data

MDPI

Article

Seaweed-Based Bioplastics: Data Mining Ingredient–Property Relations from the Scientific Literature

Fernanda Véliz [1], Thulasi Bikku [1,2], Davor Ibarra-Pérez [3], Valentina Hernández-Muñoz [4], Alysia Garmulewicz [5] and Felipe Herrera [1,6,*]

[1] Department of Physics, Universidad de Santiago de Chile, Av Victor Jara 3493, Santiago 9170124, Chile; fernanda.veliz@usach.cl (F.V.); thulasi.bikku@gmail.com (T.B.)
[2] Computer Science and Engineering, Amrita School of Computing Amaravati, Amrita Vishwa Vidyapeetham, Amaravati 522503, Andhra Pradesh, India
[3] Department of Mechanical Engineering, University of Santiago of Chile (USACH), Avenida Libertador Bernardo O'Higgins 3363, Santiago 9170022, Chile
[4] Department of Industrial Engineering, University of Santiago of Chile (USACH), Avenida Libertador Bernardo O'Higgins 3363, Santiago 9170022, Chile
[5] Department of Management, Faculty of Management and Economics, University of Santiago of Chile (USACH), Avenida Libertador Bernardo O'Higgins 3363, Estación Central 9170022, Chile
[6] Millennium Institute for Research in Optics, Concepción 4030000, Chile
* Correspondence: felipe.herrera.u@usach.cl

Abstract: Automated analysis of the scientific literature using natural language processing (NLP) can accelerate the identification of potentially unexplored formulations that enable innovations in materials engineering with fewer experimentation and testing cycles. This strategy has been successful for specific classes of inorganic materials, but their general application in broader material domains such as bioplastics remains challenging. To begin addressing this gap, we explore correlations between the ingredients and physicochemical properties of seaweed-based biofilms from a corpus of 2000 article abstracts from the scientific literature since 1958, using a supervised word co-occurrence analysis and an unsupervised approach based on the language model MatBERT without fine-tuning. Using known relations between ingredients and properties for test scenarios, we discuss the potential and limitations of these NLP approaches for identifying novel combinations of polysaccharides, plasticizers, and additives that are related to the functionality of seaweed biofilms. The model demonstrates a valuable predictive ability to identify ingredients associated with increased water vapor permeability, suggesting its potential utility in optimizing formulations for future research. Using the model further revealed alternative combinations that are underrepresented in the literature. This automated method facilitates the mapping of relationships between ingredients and properties, guiding the development of seaweed bioplastic formulations. The unstructured and heterogeneous nature of the literature on bioplastics represents a particular challenge that demands ad hoc fine-tuning strategies for state-of-the-art language models for advancing the field of seaweed bioplastics.

Keywords: seaweed; bioplastics; natural language processing; masked language model; BERT

Academic Editor: Yongqing Cai

Received: 6 November 2024
Revised: 10 January 2025
Accepted: 14 January 2025
Published: 1 February 2025

Citation: Véliz, F.; Bikku, T.; Ibarra-Pérez, D.; Hernández-Muñoz, V.; Garmulewicz, A.; Herrera, F. Seaweed-Based Bioplastics: Data Mining Ingredient–Property Relations from the Scientific Literature. *Data* **2025**, *10*, 20. https://doi.org/10.3390/data10020020

1. Introduction

Bioplastics manufacturing is a subject of great interest due to the harmful effects of plastic film on the environment. The majority of plastic bags and single-use packaging materials, made of petrochemical materials, are not recycled, ultimately breaking down into microparticles in landfills or oceans, leading to environmental degradation and even contamination of our food supply [1]. Thus, the development of environmentally friendly films with performance comparable to traditional polymers has become increasingly relevant [2,3].

Bioplastic films made from seaweed polysaccharides have emerged as a promising solution to address the environmental concerns associated with plastic film production [4]. Seaweed-based raw materials are fully biodegradable and can be cultivated using environmentally friendly practices that support ecosystem sustainability. Agar, alginate, and carrageenan are commonly used polysaccharides for manufacturing biopolymeric films from seaweed [5–7]. However, films made from a single seaweed material often have poor properties, such as mechanical or water vapor barrier properties [8,9]. To address this issue, additives or other biomaterials can be incorporated to enhance the properties of seaweed films.

Data mining has gained great relevance in recent decades due to its potential in natural language processing and machine learning modelling techniques [10]. Bioplastics datasets and regression models have been developed for assisting the experimental development of seaweed-based bioplastics [11,12]. Data mining techniques can be used to create probabilistic models that detect multi-level word associations [13,14] to address different problems involving large corpuses of specific text, such as the extraction of technical information. These techniques involve pipelines of natural language processing (NLP) tasks that have focused primarily on biomedical tasks [15] but, more recently, NLP and Large Language Models (LLMs) have found relevant applications in chemistry [16] and materials science [17,18]. While previous applications in materials science have focused on material classes such as inorganic glasses, ceramics, and alloys [17], our study is the first to apply these techniques to biopolymeric materials.

In this work, we build a corpus of 405,404 words based on 2000 scientific abstracts on seaweed biopolymer to analyse frequencies and co-occurrences of polysaccharides, plasticizes, and additives and physical properties of reported films. We use a Bag-of-Words (BoW) approach to obtain a co-occurrence matrix that identifies combinations of common and rare ingredients used in the literature, without assigning metrics of performance with respect to properties. We then explore the ability of two transformer-based Large Language Models (LLMs) pre-trained on a general material science corpus to assess the potential performance of commonly used combinations of ingredients and properties. This approach was conducted using prompts using Masked Language Modelling (MLM) in sentences designed to qualitatively interpret the relationship between compounds in the BoW. The overall NLP pipeline used in this work is illustrated in Figure 1. Our findings indicate that LLMs could suggest correlations between certain ingredients and properties, as confirmed by selected literature reports, but limitations in their ability to suggest new experiments remain.

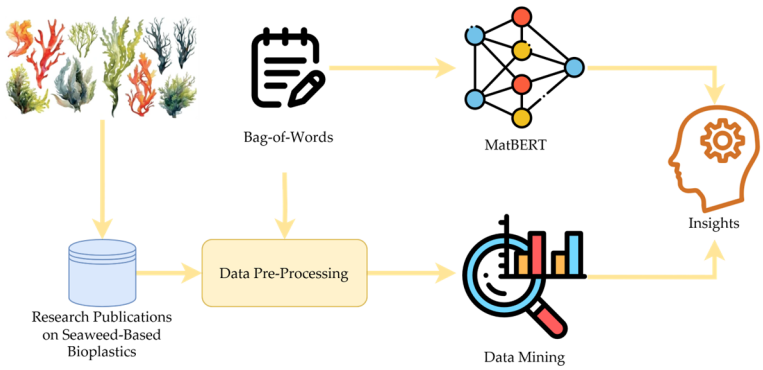

Figure 1. NLP pipeline used in this work. We extracted abstracts from various scientific publications and employed a Bag-of-Words model to analyse the co-occurrence of ingredients and properties of the bioplastics. Additionally, the Bag-of-Words model was utilized for an unsupervised approach, where different word representations were combined in the MatBERT model.

2. Materials and Methods

2.1. Abstract Corpus

A Scopus search was conducted to gather publications on seaweed biopolymers, using keywords such as "Alginate", "Agar", "Carrageenan", "Seaweed" or "Algae", and "Film" or "Packaging". Research articles and reviews from 1958 to 2022 were included, resulting in 2000 publications. The metadata and abstracts of each publication were downloaded, and publications lacking a DOI or not written in English were excluded. The number of seaweed-based bioplastic abstracts is at least two orders of magnitude lower than for other NLP studies in materials science [18]. The search keywords are listed in Table 1. The list of abstracts and search keywords can be found in the article's GitHub repository, available in [19].

Table 1. List of keywords for searching articles in Scopus related to seaweed-based materials and bioplastics or related packaging materials.

(TITLE-ABS-KEY (alginate OR agar OR carrageenan OR seaweed OR macroalgae) AND TITLE-ABS-KEY (bioplastic OR bio-plastic OR "biopolymer film" OR film OR "plastic bag" OR packaging OR biocomposite OR bio-composite))

2.2. Data Pre-Processing

To prepare the text data for analysis, pre-processing techniques such as stop word removal and lemmatization were used. Care was taken to ensure that the original meaning of the words in the abstracts was preserved after pre-processing. The resulting abstract corpus contained 276,490 words, after pre-processing.

2.3. Bag of Words and Co-Occurrence Analysis

A bag of words (BoW) was created by selecting ingredient names and properties from a list of 20 review articles covering various types of biofilms, constituent components, and characterization of their properties. Commonly reported names of ingredients and material properties were included in the BoW, giving a total of 255 ingredients, classified in 6 categories, and 111 material properties classified in 10 categories. To process the abstracts using the BoW as input to obtain word frequencies and word co-occurrences, the following steps were taken: tokenize the abstracts to split them into individual words, create a vocabulary of unique words using a set data structure, count the frequency of each word in each abstract using a dictionary, create a co-occurrence matrix that shows how often

each word co-occurs with every other word in an abstract, count co-occurrences by iterating through each abstract, and finally normalize the matrix by dividing each entry by the total number of co-occurrences to make the values interpretable and comparable across different abstracts. Co-occurrence matrices help to visualize relationships and patterns between words. The BoW dictionary could be updated and refined over time as new insights and knowledge are gained in the field.

2.4. Masked Language Modelling

MLM is a pre-training method and is utilized for how BERT is pre-training, which involves selectively masking (hiding) 15% of the words or tokens in the input within a text and then training a language model to predict what those masked words should be. This approach helps the model learn contextual information and relationships between words in a given language [20]. MatBERT is a pre-trained language model that has been trained using MLM and next-sentence prediction as the unsupervised training objectives. The model has been trained on a general materials science corpus biased towards experimental synthesis topics such as oxides, energetic materials, magnetic materials, and synthesis techniques [21]. The corpus of this training contains 2 million papers of materials science literature, it has a maximum 512 input token size with 768 hidden dimensions, and the vocabulary size for the tokenizer is 30.522 [12].

In using MatBERT with the MLM technique, prompts were generated to operate differently in analysing the relationships between ingredients and water vapor permeability in the context of film manufacturing. By masking the adjective in the prompts using [MASK], the importance of that adjective in the relationship between ingredients and the properties present in the bag of words is emphasized. Additionally, a score distribution method was applied to the qualifier word to evaluate how meaningful it was. However, it is crucial to acknowledge certain limitations associated with employing different prompts. Variability in prompt structures may introduce biases or limitations in the model responses, potentially influencing the overall findings [22].

3. Results and Discussion

3.1. Word Frequencies for Ingredients and Properties

Figure 2 shows the frequency of ingredient occurrence in a collection of documents without repetition. The probability is calculated as the ratio between the number of documents in which each ingredient appears and the total number of analysed documents. The ingredients are grouped into different categories, indicated by a color coding that facilitates visualization. The percent probabilities are shown, providing an overview of the distribution of ingredients throughout the corpus. The color coding corresponds to the six categories to which the ingredients in the BoW belong: organic, polysaccharide, inorganic, protein, plasticizer, and synthetic polymer.

The percent distribution of ingredient classes is shown in Figure 2, inset. Polysaccharide ingredients have the highest occurrence in the corpus. Additionally, both organic and inorganic ingredients used as additives in various studies are identified, with inorganic ingredients being more frequent.

Figure 3 shows the percent probability of material property occurrences in a collection of documents without repetition. The color coding indicates the categories to which the material properties belong, which are listed in the inset. The properties categorized as chemical, mechanical, antimicrobial, and optical are distributed relatively homogeneously. The predominance of tensile strength suggests that this a focal property when evaluating the performance of materials in various film applications. This relatively uniform

representation of categories illustrates the balanced study of different types of material properties in the field.

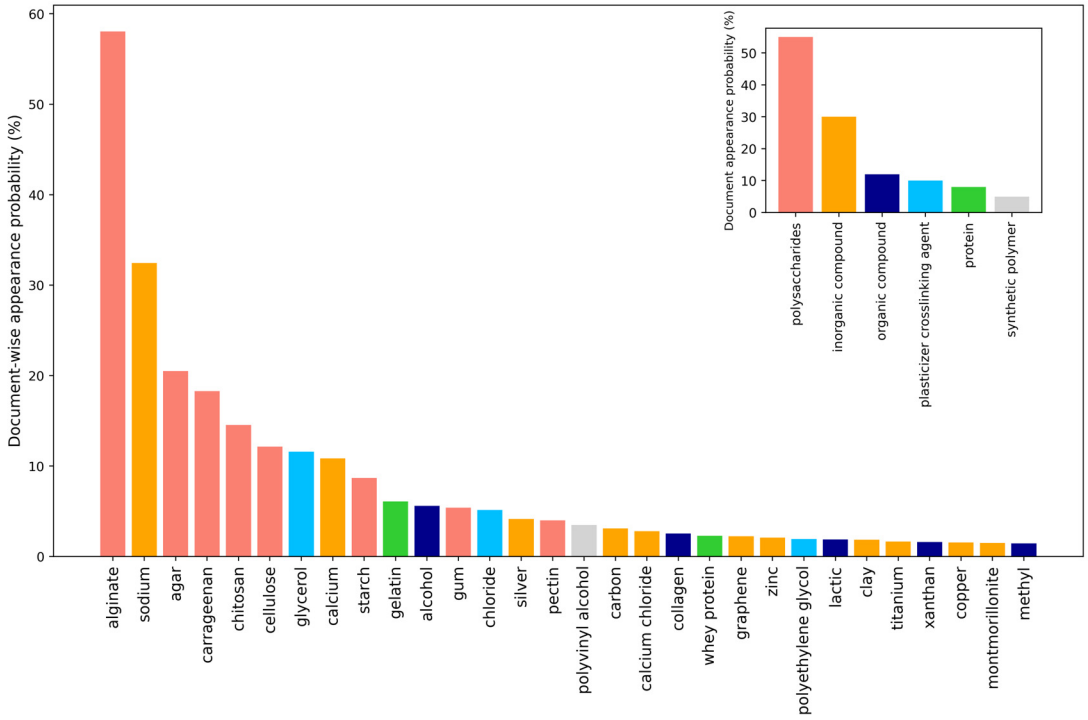

Figure 2. Document-wise percent occurrence probability for ingredients from the bag of words (BoW) in the corpus of 2000 scientific literature abstracts. Inset: percentage of classes of materials present in the BoW that occur in the abstract corpus.

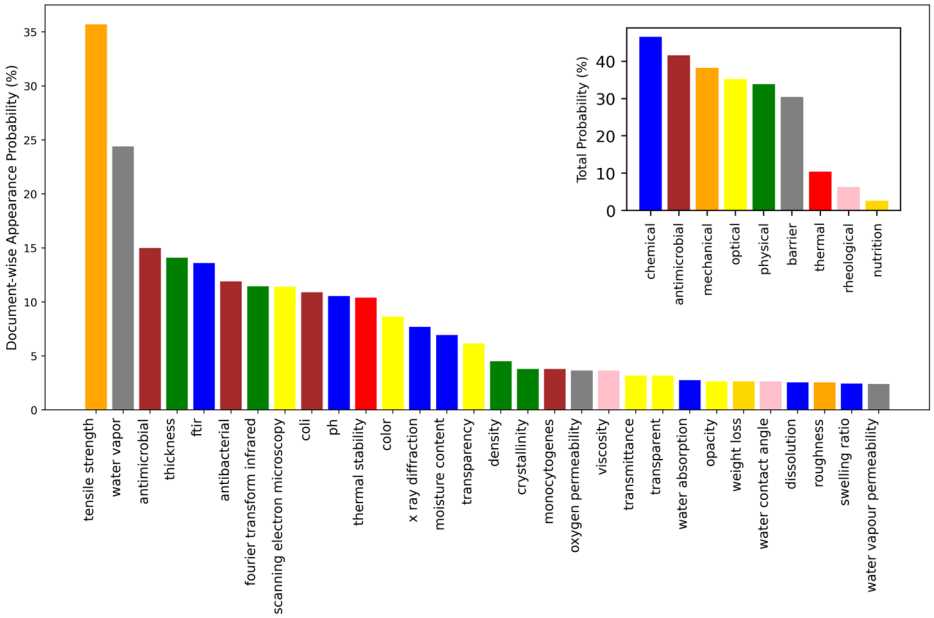

Figure 3. Document-wise occurrence probability of material properties the BoW in the corpus of 2000 scientific literature abstracts. Inset: percentage of classes of material properties present in the BoW.

3.2. Co-Occurrence Visualization

Figure 4 shows the matrix of ingredient–ingredient co-occurrences, given by the instances in which two ingredients appear together in the dataset of 2000 abstracts. This

matrix is valuable for visualizing the data related to the co-use of ingredients in the literature. By identifying pairs of ingredients in frequent associations, researchers can assess feasible relationships for exploring potential film formulations. The matrix shows that alginate is a central component in many combinations of ingredients, indicating its versatility and wide application in various formulations. The dominant presence of polysaccharides in combination with other ingredients reflects their importance in the publication record.

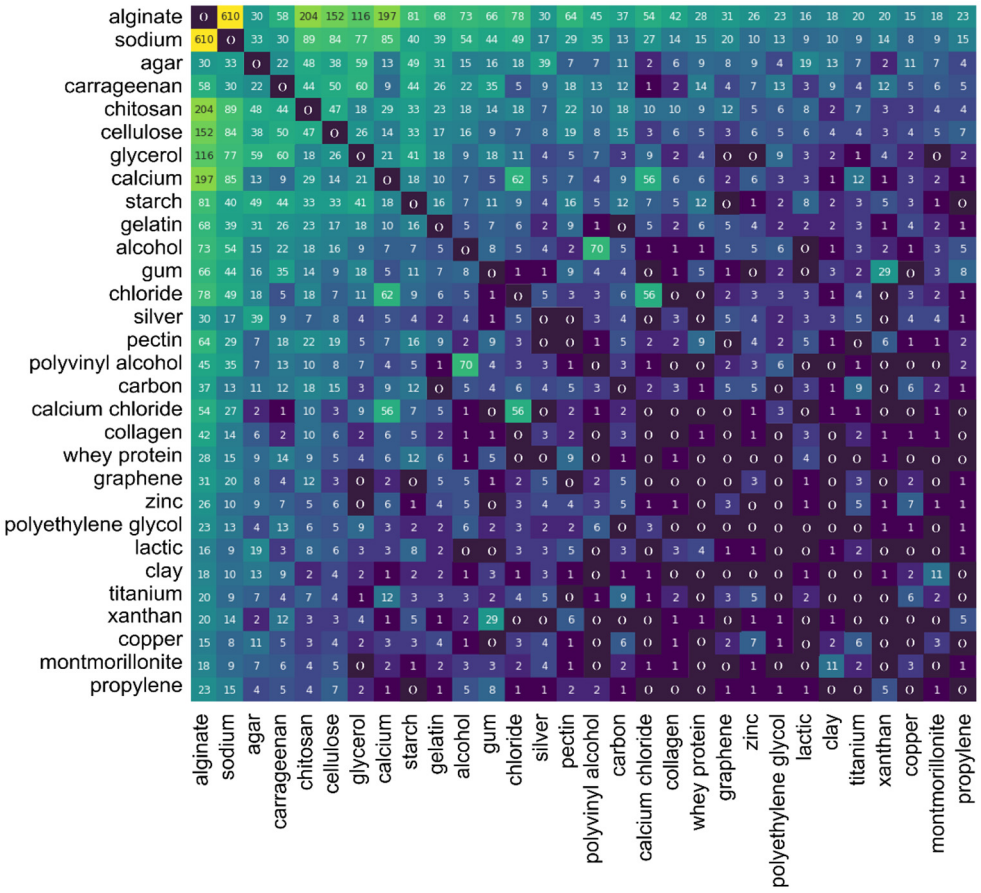

Figure 4. Ingredient–ingredient co-occurrence matrix as a heatmap**.** The 30 most commonly occurring ingredients in the dataset of abstracts are included. Each matrix entry contains the number of co-occurrences. The color scale indicates the values of co-ocurrence from yellow (high values) to dark blue (low values), as marked.

The lower right corner of the heatmap shows combinations of ingredients with less co-occurrence, such as titanium/zinc or clay/montmorillonite, which may indicate a possible relationship between them. Information about rate combinations could be valuable for identifying research niches where the potential of these ingredients can be explored for new applications or in improving the properties of existing materials.

Figure 5 shows the co-occurrence matrix of ingredients and material properties, obtained by the number of times each combination of ingredient and material property occurs in the dataset of article abstracts. Tensile strength, barrier properties, and antimicrobial activity are some of the most frequency studied properties with a broad range of ingredients. These correlations could be used to find trends in the data corpus for extracting approximate insights about seaweed-based bioplastics. However, there are limitations to working with statistical word trends, which suggests the need for more advanced NLP approaches.

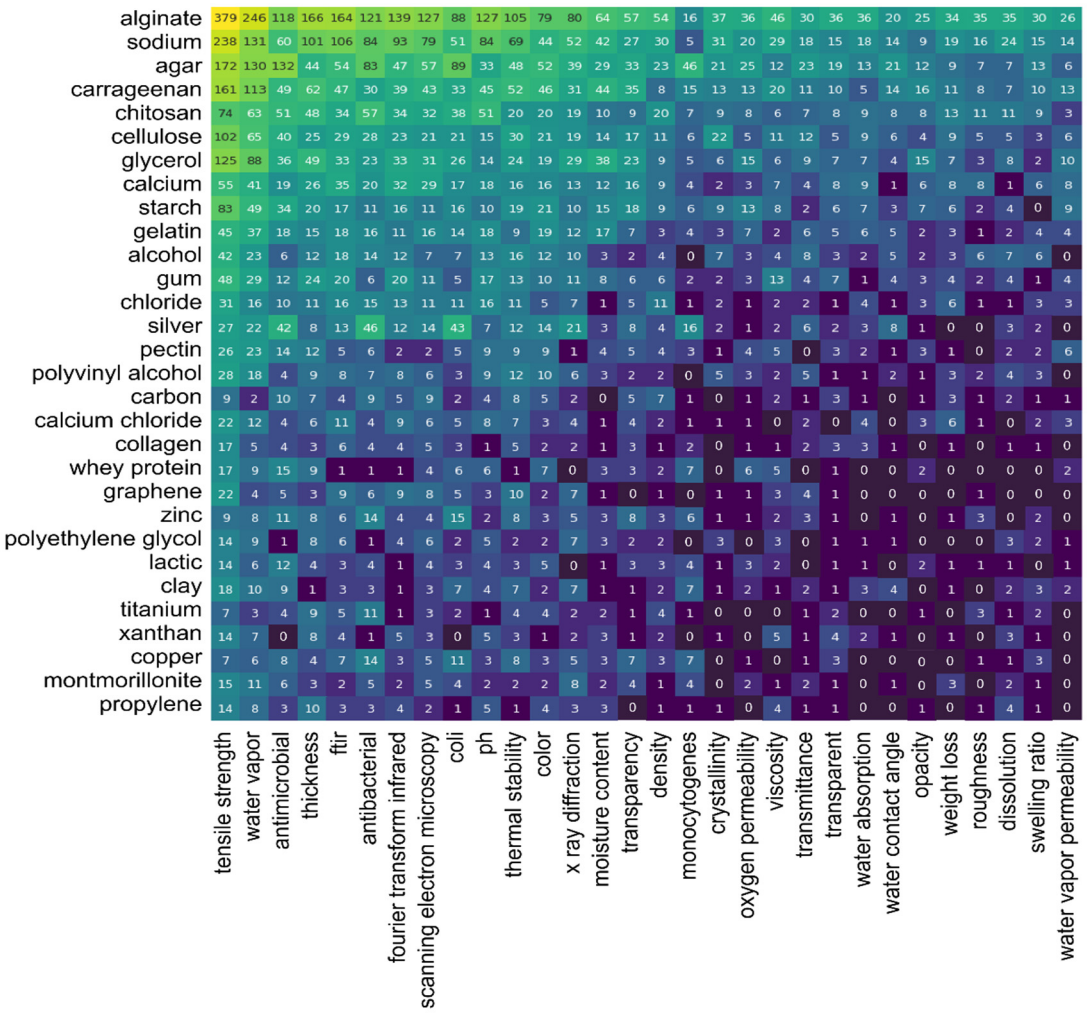

Figure 5. Ingredient–property co-occurrence matrix as a heatmap**.** The 30 most commonly occurring ingredients in the dataset of abstracts are included. Each matrix entry contains the number of co-occurrences. The color scale indicates the values of co-ocurrence from yellow (high values) to dark blue (low values), as marked.

3.3. Ingredients and Properties from Masked Language Models

In what follows, we use an unsupervised approach based on LLMs without fine-tuning, which are pre-trained to identify materials. We explore the ability of this approach to uncover relationships between different sets of ingredients and specific properties, which can potentially lead to predictions for enhanced physical characteristics in bioplastics. The advantage of using a pre-trained model over the BoW word-counting approach used above is the ability of LLMs to benefit from contextual information in the corpus.

Specifically, we explore how organic and inorganic additives correlate with properties of bioplastic films, particularly the water vapor permeability, using two Masked Language Models (see Section 2). We adopt the Fill Mask method, which consists of filling in a [MASK] in the sentence to predict possible replacements. The model is designed to describe the way compounds influence specific properties, using natural language to focus on how the combination of these compounds impacts the property depending on the output scores.

We assume hypothetical use cases of alginate membranes and films combined with glycerol as a plasticizer. The "Additive" word was extracted from a predefined bag of words containing a total of 185 organic and inorganic additives. We explored how the model MatBERT suggests, based on the sentence context, the effect of adding a third

compound as an additive by assessing how this incorporation influences the water vapor permeability of the resulting bioplastics. Table 2 shows the four sentences (S1–S4) that were used in this Fill Mask test. We found that other sentence formulations with similar meaning gave similar conclusions.

Table 2. Input sentences used in the MatBERT language model to interpret the relationship between the different compounds and water vapor permeability. The ingredient {Compound 3} is taken from a predefined bag of words of 185 organic and inorganic additives. On output, the model predicts a [MASK] word with a score.

Masked Sentence
<S1> Membranes were prepared using alginate, a polysaccharide derived from seaweed, combined with glycerol as a plasticizer. When {**Additive**} was incorporated as a secondary additive, the water vapor permeability of the membrane [**MASK**], potentially affecting its suitability for packaging applications.
<S2> The film was produced by mixing alginate, extracted from seaweed, with glycerol to enhance flexibility. Upon addition of {**Additive**}, the water vapor permeability of the resulting film [**MASK**], which could influence its performance in moisture-sensitive environments.
<S3> By adding {**Additive**} to a film composed of alginate, a seaweed-based biopolymer, and glycerol, the water vapor permeability [MASK]. This modification aims to optimize the barrier properties of the bioplastic for specific applications.
<S4> By incorporating {**Additive**} as an additive in a film formulation based on alginate, a marine-derived biopolymer, and glycerol, the water vapor permeability [**MASK**]. Such enhancements could improve the functional properties of bioplastic films for use in sustainable packaging.

Table 3 shows the [MASK] outputs predicted by MatBERT for each input sentence (S1–S4). For sentence S1, the model identifies propyl as the additive with the highest output score for "*Decreased*" (0.54%). Propyl derivatives are chemical compounds that include the propyl group (C_3H_7) as part of their structure, such as hydroxypropyl methyl cellulose (HPMC) and hydroxypropyl cellulose (HPC), and they are used in the formulations of different types of films and membranes [23–25]. Additionally, the scientific literature mentions the use of propylene glycol (PG) as a plasticizer. The use of these compounds in both contexts is related to modifying the properties of materials, such as water vapor permeability or drug release, to enhance their performance in specific applications such as packaging or drug delivery systems. Also, in relation to sentence S1, methyl (CH_3) is found as an additive that decreases water vapor permeability. While not specific to bioplastics, studies have demonstrated the incorporation of methyl in compounds such as hydroxypropyl methyl cellulose (HPMC) and sodium carboxymethyl cellulose (Na CMC) in the fabrication of mucoadhesive films [26,27]. The graft copolymerization of methyl methacrylate (MMA) onto alginate has also been explored, which is also related to the modification of properties of polymeric materials.

Table 3. Top-scoring additives in masked sentences for modifying water vapor permeability using the MatBERT model. The output MASK in each of the sentences **S1–S4** is a qualifier on the impact on the water vapor permeability of adding a third component (additive) to a mixture of alginate and glycerol.

Sentence	Third Component	Mask 1	Mask 2	Mask 3	Mask 4
S1	propyl	decreased: 0.5401%	increased: 0.3574%	improved: 0.1314%	reduced: 0.0180%
	methyl	decreased:	increased:	improved:	reduced:

		0.5379%	0.3568%	0.0288%	0.0194%
	ethyl	decreased: 0.5327%	increased: 0.3610%	improved: 0.0291%	reduced: 0.0183%
S2	grape seed	increased: 0.6404%	decreased: 0.2639%	increases: 0.0296%	improved: 0.0115%
	organic powdered cottonii	increased: 0.6388%	decreased: 0.2562%	increases: 0.0327%	improved: 0.0141%
	apricot kernel	increased: 0.6370%	decreased: 0.2715%	increases: 0.0275%	reduced: 0.0122%
S3	watermelon	increases: 0.3895%	decreases: 0.3522%	increased: 0.1024%	decreased: 0.0905%
	gold	increases: 0.3889%	decreases: 0.3080%	increased: 0.1302%	decreased: 0.0980%
	spinach	increases: 0.3887%	decreases: 0.3051%	increased: 0.1297%	decreased: 0.1005%
S4	lysozyme	increased: 0.5653%	increases: 0.1306%	decreased: 0.1155%	improved: 0.0841%
	peroxidase	increased: 0.5588%	decreased: 0.1373%	increases: 0.1108%	improved: 0.0882%
	wheat straw	increased: 0.5558%	decreased: 0.1278%	increases: 0.1216%	improved: 0.0899%

However, it is difficult to discern whether a specific ingredient consistently decreases or increases water vapor permeability. For instance, the incorporation of propylene glycol alginate can lead to either a reduction or an increase in water vapor permeability and water solubility, depending on its concentration in the formulation [28]. This dual effect emphasizes the necessity of precise concentration control when recommending additives for bioplastic fabrication. Moreover, when applying MatBERT to S1, we observe the model's sensitivity to contextual cues like "affecting". The presence of this term may introduce a negative bias, prompting the model to predict a negative adjective such as "decreased" for the masked word.

For sentence S2, the model identifies grape seed extract as the additive with the highest score for mask "*Increased*" (0.6404%), meaning it is likely to increase water vapor permeability. This ingredient is less frequently reported in relation to membrane creation than other additives, but there are reports highlighting its benefits in plastic and bioplastic films. The scientific literature shows that the phenolic compounds present in grape seed extract have antioxidant properties and potential molecular interactions with bio*polymers* that can modify the mechanical and functional properties of the material [29]. Additionally, the use of grape seed extract as an active agent in edible films has been documented to improve water vapor permeability, confirming the mask output while also giving antiviral and antioxidant capabilities to films, suggesting its potential for enhancing performance in specific applications [30]. The model also suggests that the use of organic powdered cottonii (OPC) could influence film properties when using glycerol as one of its plasticizers, which is also in agreement with reported results [31]. OPC is a product containing carrageenan and derived from the *Eucheuma cottonii* seaweed. OPC is known for altering film characteristics such as water vapor permeability. Although its specific use as an additive for alginate has not been reported, OPC is related to seaweed as it contains carrageenan. While its effectiveness has been evaluated in applications such as food packaging and edible coatings, studies do not specify its use as an additive for alginate, nor an

exact correlation between the simultaneous use of a polysaccharide, a plasticizer, and OPC as an additive. Instead, more complex combinations of OPC together with other additives and polysaccharides have been explored, as is the case with the OPC which contains carrageenan, and its impact largely depends on the specific formulation used.

For sentence S3, the model shows a decreasing trend in the difference between output values, indicating a minimal difference between "decrease" and "increase" when it comes to additives such as watermelon extract. Upon reviewing the literature on watermelon, it was found that its use has been reported in multiple contexts, including as an active ingredient and stabilizer for silver and zinc oxide nanoparticles when extracted as melanin from watermelon seeds [32,33]. Additionally, watermelon rind has been utilized to add value by creating edible alginate/glycerol films. This suggests that the model can recommend potential ingredients for specific applications based on the desired properties, demonstrating its capability to identify suitable additives for enhancing the performance of bioplastics.

In sentence S4, the model identifies three additives that, when incorporated into polymer matrices, can alter their physical properties, whether by increasing the barrier against water vapor and oxygen or by boosting microbial growth inhibition. Lysozyme, an antimicrobial enzyme produced by animals, and peroxidase, an enzyme occurring especially in plants, milk, and white blood cells, are related to enhancing microbial growth inhibition in biomaterials [34–37]. Wheat straw helps improve the mechanical properties of biopolymer-based films made from Poly(3-hydroxybutyrate-co-3-hydroxyvalerate (PHBV), carrageenan, and alginate, with variations in its effectiveness depending on how it is integrated into the bioplastic matrix. Regarding the results of the MatBERT model, Sentence 2 has the highest score, identifying 44 relevant ingredients with scores above 0.6. In contrast, Sentence 3 shows the lowest scores, displaying less relevant ingredients, such as watermelon extract.

Figure 6 shows the distribution of the top predictive masks by cumulative score for Sentence S1. The bar chart presents the total sum of the scores obtained for each of the masks which were considered the most probable in the various combinations of components evaluated. As observed, the adjectives "decreased" and "increased" are the most common, with significantly higher scores compared to other masks such as "improved", "reduced", "declined", and "dropped". This suggests that, in the context of bioplastic film additives, the MatBERT model was more frequently able to predict changes related to the decrease or increase in water vapor permeability in biopolymers. As shown in Table 2, these predictions tend to be linked to the incorporation of certain additives to modify the mechanical properties of films that include biopolymers, such as alginate and carrageenan. We also carried out a more explicit testing of MatBERT to explore its ability to predict an ingredient that increases the water vapor permeability of a film based on sodium alginate, which is one the main ingredients in seaweed films (see Figure 2). Table 4 shows the sentences (SA, SB, and SC) used in this test.

Table 4. Input sentences used in the MatBERT language model to interpret the relationship between sodium alginate and additives for increasing water vapor permeability.

Sentences
<**SA**> By adding [MASK] to sodium alginate, the water vapor permeability increases.
<**SB**> By adding an additive such as [MASK] to a sodium alginate film, the water vapor permeability increases.
<**SC**> Adding additives such as [MASK] to a sodium alginate film increases its water vapor permeability.

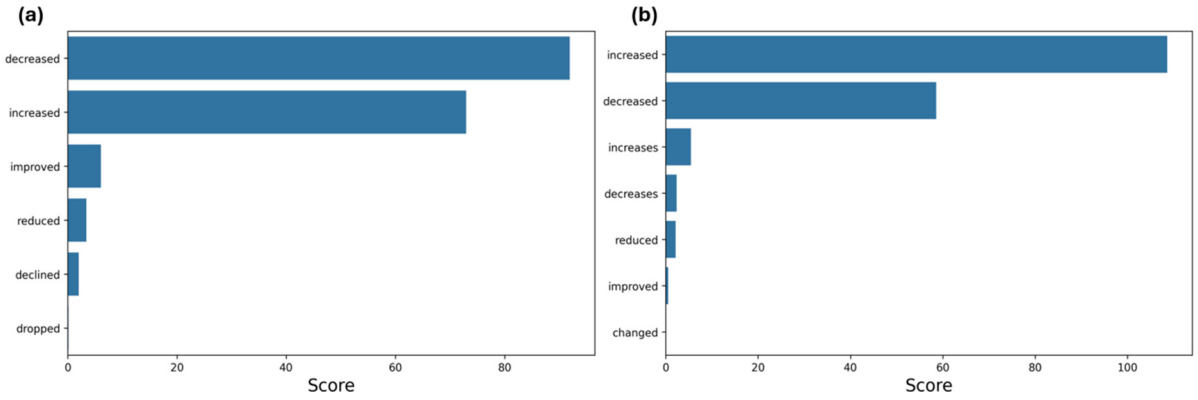

Figure 6. Distribution of the top predictive masks by cumulative score for S1 (panel **a**) and S2 (panel **b**). The chart shows the total sum of scores for each of the masks that were found to be among the most probable in the different combinations of components evaluated.

Table 5 displays the top five ingredients for each sentence. When comparing the predicted words with Figure 5, although there were variations in ingredients due to differences in sentences, we find that output ingredient words such as starch (polysaccharide) are predominant in most cases, along with chitosan (polysaccharide), gelatin (protein), and glycerol (plasticizer). Additionally, there is the presence of ethanol, an organic compound used to remove pigments and fatty acids [38], and Polyvinyl Alcohol (PVA), a synthetic polymer utilized for bioplastic preparations [39]. The relationship between sodium alginate and glycerol in bioplastic formulations is well documented, and the MatBERT output reproduces this combination. The model also predicted starch as an alternative for improving the water vapor permeability in sentences SB and SC. Starch combined with alginate is known not only for improving permeability but also for modifying the mechanical properties of biofilms [40]. This literature support for the model output is promising but also limited, given the broad generic corpus on which MatBERT was trained, primarily with inorganic chemistry literature.

Table 5. The top five predicted ingredients for increasing the water vapor permeability of seaweed-based films according to the MatBERT model, based on output scores. The mask is a second component in a mixture containing sodium alginate, as specified in sentences SA, SB, and SC from Table 4.

	Predicted Masked Words				
SA	chitosan: 0.1626% starch: 0.0539% PVA: 0.0513% gelatin: 0.0376% water: 0.0293%	**SB**	starch: 0.0757% gelatin: 0.0371% ethanol: 0.0349% PVA: 0.0344% glycerol: 0.0339%	**SC**	starch: 0.1038% gelatin: 0.0567% surfactants: 0.0498% PVP: 0.0418% glycerol: 0.0402%

Table 6 compares two BERT models used in materials science for additive prediction, averaging the top five outputs for sodium alginate, agar, and carrageenan using the masked sentences from Table 4. MatBERT predicts additives commonly cited in the scientific literature with the aim of developing applications in food packaging, food preservation, and biomedicine, using films made from polysaccharides derived from seaweed. For example, the combination of agar and PVA with chitosan in packaging films has shown that incorporating natural nanocomposites can improve water vapor permeability [41]. Similarly, the combination of gelatin with sodium alginate increases water vapor

permeability when yarrow essential oil (YEO) is added [42]. In contrast, MatSciBERT tends to predict additives associated with inorganic materials, reflecting the focus of its training data.

Table 6. Comparison of BERT models in materials science.

Model	**MatBERT**	**MatSciBERT**
Size	2,000,000 papers	150,000 papers
Dataset	Scientific publications, journal articles, and databases containing technical and academic texts in the field of materials science.	Inorganic glasses, metallic glasses, alloys, and cement and concrete from the Elsevier Science Direct Database.
Sodium Alginate	Starch: 0.0778 Chitosan: 0.0752 Gelatin: 0.0438 PVA: 0.0369 PVP: 0.0309	Sucrose: 0.0388 Glucose: 0.0242 Urea: 0.0251 Phosphate: 0.0148 Magnesium: 0.0133
Agar	Starch: 0.0853 Gelatin: 0.0605 Chitosan: 0.0408 Glycerol: 0.0312 NaCl: 0.0258	Zinc: 0.0153 Glucose: 0.014 Methanol: 0.0125 Starch: 0.0109 Glycerol: 0.01
Carrageenan	Starch: 0.0816 Chitosan: 0.0682 Glycerol: 0.0437 Gelatin: 0.0418 PVA: 0.0344	Sucrose: 0.0314 Aluminium: 0.02 Glucose: 0.0188 Magnesium: 0.0186 Glycerol: 0.0122

4. Discussion

Recent transformer-based language models for materials science such as MatBERT and MatSciBERT have not been specifically trained or fine-tuned for learning correlations between ingredients and properties in a corpus of seaweed-based bioplastics. While, in principle, it is not expected that the model outputs could be used for discussing formulations of seaweed-based films, yet some of the high-scoring outputs in Table 5 are ingredients known to be associated with water vapor permeability studies. Similar output trends are seen when testing for mechanical properties (tensile strength), but the output word distribution for ingredients (plasticizers and additives) or qualifiers (increase, decrease, good, or poor) often contain noise that needs expert assessment. The literature support found for some of the ingredient–property associations in Table 5 was limited [38–40]. However, the positive correlation suggests that the family of language models based on BERT could be valuable for the future development of bioplastic formulations after further training and fine-tuning efforts.

The output of the BERT models shows that the increase in permeability is closely related to variations in the concentrations of both the additive and the plasticizer. In this context, the model faces limitations in accurately interpreting interactions when provided with sentences containing limited context, which hinders its ability to capture the complexity of ingredient interactions. However, the model still demonstrates a valuable predictive ability to identify ingredients associated with increased water vapor permeability, suggesting its potential utility in optimizing formulations for future research.

The effectiveness of a BERT model in predicting specific elements, such as additives, largely depends on the training data corpus. For example, a model trained with data predominantly related to metallic materials tends to predict metal-related additives when

used with masked cues that explore the properties of these additives. This is because the tokenization process and learning are shaped by the dominant terms and contexts in the training data. To improve predictions in specialized areas, such as additives for seaweed polysaccharides, it is beneficial to use a diversified or specially selected corpus. Such a corpus should cover a wide variety of materials and their interactions with various additives. MatBERT, for example, is trained on a wide range of materials science literature, which provides a more complete basis for predictions in this domain.

5. Conclusions

In conclusion, our study provides ways to analyse common ingredient combinations in seaweed-based bioplastics and their relationship to properties of interest. We have identified critical ingredients such as starch, cellulose, chitosan, PLA, and their relationship to properties such as biodegradability using a word co-occurrence matrix. The application of the MatBERT model enabled us to explore new and less common combinations of polysaccharides, additives, and plasticizers. Using the model revealed alternative combinations that are underrepresented in the literature. This automated method facilitates a deeper understanding of the relationship between ingredients and properties, guiding the development of more effective seaweed bioplastic formulations. The model empowers innovators to swiftly identify ingredient combinations tailored to specific applications, enhancing the potential for experimentation with rare and underexplored combinations. This can be used to guide the development of seaweed bioplastic formulations, allowing innovators to quickly identify ingredient combinations of use to specific applications.

Our co-occurrence study has limitations with respect to the accuracy of the associations suggested between the ingredients and the properties of bioplastics, originating from the relatively small corpus size of published scientific abstracts and the bag of words. The analysis of the Masked Language Model outputs for terms within the bag of words is primarily limited by the envisioned mismatch between the general materials science corpus on which the language models were trained and the domain-specific corpus related to seaweed bioplastics. In future studies, these limitations can be addressed by expanding the text mining and data extraction processes using full-length articles including information on the fabrication and synthesis conditions of biofilms, which has been shown to be useful during the fine-tuning steps of more advanced Large Language Models [43]. Specific metrics for assessing the quality of the predicted correlations between the ingredients and the properties of biofilms and reducing the amount of expert assessment required also need to be developed before automated bioplastic formulation algorithms can be deployed. Addressing these data and model gaps is essential for advancing research and practical applications.

Author Contributions: Conceptualization, A.G. and V.H.-M.; software, F.V. and T.B.; methodology, F.V., T.B., and F.H.; validation, F.V., D.I.-P., T.B., and V.H.-M.; formal analysis, F.V., T.B., and V.H.-M.; data curation, F.V., D.I.-P., V.H.-M., and T.B.; writing—original draft preparation, F.V., T.B., and V.H.-M.; writing—review and editing, F.V., A.G., and V.H.-M.; visualization, F.V. and T.B.; supervision, V.H.-M.; project administration, V.H.-M.; funding acquisition, A.G. and V.H.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Agencia Nacional de Investigación y Desarrollo (ANID), grants Fondef IT20I0127, Fondecyt Regular 1221420 and Millennium Science Initiative Program ICN17_012. The APC was funded by Millennium Science Initiative Program ICN17_012.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The corpus data and code used for co-occurrence matrices and ingredient combinations with MatBERT are available at URL https://github.com/fherreralab/-Seaweed-Based-Bioplastics-Data-Mining-Ingredient-Property-Relations-from-the-Scientific-Literature (accessed on 16 January 2025).

Acknowledgments: V.H.-M., F.V., and T.B. were supported by ANID Fondecyt Regular 1221420 and the Millennium Science Initiative Program ICN17_012. AG would like to acknowledge support by the Department of Management and the Faculty of Management and Economics, University of Santiago of Chile. All authors were supported by ANID Fondef IT20I0127.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Thompson, R.C.; Moore, C.J.; vom Saal, F.S.; Swan, S.H. Plastics, the environment and human health: Current consensus and future trends. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2009**, *364*, 2153–2166. https://doi.org/10.1098/rstb.2009.0053.
2. Siracusa, V.; Rocculi, P.; Romani, S.; Rosa, M.D. Biodegradable polymers for food packaging: A review. *Trends Food Sci. Technol.* **2008**, *19*, 634–643. https://doi.org/10.1016/j.tifs.2008.07.003.
3. Chen, G.-Q.; Patel, M.K. Plastics Derived from Biological Sources: Present and Future: A Technical and Environmental Review. *Chem. Rev.* **2012**, *112*, 2082–2099. https://doi.org/10.1021/cr200162d.
4. Aleksanyan, K.V. Polysaccharides for Biodegradable Packaging Materials: Past, Present, and Future (Brief Review). *Polymers* **2023**, *15*, 451. https://doi.org/10.3390/polym15020451.
5. Martín-Del-Campo, A.; Fermín-Jiménez, J.A.; Fernández-Escamilla, V.V.; Escalante-García, Z.Y.; Macías-Rodríguez, M.E.; Estrada-Girón, Y. Improved extraction of carrageenan from red seaweed (*Chondracantus canaliculatus*) using ultrasound-assisted methods and evaluation of the yield, physicochemical properties and functional groups. *Food Sci. Biotechnol.* **2021**, *30*, 901–910. https://doi.org/10.1007/s10068-021-00935-7.
6. Lomartire, S.; Marques, J.C.; Gonçalves, A.M.M. An Overview of the Alternative Use of Seaweeds to Produce Safe and Sustainable Bio-Packaging. *Appl. Sci.* **2022**, *12*, 3123. https://doi.org/10.3390/app12063123.
7. Rajeswari, A.; Christy, E.J.S.; Swathi, E.; Pius, A. Fabrication of improved cellulose acetate-based biodegradable films for food packaging applications. *Environ. Chem. Ecotoxicol.* **2020**, *2*, 107–114. https://doi.org/10.1016/j.enceco.2020.07.003.
8. Escamilla-García, M.; Calderón-Domínguez, G.; Chanona-Pérez, J.J.; Mendoza-Madrigal, A.G.; Di Pierro, P.; García-Almendárez, B.E.; Amaro-Reyes, A.; Regalado-González, C. Physical, Structural, Barrier, and Antifungal Characterization of Chitosan–Zein Edible Films with Added Essential Oils. *Int. J. Mol. Sci.* **2017**, *18*, 2370. https://doi.org/10.3390/ijms18112370.
9. Dungani, R.; Sumardi, I.; Suhaya, Y.; Aditiawati, P.; Dody, S.; Rosamah, E.; Islam, N.; Hartati, S.; Karliati, T. Reinforcing Effects of Seaweed Nanoparticles in Agar-Based Biopolymer Composite: Physical, Water Vapor Barrier, Mechanical, and Biodegradable Properties. *BioResources* **2021**, *16*, 5118–5132. https://doi.org/10.15376/biores.16.3.5118-5132.
10. Cameron, J.J.; Leung, C.K. Mining Frequent Patterns from Precise and Uncertain Data," 2011, UNIFACS. Available online: http://hdl.handle.net/1993/32123 (accessed on 28 September 2024).
11. Hernández, V.; Ibarra, D.; Triana, J.F.; Martínez-Soto, B.; Faúndez, M.; Vasco, D.A.; Gordillo, L.; Herrera, F.; García-Herrera, C.; Garmulewicz, A. Agar Biopolymer Films for Biodegradable Packaging: A Reference Dataset for Exploring the Limits of Mechanical Performance. *Materials* **2022**, *15*, 3954. https://doi.org/10.3390/ma15113954.
12. Trewartha, A.; Walker, N.; Huo, H.; Lee, S.; Cruse, K.; Dagdelen, J.; Dunn, A.; Persson, K.P.; Ceder, G.; Jain, A. The Impact of Domain-Specific Pre-Training on Named Entity Recognition Tasks in Materials Science. *SSRN Electron. J.* **2021**, *3*, 100488. https://doi.org/10.2139/ssrn.3950755.
13. Liu, R.-L. Identification of conclusive association entities in biomedical articles. *J. Biomed. Semant.* **2019**, *10*, 1. https://doi.org/10.1186/s13326-018-0194-9.
14. Salloum, S.A.; Al-Emran, M.; Monem, A.A.; Shaalan, K. Using Text Mining Techniques for Extracting Information from Research Articles. In *Intelligent Natural Language Processing: Trends and Applications*; Springer: New York, NY, USA, 2017; pp. 373–397. https://doi.org/10.1007/978-3-319-67056-0_18.
15. Neumann, M.; King, D.; Beltagy, I.; Ammar, W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In Proceedings of the 18th BioNLP Workshop and Shared Task, Florence, Italy, 1 August 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019

16. Jablonka, K.M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Leveraging large language models for predictive chemistry. *Nat. Mach. Intell.* **2024**, *6*, 161–169. https://doi.org/10.1038/s42256-023-00788-1.
17. Gupta, T.; Zaki, M.; Krishnan, N.M.A. Mausam MatSciBERT: A materials domain language model for text mining and information extraction. *npj Comput. Mater.* **2022**, *8*, 102. https://doi.org/10.1038/s41524-022-00784-w.
18. Kononova, O.; Huo, H.; He, T.; Rong, Z.; Botari, T.; Sun, W.; Tshitoyan, V.; Ceder, G. Text-mined dataset of inorganic materials synthesis recipes. *Sci. Data* **2019**, *6*, 203. https://doi.org/10.1038/s41597-019-0224-1.
19. Véliz, F.A.V.; Bikku, T.; Ibarra, D.; Hernández, V.; Garmulewicz, A.; Herrera, F. *fherreralab/-Seaweed-Based-Bioplastics-Data-Mining-Ingredient-Property-Relations-from-the-Scientific-Literature: Supplementary Material*, v1.0; Zenodo: Genève, Switzerland, 2024. https://doi.org/10.5281/zenodo.13927103.
20. Tunstall, L.; von Werra, L.; Wolf, T. *Natural Language Processing with Transformers Building Language Applications with Hugging Face*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2022.
21. Trewartha, A.; Walker, N.; Huo, H.; Lee, S.; Cruse, K.; Dagdelen, J.; Dunn, A.; Persson, K.A.; Ceder, G.; Jain, A. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns* **2022**, *3*, 100488. https://doi.org/10.1016/j.patter.2022.100488.
22. Liao, W.; Liu, Z.; Dai, H.; Wu, Z.; Zhang, Y.; Huang, X.; Chen, Y.; Jiang, X.; Liu, D.; Zhu, D.; et al. Mask-guided BERT for Few Shot Text Classification. *Neurocomputing* **2023**, *610*, 128576. https://doi.org/10.1016/j.neucom.2024.128576.
23. Roda, A.; Prabhu, P.; Dubey, A. Design and evaluation of buccal patches containing combination of hydrochlorothiazide and atenolol. *Int. J. Appl. Pharm.* **2018**, *10*, 105–112. https://doi.org/10.22159/ijap.2018v10i2.24460.
24. Dawaba, A.M.; Dawaba, H.M.; Abu El-Enin, A.S.M.; Khalifa, M.K.A. Fabrication of bioadhesive ocusert with different polymers: Once a day dose. *Int. J. Appl. Pharm.* **2018**, *10*, 309–317. https://doi.org/10.22159/ijap.2018v10i6.28495.
25. Kazemi, Z.; Taghizadeh, S.M.; Keshavarz, S.T.; Lahootifard, F. Effect of composition on mechanical and physicochemical properties of mucoadhesive buccal films containing buprenorphine hydrochloride: From design of experiments to optimal formulation. *J. Drug Deliv. Sci. Technol.* **2020**, *56*, 101578. https://doi.org/10.1016/j.jddst.2020.101578.
26. Kim, B.-S.; Park, G.-T.; Park, M.-H.; Shin, Y.G.; Cho, C.-W. Preparation and evaluation of oral dissolving film containing local anesthetic agent, lidocaine. *J. Pharm. Investig.* **2017**, *47*, 575–581. https://doi.org/10.1007/s40005-016-0298-0.
27. Yang, Y.; Yu, X.; Zhu, Y.; Zeng, Y.; Fang, C.; Liu, Y.; Hu, S.; Ge, Y.; Jiang, W. Preparation and application of a colorimetric film based on sodium alginate/sodium carboxymethyl cellulose incorporated with rose anthocyanins. *Food Chem.* **2022**, *393*, 133342. https://doi.org/10.1016/j.foodchem.2022.133342.
28. Rhim, J.W.; Wu, Y.; Weller, C.L.; Schnepf, M. Physical characteristics of a composite film of soy protein isolate and propyleneglycol alginate. *J. Food Sci.* **1999**, *64*, 149–152. https://doi.org/10.1111/j.1365-2621.1999.tb09880.x.
29. Fabra, M.J.; Falcó, I.; Randazzo, W.; Sánchez, G.; López-Rubio, A. Antiviral and antioxidant properties of active alginate edible films containing phenolic extracts. *Food Hydrocoll.* **2018**, *81*, 96–103. https://doi.org/10.1016/j.foodhyd.2018.02.026.
30. Wang, S.; Marcone, M.F.; Barbut, S.; Lim, L.-T. Fortification of dietary biopolymers-based packaging material with bioactive plant extracts. *Food Res. Int.* **2012**, *49*, 80–91. https://doi.org/10.1016/j.foodres.2012.07.023.
31. Fransiska, D.; Giyatmi; Basmal, J.; Susanti, E. The effect of organic powdered cottonii concentration and types of plasticizers on the characteristics of edible film. *IOP Conf. Series: Earth Environ. Sci.* **2020**, *483*, 012008. https://doi.org/10.1088/1755-1315/483/1/012008.
32. Łopusiewicz, Ł.; Macieja, S.; Śliwiński, M.; Bartkowiak, A.; Roy, S.; Sobolewski, P. Alginate Biofunctional Films Modified with Melanin from Watermelon Seeds and Zinc Oxide/Silver Nanoparticles. *Materials* **2022**, *15*, 2381. https://doi.org/10.3390/ma15072381.
33. Wu, H.; Hu, B.; Dong, Z.; Lu, M.; Peng, Q.; Zhang, Z. Preparation and properties analysis of edible watermelon rind based film. *J. Food Sci. Biotechnol.* **2018**, *37*, 1091–1098. https://doi.org/10.3969/J.ISSN.1673-1689.2018.10.014.
34. Li, Q.; Xu, J.; Zhang, D.; Zhong, K.; Sun, T.; Li, X.; Li, J. Preparation of a bilayer edible film incorporated with lysozyme and its effect on fish spoilage bacteria. *J. Food Saf.* **2020**, *40*, 12832. https://doi.org/10.1111/jfs.12832.
35. Min, S.; Harris, L.J.; Han, J.H.; Krochta, J.M. Listeria monocytogenes Inhibition by Whey Protein Films and Coatings Incorporating Lysozyme. *J. Food Prot.* **2005**, *68*, 2317–2325. https://doi.org/10.4315/0362-028x-68.11.2317.
36. Murillo-Martínez, M.M.; Tello-Solís, S.R.; García-Sánchez, M.A.; Ponce-Alquicira, E. Antimicrobial Activity and Hydrophobicity of Edible Whey Protein Isolate Films Formulated with Nisin and/or Glucose Oxidase. *J. Food Sci.* **2013**, *78*, M560–M566. https://doi.org/10.1111/1750-3841.12078.
37. Lee, H.; Min, S.C. Antimicrobial edible defatted soybean meal-based films incorporating the lactoperoxidase system. *LWT-Food Sci. Technol.* **2013**, *54*, 42–50. https://doi.org/10.1016/j.lwt.2013.05.012.

38. Ayala, M.; Thomsen, M.; Pizzol, M. Life Cycle Assessment of pilot scale production of seaweed-based bioplastic. *Algal Res.* **2023**, *71*, 103036. https://doi.org/10.1016/j.algal.2023.103036.
39. El-Sheekh, M.M.; Alwaleed, E.A.; Ibrahim, A.; Saber, H. Preparation and characterization of bioplastic film from the green seaweed Halimeda opuntia. *Int. J. Biol. Macromol.* **2024**, *259*, 129307. https://doi.org/10.1016/j.ijbiomac.2024.129307.
40. Rajasekar, V.; Karthickumar, P.; Rose, A.H.R.; Manimmehalai, N.; Subhasri, D. Development and characterization of biodegradable film from marine red seaweed (*Kappaphycus alvarezii*). *Pigment. Resin Technol.* **2023**, *52*, 478–489. https://doi.org/10.1108/PRT-09-2021-0119/FULL/XML.
41. Zhang, L.; Wang, Z.; Jiao, Y.; Wang, Z.; Tang, X.; Du, Z.; Zhang, Z.; Lu, S.; Qiao, C.; Cui, J. Biodegradable packaging films with ε-polylysine/ZIF-L composites. *LWT* **2022**, *166*, 113776. https://doi.org/10.1016/j.lwt.2022.113776.
42. Karami, P.; Zandi, M.; Ganjloo, A. Evaluation of physicochemical, mechanical, and antimicrobial properties of gelatin-sodium alginate-yarrow (*Achillea millefolium* L.) essential oil film. *J. Food Process. Preserv.* **2022**, *46*, 16632. https://doi.org/10.1111/jfpp.16632.
43. Zheng, Z.; Zhang, O.; Borgs, C.; Chayes, J.T.; Yaghi, O.M. ChatGPT Chemistry Assistant for Text Mining and the Prediction of MOF Synthesis. *J. Am. Chem. Soc.* **2023**, *145*, 18048–18062. https://doi.org/10.1021/jacs.3c05819.