BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FIVE PAGES.**

NAME: Joachimiak, Marcin Pawel

eRA COMMONS USER NAME (credential, e.g., agency login): MJOACHIMIAK

POSITION TITLE: Staff Researcher and Software Developer

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	Completion Date MM/YYYY	FIELD OF STUDY
University of Chicago	B.A.	06/96	Mathematics
University of California, San Francisco	Ph.D.	03/02	Biophysics
University of California, Berkeley	Postdoctoral	06/2006	Human and Microbial Genomics

A. Personal Statement

I am a Staff Researcher and Software Developer at Lawrence Berkeley National Laboratory (LBNL), in the Biosystems Data Science Dept. under the Environmental Genomics and Systems Biology Division.

My background is in the areas of functional genomics, systems biology, data science, knowledge and standards modeling, and algorithm and machine learning method development in computational biology. From early on in my career I worked extensively with data from multiple experimental technologies to functionally characterize cells including gene expression profiling, high throughput phenotyping, mass spectrometry of small molecules and proteins, both in method development and data science applications which often involved interpreting results. I also developed methods for and analyzed large integrated data compendia such as from transcriptomics, gene and cell fitness assays, biological entity associations, and protein interactions, a theme that I have continued throughout my career.

Thanks to my work with KBase (the DOE Systems Biology Knowledgebase) starting in 2011, I have become an expert in data modeling and standardization as well as algorithm development in an integrated method development and user-facing platform with a rich collection of structured data types. More recently I have been co-leading the KBase Knowledge Engine prototype project, where our team developed a classification model for metagenomics samples as well as other critical machine learning components. At LBNL I am also involved with the Mungall group, which maintains the Gene Ontology and are core contributors to the Monarch Initiative, NCATS Biomedical Data Translator, and NIH Bridge2AI projects. These projects have deepened my expertise in knowledge modeling and semantic technologies and given me valuable insight into human biomedical data in the context of semantic harmonization and related methods. My earlier work at LBNL with large experimental projects has given me deep experience in omics method development and data science across many genomic and multi-omic data types. I led the development of a scalable biclustering method that outperforms the state of the art. Biclustering methods are directly relevant to our current proposal because they lead to discovery of data patterns which associate different types of features (for example genes and sample metadata in transcriptomics) in a dataset.

During my career I have successfully recruited, taught, mentored, and managed undergraduate, graduate students, research assistants, software developers, and postdocs. In addition to leading the LBNL project for microbial growth predictions, I have led or currently lead team efforts in functional genomics analysis (ENIGMA, KBase, NCATS Translator), knowledge modeling (KBase, Bridge2AI), and machine learning

algorithm development (ENIGMA, KBase, NCATS Translator). Crucially for this proposal, I have developed, led, and contributed to multiple open source projects, both standalone applications as well as large computational systems with open access for the community.

Ongoing and recently completed projects that I would like to highlight include:

DE-AC02-05CH11231 Joachimiak 10/01/2023- 09/30/2024 LBNL LDRD
CultureBot - a computational framework to support automated high throughput microbial culturing and growth assays
Role: Principal Investigator

DE-AC02-05CH11231 Arkin 10/01/2020- 09/30/2024 DOE
DOE Systems Biology KnowledgeBase
Role: Scientist and Knowledge Engine Classification and Graph Learning Project Lead

DE-AC02-05CH11231 05/01/20 - 09/30/21 LBNL LDRD
Leveraging knowledge graphs and machine learning to produce actionable knowledge for COVID-19 response
Role: Co-PI

Citations:

1. Caufield JH, Putman T, Schaper K, Unni DR, Hegde H, Callahan TJ, Cappelletti L, Moxon SAT, Ravanmehr V, Carbon S, Chan LE, Cortes K, Shefchek KA, Elsarboukh G, Balhoff J, Fontana T, Matentzoglu N, Bruskiewich RM, Thessen AE, Harris NL, Munoz-Torres MC, Haendel MA, Robinson PN, Joachimiak MP, Mungall CJ, Reese JT. KG-Hub-building and exchanging biological knowledge graphs. Bioinformatics. 2023 Jul 1;39(7). doi: 10.1093/bioinformatics/btad418 PMCID: PMC10336030. Cited by 23.
2. Joachimiak MP, Tuglus C, Salamzade R, van der Laan M, Arkin AP. Deep surveys of transcriptional modules with Massive Associative K-biclustering. bioRxiv August 26, 2022 doi:10.1101/2022.08.26.505372 (Preprint). Cited by 2.
3. Joachimiak MP, Caufield JH, Harris NL, Kim H, Mungall CJ. Gene Set Summarization using Large Language Models. arXiv [q-bio.GN]. 2023. doi:10.48550/arXiv.2305.13338 (Preprint). Cited by 2.
4. Joachimiak MP, Hegde H, Duncan WD, Reese JT, Cappelletti L, Thessen AE, Mungall CJ. KG-Microbe: A Reference Knowledge-Graph and Platform for Harmonized Microbial Information. CEUR Workshop Proceedings vol. 3073, 131-133 2021. Cited by 5.

B. Positions, Scientific Appointments, and Honors

2017- Scientist, Lawrence Berkeley National Laboratory, Mungall group, Environmental Genomics and Systems Biology Division, Berkeley, CA
2011- Scientist and Software Developer, U.S. Department of Energy Knowledgebase, Lawrence Berkeley National Laboratory, Berkeley, CA
2006-2012 Scientist and Software Developer, Virtual Institute for Microbial Stress and Survival, ESPP, ESPP2, and ENIGMA Department of Energy Scientific Focus Areas, Lawrence Berkeley National Laboratory, Berkeley, CA
2006- Staff Researcher and Software Developer, Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Prof. Adam P. Arkin Laboratory, Berkeley, CA
2003-2006 Postdoctoral Researcher, Dept. of Plant and Microbial Biology, Computational Genomics Group, Prof. Steven E. Brenner, University of California, Berkeley, CA
2002-2003 Scientist, Five Prime Therapeutics, South San Francisco, CA
1997-2002 Bioinformatics Scientific Consultant, DoubleTwist (Pangea Systems), Oakland, CA
1994-1996 Research Assistant, Dept. of Molecular Genetics and Cell Biology, Prof. Robert Haselkorn, University of Chicago, Chicago, IL

C. Contributions to Science

1. **Integrated Computational Analysis Systems and Systems Biology**
My contributions to systems biology are highlighted by my system and method design as well as software

development contributions to integrated computational systems, such as KBase, the NCATS Biomedical Data Translator, and MicrobesOnline. These platforms integrate and standardize immense amounts of genomic and functional data, along with the implementing methods that can analyze and model this data. By contributing to multiple large integrated computational systems, I have directly contributed to democratizing access to computing and data for the scientific community.

- Biomedical Data Translator Consortium. Toward a universal biomedical data translator. Clinical and translational science 12 (2), 86. 2019 PMCID: PMC6440568. Cited by 58.
- Arkin AP, et al. KBase: The United States Department of Energy Systems Biology Knowledgebase. Nat Biotechnol. 2018 Jul 6;36(7):566-569. doi: 10.1038/nbt.4163. PMCID: PMC6870991. Cited by 956.
- Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, Chivian D, Friedland GD, Huang KH, Keller K, Novichkov PS, Dubchak IL, Alm EJ, Arkin AP. MicrobesOnline: an integrated portal for comparative and functional genomics. Nucleic Acids Res. 2009 Nov; 38(Database issue). PMCID: PMC2808868. Cited by 516.
- Yang Y, Harris DP, Luo F, Xiong W, Joachimiak M, Wu L, Dehal P, Jacobsen J, Yang Z, Palumbo AV, Arkin AP, Zhou J. Snapshot of iron response in Shewanella oneidensis by gene network reconstruction. BMC Genomics. 2009 Mar 25;10:131. doi: 10.1186/1471-2164-10-131. PMCID: PMC2667191. Cited by 76.

2. **Genomic Data for Disease Understanding and Drug Discovery**

Harnessing genomic data for disease understanding and drug discovery is another significant area of my research. By integrating and analyzing genomic and phenotypic data, such as in the study of malaria and COVID-19, we aim to uncover novel therapeutic targets and deepen our understanding of disease mechanisms. This interdisciplinary approach bridges genomics, pharmacology, and clinical research, contributing to the development of new treatments and precision medicine strategies.

- Joachimiak MP. Zinc against COVID-19? Symptom surveillance and deficiency risk groups. PLoS neglected tropical diseases 15 (1), e0008895. 2021. PMCID: PMC7781367. Cited by 91.
- Reese JT, Unni D, Callahan TJ, Cappelletti L, Ravanmehr V, Carbon S, Shefchek KA, Good BM, Balhoff JP, Fontana T, Blau H, Matentzoglu N, Harris NL, Munoz-Torres MC, Haendel MA, Robinson PN, Joachimiak MP, Mungall CJ. KG-COVID-19: A Framework to Produce Customized Knowledge Graphs for COVID-19 Response. Patterns (N Y). 2021 Jan 8;2(1):100155. doi: 10.1016/j.patter.2020.100155. PMCID: PMC7444288. Cited by 87.
- Bozdech Z, Zhu J, Joachimiak M, Cohen FE, DeRisi J. Expression Profiling the Schizont and Trophozoite Stages of Plasmodium falciparum with a Long Oligonucleotide Microarray. (2003) Genome Biol. 4:R9. PMCID: PMC151308. Cited by 483.
- Joachimiak MP, Chang C, Rosenthal PJ, Cohen FE. The impact of whole genome sequence data on drug discovery: a malaria case study. Mol Med. 2001 Oct;7(10):698-710. PMCID: PMC1949995. Cited by 35.

3. **Open Source Software for Biological Research: AI/ML, Algorithms, Visualization**

I am deeply committed to the development of open-source software, exemplified by tools like JEvtrace, JColorGrid, and MAK which facilitate genomics and multi-omics research. This dedication to creating and sharing tools reflects a broader commitment to advancing scientific research through collaborative and accessible resources. By lowering barriers to making predictions, analyzing data, and generating visualizations, we empower researchers across disciplines to generate new insights and discoveries.

- Cappelletti L, Fontana T, Casiraghi E, Ravanmehr V, Callahan TJ, Cano C, Joachimiak MP, Mungall CJ, Robinson PN, Reese J, Valentini G. GRAPE for fast and scalable graph processing and random-walk-based embedding. Nature Computational Science. Nature Publishing Group; 2023 Jun 26;3(6):552–568. PMCID: PMC10768636. Cited by 9.
- Joachimiak MP, Tuglus C, Salamzade R, van der Laan M, Arkin AP. Deep surveys of transcriptional modules with Massive Associative K-biclustering. bioRxiv August 26, 2022 doi.org/10.1101/2022.08.26.505372. Cited by 1. (Preprint)
- Joachimiak MP, Weisman JL, May BC. JColorGrid: software for the visualization of biological measurements. BMC Bioinformatics. 2006 Apr;7:225. PMCID: PMC1479842. Cited by 121.
- Joachimiak MP, Cohen FE. JEvTrace: refinement and variations of the evolutionary trace in JAVA. Genome Biol. 2002;3(12):RESEARCH0077. PMCID: PMC151179. Cited by 29.

4. **Biomedical Data Integration and Knowledge Graphs**

My work in biomedical data integration and knowledge graph development, such as the KG-Hub infrastructure, KG-COVID-19 and KG-Microbe, aims to synthesize disparate data sources into coherent, structured formats that facilitate complex analyses. These efforts are crucial for understanding the multifaceted nature of diseases and for advancing precision medicine by connecting phenotypes to genotypes across species. I applied these concepts accepted in the biomedical domain to another research domains, namely microbiology and specifically microbial traits (see KG-Microbe).

- Caufield JH, Putman T, Schaper K, Unni DR, Hegde H, Callahan TJ, Cappelletti L, Moxon SAT, Ravanmehr V, Carbon S, Chan LE, Cortes K, Shefchek KA, Elsarboukh G, Balhoff J, Fontana T, Matentzoglu N, Bruskiewich RM, Thessen AE, Harris NL, Munoz-Torres MC, Haendel MA, Robinson PN, Joachimiak MP, Mungall CJ, Reese JT. KG-Hub--Building and Exchanging Biological Knowledge Graphs. Bioinformatics. 2023 Jul 1;39(7). doi: 10.1093/bioinformatics/btad418. PMCID: PMC10336030. Cited by 4.
- Unni DR, Moxon SAT, Bada M, Brush M, Bruskiewich R, Caufield JH, Clemons PA, Dancik V, Dumontier M, Fecho K, Glusman G, Hadlock JJ, Harris NL, Joshi A, Putman T, Qin G, Ramsey SA, Shefchek KA, Solbrig H, Soman K, Thessen AE, Haendel MA, Bizon C, Mungall CJ; Biomedical Data Translator Consortium. Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. Clinical and translational science 15 (8), 1848-1855. 2022. PMCID: PMC9372416. Cited by 39.
- Joachimiak MP, Hegde H, Duncan WD, Reese JT, Cappelletti L, Thessen AE, Mungall CJ. KG-Microbe: A Reference Knowledge-Graph and Platform for Harmonized Microbial Information. CEUR Workshop Proceedings vol. 3073, 131-133 2021. Cited by 5.
- Shefchek KA, Harris NL, Gargano M, Matentzoglu N, Unni D, Brush M, Keith D, Conlin T, Vasilevsky N, Zhang XA, Balhoff JP, Babb L, Bello SM, Blau H, Bradford Y, Carbon S, Carmody L, Chan LE, Cipriani V, Cuzick A, Della Rocca M, Dunn N, Essaid S, Fey P, Grove C, Gourdine JP, Hamosh A, Harris M, Helbig I, Hoatlin M, Joachimiak M, Jupp S, Lett KB, Lewis SE, McNamara C, Pendlington ZM, Pilgrim C, Putman T, Ravanmehr V, Reese J, Riggs E, Robb S, Roncaglia P, Seager J, Segerdell E, Similuk M, Storm AL, Thaxon C, Thessen A, Jacobsen JOB, McMurry JA, Groza T, Köhler S, Smedley D, Robinson PN, Mungall CJ, Haendel MA, Munoz-Torres MC, Osumi-Sutherland D. The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic acids research 48 (D1), D704-D715. 2020. PMCID: PMC7056945. Cited by 186.

5. **Metagenomics and Environmental Genomics**

Exploring the vast diversity of microbial life, my work in metagenomics and environmental genomics has significantly contributed to expanding our understanding of the microbial universe. Through large-scale sequencing efforts like the Sorcerer II Global Ocean Sampling Expedition, we have uncovered a tremendous variety of protein families, revealing the complexity and richness of microbial communities across different environments. This research not only enhances our knowledge of microbial ecology but also opens new avenues for biomedical and biotechnological applications.

- Park H, Joachimiak MP, Jungbluth SP, Yang Z, Riehl WJ, Canon RS, Arkin AP, Dehal PS. A bacterial sensor taxonomy across earth ecosystems for machine learning applications. mSystems. 2024 Jan 23;9(1):e0002623. doi: 10.1128/msystems.00026-23. PMCID: PMC3486112.
- Borglin S, Joyner D, DeAngelis KM, Khudyakov J, D'haeseleer P, Joachimiak MP, Hazen T. Application of phenotypic microarrays to environmental microbiology. Curr Opin Biotechnol. 2012 Feb;23(1):41-8. doi: 10.1016/j.copbio.2011.12.006. Epub 2012 Jan 2. Cited by 67.
- Mukhopadhyay A, Redding AM, Joachimiak MP, Arkin AP, Borglin SE, Dehal PS, Chakraborty R, Geller JT, Hazen TC, He Q, Joyner DC, Martin VJ, Wall JD, Yang ZK, Zhou J, Keasling JD. Cell-wide responses to low-oxygen exposure in Desulfovibrio vulgaris Hildenborough. J Bacteriol. 2007 Aug;189(16):5996-6010. PMCID: PMC1952033. Cited by 126.
- Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, van Belle C, Chandonia JM, Soergel DA, Zhai Y, Natarajan K, Lee S, Raphael BJ, Bafna V, Friedman R, Brenner SE, Godzik A, Eisenberg D, Dixon JE, Taylor SS, Strausberg RL, Frazier M, Venter JC. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol. 2007 Mar;5(3):e16. PMCID: PMC1821046. Cited by 996.

Complete List of Published Work:

https://www.ncbi.nlm.nih.gov/myncbi/marcin.joachimiak.2/bibliography/public/