A biomedical open knowledge network harnesses the power of AI to understand deep human biology

Sergio E. Baranzini[1,8], Katy Börner[2], John Morris[3], Charlotte A. Nelson[1], Karthik Soman[1], Erica Schleimer[1], Michael Keiser[3,4], Mark Musen[5], Roger Pearce[6], Tahsin Reza[6], Brett Smith[7], Bruce W. Herr II[2], Boris Oskotsky[8], Angela Rizk-Jackson[8], Katherine P. Rankin[1,8], Stephan J Sanders [8,9], Riley Bove[1,8], Peter W Rose[10], Sharat Israni[8], Sui Huang[7]

Affiliations:

1. Weill Institute for Neurosciences. Department of Neurology. University of California San Francisco. San Francisco, CA.
2. Department of Intelligent Systems Engineering. Indiana University. Bloomington, IN.
3. Department of Pharmaceutical Chemistry. University of California San Francisco. San Francisco, CA.
4. Institute for Neurodegenerative Diseases. University of California San Francisco. San Francisco, CA.
5. Department of Medicine (Biomedical Informatics) and of Biomedical Data Science. Stanford University School of Medicine. Stanford, CA.
6. Center for Applied Scientific Computing (CASC). Lawrence Livermore National Laboratory. Livermore, CA.
7. Institute for Systems Biology. Seattle, WA.
8. Bakar Institute for Computational Health Sciences. University of California San Francisco. San Francisco, CA.
9. Weill Institute for Neurosciences. Department of Psychiatry and Behavioral Sciences. University of California San Francisco. San Francisco, CA.
10. San Diego Supercomputer Center. University of California San Diego. La Jolla, CA.

Corresponding Author: Sergio E. Baranzini, PhD
Sergio.Baranzini@ucsf.edu
ORCID: 0000-0003-0067-194X

ACKNOWLEDGEMENTS

The development of SPOKE and its applications are being funded by grants from the National Science Foundation (NSF_2033569), NIH/NCATS (NIH_NOA_1OT2TR003450), and the Marcus Program in Precision Medicine Innovation. SEB holds the Heidrich Family and Friends Endowed Chair of Neurology at UCSF. SEB holds the Distinguished Professorship in Neurology I at UCSF.

Keywords: Knowledge graph; biomedical databases; electronic health record; drug development

■ Abstract.

Knowledge representation and reasoning (KR&R) has been successfully implemented in many fields to enable computers to solve complex problems with AI methods. However, its application to biomedicine has been lagging in part due to the daunting complexity of molecular and cellular pathways that govern human physiology and pathology. In this article we describe concrete uses of SPOKE, an open knowledge network that connects curated information from 37 specialized and human-curated databases into a single property graph, with 3 million nodes and 15 million edges to date. Applications discussed in this article include drug discovery, COVID-19 research and chronic disease diagnosis and management.

I. BACKGROUND.

Advanced machine learning (ML) has successfully been deployed for a wide range of applications. However, such ML have seen far less success in "semantically rich domains" such as biomedical sciences, where specification of knowledge is more abstract and fluid than that in other hard sciences. According to Herbert Simon, one of the founding fathers of AI, these unique domains typically lack mechanistic rules, and the complexity of the heterogeneous and deep human domain expertise cannot be statistically aggregated [1]. Big Data must be converted into Big Knowledge if we are to harness the data revolution and KR&R represents a timely and exciting avenue to achieve this goal. KR&R, a field of AI, includes work that strives to emulate human learning by creating a cognitive network of semantically related concepts on which context and previous experience determine the emergence of knowledge. [2] Early efforts to develop advanced data management systems included EBI's SRS server [3] and Kleisli,[4] somewhat anticipating the data (and information) deluge that would follow in subsequent years, and clearly highlighting the need for additional efforts to address this need.

Health care costs make up almost one-fifth of the entire U.S. GDP and affect every U.S. citizen. The opportunity--indeed, the imperative--to tap into the wisdom latent in Big Data can no longer be overlooked. The 'one-size-fits-all' approach is a major reason for patient treatment failures and costs. However, the biomedical public data and factual knowledge repositories are physically, technically, and thematically compartmentalized, posing a significant challenge when attempting to connect the dots across the domains of specialization in biomedicine.

Under the aegis of an NSF Convergence Accelerator award (Track A), we have developed concrete applications for our Biomedical Open Knowledge Network (OKN), named the Scalable PrecisiOn Medicine Knowledge Engine (SPOKE) following the hypothesis that connecting relevant information will enable the emergence of knowledge, and facilitate solutions to otherwise unattainable insights in understanding diseases, discovering drugs, and proactively improving personal health. Finally, by studying how human experts use SPOKE, we take a step towards a next generation of AI based on big knowledge, stepping beyond deep learning on data.[5]

II. GRAPH CONSTRUCTION AND CONTENT

SPOKE is a property graph containing more than 3 million nodes (of 21 types) and more than 15 million edges (of 55 types) (A detailed description of SPOKE architecture is in preparation at the time of this writing and will be published elsewhere). The OKN has so far integrated 37 data sources, listed at https://spoke.ucsf.edu/data-tools. Much of this data is composed of genomic associations with disease, chemical compounds and their binding targets, and metabolic reactions from select bacterial organisms of relevance to human health. Also included are perturbagen-gene, food-chemical and protein-celltype relationships (Figure 1). Several of the key concepts are mapped to biomedical ontologies (including Disease, molecular pathways, and taxonomy among others), to provide an organizational framework and facilitate user navigation. All ontologies in SPOKE were incorporated from NCBO's BioPortal repository, which contains more than 900 controlled vocabularies spanning various aspects of biomedicine. [6,

7]SPOKE also uses ontologies to mark up the datasets coming into the knowledge graph for consistent linking. SPOKE also strives to align with Biolink, a biomedical semantic standard currently being established by the NIH/NCATS Biomedical Translator Consortium. [8]

As a stated aim of our present NSF-CA proposal, over time we will continue to grow SPOKE by the integration of hundreds of data sources in the public domain including those from EPA, CDC, DHSS and the FDA.

Of note, to enhance its relevance to human health, SPOKE focuses on experimentally determined information. Thus, computational predictions and literature curation are not currently prioritized in SPOKE.

Some of the specific areas in which this NSF award focuses include:

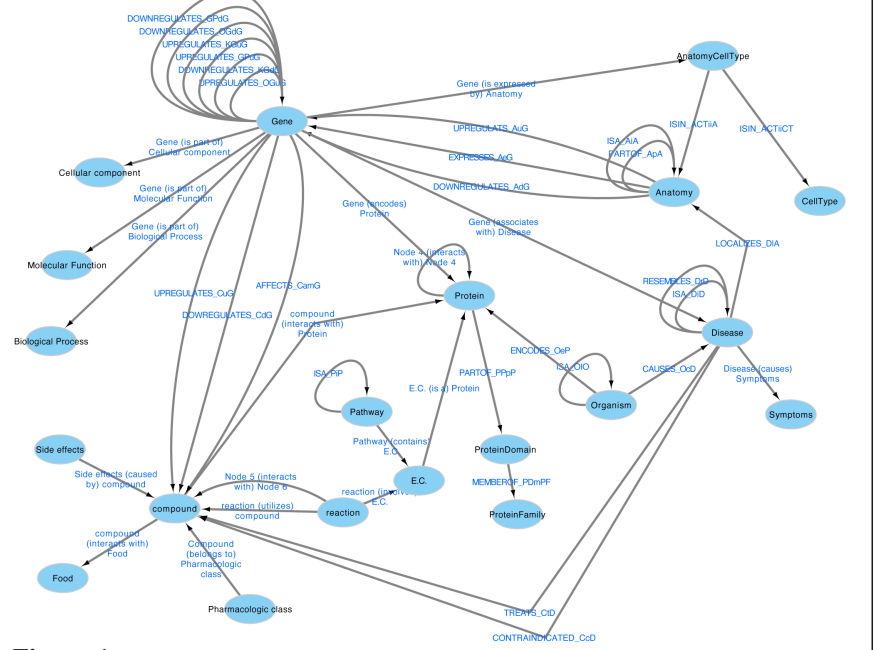

Figure 1. SPOKE Metagraph. Nodes denote biological concepts and links show how data is related and connected in the graph.

Proteins, by domain and including their 3-dimensional shapes – to answer questions such as potential targets of a drug that cause side effects, or how can an existing drug be repurposed for new indications, or whether a protein target involved in a specific disease is suitable for drug discovery (i.e., druggable).

Drug Discovery capabilities, such as adverse drug effects, drug-drug interactions, over a billion small-molecule compounds that are readily available by make-on-demand vendors and interactions between drugs and proteins, - a rich source of information for drug repurposing.

Geospatial measurement data, to bring in socio-demographic, economic, and environmental factors in health and disease.

Users can interact with the data remotely and build applications powered by the graph either interactively via Cypher queries or programmatically via one of the REST Application Programming Interfaces (APIs).

II.1. Scientific evaluation and stress-testing of the biomedical OKN

As of this writing, the network structure and balance of SPOKE has been characterized and preserved via a series of computationally intense graph-theoretical "knowledge mining" methods, including shortest path algorithm function, motif discoveries, and metabolic cycle discovery.

II.1.1. Scientific Validation – The Road Ahead

In order for SPOKE to be the basis of further scientific inquiry or new products, a series of “stress-tests” simulating real world utility need to be conducted. While anecdotal accounts of successful drug discovery guided by smaller knowledge networks reveal the potential utility of biomedical OKNs, the very concept of biomedical OKNs still must be subject to a systematic, scientific evaluation. [9] As SPOKE continues to

grow, evaluation will take place both at the structure level of the knowledge network as well as by benchmarking specific queries and use cases against medical reality.

II.1.2. Empirical relationships between Graph node concepts and paths

In addition to the generic graph-theoretical analysis (e.g. centrality, degree, etc.), we tested the utility of the specific node content using empirical data. For a set *S* of *N* concepts represented by nodes in SPOKE (e.g., "blood glucose", "gene variant X", "protein Y"), we asked whether their values measured in real life exhibits a statistical relationship to a particular structure of the subgraph in SPOKE spanned by these nodes in *S*. In the simplest case of sets of *N*=2 nodes we ask: "Are two blood metabolites observed to be highly correlated in a cohort, on average connected by a shorter path in the graph than any random pair of nodes?" (Figure 2)

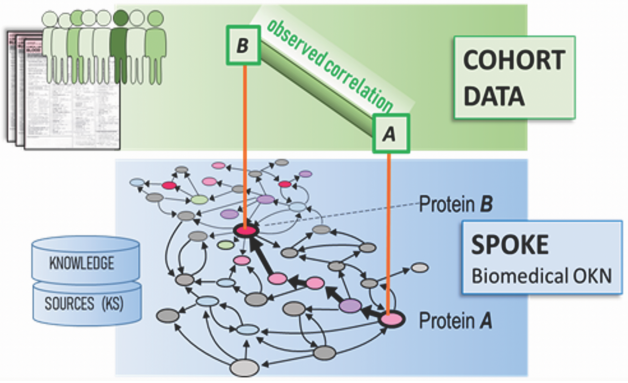

Figure 2. Mapping paths in SPOKE to empirical observations in patients

To address this question, we took advantage of a recent wellness study that collects "**multi-omics**" data in a cohort 108 healthy individuals, in which thousands of omics-data points (genomics, blood proteomics, metabolomics, clinical phenotype) were measured. [10]In this study thousands of blood analytes (abundance of circulating proteins or metabolites) were measured for each individual. In total, 8,888 pairs of these variables were found to be correlated with high statistical significance (r^2>0.9) [10]We next mapped these correlated proteins or metabolites onto nodes in the SPOKE OKN and found that, remarkably, they were connected by a path that was significantly shorter than that connecting two random nodes of the same type (Figure 3). This result offers the first empirical evidence that the graph structure of the SPOKE network that was computationally assembled from diverse biomedical medical databases preserves meaningful information about mechanistic pathways that traverse various domains, most of them never explicitly mentioned in the literature.

Based on our preliminary data, we argue that SPOKE use-cases themselves serve as stress tests; we illustrate some such AI applications below.

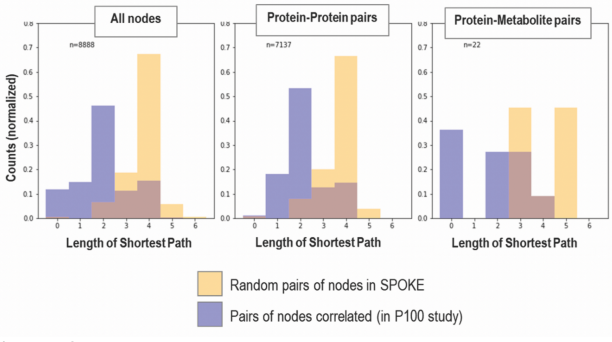

Figure 3. Analysis of blood proteomics and metabolomic data in healthy participants shows that pairs of blood analytes (protein or metabolite levels in circulation) that are correlated are connected by on

II.2. Network visualization

A complex knowledge network like SPOKE can be visualized through the Neighborhood Explorer Tool[11] to support interactive exploration by experts and citizen scientists in support of knowledge exploration (e.g., to support basic research), optimization (e.g., to resolve data problems), and communication (e.g. to better inform patients and physicians).

While standard network visualizations of large real-world networks often resemble "hairballs" that provide little actionable insight, these interactive, multi-level SPOKE visualizations compute and display clusters of related nodes and backbones between major nodes at each level of detail. [12]These additional visualizations (now under construction) resemble geospatial maps at mid-fidelity resolutions (Figure 4) with continents of similar nodes and real paths

(backbones) for each level, similar to geographic maps that show real cities and real roads at every level of detail.

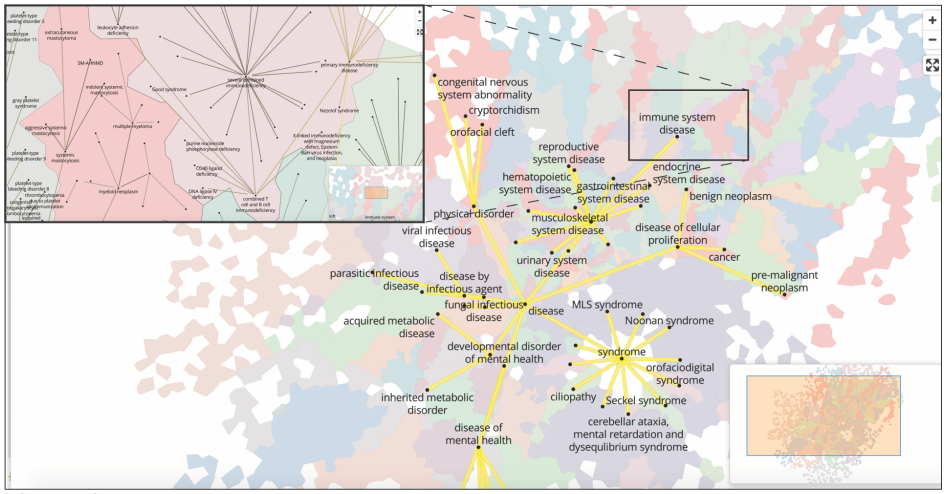

Figure 4: Initial rendering of a subgraph of SPOKE using a multi-level, map-like network visualization. Diseases are denoted in the top layer and they cluster by symptom and genetic similarity. The inset shows how additional details appear when zooming over an area (e.g., zooming on Immune system disease uncovers more details about additional diseases that belong to that category).

II.3. Knowledge graph analysis

For the contemporary biomedical researcher, in need of accessing vast amounts of trusted information, SPOKE provides the **Neighborhood Explorer** (NE, Figure 5). For example, ClinicalTrials.gov links diseases with drugs; the GWAS Catalog contains genetic associations for thousands of phenotypes and diseases; and ChEMBL contains binding information of pharmacological compounds to their protein targets. However, if an investigator seeks to identify all existing (approved and non-approved) drugs that target proteins encoded by genes containing SNPs associated with a given disease (to repurpose drugs for rare genetic disease, for instance), this will involve cumbersome manual search in a number of pertinent databases separately. Furthermore, serial queries for a group of diseases or drugs would require repeated and complicated programmatic queries in various databases and assembling the results. NE solves this need. In the future, a robust, well-supported commercial product, powered by SPOKE, with a superior UI and performance, will enable investigators to perform smart queries and return actionable information, either for hypothesis generation or to inform concrete experimental approaches.

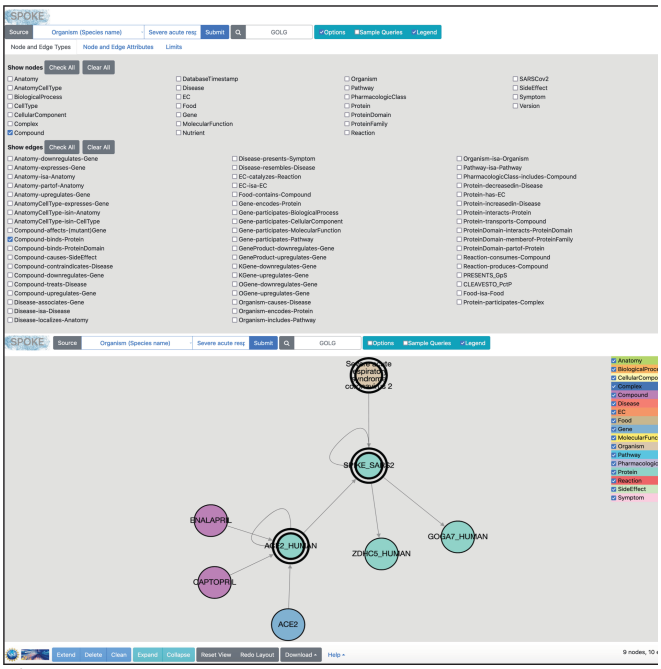

Figure 5. A view of the SPOKE Neighborhood explorer. The top panel shows the controls that allow a user to select nodes/edges for expansion as well as other key parameters. The bottom panel shows an example of the graph neighbors of the SARS-CoV-2 Spike protein (light blue), which includes 3 human proteins (green) and the genes encoding them (blue). One such protein (ACE2_HUMAN) has edges connecting it to 3 compounds (2 of them approved and one -ORE-100- in experimental phase).

III. FROM KNOWLEDGE TO INSIGHTS: AI APPLICATIONS

We envision a vast and integrated knowledge network connecting up to hundreds of millions of biomedical facts, with potential utility in a broad diversity of practical applications for specialists and informed general public alike. Its value is best harnessed by apps that are designed to extract useful information (e.g., mine the OKN) for specific applications.

SPOKE was used to predict a possible treatment to reduce mortality of COVID-19 patients placed on mechanical ventilation. [13]We constructed a chain of causation, a path in the SPOKE network that connects the ACE2 protein, the cell surface protein used by the SARS-CoV-2 virus to enter the host, to the use of Dexamethasone (a corticosteroid). SPOKE exposed a pharmacological connection that no literature or Google search would have unearthed: Through the analysis of gene expression profiles, we discovered that mechanical tissue stress caused by ventilation caused upregulation of ACE2 (Figure 6) and that dexamethasone suppresses the tissue hormone midkine (MK), that is critically involved in transducing mechanical stress to further upregulation of ACE2. Therefore, there exists a vicious cycle: mechanical ventilation used to combat respiratory distress caused by the virus would itself also facilitate the spread of the virus in the lungs. These results suggest that administration of corticosteroids, which was debated in the early days of the pandemic, could improve outcome of severe (i.e., ventilated) COVID-19 cases. Indeed, clinical studies have since reported that corticosteroids reduced the mortality of ICU specifically for patients on ventilators by 30%. [14, 15] Here SPOKE, allowing seamless search across domains of knowledge, showed its unique power in "connecting the dots", alleviating the core problem of "database selection" in complex disciplines with countless specialties.

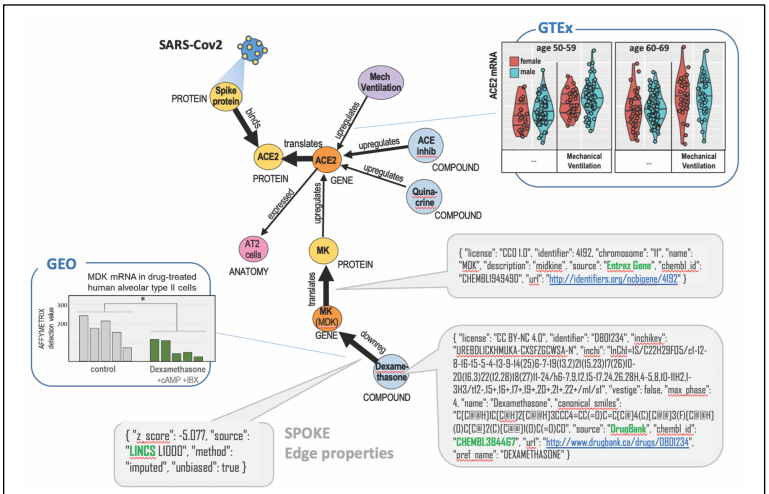

Figure 6. SPOKE-enabled reconstruction of the hypothesis that dexamethasone might help recovery of patients with COVID-19. Multiple sources of evidence were required to formulate this scenario without human intervention.

Another example of "connecting dots" is provided by integrating the role of bradykinin in COVID-19. Again, the entry point for the virus is ACE2, which has a direct connection to the bradykinin receptor BRKB2, and hence to its protein BKRB1_HUMAN, which represents the intersection between endocrine and immune regulation systems. This triggers proteolysis of the KNG1_HUMAN protein, which gets cleaved into kininogen. Kininogen has a large number of connections and effects, one of which is bradykinins, which have a potent vasopressor activity[16]. Thus, elevated bradykinin levels likely cause increases in vascular dilation, vascular permeability and hypotension, all features observed in severe COVID-19 patients.

III.1. Repurposing pharmaceutical drugs

Pharmaceutical and Biotechnology drug development is an expensive endeavor, and some estimates put the current cost of a new drug at $2.6 billion.[17] Only one for every 20 products that enter Phase I clinical trials ever becomes a commercialized product; fully 50% fail in the costly, last stage of clinical trials - or fail to meet the proposed clinical endpoints on a significant part of the patient population.

SPOKE shows promise in repurposing existing drugs or discovering new therapeutic applications for them. Its predecessor, HetioNet, was stress-tested to find concrete examples of drug repurposing, in two *retrospective* studies:

A) Bupropion, first approved for depression in 1985, was approved for smoking cessation in 1997. [18] Predictions based on SPOKE clearly highlight this new indication. [19]

B) SPOKE evaluated the top 100 scoring compounds for epilepsy seizure control, successfully classifying 77 compounds as anti-ictogenic (seizure suppressing), 8 as unknown (no established effect on the seizure threshold), and 15 as orictogenic (seizure generating). Notably, the predictions contained 23 of the 25 disease-modifying antiepileptics in PharmacotherapyDB v1.0. [19]

The therapeutic effect at genomic, metabolomic, proteomic, physiological or toxicological level may help identify additional uses for an existing drug. SPOKE can also determine ideal patient profiles and population targets for new therapeutic drugs prior to entering late-stage clinical trials.

III.2. Predicting new chemical biology from a small molecule's OKN neighborhood

In another planned application, we plan to encode a small molecule's OKN-derived *biological context* instead of its raw chemical structure, into an "OKN fingerprint." Such small molecules are "drug-like compounds." Similar structures have been observed to exhibit similar bioactivities across a standardized panel of wet-lab assays, and this phenomenon can be exploited to identify new drugs with desired activities. Too little information exists to construct experimentally derived fingerprints, and hence computational predictions of such fingerprints have been proposed.[20]

III.3. Delivering SPOKE to the clinician: BRIDGE

For clinicians to be able to ingest the ever-expanding volumes and types of information available for their patients, data and algorithms such as those enabled by SPOKE must be delivered in a clear, actionable format that is workflow friendly and will enable them to respond adequately (and in real-time) to complex scenarios to optimize patient outcomes. BRIDGE is a platform that launches directly from a patient's chart in the EHR, and assembles relevant clinical, laboratory, imaging, and patient-generated data to visualize an individual's trajectory

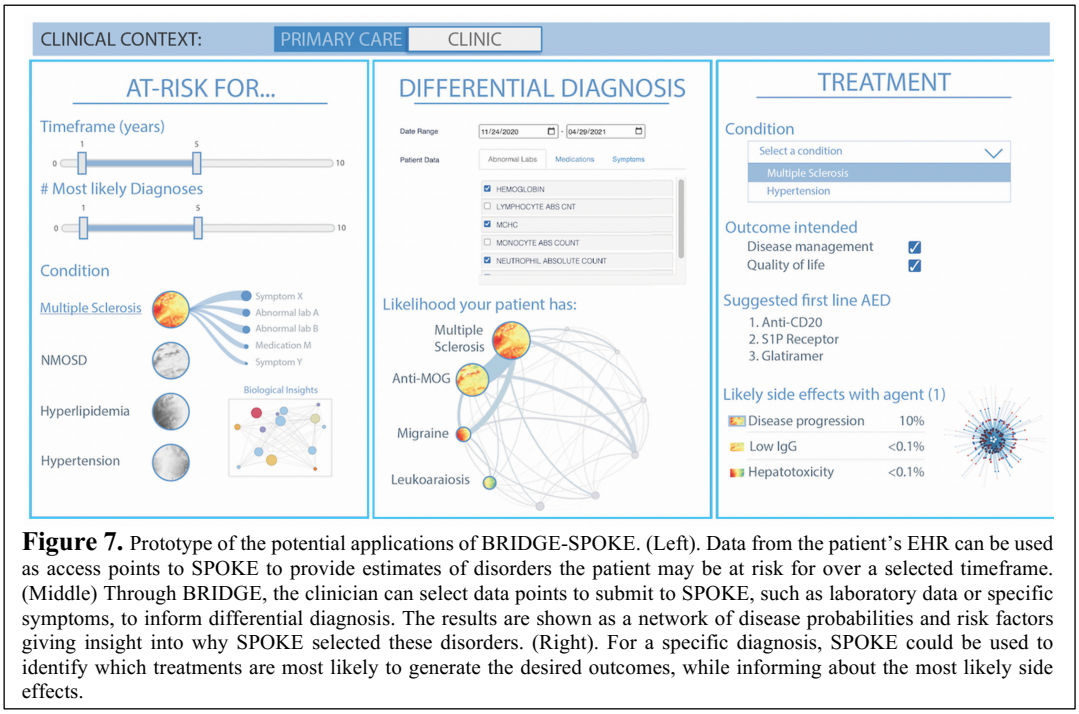

Figure 7. Prototype of the potential applications of BRIDGE-SPOKE. (Left). Data from the patient's EHR can be used as access points to SPOKE to provide estimates of disorders the patient may be at risk for over a selected timeframe. (Middle) Through BRIDGE, the clinician can select data points to submit to SPOKE, such as laboratory data or specific symptoms, to inform differential diagnosis. The results are shown as a network of disease probabilities and risk factors giving insight into why SPOKE selected these disorders. (Right). For a specific diagnosis, SPOKE could be used to identify which treatments are most likely to generate the desired outcomes, while informing about the most likely side effects.

and support clinical discussions and decision-making. Live since March 2019, it has supported a number of ongoing clinical validation projects.

The SPOKE-BRIDGE integration (Figure 7), due to complete in Fall 2022, will be thoroughly evaluated in the **neurosciences** using a research roadmap evaluating both in-clinic adoption, as well as near- and long-term key **clinical** outcomes. The integration computes personalized biomedical profiles by selecting variables from a patient's clinical record and propagating (embedding) them through the entirety of the OKN (potentially billions of concepts) to provide a deep description of the patient's health status. Such network embeddings operate by learning low-rank vector representations of graph nodes and edges that preserve the graph's inherent structure. Embedding variables from hundreds of thousands of EHR's onto SPOKE showed that new knowledge (i.e. biomedical discoveries) can emerge from such a process.[21] [22]Similar approaches have been used to analyze knowledge networks from different domains where they showed superior performance and accuracy compared to previous graph exploratory approaches.[23-26] Dimensionality reduction makes such a complex biomedical profile useful and actionable for the clinician, who is alerted only to relevant clinical processes, medications, contraindications, or differential diagnostic considerations that arise from the embeddings with the OKN. The clinician queries whether their patient's biomedical profile is mathematically closer to one of their multiple diagnostic considerations on their differential, or leverages insights from other patients to predict which medication is a more precise metabolic fit for that individual. Other models are being constructed to identify biologically similar individuals (using distance measures for multifactor data at deep granularity) to surface undiagnosed conditions, as well as for critically important disease progression predictions. This approach is also being used to study the histories of patients formally diagnosed with a complex neurological condition (e.g., Parkinson's disease) to explore how far in advance this outcome could have been predicted, and on the basis of which clinical markers.

IV. SUMMARY

Knowledge is an emergent property of the interconnected web of trusted information and known facts. To mine for "unknown knowns," we must "connect the dots" from several information sources. When heterogeneous networks are connected at a massive scale, new knowledge can be extracted as an emergent property of the network. Here, the paradigm of knowledge networks - amply proven in Search – and KR&R are applied into biomedicine, a discipline that, we argue, is inherently graph-theoretic.

Machine and deep learning models such as neural networks were traditionally "black boxes," capable of delivering new data (predictions), but in and of themselves, no new knowledge. This perceived limitation has hampered their adoption in a range of chemical and biological contexts, under the sensible argument that a recommendation, prediction or prognosis a scientist or clinician cannot understand will provide no guarantee of correctness in a true discovery context. SPOKE enables the use of explanatory (i.e., "clear box") machine learning approaches with the ability to predict biomedical outcomes in a biologically meaningful manner. It has the potential to support a host of "explainable AI" techniques (see DARPA's XAI program).

At the same time, it is important for this body of knowledge to contain all the right data to create realistic and equitable models that factor in the full diversity of population and result in better health outcomes and treatments for all members of society. We believe technology can help change the current equation of designing for the "majority," and be a great leveler.

Figures Caption.

Figure 1. SPOKE Metagraph. Nodes denote biological concepts and links show how data is related and connected in the graph.

Figure 2. Mapping paths in SPOKE to empirical observations in patients

Figure 3. Analysis of blood proteomics and metabolomic data in healthy participants shows that pairs of blood analytes (protein or metabolite levels in circulation) that are correlated are connected by on average a shorter path in the SPOKE graph than any pairs of randomly chosen nodes

Figure 4: Initial rendering of a subgraph of SPOKE using a multi-level, map-like network visualization. Diseases are denoted in the top layer and they cluster by symptom and genetic similarity. The inset shows how additional details appear when zooming over an area (e.g., zooming on Immune system disease uncovers more details about additional diseases that belong to that category).

Figure 5. A view of the SPOKE Neighborhood explorer. The top panel shows the controls that allow a user to select nodes/edges for expansion as well as other key parameters. The bottom panel shows an example of the graph neighbors of the SARS-CoV-2 Spike protein (light blue), which includes 3 human proteins (green) and the genes encoding them (blue). One such protein (ACE2_HUMAN) has edges connecting it to 3 compounds (2 of them approved and one -ORE-100- in experimental phase).

Figure 6. SPOKE-enabled reconstruction of the hypothesis that dexamethasone might help recovery of patients with COVID-19. Multiple sources of evidence were required to formulate this scenario without human intervention.

Figure 7. Prototype of the potential applications of BRIDGE-SPOKE. (Left). Data from the patient's EHR can be used as access points to SPOKE to provide estimates of disorders the patient may be at risk for over a selected timeframe. (Middle) Through BRIDGE, the clinician can select data points to submit to SPOKE, such as laboratory data or specific symptoms, to inform differential diagnosis. The results are shown as a network of disease probabilities and risk factors giving insight into why SPOKE selected these disorders. (Right). For a specific diagnosis, SPOKE could be used to identify which treatments are most likely to generate the desired outcomes, while informing about the most likely side effects.

REFERENCES

[1] H.A. Simon, The Sciences of the Artificial, MIT Press 1970.

[2] M. Croitoru, P. Marquis, S. Rudolph, G. Stapleton, Graph Structures for Knowledge Representation and Reasoning : 5th International Workshop, GKR 2017, Melbourne, VIC, Australia, August 21, 2017, Revised Selected Papers, in: Lecture Notes in Artificial Intelligence 10775, Springer International Publishing : Imprint: Springer,, Cham, 2018, pp. 1 online resource (VII, 139 pages).

[3] E.M. Zdobnov, R. Lopez, R. Apweiler, T. Etzold, The EBI SRS server--recent developments, Bioinformatics, 18(2) (2002) 368-373.

[4] S.Y. Chung, L. Wong, Kleisli: a new tool for data integration in biology, Trends Biotechnol, 17(9) (1999) 351-355.

[5] P.A.T. Langley, The computational support of scientific discovery, International Journal of Human-Computer Studies, 53(3) (2000) 393-410.

[6] M. Martinez-Romero, C. Jonquet, M.J. O'Connor, J. Graybeal, A. Pazos, M.A. Musen, NCBO Ontology Recommender 2.0: an enhanced approach for biomedical ontology recommendation, Journal of biomedical semantics, 8(1) (2017) 21.

[7] N.F. Noy, N.H. Shah, P.L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D.L. Rubin, M.A. Storey, C.G. Chute, M.A. Musen, BioPortal: ontologies and integrated data resources at the click of a mouse, Nucleic Acids Res, 37(Web Server issue) (2009) W170-173.

[8] T.B.D.T. Consortium, Toward A Universal Biomedical Data Translator, Clin Transl Sci, 12(2) (2019) 86-90.

[9] Y. Zhou, F. Wang, J. Tang, R. Nussinov, F. Cheng, Artificial intelligence in COVID-19 drug repurposing, Lancet Digit Health, 2(12) (2020) e667-e676.

[10] N.D. Price, A.T. Magis, J.C. Earls, G. Glusman, R. Levy, C. Lausted, D.T. McDonald, U. Kusebauch, C.L. Moss, Y. Zhou, S. Qin, R.L. Moritz, K. Brogaard, G.S. Omenn, J.C. Lovejoy, L. Hood, A wellness study of 108 individuals using personal, dense, dynamic data clouds, Nat Biotechnol, 35(8) (2017) 747-756.

[11] C. Huang, J. Morris, S.E. Branzini, The SPOKE Neighborhood Explorer, in, 2017.

[12] B. Saket, P. Simonetto, S. Kobourov, K. Borner, Node, Node-Link, and Node-Link-Group Diagrams: An Evaluation, IEEE Trans Vis Comput Graph, 20(12) (2014) 2231-2240.

[13] S. Huang, A. Kaipainen, M. Strasser, S. Baranzini, Mechanical ventilation stimulates expression of the SARS-Cov-2 receptor ACE2 in the lung and may trigger a vicious cycle, Preprints, (2020).

[14] C. Wu, X. Chen, Y. Cai, J. Xia, X. Zhou, S. Xu, H. Huang, L. Zhang, X. Zhou, C. Du, Y. Zhang, J. Song, S. Wang, Y. Chao, Z. Yang, J. Xu, X. Zhou, D. Chen, W. Xiong, L. Xu, F. Zhou, J. Jiang, C. Bai, J. Zheng, Y. Song, Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China, JAMA Intern Med, 180(7) (2020) 934-943.

[15] R.C. Group, P. Horby, W.S. Lim, J.R. Emberson, M. Mafham, J.L. Bell, L. Linsell, N. Staplin, C. Brightling, A. Ustianowski, E. Elmahi, B. Prudon, C. Green, T. Felton, D. Chadwick, K. Rege, C. Fegan, L.C. Chappell, S.N. Faust, T. Jaki, K. Jeffery, A. Montgomery, K. Rowan, E. Juszczak, J.K. Baillie, R. Haynes, M.J. Landray, Dexamethasone in Hospitalized Patients with Covid-19, N Engl J Med, 384(8) (2021) 693-704.

[16] M.R. Garvin, C. Alvarez, J.I. Miller, E.T. Prates, A.M. Walker, B.K. Amos, A.E. Mast, A. Justice, B. Aronow, D. Jacobson, A mechanistic model and therapeutic interventions for COVID-19 involving a RAS-mediated bradykinin storm, eLife, 9 (2020).
[17] J.A. DiMasi, H.G. Grabowski, R.W. Hansen, Innovation in the pharmaceutical industry: New estimates of R&D costs, Journal of Health Economics, 47 (2016) 20-33.
[18] D. Harmey, P.R. Griffin, P.J. Kenny, Development of novel pharmacotherapeutics for tobacco dependence: progress and future directions, Nicotine Tob Res, 14(11) (2012) 1300-1318.
[19] D.S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S.L. Chen, D. Hadley, A. Green, P. Khankhanian, S.E. Baranzini, Systematic integration of biomedical knowledge prioritizes drugs for repurposing, Elife, 6 (2017).
[20] E.J. Martin, D.C. Sullivan, AutoShim: empirically corrected scoring functions for quantitative docking with a crystal structure and IC50 training data, J Chem Inf Model, 48(4) (2008) 861-872.
[21] C.A. Nelson, A.J. Butte, S.E. Baranzini, Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful machine-readable embeddings, Nat Commun, 10(1) (2019) 3045.
[22] C.A. Nelson, R. Bove, A.J. Butte, S.E. Baranzini, Embedding electronic health records onto a knowledge network recognizes prodromal features of multiple sclerosis and predicts diagnosis, Journal of the American Medical Informatics Association, (In Press) (2021).
[23] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: Proceedings Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, Lake Tahoe, Nevada, 2013, pp. 2787–2795.
[24] S.K. Mohamed, V. Nováček, Link Prediction Using Multi Part Embeddings, in: M. Fernández, K. Janowicz, A. Zaveri, A.J.G. Gray, V. Lopez, A. Haller, K. Hammar (Eds.), Springer International Publishing, Cham, 2019, pp. 240-254.
[25] M. Nickel, V. Tresp, H.-P. Kriegel, A three-way model for collective learning on multi-relational data, in: Proceedings Proceedings of the 28th International Conference on International Conference on Machine Learning, Bellevue, Washington, USA, 2011, pp. 809–816.
[26] B. Yang, W.-t. Yih, X. He, J. Gao, L. Deng, Embedding entities and relations for learning and inference in knowledge bases, arXiv preprint arXiv:1412.6575, (2014).

Authors bios:

Sergio E. Baranzini is Distinguished Professor of Neurology at the University of California San Francisco (UCSF). He earned his PhD in human molecular genetics from the University of Buenos Aires, Argentina. His lab at UCSF uses a multi-disciplinary approach to science and it is composed by experimental and computational researchers. He is the principal investigator of SPOKE, a large multi-disciplinary bioinformatics approach to gather, integrate and analyze all biomedical data, currently supported by NIH and NSF.

Katy Börner is the Victor H. Yngve Distinguished Professor of Engineering and Information Science in the Departments of Intelligent Systems Engineering and Information Science at the Luddy School of Informatics, Computing, and Engineering at Indiana University Bloomington, where she is also founding director of the Cyberinfrastructure for Network Science Center, cns.iu.edu. She is the author of Atlas of Science, Atlas of Knowledge, and Atlas of Forecasts and a coauthor of Visual Insights: A Practical Guide to Making Sense of Data (all published by The MIT Press). She is a curator of the international Places & Spaces: Mapping Science exhibit, scimaps.org.

John "Scooter" Morris is the Executive Director of the Resource for Biocomputing, Visualization, and Informatics at UCSF and the Roving Engineer for the National Resource for Network Biology and holds an Adjunct Faculty appointment in the department of Pharmaceutical Chemistry at UCSF. He received his Ph.D. in Medical Information Science from UCSF, and has bachelors degrees in Physics, Biology, and Computer Science from UC Irvine. Scooter is the co-author of over 40 papers, mostly focused on the use of network visualization and analysis in biology. He is also a member of the Cytoscape core development team, and author of several Cytoscape plugins and core features.

Charlotte A. Nelson is an Associate Specialist at the University of California, San Francisco (UCSF). Her work focuses on leveraging knowledge graphs (KGs) to accelerate advances in biomedical research. Nelson aims to make the power of KGs accessible to a broader public by creating human-centric platforms tailored for academic, clinical, and pharmaceutical researchers. She received her PhD from UCSF in Biological and Medical Informatics. Nelson also holds BS degrees in Biochemistry and Bioinformatics from the University of California, Santa Cruz.

Karthik Soman works as Associate Specialist in Prof. Sergio Baranzini's lab, Department of Neurology, University of California San Francisco. His work mainly focuses on computational methods like machine learning, data integration and embedding on biomedical knowledge graph, big data analytics using high performance computing clusters, clinical data enrichment for precision medicine. Apart from this, he is also a part of the team that brings a universal schema for biomedical knowledge networks.

Erica Schleimer is a software engineer focusing on the intersection between software engineering and data science. While at UCSF, she was the technical lead for BRIDGE, a precision medicine platform integrated with electronic health records, and Open MS BioScreen, an openly available application for tracking and understanding Multiple Sclerosis. She has an engineering degree from Northwestern University and is currently in the Master's Degree Program in Information and Data Science at UC Berkeley.

Michael Keiser is an Associate Professor at UCSF with primary appointments in the Institute for Neurodegenerative Diseases and the Department of Pharmaceutical Chemistry. His lab's research combines machine and deep learning with chemical biology to quantify how small molecules perturb protein networks and induce their therapeutic effects. The Keiser Lab develops approaches to address core challenges in systems pharmacology, such as the determination of multi-target small-molecule mechanisms and the integration of phenotypic and high-content screening data. They likewise develop deep-learning based

pattern recognition tools for diverse biomedical data, including histopathological whole-slide images, pharmacological time-course data, and ligand-protein structural interactions.

Mark Musen is Professor of Biomedical Informatics at Stanford University, where he is Director of the Stanford Center for Biomedical Informatics Research. He conducts research related to intelligent systems, reusable ontologies, metadata for publication of scientific data sets, and biomedical decision support. His long-standing work on the Protégé system has led to widely used, open-source technology to build ontologies and intelligent computer systems. Dr. Musen is principal investigator of BioPortal ontology repository as well as the Center for Expanded Data Annotation and Retrieval (CEDAR), which applies semantic technology to enhance open science.

Roger Pierce is a computer scientist in the Center for Applied Scientific Computing (CASC) at Lawrence Livermore National Laboratory. His research interests center around parallel and external memory graph algorithms and data-intensive computing. Roger joined LLNL in 2008 as a Lawrence Scholar, and joined CASC in 2013.

Tahsin Reza is a member of the research staff at the Center for Applied Scientific Computing, Lawrence Livermore National Laboratory. He investigates systems techniques for emerging Big data problems that demand scalable and timely solution. His primary research interests are in the area of parallel and distributed computing, datamining and relational-data analysis. He completed the PhD degree in Electrical and Computer Engineering at The University of British Columbia, Vancouver.

Brett Smith is a Software Engineer IV in Health Data Science at the Institute for Systems Biology in Seattle, Washington.

Bruce W. Herr II is the Senior System Architect & Project Manager at the Cyberinfrastructure for Network Science Center at Indiana University. He has spent over 16 years leading software teams and designing and building data analysis and visualization software in both industry and academia.

Boris Oskotsky studied at Saint Petersburg Technical University where he earned a MS in Computers Science and a PhD in Solid State Physics. Previously worked at Stanford University in different departments within the School of Medicine including Information Resources and Technology (IRT), Biomedical Information Research (BMIR) and the Department of Neurobiology. Currently, he is the Lead Systems Administrator for the Butte Lab as well as the Information Commons Team within the Bakar Computational Health Sciences Institute at UCSF.

Angela Rizk-Jackson obtained her PhD in neuroscience from UCLA, and completed postdoctoral work at UCSF in the area of neuroimaging, using machine learning analytical methods. As a Program Manager at UCSF's Clinical and Translational Science Institute, she led Open Data efforts by helping to establish DataShare, and worked to facilitate several strategic initiatives including establishment of a multi-institution clinical trials networks, and development of infrastructure and processes that enable investigators to enhance their research programs.

Katherine P. Rankin is a Professor in Residence in the Memory and Aging division of the UCSF Department of Neurology. She trained at Yale, Fuller, and UCSF, earning a PhD in Clinical Psychology with a specialization in Neuropsychology. At the MAC she uses quantitative neuroimaging to investigate the underpinnings of human socioemotional behaviors such as empathy, self-awareness, social cue detection, and theory of mind, in healthy aging adults and patients with neurodegenerative disease. She also designs and builds informatics tools for harmonizing cross-disciplinary data and analytic processes to facilitate scientific collaboration, accelerate discovery, and improve clinical care.

Stephan J Sanders trained as a pediatric physician in the UK before undertaking a PhD and postdoctoral research position at Yale. He is now an Associate Professor at UCSF in the Department of Psychiatry. His

research focuses on using genomics and bioinformatics to understand the etiology of developmental disorders, such as Autism Spectrum Disorder (ASD).

Riley Bove is Associate Professor of Neurology at UCSF, and Director of Digital Innovation in the UCSF MS Group. She obtained her BA from Harvard College and her MD and MMSc degrees from Harvard University. Her research on digital health and sex- and gender-informed neurology has been funded by the NIH, National MS Society, California Initiative to Advance Precision Medicine, Hilton Foundation, and industry. In the digital health arena, she focuses on applying novel tools to improve clinical research and care. She is a Co-PI of the BRIDGE platform.

Peter Rose is Director of the Structural Bioinformatics Lab and Lead for Bioinformatics and Biomedical Applications at the San Diego Supercomputer Center (SDSC), UC San Diego. He has previously led bioinformatics and scientific computing departments at Pfizer and Agouron Pharmaceuticals. He led the RCSB Protein Data Bank team at UCSD. He is currently involved in projects to integrate cross-disciplinary data for novel COVID-19 diagnostic and surveillance methods, and the application of knowledge graphs to COVID-19 and precision medicine datasets. His research interests include the development of interactive and scalable platforms for data integration and machine learning in biomedicine and structural biology.

Sharat Israni is Executive Director & CTO at UCSF's Bakar Computational Health Sciences Institute. Previously, he was Executive Director, Data Science, at Stanford Medicine. A long-serving Technology executive, Sharat's teams pioneered the use of "Big Data." He served as VP of Data at Yahoo! (1999-2008) and Intuit (2010-13), which pioneered Data Science/AI to re-invent their products. He led Digital Media systems for broadcast/interactive TV at Silicon Graphics; and Data teams at IBM and HP. He was PI for the NITRD Open Knowledge Network workshop 2017 hosted by NSF and NIH. Sharat is a frequent peer reviewer of journal articles and grant proposals.

Sui Huang first studied medicine, followed by molecular biology and physical chemistry at the University of Zurich in the 1990s. He was a faculty at the Harvard Medical School/Children's Hospital and then at the University of Calgary, conducting studies on cell fate control and tumor angiogenesis. He has championed the embrace of complex systems theory by biomedical research. His current work at the Institute for Systems Biology which he joined in 2011, uses new technologies, including single-cell omics, along with the theory of non-linear stochastic dynamical systems to better understand the dynamics in health and disease, including cancer drug resistance, stem cell differentiation and wellness-disease transitions in Personal Medicine.