SYNTHETIC BIOLOGY

*Synthetic Biology*, 2024, **9(1)**, ysae010

DOI: https://doi.org/10.1093/synbio/ysae010
Advance Access Publication Date: 21 June 2024
**Perspectives**

# Data hazards in synthetic biology

Natalie R. Zelenka[1,2,*], Nina Di Cara[3], Kieren Sharma[4], Seeralan Sarvaharman[5], Jasdeep S. Ghataora[2,5], Fabio Parmeggiani[2,6,7,†], Jeff Nivala[8,†], Zahraa S. Abdallah[4,†], Lucia Marucci[2,4,†], and Thomas E. Gorochowski[2,5,†,*]

[1]Jean Golding Institute, University of Bristol, Bristol, UK
[2]BrisEngBio, University of Bristol, Bristol, UK
[3]School of Psychological Science, University of Bristol, Bristol, UK
[4]School of Engineering Mathematics and Technology, University of Bristol, Bristol, UK
[5]School of Biological Sciences, University of Bristol, Bristol, UK
[6]School of Biochemistry, University of Bristol, Bristol, UK
[7]School of Pharmacy and Pharmaceutical Sciences, Cardiff University, Cardiff, UK
[8]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA
[†]These authors contributed equally.
*Corresponding authors: E-mails: natalie.zelenka@bristol.ac.uk and thomas.gorochowski@bristol.ac.uk

## Abstract

Data science is playing an increasingly important role in the design and analysis of engineered biology. This has been fueled by the development of high-throughput methods like massively parallel reporter assays, data-rich microscopy techniques, computational protein structure prediction and design, and the development of whole-cell models able to generate huge volumes of data. Although the ability to apply data-centric analyses in these contexts is appealing and increasingly simple to do, it comes with potential risks. For example, how might biases in the underlying data affect the validity of a result and what might the environmental impact of large-scale data analyses be? Here, we present a community-developed framework for assessing data hazards to help address these concerns and demonstrate its application to two synthetic biology case studies. We show the diversity of considerations that arise in common types of bioengineering projects and provide some guidelines and mitigating steps. Understanding potential issues and dangers when working with data and proactively addressing them will be essential for ensuring the appropriate use of emerging data-intensive AI methods and help increase the trustworthiness of their applications in synthetic biology.

**Key words:** data hazards; data science; AI; synthetic biology; ethics

## 1. Introduction

Synthetic biology has seen a rapid expansion in the use of data-centric approaches for biological design over the past decade (1–3). By employing methods like deep learning trained on the vast biological datasets that are now becoming available (4–8), researchers can predict the behavior of complex biological systems and design new biological parts and circuits with unprecedented precision and control (9–12). While these advances have the potential to revolutionize various aspects of biotechnology, they also present a number of challenges and potential risks that require careful consideration.

One of the primary challenges in this area is the quality and reliability of the data used to build and validate the models (13). The accuracy and utility of data-centric models depend heavily on the underlying data that are used to build them. Data can be prone to errors, biases, and inconsistencies (14–16). As a result, models based on flawed or incomplete data can lead to unexpected results, such as the creation of a genetic circuit or synthetic organism with unpredictable behavior, or the inference of erroneous biological insights that hamper progress in fundamental and applied research.

The increasing complexity of data-centric approaches in synthetic biology also raises concerns about their interpretability and transparency (17). As models become more intricate and incorporate larger datasets [e.g. large neural networks (4, 9, 12) or whole cell models (18)], it becomes increasingly difficult for researchers to understand the underlying mechanisms driving their predictions. This lack of transparency hinders efforts to validate and improve these models, which is essential for ensuring their safe and responsible application.

The potential misuse of data-centric approaches in synthetic biology poses a further significant risk. The ease of access to data science tools may enable nefarious actors to develop harmful biological agents for purposes such as bioterrorism or to disrupt ecological systems intentionally. In addition, the rapid dissemination of synthetic biology techniques and knowledge, combined with a culture that fosters collaboration and innovation, could also increase the risk of an accidental (or willing) release of biological agents with unforeseen (or underestimated) consequences. Many of the models themselves also pose a significant environmental impact that is often unseen, with vast amounts of computing resources and electricity required to generate predictions or train models (19).

**Submitted**: 15 March 2024; **Received (in revised form)**: 17 May 2024; **Accepted**: 19 June 2024

More broadly, the increased use of data-centric approaches across all science and technology has also led to many ethical oversights and mistakes (20–22). These have often appeared avoidable retrospectively, with the general public and researchers from other disciplines raising alarms before the tools in question were deployed (23). However, data science and AI practitioners, who have the power to make decisions to improve the positive impact of their research, continue to find it difficult to engage with ethics work (24–26). In many cases they are disincentivized to do so, they often are not supported or trained appropriately, and many feel that ethics frameworks are either vague, unstructured, and difficult to apply, or worse still, just box ticking exercises that outsource the ethical judgment to committees who don't always understand what their research can do. Some initiatives attempt to overcome these issues. For example, the "AI Blindspot" project (https://aiblindspot.media.mit.edu) aims to proactively uncover potential oversights as an AI project is developed, highlighting potentially harmful unintended consequences. While hugely valuable for improving the safety of AI research, existing frameworks like this are typically focused purely on impacts that would directly affect humans. The potential of AI systems to harm the environment and wider ecosystems is often neglected, but of paramount concern when dealing with AI applied to engineered biology.

The use of data-centric approaches in synthetic biology offers exciting prospects for advancing our ability to engineer biological systems. However, it is crucial to proactively acknowledge and address the challenges, risks, and ethical considerations associated with these new methods. In this work, we present a community developed assessment framework called "Data Hazards" that aims to address some of these difficulties by supporting the more thorough consideration of potential data-related hazards that might exist as a project develops. While the framework is field agnostic, here we develop several extensions specific for synthetic biology applications, present two case studies to illustrate how the framework might be applied to protein design and whole-cell modeling tasks, and end by discussing potential mitigation strategies for issues that could arise. This work contributes to the ongoing conversations about responsible innovation in synthetic biology (27, 28) and the challenges that applications of data science bring to the field.

## 2. Materials and methods

### 2.1 Data hazards resources

The Data Hazard labels (Figure 1a) are generally applied to projects through workshops or self-assessment and, following this, the label-specific safety precautions and cross-label hazard mitiga-

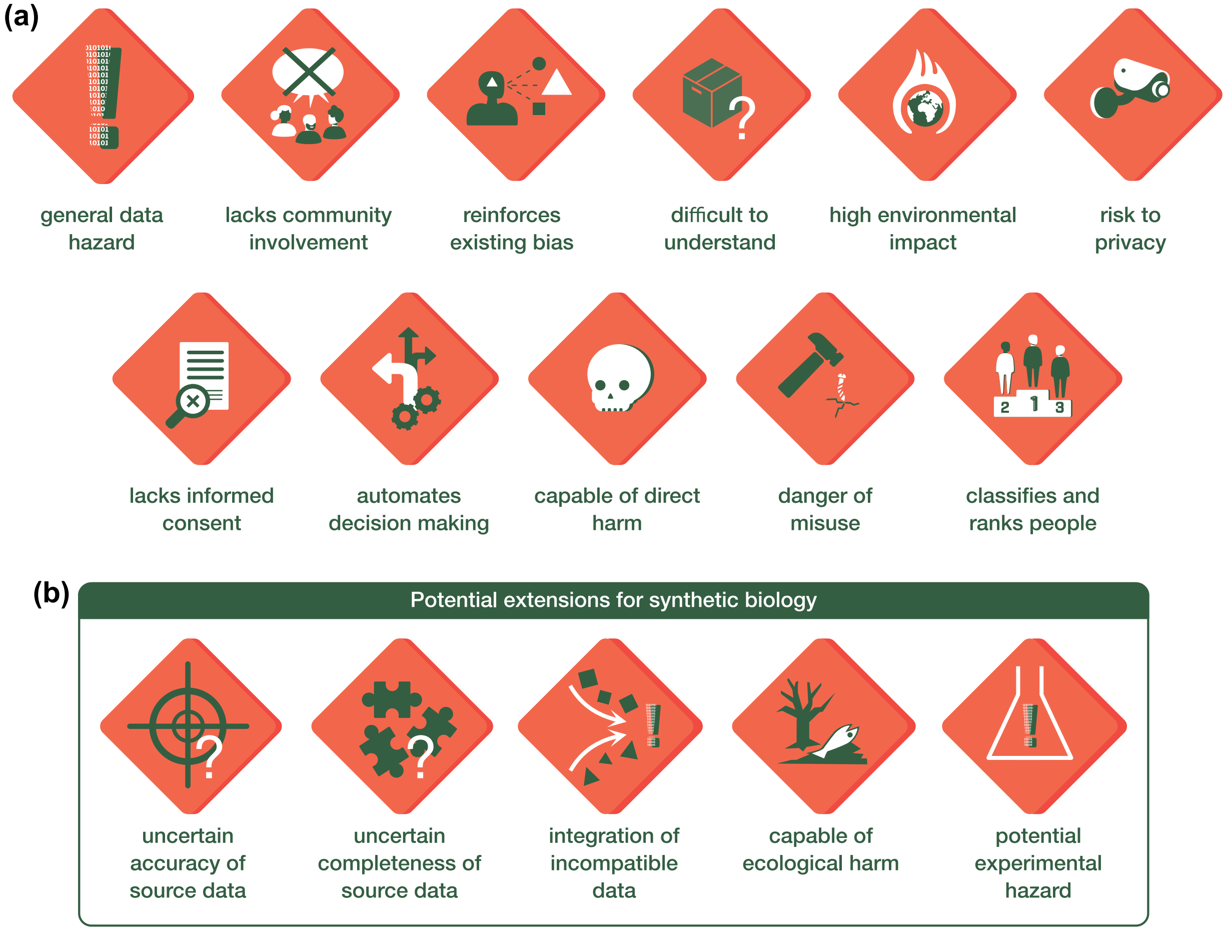
(a)
general data hazard
lacks community involvement
reinforces existing bias
difficult to understand
high environmental impact
risk to privacy
lacks informed consent
automates decision making
capable of direct harm
danger of misuse
classifies and ranks people
2
1
3
(b)
Potential extensions for synthetic biology
uncertain accuracy of source data
uncertain completeness of source data
integration of incompatible data
capable of ecological harm
potential experimental hazard

**Figure 1.** Overview of the data hazard labels. (a) Standard set of data hazard labels. Each label has been designed to clearly capture a core area of potential concern. One or many labels may apply to a piece of data science research. (b) Potential extensions to data hazard labels that address challenges common in synthetic biology research.

tion resources are used to identify potential interventions to mitigate risks. The Data Hazards website provides a user-guide for these common uses of the labels. This includes teaching materials (e.g. lesson plans and printable Hazard labels), workshops (e.g. checklists, timings, e-mail and feedback templates, slides, facilitating tips), a self-assessment guide, hazard mitigation resources, guides to displaying the labels, and finally case studies from users. All data hazard labels are available as Supplementary Information.

# 3. Results

## 3.1 A community framework for assessing data hazards

Data Hazards is an open-source, Creative Commons Attribution (CC-BY) licensed resource that aims to support data science and AI practitioners in identifying the broad risks associated with their work such as environmental concerns, misuse and algorithmic bias, and allows the to consider practical processes and activities to mitigate these (29). The resource is centered around a community-generated and evolving vocabulary of ethical risks, presented as "hazard labels" that are inspired by chemical warning signs (Figure 1a). Each label consists of an image, name, description, examples of where it applies, and safety precautions (Table 1). These act to facilitate interdisciplinary conversations and individual reflection and it is expected that they will lead to mitigating actions to address any issues raised. Without such safeguards, these labels could cause more harm than good, acting as "attention hazards" in their own right. Hazard labels associated with a project and mitigating actions can be displayed in posters, theses, ethical considerations sections of conference presentations and papers, or simply used as part of an internal process to identify necessary safeguards to improve the quality of research outputs. The hazard labels are relevant to any project that uses data, statistics, algorithms, machine learning, or AI, and have been applied to diverse projects spanning natural language processing in social media (30, 31), molecular modeling of molecules in neurons (32), and the integration of medical data sets (29, 33). Furthermore, by extending an existing safety framework that experimental scientists are already familiar with, and which covers both human and environmental impacts, we believe the barrier to adoption is lowered. When designing experiments in a laboratory, chemical hazards are assessed, and safeguards put in place; something that we feel should extend to the AI tools developed to support such research.

The project is managed through a website (https://datahazards.com), which houses the most up-to-date information regarding the aims and origin of the project, how data hazards can be best used, hints and tips for running workshops or self-assessing your own projects, options for contributing to the project, upcoming events, and examples of data hazard label use. The entire website and the associated resources are all stored in a public GitHub repository to allow for versioning control of all elements as the project develops.

## 3.2 Data hazards specific to synthetic biology research

The ability for data science to be applied to biological design means that all existing data hazards could potentially be applicable to synthetic biology research. But does synthetic biology bring further hazards to the table? We assessed some of the core challenges faced when using data-centric approaches to engineer biology and found five new data hazards that, while being relevant to synthetic biology, also touched upon key aspects of biological data more broadly.

Two such hazards relate to the nature of data typically collected from biological systems. Firstly, available biological datasets often have high levels of uncertainty associated with their measurements and may also be incomplete, providing only a limited picture of the underlying system. Both of these difficulties stem from biological processes being challenging to measure due to their complexity and dynamic nature, as well as an inability to observe these processes directly, meaning that proxies are commonly used (e.g. fluorescent reporter proteins used to measure gene expression). These factors result in inaccurate or incomplete datasets, which may have significant consequences when applying data science methods without an understanding of these limitations.

The interdisciplinary nature of synthetic biology can also lead to risks, as data of different types and from different sources may need to be integrated as part of data science pipelines. Furthermore, reproducibility of results across the life sciences remains a major challenge, and while there are efforts to improve the situation through the use of calibrants (34, 35) and minimal information standards (36, 37), large variations in measurements of even identical biological processes between different labs means that data scientists need to be keenly aware of possible incompatibilities in the data they use (e.g. measurements in different units). Perhaps even more difficult to catch are genetic differences in supposedly identical cell lines (38, 39), batch-to-batch variation in reagents (e.g. chemicals and media) (40), or the unintended variation in environmental factors when repeating experiments performed by other labs. Such information is often not captured during experiments and places questions over the quality and validity of the data produced and can potenitally impact downstream uses (e.g. for parameter fitting during modeling).

Finally, while the existing data hazard "capable of direct harm" (Figure 1a) captures impacts on other human-beings, synthetic biology opens up the potential for harm to be caused to other organisms and ecosystems more broadly [e.g. gene drives (41)], as well as the opportunity for experimental hazards that arise in the laboratory, but stem from data informed decisions with unexpected consequences (e.g. the accidental design of a new-to-nature enzyme that catalyzes an unknown ecologically harmful reaction).

For each of these cases, we developed new hazard labels that aim to capture their core features and act as an extension to the current library and recommendations (Figure 1b; Table 2; Supplementary Material).

## 3.3 Case study 1: de novo protein design

To demonstrate how the data hazards framework applies to different areas of synthetic biology, we began by exploring the use of data-centric approaches for *de novo* protein design (Figure 2a). The ability to effectively design new proteins has tremendous potential for applications across numerous fields: from catalysis via novel or engineered enzymes to the sensing of molecules and synthesis of new materials. Due to the many degrees of freedom within protein chains and the complexity of the interactions involved, computational methods have been entrenched in the protein design field as early as the 1980s, first based on physicochemical principles (such as molecular dynamics) and requiring a high level of expertise to execute. The turn of the century saw the usage of optimization algorithms combining constraints and

**Table 1.** Descriptions of data hazards with synthetic biology examples

| Data hazard | Description | Synthetic biology examples | Potential safeguards |
|---|---|---|---|
| General data hazard | Data science is being used and leading to negative outcomes. This hazard applies to all data science research outputs. | All areas that make use of data science approaches. | Proactively explore potentially negative applications and implement mitigating actions. |
| Lacks community involvement | Technology is being produced without sufficient input from the community it is designed to serve. | Proprietary ML-based algorithms developed to support a synthetic biology based therapeutic with no Patient and Public Involvement and Engagement (PPIE). | Engage with community stakeholders through consultations and participatory design processes. |
| Reinforces existing bias | Reinforces unfair treatment of individuals and groups. This may be due to input data, algorithm or software design choices, or society at large. | Focus on data collection for a limited set of model organisms. May mean our understanding and models do not translate to biology at large and lead to poor decisions when engineering non-model species. | Apply algorithms to detect bias in datasets and model outputs, helping guide new data collection/generation to alleviate found biases. |
| Difficult to understand | Danger that the technology is difficult to understand. This could arise due to a lack of interpretability (e.g. neural nets), lack of documentation, or problems with implementation details that are difficult to spot. | Deep learning models of gene regulatory sequence and proteins. Large-scale models of cellular processes (e.g. whole-cell models, metabolic models, regulatory models) | Use standardized data formats (e.g. SBOL) and seek domain expertise to apply explainable AI approaches. |
| High environmental impact | Methodologies are energy-hungry, data-hungry (requiring increasing amounts of computation), or require special hardware that require rare materials and resources that are non-sustainable. | Large deep-learning-based models require huge amounts of compute for training and often significant compute for prediction, which typically has a hidden environmental impact. Similarly whole-cell models can take days to run and generate huge data sets that require significant storage. | Explore the use of surrogate modeling to reduce computational resources required, optimize code and hardware used. |
| Risk to privacy | Possible risk to the privacy of individuals whose data is processed. | Engineering of personalized medicine applications (e.g. CAR T cell engineering). | Anonymize data where possible. |
| Lacks informed consent | Datasets or algorithms use data which have not been provided with the explicit consent of the data owner/creator. These type of data often lack other contextual information, which can also make it difficult to understand potential biases. | Bioprospecting studies of large genomic data bases often make use of sequenced samples where consent of local people may not have been given. | Develop clear guidelines for obtaining informed consent and ensure transparency in data usage. |
| Automates decision-making | Automated decision-making can be hazardous in many different ways. Important to ask: whose decisions are being automated, what automation can bring to the process, and who benefits or is harmed by this automation? | Increasing use of automation and design of experiment approaches when screening libraries and performing complex laboratory tasks. Errors in data could result in poor decisions being automatically made. | Identify areas where decisions are being automated and adapt existing safety frameworks to increase testing/validation of design choices, prior to deployment. |
| Capable of direct harm | The application area of this technology means that it is capable of causing direct physical or psychological harm to someone even if used correctly. | Many areas of synthetic biology have dual-use (e.g. toxin production, synthetic viruses, etc.) | Assess level of harm and ensure sufficient containment is in place to avoid harm. |
| Danger of misuse | There is a danger of misusing the algorithm, technology, or data collected. | Synthetic biology often has dual-use and considering new-to-nature biological parts and systems can have difficult to predict unintended consequences (e.g. gene drives, toxin production, engineering of viruses). | Ensure thorough testing of models prior to release including the identification of potential “emergent abilities” in neural network-based generative models. |
| Classifies and ranks people | Ranking and classifications of people should be handled with care. We should ask what happens when the ranking/classification is inaccurate, when people disagree with how they are ranked/classified, as well as who it serves and how it could be gamed. | Less common in synthetic biology, but may become an issue if personalized medicine becomes established. | Seek engagement with society about how classifications might cause negative outcomes and aim to build broader agreement on how issues are best handled. |

**Table 2.** Descriptions of additional data hazards relevant for synthetic biology

| Data hazard | Description | Synthetic biology examples | Potential safeguards |
|---|---|---|---|
| Uncertain accuracy of source data | The accuracy of the underlying data is not known and so its use may lead to erroneous results or introduce bias. | Metabolic modeling where inaccurately labeled conversions (e.g. due to computational prediction) might lead to unexpected products being produced by engineered pathways. | Attempt to classify uncertainty if possible to better inform decisions and understand the range of possible outcomes. |
| Uncertain completeness of source data | Underlying data are of an uncertain completeness and have missing values that causes biased results. | Whole-cell models which attempt to use all the data available, but which may be limited. Protein design often builds on sequences on those proteins so far seen, which may bias design software. | Enrich data sets with missing data or attempt to correct for known biases. |
| Integration of incompatible data | Data of different types and/or sources are being used together that may not be compatible with each other. | Models that need to integrate information about many different processes in a cell. | Convert data to compatible format where possible of collect complementary data that is compatible. |
| Capable of ecological harm | This technology has the potential to cause broad ecological harm, even if used correctly. | Gene drives used to cause extinction events and *in situ* engineering of microbiomes. | Ensure sufficient physical containment to avoid unexpected release and barriers in place if deployed. |
| Potential experimental hazard | Translating technology into experimental practice can require safety precautions | Toxin production, virus-like particles, work with potentially pathogenic microbes. | Assess possible safety issues and put in place necessary safety measures. |

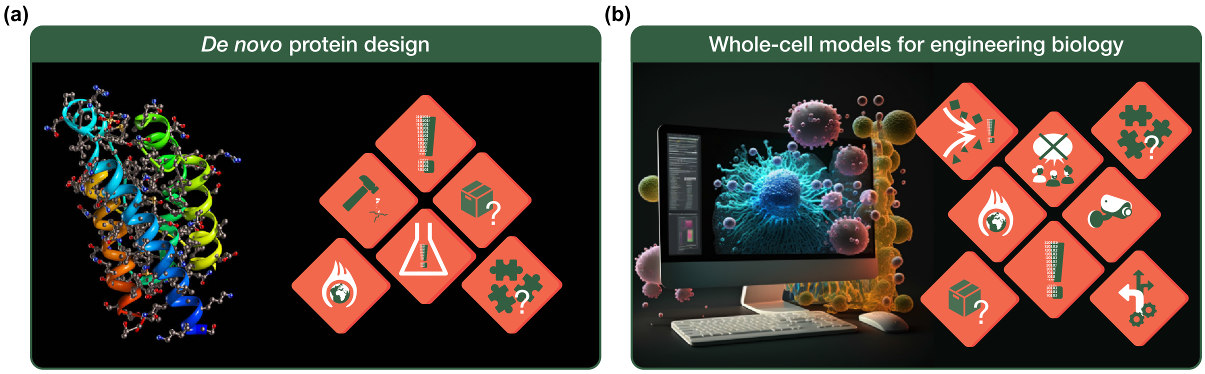

**Figure 2.** Data hazards identified for the synthetic biology case studies. (a) Data hazards associated with *de novo* protein design. Protein design is intrinsically affected by the data used to inform models and approaches, particularly from 'incompleteness of data': proteins that work are largely the ones selected by evolution, which are only a small fraction of what is possible. Traditional design methods rely on physical description of proteins as three-dimensional objects, but more recent data-intensive approaches operate on a much less intuitive level, i.e. often 'harder to understand for non-experts'. These methods can also involve millions of parameters, resulting in a 'high environmental energy cost' for training. As proteins are some of the most versatile molecules in biology and medicine, there is a great 'potential for misuse' of data and design outcomes, and, for certain designs, care should be taken to 'evaluate experimental risks' when moving from the digital world to the laboratory. (b) Data hazards associated with whole-cell models for engineering biology. Whole-cell models rely on 'integrating diverse and potentially incompatible data', as well as battling with 'gaps in the data' that is available. The scale of these models means that they often have a 'high environmental cost', both in terms of their computational execution and the storage of results. This scale and complexity also makes it 'impossible to fully understand' how they work, making verification of their predictions and results difficult; especially as many of the modeled processes cannot be directly observed. With human whole-cell models a long-term goal of the field, their future use could open up 'privacy concerns' whereby models are tailored using personal information and potentially then used to 'automate decisions' related to treatment of disease. Furthermore, whole-cell models have been so-far been developed by a relatively 'small community with little community input', while their use is potentially broad with wide impact across the entire field.

statistics-based terms. While these methods were more accessible than physics-based methods, they were often slow and didn't capture all the relevant features of proteins (42).

Recently, as with many other areas of synthetic biology, there has been a surge in the application of machine learning approaches. The need for faster and more accurate design protocols, and the large protein datasets now available because of high-throughput sequencing and structural biology techniques has provided the impetus for data-centric approaches in nearly all aspects of protein design. Neural networks, such as AlphaFold (12) and RoseTTAFold (43), and language models, such as OmegaFold (44) and ESMFold (45), initially developed for protein structure prediction, have led to the development of new protocols with increased speed and accuracy for protein–protein and protein–ligand interaction. The availability of such tools coupled with the accessibility of cloud-based computing resources is leading to the democratization of various aspects of protein design (46). While this democratization is overwhelmingly positive for the scientific community and society at large, we should be mindful of the potential risks that this ease of use brings.

One extreme example of potential misuse is the design of protein toxins, which prompted calls for regulation already in the early 2000s (47). Protein toxins that occur naturally are highly effective at interfering with cell activity. These toxins can alter the production or breakdown of molecules involved in metabolism, degrade enzymes, or cause cell lysis by creating unregulated pores (48). The difficulty of extracting toxin samples at a scale has prevented potential misuse. However, with the aforementioned methods, one can, in principle, not only design *de novo* toxins that could readily be synthesized at a scale, but also optimize them for greater affinity and specificity. For example, having access to how various protein complexes form in humans and different proteins interact in such complexes (49) could be used to identify candidates. One can then design a toxin that interferes with the natural life cycle of those candidates. Depending on the strategy, a combination of readily available deep learning toolkits can be used [e.g. quick design of protein degraders (50)]. It is precisely due to this wide availability of easy-to-use tools that we must carefully consider the potential for misuse and how it can be mitigated.

While different data-centric protein design projects may require bespoke mitigation strategies, it is also crucial to adopt general strategies (51). One of the most important is the awareness of the problem. On an individual level, this could mean adding a brief summary of the data hazards in the data availability section of manuscripts. For a more concerted effort, we should use any conference opportunities to discuss effective long-term solutions. This would be necessary to solve any implementation details of a practical strategy such as putting any data, including model parameters, behind access schemes. Such a strategy may require new technological infrastructure to be built and maintained to maximize the accessibility of good faith individuals whilst limiting any potential misuse.

### 3.4 Case study 2: whole-cell models for engineering biology

During the first decade of the 21st century, the total amount of genome sequence data being produced doubled approximately every 7 months (52). This trend has since continued, making genomics one of the largest domains within data science, with between 2 and 40 billion gigabytes of data expected to be produced annually by 2025 (52). Unlike other big data domains, which involve mostly centralized data, genome data acquisition is highly distributed across different countries, universities, and other research laboratories. This has resulted in large quantities of often heterogeneous data, which may not always be compatible.

Whole-cell computational models (WCMs) have emerged as state-of-the-art tools for integrating vast quantities of heterogeneous sequence data, generated by high-throughput measurement techniques, into a single knowledge base for a given organism (53). This unification process involves the curation of decades' worth of primary literature and experimental databases for determining parameter values, including protein half-lives, translation efficiencies and metabolic reaction constraints, to name just a few. The first model of this type was developed for *Mycoplasma genitalium* (18). More recently, work utilizing a more advanced WCM of *Escherichia coli*, which encompasses 19 119 parameters linked mechanistically by more than 10 000 interdependent mathematical equations (54). These parameters were extracted via a "deep curation" process that used over 400 publications spanning six decades and covering three lab strains of *E. coli*. The model comprises several sub-models, each focusing on distinct cellular processes, that are interconnected through shared resources and parameters, enabling a holistic representation of cellular dynamics under different environmental conditions. Simulating this model involves solving several types of mathematical equations simultaneously, such as ordinary differential equations, stochastic processes, and statistical models, producing over 200 000 time-series as output. The complexity of the model requires high-performance computing for execution, which may cause inequality of access, but may also beneficially act as a barrier against misuse. The intricacy of WCMs also complicates comprehension and interpretation of their output, which may discourage or limit community involvement. However, efforts to aid interpretability and visualization are being developed (55).

Within this complexity lies valuable predictive power that is being harnessed to design and conduct experiments *in silico*, accelerating scientific discovery (56–58). As WCMs become more complete and accurate, and genome engineering and synthesis become widely accessible, model-based genome design opens tremendous opportunities for the rapid engineering of biology for numerous types of application (59, 60). Utilizing WCMs in this way, however, has the potential to reinforce existing biases in the data used for their derivation, making experimental validation crucial.

The increased predictive power of WCMs compared to smaller-scale models, along with their ability to describe emergent behavior, facilitates more complex bioengineering tasks and could significantly accelerate synthetic biology design cycles. However, automated decision-making in this context may inadvertently introduce biases, potentially compromising the safety and efficacy of engineered biological systems. For example, the engineering of immune cells is currently being explored as a possible route toward novel cancer therapeutics (61). Whole-cell models of human cells (62, 63) could provide a foundation on which to build this technology by enabling efficient *in silico* assessment of different cellular reprogramming strategies. However, as highlighted in numerous medical case studies of automated decision-making using AI, the inherent biases present in medical datasets commonly used to train models or fit parameters are often not representative of the full range of demographics on which the therapeutic may be used (64). This bias could lead to the development of WCMs that aid the automated design of personalized therapies with a narrow operating window that have the potential to cause harm to subsets of the population. In addition to issues related to the application of WCMs, the high computational cost of running these large-scale simulations also has major environmental implications. For this reason, there are currently efforts to develop surrogate models which could help reduce computational burden and environmental impacts of such models (65).

Despite there only being a few WCMs created to date, the decreasing cost of genome sequencing, coupled with the exponential growth in computational power, is driving the development of WCMs for a variety of organisms. Recently developed kinetic WCMs developed for artificial cells aim to capture spatial features (e.g. they can account for cell geometry and ribosome distribution) (66). Such a variety of models may facilitate the design of pathogens or drug-resistant organisms for nefarious purposes. It is therefore essential that we adapt and extend existing synthetic biology safety frameworks to cater for the more predictive and capable design that WCMs support.

The Synthetic Biology Open Language (SBOL) serves as a prime example of a framework designed for standardizing the exchange of information related to biological designs (67). SBOL allows a user to more explicitly capture, not only structural data covering the DNA, RNA, proteins, and other chemical components within a design, but also information related to the functional interactions between these elements. The functional information is crucial for automating the development of models and in the context of WCMs, enables our knowledge of how biological processes are interwoven to be explicitly embedded. In the context of biosafety, capturing this information ensures that anyone assessing a design can more clearly see the potential for issues as less domain specific knowledge is hidden, allowing for more thorough testing of a model for potentially undesirable phenotypes. Furthermore, testing is assisted by simplified exchange of this information (68, 69). Beyond the underlying data, a sister standard called SBOL Visual (70) also offers a means to visualize biological designs in an explicit way, more clearly conveying information embedded within the SBOL data files (71). By improving the communication of intent in engineered biological designs, it is possible for potential issues to be more easily captured as less domain-specific knowledge is left with the designer. Extending both the SBOL data and visual standards to simplify the description of WCMs and the diverse processes they are required to include (e.g. making it easier to capture spatial elements and interactions between processes) would support many of these benefits.

Another example of an existing framework that could be adapted or expanded to enhance the safety of deploying WCM-designed organisms is the set of safety policies laid out by the International Genetically Engineered Machine (iGEM) Foundation (https://responsibility.igem.org/safety-policies/introduction). These focus on how teams should work during the iGEM competition (72, 73), but could provide broader guidelines that the entire synthetic biology community could choose to follow. In relation to WCMs, a new step could be introduced before the “release beyond containment” policy. This additional step would require that proper and rigorous analysis of *in-silico*-designed organisms had been performed to ensure their safety and functionality in controlled environments prior to any deployment into the physical world. Such screening is becoming common place for DNA synthesis, but has yet to be adapted more broadly for other areas of design in synthetic biology.

Constructing a whole-cell dynamical model for human cells is a central goal within systems biology (63). Developing and utilizing such models, however, raises privacy concerns, as they will likely necessitate the processing and storage of human genetic information, potentially exposing individuals to risks of unauthorized access, misuse, or discrimination. As an extreme example, such models could inadvertently publicize a given population’s genetic information, enabling someone to develop biological agents able to target specific genetic profiles. Transparency and interpretability within existing whole-cell modeling techniques, coupled with rigorous data privacy measures, will lay the foundations for safer and more reliable practices in the future by fostering a comprehensive understanding of the underlying assumptions, methodologies, and limitations of these models, as well as facilitating open and constructive scientific dialogue.

A summary of all of the hazards highlighted for using WCMs to engineer biology is shown in Figure 2b.

## 4. Discussion

Data science and AI have become increasingly popular in the field of synthetic biology as they enable new solutions to complex biological problems that would be difficult to solve otherwise. In this work, we have introduced the “Data Hazards” framework, which aims to broaden engagement in the responsible and ethical use of data science in the context of synthetic biology (Figure 1). Data Hazards is a relatively new initiative and as such is still evolving as it becomes established across different areas of science, engineering, and the humanities. Here, we identified five additional data hazards that are common in data-centric approaches to synthetic biology and used case studies covering *de novo* protein design and WCMs as a means to demonstrate how these hazards apply to emerging areas of biological engineering.

The case studies highlighted in this work (Figure 2) are only a few arbitrary examples and a much broader exercise would be needed to cover the full spectrum of potential synthetic biology research. To stimulate this process, we believe it would be valuable to consider how the use of data hazard labels could be integrated into existing scientific activities. For example, it becoming standard practice to display data hazard labels on posters at conferences or as part of graphical abstracts in papers with explanations for how these hazards have been mitigated. Such publicity would help to drive adoption and have the added benefit of establishing new ethical dialogs on research that are often lacking. Moreover, it could be beneficial to consider these hazards as research proposals are being developed to reduce the chance of misuse early on. Inclusion of a “Data Hazards Checklist” that must be completed as part of a grant application would highlight areas of concern before a project starts and ensure financial support is available to put safeguards in place or weed out research that should not be pursued due to issues that cannot be mitigated.

More broadly, the need to build a community within synthetic biology around data hazards and approaches to overcome data science risks is something that we believe could greatly benefit the field. Synthetic biology has historically been proactive about ethical considerations to ensure the benefits engineering biology offers are acceptable and understood by society and benefits and risks are discussed in a balanced way (74, 75). Considering the role of data science in these broader activities would be a valuable exercise moving forward. It is also important to note that while the new data hazard labels we have developed were done so with synthetic biology in mind, the often application agnostic use of data-centric methods means they may also be of relevance to other areas of science and engineering.

An interesting future direction for this work will be to explore how the integration of data hazards into synthetic biology design and implementation workflows can link to existing regulatory frameworks and initiatives. For example, there is growing activity in the area of sequence screening to ensure the synthesis of DNA with the potential for harm is avoided (76, 77) and the application of genomics surveillance is becoming more widely considered after the COVID-19 pandemic and rise of antimicrobial resistance (78). Biofoundries are also likely to play a key role in this area, due to their ability to generate the large data sets needed for data-centric biological design, and their central role in many projects as they scale beyond proof-of-concept studies in a research lab (79). This flexibility is essential as bioengineers battle with finding the most appropriate design methodology for the problem at hand (80).

Three of the data hazards we highlighted for synthetic biology that biofoundry capabilities could provide immediate mitigating actions include: uncertain accuracy and completeness of source data, as well as integration of incompatible data. Biofoundries require that experiments are explicitly described in a machine-readable format. This helps to support better reproducibility and improves overall accuracy of the data produced. Biofoundries are also ideally placed to implement the complex protocols often needed to provide more detailed and extensive measurements of an engineered biological system. The parallel application of highly quantitative sequencing, metabolomics and proteomics methods is necessary to gain a more complete picture of a biological system's inner workings. However, this is rarely done due to the costs involved, difficulties in processing the biological material, and the overall complexity of the various methods applied. Biofoundries could potentially alleviate this burden and provide validated workflows where missing data are avoided and measurements are taken in absolute (35, 81) or calibrated units (34, 82, 83) that can be easily integrated. Furthermore, the ability to run experiments in high throughput also enables better estimation of both technical and biological variability, helping to quantify uncertainty as part of the measurement process.

Incentivizing the use of such facilities and rigorous metrology remains a challenge; partly due to often limited access, but also because of the perceived additional effort they impose. These hurdles could be alleviated through automation within the biofoundries themselves, funding agencies pressing for facilities to support wider synthetic biology communities outside of their host institutions, and enforcing a requirement to meet minimal data collection standards for awarded grant funding. Together, these actions would not only improve the quality of research (i.e. reproducibility due to more explicit protocols that are run by machines), but would also provide a source of high-quality data able to support advanced modeling and for secondary use by wider research communities.

In summary, data science is sure to play a crucial role as synthetic biology develops. Using hazard labels can be a useful exercise for practitioners in this field as they help to proactively identify potential risks and stimulate discussions on ways to mitigate them. By encouraging open dialogue and promoting transparency, these labels can build trust between scientists, policymakers, and the public, ultimately leading to better-informed decisions on the use of data science and AI when engineering biological systems.

## Supplementary data

Supplementary Data are available at *SYNBIO* Online.

## Data availability

The "Data Hazards" project is a community-led initiative established in 2021 by Natalie Zelenka and Nina Di Cara. Full details about the project and all materials (e.g. Data Hazard labels, teaching materials, and a self-assessment tool) are available from https://datahazards.com under a CC-BY 4.0 license. High resolution images (PDFs) of all data hazard labels discussed in this work are also available in Supplementary Material.

## Funding

This research was supported by BrisEngBio, a UKRI Engineering Biology Transition Award grant BB/W013959/1 (N.R.Z, T.E.G. L.M.), two UKRI Breakthrough Award grants BB/W012448/1 (T.E.G. F.P., S.S.) and BB/W012235/1 (L.M.), a Royal Society University Research Fellowship grant URF\R\221008 (T.E.G), two EPSRC Fellowships grants EP/S01876X/1 (L.M.) and EP/S017542/1 (F.P.), an EPSRC DTP Scholarship (K.S.), and a Turing Fellowship from The Alan Turing Institute under EPSRC grant EP/N510129/1 (T.E.G.) The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Acknowledgments

N.R.Z., T.E.G., and L.M. conceived of the project. N.R.Z. ran all workshops to gather data for the case studies. T.E.G. developed the new hazard labels and all figures. All authors contributed to the writing and editing of the manuscript.

*Conflict of interest statement*. None declared.

## References

1. Freemont,P.S. and Bayley,H. (2019) Synthetic biology industry: data-driven design is creating new opportunities in biotechnology. *Emerg. Top. Life Sci.*, **3**, 651–657.
2. Beardall,W.A.V., Stan,G.-B. and Dunlop,M.J. (2022) Deep learning concepts and applications for synthetic biology. *GEN Biotechnol.*, **1**, 360–371.
3. Gilliot,P.-A. and Gorochowski,T.E. (2020) Sequencing enabling design and learning in synthetic biology. *Curr. Opin. Chem. Biol.*, **58**, 54–62.
4. de Boer,C.G., Vaishnav,E.D., Sadeh,R., Abeyta,E.L., Friedman,N. and Regev,A. (2020) Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.*, **38**, 56–65.
5. Tarnowski,M.J. and Gorochowski,T.E. (2022) Massively parallel characterization of engineered transcript isoforms using direct RNA sequencing. *Nat. Commun.*, **13**, 434.
6. Cambray,G., Guimaraes,J.C. and Arkin,A.P. (2018) Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in *Escherichia Coli*. *Nat. Biotechnol.*, **36**, 1005–1015.
7. Kuo,S.-T., Jahn,R.-L., Cheng,Y.-J., Chen,Y.-L., Lee,Y.-J., Hollfelder,F., Wen,J.-D. and Chou,H.-H.D. (2020) Global fitness landscapes of the Shine-Dalgarno sequence. *Genome Res.*, **30**, 711–723.
8. Kosuri,S., Goodman,D.B., Cambray,G., Mutalik,V.K., Gao,Y., Arkin,A.P., Endy,D. and Church,G.M. (2013) Composability of regulatory sequences controlling transcription and translation in *Escherichia Coli*. *Proc. Natl. Acad. Sci.*, **110**, 14024.
9. Dauparas,J., Anishchenko,I., Bennett,N., Bai,H., Ragotte,R.J., Milles,L.F., Wicky,B.I.M., Courbet,A., de Haas,R.J., Bethel,N. *et al.* (2022) Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, **378**, 49–56.
10. Kotopka,B.J. and Smolke,C.D. (2020) Model-driven generation of artificial yeast promoters. *Nat. Commun.*, **11**, 2113.
11. LaFleur,T.L., Hossain,A. and Salis,H.M. (2022) Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria. *Nat. Commun.*, **13**, 5159.
12. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
13. Bradbury,A. and Plückthun,A. (2015) Reproducibility: standardize antibodies used in research. *Nature*, **518**, 27–29.
14. Peterman,N. and Levine,E. (2016) Sort-Seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. *BMC Genomics*, **17**, 206.
15. Gilliot,P.-A. and Gorochowski,T.E. (2023) Design and analysis of massively parallel reporter assays using FORECAST. In: Selvarajoo K (ed). *Computational Biology and Machine Learning for Metabolic Engineering and Synthetic Biology*. Springer US, New York, NY, pp. 41–56.
16. Gilliot,P.-A., Gorochowski,T.E. and Martelli,P.L. (2023) Effective design and inference for cell sorting and sequencing based massively parallel reporter assays. *Bioinformatics*, **39**, btad277.
17. Gilpin,L.H., Bau,D., Yuan,B.Z., Bajwa,A., Specter,M. and Kagal,L. (2018) Explaining explanations: an overview of interpretability of machine learning. IEEE, pp 80–89.
18. Karr,J.R., Sanghvi,J.C., Macklin,D.N., Gutschow,M.V., Jacobs,J.M., Bolival,B. Jr, Assad-Garcia,N., Glass,J.I. and Covert,M.W. (2012)

A whole-cell computational model predicts phenotype from genotype. *Cell*, **150**, 389–401.

19. Dhar,P. (2020) The carbon impact of artificial intelligence. *Nat. Mach. Intell.*, **2**, 423–425.
20. Verhulst,S.G. (2021) Reimagining data responsibility: 10 new approaches toward a culture of trust in re-using data to address critical public needs. *Data Policy*, **3**, e6.
21. Reijers,H.A., Vanderfeesten,I., Plomp,M.G.A., Van Gorp,P., Fahland,D., van der Crommert,W.L.M. and Garcia,H.D.D. (2017) Evaluating data-centric process approaches: does the human factor factor in? *Softw. Syst. Model*, **16**, 649–662.
22. Vayena,E. and Blasimme,A. (2018) Health research with big data: time for systemic oversight. *J. Law Med. Ethics*, **46**, 119–129.
23. Agrawal,A., Gans,J.S. and Goldfarb,A. (2019) Artificial intelligence: the ambiguous labor market impact of automating prediction. *J. Econ. Perspect.*, **33**, 31–50.
24. Di Cara,N.H., Zelenka,N., Day,H., Bennet,E.D.S., Hanschke,V., Maggio,V., Michalec,O., Radclyffe,C., Shkunov,R., Tonkin,E. *et al.* (2022) Data ethics club: creating a collaborative space to discuss data ethics. *Patterns*, **3**, 100537.
25. Barocas,S. and Boyd,D. (2017) Engaging the ethics of data science in practice. *Commun. ACM*, **60**, 23–25.
26. Wilson,C. (2022) Public engagement and AI: a values analysis of national strategies. *Gov. Inf. Q.*, **39**, 101652.
27. Pansera,M., Owen,R., Meacham,D. and Kuh,V. (2020) Embedding responsible innovation within synthetic biology research and innovation: insights from a UK multi-disciplinary research centre. *J. Responsible Innov.*, **7**, 384–409.
28. Macnaghten,P., Owen,R., Jackson,R. and Pinheiro,V.B. (2016) Synthetic biology and the prospects for responsible innovation. *Essays Biochem.*, **60**, 347–355.
29. Zelenka,Z., Di Cara,N.H., Bennet,E., Hanschke,V., Kuwertz,E., Garcia,I.K. and Garcia,S.R. (2023) Data Hazards V1.0: an open-source vocabulary of ethical hazards for data-intensive projects. *OSF Prepr.*, **27**.
30. Maggio,V., Di Cara,N.H., Tanner,A., Haworth,C.M.A. and Davis,O.S.P. (2021) Understanding the potential and pitfalls of digital phenotypes to measure population mental health and wellbeing. *Public Health Sci.*, **398**, S10.
31. Di Cara,N., Zelenka,N., Davis,O. and Haworth,C. (2023) Using data hazards to support safe and ethical digital footprint research. *Int. J. Popul. Data Sci.*, **8**, 2279. 10.23889/ijpds.v8i3.2279
32. Garcia,S.R., Welsh,C., Di Cara,N., Sterratt,D., Romano,N. and Stefan,M. (2024) Data hazards as an ethical toolkit for neuroscience. OSF Preprints. 10.31219/osf.io/yn2j9.
33. Zelenka,N. (2021) *Phenotype and Function from Genotype: Combining Data Sources to Create Explanatory Predictions*. University of Bristol, Bristol, UK. https://nataliezelenka.github.io/phenotype_from_genotype/ (27 April 2024, date last accessed).
34. Beal,J., Haddock-Angelli,T., Baldwin,G., Gershater,M., Dwijayanti,A., Storch,M., de Mora,K., Lizarazo,M., Rettberg,R. and Olson,D., with the iGEM Interlab Study Contributors. (2018) Quantification of Bacterial Fluorescence Using Independent Calibrants. *PLoS One*, **13**, e0199432.
35. Gorochowski,T.E., Chelysheva,I., Eriksen,M., Nair,P., Pedersen,S. and Ignatova,Z. (2019) Absolute quantification of translational regulation and burden using combined sequencing approaches. *Mol. Syst. Biol.*, **15**, e8719.
36. Lee,J.A., Spidlen,J., Boyce,K., Cai,J., Crosbie,N., Dalphin,M., Furlong,J., Gasparetto,M., Goldberg,M., Goralczyk,E.M. *et al.* (2008) MIFlowCyt: the minimum information about a flow cytometry experiment. *Cytometry A*, **73A**, 926–930.
37. Taylor,C.F., Paton,N.W., Lilley,K.S., Binz,P.-A., Julian,R.K., Jones,A.R., Zhu,W., Apweiler,R., Aebersold,R., Deutsch,E.W. *et al.* (2007) The Minimum Information about a Proteomics Experiment (MIAPE). *Nat. Biotechnol.*, **25**, 887–893.
38. Ben-David,U., Siranosian,B., Ha,G., Tang,H., Oren,Y., Hinohara,K., Strathdee,C.A., Dempster,J., Lyons,N.J., Burns,R. *et al.* (2018) Genetic and transcriptional evolution alters cancer cell line drug response. *Nature*, **560**, 325–330.
39. Anon,J. (2015) Announcement: time to tackle cells' mistaken identity. *Nature*, **520**, 264.
40. Luo,Y., Pehrsson,M., Langholm,L., Karsdal,M., Bay-Jensen,A.-C. and Sun,S. (2023) Lot-to-lot variance in immunoassays—causes, consequences, and solutions. *Diagnostics*, **13**, 1835.
41. Bier,E. (2022) Gene drives gaining speed. *Nat. Rev. Genet.*, **23**, 5–22.
42. Korendovych,I.V. and DeGrado,W.F. (2020) De novo protein design, a retrospective. *Q. Rev. Biophys.*, **53**, e3.
43. Baek,M., DiMaio,F., Anishchenko,I., Dauparas,J., Ovchinnikov,S., Lee,G.R., Wang,J., Cong,Q., Kinch,L.N., Schaeffer,R.D. *et al.* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**, 871–876.
44. Wu,R., Ding,F., Wang,R., Shen,R., Zhang,X., Luo,S., Su,C., Wu,Z., Xie,Q., Berger,B. *et al.* (2022) High-resolution de novo structure prediction from primary sequence. *bioRxiv*, 2022.07.21.500999. 10.1101/2022.07.21.500999
45. Lin,Z., Akin,H., Rao,R., Hie,B., Zhu,Z., Lu,W., Dos Santos Costa,A., Fazel-Zarandi,M., Sercu,T., Candido,S. *et al.* (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, **379**, 1123–1130. 10.1126/science.ade2574
46. Mirdita,M., Schütze,K., Moriwaki,Y., Heo,L., Ovchinnikov,S. and Steinegger,M. (2022) ColabFold: making protein folding accessible to all. *Nat. Methods*, **19**, 679–682.
47. Tucker,J.B. and Hooper,C. (2006) Protein engineering: security implications. *EMBO Rep.*, **7**, S14–S17.
48. Fischer,W.B., Wang,Y.-T., Schindler,C. and Chen,C.-P. (2012) Chapter six - mechanism of function of viral channel proteins and implications for drug development. In: Jeon KW (ed). *International Review of Cell and Molecular Biology*, Vol. **294**. Cambridge, MA, USA: Academic Press, pp. 259–321.
49. Drew,K., Wallingford,J.B. and Marcotte,E.M. (2021) Hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. *Mol. Syst. Biol.*, **17**, e10016.
50. Palepu,K., Ponnapati,M., Bhat,S., Tysinger,E., Stan,T., Brixi,G., Koseki,S.R.T. and Chatterjee,P. (2022) Design of peptide-based protein degraders via contrastive deep learning. *bioRxiv*, 2022.05.23.493169. 10.1101/2022.05.23.493169
51. Ewen,C. (2024) Could AI-designed proteins be weaponized? Scientists lay out safety guidelines. *Nature*, **627**, 478.
52. Stephens,Z.D., Lee,S.Y., Faghri,F., Campbell,R.H., Zhai,C., Efron,M.J., Iyer,R., Schatz,M.C., Sinha,S. and Robinson,G.E. (2015) Big data: astronomical or genomical? *PLoS Biol.*, **13**, e1002195.
53. Carrera,J. and Covert,M.W. (2015) Why build whole-cell models? *Trends Cell Biol.*, **25**, 719–722.
54. Macklin,D.N., Ahn-Horst,T.A., Choi,H., Ruggero,N.A., Carrera,J., Mason,J.C., Sun,G., Agmon,E., DeFelice,M.M., Maayan,I. *et al.* (2020) Simultaneous cross-evaluation of heterogeneous *E. Coli* datasets via mechanistic simulation. *Science*, **369**, eaav3751.
55. Landon,S., Chalkley,O., Breese,G., Grierson,C. and Marucci,L. (2021) Understanding metabolic flux behaviour in whole-cell model output. *Front. Mol. Biosci.*, **8**, 732079.
56. Skalnik,C.J., Cheah,S.Y., Yang,M.Y., Wolff,M.B., Spangler,R.K., Talman,L., Morrison,J.H., Peirce,S.M., Agmon,E. and Covert,M.W.

(2023) Whole-cell modeling of *E. Coli* colonies enables quantification of single-cell heterogeneity in antibiotic responses. *PLoS Comput. Biol.*, **19**, e1011232.

57. Choi,H. and Covert,M.W. (2023) Whole-cell modeling of *E. Coli* confirms that *in vitro* tRNA aminoacylation measurements are insufficient to support cell growth and predicts a positive feedback mechanism regulating arginine biosynthesis. *Nucleic Acids Res.*, **51**, 5911–5930.
58. Rees-Garbutt,J., Chalkley,O., Landon,S., Purcell,O., Marucci,L. and Grierson,C. (2020) Designing minimal genomes using whole-cell models. *Nat. Commun.*, **11**, 836.
59. Marucci,L., Barberis,M., Karr,J., Ray,O., Race,P.R., de Souza Andrade,M., Grierson,C., Hoffmann,S.A., Landon,S., Rech,E. *et al.* (2020) Computer-aided whole-cell design: taking a holistic approach by integrating synthetic with systems biology. *Front. Bioeng. Biotechnol.*, **8**, 942.
60. Landon,S., Rees-Garbutt,J., Marucci,L., Grierson,C. and Szczelkun,M. (2019) Genome-driven cell engineering review: in vivo and in silico metabolic and genome engineering. *Essays Biochem.*, **63**, 267–284.
61. Labanieh,L. and Mackall,C.L. (2023) CAR immune cells: design principles, resistance and the next generation. *Nature*, **614**, 635–648.
62. Goldberg,A.P., Chew,Y.H. and Karr,J.R. (2016) Toward scalable whole-cell modeling of human cells. In: *Proceedings of the 2016 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*. SIGSIM-PADS'16. Association for Computing Machinery, New York, NY, USA, pp. 259–262.
63. Szigeti,B., Roth,Y.D., Sekar,J.A.P., Goldberg,A.P., Pochiraju,S.C. and Karr,J.R. (2018) A blueprint for human whole-cell modeling. *Future Syst. Biol. Genomics Epigenomics*, **7**, 8–15.
64. Khan,B., Fatima,H., Qureshi,A., Kumar,S., Hanan,A., Hussain,J. and Abdullah,S. (2023) Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. *Biomed. Mater. Devices*, **1**, 731–738.
65. Gherman,I.M., Abdallah,Z.S., Pang,W., Gorochowski,T.E., Grierson,C.S., Marucci,L. and Schulz,M.H. (2023) Bridging the gap between mechanistic biological models and machine learning surrogates. *PLoS Comput. Biol.*, **19**, e1010988.
66. Thornburg,Z.R., Bianchi,D.M., Brier,T.A., Gilbert,B.R., Earnest,T.M., Melo,M.C.R., Safronova,N., Sáenz,J.P., Cook,A.T., Wise,K.S. *et al.* (2022) Fundamental behaviors emerge from simulations of a living minimal cell. *Cell*, **185**, 345–360.28.
67. Buecherl,L., Mitchell,T., Scott-Brown,J., Vaidyanathan,P., Vidal,G., Baig,H., Bartley,B., Beal,J., Crowther,M., Fontanarrosa,P. *et al.* (2023) Synthetic Biology Open Language (SBOL) Version 3.1.0. *J. Integr. Bioinform.*, **20**, 20220058.
68. McLaughlin,J.A., Beal,J., Mısırlı,G., Grünberg,R., Bartley,B.A., Scott-Brown,J., Vaidyanathan,P., Fontanarrosa,P., Oberortner,E., Wipat,A. *et al.* (2020) The Synthetic Biology Open Language (SBOL) Version 3: simplified data exchange for bioengineering. *Front. Bioeng. Biotechnol.*, **8**, 1009.
69. Myers,C.J., Beal,J., Gorochowski,T.E., Kuwahara,H., Madsen,C., McLaughlin,J.A., Mısırlı,G., Nguyen,T., Oberortner,E., Samineni,M. *et al.* (2017) A standard-enabled workflow for synthetic biology. *Biochem. Soc. Trans.*, **45**, 793–803.
70. Baig,H., Fontanarossa,P., McLaughlin,J., Scott-Brown,J., Vaidyanathan,P., Gorochowski,T., Misirli,G., Beal,J. and Myers,C. (2021) Synthetic Biology Open Language Visual (SBOL Visual) Version 3.0. *J. Integr. Bioinform.*, **18**, 20210013.
71. Beal,J., Nguyen,T., Gorochowski,T.E., Goñi-Moreno,A., Scott-Brown,J., McLaughlin,J.A., Madsen,C., Aleritsch,B., Bartley,B., Bhakta,S. *et al.* (2019) Communicating structure and function in synthetic biology diagrams. *ACS Synth. Biol.*, **8**, 1818–1825.
72. Kelwick,R., Bowater,L., Yeoman,K.H., Bowater,R.P. and Fahnert,B. (2015) Promoting microbiology education through the iGEM synthetic biology competition. *FEMS Microbiol. Lett.*, **362**, fnv129.
73. Moon,H. (2022) iGEM 2021: a year in review. *Biodesign Res.*, **2022**, 9794609.
74. Ginsberg,A.D., Calvert,J., Schyfter Camacho,P., Elfick,A. and Endy,D. (2014) *Synthetic Aesthetics; Investigating Synthetic Biology's Designs on Nature*. MIT Press, Cambridge, MA.
75. Häyry,M. (2017) Synthetic biology and ethics: past, present, and future. *Camb. Q. Healthc. Ethics*, **26**, 186–205.
76. Hoffmann,S.A., Diggans,J., Densmore,D., Dai,J., Knight,T., Leproust,E., Boeke,J.D., Wheeler,N. and Cai,Y. (2023) Safety by design: biosafety and biosecurity in the age of synthetic genomics. *iScience*, **26**, 106165.
77. Millett,P., Alexanian,T., Brink,K.R., Carter,S.R., Diggans,J., Palmer,M.J., Ritterson,R., Sandbrink,J.B. and Wheeler,N.E. (2023) Beyond biosecurity by taxonomic lists: lessons, challenges, and opportunities. *Health Secur.*, **21**, 521–529.
78. NIHR Global Health Research Unit on Genomic Surveillance of AMR. (2020) Whole-genome sequencing as part of national and international surveillance programmes for antimicrobial resistance: a roadmap. *BMJ Glob. Health*, **5**, e002244.
79. Hillson,N., Caddick,M., Cai,Y., Carrasco,J.A., Chang,M.W., Curach,N.C., Bell,D.J., Le Feuvre,R., Friedman,D.C., Fu,X. *et al.* (2019) Building a global alliance of biofoundries. *Nat. Commun.*, **10**, 2040.
80. Castle,S.D., Stock,M. and Gorochowski,T.E. (2024) Engineering is evolution: a perspective on design processes to engineer biology. *Nat. Commun.*, **15**, 3640.
81. Csibra,E. and Stan,G.-B. (2022) Absolute protein quantification using fluorescence measurements with FPCountR. *Nat. Commun.*, **13**, 6600.
82. Castillo-Hair,S.M., Sexton,J.T., Landry,B.P., Olson,E.J., Igoshin,O.A. and Tabor,J.J. (2016) FlowCal: a user-friendly, open source software tool for automatically converting flow cytometry data from arbitrary to calibrated units. *ACS Synth. Biol.*, **5**, 774–780.
83. Fedorec,A.J.H., Robinson,C.M., Wen,K.Y. and Barnes,C.P. (2020) FlopR: an open source software package for calibration and normalization of plate reader and flow cytometry data. *ACS Synth. Biol.*, **9**, 2258–2266.

Synthetic Biology, 2024, 9(1), ysae010 , DOI: https://doi.org/10.1093/synbio/ysae010, Advance Access Publication Date: 21 June 2024Perspectives
**Submitted**: 15 March 2024; **Received (in revised form)**: 17 May 2024; **Accepted**: 19 June 2024