 doi:10.20944/preprints202110.0382.v1

Employing statistical machine reading for inferring key concepts of a research field from a body of abstracts and blog posts

Wenfa Ng

Department of Chemical and Biomolecular Engineering, National University of Singapore
Email: ngwenfa771@hotmail.com or ngwenfa@nus.edu.sg

Abstract

The world of science is drowned in a wealth of information. How to make sense of this wealth of published articles, blog posts and abstracts has become an important challenge given the importance of science to different aspects of societal function. At the crux of the issue lies the increasing trend where scientific discovery informs decision making at the societal level. One example, is the elucidation of the ozone hole to the promulgation of the Montreal Protocol in 1987, and documenting increasing atmospheric carbon dioxide concentration led to climate action and signing of the Paris Agreement in 2015. Hence, understanding a research field becomes an important need for many decision makers across different sectors of society. But, the scientific literature is cryptic and esoteric, and presents a significant barrier to comprehension. One approach to ameliorate the problem is statistical machine reading, which provides the critical capability of identifying key concepts that underpins a research field. Such important concepts help provide an incision point to gain further understanding of the field and initiating further conversation about the field. This work sought to validate the concept of whether applying statistical machine reading to a body of literature comprising short blog posts and abstracts of published articles help in understanding the field of metabolic engineering. One important angle pursued in this research is whether the tabulated list of terms and phrases identified by statistical machine reading could be creatively analyzed to gain a deeper understanding of the research field. For example, the most frequently occurring terms and phrases could describe key concepts of the research field. Moving down in frequency occurrence would be terms and phrases that describe methodologies and approaches of the field. Finally, less frequently occurring terms and phrases may be tools and resources used in the research field. Results validated the utility of statistical machine reading in identifying important terms and phrases associated with the research field. But the small dataset of blog posts and abstracts used in this study severely hampered the identification of most of the key concepts of metabolic engineering, which is a fairly broad field of research. Overall, statistical machine reading shows utility in identifying terms and phrases that could describe a field. However, the level of understanding is closely tied in to the breadth and depth of reading material available, which meant that the methodology is data intensive in nature. Future use of supercomputing or quantum computing could help alleviate constraints of computational capacity, and help tackle the exponential rise in computational complexity as the size of the reading material for machine reading expands.

Keywords: statistical machine reading, metabolic engineering, concepts, terms and phrases, computational complexity,

Subject areas: machine learning,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2021 doi:10.20944/preprints202110.0382.v1

Highlights

- Ensemble statistical machine reading was employed to glean key concepts in the field of metabolic engineering by profiling 20 relevant abstracts and 15 blog posts.
- Single word and two-word phrase concepts elucidated could be segregated into three tiers of concepts: (i) key concepts, (ii) approaches and methodologies, and (iii) tools, resources and targets.
- Analysis of results revealed that single word machine reading could complement two-word (phrase) machine reading is delivering useful and meaningful concepts for understanding the field of research.
- Quality of concepts elucidated by machine reading depends on the breadth and size of the reading material.
- Given that complexity of machine reading tasks scale exponentially with size of reading material, high performance computing or quantum computing need to be enlisted to push the field forward in analysis of large reading sets.

1. Introduction

The scientific literature is awash with information, and the pace at which the scientific literature is expanding is not abating. This then places significant demands on researchers on keeping up-to-date with the literature, particularly if one is to be conversant with others on cutting-edge and leading-edge research topics. Given that reading the scientific literature with the help of search engine filtering of the important search terms is already a difficult and time-consuming task, are there tools for helping us to gain at least a precursor knowledge of the literature?

From another perspective, the scientific literature does not only speaks to scientists, it also impact on decision makers at all levels of government given the importance of science in influencing many aspects of human and societal life. In short, science has a deep impact on our living world, and humans have learned to harness scientific knowledge to build a better future for the next-generation. Such a goal would necessitate some understanding of the guiding principles, processes and conclusions derived at through scientific research, and reported via the scientific literature. However, much of this scientific literature is cryptic and esoteric, and inaccessible to policy makers who may have some science training. Hence, it is imperative that some technological tool be made available to help summarize and derive some initial understanding of the literature.

Such understanding could manifest in the form of a list of keywords that describes a field of research. This is important as seen from the following usage scenario. Consider the case where a country would like to seek out ways to improve the coupling of renewable energy to the existing power grid. One solution in this direction would be the use of energy storage solutions. But, this field may not be comprehensible to policy makers, particularly from the perspective of gaining a broad understanding of the pillars that support the field. Hence, a

 doi:10.20944/preprints202110.0382.v1

technological tool that could help provide some understanding of the key concepts that underpin the field of energy storage would be significant.

One tool that could be used, and which is in development in many companies and academic labs around the world is machine reading comprehension.[1 2 3] In this approach, text analytic methods[4 5] are introduced to help organize, summarize, and provide some initial understanding of the body of literature that is being read. Much of the research in this field remains in the lab, and algorithmic details are not easily accessible on the Internet. Hence, it is difficult to gauge progress and success in the field beyond the descriptions on websites of companies and academic labs engaged in this field.

This work sought to move research along the field of machine reading comprehension by examining if ensemble statistical machine reading could help elucidate key concepts of a research field: in this case, metabolic engineering. Choice of metabolic engineering as target research field comes from its diversified array of research activities that revolve around the central concept of manipulating metabolism to yield new products or enhancing production of product from a given substrate. Such a research field requires knowledge of multiple ancillary sub-fields in order to afford adequate comprehension of the key underlying principles of the field, and is suitable for analysis by machine reading.

The key goal of using ensemble statistical machine reading is to utilize the tools of statistical machine reading over a large dataset of reading material to extract key concepts from the body of literature analyzed. Statistical machine reading, in this usage scenario, refers to the determination of the frequency of occurrence of words and phrases from the text. Terms and phrases featured in this analysis were extracted from the text, and not from human input. Hence, the computational pipeline is automatic, requiring only the supply of reading material. In this work, personal blog posts addressing different issues of metabolic engineering, as well as abstracts of articles in the journal, *Metabolic engineering*, were used as reading material. Altogether, 15 blog posts and 20 abstracts were individually used in the reading exercise to extract key concepts about the field of metabolic engineering.

In this analysis, key concepts need not be those with the highest frequency of occurrence in the reading material. Rather, the output of ensemble statistical machine reading is mined extensively to yield other aspects of the field such as helping answer questions about the approaches and methodologies that underpin the field, as well as tools, resources and targets of the field. These follow-on concepts need not be high frequency of occurrence in the final output from machine reading. Necessarily, these concepts need to be elucidated through human manual curation and inference.

Preliminary results indicate that the concepts elucidated by ensemble statistical machine reading critically depends on the size of the reading material dataset. Better and more

 doi:10.20944/preprints202110.0382.v1

diversified reading material could afford the deciphering of concepts that describe a particular nuanced aspect of the field of metabolic engineering. Given the limited dataset of blog posts and abstracts employed for this work, there is relatively little concurrence between the concepts determined by machine reading, and those obtained by our current understanding of the field.

2. Materials and Methods

2.1 Materials

Materials for feeding into statistical machine reading algorithm coded in MATLAB comprises blog posts written by myself on different aspects of metabolic engineering, as well as 20 abstracts from journal articles in the peer-reviewed journal, *Metabolic engineering*. The blog posts were originally written in Microsoft Word, and were subsequently converted to pdf format to facilitate machine reading by the algorithm. Abstracts for the machine reading task were obtained from the website of *Metabolic engineering*, and copied to Microsoft Word files. These files were subsequently converted to pdf files to facilitate machine reading by the in-house MATLAB software.

2.2 Statistical machine reading algorithm

The software used for statistical machine reading in this project was coded using MATLAB and comprises a couple of modular sub-functions. Specifically, the software first parses the text into individual words, and searches for their frequency of occurrence in the main text of the pdf used for the reading task. The software could read and search multiple pdf files containing blog posts or abstracts from journals for the task.

Given that many concepts in a field of research comprises two or more words, another module of the MATLAB software was bestowed the function of searching for “two word” phrases in the main text of the set of blog posts or abstracts used for the analysis. Frequency of occurrence of the phrases was tallied for use in understanding the significance of the phrases in the overall context of the field of research.

2.3 Data analysis approach

The most common approach for identifying major concepts in the field of research would be to look for words or phrases with high frequency of occurrence. This is true, but, in order to identify other ancillary aspects of a field of research such as approaches and methodologies as well as tools and resources, words and phrases of lower frequency of occurrence are also important. This aspect of the analysis task is supported by the output format of the MATLAB software used for this research. Specifically, all non-redundant words and phrases that exist in the text that had frequency of occurrence of at least 1 are tabulated into a database for output to an Excel file for ease of storage and information retrieval. Variables catalogued in the database include terms/phrases, Match (whether there has been a match in the main text), and Count (their frequency of occurrence). Doing so provides a platform for the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2021 doi:10.20944/preprints202110.0382.v1

user of the software to creatively and holistically assess the data to identify useful information that could describe varied aspects of a research field beyond the common goal of gaining an understanding of the major concepts of the field.

3. Results and Discussion

3.1 Developing a computational pipeline to yield a list of pertinent concepts of a research field through ensemble statistical machine reading

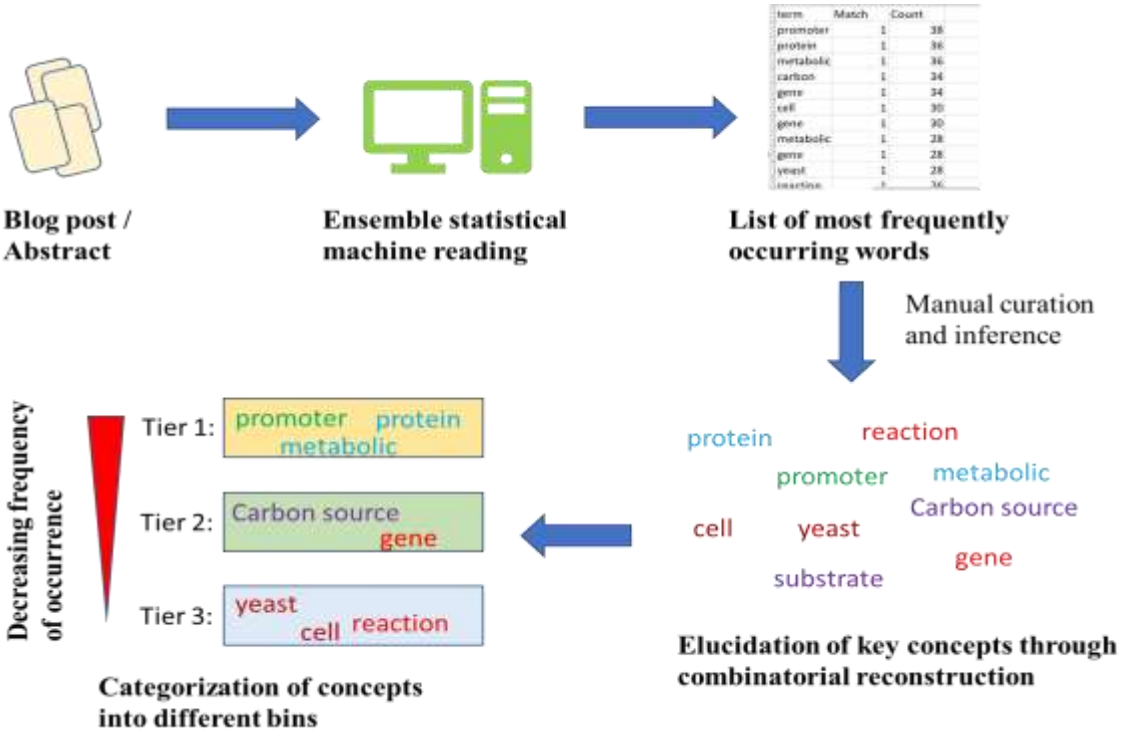

Figure 1: Schematic diagram illustrating major steps in the ensemble statistical machine reading workflow to derive meaningful concepts from a body of blog posts and abstracts as reading material.

Similar to other computational projects, the current machine reading project requires a computational pipeline to efficiently channel raw text from reading material (blog post and abstracts) to computational algorithms for text analysis to yield words or "two-word" phrases that confer meaning to a research field, which, in this case is chosen to be metabolic engineering. Figure 1 illustrates the schematic of the entire computational pipeline from text to concepts words and phrases describing the research field. Specifically, the first step in the workflow involves passing the pdf encapsulation of blog posts and abstracts to an in-house developed MATLAB based ensemble statistical machine reading software. Blog posts came from my personal collection, and the ones chosen for this project analyses critical issues impacting the field of metabolic engineering from different perspectives. On the other hand, abstracts were obtained from the websites of journal articles recently published in the peer-reviewed journal of *Metabolic engineering*. Choice of abstracts and not full-text of articles come from the difficulty of obtaining open-access articles in the field of metabolic engineering.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2021 doi:10.20944/preprints202110.0382.v1

Altogether, the compendium of blog posts and abstracts in the set of reading material that is at the core of this project provides two important contrasting viewpoints of the state of research field of metabolic engineering. Specifically, blog posts chosen for this study are short 500 to 800 words exposition on a focused facet of metabolic engineering. These writings should provide some indications of the key concepts, methodology, and approaches central to the practice of metabolic engineering. Abstracts, on the other hand, are shorter delineations of the research questions posed, methodology and approaches used to elucidate the truth, as well as description of defining results of the project. Hence, in addition to providing information of the key concepts, approaches and methodologies of a research field, abstracts also illuminate the tools and resources such as genome editing and use of *Escherichia coli* model organism common in metabolic engineering research.

Ensemble statistical machine reading is, at its core, the use of machine reading algorithm to perform a statistical analysis of the frequency of occurrence of words and phrases in the text of the reading material. Performed iteratively across the entire collection of blog posts and abstracts, the approach of ensemble statistical machine reading should yield a tabulated list of most frequently occurring words and phrases, that provide some indications of the important concepts, approaches, and methodologies that underpin a research field. But, words and phrases of high frequency occurrence in the text of the profiled blog posts and abstracts may be general and not specific to the research field of interest. Herein, lies the current challenge in statistical machine reading in that the approach is still unable to deliver a level of comprehension that affords it the ability to discriminate between different general and specialist concepts of a research field. This problem could be ameliorated by a large and diverse set of reading material, but as my personal experience with machine reading shows, the reading task rapidly becomes unmanageable as the number of articles in the collection grows, and the computational task scales exponentially with expansion of the reading material. Hence, there is an inherent limit on the size of the reading material collection, which, in this study, is limited to 15 blog posts and 20 abstracts profiled as individual set of reading material.

Lack of advanced machine reading comprehension in my algorithm necessarily meant that manual curation and inference is needed to elucidate defining concepts, approaches and methodologies from the tabulated list of most frequently occurring words and phrases. To this end, manual inspection of the list of frequently occurring words and phrases in the output from the machine reading algorithm should provide sufficient information for the author (who did a research project in metabolic engineering) to identify and verify that pertinent concepts in metabolic engineering has been selected by the machine reading algorithm. This also serves as a validation tool for the algorithm developed in this project.

The final step of the pipeline is the categorization of concepts into different bins by their different frequency of occurrence in the profiled blog posts and abstracts. To this end, three tiers of concepts serve as bins, where Tier 1 comprises the most frequently occurring terms and phrases, and Tier 3 are the list of terms of low frequency occurrence in the profiled reading material. Such categorization endows relevance to the different bins. Specifically, high

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2021 doi:10.20944/preprints202110.0382.v1

occurrence terms and phrases in Tier 1 could be the key concepts of the field of metabolic engineering. Moving down to the next tier, Tier 2 could describe the methodologies and approaches of the field, and this postulation is validated by observations in the real-world that such terms are less frequently mentioned in abstracts and articles. Finally, less frequently occurring Tier 3 terms could be describing tools and resources employed in metabolic engineering that are only mentioned sparingly in manuscripts. Overall, the approach of mining the results from a statistical machine reading exercise can help uncover less well-known facets of a research field.

3.2 Defining the research field of metabolic engineering using keywords and phrases

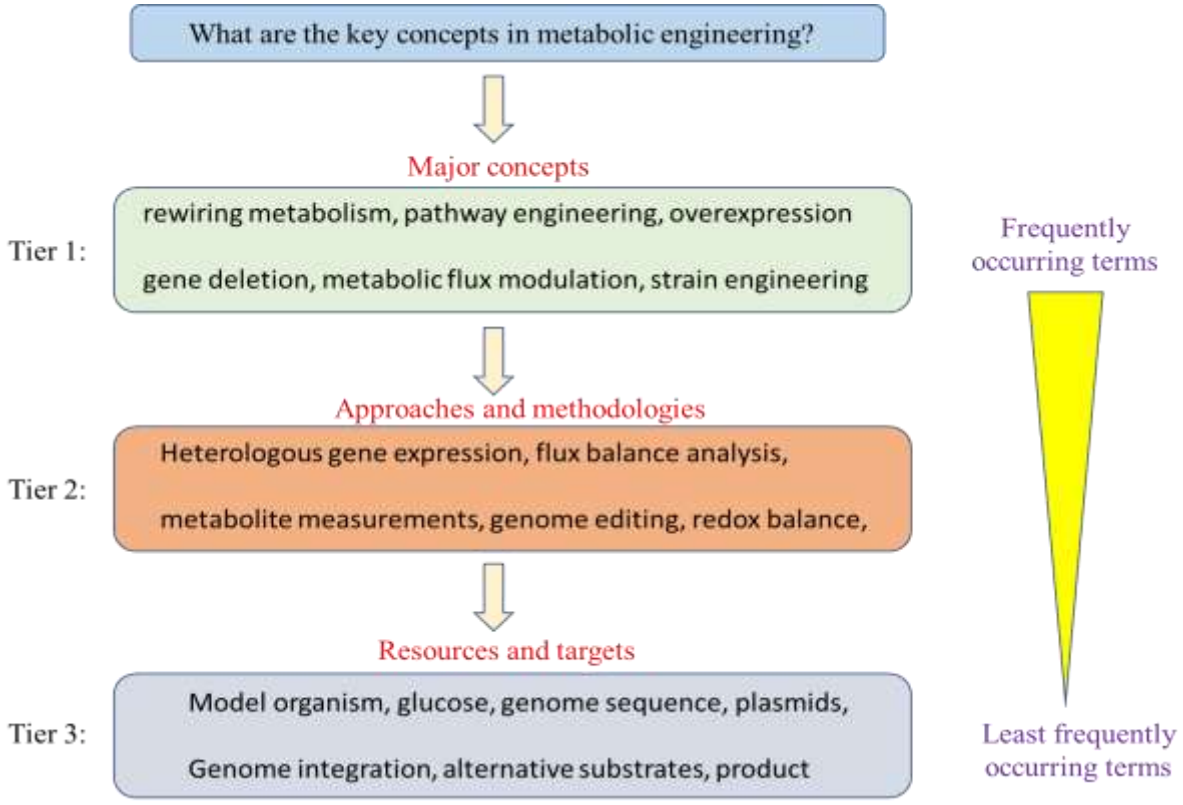

Figure 2: Current understanding of the research field of metabolic engineering. The definition of the field includes key concepts (Tier 1), approaches and methodologies (Tier 2), and resources and targets (Tier 3). In general, Tier 1 terms are the high frequency of occurrence terms, while Tier 3 terms are of low frequency of occurrence.

The object of study in this research is the research field of metabolic engineering. In general, the key concepts that underpin modern metabolic engineering are: (i) rewiring metabolism, (ii) pathway engineering, (iii) overexpression, (iv) gene deletion, (v) metabolic flux modulation, and (vi) strain engineering. These concepts are likely to be of high frequency of occurrence in the profiled blog posts and abstracts. Beyond the key concepts of a research field, it is also important to gain an appreciation of the approaches and methodologies commonly used in this field. In the case of metabolic engineering, these approaches and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2021 doi:10.20944/preprints202110.0382.v1

methodologies are: (i) heterologous gene expression, (ii) flux balance analysis, (iii) metabolite measurements, (iv) genome editing, and (v) redox balance. Given that approaches and methodologies are less frequently mentioned in abstracts, these would come out as less frequently used terms in statistical machine reading, and be classified as Tier 2 terms. The final tier of terms would be tools, resources and targets of metabolic engineering. These could include, for example: (i) model organism, (ii) glucose, (iii) genome sequence, (iv) plasmids, (v) genome integration, (vi) alternative substrates, and (vii) product. In these terms, substrate (glucose), product and model organism would be more frequently mentioned and have high frequency of occurrence. On the other hand, genome sequence, plasmids, and genome integration would be less frequently used, and have low frequency of occurrence.

3.3 Concepts derived from statistical machine reading analysis of abstracts of journal articles

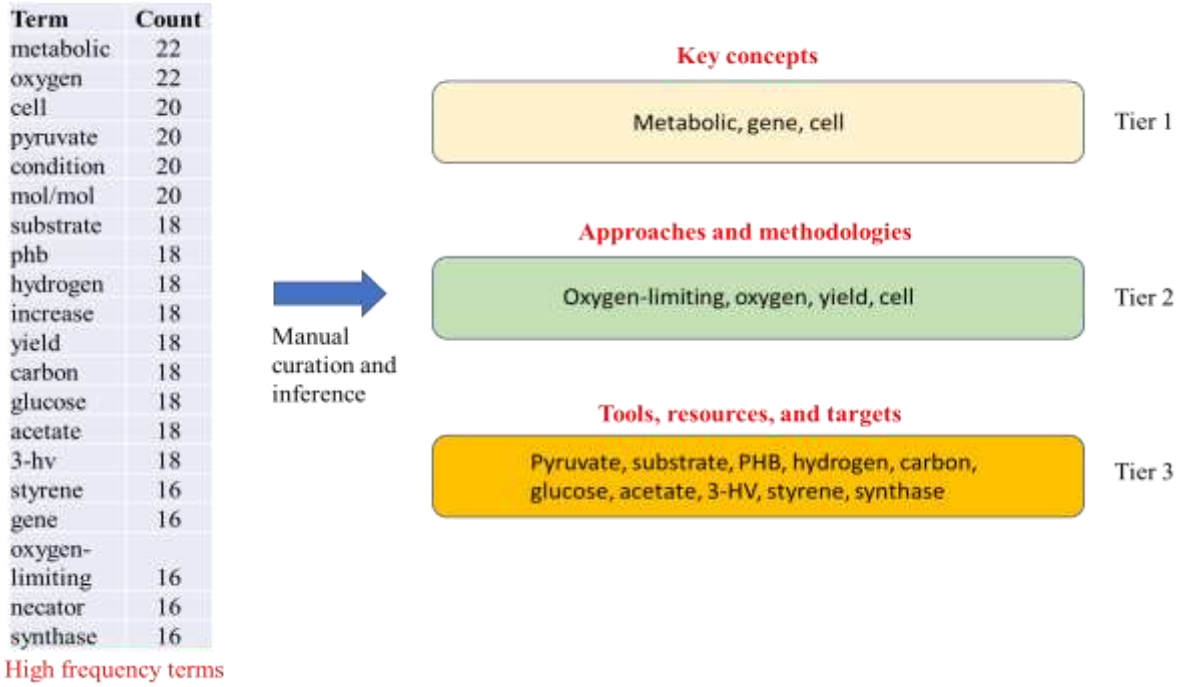

Figure 3: Important terms that describe the key concepts, approaches and methodologies, and tools, resources and targets of metabolic engineering were derived from employing ensemble statistical machine reading on 20 abstracts. Shown are results from single word machine reading analysis.

Analysis of 20 abstracts from the journal *Metabolic engineering* using single word ensemble statistical machine reading yields relatively few concepts that show a high level of correspondence with concepts depicted in Figure 2. Figure 3 shows the results from the statistical machine reading exercise on 20 abstracts at the single word level. The results showed the limitation of the technique at profiling useful overarching concepts of metabolic engineering given that such concepts are typically two-word phrases. More broadly, the results may also illustrate the trend that many introductory statements in abstracts are highly project-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2021 doi:10.20944/preprints202110.0382.v1

specific, and do not speak of the broad overarching goals of metabolic engineering. One contributing reason for this may be the tight word limit for most abstracts that led to project-specific writing on the part of authors. Similar issues of poor correspondence in approaches and methodologies of metabolic engineering between machine reading results and those accepted by the field suggest the difficulty of capturing these concepts using single word or two-word phrase. The key underlying reason that many authors are very specific in describing their methodologies, and seldom use overarching concepts to describe their methods in the abstract. But, one plus point of machine reading of abstracts is its capability at highlighting many substrates, products, tools and resources used in metabolic engineering research. Such facility likely comes from the strict requirement to report these aspects in any scientific abstract. Overall, the data yields the surprising finding that high frequency of occurrence terms could supply Tier 2 and Tier 3 concepts, and hence negate the need for profiling low frequency of occurrence terms in the dataset.

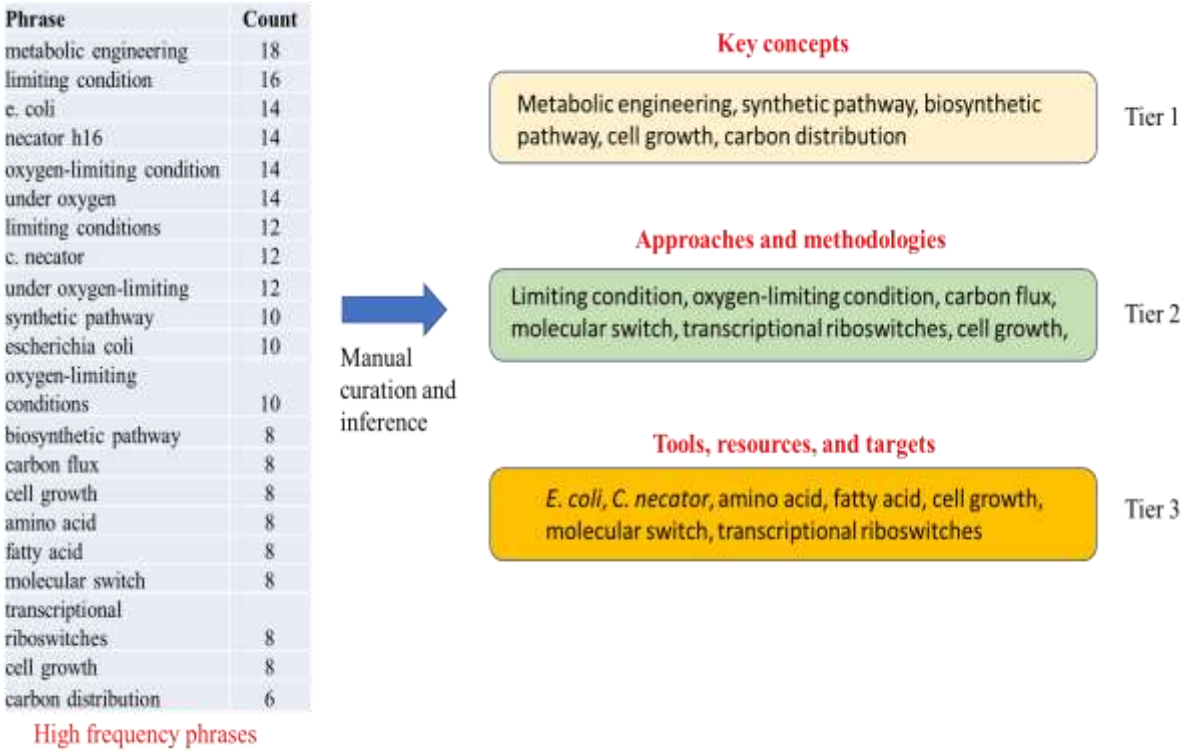

Figure 4: Important terms that describe the key concepts, approaches and methodologies, and tools, resources and targets of metabolic engineering were derived from employing ensemble statistical machine reading on 20 abstracts. Shown are results from two-word (phrase) machine reading analysis.

In comparison, two-word phrase ensemble statistical machine reading yield phrases that could better describe the key concepts of metabolic engineering compared to single word machine reading analysis. Specifically, machine reading analysis of 20 abstracts from metabolic engineering journal articles reveal relatively new concept like synthetic pathway, which is a burgeoning subfield of metabolic engineering interested in the construction of new pathways for assimilating alternative substrates like ethylene glycol and ethanol (Figure 4). In addition, machine reading also identified the key concept of carbon distribution in metabolic

 doi:10.20944/preprints202110.0382.v1

engineering, which concerns how carbon from the substrate is distributed to different pathways in central carbon metabolism. This concept is more commonly known as flux balance analysis.

Moving to Tier 2 concepts involved in approaches and methodologies of metabolic engineering, machine reading analysis of abstracts made the surprising finding of oxygen limiting condition which is unusual in metabolic engineering as most research in the field are conducted under aerobic conditions. But, the result could suggest, upon further verification, that oxygen limiting condition may be nascent trend in metabolic engineering. Further, machine reading analysis identified relatively new approaches in metabolic engineering such as the use of transcriptional riboswitches, and molecular switches in controlling gene expression in metabolic engineering research. These tools are relatively new introductions to synthetic biology and metabolic engineering.

Finally, machine reading analysis yields important tools, resources and targets for metabolic engineering such as successful identification of the preeminent role of the model bacterium, *Escherichia coli*, in metabolic engineering research. But, at the same time, the same set of data from machine reading also highlight possible emerging role of *C. necator* in metabolic engineering. In addition, amino acid and fatty acids were also identified as major targets in machine reading analysis of abstracts, which suggests that there may be growing body of research surrounding these topics. Finally, machine reading successfully identified the relatively new tools of molecular switches, and transcriptional riboswitches. Overall, machine reading yields a set of high frequency phrases that covers all three tiers of concepts profiled in this research, thereby, making the analysis of low frequency of occurrence phrases unnecessary.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2021 doi:10.20944/preprints202110.0382.v1

3.4 Concepts derived from statistical machine reading analysis of blog posts

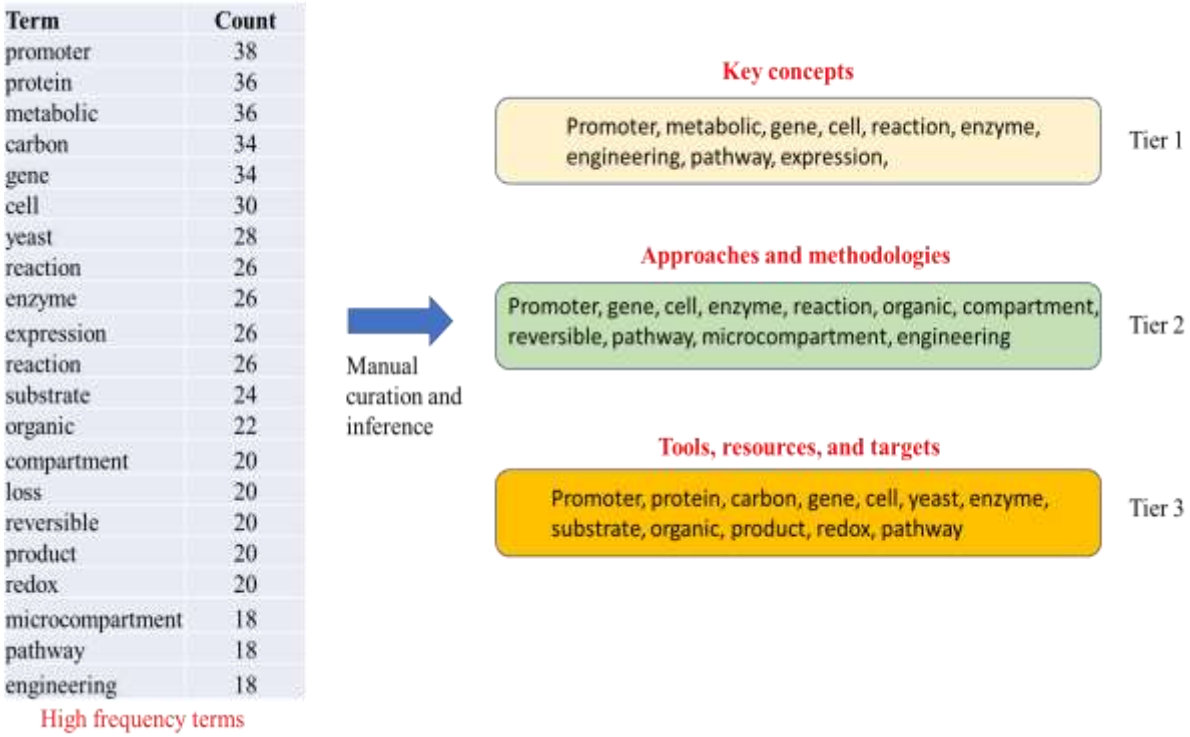

Figure 5: Important terms that describe the key concepts, approaches and methodologies, and tools, resources and targets of metabolic engineering were derived from employing ensemble statistical machine reading on 15 blog posts. Shown are results from single word machine reading analysis.

Compare to abstracts, blog posts represent a distinctly different publication medium. One major difference is that blog posts are less cryptic and more descriptive. From the standpoint of text analysis, blog posts also use less of overarching ideas, but instead uses more descriptive words to describe the same idea. In addition, blog posts are also longer than abstracts, and may be able to yield new concepts seldom able to be incorporated in abstracts.

Figure 5 depicts the concepts elucidated by single word ensemble statistical machine reading of 15 long-form blog posts whose word count ranges from 500 to 800 words. Results revealed that analysis of blog posts yielded more and different keywords compared to similar analysis of abstracts. Specifically, major concepts such as gene, cell, pathway, enzyme, reaction that describes the field of metabolic engineering could be deciphered by the single word ensemble statistical machine reading exercise. This compares favourably to the relatively few key concepts that could be determined from abstracts as reading materials. Major reason for this observation is the lack of emphasis on articulating key concepts of metabolic engineering in abstracts of professional articles given that the readership of such articles is assumed to be conversant with the major concepts of the field. Secondly, blog posts tend to explain major concepts of the field of metabolic engineering more clearly to help bring the layman reader into the topic, and thus, provide a deeper background on the key concepts of

 doi:10.20944/preprints202110.0382.v1

metabolic engineering. Finally, the result also highlights that statistical machine reading critically depends on the size and scope of the dataset in order to provide a more rounded comprehension of the field of research.

On the other hand, single word machine reading analysis of blog posts could not yield important ideas in approaches and methodologies in metabolic engineering beyond engineering of promoter, enzymes, and pathway. This likely comes about due to the use of descriptive (layman like) language in describing important approaches and methodologies in metabolic engineering. Such a writing style preclude the use of jargon that identifies an approach or methodologies, which explains why single word machine reading analysis of blog posts could not elucidate important concepts in this area.

Finally, analysis for keywords that describe tools, resources and targets of metabolic engineering yields a couple of generic terms such as substrates, products, promoter, cell, and gene. This differs significantly from the wealth of information that could be elucidated by single word machine reading analysis of abstracts. One important reason that underpin this observation is that the blog posts were designed to be generic and were used to communicate general ideas on metabolic engineering to a broad audience, and thus, did not specify the substrate or product as compared to the case in abstracts. Similarly, focus on metabolic engineering principles in the blog posts also preclude the identification of other resources and targets useful for contemporary metabolic engineering. Overall, useful concepts belonging to Tier 1 to Tier 3 concepts could be elucidated by machine reading, and low frequency of occurrence terms were not analysed in this work.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2021 doi:10.20944/preprints202110.0382.v1

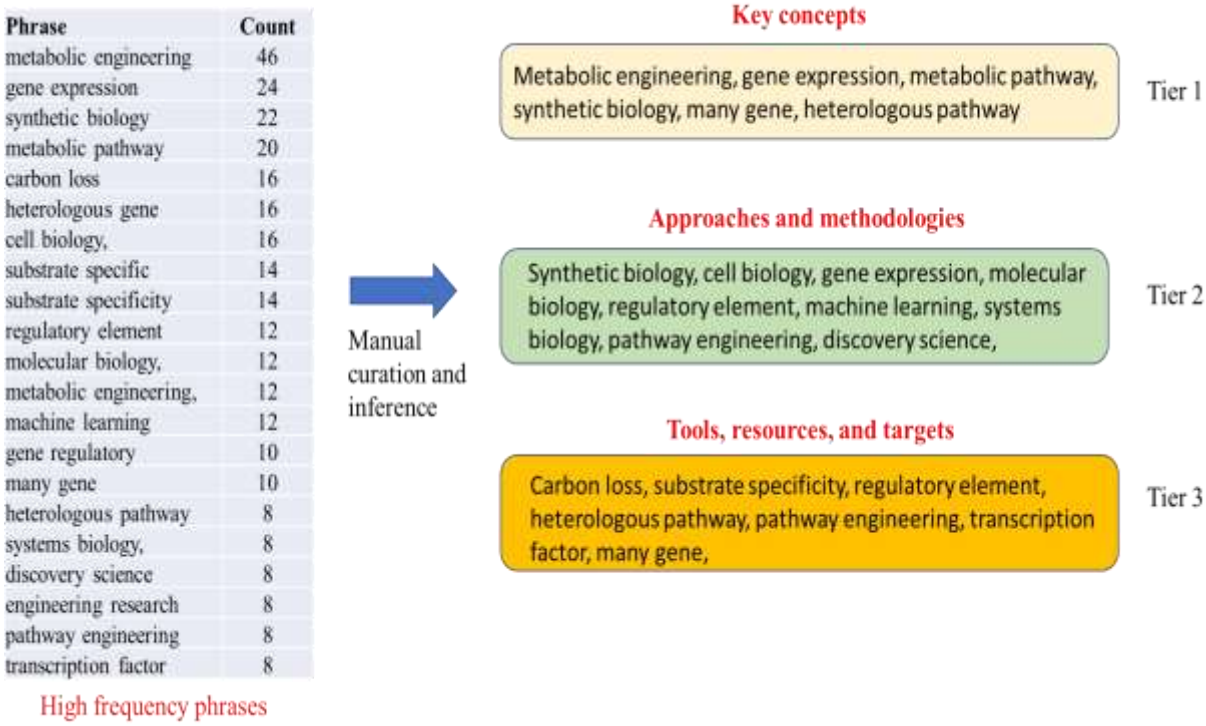

Figure 6: Important terms that describe the key concepts, approaches and methodologies, and tools, resources and targets of metabolic engineering were derived from employing ensemble statistical machine reading on 15 blog posts. Shown are results from two-word (phrase) machine reading analysis.

Figure 6 shows the results from two-word (phrase) ensemble statistical machine reading analysis of 15 blog posts on different facets of metabolic engineering. Results revealed that two-word (phrase) machine reading could elucidate many important concepts in metabolic engineering such as many gene, heterologous pathway, gene expression, and synthetic biology, which are congruent with the practice of contemporary metabolic engineering. Such a result speaks of the descriptive writing style of blog post authors who sought to lay down the foundational concepts of a field of research prior to providing further elaboration.

In terms of approaches and methodologies, two-word (phrase) machine reading could discern important tools in metabolic engineering such as machine learning and engineering of gene regulatory element. But, it also suffers from the inability to distinguish between broad concepts and specific tools. This assertion refers to the constant picking up of broad concepts such as cell biology, molecular biology, and synthetic biology by the machine reading algorithms. While these concepts are encapsulations of a broad set of tools for metabolic engineering practice, they are not specific enough to inform novice researchers of metabolic engineering which course to take or tools to use. More importantly, repeated picking up of these broad concepts by the two-word (phrase) machine reading algorithm could be ascribed to the use of these phrases as subject areas of the blog posts, which were also part of the reading material.

Finally, two-word (phrase) machine reading could identify useful tools, resources and targets of metabolic engineering from the collection of blog posts used as reading material. Tools such as pathway engineering, transcription factor, and regulatory element are commonly used in modern metabolic engineering for tuning gene expression level and modulate pathway flux. But, surprisingly, the machine reading algorithm could also pick up important trends in metabolic engineering such as building of pathways that minimize carbon loss, or expanding the substrate specificity of enzymes through protein engineering. Overall, all three tiers of concepts could be elucidated by analysis of high frequency phrases identified by two-word (phrase) machine reading, and effort was not expended in analysing low frequency of occurrence terms.

4. Conclusions

Given the wealth of information in the literature and its esoteric nature, software tools are needed to help lend organization and structure to the literature for ease of comprehension of major concepts by decision makers in all levels of government. Machine reading is one such tool. This work uses ensemble statistical machine reading to identify high frequency of occurrence terms in the field of metabolic engineering from a collection of blog posts and abstracts. Results revealed that quality of the readout from machine reading depends critically on the breadth and size of the reading material. In general, ensemble statistical machine reading is a good tool for providing some initial understanding of metabolic engineering, but, manual curation and inference remains necessary for categorizing different concepts into distinct bins such as “key concepts”, “approaches and methodologies”, and “tools, resources and targets”. Overall, the current limitation in statistical machine reading lies in its computationally intensive nature, which calls for the use of supercomputing or quantum computing resources for ameliorating the exponential scaling of computational tasks as the size of the reading material expands.

Conflicts of interest

The author declares no conflicts of interest.

Funding

No funding was used in this work.

References

1. Hu, M. *et al.* Read + Verify: Machine Reading Comprehension with Unanswerable Questions. *Proc. AAAI Conf. Artif. Intell.* **33**, 6529–6537 (2019).

 doi:10.20944/preprints202110.0382.v1

2. Sun, K., Yu, D., Yu, D. & Cardie, C. Improving Machine Reading Comprehension with General Reading Strategies. *ArXiv181013441 Cs* (2019).
3. Liu, X., Shen, Y., Duh, K. & Gao, J. Stochastic Answer Networks for Machine Reading Comprehension. *ArXiv171203556 Cs* (2018).
4. Arendt, F. & Karadas, N. Content Analysis of Mediated Associations: An Automated Text-Analytic Approach. *Commun. Methods Meas.* **11**, 105–120 (2017).
5. Anson, I. G., Moskovitz, C. & Anson, C. M. A Text-Analytic Method for Identifying Text Recycling in STEM Research Reports. *J. Writ. Anal.* **3**, (2019).