Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 October 2022 doi:10.20944/preprints202210.0247.v1

Article

The New Version of the Anddigest Tool with Improved AI-Based Short Names Recognition

Timofey V Ivanisenko[1,2*], Pavel S Demenkov[1,2], Nikolay A Kolchanov[1,2,3] and Vladimir A Ivanisenko[1,2,3]

[1] Kurchatov Genomics Center, Institute of Cytology & Genetics, Siberian Branch, Russian Academy of Sciences, Prospekt Lavrentyeva 10, Novosibirsk, 630090, Russia.
[2] Institute of Cytology & Genetics, Siberian Branch, Russian Academy of Sciences, Prospekt Lavrentyeva 10, Novosibirsk, 630090, Russia.
[3] Novosibirsk State University, st. Pirogova 1, Novosibirsk, 630090, Russia.
* Correspondence: itv@bionet.nsc.ru

Abstract

The body of scientific literature continues to grow annually. Over 1.5 million abstracts of biomedical publications were added to the PubMed database in 2021. Therefore, developing cognitive systems that provide a specialized search for information in scientific publications based on subject area ontology and modern artificial intelligence methods is urgently needed. We previously developed a web-based information retrieval system, ANDDigest, designed to search and analyze information in the PubMed database using a customized domain ontology. This paper presents an improved ANDDigest version that uses fine-tuned PubMedBERT classifiers to enhance the quality of short name recognition for molecular-genetics entities in PubMed abstracts on eight biological object types: cell components, diseases, side effects, genes, proteins, pathways, drugs, and metabolites. This approach increased average short name recognition accuracy by 13%. The new ANDDigest version (01.2022) has a web interface and is freely available to users at https://anddigest.sysbio.ru/.

Keywords: Text-mining; ANDDigest; ANDSystem; Named entity recognition; Machine learning; PubMedBERT

1. Introduction

Finding relevant information in scientific publications and patents is a significant issue when performing almost any scientific research. The number of scientific publications only in the biological sciences field has reached colossal proportions. For example, >34 million articles are stored in the PubMed database, and >1 million new biomedical articles appear annually. Modern scientific information search engines, such as those used by Google Scholar, Scopus, and PubMed [1, 2, 3], make it possible to find literature based on queries compiled through user-specified keywords. However, such systems do not provide practical tools for automatically extracting information from their search results, which can sometimes reach tens to hundreds of thousands of documents. In addition, they do not sufficiently consider the synonymy of the desired objects and their relationship with external databases.

Another strategy is to use programs based on automatic text analysis methods. Such systems automatically extract knowledge from documents and present it in graphical forms, such as semantic networks. Of particular interest are systems providing the full knowledge engineering cycle. This cycle includes automatic knowledge extraction from unstructured texts in natural language and external databases. It also includes integrating the obtained materials into the knowledge base as semantic networks, where nodes are the objects recognized in the texts, and the edges are the various established interaction types between them. In addition, such systems usually provide tools for the visualization and analysis of obtained results.

STRING [4], Pathway Studio [5], MetaCore [6, 7], and ANDSystem [8, 9, 10, 11] are well-known examples of these systems. Unlike simple search engines, these programs are based on predefined and well-validated ontologies that describe the subject area. This approach automatically considers object synonymy and relationships with external databases but limits the programs' search capacities by the size of the used ontologies. However, since their primary purpose is to establish interactions between entities and reconstruct associative networks based on the retrieved information, such tools do not attain high completeness values in finding documents or, in some cases, do not supply the user with such information.

We previously developed the ANDDigest information retrieval tool [12]. It was designed to find biomedical abstracts using complex search queries to PubMed, combining the ANDSystem ontology based on dictionaries with user-provided keywords. The ANDDigest system automatically extracts knowledge from scientific publication texts and includes tools for automated literature search and analysis.

One essential automatic text analysis step is named entity recognition (NER). A well-known problem of automatic biological entity name recognition in scientific publication texts by the dictionary relates to linguistic ambiguities associated with the intersection of object names with commonly used words and phrases, including abbreviations and various terms introduced directly by the authors [13]. This problem is especially relevant in biology and biomedicine for short gene and protein names [14, 15, 16] but is also typical for other types [17, 18]. Existing methods for overcoming the NER problem are based on three approaches: dictionaries, semantic-linguistic rules, and machine learning algorithms.

To date, the most widespread are machine learning algorithm-based methods. For example, the POSBioTM-NER system [19] uses named entities recognized based on support vector machines and a conditional random field (CRF). Chang et al. [20] used the vectorization methods of biomedical terms (word-embedding) when training their CRF model. It allowed them to significantly increase the recognition accuracy of named entities in biomedicine compared to other approaches. Wei et al. proposed a combined machine learning method involving CRF and a bidirectional long-short-term memory network [21]. This implementation showed better results than classical methods based on rules and templates and CRF-based systems. A similar approach was used in the HUNER system [22].

A turning point in natural language processing (NLP) was the invention of the transformer neural network architecture [23]. One significant difference between transformers and previous architectures was the lack of dependence on input sequence order during training, providing ample opportunities for data parallelization and facilitating transformer models trained on vast textual data arrays reaching hundreds of gigabytes. Therefore, this architecture enabled the development of many pretrained language models with multi-million and even multi-billion machine learning parameters, such as Megatron [24], GPT [25], and BERT [26]. Another essential feature of such models that distinguishes them from earlier machine learning algorithms such as word2vec [27] and GloVe [28] was their ability to generate context-sensitive word embeddings. The central concept is that the numerical representation for the same word (its vector) is not static but depends on its context. This point is significant since the meaning of a word can often change as a sentence or the whole text develops.

Another equally significant feature is the ability to perform additional transformer model training. This fine-tuning involves training one additional output layer while keeping the main model weights, reflecting the relationships between words, the same. This approach enables pre-trained models to be quickly adapted to solve diverse NLP tasks, including NER, relation extraction, and context-based object classification.

However, the main disadvantage of most pre-trained models is that they are trained in the general language domain, leading to insufficient accuracy when they are used to analyze texts from narrowly focused fields, such as biology and biomedicine. This problem mainly reflects additional linguistic ambiguities associated with the specifics of the biological and biomedical scientific language, which contains many highly specialized terms and abbreviations [13, 29].

Another BERT model appeared in 2020, entitled BioBERT [30]. It used the classic BERT model [26] trained on the data from BookCorpus [31] and Wikipedia [32], with further pretraining on open-access biomedical texts from PubMed and PubMed central. BioBERT's authors showed that it was more accurate for NLP tasks in biology and biomedicine than models trained on the larger textual corpora belonging to the general language domain.

Davagdorj et al. developed a BioBERT-based K-means model [33] that provided better biomedical document clustering accuracy than other models. The CPRiL web service [34] uses the BioBERT machine-learning model to determine the functional relationships between small molecules and proteins in biomedical literature. This product's harmonic mean of the precision and recall (F_1) score was 84.3%, reflecting 82.9% accuracy and 85.7% recall. The STRING system's authors used a fine-tuned BioBERT model to classify gene and protein names identified by their text-mining method in texts as correctly or wrongly recognized [4].

However, the BioBERT model's main disadvantage was that its training used the original BERT model's weights as the starting point. Therefore, the word embedding vocabulary was the same as the BERT model, which is specific to the general language domain and not very representative of the biomedical field.

This problem led to the development of the PubMedBERT model [35], trained from scratch using only PubMed data, with its biomedicine-specific thesaurus containing about 30% more specific terms than BioBERT. A comparative study showed that it performed the best in the biomedical domain [36].

When using machine learning NER methods, one important task is establishing links between objects identified in texts and external databases containing additional information about them. A simple name comparison is often ineffective due to synonymy. The possible solution to this problem is combining modern machine learning approaches with classical text analysis methods [4, 37], such as predefined ontologies.

In this study, we developed a new version of the ANDDigest information retrieval system (ver. 01.2022) with improved short molecular-genetic object name (≤4 characters) recognition accuracy in PubMed texts. Further trained PubMedBERT models were used to filter incorrectly recognized names mapped using ANDSystem dictionaries. We used the developed models to classify object names as correctly and incorrectly recognized based on their context in abstract texts. The classification models filter eight object types: cell components, diseases, side effects, genes, proteins, pathways, drugs, and metabolites. The developed filtering methods improved recognition accuracy for these entities by 13% on average.

2. Materials and Methods

2.1. Gold Standard

The classification accuracy of short object names with the developed models was assessed using a gold standard (see Supplementary 1) that contained molecular genetic entity names from the ANDSystem ontology marked up in scientific article abstracts and manually annotated. In total, the gold standard contained >60 thousand sentences from >35 thousand PubMed abstracts in which at least one object name (≤4 characters) was present. The following object types were considered: genes/proteins, metabolites/drugs, diseases/side effects, pathways, and cellular components. The gold standard was manually created in collaboration with experts while developing the new ANDDigest version (01.2022).

2.2. PubMed Abstract Corpus

The analysis used a corpus of >34 million English PubMed abstract texts retrieved in July 2022.

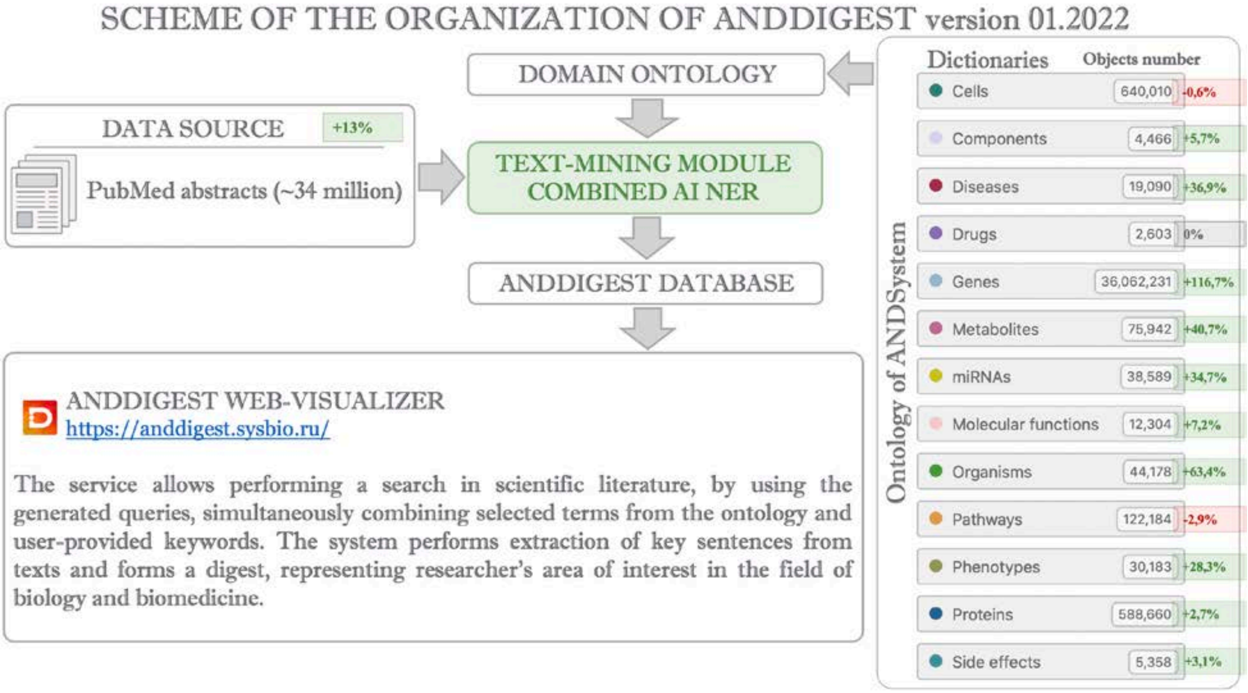
2.3. PubMedBERT

Classification models were constructed by fine-tuning the pre-trained PubMedBERT transformer model [35] using Python's transformers v4.16.2 library.

3. Results

3.1. Web-based information retrieval system ANDDigest (ver. 01.2022)

The previously developed ANDDigest software and information system [13] was designed to search and analyze information in scientific publications using a customized domain ontology. Its new version also uses an ANDSystem cognitive system ontology specific to biology and biomedicine. The general ANDDigest ver. 01.2022 module scheme is shown in Figure 1.



The service allows performing a search in scientific literature, by using the generated queries, simultaneously combining selected terms from the ontology and user-provided keywords. The system performs extraction of key sentences from texts and forms a digest, representing researcher's area of interest in the field of biology and biomedicine.

Dictionaries	Objects number	
Cells	640,010	-0,6%
Components	4,466	+5,7%
Diseases	19,090	+36,9%
Drugs	2,603	0%
Genes	36,062,231	+116,7%
Metabolites	75,942	+40,7%
miRNAs	38,589	+34,7%
Molecular functions	12,304	+7,2%
Organisms	44,178	+63,4%
Pathways	122,184	-2,9%
Phenotypes	30,183	+28,3%
Proteins	588,660	+2,7%
Side effects	5,358	+3,1%

Figure 1. A schematic illustration of ANDDigest ver. 01.2022. Green and red colors highlight new modules and data changes compared to the previous version.

The ANDDigest ver. 01.2022 uses the ANDSystem's domain ontology based on dictionaries for 13 molecular-genetic object types (cells, components, diseases, drugs, genes, metabolites, micro RNAs [miRNAs], molecular functions, organisms, pathways, phenotypes, proteins, and drug side effects). Each dictionary contains the main molecular genetic object names and synonym sets.

A search query to ANDDigest can be performed by selecting specific biological object names from the corresponding dictionaries or only their types. In addition, the user can enter additional clarifying keywords. Search queries automatically consider all synonyms of the entered object. The search is performed using all objects from the corresponding dictionary when the user specifies the object type but not its name.

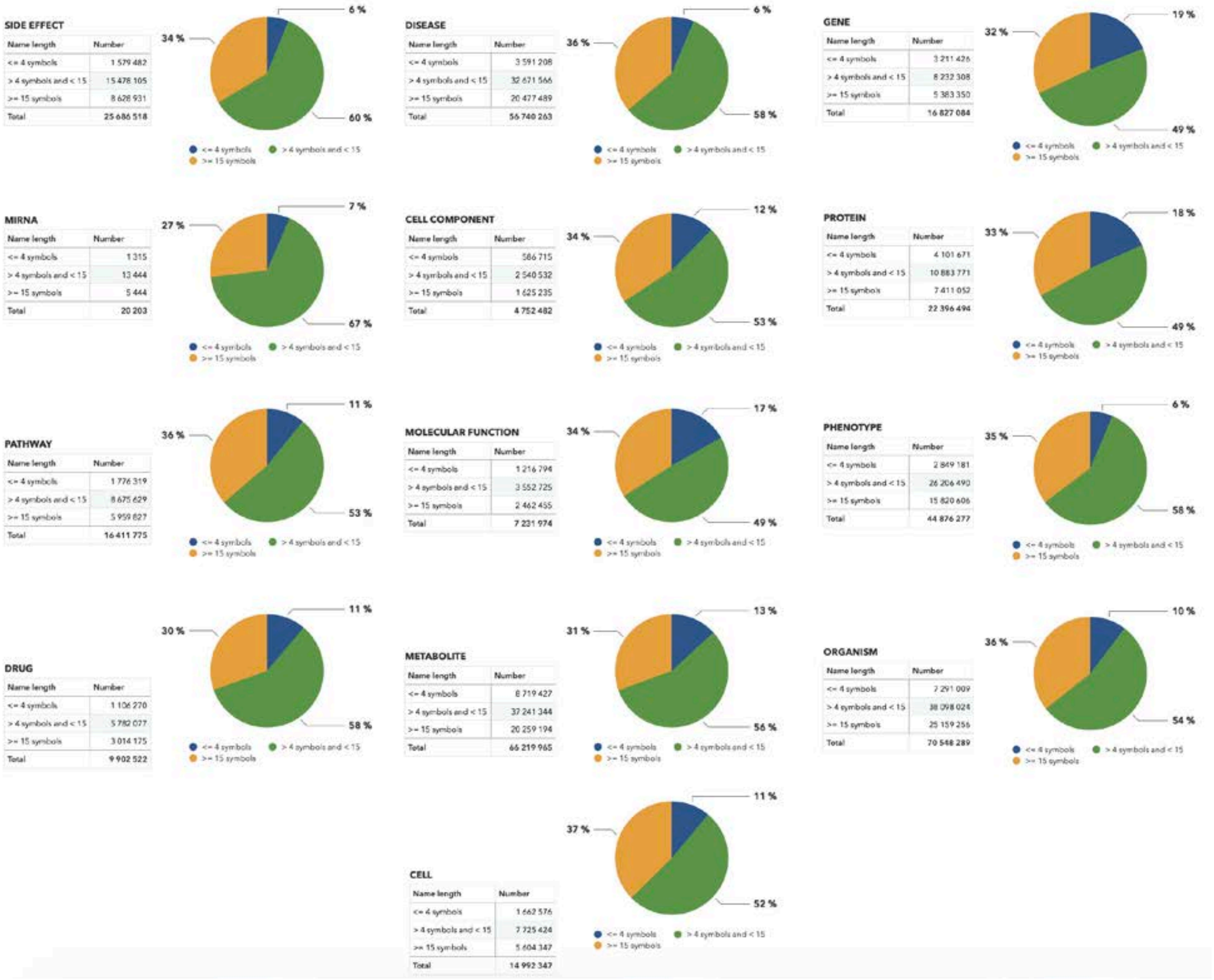
Search results are presented as a set of mapped texts containing the specified entities from the domain ontology and a graph of semantic relationships between their objects. In addition, the system provides flexible filtering and sorting functions for the identified documents, including filtering by the statistical significance level of the semantic relationships between pairs of objects, the impact factor of a scientific journal, and the publication date.

ANDDigest can calculate trends, indicating the scientific community's interest in the specific object from ANDSystem's ontology based on its number of mentions in PubMed, using the non-parametric Mann–Kendall test [38, 39]. Such dynamics are calculated in two ways: (a) standard, the total number of documents mentioning the mapped object per year; (b) normalized, the ratio of the number of documents mentioning the object to the total number of published documents per year.

A new combined NER module is implemented in ANDDigest ver. 01.2022: combined artificial intelligence (AI) (Fig. 1). This module integrates dictionary-based NER and filtering of incorrectly recognized short object names using context-based classification, which is performed by fine-tuned PubMedBERT transformer neural networks.

3.2. Dictionary-based NER

Preliminary dictionary-based mapping of the molecular genetic object names in texts is performed using the text-mining algorithms implemented in ANDSystem [11]. Then, all the recognized entities matching the corresponding dictionary are divided into three groups: terms with a length of ≤4 characters (short names), terms with a length of >4 but <15 characters, and terms with a length of ≥15 characters (long names). The distributions of object names by group, length, and type are shown in Figure 2.



SIDE EFFECT

Name length	Number
<= 4 symbols	1 579 482
> 4 symbols and < 15	15 478 105
>= 15 symbols	8 628 931
Total	25 686 518

DISEASE

Name length	Number
<= 4 symbols	3 591 208
> 4 symbols and < 15	32 671 566
>= 15 symbols	20 477 489
Total	56 740 263

GENE

Name length	Number
<= 4 symbols	3 211 426
> 4 symbols and < 15	8 232 308
>= 15 symbols	5 383 350
Total	16 827 084

MIRNA

Name length	Number
<= 4 symbols	1 315
> 4 symbols and < 15	13 444
>= 15 symbols	5 444
Total	20 203

CELL COMPONENT

Name length	Number
<= 4 symbols	586 715
> 4 symbols and < 15	2 540 532
>= 15 symbols	1 625 235
Total	4 752 482

PROTEIN

Name length	Number
<= 4 symbols	4 101 671
> 4 symbols and < 15	10 883 771
>= 15 symbols	7 411 052
Total	22 396 494

PATHWAY

Name length	Number
<= 4 symbols	1 776 319
> 4 symbols and < 15	8 675 629
>= 15 symbols	5 959 827
Total	16 411 775

MOLECULAR FUNCTION

Name length	Number
<= 4 symbols	1 216 794
> 4 symbols and < 15	3 552 725
>= 15 symbols	2 462 455
Total	7 231 974

PHENOTYPE

Name length	Number
<= 4 symbols	2 849 181
> 4 symbols and < 15	26 206 490
>= 15 symbols	15 820 606
Total	44 876 277

DRUG

Name length	Number
<= 4 symbols	1 106 270
> 4 symbols and < 15	5 782 077
>= 15 symbols	3 014 175
Total	9 902 522

METABOLITE

Name length	Number
<= 4 symbols	8 719 427
> 4 symbols and < 15	37 241 344
>= 15 symbols	20 259 194
Total	66 219 965

ORGANISM

Name length	Number
<= 4 symbols	7 291 009
> 4 symbols and < 15	38 098 024
>= 15 symbols	25 159 256
Total	70 548 289

CELL

Name length	Number
<= 4 symbols	1 662 576
> 4 symbols and < 15	7 725 424
>= 15 symbols	5 604 347
Total	14 992 347

Figure 2. Distribution of biological object names from the ANDSystem ontology in PubMed abstracts according to their length and type. This distribution is based on the primary mapping of 34 million PubMed abstracts.

The justification for filtering short names is that the most significant error associated with semantic concept ambiguity is their more frequent intersection with common words and various abbreviations [15-18]. For example, one synonym for the cyclin-dependent kinase 4 inhibitor B gene is p15. This word often occurs in texts as a page number. Another example is flu, traditionally used as a synonym for influenza. However, the UniProt database contains information on an *Escherichia coli* gene (UID: P39180) with the same name. Similarly, the *tic* term often corresponds to impaired nervous system functioning in biomedicine. However, this term was introduced as an abbreviation for tumor-initiating cells in a study on epithelial-mesenchymal transition [40].

Figure 2 shows that most references to short names in >10 million scientific publication abstracts belong to metabolites. This finding reflects the fact that this dictionary contains numerous chemical element names whose length does not exceed two letters, such as Ca (calcium), Pb (plumbum), and Mg (magnesium). Moreover, most of these terms also intersect with different abbreviations. For example, CA is also used as a short name for California, mg as milligrams, while in medicine, Pb can be short for peripheral blood. Another example is the name gold, which is often used in the context of the gold standard.

These examples highlight errors that might appear when using only dictionary-based mapping methods. One solution to this problem is the subsequent filtering of such dictionary-mapped entities according to their context.

3.3. Context-based classification of incorrectly recognized objects

We used the transformer neural network PubMedBERT [33] to filter out object names incorrectly recognized by the dictionary. The peculiarity of the chosen model was that it was trained from scratch exclusively on PubMed abstracts. This neural model was fine-tuned for each object type to classify short names as correctly and incorrectly recognized based on the context in which the authors mention them in their texts.

Five object groups were considered: (1) proteins and genes, (2) diseases and side effects, (3) drugs and metabolites, (4) cellular components, and (5) cellular pathways. The protein and gene vocabularies were combined into a common vocabulary, as were those for diseases and side effects, and for drugs and metabolites. Our analysis did not consider organisms, phenotypic traits, miRNAs, molecular functions, and cells since their object names had not been manually annotated in the gold standard.

The automated formation of training samples for each classification model was based on the following algorithm: mapped PubMed texts containing long object names of the corresponding type from the considered group were selected as positive examples. The mapping was performed using the ANDSystem’s ontology and text-mining approach. Terms of ≥15 characters were considered as long. The number of examples mentioning such names for each selected group exceeded 1 million (Fig. 2), making it possible to use them as training sets. Often several objects can be mentioned in the text, and a given name can appear multiple times in the text. Therefore, to provide the neural network with the ability to consider the context of a particular object in the specific part of the sentence, the classified term was separately replaced by a special tag: <*ANDSYSTEM-CANDIDATE*>. A schematic illustration of the algorithm is shown in Figure 3.

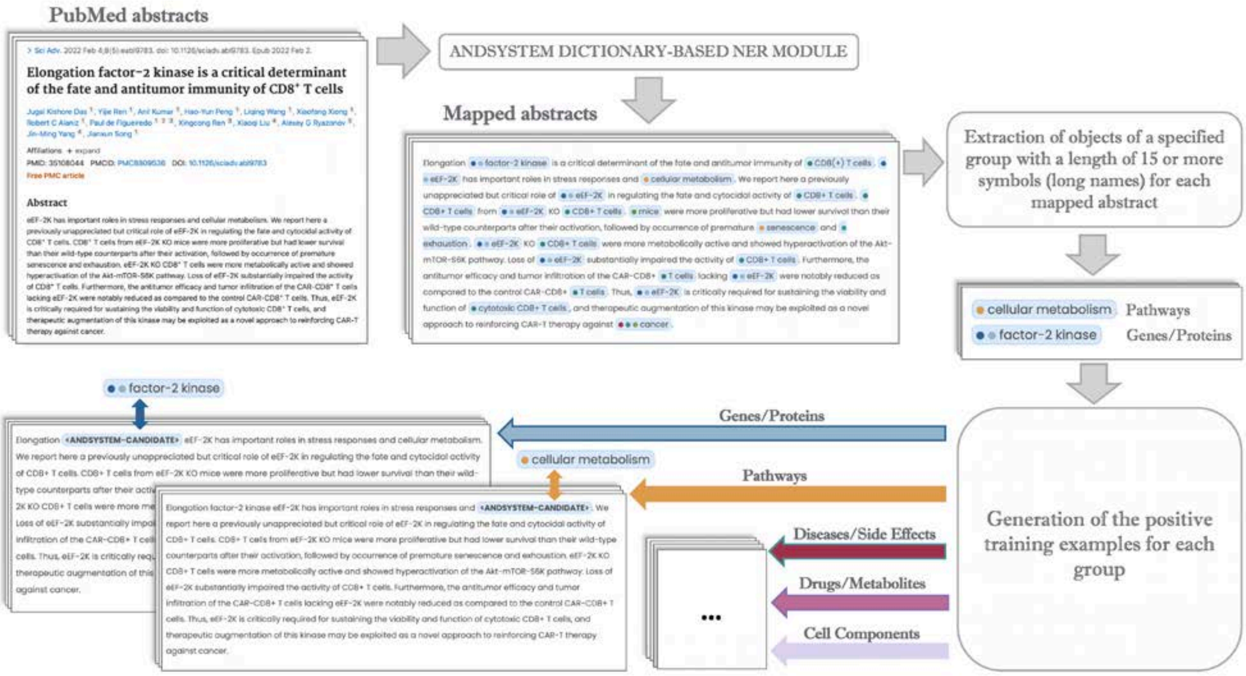
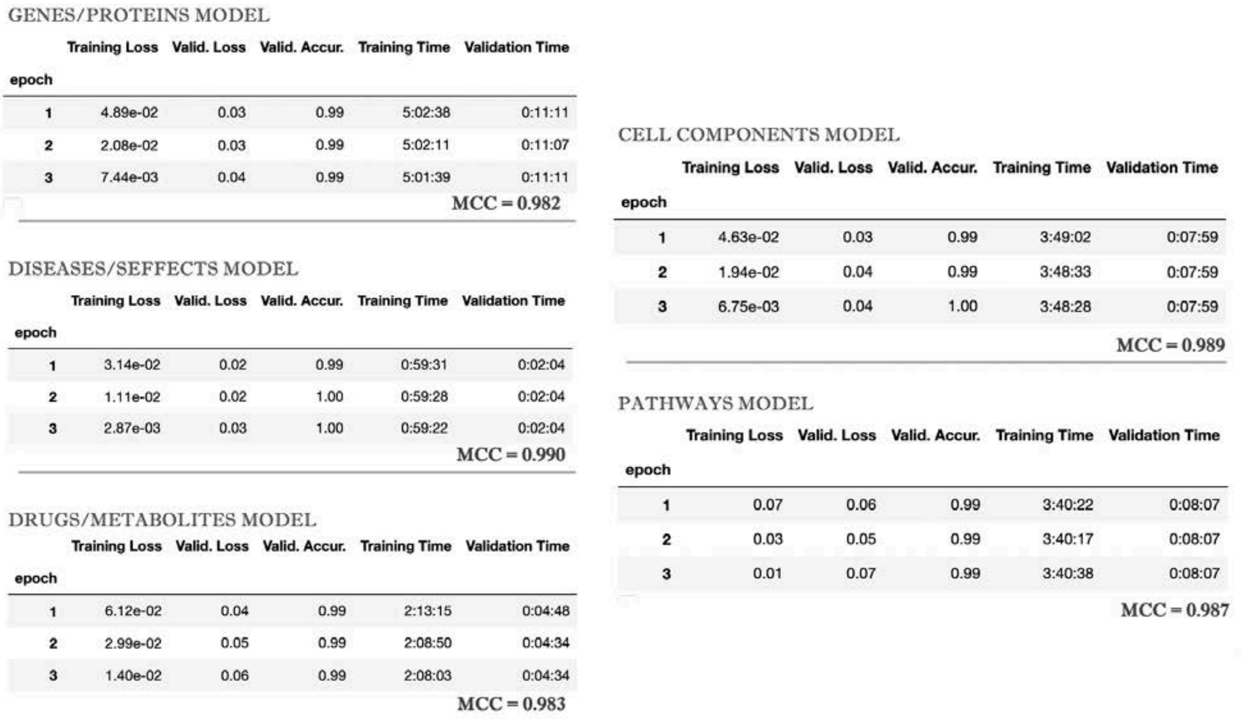

PubMed abstracts
Elongation factor-2 kinase is a critical determinant of the fate and antitumor immunity of CD8+ T cells
Abstract
ANDSYSTEM DICTIONARY-BASED NER MODULE
Mapped abstracts
Extraction of objects of a specified group with a length of 15 or more symbols (long names) for each mapped abstract
cellular metabolism
Pathways
factor-2 kinase
Genes/Proteins
Generation of the positive training examples for each group
Genes/Proteins
Pathways
Diseases/Side Effects
Drugs/Metabolites
Cell Components
factor-2 kinase
cellular metabolism
<ANDSYSTEM-CANDIDATE>
...

Figure 3. The algorithm scheme for generating training examples for fine-tuning the ANDDigest ver. 01.2022 classification models.

Positive examples for objects of another group were used as negative examples. Therefore, for the drugs/metabolites and diseases/side effects groups, data from the genes/proteins group were used as negative examples. While for the genes/proteins, cellular components, and cellular pathways groups, data from the diseases/side effects group were used as negative examples. Each model's learning set comprised 512,000 training and 50,000 validation positive and negative examples in a 1:1 ratio. All classification models were trained on the context of objects with ≥15 characters.

The datasets generated for each considered group of object types were used to fine-tune the pre-trained PubMedBERT model for the sequence classification task using Python's transformers v4.16.2 library [41] with an AdamW optimizer [42] and 2×10^{-5} learning rate. All the texts were in lowercase, and the maximum sequence length was limited to 512 words, a standard value for BERT-based models. The classification model for each group was trained as a binary classifier for three epochs. After the third epoch, each classifier's accuracy for long names was estimated by calculating Mathew's correlation coefficient (MCC). The training results for each model are shown in Figure 4.



GENES/PROTEINS MODEL

epoch	Training Loss	Valid. Loss	Valid. Accur.	Training Time	Validation Time
1	4.89e-02	0.03	0.99	5:02:38	0:11:11
2	2.08e-02	0.03	0.99	5:02:11	0:11:07
3	7.44e-03	0.04	0.99	5:01:39	0:11:11

MCC = 0.982

DISEASES/SEFFECTS MODEL

epoch	Training Loss	Valid. Loss	Valid. Accur.	Training Time	Validation Time
1	3.14e-02	0.02	0.99	0:59:31	0:02:04
2	1.11e-02	0.02	1.00	0:59:28	0:02:04
3	2.87e-03	0.03	1.00	0:59:22	0:02:04

MCC = 0.990

DRUGS/METABOLITES MODEL

epoch	Training Loss	Valid. Loss	Valid. Accur.	Training Time	Validation Time
1	6.12e-02	0.04	0.99	2:13:15	0:04:48
2	2.99e-02	0.05	0.99	2:08:50	0:04:34
3	1.40e-02	0.06	0.99	2:08:03	0:04:34

MCC = 0.983

CELL COMPONENTS MODEL

epoch	Training Loss	Valid. Loss	Valid. Accur.	Training Time	Validation Time
1	4.63e-02	0.03	0.99	3:49:02	0:07:59
2	1.94e-02	0.04	0.99	3:48:33	0:07:59
3	6.75e-03	0.04	1.00	3:48:28	0:07:59

MCC = 0.989

PATHWAYS MODEL

epoch	Training Loss	Valid. Loss	Valid. Accur.	Training Time	Validation Time
1	0.07	0.06	0.99	3:40:22	0:08:07
2	0.03	0.05	0.99	3:40:17	0:08:07
3	0.01	0.07	0.99	3:40:38	0:08:07

MCC = 0.987

Figure 4. The classification model’s training results for each selected group.

All negative sets were created using long object names of a single selected type for each classifier. Therefore, they could potentially cause the classifiers to perform well at distinguishing short gene and protein names from those specific types but not others. However, most publicly available manually curated gold standard datasets do not provide the necessary number of examples containing short object names to assess the models’ accuracy and do not cover all object types considered. Therefore, gold standards for each type were manually constructed from ANDSystem’s dictionary mapping (see Supplementary 1) to validate the obtained classifiers on the short object names (≤4 characters) of the corresponding types. Each corpus contained a short object name mapped by ANDSystem, the corresponding abstract’s PubMed ID, the sentence from which it was extracted, and a label indicating whether it was correctly or incorrectly identified.

Each model was reassessed with the gold standard, and their accuracies for the short names recognition task were calculated using receiver operating characteristic (ROC) curves. The results are shown in Figure 5.

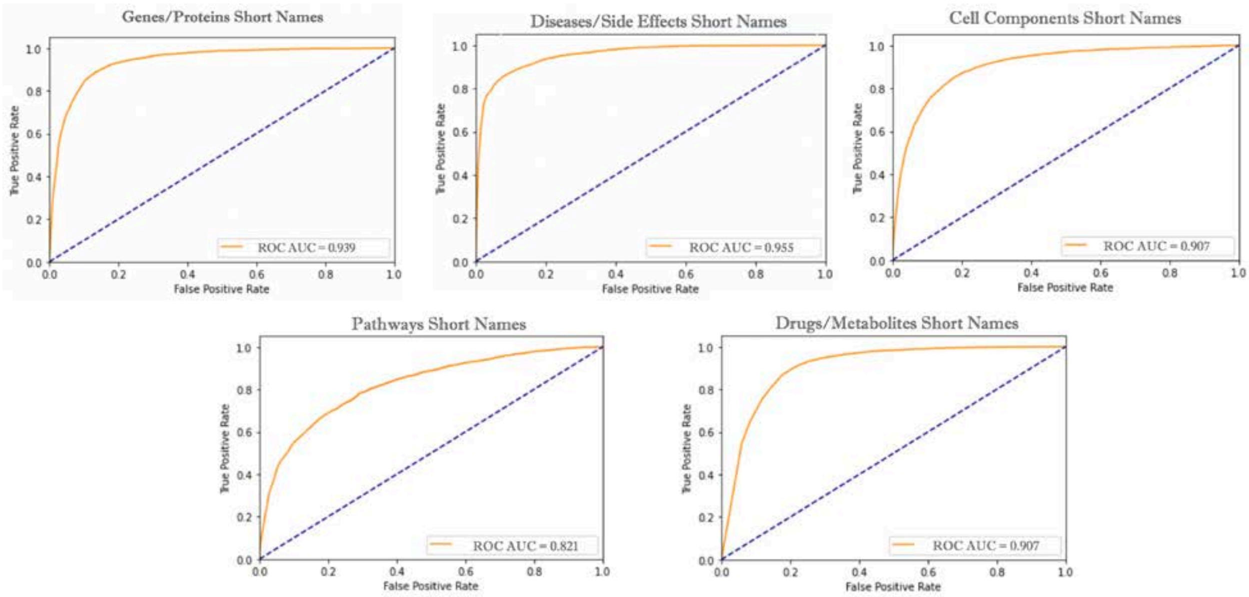


Figure 5. ROC curves illustrate short names' classification accuracy of the fine-tuned models for different object types. Genes/proteins (area under the ROC curve [AUC] = 93.9%). Diseases/side effects (AUC = 95.5%). Cell components (AUC = 90.7%). Cellular pathways (AUC = 82.1%). Drugs/metabolites (AUC = 90.7%).

The optimal thresholds for positive predictions were calculated using the reconstructed ROC curves. The curve's threshold was considered optimal when the difference between true (TPR; y-axis) and false (FPR; x-axis) positive rates was maximized. Detailed values for each model are provided in Table 1.

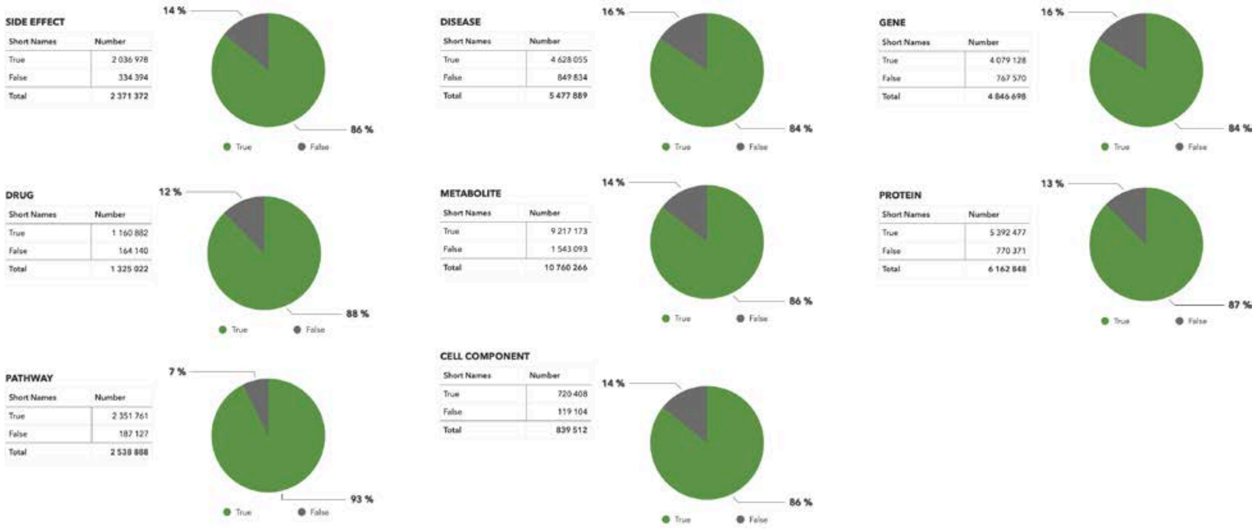
Table 1. TPR, FPR, and optimal thresholds for each classification model.

Group	TPR	FPR	Optimal threshold (positive)
Cellular components	0.85	0.17	0.9999737739562988
Diseases/side effects	0.85	0.08	0.9999943971633911
Genes/proteins	0.89	0.13	0.9999139308929443
Cellular pathways	0.80	0.21	0.9998261332511902
Drugs/metabolites	0.89	0.20	0.9999928474426270

The highest accuracy (area under the ROC curve [AUC] = 95.5%) was obtained with the diseases/side effects group. This finding can be explained by the broad specificity of PubMed abstracts, which focus mainly on biomedicine, and by the contextual peculiarities in how disease names are often used. The lowest accuracy was obtained with the cellular pathways group (AUC =

0.821). This finding likely reflects the very small number of short names in the cellular pathways dictionary (*n* = 61), the contexts in which pathway names are used, and the closeness of their names to common words.

The calculated thresholds for the developed neural networks were used to analyze the PubMed abstracts previously mapped by ANDSystem. Only abstracts containing short object names of corresponding types were considered (>10 million total documents). Statistics on the obtained results are shown in Figure 6.



SIDE EFFECT

Short Names	Number
True	2 036 978
False	334 394
Total	2 371 372

DISEASE

Short Names	Number
True	4 628 055
False	849 834
Total	5 477 889

GENE

Short Names	Number
True	4 079 128
False	767 570
Total	4 846 698

DRUG

Short Names	Number
True	1 160 882
False	164 140
Total	1 325 022

METABOLITE

Short Names	Number
True	9 217 173
False	1 543 093
Total	10 760 266

PROTEIN

Short Names	Number
True	5 392 477
False	770 371
Total	6 162 848

PATHWAY

Short Names	Number
True	2 351 761
False	187 127
Total	2 538 888

CELL COMPONENT

Short Names	Number
True	720 408
False	119 104
Total	839 512

Figure 6. Proportions of correctly and incorrectly recognized short object words in PubMed texts after dictionary-based mapping and filtering with the developed fine-tuned neural networks. Short object names' absolute numbers and proportions are based on their correctly or incorrectly recognized classification.

4. Discussion

We developed a new version of ANDDigest (ver. 01.2022; Fig. 1) incorporating a new combined text-mining AI NER module. The new module performs dictionary-based object mapping and filtering of short names erroneously recognized in texts.

Integration of the new module into ANDDigest ver. 01.2022 significantly increased the quality of object name recognition in texts. Due to the additional use of fine-tuned neural networks after the mapping stage, the recognition accuracy for short names (≤4 characters) increased by 13% on average. It should be noted that most recognition errors are traditionally associated with short names due to their linguistic ambiguity [13]. This problem leads to many false results when searching for relevant scientific literature based on user queries.

For example, one synonym for coproporphyrinogen oxidase is COX, which intersects with a Cox proportional-hazards model [43], widely used in biomedical literature. Therefore, even the previous version of ANDDigest identified >41,000 documents mentioning this object when searching only a smaller number of PubMed abstracts, most of which contained Cox. However, the new version recognized that most of those results were on the Cox regression model, returning only 1750 documents containing this term in the desired context after the context-based filtering.

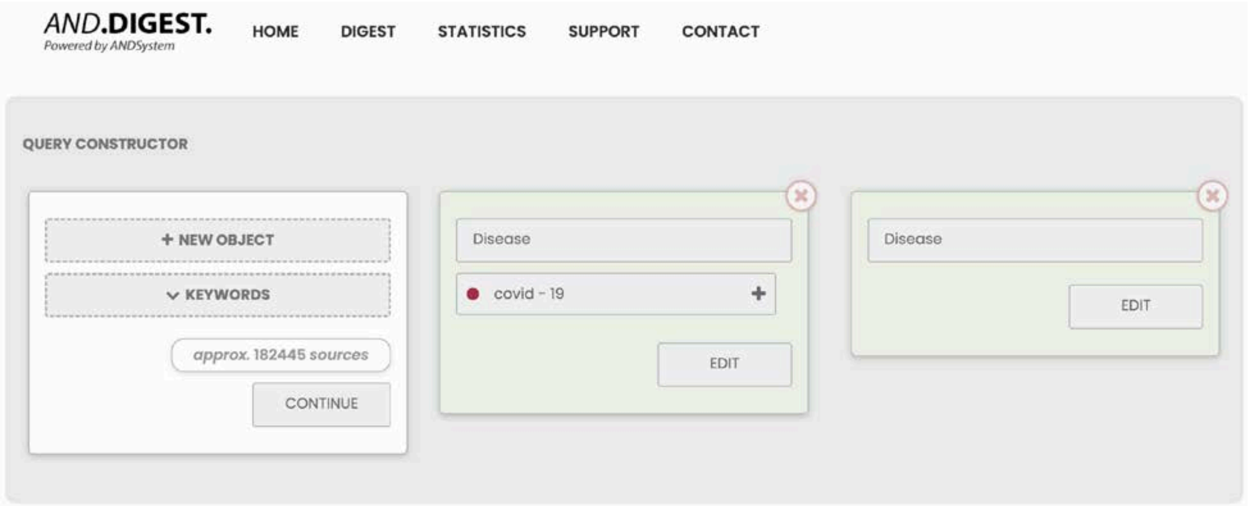
The combined AI NER text-mining module performs short name recognition filtering for eight object types: proteins, genes, drugs, metabolites, diseases, side effects, cellular components, and cellular pathways. The greatest number of incorrectly identified names filtered out using this module was for genes and diseases: 16% of all recognized short names of this type (Figure 5).

The least filtered were short names of cellular pathways (biological processes). The recognition accuracy of these objects in the gold standard without the new AI NER module was about 60%; this increased to 82% with the new AI NER module. The difficulty in identifying cellular pathway names using the proposed approach can be explained by the context in which these objects occur in the text, which is very similar to objects of other types, such as diseases.

4.1 Example use of ANDDigest ver. 01.2022 with comorbid diseases

Currently, a large proportion of the biomedical literature focuses on the problem of disease comorbidity. Comorbidity reflects the frequent joint manifestation of diseases in patients. Positive comorbidity reflects increased frequency and negative comorbidity reflects decreased frequency [44]. Our previous studies using the ANDSystem focused on molecular genetic mechanisms underlying positive disease comorbidities, such as asthma with hypertension [45, 46] and pre-eclampsia associome [47]. In addition, we explored diseases with negative comorbidities, such as asthma with tuberculosis [48]. However, widely used approaches for identifying the molecular genetic mechanisms underlying comorbid diseases search for common associated genes [49, 50]. In particular, we have shown that the proportion of genes simultaneously associated with two diseases is significantly higher for pairs of comorbid diseases compared to pairs of randomly selected diseases [48].

The study of the molecular genetic mechanisms for coronavirus disease 2019 (COVID-19) is extremely important in the context of the current pandemic [51]. We previously analyzed metabolomic data for the blood plasma of patients with COVID-19 using the ANDDigest and ANDSystem tools [52]. Therefore, we analyzed the comorbidity of COVID-19 with other diseases as a test case for applying the new ANDDigest ver. 01.2022 tool. The query formed to search for all documents mentioning COVID-19 and any other disease is shown in Figure 7.



approx. 182445 sources

Disease

● covid - 19 +

Disease

Figure 7. The search query used to find PubMed documents containing COVID-19 and mentioning at least one other disease with ANDDigest ver. 01.2022.

ANDDigest ver. 01.2022 identified 182,445 abstracts mentioning COVID-19 and at least one of 3504 other diseases after short name filtering. Next, the resulting list of diseases was filtered based on the statistical significance of their co-occurrence with COVID-19 (false discovery rate, [FDR] < 0.05) in scientific publication abstracts, identifying 84 significant diseases. The ten most common statistically significant diseases co-occurring with COVID-19 are listed in Table 2. A list of all diseases co-occurring with COVID-19, including non-significant ones, is provided in Supplementary 2.

Table 2. The top 10 diseases that significantly co-occurred with COVID-19 in the query results.

Rank	**Disease**	**Document number**	**Co-occurrence score (*p*-value)**	**FDR *p*-value (<0.05)**
1	Severe COVID-19	4584	2.08593×10^{-8}	3.496523×10^{-6}
2	Pneumonia	3944	7.00031×10^{-8}	4.968849×10^{-6}
3	Fever	3396	2.77321×10^{-8}	3.506271×10^{-6}
4	Acute respiratory distress syndrome	2600	3.20772×10^{-8}	3.593681×10^{-6}
5	Severe acute respiratory syndrome	2573	1.65303×10^{-8}	3.496523×10^{-6}
6	Infectious diseases	2524	3.42188×10^{-8}	3.627905×10^{-6}
7	Influenza	2431	3.36938×10^{-9}	2.689255×10^{-6}
8	Viral infection	2336	3.68056×10^{-8}	3.627905×10^{-6}

9	Breathlessness	1935	4.75682×10^{-8}	4.236009×10^{-6}
10	Fatigue	1738	1.01457×10^{-8}	3.203274×10^{-6}

Pneumonia, fever, and influenza were among the most frequently co-occurring diseases. The relationship between these pathologies with COVID-19 is widely discussed in the literature [53, 54]. Interestingly, one disease significantly associated with COVID-19 in the literature was delirium (33rd on the list; see Supplementary 2). Delirium is a syndrome characterized by abrupt changes in attention, awareness, and cognitive abilities. The literature discusses many factors involved in delirium's etiology. These include neuroinflammation, cerebrovascular dysfunction, altered brain metabolism, neurotransmitter imbalance, and neural network connectivity disruption [55]. In particular, some studies report that delirium is observed in elderly patients with severe COVID-19 [56, 57].

We used ANDDigest ver. 01.2022 to identify common associated genes for these two diseases using the following queries: find all publications that mention COVID-19 and at least one gene, and find all documents containing delirium and at least one gene. The first query identified 3447 genes, of which 162 significantly co-occurred with COVID-19 (FDR < 0.05). The second query identified 441 genes, of which 162 significantly co-occurred with delirium. The intersection of these two gene lists contained 230 genes common to both diseases (Supplementary 3). They included the sigma-1 receptor (FDR [COVID-19] = 3.57×10^{-5}; FDR [Delirium] = 6.00×10^{-5}) that was significant for both diseases. The sigma-1 receptor has diverse functions, including regulating neuroinflammation, neurotransmitters, neurogenesis, endoplasmic reticulum stress, and mitochondrial function. The sigma-1 receptor's significant associations with COVID-19 and delirium in the literature are consistent with its important roles in their pathologies. A graph showing the growth in publications mentioning this gene over time is shown in Figure 8.

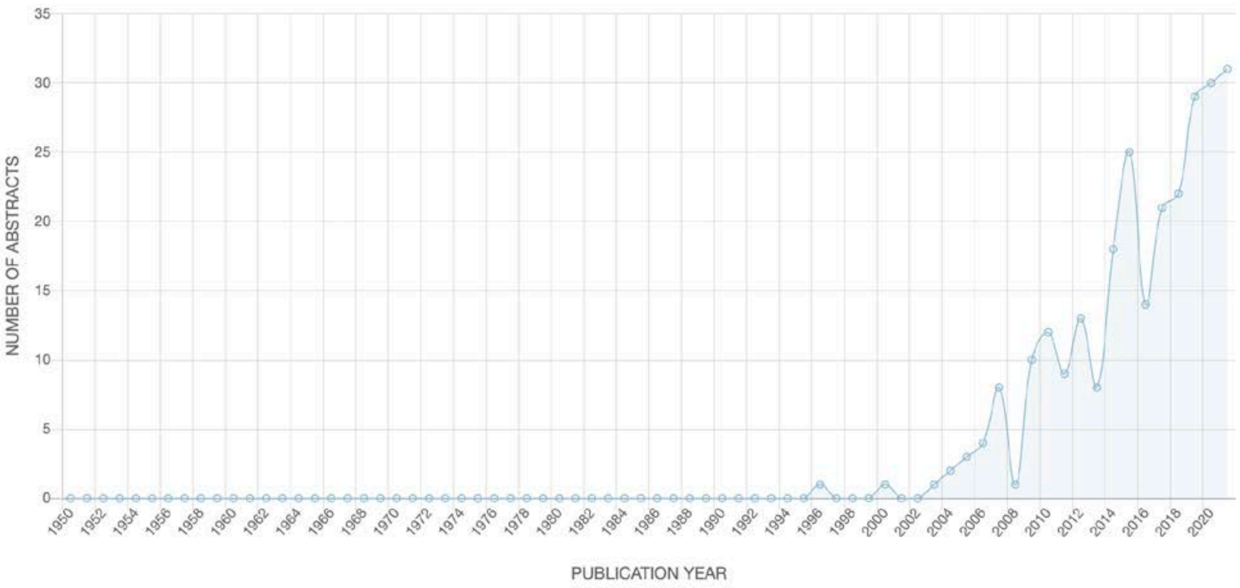


Figure 8. Graph showing the growth in PubMed abstracts mentioning the sigma-1 receptor by year generated with ANDDigest ver. 01.2022.

This gene's role in delirium has been previously discussed [58]. The role of the sigma-1 receptor as a functional host-dependency factor for the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus that causes COVID-19 has also been discussed in the literature. In particular, studies have shown that the knockout or knockdown of the sigma-1 receptor causes a consistent reduction in SARS-CoV-2 replication, suggesting that the sigma-1 receptor is important in SARS-CoV-2 replication [59].

5. Conclusions

We have shown that the developed AI NER Text-mining module integrated into ANDDigest ver. 01.2022 has high efficiency in recognizing short-named entities. A feature of the new ANDDigest version is the use of neural networks that perform binary classification of short names for biological objects in the ANDSystem ontology based only on the context in which they are mentioned. This approach makes it possible to overcome linguistic ambiguities inherent to general dictionary-based text mapping methods and the previous ANDDigest version in particular. In addition, we showed the effectiveness of our automated generation of high-quality training samples based on the context of long names for various object types. Moreover, preliminary dictionary mapping provides the user with all the necessary information about the recognized entities, such as their synonyms and links to external databases.

Author Contributions

Software development, model training and accuracy assessment were performed by I.T.V. and D.P.S.; Gold standards were created by I.T.V. and I.V.A.; Analysis of comorbidity – I.V.A.; I.T.V., D.P.S., N.A.K and I.V.A. performed writing, review and editing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

The study was funded by the Ministry of Science and Higher Education of the Russian Federation project “Kurchatov Center for World-Class Genomic Research” No. 075-15-2019-1662 from 2019-10-31. The authors are grateful to T. Pemberton from Cambridge Proofreading & Editing LLC for the careful proofreading of the manuscript.

Data Availability Statement

The developed classification models and datasets are available upon request at the following link: https://huggingface.co/Timofey.

The gold standards used for the assessment of the accuracy of the developed models are included in the supplementary materials.

Conflict of Interest

The authors declare that they have no conflict of interests.

References

1. Beel, J.; Gipp, B. *Google Scholar's Ranking Algorithm: An Introductory Overview*. In Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09), Rio de Janeiro, Brazil, 14–17 July 2009; Volume 1, pp. 230–241.

2. McEntyre, J.; Ostell, J. *The NCBI Handbook*. Bethesda (MD): National Center for Biotechnology Information (US), 2002.

3. Jacso, P. As we may search - Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Curr. Sci.* **2005**, *89*, 1537–1547.

4. Szklarczyk, D.; Gable, A.L.; Nastou K.C.; Lyon, D.; Kirsch, R.; Pyysalo, S.; Doncheva, N.T.; Legeay, M.; Fang, T.; Bork, P.; Jensen, L.J.; Mering C. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*. **2021**, 49(D1), D605–D612.

5. Nikitin, A.; Egorov, S.; Daraselia, N.; Mazo, I. Pathway studio -- the analysis and navigation of molecular networks. *Bioinformatics*. **2003**, 19(16), 2155–2157.

6. Nikolsky, Y.; Nikolskaya, T.; Bugrim, A. Biological networks and analysis of experimental data in drug discovery. *Drug discovery today*. **2005**, 10(9), 653–662.

doi:10.20944/preprints202210.0247.v1

7. Ekins, S.; Bugrim, A.; Brovold, L.; Kirillov, E.; Nikolsky, Y.; Rakhmatulin, E.; Sorokina, S.; Ryabov, A.; Serebryiskaya, T.; Melnikov, A.; Metz, J. Algorithms for network analysis in systems-ADME/Tox using the MetaCore and MetaDrug platforms. *Xenobiotica*. **2006**, 36(10–11), 877–901.

8. Demenkov, P.S.; Ivanisenko, T.V.; Kolchanov, N.A.; Ivanisenko, V.A. ANDVisio: A new tool for graphic visualization and analysis of literature mined associative gene networks in the ANDSystem. *In Silico Biol*. **2012**, 11(3–4), 149–161.

9. Ivanisenko, V.A.; Saik, O.V.; Ivanisenko, N.V.; Tiys, E.S.; Ivanisenko, T.V.; Demenkov, P.S.; Kolchanov, N.A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. *BMC Syst Biol*. **2015**, 9(S2), S2.

10. Ivanisenko, V.A.; Demenkov, P.S.; Ivanisenko, T.V.; Mishchenko, E.L.; Saik, O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinf*. **2019**, 20(1), 34.

11. Saik, O.V.; Nimaev, V.V.; Usmonov, D.B.; Demenkov, P.S.; Ivanisenko, T.V.; Lavrik, I.N.; Ivanisenko, V.A. Prioritization of genes involved in endothelial cell apoptosis by their implication in lymphedema using an analysis of associative gene networks with ANDSystem. *BMC Med Genomics*. **2019**, 12(2), 47.

12. Ivanisenko, T.V.; Saik, O.V.; Demenkov, P.S.; Ivanisenko, N.V.; Savostianov, A.N.; Ivanisenko, V.A. ANDDigest: a new web-based module of ANDSystem for the search of knowledge in the scientific literature. *BMC Bioinf*. **2020**, 21(11), 1–21.

13. Naseem, U.; Musial, K.; Eklund, P.; Prasad, M. *Biomedical named-entity recognition by hierarchically fusing biobert representations and deep contextual-level word-embedding*. In International Joint Conference on Neural Networks (IJCNN), Glasgow, United Kingdom, 19–24 July 2020; pp. 1–8.

14. Pearson, H. Biology's name game. *Nature*. **2001**, 411(6838), 631–633.

15. Wei, C.H.; Kao, H.Y.; Lu, Z. GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. *Biomed Res. Int.* **2015**, 2015.

16. Islamaj, R.; Wei, C.H.; Cissel, D.; Miliaras, N.; Printseva, O.; Rodionov, O.; Sekiya, K.; Ward, J.; Lu, Z. NLM-Gene, a richly annotated gold standard dataset for gene entities that addresses ambiguity and multi-species gene recognition. *J. Biomed. Inf.* **2021**, 118, 103779.

17. Doğan, R.I.; Leaman, R.; Lu, Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. *J. Biomed. Inf.* **2014**, 47, 1–10.

18. Krallinger, M.; Rabal, O.; Leitner, F.; Vazquez, M.; Salgado, D.; Lu, Z.; Leaman, R.; Lu, Y.; Ji, D.; Lowe, D.M.; Sayle, R.A.; Batista-Navarro, R.T.; Rak, R.; Huber, T.; Rocktaschel, T.; Matos, S.; Campos, D.; Tang, B.; Xu, H.; Munkhdalai, T.; Ryu, K.H.; Ramanan, S.V.; Nathan, S.; Zitnik, S.; Bajek, M.; Weber, L.; Irmer, M.; Akhondi, S.A.; Kors, J.A.; Xu, S.; An, X.; Sikdar, U.K.; Ekbal, A.; Yoshioka, M.; Dieb, T.M.; Choi, M.; Verspoor, K.; Khabsa, M.; Giles, C.L.; Liu, H.; Ravikumar, K.E.; Lamurias, A.; Couto, F.M.; Dai, H.J.; Tsai, R.T.H.; Ata, C.; Can, T.; Usie, A.; Alves, R.; Segura-Bedmar, I.; Martinez, P.; Oyarzabal, J.; Valencia, A. Krallinger M. et al. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J. Cheminf.* **2015**, 7(1), 1–17.

19. Song, Y.; Kim, E.; Lee, G.G.; Yi, B.K. POSBIOTM—NER: a trainable biomedical named-entity recognition system. *Bioinformatics*. **2005**, 21(11), 2794–2796.

20. Chang, F.X.; Guo, J.; Xu, W.R.; Chung, S.R. Application of word embeddings in biomedical named entity recognition tasks. *J Digit Inf Manage*. **2015**, 13(5), 321–327.

21. Wei, H.; Gao, M.; Zhou, A.; Chen, F; Qu, W.; Wang, C.; Lu, M. Named Entity Recognition From Biomedical Texts Using a Fusion Attention-Based BiLSTM-CRF. *IEEE Access*. **2019**, 7, 73627–73636.

22. Weber, L.; Munchmeyer, J.; Rocktaschel, T.; Habibi, M.; Leser, U. HUNER: improving biomedical NER with pretraining. *Bioinformatics*. **2019**, 36(1), 295–302.

23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*. **2017**, 30.

24. Shoeybi, M.; Patwary, M.; Puri, R.; LeGresley, P.; Casper, J.; Catanzaro, B. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint*. **2019**, arXiv:1909.08053.

25. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; Amodei, D. Language models are few-shot learners. *Advances in neural information processing systems*. **2020**, 33, 1877–1901.

26. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*. **2018**, arXiv:1810.04805.

27. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint*. **2013**, arXiv:1301.3781.

28. Pennington, J.; Socher, R.; Manning, C.D. *Glove: Global vectors for word representation*. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, 26–28 October 2014; pp. 1532–1543.

29. Vaidhya, T.; Kaushal, A. Domain specific BERT representation for Named Entity Recognition of lab protocol. *arXiv preprint*, **2020,** arXiv:2012.11145.

30. 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*. **2020**, 36(4), 1234–1240.

31. Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; Fidler, S. *Aligning books and movies: Towards story-like visual explanations by watching movies and reading books*. In Proceedings of the IEEE international conference on computer vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 19–27.

32. Wikipedia a free encyclopedia. https://wikipedia.org/ (archived on 01.09.2022).

33. Davagdorj, K.; Park, K.H.; Amarbayasgalan, T.; Munkhdalai, L.; Wang, L.; Li, M.; Ryu, K.H. *BioBERT Based Efficient Clustering Framework for Biomedical Document Analysis*. In the 14th International Conference on Genetic and Evolutionary Computing (ICGEC 2021), Jilin City, China, 21–23 October 2021; pp. 179–188.

34. Qaseem, A.; Gunther, S. CPRiL: compound–protein relationships in literature. *Bioinformatics*. **2022**, 38(18), 4452–4453.

35. Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*. **2021**, 3(1), 1–23.

36. Li, J.; Wei, Q.; Ghiasvand, O.; Chen, M.; Lobanov, V.; Weng, C.; Xu, H. A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora. *BMC Med Inform Decis Mak*. **2022**, 22(Suppl 3), 235.

37. Mobasher, G.; Mertová, L.; Ghosh, S.; Krebs, O.; Heinlein, B.; Müller, W. Combining dictionary-and rule-based approximate entity linking with tuned BioBERT. *bioRxiv*. **2021**.

38. Hipel, K.W.; McLeod, A.I. Time series modelling of water resources and environmental systems.; Elsevier: London, UK, 1994; pp. 1–1012.

39. Libiseller, C.; Grimvall, A. Performance of partial Mann-Kendall tests for trend detection in the presence of covariates. *Environmetrics: The official journal of the International Environmetrics Society*. **2002**, 13(1), 71–84.

40. Creighton, C.J.; Chang, J.C.; Rosen, J.M. Epithelial-Mesenchymal Transition (EMT) in Tumor-Initiating Cells and Its Clinical Implications in Breast Cancer. *J Mammary Gland Biol Neoplasia*. **2010**, 15, 253–260.

41. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T.L.; Gugger, S.; Drame, M.; Lhoest, Q.; Rush, A.M.

Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations (EMNLP 2020), Online, 16–20 November 2020; pp. 38–45.

42. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv preprint*. **2017**, arXiv:1711.05101.

43. Cox, D.R. Regression models and life-tables. *J. R. Stat. Soc. B.* **1972**, 34(2), 187– 202.

44. Gijsen, R.; Hoeymans, N.; Schellevis, F.G.; Ruwaard, D.; Satariano, W.A.; van den Bos, G.A. Causes and consequences of comorbidity: a review. *Journal of clinical epidemiology*. **2001**, 54(7), 661–674.

45. Zolotareva, O.; Saik, O.V.; Königs, C.; Bragina, E.Y.; Goncharova, I.A.; Freidin, M.B.; Dosenko, V.E.; Ivanisenko, V.A.; Hofestädt, R. Comorbidity of asthma and hypertension may be mediated by shared genetic dysregulation and drug side effects. *Sci Rep*. **2019**, 9(1), 16302.

46. Saik, O.V.; Demenkov, P.S.; Ivanisenko, T.V.; Bragina, E.Y.; Freidin, M.B.; Goncharova, I.A.; Dosenko, V.E.; Zolotareva, O.I.; Hofestaedt, R.; Lavrik, I.N.; Rogaev, E.I.; Ivanisenko, V.A. Novel candidate genes important for asthma and hypertension comorbidity revealed from associative gene networks. *BMC Med Genomics*. **2018**, 11(Suppl 1), 15.

47. Glotov, A.S.; Tiys, E.S.; Vashukova, E.S.; Pakin, V.S.; Demenkov, P.S.; Saik, O.V.; Ivanisenko, T.V.; Arzhanova, O.N.; Mozgovaya, E.V.; Zainulina, M.S.; Kolchanov, N.A.; Baranov, V.S.; Ivanisenko, V.A. Molecular association of pathogenetic contributors to pre-eclampsia (pre-eclampsia associome). *BMC Syst. Biol.* **2015**, 9(Suppl 2), S4.

48. Bragina, E.Y.; Tiys, E.S.; Freidin, M.B.; Koneva, L.A.; Demenkov, P.S.; Ivanisenko, V.A.; Kolchanov, N.A.; Puzyrev, V.P. Insights into pathophysiology of dystropy through the analysis of gene networks: An example of bronchial asthma and tuberculosis. *Immunogenetics*. **2014**, 66(7–8), 457–465.

49. Hofestädt, R.; Ivanisenko, V. Integrative Analysis of Co-Morbid Multifactorial Diseases. *J Integr Bioinform*. **2018**, 15(4), 20180088.

50. Bragina, E.Y.; Goncharova, I.A.; Garaeva, A.F.; Nemerov, E.V.; Babovskaya, A.A.; Karpov, A.B.; Semenova, Y.V.; Zhalsanova, I.Z.; Gomboeva, D.E.; Saik, O.V.; Zolotareva, O.I.; Ivanisenko, V.A.; Dosenko, V.E.; Hofestaedt, R.; Freidin, M.B. Molecular Relationships between Bronchial Asthma and Hypertension as Comorbid Diseases. *J Integr Bioinform*. **2018**, 15(4).

51. Sachs, J.D.; Karim, S.S.A.; Aknin, L.; Allen, J.; Brosbøl, K.; Colombo, F.; Barron, G.C.; Espinosa, M.F.; Gaspar, V.; Gaviria, A.; Haines, A.; Hotez, P.J.; Koundouri, P.; Bascuñán, F.L.; Lee J.-K.; Pate, M.A.; Ramos, G.; Reddy, K.S.; Serageldin, I.; Thwaites, J.; Vike-Freiberga, V.; Wang, C.; Were, M.K.; Xue, L.; Bahadur, C.; Bottazzi, M.E.; Bullen, C.; Laryea-Adjei, G.; Amor, Y.B.; Karadag, O.; Lafortune, G.; Torres, E.; Barredo, L.; Bartels, J.G.E.; Joshi, N.; Hellard, M.; Huynh, U.K.; Khandelwal, S.; Lazarus, J.V.; Michie, S. The Lancet Commission on lessons for the future from the covid-19 pandemic. *The Lancet*. **2022**.

52. Ivanisenko, V.A.; Gaisler, E.V.; Basov, N.V.; Rogachev, A.D.; Cheresiz, S.V.; Ivanisenko, T.V.; Demenkov, P.S.; Mishchenko, E.L.; Khripko, O.O.; Khripko, Yu.I.; Voevoda, S.M.; Karpenko, T.N.; Velichko, A.J.; Voevoda, M.I.; Kolchanov, N.A.; Pokrovsky, A.G. Metabolomics analysis reveals a potential role SARS-CoV-2 viral proteins in the regulation/perturbation of metabolic pathways associated with altered plasma metabolites in COVID-19 patients. *Research Square*. **2022**.

53. Gattinoni, L.; Gattarello, S.; Steinberg, I.; Busana, M.; Palermo, P.; Lazzari, S.; Romitti, F.; Quintel, M.; Meissner, K.; Marini, J.J.; Chiumello, D. COVID-19 pneumonia: Pathophysiology and management. *European Respiratory Review*. **2021**, 30(162).

54. Ozaras, R.; Cirpin, R.; Duran, A.; Duman, H.; Arslan, O.; Bakcan, Y.; Kaya, M.; Mutlu, H.; Isayeva, L.; Kebanlı, F.; Deger, B.A. Influenza and COVID-19 coinfection: report of six cases and review of the literature. *J. Med. Virol.* **2020**, 92(11), 2657–2665.

55.Wilson, J.E.; Mart, M.F.; Cunningham, C.; Shehabi, Y.; Girard, T.D., MacLullich, A.M.; Slooter, A.J.; Ely, E. Delirium. *Nat. Rev. Dis. Primers*. **2020**, 6(1), 1–26.

56. Pun, B.T.; Badenes, R.; La Calle, G.H.; Orun, O.M.; Chen, W.; Raman, R.; Simpson, B.-G.K.; Wilson-Linville, S.; Olmedillo, B.H.; de la Cueva, A.V.; van der Jagt, M.; Casado, R.N.; Sanz, P.L.; Orhun, G.; Gómez, C.F.; Vázquez, K.N.; Otero,

P.P.; Taccone, F.S.; Curto, E.G.; Caricato, A.; Woien, H.; Lacave, G.; O'Neal Jr, H.R.; Peterson, S.J.; Brummel, N.E.; Girard, T.D.; Ely, E.W.; Pandharipande, P.P. Prevalence and risk factors for delirium in critically ill patients with COVID-19 (COVID-D): a multicentre cohort study. *The Lancet Respiratory medicine*. **2021**, 9(3), 239–250.

57. Hariyanto, T.I.; Putri, C.; Hananto, J.E.; Arisa, J.; Situmeang, R.F.V.; Kurniawan, A. Delirium is a good predictor for poor outcomes from coronavirus disease 2019 (COVID-19) pneumonia: a systematic review, meta-analysis, and meta-regression. *Journal of psychiatric research*. **2021**, 142, 361–368.

58. Wang, Y.M.; Xia, C.Y.; Jia, H.M.; He, J.; Lian, W.W.; Yan, Y,; Wang, W.P.; Zhang, W.K.; Xu, J.K. Sigma-1 receptor: A potential target for the development of antidepressants. *Neurochem. Int*. **2022**, 105390.

59. Gordon, D.E.; Hiatt, J.; Bouhaddou, M.; Rezelj, V.V.; Ulferts, S.; Braberg, H.; Jureka, A.S.; Obernier, K.; Guo, J.Z.; Batra, J.; Kaake, R.M. Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. *Science*. **2020**, 370(6521), eabe9403.