A blueprint for human whole-cell modeling: Supplementary Note S1 and Figures S1-S21

Balázs Szigeti[1,2], Yosef D. Roth[1,2], John A. P. Sekar[1,2], Arthur P. Goldberg[1,2], Saahith C. Pochiraju[1,2], Jonathan R. Karr[1,2,*]

[1]Icahn Institute and [2]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

[*]Corresponding author: Karr, Jonathan R (karr@mssm.edu)

Table of contents

Note S1 and Figures S1–S2. Biomodeling community survey methodology

A. Survey questions

The survey included three sections: (a) multiple response questions about the respondent's demographics including their sector, field, and country; (b) multiple response questions about the respondent's current research goals and methods and the major challenges to their research; and (c) open-ended questions about new goals, methods, tools, and resources that could help advance the respondent's research. The multiple response questions also included options to enter custom responses. Below is a list of the survey questions and suggested responses.

Demographic information

- What sector do you work in?
  Academic research (e.g., university, non-profit, government), Industry research and development, Research sponsor (e.g., government, non-profit), and/or Other.
- What is your research field?
  Bioinformatics, Biology, Chemical engineering, Chemistry, Computational neuroscience, Computational structural biology, Computer science, Ecology, Electrical engineering, Mathematical biology, Mathematics, Mechanical engineering, Medicine, Multicellular modeling, Multiscale modeling, Statistics, Systems biology, Systems pharmacology, Physics, Physiological modeling, and/or Other.
- Where do you work?
  Australia, Austria, Canada, France, Germany, India, Israel, Italy, Japan, Korea, Netherlands, Spain, Switzerland, UK, USA, and/other Other.

Research goals, methods, and major bottlenecks

- What is the primary focus on your research?
  Methods development, Modeling and simulation, Software development, Standards development, Wet lab experimentation, and/or Other.
- What is the biological focus of your research?
  Aging, Cell cycle regulation, Individual molecules and molecular interactions, Metabolism, Microbial communities, Neurophysiology, Signaling, Tissues and higher-order structures, Transcription, Transcriptional regulation, Translation, My work is applicable to many areas of biology and has no specific biological focus, and/or Other.
- If you use models in your research, how large are the models that you typically use in your research?
  I do not use models in my research, I do not directly use model in my research and collaborate with other researchers who build and simulate models, Small (a few model elements and parameters), Large, but uniform (many model elements and parameters, but drawn from a few types), Large, and diverse (many model elements and parameters, drawn from many types), and/or Other.
- If you use models in your research, what types of data sources do you use?

Bioinformatics prediction tools such as PSORT, Databases, Text of published articles, Figures and/or tables of published articles, Supplementary materials of published articles, Other models, and/or Other.

- If you use models in your research, which data sources do you use?
  Array Express; BioCyc; BioModels; BioNumbers; BRENDA; CellML Model Repository; ECMDB, HMDB, YMDB; Gene Expression Omnibus (GEO); Google; Google Scholar; KEGG; Pathway Commons; Pax-DB; PubMed; Reactome; SABIO-RK; UniProt; and/or Other.
- If you use models in your research, which mathematical representations and simulation algorithms do you use most frequently?
  Agent-based modeling, Bayesian networks, Boolean/logical networks, Constraint-based modeling, Ordinary differential equations, Partial differential equations, Rule-based modeling, Stochastic simulation, and/or Other.
- If you use models in your research, which tools do you most frequently use to build and/or simulate models?
  BioUML; Cell Collective; Cell Designer; CellNOpt; COBRA toolbox for MATLAB; COBRApy; CompuCell3D; COPASI; E-Cell; FAME; GENESIS; Gepasi; iBioSim; iDynoMiCS; JSim; JSW Online; libRoadRunner; MASON; MCell; Mesa; MOOSE; NetLogo; Neuron; OpenCOR; PSICS; PyNN; PySB; RAVEN; RuleBender; SEEK; semanticSBML; SemSim; SIMMUNE; Spreadsheet editor such as Excel; Text editor such as Atom, NotePad++, etc.; Tellurium; VCell; and/or Other.
- If you use models in your research, which languages do you most frequently use to represent models?
  BioNetGen, CellML, Kappa, MML, NeuroML, SBML, and/or Other.
- If you use models in your research, which resources do you most frequently use to distribute models?
  BioModels, CellML Model Repository / Physiome Model Repository, FigShare, GitHub, JWS Online, Open Source Brain, SEEK, SimTk, Supplementary materials of your publications, Your website, and/or Other.
- Which programming languages do you most frequently use in your research?
  C / C++, Fortran, Java, JavaScript, MATLAB, Perl, Python, R, and/or Other.
- What additional tools do you frequently use in your research?
  Free response answers
- What are the most time-consuming aspects of your research?
  Data aggregation to support model building, Model design, Simulation, Model calibration, Model validation, Developing new methods and software tools, and/or Other.
- What do think are the main bottlenecks to building more predictive models?
  Insufficient and/or inaccessible data, Inadequate tools to design models, Inadequate languages to describe and communicate models, Inadequate tools to simulate models, Inadequate tools to calibrate models, Inadequate tools to verify models, Inadequate standards and/or incompatibility among models, Inadequate forums to interact with other modelers, and/or Other.

Methods, tools, and resources needed to advance biomodeling

- Given access to improved data and tools, what biomedical problems would you like to use models to solve (e.g. designer microorganisms, personalized cancer therapy)?

- What additional databases, methods, tools, and/or standards would help accelerate your research?
- What additional meetings, courses, and/or training opportunities would help accelerate your research?
- Are there any additional thoughts that you would like to share?

B. Survey invitations

Beginning on July 25, 2017, we invited 542 individual scientists from a broad range of biomodeling and related experimental disciplines (Figure S1) from across the world (Figure S2) to participate in a survey of the goals, bottlenecks, and needs of biomodeling. We also sent survey invitations to three community email lists. In addition, we encouraged scientists to invite additional colleagues to participate in the survey.

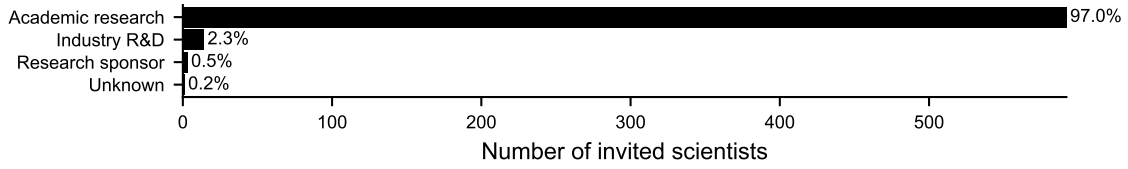

Figure S1. Sectors of the invited scientists.

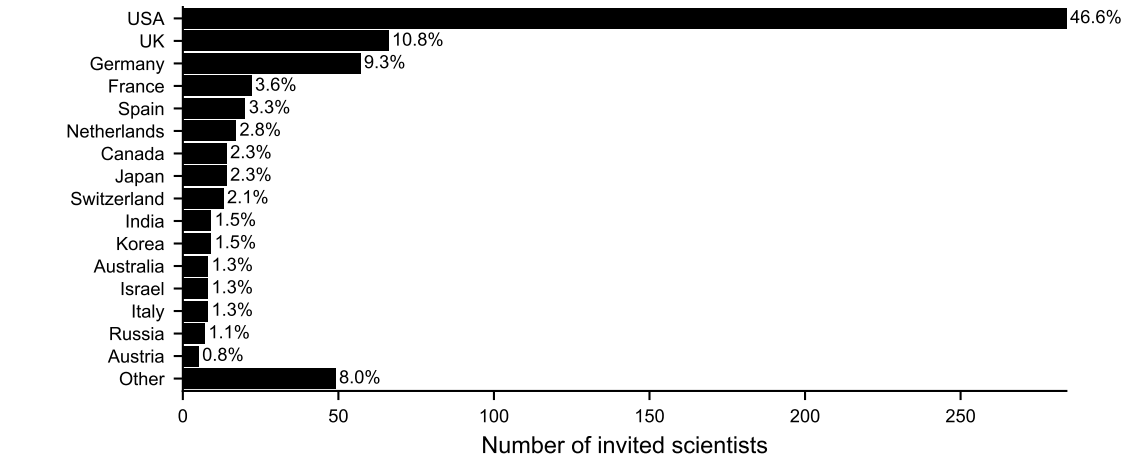

Figure S2. Countries of the invited scientists. Other: Brazil (4), Luxembourg (4), New Zealand (4), Portugal (4), Denmark (3), Norway (3), Poland (3), Singapore (3), Belgium (2), Columbia (2), Iran (2), Ireland (2), Mexico (2), South Africa (2), Sweden (2), Chile (1), China (1), Czech Republic (1), Hong Kong (1), Hungary (1), Iceland (1) and Turkey (1).

We selected the individual scientists from Europe PMC [1]; the NSF Award Search (URL: https://www.nsf.gov/awardsearch); community, conference, university websites; and our own personal contacts. We gathered contact information from Europe PMC by (a) using the Europe PMC web

service to retrieve contact information for authors who published articles about agent-based modeling, computational neuroscience, computational systems biology, mathematical biology, or multiscale modeling or related experimental studies; (b) grouping authors by their emails and ORCIDs [2]; (c) filtering out authors with few papers and few citations; and (d) manually filtering out irrelevant authors. Similarly, we gathered contact information from the NSF Award Search by (a) using the "Advanced Search" to retrieve contact information for investigators who have NSF awards within the last five years for agent-based modeling, computational neuroscience, computational systems biology, mathematical biology, or multiscale modeling or related experimental studies and (b) manually filtering out irrelevant investigators. We gathered contact information from modeling community, conference, university websites by (a) using Google to search for relevant websites and (b) manually identifying relevant scientists from each website. Notably, we gathered contact information for several computational neuroscientists from the "Computational Neuroscience on the Web" website (URL: http://home.earthlink.net/ perlewitz). Lastly, we merged the contact information that we collected from these sources and filtered out duplicate contacts.

We sent survey invitations to three community email lists: the COmputational Modeling in BIology NEtwork (COMBINE) discussion Google group (URL: http://co.mbine.org/comm), the Society for Mathematical Biology Digest (URL: http://www.smb.org/digest), and the Systems Biology Markup Language (SBML) discussion Google group (URL: http://sbml.org/Forums).

Figures S3–S21. Biomodeling community survey results

214 scientists completed the survey. Below are summaries of the scientists who participated in the survey and their responses to the survey questions.

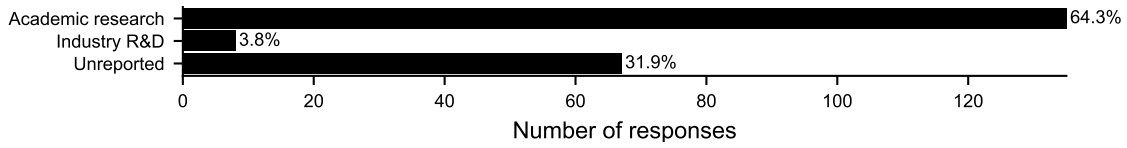

Figure S3. What sector do you work in?

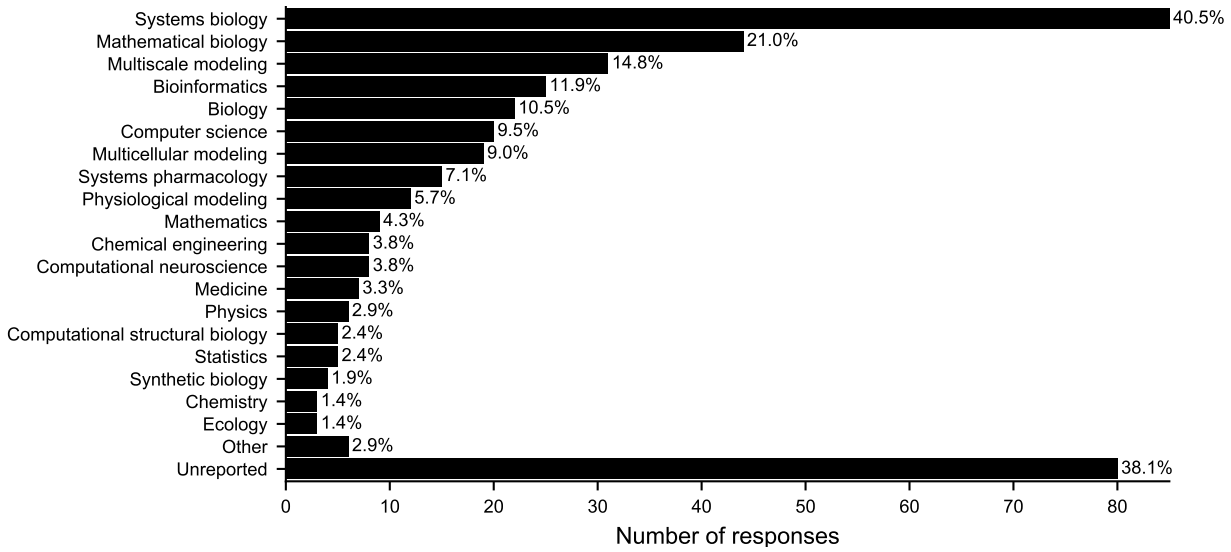

Figure S4. What is your research field? Other: Electrical engineering (2), Mechanical engineering (1), Metabolic modeling (1), Neuroscience (1) and Systems medicine (1).

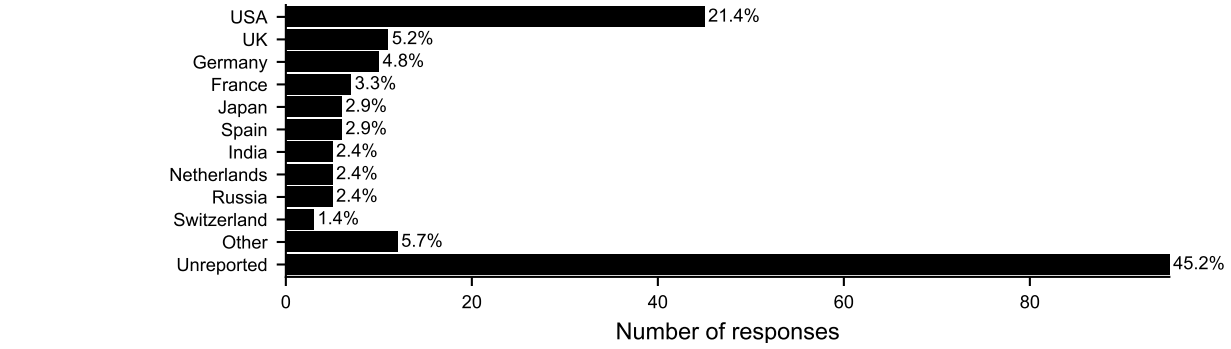

Figure S5. Where do you work? Other: Canada (2), New Zealand (2), Belgium (1), Brazil (1), Chile (1), Denmark (1), Iceland (1), Italy (1), Korea (1) and Norway (1).

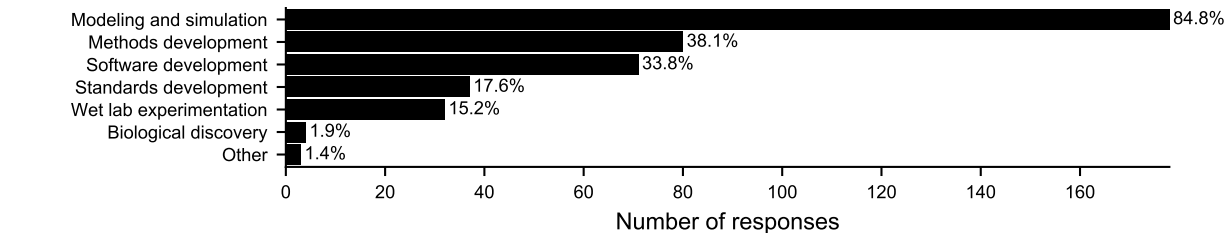

Figure S6. What is the primary focus on your research? Other: Developmental biology (1), Theory (1) and Translational medicine (1).

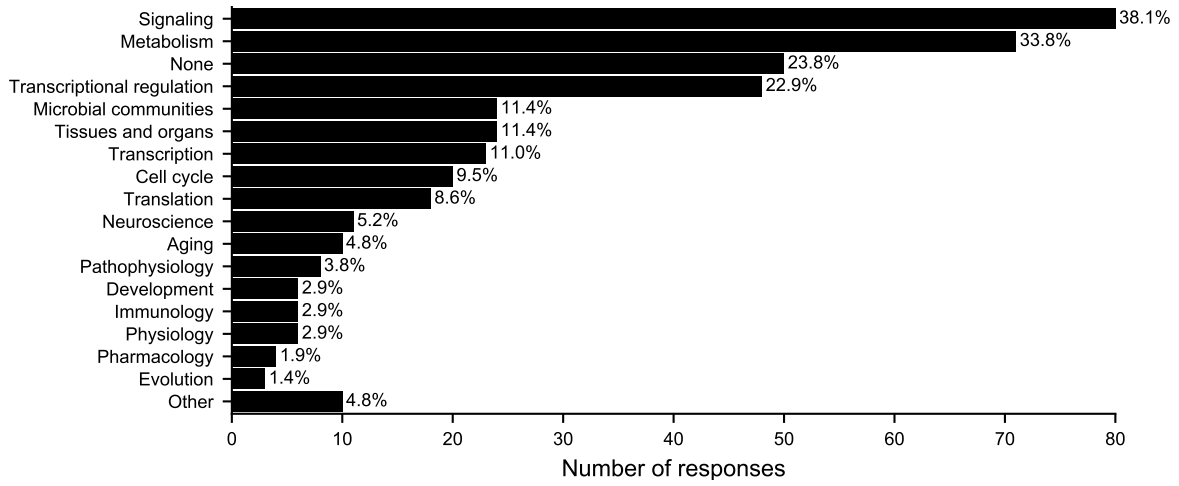

Figure S7. What is the biological focus of your research? Other: Cell physiology (2), Noise (2), Virology (2), Circadian rhythms (1), Ecology (1), Membranes (1) and Metagenomics (1).

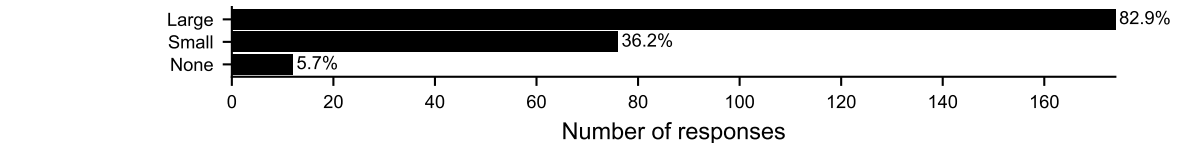

Figure S8. If you use models in your research, how large are the models that you typically use in your research?

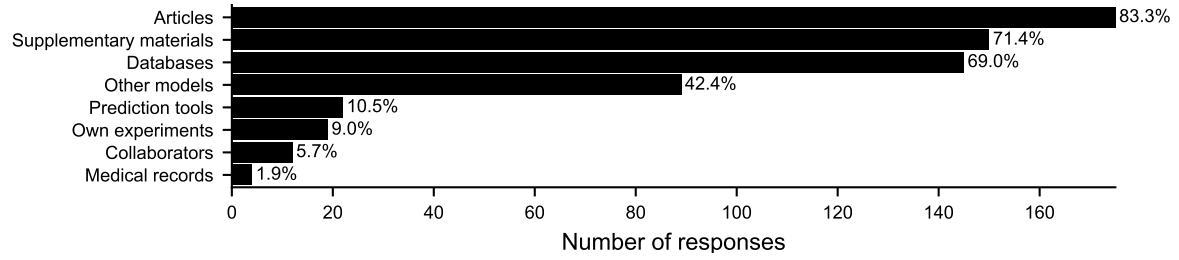

Figure S9. If you use models in your research, what types of data sources do you use?

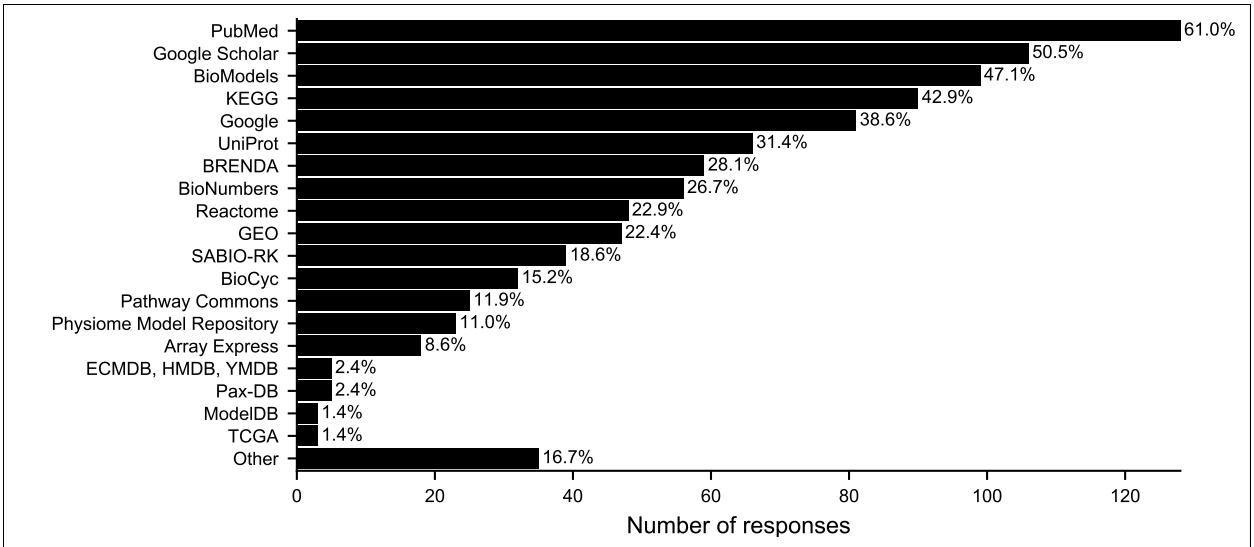

Figure S10. If you use models in your research, which data sources do you use? Other: BiGG (2), MetaNetX (2), Recon2 (2), SGD (2), ACSN (1), Allen Cell Types Database (1), BioBricks (1), Cardiac Atlas Project (1), CCLE (1), CGP (1), ChEBI (1), EEG ECoG MEG (1), ENCODE (1), Gene Ontology (1), GitHub (1), GTEX (1), HMR (1), Human Proteome Map (1), JWS Online (1), Metlin (1), MultiCellDS (1), NeuroMorpho (1), Open Source Brain (1), Panther (1), Physionet (1), ProteomicsDB (1), SEED (1), SEER (1), SignaLink (1), SIGNOR (1) and StringDB (1).

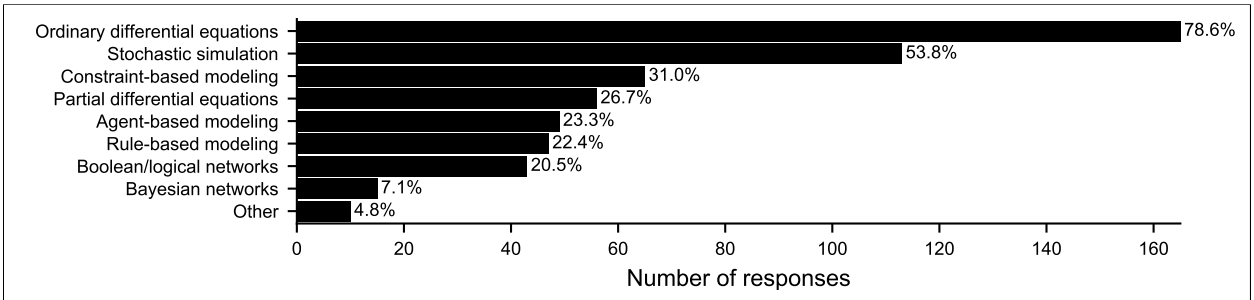

Figure S11. If you use models in your research, which mathematical representations and simulation algorithms do you use most frequently? Other: Delay differential equations (2), Discrete event simulation (2), Cellular automata (1), Cybernetic models (1), Difference equations (1), Langevin dynamics (1), Networks (1) and Petri nets (1).

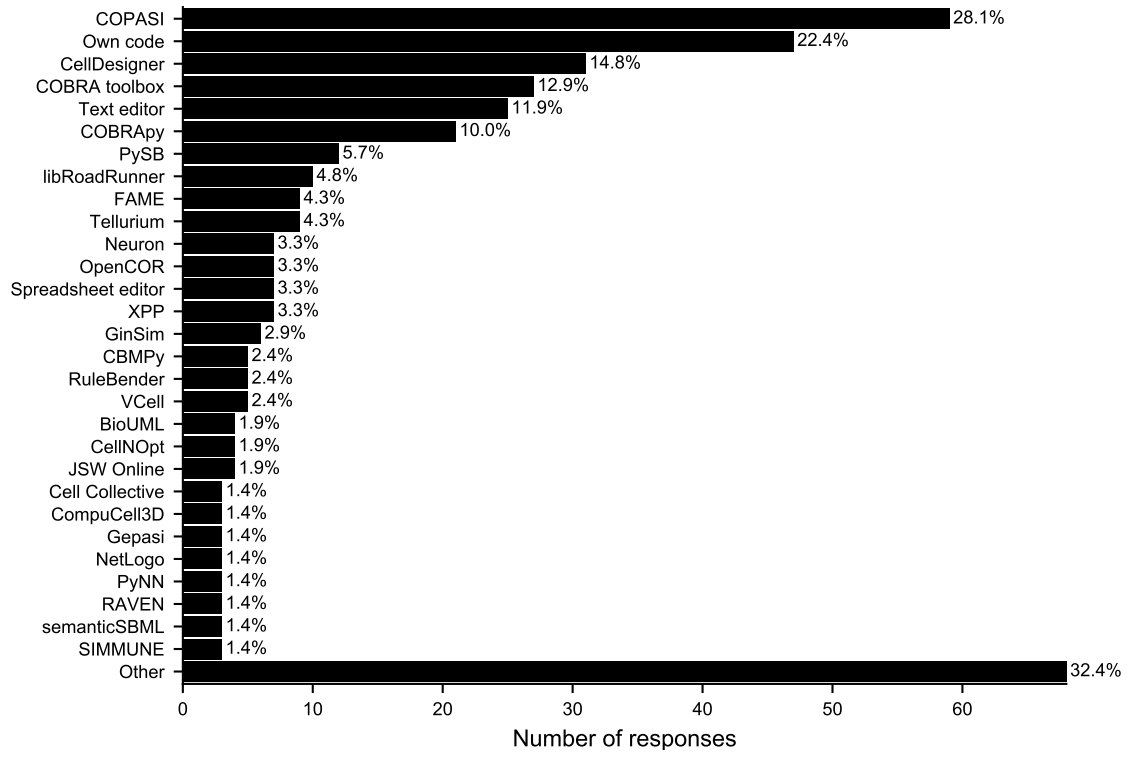

Figure S12. If you use models in your research, which tools do you most frequently use to build and/or simulate models? Other: AMIGO (2), Data2Dynamics (2), GENESIS (2), iBioSim (2), KaSim (2), MOOSE (2), Morpheus (2), PhysiCell (2), SimBiology (2), VANTED (2), AMICI (1), Animo (1), BioFVM (1), BioLayout Express (1), BooleanNet (1), Chaste (1), CLion (1), COMSOL (1), CVODE (1), DeSolve (1), E-Cell (1), Fenics (1), FlexFlux (1), fme (1), framed (1), GRO (1), IBCell (1), iggy (1), IQM toolbox (1), JSim (1), Kappa online simulator (1), MaBoSS (1), MASON (1), MCell (1), Mesa (1), Metatoolkit (1), ns-3 (1), OpenCMISS (1), optimusqual (1), Organism-Tissue Simulator (1), ProcessDB (1), PSICS (1), PyCharm (1), ROSS (1), RStudio (1), rxncon (1), SBGN-ED (1), SBMLsimulator (1), SBpipe (1), SBPOP toolbox (1), SBW (1), SEEK (1), SemSim (1), Smoldyn (1), SPiM (1), Tissue Simulation Toolkit (1), VirtualLeaf (1) and yEd (1).

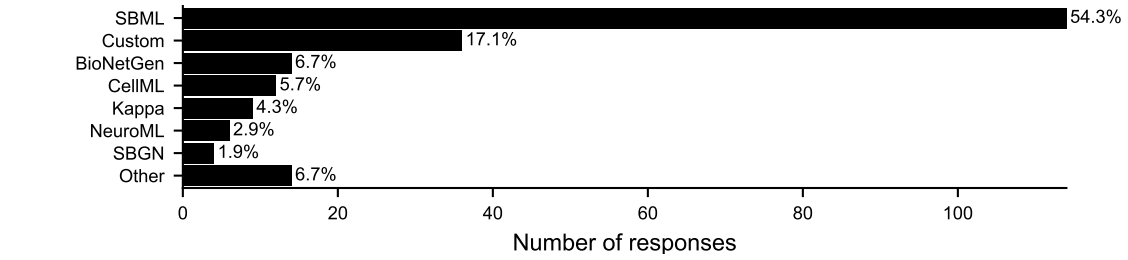

Figure S13. If you use models in your research, which languages do you most frequently use to represent models? Other: AMICI (1), CopasiML (1), Fenics (1), GinSim (1), GRO (1), mEPN (1), ML-Rules (1), MML (1), MultiCellDS (1), NEURON (1), SIMMUNE (1), Smoldyn (1), SPiM (1) and XPP (1).

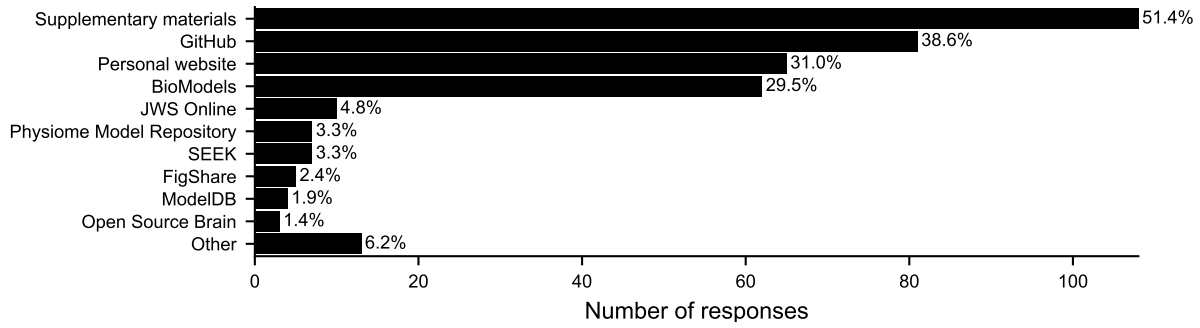

Figure S14. If you use models in your research, which resources do you most frequently use to distribute models? Other: BiGG (2), Cell Collective (2), GinSim repository (2), SimTk (2), Chaste (1), CoSBi (1), GitLab (1), SourceForge (1) and Synapse (1).

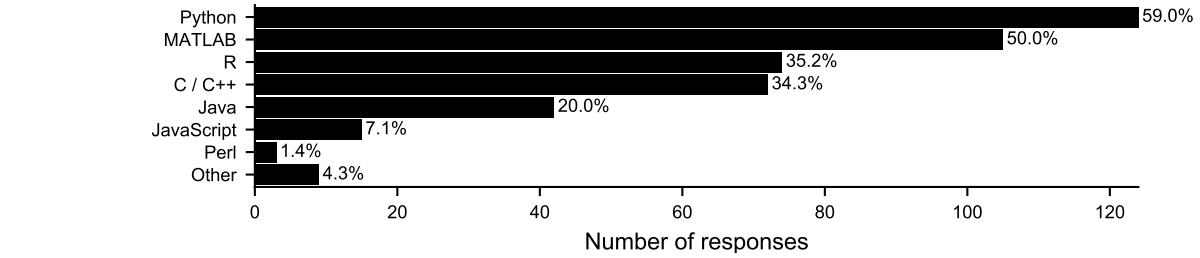

Figure S15. Which programming languages do you most frequently use in your research? Other: Fortran (2), C# (1), Cython (1), GRO (1), Julia (1), Mathematica (1), Ocaml (1) and Pascal (1).

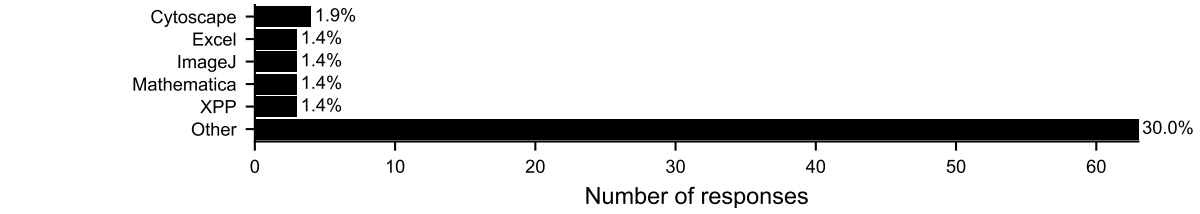

Figure S16. What additional tools do you frequently use in your research? Other: CBMPy (2), Eclipse (2), Inkscape (2), Jupyter (2), LaTeX (2), AMICI (1), AUMIC (1), Backstroke (1), BiGG (1), BLAST (1), BoolNet (1), CellDesigner (1), Chaste (1), CLC (1), CLion (1), COMBINE Archive Web (1), Emacs (1), Escher (1), Fiji (1), Git (1), GitHub (1), gnuplot (1), Google Documents (1), gPROMS (1), Icy (1), Jupyter (1), KaSim (1), libSBGN (1), Linux (1), MATLAB (1), Metatoolkit (1), MINERVA (1), NaviCell (1), NetworkX (1), NFsim (1), Octave (1), Office (1), Paxtools (1), PESTO (1), Phoenix NLME (1), PISKaS (1), PRISM (1), pyasp (1), PyCharm (1), SBGN-ED (1), SBMLsqueezer (1), SBOLDesigner (1), SBW (1), Schnitzcells (1), SED-ML Web Tools (1), SemGen (1), Slack (1), SynBioHub (1), Tissue Simulation Toolkit (1), VANTED (1), VirtualLeaf (1), yEd (1) and Zotero (1).

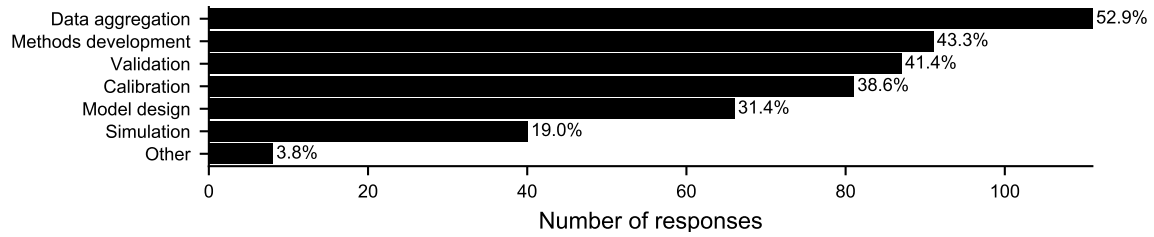

Figure S17. What are the most time-consuming aspects of your research? Other: Experimentation (2), Funding (2), Model analysis (2), Finding models (1) and Thinking (1).

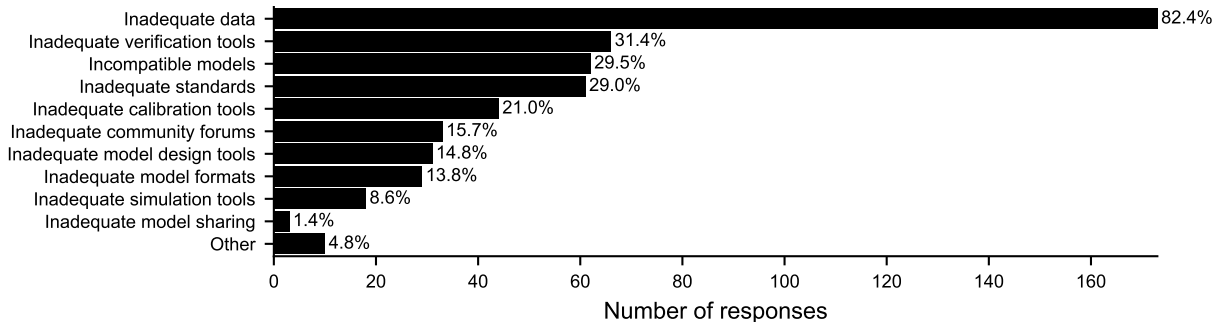

Figure S18. What do think are the main bottlenecks to building more predictive models? Other: Inadequate annotation tools (1), Inadequate documentation (1), Inadequate model annotation (1), Inadequate model selection and refinement tools (1), Inadequate tools for assessing the uncertainty of model predictions (1), Inadequate understanding of what does and doesn't need to be modeled (1), Inadequate visualization and analysis tools (1), Insufficient computational infrastructure (1), Insufficient modelers (1) and Long learning curve (1).

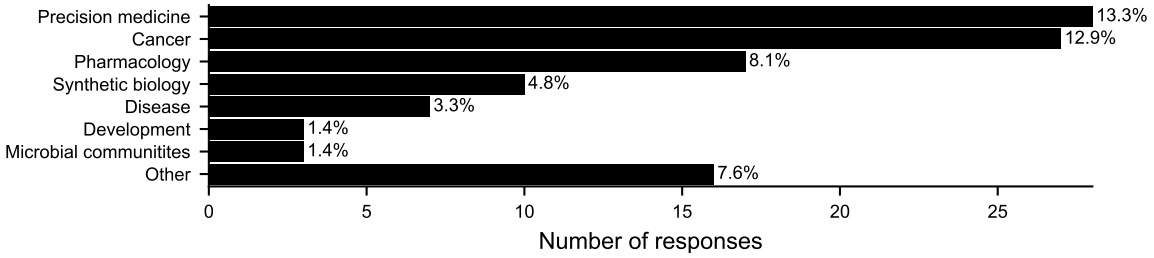

Figure S19. Given access to improved data and tools, what biomedical problems would you like to use models to solve (e.g. designer microorganisms, personalized cancer therapy)? Other: Infectious disease (2), Tissues (2), Aging (1), Antibiotic resistance (1), Cardiophysiology (1), Cell shape (1), Epigenetics (1), Evolution (1), Genetic variation (1), Microbiome (1), Phenotype prediction (1), Physiology (1), Single cell biology (1) and Translation (1).

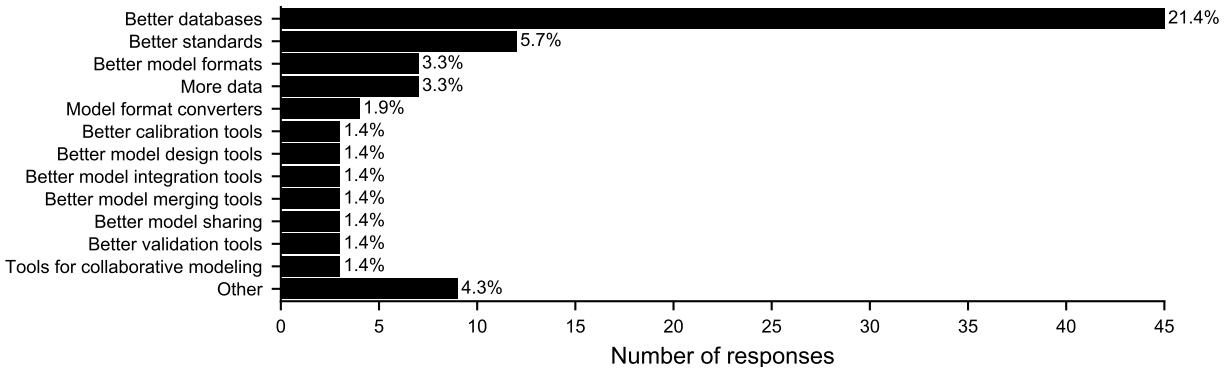

Figure S20. What additional databases, methods, tools, and/or standards would help accelerate your research? Other: Better algorithms (1), Better data analysis tools (1), Better experimental methods (1), Better model annotation (1), Better model annotation tools (1), Better model comparison tools (1), Better model selection tools (1), Better simulators (1) and More models (1).

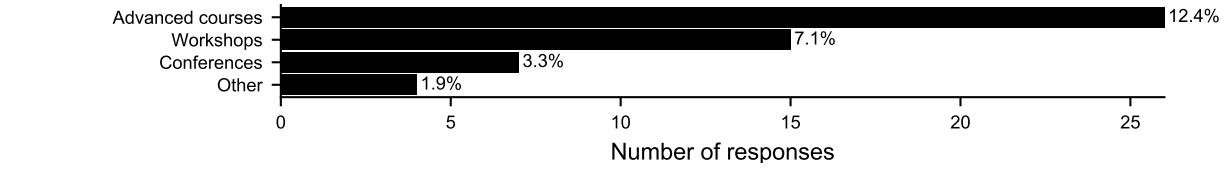

Figure S21. What additional meetings, courses, and/or training opportunities would help accelerate your research? Other: Forums (2) and Introductory courses (2).

References

1. Europe PMC Consortium *et al.*: **Europe PMC: a full-text literature database for the life sciences and platform for innovation.** *Nucleic Acids Res* 2015, **43**:D1042–D1048.

2. Haak LL, Fenner M, Paglione L, Pentz E, Ratner H: **ORCID: a system to uniquely identify researchers.** *Learned Publishing* 2012, **25**:259–264.