Appendix for "Intersections between materials science and marine plastics to address environmental degradation drivers using machine learning tools"

Henrique de Medeiros Back, Daphiny Pottmaier, Camilla Kneubl Andreussi, Orestes Estevam Alarcon

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 mechanisms[1]. 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, nonetheless[3]. 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 alcohol[4].

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 mass[5] 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 reaction[8]. 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 1960s[10]. 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 embedding[11], 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 encoded[12]. Topic Analysis allows clustering or classifying these encoded documents into distinct classes based on their vector representations[13]. Sample size should be taken into account when discussing results, as they provide an indication of the generalizability of the findings[14].

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 abstracts[14]. 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 developed[13]. 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 tokes. 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.

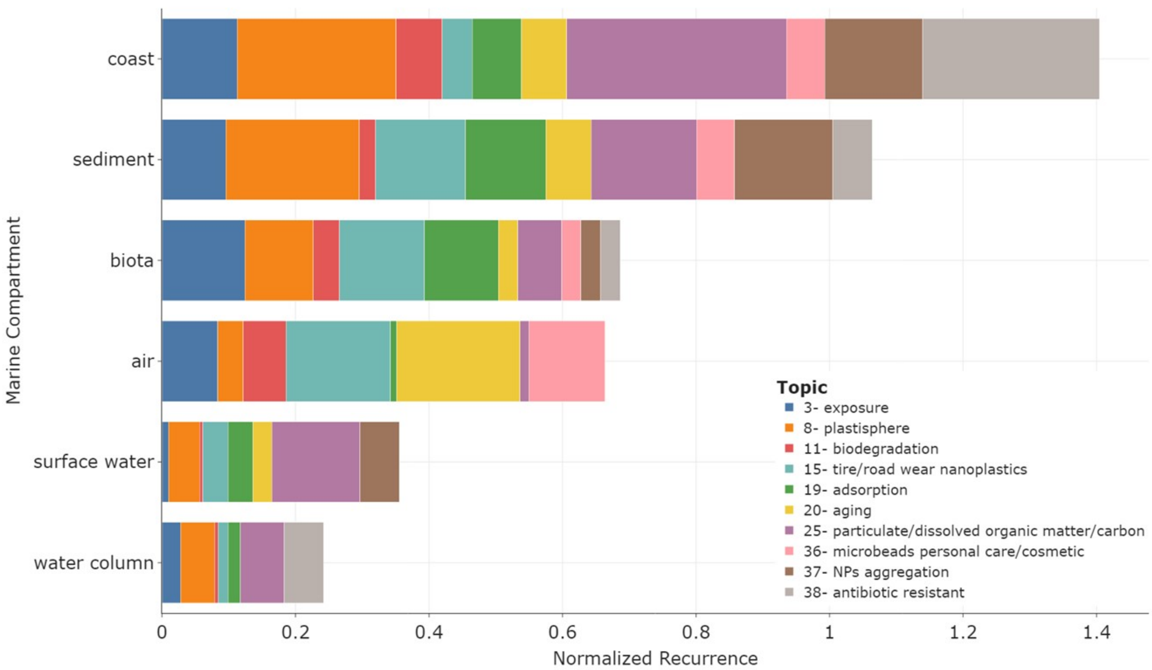

Figure S1 - Normalized recurrence of terms associated with marine compartments given by topic

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.