

## Research Article

**Cite this article:** Willocx M, Duflo J (2023). Free-text inspiration search for systematic bio-inspiration support of engineering design. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* **37**, e21, 1–13. <https://doi.org/10.1017/S0890060423000173>

Received: 16 December 2022

Revised: 24 March 2023

Accepted: 22 June 2023

### Keywords:



bio-inspired design; biologically inspired design; biomimicry; creativity and ideation; design support

### Corresponding author:

Mart Willocx;

Email: [mart.willocx@kuleuven.be](mailto:mart.willocx@kuleuven.be)

# Free-text inspiration search for systematic bio-inspiration support of engineering design

Mart Willocx<sup>1</sup>  and Joost R. Duflo<sup>1,2</sup> 

<sup>1</sup>Department of Mechanical Engineering, KU Leuven, Celestijnenlaan 300A, box 2422, Leuven 3001, Belgium and <sup>2</sup>Flanders Make, Gaston Geenslaan 8, Heverlee, Belgium

## Abstract

Current supportive bio-inspired design methods focus on handcrafting the inspiration engineers use to speed up bio-inspired design. However promising, such methods are not scalable as the time investment is shifted to an up-front investment. Furthermore, most proposed methods require the engineer to adopt a new design process. The current study presents FISh, a scalable search method based on the standard engineering design process. By leveraging machine translation between a representative corpus of biological and engineering texts, the engineer can start the search using engineering terminology, which, behind the scenes, is automatically converted to a biological query. This conversion is done using language models trained on patents and biological publications for the engineering and biology domains. Both models are aligned using the most used English words. The biological query is used to retrieve biological documents that describe the most relevant functionality for the engineering query. The presented method allows searching for bio-inspiration using a free-text query. Furthermore, updating the underlying datasets, models and organism aspects is automated, allowing the system to stay up to date without requiring interactive effort. Finally, the search functionality is validated by comparing the search results for the functionality of existing bio-inspired designs with their inspiring organisms.

## Introduction

Using bio-inspiration during the conceptual design phase increases the number of novel solutions generated (Vandevenne et al., 2016; Keshwani and Chakrabarti, 2017). In contrast with other design methods, which rely on the prior knowledge and experience of the designer, bio-inspired design offers a catalogue of proven designs, often using different working principles than the ones traditionally employed in a technical setting (Bogatyrev and Bogatyreva, 2014). Despite such promising results, bio-inspired design has not yet been widely adopted as a design strategy and only a limited number of bio-inspired products are commercially available (Wanieck et al., 2017).

Most engineers do not possess biological background knowledge. They have a hard time identifying, filtering and understanding relevant biological strategies to apply them to their design problem, turning the search for relevant biological strategies into a frustrating and time-consuming task (Goel and Helms, 2014; Fayemi et al., 2017; Kruiper et al., 2018; Graeff et al., 2019). Many methods and tools supporting engineers in retrieving relevant bio-inspiration for their design problem have been proposed (Lenau et al., 2018; Hashemi Farzaneh and Lindemann, 2019). However, most of these tools use a database of which the construction is not scalable (Vandevenne et al., 2016; Willocx et al., 2020).

In contrast with the multitude of bio-inspired design tools, only a few methods are used in a design context (Pentelovitch and Nagel, 2022). To more easily integrate with an existing design process used in a company, a tool integrated into the systematic engineering design process proposed by Pahl et al. (2007) is useful (Nagel et al., 2014). Furthermore, a free-text search system ensures that the user does not have to learn a specific vocabulary, considerably lowering the difficulty of uptake. The literature review in section two highlights that no tool currently offers free-text search, automatic translation between engineering and biology, and access to the whole biological literature.

Furthermore, Broeckhoven and du Plessis (2022) highlight that most bio-inspired designs are based on a limited number of highly publicized organisms. To counteract this, they proposed to digitize existing Natural History collections and look at more biodiverse organisms in the search for inspiration. Another opportunity is to use the already digitally available biological literature, used for example in Shu (2010) or Vandevenne et al. (2015).

The rise in open-access publishing (Laakso et al., 2011) and more permissive text-mining policies on subscription content of large publishers (Van Noorden, 2014) allow the automated extraction of interesting insights from more content. This creates an opportunity to build a bio-inspired design tool on top of a large and ever-growing collection of scientific biological literature. By

© The Author(s), 2023. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-ShareAlike licence (<http://creativecommons.org/licenses/by-sa/4.0>), which permits re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is used to distribute the re-used or adapted article and the original article is properly cited.

basing the tool on machine-processed publications and limiting the manual interaction required to update the underlying data, it would stay up to date with new developments in engineering or biology.

### Research goals and scope

This study describes a natural language search method based on the essential function that the engineering design must fulfill. Figure 1 schematically presents the Functional Inspiration Search method (FISh). The essential function of a design is defined by Pahl et al. (2007) as “the generalized crux of the problem”, and the definition thereof is part of the systematic design process. The proposed method automatically translates the functional engineering query into its biological equivalent and uses this to retrieve relevant biological publications. This allows FISh to be directly integrated into the systematic engineering design process. Furthermore, by automatically linking engineering and biology, the search tool can scale to encompass the complete biological literature and be easily updated with new literature.

Figure 1 highlights the different areas for which a research effort is required. These research efforts are driven by the following research questions:

- 1) A literature review on the currently existing bio-inspired design support tools for the search phase and the size of their underlying datasets.
- 2) Apply machine translation methods to link the engineering and biology domain.
  - a. Create a dataset that is representative of the currently available biological literature.
- 3) Automate the generation of organism aspects as described by Vandevenne et al. (2015) to allow updating the system with new biological developments.

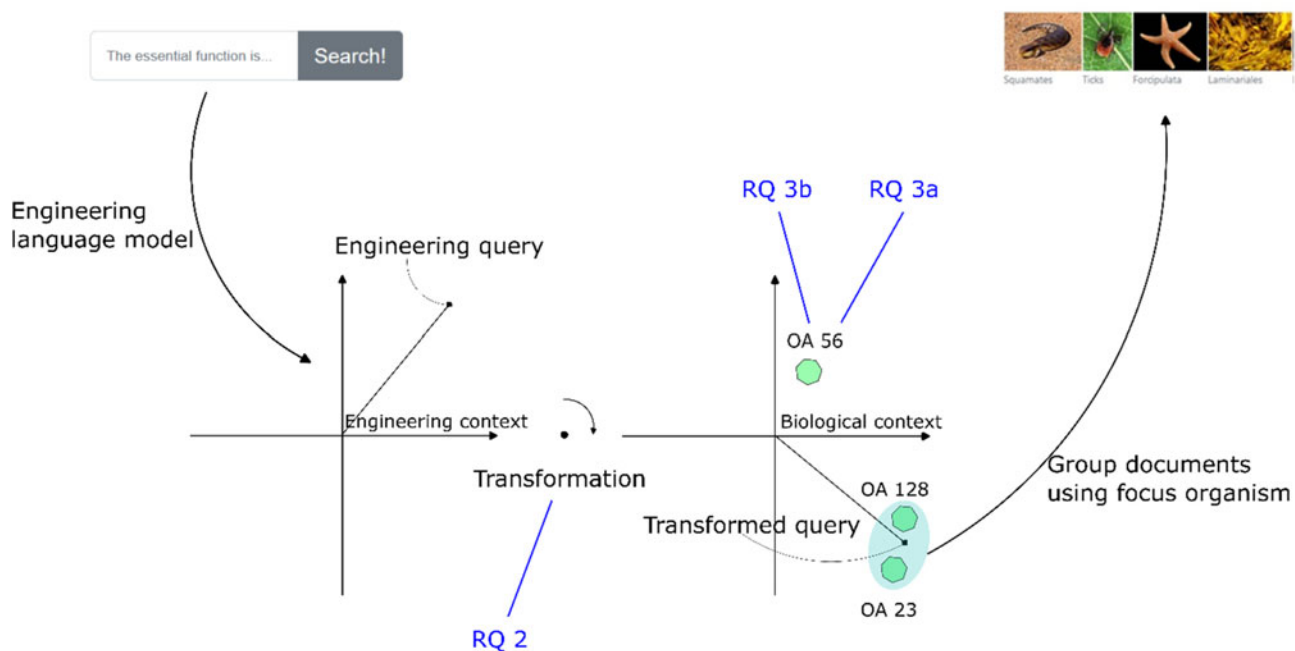
- a. Automatically determine the number of necessary organism aspects.
- b. Automatically filter the included lemmas to focus on functionality.

The scope of the tool here is limited to supporting mechanical engineers as a prototype. Furthermore, the validation of the search method is here limited to an assessment of the relevance of the retrieved biological publications and does not cover the potential to transfer this knowledge to the searching engineers.

The remainder of this research paper is structured as follows: first, the relevant literature on bio-inspired design support tools and natural language models is reviewed. Second, the proposed method of using statistical machine translation to support bio-inspired design is presented along with the used datasets to represent engineering and biology and align the spaces. Third, using existing bio-inspired designs and their inspiring organisms, the search method is validated. Finally, after providing conclusions, future research directions and possible alternative applications of the presented method are proposed.

### Background

First, the relevant bio-inspired literature is reviewed to document the steps required to integrate bio-inspiration in the design process and highlight the shortcomings and pitfalls of the current methods. To introduce the automatic linking method between biology and engineering, a short literature review focused on the natural language models used is presented. The natural language model selection is explained and the literature on employing the characteristics of the model to enable machine translation is reviewed.



**Figure 1.** Schematic representation of the approach proposed for the functional inspiration search method (FISh). Starting in the top left, a free-text functional query is converted to a vector representation (embedding) using the engineering language model and transformed to the biology domain using the linking method proposed in this work. Next, the most relevant organism aspects are selected and the documents retrieved are clustered based on their focus organisms and presented to the end user.

### Bio-inspired design tools

Systematically applying biological inspiration requires a consistent process that aids the engineer in finding and applying relevant biological inspiration to a certain, given engineering problem. This process is known as the engineering pull (Helms et al., 2009), where the engineering problem is the start of the search for bio-inspiration. While different authors have defined several phases in the systematic process (e.g., Lenau et al., 2018; Hashemi Farzaneh and Lindemann, 2019), they can be reduced to the four phases schematically presented in Figure 2 (Vandevienne et al., 2015). After formulating the problem and searching for bio-inspiration, the different retrieved strategies are analyzed and selected. Finally, the working principles need to be identified and transferred analogically to the engineering domain.

A major difficulty in bio-inspired design is retrieving and identifying relevant biological inspiration. The language used in engineering and biology is different (Fayemi et al., 2017) and engineers often lack the required biological terminology and knowledge to search for relevant biological systems effectively. This turns the search phase into a time-consuming task (Vattam and Goel, 2011). Furthermore, due to the large time investment required, engineers fixate on a single biological strategy, even when a better alternative is presented (Ahmed-Kristensen et al., 2014).

A first method to provide engineers with biological inspiration during the design process are design books with a list of interesting biological systems, for example (Nachtigall and Wissler, 2015). Another option would be to consult an expert biologist (Graeff et al., 2019). However, these time-consuming methods interrupt the engineering design process and are thus not likely to be employed.

In the following paragraphs, the existing bio-inspired design support tools will be reviewed concerning their search input or query, the type and size of the dataset they can search and their current availability. To limit the scope of the review, only recent supportive methods or tools that have an explicit contribution to supporting the search for relevant bio-inspiration are considered. Table 1 presents the results of this review.

In the function-based model tools, the engineering problem and the biological strategies are expressed in a functional model, and then, a model alignment is used to match the biological strategies with the model of the problem. The time-consuming nature of the modeling process (e.g., creating a single model takes between 40 and 100 h (Vattam et al., 2011)) causes the databases of, for example, UNO-BID to be limited to 42 biological models. Furthermore, before a biological strategy search can be started, the engineering problem will also have to be modeled.

The AskNature database uses a functional taxonomy to organize the biological strategies, which allows engineers to systematically find these by locating the corresponding function in the taxonomy. As noted by Deldin and Schuknecht (2014), the functional categories have been arbitrarily chosen. A more recent version of the website also allows searching based on keywords (Hooker and Smith, 2016). However, this runs into the difficulty that keyword-based search engines are not designed for cross-domain retrieval (Rugaber et al., 2016). Each biological strategy

that is entered into the database is pre-processed to be useful in an engineering design exercise. Again, this time-consuming process limits the number of biological strategies that can be added to the database.

BioScrabble overcomes the limited size of a prepared database by searching directly in PubMed Central, an open archive containing full-text articles from biomedical and life-sciences journals. The support tool expands the functional query using the synonyms present in WordNet and uses these extra keywords to ameliorate the query (Kaiser et al., 2013). Each query retrieves a large number of publications, requiring the design engineer to manually scan them (Kaiser et al., 2014).

The engineering to biology thesaurus provides a translation guide between the functional (engineering) basis and biological language use (Nagel et al., 2010). The functional basis consists of a limited vocabulary of 42 terms that are claimed to be a standard necessary to describe all engineering problems (Hirtz et al., 2002). This method was later expanded by using the biological keywords to search directly in natural language biological texts, extracting the relevant sentences directly (Nagel and Stone, 2012).

SEABIRD (Vandevienne et al., 2015) links biology and engineering by generating so-called product and organism aspects from patents and biological publications for representing respectively the engineering and biological domains. These aspects were then manually linked together, allowing for the input of one or more product aspects to yield the most relevant organism aspects and their associated publications. This search tool allows to access 11,000 biological publications. To formulate a query, the engineer must be familiar with the 300 product aspects and map the engineering design problem on those. Major drawbacks of the method are that, when updating the dataset with new engineering information, the product aspects will change to follow new developments in engineering (requiring retraining of the engineer) and the manual linking must be performed again.

Despite these drawbacks, the grouping property of organism aspects based on the functionality described in the biological documents is useful for the work at hand. This work will utilize the concept of organism aspects, but improve the generation by automating the lemma selection and build a theoretical foundation to automatically pick a relevant number of organism aspects. Furthermore, the automatic translation method proposed eliminates the need for product aspects and manual linking between engineering and biology.

A systematic approach that can deal with free text queries has the advantage of dealing with out-of-vocabulary words, not requiring the engineer to be familiar with a predefined vocabulary. Furthermore, when the engineering domain moves forward, the predefined vocabularies will age and not be able to represent new developments. For example, the domain of additive manufacturing is a recent development and cannot be reasonably modeled with the engineering-to-biology thesaurus. This illustrates the difficulty of keeping up to date with new engineering developments.

While Sun et al. (2022) provided a free-text search with keywords, there is no provision taken for the difference in vocabulary between engineering and biology. This is not necessary with the



Figure 2. The phases of the systematic bio-inspired design process as identified by Vandevienne et al. (2015).

**Table 1.** Overview of the different bio-inspired design support methods that describe a method for finding bio-inspiration. For each method, the input query and the search method are summarized. The type and size of the dataset that is accessible via the search method are based on the most recent publicly available data or the data that have effectively been prepared and described by the authors of the method.

Method	Search input (query)	Method	Type and size of dataset
Function-based models (SBF, SAPPHIRE, IDEA-Inspire, DANE, MBE, UNO-BID)	Function-based model of the required function	Model alignment via text similarity/similar functions	42 processed biological models (Rosa et al., 2015); 100 SAPPHIRE models (Siddharth and Chakrabarti, 2018); 40 partially completed DANE models (Goel et al., 2012)
AskNature (Deldin and Schuknecht, 2014)	Keyword/location in 3-level taxonomy	Keyword/taxonomy	1696 pre-processed biological strategies (Graeff et al., 2019)
BioScrabble (Kaiser et al., 2013)	Technical functions, properties and environments	Expands functions using synonyms from wordnet and uses this in a full-text search in the PubMed database	PubMed database (7 million publications)
Engineering-to-biology thesaurus (Nagel et al., 2010)	A limited set of functional keywords	A direct translation of engineering keywords to corresponding biological keywords. Then a keyword-based internet search	Everything indexed by the used search engine
A computational approach to bio-inspired design (Nagel and Stone, 2012)	Function-component interactions (based on the limited set of functional keywords)	Using the translation of the engineering to biology thesaurus to locate relevant sentences in a natural language biological source	One undergraduate-level biological textbook
Natural language approach to biomimicry (Chiu and Shu, 2007; Cheong and Shu, 2014)	Precomputed list of biologically meaningful keywords and causal function verbs	Using the precomputed keywords, the analogies are retrieved from the textbook	One undergraduate-level biological textbook
BioDesign (Sun et al., 2022)	Free-text keywords	BERT-based method to link to AskNature	Processed strategies from the AskNature database
SEABIRD (Vandevenne et al., 2015)	Product aspects (set of 300 keywords)	Links biology and engineering by generating Product aspects and Organism aspects and linking those	11,000 open-access biological publications

use of the pre-processed strategies for engineering application in AskNature, however, given the slow pace of growth of AskNature, it is currently unclear if this method can scale beyond the AskNature database.

In conclusion, while several search support tools have been built, no tool offers free-text search, automatic translation, and access to the whole biological literature.

### Natural language models

To automate the free-text translation from engineering vocabulary to biology, a method to mathematically represent both domains is required. This section first explains the main idea of representing language using embeddings, then reviews the recent language models created to perform this task. Next, the choice for the specific language model is underpinned. Finally, the literature on leveraging the characteristics of the chosen language model to translate between different languages is reviewed. This will later be applied between domains within the same language.

Word embeddings are a method of representing words in machine learning tasks. To create a word embedding, the given word is associated with a mathematical object which allows us to perform a natural language processing task. Often this is a vector in which each of the dimensions includes information about the word, its meaning or context (Turian et al., 2010). For example, using the naïve method of one-hot coding the words cat, dog, and house would result in the respective vectors  $\langle 1,0,0 \rangle$ ,  $\langle 0,1,0 \rangle$ , and  $\langle 0,0,1 \rangle$ .

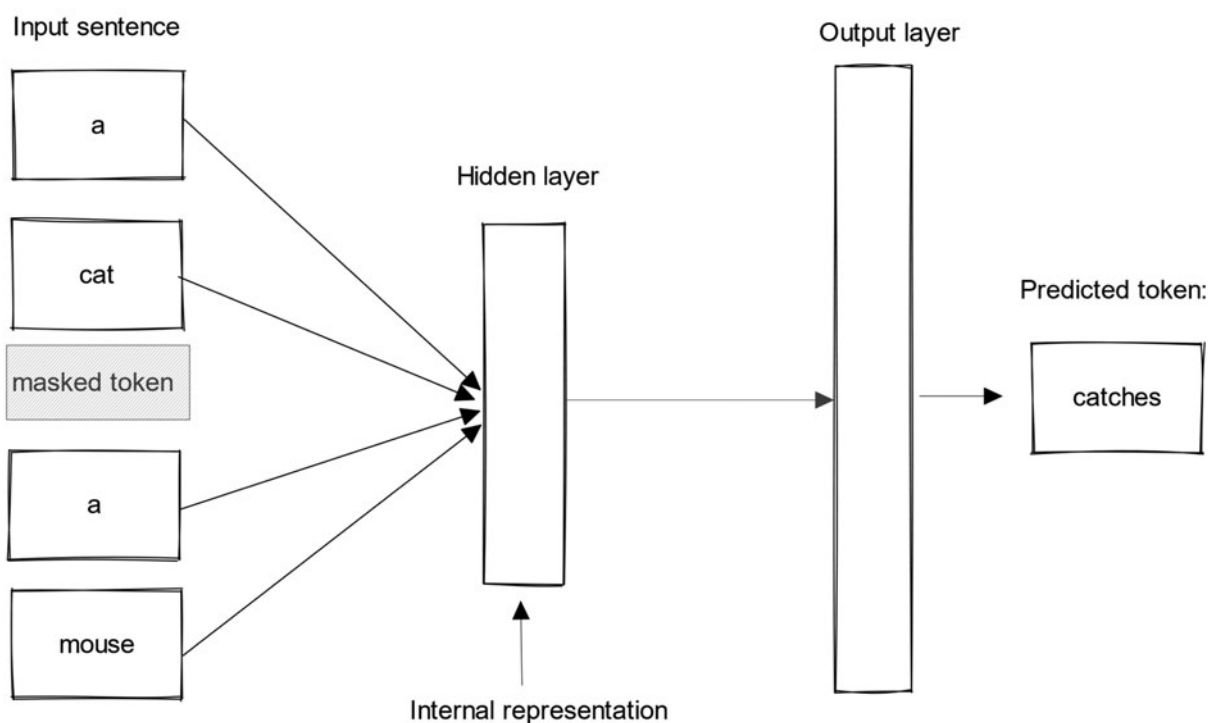
The distributional hypothesis assumes that the meaning of words is determined by the context in which they occur (Landauer et al., 2007). Recent developments in language

representation for AI models allow converting a word to a vector representation which captures the meaning using the context in which the word occurs (Karani, 2018). Different methods for creating this vector exist and a more comprehensive overview is given by Devlin et al. (2018). Here, the focus is placed on the methods that will be employed in the remainder of the study.

Mikolov et al. calculated the vector for each word by performing a word prediction task with a shallow neural network as represented in Figure 3. The network must predict the word based on the given context. The hidden projection layer for each word is assigned to be the word embedding, enabling the resulting vector to capture the information about the context in which the word is used in the text on which the training is performed. This causes synonymous words and words with a similar conceptual meaning to be grouped in the space spanned by the vectors of the language model (Mikolov et al., 2013a).

Bojanowski et al. (2016) expand the word prediction task from Mikolov et al. (2013a) to consider subword information by using character n-grams instead of the complete word. This has the consequence of extending the method to be able to deal with unseen (out-of-vocabulary) words. Due to this property and the public availability of the code to create language models, this method is selected to be used. A known disadvantage of this method is that lemmas with multiple meanings only are assigned one vector.

Here, a conscious choice to use context-free word embeddings is made. For example, the more recent BERT word representation system takes the context of the word into account when assigning a context vector to a word (Devlin et al., 2018). However, as the expected search queries will consist of functional keywords and possibly a limited set of related words, a query does not set up



**Figure 3.** Word prediction neural network as employed by Mikolov et al. The tokens surrounding a masked token in a sentence are presented to the neural network. The training objective is to predict the masked lemma, leading the neural network to form an internal representation of the masked token. This internal representation becomes the word embedding for a given token.

a context. Furthermore, by using a larger context vector, aligning the spaces (as described in the next section) would require a larger parallel corpus. Finally, comparing two vectors with the cosine similarity rapidly becomes less useful when the dimensionality increases (Houle et al., 2010).

By utilizing a small bilingual corpus shared between two different languages, the transformation necessary to link the vectors of the known words from one space to another can be calculated. Mikolov et al. (2013b) proposed to calculate the transformation matrix by solving the least squares problem with the embeddings of the word pairs in both language models. This transformation allows the translation of any word vector from one model to another. By selecting the closest known word in the target language, machine translation between both languages is achieved. Expanding on this work, Smith et al. (2017) proved that the required transformation is orthogonal and provide a robust calculation of the transformation matrix using the singular value decomposition of the product of the matrices of the embeddings of the known word pairs.

From the success of using this method to reliably translate between two similar languages (Smith et al., 2017), here it is theorized that this method can also be used to translate between two domains using different jargon in the same language. Assembling a parallel corpus to align the spaces is a challenge since there is no clear “translation” possible between the domain of engineering and biology. This challenge is tackled in the next paragraphs.

### Linking engineering and biology

As the language used in the engineering and biology domains is different (Nagel, 2014; Helfman Cohen and Reich, 2016) and

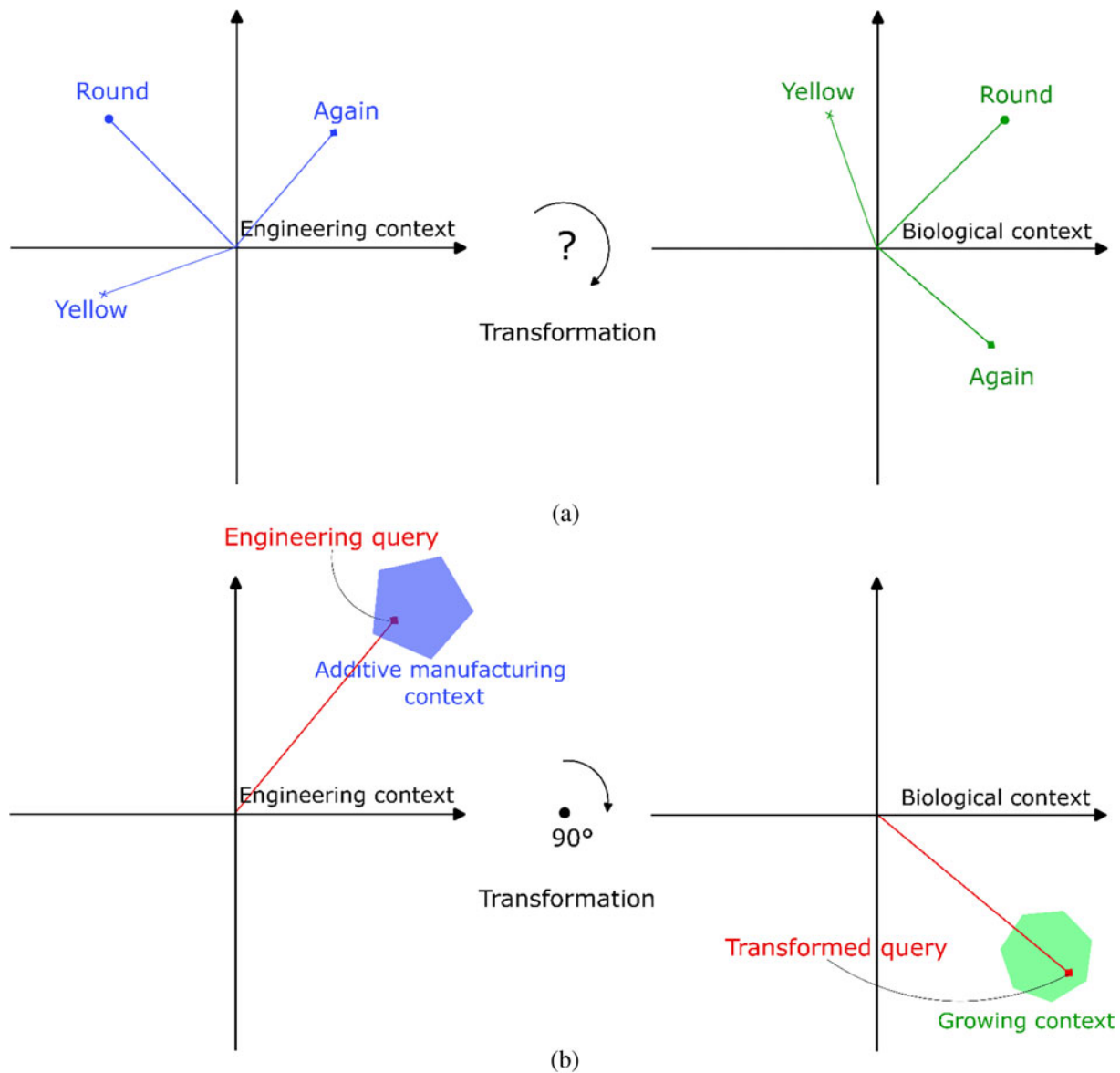
literature about both areas is prevalent, the hypothesis that machine translation between both domains can translate a query between engineering and biology is posited. It is proposed here to train a language model representing the biological domain using biological publications. Similarly, the engineering domain’s representative language model is trained on publicly available patents, representing the latest engineering innovations.

As both language models represent different domains but are rooted in English, the assumption that the most used words retain their meaning between both models is made. Now, the most used words can be used to calculate the transformation between both models, which allows the translation of the functional query to the biology domain.

Figure 4 schematically presents the proposed alignment of both spaces by using the top words and an illustrative example of the translation of an engineering query. Via the alignment, the functional query for a filtration system is transferred from the engineering language model to the conceptually similar location in the biological model.

The functions described in the biological publications are grouped by performing the organism aspect generation procedure described in Vandevienne et al. (2015). These organism aspects are then located in the biological space, completing the link. After mapping the query from engineering to biology, the closest organism aspects are selected and used in a weighted calculation to determine the most relevant biological documents. By using the organism aspects instead of the content of the publication, the focus is placed on the relevant functions.

The next sections detail the data and methods used to represent biology and engineering and the pre-processing of the language and alignment data.



**Figure 4.** Schematic representation of the alignment of the biological and engineering language models using the most frequent words in both languages. (a) Illustrates using the vectors of the most used words in English to determine the transformation required to go from model to model. Here, the alignment of both spaces is illustrated as a 90° clockwise rotation from engineering to biology. In reality, both domains are represented by a 300-dimensional space and this translation is performed by executing a matrix multiplication using the transformation matrix. (b) To illustrate the transformation of a query from engineering to biology, an example query related to additive manufacturing is presented. The colored polygons illustrate the contexts of terms related to additive manufacturing and growing for engineering and biology, respectively. By using the transformation determined in the previous step, the query vector is transferred from engineering to biology and ends up near the biological context related to growth.

### Corpora

To compile the corpus of biological documents, the reference list of biological journals from Scopus is downloaded and processed to exclude journals not published in English. Furthermore, as the scope of useful inspiration in the biology corpus is limited to mechanical engineering in this research, any journals covering molecular biology or human health are also excluded. By leveraging the metadata service Crossref, 2.5 million potential biological documents were identified. Of these documents, not all are available in electronic form, nor have a license that allows text mining. To limit the number of individual text-mining agreements that had to be reviewed, the largest publishers, Elsevier, Springer, Wiley, and Nature, were selected.

Additionally, the *Journal of Experimental Biology* is also included in the dataset. This results in a mix of open-access and subscription publications.

The next selection is based on the content of the documents. First, using a language classifier, texts not in English are excluded. Second, texts shorter than 500 words are also excluded. Third, the focus organism detection algorithm described by Vandevenne et al. (2014) is used to detect the focus organisms in the title and abstract of the publication. The organism detected in the title or most mentioned in the abstract is taken as the focus organism of the publication. Documents not containing a focus organism are finally also excluded. In the end, this procedure yields a corpus of 161,342 publications.

To compile the engineering corpus, patents were downloaded from the United States Patent Office (USPTO). To create a representative corpus which does include the latest innovations but also takes older patents into account, the dataset was built biased toward newer patents. Starting from 2001, for each year, a random month was selected to download all the patents published in that month. Next, all the patents of 2020 were also included in the dataset. To ensure that both models contain approximately the same amount of training data, patents are randomly drawn from the set of available patents until the same amount of text as contained in the biological corpus is attained. Another option would be to use engineering publications to represent the engineering domain, however, this lowered the performance during screening experiments. It is theorized that the research publications do not allow the model to represent functions as well as patents. For future research, it might be interesting to compare the effectiveness of the models for the representation of other types of queries.

### Pre-processing and model generation

Before training the language models, the corpora are pre-processed to remove organism mentions, group common bigrams and reduce the vocabulary. Organism mentions are removed by using the organism detection described in Vandevenne et al. (2014) so the language model does not use the species to define the context.

Common bigrams are joined together to better capture meaning in the resulting language model. This is done based on a statistical model where the presence of one word strongly implies the other. For example, when “additive” and “manufacturing” are together, their meaning is compounded, and they are joined together. This is done for both the patents and the biological documents based on mutual information with a cut-off value determined on a knee point analysis of the mutual information of lemmas in the biological documents (Mei et al., 2007).

Finally, lemmatization is performed by the Spacy package (Montani et al., 2022). The stemming combines different versions of the same root word, collating more word use information together, allowing the language model to better model the meaning of the word.

Using the processed texts, the language models are trained using the code provided by Bojanowski et al. (2016). Here, the dimension of 300 is chosen since this is the same as reported by Bojanowski et al. (2016) and this gave the best results in a screening experiment with the complete system.

The proposed procedures are automated and can be repeated, for example, every 6 months, to keep the corpora and models up to date with new biological research and engineering developments. By updating this model with the latest released patents, the search functionality will be able to understand queries along the state of the art.

### Organism aspects

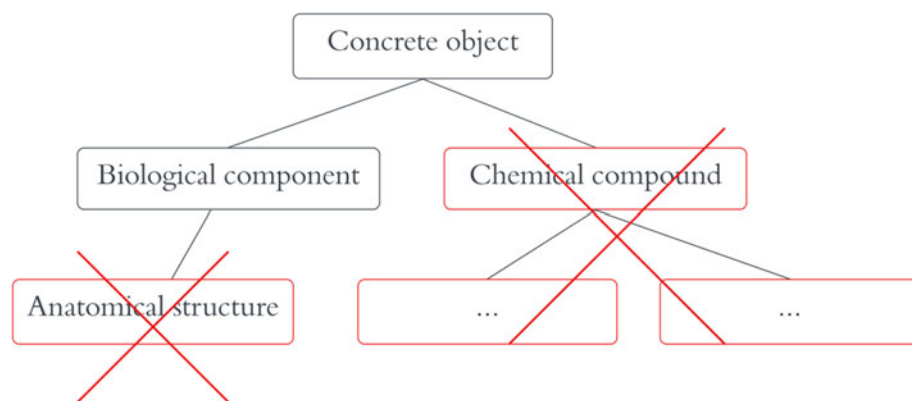
For the generation of the organism aspects, the vocabulary in the processed texts is further filtered based on the part-of-speech tagging provided by the Spacy package and the usefulness to explaining functions of the words: verbs (Shu, 2010), nouns (Vandevenne et al., 2015), and adjectives (Ke et al., 2010) are kept. Adverbs are explicitly removed as they tend to dominate an aspect without adding information. Lemmas that appear less than ten times in the corpus are also excluded.

While Vandevenne et al. (2015) relied on a further manual categorization of the lemmas that were useful for the transfer from the biological domain to further filter the lemmas included in the organism aspect generation, here an automatic categorization method is proposed by leveraging the hierarchy of concepts found in WikiData (Wikimedia Foundation, 2020). A hierarchy of concepts groups words into their respective conceptual categories. For instance, the abdomen of an arthropod is, climbing the taxonomy, a tagma, an anatomical structure, a biological component ...

The hierarchy of concepts allows defining a few rules to filter out unwanted categories of lemmas, as visualized in Figure 5. Using the manual annotation work of Vandevenne allowed to start finding the categories that must be excluded. This resulted in a list of 35 categories and their subcategories that are excluded from the vocabulary used to generate the organism aspects. This list was compiled in September 2020 and is included in Appendix A.

Finally, it is important to determine the number of organism aspects that can be generated, balancing ignoring nonsense aspects and leaving behind aspects that describe real functions. Vandevenne et al. (2015) arbitrarily chose 300 aspects, however, the number has to be adjusted to the information contained in the dataset. The optimum is thus expected to be dependent on the number of biological publications present in the dataset.

To determine an upper limit on the number of organism aspects to generate, a parallel analysis using the real data and synthetic data based on the real dataset is used (Franklin et al., 1995). These synthetic data are generated based on the frequencies of the lemmas in the real data, but they are randomly spread over the different documents. This causes the synthetic data to represent



**Figure 5.** Lemma filtering based on a taxonomy of concepts. The taxonomy shown is based on the WikiData taxonomy, excluding chemical compounds and anatomical structures, based on the lemmas excluded by Vandevenne et al. (2016).

documents which are indiscernible from noise but are drawn from the same distribution as the real data.

The generation of the organism aspects is based on a principal component analysis, where cumulatively each additional aspect captures more variance than the last. By comparing the variance captured in the organism aspect generated by the real data and the synthetic data, the cut-off point where the aspects start to capture noise is determined. When the corrected variance captured from the generated data is more than that of the real data (Jolliffe, 2002), the cut-off number of organism aspects is reached (Horn, 1965; Buja and Eyuboglu, 1992). This analysis is repeated several times to determine the 5% quantile where the cut-off is placed. For the current corpus of biological documents, this cut-off is determined to be at 3403 organism aspects.

Each organism aspect is loaded on a limited number of functional lemmas, which are used to represent the organism aspect in the biological space. The weighting of the lemmas is considered when determining the embedding by weighting the contribution of each lemma's vector. The matching between an engineering query and the organism aspects is performed based on the cosine distances between the translated query and the vectorial representations of the organism aspects.

By presenting the automatic selection method of the vocabulary that is relevant to represent the functionality in the biological literature and selecting the number of organism aspects that must be generated, this work contributes an automatic method of recalculating the organism aspects. Furthermore, this method could also be applied to the product aspects presented in Verhaegen et al. (2011) to keep them up to date.

### Space alignment and representation

The 1500 most used lemmas in the Corpus of Contemporary American English (COCA) (Davies, 2019) are used to align the engineering and biology space. While approximately 40% of the most used lemmas in engineering and biology are also present in the most used words overall, the remainder of the most used lemmas differ between both domains. This motivates the use of the independent COCA corpus (Davies, 2019) to determine the lemmas used to determine the transformation between both models. The transformation matrix is determined using the procedure presented by Smith et al. (2017).

The assumption that these lemmas are mapped to the same context in both domains was verified by using a 20/80% test-train split where the test set was mapped with a precision@5

(Manning et al., 2009) score near the same lemmas of 89% using the transformation trained on the train set.

The contributions mentioned above allow the translation system to be updated without manual intervention other than supplying the new source documents for both domains. This is important to keep up with the ever-changing language use. Delobelle et al. (2022) for example, find that a 3-year-old language model underperforms on more recent benchmarks. However, this could be solved by adding new tokens to the model (reflecting the changing word use) and training the model with newer data.

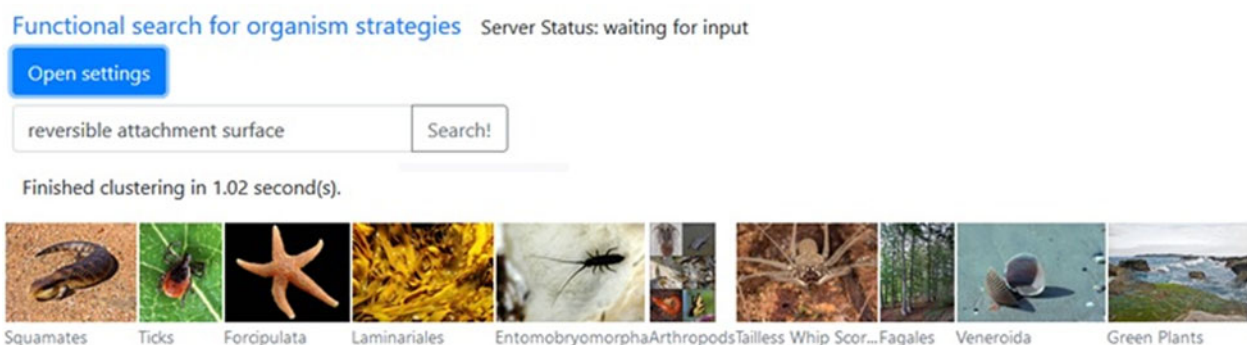
### Interface

The search method presented in this work returns multiple biological publications. As this can cause information overload on the processing engineer (Willocx et al., 2020), an interface based on the work of (Vandevenne et al., 2015) was added to the tool.

Vandevenne et al. (2015) assumed that the publications on the same focus organism retrieved for a search query deal with the same biological strategy. To exploit this, the documents are clustered based on the detected focus organism and the location of the organism in the NCBI taxonomy (Schoch et al., 2020). To avoid having too specific clusters with too few documents, the clustering per organism is performed on the family level. If a focus organism is specified more specifically, the taxonomy tree is climbed until the family level is reached. This groups the more specific levels of the tree together. For broad publications located higher in the taxonomy, for example, treating a whole class of organisms, the tree is not descended. This has the drawback of creating small clusters containing broad publications.

Furthermore, this clustering method does not group organisms that have arrived at the same working principle by parallel evolution, nor the working principles that are broadly adopted in the taxonomy (Willocx and Duflou, 2021). Organisms that employ multiple working principles are also grouped together, as shown in the discussion and in Willocx et al. (2020). However, for the current study, the authors feel that this is a better balance than not clustering the documents at all. An interesting future research direction can be to cluster the documents based on the working principle and not the focus organism.

In the interface, these clusters are represented by an image retrieved from Wikimedia Commons, in contrast to the previous work by Vandevenne et al. This image does not necessarily present the biological strategy employed by the organism, however, serves to give the designer a visual cue. Figure 6 presents the



**Figure 6.** Representation of the focus organisms each representing a cluster of documents retrieved for the search query for "reversible attachment surface". The organism images were retrieved from Wikimedia Commons under either a public domain or Creative Commons license.<sup>1</sup>



results of the search query for “reversible attachment surface”. By clicking on the image of the focus organisms, the designer is presented with the list of biological publications linked to the focus organism, along with links to read them.

### Validation using existing biologically inspired designs

The proposed search tool is validated by performing a search for the functionality provided by known bio-inspired designs and comparing the output of the tool with the original inspiring organism. The resulting rank of the inspiring organism is reported in the search output for the different functional queries aiming for the functionality of the bio-inspired design. Note that the goal here is not to place those organisms directly at rank one: other strategies might also be (more) relevant for the targeted function. The original inspiring organism should be returned with a reasonable rank to assert that the proposed mapping method returns functionally relevant documents linked to known solutions.

As noted previously, the query is the essential function that is performed by the bio-inspired design. The essential function was determined after consulting (Nachtigall and Wisser, 2015), which contains examples of bio-inspired design and explains the function that is taken from the biological system. From this explanation, one or two essential functions were distilled. To avoid overly specific queries biased toward the bio-inspired solutions, care was taken to keep the functional queries used in the validation simple. Table 2 presents the resulting essential functions that were extracted.

Furthermore, the label of the functional category that contains the inspiring strategy in the AskNature database is also taken as a functional query. The functional labels used in the AskNature taxonomy are quite broad and were assigned by independent contributors (Deldin and Schuknecht, 2014). This makes the label of the

functional categories an unbiased functional query. The resulting rank of using the label as a query is reported in Table 2.

The bio-inspired designs used in the validation were chosen for their prevalence as examples given when discussing bio-inspired design. As they have been used many times as an example of bio-inspired design, they stood the test of time. Table 2 lists the selected existing bio-inspired designs, their essential function, the inspiring organisms, and the rank of these organisms in the clustered search results.

The ranks of the inspiring organisms in most of the validation cases are within a reasonable number of organism clusters to be checked by the engineer. Again, the goal of the search tool is not to present the single best biological strategy but to offer a list of strategies that the designer can use to quickly select the strategy that resonates most with the problem at hand, causing them to quickly go through 10–15 groups of organisms.

The last column reports on the rank of the organism’s cluster when using the relatively broad labels of the functional categories presented in the AskNature taxonomy. These ranks are well within the number of strategies already present in some categories in the AskNature database. Since a search using the taxonomy also requires the designer to evaluate an equal number of strategies, this is a reasonable result.

To illustrate the contents of the formed organism clusters and highlight the difficulties that still arise when processing this bio-inspiration, Tables 3 and 4 present the clusters formed for the butterfly family for the queries “modify color” and “display color”, respectively. These clusters were formed by grouping the documents based on the family of the focus organism, as described in the previous section. Here, it is noted that while some documents describe the coloration of butterflies and how to control it on a biological level, other documents also describe using colors to perform cooling and color vision in the focus organism. The grouping based on the focus organism mixes different

**Table 2.** A list of bio-inspired designs, their inspiring organisms, the extracted essential function, and the rank of the inspiring organism’s cluster in the results for the given queries. Furthermore, the label of the AskNature functional category containing the strategy was also used as a query and the rank of the organism’s cluster for this query is also reported.

Bio-inspired design	Inspiring organism(s)	Essential function (query)	Rank of the organism’s cluster	Corresponding functional label in the AskNature Taxonomy	Rank of organism’s cluster using the AskNature functional label as query
Mirasol displays show vibrant colors with structures (Vandevenne et al., 2015)	Morpho butterfly	Display color	12	Modify color	10
		Render image	13		
The Eastgate centre uses thermal gradients based on termite hills to provide ventilation (Vandevenne et al., 2015)	Termites	Ventilation	15	Distribute gasses	49
Self-sharpening shredder stays sharp while cutting by using an erosion pattern based on rodent teeth (Shu et al., 2011)	Rodents	Self-sharpening	29	Manage mechanical wear	8
		Resist wear	4		
Hydrophobic surfaces based on lotus leaves stay clean (Shu et al., 2011)	Lotus leaves	Hydrophobic	16	Protect from excess liquids	55
		Stay clean	40		
Needle based on mosquito is painless (Shu et al., 2011)	Mosquito	Inject	9	Move in/on solids	43
Woodwasps inspired a drill that does not induce a large axial force (Shu et al., 2011)	Woodwasp	Drilling	51	Move in/on solids	37
		Pierce	5		
Winglets and other drag-reducing elements of modern aeroplane wings are based on the wings of birds of prey (Slosar, 2021)	Birds of prey	Reduce drag	16	Move in/through gases	27
		Reduce vortices	16		

**Table 3.** Documents contained in cluster 10 retrieved for the query “modify color”

Rank	DOI	Title
1	10.1016/0022-1910(87)90030-8	Hormonal control of seasonal morphs by the timing of ecdysteroid release in <i>Araschnia levana</i> L. (Nymphalidae: Lepidoptera)
2	10.1186/s40851-016-0040-9	Distal-less induces elemental color patterns in <i>Junonia</i> butterfly wings
3	10.1016/j.jtherbio.2011.06.002	Heat-shock-induced color-pattern changes of the blue pansy butterfly <i>Junonia orithya</i> : physiological and evolutionary implications
4	10.1007/s10905-015-9519-z	The relative importance of flower color and shape for the foraging monarch butterfly (Lepidoptera: Nymphalidae)
5	10.1016/j.jtherbio.2016.01.007	Cool bands: wing bands decrease rate of heating, but not equilibrium temperature in <i>Anartia fatima</i>
6	10.1016/0003-3472(70)90071-0	Colour selection and learned feeding preferences in the butterfly, <i>Heliconius charitonius</i> Linn

This cluster was formed based on the NCBI taxonomy.

**Table 4.** Documents contained in cluster 12 retrieved for the query “display color”

Rank	DOI	Title
1	10.1186/s40851-016-0040-9	Distal-less induces elemental color patterns in <i>Junonia</i> butterfly wings
2	10.1007/s10905-015-9519-z	The relative importance of flower color and shape for the foraging monarch butterfly (Lepidoptera: Nymphalidae)
3	10.1016/0003-3472(70)90071-0	Colour selection and learned feeding preferences in the butterfly, <i>Heliconius charitonius</i> Linn
4	10.1016/0022-1910(72)90038-8	The neural basis of colour vision in the butterfly, <i>Heliconius erato</i>
5	10.1007/s00265-019-2648-1	Spontaneous colour preferences and colour learning in the fruit-feeding butterfly, <i>Mycalesis mineus</i>

working principles. Furthermore, the current presentation format of a list of documents requires a lot of processing for the designer to find out what documents apply to their design problem.

A further validation where the output of the FISH search tool is compared with the bio-inspiration found by an expert biologist and AskNature is presented in Willocx et al. (2020). The results of that evaluation indicate that the FISH search tool does retrieve all of the biological strategies present in AskNature and proposed by the biologist, but that there is too much information for the designer to identify all of the relevant strategies on the first pass (Willocx et al., 2020). A second attempt at interpreting the documents did reveal the strategies mentioned in AskNature.

An interesting avenue to support the engineer to better identify the functional relevance of the biological strategy could consist of a summary generated from the biological documents in the organism cluster. By allowing the engineer to grasp the content quickly, the time investment for each strategy can be reduced.

## Conclusion

The presented FISH system allows a free-text search with an engineering functional query in the biological literature. Each component can be updated automatically, allowing the system to be kept up to date with new engineering developments and biological publications.

Key contributions are the transformation between two domain-specific languages based on the most common English words, the automatic filtering of relevant lemmas and automatically determining the required number of organism aspects. Finally, the proposed system was validated using already existing biologically inspired designs.

Care should be taken to comply with all the different copyright restrictions, however, with the rise of open-access publishing with

permissive licenses, the automated remixing of biological publications will open exciting avenues for bio-inspired design.

## Outlook

As noted previously and in Willocx et al. (2020), the search method returns a list of biological documents. While they are relevant to the query, this is often unclear to the designer due to the information overload of receiving too many documents. A future research goal is to process these documents further into a more manageable summary for the engineer. However, care should be taken to comply with the “No Derivatives” (ND) licence some papers are published under Willocx (2021).

The developed method could also be applied to different domains. For example, palaeontology can be considered a source domain for other analogies, resulting in the recently coined Paleomimetics. Perricone et al. (2022) describe a conceptual framework to transfer biomimetics to the palaeontology domain. This method still suffers from difficulties in identifying the organism from which the principles will be used. The authors propose to create a database consisting of extinct organisms, but with a corpus consisting of palaeontology publications, the proposed search system could be integrated and the time investment of creating a new database avoided.

Furthermore, with a focus on different innovations, we speculate that a transfer between the chemical engineering domain and the – here explicitly excluded – domain of biochemistry might yield some interesting results, certainly in the domain of catalysis.

Finally, the very recent developments in the field of chatbots powered by large language models open some interesting avenues for use in bio-inspired design. Despite the enormous potential, the currently available chatGPT is a black box model that might supply wrong answers (Kitamura, 2023). While asking the

model directly for an analogy risks running into a confident, but untrue answer, asking the model to extract the working principles from the retrieved documents is a more certain route. By carefully crafting the right prompt, the chatbot might be able to annotate the summary with a reference to the source of the idea. A similar approach is already in use in the perplexity.ai search engine (Srinivas, 2022), which generates a short answer or summary annotated with references from the search results for a general search query.

**Data availability statement.** The code, DOI identifiers, and other links to the content used in this study are available upon reasonable request from the corresponding author. The full-text content used to generate the models presented in this study is available from their respective publishers. Restrictions apply to the availability of these data, which were used under licence for this study.

**Acknowledgements.** This work was supported by the Fraunhofer-Gesellschaft Think Tank project BioMANU II.

**Competing interests.** The authors declare no competing interests.

## References

- Ahmed-Kristensen S, Christensen BT and Lenau TA (2014) Naturally original: stimulating creative design through biological analogies and random images. In DS77: Proceedings of the DESIGN 2014 13th International Design Conference, Vol. 13. Dubrovnik: Design, pp. 427–436.
- Bogatyrev N and Bogatyreva O (2014) BioTRIZ: a win-win methodology for eco-innovation. In Azevedo SG, Brandenburg M, Carvalho H and Cruz-Machado V (eds), *Eco-Innovation and the Development of Business Models*. Cham: Springer International Publishing, pp. 297–314. doi:10.1007/978-3-319-05077-5\_15.
- Bojanowski P, Grave E, Joulin A and Mikolov T (2016) Enriching word vectors with subword information. ArXiv:1607.04606 [Cs], July. <http://arxiv.org/abs/1607.04606>.
- Broeckhoven C and du Plessis A (2022) Escaping the labyrinth of bioinspiration: biodiversity as key to successful product innovation. *Advanced Functional Materials* 32, 2110235. doi:10.1002/adfm.202110235
- Buja A and Eyuboglu N (1992) Remarks on parallel analysis. *Multivariate Behavioral Research* 27, 509–540. doi:10.1207/s15327906mbr2704\_2
- Cheong H and Shu LH (2014) Retrieving causally related functions from natural-language text for biomimetic design. *Journal of Mechanical Design* 136, 081008. doi:10.1115/1.4027494
- Chiu I and Shu LH (2007) Biomimetic design through natural language analysis to facilitate cross-domain information retrieval. *AI EDAM* 21, doi:10.1017/S0890060407070138
- Davies M (2019) Word frequency data from The Corpus of Contemporary American English (COCA). <https://www.wordfrequency.info>.
- Deldin J-M and Schuknecht M (2014) The AskNature database: enabling solutions in biomimetic design. In Goel AK, McAdams DA and Stone RB (eds), *Biologically Inspired Design: Computational Methods and Tools*. London: Springer London, pp. 17–27. doi:10.1007/978-1-4471-5248-4\_2.
- Delobelle P, Winters T and Berendt B (2022) RobBERT-2022: updating a Dutch language model to account for evolving language use. arXiv. <http://arxiv.org/abs/2211.08192>.
- Devlin J, Chang M-W, Lee K and Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv:1810.04805 [Cs], October. <http://arxiv.org/abs/1810.04805>.
- Fayemi P-E, Wanieck K, Zollfrank C, Maranzana N and Aoussat A (2017) Biomimetics: process, tools and practice. *Bioinspiration & Biomimetics* 12, 011002. doi:10.1088/1748-3190/12/1/011002
- Franklin SB, Gibson DJ, Robertson PA, Pohlmann JT and Fralish JS (1995) Parallel analysis: a method for determining significant principal components. *Journal of Vegetation Science* 6, 99–106. doi:10.2307/3236261
- Goel AK and Helms ME (2014) Theories, models, programs, and tools of design: views from artificial intelligence, cognitive science, and human-centered computing. In Chakrabarti A and Blessing LTM (eds), *An Anthology of Theories and Models of Design: Philosophy, Approaches and Empirical Explorations*. London: Springer, pp. 417–432. doi:10.1007/978-1-4471-6338-1\_20.
- Goel AK, Vattam S, Wiltgen B and Helms M (2012) Cognitive, collaborative, conceptual and creative — four characteristics of the next generation of knowledge-based CAD systems: a study in biologically inspired design. *Computer-Aided Design* 44, 879–900. doi:10.1016/j.cad.2011.03.010
- Graeff E, Maranzana N and Aoussat A (2019) Engineers' and biologists' roles during biomimetic design processes, towards a methodological symbiosis. *Proceedings of the Design Society: International Conference on Engineering Design I*, 319–328. doi:10.1017/dsi.2019.35
- Hashemi Farzaneh H and Lindemann U (2019) *A Practical Guide to Bio-Inspired Design*. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-662-57684-7.
- Helfman Cohen Y and Reich Y (2016) *Biomimetic Design Method for Innovation and Sustainability*. Springer International Publishing. doi:10.1007/978-3-319-33997-9.
- Helms M, Vattam SS and Goel AK (2009) Biologically inspired design: process and products. *Design Studies* 30, 606–622. doi:10.1016/j.destud.2009.04.003
- Hirtz J, Stone RB, McAdams DA, Szykman S and Wood KL (2002) A functional basis for engineering design: reconciling and evolving previous efforts. *Research in Engineering Design* 13, 65–82. doi:10.1007/s00163-001-0008-3
- Hooker G and Smith E (2016) Asknature and the biomimicry taxonomy. *INSIGHT* 19, 46–49. doi:10.1002/inst.12073
- Horn JL (1965) A rationale and test for the number of factors in factor analysis. *Psychometrika* 30, 179–185. doi:10.1007/BF02289447
- Houle ME, Kriegel H-P, Kröger P, Schubert E and Zimek A (2010) Can shared-neighbor distances defeat the curse of dimensionality? In Gertz M and Ludäscher B (eds), *Scientific and Statistical Database Management*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 482–500. doi:10.1007/978-3-642-13818-8\_34.
- Jolliffe, I. 2002. *Principal Component Analysis*. 2nd Edn. Springer. doi:10.1007/b98835.
- Kaiser MK, Farzaneh HH and Lindemann U (2013) BIOscrabble: extraction of biological analogies out of large text sources. In Proceedings of International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, p. 11. <https://mediatum.ub.tum.de/doc/1188707/1188707.pdf>.
- Kaiser MK, Hashemi Farzaneh H and Lindemann U (2014) Bioscrabble – the role of different types of search terms when searching for biological inspiration in biological research articles.
- Karani D (2018) Introduction to word embedding and Word2Vec. *Towards Data Science*. September 1, 2018. <https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa>.
- Ke J, Chiu I, Wallace JS and Shu LH (2010) Supporting biomimetic design by embedding metadata in natural-language Corpora. In Volume 5: 22nd International Conference on Design Theory and Methodology; Special Conference on Mechanical Vibration and Noise. Montreal, Quebec, Canada: ASME, pp. 167–74. doi:10.1115/DETC2010-29057.
- Keshwani S and Chakrabarti A (2017) Towards automatic classification of description of analogies into SAPPPhIRE constructs. In Chakrabarti A and Chakrabarti D (eds), *Research into Design for Communities*, Vol. 2. Smart Innovation, Systems and Technologies. Singapore: Springer, pp. 643–655. doi:10.1007/978-981-10-3521-0\_55.
- Kitamura F (2023) ChatGPT is shaping the future of medical writing but still requires human judgment. *Radiology*, 230171. doi:10.1148/radiol.230171.
- Kruiper R, Vincent JFV, Abraham E, Soar RC, Konstas I, Chen-Burger J and Desmulliez MPY (2018) Towards a design process for computer-aided biomimetics. *Biomimetics* 3, 14. doi:10.3390/biomimetics3030014
- Laakso M, Welling P, Bukvova H, Nyman L, Björk B-C and Hedlund T (2011) The development of open access journal publishing from 1993 to 2009. *PLoS One* 6, e20961. doi:10.1371/journal.pone.0020961
- Landauer TK, McNamara DS, Dennis S and Kintsch W (2007) *Handbook of Latent Semantic Analysis*. Taylor & Francis. doi:10.4324/9780203936399.
- Lenau T, Metz A-L and Hesselberg T (2018) Paradigms for biologically inspired design. In A. Lakhtakia (ed.), *Bioinspiration, Biomimetics, and Bioreplication VIII*, Vol. 1. Denver, USA: SPIE. doi:10.1117/12.2296560.

- Manning C, Raghavan P and Schuetze H** (2009) Introduction to information retrieval. <https://nlp.stanford.edu/IR-book/>.
- Mei Q, Shen X and Zhai C** (2007) Automatic labeling of multinomial topic models. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '07. New York, NY, USA: ACM, pp. 490–499. doi:10.1145/1281192.1281246.
- Mikolov T, Chen K, Corrado G and Dean J** (2013a) Efficient estimation of word representations in vector space. ArXiv:1301.3781 [Cs], January. <http://arxiv.org/abs/1301.3781>.
- Mikolov T, Le Quoc V and Sutskever I** (2013b) Exploiting similarities among languages for machine translation. ArXiv:1309.4168 [Cs], September. <http://arxiv.org/abs/1309.4168>.
- Montani I, Honnibal M, Honnibal M, Van Landeghem S, Boyd A, Peters H, McCann PO, et al.** (2022) Explosion/SpaCy: v3.4.2: Latin and Luganda support, python 3.11 wheels and more. *Zenodo*. doi:10.5281/zenodo.7228125
- Nachtigall W and Wisser A** (2015) *Bionics by Examples*. Cham: Springer International Publishing. doi:10.1007/978-3-319-05858-0.
- Nagel JKS** (2014) A thesaurus for bioinspired engineering Design. In Goel AK, McAdams DA and Stone RB (eds), *Biologically Inspired Design*. London: Springer London, pp. 63–94. doi:10.1007/978-1-4471-5248-4\_4.
- Nagel JKS and Stone RB** (2012) A computational approach to biologically inspired design. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 26, 161–176. doi:10.1017/S0890060412000054
- Nagel JKS, Stone RB and McAdams DA** (2010) An engineering-to-biology thesaurus for engineering design. In Volume 5: 22nd International Conference on Design Theory and Methodology; Special Conference on Mechanical Vibration and Noise. Montreal, Quebec, Canada: ASME, pp. 117–128. doi:10.1115/DETC2010-28233.
- Nagel JKS, Stone RB and McAdams DA** (2014) Function-based biologically inspired design. In Goel AK, McAdams DA and Stone RB (eds), *Biologically Inspired Design: Computational Methods and Tools*. London: Springer London, pp. 95–125. doi:10.1007/978-1-4471-5248-4\_5.
- Pahl G, Beitz W, Feldhusen J and Grote K** (2007) *Engineering Design: A Systematic Approach*, 3rd Edn. London: Springer. doi:10.1007/978-1-84628-319-2.
- Pentelovitch N and Nagel JK** (2022) Understanding the use of bio-inspired design tools by industry professionals. *Biomimetics* 7, 63. doi:10.3390/biomimetics7020063
- Perricone V, Grun T, Raia P and Langella C** (2022) Paleomimetics: a conceptual framework for a biomimetic design inspired by fossils and evolutionary processes. *Biomimetics* 7, 89. doi:10.3390/biomimetics7030089
- Rosa F, Cascini G and Baldussu A** (2015) UNO-BID: unified ontology for causal-function modeling in biologically inspired design. *International Journal of Design Creativity and Innovation* 3, 177–210. doi:10.1080/21650349.2014.941941
- Rugaber S, Bhati S, Goswami V, Spiliopoulou E, Azad S, Koushik S, Kulkarni R, Kumble M, Sarathy S and Goel A** (2016) Knowledge Extraction and Annotation for Cross-Domain Textual Case-Based Reasoning in Biologically Inspired Design. In Goel A, Belén Diaz-Agudo M and Roth-Berghofer T (eds), *Case-Based Reasoning Research and Development*. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 342–355. doi:10.1007/978-3-319-47096-2\_23.
- Schoch CL, Ciufu S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, et al.** (2020) NCBI taxonomy: a comprehensive update on curation, resources and tools. *Database: The Journal of Biological Databases and Curation* 2020. doi:10.1093/database/baaa062
- Shu LH** (2010) A natural-language approach to biomimetic design. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 24, 507–519. doi:10.1017/S0890060410000363
- Shu LH, Ueda K, Chiu I and Cheong H** (2011) Biologically inspired design. *CIRP Annals* 60, 673–693. doi:10.1016/j.cirp.2011.06.001
- Siddharth L and Chakrabarti A** (2018) Evaluating the impact of idea-inspire 4.0 on analogical transfer of concepts. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 32, 431–448. doi:10.1017/S0890060418000136
- Slosar N** (2021) Avians to airplanes: biomimicry in flight and wing design. *Berkeley Scientific Journal* 25. doi:10.5070/BS325254487
- Smith SL, Turban DHP, Hamblin S and Hammerla NY** (2017) Offline bilingual word vectors, orthogonal transformations and the inverted softmax. ArXiv:1702.03859 [Cs], February. <http://arxiv.org/abs/1702.03859>.
- Srinivas A** (2022) Perplexity AI: ask anything. Perplexity.Ai. 2022. <https://www.perplexity.ai/>.
- Sun F, Xu H, Meng Y and Lu Z** (2022) A BERT-based model for coupled biological strategies in biomimetic design. *Neural Computing and Applications*. September. doi:10.1007/s00521-022-07734-z.
- Turian J, Ratino L-A and Bengio Y** (2010) Word representations: a simple and general method for semi-supervised learning. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden: Association for Computational Linguistics, pp. 384–394.
- Vandevenne D, Verhaegen P-A and Joost Duflo R** (2014) Mention and focus organism detection and their applications for scalable systematic bio-ideation tools. *Journal of Mechanical Design* 136, 111104. doi:10.1115/1.4028278
- Vandevenne D, Verhaegen P-A, Dewulf S and Duflo JR** (2015) SEABIRD: scalable search for systematic biologically inspired design. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 30, 78–95. doi:10.1017/S0890060415000177
- Vandevenne D, Pieters T and Duflo JR** (2016) Enhancing novelty with knowledge-based support for biologically-inspired design. *Design Studies* 46, 152–173. doi:10.1016/j.destud.2016.05.003
- Van Noorden R** (2014) Elsevier opens its papers to text-mining. *Nature* 506, 17–17. doi:10.1038/506017a
- Vattam SS and Goel AK** (2011) Foraging for inspiration: understanding and supporting the online information seeking practices of biologically inspired designers. In *Proceedings of the ASME Design Engineering Technical Conference*, Vol. 9, pp. 177–186. doi:10.1115/DETC2011-48238
- Vattam SS, Wiltgen B, Helms M, Goel AK and Yen J** (2011) DANE: fostering creativity in and through biologically inspired design. In Taura T and Nagai Y (eds), *Design Creativity 2010*, pp.115–122. London: Springer London. doi:10.1007/978-0-85729-224-7\_16.
- Verhaegen P-A, D'hondt J, Vandevenne D, Dewulf S and Duflo JR** (2011) Automatically characterizing products through product aspects. In Bernard A (ed.), *Global Product Development*. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-15973-2\_60.
- Wanieck K, Fayemi P-E, Maranzana N, Zollfrank C and Jacobs S** (2017) Biomimetics and its tools. *Bioinspired, Biomimetic and Nanobiomaterials* 6, 53–66. doi:10.1680/jbibn.16.00010.
- Wikimedia Foundation.** (2020) Wikidata. [https://www.wikidata.org/wiki/Wikidata:Main\\_Page](https://www.wikidata.org/wiki/Wikidata:Main_Page).
- Willocx M** (2021) Building a paper bridge between biology and engineering. In KU Leuven Open Science Day. doi:10.21428/1192f2f8.fa594606.
- Willocx M and Duflo JR** (2021) Metrics for bio-inspiration based on taxonomies. *Proceedings of the Design Society* 1, 2087–2096. doi:10.1017/pds.2021.470
- Willocx M, Ayali A and Duflo JR** (2020) Where and how to find bio-inspiration?: A comparison of search approaches for Bio-inspired design. *CIRP Journal of Manufacturing Science and Technology* 31, 61–67. doi:10.1016/j.cirpj.2020.09.013

**Mart Willocx** is a researcher at the Centre for Industrial Management at KU Leuven. After finishing his studies as a mechanical engineer, he was promptly convinced by the inspiring genius of nature. His current research focuses on bridging the gap between engineering and biology, giving mechanical engineers the tools to find relevant biological strategies.

**Joost R. Duflo** is a Professor in the Department of Mechanical Engineering at KU Leuven. He has master degrees in architectural and electromechanical engineering and a PhD in engineering from KU Leuven. He is a member of CIRP and has been published in over 200 international publications. His principal research activities are in the field of design support methods and methodologies, with special attention for systematic innovation, ecodesign, and life cycle engineering.

## Appendix A

To automate the generation of the organism aspects, a rule-based approach is presented based on the location of the lemma in the concept hierarchy in

WikiData. Based on the target of identifying strategies and the manual annotation by Vandevenne, the entities presented in Table A1 and their descendants are excluded from the organism aspects. This list was compiled in September 2020.

**Table A1.** List of entities which along with their descendants will be excluded in the generation of the organism aspects. This list was built based on the contents available in WikiData in September 2020.

Entity ID	Description	Entity ID	Description	Entity ID	Description
Q729	Animal	Q13442814	Scholarly article	Q28845870	Anatomical structure
Q7239	Organism	Q4167410	Disambiguation page	Q4936952	Organic compound
Q1811014	Phase of life	Q215380	Band	Q174211	Concrete object
Q39546	Tool	Q7187	Gene	Q4406616	Organ
Q6671777	Structure	Q5	Human	Q712378	Organ
Q52948	Interaction	Q134556	Single	Q12136	Disease
Q8171	Word	Q81163	Polymer	Q618123	Geographical object
Q12767945	Language device	Q11424	Film	Q231002	Nationality
Q1150070	Change	Q1310239	Component	Q215627	Person
Q1194058	Disposable product	Q15989253	Part	Q10856962	Anthroponym
Q16521	Taxon	Q28732711	Physical substance	Q11173	Chemical compound