

Rajarshi Guha
231 Informatics
901 E. 10th Street
Bloomington, IN 47408
Ph: (812) 856 0386

3209 E. 10th Street, Apt S6
Bloomington, IN 47408
Ph: (814) 404 5449

rguha@indiana.edu

Professional Experience

- 2007 - 2009 • Visiting Assistant Professor, School of Informatics, Indiana University
- 2006 - 2007 • Post-doctoral scholar working with Prof. D. Wild at Indiana University
- 2005 - 2006 • Post-doctoral scholar working with Prof. P. C. Jurs, Pennsylvania State University

Education

- 2001 – 2005 • Ph.D. (Chemistry), Department of Chemistry, The Pennsylvania State University.
- 1999 – 2001 • M.Sc. (Chemistry), Indian Institute of Technology, Kharagpur, India.
- 1995 – 1998 • B.Sc. (Chemistry), Presidency College, Calcutta, India.

Research Experience

- 2007 – ...
 - Applications of density of space measurements to compound selection in combinatorial libraries (in collaboration with Dr. J. Medina-Franco of the Torrey Pines Institute).
 - Exploring promiscuity and polypharmacology in PubChem bioassays using network models. Profiling compounds over the PubChem bioassay collection.
 - Characterization and use of structure-activity landscapes (in collaboration with Dr. J. H. Van Drie).
 - Virtual screening and QSAR modeling for anti-malarials (in collaboration with Prof. J.-C. Bradley of Drexel University).
 - Development of methods to estimate model domain applicability as applied to individual compounds and entire datasets (in collaboration with Dr. D. Stanton of Procter & Gamble).
 - Analysis of HTS cytotoxicity assay data (in collaboration with Dr. S. Schürer of Scripps, Florida).
 - Fast 3D searching and pharmacophore matching in very large compound databases.
- 2006 – 2007
 - Investigation of ensemble feature selection methods for the purpose of QSAR descriptor selection suitable for the simultaneous development of linear and non-linear models.
 - Development of a cheminformatics web-service and database enabled infrastructure. Integration of statistical environments and cheminformatics toolkits.
- 2005 – 2006
 - Applications of approximate nearest neighbor methods to the analysis of large molecular datasets and the use of the *R*-nearest neighbor technique to the spatial analysis of chemical spaces for diversity analysis (in collaboration with Dr. D. Dutta at the University of Southern California).

Research Experience (continued)

- Development of an automatic QSAR pipeline suitable for handling large datasets. The function of the pipeline is to serve as a framework for other projects including a generalized method to test model applicability (in collaboration with Procter & Gamble).
- Development of a tiered high-throughput filtering protocol based on consensus methods applied to HIV integrase inhibitors.

2001 – 2005

- Graduate research under Prof. Peter C. Jurs in the field of computer aided chemistry. The research consisted of both model and methodology development. Specific projects included:
- Development of methods to provide interpretability to neural network based QSAR models (in collaboration with Procter & Gamble).
- Investigation of methods to quantify the applicability of linear QSAR models to unknown molecules.
- Development of QSAR models to predict the activity of PDGFR tyrosine kinase inhibitors and artemisinin analogs.
- Development of a strategy to use a Kohonen self organizing map to create representative training, cross validation and prediction sets for QSAR studies.

Computing Experience

Languages

- Extensive experience with Java, Python, C and R. Familiar with C++, Fortran and Lisp

Platforms

- Experienced user of the OpenEye toolkit, familiar with Pipeline Pilot, KNIME. Comfortable on various OS platforms (Linux, OS X, Windows).

Contributions

- Major contributor to the Chemistry Development Kit, a Java toolkit for cheminformatics. Examples include contributions in method development (descriptors, pharmacophores, rigid alignments, I/O), build system, documentation and unit tests. Significant contributions to performance improvements.
- Various standalone utilities for SALI analysis, pharmacophore searching and descriptor calculations.
- Developed a set of cheminformatics and statistical web services. These exposed a variety of core cheminformatics functionality that were then used in various web applications and pipeline tools. The statistical web service were based on R and provide core methods (modeling and sampling) and also supported deployment of predictive models. Developed REST interfaces for many of the web services.
- Extensive experience with R for modeling and algorithm prototyping. Published and maintain packages that incorporate cheminformatics within R, handle molecular fingerprint data and interface R to the PubChem bioassay and compound collections. Extended support for PMML in R.
- The use of SQL databases (primarily Postgres) to store and manipulate chemical information, using cheminformatics cartridges. These were interfaced via web service frontends. The primary focus was on the use of spatial indexing for fast shape searching. Also implemented mirroring schemes for a local PubChem (compound and bioassay) mirror.

Publications

- Singh, N.; **Guha, R.**; Guilianotti, M.; Houghten, R.; Medina-Franco, J.L.; “Chemoinformatic Analysis of Drugs, Natural Products, Molecular Libraries Small Molecule Repository and Combinatorial Libraries”, *J. Chem. Inf. Model.*, **2009**, in press
- **Guha, R.**; Gilbert, K.; Fox, G.C.; Pierce, M.; Wild, D.; Yuan, H.; “Advances in Cheminformatics Methodologies and Infrastructure to Support the Data Mining of Large, Heterogeneous Chemical Datasets”, *Curr. Comp. Aid. Drug Des.*, **2008**, submitted
- Bajorath, J.; Peltason, L.; Wawer, M.; **Guha, R.**; Lajiness, M.S.; van Drie, J.H.; “Navigating Structure Activity Landscapes”, *Drug Discov. Today*, **2008**, submitted
- **Guha, R.**; Wiggins, G.D.; Wild, D.J.; Baik, M.H.; Pierce, M.E.; Fox, G.C.; “Improving Usability and Accessibility of Cheminformatics Tools for Chemists Through Cyberinfrastructure and Education”, *Cheminformatics*, **2008**, in press
- **Guha, R.**; Van Drie, J.H.; “Pharmacophore Representation and Searching”, *CDK News*, **2008**, ASAP
- **Guha, R.**; Van Drie, J.H.; “Assessing How Well a Modeling Protocol Captures a Structure-Activity Landscape”, *J. Chem. Inf. Model.*, **2008**, *48*, 1716–1728
- **Guha, R.**; Van Drie, J.H.; “The Structure-Activity Landscape Index: Identifying and Quantifying Activity-Cliffs”, *J. Chem. Inf. Model.*, **2008**, *48*, 646–658
- **Guha, R.**; “A Flexible Web Service Infrastructure for the Development and Deployment of Predictive Models”, *J. Chem. Inf. Model.*, **2008**, *48*, 456–464
- **Guha, R.**; “On the Interpretation and Interpretability of QSAR Models”, *J. Comp. Aid. Molec. Des.*, **2008**, *22*, 857–871
- **Guha, R.**; Schürer, S.C.; “Utilizing High Throughput Screening Data for Predictive Toxicology Models: Protocols and Application to MLSCN Assays”, *J. Comp. Aid. Molec. Des.*, **2008**, *22*, 367–384
- Willighagen, E.L.; O’Boyle, N.; Gopalakrishnan, H.; Jiao, D.; **Guha, R.**; Steinbeck, C.; Wild, D.J.; “Userscripts for the Life Sciences”, *BMC Bioinformatics*, **2007**, *8*, 487
- Wang, H.; Klinginsmith, J.; Dong, X.; Lee, A.; **Guha, R.**; Wu, Y.; Crippen, G.; Wild, D.J.; “Chemical Data Mining of the NCI Human Tumor Cell Line Database”, *J. Chem. Inf. Model.*, **2007**, *47*, 2063–2076
- **Guha, R.**; Dutta, D.; Chen, T.; Wild, D.J.; “Counting Clusters Using R-NN Curves”, *J. Chem. Inf. Model.*, **2007**, *47*, 1308–1318
- Dong, X.; Gilbert, K.; **Guha, R.**; Heiland, R.; Kim, J.; Pierce, M.; Fox, G.; Wild, D.J.; “A Web Service Infrastructure for Cheminformatics”, *J. Chem. Inf. Model.*, **2007**, *47*, 1303–1307
- Dutta, D.; **Guha, R.**; Chen, T.; Wild, D.J.; “Ensemble Feature Selection: Consistent Descriptor Subsets for Multiple QSAR Models”, *J. Chem. Inf. Model.*, **2007**, *47*, 989–997
- **Guha, R.**; “Chemical Informatics Functionality in R”, *J. Stat. Soft.*, **2007**, *18*,
- **Guha, R.**; Dutta, D.; Jurs, P.C.; Chen, T.; “Local Lazy Regression: Making Use of the Neighborhood to Improve QSAR Predictions.”, *J. Chem. Inf. Model.*, **2006**, *46*, 1836–1847
- **Guha, R.**; Dutta, D.; Jurs, P.C.; Chen, T.; “R-NN Curves: An Intuitive Approach to Outlier Detection Using a Distance Based Method”, *J. Chem. Inf. Model.*, **2006**, *46*, 1713–1722
- **Guha, R.**; Howard, M.T.; Hutchison, G.R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E.L.; “The Blue Obelisk–Interoperability in Chemical Informatics.”, *J. Chem. Inf. Model.*, **2006**, *46*, 991–998

Publications (continued)

- Dutta, D.; **Guha, R.**; Jurs, P.C.; Chen, T.; “Scalable Partitioning and Exploration of Chemical Spaces using Geometric Hashing”, *J. Chem. Inf. Model.*, **2006**, *46*, 321–333
- **Guha, R.**; “Generating, Using and Visualizing Molecular Information in R”, *R News*, **2006**, *3*, 28–33
- Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; **Guha, R.**; Willighagen, E.L.; “Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics”, *Curr. Pharm. Des.*, **2006**, *12*, 2110–2120
- **Guha, R.**; Stanton, D.T.; Jurs, P.C.; “Interpreting Computational Neural Network QSAR Models: A Detailed Interpretation of the Weights and Biases”, *J. Chem. Inf. Model.*, **2005**, *45*, 1109–1121
- **Guha, R.**; Jurs, P.C.; “Interpreting Computational Neural Network QSAR Models: A Measure of Descriptor Importance”, *J. Chem. Inf. Model.*, **2005**, *45*, 800–806
- **Guha, R.**; Jurs, P.C.; “Determining the Validity of a QSAR Model—A Classification Approach”, *J. Chem. Inf. Model.*, **2005**, *45*, 65–73
- **Guha, R.**; “Using R to Provide Statistical Functionality for QSAR Modeling in CDK to Provide Statistical Functionality for QSAR Modeling in CDK”, *CDK News*, **2005**, *2*, 7–13
- **Guha, R.**; Jurs, P.C.; “Development of Linear, Ensemble, and Nonlinear Models for the Prediction and Interpretation of the Biological Activity of a Set of PDGFR Inhibitors.”, *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 2179–2189
- **Guha, R.**; Jurs, P.C.; “The Development of QSAR Models To Predict and Interpret the Biological Activity of Artemisinin Analogues”, *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 1440–1449
- **Guha, R.**; Serra, J.R.; Jurs, P.C.; “Generation of QSAR Sets with a Self-Organizing Map.”, *J. Mol. Graph. Model.*, **2004**, *23*, 1–14

Books and Book Chapters

- **Guha, R.**; *A Handbook of Cheminformatics Algorithms*, Eds. Faulon, J.-L.; Bender, A.; chapter Open Source Cheminformatics Software & Database Technologies, John Wiley & Sons, New York, NY. submitted
- **Guha, R.**; Bender, A. (Eds.); *Computational Approaches in Cheminformatics and Bioinformatics*, in preparation, John Wiley & Sons
- Fox, G.C.; **Guha, R.**; McMullen, D.F.; Mustacoglu, A.F.; Pierce, M.; Topcu, A.E.; Wild, D.J.; “Web 2.0 for Grids and e-Science” in *Proc. INGRID 2007 - Instrumenting the Grid*, **2007**

Invited Presentations

- | | |
|---------------|---|
| December 2008 | • <i>Networks - More Than Just Pretty Pictures</i> at Vertex Pharmaceuticals, Cambridge, MA. |
| December 2008 | • <i>A Network View of Structure-Activity Landscapes</i> at Drexel University, Philadelphia, PA. |
| December 2008 | • <i>Networks - More Than Just Pretty Pictures</i> at the CGB Roundtable, Indiana University, Bloomington. IN. |
| October 2008 | • <i>Structure-Activity Relationships and Networks: A Generalized Approach to Exploring Structure-Activity Landscapes</i> at the NIH Chemical Genomics Center, Rockville, MD. |
| August 2008 | • <i>Defining and Using Structure Activity Landscapes</i> at Simulations Plus, Lancaster, CA. |

Invited Presentations (continued)

- June 2008 • *Data Mining in Drug Discovery* at the 31st National Medicinal Chemistry Symposium, Pittsburgh, PA.
- May 2008 • *Aspects of Model Quality & Applicability* at Pfizer Inc., Groton, CT.
- March 2008 • *The Structure Activity Landscape Index: Visualization and Applications* at Eli Lilly & Co., Indianapolis, IN
- January 2008 • *Characterizing and Utilizing Structure Activity Landscapes* at Abbott Laboratories, Chicago, IL.
- August 2007 • *Characterizing the Density of Chemical Spaces and its Use in Outlier Analysis and Clustering* at the Novartis Institute for Biomedical Research, Cambridge, MA.
- May 2007 • *The Development and Deployment of Predictive Toxicology Models* at the MLSCN Steering Committee Meeting, Philadelphia, PA.
- February 2007 • *Making the Most of Predictive Models* at the Openeye Cup 8, Santa Fe, NM.
- January 2007 • *The Role of the Neighborhood in QSAR Modeling and Cheminformatics* at the Dept. of Pharmaceutical Technology, Jadavpur University, Calcutta.
- August 2006 • *Chemical Spaces: Modeling, Exploration & Understanding* at the CICC-MACE-Lilly Cheminformatics Workshop, Indianapolis, IN.
- August 2006 • *Writing & Using Web Services* at the CICC-MACE-Lilly Cheminformatics Workshop, Indianapolis, IN.
- April 2006 • *Navigating Molecular Haystacks: Tools & Applications* at the School of Informatics, Indiana University, Bloomington, IN.
- March 2006 • *Computational Tools & Protocols For Drug Discovery* at the School of Pharmacy, University of Maryland, Baltimore, MD.
- September 2005 • *Extending Validation and Providing Interpretability for QSAR Models* at the Jet Propulsion Laboratory, Pasadena, CA.
- February 2005 • *The Validation and Interpretation of QSAR Models* at NCI, Frederick.

Contributed Presentations

- December 2008 • Talk titled *SQMD: Architecture for Scalable, Distributed Database System built on Virtual Private Servers* (with K. Kim and M. Pierce) at the 4th Intl. Conf on e-Science, Indianapolis, IN.
- December 2008 • Talk titled *Open Drug Discovery in Malaria Research* (with J.-C. Bradley, P. Rosenthal, K. Mirza and J. Gut) at the 4th Intl. Conf on e-Science, Indianapolis, IN.
- August 2008 • Talk titled *PubChem Bioassays as a Source of Polypharmacology* (with B. Chen and D. J. Wild) at 236th ACS National Meeting at Philadelphia, PA.
- June 2008 • Talk titled *Characterizing the Structure Activity Landscape and Implications for Predictive Modeling and Molecular Representations* at the 5th Indo-US Workshop on Mathematical Chemistry at Duluth, MN.
- April 2008 • Talk titled *Combining Global and Local Approaches to Model Domain Applicability* (with D. Stanton) at the 235th ACS National Meeting at New Orleans, LA.

Contributed Presentations (continued)

- April 2008 • Talk titled *I Don't Care Where My Data and Methods Are: A Web-Service Approach for Distributed Access to Methods, Data and Models* at the 235th ACS National Meeting at New Orleans, LA.
- August 2007 • Talk titled *Random Forest Ensembles Applied to MLSCN Screening Data for Prediction and Feature Selection* (with S. Schürer) at the 234th ACS National Meeting at Boston, MA.
- August 2007 • Poster titled *Using Semantic Information for Feature Selection* at the Gordon Research Conference on Computer Aided Drug Design at Tilton, NH.
- May 2007 • Talk titled *Integrating R with the CDK: Enhanced Chemical Data Mining* at the Central Regional ACS Meeting, Covington, KY.
- March 2007 • Talk titled *Spectral Clustering of Chemical Datasets* at the 233rd ACS National Meeting at Chicago, IL.
- March 2007 • Talk titled *A Tiered Screen Protocol for the Discovery of Structurally Diverse HIV Integrase Inhibitors* at the 233rd ACS National Meeting at Chicago, IL.
- September 2006 • Talk titled *Local Lazy Regression: Making Use of the Neighborhood to Improve QSAR Predictions* at the 232nd ACS National Meeting at San Francisco, CA.
- September 2006 • Poster titled *R-NN Curves: A Method for Diversity Analysis and Cluster Identification* at the 232nd ACS National Meeting at San Francisco, CA.
- March 2006 • Talk titled *Scalable Partitioning & Exploration of Chemical Spaces Using Geometric Hashing* at the 231st ACS National Meeting at Atlanta, GA.
- February 2006 • Poster titled *A Tiered Screening Protocol for the Discovery of Structurally Diverse HIV Integrase Inhibitors* at the 2nd Annual Computation Day, Pennsylvania State University, University Park, PA.
- August 2005 • Talk titled *Integrating R with the CDK for QSAR Modeling* at the 230th ACS National Meeting at Washington D.C.
- August 2005 • Poster titled *Applications of Spectral Clustering to Chemical Datasets* at the 2005 Gordon Research Conference on Computer Aided Drug Design, Tilton, NH.
- March 2005 • Talk titled *The Interpretation of Neural Network QSAR Models Using Weights & Biases* at the 229th ACS National Meeting at San Diego, CA.
- August 2004 • Poster titled *How Well Can a QSAR Model Handle New Datasets?* at the 228th ACS National Meeting at Philadelphia, PA.
- August 2003 • Poster titled *Generation of QSAR Sets With A Self Organizing Map* at the 226th ACS National Meeting at New York City, NY.

Teaching Experience

- Fall 2008 • Instructor for 'Information Infrastructure II' and 'Advanced Seminar in Cheminformatics (Predictive Modeling)'
- Spring 2008 • Instructor for 'Programming for Life Sciences'
- Fall 2007 • Instructor for 'Information Infrastructure II'
- Spring 2002, Spring 2003 • TA for 'Experimental Physical Chemistry'
- Fall 2004 & Spring 2005 • TA for 'Chemical Principles'
- Fall 2001 • TA for 'Chemical Principles'

Funding

- Submitted (NSF) • “Chemistry and Modeling Crowdsourcing using Open Notebook Science”, (Co-PI, \$223,000 requested)
- Submitted (NIH) • “Continued Maintenance and Development of the Chemistry Development Kit (CDK)”, (Co-PI, \$436,000 requested)
- Submitted (NIH) • “A Framework for Data Mining and Network Modeling of Bioassay Datasets”, (PI, \$655,000 requested)
- 2007 - 2008 • P20 HG003894-02S1 - “Chemical Informatics Cyberinfrastructure”. (Prof. G.C. Fox, PI)

Awards

- 2007 • Jacques-Emile Dubois Grant
- 2006 • Best Poster Award, ICS Computation Day, Pennsylvania State University.
- 2005 • Graduate Student Travel Award, Pennsylvania State University.
- 2004 • Dalalian Fellowship, Pennsylvania State University.
- 2004 • ACS COMP Division CCG Excellence Award.
- 2002 • Braddock Graduate Fellowship, Pennsylvania State University.
- 2001 • Roberts Graduate Fellowship, Pennsylvania State University.

Service

- Program Chair (2009-2010) for the ACS CINF division
- Assistant Program Chair (2007-2009) for the ACS CINF division
- Co-organizer and co-chair for the symposium titled *Systems Chemical Biology: Integrating Chemistry and Biology for Network Models* at the Fall 2008 ACS National Conference
- Organizer & chair for the symposium titled *ADAPT'ing to Retirement: A Symposium Honoring Peter C. Jurs* at the Spring 2008 ACS National Conference
- Organizer & chair for the symposium titled *Cheminformatics Techniques in Bioinformatics* at the Fall 2007 ACS National Conference
- Member of the editorial boards of *Chemistry Central Journal* and *J. Cheminformatics*
- Reviewer for *Nat. Rev. Drug Discov.*, *J. Chem. Inf. Model.*, *QSAR. Comb. Sci.*, *J. Comp. Aided Molec. Des.*, *Chemistry Central Journal*, *Statistical Analysis & Data Mining*, *Chemosphere* and *Molecular Diversity*