W.J., Hong, S.H., Eo, S.E., 2017. Identification methods in microplastic analysis: a review. *Analytical Methods* 9, 1384–1391. <https://doi.org/10.1039/c6ay02558g>. sklearn.calibration.CalibratedClassifierCV, 2021. viewed 08 June. <https://scikit-learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html#sklearn.calibration.CalibratedClassifierCV>.
- Song, Y.K., Hong, S.H., Jang, M., Han, G.M., Shim, W.J., 2015. Occurrence and distribution of microplastics in the sea surface microlayer in Jinhae bay, South Korea. *Arch. Environ. Contam. Toxicol.* 69 <https://doi.org/10.1007/s00244-015-0209-9>.
- Thompson, R.C., Swan, S.H., Moore, C.J., vom Saal, F.S., 2009. Our plastic age. *Philosophical Transactions of the Royal Society B: Biological Sciences*. <https://doi.org/10.1098/rstb.2009.0054>.
- Trunk, G. v, 1979. A problem of dimensionality: a simple example given. *IEEE Trans. Pattern Anal. Mach. Intell.* 306–307.
- Turner, A., Holmes, L., 2011. Occurrence, distribution and characteristics of beached plastic production pellets on the island of Malta (central Mediterranean). *Mar. Pollut. Bull.* 62 <https://doi.org/10.1016/j.marpolbul.2010.09.027>.
- Vapnik, V., 2006. Estimation of dependencies based on empirical data. *Empirical Inference*.
- Vapnik, V., 2000. *The Nature of Statistical Learning*. Information Science and Statistics. Springer-Verlag, New York.
- Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, I., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A.P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C.N., Fulton, C.,

- Masson, C., Haggstrom, C., Fitzgerald, C., Nicholson, D.A., Hagen, D.R., Pasechnik, D.v., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G.A., Ingold, G.L., Allen, G.E., Lee, G.R., Audren, H., Probst, I., Dietrich, J.P., Silterra, J., Webber, J.T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schonberger, J.L., de Miranda Cardoso, J.V., Reimer, J., Harrington, J., Rodríguez, J.L.C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N.J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P.A., Lee, P., McGibbon, R.T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T.J., Robitaille, T.P., Spura, T., Jones, T.R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y.O., Vázquez-Baeza, Y., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 17. <https://doi.org/10.1038/s41592-019-0686-2>.
- Wang, J., Liu, X., Li, Y., Powell, T., Wang, X., Wang, G., Zhang, P., 2019. Microplastics as contaminants in the soil environment : a mini-review. *Sci. Total Environ.* 691, 848–857. <https://doi.org/10.1016/j.scitotenv.2019.07.209>.
- Wolpert, D.H., 1996. The lack of A priori distinctions between learning algorithms. *Neural Comput.* 8 (7), 1341–1390.
- Xu, Q.S., Liang, Y.Z., 2001. Monte Carlo cross-validation. *Chemometr. Intell. Lab. Syst.* 56 [https://doi.org/10.1016/S0169-7439\(00\)00122-2](https://doi.org/10.1016/S0169-7439(00)00122-2).
- Zettler, E.R., Mincer, T.J., Amaral-Zettler, L.A., 2013. Life in the “plastisphere”: microbial communities on plastic marine debris. *Environ. Sci. Technol.* 47 <https://doi.org/10.1021/es401288x>.
- Zhang, Y., Kang, S., Allen, S., Allen, D., Gao, T., Sillanpaa, M., 2020. Atmospheric microplastics: a review on current status and perspectives. *Earth Sci. Rev.* 203, 103–118. <https://doi.org/10.1016/j.earscirev.2020.103118>.
- Zimmermann, B., Kohler, A., 2013. Optimizing savitzky-golay parameters for improving spectral resolution and quantification in infrared spectroscopy. *Appl. Spectrosc.* 67, 892–902. <https://doi.org/10.1366/12-06723>.

6 SÍNTESE

É sensato reconhecer os muitos benefícios sociais, econômicos e até ambientais da utilização generalizada de polímeros na produção de bens de consumo. Com uma história intimamente ligada ao desenvolvimento técnico desde a pré-história, com os polímeros naturais, essa classe de materiais tem permitido ganhos de eficiência em muitas áreas e tem permitido o acesso a muitos bens de consumo por classes econômicas mais baixas com a produção e fornecimento de produtos a baixo custo. Permitiu também a redução da predação sobre o ambiente natural, substituindo materiais já escassos, como o marfim, na produção de determinados produtos. No entanto, existe uma clara ambivalência na sua utilização. A poluição plástica está claramente ligada a múltiplas crises relacionadas com o desenvolvimento e a globalização, desde as Grandes Guerras até às crises climáticas e ecológicas. Embora os plásticos possam ter ajudado a reduzir os nossos impactos em algumas áreas (por exemplo, reduzir as emissões de carbono nos transportes), contribuíram para a Grande Aceleração e, portanto, para o nosso crescente impacto global no ambiente natural, e aprofundaram as relações capitalistas e a nossa dependência do petróleo.

Esta tese baseia-se na premissa de que o destino dos plásticos no ambiente não é completamente compreendido. Conforme discutido, existe uma grande inconsistência no balanço de massa dos plásticos no oceano e a sua distribuição em três dimensões nas diferentes bacias oceânicas, além disso a sua distribuição de tamanho ainda não está resolvida. Isso se deve, em parte, à falta de dados. O financiamento e a infraestrutura limitados para investigar e monitorar a poluição dificultam amplamente a produção de dados, diante da enorme escala dos oceanos e a onipresença da poluição plástica. Esta tese aborda essas limitações, melhorando os métodos de análise dos dados atuais, destacando lacunas na literatura e facilitando a produção de dados de alta qualidade sem a necessidade de infraestrutura adicional.

Algumas áreas, particularmente as do Sul Global, estão sub-representadas na literatura, levando a vieses na nossa compreensão da distribuição global do MPL, conforme discutido no Capítulo 3. Essas desigualdades na produção científica

podem criar disparidades espaciais no conhecimento na área. Os resultados apresentados na Figura 2 do Capítulo 3 mostram a distribuição regional dos estudos e podem ajudar pesquisadores a abordar as desigualdades nas estratégias de mitigação e adaptação ao MPL. Essas podem ser postas em prática por meio do direcionamento de recursos e do compartilhamento de conhecimentos e infraestruturas com pesquisadores locais nas regiões menos estudadas. Isso pode ajudar a reduzir distorções nas estimativas do balanço de massa global e produzir um entendimento comum das principais ameaças e soluções para o MPL.

Ainda assim, a maioria dos dados é coletada a partir de amostras feitas na superfície, enquanto a coluna d'água e os sedimentos são muito menos estudados. Isso representa mais uma fonte de viés e incompletude nos dados observacionais. Isso porque, embora a maioria dos plásticos flutuem na superfície imediatamente após sua emissão, a sua flutuabilidade pode ser afetada pelas intempéries, pela fragmentação e pelo crescimento de organismos na sua superfície. A discussão sobre os fatores e mecanismos de degradação, apresentada no Capítulo 4, contribui para uma melhor compreensão da distribuição do plástico nos compartimentos em três dimensões e como é condicionada pela degradação e pelas consequentes alterações nas propriedades dos plásticos. Esse debate traz à luz importantes propriedades dos plásticos que têm sido geralmente negligenciadas em estudos de degradação e fragmentação na literatura de MPL. Uma compreensão mais aprofundada desses fatores de degradação pode apoiar modelos conceituais e teóricos que estimam os fluxos de plásticos em compartimentos marinhos. Isso pode ajudar os investigadores a compreender melhor as concentrações de plásticos no ambiente, particularmente, os processos que levam a um estoque observado de plásticos na superfície menor do que o esperado. Além disso, cada polímero se comporta de maneira diferente de acordo com as suas características, incorrendo em taxas variadas de fragmentação, bioincrustação etc. Assim, modelos do tipo podem ajudar, ainda, a distinguir os destinos de tipos de plástico diferentes, o que é importante para avaliações de impacto.

Finalmente, o Capítulo 5 descreve um método para melhorar a velocidade e a confiabilidade na caracterização de polímeros em amostras ambientais. Identificar todas as partículas presentes em amostras ambientais é um desafio, em razão da quantidade em que são encontradas. Por exemplo, as pesquisas tradicionais em bases de dados são computacionalmente exigentes e imprecisas para plásticos

degradados, o que requer confirmação manual por um especialista. Técnicas de imagens espectrais têm sido utilizadas para analisar vários itens ao mesmo tempo, principalmente MPs pequenos, e são ótimas alternativas. No entanto, o equipamento necessário nem sempre está disponível, motivo pelo qual ainda são necessários métodos que aprimorem as técnicas convencionais de FTIR.

Uma metodologia rigorosa foi aplicada no artigo 3 (Capítulo 5) para treinar e avaliar diferentes modelos de ML na caracterização de espectros de FTIR de plásticos. O modelo escolhido como o modelo de melhor desempenho demonstrou melhorar significativamente o fluxo de trabalho de caracterização de partículas. Proporciona avanços significativos nas ferramentas de análise, considerando as dificuldades na geração desse tipo de dado. Embora ainda não seja capaz de classificar automaticamente todas as amostras concebíveis, espera-se que funcione na maioria dos casos, desde que as amostras sejam coletadas em condições semelhantes. Em última análise, a utilização de modelos de AM, como o proposto, pode aumentar a disponibilidade de dados de alta qualidade sobre plásticos no ambiente.

Esta tese é apresentada como um esforço composto para abraçar a Inteligência Artificial no estudo do Lixo Plástico Marinho buscando os benefícios que ela pode proporcionar para uma melhor compreensão de um dos principais problemas sociais e ambientais que enfrentamos no nosso tempo. Abordar as lacunas e os vieses nos dados é crucial para o desenvolvimento de estratégias eficazes para mitigar o impacto do lixo plástico marinho nos ecossistemas marinhos e na saúde humana. Esforços de pesquisa colaborativos e interdisciplinares ou transdisciplinares de pesquisadores de várias regiões, além de novas disciplinas, são essenciais para preencher essas lacunas e avançar a nossa compreensão do lixo plástico marinho, como prioridade para preservação do meio ambiente oceânico.

O aprendizado de máquina, em particular, tem o potencial de acelerar significativamente o ritmo da descoberta científica, permitindo aos investigadores analisar grandes quantidades de dados, revelar padrões e tomar decisões mais informadas.

7 PERSPECTIVAS FUTURAS

Independentemente das lacunas de conhecimento em MPL, já existem evidências suficientes para apoiar ações para reduzir a contaminação por plástico no meio ambiente. Para transicionarmos para além dos plásticos, é necessário que haja um esforço coordenado entre os setores público e privado em âmbito internacional. As discussões sobre soluções para essa crise, como as que atualmente precedem um Acordo Internacional Juridicamente Vinculante sobre Plásticos, devem ser abrangentes e evitar limitarem-se a meras soluções tecnológicas e alternativas aos plásticos convencionais. Reduzir o consumo é tão importante como melhorar as infraestruturas de tratamento de resíduos em todo o mundo. As dimensões do problema deixam espaço para estratégias de limpeza, novos modelos de valorização de resíduos, recuperação de modelos de negócios da era “pré-plásticos”, desenvolvimento de melhores materiais, entre outras transformações.

Dentre as possibilidades de expansão e aprofundamento dos trabalhos desenvolvidos nesta tese, pode-se destacar:

- Aplicar os métodos do Artigo 1 às diferentes áreas da ciência, buscando avaliar a distribuição e o desenvolvimento da pesquisa científica no mundo. Análises semelhantes já foram feitas observando disciplinas inteiras, como “Física e Astronomia” ou “Biociências”, mas a modelagem de tópicos pode trazer flexibilidade e insights ainda não obtidos;
- Investigar as concentrações de tipos de polímeros no oceano. Sobretudo em metanálises e estudos mais abrangentes, essa distinção é pouco considerada, possivelmente por dificuldades metodológicas, mas que podem evoluir com o uso de novas tecnologias, inclusive com a difusão da Inteligência Artificial;
- Realizar experimentos em taxas de degradação e afundamento para diferentes polímeros. Esses são temas já abordados na área, mas com literaturas ainda modestas que requerem maior investigação em variedade de polímeros e condições. Podem contribuir com dados importantes sobre as taxas de fluxo entre compartimentos ambientais;

- Aplicar o *pipeline* de treinamento e avaliação de modelos de ML (Artigo 3) a diferentes técnicas de análise química na caracterização de microplásticos, como microFTIR e raman. Tem potencial de melhorar os fluxos de trabalho em laboratórios que operam sob diferentes protocolos e aumentar a confiabilidade dos resultados. Ainda, a metodologia de treino e teste descrita no Artigo, traz discussões importantes sobre o rigor necessário para explorar ao máximo aplicações de ML;
- Desenvolver uma interface de fácil utilização para caracterização de plásticos, para que o modelo possa ser usado por todos, mesmo sem conhecimentos prévios em programação.

Ainda, outros estudos estratégicos complementares aos desenvolvidos até aqui são possíveis:

- Padronizar os métodos de coleta e análise de microplásticos e/ou avaliações rigorosas sobre a comparabilidade entre os métodos. Avanços nesse sentido são fundamentais para aproximarmos o balanço de massa e entendermos o destino dos plásticos no oceano;
- Melhorar os métodos de detecção de microplásticos pequenos pode abrir novas linhas de pesquisa e preencher lacunas de conhecimento sobre a fragmentação e a persistência desses contaminantes no ambiente;
- Ampliar espaço-temporalmente a distribuição de dados. O acervo de amostras do Continuous Plankton Recorder (CPR) pode ser o registro mais amplo, espaço-temporalmente, do acúmulo de plásticos nos oceanos. Entretanto, o baixo volume de filtragem e a pequena entrada de água significam que deve haver muito poucos e muito pequenos microplásticos nessas amostras, o que torna a análise demorada. A IA pode ajudar na detecção por meio da visão computacional hiperespectral;
- Avaliar a extensão e o impacto da utilização da IA no meio acadêmico é importante para entender e melhor gerenciar os riscos da sua utilização, minimizando os potenciais impactos negativos.

A amostragem do ambiente marinho para avaliar o estado da poluição não é uma tarefa fácil e normalmente requer infraestruturas dispendiosas e recursos substanciais. Assim, permitir que toda a comunidade se beneficie desses esforços é benéfico para o campo e garante que os recursos empregados, que na sua maioria são financiados publicamente, sejam gastos de forma eficiente. Essa é uma tendência observada em toda a pesquisa científica. Para garantir novos avanços, foram feitos esforços consideráveis na ciência para facilitar o acesso aos resultados das pesquisas. Iniciativas como o acesso aberto e o compartilhamento de dados têm sido promovidas e são cada vez mais exigidas pelas agências de fomento.

A produção de dados é uma tarefa laboriosa, por isso, há uma necessidade de promover a reutilização de dados acadêmicos. Recentemente, um consórcio de cientistas e organizações propôs uma estrutura de princípios para promover esse objetivo, chamada FAIR. Seu objetivo é tornar os dados “encontráveis”, acessíveis, interoperáveis e reutilizáveis (*Findable, Accessible, Interoperable and Reusable*, em inglês). “Os Princípios FAIR colocam ênfase específica no aprimoramento da capacidade das máquinas de encontrar e usar automaticamente os dados” (WILKINSON et al., 2016).

Grande parte das informações compartilhadas entre a comunidade científica, porém, vem na forma de publicações científicas. São documentos de texto que incluem não apenas medições, mas também considerações e inferências sobre os dados em que se baseiam. Embora normalmente não sejam considerados dados em si, os textos são um tipo de dado não estruturado que não se enquadra bem nos princípios FAIR, mas complementa os dados estruturados, como tabelas e arquivos. Os capítulos 3 e 4 servem como exemplos de quanta informação pode ser extraída de documentos com o uso de métodos de IA. Por exemplo, o Capítulo 3 contém a análise de mais de 9 mil publicações e mostra a diversificação de tópicos de pesquisa na última década, bem como mudanças em interesses particulares dentro desses tópicos. O desenvolvimento adicional de estruturas para incorporar algoritmos de ML em métodos de extração de informações tem o potencial de transformar a maneira como a pesquisa científica é conduzida e acessada, oferecendo oportunidades para aumentar a produtividade, promover a inovação e melhorar os processos de tomada de decisão. Ao automatizar tarefas como buscas e seleção de artigos e extração de dados, os algoritmos de aprendizado de máquina

podem agilizar os estágios iniciais das revisões sistemáticas, permitindo que os pesquisadores concentrem seus esforços em análises de nível superior. Da mesma forma, os sistemas alimentados por IA podem servir como ferramentas valiosas para consultores em políticas públicas embasadas na ciência, fornecendo resumos e as principais conclusões da pesquisa científica em um campo. No entanto, existem desafios relacionados com a falta de consenso e de referenciamento, que podem afetar a precisão e a credibilidade da informação recuperada pelos sistemas de IA. Embora a IA conversacional possa ajudar os investigadores a debater ideias e a extrair informações relevantes de textos, tirar conclusões a partir das provas acumuladas, ainda, normalmente requer julgamento humano, motivo pelo qual manter um equilíbrio entre a assistência da IA e a experiência humana será crucial para métodos com credibilidade no futuro.

Num futuro próximo, a IA tem potencial para aumentar a produtividade científica e a inovação, automatizando tarefas repetitivas e ajudando os investigadores a superar a sobrecarga de informação, que, como visto no Capítulo 3, pode levar à estagnação nos campos científicos e impedir descobertas. Ao analisar dados e identificar tendências, a IA pode ajudar os tomadores de decisão a tomarem decisões informadas e fundamentadas em provas científicas. Preocupações com essas tecnologias, desde a perda de empregos e os efeitos da dependência excessiva da assistência da IA, até à desinformação gerada pela IA que mina as democracias, realçam a necessidade de uma consideração e regulamentação cuidadosas. Deve ser feito um esforço colaborativo para avaliar rapidamente os numerosos impactos na sociedade.

Em vários campos científicos, os modelos de ML avançaram significativamente na modelagem preditiva, permitindo aos pesquisadores prever resultados com base em observações anteriores. No entanto, um desafio fundamental para a sua utilização generalizada na ciência reside na compreensão da causalidade por detrás das previsões feitas pelos modelos de IA. Embora a IA possa identificar correlações e prever categorias como visto ao longo desta tese, ela não pode explicitar os mecanismos causais subjacentes, o que é crucial para a compreensão científica. Ainda assim, pode ajudar os investigadores a gerar novas hipóteses de investigação.

REFERÊNCIAS

ALLEN, S. et al. Examination of the ocean as a source for atmospheric microplastics. **PLoS ONE**, v. 15, n. 5, 2020.

AUTA, H. S.; EMENIKE, C. U.; FAUZIAH, S. H. Distribution and importance of microplastics in the marine environment: A review of the sources, fate, effects, and potential solutions. **Environment International**, v. 102, p. 165–176, maio 2017.

AVIO, C. G.; GORBI, S.; REGOLI, F. Plastics and microplastics in the oceans: From emerging pollutants to emerged threat. **Marine Environmental Research**, Blue Growth and Marine Environmental Safety. v. 128, p. 2–11, 1 jul. 2017.

BORNMANN, L.; HAUNSCHILD, R.; MUTZ, R. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. **Humanities and Social Sciences Communications**, v. 8, n. 1, p. 1–15, 7 out. 2021.

BOUCHER, J.; FRIOT, D. **Primary microplastics in the oceans**. [s.l.] IUCN, 2017.

BRANDON, J.; GOLDSTEIN, M.; OHMAN, M. D. Long-term aging and degradation of microplastic particles: Comparing in situ oceanic and experimental weathering patterns. **Marine Pollution Bulletin**, v. 110, n. 1, p. 299–308, 15 set. 2016.

CALLISTER, W. **Chapter 15: Characteristics, Applications, and Processing of Polymers, Materials Science and Engineering: An Introduction**. 6. ed. [s.l.: s.n.].

CAPOLUPO, M. et al. Chemical composition and ecotoxicity of plastic and car tire rubber leachates to aquatic organisms. **Water Research**, v. 169, p. 115270, 1 fev. 2020.

COSGROVE. **‘Throwaway Living’: When Tossing Out Everything Was All the Rage**. Disponível em: <<https://time.com/3879873/throwaway-living-when-tossing-it-all-was-all-the-rage/>>. Acesso em: 10 abr. 2024.

COSTIGAN, E. et al. Adsorption of organic pollutants by microplastics: Overview of a dissonant literature. **Journal of Hazardous Materials Advances**, v. 6, p. 100091, 1 maio 2022.

CÓZAR, A. et al. Plastic debris in the open ocean. **Proceedings of the National Academy of Sciences**, v. 111, n. 28, p. 10239–10244, 15 jul. 2014.

CRAWFORD, C. B.; QUINN, B. 1 - The emergence of plastics. Em: CRAWFORD, C. B.; QUINN, B. (Eds.). **Microplastic Pollutants**. [s.l.] Elsevier, 2017. p. 1–17.

DE HAAN, W. P.; SANCHEZ-VIDAL, A.; CANALS, M. Floating microplastics and aggregate formation in the Western Mediterranean Sea. **Marine Pollution Bulletin**, v. 140, p. 523–535, 1 mar. 2019.

DRIS, R.; GASPERI, J.; TASSIN, B. Sources and Fate of Microplastics in Urban Areas: A Focus on Paris Megacity. Em: WAGNER, M.; LAMBERT, S. (Eds.). **Freshwater Microplastics : Emerging Environmental Contaminants?** Cham: Springer International Publishing, 2018. p. 69–83.

ERIKSEN, M. et al. Plastic Pollution in the World's Oceans: More than 5 Trillion Plastic Pieces Weighing over 250,000 Tons Afloat at Sea. **PLOS ONE**, v. 9, n. 12, p. e111913, 10 dez. 2014.

ERNI-CASSOLA, G. et al. Distribution of plastic polymer types in the marine environment; A meta-analysis. **Journal of Hazardous Materials**, v. 369, p. 691–698, 5 maio 2019.

FERNÁNDEZ-GONZÁLEZ, V. et al. Misidentification of PVC microplastics in marine environmental samples. **TrAC Trends in Analytical Chemistry**, v. 153, p. 116649, 1 ago. 2022.

GALL, S. C.; THOMPSON, R. C. The impact of debris on marine life. **Marine Pollution Bulletin**, v. 92, n. 1, p. 170–179, 15 mar. 2015.

GAN, X. et al. Network medicine framework reveals generic herb-symptom effectiveness of traditional Chinese medicine. **Science Advances**, v. 9, n. 43, p. eadh0215, 27 out. 2023.

GÉRON, A. 1. **The Machine Learning Landscape, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow**. 2. ed. [s.l: s.n.].

GESAMP. **Sources, Fate and Effects of Microplastics in the Marine Environment (Part 2)**. [s.l: s.n.]. Disponível em: <<http://www.gesamp.org/publications/microplastics-in-the-marine-environment-part-2>>. Acesso em: 22 nov. 2023.

GEYER, R. A Brief History of Plastics. Em: STREIT-BIANCHI, M.; CIMADEVILA, M.; TRETTNAK, W. (Eds.). **Mare Plasticum - The Plastic Sea: Combatting Plastic Pollution Through Science and Art**. Cham: Springer International Publishing, 2020. p. 31–47.

GEYER, R.; JAMBECK, J. R.; LAW, K. L. Production, use, and fate of all plastics ever made. **Science Advances**, v. 3, n. 7, p. e1700782, 19 jul. 2017.

GILMAN, E. et al. Highest risk abandoned, lost and discarded fishing gear. **Scientific Reports**, v. 11, n. 1, p. 7195, 30 mar. 2021.

HARTMANN, N. B. et al. Microplastics as vectors for environmental contaminants: Exploring sorption, desorption, and transfer to biota. **Integrated Environmental Assessment and Management**, v. 13, n. 3, p. 488–493, 2017.

HASSAN, F. et al. Microplastic contamination in sewage sludge: Abundance, characteristics, and impacts on the environment and human health. **Environmental Technology & Innovation**, v. 31, p. 103176, 1 ago. 2023.

ISO/TR 21960:2020. , 2020. Disponível em:

<<https://www.iso.org/standard/72300.html>>. Acesso em: 10 abr. 2024

JAMBECK, J. R. et al. Plastic waste inputs from land into the ocean. **Science**, v. 347, n. 6223, p. 768–771, 2015.

JORDAN, M. I. Artificial Intelligence—The Revolution Hasn't Happened Yet. **Harvard Data Science Review**, v. 1, n. 1, 3 jul. 2019.

JUNG, M. R. et al. Validation of ATR FT-IR to identify polymers of plastic marine debris, including those ingested by marine organisms. **Marine Pollution Bulletin**, v. 127, p. 704–716, 1 fev. 2018.

JURADO, E.; DACHS, J. Seasonality in the “grasshopping” and atmospheric residence times of persistent organic pollutants over the oceans. **Geophysical Research Letters**, v. 35, n. 17, 2008.

KANE, I. A.; CLARE, M. A. Dispersion, Accumulation, and the Ultimate Fate of Microplastics in Deep-Marine Environments: A Review and Future Directions. **Frontiers in Earth Science**, v. 7, 30 abr. 2019.

KOELMANS, A. A. et al. All is not lost: deriving a top-down mass budget of plastic at sea. **Environmental Research Letters**, v. 12, n. 11, p. 114028, nov. 2017.

KOOI, M. et al. Ups and Downs in the Ocean: Effects of Biofouling on Vertical Transport of Microplastics. **Environmental Science & Technology**, v. 51, n. 14, p. 7963–7971, 18 jul. 2017.

LAM, R. et al. Learning skillful medium-range global weather forecasting. **Science**, 22 dez. 2023.

LANGLEY, P. The changing science of machine learning. **Machine Learning**, v. 82, n. 3, p. 275–279, 1 mar. 2011.

LAU, J. **Google Maps 101: How AI helps predict traffic and determine routes.** Disponível em: <<https://blog.google/products/maps/google-maps-101-how-ai-helps-predict-traffic-and-determine-routes/>>. Acesso em: 11 abr. 2024.

LAVAN, L.; SWIFT, E.; PUTZEYS, T. Material Spatiality In Late Antiquity: Sources, Approaches And Field Methods. Em: **Objects in Context, Objects in Use**. [s.l.] Brill, 2008. p. 1–42.

LAW, K. L. et al. Plastic Accumulation in the North Atlantic Subtropical Gyre. **Science**, v. 329, n. 5996, p. 1185–1188, 3 set. 2010.

LEBRETON, L. et al. Evidence that the Great Pacific Garbage Patch is rapidly accumulating plastic. **Scientific Reports**, v. 8, n. 1, p. 4666, 22 mar. 2018.

LEBRETON, L. C. M. et al. River plastic emissions to the world's oceans. **Nature Communications**, v. 8, n. 1, p. 15611, 7 jun. 2017.

LEBRETON, L.; EGGER, M.; SLAT, B. A global mass budget for positively buoyant macroplastic debris in the ocean. **Scientific Reports**, v. 9, n. 1, p. 12922, 12 set. 2019.

LEHNER, R. et al. Emergence of Nanoplastic in the Environment and Possible Impact on Human Health. **Environmental Science & Technology**, v. 53, n. 4, p. 1748–1765, 19 fev. 2019.

LESLIE, H. A. et al. Discovery and quantification of plastic particle pollution in human blood. **Environment International**, v. 163, p. 107199, 1 maio 2022.

MATERIĆ, D. et al. Micro- and Nanoplastics in Alpine Snow: A New Method for Chemical Identification and (Semi)Quantification in the Nanogram Range. **Environmental Science & Technology**, v. 54, n. 4, p. 2353–2359, 18 fev. 2020.

MOON, W. et al. **Path Planning of Cleaning Robot with Reinforcement Learning**. arXiv, , 17 ago. 2022. Disponível em: <<http://arxiv.org/abs/2208.08211>>. Acesso em: 11 abr. 2024

NAPPER, I. E.; THOMPSON, R. C. Release of synthetic microplastic plastic fibres from domestic washing machines: Effects of fabric type and washing conditions. **Marine Pollution Bulletin**, v. 112, n. 1–2, p. 39–45, 15 nov. 2016.

NATIONAL HISTORIC CHEMICAL LANDMARK. **Bakelite® First Synthetic Plastic**. Disponível em: <<https://www.acs.org/education/whatischemistry/landmarks/bakelite.html>>. Acesso em: 10 abr. 2024.

OECD. **Global Plastics Outlook: Economic Drivers, Environmental Impacts and Policy Options**. Paris: Organisation for Economic Co-operation and Development, 2022.

PATHAK, G.; MARK, N. Ecocommunicability, citizenship, and discourses on plastic control in India - ScienceDirect. **Geoforum**, v. 125, out. 2021.

PERKINS, T. ‘A sea of misinformation’: FTC to address industry greenwashing complaints. **The Guardian**, 4 maio 2023.

Plastics - the Facts 2021. . [s.l.] Plastics Europe, 2021. Disponível em: <<https://plasticseurope.org/knowledge-hub/plastics-the-facts-2021/>>. Acesso em: 10 abr. 2024.

Plastics, EDCs & Health: A guide for public interest organizations and policy-makers on endocrine disrupting chemicals & plastics. . [s.l.] The Endocrine Society, International Pollutants Elimination Network, 10 dez. 2020. Disponível em: <<https://www.endocrine.org/topics/edc/plastics-edcs-and-health>>. Acesso em: 10 abr. 2024.

PLASTICS EUROPE. **Plastics – the fast Facts 2023**. [s.l.: s.n.]. Disponível em: <<https://plasticseurope.org/knowledge-hub/plastics-the-fast-facts-2023/>>. Acesso em:

10 abr. 2024.

PRATA, J. C. Airborne microplastics: Consequences to human health? **Environmental Pollution (Barking, Essex: 1987)**, v. 234, p. 115–126, mar. 2018.

RITCHIE, H.; SAMBORSKA, V.; ROSER, M. “**Data Page: Global plastics production**”. Disponível em: <<https://ourworldindata.org/grapher/global-plastics-production>>. Acesso em: 10 abr. 2024.

SAUVÉ, S.; DESROSIERS, M. A review of what is an emerging contaminant. **Chemistry Central Journal**, v. 8, n. 1, p. 15, 26 fev. 2014.
Science-Policy Interface for Plastic Pollution. . [s.l.] GRID Arendal, 2023.
Disponível em: <<https://www.grida.no/publications/1007>>. Acesso em: 22 nov. 2023.

SEBILLE, E. VAN et al. A global inventory of small floating plastic debris. **Environmental Research Letters**, v. 10, n. 12, p. 124006, dez. 2015.

SLADE, G. **Made to Break: Technology and Obsolescence in America**. [s.l.: s.n.].

SMITH, J. K. World War II and the Transformation of the American Chemical Industry. Em: MENDELSON, E.; SMITH, M. R.; WEINGART, P. (Eds.). **Science, Technology and the Military**. Dordrecht: Springer Netherlands, 1988. p. 307–322.

SPRING, M. et al. **Policy Brief: Creating a strong interface between science, policy and society to tackle global plastic pollution**. [s.l.: s.n.].

STEFFEN, W. et al. The trajectory of the Anthropocene: The Great Acceleration. **The Anthropocene Review**, v. 2, n. 1, p. 81–98, abr. 2015.

SU, L. et al. Global transportation of plastics and microplastics: A critical review of pathways and influences. **The Science of the Total Environment**, v. 831, p. 154884, 20 jul. 2022.

SUTTON, R. T. et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. **npj Digital Medicine**, v. 3, n. 1, p. 1–10, 6 fev. 2020.

THOMPSON, R. C. et al. Lost at Sea: Where Is All the Plastic? **Science**, v. 304, n. 5672, p. 838–838, 7 maio 2004.

TYLER, C. et al. AI tools as science policy advisers? The potential and the pitfalls. **Nature**, 2023.

UNDERWOOD, A. J.; CHAPMAN, M. G.; BROWNE, M. A. Some problems and practicalities in design and interpretation of samples of microplastic waste. **Analytical Methods**, v. 9, n. 9, p. 1332–1345, 2 mar. 2017.

VAN NOORDEN, R.; PERKEL, J. M. AI and science: what 1,600 researchers think. **Nature**, v. 621, n. 7980, p. 672–675, 27 set. 2023.

VASWANI, A. et al. **Attention Is All You Need**. arXiv, , 1 ago. 2023. Disponível em: <<http://arxiv.org/abs/1706.03762>>. Acesso em: 22 nov. 2023

VICTOR, D. Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk. **The New York Times**, 2016.

WAYMAN, C.; NIEMANN, H. The fate of plastic in the ocean environment – a minireview. **Environmental Science: Processes & Impacts**, v. 23, n. 2, p. 198–212, 2021.

WEI, J. et al. **Emergent Abilities of Large Language Models**. arXiv, , 26 out. 2022. Disponível em: <<http://arxiv.org/abs/2206.07682>>. Acesso em: 11 abr. 2024

WEINSTEIN, J. E. et al. Microplastics: A Global Water Pollution Problem. Em: **Encyclopedia of Water**. [s.l.] John Wiley & Sons, Ltd, 2019. p. 1–15.

Why Nature will not allow the use of generative AI in images and video. **Nature**, v. 618, n. 7964, p. 214–214, 7 jun. 2023.

WILKINSON, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, v. 3, n. 1, p. 160018, 15 mar. 2016.

WRIGHT, S. L.; KELLY, F. J. Plastic and Human Health: A Micro Issue? **Environmental Science & Technology**, v. 51, n. 12, p. 6634–6647, 20 jun. 2017.

YADAV, P. et al. Biomedical Biopolymers, their Origin and Evolution in Biomedical Sciences: A Systematic Review. **Journal of Clinical and Diagnostic Research : JCDR**, v. 9, n. 9, p. ZE21–ZE25, set. 2015.

ZALASIEWICZ, J. et al. The geological cycle of plastics and their use as a stratigraphic indicator of the Anthropocene. **Anthropocene**, v. 13, p. 4–17, 1 mar. 2016.