Yingyao Zhou, Ph.D.

yingyao.zhou@gmail.com San Diego, CA
(858) 336-7089

PROFILE

- Over 20 years' experience in applying data science and data engineering to drug discovery.
- Key architect in an application ecosystem covering the complete pre-clinical research spectrum.
- Enthusiastic leader for applying machine learning to drug discovery challenges.
- Strong motivation to improve public health through contributing open-access bioinformatics tools.

PROFESSIONAL EXPERIENCE

Apr 1999-Present Novartis Institute of Biomedical Research (NIBR)
Genomics Institute of the Novartis Research Foundation (GNF)
San Diego, CA

Director, Data Science 2022-present

- Lead the global imaging and deep learning team (12 Ph.D. data scientists) at Cell Biology Technology (CBT), NIBR.
- Apply deep learning to image segmentation and clustering, compound target prediction, and compound activity imputation.

Director, Data Science & Data Engineering 2016-2022

- Responsible for building up Data Science (machine learning, bio/cheminformatics, imaging, and deep learning) and Data Engineering (big data and cloud computing) capabilities at GNF.
- Mentor and retain talents; lead 12 Ph.D.-level data scientists.
- In-depth machine learning knowledge (online book). Championed the winning team in the Novartis Data Science AI Challenge 2019 to predict the success rate of clinical trials.
- Developed a popular bioinformatics web site Coronascape for public COVID research.

Director, Informatics & IT 2005-2016

- Headed GNF Lead Discovery Database (LDDB) development. LDDB is an in-house developed comprehensive drug discovery informatics platform that consists of modules for compound management, HTS screening, lead tracking, program management, analytical chemistry, medicinal chemistry and pharmacology.
- Developed Metascape, an high-traffic open-access bioinformatics tool that enables experimental biologists to analyze OMICs-based gene lists with a single click. Metascape empowers 5000 gene list analyses/day for biomedical research community (Nature Communication, 100 citations/month).
- Transformed IT function (20 members) to provide robust and research-friendly data infrastructure for the institute.

Bioinformatics Scientist 1999-2005

- Contributed to the successful antimalarial drug development by applying data science to analyze parasite life cycle expression data, regulatory motifs, genetic variants, and drug screening. Over 27 top-tier peer-reviewed publications.
- Developed multiple high-impact bioinformatics algorithms: Redundant siRNA Activity (RSA) analysis algorithm for genomics screen hit selection (Nature Method, citation 341); Gene Enrichment Motif Search (GEMS) algorithm (BMC Genomics, citation 143); Ontology-based Pattern Identification (OPI) algorithm for gene expression analysis (Bioinformatics, citation 97); Match-only Integral Distribution (MOID) algorithm (BMC Bioinformatics, citation 84).

EDUCATION

- New York University, New York — Ph.D. Biophysics
- New York University, New York — M.S. Computer Science
- Fudan University, Shanghai — B.S. Physics

CONSULTANCIES

- Wildcat Discovery Technologies, San Diego, CA — 2008-2009
- Molsoft, LLC, San Diego, CA — 2001-2002

TRAINING AND SKILLS

Machine Learning/Deep Learning

- Coursera: Machine Learning Specialization, Deep Learning Specialization, How to Win a Data Science Competition, Bayesian Methods for Machine Learning, Probabilistic Graphics Models, TensorFlow in Practice Specialization.
- Fluent in Python (numpy, pandas, scikit-learn, matplotlib), Pytorch, Fast.ai, TensorFlow, Jupyter Notebook, etc.

Biology & Bioinformatics

- Fundamentals of Immunology Specialization (Rice University)
- Bioinformatics algorithms (UCSD Extension)
- Ph.D. research in protein structure prediction and compound docking.
- Deep know-how about preclinical drug discovery process, including compound management, high-throughput screening, high-content imaging, flow cytometry, hit identification, analytical chemistry, medicinal chemistry, ADME/Tox profiling, pharmacokinetics.

Computer science

- Rigorous computer science training. Fluent in Linux, high-performance computing, SQL language, MySQL, Oracle. Hands-on experience in using C#, Perl, JavaScript, C++, Java, R, C.
- Maintaining Metascape web site using cloud-computing technology including AWS, docker, and Kubernetes.

HONORS AND AWARDS

- The winning team of the first Novartis Data Science & Artificial Intelligence Competition, 2020

- Referee for Bioinformatics, BMC Bioinformatics, BMC Biotechnology, BMC Genomics, BioTechniques, International Journal for Parasitology, etc.
- "The First Matthew Smosna Prize" for Excellence in Computer Science, Courant Institute of Mathematical Science, 1998
- Mini-CUSPEA Scholar (Exam Year 1993, https://en.wikipedia.org/wiki/CUSPEA).
- "The T.D.Lee Physics Golden Award", of which the founder, Prof. T.D.Lee is a 1957 Nobel Physics Prize laureate, 1993

BOOK CHAPTERS

1. Mining high-throughput screening data by novel knowledge-based optimization analysis. Yan, SF, King FJ, Chanda SK, Caldwell JS, Winzeler EA, Zhou Y. Pharmaceutical Data Mining: Approaches and Applications for Drug Discovery, Edited by Konstantin V. Balakin. John Wiley & Sons, Inc. (2010) 205-233. 2009. [ISBN: 978-0-470-19608-3]
2. Efficient Stochastic Global Optimization for Protein Structure Prediction, Yingyao Zhou and Ruben Abagyan, Rigidity Theory and Application, edited by M.F. Thorpe and P. M. Duxbury (Plenum Publishing), (1999) 345-356.
3. Efficient Stochastic Global Optimization for Protein Structure Prediction, Yingyao Zhou and Ruben Abagyan, Inquiry Program into the Fields of Simulation of Biological Functions – Computational Biology, March 1998, The Science and Technology Advancement Association, Japan.

PATENTS

1. Method and system for enterprise data access, annotation and sharing (US 2006/0200453 A1). Andrey Santrosyan and Yingyao Zhou.
2. A Novel Statistical Approach for Primary High-Throughput Screening Hit Selection (US 2006/0200315 A1). Yingyao Zhou, S. Frank Yan, Hayk Asatryan.

SOFTWARE DEVELOPED

1. Metascape – a gene annotation & analysis resource
   https://metascape.org
2. Coronascape – a central gene list analysis resource for COVID-19 research
   https://coronascape.org
3. Author of the Perl CPAN Data::Table.pm package
   https://metacpan.org/pod/Data::Table

SELECTED PUBLICATIONS (74 Total)

<u>Machine Learning Applications</u>

1. Predicting Drug Approvals: The Novartis Data Science and Artificial Intelligence Challenge. Siah KW, Kelley N, Ballerstedt S, Holzhauer B, Lyu T, Mettler D, Sun S, Wandel S, Zhong Y, Zhou B, Pan S, Zhou Y, Lo AW. Cell Patterns. 2021. Jul 21;2(8):100312. [PMID: 34430930]

2. A cell-level quality control workflow for high-throughput image analysis. Qiu M, Zhou B, Lo F, Cook S, Chyba J, Quackenbush D, Matzen J, Li Z, Mak PA, Chen K, Zhou Y. BMC Bioinformatics. 2020 21(1):280. [PMID: 32615917]

Data Science Applications (Bioinformatics & Cheminformatics)

1. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C, Chanda SK. Nature Communications. 2019 10(1):1523. [PubMed: 30944313] (>3000 citations)
2. Human host factors required for influenza virus replication. König R, Stertz S, Zhou Y, Inoue A, Hoffmann HH, Bhattacharyya S, Alamares JG, Tscherne DM, Ortigoza MB, Liang Y, Gao Q, Andrews SE, Bandyopadhyay S, De Jesus P, Tu BP, Pache L, Shih C, Orth A, Bonamy G, Miraglia L, Ideker T, García-Sastre A, Young JA, Palese P, Shaw ML, Chanda SK. Nature. 2009 Dec 21. [PMID: 20027183]
3. Global analysis of host-pathogen interactions that regulate early stage HIV-1 replication. König R, Zhou Y, Elleder D, Diamond TL, Bonamy GM, Irelan JT, Chiang CY, Tu BP, De Jesus PD, Lilley CE, Seidel S, Opaluch AM, Caldwell JS, Weitzman MD, Kuhen KL, Bandyopadhyay S, Ideker T, Orth AP, Miraglia LJ, Bushman FD, Young JA, Chanda SK. Cell 2008, 135:49-60. [PMID: 18854154]

Drug Discovery – Antimalarial Research

1. Open-source discovery of chemical leads for next-generation chemoprotective antimalarials. Antonova-Koch Y, Meister S, Abraham M, Luth MR, Ottilie S, Lukens AK, Sakata-Kato T, Vanaerschot M, Owen E, Jado JC, Maher SP, Calla J, Plouffe D, Zhong Y, Chen K, Chaumeau V, Conway AJ, McNamara CW, Ibanez M, Gagaring K, Serrano FN, Eribez K, Taggard CM, Cheung AL, Lincoln C, Ambachew B, Rouillier M, Siegel D, Nosten F, Kyle DE, Gamo FJ, Zhou Y, Llinás M, Fidock DA, Wirth DF, Burrows J, Campo B, Winzeler EA. Science. 2018 Dec 7;362(6419). [PubMed: 30523084].
2. Imaging of Plasmodium Liver Stages to Drive Next-Generation Antimalarial Drug Discovery. Meister S, Plouffe DM, Kuhen KL, Bonamy GM, Wu T, Barnes SW, Bopp SE, Borboa R, Bright AT, Che J, Cohen S, Dharia NV, Gagaring K, Gettayacamin M, Gordon P, Groessl T, Kato N, Lee MC, McNamara CW, Fidock DA, Nagle A, Nam TG, Richmond W, Roland J, Rottmann M, Zhou B, Froissard P, Glynne RJ, Mazier D, Sattabongkot J, Schultz PG, Tuntland T, Walker JR, Zhou Y, Chatterjee A, Diagana TT, Winzeler EA. Science, 2011 Nov 17. [PMID: 22096101]

Data Engineering Applications

1. Chemical-text hybrid search engines. Zhou Y, Zhou B, Jiang S, King FJ. J Chem Inf Model. 2010, 50:47-54. [PMID: 20047295]
2. Large-scale annotation of small-molecule libraries using public databases. Zhou Y, Zhou B, Chen K, Yan SF, King FJ, Jiang S, Winzeler EA. J. Chem. Inf. & Model. 2007, 47:1386-1394. [PMID: 17608408]

Algorithm Development

1. Probability-based approach for the analysis of large-scale RNAi screens. König R, Chiang CY, Tu BP, Yan SF, DeJesus PD, Romero A, Bergauer T, Orth A, Krueger U, Zhou Y, Chanda SK. Nature Method. 2007. 4:847-849. [PMID: 17828270]
2. In silico gene function prediction using ontology-based pattern identification. Zhou Y, Young JA, Santrosyan A, Chen K, Yan SF, Winzeler EA. Bioinformatics, 2005, 191:1196-1203. [PMID: 15531612]