scientific reports

Check for updates

OPEN

Identifying protein subcellular localisation in scientific literature using bidirectional deep recurrent neural network

Rakesh David[1]✉, Rhys-Joshua D. Menezes[2], Jan De Klerk[2], Ian R. Castleden[3], Cornelia M. Hooper[3], Gustavo Carneiro[2] & Matthew Gilliham[1]

The increased diversity and scale of published biological data has to led to a growing appreciation for the applications of machine learning and statistical methodologies to gain new insights. Key to achieving this aim is solving the Relationship Extraction problem which specifies the semantic interaction between two or more biological entities in a published study. Here, we employed two deep neural network natural language processing (NLP) methods, namely: the continuous bag of words (CBOW), and the bi-directional long short-term memory (bi-LSTM). These methods were employed to predict relations between entities that describe protein subcellular localisation in plants. We applied our system to 1700 published *Arabidopsis* protein subcellular studies from the SUBA manually curated dataset. The system combines pre-processing of full-text articles in a machine-readable format with relevant sentence extraction for downstream NLP analysis. Using the SUBA corpus, the neural network classifier predicted interactions between protein name, subcellular localisation and experimental methodology with an average precision, recall rate, accuracy and F1 scores of 95.1%, 82.8%, 89.3% and 88.4% respectively (n = 30). Comparable scoring metrics were obtained using the CropPAL database as an independent testing dataset that stores protein subcellular localisation in crop species, demonstrating wide applicability of prediction model. We provide a framework for extracting protein functional features from unstructured text in the literature with high accuracy, improving data dissemination and unlocking the potential of big data text analytics for generating new hypotheses.

Experimental techniques to characterize proteins at a biochemical, structural and physiological level have improved considerably in the last 25 years providing researchers with the tools necessary to understand protein function at a cellular and an organism level. Combined with detailed functional data, large-scale genome sequencing efforts have also greatly increased the scale of proteomic data available from model and non-model species. However, a major challenge facing researchers today is simply keeping pace with the sheer volume of low-throughput and high-throughput data being generated. Although scientific publications are used to disseminate research findings to the wider community, manually identifying, curating and collating individual experiments is a time and labour intense process. In addition, the large number of papers published, and the unstructured and versatile format of the content make the data difficult to integrate and analyse. The development of biological databases to curate protein experimental data from published literature have, as a result, become indispensable research tools for researchers. However, manual curation makes it difficult for databases to keep up with the ever-increasing amounts of data being generated.

To address these issues, automated and augmented curation systems for extracting protein functional data from scientific literature are becoming increasingly desired. In particular, Machine Learning and Natural Language Processing techniques are beginning to be employed for biocuration efforts[1,2] for extracting and organising unstructured biological information into a structured form that is accessible to biologists. Central to these

[1]School of Agriculture, Food and Wine, The Waite Research Institute, ARC Centre of Excellence in Plant Energy Biology, Waite Campus, The University of Adelaide, Adelaide, SA, Australia. [2]School of Computer Science, Australian Institute for Machine Learning, The University of Adelaide, Adelaide, SA, Australia. [3]ARC Centre of Excellence in Plant Energy Biology, The University of Western Australia, Perth, WA, Australia. ✉email: rakesh.david@adelaide.edu.au

automated systems, is the process of unambiguously extracting semantic relationships between two or more biological entities in the literature[3]. When referring to proteins or genes, entity relationships can describe protein–protein interaction, drug interactions, physicochemical properties or functional motifs, disease/trait interaction, and are very useful for biological network construction[4–7]. In addition, ML techniques have the potential to assist in protein annotation efforts such as the manually curated UniProt database by automatically extracting protein knowledge from research articles[2].

Initial methods for entity relationship extractions (RE) relied on dictionary-based and rule-based methods and parsers[8]. This type of system worked well for entity detection if the dictionary was big but still struggled to extract relations especially if there was a large distance between related words. Further improvements to biological entity RE were described by Fundel et al. using a multi-step process[9]. The approach relies on first preprocessing data and then creating a parse tree of the sentence that can be filtered against extraction rules. Other methods rely on Support Vector Machines (SVM) and kernel-based solutions or by combining the two approaches[10]. In the combined approach, kernels were utilized by converting the input text to a vector format and then finding the similarity between two entities and the sum of their sub structures similar to a "bag of words" approach. A SVM was then used for entity classification. This gave both the benefits of kernels such as being able to search a large feature space and regularization methods such as boundary detection between classes from SVM's. These types of methods improved the task of relation classification but ultimately were suboptimal in the field of relation detection.

With the advent of machine learning, many methods started to treat words as sequence data and applied deep learning models to find solutions. In the study by Li et al. the authors suggest creating a dependency tree which then feeds into a Convolutional Neural Network (CNN) to pretrain character level embeddings from words in the tree[11]. The tree was then passed into a bi-directional Long Short-Term Memory (bi-LSTM) model with the word embeddings and eventually to a dense layer. The end result was then put through a softmax layer and returned a binary classification corresponding to if the sentence contains an entity relation or not. It is shown that this method gets good results, but a major disadvantage is that due to the dependency tree, inter-sentence relations are not able to be extracted, limiting the scope of the biological entity integration.

Furthermore, text analytic methods described for biological research are largely optimised for extracting binary entity relationships such as interaction between two proteins or association between proteins and diseases[12]. Often, the experimental methodology that was used to verify the result, although mentioned in the sentence, is not processed as an entity type and hence not linked to the protein. The experimental methodology is an important factor for interpreting the results. For example, when analysing protein subcellular localisation, two commonly used approaches are fluorophore tagging of the protein (e.g. Green Fluorescent Protein) or by a proteomic approach of identifying all the proteins within a specific cellular compartment (e.g. Mass Spectrometry). Each approach has their advantages as the fluorophore approach provides better spatial resolution of protein location whereas MS methods allow for better quantitative analysis of cellular proteomes. Knowing what technique was used together with protein subcellular information can guide researchers in designing future experiments.

In this study, we describe the development and implementation of a pipeline to extract protein subcellular information and the experimental methodology from published studies using deep learning techniques. Protein subcellular location was chosen as a feature to extract as knowing where a protein resides within a cell provides important clues to its cellular function and represents a fundamental unit of how proteins function in nature. We tested the deep learning system across two datasets, the SUBcellular location database for Arabidopsis proteins (SUBA)[13] and another dataset that includes Crop Proteins with Annotated Locations (CropPAL)[14]. We demonstrate, through this method, triplet entity relationships can be predicted with high accuracy. Thus, the pipeline provides a framework for high-throughput extraction and linking of biological entities from unstructured text giving researchers access to the latest scientific information from diverse datasets.

Results and discussion

Outline of the approach. We describe a semi-automated pipeline for extraction of complex protein features from unstructured published literature (Fig. 1). We analysed articles describing subcellular localisation of proteins experimentally determined by fluorescence tagging and mass spectrometry, two commonly used techniques to resolve spatial and quantitative properties of protein in living cells[13]. Three biological entity types were considered for this classification: protein name, experimental methodology and subcellular location. Following retrieval of full-text articles, subsections and sentences that contain relevant biological entities were identified by text parsing, followed by annotation of entities and relationship types. By using full-text for pre-training vocabulary (continuous bag of words) and a set of annotated sentences for training the classifier (bi-directional long short-term memory), the deep learning model was subsequently assessed for its ability to classify 'True' or 'False' relationships between these three entity types using two independent and non-overlapping testing datasets (SUBA, CropPAL). A 'True' relationship between the entities indicates the given protein was experimentally verified to be located within a specific cellular organelle and a 'False' relationship indicates the co-occurrence of these entity types without being related. Implementation details and source codes of the pipeline are available at GitHub: https://github.com/RhysMenezes/find-a-protein.

SUBA dataset and literature retrieval. The SUBA collection (version 4.0) (http://suba.live/) includes subcellular information relating to 11,740 proteins in Arabidopsis that was manually extracted from 1768 published studies by expert curators in the plant biology field[13]. The database contains experimentally verified and predicted subcellular information relating to *Arabidopsis* proteins. For the purposes of this study however, we only selected proteins with experimentally verified information as this is considered as a high confidence dataset for *Arabidopsis* protein subcellular locations. Experimental evidence is based on fluorescence protein visualisa-

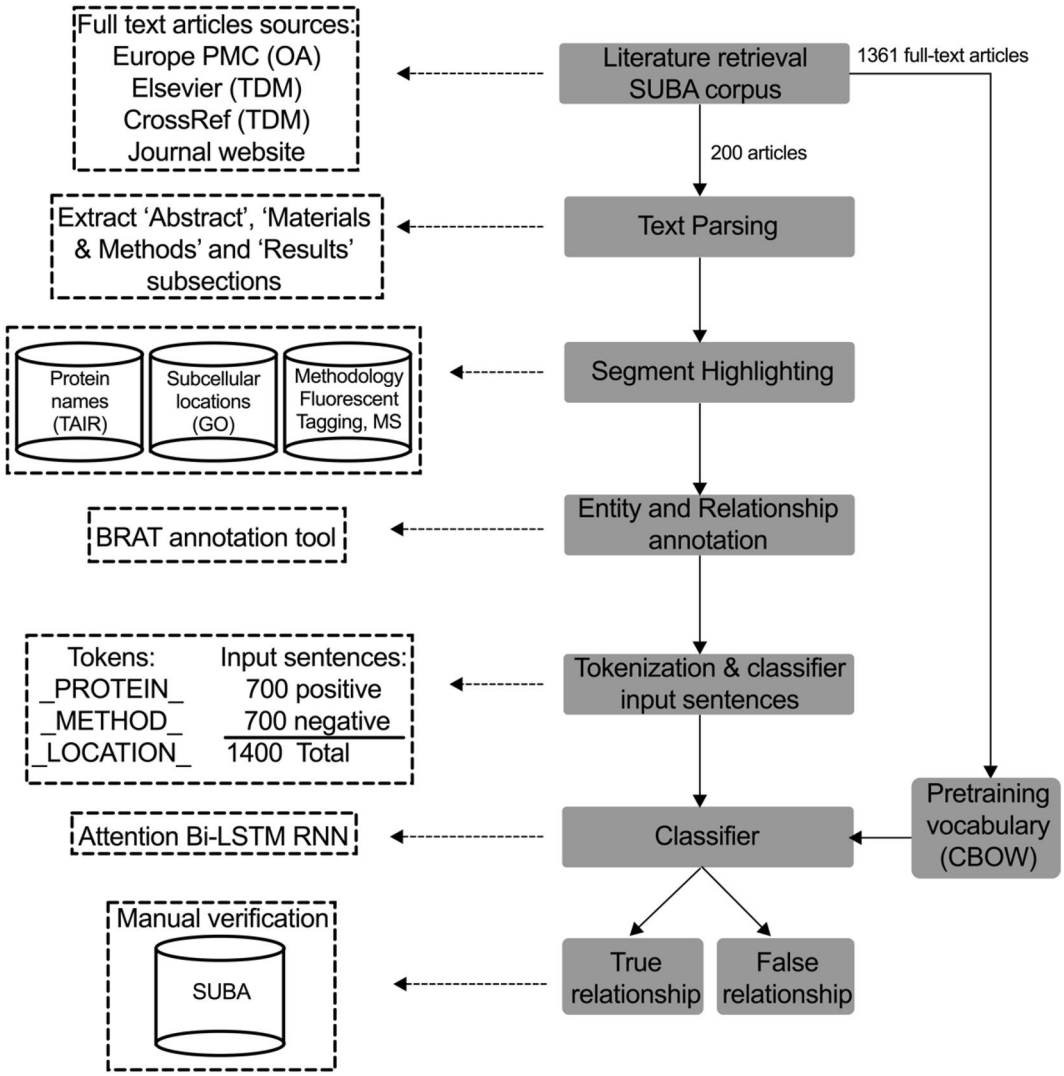

Figure 1. Overview of the deep learning approach to extract protein subcellular information from scientific literature. A summarised workflow of the system is shown on the right (grey boxes). Technical details for the implementation of each step including data sources, methodology and input examples is shown on the left (dashed boxes). The system was trained and tested using published studies from the SUBA corpus of protein subcellular data in *Arabidopsis*. Using the Pubmed ID in SUBA, full text articles were bulk retrieved from Europe PMC open access (OA) repository, text and data mining (TDM) services from Elsevier and CrossRef as well as directly from journal publisher website. A subset of 200 articles were used for annotating protein subcellular information to be used subsequently for classifier training and testing purposes. Triplet entities and relationships were annotated using the BRAT tool, following which protein names, experimental methodology and subcellular location were replaced with tokens to create input sentences for the RNN classifier. A second neural network, Continuous Bag of Words (CBOW) was employed to pretrain the classifier using the complete 1700 full-text articles from SUBA. The classifier determines a binary true or false classification for an input sentence to contain a valid triplet entity relationship. The results of the classifier were manually verified using the input sentences extracted from the SUBA articles.

tion and large-scale subcellular proteomics extracted from published studies and organised in a MySQL database format. Access to full-text was a requirement for the NLP analysis as often the protein subcellular information, and particularly the supporting experimental evidence are not described in the abstracts accessible in PubMed. As SUBA only stores citation details for proteins with experimentally verified subcellular information and not the associated full-text, we used a combination of strategies including an in-house developed NLP publication retrieval and processing tool (outlined in the "Methods" section) to analyse 1700 full-text articles for protein subcellular entity relationship prediction.

Parsing and segment highlighting. Published full-text articles were processed into XML-tagged documents, and the 'Abstract', 'Results' (including figure captions) and 'Materials and Methods' subsections were

extracted as these are likely to include protein subcellular information and experimental methods (pre-processed articles available in the GitHub repository, https://github.com/RhysMenezes/find-a-protein/tree/master/Data). Given the time-consuming process of manual annotation, we chose a random subset of 200 studies from the SUBA corpus to annotate entities and relationships. To identify relevant text, article subsections were parsed using regular expressions terms to find co-occurrences of three biological entities; protein names, subcellular locations and experimental methodology in a forty-word sliding window. Entity names were matched against a list of all Arabidopsis protein names (including synonyms) from The Arabidopsis Information Resource (TAIR), subcellular compartment names derived from the Gene Ontology cellular components annotations[15] and fluorescence tagging methodologies (Supplementary Tables S1–S2). Using this strategy, passages of text that include one or more triplet entity groups as well as neighbouring contextual words were extracted and used as the primary data for testing and training the recursive classifier.

Annotation of entity types and relationships. From the 200 pre-processed papers, we retrieved 1400 sentences from the segment highlighting approach and these were subsequently annotated for one or more triplet entity groups. In order to validate the correct triplet entity was being used for the deep learning prediction model, entities were labelled and the relationship between them were specified using the brat rapid annotation tool (BRAT)[16] (Fig. 2, Supplementary Fig. S1). Protein name was used as the common entity type to specify the relationship between the methodology and subcellular location described in the text. As the description of protein names can vary in individual studies, three entity types were used for annotation; abbreviated names of proteins including synonyms ('protein') (TAIR), modified proteins as result of mutations, truncations ('protein_variant') and proteins that are described as fluorescent tagged versions ('tagged_protein'). With relation to subcellular studies, we found tagged proteins in which both a protein name and methodology is concatenated in the same word, (e.g. SUVR2a-GFP[17]), were the most commonly used protein descriptors in the SUBA corpus. In such cases, the concatenated word gets labelled as one entity type called 'tagged protein' and sub labels are used to differentiate between protein and methodology (Supplementary Fig. S1, Supplementary Table S3)[18,19].

Annotation cleaning and tokenization. The RNN model's task is to classify an input sentence as either 'True' relationship or 'False' relationship, the output is discrete either a 1 (true) or a 0 (false). For all input sentences annotated, a 'True' relationship is defined as one that satisfies both semantic and syntactic relationship between a triplet entity group. The entities are considered to have a semantic relationship if they are structured such that the protein's subcellular location and the methodology used can be logically discerned from the sentence (Fig. 2).

For the input data used in the RNN classifier, the actual protein names, methodologies and subcellular locations were not needed as we investigated the possibility of classifying relationships using the entities neighbouring context words. The names in the input sentences were replaced by tokens that represent the entity types. All protein names were replaced with a "_PROTEIN_" taken, methodologies were replaced with a "_METHOD_" token and subcellular locations were replaced with a "_LOCATION_" token. As the model classifies if a given relationship is true or false, the input sentences will have to be presented such that only one relationship is present in a sentence. If a sentence contains multiple tokens then a valid triplet will be tokenized, and the rest blanked out using _BLANK_ tokens. For example, the sentence from Tanz et al., "Both Deg1 and Deg9 have been localized to the plastid and nucleus using mass spectrometry"[21], four separate true states, one for each triplet group (underlined) can be presented to the classifier:

1. "Both ***_PROTEIN_*** and ***_BLANKP_*** have been localized to the ***_LOCATION_*** and ***_BLANKL_*** using mass ***_METHOD_***."
2. "Both ***_PROTEIN_*** and ***_BLANKP_*** have been localized to the ***_BLANKL_*** and ***_LOCATION_*** using ***_METHOD_***."
3. "Both ***_BLANKP_*** and ***_PROTEIN_*** have been localized to the ***_LOCATION_*** and ***_BLANKL_*** using ***_METHOD_***."
4. "Both ***_BLANKP_*** and ***_PROTEIN_*** have been localized to the ***_BLANKL_*** and ***_LOCATION_*** using ***_METHOD_***."

Lastly to ensure all input data was consistent and the dimensions remained the same in the classifier model, sentences were padded using the PAD term to the maximum sequence length. From 1400 tokenized text input sentences, we created 700 positive examples that satisfied both semantic and syntactic criteria for entity relationships. To generate negative cases, we randomly permutated the entities within the sentences to create 700 negative examples.

Pretraining: CBOW. In order to utilize the SUBA data more effectively and due to the relatively small number of annotated sentences used for training the classifier, we pretrained the network's vocabulary on the SUBA dataset comprising of 1361 full-text articles. Pretraining the model with plant biology specific domain knowledge has the advantage of improving the performance of the classifier in entity relationship prediction. Pretraining was achieved using CBOW NLP model that represents the meaning of each word in a sentence as a single weighted vector representations[22]. This allows the CBOW model to predict the next word in the sentence given the target word's neighbouring context words. The trained word representations/embeddings were extracted and integrated into the classifier model.

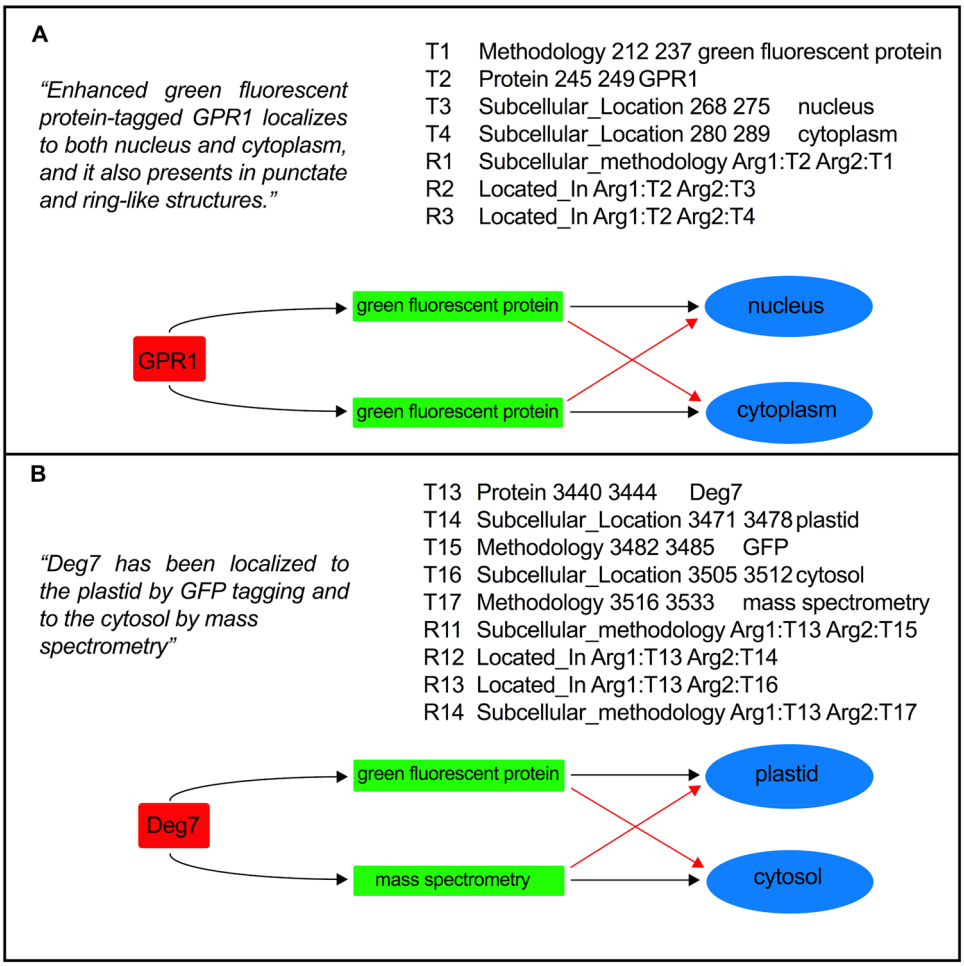

Figure 2. Two examples of protein subcellular information described in the SUBA corpus text and mapping of entity relationships. For both examples, the sentence from the article is quoted (top left), BRAT annotation defining the entities and relationships for that sentence in standoff format (top right) and a visual network representing the interaction between protein (red), experimental methodology (green) and cellular location (blue) is shown (bottom). The example shown in (**A**) is from Yang et al. and illustrates a simple semantic and syntactic relationship between the protein, GPR1, experimental methodology using green fluorescent protein (GFP) and subcellular locations, nucleus and cytoplasm[20]. In this example, the classifier predicting the relationship for each triplet entity group is straightforward as both groups use GFP as the common methodology. In this case, an incorrect classifier prediction between the two entity groups (red arrows in both panels) would still result in an overall correct relationship being defined. The example shown in (**B**) is from Tanz et al. and illustrates a more complex sentence in which Deg7 protein is demonstrated to be located in the plastid and cytosol using GFP and mass spectrometry, respectively[21]. Here, the classifier would need to use the contextual words in the input sentence to correctly identify the syntactic relationship between the three entities in each group.

Bi-directional LSTM classifier model. The classifier model's architecture is that of a bidirectional LSTM network, which is a type of Recurrent Neural Network (RNN) and has the advantage of capturing long-term contextual information and is suitable for NLP tasks involving long sequence of words. Similar models have been used for biomedical named entity recognition (BNER) to recognize proteins and gene names and for Relation Extraction tasks from unstructured text[23]. However, relation extraction in the biomedical domain has been mainly restricted to binary relations between entities, such as protein–protein interaction[24], drug–drug interaction[25]. Here, we evaluate the accuracy of the model to extract relations from three entities incorporating both biological information and experimental methodology into the prediction model. The model takes annotated sentences as inputs and reads it word by word to determine a binary classification of a "yes" or a "no" (1 or 0) if a sentence contains a valid entity relationship between the protein, methodology and subcellular location. The input data was split into a ratio of 1:4, 20% for testing and 80% for training the classifier, making sure the data was separated so that papers did not overlap between the testing and training sets. This prevents the classifier model from analysing very similar input sentences both within the training and testing data sets.

We achieved an average accuracy score of 89.3% (n = 30 experimental runs, standard deviation = 0.028) for all test input sentences in the SUBA testing data set (Table 1, Supplementary Table S3–S4) (Fig. 3). Interestingly,

Dataset	LSTM model	Precision (%)	Recall (%)	Accuracy (%)	F-score (%)
SUBA	Bidirectional	95.17 ± 2.4	82.80 ± 6.1	89.39 ± 2.8	88.40 ± 3.6
	Unidirectional	95.32 ± 1.9	79.26 ± 7.1	87.80 ± 3.0	86.35 ± 4.0
CropPAL	Bidirectional	89.44 ± 2.6	92.56 ± 6.7	89.79 ± 3.7	90.82 ± 3.8
	Unidirectional	89.20 ± 2.0	89.87 ± 5.6	88.37 ± 3.2	89.53 ± 3.2

Table 1. Mean and standard deviation of the precision, recall, accuracy and F1-score for the tested uni- and bi-directional LSTM models on the SUBA and CropPAL datasets.

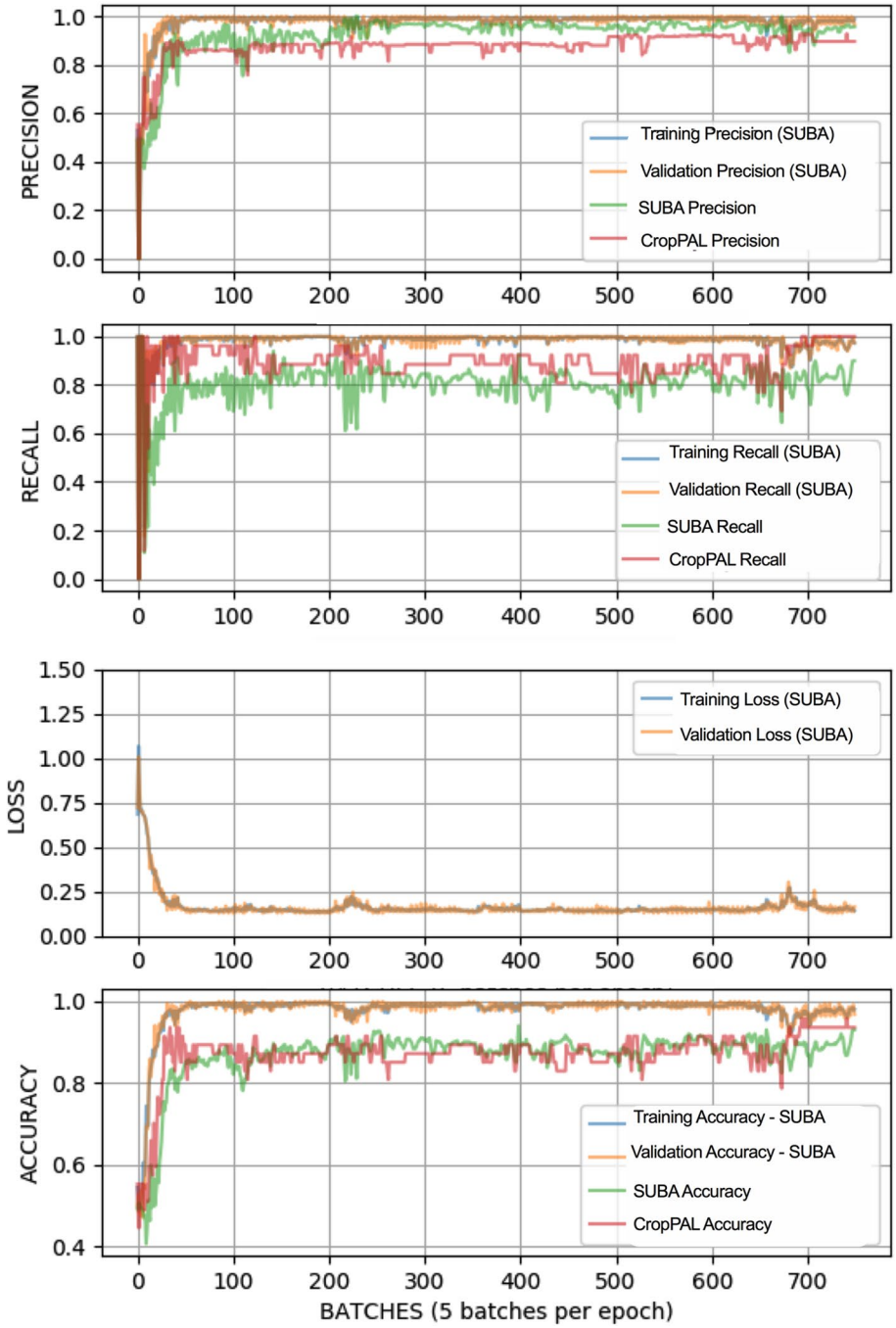

Figure 3. Precision, recall, loss and accuracy progression during the training for the Bi-LSTM to classify entity relationships between proteins, experimental methodology and subcellular locations from input sentences. Training, validation and testing of the model was undertaken using the SUBA dataset and independently tested using the CropPAL dataset.

the classifier was able to provide more accurate predictions for negative input sentences in which the entities did not satisfy semantic and syntactic criteria (95.5%) compared to positive input statements (86.5%). Furthermore, when comparing input sentences with one or more triplet entity groups, the prediction accuracy was observed to be higher for single triplet groups (95.6%), containing a single protein, a subcellular location and an experimental methodology compared to sentences with multiple triplet groups (90.8%). Nevertheless, an overall accuracy of 89.3% is still, to the best of our knowledge, a significant improvement among similar neural network and machine learning approaches used previously for predicting relationships for triplet entity groups using unstructured biomedical text. When evaluating drug-gene-mutation interactions, a graph-based LSTM model achieved an average accuracy of 80.7% for triplet entity relation extraction[26]. Although our model improvedaccuracy by 8.6% compared to this study, the different dataset, triplet entity group and annotation used in this study make any direct comparison difficult. A comparable study using a distant supervised learning approach for predicting relation between a protein and a subcellular location, achieved an accuracy of 82%[27], a drop in 7.3% compared with our model and crucially focussed only on binary relations without including experimental methodology as we did in our study.

Independent testing of the classifier performance using the CropPAL database. We next tested the Bi-LSTM RNN for its accuracy in predicting protein subcellular localisation using an independent dataset. The CropPAL (ver 1.0) (http://crop-pal.org/) database stores published protein subcellular information for the crop species rice, wheat, barley and maize using a similar manual curation process to the SUBA database. Although both databases address similar problems in curating protein subcellular information, there was no overlap in the published articles or input sentences used in training and testing from the SUBA corpus and the subsequent independent testing from the CropPAL corpus. Literature retrieval and pre-processing of the full-text article and input statements for CropPAL were carried out as described for the SUBA dataset (Fig. 1) (pre-processed articles available in the GitHub repository, https://github.com/RhysMenezes/find-a-protein/tree/master/Data). Using a smaller subset of randomly selected papers from the CropPAL dataset, 65 input sentences were extracted, annotated and tagged using the same process described above. Using the Bi-LSTM classifier, we achieved a similar accuracy score of 89.7% (n = 30 experimental runs, standard deviation = 0.037) as the SUBA dataset demonstrating transfer learning capacity of the model (Table 1, Fig. 3).

Bi-directional versus Uni-directional LSTM. As a comparison to the bidirectional LSTM, we tested the performance of a unidirectional LSTM (Uni-LSTM) on both SUBA and CropPAL datasets (Table 1). We found the performance dropped marginally for the SUBA dataset, with an average accuracy score of 87.8%, a decrease of 1.5% compared to Bi-LSTM. For CropPAL dataset, the accuracy score was also marginally lower, 88.3%, a decrease of 1.4% compared to Bi-LSTM. Overall, our analysis suggests that Bi-directional LSTM classifier offers a small improvement in predicting protein subcellular localisation over a uni-directional classifier. However, these results may depend on the type of data used in this study and future analysis will require testing of input sentences that describe other protein characteristics such as protein–protein interaction or protein–protein-disease/trait interaction to analyse the RNN classifier that offers the better overall predictive ability for biological entity relationship extraction.

Conclusion

We demonstrate the Bi-LSTM RNN classifier model uses contextual information in the input sentence to predict triplet entity relationships with a high degree of accuracy and can distinguish from input sentences in which the three entities co-occur by chance. This approach can be effective for extracting biological information from unstructured text in a meaningful way that can be made accessible to the research community. Earlier reports of Machine Learning and neural network models have primarily focussed on amino acid sequence information for predicting protein subcellular localisation with varying degrees of success[28–30]. While these approaches are important for uncharacterised proteins and those predicted from large-scale sequencing datasets, there is a wealth of published protein subcellular information that can be integrated and exploited for practical use by researchers. A Natural Language Processing system that relies on deep learning Technique such as the one described here will help bridge the gap between the data generation process and integrating the vast amount of published molecular and physiological information available about proteins. In particular, a future opportunity exists in improving the sustainability of biological databases by augmenting manual curation efforts with semi or fully automated neural network approaches. Despite their importance in primary research, biological databases face a survivability challenge due to funding cycles typically lasting 3–5 years and essential services such as data curation and maintenance becoming increasingly difficult to sustain[31]. This is best illustrated with the number of biological databases that are listed as obsolete in the annual database issues published by Nucleic Acid Research (NAR) journal. In 2020, NAR reported 125 discontinued databases, with the number of obsolete databases increasing five times over the last five years. In contrast, 65 new database resources were published in the same database issue, with the number remaining steady over the same five year period[32]. While the study here focuses on protein subcellular information, the pipeline developed provides a framework for capturing key functional properties of proteins from published studies and assisting in biocuration efforts. In addition to improving the overall prediction accuracy, we envisage the Bi-LSTM RNN-based pipeline to be applied for extracting relationships involving two or more entities that describe complex protein features such as protein interaction with other biomolecules such as proteins, DNA and RNA as well as for protein-phenotype associations. The transition from manual to semi-automated to potentially, fully automated biocuration systems that rely on deep-learning processes in the future can be exploited by researchers in expanding proteome knowledgebases and for predicting novel functions for individual proteins.

Entity 1	Entity 2	Relationship type
Protein	Subcellular_Location	Located_In
Tagged_Protein	Subcellular_Location	Located_In
Protein	Methodology	Subcellular_methodology
Protein_Locus	Subcellular_Location	Located_In
Protein_Locus	Methodology	Subcellular_methodology
Protein_variant	Subcellular_Location	Located_In
Protein_variant	Methodology	Subcellular_methodology
Protein	Identifier	Protein_identifier
Protein	Gene_Locus	Locus_link
Protein	Subcellular_Location	Not_Located_In
Tagged_Protein	Subcellular_Location	Not_Located_In

Table 2. BRAT annotation entity and relationship types defined for the SUBA and CropPAL corpora.

Methods

Bulk retrieval of full-text articles and RNN data split. As SUBA and CropPAL does not store full-text articles, we used PubMed ID (or DOI, when PubMed ID was not available) (http://suba.live/stats.html) to retrieve full-text articles from various sources from both datasets. For the SUBA database, 178 articles were retrieved from the Europe PMC repository open access full-text RESTful API service (https://europepmc.org/RestfulWebService). In addition, full-text articles from the SUBA dataset were also retrieved from the Elsevier text mining service and the CrossRef association of scholarly publishers bringing the total number of articles to 227. The remaining full-text journal articles were retrieved and processed using an in-house developed tool called 'NLP-ready' bringing the total number included in the analysis to 1700 published studies. For the CropPAL dataset, a smaller subset of 10 randomly selected full-text articles from the CropPAL dataset were retrieved from which 65 input sentences were extracted for testing the model performance. All pre-processed articles from the SUBA and CropPAL corpora and extracted input sentences used for training and testing purposes are available in the data folder of GitHub repository: https://github.com/RhysMenezes/find-a-protein/tree/master/Data. Details for the automated retrieval and pre-processing tool can be accessed from: https://github.com/arabidopsis/NLP-ready.

A smaller subset of 200 randomly selected articles from the SUBA corpus were used for the annotation of entity types and classifying entity relationships. The annotated data from this group was used for training and testing the Bi-LSTM recursive neural network classifier. In addition, 1361 articles were also pre-processed by passing the full text through a Continuous Bag of Words (CBOW) artificial neural network model from Word2Vec[33] for pretraining the vocabulary as an additional input for the RNN training.

BRAT annotation. Protein subcellular information and associated methodology in the text was manually annotated with BRAT[16], using standoff format in which annotation information is kept separate from the text (Fig. 2, Supplementary Fig. S1). Each annotation between a protein, a subcellular location and a methodology include pairwise relations specified through the type of relation and the two entities (Table 2). Annotated input sentences for SUBA and CropPAL are available from the GitHub repository, https://github.com/RhysMenezes/find-a-protein/tree/master/Data.

Bi-LSTM model. We propose a bidirectional LSTM model that contains an Input layer, Embedding layer, Bi-directional LSTM layer, Hidden layer and an Output layer (Fig. 4).

A simpler LSTM is the unidirectional, which is limited because it only considers past and disregard context. When reading sequences, it is helpful to get both past and future word context, such as the one provided by Bi-LSTM. As shown in Fig. 4, after the input, CBOW and Bi-LSTM layers, the hidden states from the Bi-LSTM layers are combined at the end using concatenation using drop out at rate of 0.5 as regularization for the network training. We compare the functionality of the two kinds of LSTMs in the section 'Bi-directional versus Uni-directional LSTM'.

Output layer. The output layer contains a dense layer with a Tanh activation. This is then passed through a log softmax and then an argmax which gave us the binary classification of 'True' relationship and 'False' relationship for every input sentence, as in:

$$y^* = \arg\max log\left(\frac{exp(x_i)}{\Sigma_j exp(x_i)}\right),$$

where, x_i represents individual words as integer vector representation.

Experiments. On the training set we performed k-fold cross validation with fivefolds, to achieve the validation and training split ratio of 1:4. When training, the training set was shuffled after each epoch, so that we would

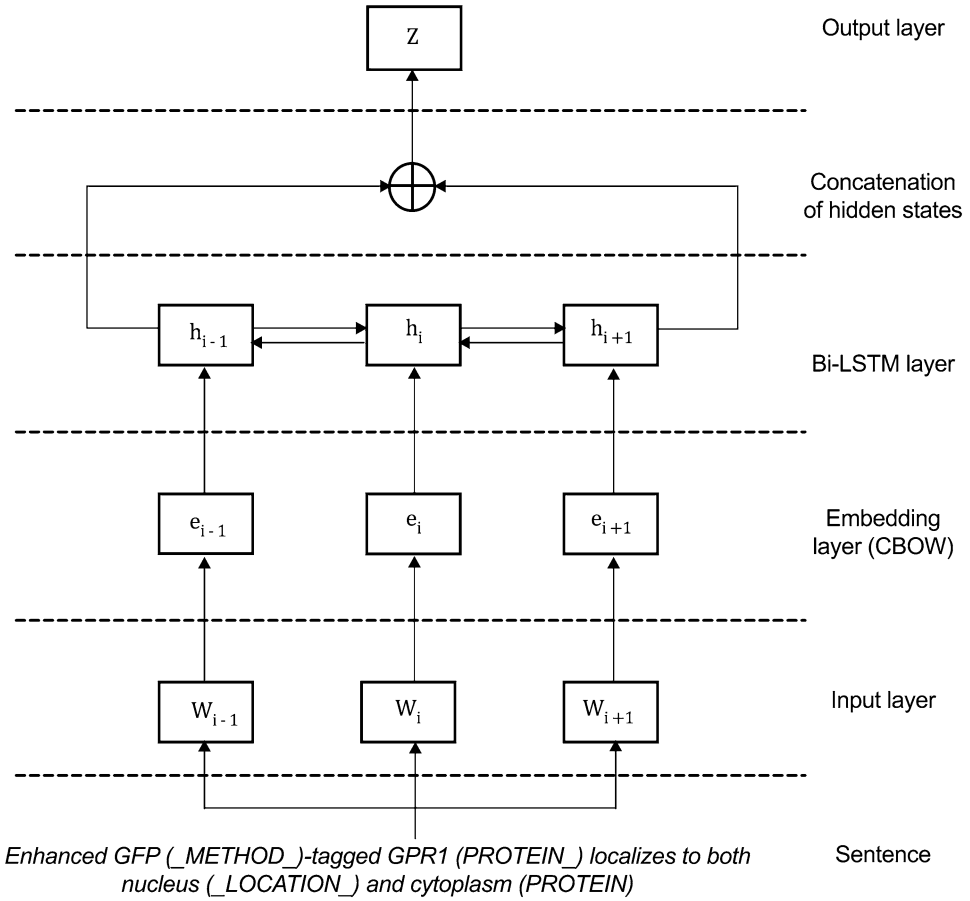

Figure 4. The bi-directional LSTM model to recognize protein subcellular localisation from published literature.

be able to perform the k-fold split such that we would not get repeating folds when training. For the fivefolds, each fold was a validation set, making each epoch 5 experiments long.

Model parameters used to generate the results:
Learn rate = 0.01.
Dropout rate = 0.50.
Hidden nodes = 150.
Weight decay = 0.0001.
Learn rate decay = 500.
Optimizer = ADAM.

The ADAM optimizer was used to perform back propagation and train the network with a learning rate of 0.01, dropout rate of 0.50, weight decay of 0.0001. The hidden state of the LSTM had a feature size of 150.

K-folds cross validation was also utilized since there was a sparse amount of data. This was coupled with cross entropy loss to achieve the best results. We measured the performance of the network with four metrics: accuracy, precision, recall, and F1 score, using the network's predictions' true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). To test our result, we use an F1 score which is the balance between recall and precision.

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall},$$

where

$$Precision = \frac{True\ Positive}{True\ Positive + False\ positve},$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}.$$

The scores of Recall, Precision, Loss and F1 for 30 experimental runs for both SUBA and CropPAL testing dataset is provided in the GitHub repository: https://github.com/RhysMenezes/find-a-protein. The output for

the Bi-LSTM and Uni-LSTM prediction was independently verified through manual inspection by an expert in the field using the unannotated input sentences extracted from articles.

Data availability

The datasets generated during the current study are available in the GitHub repository: https://github.com/RhysMenezes/find-a-protein/tree/master/Data. https://github.com/arabidopsis/NLP-ready

Received: 14 September 2020; Accepted: 17 December 2020
Published online: 18 January 2021

References

1. König, C. *et al.* Using machine learning tools for protein database biocuration assistance. *Sci. Rep.* **8**, 10148. https://doi.org/10.1038/s41598-018-28330-z (2018).
2. Teodoro, D. *et al.* UPCLASS: A deep learning-based classifier for UniProtKB entry publications. *Database (Oxford).* https://doi.org/10.1093/database/baaa026 (2020).
3. Altman, R. B. *et al.* Text mining for biology--The way forward: opinions from leading scientists. *Genome Biol.* **9**(Suppl 2), S7. https://doi.org/10.1186/gb-2008-9-s2-s7 (2008).
4. Islamaj Dogan, R. *et al.* Overview of the BioCreative VI Precision Medicine Track: Mining protein interactions and mutations for precision medicine. *Database (Oxford)* https://doi.org/10.1093/database/bay147 (2019).
5. Xing, W. *et al.* A gene–phenotype relationship extraction pipeline from the biomedical literature using a representation learning approach. *Bioinformatics (Oxford, England)* **34**, i386–i394. https://doi.org/10.1093/bioinformatics/bty263 (2018).
6. Zhang, W. *et al.* Recent advances in the machine learning-based drug-target interaction prediction. *Curr. Drug Metab.* **20**, 194–202. https://doi.org/10.2174/1389200219666180821094047 (2019).
7. Cunningham, J. M., Koytiger, G., Sorger, P. K. & AlQuraishi, M. Biophysical prediction of protein–peptide interactions and signaling networks using machine learning. *Nat. Methods* **17**, 175–183. https://doi.org/10.1038/s41592-019-0687-1 (2020).
8. Ono, T., Hishigaki, H., Tanigami, A. & Takagi, T. Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics (Oxford, England)* **17**, 155–161. https://doi.org/10.1093/bioinformatics/17.2.155 (2001).
9. Fundel, K., Küffner, R. & Zimmer, R. RelEx–relation extraction using dependency parse trees. *Bioinformatics (Oxford, England)* **23**, 365–371. https://doi.org/10.1093/bioinformatics/btl616 (2007).
10. Culotta, A. & Sorensen, J. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04).* 423–429.
11. Li, F., Zhang, M., Fu, G. & Ji, D. A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinform.* **18**, 198. https://doi.org/10.1186/s12859-017-1609-9 (2017).
12. Zhou, D., Zhong, D. & He, Y. Biomedical relation extraction: From binary to complex. *Comput. Math. Methods Med.* **2014**, 298473. https://doi.org/10.1155/2014/298473 (2014).
13. Hooper, C. M., Castleden, I. R., Tanz, S. K., Aryamanesh, N. & Millar, A. H. SUBA4: The interactive data analysis centre for Arabidopsis subcellular protein locations. *Nucleic Acids Res.* **45**, D1064-d1074. https://doi.org/10.1093/nar/gkw1041 (2017).
14. Hooper, C. M., Castleden, I. R., Aryamanesh, N., Jacoby, R. P. & Millar, A. H. Finding the subcellular location of barley, wheat, rice and maize proteins: The compendium of crop proteins with annotated locations (cropPAL). *Plant Cell Physiol.* **57**, e9. https://doi.org/10.1093/pcp/pcv170 (2016).
15. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29. https://doi.org/10.1038/75556 (2000).
16. Stenetorp, P. *et al.* In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 102–107.
17. Thorstensen, T. *et al.* The Arabidopsis SUVR4 protein is a nucleolar histone methyltransferase with preference for monomethylated H3K9. *Nucleic Acids Res.* **34**, 5461–5470. https://doi.org/10.1093/nar/gkl687 (2006).
18. Vorwerk, S. *et al.* EDR2 negatively regulates salicylic acid-based defenses and cell death during powdery mildew infections of *Arabidopsis thaliana*. *BMC Plant Biol.* **7**, 35. https://doi.org/10.1186/1471-2229-7-35 (2007).
19. Chi, Y. H. *et al.* AtSRP1, small rubber particle protein homolog, functions in pollen growth and development in Arabidopsis. *Biochem. Biophys. Res. Commun.* **475**, 223–229. https://doi.org/10.1016/j.bbrc.2016.05.083 (2016).
20. Yang, X. *et al.* The Arabidopsis GPR1 gene negatively affects pollen germination, pollen tube growth, and gametophyte senescence. *Int. J. Mol. Sci.* https://doi.org/10.3390/ijms18061303 (2017).
21. Tanz, S. K., Castleden, I., Hooper, C. M., Small, I. & Millar, A. H. Using the SUBcellular database for Arabidopsis proteins to localize the Deg protease family. *Front. Plant Sci.* **5**, 396. https://doi.org/10.3389/fpls.2014.00396 (2014).
22. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint* arXeiv:1301.3781 (2013).
23. Lyu, C., Chen, B., Ren, Y. & Ji, D. Long short-term memory RNN for biomedical named entity recognition. *BMC Bioinform.* **18**, 462. https://doi.org/10.1186/s12859-017-1868-5 (2017).
24. Quan, C., Luo, Z. & Wang, S. A hybrid deep learning model for protein–protein interactions extraction from biomedical literature. *Appl. Sci.* **10**, 2690 (2020).
25. Sahu, S. K. & Anand, A. Drug–drug interaction extraction from biomedical texts using long short-term memory network. *J. Biomed. Inform.* **86**, 15–24. https://doi.org/10.1016/j.jbi.2018.08.005 (2018).
26. Peng, N., Poon, H., Quirk, C., Toutanova, K. & Yih, W.-T. Cross-sentence n-ary relation extraction with graph lstms. *Trans. Assoc. Comput. Linguist.* **5**, 101–115 (2017).
27. Zheng, W. & Blake, C. Using distant supervised learning to identify protein subcellular localizations from full-text scientific articles. *J. Biomed. Inform.* **57**, 134–144. https://doi.org/10.1016/j.jbi.2015.07.013 (2015).
28. Cheng, X., Xiao, X. & Chou, K. C. pLoc-mPlant: Predict subcellular localization of multi-location plant proteins by incorporating the optimal GO information into general PseAAC. *Mol. Biosyst.* **13**, 1722–1727. https://doi.org/10.1039/c7mb00267j (2017).
29. Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H. & Winther, O. DeepLoc: Prediction of protein subcellular localization using deep learning. *Bioinformatics (Oxford, England)* **33**, 3387–3395. https://doi.org/10.1093/bioinformatics/btx431 (2017).
30. Zhang, N. *et al.* MU-LOC: A machine-learning method for predicting mitochondrially localized proteins in plants. *Front. Plant Sci.* **9**, 634. https://doi.org/10.3389/fpls.2018.00634 (2018).
31. Reiser, L. *et al.* Sustainable funding for biocuration: The Arabidopsis Information Resource (TAIR) as a case study of a subscription-based funding model. *Database (Oxford)* https://doi.org/10.1093/database/baw018 (2016).
32. Rigden, D. J. & Fernández, X. M. The 27th annual Nucleic Acids Research database issue and molecular biology database collection. *Nucleic Acids Res.* **48**, D1-d8. https://doi.org/10.1093/nar/gkz1161 (2020).
33. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. In *Advances in Neural Information Processing Systems.* 3111–3119.

Acknowledgements
This research was supported by University of Adelaide Interdisciplinary Research Funding Scheme awarded to M.G. and Australian Research Council through CE140100008 to M.G.

Author contributions
R.D., G.C., and M.G. designed the study. R.D., J.D.K, R.D.M., I.R.C., and C.M.H. performed the experiments. R.D., J.D.K, and R.D.M. performed data analysis. R.D., J.D.K, R.D.M., G.C., and M.G. prepared and edited the manuscript. All authors approved and edited the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-020-80441-8.

Correspondence and requests for materials should be addressed to R.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.