Application of Systems Engineering Principles and Techniques in Biological Big Data Analytics: A Review

Authors:

Q. Peter He, Jin Wang

Date Submitted: 2020-12-17

Keywords: dynamic analysis, overfitting, feature engineering, systems engineering, Machine Learning, biological big data

Abstract:

In the past few decades, we have witnessed tremendous advancements in biology, life sciences and healthcare. These advancements are due in no small part to the big data made available by various high-throughput technologies, the ever-advancing computing power, and the algorithmic advancements in machine learning. Specifically, big data analytics such as statistical and machine learning has become an essential tool in these rapidly developing fields. As a result, the subject has drawn increased attention and many review papers have been published in just the past few years on the subject. Different from all existing reviews, this work focuses on the application of systems, engineering principles and techniques in addressing some of the common challenges in big data analytics for biological, biomedical and healthcare applications. Specifically, this review focuses on the following three key areas in biological big data analytics where systems engineering principles and techniques have been playing important roles: the principle of parsimony in addressing overfitting, the dynamic analysis of biological data, and the role of domain knowledge in biological data analytics.

Record Type: Published Article

Submitted To: LAPSE (Living Archive for Process Systems Engineering)

Citation (overall record, always the latest version):	LAPSE:2020.1226
Citation (this specific file, latest version):	LAPSE:2020.1226-1
Citation (this specific file, this version):	LAPSE:2020.1226-1v1

DOI of Published Version: https://doi.org/10.3390/pr8080951

processes

MDPI

Review

Application of Systems Engineering Principles and Techniques in Biological Big Data Analytics: A Review

Q. Peter He * and Jin Wang

Department of Chemical Engineering, Auburn University, Auburn, AL 36849, USA; wang@auburn.edu
* Correspondence: qhe@auburn.edu

Received: 29 June 2020; Accepted: 30 July 2020; Published: 7 August 2020

Abstract: In the past few decades, we have witnessed tremendous advancements in biology, life sciences and healthcare. These advancements are due in no small part to the big data made available by various high-throughput technologies, the ever-advancing computing power, and the algorithmic advancements in machine learning. Specifically, big data analytics such as statistical and machine learning has become an essential tool in these rapidly developing fields. As a result, the subject has drawn increased attention and many review papers have been published in just the past few years on the subject. Different from all existing reviews, this work focuses on the application of systems, engineering principles and techniques in addressing some of the common challenges in big data analytics for biological, biomedical and healthcare applications. Specifically, this review focuses on the following three key areas in biological big data analytics where systems engineering principles and techniques have been playing important roles: the principle of parsimony in addressing overfitting, the dynamic analysis of biological data, and the role of domain knowledge in biological data analytics.

Keywords: biological big data; systems engineering; machine learning; feature engineering; overfitting; dynamic analysis

1. Introduction

Massive quantities of data are being generated in biology, the life sciences and healthcare industries and institutions, which hold the promise of advancing our understandings of various biological systems and diseases, developing new biocatalysts and drugs, as well as delivering more affordable and effective patient care. These massive data collected from high throughput instruments are often referred to as "big data", which are generally characterized by their 4V characteristics: Volume (size or scale), Variety (multitype), Velocity (batch or streaming), and Veracity (uncertainty) [1,2]. Big data have inspired revolutionary breakthroughs in a variety of fields and many journal special issues have been published in the past a few years, such as "Scalable Computing for Big Data" in *Big Data Research* [3], "Big Data Analytics for Business Intelligence" in *Expert Systems with Applications* [4], and "Big Data and Natural Disasters" in *Computers and Geosciences* [5]. Driven by the needs, we have also witnessed rapid development in programming languages and integrated development environments (IDEs) especially suited for analyzing big data such as R and Python [6–8]. With the explosion in big data-related research, there are also concerns about quality and integrity of the publications such as reproducibility [9–11] and ethics [12]. To get a big picture of the research in the biological big data analytics field, we conducted a search on the Web of Science using the exact phrase: "big data" and any of the following words or phrases: biology, "life science", healthcare, "health care", biomedical, disease, and cancer. We also added the additional language constraint of "English only", document type constraint of "Article only",

and year constraint of "2010–2019". The search returned 2913 records, which are shown as the vertical bars in Figure 1. As can be seen from Figure 1, the big data analytics in biology, life science and healthcare is really a new area, which, unsurprisingly, coincides with the emergence and rapid advancements of high-throughput technologies such as next-generation sequencing (RNA-Seq). In the past five years, the increase in publications is almost linear with the rate of 135 papers per year, indicating fast growing of the field. The number of citing articles increases even faster as indicated by the line chart in Figure 1, suggesting strong and growing influence of the topic. It is worth noting that these numbers are conservative as not all articles on the subject have used the phrase "big data". Nevertheless, the numbers do provide a glimpse of the field and capture its general upward trend.

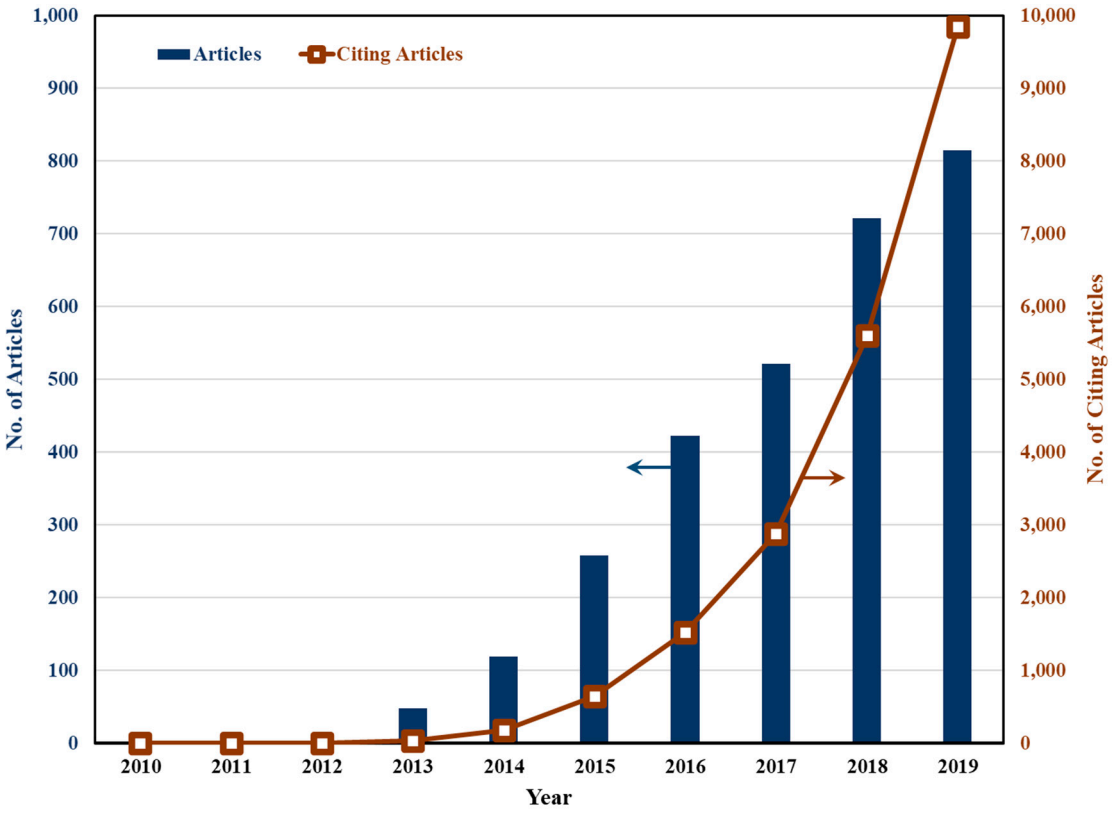

Figure 1. Journal articles of big data analytics in biology, life sciences and healthcare, and their citation numbers in the past decade based on a Web of Science search.

Given the fast-growing nature of the field, many review papers have been published in just the past few years. For example, there are many broad overview of big data analytics in the fields of biology, biomedical, and healthcare, such as what are the big data, where are the big data sources in the relevant field, what are the characteristics of these big data (e.g., volume, velocity, variety and veracity or 4V's of big data), and broad discussions on where or what are the opportunities and potential challenges, and future trends or perspectives [13–18]. There are also reviews in more specific areas, such as data sources and databases [19,20], data fusion and integration [21–23], data mining and machine algorithms [24–27], deep learning [28–30], imaging informatics [31], high-performance computing platforms [32,33], cloud and parallel bioinformatics tools [34,35], security, privacy and ethics [12,36]. Most technical reviews compare different methods and discuss their pros and cons. They also point to available commercial or open source techniques or tools such as R packages and/or Python libraries.

Different from all existing reviews, this work focuses on the application of systems engineering principles and techniques in addressing some of the important challenges in big data analytics for biological, biomedical and healthcare applications. A particular focus is given to research from the

systems engineering community in the chemical engineering field and others. Specifically, the following subjects are reviewed: the principle of parsimony in addressing overfitting, the advancement of dynamic modeling and analysis, and the domain knowledge-guided data analytics. In this work, we use (data) points, measurements and samples interchangeably, while variables and features are interchangeable unless explicitly stated otherwise. In addition, strictly speaking, there are distinctions between statistical learning and machine learning. In general, for statistical learning, a certain sample distribution is assumed, while for machine learning, data distribution is usually not assumed to be known or to follow certain form. However, for ease of reading, these distinctions are largely ignored in this work.

2. Principle of Parsimony in Addressing Overfitting

The principle of parsimony or principle of simplicity, a.k.a. Occam's razor, is one of the fundamental systems engineering principles. It suggests that one should choose the simplest explanation of a phenomenon (i.e., a model), which requires the fewest assumptions and/or fewest parameters. This principle directly leads to the concept of avoiding overfitting in data-based modeling. In other words, overfitting is the use of models or procedures that violate parsimony. Specifically, overfitting is the use of models that include more terms than are necessary or the use of more complicated models or approaches than are necessary [37]. There are two types of overfitting. One is the use of a model that is more flexible than it needs to be. For example, a neural net instead of a linear regression is used when the data set conforms to a linear model. The other is the use of a model that includes irrelevant components. For example, a polynomial of excessive degree or a multiple linear regression containing irrelevant predictors. Overfitting is undesirable for a number of reasons, including worse predictions on new data, identifying the wrong target in drug discovery, wasting resources on measuring irrelevant variables, and reducing model portability [37].

Overfitting can occur in all data analytics or modeling applications, either statistical learning where certain sample distribution is assumed or general machine learning where data distribution are not assumed to be known or to follow certain form. In this work, we focus on the following applications: classification (e.g., given measurements, to assign classes such as normal/control group and disease/patient group), modeling (e.g., linear or nonlinear regression/prediction, metabolic network modeling, etc.), and biomarker identification (e.g., given classes/regressands, to identify important biomarkers/regressors, etc.). In the following, we first review methods that enable us to check for overfitting, then we review methods to reduce the risk of overfitting.

2.1. Checking for Overfitting

There are direct and indirect ways that can be employed to check for overfitting. In the direct way, overfitting is not viewed as an absolute measure but involves a comparison. For example, one model is an overfit if its predictions are no better than those of another simpler model [37]. Besides directly comparing predictions from models of different complexity to detect overfitting, there are indirect indications of overfitting even with a single model. In this case, the most commonly deployed strategies are different forms of cross-validation, including leave-one-out (LOO), multi-fold cross-validation, and Monte Carlo cross-validation [37–40]. No matter which form of cross-validation is used, the procedure reuses the training samples as validation samples. Because of that, cross-validation is also known as internal validation [39]. This is not a serious issue if model hyperparameters (e.g., number of principal components (PCs) in a partial least squares (PLS) model, regularization parameter in a Lasso regression, number of neurons in each layer of a neural network model) are not optimized or determined through cross-validation. In other words, the validation samples have no influence on model building in any way. This is the case for some traditional machine learning algorithms where there is no model hyperparameter, such as multiple linear regression and linear discriminant analysis. However, in most modern machine learning algorithms, there are model hyperparameters that need to be determined or optimized. This is often carried out during the cross-validation step in most studies. In these cases, cross-validation is actually part of the model calibration or tuning,

making it a poor method for verifying the fit of a model. Or perhaps cross-validation is a misnomer that is incorrectly used in this situation and a better term is hyperparameter tuning or optimization as part of the model training. In these cases, having independent test samples that are held out throughout the validation process is essential to get a fair evaluation of a model predictive capability. In addition, this independent test set needs to be large to be trustworthy. A small test set is not reliable because it leaves room for chance (e.g., a particular test set containing an extreme sample or outlier). However, having a large independent test set is a luxury that may not be available in some applications. In these cases, we recommend Monte Carlo validation and testing (MCVT) that we proposed recently [41]. MCVT randomly select training samples, validation samples (or more properly, tuning samples for optimizing model hyperparameters), and set-aside test samples from all samples. In MCVT, model parameters (e.g., coefficients of a PLS model or weights of a neural net) are determined based on the training samples with model hyperparameters (e.g., number of PCs, regularization parameters, number of neurons, etc.) tuned using the tuning (cross-validation) samples, while the model performance is assessed using the test samples. We recommend using the mean and standard deviations of normalized root-mean-square error (NRMSE) of the predictions on the test samples from all MCs to evaluate the accuracy and precision of the model, respectively. Due to the multiple MC runs, the assessment bias due to a particular division of training, validation and testing samples can be eliminated. If the mean and/or standard deviation of the normalized RMSE from training or validation set is much smaller than those of the test set, it suggests high model flexibility and is a warning sign of overfitting.

2.2. Reducing or Avoiding Overfitting

As shown in Figure 2, there are three ways to reduce the risk of overfitting: (1) reducing feature space through feature selection or reduction, and feature combination or extraction; (2) reducing parameter space by selecting a model with a small number of parameters, imposing regularization on model parameters, or removing/fixing non-identifiable or insensitive model parameters; and (3) increasing sample space through data augmentation, oversampling, or upsampling. These strategies are not mutually exclusive and can be used in any combination when possible to obtain the maximum benefit. In the following we review some representative studies that have employed these strategies.

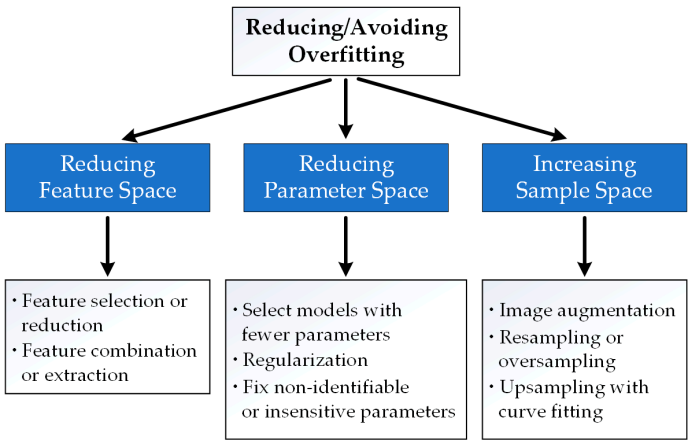

Figure 2. Strategies for reducing/avoiding overfitting.

2.2.1. Reducing Feature Space

Feature Selection or Reduction

When dealing with high dimensional data, especially when the features are highly correlated, the most common strategy to reduce overfitting is through explicit feature selection or reduction. Eliminating irrelevant features also improves model interpretability. In certain settings, such as

bioinformatics, biomarker discovery, or disease detection, determining the most descriptive features for the system under study in order to gain insights and a greater understanding is at least as important as obtaining a high performance model [42]. In this strategy, the most informative features are selected or retained to be included in the model, while the redundant or less informative features are eliminated. There are many feature selection methods proposed in the literature. For more comprehensive discussions on various feature selection techniques, the interested readers are referred to some review papers on the subject [43–51]. Here we only review a few examples used in the biological data analysis. Yang et al. [52] proposed a consensus feature elimination (CFE) approach where 100 MC runs were performed to generate 5-fold random splits of samples. A total of 500 classifiers were constructed based on the random splits with their AUCs (Area Under Receiver Operating Characteristic Curves) and weights of the features. Each feature is then ranked by its average square weight. The lowest ranking feature was removed backward until the maximum average AUC was achieved. The procedure is repeated 100 times, and the most frequently occurring feature set was regarded to be the ultimate feature set. The authors used CFE to identify genes directly correspond to breast cancer metastasis (i.e., the "driver" genes) using three breast cancer datasets. The biomarkers identified by the proposed method exhibited 10-fold higher reproducibility than other methods, with up to 30-fold greater enrichment for known cancer-related genes, and 4-fold enrichment for known breast cancer susceptible genes. Guzman et al. [53] proposed a mixed-integer linear programming (MILP) based approach where feature selection was formulated as an optimization problem. The proposed method is applied to an untargeted, high-throughput proteomics data of gingival crevicular fluid samples collected from patients after treatment of chronic periodontitis. Dean et al. [54] used p-value, q-value and fold-change to select potential biomarkers based on multiple classification methods. The approach was used to reduce over one million initial features (from multi-omics data including genomics, transcriptomics, proteomics, methylomics, lipidomics and metabolomics) down to 343 candidate biomarkers and later down to a final 28 biomarker panel. Lee et al. [55] used least absolute shrinkage and selection operator (Lasso) to select 20 features out of a total of 324 features in an electroencephalography (EEG) dataset. The selected features were then used to build a logistic regression (LR) model for detecting drivers' anxiety invoked by driving situations. In addition to pure data-driven feature selection approaches, domain knowledge can also be utilized to help with feature selection to avoid overfitting. For example, in developing a multiple linear regression model to identify significant parameters to be included for the estimation of lipid induction in *Chlorella sorokiniana* HS1, Oh et al. [56] used domain knowledge to identify highly correlated variables such as chlorophyll and carotenoid, then selected representative variables before developing the model.

Feature Combination or Extraction

Besides feature selection, feature combination or feature extraction has also been used to reduce the dimension of a feature space. Under this category, principal component analysis (PCA) has been the most popular method due to its theoretical simplicity and computational efficiency. For example, in [57], PCA was used to extract 80 principal components (PCs) from a protein dataset of 260 features. The PCs were then used as the input to a neural network model, which achieved superior protein secondary structure prediction performance compared to when the raw data were used as the input. In another study, a sparse PCA (SPCA) was proposed to select combinations of miRNAs in the blood that discriminate between healthy controls and diseases [58]. Compared to regular PCA, which employs all original features to construct a new, lower dimensional feature space, SPCA aims to construct lower dimensional features by a smaller number of the original features. In other words, SPCA is a feature extraction method that integrates feature selection and feature combination. PCA and its variants have also been applied for feature combination/extraction in autism spectrum disorder (ASD) biomarker discovery using folate-dependent one-carbon metabolism (FOCM) and transsulfuration (TS) metabolites [59] and urinary toxic metals [60], in silico drug discovery for posttraumatic stress disorder-mediated heart disease [61], diagnosis of valvular heart diseases using the Doppler heart

sounds [62], single-cell gene expression analysis [63], antimicrobial resistance analysis of clinical pathogens [64], and diagnosis of bioprocess cell growth [65]. Besides feature combination/extraction, PCA has also been used to handle missing features [66]. For more applications of PCA and other feature combination/extraction techniques, the interested readers are referred to some review papers on the subject [67–70].

2.2.2. Reducing Parameter Space

Because of the direct relationship between features and their corresponding parameters, reduction in one leads to reduction in the other. In the previous section, we discussed developing parsimonious models through reducing feature space. In this section, we review studies where parsimonious models are developed by reducing parameter space. This can be achieved by selecting a model (structure) with a small number of parameters, imposing a regularization on model parameters, or parameter identifiability analysis and sensitivity analysis to identify and eliminate non-identifiable, poorly identifiable, and insensitive model parameters.

Selecting a Model with a Small Number of Parameters

To avoid overfitting by reducing parameter space, an intuitive approach is to select a model with a small number of parameters. The reduced-order model can be a simplified mechanistic model based on domain knowledge, or a data-driven model capturing major relationships that have been revealed in previous studies. For example, in [71], a linear model was selected to capture the insulin–glucose and carbohydrate–glucose relationships. The linear model ignores the high order non-linear relationships such as the effect of physiological exercise, stress, and hormonal variations on the subcutaneous glucose level. However, the advantage of using this parsimonious model is that it is more robust and also easily identifiable in clinical practice [71]. In another example, a simpler but more robust version of the metabolic network models was chosen to address the limited number of points or measurements [72]. Grivas et al. [73] also noted the importance of reducing model complexity by selecting models with fewer parameters, especially in clinical studies where participant recruitment is an obstacle or when data collection has already concluded. An extension to this idea is to determine or estimate some model parameters using literature data. For example, in [74], to avoid overfitting of an integrated posttraumatic stress disorder (PTSD) model, a key model parameter was estimated based on data reported in the literature to reduce the number of parameters to be estimated using their own data. There are methods proposed for systematic simplification of general dynamic models to obtain reduced-order mechanistic models. For instance, Bastin and Dochain [75] studied online estimation and adaptive control of bioreactors extensively. They have shown that although a bioreactor dynamic model can be fairly complex involving a large number of differential equations, a simplified reduced-order model is often sufficient for many practical applications from an engineering viewpoint. They demonstrated the application of singular perturbation technique to transform differential equations into algebraic equations when the dynamics of certain reactions or processes can be neglected. The readers interested in more thorough discussion on model reduction for biological systems are referred to a recent review paper [76].

Regularization

Regularization is another strategy for developing parsimonious models. Instead of selecting a known model with a small number of parameters, this strategy starts with a full parameter space (and hence full feature space) with one or multiple regularization term(s) imposed on the parameter space. To illustrate the concept, we adopt the approach taken in [77] where learning is viewed as a multivariate function approximation. The goal is to minimize the following function based on a set of N observations $\left\{\left(\boldsymbol{x}_i, \boldsymbol{y}_i\right)\right\}_{i=1}^{N}$ obtained by random sampling of a function f:

$$H[f] = \sum_{i=1}^{N} \left(f(\boldsymbol{x}_i) - \boldsymbol{y}_i \right)^2 + \lambda \phi[f] \tag{1}$$

where λ is a positive number known as the regularization parameter. The first term in Equation (1) enforces closeness to the data, and the second smoothness. The regularization parameter λ controls the tradeoff between these two terms [77]. For example, in multiple linear regression, the regularization imposed on the regression model parameters $\beta_i (i = 1 \cdots p)$ is defined as $\lambda \sum_{i=1}^{p} |\beta_i|$ for L1 regularization (a.k.a. Lasso regression), $\lambda \sum_{i=1}^{p} \beta_i^2$ for L2 regularization (a.k.a. ridge regression), and $\lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p} \beta_i^2$ for elastic net regularization, respectively. The regularization parameter(s) λ_i are hyperparameters determined through cross-validation as discussed in Section 2.1. There are many examples where regularization is utilized to avoid overfitting. For example, an expectation maximization sparse discriminant analysis (EM-SDA) was proposed in [66] to produce a sparse LDA model for classification of high-dimensional biological datasets, where a regularization penalty term was included in SDA to ensure sparsity and therefore to avoid overfitting. It has long been recognized that in the cases where statistical methods have failed to give a robust solution, neural networks can suffer from the same issues such as collinearity and overfitting [78]. As a result, regularization has also been commonly used in deep learning to address overfitting where the objective function to be minimized often includes a loss function that penalizes the prediction error, and a regularization function that penalizes model complexity. These applications can be found in many references [42,65,73,79–81]. A more detailed discussion on various regularization techniques employed in the literature can be found in [82].

Model Parameter Identifiability Analysis and Sensitivity Analysis

Mathematical modeling of biological systems often starts with a large model whose parameters are tuned to fit experimental data. As a result, a model can be over-parameterized with non-identifiable or poorly identifiable model parameters. The values of these non-identifiable or poorly identifiable parameters cannot be uniquely and/or reliably determined based on the available data. The possible reasons include redundancy and/or interdependence among parameters (i.e., structural non-identifiability), as well as limited data and/or poor data quality (i.e., practical non-identifiability) [83–85]. Therefore, it is desirable to detect these non-identifiable parameters so that a parsimonious model can be derived.

Generally speaking, model parameter identifiability analysis quantifies how accurate the value of a model parameter can be determined by the experimental data. There are several approaches for performing model parameter identifiability analysis, such as power series approach [86,87], similarity transformation [88], and differential algebraic methods [89]. A more comprehensive review of these methods can be found in [83]. In this work, we focus on a data-driven approach introduced by Raue et al. [83], in which prior knowledge is not required, although it can be incorporated if available.

Let $\boldsymbol{x}(t)$ denote the internal states and $\boldsymbol{y}(t)$ the observed outputs of a biological system whose behaviors are described by the following ordinary differential equations (ODEs)

$$\dot{\boldsymbol{x}}(t) = g(\boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{p}) \tag{2}$$

$$\dot{\boldsymbol{y}}(t) = h(\boldsymbol{x}(t), \boldsymbol{s}) + \boldsymbol{\epsilon}(t) \tag{3}$$

where g and h are functions describing the interactions among the states, and between the states and outputs, respectively. $\boldsymbol{u}(t)$ denotes system inputs/stimuli, $\boldsymbol{\epsilon}(t)$ measurement noise, $\boldsymbol{p}$ and $\boldsymbol{s}$ parameters of functions g and h, respectively. The model is fully specified with the parameter set

$$\theta = \{\boldsymbol{p}, \boldsymbol{S}, \boldsymbol{x}(0)\} \tag{4}$$

where $\boldsymbol{x}(0)$ denotes the initial states.

The confidence interval of a parameter estimate $\hat{\theta}_i$ can be estimated asymptotically or based on maximum likelihood estimation [83,90]. In the case of finite number of samples, it has been suggested that likelihood-based confidence intervals are more appropriate than asymptotic confidence intervals [83,91]. A parameter θ_i is identifiable if the confidence interval of its estimate $\hat{\theta}_i$ is finite. Based on this idea, Raue and co-workers have developed an approach based on profile likelihood [83], and a model reduction procedure that iteratively eliminates non-identifiable parameters [84].

It is also possible that a model parameter has limited influence on measured outputs. In this case, the model parameter is poorly identifiable because the system outputs are insensitive to the uncertainty of its estimated value. Sensitivity analysis can be applied to detect insensitive parameters and their values can then be fixed to nominal values based on literature and related processes or steps may be simplified or eliminated to obtain a parsimonious model [92]. Sensitivity analysis can be applied on either inputs $\boldsymbol{u}(t)$ or parameters θ for different purposes. For model reduction, sensitivity analysis is performed with respect to parameters, which is also known as parametric sensitivity. There are two types of parametric sensitivity: local and global sensitivity. Local sensitivity calculates the following partial derivative around the nominal values of θ:

$$\boldsymbol{S}_i(t) = \frac{\partial \boldsymbol{y}(t)}{\partial \theta_i} = \lim_{\Delta\theta_i \to 0} \frac{\boldsymbol{y}(t, \theta_i + \Delta\theta_i) - \boldsymbol{y}(t, \theta_i)}{\Delta\theta_i} \tag{5}$$

while keeping $\theta_j (j \neq i)$ at their nominal values. Local sensitivities can be viewed as the gradients around the nominal θ. Global sensitivity takes into account the range and/or statistical properties of the parameter uncertainty, and calculate an average, $\langle \partial \boldsymbol{y}(t)/(\partial \theta_i) \rangle$, over the region of parameter uncertainty. As noted in [93], local sensitivities provide more details with less computation, while global sensitivities are better at handling large variations in θ. The choice depends on the system characteristics and parameters under consideration. Thorough treatments of parameter sensitivity analysis and its application to systems biology and chemical kinetics models can be found in [92–94].

2.2.3. Increasing Sample Space

To reduce overfitting, previously we discussed strategies of reducing parameter space directly by selecting a simpler model with fewer parameters or imposing regularization on model parameters, or indirectly by reducing feature space (i.e., feature selection or combination). An alternative or complementary approach is to artificially increase the sample space (i.e., the number of samples) through so-called data augmentation. This sample space solution to overfitting is the process of supplementing a dataset with similar data that is created from the information in that dataset [95]. When implemented properly, this technique is very effective in coping with small datasets and/or limited labeled samples. The use of augmentation in deep learning is ubiquitous because of its heavy reliance on big data to avoid overfitting. This is especially true when dealing with images, where rotation, translation, blurring and other modifications to existing images are often used to generate augmented images for reducing overfitting [96]. For example, to facilitate feedforward control in artificial pancreas systems, Chakrabarty et al. [79] proposed a deep-learning-assisted macronutrient estimation technique to automatically estimate the macronutrient content via real-time image recognition. Image augmentation was performed in order to reduce overfitting of a deep convolutional neural network (CNN) model. Specifically, the augmentation was implemented by translating image pixels by up to 20% of their respective heights and widths, flipping images horizontally, shifting color channels by up to 30%, and zooming out by up to 20%. Frid-Adar et al. [97] proposed a data augmentation method using generative adversarial networks (GANs) to generate synthetic computed tomography (CT) images of liver lesions. The classification performance with the data augmentation showed significantly improved sensitivity and specificity. It is worth noting that image augmentation through rotation, translation and other modifications is a straightforward and very effective way to enlarge sample space. However, data augmentation beyond the field of image classification/recognition has drawn limited attention. This is partially due to the fact that the

implantation of data augmentation to other datasets, such as omics data, is not that straightforward and further investigation is needed to study any unexpected side effects that data augmentation may introduce to the modeling and analysis of these datasets. There is a hope, though, in the case of dynamic data (a.k.a. time-series or time-course data) where limited measurements over time are smoothed and fitted by a linear, spline or polynomial function, and more frequent data points can be obtained from the fitted line or curve to increase temporal resolution [98,99].

2.3. Summary and Discussion

When applying the principle of parsimony to address overfitting, special attentions have been paid to high flexibility models such as nonlinear, kernel, or neural nets based modeling approaches. For example, in [100] when performing classification to differentiate individuals with autism spectrum disorder (ASD) from typically developing peers, both a linear/Fisher discriminant analysis (LDA or FDA) and nonlinear kernel-based FDA (KFDA) classifiers were tested on plasma amino acid measurements. Considering that KFDA is significantly more prone to model overfitting, the authors used only combinations of up to five plasma amino acids for KFDA. In other words, variable sets containing more than five plasma amino acids were not considered. Similar approaches have been used in [59,60]. In addition, most studies reviewed in this work used independent test samples to check for overfitting. By reserving a subset of data for model prediction, the procedure provides a statistically independent assessment of model performance. However, the size of the test set used in some studies are small, which could subject their results to chance of extreme samples, outliers or other data artifacts.

3. Dynamic Analysis of Biological Data

The field of control and systems has been connected to biological systems and biotechnology for many decades, such as Norbert Wiener on cybernetics in 1965, Walter Cannon on homeostasis in 1929, and Claude Bernard on the *milieu interieur* in 1865 [101]. This connection has led to the recent birth and blossom of systems biology, where systems engineering principles have been instrumental and PSE community have been an integral component of this interdisciplinary field. However, the focus of this review is on how modeling or analyzing dynamic biological data can extract more meaningful information compared to studying steady state alone. For the broad systems biology field, interested reads are referred to some excellent papers discussing how the principles of dynamics and control have been contributing to the field [102–107].

In recent years, high-throughput technologies enabled the large-scale analysis of many of the cellular components. These various '-omes' are inventories or descriptive lists of (isolated) parts or components, and the enormous quantity of data generated by the -omics disciplines represents a 'snapshot' or static description of the investigated system [108]. For example, RNA-seq-based transcriptome profiling has been used to enhance the understanding of the genome-scale response of the organism to different stimuli. While these studies have provided insightful findings, they are most often limited to studying one or few steady-state conditions. For example, by comparing the whole gene expression profiles at two steady states, one can identify the key genes that show large changes in gene expression levels and postulate possible gene regulatory mechanisms. However, such steady state comparison completely misses the genes that only show differential expression during the transition between the two steady states, therefore missing important insights on the regulatory mechanism of cellular metabolism [109,110]. Similarly, physiological measurements for biomarkers are usually just collected at one point in time or at most a few selected time points during a medical intervention [111]. In these cases, biomarker measurements taken at specific time points, such as "before" and "after," are commonly used to evaluate the effects of treatment. As a result, models used as biomarkers are generally time-invariant and depict a steady-state system [73].

Recently, we started to see some change in the field, especially with more participation from the systems engineering community. For example, it has been recognized that cellular systems are networks of interacting components that change with time in response to external and internal

events. Several early publications have demonstrated that the pairing of the systematic application of experimental perturbations with high-throughput data collection methods and the observation of the temporal changes as the consequence of the perturbations can provide systematic insights [112–115]. Therefore, to reveal the dynamic behavior of these networks, and hence their functionalities, it is important to record many snapshots at different time points. It is now recognized that quantitative time-series data can lead to more meaningful models to improve our knowledge of human physiology in health and disease, and aid the search for earlier diagnoses, better therapies and a healthier life [108]. From a control perspective, it makes perfect sense because cellular metabolism is a highly complex dynamic system. As illustrated in Figure 3, the transient response trajectory could offer significantly more information on the system dynamics, especially the internal control mechanism, such as the accumulation or depletion of a compound, oscillations, switches (on/off)and delays [108]. On the other hand, it is impossible to reveal the transient behavior if only the two (e.g., "before" and "after") steady state data points were recorded and analyzed. Therefore, studying the dynamic behavior of these networks ought to be the basis for understanding cellular functions and disease mechanisms. Since gene regulatory networks, metabolic networks, and signal transduction networks are some of the most important cellular networks in biological systems, in this work we review the progress made towards dynamic analysis of these networks.

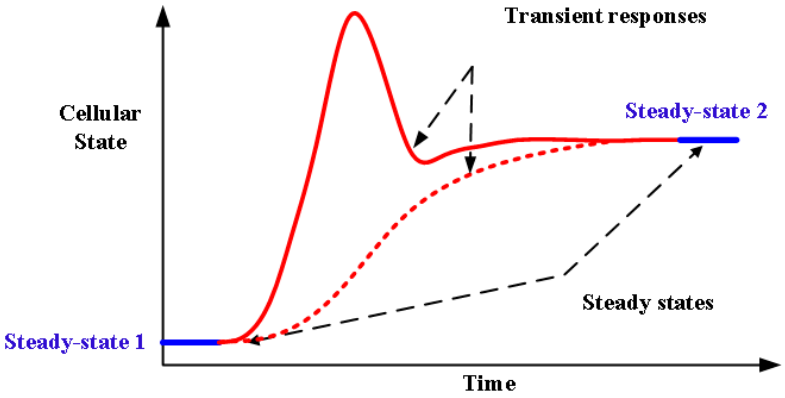

Figure 3. Example of a cellular state transitioning between two steady states via two different transient responses where the steady states alone cannot reveal the transient response.

3.1. Dynamic Analysis of Genomics Data

In the following, we first use a specific example to illustrate the point discussed at the beginning of this section. Then we review similar work that has been carried out recently in advancing dynamic analysis of gene regulatory networks. We recently investigated key gene regulatory mechanisms that *Scheffersomyces stipitis* (an important yeast in biorenewables) utilizes to cope with oxygen limitation during a transition from aerobic growth to oxygen-limited fermentation. Guided by the principles discussed above, we designed experiments to obtain its transcriptomic profiles during the transition from aerobic growth to oxygen-limited fermentation, and analyzed the dynamic transcriptomic data to gain better understanding on its cellular metabolism. Two samples were collected during a controlled chemostat of aerobic growth. Eighteen samples were collected at 10-minute intervals for three hours during the transition period. Two samples were collected at the new, oxygen-limited steady state. Biomass was maintained at a relatively constant level by adjusting dilution rate in order to keep the cellular properties comparable across different conditions [116]. The experiment was repeated four times. Therefore, 22 cell samples were collected for each experiment, resulting in 88 samples.

Here, we present several key findings from the study. More details on the experiments, data collection, data pre-processing, and data analyses can be found in [110]. First, we performed principal component analysis (PCA) to analyze the transcripts per million (TPM) value for all 88 samples (22 samples per trial). The first two principal components (PCs) captured about 80% of variation amongst all 88 samples,

indicating highly correlated responses among all genes. The scatter plots of these first 2 PC scores are plotted in Figure 4. The circles and stars indicate the initial aerobic and final oxygen-limited steady states (SS), respectively, while the triangles indicate the transition states. Figure 4 clearly shows that, despite the fact that the four trials follow a somewhat different trajectory from each other, their aerobic and oxygen-limited steady states are clustered closely together, indicating similar initial and final states of the dynamic transition process. The different transient trajectories of the four trials is likely due to the following reason: (1) the four trials are not truly biological replicates starting with the same inoculum; (2) similar but not exactly the same conditions across trials (e.g., cell density, dilution rate), and (3) stochastic nature of gene regulatory events. One important observation of Figure 4 is that, compared to the overall gene expression difference between the two steady states, the transient states of the gene expressions show significantly higher variations. The observation suggests that if only the two steady states were considered, key information on regulatory mechanism that govern the transition between the two steady states could be missed. This is illustrated by the time-series plots of the TPM of two genes PICST 66442 (Figure 5a) and PICST 76518 (Figure 5b). Figure 5a shows the gene PICST 66442 was significantly down-regulated during the transition, then returns to a final expression level not far from its initial level. Despite the variations among trials, the overall trends are consistent across trials. Similarly, Figure 5b shows the gene PICST 76518 was significantly upregulated during the initial stage of the transition right after the disturbance was introduced, then gradually returns to a final expression level very close to its initial level. Again, despite the variations among trials, the overall trends are consistent across trials. If only the steady state expression levels were measured and fold change were used to detect differentially expressed genes, both genes would have been missed. Further analysis shows that more than half of the differentially expressed genes show changes during the transition period only [110]. If experiments and analysis were performed on the two steady states only, these genes would be completely missed, and any gene regulatory mechanisms derived from this incomplete gene list would surely be an incorrect or incomplete picture of the true regulation. Even for the genes that show differential expression at the new steady state, many of them experienced much larger or reverse changes during the transient states compared to the new steady state. Again, if only the transcriptome profiles at the two steady states were compared, important information would have been missed. Finally, by integrating the analysis results obtained from dynamic transcriptomic data and the cultivation data with genome-scale modelling [117], we were able to identify potential short- and long-term strategies that the cells utilize to cope with oxygen limitation. More details can be found in [109,110,117].

This study highlights the importance of studying the dynamic transient response of RNA-Seq data in order to understand the global response triggered by the perturbation and potential gene-regulatory mechanism. While the new steady state represents the new phenotype expressed by the cells, it is likely a direct result of the path taken during the transition period. In other words, the genes differentially expressed during the transition period, can have significant influence on the final steady state. Therefore, it is of significant interest to identify the key genes affected during the transition period and discover their roles in regulating the final steady state phenotype. Our study also showed that for the genes that dominate the transition period, most of them returned to the proximity of the initial steady state (i.e., aerobic growth), and such recovery is not growth associated. The closely related initial and final steady states suggest that the steady state gene expression profile might be a highly conserved optimal state which could enable cell's fast response to different disturbances, analogous to the ready position of many sports such as tennis and baseball [110].

There are other examples that highlight the importance of dynamic analysis of gene expression data. For example, McDowell et al. [118] proposed an infinite Gaussian process mixture model for clustering gene expression time-series data. Using time-series microarray gene expression data, Cheng et al. [119] evaluated various machine learning techniques, including logistic regression (LR), random forest (RF) and support vector machine (SVM), to predict cell cycle-regulated genes in budding yeast. Bar-Joseph et al. [120] reviewed some early work concerning the basic experimental

considerations and computational methods for representing, clustering and classifying time-series gene expression data. Other applications of dynamic gene expression data analysis have enabled insights into cell response to environmental stress [121], peripheral circadian gene regulation in mouse liver and heart [122], and genes periodically expressed in tumors [123]. As experimental and computational challenges being adequately addressed, generating and analyzing time-series expression data has become one of the most fundamental methods for understanding a variety of biological processes.

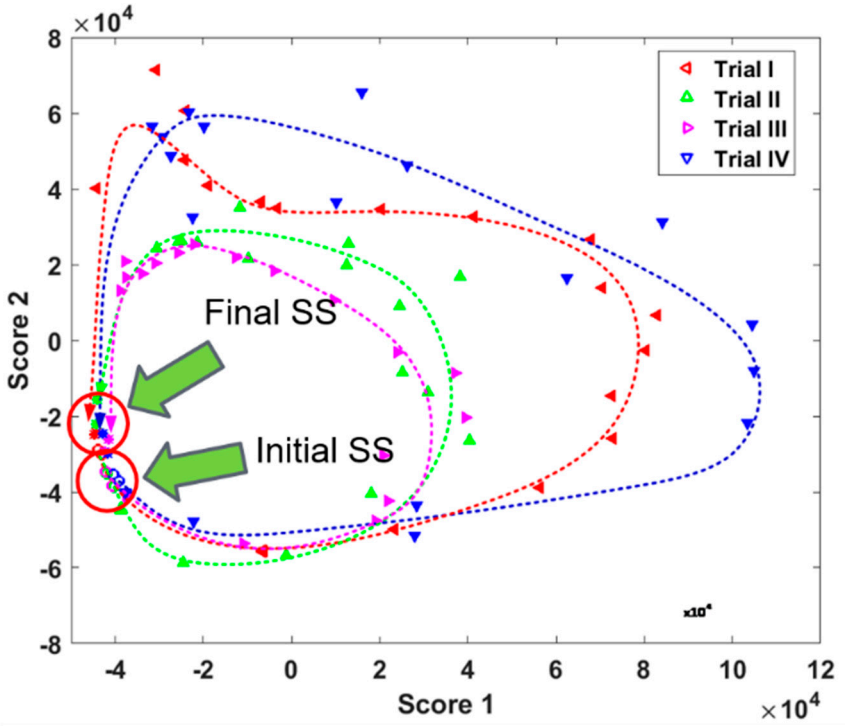

Figure 4. Score plot of TPM values of all four trials. Circles and stars denote initial (i.e., aerobic) and final (i.e., oxygen-limited) steady states (SS), respectively; triangles denote transition states. Samples in total. These samples were analyzed using Illumina Next-Generation Sequencing (Illumina Inc., San Diego, CA, USA).

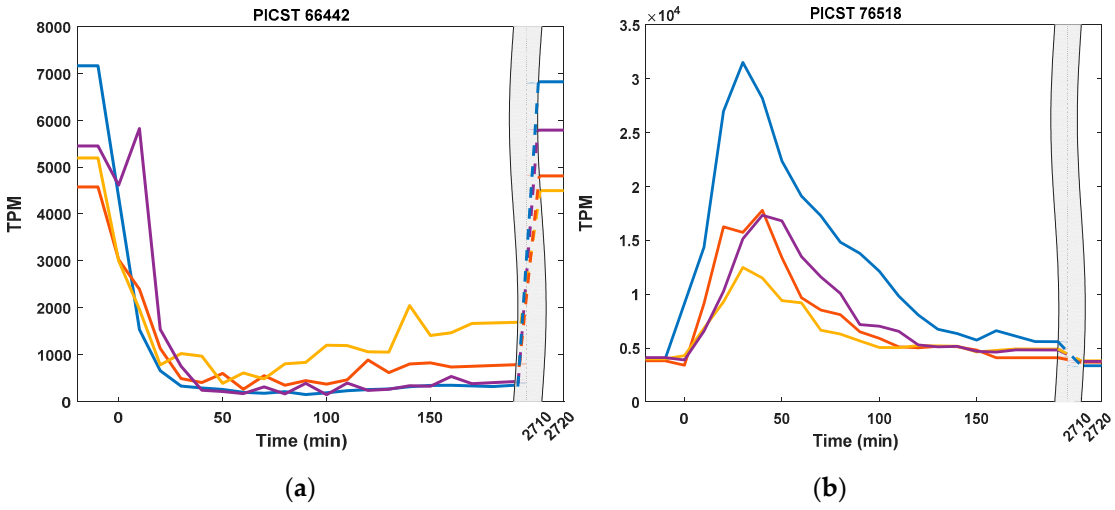

Figure 5. Time-series data for gene PICST 66442 (**a**) and PICST 76518 (**b**) show significant changes during the transition period only. The final oxygen limited fermentation steady state expression levels (i.e., the last two points) are not far from their initial aerobic growth steady state levels (i.e., the first two points).

3.2. Dynamic Metabolic Flux Analysis

Metabolic networks enable the conversion between small biochemical molecules (i.e., metabolites) to provide essential building blocks and energy that are critical for cell function and growth. Mathematical analysis of flux distribution in a metabolic network can guide the metabolic engineering process. Metabolic flux analysis (MFA) and flux balance analysis (FBA) were initially proposed based on balancing fluxes around intracellular metabolites within an assumed network stoichiometry, where external rate measurements such as carbon uptake rate, biomass growth rate were used as the constraints [124,125]. The major difference between MFA and FBA is that MFA is to minimize error in predicted and measured external metabolic rates (e.g., minimizing variance weighted sum of squared residuals (SSR) through least square regression) using stoichiometry of a small-scale or simplified model as constraints, while FBA maximizes an objective function (e.g., cell growth) using stoichiometry of a large-scale (e.g., genome-scale) model as constraints [126]. Since FBA aims to quantify fluxes in often underdetermined systems, additional constraints such as upper and lower bounds of fluxes are needed. ^{13}C-based MFA (^{13}C-MFA) is a more advanced MFA that makes use of ^{13}C-labeled tracers, combined with isotopomer balancing, metabolite balancing, and isotopic labeling measurements, to estimate fluxes [99,127–130]. In addition, flux variability analysis (FVA) has been proposed to find the range of flux variability for reactions in the metabolic network while achieving optimal or suboptimal objective states [131]. In early studies, due to the lack of in vivo intracellular measurements, the system under study is often required to be maintained at steady state. In other words, MFA and FBA were typically used to determine the metabolic flux of a system at a particular steady state. However, in 'non-laboratory' biotechnological conditions, such as large-scale batch or semi-batch cultivations or fermentations, the systems, and hence the cellular metabolic networks, are rarely at steady state. Therefore, the dynamics of these metabolic networks have to be studied, which lead to the development of dynamic metabolic flux analysis (DMFA) [98,132–134], ^{13}C-DMFA [135], and dynamic flux balance analysis (DFBA) [136–138] in recent years. It is worth noting that many of these approaches, especially the dynamic extensions, were proposed and further improved by researchers from the systems engineering community. For more thorough discussions on recent advances and remaining challenges of this rapidly developing field, the interested readers are referred to several review papers [99,116,126,139].

3.3. Dynamic Analysis of Signal Transduction Networks

Signal transduction pathways regulate many cellular processes (e.g., gene expression for the target proteins) and are also involved in extracellular communications. Understanding signal transduction mechanisms has many potential applications such as improved disease treatment. Modelling of signal transduction networks can be based on biological understanding of the molecular mechanisms involved [140]. Alternatively, data-driven models can be constructed [141]. As dynamic behavior of some proteins, such as transcription factors, have a direct effect on the response of a cell to a stimulus, analyzing the steady state behavior alone is insufficient for characterizing the response [142]. Here we focus on dynamic analysis in helping understanding of signal transduction networks. As cellular signals are not static but dynamic [143,144], systems engineering has had a pervasive influence on the understanding of their behavior such as chemotaxis [145,146], robustness [147,148], and control mechanisms [102,149–152]. Compared to experimental data for metabolic and gene regulatory networks, there have been lower amounts of quantitative data available on signal transduction networks. Furthermore, information about transient dynamics is required for signal transduction pathways whereas steady state analysis is extensively implemented for metabolic and gene regulatory networks [142]. The increasing availability of high-throughput and multiplex techniques for quantifying signaling and cellular responses have enabled large-scale quantitative studies of signal transduction networks. However, these data have been found difficult to be understood completely by inspection and intuition. Data-driven statistical modelling approaches such as principal components analysis (PCA), partial least squares (PLS), and systems identification techniques have been found useful

in deriving biological insights from large-scale experiments [141]. These models are emerging as standard tools for systems-level research in signaling networks. For example, Gadkar et al. [153,154] proposed an iterative approach with a state regulator algorithm for model identification of signal transduction networks from data. Another issue that has drawn attention from the systems engineering community is the parameter estimation in the modeling of signal transduction networks. As these models commonly contain a large number of parameters, while there is usually limited amount of data available for their estimation, regularization techniques are require to avoid overfitting. Howsmon and Hahn [82] provided a tutorial of commonly used regularization techniques and demonstrated their effectiveness on an interleukin-6 (IL-6) signaling network.

3.4. Integrated Dynamic Analysis of Multi-Omics Data

There are emerging studies that investigate the integrated dynamic analysis of multi-omics data. For example, personalized medicine is expected to benefit from combining genomic information with regular monitoring of physiological states by multiple high-throughput methods. To this end, Chen et al. [155] presented an integrative personal omics profile (iPOP) analysis that combines genomic, transcriptomic, proteomic, metabolomic, and autoantibody profiles from a single individual over a 14 month period. The iPOP analysis revealed various medical risks, including type 2 diabetes. It also uncovered extensive, dynamic changes in diverse molecular components and biological pathways across healthy and diseased conditions. Mias et al. [156] proposed a mathematical tool to integrate multiple omics information arising from dynamic profiling in a personalized medicine approach. Nakanishi et al. [157] used dynamic omics approaches, including time-lapse 2D-nuclear magnetic resonance (NMR) metabolic profiling, transcriptomic, and proteomic analyses, to identify nutrition-mediated microbial interactions in a minimum microbial ecosystem. Przytycka et al. [158] reviewed some recent computational attempts to obtain a dynamic view of the interactome. It is anticipated that incorporating multiple data sources, resulting from advances in mass spectrometry, next-generation sequencing and other high-throughput experimental methods, will help to further elucidate protein interactions and network dynamics.

3.5. Other Applications of Dynamic Data Analysis

While this review focuses more on high-throughput omics data analysis, there are other applications of dynamic data analysis. For example, Zeger et al. [159] reviewed time-series analysis methods, in both time and frequency domains, applied in public health and biomedical applications. Dynamic shape constrained splines (SCS) were proposed as transparent black-box models for bioprocess modeling. It was demonstrated that the dynamic SCS rate laws exhibit the flexibility of typical data-driven black-box models, while offering a transparent interpretation akin to conventionally applied rate laws such as Monod and Haldane. In another study [53], to identify temporal proteomic profiles of chronic periodontitis, a number of filters including temporal pattern matching, logistic function fitting and mixed-integer linear optimization were applied to temporal proteomic profiles to identify a small subset of proteins from a large set containing substantial number of proteins. Finally, dynamics are taken into account in classification and diagnosis of bioprocess cell growth productions using early-stage data [65]. Transient response of heat-shock (HS) response in *E. coli* has also been studied using systems engineering control theories [160,161].

3.6. Summary and Discussion

It is now well recognized that biological systems and their cellular networks should be viewed dynamically. In other words, they are networks of interacting components that change with time in response to external and internal events. Therefore, studying the dynamic behavior of these networks is the basis for an understanding of cellular functions and disease mechanisms. However, obtaining genome-scale measurements (e.g., abundance of RNAs, proteins, and metabolites) can be costly and laborious. Time-series data required by dynamic analysis are even more costly and

laborious to obtain than steady state data used in steady state studies. There is no obvious consensus regarding design of experiment to generate time-series data. Nevertheless, there are few discussions on the relevant technical considerations to generate reproducible, statistically sound, and broadly useful genome-scale data [162]. For example, for oscillatory or periodic processes such as the mammalian circadian system, process dynamics have to be considered. In this case, sampling rate or frequency is crucial. For genome-scale analysis of biological rhythms, it is recommended to sample at least 12 time points per cycle across two full cycles to optimize statistical power [162]. To overcome the cost and/or technical challenges of obtaining sufficient time-course data, data augmentation has been implemented in some studies to artificially increase dynamic or time-series measurements. In this case, the limited time-series data are smoothed and fitted by an appropriate linear, spline or polynomial function, and more frequent data points can be obtained from the fitted line or curve to increase temporal resolution [98,99].

4. The Role of Domain Knowledge in Biological Data Analytics

Rapid advancements in computing power, assisted by recent algorithmic breakthroughs in machine learning (ML), especially deep learning (DL), have led to the explosion of artificial intelligence (AI) applications. High-profile examples include AlphaGo and AlphaGo Zero from Google and the associated high-profile publications in *Nature* with titles such as "mastering the game of go without human knowledge" [163,164]. These well-publicized breakthroughs have led to publications and internet posts where authors claim that anything can be inferred by detecting patterns within huge databases, and there is no point of modelling anymore. This extreme stance is summarized in Anderson's provocative statement published in *Wired* magazine: "The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all" [165]. Alarmed by these provocative statements, there have been several important papers to caution the funding and promotion of "blind" big data projects and provided evidence that the successful use of big data in many applications depends on more than the quantity of data alone and are skeptical that a purely data-driven approach—'blind big data'—can deliver the high expectations of some of its most passionate proponents [166,167]. They also include ample examples to demonstrate that, instead of rendering theory, modeling and simulation obsolete, big data should and will ultimately be used to complement and enhance them [167]. These viewpoints are consistent with systems engineering principles.

Systems engineers usually work within a specific domain such as chemical, mechanical and electrical engineering, who utilize processes and methods that are tailored to their domain's unique problems, constraints, risks and opportunities. Therefore, domain knowledge has always been an integral part of systems engineering principles and techniques. This is in drastic contrast to other purely data-driven or data-centric approaches where the analysis without domain knowledge is touted as a significant advantage of these approaches. Here, we express our strong disagreement with these purely data-driven approaches when applied to analyze high dimensional biological data such as genome-scale omics data. The single most important reason is that the extreme complexity of the system or process in conjunction with severely incomplete knowledge of the system/process lead to models with very high degree of freedom. A good agreement between model predictions and experimental measurements is not sufficient to guarantee the predictive capability of a model because it is not difficult to change a few parameters and/or constraints among hundreds or even thousands to match the usually small number of experimental measurements for post hoc explanations. As illustrated in the next section, point-matching alone is not a reliable approach for validating a model with high degree of freedom. If a model offers no structural explanations of the correlations it reveals, i.e., matching the known understanding of the system, many of these correlations are likely to be false positive [166,168]. In this section, we review studies where knowledge has played critical role in model validation, unsupervised and supervised learning, and feature engineering and selection.

4.1. Knowledge Matching vs. Point Matching for Model Validation

The use of computational models for hypothesis testing has long been recognized. For example, Kitano, an expert in artificial intelligence and robotics and one of the earliest pioneers of systems biology, divides computational biology into two distinct branches: knowledge discovery, or data mining, which extracts the hidden patterns from huge quantities of experimental data, forming hypotheses as a result; and simulation-based analysis, which tests hypotheses with in silico experiments, providing predictions to be tested by in vitro and in vivo studies [105,106]. While data mining or machine learning-based knowledge discovery has been widely utilized, simulation-based hypothesis testing or model validation has not been fully explored.

The conventional approach for biological model validation is to compare model predictions with experimental data under different conditions [117,169,170]. For example, for genome-scale metabolic network model (GEM) validation, most often the experimental data consist of measured cross-membrane fluxes, i.e., various substrate uptake rates, product excretion rates, and cell growth rate. Such a validation approach is deemed as the gold standard for evaluating the quality of a GEM. We term these approaches "point-matching" approaches because each experimental condition represents a single point (although potentially high dimensional) in the phenotypic space. For well-characterized organisms, point-matching approaches work well, because their metabolic network structures have been well studied and defined. However, given the fact that a GEM, especially a less studied one, is usually severely underdetermined (i.e., with high degree of freedom), matching numerical experimental data over a few limited conditions does not necessarily indicate a high-quality GEM and can result in very misleading conclusions. This was clearly demonstrated in our recent study on the evaluation of the two GEMs of *S. stipitis* iSS884 and iBB814 [170]. In that study, although iSS884 consistently showed much better agreement with experimental measurements than iBB814 across multiple data sets in terms of predicting product secretion rates, its predictions for several mutant strains are incorrect. Such a lack of predictability suggests that model iSS884 contains internal errors that were not revealed by point matching.

To address the shortcomings of point-matching validation approaches, we proposed a system identification (SID)-based framework for GEM validation. In the SID framework, biological knowledge embedded in a GEM is first extracted from a series of designed in silico experiments through multivariate analysis methods such as principal component analysis (PCA); next, the extracted knowledge, such as how cells respond under a given stimulus, is visualized and compared with the existing knowledge for model validation and analysis [117,171]. We term the proposed approach "knowledge-matching" as the simulation results are not directly compared with experimental data; instead, the knowledge captured by the model is compared with available knowledge. Although rooted in simulations, the SID-based approach is more of a qualitative validation, instead of a quantitative approach, and offers additional robustness against measurement errors. In [117], through the knowledge-matching based validation, we have shown that although iSS884 exhibited much better agreement with experimentally measured cross-membrane fluxes, it contains some significant errors. On the other hand, although iBB814 shows poorer performance in quantitative point-matching validations, it captures the knowledge that aligns better with existing knowledge on *S. stipitis* [117,170].

The work by Somvanshi et al. [74] is another example of knowledge matching where the model-based observations are compared with the interplay of oxidative stress, inflammation, insulin resistance, and energy deficit observed from the data. In addition, the trends predicted by the model are qualitatively compared to the observed physiological responses reported in the literature. Zomorrodi et al. [139] also paid special attention to knowledge matching when it comes to optimizing metabolic models, where it recommends against conditional modifications to resolve an inconsistency in prediction.

4.2. Knowledge-Guided Unsupervised Learning

Clustering methods are the most commonly utilized unsupervised learning methods for analyzing biological data such as gene expression [172,173]. In pure data-driven clustering, the algorithms focus on mathematical similarity of genes and conditions, while the biological meanings of the clusters are neglected. As a result, genes are usually clustered into disjoint groups, which for example, do not capture the biological fact that many gene products participate in more than one biological process [174]. To overcome these limitations, many biological knowledge-guided clustering methods have been proposed in the past decade. For example, Fang et al. [174] proposed a knowledge-guided clustering method for microarray expression data clustering based on gene ontology (GO). It is shown that the GO knowledge-guided clustering is more advantageous than pure data-driven clustering in both the quality of clusters and the precision of biological annotations. In a more recent study, a multi-objective clustering algorithm guided by a priori biological knowledge was proposed to find clusters of genes with high levels of co-expression, biological coherence, and also good compactness and separation [175]. In the proposed method, cluster quality indexes are used to simultaneously optimize gene relationships at expression level and biological functionality. Specifically, the information of genetic elements regarding their levels of expression and biological functions was integrated into clustering to optimize the biology- and expression-based distances. In another study, Yang et al. [176] proposed to improve clustering of microRNA microarray data by incorporating functional similarity. The clustering of microRNA expression profiles was improved by incorporating the gene ontology information of the target genes of miRNAs to obtain more functionally compact clusters, which benefits the identification of potential miRNA biomarkers and the construction of miRNA co-regulation networks. Schwaber et al. [177] also proposed a knowledge-based clustering, where intelligent software agents are implemented to gather knowledge automatically from pathway databases for clustering.

Besides clustering, domain knowledge has been used to enhance other unsupervised machine learning, such as widely used PCA for dimension reduction or feature extraction, where regular PCA can fail to accurately capture the characteristics of a biology system or process. For example, regular PCA assumes a multivariate normal distribution. However, studies have demonstrated that microarray gene expression measurements follow a super-Gaussian distribution instead [178]. In addition, PCA decomposes the data based on the maximization of its variance. However, in some cases, the biological question may not be related to the highest variance in the data [179,180]. To address these limitations, various variations of knowledge-guided PCA have been proposed, such as ontology-guided PCA [181], knowledge-guided gene ranking by coordinative component analysis [182]. In addition, maximum likelihood PCA (MLPCA) has been proposed that can integrate process knowledge into PCA such as measurement errors, nonstationary behavior, and heterogeneous variances and correlations [183–186]. There are other knowledge-guided or enhanced unsupervised learning techniques, the interested readers are referred to two recent review papers [187,188].

4.3. Knowledge-Guided Supervised Learning

Domain knowledge can also help guide supervised machine learning. In a recent work [189], we examined whether the so-called focused factory theory (i.e., factories that concentrate on narrow range of services or operations produce better products at low costs) is applicable to hospital operations. Specifically, we examine whether the hospitals that are specialized in certain diseases achieve better results in terms of costs and patient outcomes using a large national healthcare cost and utilization project (HCUP) dataset. Pure data-driven ML approaches based on multiple linear regression (MLR), principal component regression (PCR), partial least squares (PLS), Fisher discriminant analysis (FDA), ordinal regression (OR) and logistic regression (LR) were used to investigate the effects of hospital specialization on hospital performance in terms of cost (measured by total charge) and patient outcome (measured by death of patient during hospitalization). We show that, without domain knowledge, pure data-driven ML approaches only reveal direct effect of specialization on total charge, as illustrated in Figure 6a. To address the limitation, we propose a knowledge-guided ML approach by defining

model structures based on domain knowledge and hypothesis. The results from knowledge-guided ML indicate that specialization reduces total charge both directly and indirectly (i.e., through reducing number of diagnosis and length of hospital stay) as illustrated in Figure 6b, which is consistent with the common belief that operational efficiency and physician effectiveness are two important factors determining the total charge.

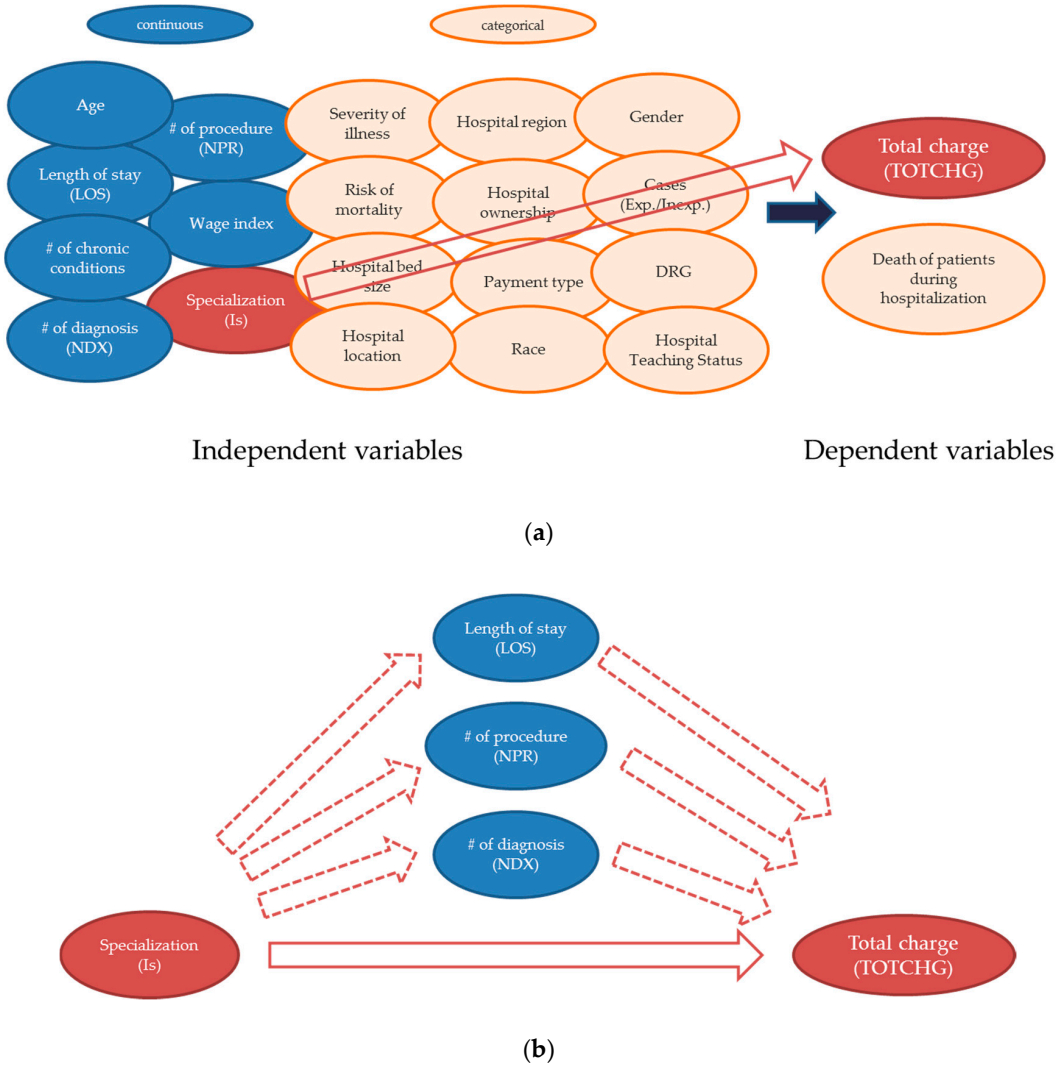

Figure 6. (**a**) Pure data-driven ML models only reveal direct effects between independent variables and dependent variables. (**b**) Domain knowledge helps identify three independent variables (i.e., LOS, NPR and NDX) that mediate the indirect effects between Is and TOTCHG. Although pure data-driven methods can quantify the correlations among independent variables, they cannot reveal the causal relationships such as mediation.

There are a variety of other studies where domain knowledge has been explicitly used to guide supervised learning. For example, Shen et al. [190] propose a knowledge-guided bioinformatics model for identifying autism spectrum disorder diagnostic microRNA biomarkers. An autism-specific miRNA–mRNA network was constructed, and candidate autism biomarker miRNAs were inferred based on their regulatory modes and functions. Hvidsten et al. [191] introduced a supervised clustering methodology for functional classification of gene expressions from microarray hybridization experiments. The methodology is different from the commonly used unsupervised clustering approaches in that it exploits background knowledge of gene function in a supervised manner. Genes are annotated using Ashburner's Gene Ontology and the functional classes used for learning are mined

from these annotations [191]. In another study, to achieve accurate prostate segmentation, which is necessary for maximizing the effectiveness of radiation therapy of prostate cancer, Park et al. [192] propose a semi-supervised learning approach that utilizes a priori knowledge of training data and domain knowledge through user interactions. More knowledge-guided supervised learning techniques and their applications to biological big data can be found in some recent review papers [193,194].

4.4. Knowledge-Guided Feature Engineering and Feature Selection

Besides direct utilization, domain knowledge has been extensively used in feature engineering and selection to indirectly enhance statistical and machine learning. From PSE perspective, what features or variables to be measured and used as inputs is the very first step in developing a model. For example, in a multi-omic biomarker identification and validation study for diagnosing warzone-related post-traumatic stress disorder [54], domain area expertise of multiple researchers were utilized in conjunction with multiple machine learning approaches for drastically reducing over a million features or biomarkers (molecular datasets from GWAS, DNA methylation, miRNA, protein, metabolites, small molecules, endocrine, clinical labs, and biometrics/physiological data) to 343 before further refinement to 28.

Knowledge-guided feature engineering and selection is particularly important for mining clinical data. For these application, it has been recognized that without using predictors (i.e., features) rooted in (patho)physiology for clinical prediction, there is a risk of capturing a factor that may change at any time, invalidating a model that appears to produce computationally reproducible results [195]. This is supported by our recent study showing that knowledge-guided feature selection methods are significantly less sensitive to the training samples used for feature selection [196]. Specifically, we show that, unlike existing data-driven feature/variable selection methods, consistency-enhanced evolution for variable selection (CEEVS) is able to identify the underlying chemical information of a sample from its NIR spectrum, i.e., the wavelengths corresponding to the chemical bounds or functional groups that determine the sample properties of interest. We show that the knowledge-associated features are significantly less sensitive to the training samples, which in turn significantly improves model accuracy and robustness.

Knowledge-guided feature engineering and selection has also drawn increased attention in computer vision, including medical image analysis, in recent years. For example, approaches have been proposed to mimic the human intelligence capabilities by combining prior sematic and contextual knowledge and visual information in knowledge based vision systems, which is a step towards cognitive vision. [197]. Several studies [197–200], including automated prostate segmentation in magnetic resonance imaging (MRI) images and colon segmentation in computed tomography (CT) colonography images, have shown that knowledge based vision systems resulted in increased accuracy, and were more robust and less dependent on data.

There are other studies where domain knowledge has played critical role in feature engineering and feature selection. For example, in the study of arrayCGH [65,66], the features (i.e., the DNA copy numbers along the genome) have the natural spatial order, and incorporating the structure information using an extension of the $\ell 1$-norm outperforms the Lasso in both classification and feature selection. In other studies, Garla and Brandt [201] presented a feature engineering technique, namely ontology-guided feature engineering, that leverages the biomedical domain knowledge encoded in the Unified Medical Language System (UMLS) to improve machine-learning based clinical text classification. Yao et al. [202] proposed a knowledge-guided feature engineering and deep learning models for effective disease classification. For more discussions in this area, interested readers are referred to several review papers [48,195,197].

4.5. Summary and Discussion

In summary, there are concerns about the recent overhype of deep learning without domain knowledge or theory. Although there are applications where domain knowledge may have appeared

to be playing a lesser role, we demonstrated with ample examples that, for analyzing biological big data, domain knowledge has been and will continue to be playing a significant role. Studies have also shown that domain knowledge is particularly useful when developing and validating highly flexible models (i.e., models with high degree of freedom) that are often underdetermined. In these cases, domain knowledge or theory can help select an appropriate model structure, serve as constraints to ensure qualitatively correct model behavior, or guide the development of biologically meaningful features/predictors. Specifically, we show that the gold standard of point-matching has severe limitations when it comes to validating biological models with a high degree of freedom. In contrast, knowledge-matching is a much more reliable alternative. Studies have also shown that knowledge-guided unsupervised and supervised learning usually performs significantly better than their pure data-driven counterparts. In these knowledge-guided approaches, domain knowledge or theory is usually used as a guide for model selection and simplification, or serves as constraints, or incorporated into a regularization term on model parameters. Alternatively, domain knowledge can also assist feature engineering for developing features that are rooted in biological mechanisms such as (patho)physiology for biomedical and clinical applications. Last but not least, data-driven machine learning models are often tuned based on, and evaluated by, some numerical metrics (e.g., mean-square-error, sensitivity, specificity, AUC, etc.), which could lead to suboptimal or even misleading results. For example in biomedical and clinical research, without domain knowledge, classification errors (i.e., false positive or type I error, and false negative or type II error) are often treated the same without taking into account the often different clinical and economic consequences of these two different types of errors [203,204]. Therefore, we argue that domain knowledge plays important roles at every stage of biological data analytics, and should be incorporated into decision making wherever possible.

5. Conclusions

The discipline of process systems engineering has contributed significantly to the field of biological big data analytics. In this work, we reviewed and discussed the application of systems engineering principles and techniques in addressing some of the most important challenges in big data analytics for biological, biomedical and healthcare applications. The systems engineering principles emphasize parsimonious modeling. In this work, we discussed direct and indirect means of checking for overfitting, and reducing the risk of overfitting through the principle of parsimony. We classified strategies of reducing overfitting into three categories: (1) reducing parameter space; (2) reducing feature space; and (3) increasing sample space. Reducing parameter space is a direct and top-down approach for reducing overfitting. It starts with the selection of a model (structure) with a small number of parameters or imposing regularizations on model parameters without explicitly evaluating the features. The selected models are often based on knowledge from domain expertise or prior studies. The data are used for estimating model parameters, as well as for model validation. The second strategy is to reduce feature space, which is an indirect and bottom-up approach for reducing overfitting. In this strategy, experimental data are used to evaluate the importance of features. This step is often carried out without the involvement of domain knowledge other than an objective function to be optimized. The goal is to retain only the important features (through feature reduction or combination) in the final prediction or classification models, which indirectly, but effectively, reduces model parameters. The third strategy is to increase sample space through data augmentation, which is often a complementary, rather than stand-alone, approach to the first two strategies. When implemented properly, data augmentation can improve the accuracy and precision of model parameter estimation and enhance the performance of feature selection or extraction. In particular, this strategy can be very effective in coping with small datasets and/or datasets with limited labeled samples.

The systems engineering principles also promote dynamic analysis of biological data, where cellular systems are treated as networks of interacting components that change with time in response to external and internal events. It has been gradually recognized that studying the dynamic behavior of these

networks ought to be the basis for understanding cellular functions and disease mechanisms. In this work, we reviewed the progress made towards dynamic analysis of gene regulatory networks, metabolic networks, and signal transduction networks. Progress has also been made in the integrated dynamic analysis of these networks simultaneously. For gene regulatory networks, several recent studies have highlighted the importance of studying the dynamic transient response of gene expression data in order to understand the global response triggered by various perturbations, and to gain knowledge on potential gene regulatory mechanisms. For metabolic networks, mathematical analysis of metabolite flux distribution in a metabolic network can guide the metabolic engineering process. Since cellular metabolic networks are rarely at steady state under non-laboratory conditions, the dynamics of these metabolic networks have to be studied for biotechnology development. This has led to the rapid development of dynamic approaches such as dynamic metabolic flux analysis and dynamic flux balance analysis. Similarly, for signal transduction networks, the cellular signals are not static but dynamic. Therefore, analyzing the steady-state behavior alone is insufficient for characterizing the response. There are also emerging studies that investigate the integrated dynamic analysis of multi-omics data. Finally, there are other applications of dynamic data analysis, such as in public health and biomedical applications. As experimental and computational challenges being adequately addressed, generating and analyzing time-series expression data has become one of the most fundamental methods for understanding a variety of biological processes. One outstanding challenge in this area is that obtaining time-series genome-scale measurements is still costly and laborious. In addition, more studies and standards/protocols are needed for technical considerations related to data acquisition such as experimental design and data reproducibility.

Finally, the systems engineering principles advocate the integration of domain knowledge with data-driven machine learning in analyzing biological big data. In this work, we reviewed the role of domain knowledge in model validation, unsupervised and supervised machine learning, and feature engineering. For model validation, many studies, including our own, have demonstrated advantages of knowledge matching in model validation compared to the conventional point matching approach. This is particularly important when we attempt to validate complex models with very high degree of freedoms, such as genome-scale metabolic network models. We also reviewed studies that integrate biological or (patho)physiological knowledge with unsupervised learning. Generally speaking, domain knowledge has been found instrumental in the success of many of these applications. There are also various studies where domain knowledge has been used explicitly to guide supervised learning, particularly in determining appropriate model structures. Besides direct utilization, domain knowledge has been extensively used in feature engineering and selection to indirectly enhance machine learning. Many studies, including our own, have demonstrated that knowledge-associated features are more robust and less dependent on data (i.e., less sensitive to the training samples), which in turn significantly improves model accuracy and robustness. This is particularly important for clinical data modeling where predictive features rooted in (patho)physiology are important for model interpretation and clinical prediction.

Similar to the field of systems biology where researchers from PSE community have been integral and essential components of many interdisciplinary research teams, PSE community has played important role in the field of biological big data analytics. Despite many challenges, we envision that synergistic collaborations between researchers from PSE community and biologists, oncologists and physicians will continue to evolve and open new research avenues to address important questions in biology, life sciences, and healthcare.

Author Contributions: Conceptualization, Q.P.H. and J.W.; writing—original draft preparation, J.W.; writing—review and editing, Q.P.H.; supervision, Q.P.H.; project administration, Q.P.H. Both authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Science Foundation, grant number NSF-CBET #1805950, and U.S. Department of Energy, grant number DE-SC0019181.

Acknowledgments: The authors would like to thank the financial support from National Science Foundation under the grant NSF-CBET #1805950 and U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomic Science Program under Award Number DE-SC0019181.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zikopoulos, P.; Eaton, C. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*; McGraw-Hill Osborne Media: New York, NY, USA, 2011.
2. Zikopoulos, P.C.; Deroos, D.; Parasuraman, K. *Harness the Power of Big Data: The IBM Big Data Platform*; McGraw-Hill: New York, NY, USA, 2013; ISBN 0071808183.
3. Yang, L.T.; Chen, J. Special Issue on Scalable Computing for Big Data. *Big Data Res.* **2014**, *100*, 2–3. [CrossRef]
4. Liang, T.-P.; Guo, X.; Shen, K. Special Issue: Big data analytics for business intelligence. *Expert Syst. Appl.* **2018**, *111*, 1. [CrossRef]
5. Martínez–Álvarez, F.; Morales–Esteban, A. Big data and natural disasters: New approaches for spatial and temporal massive data analysis. *Comput. Geosci.* **2019**, *129*, 38–39. [CrossRef]
6. Bassi, S. A primer on python for life science researchers. *PLoS Comput. Biol.* **2007**, *3*, e199. [CrossRef] [PubMed]
7. Ekmekci, B.; McAnany, C.E.; Mura, C. An introduction to programming for bioscientists: A Python-based primer. *PLoS Comput. Biol.* **2016**, *12*, e1004867. [CrossRef] [PubMed]
8. Charalampopoulos, I. The R Language as a Tool for Biometeorological Research. *Atmosphere* **2020**, *11*, 682. [CrossRef]
9. Peng, R.D. Reproducible research and biostatistics. *Biostatistics* **2009**, *10*, 405–408. [CrossRef]
10. Peng, R.D. Reproducible research in computational science. *Science (80-)* **2011**, *334*, 1226–1227. [CrossRef]
11. Stodden, V. Reproducible research: Tools and strategies for scientific computing. *Comput. Sci. Eng.* **2012**, *14*, 11–12. [CrossRef]
12. Mittelstadt, B.D.; Floridi, L. The ethics of big data: Current and foreseeable issues in biomedical contexts. *Sci. Eng. Ethics* **2016**, *22*, 303–341. [CrossRef]
13. Raghupathi, W.; Raghupathi, V. Big data analytics in healthcare: Promise and potential. *Heal. Inf. Sci. Syst.* **2014**, *2*, 3. [CrossRef]
14. Feldman, B.; Martin, E.M.; Skotnes, T. Big data in healthcare hype and hope. *Dr. Bonnie 360* **2012**, 122–125.
15. Mehta, N.; Pandit, A. Concurrence of big data analytics and healthcare: A systematic review. *Int. J. Med. Inform.* **2018**, *114*, 57–65. [CrossRef]
16. Senthilkumar, S.A.; Rai, B.K.; Meshram, A.A.; Gunasekaran, A.; Chandrakumarmangalam, S. Big data in healthcare management: A review of literature. *Am. J. Theor. Appl. Bus.* **2018**, *4*, 57–69.
17. Alyass, A.; Turcotte, M.; Meyre, D. From big data analysis to personalized medicine for all: Challenges and opportunities. *BMC Med. Genomics* **2015**, *8*, 33. [CrossRef]
18. Luo, J.; Wu, M.; Gopukumar, D.; Zhao, Y. Big data application in biomedical research and health care: A literature review. *Biomed. Inform. Insights* **2016**, *8*, BII-S31559. [CrossRef]
19. Alonso, S.G.; de la Torre Diez, I.; Rodrigues, J.J.P.C.; Hamrioui, S.; Lopez-Coronado, M. A systematic review of techniques and sources of big data in the healthcare sector. *J. Med. Syst.* **2017**, *41*, 183. [CrossRef] [PubMed]
20. Herland, M.; Khoshgoftaar, T.M.; Wald, R. A review of data mining using big data in health informatics. *J. Big Data* **2014**, *1*, 1–35. [CrossRef]
21. Andrew, C.; Heegaard, E.; Kirk, P.M.; Bässler, C.; Heilmann-Clausen, J.; Krisai-Greilhuber, I.; Kuyper, T.W.; Senn-Irlet, B.; Büntgen, U.; Diez, J. Big data integration: Pan-European fungal species observations' assembly for addressing contemporary questions in ecology and global change biology. *Fungal Biol. Rev.* **2017**, *31*, 88–98. [CrossRef]
22. Heart, T.; Ben-Assuli, O.; Shabtai, I. A review of PHR, EMR and EHR integration: A more personalized healthcare and public health policy. *Heal. Policy Technol.* **2017**, *6*, 20–25. [CrossRef]
23. Ritchie, M.D.; Holzinger, E.R.; Li, R.; Pendergrass, S.A.; Kim, D. Methods of integrating data to uncover genotype–phenotype interactions. *Nat. Rev. Genet.* **2015**, *16*, 85–97. [CrossRef]
24. Tomar, D.; Agarwal, S. A survey on Data Mining approaches for Healthcare. *Int. J. Bio-Sci. Bio-Technol.* **2013**, *5*, 241–266. [CrossRef]

25. Yoo, I.; Alafaireet, P.; Marinov, M.; Pena-Hernandez, K.; Gopidi, R.; Chang, J.-F.; Hua, L. Data mining in healthcare and biomedicine: A survey of the literature. *J. Med. Syst.* **2012**, *36*, 2431–2448. [CrossRef] [PubMed]
26. Shukla, D.P.; Patel, S.B.; Sen, A.K. A literature review in health informatics using data mining techniques. *Int. J. Softw. Hardw. Res. Eng.* **2014**, *2*, 123–129.
27. König, I.R.; Auerbach, J.; Gola, D.; Held, E.; Holzinger, E.R.; Legault, M.-A.; Sun, R.; Tintle, N.; Yang, H.-C. Machine learning and data mining in complex genomic data—A review on the lessons learned in Genetic Analysis Workshop 19. *BMC Genet.* **2016**, *17*, S1.
28. Miotto, R.; Wang, F.; Wang, S.; Jiang, X.; Dudley, J.T. Deep learning for healthcare: Review, opportunities and challenges. *Brief. Bioinform.* **2018**, *19*, 1236–1246. [CrossRef]
29. Min, S.; Lee, B.; Yoon, S. Deep learning in bioinformatics. *Brief. Bioinform.* **2017**, *18*, 851–869. [CrossRef]
30. Baldi, P. Deep learning in biomedical data science. *Annu. Rev. Biomed. Data Sci.* **2018**, *1*, 181–205. [CrossRef]
31. Belle, A.; Thiagarajan, R.; Soroushmehr, S.M.; Navidi, F.; Beard, D.A.; Najarian, K. Big data analytics in healthcare. *Biomed Res. Int.* **2015**, *2015*. [CrossRef]
32. Schadt, E.E.; Linderman, M.D.; Sorenson, J.; Lee, L.; Nolan, G.P. Computational solutions to large-scale data management and analysis. *Nat. Rev. Genet.* **2010**, *11*, 647–657. [CrossRef]
33. Hashem, I.A.T.; Yaqoob, I.; Anuar, N.B.; Mokhtar, S.; Gani, A.; Khan, S.U. The rise of "big data" on cloud computing: Review and open research issues. *Inf. Syst.* **2015**, *47*, 98–115. [CrossRef]
34. O'Driscoll, A.; Daugelaite, J.; Sleator, R.D. "Big data", Hadoop and cloud computing in genomics. *J. Biomed. Inform.* **2013**, *46*, 774–781. [CrossRef] [PubMed]
35. Dai, L.; Gao, X.; Guo, Y.; Xiao, J.; Zhang, Z. Bioinformatics clouds for big data manipulation. *Biol. Direct* **2012**, *7*, 43. [CrossRef] [PubMed]
36. Abouelmehdi, K.; Beni-Hssane, A.; Khaloufi, H.; Saadi, M. Big data security and privacy in healthcare: A Review. *Procedia Comput. Sci.* **2017**, *113*, 73–80. [CrossRef]
37. Hawkins, D.M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1–12. [CrossRef]
38. Xu, Q.-S.; Liang, Y.-Z. Monte Carlo cross validation. *Chemom. Intell. Lab. Syst.* **2001**, *56*, 1–11. [CrossRef]
39. Faber, N.M.; Rajko, R. How to avoid over-fitting in multivariate calibration—The conventional validation approach and an alternative. *Anal. Chim. Acta* **2007**, *595*, 98–106. [CrossRef] [PubMed]
40. Cook, R.R.P.R.D. Cross-Validation of Regression Models. *J. Am. Stat. Assoc.* **1984**, *79*, 575–583.
41. Shah, D.; Wang, J.; He, Q.P. A feature-based soft sensor for spectroscopic data analysis. *J. Process Control* **2019**, *78*, 98–107. [CrossRef]
42. Guzman, Y.A. Theoretical Advances in Robust Optimization, Feature Selection, and Biomarker Discovery. Ph.D. Thesis, Princeton University, Princeton, NJ, USA, 2016.
43. Mehmood, T.; Liland, K.H.; Snipen, L.; Saebo, S. A review of variable selection methods in Partial Least Squares Regression. *Chemom. Intell. Lab. Syst.* **2012**, *118*, 62–69. [CrossRef]
44. O'Hara, R.B.; Sillanpää, M.J. A review of Bayesian variable selection methods: What, how and which. *Bayesian Anal.* **2009**, *4*, 85–117. [CrossRef]
45. May, R.; Dandy, G.; Maier, H. Review of input variable selection methods for artificial neural networks. *Artif. Neural Netw. Methodol. Adv. Biomed. Appl.* **2011**, *10*, 16004.
46. Peres, F.A.P.; Fogliatto, F.S. Variable selection methods in multivariate statistical process control: A systematic literature review. *Comput. Ind. Eng.* **2018**, *115*, 603–619. [CrossRef]
47. Heinze, G.; Wallisch, C.; Dunkler, D. Variable selection—A review and recommendations for the practicing statistician. *Biom. J.* **2018**, *60*, 431–449. [CrossRef] [PubMed]
48. Tang, J.; Alelyani, S.; Liu, H. Feature selection for classification: A review. *Data Classif. Algorithms Appl.* **2014**, 37–64. [CrossRef]
49. Vergara, J.R.; Estévez, P.A. A review of feature selection methods based on mutual information. *Neural Comput. Appl.* **2014**, *24*, 175–186. [CrossRef]
50. Kumar, V.; Minz, S. Feature selection: A literature review. *SmartCR* **2014**, *4*, 211–229. [CrossRef]
51. Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [CrossRef]
52. Yang, R.; Daigle, B.J.; Petzold, L.R.; Doyle, F.J. Core module biomarker identification with network exploration for breast cancer metastasis. *BMC Bioinform.* **2012**, *13*, 12. [CrossRef]

53. Guzman, Y.A.; Sakellari, D.; Papadimitriou, K.; Floudas, C.A. High-throughput proteomic analysis of candidate biomarker changes in gingival crevicular fluid after treatment of chronic periodontitis. *J. Periodontal Res.* **2018**, *53*, 853–860. [CrossRef]
54. Dean, K.R.; Hammamieh, R.; Mellon, S.H.; Abu-Amara, D.; Flory, J.D.; Guffanti, G.; Wang, K.; Daigle, B.J.; Gautam, A.; Lee, I. Multi-omic biomarker identification and validation for diagnosing warzone-related post-traumatic stress disorder. *Mol. Psychiatry* **2019**, 1–13. [CrossRef]
55. Lee, S.; Lee, T.; Yang, T.; Yoon, C.; Kim, S.-P. Detection of Drivers' Anxiety Invoked by Driving Situations Using Multimodal Biosignals. *Processes* **2020**, *8*, 155. [CrossRef]
56. Oh, S.H.; Chang, Y.K.; Lee, J.H. Identification of significant proxy variable for the physiological status affecting salt stress-induced lipid accumulation in Chlorella sorokiniana HS1. *Biotechnol. Biofuels* **2019**, *12*, 242. [CrossRef]
57. Melo, J.C.B.; Cavalcanti, G.D.C.; Guimaraes, K.S. PCA feature extraction for protein structure prediction. In Proceedings of the International Joint Conference on Neural Networks, IEEE, Portland, OR, USA, 20–24 July 2003; Volume 4, pp. 2952–2957.
58. Taguchi, Y.H.; Murakami, Y. Principal component analysis based feature extraction approach to identify circulating microRNA biomarkers. *PLoS ONE* **2013**, *8*, e66714. [CrossRef] [PubMed]
59. Howsmon, D.P.; Vargason, T.; Rubin, R.A.; Delhey, L.; Tippett, M.; Rose, S.; Bennuri, S.C.; Slattery, J.C.; Melnyk, S.; James, S.J. Multivariate techniques enable a biochemical classification of children with autism spectrum disorder versus typically-developing peers: A comparison and validation study. *Bioeng. Transl. Med.* **2018**, *3*, 156–165. [CrossRef] [PubMed]
60. Adams, J.; Howsmon, D.P.; Kruger, U.; Geis, E.; Gehn, E.; Fimbres, V.; Pollard, E.; Mitchell, J.; Ingram, J.; Hellmers, R. Significant association of urinary toxic metals and autism-related symptoms—A nonlinear statistical analysis with cross validation. *PLoS ONE* **2017**, *12*, e0169526. [CrossRef]
61. Taguchi, Y.H.; Iwadate, M.; Umeyama, H. Principal component analysis-based unsupervised feature extraction applied to in silico drug discovery for posttraumatic stress disorder-mediated heart disease. *BMC Bioinform.* **2015**, *16*, 139. [CrossRef] [PubMed]
62. Sengur, A. An expert system based on principal component analysis, artificial immune system and fuzzy k-NN for diagnosis of valvular heart diseases. *Comput. Biol. Med.* **2008**, *38*, 329–338. [CrossRef]
63. Taguchi, Y. Principal component analysis-based unsupervised feature extraction applied to single-cell gene expression analysis. In Proceedings of the International Conference on Intelligent Computing, Bengaluru, India, 25–27 October 2018; pp. 816–826.
64. Li, K.; Zheng, J.; Deng, T.; Peng, J.; Daniel, D.; Jia, Q.; Huang, Z. An Analysis of Antimicrobial Resistance of Clinical Pathogens from Historical Samples for Six Countries. *Processes* **2019**, *7*, 964. [CrossRef]
65. Jin, Y.; Qin, S.J.; Huang, Q.; Saucedo, V.; Li, Z.; Meier, A.; Kundu, S.; Lehr, B.; Charaniya, S. Classification and Diagnosis of Bioprocess Cell Growth Productions Using Early-Stage Data. *Ind. Eng. Chem. Res.* **2019**, *58*, 13469–13480. [CrossRef]
66. Severson, K.A.; Monian, B.; Love, J.C.; Braatz, R.D. A method for learning a sparse classifier in the presence of missing data for high-dimensional biological datasets. *Bioinformatics* **2017**, *33*, 2897–2905. [CrossRef]
67. Hira, Z.M.; Gillies, D.F. A review of feature selection and feature extraction methods applied on microarray data. *Adv. Bioinform.* **2015**, *2015*. [CrossRef]
68. Azlan, W.A.W.; Low, Y.F. Feature extraction of electroencephalogram (EEG) signal-A review. In Proceedings of the 2014 IEEE Conference on Biomedical Engineering and Sciences (IECBES); IEEE, Miri, Malaysia, 8–10 December 2014; pp. 801–806.
69. Rathore, S.; Habes, M.; Iftikhar, M.A.; Shacklett, A.; Davatzikos, C. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *Neuroimage* **2017**, *155*, 530–548. [CrossRef]
70. Taguchi, Y.H.; Iwadate, M.; Umeyama, H.; Murakami, Y. Principal component analysis based unsupervised feature extraction applied to bioinformatics analysis. *Comput. Methods Appl. Bioinforma. Anal.* **2017**, *8*, 153–182.
71. Mahmoudi, Z.; Cameron, F.; Poulsen, N.K.; Madsen, H.; Bequette, B.W.; Jørgensen, J.B. Sensor-based detection and estimation of meal carbohydrates for people with diabetes. *Biomed. Signal Process. Control* **2019**, *48*, 12–25. [CrossRef]

72. Panagiotou, G.; Andersen, M.R.; Grotkjaer, T.; Regueira, T.B.; Nielsen, J.; Olsson, L. Studies of the production of fungal polyketides in Aspergillus nidulans by using systems biology tools. *Appl. Environ. Microbiol.* **2009**, *75*, 2212–2220. [CrossRef] [PubMed]
73. Grivas, G.; Vargason, T.; Hahn, J. Biomarker Identification of Complex Diseases/Disorders: Methodological Parallels to Parameter Estimation. *Ind. Eng. Chem. Res.* **2019**, *59*, 2366–2377. [CrossRef]
74. Somvanshi, P.R.; Mellon, S.H.; Flory, J.D.; Abu-Amara, D.; Consortium, P.S.B.; Wolkowitz, O.M.; Yehuda, R.; Jett, M.; Hood, L.; Marmar, C. Mechanistic inferences on metabolic dysfunction in posttraumatic stress disorder from an integrated model and multiomic analysis: Role of glucocorticoid receptor sensitivity. *Am. J. Physiol. Metab.* **2019**, *317*, E879–E898.
75. Bastin, G.; Dochain, D. *On-line Estimation and Adaptive Control of Bioreactors*; Elsevier: Amsterdam, The Netherlands, 2013; Volume 1, ISBN 1483290980.
76. Snowden, T.J.; van der Graaf, P.H.; Tindall, M.J. Methods of model reduction for large-scale biological systems: A survey of current methods and trends. *Bull. Math. Biol.* **2017**, *79*, 1449–1486. [CrossRef]
77. Girosi, F.; Jones, M.; Poggio, T. Regularization theory and neural networks architectures. *Neural Comput.* **1995**, *7*, 219–269. [CrossRef]
78. Qin, S.J. A statistical perspective of neural networks for process modeling and control. In Proceedings of the 8th IEEE International Symposium on Intelligent Control, IEEE, Chicago, IL, USA, 25–27 August 1993; pp. 599–604.
79. Chakrabarty, A.; Doyle, F.J.; Dassau, E. Deep learning assisted macronutrient estimation for feedforward-feedback control in artificial pancreas systems. In Proceedings of the 2018 Annual American Control Conference (ACC), IEEE, Milwaukee, WI, USA, 27–29 June 2018; pp. 3564–3570.
80. Vargason, T.; Howsmon, D.P.; Melnyk, S.; James, S.J.; Hahn, J. Mathematical modeling of the methionine cycle and transsulfuration pathway in individuals with autism spectrum disorder. *J. Theor. Biol.* **2017**, *416*, 28–37. [CrossRef] [PubMed]
81. Sun, M.; Min, T.; Zang, T.; Wang, Y. CDL4CDRP: A Collaborative Deep Learning Approach for Clinical Decision and Risk Prediction. *Processes* **2019**, *7*, 265. [CrossRef]
82. Howsmon, D.P.; Hahn, J. Regularization Techniques to Overcome Overparameterization of Complex Biochemical Reaction Networks. *IEEE Life Sci. Lett.* **2016**, *2*, 31–34. [CrossRef]
83. Raue, A.; Kreutz, C.; Maiwald, T.; Bachmann, J.; Schilling, M.; Klingmüller, U.; Timmer, J. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **2009**, *25*, 1923–1929. [CrossRef] [PubMed]
84. Maiwald, T.; Hass, H.; Steiert, B.; Vanlier, J.; Engesser, R.; Raue, A.; Kipkeew, F.; Bock, H.H.; Kaschek, D.; Kreutz, C. Driving the model to its limit: Profile likelihood based model reduction. *PLoS ONE* **2016**, *11*, e0162366. [CrossRef] [PubMed]
85. Rateitschak, K.; Winter, F.; Lange, F.; Jaster, R.; Wolkenhauer, O. Parameter identifiability and sensitivity analysis predict targets for enhancement of STAT1 activity in pancreatic cancer and stellate cells. *PLoS Comput. Biol.* **2012**, *8*, e1002815. [CrossRef]
86. Pohjanpalo, H. System identifiability based on the power series expansion of the solution. *Math. Biosci.* **1978**, *41*, 21–33. [CrossRef]
87. Lecourtier, Y.; Lamnabhi-Lagarrigue, F.; Walter, E. Volterra and generating power series approaches to identifiability testing. *Identifiability Parametr. Model.* **1987**, 50–66.
88. Vajda, S.; Godfrey, K.R.; Rabitz, H. Similarity transformation approach to identifiability analysis of nonlinear compartmental models. *Math. Biosci.* **1989**, *93*, 217–248. [CrossRef]
89. Ljung, L.; Glad, T. On global identifiability for arbitrary model parametrizations. *Automatica* **1994**, *30*, 265–276. [CrossRef]
90. Meeker, W.Q.; Escobar, L.A. Teaching about approximate confidence regions based on maximum likelihood estimation. *Am. Stat.* **1995**, *49*, 48–53.
91. Neale, M.C.; Miller, M.B. The use of likelihood-based confidence intervals in genetic models. *Behav. Genet.* **1997**, *27*, 113–120. [CrossRef]
92. Zi, Z. Sensitivity analysis approaches applied to systems biology models. *IET Syst. Biol.* **2011**, *5*, 336–346. [CrossRef] [PubMed]
93. Rabitz, H.; Kramer, M.; Dacol, D. Sensitivity analysis in chemical kinetics. *Annu. Rev. Phys. Chem.* **1983**, *34*, 419–461. [CrossRef]

94. Ingalls, B. Sensitivity analysis: From model parameters to system behaviour. *Essays Biochem.* **2008**, *45*, 177–194.
95. Lemley, J.; Bazrafkan, S.; Corcoran, P. Smart augmentation learning an optimal data augmentation strategy. *IEEE Access* **2017**, *5*, 5858–5869. [CrossRef]
96. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
97. Frid-Adar, M.; Klang, E.; Amitai, M.; Goldberger, J.; Greenspan, H. Synthetic data augmentation using GAN for improved liver lesion classification. In Proceedings of the 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), IEEE, Washington, DC, USA, 4–7 April 2018; pp. 289–293.
98. Niklas, J.; Schräder, E.; Sandig, V.; Noll, T.; Heinzle, E. Quantitative characterization of metabolism and metabolic shifts during growth of the new human cell line AGE1. HN using time resolved metabolic flux analysis. *Bioprocess Biosyst. Eng.* **2011**, *34*, 533–545. [CrossRef]
99. Antoniewicz, M.R. Methods and advances in metabolic flux analysis: A mini-review. *J. Ind. Microbiol. Biotechnol.* **2015**, *42*, 317–325. [CrossRef]
100. Vargason, T.; Kruger, U.; McGuinness, D.L.; Adams, J.B.; Geis, E.; Gehn, E.; Coleman, D.; Hahn, J. Investigating plasma amino acids for differentiating individuals with autism spectrum disorder and typically developing peers. *Res. Autism Spectr. Disord.* **2018**, *50*, 60–72. [CrossRef]
101. Doyle, F.J., III; Bequette, B.W.; Middleton, R.; Ogunnaike, B.; Paden, B.; Parker, R.S.; Vidyasagar, M. Control in biological systems. In *The Impact of Control Technology*; Samad, T., Annaswamy, A., Eds.; IEEE Control Systems Society: Piscataway, NJ, USA, 2011.
102. Doyle, F.J., III. Robust control in biology: From genes to cells to systems. *IFAC Proc. Vol.* **2008**, *41*, 3470–3479. [CrossRef]
103. Doyle, F.J., III. Control and Biology. *IEEE Control Syst. Mag.* **2016**, *30*, 8–10.
104. Csete, M.E.; Doyle, J.C. Reverse engineering of biological complexity. *Science (80-)* **2002**, *295*, 1664–1669. [CrossRef] [PubMed]
105. Kitano, H. Systems biology: A brief overview. *Science (80-)* **2002**, *295*, 1662–1664. [CrossRef] [PubMed]
106. Kitano, H. Computational systems biology. *Nature* **2002**, *420*, 206–210. [CrossRef] [PubMed]
107. Chuang, H.-Y.; Hofree, M.; Ideker, T. A decade of systems biology. *Annu. Rev. Cell Dev. Biol.* **2010**, *26*, 721–744. [CrossRef]
108. Assmus, H.E.; Herwig, R.; Cho, K.-H.; Wolkenhauer, O. Dynamics of biological systems: Role of systems biology in medical research. *Expert Rev. Mol. Diagn.* **2006**, *6*, 891–902. [CrossRef]
109. Hilliard, M.; Wang, J.; He, Q.P. Dynamic Transcriptomic Data Analysis by Integrating Data-driven and Model-guided Approaches. *IFAC-PapersOnLine* **2018**, *51*, 104–107. [CrossRef]
110. Hilliard, M.; He, Q.P.; Wang, J. Dynamic Transcriptomic Data Reveal Unexpected Regulatory Behavior of Scheffersomyces stipitis. *IFAC-PapersOnLine* **2019**, *52*, 538–543. [CrossRef]
111. Strimbu, K.; Tavel, J.A. What are biomarkers? *Curr. Opin. HIV AIDS* **2010**, *5*, 463. [CrossRef]
112. Iyer, V.R.; Eisen, M.B.; Ross, D.T.; Schuler, G.; Moore, T.; Lee, J.C.F.; Trent, J.M.; Staudt, L.M.; Hudson, J.; Boguski, M.S. The transcriptional program in the response of human fibroblasts to serum. *Science (80-)* **1999**, *283*, 83–87. [CrossRef]
113. Ideker, T.; Thorsson, V.; Ranish, J.A.; Christmas, R.; Buhler, J.; Eng, J.K.; Bumgarner, R.; Goodlett, D.R.; Aebersold, R.; Hood, L. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science (80-)* **2001**, *292*, 929–934. [CrossRef]
114. Kholodenko, B.N.; Kiyatkin, A.; Bruggeman, F.J.; Sontag, E.; Westerhoff, H.V.; Hoek, J.B. Untangling the wires: A strategy to trace functional interactions in signaling and gene networks. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12841–12846. [CrossRef] [PubMed]
115. Nicholson, J.K.; Holmes, E.; Lindon, J.C.; Wilson, I.D. The challenges of modeling mammalian biocomplexity. *Nat. Biotechnol.* **2004**, *22*, 1268–1274. [CrossRef] [PubMed]
116. Vasilakou, E.; Machado, D.; Theorell, A.; Rocha, I.; Nöh, K.; Oldiges, M.; Wahl, S.A. Current state and challenges for dynamic metabolic modeling. *Curr. Opin. Microbiol.* **2016**, *33*, 97–104. [CrossRef] [PubMed]
117. Hilliard, M.; Damiani, A.; He, Q.P.; Jeffries, T.; Wang, J. Elucidating redox balance shift in Scheffersomyces stipitis' fermentative metabolism using a modified genome-scale metabolic model. *Microb. Cell Fact.* **2018**, *17*, 140. [CrossRef] [PubMed]

118. McDowell, I.C.; Manandhar, D.; Vockley, C.M.; Schmid, A.K.; Reddy, T.E.; Engelhardt, B.E. Clustering gene expression time series data using an infinite Gaussian process mixture model. *PLoS Comput. Biol.* **2018**, *14*, e1005896. [CrossRef]
119. Cheng, C.; Fu, Y.; Shen, L.; Gerstein, M. Identification of yeast cell cycle regulated genes based on genomic features. *BMC Syst. Biol.* **2013**, *7*, 70. [CrossRef]
120. Bar-Joseph, Z.; Gitter, A.; Simon, I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.* **2012**, *13*, 552–564. [CrossRef]
121. Gasch, A.P.; Spellman, P.T.; Kao, C.M.; Carmel-Harel, O.; Eisen, M.B.; Storz, G.; Botstein, D.; Brown, P.O. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **2000**, *11*, 4241–4257. [CrossRef]
122. Storch, K.-F.; Lipan, O.; Leykin, I.; Viswanathan, N.; Davis, F.C.; Wong, W.H.; Weitz, C.J. Extensive and divergent circadian gene expression in liver and heart. *Nature* **2002**, *417*, 78–83. [CrossRef]
123. Whitfield, M.L.; Sherlock, G.; Saldanha, A.J.; Murray, J.I.; Ball, C.A.; Alexander, K.E.; Matese, J.C.; Perou, C.M.; Hurt, M.M.; Brown, P.O. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **2002**, *13*, 1977–2000. [CrossRef] [PubMed]
124. Vangulik, W.M.; Antoniewicz, M.R.; Delaat, W.; Vinke, J.L.; Heijnen, J.J. Energetics of growth and penicillin production in a high-producing strain of Penicillium chrysogenum. *Biotechnol. Bioeng.* **2001**, *72*, 185–193. [CrossRef]
125. Orth, J.D.; Thiele, I.; Palsson, B.Ø. What is flux balance analysis? *Nat. Biotechnol.* **2010**, *28*, 245. [CrossRef] [PubMed]
126. Antoniewicz, M.R. Dynamic metabolic flux analysis—Tools for probing transient states of metabolic networks. *Curr. Opin. Biotechnol.* **2013**, *24*, 973–978. [CrossRef]
127. Foster, C.J.; Gopalakrishnan, S.; Antoniewicz, M.R.; Maranas, C.D. From Escherichia coli mutant 13C labeling data to a core kinetic model: A kinetic model parameterization pipeline. *PLoS Comput. Biol.* **2019**, *15*, e1007319. [CrossRef] [PubMed]
128. Hendry, J.I.; Gopalakrishnan, S.; Ungerer, J.; Pakrasi, H.B.; Tang, Y.J.; Maranas, C.D. Genome-scale fluxome of Synechococcus elongatus UTEX 2973 using transient 13C-labeling data. *Plant Physiol.* **2019**, *179*, 761–769. [CrossRef]
129. Cheah, Y.E.; Young, J.D. Isotopically nonstationary metabolic flux analysis (INST-MFA): Putting theory into practice. *Curr. Opin. Biotechnol.* **2018**, *54*, 80–87. [CrossRef]
130. Young, J.D. INCA: A computational platform for isotopically non-stationary metabolic flux analysis. *Bioinformatics* **2014**, *30*, 1333–1335. [CrossRef]
131. Mahadevan, R.; Schilling, C.H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* **2003**, *5*, 264–276. [CrossRef]
132. Ahn, W.S.; Antoniewicz, M.R. Towards dynamic metabolic flux analysis in CHO cell cultures. *Biotechnol. J.* **2012**, *7*, 61–74. [CrossRef]
133. Lequeux, G.; Beauprez, J.; Maertens, J.; Van Horen, E.; Soetaert, W.; Vandamme, E.; Vanrolleghem, P.A. Dynamic metabolic flux analysis demonstrated on cultures where the limiting substrate is changed from carbon to nitrogen and vice versa. *Biomed Res. Int.* **2010**, *2010*. [CrossRef]
134. Llaneras, F.; Picó, J. A procedure for the estimation over time of metabolic fluxes in scenarios where measurements are uncertain and/or insufficient. *BMC Bioinform.* **2007**, *8*, 421. [CrossRef] [PubMed]
135. Antoniewicz, M.R.; Kraynie, D.F.; Laffend, L.A.; González-Lergier, J.; Kelleher, J.K.; Stephanopoulos, G. Metabolic flux analysis in a nonstationary system: Fed-batch fermentation of a high yielding strain of E. coli producing 1, 3-propanediol. *Metab. Eng.* **2007**, *9*, 277–292. [CrossRef] [PubMed]
136. Mahadevan, R.; Edwards, J.S.; Doyle, F.J., III. Dynamic flux balance analysis of diauxic growth in Escherichia coli. *Biophys. J.* **2002**, *83*, 1331–1340. [CrossRef]
137. Hanly, T.J.; Urello, M.; Henson, M.A. Dynamic flux balance modeling of S. cerevisiae and E. coli co-cultures for efficient consumption of glucose/xylose mixtures. *Appl. Microbiol. Biotechnol.* **2012**, *93*, 2529–2541. [CrossRef]
138. Gomez, J.A.; Höffner, K.; Barton, P.I. DFBAlab: A fast and reliable MATLAB code for dynamic flux balance analysis. *BMC Bioinform.* **2014**, *15*, 409. [CrossRef]
139. Zomorrodi, A.R.; Suthers, P.F.; Ranganathan, S.; Maranas, C.D. Mathematical optimization applications in metabolic networks. *Metab. Eng.* **2012**, *14*, 672–686. [CrossRef]

140. Aldridge, B.B.; Burke, J.M.; Lauffenburger, D.A.; Sorger, P.K. Physicochemical modelling of cell signalling pathways. *Nat. Cell Biol.* **2006**, *8*, 1195–1203. [CrossRef]
141. Janes, K.A.; Yaffe, M.B. Data-driven modelling of signal-transduction networks. *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 820–828. [CrossRef]
142. Huang, Z. *A Systems Biology Approach to Develop Models of Signal Transduction Pathways*; Texas A&M University: College Station, TX, USA, 2010; ISBN 1124381996.
143. Hunter, T. Signaling—2000 and beyond. *Cell* **2000**, *100*, 113–127. [CrossRef]
144. Pawson, T. Specificity in signal transduction: From phosphotyrosine-SH2 domain interactions to complex cellular systems. *Cell* **2004**, *116*, 191–203. [CrossRef]
145. Korobkova, E.; Emonet, T.; Vilar, J.M.G.; Shimizu, T.S.; Cluzel, P. From molecular noise to behavioural variability in a single bacterium. *Nature* **2004**, *428*, 574–578. [CrossRef]
146. Rao, C.V.; Kirby, J.R.; Arkin, A.P. Design and diversity in bacterial chemotaxis: A comparative study in Escherichia coli and Bacillus subtilis. *PLoS Biol.* **2004**, *2*, e49. [CrossRef]
147. Stelling, J.; Sauer, U.; Szallasi, Z.; Doyle, F.J., III; Doyle, J. Robustness of cellular functions. *Cell* **2004**, *118*, 675–685. [CrossRef] [PubMed]
148. Huang, C.-Y.; Ferrell, J.E. Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 10078–10083. [CrossRef] [PubMed]
149. Sontag, E.D. Asymptotic amplitudes and Cauchy gains: A small-gain principle and an application to inhibitory biological feedback. *Syst. Control Lett.* **2002**, *47*, 167–179. [CrossRef]
150. Sourjik, V.; Berg, H.C. Functional interactions between receptors in bacterial chemotaxis. *Nature* **2004**, *428*, 437–441. [CrossRef]
151. Cluzel, P.; Surette, M.; Leibler, S. An ultrasensitive bacterial motor revealed by monitoring signaling proteins in single cells. *Science (80-)* **2000**, *287*, 1652–1655. [CrossRef]
152. Almogy, G.; Stone, L.; Ben-Tal, N. Multi-stage regulation, a key to reliable adaptive biochemical pathways. *Biophys. J.* **2001**, *81*, 3016–3028. [CrossRef]
153. Gadkar, K.G.; Varner, J.; Doyle, F.J., III. Model identification of signal transduction networks from data using a state regulator problem. *Syst. Biol. (Stevenage)* **2005**, *2*, 17–30. [CrossRef]
154. Gadkar, K.G.; Gunawan, R.; Doyle, F.J. Iterative approach to model identification of biological networks. *BMC Bioinform.* **2005**, *6*, 155. [CrossRef] [PubMed]
155. Chen, R.R.; Mias, G.I.; Li-Pook-Than, J.; Jiang, L.; Lam, H.Y.K.; Chen, R.R.; Miriami, E.; Karczewski, K.J.; Hariharan, M.; Dewey, F.E. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **2012**, *148*, 1293–1307. [CrossRef] [PubMed]
156. Mias, G.I.; Yusufaly, T.; Roushangar, R.; Brooks, L.R.K.; Singh, V.V.; Christou, C. MathIOmica: An integrative platform for dynamic omics. *Sci. Rep.* **2016**, *6*, 37237. [CrossRef] [PubMed]
157. Nakanishi, Y.; Fukuda, S.; Chikayama, E.; Kimura, Y.; Ohno, H.; Kikuchi, J. Dynamic omics approach identifies nutrition-mediated microbial interactions. *J. Proteome Res.* **2011**, *10*, 824–836. [CrossRef] [PubMed]
158. Przytycka, T.M.; Singh, M.; Slonim, D.K. Toward the dynamic interactome: It's about time. *Brief. Bioinform.* **2010**, *11*, 15–29. [CrossRef]
159. Zeger, S.L.; Irizarry, R.; Peng, R.D. On time series analysis of public health and biomedical data. *Annu. Rev. Public Heal.* **2006**, *27*, 57–79. [CrossRef]
160. El-Samad, H.; Prajna, S.; Papachristodoulou, A.; Doyle, J.; Khammash, M. Advanced methods and algorithms for biological networks analysis. *Proc. IEEE* **2006**, *94*, 832–853. [CrossRef]
161. El-Samad, H.; Kurata, H.; Doyle, J.C.; Gross, C.A.; Khammash, M. Surviving heat shock: Control strategies for robustness and performance. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 2736–2741. [CrossRef]
162. Hughes, M.E.; Abruzzi, K.C.; Allada, R.; Anafi, R.; Arpat, A.B.; Asher, G.; Baldi, P.; De Bekker, C.; Bell-Pedersen, D.; Blau, J. Guidelines for genome-scale analysis of biological rhythms. *J. Biol. Rhythms* **2017**, *32*, 380–393. [CrossRef]
163. Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A. Mastering the game of go without human knowledge. *Nature* **2017**, *550*, 354–359. [CrossRef]
164. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484. [CrossRef] [PubMed]

165. Anderson, C. The end of theory: The data deluge makes the scientific method obsolete. *Wired Mag.* **2008**, *16*, 7–16. [CrossRef]
166. Coveney, P.V.; Dougherty, E.R.; Highfield, R.R. Big data need big theory too. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2016**, *374*, 20160153. [CrossRef] [PubMed]
167. Succi, S.; Coveney, P.V. Big data: The end of the scientific method? *Philos. Trans. R. Soc. A* **2019**, *377*, 20180145. [CrossRef] [PubMed]
168. Silver, N. *The Signal and the Noise: Why so Many Predictions Fail–but Some Don't*; Penguin: London, UK, 2012, ISBN 159420411X.
169. Sánchez, B.J.; Nielsen, J. Genome scale models of yeast: Towards standardized evaluation and consistent omic integration. *Integr. Biol.* **2015**, *7*, 846–858. [CrossRef]
170. Damiani, A.L.; He, Q.P.; Jeffries, T.W.; Wang, J. Comprehensive evaluation of two genome-scale metabolic network models for Scheffersomyces stipitis. *Biotechnol. Bioeng.* **2015**, *112*, 1250–1262. [CrossRef]
171. Wang, J.; He, Q.P.; Damiani, A.; He, Q.P.; Wang, J. A System Identification Based Framework for Genome-Scale Metabolic Model Validation and Refinement. In Proceedings of the Foundations of Systems Biology in Engineering, Boston, MA, USA, 9–12 August 2015; 2017; pp. 13013–13018.
172. Eisen, M.B.; Spellman, P.T.; Brown, P.O.; Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 14863–14868. [CrossRef] [PubMed]
173. Herwig, R.; Poustka, A.J.; Müller, C.; Bull, C.; Lehrach, H.; O'Brien, J. Large-scale clustering of cDNA-fingerprinting data. *Genome Res.* **1999**, *9*, 1093–1105. [CrossRef]
174. Fang, Z.; Yang, J.; Li, Y.; Luo, Q.; Liu, L. Knowledge guided analysis of microarray data. *J. Biomed. Inform.* **2006**, *39*, 401–411. [CrossRef]
175. Parraga-Alava, J.; Dorn, M.; Inostroza-Ponta, M. A multi-objective gene clustering algorithm guided by apriori biological knowledge with intensification and diversification strategies. *BioData Min.* **2018**, *11*, 16. [CrossRef]
176. Yang, Y.; Wu, Z.; Kong, W. Improving clustering of microrna microarray data by incorporating functional similarity. *Curr. Bioinform.* **2018**, *13*, 34–41. [CrossRef]
177. Schwaber, J.S.; Doyle, F.J.; Zak, D.E. Controlled Biological Processes and Computational Genomics. In *Proceedings of the Chemical Process Control VI*; American Institute of Chemical Engineers: New York, NY, USA, 2001; pp. 75–80.
178. Purdom, E.; Holmes, S.P. Error distribution for gene expression data. *Stat. Appl. Genet. Mol. Biol.* **2005**, *4*. [CrossRef]
179. Scholz, M.; Gatzek, S.; Sterling, A.; Fiehn, O.; Selbig, J. Metabolite fingerprinting: Detecting biological features by independent component analysis. *Bioinformatics* **2004**, *20*, 2447–2454. [CrossRef]
180. Yao, F.; Coquery, J.; Lê Cao, K.-A. Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. *BMC Bioinform.* **2012**, *13*, 24. [CrossRef] [PubMed]
181. Wartner, S.; Girardi, D.; Wiesinger-Widi, M.; Trenkler, J.; Kleiser, R.; Holzinger, A. Ontology-guided principal component analysis: Reaching the limits of the doctor-in-the-loop. In Proceedings of the International Conference on Information Technology in Bio-and Medical Informatics, Porto, Portugal, 5–8 September 2016; pp. 22–33.
182. Wang, C.; Xuan, J.; Li, H.; Wang, Y.; Zhan, M.; Hoffman, E.P.; Clarke, R. Knowledge-guided gene ranking by coordinative component analysis. *BMC Bioinform.* **2010**, *11*, 162. [CrossRef] [PubMed]
183. Wentzell, P.D.; Andrews, D.T.; Hamilton, D.C.; Faber, K.; Kowalski, B.R. Maximum likelihood principal component analysis. *J. Chemom. A J. Chemom. Soc.* **1997**, *11*, 339–366. [CrossRef]
184. Choi, S.W.; Martin, E.B.; Morris, A.J.; Lee, I.-B. Fault detection based on a maximum-likelihood principal component analysis (PCA) mixture. *Ind. Eng. Chem. Res.* **2005**, *44*, 2316–2327. [CrossRef]
185. Theobald, D.L.; Wuttke, D.S. Accurate structural correlations from maximum likelihood superpositions. *PLoS Comput. Biol.* **2008**, *4*, e43. [CrossRef]
186. Mailier, J.; Remy, M.; Wouwer, A. Vande Stoichiometric identification with maximum likelihood principal component analysis. *J. Math. Biol.* **2013**, *67*, 739–765. [CrossRef]
187. Zhao, Y.; Chang, C.; Long, Q. Knowledge-guided statistical learning methods for analysis of high-dimensional-omics data in precision oncology. *JCO Precis. Oncol.* **2019**, *3*, 1–9. [CrossRef]

188. McDermott, J.E.; Wang, J.; Mitchell, H.; Webb-Robertson, B.-J.; Hafen, R.; Ramey, J.; Rodland, K.D. Challenges in biomarker discovery: Combining expert insights with statistical analysis of complex omics data. *Expert Opin. Med. Diagn.* **2013**, *7*, 37–51. [CrossRef] [PubMed]
189. Lee, J.; He, Q.P. Understanding the effect of specialization on hospital performance through knowledge-guided machine learning. *Comput. Chem. Eng.* **2019**, *125*, 490–498. [CrossRef]
190. Shen, L.; Lin, Y.; Sun, Z.; Yuan, X.; Chen, L.; Shen, B. Knowledge-guided bioinformatics model for identifying autism spectrum disorder diagnostic MicroRNA biomarkers. *Sci. Rep.* **2016**, *6*, 39663. [CrossRef] [PubMed]
191. Hvidsten, T.R.; Komorowski, J.; Sandvik, A.K.; Lægreid, A. Predicting gene function from gene expressions and ontologies. In *Biocomputing 2001*; World Scientific: Singapore, 2000; pp. 299–310.
192. Park, S.H.; Gao, Y.; Shi, Y.; Shen, D. Interactive prostate segmentation using atlas-guided semi-supervised learning and adaptive feature selection. *Med. Phys.* **2014**, *41*, 111715. [CrossRef] [PubMed]
193. Libbrecht, M.W.; Noble, W.S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **2015**, *16*, 321–332. [CrossRef] [PubMed]
194. Li, Y.; Wu, F.-X.; Ngom, A. A review on machine learning principles for multi-view biological data integration. *Brief. Bioinform.* **2018**, *19*, 325–340. [CrossRef]
195. Yadav, P.; Steinbach, M.; Kumar, V.; Simon, G. Mining Electronic Health Records (EHRs) A Survey. *ACM Comput. Surv.* **2018**, *50*, 1–40. [CrossRef]
196. Lee, J.; Flores-Cerrillo, J.; Wang, J.; He, Q.P. Consistency-Enhanced Evolution for Variable Selection Can Identify Key Chemical Information from Spectroscopic Data. *Ind. Eng. Chem. Res.* **2020**, *59*, 3446–3457. [CrossRef]
197. de Souza Alves, T.; de Oliveira, C.S.; Sanin, C.; Szczerbicki, E. From knowledge based vision systems to cognitive vision systems: A review. *Procedia Comput. Sci.* **2018**, *126*, 1855–1864. [CrossRef]
198. Li, A.; Li, C.; Wang, X.; Eberl, S.; Feng, D.D.D.; Fulham, M. Automated segmentation of prostate MR images using prior knowledge enhanced random walker. In Proceedings of the 2013 International Conference on Digital Image Computing: Techniques and Applications (DICTA); IEEE, Hobart, Australia, 26–28 November 2013; pp. 1–7.
199. de Andrade, M.L.S.C.L.S.C.; Skeika, E.; Aires, S.B.K.B.K. Segmentation of the Prostate Gland in Images Using Prior Knowledge and Level Set Method. In Proceedings of the 2017 Workshop of Computer Vision (WVC), IEEE, Rio Grande do Norte, Brazil, 30 October–1 November 2017; pp. 31–36.
200. Manjunath, K.N.N.; Prabhu, K.G.G.; Siddalingaswamy, P.C.C. A knowledge based approach for colon segmentation in CT colonography images. In Proceedings of the 2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), IEEE, Pullman, DC, USA, 19–21 October 2015; pp. 65–70.
201. Garla, V.N.; Brandt, C. Ontology-guided feature engineering for clinical text classification. *J. Biomed. Inform.* **2012**, *45*, 992–998. [CrossRef]
202. Yao, L.; Mao, C.; Luo, Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 71. [CrossRef] [PubMed]
203. Rodger, J.A. Discovery of medical Big Data analytics: Improving the prediction of traumatic brain injury survival rates by data mining Patient Informatics Processing Software Hybrid Hadoop Hive. *Inform. Med. Unlocked* **2015**, *1*, 17–26. [CrossRef]
204. Hand, D.J. Evaluating diagnostic tests: The area under the ROC curve and the balance of errors. *Stat. Med.* **2010**, *29*, 1502–1510. [CrossRef] [PubMed]