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: Vitek, Olga

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

POSITION TITLE: Professor of Computer Science

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 Geneva, Geneva, Switzerland	B.Sc.	06/1995	Econometrics
University of Geneva, Geneva, Switzerland	M.Sc.	06/1997	Econometrics
Purdue University, West Lafayette IN	M.Sc.	05/2001	Mathematical statistics
Purdue University, West Lafayette IN	Ph.D.	05/2005	Statistics
Institute for Systems Biology, Seattle WA	Post-doctoral	07/2006	Proteomics

A. Personal Statement

My research explores synergies between statistical science and machine learning, as applied to quantitative large-scale mass spectrometry-based investigations, to understand the functioning of living organisms. Our work covers (1) statistical experimental design, (2) detecting analyte's signals in large and complex outputs produced by the instruments, (3) statistical selection of the relevant signals, and (4) inference of causal associations among the analytes that are informative of their biological function.

My group develops methods and open-source R-based software, that are broadly used by scientists in academia and industry. These include MSstats (relative quantification of proteins in mass spectrometric experiments, > 68K downloads), Cardinal (a comprehensive software suite for analysis of mass spectrometry imaging experiments, > 26K downloads) and matter (memory-efficient reading, writing, and manipulation of binary larger-than-memory datasets in complex formats, 3,400 unique IP downloads.) This also includes the infrastructure within the public repository MassIVE, called MassIVE.quant, for reproducible documentation and reanalyses of quantitative proteomic experiments. The work was recognized with the 2019 Chan Zuckerberg Essential Open Source Software award, and with the 2021 Gilbert S. Omenn Computational Proteomics Award of the US Human Proteome Organization. In 2021 was elected Fellow of the American Statistical Association.

I am committed to statistical education in life sciences. While at Purdue, in 2008 I was recognized this with the Teaching for Tomorrow Award, given to assistant professors with a high potential in teaching. In 2010, I was recognized by the Department of Statistics with the Outstanding Assistant Professor Teaching Award and the Graduate Student Mentoring Award. I am president of the Boston Chapter of the American Statistical Association, which organizes extensive educational events for data science professionals. I co-taught or co-organized over 40 short courses for life scientists world-wide. In Fall 2016 I received support from the NIH "Big data to knowledge" (BD2K) program to organize, on the campus of Northeastern University, an educational event called *May Institute on Computation and Statistics in Mass Spectrometry and Proteomics*. The program reaches out to the members of mass spectrometry and proteomics communities who are often left out of training in statistics and computing. It also reaches out to statisticians and computational scientists interested in mass spectrometry.

B. Positions and Honors

Positions and Employment

1996-1999	Teaching assistant, University of Geneva, Switzerland
1997-1998	Statistician, Department of Epidemiology, University Hospitals of Geneva, Switzerland
2001 - 2005	Research Assistant, Purdue University, IN
2004	Intern, Mass spectrometry-based proteomics, Eli Lilly and Co, Indianapolis IN
2005	Member of the Statistical Consulting Service, Purdue University, IN
2006 - 2011	Assistant Professor, Departments of Statistics and Computer Science, Purdue University, IN
2012-2013	Visiting Associate Professor, Department of Radiology, Stanford University, CA
2011-2014	Associate Professor with tenure, Departments of Statistics (95%) and Computer Science (5%), Purdue University, IN
2014-2017	Sy and Laurie Sternberg Interdisciplinary Associate Professor, Chemistry and Chemical Biology, College of Science (50%), and College of Computer and Information Science (50%), Northeastern University, MA
2018-2019	Associate Professor, College of Computer and Information Science, Northeastern University, M
2019-	Professor, Khoury College of Computer Sciences, Northeastern University, MA
2022-	Affiliated Faculty, Chemistry and Chemical Biology, College of Science

Other Experience and Professional Memberships

2014-	Board of Directors member, US Human Proteome Organization (USHUPO)
2016-2019	Executive Committee member, Board of Directors, US HUPO
2019-	Member, Council of the International Human Proteome Organization (HUPO)
2019-2020	Elected Officer and Program Chair, Boston Chapter of the American Statistical Association
2019-	Senior Member, International Society for Computational Biology
2020-	Elected President, Boston Chapter of the American Statistical Association
2020-	Associate Editor, *Bioinformatics*

Honors

2008	Teaching for Tomorrow Award, Office of the Provost, Purdue University
2010	Graduate Student Mentoring Award, College of Science, Purdue University
2010	Outstanding Assistant Professor Teaching Award, Department of Statistics, Purdue University
2011	NSF CAREER Award
2013	Purdue University Faculty Scholar
2015	Sy and Laurie Sternberg Interdisciplinary chair, Northeastern University
2020	Northeastern University Excellence in Research and Creative Activity Award
2020	Essential Open Source Software for Science Award, Chan-Zuckerberg Initiative
2021	Gilbert S. Omenn Computational Proteomics Award of the US Human Proteome Organization
2021	Elected Fellow of the American Statistical Association
2022	Indigo BioAutomation Females in Mass Spectrometry Distinguished Contribution Award

C. Contributions to Science

C.1. Experimental designs for mass spectrometric systems biology investigations

One way to maximize biological insight from the planned experiment is to select for quantification an optimal set of proteins, conditions and perturbations. We developed a Bayesian active learning experimental design strategy that selects optimal perturbations for causal inference of signaling networks (Ref. a). In (Ref. b), we advocated for the use of structural causal models for counterfactual inference of outcomes of interventions on biomolecular networks. In (Ref. c), we proposed an approach for integrating stochastic differential equations with structural causal models, to evaluate outcomes of hypothetical mutually exclusive interventions. Another way to maximize biological insight from a planned experiment is to optimize its technological aspects, and reduce the unwanted artifacts and variation. We proposed an informative experimental design for one such mass spectrometric experiment type, which utilizes targeted mass spectrometry and stable isotope labeled reference

peptides (Ref. d). We overviewed the fundamental principles of statistical experimental design, as applied to mass spectrometry-based proteomics, in (Ref. e).

a. R. O. Ness, K. Sachs, P. Mallick, **O. Vitek**. "A Bayesian active learning experimental design for inferring signaling networks". *In: Sahinalp S. (eds) Research in Computational Molecular Biology (RECOMB). Lecture Notes in Computer Science,* 10229:134, 2017, acceptance rate 20%. PMID: 29927613
b. J. Zucker, K. Paneri, S. Mohammad-Taheri, S. Bhargava, P. Kolambkar, C. Bakker, J. Teuton, C. T. Hoyt, K. Oxford, R. Ness, **O. Vitek**. "Leveraging structured biological knowledge for counterfactual inference: a case study of viral pathogenesis". *IEEE Transactions on Big Data*, *in press*, 2020. PMID: NA
c. R. O. Ness, K. Paneri, **O. Vitek**. "Integrating mechanistic and structural causal models enables counterfactual inference in complex systems" In: *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 14211, 2019, acceptance rate 21%. PMID: *NA*.
d. C.-Y. Chang, E. Sabidó, R. Aebersold, **O. Vitek**. "Targeted protein quantification using sparse reference labeling". *Nature Methods*, 11:301, 2014. PMID: 24441934.
e. A. L. Oberg, **O. Vitek**. "Statistical design of quantitative mass spectrometry-based proteomic experiments". *Journal of Proteome Research*, 8:2144, 2009. PMID: 19222236

C.2. Inferential relative quantification of proteins for mass spectrometry-based proteomics

To address statistical challenges of quantitative mass spectrometry-based proteomics, my group developed a general and flexible statistical methodology for relative quantification of proteins in mass spectrometry-based experiments. Specifically, we developed a family of linear mixed effects models that is applicable to a broad variety of workflows (shotgun, targeted or data-independent; label-free or label-based) relying on liquid chromatography for quantification of spectral features. The methods are accompanied with open-source software MSstats www.msstats.org, which automatically recognizes arbitrary complex experimental designs, fits the appropriate model, and derives model-based conclusions. MSstats is distributed via Bioconductor (38,992 unique IP downloads since 2013, top 20% of all packages), and as an external tool in the computational framework Skyline (20,313 unique IP downloads since 2015). It offers a general statistical framework for storing and reanalyzing public quantitative proteomic datasets in the repository MassIVE (Ref a.) MSstats implements advanced statistical methods such as modeling of multiplexed experiments (Ref b.) or selection of protein features with consistent quantitative profiles (Ref. c). We further extended the statistical methodology to tasks beyond relative protein quantification. We proposed non-linear regression models for characterization of multiplexed mass spectrometric assays (Ref. d), and statistical methodology for longitudinal system suitability monitoring of large-scale experiments (Ref. e).

a. M. Choi, J. Carver, C. Chiva, M. Tzouros, T. Huang, T.-H. Tsai, B. Pullman, O. M. Bernhardt, R. Hüttenhain, G. C. Teo, Y. Perez-Riverol, J. Muntel, M. Müller, S. Goetze, M. Pavlou, E. Verschueren, B. Wollscheid, A. I. Nesvizhskii, L. Reiter, T. Dunkley, E. Sabido, N. Bandeira, **O. Vitek**. "MassIVE.quant: a community resource of quantitative mass spectrometry–based proteomics datasets", *Nature Methods*, 17:981, 2020. PMID: 32929271
b. T. Huang, M. Choi, M. Tzouros, S. Golling, N. J. Pandya, B. Banfai, T. Dunkley, **O. Vitek**. "MSstatsTMT: Statistical detection of differentially abundant proteins in experiments with isobaric labeling and multiple mixtures", *Molecular & Cellular Proteomics*, mcp.RA120.002105, 2020. PMID: 32680918
c. T.-H. Tsai, M. Choi, B. Banfai, Y. Liu, B. X. MacLean, T. Dunkley, **O. Vitek**. "Selection of features with consistent profiles improves relative protein quantification in mass spectrometry experiments". *Molecular & Cellular Proteomics*, mcp.RA119.001792, 2020. PMC7261813
d. C. Galitzine, J. D. Egertson, S. Abbatiello, C. M. Henderson, A. N. Hoofnagle, M. MacCoss, **O. Vitek**. "Nonlinear regression improves accuracy of characterization of multiplexed mass spectrometric assays", *Molecular & Cellular Proteomics*, RA117.000322, 2018. PMC5930407
e. E. Dogu, S. Mohammad-Taheri, S. E. Abbatiello, M. S. Bereman, B. MacLean, B. Schilling, **O. Vitek**. "MSstatsQC: Longitudinal system suitability monitoring and quality control for targeted proteomic experiments". *Molecular & Cellular Proteomics*, M116. 064774, 2017. PMC5500765

C.3. Mass spectrometry-based imaging

This new technology investigates the spatial chemical composition of biological samples such as tissues. The data are thousands of spectral features, characterized by spatial, within- and between-tissue variation. There is currently no generally accepted methodology for data interpretation. We developed spatial shrunken centroids, a statistical framework for both supervised classification (Ref. a,b) and unsupervised segmentation (Ref. c) of the images, in presence of spatial structure and complex experimental designs. We implemented Cardinal, an efficient, R-based software package for interpretation of mass spectrometry imaging (Ref. d). Cardinal implements a set of spectral processing algorithms, as well as machine learning methods developed in our group. It is the only open-source tool for mass spectrometry imaging with such a broad scope currently available. Cardinal was recognized with the John M. Chambers Statistical Software Award, the highest award given by the American Statistical Association to a student-led project. We expanded our work to support large-scale datasets with Matter, an R-based package for rapid development of out-of-core algorithms (Ref. e). It directly accesses larger-than-memory data stored in custom binary formats. Since 2014 Cardinal and Matter had 18,158 unique IP downloads and is in top 20% of all Bioconductor packages.

a. K. D. Bemis, A. Harry, L. S. Eberlin, C. R. Ferreira, S. M. van de Ven, P. Mallick, M. Stolowitz, **O. Vitek**. "Probabilistic segmentation of mass spectrometry images helps select important ions and characterize confidence in the resulting segments". *Molecular & Cellular Proteomics*, mcp.O115.053918, 2016. PMID: 26796117
b. D. Guo, M. Föll, V. Volkmann, K. Enderle-Ammour, P. Bronsert, O. Schilling, **O. Vitek**. "Deep multiple instance learning classifies subtissue locations in mass spectrometry images from tissue-level annotations", In: *Proceedings of Intelligent Systems for Molecular Biology (ISMB)*, 36:i300, 2020, acceptance rate 20%. PMID: NA
c. D. Guo, K. Bemis, C. Rawlins, J. Agar, **O. Vitek**. "Unsupervised segmentation of mass spectrometric ion images characterizes morphology of tissues", *In: Proceedings of International Society for Computational Biology (ISMB)*, 2019, acceptance rate 18%. PMID: NA.
d. K.D. Bemis, A. Harry, L. S. Eberlin, C. Ferreira, S. M. van de Ven, P. Mallick, M. Stolowitz, **O. Vitek**. "Cardinal: an R package for statistical analysis of mass spectrometry-based imaging experiments". *Bioinformatics*, 31:2418, 2015. PMC4495298
e. K. A. Bemis, **O. Vitek**. "Matter: an R package for rapid prototyping with larger-than-memory datasets on disk", *Bioinformatics*, 19:3142, 2017 PMC5870624

C.4. Statistical detection of hits in high-throughput screens of elemental profiles
Inductively coupled plasma spectroscopy combined with mass spectroscopy (ICP-MS) can be used to quantify mineral nutrient and trace elements (e.g. P, Ca, K, Mg, Fe, Zn) in various biological samples, and in particular in large-scale perturbation screens. We proposed a linear mixed-effects modeling and an Empirical Bayes procedure for relative element quantification in large-scale perturbation screens, that accounts for experimental artifacts and detects elements and conditions that are affected by the perturbations. We evaluated the approach using two comprehensive screens of *S. cerevisiae*, which involve 4965 single-gene knock-out mutants and 5825 single-gene over-expressed mutants (Ref. a), and showed that it leads to biologically meaningful results (Ref. b). Our related work appeared in (Ref. c, d).

a. D. Yu, J. Danku, I. Baxter, S. Kim, O. Vatamaniuk, D. E. Salt, **O. Vitek**. "Noise reduction in genome-wide perturbation screens using linear mixed-effects models", *Bioinformatics*, 27:2173, 2011. PMID: 21685046.
b. D. Yu, J. Danku, I. Baxter, S. Kim, O. Vatamaniuk, **O. Vitek**, D. E. Salt. "High-resolution genome-wide scan of genes, gene-networks and cellular systems impacting the yeast ionome". *BMC Genomics*, 13:623, 2012. PMID: 23151179.
c. I. R. Baxter, **O. Vitek**, B. Lahner, B. Muthukumar, M. Borghi, J. Morrissey, M. L. Guerinot, D. E. Salt. "The leaf ionome as a multivariable system to detect plant's physiological status", *Proceedings of the National Academy of Sciences*, 105:12081, 2008. PMID: 18697928.
d. I. R. Baxter, J. N. Brazelton, D. Yu, Y. S. Huang, B. Lahner, E. Yakubova, Y. Li, J. Bergelson, J. O. Borevitz, M. Nordborg, **O. Vitek**, D. E. Salt. "A costal cline in sodium accumulation in *Arabidopsis thaliana* is driven by natural variation of the sodium transporter AtHKT1;1", *PLoS Genetics*, 6:e1001193, 2010. PMID: 21085628.

5. Inferential identification and relative quantification of metabolites:

Although the chemical diversity and experimental characteristics of metabolites differs from that of proteins, metabolomics also faces great challenges of statistical experimental design and analysis. To simultaneously identify and quantify the metabolites in ^{1}H Nuclear Magnetic Resonance (NMR) spectra (Ref. a), we viewed the observed spectra as a linear combination of signals from pure compounds available in an experimental database. We modeled these mixtures using a Bayesian hierarchical probability model, and converted the problem of identifying the metabolites that are present in the sample into a problem of variable selection. We utilized controlled experiments (Ref. b) to demonstrate that the approach is superior in accuracy and reproducibility to the alternatives that existed at that time. Our related work on relative metabolite quantification in Snapdragon flowers appeared in (Ref. c).

a. C. Zheng, S. Zhang, S. Ragg, D. Raftery, **O. Vitek**. "Identification and quantification of metabolites in 1H NMR spectra by Bayesian model selection", *Bioinformatics*, 27:1637, 2011. PMC3106181.
b. S. Zhang, C. Zheng, I. R. Lanza, K. S. Nair, D. Raftery, **O. Vitek**. "Interdependence of signal processing and analysis of urine 1H NMR spectra for metabolomics profiling", *Analytical Chemistry*, 81:6080, 2009. PMC2789356
c. J. K. Muhlemann, H. Maeda, C.-Y. Chang, P. San Miguel, I. Baxter, B. Cooper, M. A. Perera, B. J. Nikolau, **O. Vitek**, J. A. Morgan, N. Dudareva. "Developmental changes in the metabolic network of snapdragon flowers", *PLoS One*, 7(7):e40381, 2012. PMC3394800.

D. Additional Information: Research Support and/or Scholastic Performance

Ongoing Research Support

O. Vitek (PI): Chan-Zuckerberg foundation. 12/01/2019 - 11/30/2021
MSstats + Cardinal: Next Generation Statistical Mass Spectrometry in R.

O. Vitek (PI): NSF-BIO/DBI 1950412 05/01/2020-05/01/2021
REU Supplement to ABI Innovation: Scalable and Agile Analysis of Mass Spectrometry Experiments

N. Bandeira (PI), O. Vitek (subcontract): NIH-NLM-R01 1R01LM013115 06/01/2019 - 05/31/2023
MassIVE.quant: a curated and scalable community resource for quantitative proteomics

J. Vitek (PI), O. Vitek (co-PI): NSF-CISE/CCRI 1925507 10/01/2019 - 09/30/2022
ENS: Collaborative research: Enhancing R for scalability and deployment

O. Vitek (PI), K. Bemis (co-PI), J. Vitek (co-PI): NSF-BIO/DBI 1759736 08/01/2018-07/31/2021
ABI Innovation: Scalable and agile analysis of mass spectrometry experiments