

Centralizing data to unlock whole-cell models

Chew, Yin Hoon; Karr, Jonathan R.

DOI:

[10.1016/j.coisb.2021.06.004](https://doi.org/10.1016/j.coisb.2021.06.004)

License:

Creative Commons: Attribution-NonCommercial-NoDerivs (CC BY-NC-ND)

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>

[Link to publication on Research at Birmingham portal](#)

General rights

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

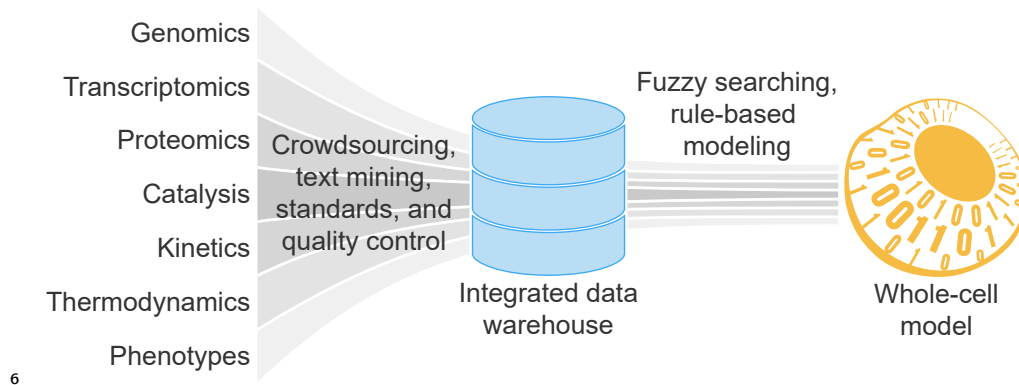
If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

1 Centralizing data to unlock whole-cell models

2 Yin Hoon Chew and Jonathan R. Karr

3 Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai,
4 1425 Madison Avenue, New York, NY 10029, USA

5 Graphical abstract



7 Highlights

- 8 • Whole-cell models require data about each molecule and molecular interaction
9 • Data is increasingly available, but its scattered organization hinders modeling
10 • A central database of data and knowledge would accelerate whole-cell modeling
11 • Such a database requires collaboration and standardization
12 • New experimental methods and automation are also needed to broaden and deepen our
13 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

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

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

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

52 **The mountain of data needed to model an entire cell**

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

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

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

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

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

77 Exemplary data resources that we believe can be repurposed for whole-cell modeling include,
78 **but are not limited to, the Protein Data Bank (PDB) [29], ECMDB [30], YMDB [31], PaxDB**
79 **[32], PSORTdb [33], BRENDA [34], and SABIO-RK [35] (Table 1).** ECMDB and YMDB
80 contain thousands of measurements of the concentrations of metabolites in *E. coli* and *S.*
81 *cerevisiae*. PaxDB contains over 1 million measurements of the abundances of proteins
82 in over 50 organisms. PSORTdb contains over 10,000 measurements of the localization of
83 proteins in over 400 organisms, as well as predicted localizations for over 15,000 organisms.
84 Together, BRENDA and SABIO-RK contain over 300,000 kinetic parameters for thousands
85 of metabolic reactions. In our experience, BioNumbers [36] is also a valuable **resource for data**
86 **that is outside the scope of repositories for specific types of data. For example, BioNumbers**
87 **contains data about the rates of non-metabolic processes such as DNA damage and RNA**
88 **polymerization; the fluxes of the exchange of nutrients into and out of cells; and the sizes,**
89 **densities, and growth rates of cells, which are not contained in other repositories.**

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

96 **The challenges to reusing data for whole-cell modeling**

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

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

110 Third, many databases and articles only provide minimal metadata or minimally structured
111 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], IncATLAS [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.

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

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

122 **Emerging tools for sharing, discovering, and reusing data**

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

127 *Formats for exchanging data for whole-cell modeling*

128 Three notable formats for capturing some of the data and knowledge needed for whole-cell
129 modeling include the Investigation/Study/Assay tabular (ISA-Tab) format [53], the Mul-
130 ticellular Data Standard (MultiCellDS) [76], and BioPAX [66]. ISA-Tab is ideal for high-
131 dimensional data, such as transcriptome-wide measurements of RNA turnover rates, which
132 lack more specific formats. MultiCellDS is an emerging format intended to capture a digi-
133 tal “snapshot” of a cell line, encompassing measurements of its metabolome, transcriptome,
134 proteome, and phenotype, as well as metadata about the environmental context of each mea-
135 surement and the methods used to collect it. BioPAX is a format for describing knowledge
136 about the molecules and molecular interactions inside cells.

137 In our experience, whole-cell modeling requires both quantitative and relational data about
138 multiple aspects of a cell. To capture this information for our first models, we developed
139 the WholeCellKB schema [77]. Simultaneously, Lubitz and colleagues developed SBTab
140 [78], a tabular format with similar goals. As we began to explore additional models, we
141 realized that many modelers both want to be able to use spreadsheets to quickly assemble
142 datasets and use computer programs to quality control their datasets and incorporate them
143 into models. To meet this need, we recently merged the concepts behind WholeCellKB and
144 SBTab into ObjTables [79], a set of tools that make it easy for modelers to use user-friendly
145 spreadsheets to integrate data, define schemas for rigorously validating their data, and parse
146 linked spreadsheets into data structures that are conducive to modeling. SEEK provides an
147 online environment for managing datasets organized as spreadsheets [80].

148 *Formats for critical metadata for whole-cell modeling*

149 As we discussed above, structured metadata is critical for understanding and merging data.
150 Because cells contain millions of distinct molecular species [81] due to combinatorial bio-
151 chemical processes such as post-transcriptional and post-translational modification and com-
152 plexation, we think that it is particularly important for datasets to concretely describe the
153 molecules and molecular interactions that they characterize. Small molecules can be de-
154 scribed using several formats such as the Chemical Markup Language (CML) [63] and IU-
155 PAC International Chemical Identifier (InChI) [44] formats. Sequences of unmodified DNAs,
156 RNAs, and proteins can be described using the FASTA format. Sequences of modified DNAs,
157 RNAs, and proteins can be described using BpForms [82] and HELM [48]. BpForms general-
158 izes the IUPAC and IUBMB formats commonly used to describe unmodified DNAs, RNAs,
159 and proteins to capture physiological polymers with modifications, crosslinks, and nicks.
160 Macromolecular complexes can be described using BcForms [82] and HELM.

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

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

170 *Centralized knowledgebases of information for whole-cell modeling*

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

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

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

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

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

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

216 Roadmap to data for whole-cell modeling

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

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

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

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

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

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

250 Fifth, the community could develop additional primary databases for types of data that
251 are not covered by the existing primary databases. For example, a group of researchers
252 is beginning to assemble a database of the thermodynamics of biochemical reactions. Each
253 database could be initiated by a small team of curators who seed the database by aggregating
254 their own data and data from the literature. Beyond this initial phase, these databases
255 could allow the community to submit data directly. In some cases, text mining could also
256 be used to automatically or semi-automatically extract data from the literature. One area
257 where text mining has been successful is collating interactions between genes and drugs
258 [97]. Foundational tools for text mining include the Natural Language Toolkit [98] and
259 spaCy. Collectively, multiple such primary databases would be able to support a broad
260 range of formats for different types of data. These primary databases would also be well-
261 positioned for expert curators to quality control specific types of data. Furthermore, such
262 primary databases might be able to assemble the critical mass of investigators needed to
263 lobby journals to require public deposition of specific types of data.

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

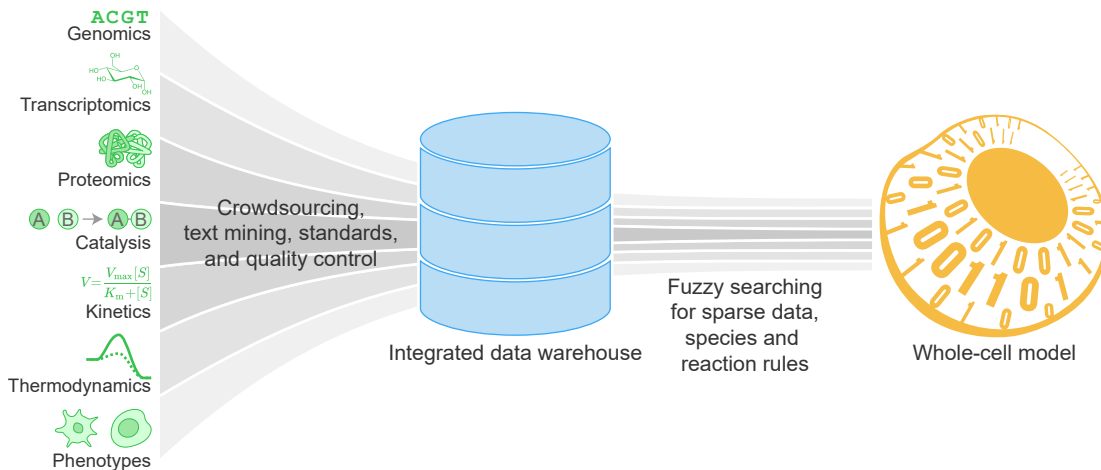


Figure 1: An integrated warehouse of molecular data and knowledge is needed to accelerate whole-cell modeling. This warehouse could be assembled by combining multiple crowdsourced databases for different types of data with data automatically mined from the literature. Models could be systematically constructed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

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

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

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

290 Conclusions

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

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

299 Declaration of competing interest

300 None.

301 Acknowledgments

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

306 References

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

309 * of special interest

310 ** of outstanding interest

- 311 1. Carrera J, Covert MW: **Why build whole-cell models?** *Trends Cell Biol* 2015,
312 **25**:719–722.
- 313 2. Tomita M: **Whole-cell simulation: a grand challenge of the 21st century.** *Trends*
314 *Biotechnol* 2001, **19**:205–210.
- 315 3. Marucci L, Barberis M, Karr J, Ray O, Race PR, Souza Andrade M de, Grierson C,
316 Hoffmann SA, Landon S, Rech E, *et al.*: **Computer-aided whole-cell design: taking**
317 **a holistic approach by integrating synthetic with systems biology.** *Front Bioeng*
318 *Biotechnol* 2020, **8**:942.
- 319 4. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival Jr B, Assad-
320 Garcia N, Glass JI, Covert MW: **A whole-cell computational model predicts phe-**
321 **notype from genotype.** *Cell* 2012, **150**:389–401.
- 322 5. Burke PE, Claudia BdL, Costa LdF, Quiles MG: **A biochemical network modeling**
323 **of a whole-cell.** *Sci Rep* 2020, **10**:1–14.

- 324 6. Thornburg ZR, Melo MC, Bianchi D, Brier TA, Crotty C, Breuer M, Smith HO, Hutchi-
325 son III CA, Glass JI, Luthey-Schulten Z: **Kinetic modeling of the genetic informa-**
326 **tion processes in a minimal cell.** *Front Mol Biosci* 2019, **6**:130.
- 327 7. Thiele I, Jamshidi N, Fleming RM, Palsson BØ: **Genome-scale reconstruction of**
328 **Escherichia coli’s transcriptional and translational machinery: a knowledge**
329 **base, its mathematical formulation, and its functional characterization** *PLoS*
330 *Comput Biol* 2009, **5**:e1000312.
- 331 8. Roberts E, Magis A, Ortiz JO, Baumeister W, Luthey-Schulten Z: **Noise contributions**
332 **in an inducible genetic switch: a whole-cell simulation study.** *PLoS Comput Biol*
333 2011, **7**:e1002010.
- 334 9. Carrera J, Estrela R, Luo J, Rai N, Tsoukalas A, Tagkopoulos I: **An integrative, multi-**
335 **scale, genome-wide model reveals the phenotypic landscape of Escherichia**
336 **coli.** *Mol Syst Biol* 2014, **10**:735.
- 337 10. Macklin DN, Ahn-Horst TA, Choi H, Ruggero NA, Carrera J, Mason JC, Sun G, Ag-
338 mon E, DeFelice MM, Maayan I, *et al.*: **Simultaneous cross-evaluation of hetero-**
339 **geneous E. coli datasets via mechanistic simulation.** *Science* 2020.
- 340 11. Münzner U, Klipp E, Krantz M: **A comprehensive, mechanistically detailed, and**
341 **executable model of the cell division cycle in Saccharomyces cerevisiae.** *Nat*
342 *Commun* 2019, **10**:1–12.
- 343 12. Ye C, Xu N, Gao C, Liu G, Xu J, Zhang W, Chen X, Nielsen J, Liu L: **Comprehensive**
344 **understanding of Saccharomyces cerevisiae phenotypes with whole-cell model**
345 **WM_S288C.** *Biotechnol Bioeng* 2020, **117**:1562–1574.
- 346 13. Ghaemi Z, Peterson JR, Gruebele M, Luthey-Schulten Z: **An in-silico human cell**
347 **model reveals the influence of spatial organization on RNA splicing.** *PLoS*
348 *Comput Biol* 2020, **16**:e1007717.
- 349 14. Bordbar A, McCloskey D, Zielinski DC, Sonnenschein N, Jamshidi N, Palsson BO: **Per-**
350 **sonalized whole-cell kinetic models of metabolism for discovery in genomics**
351 **and pharmacodynamics.** *Cell Syst* 2015, **1**:283–292.
- 352 15. Purcell O, Jain B, Karr JR, Covert MW, Lu TK: **Towards a whole-cell modeling**
353 **approach for synthetic biology.** *Chaos* 2013, **23**:025112.
- 354 16. Rees-Garbutt J, Chalkley O, Landon S, Purcell O, Marucci L, Grierson C: **Designing**
355 **minimal genomes using whole-cell models.** *Nat Commun* 2020, **11**:1–12.
- 356 17. Takahashi K, Yugi K, Hashimoto K, Yamada Y, Pickett CJ, Tomita M: **Computational**
357 **challenges in cell simulation: a software engineering approach.** *IEEE Intell Syst*
358 2002, **17**:64–71.
- 359 18. Im W, Liang J, Olson A, Zhou HX, Vajda S, Vakser IA: **Challenges in structural**
360 **approaches to cell modeling.** *J Mol Biol* 2016, **428**:2943–2964.

- 361 19. Luthey-Schulten Z: **Integrating experiments, theory and simulations into whole-**
362 **cell models.** *Nat Methods* 2021, **18**:446–447.
- 363 20. Goldberg AP, Chew YH, Karr JR: **Toward scalable whole-cell modeling of human**
364 **cells.** *Proc 2016 ACM SIGSIM Conf Princip Adv Discrete Simul* 2016, 259–262.
- 365 21. Babbie AC, Stumpf MPH: **How to deal with parameters for whole-cell modelling.**
366 *J R Soc Interface* 2017, **14**:20170237.
- 367 22. Stumpf MPH: **Statistical and computational challenges for whole cell mod-**
368 **elling.** *Curr Opin Syst Biol* 2021.
- 369 23. Macklin DN, Ruggero NA, Covert MW: **The future of whole-cell modeling.** *Curr*
370 *Opin Biotechnol* 2014, **28**:111–115.
- 371 24. Feig M, Sugita Y: **Whole-cell models and simulations in molecular detail.** *Annu*
372 *Rev Cell Dev Biol* 2019, **35**:191–211.
- 373 25. Singla J, White KL: **A community approach to whole-cell modeling.** *Curr Opin*
374 *Syst Biol* 2021.
- 375 26. Waltemath D, Karr JR, Bergmann FT, Chelliah V, Hucka M, Krantz M, Liebermeister
376 W, Mendes P, Myers CJ, Pir P, *et al.*: **Toward community standards and**
377 **software for whole-cell modeling.** *IEEE Trans Biomed Eng* 2016, **63**:2007–2014.
- 378 27. Goldberg AP, Szigeti B, Chew YH, Sekar JA, Roth YD, Karr JR: **Emerging whole-cell**
379 **modeling principles and methods.** *Curr Opin Biotechnol* 2018, **51**:97–102.
- 380 28. Szigeti B, Roth YD, Sekar JA, Goldberg AP, Pochiraju SC, Karr JR: **A blueprint for**
381 **human whole-cell modeling.** *Curr Opinion Systems Biol* 2018, **7**:8–15.
- 382 29. wwPDB consortium: **Protein Data Bank: the single global archive for 3D**
383 **macromolecular structure data** *Nucleic Acids Res* 2019, **47**:D520–D528.
- 384 30. Sajed T, Marcu A, Ramirez M, Pon A, Guo AC, Knox C, Wilson M, Grant JR, Djoumbou
385 Y, Wishart DS: **ECMDB 2.0: A richer resource for understanding the**
386 **biochemistry of E. coli.** *Nucleic Acids Res* 2016, **44**:D495–D501.
- 387 31. Ramirez-Gaona M, Marcu A, Pon A, Guo AC, Sajed T, Wishart NA, Karu N, Djoumbou
388 Feunang Y, Arndt D, Wishart DS: **YMDB 2.0: a significantly expanded ver-**
389 **sion of the yeast metabolome database.** *Nucleic Acids Res* 2017, **45**:D440–D445.
- 390 32. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, Mering C von: **Version 4.0 of**
391 **PaxDb: protein abundance data, integrated across model organisms, tissues,**
392 **and cell-lines.** *Proteomics* 2015, **15**:3163–3168.
- 393 33. Lau WYV, Hoad GR, Jin V, Winsor GL, Madyan A, Gray KL, Laird MR, Lo R,
394 Brinkman FSL: **PSORTdb 4.0: expanded and redesigned bacterial and arch-**
395 **aeal protein subcellular localization database incorporating new secondary**
396 **localizations.** *Nucleic Acids Res* 2021, **49**:D803–D808.

- 397 34. Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitiz J, Schomburg I, Neumann-Schaal M,
398 Jahn D, Schomburg D: **BRENDA, the ELIXIR core data resource in 2021: new**
399 **developments and updates.** *Nucleic Acids Res* 2021, **49**:D498–D508.
- 400 35. Wittig U, Rey M, Weidemann A, Kania R, Müller W: **SABIO-RK: an updated**
401 **resource for manually curated biochemical reaction kinetics.** *Nucleic Acids Res*
402 2018, **46**:D656–D660.
- 403 36. Milo R, Jorgensen P, Moran U, Weber G, Springer M: **BioNumbers—the database of**
404 **key numbers in molecular and cell biology.** *Nucleic Acids Res* 2010, **38**:D750–D753.
- 405 37. Harrison PW, Ahamed A, Aslam R, Alako BT, Burgin J, Buso N, Courtot M, Fan J,
406 Gupta D, Haseeb M, *et al.*: **The European Nucleotide Archive in 2020** *Nucleic*
407 *Acids Res* 2021, **49**:D82–D85.
- 408 38. Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST, Karsch-
409 Mizrachi I: **GenBank** *Nucleic Acids Res* 2021, **49**:D92–D96.
- 410 39. Field D, Sterk P, Kottmann R, De Smet JW, Amaral-Zettler L, Cochrane G, Cole JR,
411 Davies N, Dawyndt P, Garrity GM, *et al.*: **Genomic Standards Consortium**
412 **projects** *Standards Genomic Sci* 2014, **9**:599–601.
- 413 40. Sood AJ, Viner C, Hoffman MM: **DNAmoD: the DNA modification database** *J*
414 *Cheminform* 2019, **11**:1–10.
- 415 41. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swain-
416 ston N, Mendes P, Steinbeck C: **ChEBI in 2016: Improved services and an ex-**
417 **panding collection of metabolites** *Nucleic Acids Res* 2016, **44**:D1214–D1219.
- 418 42. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA,
419 Yu B, *et al.*: **PubChem in 2021: new data content and improved web interfaces**
420 *Nucleic Acids Res* 2021, **49**:D1388–D1395.
- 421 43. Murray-Rust P, Rzepa HS, Wright M: **Development of chemical markup language**
422 **(CML) as a system for handling complex chemical content** *New J Chem* 2001,
423 **25**:618–634.
- 424 44. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D: **InChI, the IUPAC**
425 **international chemical identifier** *J Cheminform* 2015, **7**:1–34.
- 426 45. Fiehn O, Robertson D, Griffin J, Werf M van der, Nikolau B, Morrison N, Sumner LW,
427 Goodacre R, Hardy NW, Taylor C, *et al.*: **The metabolomics standards initiative**
428 **(msi)** *Metabolomics* 2007, **3**:175–178.
- 429 46. Chen C, Huang H, Ross KE, Cowart JE, Arighi CN, Wu CH, Natale DA: **Protein**
430 **Ontology on the semantic web for knowledge discovery** *Sci Data* 2020, **7**:1–12.
- 431 47. Lang PF, Chebaro Y, Zheng X, P. Sekar JA, Shaikh B, Natale DA, Karr JR: **BpForms**
432 **and BcForms: a toolkit for concretely describing non-canonical polymers and**
433 **complexes to facilitate global biochemical networks** *Genome Biol* 2020, **21**:1–21.

- 434 48. Zhang T, Li H, Xi H, Stanton RV, Rotstein SH: **HELM: a hierarchical notation**
435 **language for complex biomolecule structure representation.** *J Chem Inf Model*
436 2012, **52**:2796–2806.
- 437 49. Westbrook JD, Fitzgerald P: **The PDB format, mmCIF, and other data formats**
438 *Methods Biochem Anal* 2003, **44**:161–179.
- 439 50. Sivade M, Alonso-López D, Ammari M, Bradley G, Campbell NH, Ceol A, Cesareni G,
440 Combe C, De Las Rivas J, Del-Toro N, *et al.*: **Encompassing new use cases-level**
441 **3.0 of the HUPO-PSI format for molecular interactions** *BMC Bioinformatics*
442 2018, **19**:1–8.
- 443 51. Pierleoni A, Martelli PL, Fariselli P, Casadio R: **eSLDB: eukaryotic subcellular**
444 **localization database** *Nucleic Acids Res* 2007, **35**:D208–D212.
- 445 52. Thul PJ, Lindskog C: **The Human Protein Atlas: a spatial map of the human**
446 **proteome** *Protein Sci* 2018, **27**:233–244.
- 447 53. Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, Begley K, Field D, Harris S,
448 Hide W, Hofmann O, *et al.*: **ISA software suite: supporting standards-compliant**
449 **experimental annotation and enabling curation at the community level.** *Bioin-*
450 *formatics* 2010, **26**:2354–2356.
- 451 54. Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang WH,
452 Römpp A, Neumann S, Pizarro AD, *et al.*: **mzML—a community standard for mass**
453 **spectrometry data** *Mol Cell Proteomics* 2011, **10**:R110–000133.
- 454 55. Boccaletto P, Machnicka MA, Purta E, Piątkowski P, Bagiński B, Wirecki TK, Crécy-
455 Lagard V de, Ross R, Limbach PA, Kotter A, *et al.*: **MODOMICS: a database of**
456 **RNA modification pathways. 2017 update** *Nucleic Acids Res* 2018, **46**:D303–D307.
- 457 56. Zhang T, Tan P, Wang L, Jin N, Li Y, Zhang L, Yang H, Hu Z, Zhang L, Hu C, *et al.*:
458 **RNAlocate: a resource for RNA subcellular localizations** *Nucleic Acids Res*
459 2017, **45**:D135–D138.
- 460 57. Mas-Ponte D, Carlevaro-Fita J, Palumbo E, Pulido TH, Guigo R, Johnson R: **IncAT-**
461 **LAS database for subcellular localization of long noncoding rnas** *RNA* 2017,
462 **23**:1080–1087.
- 463 58. Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, Snow C, Fonseca NA,
464 Petryszak R, Papatheodorou I, *et al.*: **ArrayExpress update—from bulk to single-**
465 **cell expression data** *Nucleic Acids Res* 2019, **47**:D711–D715.
- 466 59. Clough E, Barrett T *The Gene Expression Omnibus database in: Statistical Genomic-*
467 *sSpringer, 2016pp. 93–110.*
- 468 60. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
469 Durbin R: **The sequence alignment/map format and SAMtools** *Bioinformatics*
470 2009, **25**:2078–2079.

- 471 61. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM: **The Sanger FASTQ file format**
472 **for sequences with quality scores, and the Solexa/Illumina FASTQ variants**
473 *Nucleic Acids Res* 2010, **38**:1767–1771.
- 474 62. Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, Keseler IM,
475 Krummenacker M, Midford PE, Ong Q, *et al.*: **The BioCyc ollection of microbial**
476 **genomes and metabolic pathways** *Brief Bioinform* 2019, **20**:1085–1093.
- 477 63. Meldal BHM, Bye-A-Jee H, Gajdoš L, Hammerová Z, Horáčková A, Melicher F, Per-
478 fetto L, Pokorný D, Lopez MR, Tůrková A, *et al.*: **Complex Portal 2018: extended**
479 **content and enhanced visualization tools for macromolecular complexes** *Nu-*
480 *cleic Acids Res* 2019, **47**:D550–D558.
- 481 64. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M: **KEGG: inte-**
482 **grating viruses and cellular organisms** *Nucleic Acids Res* 2021, **49**:D545–D551.
- 483 65. Moretti S, Tran VDT, Mehl F, Ibberson M, Pagni M: **MetaNetX/MNXref: uni-**
484 **fied namespace for metabolites and biochemical reactions in the context of**
485 **metabolic models** *Nucleic Acids Res* 2021, **49**:D570–D574.
- 486 66. Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, Wu G, D’eustachio P,
487 Schaefer C, Luciano J, *et al.*: **The BioPAX community standard for pathway**
488 **data sharing.** *Nat Biotechnol* 2010, **28**:935–942.
- 489 67. Gardossi L, Poulsen PB, Ballesteros A, Hult K, Švedas VK, Vasić-Rački Đ, Carrea G,
490 Magnusson A, Schmid A, Wohlgemuth R, *et al.*: **Guidelines for reporting of bio-**
491 **catalytic reactions** *Trends Biotechnol* 2010, **28**:171–180.
- 492 68. Zhang Z, Shen T, Rui B, Zhou W, Zhou X, Shang C, Xin C, Liu X, Li G, Jiang J,
493 *et al.*: **CeCaFDB: a curated database for the documentation, visualization and**
494 **comparative analysis of central carbon metabolic flux distributions explored**
495 **by 13c-fluxomics** *Nucleic Acids Res* 2015, **43**:D549–D557.
- 496 69. Medvedeva YA, Lennartsson A, Ehsani R, Kulakovskiy IV, Vorontsov IE, Panahan-
497 deh P, Khimulya G, Kasukawa T, Drabløs F, Consortium F, *et al.*: **EpiFactors: a**
498 **comprehensive database of human epigenetic factors and complexes** *Database*
499 2015.
- 500 70. Fornes O, Castro-Mondragon JA, Khan A, Lee R Van der, Zhang X, Richmond PA,
501 Modi BP, Correard S, Gheorghe M, Baranašić D, *et al.*: **JASPAR 2020: update of**
502 **the open-access database of transcription factor binding profiles** *Nucleic Acids*
503 *Res* 2020, **48**:D87–D92.
- 504 71. Wingender E, Dietze P, Karas H, Knüppel R: **TRANSFAC: a database on tran-**
505 **scription factors and their dna binding sites** *Nucleic Acids Res* 1996, **24**:238–241.
- 506 72. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE,
507 Bickel P, Brown JB, Cayting P, *et al.*: **ChIP-seq guidelines and practices of the**
508 **ENCODE and modENCODE consortia** *Genome Res* 2012, **22**:1813–1831.

- 509 73. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M,
510 Dumousseau M, Feuermann M, Hinz U, *et al.*: **The IntAct molecular interaction**
511 **database in 2012** *Nucleic Acids Res* 2012, **40**:D841–D846.
- 512 74. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT,
513 Legeay M, Fang T, Bork P, *et al.*: **The STRING database in 2021: customizable**
514 **protein–protein networks, and functional characterization of user-uploaded**
515 **gene/measurement sets** *Nucleic Acids Res* 2021, **49**:D605–D612.
- 516 75. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A,
517 Blomberg N, Boiten JW, Silva Santos LB da, Bourne PE, *et al.*: **The FAIR Guid-**
518 **ing Principles for scientific data management and stewardship.** *Sci Data* 2016,
519 **3**:160018.
- 520 76. Friedman SH, Anderson AR, Bortz DM, Fletcher AG, Frieboes HB, Ghaffarizadeh A,
521 Grimes DR, Hawkins-Daarud A, Hoehme S, Juarez EF, *et al.*: **MultiCellDS: a stan-**
522 **dard and a community for sharing multicellular data.** *bioRxiv* 2016, 090696.
- 523 77. Karr JR, Sanghvi JC, Macklin DN, Arora A, Covert MW: **WholeCellKB: model**
524 **organism databases for comprehensive whole-cell models.** *Nucleic Acids Res*
525 2012, **41**:D787–D792.
- 526 78. Lubitz T, Hahn J, Bergmann FT, Noor E, Klipp E, Liebermeister W: **SBtab: a flexible**
527 **table format for data exchange in systems biology.** *Bioinformatics* 2016, **32**:2559–
528 2561.
- 529 79. Karr JR, Liebermeister W, Goldberg AP, Sekar JA, Shaikh B: **Structured spread-**
530 **sheets with ObjTables enable data reuse and integration.** *arXiv* 2020,
531 2005.05227.
- 532 80. Wolstencroft K, Owen S, Krebs O, Nguyen Q, Stanford NJ, Golebiewski M, Weide-
533 mann A, Bittkowski M, An L, Shockley D, *et al.*: **SEEK: a systems biology data**
534 **and model management platform.** *BMC Syst Biol* 2015, **9**:1–12.
- 535 81. Aebersold R, Agar JN, Amster IJ, Baker MS, Bertozzi CR, Boja ES, Costello CE,
536 Cravatt BF, Fenselau C, Garcia BA, *et al.*: **How many human proteoforms are**
537 **there?** *Nat Chem Biol* 2018, **14**:206–214.
- 538 82. Lang PF, Chebaro Y, Zheng X, P Sekar JA, Shaikh B, Natale DA, Karr JR: **BpForms**
539 **and BcForms: a toolkit for concretely describing non-canonical polymers and**
540 **complexes to facilitate global biochemical networks.** *Genome Biol* 2020, **21**:117.
- 541 83. Schoch CL, Ciufu S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D,
542 Mcveigh R, O’Neill K, Robbertse B, *et al.*: **NCBI Taxonomy: a comprehensive**
543 **update on curation, resources and tools** *Database* 2020.
- 544 84. Sarntivijai S, Lin Y, Xiang Z, Meehan TF, Diehl AD, Vempati UD, Schürer SC, Pang C,
545 Malone J, Parkinson H, *et al.*: **CLO: the Cell Line Pntology** *J Biomed Semant* 2014,
546 **5**:1–10.

- 547 85. Dunnen JT den, Dagleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-
548 Jordan J, Roux AF, Smith T, Antonarakis SE, Taschner PE, *et al.*: **HGVS recom-**
549 **mendations for the description of sequence variants: 2016 update** *Hum Mutat*
550 2016, **37**:564–569.
- 551 86. Oberhardt MA, Zarecki R, Gronow S, Lang E, Klenk HP, Gophna U, Ruppin E: **Har-**
552 **nessing the landscape of microbial culture media to predict new organism–**
553 **media pairings** *Nat Commun* 2015, **6**:1–14.
- 554 87. Richards MA, Cassen V, Heavner BD, Ajami NE, Herrmann A, Simeonidis E, Price ND:
555 **MediaDB: a database of microbial growth conditions in defined media** *PLoS*
556 *One* 2014, **9**:e103548.
- 557 88. Sansone SA, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL,
558 Thurston M: **FAIRsharing as a community approach to standards, reposito-**
559 **ries and policies** *Nat Biotechnol* 2019, **37**:358–367.
- 560 89. Sundararaj S, Guo A, Habibi-Nazhad B, Rouani M, Stothard P, Ellison M, Wishart DS:
561 **The CyberCell Database (CCDB): a comprehensive, self-updating, relational**
562 **database to coordinate and facilitate in silico modeling of Escherichia coli.**
563 *Nucleic Acids Res* 2004, **32**:D293–D295.
- 564 *The CCDB (<http://ccdb.wishartlab.com>) is a pioneering database that was developed
565 to facilitate models of *E. coli*. By centralizing information about the structure and abun-
566 dance of metabolites, RNAs, and proteins, the CCDB enables modelers to focus on
567 creating models rather than on aggregating data.
- 568 90. Keseler IM, Mackie A, Santos-Zavaleta A, Billington R, Bonavides-Martínez C, Caspi R,
569 Fulcher C, Gama-Castro S, Kothari A, Krummenacker M, *et al.*: **The EcoCyc**
570 **database: reflecting new knowledge about Escherichia coli K-12.** *Nucleic Acids*
571 *Res* 2017, **45**:D543–D550.
- 572 **EcoCyc (<http://ecocyc.org>) and the broader BioCyc (<http://biocyc.org>) collection of
573 pathway-genome databases are some of the most comprehensive and highest quality re-
574 sources for qualitative and relational information for whole-cell modeling. For example,
575 EcoCyc has been a key source of data for models of the metabolism of *E. coli*. The Path-
576 way Tools software used to build EcoCyc and BioCyc could also be useful for organizing
577 data for specific models.
- 578 91. Crasto CJ, Marenco LN, Liu N, Morse TM, Cheung KH, Lai PC, Bahl G, Masiar P,
579 Lam HY, Lim E, *et al.*: **SenseLab: new developments in disseminating neuro-**
580 **science information.** *Brief Bioinformatics* 2007, **8**:150–162.
- 581 *CellPropDB and NeuronDB (<https://senselab.med.yale.edu>) are pioneering databases
582 that were developed to facilitate models of neurons. By providing data about the ex-
583 pression of membrane channels, receptors, and neurotransmitters, the databases enable
584 modelers to focus on building better models.

- 585 92. Latendresse M, Krummenacker M, Trupp M, Karp PD: **Construction and completion**
586 **of flux balance models from pathway databases.** *Bioinformatics* 2012, **28**:388–396.
- 587 93. Mondeel TD, Crémazy F, Barberis M: **GEMMER: GEnome-wide tool for Multi-**
588 **scale Modeling data Extraction and Representation for Saccharomyces cere-**
589 **visiae.** *Bioinformatics* 2018, **34**:2147–2149.
- 590 94. Perez-Riverol Y, Bai M, Veiga Leprevost F da, Squizzato S, Park YM, Haug K, Car-
591 roll AJ, Spalding D, Paschall J, Wang M, *et al.*: **Discovering and linking public**
592 **omics data sets using the Omics Discovery Index.** *Nat Biotechnol* 2017, **35**:406–
593 409.
- 594 ****OmicsDI (<https://www.omicsdi.org>) is one of the most comprehensive search engines**
595 **for quantitative omics data. OmicsDI encompasses data for a wide range of organisms**
596 **and cell types. OmicsDI’s distributed approach to data aggregation both enables many**
597 **investigators to contribute to OmicsDI and enables experts to quality control each type**
598 **of data contained in the database.**
- 599 95. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur Ö, Anwar N, Schultz N,
600 Bader GD, Sander C: **Pathway Commons, a web resource for biological pathway**
601 **data.** *Nucleic Acids Res* 2010, **39**:D685–D690.
- 602 96. Roth YD, Lian Z, Pochiraju S, Shaikh B, Karr JR: **Datanator: an integrated**
603 **database of molecular data for quantitatively modeling cellular behavior.**
604 *Nucleic Acids Res* 2021, **49**:D516–D522.
- 605 ****Datanator (<https://datanator.info>) is an integrated database of several key types of**
606 **data for modeling cells. To help investigators best leverage the limited data available**
607 **for modeling, Datanator provides tools for assembling clouds of measurements centered**
608 **around specific molecules and molecular interactions in a specific organism. As a data**
609 **warehouse, Datanator also provides this data in a consistent format.**
- 610 97. Percha B, Garten Y, Altman RB: **Discovery and explanation of drug-drug interactions via**
611 **text miningin: BiocomputingWorld Scientific, 2012pp. 410–421.**
- 612 98. Bird SNLTK: the Natural Language Toolkitin: **Proceedings of the COLING/ACL 2006**
613 **Interactive Presentation Sessions2006pp. 69–72.**