Accelerating COVID-19 Research with Graph Mining and Transformer-Based Learning

Ilya Tyagin[1], Ankit Kulshrestha[2], Justin Sybrandt*[3], Krish Matta[4], Michael Shtutman[5], Ilya Safro[2]

[1] Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE
[2] Computer and Information Sciences, University of Delaware, Newark, DE
[3] School of Computing, Clemson University, Clemson, SC
[4] Charter School of Wilmington, Wilmington, DE
[5] Drug Discovery and Biomedical Sciences, University of S. Carolina, Columbia, SC
tyagin@udel.edu, akulshr@udel.edu, jsybran@clemson.edu,
matta.krish@charterschool.org, shtutmanm@sccp.sc.edu, isafro@udel.edu

Abstract

In 2020, the White House released the "Call to Action to the Tech Community on New Machine Readable COVID-19 Dataset," wherein artificial intelligence experts are asked to collect data and develop text mining techniques that can help the science community answer high-priority scientific questions related to COVID-19. The Allen Institute for AI and collaborators announced the availability of a rapidly growing open dataset of publications, the COVID-19 Open Research Dataset (CORD-19). As the pace of research accelerates, biomedical scientists struggle to stay current. To expedite their investigations, scientists leverage hypothesis generation systems, which can automatically inspect published papers to discover novel implicit connections. We present automated general purpose hypothesis generation systems AGATHA-C and AGATHA-GP for COVID-19 research. The systems are based on the graph mining and transformer models. The systems are massively validated using retrospective information rediscovery and proactive analysis involving human-in-the-loop expert analysis. Both systems achieve high-quality predictions across domains in fast computational time and are released to the broad scientific community to accelerate biomedical research. In addition, by performing the domain expert curated study, we show that the systems are able to discover ongoing research findings such as the relationship between COVID-19 and oxytocin hormone.

All code, details, and pre-trained models are available at https://github.com/IlyaTyagin/AGATHA-C-GP.

Introduction

Development of vaccines for COVID-19 is a major triumph of modern medicine and humankind's ability to accelerate scientific research. While we are all hoping to see large-scale positive changes from fast mass adoption of the existing vaccines, there remain significant open research questions around COVID-19. The scientific community has a responsibility to do everything possible to block the ongoing transmission of the dangerous virus and *accelerate research to mitigate its consequences*. We present the following automated knowledge discovery system in order to propose new tools that could compliment the existing arsenal of techniques to accelerate biomedical and drug discovery research for events like COVID-19.

The COVID-19 pandemic became one of the most important events in the information space since the end of 2019. The pace of published scientific information is unprecedented and spans all resolutions, from the news and pop-science articles to drug design at the molecular level. The pace of scientific research has already been a significant problem in science for years (Spangler 2015), and under current circumstances this factor becomes even more pronounced. Several thousands papers are being added *weekly* to CORD-19[1] (the dataset of publications related to COVID-19) and even more in MEDLINE[2]. As a result, groups working on similar problems may not be immediately aware of the other's findings, which can lead to inefficient investments and production delays.

Although, there are quite a few hypothesis generation (HG) systems (Gopalakrishnan et al. 2019) including those we have previously proposed (Sybrandt, Shtutman, and Safro 2017; Sybrandt et al. 2020), *none of them is currently COVID-19 customized and available in the* **open domain** *to massively process related queries*. In addition to the traditional requirements for HG systems, such as high-quality results of hypotheses, interpretability and availability for broad scientific community, a specific demand for COVID-19 data analysis requires: (1) customization of the vocabulary and other logical units such as subject-verb-object predicates; (2) customization of the training data that in the reality of urgent research contains a lot of controversial and incorrect information; (3) multiple models for different information resolutions (e.g., microscopic for drug design, and macroscopic for the population related conclusions); and (4) validation on the on-going domain-specific discovery.

Our contribution: In this work we bridge this gap by releasing, AGATHA-C and AGATHA-GP , reliable and easy to use HG systems that demonstrate state-of-the art performance and validate their inference capabilities on both COVID-19 related and general biomedical data. To

*Now with Google Brain. Contact: jsybrandt@google.com.

[1] https://www.semanticscholar.org/cord19

[2] https://www.nlm.nih.gov/bsd/stats/cit_added.html

make them closely related to different goals of COVID-19 research, they correspond to micro- (AGATHA-C, for COVID-19) and macroscopic (AGATHA-GP, for general purpose) scales of knowledge discovery. Both systems are trained on all existing biomedical literature available through NIH and CORD-19 and able to process any queries to connect biomedical concepts but AGATHA-C exhibits better results on the molecular scale queries, e.g., those that are relevant to drug design, and AGATHA-GP works better for general queries, e.g., establishing connections between certain profession and COVID-19 transmission. As it will be explained later, we emphasize that AGATHA is not a traditional *information retrieval* system that effectively searches for *existing* information and thus cannot be compared to them. Instead, AGATHA generates novel hypotheses.

Both systems are the next generation of the AGATHA knowledge network mining transformer model (Sybrandt et al. 2020). *They substantially improve the quality of the previous AGATHA by introducing new information layers into multi-layered semantic knowledge network pipeline, and expanding new information retrieval techniques that facilitate inference.* We present the deep learning transformer-based AGATHA-C/GP models trained with up-to date datasets and provide easy to use interface to broad scientific community to conduct COVID-19 research. We validate the system via candidate ranking (Sybrandt, Shtutman, and Safro 2018; Sybrandt et al. 2020) using very recent scientific publications containing findings absent in the training set. While the original AGATHA has demonstrated state-of-the-art performance for the time of its release, AGATHA and other systems were found to perform with notably lower quality on extremely rapidly changing COVID-19 research. We demonstrate a remarkable improvement on different types of queries with very fast query process that allows massive validation. In addition, we demonstrate that the proposed system can identify recently uncovered gene (BST2) and hormone (oxytocin and melatonin) relationships to COVID-19, using only papers published before these connections were discovered.

More technical details are available in our extended preprint version of the paper at arXiv (Tyagin et al. 2021).

Background and Related Work

A number of works have been proposed to organize the CORD-19 literature into structured graphs for different purposes (Basu et al. 2020; Oniani et al. 2020). A major shortcoming of these approaches is that they are limited to either specific kind of entities or relations or both and as a result not only the scope of possible new literature is narrowed but a lot of additional useful knowledge is filtered out of the system.

A major interest of constructing knowledge graphs is to allow medical researchers to re-purpose existing drugs for treating COVID-19. Zhang *et al.* (Zhang et al. 2020) develop a system that uses combined semantic predicates from SemMedDB and CORD-19 (extracted using SemRep) to recommend drugs for COVID-19 treatment. To improve the predications from CORD-19, the authors fine tune various transformer based models on a manually annotated internal dataset.

The most similar to our current work is the system proposed by (Nordon et al. 2019). The authors use Electronic Medical Record (EMR) to generate candidates for drug repurposing and then use a knowledge graph constructed from MEDLINE documents to validate those candidates. In context of recent diseases like COVID-19 for which EMRs are not as readily available, the system will be limited in their candidate generation. Moreover, validation using only SemMedDB publication data may yield subpar results since it is not updated as frequently. In contrast, our method incorporates new sources of literature during the graph construction phase itself and can be readily adapted to new and emerging challenges in medicine. Other systems that are built for COVID-19 drug discovery include systems by (Martinc et al. 2020) and Kinderminer (Kuusisto et al. 2017). The former tool uses fine-tuned SciBERT model to generate contexualized embeddings given an initial seed set of words and the latter system uses a keyword co-count algorithm to propose candidates for COVID-19. We observe that our graph contains a larger variety of data sources data than any of these tools and thus can produce broader set of hypotheses.

The lack of broader applicability of systems like these in the situation with COVID-19 pandemic demonstrates that several major issues exist and require immediate attention:
(1) Most of the existing HG systems are domain-specific (e.g., gene-disease interactions) that is usually expressed in limiting the processed information (e.g., significant filtering vocabulary and papers to a specific domain in probabilistic topic modeling (Wang et al. 2011));
(2) A proper validation of HG system remains a technical problem because multiple large-scale models have to be trained with all heterogeneous data carefully eliminated several years back;
(3) Moreover, a large number of HG systems are not massively validated at all except of very old findings rediscovery (Smalheiser 2017) or demonstrating just a few proactive examples in manually curated investigation; and
(4) Interpretability and explainability of generated hypotheses remain a major issue.

Pipeline Summary

We briefly summarize the AGATHA semantic graph construction pipeline. It is described in greater details in (Sybrandt et al. 2020).

Text pre-processing. The input for our system is *a corpora of scientific citations* from the MEDLINE and CORD-19 datasets. These files contain titles and abstracts for millions of biomedical papers. We filter non-English documents, using the FastText Language Identification model (Joulin et al. 2016) if the language is not provided. After that we split all abstracts into sentences and process all sentences with ScispaCy library. From each sentence we extract POS-annotated lemmas, entities and perform n-gram mining, where $n \in [2, 3, 4]$ and n-grams are composed of frequently co-occurring lemmas. Additionally, we associate all sentences with any relevant metadata, such as the

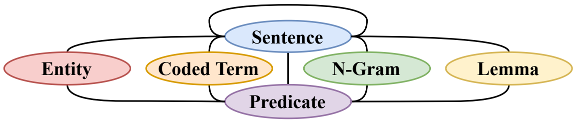

Figure 1: AGATHA multi-layered graph schema.

MeSH/UMLS (Bodenreider 2004) keywords provided along with the citation.

Semantic Graph Construction. We construct a semantic graph containing different types of nodes, namely, sentences, entities, coded terms (from UMLS and MeSH), n-grams, lemmas, and predicates following the schema depicted in Figure 1. Edges between sentences are induced from the nearest-neighbors network of sentence embeddings. We also include an edge between two sentences that appear sequentially within the same abstract, counting the title as the first sentence. Other edges can be inferred directly from the recorded metadata. For instance, the node representing the entity "COVID-19" is connected to every sentence and predicate that discuss COVID-19.

NLM UMLS implementation. The prior AGATHA semantic network only includes UMLS terms that appear in SemMedDB predicates (Kilicoglu et al. 2012) which is a major limitation. In this work we enrich the "Coded Term" layer by introducing an additional preprocessing phase wherein we run the SemRep tool with full-fielded output option ourselves *on the entire input corpora*. This phase would be necessary as CORD-19 and most recent MEDLINE citations are not represented within slowly updated SemMedDB. However, we find that we can substantially increase the quality of recovered terms by applying these tools ourselves. By doing that we not only enrich the "Coded Terms" semantic network layer, but also introduce a significant number of uncovered previously semantic predicates.

Graph Embedding. The resulting semantic graph is a large undirected heterogeneous network, where each node has its own type (as shown in Figure 1) and each edge between nodes with types u and v corresponds to type uv. At this point we additionally clarify that the constructed network is *not a traditional homogeneous knowledge graph.* We embed the network using a heterogeneous technique that captures node similarity through a *biased transformed dot product*. By explicitly including a bias term for each node, we capture a concepts overall affinity within the network that is critical for such general terms as "coronavirus." By learning transformations between each pair of node types (e.g., between sentences and lemmas), we enable each type to occupy embedding spaces with differing characteristics. Specifically, we fit an embedding model that optimizes the following similarity measure:

$$\mathcal{S}(u, v) = \hat{u}_1 + \hat{v}_1 + T_1^{uv} + \sum_{i=2}^{d} \hat{u}_i(\hat{v}_i + T_i^{uv}), \qquad (1)$$

where d - space dimensionality, u, v are nodes in the semantic graph with embeddings $\hat{u}, \hat{v}$, and T^{uv} is the directional transformation vector between nodes of u's type to nodes of v's.

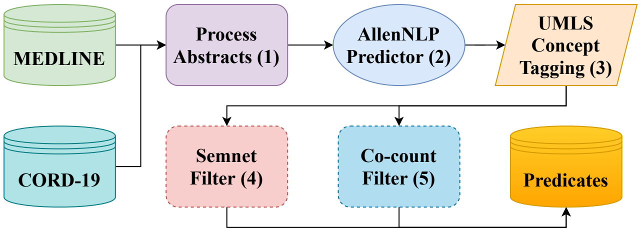

Figure 2: Predicate Extraction pipeline with Deep Learning based Open IE system.

Ranking Semantic Predicates (Transformer model). After we obtain embeddings per node in the semantic graph, we train AGATHA system ranking model. This model is trained to rank published subject-object pairs above randomly composed pairs of UMLS concepts (negative samples). Two coded terms, along with a fixed-size random subsample of predicates containing each term are input to this model. Graph embeddings for each term and predicate are fed into stacked transformer encoder layers, which apply multi-headed self-attention across the embedding set. The last set of encodings are averaged and the result is projected to the unit interval, forming a scalar prediction for the input's "plausibility."

Augmenting Semantic Predicates with Deep Learning

We used SemRep predicate extraction system in the first system, AGATHA-C , to extract predicates from the abstracts. However, SemRep relies on expert coded rules and heuristics to extract biomedical relations leading to significantly fewer predicates for training. Thus, in order to augment the predicates (for the second system, AGATHA-GP) we decided to use a deep learning based information extraction system by Stanvosky *et al.* (Stanovsky et al. 2018). Figure 2 shows our overall predicate extraction pipeline. We index our pipeline by box numbers and describe them below.

Abstract Pre-processing Box [Box 1]. We use SemRep tool described in previous sections to process the abstracts and mine information about the biomedical entities and potential semantic relations. This information is processed into a record-like data structure and is augmented throughout the rest of the pipeline.

Raw Predicate Extraction [Boxes 2 & 3]. We use the model described in (Stanovsky et al. 2018) as the deep learning model to extract semantic predicates. The model is provided as a prediction endpoint by AllenNLP and is trained to predict the *beginning-input-output* (BIO) tags for a particular sentence and classify them into subject, verb and object. We extract and match the UMLS concepts contained in the phrases and store these as "raw predicates" for further filtering.

Semnet Filtering [Box 4]. The raw predicates extracted from previous stage contain superfluous and spurious connections between concepts that can only serve to increase the noise in the training dataset. Hence we design two distinct filtering strategies to extract the most relevant predicates. Semnet filtering takes advantage of the UMLS concept's semantic types to construct a relationship graph between all known types. If the subject and object terms are connected via some path in this graph then the predicate is retained and otherwise removed.

Co-count Filtering [Box 5]. A second strategy of pruning is to get information about the terms that co-occur the most in our corpus. A larger degree of co-occurrence implies a greater correlation between concepts. We use a normalized frequency count as a scoring measure of co-occurrence. All predicates that contain subject and object terms below a certain threshold are pruned from the list of candidate predicates.

Validation

A fair validation of HG systems is extremely challenging, as these models are designed to predict *novel* connections that are unknown to even those who evaluate the system (Sybrandt et al. 2018). In addition, even if validated by rediscovering findings using historical data, the process is computationally expensive because of the need to train multiple models to understand how many months (or years) back the HG system can predict the findings, which requires careful filtering of the used papers, vocabulary and other types of information. To present our results in terms of its usefulness for urgent CORD-19-related HG, we use a historical benchmark, which is conceptually described in (Sybrandt et al. 2020). This technique is fully automated and does not require any domain experts intervention.

Positive samples collection. We use SemRep tool and proposed in the previous section approach to process the most recent CORD-19 citations, which were published after the specific cut date making sure that the citations are not included in the training set. After that we extract all subject-object pairs from the obtained results and explicitly check that none of these pairs are presented in the training set. Pairs mentioned in the CORD-19 less than twice are filtered out from the validation set. Almost all of them are either noisy or represent information that already appears in other pairs (e.g., because of the difference in grammar).

We also use the strategy of **subdomain recommendation**. This strategy works in the following way. For each UMLS term we collect its semantic type (which is a part of the metadata provided in UMLS metathesaurus) and group all extracted SemRep pairs by the term-pair criteria (combination of subject and object types). Then we identify the top-10 most common term-pairs subdomains and construct the validation set from pairs belonging to these 10 subdomains.

Negative samples generation. To generate negative samples per domain, the random sampling is used, that is, for each positive sample we keep its subject and randomly sample the object belonging to the same semantic type as the object of the source pair. We do this 10 times, thus having 10 negative domain-specific samples for each positive sample. When the validation set is generated, we apply our ranking criteria to it, obtaining a numerical score value s per each sample, where $s \in [0, 1]$.

Evaluation metrics. We propose our approach as a recommendation system and to report our results we use a combination of the following classification and recommendation metrics:

- Classification metrics: (1) Area under the receiver-operating-characteristic curve (AUC ROC); (2) Area under the precision-recall curve (AUC PR).
- Recommendation metrics: (1) Top-k precision (P.@k); (2) Average precision (AP.@k).

We report these numbers in per subdomain manner to better understand how the system performs with respect to specific task (e.g. drug repurposing).

Results

To report the results, we provide the performance measures for three AGATHA models trained on the same input data (MEDLINE corpus and CORD-19 abstracts dataset):

1. AGATHA-O : Baseline AGATHA model (Sybrandt et al. 2020);
2. AGATHA-C : AGATHA-O with new UMLS layer and SemRep enrichment;
3. AGATHA-GP : AGATHA-C with additional deep learning-based extracted and further filtered predicates.

It is done in this particular manner because the major role in learning the proposed ranking criteria depends heavily on the quality of extracted semantic predicates and their number, as they form the training set for the AGATHA ranking module. *At the moment of writing, no other general purpose and available for public use HG system compliant with the three validation criteria, namely, (a) ability to run thousands of queries in a reasonable time, (b) ability to process COVID-19 related vocabulary, and (c) ability to operate in multiple domains was available for comparison.* Comparison of the baseline AGATHA-O is discussed in (Sybrandt et al. 2020).

The performance of both AGATHA-C and AGATHA-GP allows to run thousands of queries in a very short time (in the order of minutes), making the validation on a large number of samples possible. Unfortunately, given the current circumstances, large-scale validation for the specific scientific subdomain (COVID-19 related hypotheses) is hard to implement, because well-established and reliable factual base is being actively developed at the moment and big historic gap for the vocabulary simply does not exist (e.g., the COVID-19 term is less than two years old). We, however, provide the validation set of positive connections extracted from CORD-19 dataset citations added within the time frame from October 28, 2020 to January 21, 2021, which numbered at 77 thousand abstracts.

The overall training dataset contains 190.6 million sentences, which results in 287 million nodes and 13.5 billion edges (AGATHA-C model).

In Table 1, we compare aforementioned models using the metrics described in the previous section. We present predicate types with NLM semantic type codes (McCray, Burgun, and Bodenreider 2001) due to space restrictions. *Both AGATHA-C and AGATHA-GP models show significant gains when compared to AGATHA-O baseline model*. Benefits in the most problematic for the baseline model areas

	ROC AUC			PR AUC			P.@10			P.@100			AP.@10			AP.@100		
	O	C	GP	O	C	GP	O	C	GP	O	C	GP	O	C	GP	O	C	GP
dsyn:dsyn	0.83	0.88	0.88	0.35	0.41	0.42	0.40	0.60	0.70	0.48	0.51	0.56	0.57	0.83	0.74	0.54	0.57	0.60
phsu:dsyn	0.86	0.91	0.91	0.36	0.41	0.46	0.50	0.10	0.80	0.45	0.47	0.57	0.49	0.33	0.66	0.46	0.42	0.64
fndg:dsyn	0.85	0.93	0.92	0.40	0.54	0.54	0.70	0.80	0.90	0.52	0.59	0.62	0.84	0.82	0.95	0.64	0.67	0.74
dsyn:fndg	0.81	0.89	0.90	0.32	0.43	0.45	0.30	0.80	0.60	0.45	0.48	0.50	0.50	0.96	0.57	0.49	0.60	0.57
fndg:humn	0.81	0.89	0.89	0.37	0.45	0.48	0.80	0.60	1.00	0.44	0.56	0.55	0.95	0.72	1.00	0.68	0.59	0.73
dsyn:humn	0.77	0.84	0.85	0.27	0.33	0.37	0.50	0.50	0.50	0.35	0.40	0.52	0.50	0.35	0.40	0.50	0.47	0.54
topp:dsyn	0.87	0.92	0.92	0.38	0.52	0.50	0.30	0.60	0.70	0.50	0.60	0.57	0.29	0.77	0.76	0.47	0.67	0.62
orch:dsyn	0.87	0.89	0.88	0.34	0.45	0.45	0.40	0.80	0.40	0.36	0.45	0.51	0.65	0.98	0.70	0.45	0.63	0.61
geoa:spco	0.75	0.74	0.90	0.22	0.20	0.44	0.20	0.30	0.50	0.30	0.25	0.53	0.33	0.40	0.51	0.32	0.26	0.57
aapp:dsyn	0.86	0.93	0.92	0.39	0.44	0.49	0.50	0.40	0.50	0.45	0.48	0.48	0.71	0.40	1.00	0.53	0.45	0.58
Mean	0.83	0.88	**0.90**	0.34	0.42	**0.46**	0.46	0.55	**0.66**	0.43	0.48	**0.54**	0.58	0.66	**0.73**	0.51	0.53	**0.62**

Table 1: Classification and recommendation quality metrics across recently popular COVID-19-related biomedical subdomains. Labels O, C and GP stand for AGATHA-O , AGATHA-C and AGATHA-GP models, respectively. Used abbreviations: *dsyn*: Disease or Syndrome; *topp*: Therapeutic or Preventive Procedure; *humn*: Human; *aapp*: Amino Acid, Peptide, or Protein; *phsu*: Pharmacologic Substance; *orch*: Organic Chemical; *spco*: Spatial Concept; *fndg*: Finding; *geoa*: Geographic Area.

(e.g., *(Geographic Area) → (Spatial Concept)* denoted by *(geoa,spco)*) serve the best illustration for that, showing up to 0.25 advantage in ROC AUC (AGATHA-GP). Important biomedical domains, such as *(Amino Acid, Peptide or Protein) → (Disease or Syndrome)* denoted by *(aapp,dsyn)* also show noticeable improvements (0.07 for AGATHA-C). Due to space limit we include the results for only top-10 most popular subdomains. More can be found in the long preprint version (Tyagin et al. 2021). Average ROC AUC value is increased by 0.07.

Our validation strategy involves a big number of many-to-many queries, making the area under precision-recall curve another very illustrative metric. This is where the newly proposed models show even more drastic improvements over the baseline AGATHA-O . For some subdomains, like *(Organic Chemical) → (Disease or Syndrome) (orch,dsyn)* we observe that new models improve the PR AUC score on more than 0.1. Average PR AUC value is increased by 0.12.

Emergent Discovery Case Study

The proactive discovery of ongoing research findings is an important component in the validation of hypothesis generation systems (Sybrandt, Shtutman, and Safro 2018). In particular, in the current uncertain situation with COVID-19 when a lot of unintentionally incorrect discoveries are published, the validation must include human-in-the-loop part even in limited capacity such as in (Aksenova et al. 2019; Spangler et al. 2014). To demonstrate the predictive potential of AGATHA-C , we perform a case study on three COVID-19-related novel connections manually selected by the domain expert. These connections were published after the cut date before which any data used in training was available to download at NIH.

At a low level, all AGATHA models use entity subsampling to calculate pairwise ranking criteria, which means that the absolute numbers may fluctuate slightly. Thus, to present the numeric scores, each experiment was repeated 100 times to compute the average and standard deviation that we present in Figure 3.

AGATHA-C was tested whether it would be able to predict compounds potentially applicable for the treatment of COVID-19 and the genes involved in the SARS-CoV-2 pathogenesis. The data confirming cardiovascular protective effects of hormone oxytocine were published recently (Diep 2021; Wang and Wang 2021). The protective effect is linked to anti-inflammatory activity of the hormone. AGATHA-C ranked this connection at top 1.4 percent.

Similarly, we tested the prediction of the effects of the other hormone, melatonin. Several publications, started from November 2020 (Cardinali, Brown, and Pandi-Perumal 2020; Zimmermann and Curtis 2020; Alschuler et al. 2020; Ho et al. 2021) show the protective effects of melatonin, specifically for COVID-19 neurological complications. The activity was linked to anti-oxidative effects of the melatonin. This connection was ranked at top 5.6 percent.

Our system ranked at top 7.6 percent the involvement of tetherin (BST2). The results published in 2021 (Stewart et al. 2021) show that tetherin restricts the secretion of SARS-CoV-2 viral particles and is downregulated by SARS-CoV-2. Therefore, pharmacological activation of tetherin expression, or inhibition of the degradation could be a promising direction of the development of SARS-CoV-2 treatment.

To demonstrate AGATHA-C ranking capabilities, we use similar strategy to what we proposed in the validation section, but now we randomly generate 500 negative samples for each pair of interest, maintaining the ratio of 1:500 between real-world connections and random noise. The goal of this experiment is to rank the *real* connections above randomly sampled pairs of the same semantic network types. Illustration of this experiment is presented in Figure 3. In each of the three sub-figures, there are 501 results of scores for 501 pairs. One of the results is now known to be correct, while other 500 are unknown. For example, for the first known meaningful pair COVID-19-melatonin, we generated 500 pairs COVID-19-X pairs, where X is a randomly chosen hormone. Because we are dealing with a medical domain, nobody can be 100% sure that all these 500 are irrelevant. However, most of them are not likely to be related to COVID-19. In the histograms of Fig. 3 we show a dis-

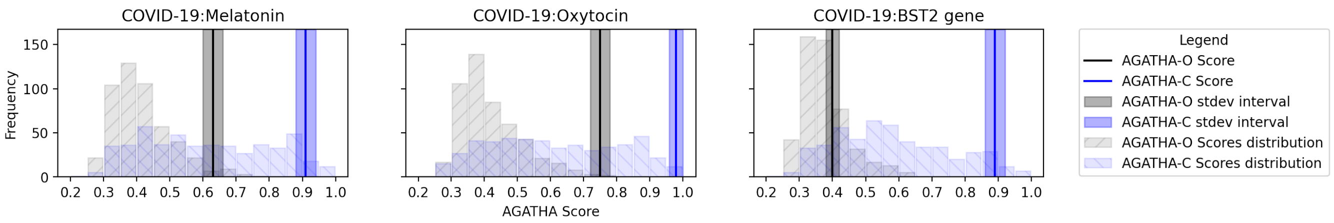

Figure 3: Score distributions in case study experiment. Presented scores are obtained with AGATHA-O and AGATHA-C models.

tribution of scores for each 501 experiments for two models: baseline AGATHA-O and newly proposed AGATHA-C . The solid lines indicate actual scores for the known pairs. To conclude, if lab experiments were required to confirm the connections, the known connections would be confirmed in the top 10 percent of predictions which significantly reduces the research time and cost.

Model	ROC AUC	PR AUC	AP.@10
C (Oct 2020)	0.88	0.49	0.67
C (June 2021)	0.90	0.55	0.77

Table 2: Comparison between AGATHA-C models trained with different cut dates and validated with recently discovered pairs.

We also test how the inclusion of more recent training data affects AGATHA performance. For that we take two models trained with the same methodology (C models), but one model contains only training data limited to October 28th 2020 and another model is larger and contains more recent data (threshold: June 23rd 2021). Both models were validated with the pairs firstly introduced between June 24th 2021 and August 11th 2021. Results of this experiment are presented in Table 2. It shows that retraining the model using the proposed method and more recent data yields in slightly better scores in all basic metrics.

Lessons Learned and Open Problems

Quality of the information retrieval pipelines. Information retrieval is an important part of any HG pipeline. In order to uncover *implicit* connections, the system should be able to capture existing *explicit* connections with as high quality as possible.

We observed that the SemRep system performs better concept and relation recognition when full abstracts are used as input data instead of single sentences. SemRep also allows to perform optional sortal anaphora resolution to extract co-references to the entities from neighbouring sentences, which was shown to be useful in (Kilicoglu et al. 2016) and is used in this work.

"Positive" research bias. The absence of published negative research results is a big problem for the HG field. With mostly positive results available, we often have to generate negative examples using some kind of random sampling. These negative samples likely do not adequately represent the real nature of negatively confirmed scientific findings. Consequently, one of the most important future work directions in the area of HG is to accurately distinguish and leverage positive and negative proposed results.

The nature of input corpora. The question of what should be used as input to a topic-modeling based hypothesis generation system is raised in (Sybrandt et al. 2018). Using full-text papers shows an improvement, but the trade-off between run time and output quality was barely justifiable. However, deep learning models have a greater potential for extracting useful information from large input sources, and as it was demonstrated in our previous work (Sybrandt et al. 2020), show significant performance advancements. Thus the question of using full-text papers in deep learning-based hypothesis generation systems should be addressed.

Knowledge resolution. Our newly proposed systems showed that the knowledge resolution plays a major role in subdomain recommendation. To increase the scope of model expertise (and the scope of potential applications beyond the biomedical fields) we deliberately incorporate a general-purpose information retrieval system RnnOIE into AGATHA-GP . Although, both systems process all types of queries, the general purpose predicates participated in training significantly improve "macroscopic" types of queries.

Predicate Extraction. One of the most important aspects of any hypothesis generation system is to give it the ability to *reject* hypothesis which are not backed by any research. This task becomes difficult when we consider the positive research bias of the existing literature. We aim to address this enhancement in our future work.

Conclusions

We present two graph mining transformer-based models AGATHA-C and AGATHA-GP , for micro- and macroscopic scales of queries respectively, which are designed to help domain experts solve high-priority research problems and accelerate scientific discovery. We perform per-subdomain validation of these new models on a rapidly changing COVID-19 focused dataset, composed of recently published concept pairs and demonstrate that the proposed models achieve state-of-the-art prediction quality. Both models significantly outperform the existing baselines. We deploy the proposed models to the broad scientific community and believe that our contribution can raise more interest in prospective hypothesis generation applications.

References

Aksenova, M.; Sybrandt, J.; Cui, B.; Sikirzhytski, V.; Ji, H.; Odhiambo, D.; Lucius, M. D.; Turner, J. R.; Broude, E.; Peña, E.; et al. 2019. Inhibition of the Dead Box RNA Helicase 3 prevents HIV-1 Tat and cocaine-induced neurotoxicity by targeting microglia activation. *Journal of Neuroimmune Pharmacology*, 1–15.

Alschuler, L.; Chiasson, A. M.; Horwitz, R.; Sternberg, E.; Crocker, R.; Weil, A.; and Maizes, V. 2020. Integrative medicine considerations for convalescence from mild-to-moderate COVID-19 disease. *Explore*.

Basu, S.; et al. 2020. ERLKG: Entity Representation Learning and Knowledge Graph based association analysis of COVID-19 through mining of unstructured biomedical corpora. In *Proceedings of the First Workshop on Scholarly Document Processing*, 127–137.

Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*, 32(Database issue): D267–270.

Cardinali, D. P.; Brown, G. M.; and Pandi-Perumal, S. R. 2020. Can Melatonin Be a Potential "Silver Bullet" in Treating COVID-19 Patients? *Diseases*, 8(4): 44.

Diep, P.-T. 2021. Is there an underlying link between COVID-19, ACE2, oxytocin and vitamin D? *Medical Hypotheses*, 146: 110360.

Gopalakrishnan, V.; Jha, K.; Jin, W.; and Zhang, A. 2019. A survey on literature based discovery approaches in biomedical domain. *Journal of biomedical informatics*, 93: 103141.

Ho, P.; et al. 2021. Perspective Adjunctive Therapies for COVID-19: Beyond Antiviral Therapy. *International Journal of Medical Sciences*, 18(2): 314.

Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; and Mikolov, T. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Kilicoglu, H.; Shin, D.; Fiszman, M.; Rosemblat, G.; and Rindflesch, T. C. 2012. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinform.*, 28(23): 3158–3160.

Kilicoglu, H.; et al. 2016. Sortal anaphora resolution to enhance relation extraction from biomedical literature. *BMC Bioinformatics*, 17: 163.

Kuusisto, F.; Steill, J.; Kuang, Z.; Thomson, J.; Page, D.; and Stewart, R. 2017. A Simple Text Mining Approach for Ranking Pairwise Associations in Biomedical Applications. *AMIA Jt Summits Transl Sci Proc*, 2017: 166–174.

Martinc, M.; et al. 2020. COVID-19 Therapy Target Discovery with Context-Aware Literature Mining. In *Discovery Science*, 109–123. Cham: Springer International Publishing. ISBN 978-3-030-61527-7.

McCray, A. T.; Burgun, A.; and Bodenreider, O. 2001. Aggregating UMLS semantic types for reducing conceptual complexity. *Stud Health Technol Inform*, 84(Pt 1): 216–220.

Nordon, G.; Koren, G.; Shalev, V.; Horvitz, E.; and Radinsky, K. 2019. Separating Wheat from Chaff: Joining Biomedical Knowledge and Patient Data for Repurposing Medications. In *AAAI*.

Oniani, D.; Jiang, G.; Liu, H.; and Shen, F. 2020. Constructing co-occurrence network embeddings to assist association extraction for COVID-19 and other coronavirus infectious diseases. *Journal of the American Medical Informatics Association*, 27(8): 1259–1267.

Smalheiser, N. R. 2017. Rediscovering Don Swanson: The past, present and future of literature-based discovery. *Journal of Data and Information Science*, 2(4): 43–64.

Spangler, S. 2015. *Accelerating Discovery: Mining Unstructured Information for Hypothesis Generation*. Chapman and Hall/CRC.

Spangler, S.; et al. 2014. Automated hypothesis generation based on mining scientific literature. In *Proceedings of the 20th ACM SIGKDD*, 1877–1886.

Stanovsky, G.; Michael, J.; Zettlemoyer, L.; and Dagan, I. 2018. Supervised Open Information Extraction. In *Proceedings of The 16th Annual Conference of NAACL HLT*.

Stewart, H.; Johansen, K. H.; McGovern, N.; Palmulli, R.; Carnell, G. W.; Heeney, J. L.; Okkenhaug, K.; Firth, A.; Peden, A. A.; and Edgar, J. R. 2021. SARS-CoV-2 spike downregulates tetherin to enhance viral spread. *bioRxiv*, 2021–01.

Sybrandt, J.; Carrabba, A.; Herzog, A.; and Safro, I. 2018. Are Abstracts Enough for Hypothesis Generation? In *2018 IEEE International Conference on Big Data*, 1504–1513.

Sybrandt, J.; Shtutman, M.; and Safro, I. 2017. MOLIERE: Automatic Biomedical Hypothesis Generation System. In *Proceedings of the 23rd ACM SIGKDD*, KDD '17, 1633–1642. New York, NY, USA: ACM. ISBN 978-1-4503-4887-4.

Sybrandt, J.; Shtutman, M.; and Safro, I. 2018. Large-Scale Validation of Hypothesis Generation Systems via Candidate Ranking. In *2018 IEEE International Conference on Big Data*, 1494–1503.

Sybrandt, J.; Tyagin, I.; Shtutman, M.; and Safro, I. 2020. *AGATHA: Automatic Graph Mining And Transformer Based Hypothesis Generation Approach*, 2757–2764. New York, NY, USA: Association for Computing Machinery. ISBN 9781450368599.

Tyagin, I.; Kulshrestha, A.; Sybrandt, J.; Matta, K.; Shtutman, M.; and Safro, I. 2021. Accelerating COVID-19 research with graph mining and transformer-based learning. arXiv:2102.07631.

Wang, H.; Ding, Y.; Tang, J.; Dong, X.; He, B.; Qiu, J.; and Wild, D. J. 2011. Finding complex biological relationships in recent PubMed articles using Bio-LDA. *PloS one*, 6(3): e17243.

Wang, S. C.; and Wang, Y.-F. 2021. Cardiovascular protective properties of oxytocin against COVID-19. *Life Sciences*, 119130.

Zhang, R.; Hristovski, D.; Schutte, D.; Kastrin, A.; Fiszman, M.; and Kilicoglu, H. 2020. Drug Repurposing for COVID-19 via Knowledge Graph Completion. arXiv:2010.09600.

Zimmermann, P.; and Curtis, N. 2020. Why is COVID-19 less severe in children? A review of the proposed mechanisms underlying the age-related difference in severity of SARS-CoV-2 infections. *Archives of Disease in Childhood*.