COVID19 Initiative: A curated network of protein-drug-gene data and virus-host interactions for drug-repurposing research

Aim and Scope

The report describes a resource released to the community by the CLAIRE-COVID19 Bioinformatics working group. The initiative joins forces from AI, clinical and life-sciences experts working on the analysis of complex and multi-sourced biomedical data integrating clinical evidence on COVID-19 with genomic and proteomic information, as well as molecular data. We are exploring data-driven AI methodologies and bioinformatics approaches covering network data analysis, machine learning, and deep learning for graphs, predictive modeling, and feature selection of Omics data. Our primary goal is to support the community with the release of resources for:

- characterising the disease from its related structural information, including prediction of viral protein folding;
- studying interactions between the virus and human hosts, including analysing protein-protein interaction data;
- filtering, retrieval, and generation of targeted drugs leveraging molecular and well as proteomic information;
- delivering predictive insights onto the genetic features of the virus.

To enable these objectives, as first task we are assembling a resource that fuses information from heterogeneous sources and different studies from the literature into a unique network-based representation, facilitating the use of relational and graph-based learning methods. The document is organised as follows. First, we present the list of the identified resources, providing for each one of them the descriptions, the license and the reference source. Then we provide details on the methodology adopted to reduce the numerosity and the specificity of the GO Terms. Lastly, we give details to easily access the collected resources. We conclude, by giving information about the participants of this subtask.

Identified Resources

ID	Used Information	Source	License
[R1]	Protein-Protein	Network-based prediction of drug combinations [4]	CC BY 4.0
[R2]	Protein domains	Uniprot [5]	CC BY 4.0
[R3]	Protein families	Uniprot [5]	CC BY 4.0
[R4]	Protein pathways	Reactome [8]	CC BY 4.0
[R5]	GO-Terms	Gene Ontology (GO) [1, 15]	CC BY 4.0
[R6]	Drug-Protein	Network-based prediction of drug combinations [4]	CC BY 4.0
[R7]	Drug Structures	DrugBank [17]	CC BY-NC 4.0
[R8]	Drug-Drug	Network-based prediction of drug combinations [4]	CC BY 4.0
[R9]	Disease-Gene	DisGeNET [14]	CC BY 4.0
[R10]	Virus-Host interactions	BioGrid [13]	MIT

Table 1: License and references of the identified resources.

Hereafter we explain the rationale and the motivations that had driven us to select the resources listed in table 1.

Protein-Protein [R1]: Protein-Protein Interactions (PPIs) are physical interaction between two or more proteins. A collection of PPIs, namely PPI network or, is called more broadly interactome. A relevant finding of the interactome is that proteins involved in the same processes can cluster together in the network. Protein-protein interactions are important because it allows us to understand a protein's function and its behavior. Actually, only a small portion of the human protein-protein interactions are studied by in lab experiments. The host-host interactive collected here consists of 217.161 interactions among the 15.970 human proteins.

Domains [R2]: Domains are distinct functional and/or structural units in a protein. Usually, they are responsible for a particular function or interaction, contributing to the overall role of a protein. Domains may exist in a variety of biological contexts, where similar domains can be found in proteins with different functions. We collected from the Uniprot dataset [5], 15.648 genes-domains associations.

Families [R3]: A protein family is a group of proteins that share a common evolutionary origin. This origin is reflected by their functions, and thus is possible to notice similarities in their sequence or structure. We collected from the Uniprot dataset [5], 15.648 genes-families associations.

Pathways [R4]: A biological pathway is an ordered series of molecular events occurring among molecules in a cell, and that leads to producing a certain biological product, or change in the involved cell. We retrieved from [8] a total of 15.648 genes-pathways associations.

GO-Terms [R5]: Go-Terms are biological terms, or concepts, related to the genes. There exist three biological ontologies of GO-Terms [1]: *Biological process*, where terms represent a series of molecular events or functions; *Molecular function*, contains activities performed by individual gene products at the molecular level; *Cellular component*, describes the parts of the cell and the extracellular environment in which a gene product may be localized. We collected from [1] 330.519 relations between genes and GO-Terms. Moreover, we designed a methodology (see able to reduce the high numerosity and specificity of the GO-Terms. The result of the aforementioned process is readily available with the other resources.

Drug-Host [R6]: A drug is designed to produce a specific desirable therapeutic effect on the target organism. The relation between a drug and the target molecules of the organism, usually a protein, is named drug-target association or interaction. We provide 15.052 drug-host interactions yielded by 4.428 drugs, as reported in [4].

Drug Structures [R7]: Drug structures provide information about the topological structure of the drug molecules, such as spatial coordinates of the atoms and their bonds. We collect 10.674 different drug structures[1] with their unique DrugBank Identifier, taken from the open dataset [17].

Drug-Drug [R8]: A Drug-Drug Interaction (DDI) is an alteration of the drug's expected effect - on the target organism - if administered with another drug product. Knowing whether a DDI produces a therapeutic or an adverse effect on the target organism is of paramount importance to repurpose multiple drugs together. In this resource, we collected 14.079 drug-drug interactions taken from [4].

Disease-Gene [R9]: According to [2], in the molecular network context «a disease is rarely a consequence of an abnormality in a single gene, but reflects the disruptions of the complex intracellular network». Following this perspective, a gene or a gene product (e.g. protein) can be linked to a disease (i.e. disease gene). We provide a collection of $1.134.942$ disease-gene associations gathered through DisGeNET [14] and 299 gene-disease associations taken from [11].

Virus-Host [R10]: The virus-host interaction represents a physical interaction between a virus molecule and a host (e.g., human) protein. We collected the information about virus-virus, virus-host, and human-human interactions related to 20 different viruses. For each virus, the number of interactions can range from 2 to 2060. Note that virus-host interactions could be seen as disease-gene interactions where the virus is considered as a disease. Moreover, many Virus-Host interactions could be included in the disease-gene associations' collection [R9], but we chose to leave them directly available on their own.

Methodology to process the GO-Terms

The high number ($\simeq$ 74700 terms divided into three : Cellular Component, Molecular Function and Biological Process) and specificity (each protein/gene product is associated on average to 16 ± 17 terms) of GO-Terms [R5] lead to two main drawbacks: *i)* the high numerosity will produce high demand for memory and computational resources when neural network methods will be used; *ii)* the high specificity will lead

[1] 4.283 of which representing drugs from [R6].

to learning a smaller amount of patterns and thus can affect the performances of the employed methods. In order to solve such drawbacks, we decided to preprocess the GO-Terms to reduce their numerosity without losing their intrinsic semantics. Thus we cluster those terms that expose similar biological functionalities. Since the GO-Terms are complemented by a textual description and they are also organized in a taxonomy (DAG), we can apply the following approaches:

- *Handcrafted selection:* Domain experts can select and group together those terms that they consider to be highly correlated, or they can decide to edit the original taxonomy in order to produce a reduced version of it.
- *Text based:* Textual information, as the GO term description of what the term represents in its biological context, attached to each GO-term can be employed to produce a dense representation [10] used to cluster together them [9].
- *Connectivity based:* The connectivity exposed by the taxonomy can be used to build a dense representation [7] lately employed to cluster together with the GO-terms.

The *handcrafted selection*[2] can be biased, and it is a time-consuming task based on specific knowledge. *Text based* clustering looks to be a promising solution to our problem but suffers from the drawback to not incorporate the connectivity information (as a result of the aforementioned human-curated process) contained in the taxonomy. Lastly, we have noted that Gene Ontology is a human-curated resource based on expert knowledge, and also the textual information of GO reflects the same domain experts' knowledge. Our intuition, then, is that all this information is already hiddenly encoded in the taxonomy, and thus we decided to apply the *connectivity based* approach as the more complete and efficient.

To implement the connectivity-based approach, we decide to use the Node2Vec [7] algorithm. Node2Vec learns the embedding of a given node by performing a certain number of random walks starting from the node of interest (the *target* node). Each walk results in a sequence of visited nodes (or *context* nodes). These node sequences are then used to learn node embeddings by optimizing a skip-gram objective [12]. Node2Vec has 4 main hyper-parameters: the number of random walks r to take for each node, length ℓ of the random walk, a so-called "return" parameter p, which specifies the probability of returning to an already visited node, and an "in-out" parameter q, which controls the probability of visiting nodes farther away from the target node. Other hyper-parameters are specific to the skip-gram objective, such as context window size and the number of Stochastic Gradient Descent (SGD) iterations.

Our choice has been to use three different Directed Acyclic Graphs (DAGs) to represent the three GO namespaces. We applied Node2Vec independently on each DAG. Importantly, we did not consider edge orientation but treat the DAG as an undirected graph instead, so to capture relationships between nodes in both directions. For our purposes, we set the hyper-parameters as follows: r has been set to 10 times the maximum node degree of the Gene Ontology DAG; ℓ has been set to the diameter of the graph; p and q have been assigned the default value of 1. As regards the skip-gram hyper-parameters, we set the context window size to 7 and the number of SGD iterations to 5.

An example of the computed embeddings is depicted in Figure 1, where the original space of 128 dimensions was reduced (just for viewing purpose) to 2 using the T-distributed Stochastic Neighbor Embedding (t-SNE) algorithm [16]. As it is possible to notice the GO-Terms (using the generated embeddings) shows a natural tendency to be organized in clusters (see Figure 1), and by following the intrinsic nature of the Gene Ontology they expose variable local densities.

In this case, our choice has been to use the Hierarchical Density-Based Spatial Clustering (HDBSCAN)[3], an evolution of DBSCAN[6] algorithm, where several ϵ values are integrated to find a clustering that gives the best stability. The above integration strategy allows HDBSCAN to find clusters of varying densities, and be more robust to parameter selection, that it is exactly the result that we want to obtain. One characteristic of HDBSCAN is that not all the instances (GO-Terms in our case) are clustered, but part of them are considered noise and then grouped all together. A researcher that uses this resource can then decide to recover the "noisy" data points by adding them as singleton or keep ignoring them. The clustering obtained by applying HDBSCAN, with a Euclidean distance measure, have produced $276, 85, 781$ clusters and $6709, 2887, 23515$ noisy terms, for the namespaces of molecular function, cellular component, and biological process respectively. Both, the produced GO embeddings and their clustering are available in the repository (see Table 2 for their exact location).

[2] if is not performed by several domain experts

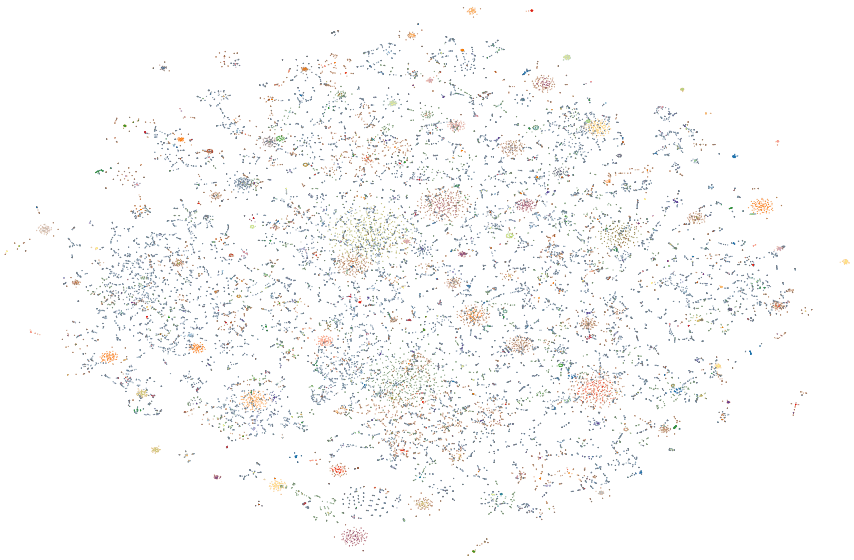

Figure 1: Bi-dimensional representation of the embeddings obtained from the Cellular Components GO-Terms .

How to access the resources

The online repository[3] is organised in order to directly reflect the information and nomenclature in this document. More specifically, Table 2 reports the relative locations of the resources in the online repository:

Resource	Location	Resource	Location
[R1] **Protein-Protein**	protein-protein.tab	[R6] **Drug-Host**	drug-host.tab, drug-host_uniprot.tab
[R2] **Domains**	proteins-info.tab	[R7] **Drug Structures**	drug-structures.sdf
[R3] **Families**	proteins-info.tab	[R8] **Drug-Drug**	drug-drug.tab
[R4] **Pathways**	proteins-pathways.tab	[R9] **Disease-Gene**	disease-gene/
[R5] **GO-Terms**	go-terms.tab	[R10] **Virus-Host**	virus-host/
GO Embeddings	GO_terms/*_emb_128.txt	**GO Clutering**	GO_terms/*_cluster_*.txt

Table 2: Pointers of the identified resources in the repository.

For a detailed description of data and formats look at the README.md file.

Contributors to the Resource

Davide Bacciu is Assistant Professor at the Department of Computer Science, University of Pisa. His research spans several fundamental and applied aspects of machine learning, including the design of neural and generative learning models, graph and relational data processing, and distributed learning systems with applications to IoT, smart living, transportation, robotics and health. He has been the coordinator of European H2020, Italian national and industrial research projects. He received the 2009 Caianiello Award for the best Italian Ph.D. thesis in neural networks. He is Secretary and board member of the Italian Association for AI, a member of the IEEE Neural Networks committee and the chair of the IEEE task force on Learning for Structured Data. He coordinates the CLAIRE COVID19 working group on Bioinformatics. Available at: bacciu@di.unipi.it

Federico Errica is a Ph.D. student at the University of Pisa, Italy. He received his Bachelor and Master's degrees in Computer Science in 2015 and 2018, respectively. His interests are machine learning for structured

[3] https://github.com/CLAIRE-COVID-T4/covid-data

data with particular emphasis on probabilistic and neural models for graphs. In 2019 he was a research intern at Facebook AI London. He has published at top-tier conferences (ICML, ICLR), and he currently volunteers as a reviewer for top-tier journals (TNNLS, TPAMI). Available at: federico.errica@phd.unipi.it

Alessio Gravina received his Bachelor and Master of Science in Computer Science from University of Pisa, Italy, in 2018 and 2020, respectively. His interests are related to the area of Machine Learning, and he has experience in its application to the biomedical domain. In 2018 he was one of the three winners of the Fujistu AI-NLP Challenge, while in 2019 he was a visiting student at University College Dublin (UCD). In the same year, he was a visiting student researcher at Stanford University researching, in collaboration with SPARK, on graph learning for Schizophrenia treatment. Available at: gravina.alessio@gmail.com

Francesco Landolfi is a Ph.D. student at the University of Pisa, Italy. He received his Master's degree in Computer Science in 2019. His research interests are machine learning and graph-theoretical approaches for geometric deep learning. Available at: francesco.landolfi@phd.unipi.it

Lorenzo Madeddu is a PhD student at the Department of Translational and Precision Medicine at Sapienza University of Rome with a Computer Science Master Degree. His research interests focus on machine learning, graph mining and Network Medicine. He is involved in interdisciplinary projects in the fields of Healthcare and Precision Medicine and is supported by the "Sapienza information-based Technology InnovaTion Center for Health - STITCH". He received his master degree in Computer Science from the Sapienza University of Rome in 2018. Available at: madeddu@di.uniroma1.it

Marco Podda is a Ph.D. student at the University of Pisa, Italy. He received his Master's degree in Computer Science in 2017. His research interests are machine learning and deep learning for graph data, with emphasis on deep generative models. His research finds application in the bio-medical field. Available at: marco.podda@di.unipi.it

Giovanni Stilo is an Assistant Professor in the Department of Information Engineering, Computer Science and Mathematics at the University of L'Aquila. He received his Ph.D. in Computer Science in 2013, and in 2014 he was a visiting researcher at Yahoo! Labs in Barcelona. Between 2015 and 2018, he was a researcher in the Computer Science Department at La Sapienza University, in Rome. His research interests are in the areas of machine learning and data mining, and specifically temporal mining, social network analysis, network medicine, semantics-aware recommender systems, and anomaly detection. He has organised several international workshops, held in conjunction with top-tier conferences (ICDM, CIKM, and ECIR), and he is involved as editor and reviewer of top-tier journals, such as TITS, TKDE, DMKD, AI, KAIS, and AIIM. His research is supported by the MIUR under grant "Dipartimenti di eccellenza 2018-2022" of the Department of Computer Science of Sapienza University, by the "Sapienza information-based Technology InnovaTion Center for Health - STITCH" and by the "Territori Aperti project" funded by "Fondo Territori Lavoro e Conoscenza CGIL, CSIL and UIL". Available at: giovanni.stilo@univaq.it

References

[1] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[2] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: A network-based approach to human disease. *Nature reviews. Genetics*, 12:56–68, 01 2011. doi:10.1038/nrg2918.

[3] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[4] Feixiong Cheng, István A Kovács, and Albert-László Barabási. Network-based prediction of drug combinations. *Nature communications*, 10(1):1–11, 2019.

[5] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 11 2018. URL: https://www.uniprot.org/, doi:10.1093/nar/gky1049.

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996.

[7] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864, 2016.

[8] Bijay Jassal, Lisa Matthews, Guilherme Viteri, Chuqiao Gong, Pascual Lorente, Antonio Fabregat, Konstantinos Sidiropoulos, Justin Cook, Marc Gillespie, Robin Haw, Fred Loney, Bruce May, Marija Milacic, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Research*, 48(D1):D498–D503, 11 2019. URL: https://reactome.org/, doi:10.1093/nar/gkz1031.

[9] Jey Han Lau and Timothy Baldwin. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 78–86, Berlin, Germany, August 2016. Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/W16-1609, doi:10.18653/v1/W16-1609.

[10] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

[11] Jörg Menche, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiassian, Marc Vidal, Joseph Loscalzo, and Albert-László Barabási. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224):1257601, 2015.

[12] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013. URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

[13] Rose Oughtred, Chris Stark, Bobby-Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, Nadine Kolas, Lara O'Donnell, Genie Leung, Rochelle McAdam, Frederick Zhang, Sonam Dolma, Andrew Willems, Jasmin Coulombe-Huntington, Andrew Chatr-aryamontri, Kara Dolinski, and Mike Tyers. The BioGRID interaction database: 2019 update. *Nucleic Acids Research*, 47(D1):D529–D541, 11 2018. URL: https://thebiogrid.org, doi:10.1093/nar/gky1079.

[14] Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I Furlong. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1):D845–D855, 11 2019. URL: https://www.disgenet.org/, doi:10.1093/nar/gkz1021.

[15] The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338, 11 2018. URL: http://geneontology.org/, doi:10.1093/nar/gky1055.

[16] L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[17] David S. Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, 36(suppl_1):D901–D906, 11 2007. URL: https://www.drugbank.ca/, doi:10.1093/nar/gkm958.