Article

MetaboListem and TABoLiSTM: Two Deep Learning Algorithms for Metabolite Named Entity Recognition

Cheng S. Yeung [1], Tim Beck [2,3,*] and Joram M. Posma [1,3,*]

[1] Section of Bioinformatics, Division of Systems Medicine, Department of Metabolism, Digestion and Reproduction, Faculty of Medicine, Imperial College London, London SW7 2AZ, UK; c.yeung20@imperial.ac.uk

[2] Department of Genetics and Genome Biology, University of Leicester, Leicester LE1 7RH, UK

[3] Health Data Research (HDR), London NW1 2BE, UK

[*] Correspondence: timbeck@leicester.ac.uk (T.B.); jmp111@imperial.ac.uk (J.M.P.)

Abstract: Reviewing the metabolomics literature is becoming increasingly difficult because of the rapid expansion of relevant journal literature. Text-mining technologies are therefore needed to facilitate more efficient literature reviews. Here we contribute a standardised corpus of full-text publications from metabolomics studies and describe the development of two metabolite named entity recognition (NER) methods. These methods are based on Bidirectional Long Short-Term Memory (BiLSTM) networks and each incorporate different transfer learning techniques (for tokenisation and word embedding). Our first model (MetaboListem) follows prior methodology using GloVe word embeddings. Our second model exploits BERT and BioBERT for embedding and is named TABoLiSTM (Transformer-Affixed BiLSTM). The methods are trained on a novel corpus annotated using rule-based methods, and evaluated on manually annotated metabolomics articles. MetaboListem (F1-score 0.890, precision 0.892, recall 0.888) and TABoLiSTM (BioBERT version: F1-score 0.909, precision 0.926, recall 0.893) have achieved state-of-the-art performance on metabolite NER. A training corpus with full-text sentences from >1000 full-text Open Access metabolomics publications with 105,335 annotated metabolites was created, as well as a manually annotated test corpus (19,138 annotations). This work demonstrates that deep learning algorithms are capable of identifying metabolite names accurately and efficiently in text. The proposed corpus and NER algorithms can be used for metabolomics text-mining tasks such as information retrieval, document classification and literature-based discovery and are available from the omicsNLP GitHub repository.

Keywords: deep learning; named entity recognition; natural language processing

1. Introduction

Since the late 1990s, metabolomics has contributed to better the understanding of the roles that small molecules play in health and disease [1]. The on-going developments of high-throughput data acquisition ensures that the technology is accessible to more researchers [2], evident by the increasing number of studies year-on-year for studies in any organism as well as those only done in humans (Appendix A). With the growth of the number of publications comes the difficulty of literature review, and performing a literature review of all papers relating to a specific phenotype has become an almost impossibly difficult task for researchers [3]. To facilitate more efficient review processes, computational literature mining tools are needed.

Natural language processing (NLP) is a branch of computer science that aims to understand natural human language computationally; in the biomedical field, NLP often refers to the technologies that allow computational tools to interpret scientific texts written by humans. In recent years, NLP has been successfully applied in many biomedical text mining tasks, including information retrieval, document classification and literature-based discovery [4–7]. By identifying biologically important entities in articles and computationally inferring the connection between them, NLP facilitates understanding of biological

relationships from textual information. Such efficient and automatic knowledge mining systems can provide a comprehensive viewpoint to researchers by analysing a vast amount of articles.

Some advances in metabolomics have been made for using computer algorithms to read scientific text [8–11], however these suffer from a few drawbacks such as using rule-based annotation methods and dictionary matching and corpora and code/models being unavailable. While these methods can indeed correctly capture a vast amount of metabolite names, they are dependent on the library of compound names used—and no library is exhaustive. The latest version of the Human Metabolome Database (HMDB) [12] reports over 18,557 metabolites have been detected and quantified in the human body, and 91,822 more are expected or predicted. Moreover, these numbers are still increasing, from 2180 (HMDB 1.0 in 2007) to 114,100 (HMDB 4.0 in 2017) in a decade [13] and predicted to be an order of magnitude higher [1]. Hence there is a need for generalisable approaches for automated metabolite recognition.

Within NLP, Named Entity Recognition (NER) is a field concerned with automatically detecting specific entities from texts. To facilitate biomedical text-mining, and NER in particular, BioCreative (Critical Assessment of Information Extraction in Biology) [14] have proposed multiple tasks/challenges and provided corpora for chemical entity recognition, such as the CHEMDNER (chemical compound and drug name recognition, BioCreative V4) [15] and CEMP (Chemical Entity Mention in Patents, part of BioCreative V5) [16] challenges. The use of machine learning techniques such as Conditional Random Fields (CRF) has shown promising results with F1-scores > 0.87 of tmChem [17] in the V4 challenge, however the CRF methods were surpassed by a Bidirectional Long Short-Term Memory (BiLSTM) recurrent neural networks CollaboNet [18] (F1-score of 0.88) and LSTMVoter [19] (F1-score of 0.90), and an attention-based BiLSTM [20] (Att-ChemdNER, F1-score of 0.91). For the V5 challenge, and the CEMP task, LSTMVoter achieved an F1-score of 0.89 and another deep learning (DL) system using BiLSTMs called ChemListem [21] achieved an F1-score of 0.90.

Adding a word embedding layer to NER DL models is well known to improve performance [22], and ChemListem used context-free word embeddings/representation using global vectors (GloVe) [23] for this purpose. However, in recent years big improvements have been made in NLP by using contextual embedding with transformers such as BERT (Bidirectional Encoder Representations from Transformers) [24]. BERT was trained on a corpus containing 3.3 billion words mainly from the English Wikipedia and BooksCorpus. Unlike the context-free embedding techniques, BERT's contextual embedding makes the vector representation of a word dependent on its context. Therefore, for biomedical text mining the BioBERT model [25] was pretrained on PubMed abstracts and PubMed Central (PMC) articles [26], and has been used to obtain F1-scores on the BioCreative V4 and V5 of 0.93 and 0.94, respectively, and F1-scores of 0.78–0.85 for recognising genes and proteins [27]. However, fine-tuned BERT-based models do not always yield state-of-the-art results as they greatly depend on the training data size [28].

There are three main issues with using these chemical NER methods for metabolite NER. First, as mentioned, with these methods the context matters and therefore training algorithms to recognise chemical entities in abstracts does not necessarily mean they work as well for full-text articles. Second, while a metabolite is a chemical entity, a chemical entity is not necessarily a metabolite, therefore these existing corpora and algorithms may result in many false positives when using them in metabolomics. Last, each metabolite can have multiple (sometimes ten or even a hundred) different synonyms—over 1.2 million synonyms are reported in HMDB)—adding to the complexity of using a dictionary of synonyms as new names can be used by different authors.

Our contribution is the development of a standardised, machine-readable metabolomics corpus of full-text Open Access (OA) PMC articles by using the Auto-CORPus (Automated and Consistent Outputs from Research Publications) package [29] for text standardisation in conjunction with semi-automatic annotation of metabolites in full texts using a combination of dictionary searching (using HMDB to stay in line with prior work [10]), regular

expression matching and rule-based approaches. We then use this corpus to train three DL-based algorithms (with different word embedding layers) to perform metabolite NER with the aim to obtain a generalisable model that can be used to speed up metabolomics literature review. Our work addresses these issues by first creating a publicly available corpus of full-text articles with annotated metabolites and, second, developing new DL algorithms trained to focus specifically on recognising metabolites as well as not constraining these algorithm to abstracts (like prior work), but on full-text paragraphs.

2. Results

Here we describe the characteristics of the automatically annotated metabolomics corpus. This corpus was then used to train three NER algorithms and the manually annotated test set used to evaluate the accuracy of the annotation pipeline, a prior algorithm for chemical NER and the three NER algorithms. Last, we applied the best performing model on all cancer metabolomics articles in the corpus to demonstrate how the algorithms can be used for literature review.

2.1. Metabolomics Corpus

The metabolomics corpus generated by means of the annotation pipeline (see Section 4) is of similar size to that of the BioCreative V5 dataset. We extracted 57,684 sentences that contain metabolites from 1217 metabolomics articles, and in these have identified a total of 123,658 metabolites. For comparison, the BioCreative V5 dataset contains 21,000 abstracts and 99,632 chemical entities. Tables 1 and 2 detail the metabolomics corpus and the number of annotated unique entities for journals and traits, respectively, and illustrates that most metabolomics articles report between 20 and 30 unique metabolites. While metabolomics articles on cancer are the largest single group within the corpus, the average number of metabolites reported in cancer articles is similar to that of other disease areas. The workflow with the structure of the metabolomics corpus is shown for an abstract of a publication in Figure 1.

Table 1. Summary of number (#) of unique metabolites mentioned per journal. Each metabolite is only counted once per article.

Journal	# of Metabolites	# of Articles	Average # of (Unique) Metabolites per Article
Analytical and Bioanalytical Chemistry	716	34	21.1
Analytical Chemistry	863	48	18.0
Journal of Chromatography A	220	9	24.4
Journal of Proteome Research	694	24	28.9
Metabolites	1279	50	25.6
Metabolomics	1267	43	29.5
PLOS One	5400	175	30.9
PNAS USA	429	14	30.6
Scientific Reports	4278	126	34.0

Table 2. Summary of the number (#) of unique metabolites annotated per trait. Each metabolite is only counted once per article.

Trait	# of Metabolites	# of Articles	Average # of (Unique) Metabolites per Article
cancer	12,662	492	25.7
gastrointestinal	876	37	23.7
liver disease	1959	121	16.2
metabolic syndrome	7081	286	24.8
neurodegenerative, psychiatric, and brain illnesses	2546	113	22.5
respiratory diseases	398	37	10.8
sepsis	552	22	25.1
smoking	1360	124	11.0

Figure 1. Data processing and corpus generation. (**A**) The Auto-CORPus output (BioC JSON format) for the abstract of PMC6116856. (**B**) Extracted sentences containing annotations in the abstract (part of the metabolomics corpus, training set). (**C**) Rule-based annotations for the abstract sentences (part of the metabolomics corpus, training set).

2.2. Model Performance

The metabolomics corpus was split into a training set (1034 articles, 49,007 sentences, 105,335 annotations) and a test set (183 articles). The test set was manually annotated and consists of 8950 sentences with 19,138 annotations. We first evaluated the application of the rule-based annotation pipeline on the independent test set articles. Our pipeline yields an F1-score of 0.8893, with precision and recall of 0.8850 and 0.8937 respectively (see Table 3). Next, we evaluated the performance of ChemListem [21], trained on the BioCreative V5 dataset, on the metabolomics corpus test set where it achieves an F1-score of 0.7669. Our first DL model uses the same architecture as ChemListem but was trained on the metabolomics training corpus, therefore this model was named MetaboListem, and it achieved an F1-score of 0.8900 (precision of 0.8923 and recall of 0.8877). Our second DL model uses a different embedding layer (BERT or BioBERT) to MetaboListem, but was trained on the same training dataset. The TABoLiSTM (Transformed-Affixed Bidirectional LSTM) model with BERT embeddings achieved an F1-score of 0.9004, with precision and recall rates of 0.9187 and 0.8829 respectively; the same model with BioBERT embeddings results in a 0.85% improvement of the F1-score.

Table 3. Summary of model performance. Models are evaluated in terms of F1-score, precision and recall rate, and F*-score using the metabolomics corpus test set (183 manually annotated full-text articles). The best results for each metric is highlighted in bold.

Model	Training Data	Embedding	F1-Score	Precision	Recall	F*-Score	Model Size
Annotation pipeline			0.8893	0.8850	**0.8936**	0.8006	
ChemListem [21]	CEMP BioCreative V.5	GloVe	0.7669	0.7301	0.8075	0.6219	26 MB
MetaboListem	metabolomics corpus	GloVe	0.8900	0.8923	0.8877	0.8018	26 MB
TABoLiSTM	metabolomics corpus	BERT	0.9004	0.9187	0.8829	0.8189	827 MB
TABoLiSTM	metabolomics corpus	BioBERT	**0.9089**	**0.9255**	0.8928	**0.8329**	827 MB

The workflow including the structure of the metabolomics corpus, the annotation and result of the TABoLiSTM application on the abstract of a publication is shown in Figure 1. We exemplify the differences in the output from the TABoLiSTM model and our annotation pipeline using three snippets of text in Figure 2 that illustrate the typical examples of how the DL model predictions differ from the rule-based annotations, including the abstract

from Figure 1 shown in Figure 2A. These examples include spelling errors, metabolites not included in the dictionary and metabolite abbreviations.
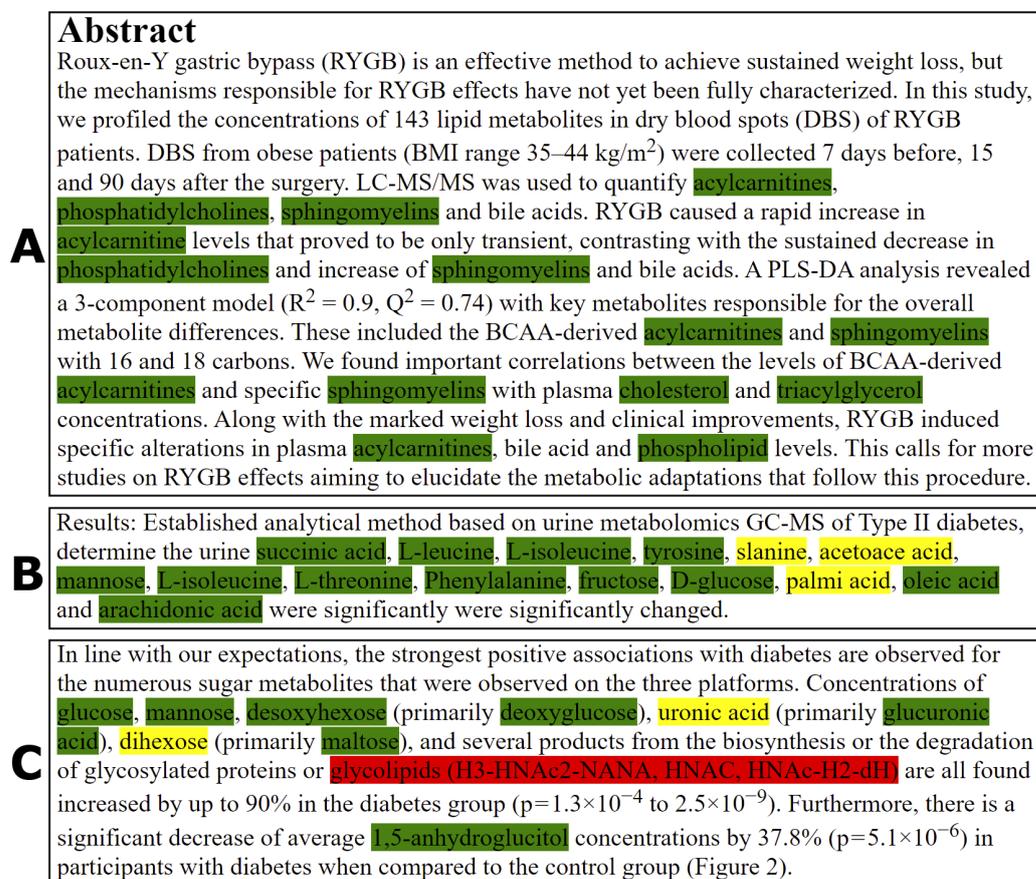
A

Abstract

Roux-en-Y gastric bypass (RYGB) is an effective method to achieve sustained weight loss, but the mechanisms responsible for RYGB effects have not yet been fully characterized. In this study, we profiled the concentrations of 143 lipid metabolites in dry blood spots (DBS) of RYGB patients. DBS from obese patients (BMI range 35–44 kg/m^2) were collected 7 days before, 15 and 90 days after the surgery. LC-MS/MS was used to quantify acylcarnitines, phosphatidylcholines, sphingomyelins and bile acids. RYGB caused a rapid increase in acylcarnitine levels that proved to be only transient, contrasting with the sustained decrease in phosphatidylcholines and increase of sphingomyelins and bile acids. A PLS-DA analysis revealed a 3-component model (R^2 = 0.9, Q^2 = 0.74) with key metabolites responsible for the overall metabolite differences. These included the BCAA-derived acylcarnitines and sphingomyelins with 16 and 18 carbons. We found important correlations between the levels of BCAA-derived acylcarnitines and specific sphingomyelins with plasma cholesterol and triacylglycerol concentrations. Along with the marked weight loss and clinical improvements, RYGB induced specific alterations in plasma acylcarnitines, bile acid and phospholipid levels. This calls for more studies on RYGB effects aiming to elucidate the metabolic adaptations that follow this procedure.

B

Results: Established analytical method based on urine metabolomics GC-MS of Type II diabetes, determine the urine succinic acid, L-leucine, L-isoleucine, tyrosine, slanine, acetoace acid, mannose, L-isoleucine, L-threonine, Phenylalanine, fructose, D-glucose, palmi acid, oleic acid and arachidonic acid were significantly were significantly changed.

C

In line with our expectations, the strongest positive associations with diabetes are observed for the numerous sugar metabolites that were observed on the three platforms. Concentrations of glucose, mannose, desoxyhexose (primarily deoxyglucose), uronic acid (primarily glucuronic acid), dihexose (primarily maltose), and several products from the biosynthesis or the degradation of glycosylated proteins or glycolipids (H3-HNAc2-NANA, HNAC, HNAc-H2-dH) are all found increased by up to 90% in the diabetes group (p=1.3×10^{-4} to 2.5×10^{-9}). Furthermore, there is a significant decrease of average 1,5-anhydroglucitol concentrations by 37.8% (p=5.1×10^{-6}) in participants with diabetes when compared to the control group (Figure 2).

Figure 2. Examples of TABoLiSTM and rule-based annotation results visualised in annotated HTML files. Entities marked in green are recognised by both TABoLiSTM and the rule-based annotation pipeline, yellow entities are only recognised by TABoLiSTM, and red entities only by the rule-based annotation. (**A**) Excerpted from PMC6116856. (**B**) Excerpted from PMC4969426 and exemplifies typos. (**C**) Excerpted from PMC2978704.

Beyond prediction accuracy, another aspect to evaluate is the model sizes. Despite the performance advantages, TABoLiSTM is much larger than both ChemListem and MetaboListem; the sizes of the latter two models are around 26 MB, whereas the size of the TABoLiSTM weight file exceeds 412 MB. Furthermore, pretrained BERT/BioBERT (cased versions) are required to employ the TABoLiSTM model, which adds an extra 415 MB.

The performance of each of the 5 models was evaluated for separate sections (abstract, materials/methods, results, results&discussion, discussion) in the test set (Figure 3). Across all models, the best performance was achieved on separate results and discussion sections, with lowest performance for material/methods and results&discussion sections.

2.3. Automated Literature Review of the Metabolomics Corpus

We applied TABoLiSTM on the abstract, results (including table data) and discussion sections of each of the 8 traits (see Table 2) of the metabolomics corpus to summarise the application of the algorithm in literature review. Figure 4 visualises the results for the 492 cancer articles in the corpus, where only metabolites found in at least 3 different articles are visualised. The size of the text of each metabolites is proportional to the number of articles that reported the metabolite and the position is based on how frequent metabolites are found together. In total, 664 metabolites are recognised in at least three of the 492 articles on cancer (compared to 12,662 unique metabolite names across all 492 articles).

The five most frequent metabolites amongst these are glucose, glutamine, lactate, alanine and glutamate, which are mentioned in 22–29% of articles (see Table 4).
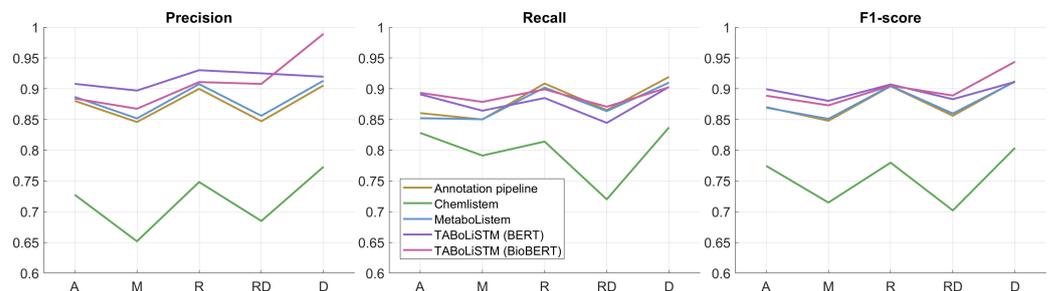

Figure 3. Model performance per publication section. The model performance (precision, recall and F1-score) are given for 5 sections (A, abstract; M, materials/methods; R, results; RD, results and discussion (as single section); D, discussion) for the 5 models.
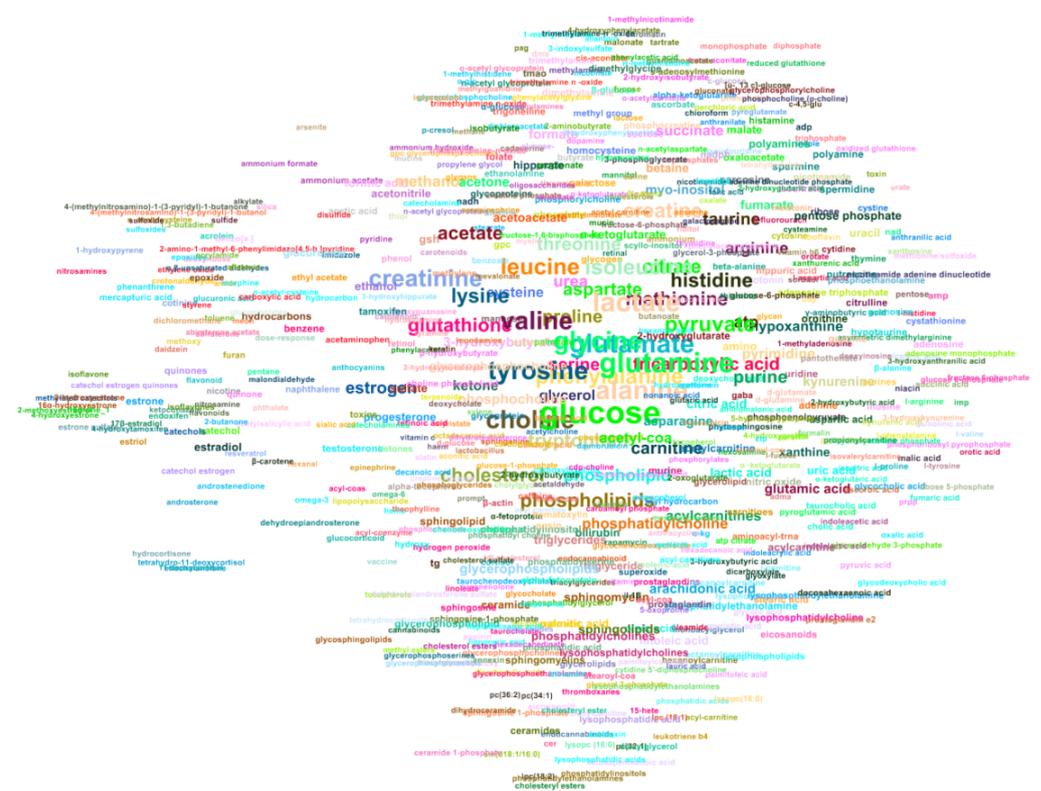

Figure 4. Graphical representation of metabolites mentioned in the abstract, results and discussion sections of 492 cancer-related metabolomics articles. Each node represents a metabolite and is displayed as the (lower-cased) metabolite name if it appeared in 3 or more articles. The closer metabolites are in the graph, they are more likely to co-occur in the same article. The text is sized proportionally to the number of articles that report the metabolite.

In the cancer summary graph, no apparent patterns or clusters are observable based on the node colours due to the three-article threshold applied to the graph. This is in contrast to other networks such as one created for the smoking trait where the restriction of 3 articles is not applied and all 1360 metabolites reported in 124 articles are visualised (see Appendix B). This reflects the low number of common metabolites for this trait in the corpus (Table 4). Glucose is the only metabolite that appears in the top 10 for all traits, with it being reported in 62% of articles on metabolic syndrome.

Table 4. Top ten metabolites reported for each trait in the metabolomics corpus. The number of articles that report the metabolite are given including the percentage of the total. The total number of articles in each category can be found in Table 2.

Cancer		Gastrointestinal		Liver Disease		Metabolic Syndrome	
glucose	142 (29%)	glucose	12 (32%)	glucose	29 (24%)	glucose	178 (62%)
glutamine	121 (25%)	tyrosine	12 (32%)	cholesterol	25 (21%)	cholesterol	108 (38%)
lactate	117 (24%)	lactate	11 (30%)	creatinine	21 (17%)	valine	83 (29%)
alanine	111 (23%)	acetate	10 (27%)	tyrosine	20 (17%)	leucine	73 (26%)
glutamate	109 (22%)	phenylalanine	9 (24%)	glycine	18 (15%)	tyrosine	71 (25%)
tyrosine	105 (21%)	leucine	9 (24%)	phenylalanine	17 (14%)	triglycerides	70 (24%)
glycine	105 (21%)	tryptophan	9 (24%)	lactate	16 (13%)	alanine	67 (23%)
valine	104 (21%)	alanine	8 (22%)	bilirubin	16 (13%)	isoleucine	65 (23%)
choline	103 (21%)	arginine	8 (22%)	pyruvate	15 (12%)	phenylalanine	62 (22%)
creatinine	95 (19%)	citrate	8 (22%)	valine	15 (12%)	glycine	59 (21%)
Neurodegenerative, Psychiatric, and Brain Illnesses		**Respiratory**		**Sepsis**		**Smoking**	
tryptophan	31 (27%)	lactate	8 (22%)	phenylalanine	11 (50%)	creatinine	24 (19%)
glucose	26 (23%)	glycine	6 (16%)	glucose	9 (41%)	cotinine	15 (12%)
glycine	25 (22%)	leucine	6 (16%)	lactate	9 (41%)	nicotine	14 (11%)
glutamate	22 (19%)	pyruvate	6 (16%)	arginine	8 (36%)	glucose	14 (11%)
kynurenine	22 (19%)	glucose	6 (16%)	urea	6 (27%)	leucine	11 (9%)
phenylalanine	21 (19%)	lysine	5 (14%)	nitric oxide	6 (27%)	isoleucine	11 (9%)
serotonin	18 (16%)	creatine	5 (14%)	methionine	6 (27%)	valine	11 (9%)
tyrosine	18 (16%)	isoleucine	5 (14%)	pyruvate	6 (27%)	cholesterol	11 (9%)
creatinine	18 (16%)	glutamine	5 (14%)	creatinine	6 (27%)	lactate	10 (8%)
lactate	17 (15%)	acetate	4 (11%)	ATP	5 (23%)	glutamate	9 (7%)

3. Discussion

We have described three new DL models, using different word embedding layers, that have both achieved state-of-the-art performance (F1-score > 0.89) for metabolite NER on a manually annotated dataset with OA metabolomics articles. Our DL models (F1-scores of 0.91, 0.90 and 0.89 for TABoLiSTM (BioBERT), TABoLiSTM (BERT) and MetaboListem, respectively) surpass by a considerable margin previous methods (using CRFs) for metabolite NER which achieved maximum F1-scores of 0.78 [8] and 0.87 [9], albeit these methods were trained and evaluated on different corpora. The full corpora for both of these methods are not available, only the annotations from the manually annotated test corpus were available for one [8]. The models themselves were not made available so could not be evaluated on our corpus. TABoLiSTM (BioBERT) evaluated on the test annotations (1856 entities) from [8] resulted in a recall of 0.81 which outperforms the recall of 0.74 obtained by the authors. However, the test annotations also include several entities that are not metabolites, such as chemical elements/ions (calcium, Fe^{2+}, H(+), nitrogen, oxygen, potassium, sodium and sulfur) and enzymes (aminoacyl-tRNA synthetases, asparagine synthetase, lactaldehyde dehydrogenase and vacuolar proton ATPase); Excluding these gives a total of 1815 entities that are metabolites (including abbreviations) and applying TABoLiSTM resulted in a recall of 0.84 for these data. Excluding the 317 abbreviations of the test annotations resulted in the TABoLiSTM recall of 0.92 (evaluated on 1498 entities). This exemplifies two aspects, first, while the training data for TABoLiSTM does include some abbreviations (insofar the definition was found as a metabolite using the Auto-CORPus abbreviation file) it is less capable of recognising these in the absence of the definitions. This highlights further improvements can be made with better abbreviation detection algorithms. Second, the manual test annotations from [8] include several entities that are not metabolites. Our metabolomics test corpus contains over 19K metabolites, each of

which are listed in HMDB [12] or found by rule-based methods, and the corpora (training and test) contribute both the sentences and annotations.

Our methods, compared to similar methods for chemical entity recognition [17,21,27] and metabolite NER [8,9], have the benefit of having been trained not only on abstracts and titles, but also on full-text paragraphs that are relevant for information retrieval from studies (results and discussion sections). The corpora these algorithms were trained and evaluated on are made available for future re-use and algorithm development, and to the best of our knowledge is the first of its kind for metabolomics. In addition, we developed a semi-automated annotation pipeline combining regular expression and rule-based approaches with dictionary searching, both for processing the training data for and as an alternative to DL methods, which too achieved outstanding performance (F1-score > 0.88) and also surpassed prior metabolite NER work.

3.1. Annotation Pipeline

Recent work that used NLP methods via dictionary searching for text mining in metabolomics highlighted the considerable user effort (1–4 h as per [10]) required to run these types of analyses. Our annotation pipeline consists of a dictionary of metabolite names and synonyms (from HMDB) which can easily be added to by including data from other databases such as ChemSpider [30] and ChEBI [31]. However, any entity added to the dictionary must undergo some form of quality control to ensure it is a metabolite. Here, we have limited ourselves to only including metabolites listed in HMDB that were classified as either 'quantified' or 'detected' and have not included those that are 'expected' or 'predicted'.

The annotation pipeline also includes regular expression and rule-based methods used to recognise metabolites in text and these add additional advantages: they are time-efficient, easily customisable and explainable. While it requires minutes or even hours for a skilled expert to annotate paragraphs and entire full-text articles, the time to process one full article using these methods is less than 15 s on average. Once experts establish additional rules focusing on recognising naming conventions, then these can easily be added into the pipeline. Moreover, for each identified entity it is immediately obvious which rule(s) resulted in it being predicted as a metabolite entity, and therefore more interpretable than a DL model. For this reason, any false positives or false negatives can be analysed and new sets of rules established to correctly annotate the metabolites in text. This also allows for developing a user-feedback pipeline in which new rules can be shared to obtain better results. We have used this semi-automated annotation pipeline to prepare the training corpus for the DL methods, therefore any improvement in the rules for annotation may also improve DL models. Another application of the pipeline is to run it in parallel with the DL-based methods and combine the results as a post-processing step.

The main limitation of the annotation system is that the system assumes that the dictionary of metabolite names contains all known metabolites with irregular names. We used the HMDB as dictionary because it is the most comprehensive database of human metabolites thus far, however our annotation approach may still result in missing metabolites. One example is endoxifen, which is an active metabolite of the breast cancer drug tamoxifen. Although its synonym 4-hydroxy-*N*-desmethyltamoxifen can be recognised using regular expressions (for example because of the prefix, see Appendix C), the word endoxifen itself is not recognised by the pipeline because of its irregularity and not being included in our dictionary (it exists in HMDB as 'expected'). Augmenting the dictionary with the metabolites listed as 'expected'/'predicted' may appear to be a feature and obvious solution, however adding all of these terms would generate a dictionary 4 times larger than the current one and therefore drastically increase the run-time for the rule-based annotation. Also, >90% of the 'expected' metabolites are lipids [12], and these names are usually recognisable using regular expressions.

Overall, the annotation pipeline is dependent on its data and rules, it can therefore not detect metabolites that have been misspelled, nor will it flag certain terms as false positives if they appear in the dictionary as synonyms for a metabolite. For example, common words

or words with multiple meanings (depending on the context) would be recognised by the system such as 'result' (synonym of 'omeprazole') and 'retinal' (both a vitamin-derivative and a medical term for 'eye'). HMDB also includes chemical entities such as 'ammonium acetate' and 'silica', which are commonly mentioned in methods sections of metabolomics articles. As the HMDB includes more than 18,000 'detected' and 'quantified' metabolites and their synonyms, it was not feasible to examine all entities as pre-processing in this study and only those with 5 or less characters were manually checked; therefore, further cleaning of the dictionary would likely improve the precision of the annotation pipeline. Our results indicate that the combination of regular expressions, dictionary searching and rules can indeed correctly capture a vast amount of metabolite names as also reported by Majumder et al. [10].

3.2. Metabolomics Corpus

To the best of our knowledge, our metabolomics corpus, which consists of the abstract, method, result and discussion sections of 183 manually annotated (test set) and 1034 semi-automatically annotated (training set) PMC articles, is the first available corpus that is designed for human metabolomics research. While some databases include some references to articles where metabolites are mentioned, the databases do not contain enough information (e.g., annotated texts) for NLP model development. An advantage of our metabolomics corpus, compared to common chemical NER corpora [15,16], is that it contains not only abstracts but also sentences from the method, result and discussion sections which is relevant for context-dependent NER. While abstracts indeed provide a rich set of context, the structure of sentences is often different from the main text and therefore training and employing contextual NLP models only on abstracts could potentially result in overlooking valuable information while mining the full text of articles. Methods sections are included since these can include information on targeted metabolites with different assays.

The limitation of the current metabolomics corpus is that it was designed only to label metabolites that appear in full in sentences; that is, if the recognition of a metabolite requires semantic interpretation then it would not be correctly detected. For example, if the string '1- and 3-methylhistidine' is encountered by the pipeline, then only '3-methylhistidine' is annotated, whereas the metabolite '1-methylhistidine' is also in the string. To overcome this limitation of the corpus, it needs to be processed further using semantic analysis techniques such as parse trees [32,33].

We are currently working on addressing a limitation of the Auto-CORPus package [29] that we used to process the full-text. In investigating our results, we found that any abbreviations that cannot be mapped to full names are not annotated, this includes for example cases where Greek letters are abbreviated by letters of Latin alphabet (α as 'a'). We also encountered instances where metabolites are abbreviated without the abbreviation (PC, PE, SM, and others) being defined, because the abbreviations are commonly used in the (lipidomics) community. These abbreviations were also not annotated in the (semi-automatically annotated) metabolomics corpus, but could easily be added as a set of rules to the algorithm. Lastly, superscripts and subscripts in the corpus are encoded differently from normal text by Auto-CORPus [29] and this causes negative impact on the annotation algorithms. We therefore anticipate periodically updating the metabolomics corpus based on user feedback and further research.

3.3. Deep Learning Models

Four LSTM-based models were evaluated on the same manually annotated test set as the annotation pipeline. ChemListem [21], which achieved an F1-score > 0.90 on the BioCreative V5 [16] challenge data, achieved a lower F1-score of 0.77 when evaluated on the metabolomics corpus test set. This is in contrast with our MetaboListem model that used the same neural network architecture as ChemListem, but was trained on the metabolomics corpus, achieving an F1-score of ~0.89. The lower precision and recall rates of ChemListem, when applied to our corpus, are the results of higher false positives and false negatives,

respectively. The false positives are expected due to the fact ChemListem was trained to recognise chemical entities and not metabolites specifically. The higher false negative rate is indicative of limited power for detecting metabolites—despite these also being chemical entities. The recall rates of ChemListem evaluated on the CEMP and metabolomics corpora are 0.89 [21] and 0.81 (here), respectively, which is indicative of the fact that metabolites appear less frequently in the BioCreative V5 dataset. Our MetaboListem model outperforms ChemListem on all three metrics for the metabolomics corpus, thus indicating that the BiLSTM network structure is generalisable to finer NER tasks and that training on relevant data further improves performance. This allows the algorithms to deal with changes in nomenclature and overcome the issue of having to define complex domain knowledge with limited sets of rules which was observed in NLP tasks for recognising genes and proteins previously [34]. Moreover, we have shown how each algorithm achieves different levels of performance on individual sections of a publication, therefore algorithms developed only on abstracts (as these are easily available) may not perform as well on full-text sections. Our algorithms have been trained on a variety of sections (including abstracts) and have shown good performance across all sections.

Our second model, TABoLiSTM, replaced the GloVe embedding [23] that was used in MetaboListem with BERT [24] and BioBERT [25]. The TABoLiSTM model with BioBERT achieved the highest F1-score (0.91) of all models evaluated and improves on all metrics compared to MetaboListem. This improvement is directly attributable to using contextualised token embeddings—enhancing the contextual sensitivity—and using smaller token sequences that are in between character- and word-level embeddings. Smaller token sequences are known to be more suitable for NER tasks where segmenting words is difficult [35], for example with metabolites.

While the annotation pipeline has the highest recall rate, we observed that for many instances with spelling errors in metabolite names that these were recognised by the MetaboListem and TABoLiSTM models but were missed by the rule-based annotation pipeline. This includes for example misspelled entities such as 'slanine' (typo of 'alanine'), 'acetoace acid' (typo of 'acetoacetic acid') and 'palmi acid' (typo of 'palmitic acid') indicating that the DL models are capable to generalise and identify unseen metabolite names, including the misspelled ones, by recognising the token structures and the contexts. In light of earlier discussion on the annotation pipeline not recognising endoxifen, the DL models do correctly recognise endoxifen as metabolite which is likely because the token structure of endoxifen is similar to that of tamoxifen in addition to the similar context embedding of metabolites.

MetaboListem outperforms the precision of the annotation pipeline by 0.7% which may be explained by the ambiguous synonyms falsely identified by the annotation pipeline. For the TABoLiSTM model (BioBERT embedding achieving higher precision by 4% compared to the annotation pipeline) this may be to learning contexts rather than learning the rules and regular structures designated to the annotation pipeline. The algorithms were (trained and) evaluated on the full text output from Auto-CORPus [29], however Auto-CORPus also provides separate JSON output files for table data and abbreviations and these files mostly contain single terms without context. Empirically, we found that although the DL models are context sensitive by construction (BiLSTM network and BioBERT embedding) they detect entities in tables and abbreviation lists with high accuracy comparable to the full text results.

Overall, the semi-automatic annotation pipeline is ideally used in parallel with the DL methods as combining these can to provide wider coverage of features that are not yet learned by the DL algorithms due to limited training examples. Likewise, this is also relevant for addressing some limitations of the DL methods identified from their false positives. We found that while most bacteria are not recognised, a very small number of bacterial species are picked up as 'metabolites' by the DL methods. These terms all include tokens that can also be found in metabolites such as 'lact' (*Lactobacillus*, *Lactococcus*), 'butyr' (*Butyricimonas*, *Pseudobutyrivibrio*, *Butyrivibrio*), 'chloro' (*Chloroflexi*), 'succini' (*Succinivibrio*), 'sulfo' (*Desulfovibrio*), and 'acid' and 'amino' (*Acidaminobacter*). As part of a wider effort

to provide NLP tools to the omics community, we are developing an NER algorithm for microbiome studies which can also be used to filter out false positives from our DL models when complete. Other metabolites are identified (by both the DL and rule-based methods) that are in fact part of an enzyme or pathway name. In our DL models we have not included negative training examples of terms in these contexts, however more elaborate rules can also be developed to filter these false positives out in a post-processing step. For example, we observed in the training set that the words 'process' and 'success' are falsely recognised as metabolites in a certain context. Based on the dictionary list, no known metabolite name/synonym ends with 'ss', so therefore all terms that match the regular pattern 'ss$' could be excluded. This type of rule-based post-processing was not used for our models because of the risk of introducing bias in model evaluation, where an n-gram of 'success' is found in metabolite names such as succinic acid/succinate.

In terms of usability of the DL algorithms, the model size is an important factor. The TABoLiSTM models require 827 MB of storage, whereas ChemListem and MetaboListem only need 26 MB. Therefore, TABoLiSTM may be too heavy to be leveraged on web services and would ideally be used on local systems. Therefore, employing the lighter MetaboListem model in an online environment is more feasible in future, regardless of the cost of 1.8% in the F1-score relative to the TABoLiSTM (BioBERT) model (which outperforms MetaboListem on all evaluation metrics). On average, a full-text article can be annotated by MetaboListem in under 10s, whereas for TABoLiSTM this is close to 40 seconds, hence users may find MetaboListem easier to use. The annotation pipeline and TABoLiSTM (BioBERT) could be used together to increase the confidence of automated annotations (where both methods are in agreement) as they are based on separate methodologies. However, the recall rates are virtually identical and the precision of TABoLISTM is over 4% higher compared to the pipeline, hence it is doubtful this will increase the overall performance of a 'hybrid' algorithm. Moreover, the annotation pipeline requires users using the most up-to-date dictionary at all times to achieve the best results whereas our DL methods have already shown they are capable of generalisation (e.g., correctly recognising metabolites with spelling errors). Hence, the contribution of TABoLiSTM (and MetaboListem) here is to provide an alternative to the reliance upon dictionaries of other NER algorithms [10]. However, as we have shown by using different abbreviation detection algorithms in Auto-CORPus [29] this can be used to increase the number of annotations, therefore an open line of investigation is to evaluate whether the confidence of users in the accuracy of automatic (opposed to manual) annotations is increased if the annotations are identified with multiple (different) methods.

3.4. Biological Interpretation

We have exemplified how the algorithms can be used to speed up literature review by providing a summary of all identified metabolites in the full text. The most common metabolites reported appear to be those that can be measured using virtually all major assays. Glucose is reported most frequently for 4 of the traits, and as one of the top 5 metabolites for the other traits. For cancer studies, the high co-occurence of glucose and lactate is indicative of the Warburg effect [36] where glucose is converted to lactate for cancer cell energy metabolism. However, the remaining top reported metabolites are mostly amino acids, which is also observed for other traits. At the same time, some trait-specific metabolites are found such as bilirubin and cholesterol for liver diseases, cholesterol and triglycerides for metabolic syndrome, cotinine and nicotine for smoking, and tryptophan, kynurenine and serotonin for neurological/psychiatric/brain diseases. The latter being reflective of the serotonin hypothesis [37] that has resulted in antidepressant drugs such as kynurenine uptake inhibitors. The proximities between nodes in the network graphs are known to express community structures [38], hence it is not surprising to find so many amino acids and other common metabolites near the centre of the graphs.

4. Materials and Methods

4.1. Dataset

The data used in the study consists of Open Access metabolomics publications from PMC (n = 1217). The metabolomics corpus is constituted of human-related articles in 18 categories; 10 are by journals (see Table 1) and 8 of which are selected by traits (see Table 2). The complete search terms are provided in Appendix D.

The metabolomics articles were stored in HyperText Markup Language (HTML) format, which were then processed by the Auto-CORPus package [29], and standardised into machine-readable JavaScript Object Notation (JSON) documents.

Auto-CORPus outputs three JSON files based on the input of an HTML file of an article: 'maintext', 'table' and 'abbreviation'. In the maintext file, the textual content in the HTML is split into subsections, and each subsections have five attributes: 'section_heading', 'subsection_heading', 'body', 'IAO_term' and 'IAO_ID', where IAO stands for Information Artifact Ontology [39]. Here we focused on four section types based on standardised IAO terms, namely the textual abstract ('A'),methods ('M'), results ('R') and discussion ('D') sections. We did not use the introduction sections as these do not typically contain information pertaining to the study itself or (how) new results (were acquired).

4.2. Rule-Based Annotation Pipeline

The metabolomics corpus was annotated using a semi-automated pipeline to deal with the large volume of texts and prepare it for supervised learning. The pipeline takes the Auto-CORPus-processed articles as JSON input, and outputs the location (character number) of metabolites in each text (an example can be found in Figure 1 in the Results section). The annotation pipeline is comprised of three steps: pre-processing, identifying metabolites and locating recognised entities.

4.2.1. Pre-Processing

(Sub)sections are split into sentences using sentence tokenisation with the BioBERT [25] model (version biobert-base-cased-v1.2) using the Huggingface Transformer package [40]. Each sentence is uniquely identifiable by two attributes: the PMC identifier (PMCID) of the article that the sentence belongs to, and a sentence identifier. A sentence identifier is a string that contains a letter and five digits (Figure 5) where the letter indicates the relevant sections types (textual abstract (A), methods (M), results (R) and discussion (D)). These section types are determined by the IAO_term key in the Auto-CORPus full-text JSON file.
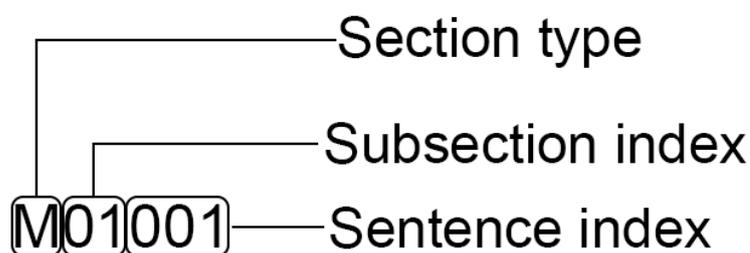

Figure 5. An example of sentence ID. The ID refers to the second sentence in the second subsection of the method section. Indexes start with 0.

4.2.2. Metabolite Identification

The next step is to identify any metabolite mentioned in each tokenised sentence using both dictionary searching and regular expression matching. The dictionary used here was based on the list of metabolites (and their synonyms) that are classified in HMDB [12] as 'quantified' or 'detected'. A list of all metabolites and synonyms was first downloaded on 25 June 2020, and subsequently refreshed on 16 August 2021, after which it was filtered for the 'quantified' or 'detected' metabolites. Short abbreviations (up to 5 characters) that are not exclusive for metabolites were removed (e.g., 'PC' can refer to the metabolite

'phosphatidylcholine', but is also a common abbreviation for 'principal component' and 'pancreatic cancer'), and it was then further cleaned of other non-specific words such as 'result' (a synonym of 'omeprazole').

To complement dictionary searching, the regular expressions are designed to provide partial matches for the metabolite entities (Table A1). It is expected that more metabolites can be captured in this manner compared to complete matching. Entities with a partial match are then expanded to include the entire word and included a full metabolite name or part of a metabolite name; the latter scenario is treated further in post-processing. To craft the regular expressions, the HMDB dictionary was used as a reference. From the dictionary, the entities with one or more spaces and the ones with symbols like hyphen ('-') and colon (':') were selected to form a 'reference set'. The reference set meant to include metabolite names that are less trivial (and thereby more regular). Trivial terms such as singletons are usually more difficult to be captured with regular expressions without drastically increasing the number of false positives by including common words. The regular expression approach is used to complement the HMDB dictionary, therefore trivial terms are omitted from the reference set as they are assumed to be listed in HMDB already. This was used to create a set of regular expressions such that most, if not all, of the reference set can be (at least partially) matched by at least one regular pattern in the set, while avoiding matching unwanted terms.

The regular expressions were created in a recursive manner: while not all terms in the reference set are matched and there remain some observable patterns in the remaining terms, either a new expression was created or an existing expression augmented in order to match the remaining regular terms. Examples of regular expressions and the terms that they are expected to match are shown in Appendix C.

4.2.3. Post-Processing

Post-processing of the matches was utilised to integrate the result of dictionary searching and regular expression matching, including those from partial matching. This step is comprised of three components: combining adjacent words, balancing the number of brackets, and merging overlapped entities.

For each metabolite entity that has been recognised its adjacent words are evaluated. The term before the entity is prepended to the annotation only if certain rules are satisfied: the preceding word ends with a hyphen ('-'), ends with a number, the recognised entity itself starts with a comma and number (e.g., ',3') or the current word starts with a hyphen. If the preceding word is a stop word, or if none of the rules can be met, then the procedure is complete an no term is prepended to the entity. The rules for the next word are the same as described above, except now the entity becomes the 'preceding' word in the above scenario and one more rule is evaluated (whether the next word is one of either 'acid', 'isomer', 'ester' or 'ether'). This process is recursive and is done until no adjacent words satisfy any of the rules.

For all entities the number of brackets is evaluated to ensure all parentheses, square brackets and curly brackets have a starting and ending bracket in the entity. If an opening or closing bracket is missing then the entity is enclosed by the corresponding missing bracket.

All entities that overlap or separated only by a (white)space character are merged into a single entity. The entire post-processing step is executed recursively, i.e., until there is no more change in the list of identified metabolite entities. Once this has finished, two files are created that are used to train the DL models. The first file contains all sentences that have at least one recognised metabolite, along with their sentence identifiers (see Figure 5) and the PMCIDs of the articles the sentences originate from. The second file is in table with 5 fields: PMCID, sentence identifier, position of the start character of the entity, position of the end character of the entity, and the entity itself. These two files are exemplified in Figure 1B,C.

4.3. Training, Validation and Test Set Generation and Evaluation

The 1218 metabolomics publications were split at random into a training, validation and test set a 75:10:15 ratio, corresponding to 913 training publications, 122 publications for

validation and 183 publications for the test set. The publications in the training and validation set were processed using the rule-based annotation pipeline described above, creating a semi-automatically annotated corpus. The TrainingSet.txt and TrainingSetAnnot.tsv (see https://github.com/omicsNLP/MetaboliteNER, accessed on 15 March 2022) files contain both the training and validation partitions as these are split within the training process by the algorithms. The publications in the test set, this included entities in the full text as well as in tables, were manually annotated to avoid bias and to provide a ground truth to evaluate the algorithms on.

Although metabolite abbreviations and entities in tables were not deliberately annotated in the training set with the annotation pipeline, they were manually annotated in the test set for completeness; these types of metabolite names were considered separately while employing models and evaluating performances (detailed in Post-processing). The performance of algorithms was assessed by evaluating the number of true positives (*TP*), false positives (*FP*) and false negatives (*FN*) and calculating the precision ($\frac{TP}{TP+FP}$), recall ($\frac{TP}{TP+FN}$), F1-score ($\frac{2 \times precision \times recall}{precision+recall}$) and F*-score [41] ($\frac{precision \times recall}{precision+recall-precision \times recall}$) on the test set.

4.4. Metabolite NER Using LSTM

A BiLSTM model architecture that was used previously on chemical NER task achieving F1-scores > 0.9 [21] was used here for metabolite NER. We first evaluated the original ChemListem model on the corpus, we then developed two new methods (MetaboListem, TABoLiSTM) based on the prior methodology to improve the performance and described below. Our new models were trained in 50 epochs, and the epoch that achieved the highest F1-score on the validation set was used for evaluation of the test set. All F1-scores reported here are for the test set only.

4.4.1. Pre-Processing

Following prior work [21], a tokeniser designed for chemical text-mining (OSCAR4) [42] was employed to tokenise words in the metabolomics corpus for MetaboListem. For the TABoLiSTM model, we used BioBERT as tokeniser [40] because TABoLiSTM uses the BioBERT token embeddings.

For both word tokenisation systems, the BIOES tag scheme was employed to label token sequences. BIOES tag scheme is a standard sequence labeling technique that classifies tokens into five categories: 'B' labels the tokens at the beginning of entities, 'I' labels internal tokens of entities, 'E' labels the ending tokens of entities, 'S' labels singletons (i.e., tokens that are a complete entity on their own), and 'O' the tokens that do not belong to an entity.

ChemListem used a pre-classifier random forest to predict the probability of a token being (part of) a chemical entity ('B', 'I', 'E' or 'S') based on a set of ChEBI-derived [31] chemicals and chemical elements. For MetaboListem and TABoLiSTM we implemented the same pre-classifier approach using the set of metabolites from HMDB [12] rather than chemicals from ChEBI, where tokens were segmented using OSCAR4 and BioBERT for MetaboListem and TABoLiSTM, respectively. The pre-classifier produces name-internal features that were used in the model training.

4.4.2. ChemListem and MetaboListem

For ChemListem [21], name-internal features for each token were computed in the pre-processing stage using the pre-classifier and treated as one of the two inputs to the model (Figure 6). The features go through a convolutional layer with width of 3 and are concatenated with the numeric representation of words (i.e., a unique number assigned to each word) in the GloVe [23] vector space. Each token now has 256 + 300 = 556 dimensions, where 256 is the output neuron number of the convolutional layer, and 300 the GloVe vector representation. The concatenated tensor is then fed into the BiLSTM layer, before outputting a 5-dimensional result (number of tags used in BIOES tagging) via a TimeDistributed Dense neuron.
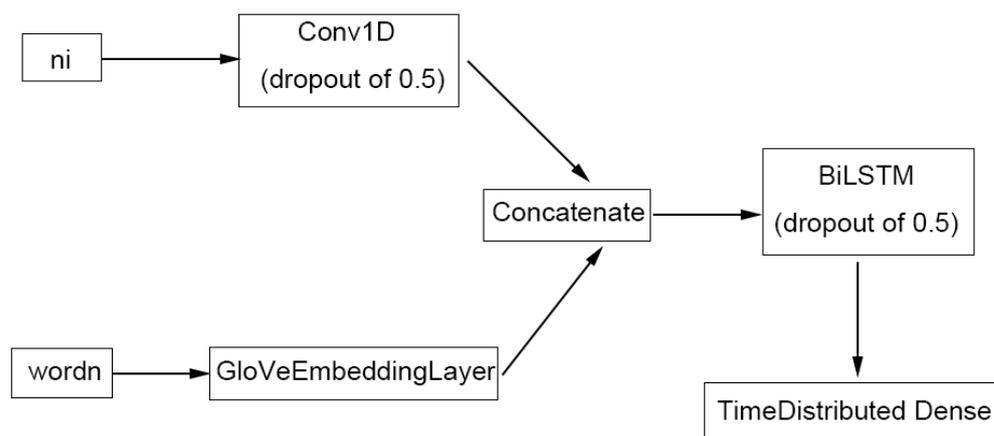
Figure 6. The architecture of the ChemListem network. The input tensors 'wordn' and 'ni' correspond to the numerical representations of word (token) sequences and the name-internal features of word (token) sequence determined by the pre-classifier, respectively.

Our MetaboListem algorithm is based on the same neural network architecture (and GloVe embedding) as ChemListem, hence we named the algorithm in the same manner. It differs from ChemListem only in the training data (metabolomics corpus opposed to BioCreative V5 [16]), pre-classifier data (HMDB instead of ChEBI) and TensorFlow implementation (version 2.0 versus 1.3).

4.4.3. TABoLiSTM (Transformer-Affixed Bidirectional LSTM)

Our TABoLiSTM model differs from MetaboListem mainly in the use of pre-trained BERT [24] (bert-base-cased) and BioBERT [25] (biobert-base-cased-v1.2) embedding. The name-internal features and the numeric presentation of words were all re-computed using (Bio)BERT tokens instead of words and the last hidden layer of the (Bio)BERT model replaces the GloVeEmbeddingLayer in Figure 6. The (Bio)BERT layer provides contextualised word embeddings, and the embeddings were concatenated with the convolution output of name-internal features as previously, except for larger tensors of 256 + 768 = 1214 dimensions. A SpatialDropout1D layer with dropout rate 0.25 is applied on top of the (Bio)BERT layer to prevent overfitting and improve model performance by embedding the dropout [43].

4.4.4. Post-Processing

Post-processing of our DL models involves the same procedure as that of the annotation pipeline (Section 4.2.3) to fix pathological terms such as incomplete entities: to check the neighbouring entities, to fix unbalanced parentheses if possible, and to merge overlapped positions. At this stage further rule-based approaches can be employed to further improve models, for MetaboListem and TABoLiSTM these were not used, but rather represent potential further improvements that can be made to the post-processing.

The DL algorithms were trained on sentences with full metabolite names only, and were not trained on abbreviations or data in tables. In the post-processing step, the algorithms were applied to the Auto-CORPus JSON output of the list of abbreviations (abbreviations and definitions) and any recognised entity (definition) was then linked to its abbreviation and all occurrences in the full text marked as a recognised entity. This was only done when the entire definition was recognised as metabolite entity, i.e., the rules to fix incomplete entities are not applied here. A direct approach to replace all abbreviations in the text with their definitions was investigated but discarded since the current implementation of the abbreviation recognition does not detect all types of abbreviations. The same approach was undertaken on the Auto-CORPus JSON output of all tabular data,

where each cell was evaluated by the DL models as a sentence. Any recognised entity was annotated and evaluated against the manually annotated entities in the test set.

4.5. Metabolite Co-Occurence Network

TABoLiSTM (BioBERT) was applied to all articles in the corpus related to cancer and smoking, and only on the abstract, result (including those in tables) and discussion sections. For each pair of (unique) metabolites mentioned in an article the edge is weighted, likewise the node weight is increased for each (unique) metabolite found. Using these data an adjacency list was created and imported into Gephi [44]. The graph layout was optimised using the ForceAtlas2 [45] algorithm and visualised without edges for the sake of visibility. Nodes placed closer together are more likely to be mentioned in the same articles, and the size of the text is directly proportional to the number of times the metabolite is mentioned across all articles of the same trait. For the cancer graph only the metabolites that appear in three or more articles are used for the graph construction because of the large number of unique metabolites reported (see Table 2). For the smoking graph all metabolites (including those only found by a single study) are displayed. The node names are coloured based on the article that reports the metabolite most (if such article is not unique, the one with lower PMCID number is selected).

5. Conclusions

In summary, this work contributed the first annotated metabolomics corpus of full-text articles to facilitate future NER development for metabolomics research. This corpus is split into two parts: a training set with 49,007 sentences from different sections (abstract, methods, results, discussion) that contains 105,335 annotated metabolite entities, and a (gold-standard) test set with 8950 sentences and 19,138 manually annotated metabolites. This will allow future work on metabolite NER to be trained and/or evaluated on the same data to allow direct comparison of performance between different algorithms. Our four novel NER algorithms, including a rule-based annotation algorithm and three DL NER systems (each with different word embedding layers), are implemented in Python. These algorithms have all shown state-of-the-art performance (F1-score > 0.88) for metabolite NER tasks with the best model (TABoLiSTM with BioBERT word embeddings) achieving an F1-score of 0.91, thereby facilitating future effort in text-mining metabolomics literature. Current on-going work by us focuses on developing new corpora and algorithms for enzyme, pathway and microbiota NER, each of which will allow more effective data extraction and these are intended to further improve the performance of TABoLiSTM by filtering out any false positives or better indicate the context of the metabolite (part of an enzyme or pathway name opposed to a metabolite on its own). Fine-tuning a bidirectional transformer model such as BioBERT for metabolite NER is another potential avenue worth exploring using the metabolomics corpus. All code and data is made available to allow other researchers to use or build upon our work to enable researcher to more effectively perform literature review for metabolomics research.

Author Contributions: Conceptualization, J.M.P. and T.B.; data curation, C.S.Y. and J.M.P.; formal analysis, C.S.Y.; funding acquisition, J.M.P. and T.B.; investigation, C.S.Y.; methodology, C.S.Y. and J.M.P.; project administration, J.M.P. and T.B.; resources, J.M.P.; software, C.S.Y.; supervision, J.M.P. and T.B.; validation, J.M.P.; visualization, C.S.Y.; writing—original draft preparation, C.S.Y.; writing—review and editing, J.M.P. and T.B. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Abbreviations

The following abbreviations are used in this manuscript:

Auto-CORPus	Automated and Consistent Outputs from Research Publications
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
BioBERT	BERT for Biomedical Text Mining
CRF	Conditional Random Fields
DL	Deep Learning
GloVe	Global Vectors for word representation
HMDB	Human Metabolome DataBase
HTML	HyperText Markup Language
IAO	Information Artefact Ontology
JSON	JavaScript Object Notation
NER	Named Entity Recognition
NLP	Natural Language Processing
OA	Open Access
PMC	PubMed Central
PMCID	PMC Identifier
TABoLiSTM	Transformer-Affixed Bidirectional LSTM

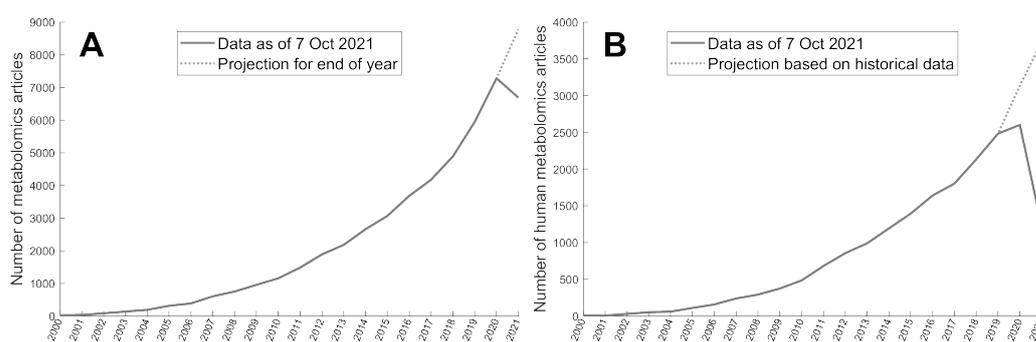## Appendix A. Metabolomics Studies PubMed Search

Figure A1. PubMed search results for metabolomics studies (2000–2021), performed on 7 October 2021. PubMed was searched using the following terms in the title and/or abstract: metabolomics, metabonomics, metabolome-wide association, metabolic phenotyping, metabolic fingerprinting, metabolomic, metabonomic, metabolome, metabonome, metabolic profiling, MWAS. The number of projected metabolomics articles in 2021 was calculated as the number on 7 October 2021 (day 278) multiplied by 365/278 to obtain an estimate at the end of the year. (**A**) All metabolomics studies. (**B**) Metadata mining by PubMed, such as the studies species (e.g., human), is updated as data becomes available which can be up to 2 years. Here, we estimated the proportion of human studies for 2020 and 2021 based on a robust second-order polynomial regression on the data from 2000–2019 from the total. From 2006–2019 this proportion has been relatively stable between 39–46%.

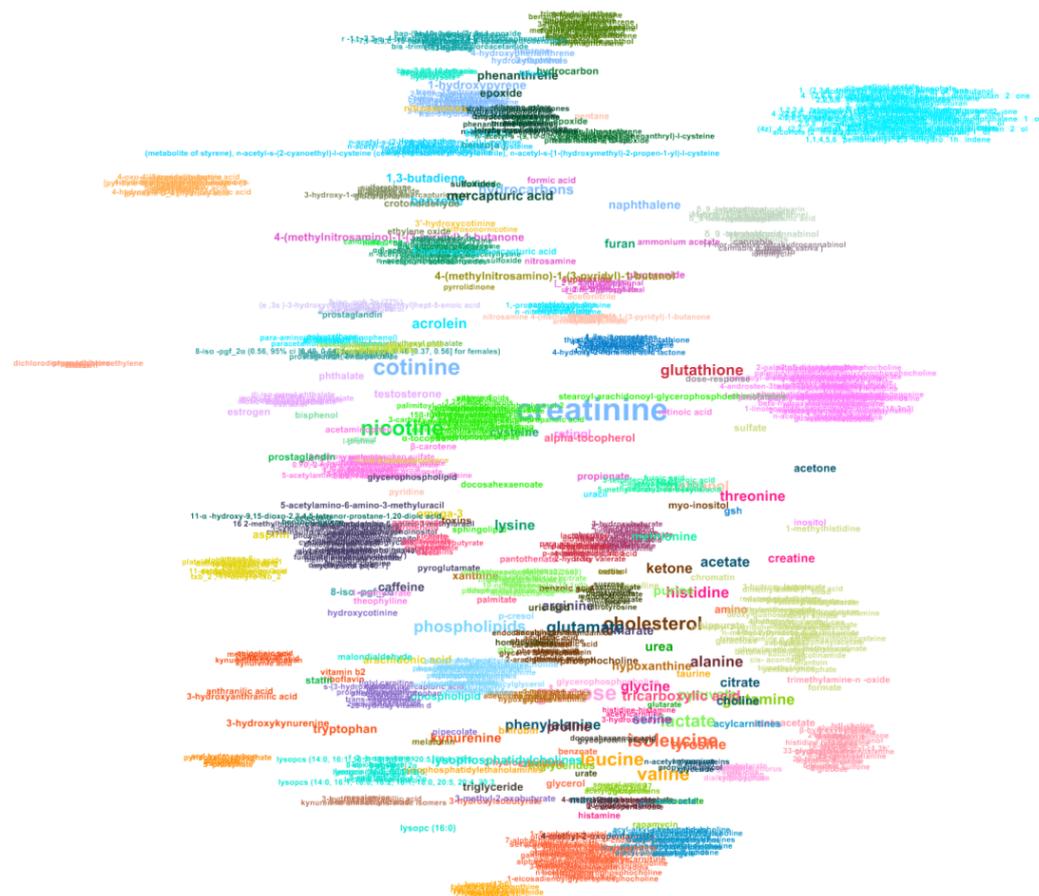## Appendix B. Network Representing the Co-Occurrence of Metabolites in Smoking Articles

Figure A2. Graphical representation of metabolites mentioned in 124 smoking-related metabolomics articles. Each node represents a metabolite and the closer metabolites appear in the graph, the more likely they are to appear together in articles. The displayed metabolite names are sized proportionally to the number of articles that report the metabolite, and are coloured based on the article that reports the metabolite most.

Appendix C. Regular Expression Table

Table A1. Examples of regular expressions used in the rule-based annotation pipeline. Example metabolites that each regular expression matches are shown on the right; the matching parts of the entities are underlined.

Regular Expression	Example Metabolite
(\(?\)?-?\w+(oxy|yl|ol|ine|ene|ose|ane|ate|ester |isomer|ium|ino|ide|one|oxo|en|estra))+ (-\w+)s?\(?\)?(ic)?(?=[-,\.\!\?;])	3-Methoxy-4-hydroxy-phenyl-acetate
([a-z]|-|[0-9])*([0-9]+([a-z]|alpha|beta|sigma| omega|tau|epsilon|zeta|[α-ω])* [\),\+ \- \] \}])+((([a-z]|[0-9])+-)+	5alpha-Androstan-3alpha,17beta-diol disulfate
((([a-z]?[0-9]+(:)[0-9]+)(/[a-z]?[0-9]+(:)[0-9]+)*)	d18:1/16:0

Table A1. *Cont.*

Regular Expression	Example Metabolite
\d+[a-z]?(,\d+[a-z]?)+\)?-	(1S,10R,11S,15S)-5-hydroxy-<u>4</u>-methoxy -<u>15</u>-methyltetracycloheptadeca -2(7),<u>3</u>,5-trien-<u>14</u>-one
([1-9]|\()((alpha|beta|sigma|omega|tau|epsilon| zeta|[α-ω]|Z|E|L|R|S|-|\+),?)+\)-	(<u>11E</u>)-Octadecenoic acid
(oline|dino|etin|olin|ecine|idine|noid|oate|gmine| [a-z]dine|none|igine|zole|pine|nime|anide|acene| prost|sone|idone|hane|cine|cin|tril|erol| [a-wy-z]amine|zime|[a-z]toxin(e?)|otin(e?)|ione| oside|zine|olone|xone|[a-z]dene|lene|lasin(e?)| atin(e?)|ffin(e?)|[a-np-z]yline|xin(e?)|zin(e?)| bene|stine|ythin(e?)|astin(e?)|xitin(e?))(s?)$	Ribofuranosylcre<u>atine</u>

Appendix D. PMC Search Terms for Metabolomics

The following search terms were used for filtering the metabolomics corpus:

(metabolomics[Abstract] OR metabolome[Abstract] OR metabolomic[Abstract] OR metabolome-wide[Abstract] OR metabonomics[Abstract] OR metabonome[Abstract] OR metabonomic[Abstract] OR metabonome-wide[Abstract] OR lipidomics[Abstract] OR lipidome[Abstract] OR lipidomic[Abstract] OR lipidome-wide[Abstract] OR metabolites [Abstract] OR metabolite[Abstract] OR metabolic profiling[Abstract] OR metabolic phenotyping[Abstract] OR metabolomics[Title] OR metabolome[Title] OR metabolomic[Title] OR metabolome-wide[Title] OR metabonomics[Title] OR metabonome[Title] OR metabonomic[Title] OR metabonome-wide[Title] OR lipidomics[Title] OR lipidome[Title] OR lipidomic[Title] OR lipidome-wide[Title] OR metabolites[Title] OR metabolite[Title] OR metabolic profiling[Title] OR metabolic phenotyping[Title]) **AND** (urine[Abstract] OR urinary[Abstract] OR blood[Abstract] OR serum[Abstract] OR plasma[Abstract] OR faecal[Abstract] OR faeces[Abstract] OR cerebrospinal fluid[Abstract] OR CSF[Abstract] OR biofluid[Abstract] OR stool[Abstract] OR feces[Abstract] OR fecal[Abstract] OR stool[Title] OR feces[Title] OR fecal[Title] OR urine[Title] OR urinary[Title] OR blood[Title] OR serum [Title] OR plasma[Title] OR faecal[Title] OR faeces[Title] OR cerebrospinal fluid[Title] OR CSF[Title] OR biofluid[Title]) **AND** (nmr spectroscopy[Abstract] OR nuclear magnetic resonance[Abstract] OR NMR[Abstract] OR mass spectrometry[Abstract] OR LCMS[Abstract] OR GCMS[Abstract] OR UPLCMS[Abstract] OR LC-MS[Abstract] OR GC-MS[Abstract] OR UPLC-MS[Abstract] OR CEMS[Abstract] OR CE-MS[Abstract] OR nmr spectroscopy[Title] OR nuclear magnetic resonance[Title] OR NMR[Title] OR mass spectrometry[Title] OR LCMS[Title] OR GCMS[Title] OR UPLCMS[Title] OR LC-MS[Title] OR GC-MS[Title] OR UPLC-MS[Title] OR CEMS[Title] OR CE-MS[Title]) **AND** (human[Abstract] OR patients [Abstract] OR subjects[Abstract] OR participants[Abstract] OR human[Title] OR patients[Title] OR subjects[Title] OR participants[Title]) **NOT** (mouse[Abstract] OR mice [Abstract] OR rat[Abstract] OR rats[Abstract] OR dog[Abstract] OR dogs[Abstract] OR animal[Abstract] OR cell culture[Abstract] OR dose[Abstract] OR review[Title] OR proteomics[Abstract] OR diet[Abstract] OR proteomic[Abstract] OR proteome[Abstract] OR transcriptomics[Abstract] OR transcriptomic[Abstract] OR transcriptome[Abstract] OR mouse[Title] OR mice[Title] OR rat[Title] OR rats[Title] OR dog[Title] OR dogs[Title] OR animal[Title] OR cell culture[Title] OR dose[Title] OR proteomics[Title] OR diet[Title] OR proteomic[Title] OR proteome[Title] OR transcriptomics[Title] OR transcriptomic[Title] OR transcriptome[Title]) **AND** ("01/01/2000"[Publication Date] : "3000"[Publication Date]).

The publications were extracted on the following dates: cancer (27 November 2020), gastrointestinal (15 April 2020), liver disease (4 June 2021), metabolic syndrome (27 November 2020), neurodegenerative, psychiatric, and brain illnesses (27 November 2020), respiratory (l16 April 2020), sepsis (8 April 2020), smoking (4 June 2021); *Analytica Chimica Acta*

(14 April 2020), *Analytical and Bioanalytical Chemistry* (15 April 2020), *Analytical Chemistry* (22 April 2020), *Journal of Chromatography A* (14 April 2020), *Journal of Proteome Research* (8 April 2020), *Metabolites* (11 September 2020), *Metabolomics* (15 April 2020), *PLOS One* (27 November 2020), *Proceeding of the National Academy of Sciences of the United States of America* (14 April 2020), *Scientific Reports* (15 April 2020).

The search terms above are formatted in segments, and the segments were to filter PMC articles by: types of metabolomics studies, type of sample (biofluid), technology/assay, human study, avoiding non-human and non-metabolomics studies and publication date respectively. After we filtered for metabolomics articles, a list of search terms was used to search by phenotypes/journals. Journals were filtered by adding a term such as '(Metabolites[Journal])', whereas the phenotypes were filtered by adding the following:

Neurodegenerative, psychiatric, and brain illnesses:(Alzheimer[Abstract] OR Dementia[Abstract] OR Creutzfeldt-Jakob[Abstract] OR Stroke[Abstract] OR Multiple Sclerosis[Abstract] OR Guillain-Barre[Abstract] OR epilepsy[Abstract] OR Parkinson[Abstract] OR migraine[Abstract] OR headache[Abstract] OR Schizophrenia[Abstract] OR Huntington[Abstract] OR autism[Abstract] OR depression[Abstract] OR Alzheimer[Title] OR Dementia[Title] OR Creutzfeldt-Jakob[Title] OR Stroke[Title] OR Multiple Sclerosis[Title] OR Guillain-Barre[Title] OR epilepsy[Title] OR Parkinson[Title] OR migraine[Title] OR headache[Title] OR Schizophrenia[Title] OR Huntington[Title] OR autism[Title] OR depression[Title]).

Metabolic syndrome: (obesity[Abstract] OR adiposity[Abstract] OR BMI[Abstract] OR fat mass[Abstract] OR obese[Abstract] OR hypertension[Abstract] OR blood pressure[Abstract] OR diabetes[Abstract] OR CVD[Abstract] OR lipidemia[Abstract] OR waist circumference[Abstract] OR insulin resistance[Abstract] OR fasting glucose[Abstract] OR obesity[Title] OR adiposity[Title] OR BMI[Title] OR fat mass[Title] OR obese[Title] OR hypertension[Title] OR blood pressure[Title] OR diabetes[Title] OR CVD[Title] OR lipidemia[Title] OR waist circumference[Title] OR insulin resistance[Title] OR fasting glucose[Title]).

Sepsis: (Sepsis[Abstract] OR Septic shock[Abstract] OR Septic[Abstract] OR Systemic inflammatory response syndrome[Abstract] OR Multiple organ dysfunction syndrome[Abstract] OR Acute pancreatitis[Abstract] OR toxic shock syndrome[Abstract] OR atypical Kawasaki disease[Abstract] OR Sepsis[Title] OR Septic shock[Title] OR Septic[Title] OR Systemic inflammatory response syndrome[Title] OR Multiple organ dysfunction syndrome[Title] OR Acute pancreatitis[Title] OR toxic shock syndrome[Title] OR atypical Kawasaki disease[Title]).

Cancer: (cancer[Abstract] OR tumor[Abstract] OR tumour[Abstract] OR cancerous[Abstract] OR carcinogen[Abstract] OR carcinogenic[Abstract] OR carcinoma[Abstract] OR leukaemia[Abstract] OR leukemia[Abstract] OR leukaemic[Abstract] OR leukemic[Abstract] OR lymphoma[Abstract] OR malignancy[Abstract] OR premalignancy[Abstract] OR pre-malignancy[Abstract] OR melanoma[Abstract] OR metastasis[Abstract] OR sarcoma[Abstract] OR adjuvant[Abstract] OR neoadjuvant[Abstract] OR chemotherapy[Abstract] OR chemo therapy[Abstract] OR chemo-therapy[Abstract] OR malignant[Abstract] OR premalignant[Abstract] OR pre-malignant[Abstract] OR precancerous[Abstract] OR pre-cancerous[Abstract] OR adenocarcinoma[Abstract] OR metastatic[Abstract] OR cancer[Title] OR tumor[Title] OR tumour[Title] OR cancerous[Title] OR carcinogen[Title] OR carcinogenic[Title] OR carcinoma[Title] OR leukaemia[Title] OR leukemia[Title] OR leukaemic[Title] OR leukemic[Title] OR lymphoma[Title] OR malignancy[Title] OR premalignancy[Title] OR pre-malignancy[Title] OR melanoma[Title] OR metastasis[Title] OR sarcoma[Title] OR adjuvant[Title] OR neoadjuvant[Title] OR chemotherapy[Title] OR chemo therapy[Title] OR chemo-therapy[Title] OR malignant[Title] OR premalignant[Title] OR pre-malignant[Title] OR precancerous[Title] OR pre-cancerous[Title] OR adenocarcinoma[Title] OR metastatic[Title]).

Gastrointestinal: (IBS[Abstract] OR IBD[Abstract] OR Crohn[Abstract] OR ulcerative colitis[Abstract] OR inflammatory bowel disease[Abstract] OR irritable bowel syndrome[Abstract] OR microscopic colitis[Abstract] OR collagenous colitis[Abstract] OR lymphocytic colitis[Abstract] OR bile acid malabsorption[Abstract] OR celiac[Abstract]

OR diversion colitis[Abstract] OR Behçet[Abstract] OR Behcet[Abstract] OR indeterminate colitis[Abstract] OR IBS[Title] OR IBD[Title] OR Crohn[Title] OR ulcerative colitis[Title] OR inflammatory bowel disease[Title] OR irritable bowel syndrome[Title] OR microscopic colitis[Title] OR collagenous colitis[Title] OR lymphocytic colitis[Title] OR bile acid malabsorption[Title] OR celiac[Title] OR diversion colitis[Title] OR Behçet[Title] OR Behcet[Title] OR indeterminate colitis[Title]).

Smoking: (smoking[Abstract] OR cigarette[Abstract] OR tobacco[Abstract] OR vaper [Abstract] OR vaping[Abstract] OR smoker[Abstract] OR smokers[Abstract] OR e-cig[Abstract] OR e-cigarette[Abstract] OR e-smoking[Abstract] OR e-smoker[Abstract] OR nicotine patch[Abstract] OR nicotine dependence[Abstract] OR cigar[Abstract] OR pack-year[Abstract] OR pack-years[Abstract] OR pack year[Abstract] OR packyear[Abstract] OR nicotine use[Abstract] OR cannabis use[Abstract] OR smoking[Abstract] OR cigarette[Abstract] OR tobacco[Abstract] OR vaper[Abstract] OR vaping[Abstract] OR smoker[Title] OR smokers[Title] OR e-cig[Title] OR e-cigarette[Title] OR e-smoking[Title] OR e-smoker[Title] OR nicotine patch[Title] OR nicotine dependence[Title] OR cigar[Title] OR pack-year[Title] OR pack-years[Title] OR pack year[Title] OR packyear[Title] OR nicotine use[Title] OR cannabis use[Title]).

Respiratory: (pneumonia[Abstract] OR respiratory tract infection[Abstract] OR COVID [Abstract] OR SARS[Abstract] OR MERS[Abstract] OR respiratory syndrome[Abstract] OR coronavirus[Abstract] OR pulmonary fibrosis[Abstract] OR COPD[Abstract] OR Chronic Obstructive Pulmonary Disease[Abstract] OR SARS-CoV-2[Abstract] OR Acute Respiratory Syndrome[Abstract] OR Middle East respiratory syndrome[Abstract] OR severe acute respiratory syndrome[Abstract] OR respiratory virus[Abstract] OR pneumonia[Title] OR respiratory tract infection[Title] OR COVID[Title] OR SARS[Title] OR MERS[Title] OR respiratory syndrome[Title] OR coronavirus[Title] OR pulmonary fibrosis[Title] OR COPD[Title] OR Chronic Obstructive Pulmonary Disease[Title] OR SARS-CoV-2[Title] OR Acute Respiratory Syndrome[Title] OR Middle East respiratory syndrome[Title] OR severe acute respiratory syndrome[Title] OR respiratory virus[Title]).

Liver disease: (hepatitis[Abstract] OR haemochromatosis[Abstract] OR hemochromatosis[Abstract] OR cirrhosis[Abstract] OR non-alcoholic fatty liver disease[Abstract] OR NAFLD[Abstract] OR nonalcoholic fatty liver disease[Abstract] OR nonalcoholic steatohepatitis[Abstract] OR non-alcoholic steatohepatitis[Abstract] OR liver fibrosis[Abstract] OR hepatocellular carcinoma[Abstract] OR HCC[Abstract] OR hepatocellular cancer[Abstract] OR acute fatty liver disease[Abstract] OR hepatic encephalopathy [Abstract] OR primary biliary cholangitis[Abstract] OR primary sclerosing cholangitis[Abstract] OR chronic liver disease[Abstract] OR alcoholic liver disease[Abstract] OR ALD[Abstract] OR alcohol-related liver disease[Abstract] OR ARLD[Abstract] OR steatosis[Abstract] OR hepatitis[Title] OR haemochromatosis[Title] OR hemochromatosis[Title] OR cirrhosis[Title] OR non-alcoholic fatty liver disease[Title] OR NAFLD[Title] OR nonalcoholic fatty liver disease[Title] OR nonalcoholic steatohepatitis[Title] OR non-alcoholic steatohepatitis[Title] OR liver fibrosis[Title] OR hepatocellular carcinoma[Title] OR HCC[Title] OR hepatocellular cancer[Title] OR acute fatty liver disease[Title] OR hepatic encephalopathy [Title] OR primary biliary cholangitis[Title] OR primary sclerosing cholangitis[Title] OR chronic liver disease[Title] OR alcoholic liver disease[Title] OR ALD[Title] OR alcohol-related liver disease[Title] OR ARLD[Title] OR steatosis[Title]).

References

1. Wishart, D.S. Metabolomics for Investigating Physiological and Pathophysiological Processes. *Physiol. Rev.* **2019**, *99*, 1819–1875. [CrossRef] [PubMed]
2. Miggiels, P.; Wouters, B.; van Westen, G.J.; Dubbelman, A.C.; Hankemeier, T. Novel technologies for metabolomics: More for less. *TrAC Trends Anal. Chem.* **2019**, *120*, 115323. [CrossRef]
3. Bornmann, L.; Mutz, R. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references: Growth Rates of Modern Science: A Bibliometric Analysis Based on the Number of Publications and Cited References. *J. Assoc. Inf. Sci. Technol.* **2015**, *66*, 2215–2222. [CrossRef]

4. Jackson, R.G.; Patel, R.; Jayatilleke, N.; Kolliakou, A.; Ball, M.; Gorrell, G.; Roberts, A.; Dobson, R.J.; Stewart, R. Natural language processing to extract symptoms of severe mental illness from clinical text: The Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. *BMJ Open* **2017**, *7*, e012012. [CrossRef]

5. Sheikhalishahi, S.; Miotto, R.; Dudley, J.T.; Lavelli, A.; Rinaldi, F.; Osmani, V. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Med. Inform.* **2019**, *7*, e12239. [CrossRef]

6. Milosevic, N.; Gregson, C.; Hernandez, R.; Nenadic, G. A framework for information extraction from tables in biomedical literature. *Int. J. Doc. Anal. Recognit. (IJDAR)* **2019**, *22*, 55–78. [CrossRef]

7. Kuleshov, V.; Ding, J.; Vo, C.; Hancock, B.; Ratner, A.; Li, Y.; Ré, C.; Batzoglou, S.; Snyder, M. A machine-compiled database of genome-wide association studies. *Nat. Commun.* **2019**, *10*, 3341. [CrossRef]

8. Nobata, C.; Dobson, P.D.; Iqbal, S.A.; Mendes, P.; Tsujii, J.; Kell, D.B.; Ananiadou, S. Mining metabolites: Extracting the yeast metabolome from the literature. *Metabolomics* **2010**, *7*, 94–101. [CrossRef]

9. Kongburan, W.; Padungweang, P.; Krathu, W.; Chan, J.H. Metabolite Named Entity Recognition: A Hybrid Approach. In *International Conference on Neural Information Processing: Neural Information Processing*; Hirose, A., Ozawa, S., Doya, K., Ikeda, K., Lee, M., Liu, D., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 451–460.

10. Majumder, E.L.W.; Billings, E.M.; Benton, H.P.; Martin, R.L.; Palermo, A.; Guijas, C.; Rinschen, M.M.; Domingo-Almenara, X.; Montenegro-Burke, J.R.; Tagtow, B.A.; et al. Cognitive analysis of metabolomics data for systems biology. *Nat. Protoc.* **2021**, *16*, 1376–1418. [CrossRef]

11. Kodra, D.; Pousinis, P.; Vorkas, P.A.; Kademoglou, K.; Liapikos, T.; Pechlivanis, A.; Virgiliou, C.; Wilson, I.D.; Gika, H.; Theodoridis, G. Is Current Practice Adhering to Guidelines Proposed for Metabolite Identification in LC-MS Untargeted Metabolomics? A Meta-Analysis of the Literature. *J. Proteome Res.* **2022**, *21*, 590–598. [CrossRef]

12. Wishart, D.S.; Feunang, Y.D.; Marcu, A.; Guo, A.C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; et al. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res.* **2018**, *46*, D608–D617. [CrossRef] [PubMed]

13. Wishart, D.S.; Knox, C.; Guo, A.C.; Eisner, R.; Young, N.; Gautam, B.; Hau, D.D.; Psychogios, N.; Dong, E.; Bouatra, S.; et al. HMDB: A knowledgebase for the human metabolome. *Nucleic Acids Res.* **2009**, *37*, D603–D610. [CrossRef]

14. Hirschman, L.; Yeh, A.; Blaschke, C.; Valencia, A. Overview of BioCreAtIvE: Critical assessment of information extraction for biology. *BMC Bioinform.* **2005**, *6*, S1. [CrossRef] [PubMed]

15. Krallinger, M.; Leitner, F.; Rabal, O.; Vazquez, M.; Oyarzabal, J.; Valencia, A. CHEMDNER: The drugs and chemical names extraction challenge. *J. Cheminform.* **2015**, *7*, S1. [CrossRef] [PubMed]

16. Wang, Q.; S. Abdul, S.; Almeida, L.; Ananiadou, S.; Balderas-Martínez, Y.I.; Batista-Navarro, R.; Campos, D.; Chilton, L.; Chou, H.J.; Contreras, G.; et al. Overview of the interactive task in BioCreative V. *Database* **2016**, *2016*, baw119. [CrossRef]

17. Leaman, R.; Wei, C.-H.; Lu, Z. tmChem: A high performance approach for chemical named entity recognition and normalization. *J. Cheminform.* **2015**, *7*, S3. [CrossRef] [PubMed]

18. Yoon, W.; So, C.H.; Lee, J.; Kang, J. CollaboNet: Collaboration of deep neural networks for biomedical named entity recognition. *BMC Bioinform.* **2019**, *20*, 85. [CrossRef] [PubMed]

19. Hemati, W.; Mehler, A. LSTMVoter: Chemical named entity recognition using a conglomerate of sequence labeling tools. *J. Cheminform.* **2019**, *11*, 3. [CrossRef]

20. Luo, L.; Yang, Z.; Yang, P.; Zhang, Y.; Wang, L.; Lin, H.; Wang, J. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics* **2017**, *34*, 1381–1388. [CrossRef]

21. Corbett, P.; Boyle, J. Chemlistem: Chemical named entity recognition using recurrent neural networks. *J. Cheminform.* **2018**, *10*, 59. [CrossRef]

22. Habibi, M.; Weber, L.; Neves, M.; Wiegandt, D.L.; Leser, U. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* **2017**, *33*, i37–i48. [CrossRef]

23. Pennington, J.; Socher, R.; Manning, C. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543. [CrossRef]

24. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North, Minneapolis, MN, USA, 19 June 2019; pp. 4171–4186. [CrossRef]

25. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2019**, *36*, 1234–1240. [CrossRef] [PubMed]

26. Beck, J. Report from the Field: PubMed Central, an XML-based Archive of Life Sciences Journal Articles. In Proceedings of the International Symposium on XML for the Long Haul: Issues in the Long-term Preservation of XML, Montréal, QC, Canada, 2 August 2010.

27. Sun, C.; Yang, Z.; Wang, L.; Zhang, Y.; Lin, H.; Wang, J. Biomedical named entity recognition using BERT in the machine reading comprehension framework. *J. Biomed. Inform.* **2021**, *118*, 103799. [CrossRef] [PubMed]

28. Islamaj, R.; Leaman, R.; Kim, S.; Kwon, D.; Wei, C.-H.; Comeau, D.C.; Peng, Y.; Cissel, D.; Coss, C.; Fisher, C.; et al. NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. *Sci. Data* **2021**, *8*, 91. [CrossRef] [PubMed]

29. Beck, T.; Shorter, T.; Hu, Y.; Li, Z.; Sun, S.; Popovici, C.M.; McQuibban, N.A.R.; Makraduli, F.; Yeung, C.S.; Rowlands, T.; et al. Auto-CORPus: A Natural Language Processing Tool for Standardising and Reusing Biomedical Literature. *Front. Digit. Health* **2022**, *4*, 788124. [CrossRef]

30. Williams, A.J. Chemspider: A Platform for Crowdsourced Collaboration to Curate Data Derived From Public Compound Databases. In *Collaborative Computational Technologies for Biomedical Research*; Ekins, S., Hupcey, M.A.Z., Williams, A.J., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2011; pp. 363–386. [CrossRef]

31. de Matos, P.; Alcántara, R.; Dekker, A.; Ennis, M.; Hastings, J.; Haug, K.; Spiteri, I.; Turner, S.; Steinbeck, C. Chemical Entities of Biological Interest: An update. *Nucleic Acids Res.* **2010**, *38*, D249–D254. [CrossRef]

32. Fundel, K.; Küffner, R.; Zimmer, R. RelEx—Relation extraction using dependency parse trees. *Bioinformatics* **2006**, *23*, 365–371. [CrossRef]

33. Zhou, G.; Zhang, M.; Ji, D.; Zhu, Q. Tree Kernel-Based Relation Extraction with Context-Sensitive Structured Parse Tree Information. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech Republic, 7 June 2007; pp. 728–736.

34. Erhardt, R.A.A.; Schneider, R.; Blaschke, C. Status of text-mining techniques applied to biomedical text. *Drug Discov. Today* **2006**, *11*, 315–325. [CrossRef]

35. Gridach, M. Character-level neural network for biomedical named entity recognition. *J. Biomed. Inform.* **2017**, *70*, 85–91. [CrossRef]

36. Vander Heiden, M.G.; Cantley, L.C.; Thompson, C.B. Understanding the Warburg Effect: The Metabolic Requirements of Cell Proliferation. *Science* **2009**, *324*, 1029–1033. [CrossRef]

37. Lapin, I.; Oxenkrug, G. Intensification Of The Central Serotoninergic Processes As A Possible Determinant Of The Thymoleptic Effect. *Lancet* **1969**, *293*, 132–136. [CrossRef]

38. Noack, A. Modularity clustering is force-directed layout. *Phys. Rev. E* **2009**, *79*, 026102. [CrossRef] [PubMed]

39. Ceusters, W. An information artifact ontology perspective on data collections and associated representational artifacts. *Stud. Health Technol. Inform.* **2012**, *180*, 68–72.

40. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online Conference, 16–20 November 2020; pp. 38–45. Available online: https://aclanthology.org/2020.emnlp-demos.6/ (accessed on 24 February 2022). [CrossRef]

41. Hand, D.J.; Christen, P.; Kirielle, N. F*: An interpretable transformation of the F-measure. *Mach. Learn.* **2021**, *110*, 451–456. [CrossRef]

42. Jessop, D.M.; Adams, S.E.; Willighagen, E.L.; Hawizy, L.; Murray-Rust, P. OSCAR4: A flexible architecture for chemical text-mining. *J. Cheminform.* **2011**, *3*, 41. [CrossRef]

43. Gal, Y.; Ghahramani, Z. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, Barcelona, Spain, 5–10 December 2016*; Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R., Eds.; Curran Associates Inc.: Red Hook, NY, USA, 2016 ; pp. 1019–1027.

44. Bastian, M.; Heymann, S.; Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. In Proceedings of the Third International AAAI Conference on Weblogs and Social Media, San Jose, CA, USA, 17–20 May 2009. [CrossRef]

45. Jacomy, M.; Venturini, T.; Heymann, S.; Bastian, M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE* **2014**, *9*, e98679. [CrossRef] [PubMed]