Centralizing data to unlock whole-cell models

Chew, Yin Hoon; Karr, Jonathan R.

DOI:
10.1016/j.coisb.2021.06.004

Document Version
Peer reviewed version

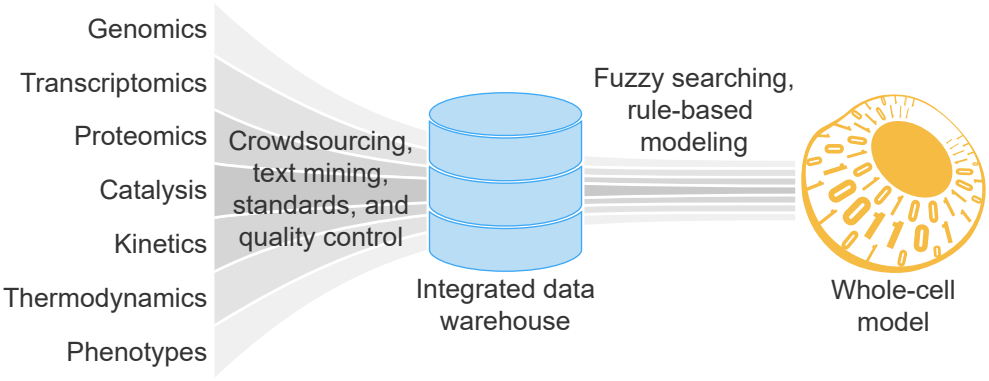
Citation for published version (Harvard):
Chew, YH & Karr, JR 2021, 'Centralizing data to unlock whole-cell models', *Current Opinion in Systems Biology*, vol. 27, 100353. https://doi.org/10.1016/j.coisb.2021.06.004

Centralizing data to unlock whole-cell models

Yin Hoon Chew and Jonathan R. Karr

Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, 1425 Madison Avenue, New York, NY 10029, USA

Graphical abstract



Highlights

- Whole-cell models require data about each molecule and molecular interaction
- Data is increasingly available, but its scattered organization hinders modeling
- A central database of data and knowledge would accelerate whole-cell modeling
- Such a database requires collaboration and standardization
- New experimental methods and automation are also needed to broaden and deepen our data

Centralizing data to unlock whole-cell models

Yin Hoon Chew and Jonathan R. Karr

Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, 1425 Madison Avenue, New York, 10029, NY, USA

Abstract

Despite substantial potential to transform bioscience, medicine, and bioengineering, whole-cell models remain elusive. One of the biggest challenges to whole-cell models is assembling the large and diverse array of data needed to model an entire cell. Thanks to rapid advances in experimentation, much of the necessary data is becoming available. Furthermore, investigators are increasingly sharing their data due to growing recognition of the importance of research that is transparent and reproducible to others. However, the scattered organization of this data continues to hamper modeling. Toward more predictive models, we highlight the challenges to assembling the data needed for whole-cell modeling and outline how we can overcome these challenges by working together to build a central data warehouse.

Introduction

More comprehensive and more predictive models of cells are broadly perceived as vital for understanding, controlling, and designing biology. For example, whole-cell models would likely help scientists conduct experiments in silico with unprecedented control and resolution [1], help physicians precisely treat each patient's unique genomics [2], and help bioengineers rationally design synthetic cells [3].

Recently, scientists have taken several steps toward whole-cell models, producing large-scale models of *Mycoplasma genitalium* [4, 5], *Mycoplasma mycoides* [6], *Escherichia coli* [7–10], *Saccharomyces cerevisiae* [11, 12], and human epithelial cells [13] among others. Researchers have also begun to explore how whole-cell models could help guide personalized medical decisions [14] and design synthetic cells [15, 16].

Despite substantial interest, whole-cell models remain elusive due to numerous challenges, including integrating vast information about diverse biochemical processes [17], accounting for the structure and organization of cells and their numerous components [18, 19]; simulating [20], calibrating [21, 22], visualizing [23, 24], and validating [23, 24] high-dimensional, computationally-expensive, hybrid models; and developing models collaboratively [25, 26]. Toward a framework for whole-cell modeling, we and others have summarized these challenges [23, 24, 27, 28].

To help focus efforts to accelerate whole-cell modeling, we recently surveyed the community

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

about the bottlenecks to progress [28]. Most respondents expressed that the main immediate barrier to more predictive models is insufficient experimental data and knowledge.

Undeniably, we do not yet have enough data to completely model a cell. As a result, complete models of entire cells are not presently feasible. Nevertheless, we believe that significantly more comprehensive models can already be constructed by leveraging the substantial data that is already available. Thus, in our opinion, the practical bottleneck to better models is not our limited experimental capabilities, but the scattered organization of our existing data. Furthermore, as our experimental capabilities continue to expand rapidly, we believe that it is critical to begin to develop whole-cell modeling capabilities now so that we are prepared to realize whole-cell models when sufficient data is available.

To focus efforts to address this bottleneck, here we explore the data that is already available and how we can best leverage it for whole-cell modeling. First, we outline the data that is needed for whole-cell modeling. Second, we highlight exemplary resources that already provide key data. Third, we assess the challenges to moving beyond these resources. Finally, we present a roadmap to assembling a data warehouse for whole-cell modeling. We firmly believe that such a warehouse would accelerate the development of more predictive models.

The mountain of data needed to model an entire cell

Modeling an entire cell will likely require similarly comprehensive experimental data. At a minimum, this will likely include (a) the sequence of the cell's genome; (b) data about the structure of its genome, such as the location of each replication origin, promoter, and terminator; (c) information about the structure, abundance, turnover, and spatial distribution of each molecule in the cell; (d) information about each molecular interaction that can occur in the cell, including the molecules that participate in each interaction and the catalysis, rate, thermodynamics, and duration of each interaction; and (e) global information about the temporal dynamics and spatial organization of the cell, such as the organization of its life cycle, its size, shape, and subcellular organization.

To enable modelers to best leverage this data, this data should be accompanied by detailed metadata about its semantic meaning and provenance. At a minimum, each experimental observation should be accompanied by metadata about the molecule or molecular process which was measured, the genetic and environmental context in which the measurement was conducted, the methods used to collect and reduce the data, the individuals who collected and processed the data, and the dates when the data was collected and reduced.

The sea of data that could be repurposed for whole-cell modeling

Compared to the experimental capabilities of an individual lab or even a consortium, this laundry list of data seems insurmountable. Without a quantum leap forward in automation or a massive increase in funding, we expect the data needed for whole-cell modeling to exceed the experimental capabilities of most labs for the foreseeable future.

Although little data has been explicitly collected for whole-cell modeling, the scientific literature already contains substantial relevant data. Furthermore, much of this data is already publicly accessible due to an increasing culture of data sharing. Taken together, we believe that substantial data can be repurposed for more comprehensive models.

Exemplary data resources that we believe can be repurposed for whole-cell modeling include, but are not limited to, the Protein Data Bank (PDB) [29], ECMDB [30], YMDB [31], PaxDB [32], PSORTdb [33], BRENDA [34], and SABIO-RK [35] (Table 1). ECMDB and YMBD contain thousands of measurements of the concentrations of metabolites in *E. coli* and *S. cerevisiae*. PaxDB contains over 1 million measurements of the abundances of proteins in over 50 organisms. PSORTdb contains over 10,000 measurements of the localization of proteins in over 400 organisms, as well as predicted localizations for over 15,000 organisms. Together, BRENDA and SABIO-RK contain over 300,000 kinetic parameters for thousands of metabolic reactions. In our experience, BioNumbers [36] is also a valuable resource for data that is outside the scope of repositories for specific types of data. For example, BioNumbers contains data about the rates of non-metabolic processes such as DNA damage and RNA polymerization; the fluxes of the exchange of nutrients into and out of cells; and the sizes, densities, and growth rates of cells, which are not contained in other repositories.

In addition to repurposing data for whole-cell modeling, foundational research is also needed to expand our experimental capabilities. While our capabilities to characterize the transcriptome and proteome have advanced rapidly over the past 20 years, our capabilities to characterize the metabolome, single cell variation, and temporal dynamics continue to lag. For example, additional capabilities to characterize the composition and dynamics of the metabolome could enable more complete flux balance analysis models.

The challenges to reusing data for whole-cell modeling

While substantial data is already available for whole-cell modeling, unfortunately, most of this data is not readily accessible. The challenges to utilizing the existing data are severalfold. First, the existing data is distributed over a wide range of organisms and experimental conditions. As a result, only a small amount of data is available for each organism and experimental condition. One potential solution to this data sparsity is to leverage data from closely related organisms and conditions. However, few databases have been designed to help investigators search for such related data. Literature search engines such as Google Scholar and PubMed have also not been designed to help investigators find such related data.

Second, our existing data is organized heterogeneously. Our existing data is scattered across many databases, as well as many individual journal articles. Additionally, the existing databases provide different interfaces and APIs. Furthermore, the existing data is described with many different formats, identifier systems, and ontologies. The effort required to deal with this heterogeneity distracts investigators from modeling.

Third, many databases and articles only provide minimal metadata or minimally structured metadata. The lack of detailed metadata is part of why it is difficult to find measurements

Type	Key sources	Relevant standards
Annotated genomes	ENA [37], GenBank [38]	BED, FASTA, GenBank, GFF, GSC [39]
DNA modifications	DNAmod [40]	
Metabolite structures	ChEBI [41], PubChem [42]	CML [43], InChI [44]
Metabolite concentrations	ECMDB [30], YMDB [31]	MSI [45]
Protein modifications	Protein Ontology [46]	BpForms [47], HELM [48], PDB format [49]
Protein structures	Protein Data Bank [29]	PDBx/mmCIF [49], PDB format [49], PSI [50]
Protein localizations	eSLDB [51], Human Protein Atlas [52], PSORTdb [53]	
Protein abundances	PaxDB [32]	mzML [54], PSI [50]
Protein half-lives	Literature	
RNA modifications	MODOMICS [55]	BpForms [47], HELM [48], MODOMICS [55]
RNA localizations	RNALocate [56], lncATLAS [57]	
RNA abundances	ArrayExpress [58], GEO [59]	BAM [60], FASTQ, [61], MINSEQE
RNA half-lives	Literature	
Composition of complexes	BioCyc [62], Complex Portal [63]	BcForms [47], PDBx/mmCIF [49], PDB format [49], PSI [50]
Reaction equations and catalysis	BioCyc [62], KEGG [64], MetaNetX [65]	BioPAX [66], EC, STRENDA [67]
Reaction rate constants	BRENDA [34], SABIO-RK [35]	EC, STRENDA [67]
Reaction fluxes	CeCaFDB [68]	
DNA-protein binding	EpiFactors [69], JASPAR [70], TRANSFAC [71]	ENCODE standards [72]
Protein-protein interactions	IntAct [73], STRING [74]	PSI [50]
Physiological parameters	BioNumbers [36]	

Table 1: Key types and sources of data for whole-cell modeling and relevant formats and metadata standards for this data.

of related organisms and conditions. The lack of detailed, consistently structured metadata also makes it challenging to interpret and integrate data accurately.

Fourth, a significant amount of data is not available in any reusable form. Despite increasing emphasis on data sharing and reuse [75], many results are still reported without their underlying data. One contributing factor is the lack of domain-specific formats and databases for many types of data. Such shared infrastructure makes it easier for authors to share data and easier for other investigators to reuse it. In the absence of such infrastructure, authors often have little incentive to share data, and reviewers often have low expectations for data sharing. Furthermore, with notable exceptions for genetic and structural data, many journals still have porous guidelines that permit publication without sharing the underlying data.

Emerging tools for sharing, discovering, and reusing data

Efforts to make data easier to share, discover, and reuse for whole-cell modeling and other research are underway. This includes the development of standard formats and ontologies for describing data, central databases for storing data, and tools for discovering specific data. Here, we highlight some of the most relevant emerging resources for whole-cell modeling.

Formats for exchanging data for whole-cell modeling

Three notable formats for capturing some of the data and knowledge needed for whole-cell modeling include the Investigation/Study/Assay tabular (ISA-Tab) format [53], the Multicellular Data Standard (MultiCellDS) [76], and BioPAX [66]. ISA-Tab is ideal for high-dimensional data, such as transcriptome-wide measurements of RNA turnover rates, which lack more specific formats. MultiCellDS is an emerging format intended to capture a digital "snapshot" of a cell line, encompassing measurements of its metabolome, transcriptome, proteome, and phenotype, as well as metadata about the environmental context of each measurement and the methods used to collect it. BioPAX is a format for describing knowledge about the molecules and molecular interactions inside cells.

In our experience, whole-cell modeling requires both quantitative and relational data about multiple aspects of a cell. To capture this information for our first models, we developed the WholeCellKB schema [77]. Simultaneously, Lubitz and colleagues developed SBTab [78], a tabular format with similar goals. As we began to explore additional models, we realized that many modelers both want to be able to use spreadsheets to quickly assemble datasets and use computer programs to quality control their datasets and incorporate them into models. To meet this need, we recently merged the concepts behind WholeCellKB and SBTab into ObjTables [79], a set of tools that make it easy for modelers to use user-friendly spreadsheets to integrate data, define schemas for rigorously validating their data, and parse linked spreadsheets into data structures that are conducive to modeling. SEEK provides an online environment for managing datasets organized as spreadsheets [80].

Formats for critical metadata for whole-cell modeling

As we discussed above, structured metadata is critical for understanding and merging data. Because cells contain millions of distinct molecular species [81] due to combinatorial biochemical processes such as post-transcriptional and post-translational modification and complexation, we think that it is particularly important for datasets to concretely describe the molecules and molecular interactions that they characterize. Small molecules can be described using several formats such as the Chemical Markup Language (CML) [63] and IUPAC International Chemical Identifier (InChI) [44] formats. Sequences of unmodified DNAs, RNAs, and proteins can be described using the FASTA format. Sequences of modified DNAs, RNAs, and proteins can be described using BpForms [82] and HELM [48]. BpForms generalizes the IUPAC and IUBMB formats commonly used to describe unmodified DNAs, RNAs, and proteins to capture physiological polymers with modifications, crosslinks, and nicks. Macromolecular complexes can be described using BcForms [82] and HELM.

Resources for capturing metadata about the genetic context of measurements include the NCBI Taxonomy database [83], the Cell Line Ontology [84], and standard nomenclatures for genetic variants, such as the HGVS standard [85] for human or the MGI standard for mouse and rat. Resources for capturing metadata about the environmental context of measurements including databases such as the Known Media Database [86] and MediaDB [87].

Numerous formats have been developed to capture detailed information about how specific types of data are collected. FAIRSharing [88] is an excellent resource for finding formats for specific types of data. ORCID is increasingly being used to capture information about the investigators who conducted an experiment.

Centralized knowledgebases of information for whole-cell modeling

Because whole-cell modeling requires multiple types of data, we believe that centralized databases are also needed to help investigators find and obtain data. Three pioneering efforts to centralize data for modeling cells were the CyberCell Database (CCDB) for quantitative data about *E. coli* [89*], EcoCyc for qualitative and relational information about *E. coli* [90**], and NeuronDB and CellPropDB for quantitative data about membrane channels, receptors, and neurotransmitters [91*]. EcoCyc continues to be a valuable resource, particularly for the development of genome-scale metabolic models [92]. GEMMER is a newer database that aims to facilitate models of *S. cerevisiae* [93].

More recent efforts to aggregate data for modeling have refined and expanded the concepts pioneered by the CCDB, CellPropDB, EcoCyc, NeuronDB, and others. One additional concept which we believe is essential is crowdsourcing. Crowdsourcing data aggregation addresses the problem that no single lab can curate the entire literature, and it can help avoid duplicate efforts by multiple researchers to curate similar data. Two exemplary resources that embody this philosophy are the Omics Discovery Index (OmicsDI) [94**], which provides a search engine to discover over 20 different types of quantitative molecular data curated by more than 20 different communities, and Pathway Commons [95], which provides a search engine for information about molecular interactions curated by more than 22 groups of curators.

To make it easy to contribute to OmicsDI and Pathway Commons, contributors only need to contribute a small amount of information about each dataset (OmicsDI) and pathway (Pathway Commons). However, this strategy pushes the onerous work of aggregating and normalizing data from the developers of these resources to their users.

To further help modelers obtain data for whole-cell modeling, we developed Datanator [96**], an integrated database of data for modeling the biochemical activity of a cell. Presently, Datanator contains several key types of data for whole-cell modeling, including data about metabolite structures and concentrations; RNA modifications, localizations, and half-lives; protein modifications, localizations, abundances, and half-lives; and reaction rate constants, each for a broad range of organisms. In addition, Datanator provides a search engine tailored to the sparse nature of our existing data. This search engine can help modelers compensate for the absence of direct measurements with measurements of similar molecules, molecular interactions, organisms, or experimental conditions.

Datanator builds on many of the ideas pioneered by the CCDB, OmicsDI, and other databases. Like OmicsDI, Datanator is a meta database that leverages the curation efforts and expertise of several primary databases. Like the CCDB, Datanator provides data in a consistent format that is convenient for modelers.

To provide all of the data needed for whole-cell modeling, Datanator must be expanded to fill in gaps in the types of data that Datanator already captures and to capture additional types of data. This will require integrating many more databases into Datanator and aggregating additional types of data directly from the literature. One key gap in the data already captured by Datanator is the limited measurements of the intracellular concentrations of metabolites. Unfortunately, limited data is available in the literature. Additional experiments are needed to measure additional metabolites and to generate data for a wider range of organisms. One key type of data that should be added to Datanator is measurements of RNA abundances. Abundant data is available from ArrayExpress [58]. A second type of data that we believe is critical to add to Datanator is measurements of reaction fluxes. This information could be imported from CeCaFDB [68].

Roadmap to data for whole-cell modeling

Despite progress, we still only have a fraction of the data that will likely be needed for whole-cell modeling, and it remains tedious to gather the data that does exist. Ultimately, new experimental methods will be needed to fill the gaps in our understanding of the individual molecules and molecular interactions in cells. To enable investigators to independently train and test their models, increased automation will also be needed to generate data about a wider range of genotypes and environmental conditions. Most importantly, investigators need to pool their efforts so that everyone has access to more data. Here, we outline one way the community could work together to assemble the data that many modelers need (Figure 1).

To facilitate the density of data needed for more comprehensive models, the community could first focus on a small number of organisms and cell types such as *E. coli*, *S. cerevisiae*, and *H.*

sapiens stem cells. Similarly, the community could focus on a specific set of environmental conditions, such as minimal media for microbes.

Second, the community could develop a central database of the most essential types of data that need to be collected for these cells. This database could both allow individual investigators to suggest specific types of data that they believe should be collected, and allow the community to vote for the data that they believe would be most valuable. Ideally, investigators would then consider these votes when deciding which data to generate, focusing on the most frequently requested data. A large number of votes for a type of data would also likely be powerful support for proposals for funding to collect the data.

Third, the community could coordinate the generation of this data to ensure that these cells are characterized deeply and avoid redundant efforts to generate similar data. The database outlined above could help facilitate this by enabling investigators to submit information about data they plan to generate. Experimentalists could then use this information to focus on generating unique data, and computational scientists could use this information to learn about upcoming experiments and contribute to their design to ensure they produce data that is well-suited and annotated for modeling.

Fourth, the community could align on common formats, metadata, and quality control mechanisms for each type of data. Importantly, this metadata should include common formats for describing the genotype of each sample, the structure of each measured molecule, and the composition of each measured media condition. User-friendly and automated software tools could be created to make it easy for investigators to embrace these formats and rigorously assess the quality of their data.

Fifth, the community could develop additional primary databases for types of data that are not covered by the existing primary databases. For example, a group of researchers is beginning to assemble a database of the thermodynamics of biochemical reactions. Each database could be initiated by a small team of curators who seed the database by aggregating their own data and data from the literature. Beyond this initial phase, these databases could allow the community to submit data directly. In some cases, text mining could also be used to automatically or semi-automatically extract data from the literature. One area where text mining has been successful is collating interactions between genes and drugs [97]. Foundational tools for text mining include the Natural Language Toolkit [98] and spaCy. Collectively, multiple such primary databases would be able to support a broad range of formats for different types of data. These primary databases would also be well-positioned for expert curators to quality control specific types of data. Furthermore, such primary databases might be able to assemble the critical mass of investigators needed to lobby journals to require public deposition of specific types of data.

Sixth, more of these primary databases could be integrated into Datanator. This would make all of this data accessible from a single interface and discoverable with Datanator's tools for extracting clouds of potentially relevant data from sparse data sets. This process could be simplified and accelerated by aligning the primary databases on a common export format. In particular, the primary databases would need to align on a common scheme for representing metadata about the meaning and provenance of each measurement. In addition,

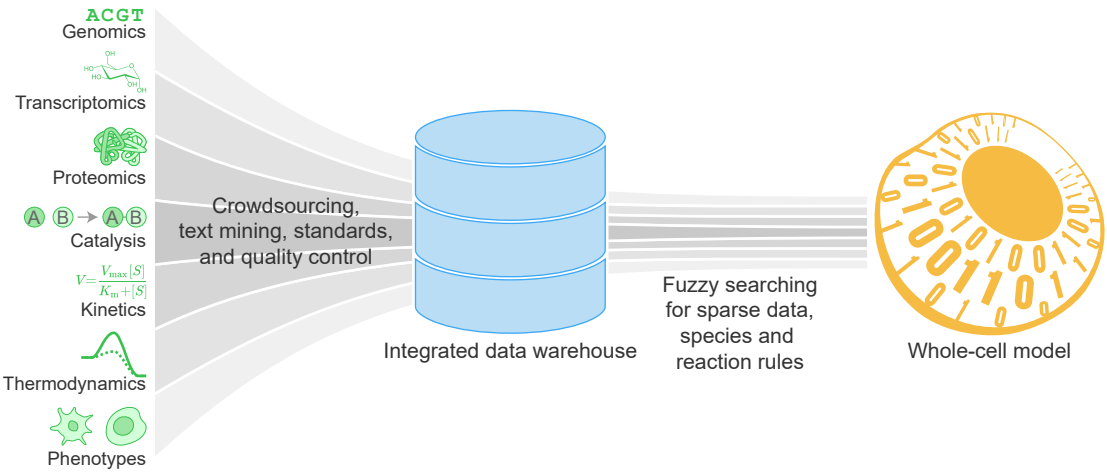


Figure 1: An integrated warehouse of molecular data and knowledge is needed to accelerate whole-cell modeling. This warehouse could be assembled by combining multiple crowdsourced databases for different types of data with data automatically mined from the literature. Models could be systematically constructed from this warehouse using sets of rules that encode biochemical processes and physical laws.

Datanator could be expanded to directly accept data. This would enable any type of data to be integrated into Datanator, including data that falls outside the scope of all of the primary databases. Furthermore, automated programs could be developed to identify potential issues with the data integrated into Datanator by examining the consistency of different sources and types of data. We invite the community to contribute data to Datanator, and we welcome input into its goals, design, and implementation.

In addition, Datanator could be further integrated with databases of relational and descriptive information such as EcoCyc and Pathway Commons. Ideally, a team of curators would be established to quality control this final integrated database.

Once this data warehouse is available, additional methods and tools will be needed to use it to construct models. One possible way to use the data will be to devise rules, or templates, for generating species, reactions, rate laws, and rate parameters for specific types of data. For example, a rule could be created that generates protein species and translation and protein turnover reactions based on sequenced genomes, computed locations of start and stop codons, and measured protein abundances and half-lives. Such rules could encode biochemical processes such as translation and physical laws such as mass-action kinetics. Potentially, entire models could be constructed from such rules. This workflow would enable complex, detailed models to be systematically and transparently constructed from comparatively small sets of rules. We are building a system that will enable such rules. We anticipate it will accelerate the construction of large models.

Conclusions

Despite the challenges to assembling the data needed for whole-cell modeling, we are confident that the combination of technology development, standardization, and collaboration

outlined above will enable substantially more comprehensive, predictive, and credible models. Our Datanator database implements many of these ideas. To illustrate their potential, we are currently using Datanator to help construct a higher resolution model of the metabolism of *E. coli*. To move forward, we encourage the community to join existing efforts to aggregate data such as Datanator, EcoCyc, and OmicsDI by helping to gather, integrate, or quality control data, or develop formats and tools that could facilitate these efforts.

Declaration of competing interest

None.

Acknowledgments

We thank Paul Lang, Zhouyang Lian, Wolfram Liebermeister, Saahith Pochiraju, Yosef Roth, and David Wishart for enlightening discussions about data for whole-cell modeling. This work was supported by the National Institutes of Health [grant numbers R35GM119771, P41EB023912].

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. Carrera J, Covert MW: **Why build whole-cell models?** *Trends Cell Biol* 2015, **25**:719–722.

2. Tomita M: **Whole-cell simulation: a grand challenge of the 21st century.** *Trends Biotechnol* 2001, **19**:205–210.

3. Marucci L, Barberis M, Karr J, Ray O, Race PR, Souza Andrade M de, Grierson C, Hoffmann SA, Landon S, Rech E, *et al.*: **Computer-aided whole-cell design: taking a holistic approach by integrating synthetic with systems biology.** *Front Bioeng Biotechnol* 2020, **8**:942.

4. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival Jr B, Assad-Garcia N, Glass JI, Covert MW: **A whole-cell computational model predicts phenotype from genotype.** *Cell* 2012, **150**:389–401.

5. Burke PE, Claudia BdL, Costa LdF, Quiles MG: **A biochemical network modeling of a whole-cell.** *Sci Rep* 2020, **10**:1–14.

6. Thornburg ZR, Melo MC, Bianchi D, Brier TA, Crotty C, Breuer M, Smith HO, Hutchison III CA, Glass JI, Luthey-Schulten Z: **Kinetic modeling of the genetic information processes in a minimal cell.** *Front Mol Biosci* 2019, **6**:130.

7. Thiele I, Jamshidi N, Fleming RM, Palsson BØ: **Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization** *PLoS Comput Biol* 2009, **5**:e1000312.

8. Roberts E, Magis A, Ortiz JO, Baumeister W, Luthey-Schulten Z: **Noise contributions in an inducible genetic switch: a whole-cell simulation study.** *PLoS Comput Biol* 2011, **7**:e1002010.

9. Carrera J, Estrela R, Luo J, Rai N, Tsoukalas A, Tagkopoulos I: **An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of Escherichia coli.** *Mol Syst Biol* 2014, **10**:735.

10. Macklin DN, Ahn-Horst TA, Choi H, Ruggero NA, Carrera J, Mason JC, Sun G, Agmon E, DeFelice MM, Maayan I, *et al.*: **Simultaneous cross-evaluation of heterogeneous E. coli datasets via mechanistic simulation.** *Science* 2020.

11. Münzner U, Klipp E, Krantz M: **A comprehensive, mechanistically detailed, and executable model of the cell division cycle in Saccharomyces cerevisiae.** *Nat Commun* 2019, **10**:1–12.

12. Ye C, Xu N, Gao C, Liu G, Xu J, Zhang W, Chen X, Nielsen J, Liu L: **Comprehensive understanding of Saccharomyces cerevisiae phenotypes with whole-cell model WM_S288C.** *Biotechnol Bioeng* 2020, **117**:1562–1574.

13. Ghaemi Z, Peterson JR, Gruebele M, Luthey-Schulten Z: **An in-silico human cell model reveals the influence of spatial organization on RNA splicing.** *PLoS Comput Biol* 2020, **16**:e1007717.

14. Bordbar A, McCloskey D, Zielinski DC, Sonnenschein N, Jamshidi N, Palsson BO: **Personalized whole-cell kinetic models of metabolism for discovery in genomics and pharmacodynamics.** *Cell Syst* 2015, **1**:283–292.

15. Purcell O, Jain B, Karr JR, Covert MW, Lu TK: **Towards a whole-cell modeling approach for synthetic biology.** *Chaos* 2013, **23**:025112.

16. Rees-Garbutt J, Chalkley O, Landon S, Purcell O, Marucci L, Grierson C: **Designing minimal genomes using whole-cell models.** *Nat Commun* 2020, **11**:1–12.

17. Takahashi K, Yugi K, Hashimoto K, Yamada Y, Pickett CJ, Tomita M: **Computational challenges in cell simulation: a software engineering approach.** *IEEE Intell Syst* 2002, **17**:64–71.

18. Im W, Liang J, Olson A, Zhou HX, Vajda S, Vakser IA: **Challenges in structural approaches to cell modeling.** *J Mol Biol* 2016, **428**:2943–2964.

19. Luthey-Schulten Z: **Integrating experiments, theory and simulations into whole-cell models.** *Nat Methods* 2021, **18**:446–447.

20. Goldberg AP, Chew YH, Karr JR: **Toward scalable whole-cell modeling of human cells.** *Proc 2016 ACM SIGSIM Conf Princip Adv Discrete Simul* 2016, 259–262.

21. Babtie AC, Stumpf MPH: **How to deal with parameters for whole-cell modelling.** *J R Soc Interface* 2017, **14**:20170237.

22. Stumpf MPH: **Statistical and computational challenges for whole cell modelling.** *Curr Opin Syst Biol* 2021.

23. Macklin DN, Ruggero NA, Covert MW: **The future of whole-cell modeling.** *Curr Opin Biotechnol* 2014, **28**:111–115.

24. Feig M, Sugita Y: **Whole-cell models and simulations in molecular detail.** *Annu Rev Cell Dev Biol* 2019, **35**:191–211.

25. Singla J, White KL: **A community approach to whole-cell modeling.** *Curr Opin Syst Biol* 2021.

26. Waltemath D, Karr JR, Bergmann FT, Chelliah V, Hucka M, Krantz M, Liebermeister W, Mendes P, Myers CJ, Pir P, *et al.*: **Toward community standards and software for whole-cell modeling.** *IEEE Trans Biomed Eng* 2016, **63**:2007–2014.

27. Goldberg AP, Szigeti B, Chew YH, Sekar JA, Roth YD, Karr JR: **Emerging whole-cell modeling principles and methods.** *Curr Opin Biotechnol* 2018, **51**:97–102.

28. Szigeti B, Roth YD, Sekar JA, Goldberg AP, Pochiraju SC, Karr JR: **A blueprint for human whole-cell modeling.** *Curr Opinion Systems Biol* 2018, **7**:8–15.

29. wwPDB consortium: **Protein Data Bank: the single global archive for 3D macromolecular structure data** *Nucleic Acids Res* 2019, **47**:D520–D528.

30. Sajed T, Marcu A, Ramirez M, Pon A, Guo AC, Knox C, Wilson M, Grant JR, Djoumbou Y, Wishart DS: **ECMDB 2.0: A richer resource for understanding the biochemistry of E. coli.** *Nucleic Acids Res* 2016, **44**:D495–D501.

31. Ramirez-Gaona M, Marcu A, Pon A, Guo AC, Sajed T, Wishart NA, Karu N, Djoumbou Feunang Y, Arndt D, Wishart DS: **YMDB 2.0: a significantly expanded version of the yeast metabolome database.** *Nucleic Acids Res* 2017, **45**:D440–D445.

32. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, Mering C von: **Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines.** *Proteomics* 2015, **15**:3163–3168.

33. Lau WYV, Hoad GR, Jin V, Winsor GL, Madyan A, Gray KL, Laird MR, Lo R, Brinkman FSL: **PSORTdb 4.0: expanded and redesigned bacterial and archaeal protein subcellular localization database incorporating new secondary localizations.** *Nucleic Acids Res* 2021, **49**:D803–D808.

34. Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, Neumann-Schaal M, Jahn D, Schomburg D: **BRENDA, the ELIXIR core data resource in 2021: new developments and updates.** *Nucleic Acids Res* 2021, **49**:D498–D508.

35. Wittig U, Rey M, Weidemann A, Kania R, Müller W: **SABIO-RK: an updated resource for manually curated biochemical reaction kinetics.** *Nucleic Acids Res* 2018, **46**:D656–D660.

36. Milo R, Jorgensen P, Moran U, Weber G, Springer M: **BioNumbers–the database of key numbers in molecular and cell biology.** *Nucleic Acids Res* 2010, **38**:D750–D753.

37. Harrison PW, Ahamed A, Aslam R, Alako BT, Burgin J, Buso N, Courtot M, Fan J, Gupta D, Haseeb M, *et al.*: **The European Nucleotide Archive in 2020** *Nucleic Acids Res* 2021, **49**:D82–D85.

38. Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST, Karsch-Mizrachi I: **GenBank** *Nucleic Acids Res* 2021, **49**:D92–D96.

39. Field D, Sterk P, Kottmann R, De Smet JW, Amaral-Zettler L, Cochrane G, Cole JR, Davies N, Dawyndt P, Garrity GM, *et al.*: **Genomic Standards Consortium projects** *Standards Genomic Sci* 2014, **9**:599–601.

40. Sood AJ, Viner C, Hoffman MM: **DNAmod: the DNA modification database** *J Cheminform* 2019, **11**:1–10.

41. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C: **ChEBI in 2016: Improved services and an expanding collection of metabolites** *Nucleic Acids Res* 2016, **44**:D1214–D1219.

42. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, *et al.*: **PubChem in 2021: new data content and improved web interfaces** *Nucleic Acids Res* 2021, **49**:D1388–D1395.

43. Murray-Rust P, Rzepa HS, Wright M: **Development of chemical markup language (CML) as a system for handling complex chemical content** *New J Chem* 2001, **25**:618–634.

44. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D: **InChI, the IUPAC international chemical identifier** *J Cheminform* 2015, **7**:1–34.

45. Fiehn O, Robertson D, Griffin J, Werf M van der, Nikolau B, Morrison N, Sumner LW, Goodacre R, Hardy NW, Taylor C, *et al.*: **The metabolomics standards initiative (msi)** *Metabolomics* 2007, **3**:175–178.

46. Chen C, Huang H, Ross KE, Cowart JE, Arighi CN, Wu CH, Natale DA: **Protein Ontology on the semantic web for knowledge discovery** *Sci Data* 2020, **7**:1–12.

47. Lang PF, Chebaro Y, Zheng X, P. Sekar JA, Shaikh B, Natale DA, Karr JR: **BpForms and BcForms: a toolkit for concretely describing non-canonical polymers and complexes to facilitate global biochemical networks** *Genome Biol* 2020, **21**:1–21.

48. Zhang T, Li H, Xi H, Stanton RV, Rotstein SH: **HELM: a hierarchical notation language for complex biomolecule structure representation.** *J Chem Inf Model* 2012, **52**:2796–2806.

49. Westbrook JD, Fitzgerald P: **The PDB format, mmCIF, and other data formats** *Methods Biochem Anal* 2003, **44**:161–179.

50. Sivade M, Alonso-López D, Ammari M, Bradley G, Campbell NH, Ceol A, Cesareni G, Combe C, De Las Rivas J, Del-Toro N, *et al.*: **Encompassing new use cases-level 3.0 of the HUPO-PSI format for molecular interactions** *BMC Bioinformatics* 2018, **19**:1–8.

51. Pierleoni A, Martelli PL, Fariselli P, Casadio R: **eSLDB: eukaryotic subcellular localization database** *Nucleic Acids Res* 2007, **35**:D208–D212.

52. Thul PJ, Lindskog C: **The Human Protein Atlas: a spatial map of the human proteome** *Protein Sci* 2018, **27**:233–244.

53. Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, Begley K, Field D, Harris S, Hide W, Hofmann O, *et al.*: **ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level.** *Bioinformatics* 2010, **26**:2354–2356.

54. Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang WH, Römpp A, Neumann S, Pizarro AD, *et al.*: **mzML–a community standard for mass spectrometry data** *Mol Cell Proteomics* 2011, **10**:R110–000133.

55. Boccaletto P, Machnicka MA, Purta E, Piątkowski P, Bagiński B, Wirecki TK, Crécy-Lagard V de, Ross R, Limbach PA, Kotter A, *et al.*: **MODOMICS: a database of RNA modification pathways. 2017 update** *Nucleic Acids Res* 2018, **46**:D303–D307.

56. Zhang T, Tan P, Wang L, Jin N, Li Y, Zhang L, Yang H, Hu Z, Zhang L, Hu C, *et al.*: **RNALocate: a resource for RNA subcellular localizations** *Nucleic Acids Res* 2017, **45**:D135–D138.

57. Mas-Ponte D, Carlevaro-Fita J, Palumbo E, Pulido TH, Guigo R, Johnson R: **lncATLAS database for subcellular localization of long noncoding rnas** *RNA* 2017, **23**:1080–1087.

58. Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, Snow C, Fonseca NA, Petryszak R, Papatheodorou I, *et al.*: **ArrayExpress update–from bulk to single-cell expression data** *Nucleic Acids Res* 2019, **47**:D711–D715.

59. Clough E, Barrett TThe Gene Expression Omnibus databasein: Statistical Genomics Springer, 2016pp. 93–110.

60. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The sequence alignment/map format and SAMtools** *Bioinformatics* 2009, **25**:2078–2079.

61. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM: **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants** *Nucleic Acids Res* 2010, **38**:1767–1771.

62. Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, Keseler IM, Krummenacker M, Midford PE, Ong Q, *et al.*: **The BioCyc ollection of microbial genomes and metabolic pathways** *Brief Bioinform* 2019, **20**:1085–1093.

63. Meldal BHM, Bye-A-Jee H, Gajdoš L, Hammerová Z, Horáčková A, Melicher F, Perfetto L, Pokornỳ D, Lopez MR, Türková A, *et al.*: **Complex Portal 2018: extended content and enhanced visualization tools for macromolecular complexes** *Nucleic Acids Res* 2019, **47**:D550–D558.

64. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M: **KEGG: integrating viruses and cellular organisms** *Nucleic Acids Res* 2021, **49**:D545–D551.

65. Moretti S, Tran VDT, Mehl F, Ibberson M, Pagni M: **MetaNetX/MNXref: unified namespace for metabolites and biochemical reactions in the context of metabolic models** *Nucleic Acids Res* 2021, **49**:D570–D574.

66. Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, Wu G, D'eustachio P, Schaefer C, Luciano J, *et al.*: **The BioPAX community standard for pathway data sharing.** *Nat Biotechnol* 2010, **28**:935–942.

67. Gardossi L, Poulsen PB, Ballesteros A, Hult K, Švedas VK, Vasić-Rački Đ, Carrea G, Magnusson A, Schmid A, Wohlgemuth R, *et al.*: **Guidelines for reporting of biocatalytic reactions** *Trends Biotechnol* 2010, **28**:171–180.

68. Zhang Z, Shen T, Rui B, Zhou W, Zhou X, Shang C, Xin C, Liu X, Li G, Jiang J, *et al.*: **CeCaFDB: a curated database for the documentation, visualization and comparative analysis of central carbon metabolic flux distributions explored by 13c-fluxomics** *Nucleic Acids Res* 2015, **43**:D549–D557.

69. Medvedeva YA, Lennartsson A, Ehsani R, Kulakovskiy IV, Vorontsov IE, Panahandeh P, Khimulya G, Kasukawa T, Drabløs F, Consortium F, *et al.*: **EpiFactors: a comprehensive database of human epigenetic factors and complexes** *Database* 2015.

70. Fornes O, Castro-Mondragon JA, Khan A, Lee R Van der, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D, *et al.*: **JASPAR 2020: update of the open-access database of transcription factor binding profiles** *Nucleic Acids Res* 2020, **48**:D87–D92.

71. Wingender E, Dietze P, Karas H, Knüppel R: **TRANSFAC: a database on transcription factors and their dna binding sites** *Nucleic Acids Res* 1996, **24**:238–241.

72. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, *et al.*: **ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia** *Genome Res* 2012, **22**:1813–1831.

73. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, *et al.*: **The IntAct molecular interaction database in 2012** *Nucleic Acids Res* 2012, **40**:D841–D846.

74. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, *et al.*: **The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets** *Nucleic Acids Res* 2021, **49**:D605–D612.

75. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, Silva Santos LB da, Bourne PE, *et al.*: **The FAIR Guiding Principles for scientific data management and stewardship.** *Sci Data* 2016, **3**:160018.

76. Friedman SH, Anderson AR, Bortz DM, Fletcher AG, Frieboes HB, Ghaffarizadeh A, Grimes DR, Hawkins-Daarud A, Hoehme S, Juarez EF, *et al.*: **MultiCellDS: a standard and a community for sharing multicellular data.** *bioRxiv* 2016, 090696.

77. Karr JR, Sanghvi JC, Macklin DN, Arora A, Covert MW: **WholeCellKB: model organism databases for comprehensive whole-cell models.** *Nucleic Acids Res* 2012, **41**:D787–D792.

78. Lubitz T, Hahn J, Bergmann FT, Noor E, Klipp E, Liebermeister W: **SBtab: a flexible table format for data exchange in systems biology.** *Bioinformatics* 2016, **32**:2559–2561.

79. Karr JR, Liebermeister W, Goldberg AP, Sekar JA, Shaikh B: **Structured spreadsheets with ObjTables enable data reuse and integration.** *arXiv* 2020, 2005.05227.

80. Wolstencroft K, Owen S, Krebs O, Nguyen Q, Stanford NJ, Golebiewski M, Weidemann A, Bittkowski M, An L, Shockley D, *et al.*: **SEEK: a systems biology data and model management platform.** *BMC Syst Biol* 2015, **9**:1–12.

81. Aebersold R, Agar JN, Amster IJ, Baker MS, Bertozzi CR, Boja ES, Costello CE, Cravatt BF, Fenselau C, Garcia BA, *et al.*: **How many human proteoforms are there?** *Nat Chem Biol* 2018, **14**:206–214.

82. Lang PF, Chebaro Y, Zheng X, P Sekar JA, Shaikh B, Natale DA, Karr JR: **BpForms and BcForms: a toolkit for concretely describing non-canonical polymers and complexes to facilitate global biochemical networks.** *Genome Biol* 2020, **21**:117.

83. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, *et al.*: **NCBI Taxonomy: a comprehensive update on curation, resources and tools** *Database* 2020.

84. Sarntivijai S, Lin Y, Xiang Z, Meehan TF, Diehl AD, Vempati UD, Schürer SC, Pang C, Malone J, Parkinson H, *et al.*: **CLO: the Cell Line Pntology** *J Biomed Semant* 2014, **5**:1–10.

85. Dunnen JT den, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, Roux AF, Smith T, Antonarakis SE, Taschner PE, *et al.*: **HGVS recommendations for the description of sequence variants: 2016 update** *Hum Mutat* 2016, **37**:564–569.

86. Oberhardt MA, Zarecki R, Gronow S, Lang E, Klenk HP, Gophna U, Ruppin E: **Harnessing the landscape of microbial culture media to predict new organism–media pairings** *Nat Commun* 2015, **6**:1–14.

87. Richards MA, Cassen V, Heavner BD, Ajami NE, Herrmann A, Simeonidis E, Price ND: **MediaDB: a database of microbial growth conditions in defined media** *PLoS One* 2014, **9**:e103548.

88. Sansone SA, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, Thurston M: **FAIRsharing as a community approach to standards, repositories and policies** *Nat Biotechnol* 2019, **37**:358–367.

89. Sundararaj S, Guo A, Habibi-Nazhad B, Rouani M, Stothard P, Ellison M, Wishart DS: **The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate in silico modeling of Escherichia coli.** *Nucleic Acids Res* 2004, **32**:D293–D295.

    *The CCDB (http://ccdb.wishartlab.com) is a pioneering database that was developed to facilitate models of *E. coli*. By centralizing information about the structure and abundance of metabolites, RNAs, and proteins, the CCDB enables modelers to focus on creating models rather than on aggregating data.

90. Keseler IM, Mackie A, Santos-Zavaleta A, Billington R, Bonavides-Martínez C, Caspi R, Fulcher C, Gama-Castro S, Kothari A, Krummenacker M, *et al.*: **The EcoCyc database: reflecting new knowledge about Escherichia coli K-12.** *Nucleic Acids Res* 2017, **45**:D543–D550.

    **EcoCyc (http://ecocyc.org) and the broader BioCyc (http://biocyc.org) collection of pathway-genome databases are some of the most comprehensive and highest quality resources for qualitative and relational information for whole-cell modeling. For example, EcoCyc has been a key source of data for models of the metabolism of *E. coli*. The Pathway Tools software used to build EcoCyc and BioCyc could also be useful for organizing data for specific models.

91. Crasto CJ, Marenco LN, Liu N, Morse TM, Cheung KH, Lai PC, Bahl G, Masiar P, Lam HY, Lim E, *et al.*: **SenseLab: new developments in disseminating neuroscience information.** *Brief Bioinformatics* 2007, **8**:150–162.

    *CellPropDB and NeuronDB (https://senselab.med.yale.edu) are pioneering databases that were developed to facilitate models of neurons. By providing data about the expression of membrane channels, receptors, and neurotransmitters, the databases enable modelers to focus on building better models.

92. Latendresse M, Krummenacker M, Trupp M, Karp PD: **Construction and completion of flux balance models from pathway databases.** *Bioinformatics* 2012, **28**:388–396.

93. Mondeel TD, Crémazy F, Barberis M: **GEMMER: GEnome-wide tool for Multi-scale Modeling data Extraction and Representation for Saccharomyces cerevisiae.** *Bioinformatics* 2018, **34**:2147–2149.

94. Perez-Riverol Y, Bai M, Veiga Leprevost F da, Squizzato S, Park YM, Haug K, Carroll AJ, Spalding D, Paschall J, Wang M, *et al.*: **Discovering and linking public omics data sets using the Omics Discovery Index.** *Nat Biotechnol* 2017, **35**:406–409.

**OmicsDI (https://www.omicsdi.org) is one of the most comprehensive search engines for quantitative omics data. OmicsDI encompasses data for a wide range of organisms and cell types. OmicsDI's distributed approach to data aggregation both enables many investigators to contribute to OmicsDI and enables experts to quality control each type of data contained in the database.

95. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur Ö, Anwar N, Schultz N, Bader GD, Sander C: **Pathway Commons, a web resource for biological pathway data.** *Nucleic Acids Res* 2010, **39**:D685–D690.

96. Roth YD, Lian Z, Pochiraju S, Shaikh B, Karr JR: **Datanator: an integrated database of molecular data for quantitatively modeling cellular behavior.** *Nucleic Acids Res* 2021, **49**:D516–D522.

**Datanator (https://datanator.info) is an integrated database of several key types of data for modeling cells. To help investigators best leverage the limited data available for modeling, Datanator provides tools for assembling clouds of measurements centered around specific molecules and molecular interactions in a specific organism. As a data warehouse, Datanator also provides this data in a consistent format.

97. Percha B, Garten Y, Altman RBDiscovery and explanation of drug-drug interactions via text miningin: BiocomputingWorld Scientific, 2012pp. 410–421.

98. Bird SNLTK: the Natural Language Toolkitin: Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions2006pp. 69–72.