Review

From Observational to Actionable: Rethinking Omics in Biologics Production

Helen O. Masson[1], Karen Julie la Cour Karottki[2], Jasmine Tat[1,3], Hooman Hefzi[4,*] and Nathan E. Lewis[1,2,*]

[1] Dept of Bioengineering, UC San Diego
[2] Dept of Pediatrics, UC San Diego
[3] Amgen, Inc.
[4] Biogen, Inc.
* Correspondence: nlewisres@ucsd.edu (H.H.); hooman.hefzi@biogen.com (N.E.L.)

Abstract: As the era of omics continues to expand, omics-based experiments have become popular in industrial biotechnology. They promise deeper biological understanding, which may be leveraged to develop novel solutions to bioprocess optimization and cell line engineering strategies. Despite this expansion, naivety about what omic data offers and how to best handle such data can challenge the extraction of actionable value. However, the value of omic experiments in biotechnology research and development can be maximized with deliberate application of omic approaches and forethought about analysis techniques. Here we describe important considerations when designing and implementing omic-based experiments, and discuss how systems biology analysis strategies can enhance efforts to obtain actionable insights in biomanufacturing.

Keywords: biopharmaceutical protein production; cell line development; bioprocess optimization; omics; statistical inference; mechanistic models

---

Highlights:

- Omic experiments have become increasingly popular in biotechnology, however, project timelines and system complexity make extracting actionable insights challenging.
- Current applications of omic technologies within the CHO community provide predominantly observational data, however these can become more actionable through the use of careful experimental design and/or emerging systems biology methods.
- Mechanistic models for diverse cellular systems have recently been developed and statistical analysis techniques for large datasets are increasingly commonplace but have yet to be widely applied to CHO omic data.

1. Towards actionable omics

Biotechnology is an evolving research field that thrives off our ability to harness living organisms to develop invaluable products and technologies. This interdisciplinary manufacturing approach is employed across a range of industries including energy, material, food, agriculture, cosmetics, and pharmaceuticals[1]. The biopharmaceutical industry in particular has been successful in manufacturing life-saving recombinant biotherapeutics in mammalian cells for over 40 years. While significant advances have been made–with product titers increasing from ~50 mg/L to >10 g/L for monoclonal antibodies, it is perhaps surprising to see that the fundamental steps underpinning the cell line generation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 February 2023 doi:10.20944/preprints202302.0037.v1

and bioprocess development processes have remained nearly the same, perhaps differing only in scale (Box 1).

Chinese hamster ovary (CHO) cells–the primary host system used for the manufacturing of biologics–are faced with challenges concerning the production of complex large molecules that meet stringent product quality (PQ) requirements. To meet industrial demands various strategies have emerged to manipulate the physiological functions of cell factories at the gene level (genetic engineering) and modify the culture environment for optimal growth and production (bioprocess optimization). The rational design and optimization of such strategies require a functional understanding of the molecular components driving bioproduction. Omic technologies provide large-scale, systems-wide monitoring of a broad range of such molecular components.

The sequencing of CHO and Chinese hamster genomes, the dramatic decrease in NGS costs, the growing ease of generating diverse omic data, and the development of mammalian genome editing tools provide a valuable toolbox for a new era of rationally driven cell line generation and process development[2–5]. Indeed, decades of omics research have yielded examples demonstrating the value of these data; however, while some omics strategies are inherently actionable by providing targets directly linked to a phenotype of interest (e.g., CRISPR screens[6–10]), many omic technologies can prove observational (i.e., merely providing a momentary snapshot of the molecular composition of the cells) when analyzed using basic approaches. Thus, it can be challenging to deploy strategies to maximize actionable value from the experiments. By actionable we mean the ability (1) to identify a minimal set of targets driving the phenotype of interest, (2) to direct a clear downstream strategy for engineering or optimization towards the phenotype of interest, and (3) to deploy the strategies in a project-compatible timeframe.

2. Challenges to actionability in biopharmaceutical process development

The timeline and platform-driven nature of the biologics industry has shaped how omic tools have been applied, and it is important to assess how it can be adapted to accommodate actionable omics strategies. The long, resource-intensive nature of cell line generation nearly necessitates that genetic engineering efforts be applied to the host cell—rather than introduced into a producing clone—as genetic modifications require further verification of clonality, cell line stability, and process optimization. Consequently, process optimization often remains the intervention of choice in the biopharmaceutical industry.

While direct interventional studies such as media optimization design of experiments (DoEs) are effective, they do not answer the 'why' or 'how' of success and thus provide limited understanding for future work. In contrast, omic studies, where molecular profiles of a given clone in a given process are measured, provide information about the cellular state tied to a given phenotype. These omic based approaches can thus inform future cell line engineering or process optimization efforts, but only when applied with careful forethought.

The difficulty, regardless of approach, lies in the complexity of the system being engineered (Figure 1). The CHO genome is significantly larger than genomes of other industrial lines (e.g., *E. coli*, yeast, etc.) and most genes remain uncharacterized in the context of protein production. Furthermore, the genomic plasticity exploited to generate highly productive cell lines through high-throughput screening means that a new clone may not respond as expected to a given bioprocess.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 February 2023 doi:10.20944/preprints202302.0037.v1

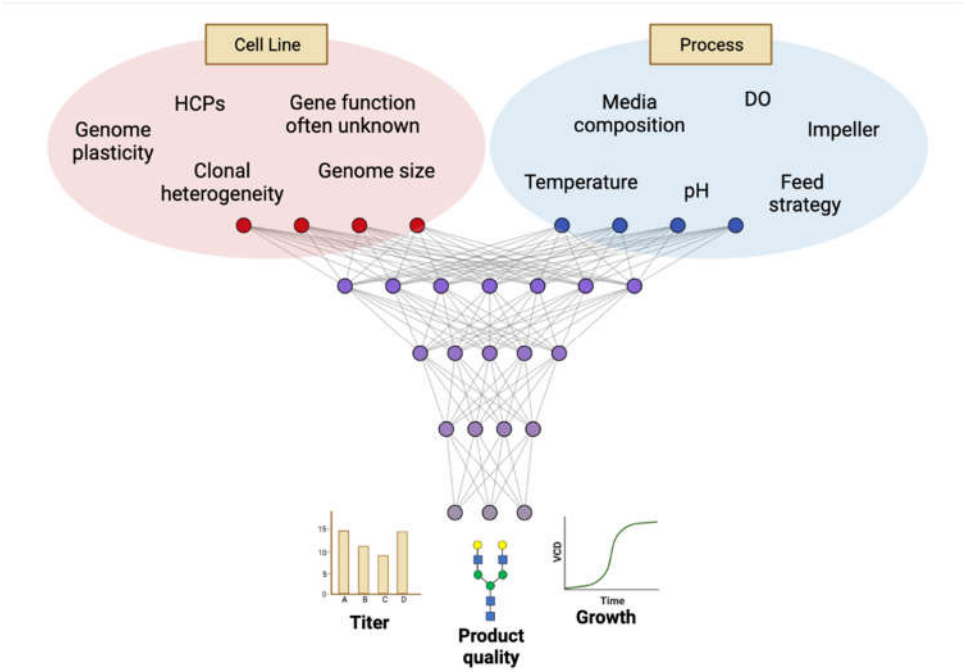

Figure 1. Complexity of biotherapeutic protein production. In biomanufacturing, there are multiple parameters and sources of complexity that must be managed and optimized for a given cell culture process for a given process development project (top). The end goal (bottom) for any process development effort is to generate large amounts of high quality product. Changes can be made to either the cell line or to the cell culture process, however these changes ultimately have to be evaluated based on their effect on titer, product quality, and growth. While it is tempting to view each point of implementation as independent, the effect of modifying either cell line or process cannot be understood in a vacuum–their interconnectedness represents the ultimate complexity in trying to assess, understand, and engineer this system. Since the responsiveness of a clone to a given process will vary depending on the underlying genotype, and the optimal process may vary considerably from clone to clone, it is critical to consider this complexity when designing and interpreting omic experiments. [HCPs: host cell proteins, DO: dissolved oxygen].

Given these complexities, extracting actionable insights from classical observational omic experiments (e.g., high feature number, low sample number, high background variability–present in an experiment looking at transcriptomic data from a high producer vs low producer clone) is challenging. When comparisons yield 100s to 1000s of potential targets (e.g., differentially expressed genes), it becomes difficult to decipher which gene(s) are actually driving the variation in phenotype, since many transcriptional differences may be superfluous for the phenotype of interest. Fortunately, depending on the desired objective, capabilities, and timelines, a range of different approaches may be suitable to extract actionable insights from omic experiments. Here we highlight types of analysis tools and strategies that can extract actionable insights from observational omics. Specifically, the integration of multi-omic approaches with systems biology computational models can help drive actionability.

3. Extracting actionable value from omics

To maximize actionability from omics, we recommend designing and executing stages of omic studies under the framework and considerations as described in Figure 2 (Key Figure). Prior to any data generation and analysis, one should define a clear implementation strategy by connecting analysis approaches to intervention objectives under known technical and logistical constraints. We emphasize that the implementation strategy should influence not only the selection of omic tools, but also the analysis methods employed and the prioritization of enriched targets. To guide systematic analysis of otherwise observational omic data, we describe, as follows, 4 approaches to extract actionable information from omic experiments, ranging from rational experimental design to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 February 2023 doi:10.20944/preprints202302.0037.v1

computational models. Depending on timelines, desired objectives, budgets, and tool availability, some approaches may be more desirable or could be infeasible. We provide explanations and examples of successful applications for each approach below.

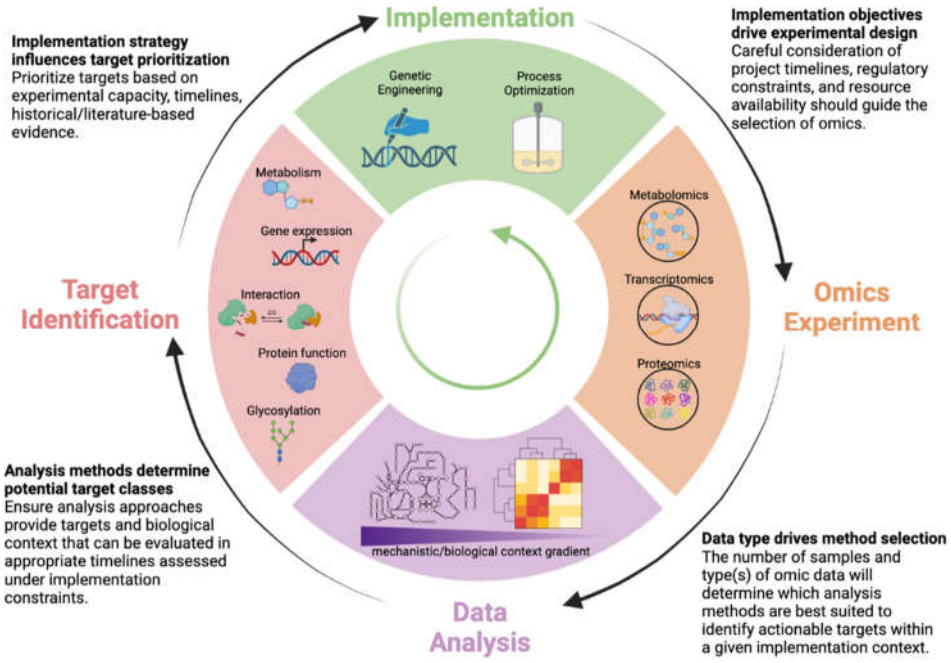

Figure 2. (Key Figure). Industrial application of actionable omics and key considerations. To design and implement effective cell line and bioprocess based omic studies, there are multiple considerations to maximize actionability. The outer arrow sequence represents the general lifecycle of omic study conceptualization, experimentation, and implementation; iterations of this process may be required to narrow down and validate biological targets. In addition to these considerations, we note through the inner green arrow that the implementation objectives guide all stages in the lifecycle and that stage interdependency can constrain the methods toolbox and expected outcomes. For example, the implementation strategy can inform both the expected target species to be intervened through cell line engineering or process optimization–which informs the omic type(s)--but also the respective analysis methods that can account for such biological scales or systems dynamics.

Rational experimental design

When the research question of interest is clearly defined and the root cause can be narrowed down to a reasonable set of molecular entities (e.g., dozens or a few hundred genes), it can be possible to derive actionable conclusions from small, rationally designed studies. As an example, a targeted proteomic study on high and low aggregation cell lines identified molecular mechanisms underlying cell aggregation[11,12]. By looking solely at the 'surfaceome' (surface protein sub-proteome) as the molecular entity(s) underpinning aggregation, the authors found proteins with differential abundance between the two cell lines and were able to decrease aggregation by knockdown of a surface protein. Other studies quantified the CHO secretome using proteomics to identify host cell proteins for removal to improve product quality or increase cell performance[13,14]. However, it is critical to predefine and selectively study only the responsible entities (e.g.,, surface proteins or host cell proteins) when sample numbers are low. For example, in classic 'high producer vs low producer' studies, the lack of actionability stems from high numbers of differentially expressed targets without practical ways to reduce the target list to a manageable number of testable hypotheses, nor define which genes are directly connected to productivity, as opposed irrelevant genes that happen to have altered expression for other reasons (e.g., sharing a transcription factor). Furthermore, when analyzing low numbers

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 February 2023 doi:10.20944/preprints202302.0037.v1

of clones, it is quite likely that differences between clones stem from clonal variability rather than being causally linked to the phenotype of interest.

Knowledge-based parametric models

Knowledge-based parametric models can link genotype to phenotype on a mechanistic level to elucidate biological causation from omic data[15–17]. These network models employ carefully curated biochemical, genetic, and genomic data into a knowledgebase of an organism's molecular components and their interactions[18]. With the integration of omic data, these models can guide the rational design of systems-level engineering targets for bioproduction. Here we describe two such classes of models with promising applications in the biopharmaceutical industry: (1) genome scale models and (2) kinetic models (Box 2).

Knowledgebases can help construct diverse biological networks, and success has been demonstrated with genome scale metabolic models (GeMs), thanks to decades of legacy biochemical research in metabolism[19]. The integration of omic data with GeMs to identify cell line engineering targets has been demonstrated. For example, a multi-omic and GeM flux analysis helped identify metabolic bottlenecks and potential engineering targets in mAb-producing CHO cells[20]. Other studies[21,22] successfully integrated omics with GeM simulations to design optimal bioprocessing conditions in mAb-producing CHO cells, resulting in an average titer increase of ~11.8% and approximate two-fold increase in total mAb expression respectively. Another study used omics data to obtain GeM simulations that guided media optimization to reduce growth inhibitory metabolite section[23]. These examples demonstrate how overlaying omics on a functional network contextualizes the data in terms of underlying biochemistry, and comparison of experimental data with model simulations can validate or refine hypotheses. Importantly, the systems-level metabolic response and flux simulations permitted by these networks go beyond any type of analysis possible with generic metabolic pathway databases such as KEGG[24].

Over the years, GeMs have expanded to include additional cellular processes. A desire to model the biochemical reactions underlying gene expression (transcription and translation) resulted in genome-scale models of metabolism and macromolecular expression (ME-models)[25,26]. Much like GeMs, ME-models can be platforms for the mechanistic integration of transcriptomic and proteomic data. In addition to expanding GeMs to include macromolecular expression networks, models can include core components of conventional protein secretion[27]. Many therapeutic proteins are clients of the secretory pathway, therefore a mechanistic understanding of pathway usage can provide novel insight into targeted engineering strategies. Furthermore, different biotherapeutics may utilize unique sets of secretory machinery exerting non-negligible metabolic demands on the host cell[13,27]. The resulting protein-specific models are able to calculate energetic costs and machinery demands for each secreted protein. Additional omic data can be integrated with the models and it was found that highly secretory cells have adapted to downregulate the expression and secretion of expensive native proteins[27]. Identification of costly native proteins that compete for cell resources present targeted engineering strategies for cell line engineering. Furthermore, the models were used to successfully optimize the production of monoclonal antibodies[27] and feeding strategies[28].

Complementary to the abovementioned constraint-based genome-scale stoichiometric models, kinetic models can effectively describe the dynamic character of mammalian cell culture and protein production[29,30]. These systems are non-stationary in nature, and depend on time and system history. Kinetic models can mechanistically model this dynamic behavior using mathematical expressions for the biochemical reaction rates of the system. While computationally intensive, kinetic models can be used to understand, predict, and evaluate the effects of targeted bioprocess manipulations and support the design of enhanced bioprocessing systems. Recent kinetic models of eukaryotes have been developed and can aid in cell line optimization[29,31]. Mammalian N-linked glycosylation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 February 2023 doi:10.20944/preprints202302.0037.v1

was integrated with a metabolic kinetic model [32] to rationally manipulate the glycoprofile of a secreted IgG in CHO[33]. Kinetic modeling was employed to predict cell culture performance and screen optimal temperature shift strategies[34] and predict the impact of select amino acids on cell growth, metabolism, and mAb production and optimization of fed-batch culture feeding in monoclonal antibody-producing CHO cells[34,35].

Data-driven inference models

Knowledge-based parametric models offer insightful contextualization to omic data, but they can be computationally intensive and require technical expertise. If these models are not accessible or not compatible with the omic data type(s), we suggest the use of data-driven inference models. Data-driven models for omic analyses are built upon statistical methods to interpret and extract useful information from high-dimensional data[36–38]. To increase specificity, some models have evolved to integrate biological assumptions and context. These methods are valuable when the studied problem contains many unknowns and/or the number of samples capturing the expected biological variability is present. We provide an overview of the following classes of data-driven models, ranging from biology-independent to biology-dependent methodologies, and how they are applied for omic data analysis: (1) unsupervised and supervised, (2) correlation networks, and (3) empirical-based, interactome inference models.

Unsupervised and supervised techniques are built upon statistical algorithms lacking mechanistic or biological considerations but can enrich meaningful and potentially accurate underlying patterns in omic data. Unsupervised techniques are predominantly useful in the validation of sample profiles, identification of subpopulations, detection of biological patterns, and integration of multi-omic data (Box 3). Principal component analysis (PCA) is commonly applied as quality control to validate sample similarities and to detect obvious technical factors such as batch effects[39–45], while k-means or hierarchical clustering can facilitate the identification of subject subpopulations or feature groups that exhibit similar behaviors[42,45,46]. Independent component analysis (ICA) and Markov clustering can identify biologically meaningful interactions between molecular species as demonstrated in *E.coli[47]* and human[48] omic data. Meanwhile, supervised techniques leverage high-dimensional data to predict continuous phenotypes like endpoint titer or cell density (regression; Box 3) or to predict whether a cell line will be high or low producing (classification; Box 3). End-point titer was predicted by PLS models trained with time-series metabolomic data[40] and process data[41]. Other studies[39,49,50] obtained strong correlations between key metabolite concentrations and performance attributes to identify targets for potential cell line screening and media optimization. Similar to clinical case studies predicting patient phenotypes[51,52], supervised learning across multiple omic levels—such as copy number, mutation data, transcriptomics, and metabolomics–can predict cell line phenotypes. While these statistical methods can deliver accurate predictions or global insight on groups of features, low interpretability of latent variables and the lack of biological context can hinder actionability when the studies aim to identify engineering targets or elucidate mechanisms causing the phenotype of interest.

Correlation networks combine unsupervised and supervised techniques with mechanistic considerations to infer association of network-based target groups with phenotype (Box 3). Weighted gene co-expression network analysis (WGCNA) can validate or discover co-expressed biological networks. For example, WGCNA has been used on scRNA-Seq data to resolve gene modules associated with an ER-stressed condition and to confirm biological pathways enriched in respective subpopulations[53]. Meanwhile, gene regulatory network (GRN) inference algorithms, such as SCENIC[54] and scGRNom[55] for single-cell RNA-seq and ChEA3[56] for bulk RNA-seq, assume association, or co-expression, of gene targets with transcription factors to predict gene network regulons and assess regulon enrichment. While correlation network techniques can reduce thousands of targets to dozens or hundreds of grouped targets, they also inevitably output

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 February 2023 doi:10.20944/preprints202302.0037.v1

false positive and indirect targets. These limitations, however, can be addressed with the integration of additional omic layers to correlate dynamics across hierarchical scales.

The integration of empirical and interaction context has produced network-based data-inference tools leaning on the biology-dependent end of the spectrum. Ingenuity Pathway Analysis (IPA)[57] and protein-protein Interaction (PPI) databases[58–60] are accessible tools in this realm (Box 3). IPA can be a hypothesis generation tool using pathway activity insight and as a targeted tool to identify testable regulator targets and validate downstream effects within pathways of interest. IPA was leveraged to identify and knockout two repressors to improve viral resistance in CHO cells[61]. Meanwhile, overlaying omic data on PPI networks—such as PCNet[59], STRING[58], or self-constructed networks[62]—enhances mechanistic understanding and target identification through contextualizing experimentally identified or computationally inferred biological interactions. Transcriptome data were supplemented with PPI networks to quantify secretory fitness variation across tissues and elucidate the role of perturbed secretory machinery in human amyloidogenesis[62]. Similar methods can provide mechanistic insight on secretory fitness in CHO and unravel key regulators.

Hybrid models

There are challenges that accompany mechanistic models. GeMs are exceptional at describing cellular mechanisms based on reaction stoichiometry and also cover all metabolic pathways; however, they can suffer from their steady-state assumption and miss cellular dynamics. On the other hand, kinetic models capture mechanistic details of dynamic cellular processes, but suffer from the computational burden associated with parameter estimation which limits full genomic coverage. Recent applications have combined machine learning and parametric mechanistic modeling frameworks to overcome some of these inherent challenges. Machine learning (ML) can help restructure and modify mechanistic models, or it can facilitate and tune model parameterization. For example, ML methods can estimate model parameters from experimental omic data[63–66]. Alternatively, flux solutions obtained from constraint-based models can provide an additional "omic layer" and be integrated into ML approaches to predict growth conditions[67], pathway engineering for optimized tryptophan production[68], and even drug side-effects[69]. Similar approaches are being applied towards creating digital twins of bioreactors[30,70,71] to enable *in silico* bioprocess optimization–aiming to decrease the number of iterations needed for process optimization. Recently a hybrid kinetic/artificial neural network (ANN) glycosylation model successfully predicted the glycoform distribution in monoclonal antibodies[72]. The hybrid model consists of two kinetic modules describing CHO cell metabolism and nucleotide sugar donor synthesis, feeding into a novel ANN model of glycosylation[73]. These types of synergistic approaches allow incorporation of key mechanistic information in otherwise biologically agnostic learning processes[74].

4. Concluding remarks and future perspectives

There is a need to clearly define where, when, and how omics can be used to effectively improve cell line development and optimize bioprocesses. Many factors contribute to the actionability of omics approaches, but we find systems biology modeling strategies are particularly poised to enable biotechnology researchers to enrich omic experiments with interpretable results. Importantly, they can make results from omics studies more *actionable* to enhance cell factory design. However, while we endorse the implementation of inherently more mechanistic models such as GeMs, we recognize that these and emerging tools need to improve upon accessibility, ease of execution, and accuracy for more widespread integration[75,76].

We focused on selecting downstream analysis tools for integration and interpretation of omic data; however, it is important to establish implementation strategies prior to experimentation. Researchers should clearly identify the biological questions they wish to interrogate and choose relevant experimental design and omic approaches accordingly.

 doi:10.20944/preprints202302.0037.v1

The experimental design should consider the number of features output by the omic tool(s) and subsequent analysis method when deciding on sample size and conditions. Accordingly, we strongly suggest that feature rich datasets be analyzed via mechanistic models if large sample numbers are not available to enable purely data-driven inference based approaches. Meanwhile, the omic approaches employed should be prioritized based on the ability to allow for the earliest, fastest, and most likely to succeed interventions.

Finally, we see a need to develop new resources in the industrial community to increase the size and diversity of omic datasets (see Outstanding Questions). Compared to disease and academic research, the CHO space lacks dedicated species-specific open-source databases such as GEO[77,78], TCGA[79], ICGC[80], ENCODE[81,82], etc., along with phenotypic databases[83], to facilitate benchmarking and transparency in analytical best practices. Public data resources such as these also increase statistical power in resolving biological patterns and targets. Reticence around data sharing is understandable as data from industrial cell lines contain proprietary molecule sequences and information supporting competitive advantages. However, the industry could significantly accelerate the progress towards actionable omics with freedom to access a lot more data relevant to the cell system, in addition to harnessing a more cohesive omics industrial community, and integrating omic approaches with appropriate analytical tools.

Outstanding Questions

Is there a path towards shared public omic databases in the biopharmaceutical space to create rich datasets similar to those available in human and other model organisms (e.g., The Cancer Genome Atlas [TCGA], Cancer Dependency Map, or ENCODE)?

- We acknowledge significant hurdles exist to ensure that proprietary sequence information about pre-clinical therapeutic compounds are not inadvertently shared, however, pre-filtering raw sequencing data before dissemination could be a potential solution.
- Can we establish best practices for data generation and sharing?

Is there a way to assess the likelihood of a genetic engineering or process modification strategy being universally applicable (e.g., into the host or into a platform process) based on the results of introduction into producing clones?

- Clonal and molecule variability means there may not be a “one size fits all” solution.

What is the role of academic/industrial partnerships and/or consortiums in driving innovation?

- Translatability of findings from academic studies is often unknown due to limitations on
- industrially-relevant domain knowledge and resources.
- Is there a route toward more collaborative partnerships (e.g., resource sharing or validation in an industrial context)?

Acknowledgments: This work was supported by generous funding from NIGMS (R35 GM119850), NIAID (UH2 AI153029), NSF (CBET-2030039) and the Novo Nordisk Foundation (NNF20SA0066621).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 February 2023 doi:10.20944/preprints202302.0037.v1

Text Boxes

Text Box 1: Overview of the biopharmaceutical protein production process

Several steps are taken in the development of a production cell line. A few common steps are as follow:

1) The sequence coding for the protein of interest (e.g. monoclonal antibody) is introduced to the cell along with a selectable marker (e.g. dihydrofolate reductase or glutamine synthetase) and integrated into the genome.
2) The resulting pool of cells is highly heterogeneous: some cells have integrated multiple copies of the antibody gene and are producing appreciable quantities, while others may be producing little to no product. Due to this and regulatory requirements, 100s to 1000s of single cells are isolated, grown, and evaluated to find the 'winners' which produce high quantities and quality of the desired product.
3) The winners are subjected to bespoke, intensified process development to maximize the amount of biotherapeutic generated. Owing to clonal variability and the plasticity of the CHO genome, the optimal process will almost certainly differ for each cell line.
4) Additional care must be taken–either via process modification or clone selection– to ensure that critical quality attributes (CQAs) such as glycosylation or aggregation are maintained at appropriate levels.

This process can take upwards of a year and is repeated for every biotherapeutic protein in a company's pipeline that aims to progress into clinical trials. Many steps of this process are critical path activities, thus delays must be avoided and innovations that shorten timelines are highly valuable.

Text Box 2: Mechanistic modeling approaches

Genome-scale metabolic models (GeMs) : GeM reconstructions are mathematical representations of an organism's stoichiometry-based, mass-balanced metabolic reactions using gene-protein-reaction (GPR) associations that have been formulated based on carefully curated genome annotations and experimental data[84–86]. Since the first GeM of *Haemophilus influenzae* RD was reported in 1999[87], numerous other GeMs have been published for a variety of organisms including industrially relevant *Escherichia coli[88]*, *Saccharomyces cerevisiae[89]*, *Bacillus subtilis[90,91]*, *Pichia pastoris[92]*, human[93], and CHO[94], most of which are accessible via the BiGG Models database (http://bigg.ucsd.edu/)[95].

ME models: Genome-scale models of metabolism and macromolecular expression (ME-models) integrate Metabolism and Expression on a genome scale, permitting calculation of the cellular cost of enzyme synthesis, in addition to stoichiometric balancing of the reaction(s) they catalyze[96]. These computational ME models provide a framework to determine a cell's most protein-cost-effective way of carrying out its required biological functions. Due to challenges regarding computational resources and model development, ME models have only been constructed for 3 organisms thus far: *Thermotoga maritima*, *Escherichia coli*, and *Clostridium ljungdahlii*.

Genome scale models of protein secretion: These computational models represent genome-scale stoichiometric reconstructions of metabolism coupled to protein secretion. In 2013, the first genome-scale model for yeast secretory machinery was constructed[97]. Following these efforts, models of mammalian metabolism coupled to protein secretion were developed for human, mouse, and CHO cells[27]. These models implement a protein-specific information matrix (PSIM) which quantifies select protein attributes (e.g: disulfide bonds, N-linked and O-linked glycans, transmembrane domains, protein length) for proteins of the secretome, enabling the construction of protein-specific secretory models using the template reactions in the reconstruction.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 February 2023 doi:10.20944/preprints202302.0037.v1

Kinetic models: Kinetic models use mathematical expressions of biochemical reaction rates, which form mass balance equations to capture the temporal behavior of the system. Unlike stoichiometric models that only require stoichiometry and directionality constraints, kinetic models require a considerable upfront investment for parameterization. Typically, enzyme characterization experiments must be performed to experimentally determine these parameters. Various kinetic parameterization approaches exist to determine the "best-fit" model that most closely emulates experimental data. With the emergence of simulation-based methods of parameter fitting such as Monte Carlo, the modeling community has advocated the use of multi-omic data sets to precisely fit these coarse-grained models[98].

Text Box 3: Data inference modeling approaches

Unsupervised learning: Unsupervised techniques are statistical methods that reduce the feature (e.g. genes or other molecular species) dimensions and to resolve patterns in unlabeled data that can correlate with a phenotype of interest. Matrix factorization (e.g. PCA, ICA, NMF), clustering (e.g. k-means, hierarchical), and autoencoders (ANN) identify sources of variation or separation in data. These approaches have been broadly applied and underpin many machine learning tools. More recently, methods like canonical correlation analysis (CCA) have been leveraged to integrate multi-omic data through the conservation of complex data patterns across layers of high-dimensional data.

Supervised learning: Supervised techniques generally adapt the same statistical foundation as unsupervised techniques and incorporate known data labels to consider association with dependent variables. Also comprising the basis of many machine-learning predictive models, these techniques fall under regression (e.g. PLS, gradient descent) and classification (e.g. SVM, k-NN). Supervised techniques range in transparency and explainability–tools like PLS and Random Forest can provide feature importance metrics while neural networks and other black box approaches offer lower transparency and explainability of the input features but capture non-linear or complex relationships.

Correlation Networks: Correlation networks combine unsupervised and supervised techniques with mechanistic considerations to infer association of network-based target groups with phenotype. Tools like Weighted Gene Co-expression Network Analysis (WGCNA) and gene regulatory network (GRN) inference implement dimension reduction techniques under biological topology based assumptions. WGCNA can be used to find groups of co-expressed genes aggregated into "eigengenes" with loadings that can be quantitatively correlated with phenotype metrics. GRN methods assume association, or co-expression, of genes with transcription factors to predict gene network regulons. GRN methods range from leveraging single-omic data (e.g. SCENIC built for scRNA-seq data[54]), to multi-omic data (e.g. scGRNom[55]) where integration of epigenomics provides an additional layer of mechanistic context to increase inference accuracy.

Empirical-based, interactome models: Empirically supported databases and protein-protein interaction networks offer mechanistic contextualization over conventional statistical analysis. Ingenuity pathway analysis (IPA) is a user-friendly platform and rich database that infers causal relationships and regulators, adding mechanistic and directional context over standard overrepresentation pathway analysis and GSEA against GO or KEGG. Protein-protein interaction (PPI) networks are experimentally identified, or statistically inferred physical or functional interactions between proteins and can be visualized or created using mass transfer dictated network propagation or user-friendly tools such as Cytoscape (http://cytoscape.org/).

Glossary

ANN: Artificial neural networks constitute a variety of deep learning technology inspired by the biological neural networks of the human brain. These networks consist of an input layer, one or more hidden node layers, and an output layer. Each node (artificial neuron) has an associated weight and threshold, and if the output of any individual

 doi:10.20944/preprints202302.0037.v1

node is above the given threshold then that node is activated and sends data to the next layer in the network.

CHO: Chinese hamster ovary cells are an epithelial cell line derived from the ovary of the Chinese hamster and have emerged as the gold standard production system in the biologics industry.

CQA: Critical quality attributes are physical, chemical, or biological properties or characteristics that must be within an appropriate limit, range, or distribution to ensure the desired product quality. A Quality-by-design framework is generally implemented to identify and define CQAs per molecule program's Quality Target Product Profile (QTPP).

DoE: Design of experiment is a structured data collection and analysis method used to study the relationship between various factors hypothesized to affect key output variables.

GeM: Genome-scale metabolic models are mathematical network representations of the metabolism formulated based on carefully curated genome annotations and experimental data.

GO: The Gene Ontology is a major bioinformatics initiative to unify the annotation of gene and gene product attributes across species.

GRN: Gene regulatory networks are groups of genes identified or inferred to interact with each other and possibly other molecular species to control cellular functions.

GSEA: Gene set enrichment analysis is a computational method to identify classes of genes that show significant, concordant differences between two biological states (e.g., phenotypes).

Matrix Factorization: Methods such as Principal Component Analysis (**PCA**) Independent Component Analysis (**ICA**) and Non-negative Matrix Factorization (**NMF**) can extract latent variables that represent biologically meaningful patterns to allow data interpretation or visualization of high-dimensional data.

KEGG: Kyoto Encyclopedia of Genes and Genomes consists of a collection of databases related to genomes, biological pathways, diseases, drugs, and chemical substances. According to the developer, KEGG is a "computer representation" of the biological system integrating genetic building blocks (genes/proteins), chemical building blocks (small molecules and reactions), and wiring diagrams of molecular interaction and reaction networks.

NGS: Next-generation sequencing offers ultra-high throughput sequencing technology that has revolutionized genomic research.

PLS: Partial least squares is a supervised regression method that decomposes data into principal components as in PCA, except that the components maximize correlation with the dependent variable. PLS and its variations are widely used in predictive machine learning models and can be easily implemented for multivariate omic data.

PQ: Product quality refers to physical and chemical molecule attributes that may affect the identity, efficacy, safety, or purity of the molecule and are closely monitored during cell line and bioprocess experimentation. Examples of product quality attributes typically quantified are glycan species, molecular size variants (such as high, medium, low molecular weight species), charged-base variants.

References

1 Zhang, Y.-H.P. *et al.* (2017) Biomanufacturing: history and perspective. *J. Ind. Microbiol. Biotechnol.* 44, 773–784

2 Amer, B. and Baidoo, E.E.K. (2021) Omics-Driven Biotechnology for Industrial Applications. *Front Bioeng Biotechnol* 9, 613307

3 Samoudi, M. *et al.* From omics to cellular mechanisms in mammalian cell factory development. , *Current Opinion in Chemical Engineering*, 32. (2021) , 100688

4 Stolfa, G. *et al.* (2018) CHO-Omics Review: The Impact of Current and Emerging Technologies on Chinese Hamster Ovary Based Bioproduction. *Biotechnol. J.* 13, e1700227

5 Hefzi, H. and Lewis, N.E. From random mutagenesis to systems biology in metabolic engineering of mammalian cells. , *Pharmaceutical Bioprocessing*, 2. (2014) , 355–358

6 Xiong, K. *et al.* (2021) An optimized genome-wide, virus-free CRISPR screen for mammalian cells. *Cell Rep Methods* 1,

7 Karottki, K.J. la C. *et al.* (2021) A metabolic CRISPR-Cas9 screen in Chinese hamster ovary cells identifies glutamine-sensitive genes. *Metab. Eng.* 66, 114–122

8 Schmieder, V. *et al.* (2021) A pooled CRISPR/AsCpf1 screen using paired gRNAs to induce genomic deletions in Chinese hamster ovary cells. *Biotechnol Rep (Amst)* 31, e00649

9 Bauer, N. *et al.* (2022) An arrayed CRISPR screen reveals *Myc* depletion to increase productivity of difficult-to-express complex antibodies in CHO cells. *Synth. Biol.* DOI: 10.1093/synbio/ysac026

10 Kretzmer, C. *et al.* (2022) De novo assembly and annotation of the CHOZN® GS genome supports high-throughput genome-scale screening. *Biotechnol. Bioeng.* 119, 3632–3646

11 Klingler, F. *et al.* (2021) Unveiling the CHO surfaceome: Identification of cell surface proteins reveals cell aggregation-relevant mechanisms. *Biotechnol. Bioeng.* 118, 3015–3028

12 Jerabek, T. *et al.* (2022) The potential of emerging sub-omics technologies for CHO cell engineering. *Biotechnol. Adv.* 59, 107978

13 Kol, S. *et al.* (2020) Multiplex secretome engineering enhances recombinant protein production and purity. *Nat. Commun.* 11, 1908

14 Valente, K.N. *et al.* (2018) Applications of proteomic methods for CHO host cell protein characterization in biopharmaceutical manufacturing. *Curr. Opin. Biotechnol.* 53, 144–150

15 Lewis, N.E. *et al.* (2009) Gene expression profiling and the use of genome-scale in silico models of Escherichia coli for analysis: providing context for content. *J. Bacteriol.* 191, 3437–3444

16 Gutierrez, J.M. and Lewis, N.E. (2015) Optimizing eukaryotic cell hosts for protein production through systems biotechnology and genome-scale modeling. *Biotechnol. J.* 10, 939–949

17 Lu, J. *et al.* (2022) In silico cell factory design driven by comprehensive genome-scale metabolic models: development and challenges. *Systems Microbiology and Biomanufacturing*

18 Lewis, N.E. *et al.* (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.* 10, 291–305

19 Hyduke, D.R. *et al.* (2013) Analysis of omics data with genome-scale models of metabolism. *Mol. Biosyst.* 9, 167–174

20 Huang, Z. and Yoon, S. Identifying metabolic features and engineering targets for productivity improvement in CHO cells by integrated transcriptomics and genome-scale metabolic model. , *Biochemical Engineering Journal*, 159. (2020) , 107624

21 Huang, Z. *et al.* CHO cell productivity improvement by genome-scale modeling and pathway analysis: Application to feed supplements. , *Biochemical Engineering Journal*, 160. (2020) , 107638

22 Fouladiha, H. *et al.* (2020) A metabolic network-based approach for developing feeding strategies for CHO cells to increase monoclonal antibody production. *Bioprocess Biosyst. Eng.* 43, 1381–1389

23 Hoang, D. *et al.* (2022) Modulation of Nutrient Precursors for Controlling Metabolic Inhibitors by Genome-Scale Flux Balance Analysis. *Biotechnol. Prog.*

24 Zhang, C. and Hua, Q. (2015) Applications of Genome-Scale Metabolic Models in Biotechnology and Systems Medicine. *Front. Physiol.* 6, 413

25 Lerman, J.A. *et al.* (2012) In silico method for modelling metabolism and gene product expression at genome scale. *Nat. Commun.* 3, 929

26 Dahal, S. *et al.* (2021) Recent advances in genome-scale modeling of proteome allocation. *Curr. Opin. Syst. Biol.* 26, 39–45

27 Gutierrez, J.M. *et al.* (2020) Genome-scale reconstructions of the mammalian secretory pathway predict metabolic costs and limitations of protein secretion. *Nat. Commun.* 11, 68

28 Schinn, S.-M. *et al.* (2021) A genome-scale metabolic network model and machine learning predict amino acid concentrations in Chinese Hamster Ovary cell cultures. *Biotechnol. Bioeng.* 118, 2118–2123

29 Kyriakopoulos, S. *et al.* (2018) Kinetic Modeling of Mammalian Cell Culture Bioprocessing: The Quest to Advance

Biomanufacturing. *Biotechnol. J.* 13, e1700229

30 Park, S.-Y. *et al.* Bioprocess digital twins of mammalian cell culture for advanced biomanufacturing. , *Current Opinion in Chemical Engineering*, 33. (2021) , 100702

31 Almquist, J. *et al.* (2014) Kinetic models in industrial biotechnology - Improving cell factory performance. *Metab. Eng.* 24, 38–60

32 Hossler, P. *et al.* (2007) Systems analysis of N-glycan processing in mammalian cells. *PLoS One* 2, e713

33 Stach, C.S. *et al.* (2019) Model-Driven Engineering of N-Linked Glycosylation in Chinese Hamster Ovary Cells. *ACS Synth. Biol.* 8, 2524–2535

34 Xu, J. *et al.* (2019) Systematic development of temperature shift strategies for Chinese hamster ovary cells based on short duration cultures and kinetic modeling. *MAbs* 11, 191–204

35 Ben Yahia, B. *et al.* (2021) Predictive macroscopic modeling of cell growth, metabolism and monoclonal antibody production: Case study of a CHO fed-batch production. *Metab. Eng.* 66, 204–216

36 Clarke, C. *et al.* Statistical methods for mining Chinese hamster ovary cell 'omics data: from differential expression to integrated multilevel analysis of the biological system. , *Pharmaceutical Bioprocessing*, 2. (2014) , 469–481

37 Mowbray, M. *et al.* Machine learning for biochemical engineering: A review. , *Biochemical Engineering Journal*, 172. (2021) , 108054

38 Antonakoudis, A. *et al.* (2020) The era of big data: Genome-scale modelling meets machine learning. *Comput. Struct. Biotechnol. J.* 18, 3287–3300

39 Alden, N. *et al.* (2020) Using Metabolomics to Identify Cell Line-Independent Indicators of Growth Inhibition for Chinese Hamster Ovary Cell-based Bioprocesses. *Metabolites* 10,

40 Barberi, G. *et al.* Anticipated cell lines selection in bioprocess scale-up through machine learning on metabolomics dynamics. , *IFAC-PapersOnLine*, 54. (2021) , 85–90

41 Barberi, G. *et al.* (2022) Integrating metabolome dynamics and process data to guide cell line selection in biopharmaceutical process development. *Metab. Eng.* 72, 353–364

42 Heffner, K. *et al.* (2020) Expanded Chinese hamster organ and cell line proteomics profiling reveals tissue-specific functionalities. *Sci. Rep.* 10, 15841

43 Budge, J.D. *et al.* Engineering of Chinese hamster ovary cell lipid metabolism results in an expanded ER and enhanced recombinant biotherapeutic protein production. , *Metabolic Engineering*, 57. (2020) , 203–216

44 Torres, M. *et al.* (2021) Metabolic profiling of Chinese hamster ovary cell cultures at different working volumes and agitation speeds using spin tube reactors. *Biotechnol. Prog.* 37, e3099

45 Lin, D. *et al.* (2020) CHOmics: A web-based tool for multi-omics data analysis and interactive visualization in CHO cell lines. *PLoS Comput. Biol.* 16, e1008498

46 Dhiman, H. *et al.* (2019) Genetic and Epigenetic Variation across Genes Involved in Energy Metabolism and Mitochondria of Chinese Hamster Ovary Cell Lines. *Biotechnol. J.* 14, e1800681

47 Choudhary, K.S. *et al.* (2020) Elucidation of Regulatory Modes for Five Two-Component Systems in Escherichia coli Reveals Novel Relationships. *mSystems* 5,

48 Huttlin, E.L. *et al.* (2017) Architecture of the human interactome defines protein communities and disease networks. *Nature* 545, 505–509

49 Yeo, H.C. *et al.* Combined multivariate statistical and flux balance analyses uncover media bottlenecks to the growth and productivity of Chinese hamster ovary cell cultures. , *Biotechnology and Bioengineering*, 119. (2022) , 1740–1754

50 Yao, G. *et al.* (2021) A Metabolomics Approach to Increasing Chinese Hamster Ovary (CHO) Cell Productivity. *Metabolites* 11,

51 Ding, M.Q. *et al.* (2018) Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics. *Mol. Cancer Res.* 16, 269–278

52 Tan, K. *et al.* (2020) A multi-omics supervised autoencoder for pan-cancer clinical outcome endpoints prediction. *BMC Med.*

Inform. Decis. Mak. 20, 129

53 Tzani, I. *et al.* 31-Mar-(2022) , Understanding the transcriptional response to ER stress in Chinese hamster ovary cells using multiplexed single cell RNA-seq. , *bioRxiv*, 2022.03.31.486542

54 Aibar, S. *et al.* (2017) SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* 14, 1083–1086

55 Jin, T. *et al.* (2021) scGRNom: a computational pipeline of integrative multi-omics analyses for predicting cell-type disease genes and regulatory networks. *Genome Med.* 13, 95

56 Keenan, A.B. *et al.* (2019) ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Res.* 47, W212–W224

57 Krämer, A. *et al.* (2014) Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* 30, 523–530

58 Szklarczyk, D. *et al.* (2021) The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* 49, D605–D612

59 Huang, J.K. *et al.* (2018) Systematic Evaluation of Molecular Networks for Discovery of Disease Genes. *Cell Syst* 6, 484–495.e5

60 Zhang, Q.C. *et al.* (2013) PrePPI: a structure-informed database of protein-protein interactions. *Nucleic Acids Res.* 41, D828–33

61 Chiang, A.W.T. *et al.* (2019) Combating viral contaminants in CHO cells by engineering innate immunity. *Sci. Rep.* 9, 8827

62 Kuo, C.-C. *et al.* (2021) Dysregulation of the secretory pathway connects Alzheimer's disease genetics to aggregate formation. *Cell Syst* 12, 873–884.e4

63 Sriyudthsak, K. *et al.* (2016) Mathematical Modeling and Dynamic Simulation of Metabolic Reaction Systems Using Metabolome Time Series Data. *Front Mol Biosci* 3, 15

64 Saa, P.A. and Nielsen, L.K. (2016) Construction of feasible and accurate kinetic models of metabolism: A Bayesian approach. *Sci. Rep.* 6, 29635

65 Girbig, D. *et al.* (2012) Systematic analysis of stability patterns in plant primary metabolism. *PLoS One* 7, e34686

66 Andreozzi, S. *et al.* (2016) iSCHRUNK--In Silico Approach to Characterization and Reduction of Uncertainty in the Kinetic Models of Genome-scale Metabolic Networks. *Metab. Eng.* 33, 158–168

67 Sridhara, V. *et al.* Predicting growth conditions from internal metabolic fluxes in an in-silico model of E. coli. .

68 Zhang, J. *et al.* (2020) Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism. *Nat. Commun.* 11, 4880

69 Shaked, I. *et al.* Metabolic Network Prediction of Drug Side Effects. , *Cell Systems*, 2. (2016) , 209–213

70 Narayanan, H. *et al.* Hybrid Models Based on Machine Learning and an Increasing Degree of Process Knowledge: Application to Cell Culture Processes. , *Industrial & Engineering Chemistry Research*, 61. (2022) , 8658–8672

71 Tsopanoglou, A. and del Val, I.J. Moving towards an era of hybrid modelling: advantages and challenges of coupling mechanistic and data-driven models for upstream pharmaceutical bioprocesses. , *Current Opinion in Chemical Engineering*, 32. (2021) , 100691

72 Kotidis, P. and Kontoravdi, C. (2020) Harnessing the potential of artificial neural networks for predicting protein glycosylation. *Metab Eng Commun* 10, e00131

73 Kotidis, P. *et al.* (2019) Model-based optimization of antibody galactosylation in CHO cell culture. *Biotechnol. Bioeng.* 116, 1612–1626

74 Zampieri, G. *et al.* (2019) Machine and deep learning meet genome-scale metabolic modeling. *PLoS Comput. Biol.* 15, e1007084

75 Richelle, A. *et al.* (2021) Model-based assessment of mammalian cell metabolic functionalities using omics data. *Cell Rep Methods* 1,

76 Richelle, A. *et al.* (2019) Assessing key decisions for transcriptomic data integration in biochemical networks. *PLoS Comput. Biol.* 15, e1007185

77 Edgar, R. *et al.* (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210

 doi:10.20944/preprints202302.0037.v1

78 Barrett, T. *et al.* (2013) NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* 41, D991–5

79 Tomczak, K. *et al.* (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* 19, A68–77

80 International Cancer Genome Consortium *et al.* (2010) International network of cancer genome projects. *Nature* 464, 993–998

81 ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74

82 Luo, Y. *et al.* (2020) New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* 48, D882–D889

83 Golabgir, A. *et al.* (2016) Quantitative feature extraction from the Chinese hamster ovary bioprocess bibliome using a novel meta-analysis workflow. *Biotechnol. Adv.* 34, 621–633

84 Kim, H.U. *et al.* (2008) Metabolic flux analysis and metabolic engineering of microorganisms. *Mol. Biosyst.* 4, 113–120

85 Thiele, I. and Palsson, B.Ø. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* 5, 93–121

86 Gu, C. *et al.* (2019) Current status and applications of genome-scale metabolic models. *Genome Biol.* 20, 121

87 Edwards, J.S. and Palsson, B.O. (1999) Systems properties of the Haemophilus influenzae Rd metabolic genotype. *J. Biol. Chem.* 274, 17410–17416

88 Edwards, J.S. and Palsson, B.O. (2000) The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. U. S. A.* 97, 5528–5533

89 Förster, J. *et al.* (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res.* 13, 244–253

90 Henry, C.S. *et al.* (2009) iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations. *Genome Biol.* 10, R69

91 He, M. *et al.* (2021) Metabolic engineering of based on genome-scale metabolic model to promote fengycin production. *3 Biotech* 11, 448

92 Theron, C.W. *et al.* (2018) Integrating metabolic modeling and population heterogeneity analysis into optimizing recombinant protein production by Komagataella (Pichia) pastoris. *Appl. Microbiol. Biotechnol.* 102, 63–80

93 Duarte, N.C. *et al.* (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U. S. A.* 104, 1777–1782

94 Hefzi, H. *et al.* (2016) A Consensus Genome-scale Reconstruction of Chinese Hamster Ovary Cell Metabolism. *Cell Syst* 3, 434–443.e8

95 Norsigian, C.J. *et al.* (2020) BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. *Nucleic Acids Res.* 48, D402–D406

96 Fang, X. *et al.* (2020) Reconstructing organisms in silico: genome-scale models and their emerging applications. *Nat. Rev. Microbiol.* 18, 731–743

97 Feizi, A. *et al.* (2013) Genome-scale modeling of the protein secretory machinery in yeast. *PLoS One* 8, e63284

98 Islam, M.M. *et al.* Kinetic modeling of metabolism: Present and future. , *Current Opinion in Systems Biology*, 26. (2021) , 72–78